Речници у дигиталном добу - информатичка подршка за српски језик

Биљана Рујевић

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЛОШКИ ФАКУЛТЕТ

Биљана Ђ. Рујевић

РЕЧНИЦИ У ДИГИТАЛНОМ ДОБУ – ИНФОРМАТИЧКА ПОДРШКА ЗА СРПСКИ ЈЕЗИК

Докторска дисертација

Београд, 2022.

UNIVERSITY OF BELGRADE
FACULTY OF PHILOLOGY

Biljana Đ. Rujević

DICTIONARIES IN DIGITAL AGE – INFORMATION TECHNOLOGY SUPPORT FOR SERBIAN LANGUAGE

Doctoral Dissertation

Belgrade, 2022

УНИВЕРСИТЕТ В БЕЛГРАДЕ
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

Биляна Д. Рујевич

СЛОВАРИ В ЦИФРОВОМ ВОЗРАСТЕ - ИНФОРМАЦИОННАЯ ПОДДЕРЖКА ДЛЯ СЕРБСКИЙ ЯЗЫК

Докторская диссертация

Белград, 2022.

Ментор:

Проф. др Цветана Крстев, редовни професор

Универзитет у Београду, Филолошки факултет

Чланови комисије:

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Датум одбране:

_____

Неизмерну захвалност изражавам менторки, проф. др Цветани Крстев, као и проф. др Ранки Станковић на подстицајима без којих ова дисертација не би била завршена. Велико хвала и на свим идејама, стрпљењу и подршци.

Посебну захвалност дугујем др Милошу Утвићу, доценту Филолошког факултета Универзитета у Београду, др Јовану Чудомировићу, доценту Филолошког факултета Универзитета у Београду, и др Александри Марковић, научном сараднику Института за српски језик САНУ, на пажљивом читању текста и корисним сугестијама.

Захваљујем се колеги Михаилу Шкорићу уз чију стручну и техничку помоћ је апликација Лексимирка настала и развија се.

На волонтерској евалуацији терминологије из области геологије, захваљујем се колегиницама са Рударско-геолошког факултета – проф. др Јелени Миливојевић, Милици Пешић Георгиадис и др Алени Здравковић.

Речници у дигиталном добу – информатичка подршка за српски језик

Сажетак

Морфолошки речници српског језика представљају електронски језички ресурс који има значајну историју развоја и коришћења за потребе обраде природних језика. С обзиром на то да су чувани у облику датотека чији је број нарастао па је самим тим управљање речницима постало отежано јавила се потреба за смештањем информација из речника у облик лексикографске базе. Како би се омогућио симултани рад на развоју речника за више корисника јавила се потреба за веб-апликацијом заснованој на лексикографској бази.

Како би се размотриле функционалности које пружају речници у дигиталном окружењу у циљу проналаска најбољег решења за развој апликације, дескриптивном методом су анализирани различити примери дигиталних речника неколико језика.

Са циљем одабира адекватног модела за развој лексикографске базе разматрана су три стандардизована модела за представљање информација из речника: *TEI*, *LMF* и *lemon*. Модел развијене лексикографске базе се заснива на комбинацији модела *LMF* и *lemon*. Током разматрања и развоја модела лексикографске базе коришћене су дескриптивна и информатичка научна метода. Употреба лексикографске базе је омогућила напредну претрагу као и успостављање релација између лексичких записа. Успостављање релација се заснива на дефинисању група правила које лексички записи за повезивање треба да задовоље.

Захваљујући употреби лексикографске базе и апликације за преглед и у управљање речницима појавила се могућност надградње *Морфолошких речника за српски језик* као ресурса. Лексички записи су допуњени везама са екстерним лексичким ресурсима као што су *Ворднет*, *Терми*, *BabelNet*, *Glosbe* и *Wikidata*. Осим тога, омогућено је повезивање са записима из дигитализованих традиционалних речника српског језика које би могло бити доступно групама корисника који имају право на њихово коришћење у дигиталном облику.

Лексички записи су применом регуларних израза и коначних аутомата повезани са корпусима у виду могућности претраге конкорданци које садрже лему записа или предефинисане обрасце појављивања речи. Записима су придружене и информације о фреквенцији појављивања лема и облика речи у одређеним корпусима.

Развијене апликација и база су тестиране на речницима ексцерпираним из корпуса из геолошког домена *ГеоСрпКор* развијеном за потребе овог истраживања.

Кључне речи: електронски речници, лексикографска база података, лексички ресурси, српски језик

Научна област: Филолошке науке, Библиотекарство и информатика

Ужа научна област: Библиотечка информатика, Обрада природних језика, Електронска лексикографија

УДК број: 811.163.41’322.2:004.22(043.3)

Dictionaries in Digital Age – Information Technology Support for Serbian Language

Abstract

Serbian morphological dictionaries represent an electronic language resource with significant history of development and use in natural language processing. Since they were kept in form of files whose number grew, and thus the management of dictionaries became more difficult, it was necessary to store information from the dictionary in the form of a lexicographic database. In order to enable the dictionary development based on simultaneous work for several users, a web application based on a lexicographic database was needed.

In order to consider the functionalities provided by dictionaries in the digital environment towards finding the best solution for application development, various examples of digital dictionaries of several languages were analyzed using the descriptive method.

To establish an adequate model for development of the lexicographic database, three standardized models for presenting information from the dictionary were considered: *TEI*, *LMF* and *lemon*. The model of the developed lexicographic database is based on a combination of the *LMF* and *lemon* models. During the process of the lexicographic database model development, descriptive and informatics scientific methods were used. The use of lexicographic base enabled advanced search as well as the establishment of relations between lexical entries. Establishing lexical relations is based on the set of rules that define which criteria the lexical entries should meet.

The upgrade of Serbian morphological dictionaries came as a result of using the lexicographic database and the application for browsing and managing dictionaries Lexical entries are enriched by links to external lexical resources, some of which are: *Wordnet*, *Termi*, *BabelNet*, *Glosbe* and *Wikidata*. It is also possible to set up the connection with lexical entries from digitized printed dictionaries of the Serbian language. This could be available to groups of users who have access to these dictionaries in digital form.

Lexical entries are linked with corpora using regular expressions and finite automata. There is a possibility of searching for concordances that contain a lemma of lexical entries or predefined patterns of word occurrence. The lexical entries are extended by information on the frequency of occurrence of lemmas and word forms in certain corpora.

The developed application and database were tested on dictionaries excerpted from the corpus from the geological domain - *GeoSrpKor* that was developed for the purpose of this research.

Key words: electronic dictionaries, lexicographic data-base, lexical resources, Serbian language

Scientific field: Philological Sciences, Library and Information Science

Scientific subfield: Library informatics, Natural Language Processing, Electronic Lexicography

UDC Number: 811.163.41’322.2:004.22(043.3)

Речници у дигиталном добу – информатичка подршка за српски језик

1. Увод ... 1

2. Поглед на традиционалне и електронске речнике ... 3

2.1 Од традиционалних ка дигиталним речницима ... 3

2.2 Основа за развој дигиталних речника српског језика ... 4

2.3 Врсте речника ... 7

2.3.1 Традиционални речници ... 7

2.3.2 Речници у дигиталном облику ... 8

Дигитализовани речници ... 8

Речници настали у дигиталном окружењу ... 9

Методе електронске лексикографије ... 14

Преглед угледних речника доступних у дигиталном окружењу ... 15

2.3.3 Информатичка подршка изради речника ... 25

Лексикографска база ... 25

Системи за писање речника ... 26

3. Језички ресурси за српски језик ... 30

3.1 Морфолошки речници ... 30

3.2 Ворднет ... 33

3.3 Корпуси ... 34

4. Модели лексикографских података и развој модела лексикографске базе *Лексимирка* ... 37

4.1 Стандардизовани модели за постављање лексикографске базе ... 37

4.1.1 Иницијатива за обележавање текста - TEI ... 37

4.1.2 Оквир за лексичко обележавање - LMF ... 44

Пакет за опис морфолошких информација ... 47

Пакет проширења за машински читљиве речнике ... 49

Проширење за морфолошке обрасце за потребе обраде природних језика ... 51

Пакет проширења за синтаксу у обради природних језика ... 55

Пакет проширења за семантику у обради природних језика ... 57

Пакет за вишечлане лексичке изразе у обради природних језика ... 60

Пакет проширења за обележавање вишејезичности ... 63

4.1.3 Lemon ... 64

Синтакса и семантика (synsem) ... 72

Модул декомпозиција (decomp) ... 74

Модул варијације и превод - vartrans ... 76

Модул лингвистичких метаподатака - *lime* ... 78

4.1.4 Поређење представљених модела ... 81

4.2 Одабир адекватног модела лексикографске базе и њен развој ... 82

4.2.1 Модел развијене лексикографске базе ... 82

Примена лексикографске базе на француски језик ... 88

4.2.2 Доменске онтологије ... 90

Мапирање категорија података из *Морфолошких речника српског језика* ... 93

Мапирање значења једнојезичних речника ... 95

5. Формирање лексикографске базе српског језика у *Лексимирки* и успостављање унутрашњих и спољашњих веза ... 98

5.1 Формирање лексикографске базе ... 98

5.2 Развој модела повезивања речи у лексикографској бази и креирање механизама за полуаутоматско хармонизовање и усклађивање ... 99

5.2.1 Повезивање одредница у е-речнику – врсте веза ... 99

5.2.2 Повезивање са *Ворднетом* ... 104

5.2.3 Повезивање са корпусом 107
Приказ конкорданци 107
Фреквенције речи у различитим доменима 110
5.2.4 Профил речи 113
5.2.5 Повезивање са екстерним ресурсима 113
Повезане базе података 114
Ресурси на вебу 115

6. Постављање презентације на веб 118

6.1 Опис корисничког сучеља 118

6.2 Претраживање лексикографске базе 119

6.3 Функционалности 122
6.3.1 Извоз речника у формату према потреби корисника 122
6.3.2 Веб-сервиси засновани на речнику 122
6.3.3 Примери имплементације веб-сервиса 123

7. Евалуација резултата на лексици (терминологији) из геолошког домена 127

7.1 Екстракција терминологије 127

7.2 Обогаћивање речника маркерима 134

7.3 Обогаћивање речника везама 134

7.4 Обрада текстова из домена рударства и геологије са применом и без примене развијених речника 135

8. Закључак 139

8.1 Допринос 139

8.2 Правци за даљи рад 139

9. Литература 141

Прилог 1. Поређење категорија података из Морфолошких речника српског језика у односу на скраћенице из Речника САНУ и концепте из онтологије SUMO i

Прилог 2. Упутство за коришћење апликације *Лексимирка* vi

Прилог 3. Списак успостављених релација са бројем правила, бројем успостављених веза и примерима xxxii

Списак слика xli

Списак табела xlii

Биографија аутора xliii

1. Увод

Предмет истраживања ове докторске дисертације јесу информатичка подршка и језички ресурси за српски језик у процесу израде, одржавања и презентације електронских речника српског језика у дигиталном окружењу намењених људима и рачунарским апликацијама.

Крајњи циљ дисертације јесте добијање двојако функционалног речника српског језика доступног на вебу. За те потребе је развијена веб-апликација заснована на лексикографској бази података.[1] Апликација ће служити за одржавање и проширење садашњег електронског морфолошког речника српског језика и управљање њим, али ће је користити и широк круг корисника којима ће пружати могућност претраге речника према различитим критеријумима. Како би се постигао овај циљ, у раду ће бити изложени поступци трансформације постојећег електронског морфолошког речника српског језика из неструктурираних датотека у облик лексикографске базе.

Како бисмо се упознали са речницима у дигиталном окружењу и испитали шта он треба да садржи, у другом поглављу је дат кратак историјски преглед развоја лексикографије уопште и лексикографије у Србији, као и анализа садржаја изабраних дигиталних речника. Потом су описани алати који се користе као информатичка подршка за развој дигиталних речника.

У трећем поглављу је дат опис језичких ресурса за српски језик коришћених као основа за развој лексикографске базе и апликације. Посебно су описани *Морфолошки речници за српски језик* који су примарна тема дисертације.

Како би се задовољила потреба за стандардизацијом у управљању лексикографским информацијама похрањеним у лексикографској бази, у четвртом поглављу су описана и размотрена три модела за представљање лексикографских информација: *Иницијатива за обележавање текста – TEI, Оквир за лексичко обележавање – LMF и lemon.* Потом је представљен модел лексикографске базе *Лексимирка*, развијен током рада на овој дисертацији. Приказан је и пример примене исте лексикографске базе за потребе подршке електронском речнику француског језика, чиме се показује да је њена примена језички независна. Посебна пажња је посвећена употреби информатичких онтологија које представљају однос концепата унутар неког домена употребе. Приказан је значај мапирања категорија података из *Морфолошких речника за српски језик* са подацима из онтологија и категоријама података из других речника.

У петом поглављу је дат опис извршене трансформације података из *Морфолошких речника за српски језик* у новонасталу лексикографску базу. Посебно је описан нови модел за повезивање одредница у оквиру *Морфолошких речника*, као и повезивање са подацима из Ворднета и корпусним информацијама. Кроз ово поглавље је такође дат опис повезивања лексичких записа са лексичким ресурсима похрањеним у локалној бази података, као и са електронским језичким ресурсима на вебу.

У шестом поглављу је приказан кратак опис корисничког сучеља развијене апликације *Лексимирка* које се користи за управљање речницима, али и за представљање лексичких записа. Детаљно упутство за употребу апликације се налази у Прилогу 2 ове дисертације. Кроз шесто поглавље су описани и начини претраге развијене лексикографске базе података, као и неке значајније функције које омогућава

[1] У раду ће бити коришћен термин *лексикографска база* јер садржи податке намењене изради речника, али би био исправан и термин *лексичка база* јер база садржи податке који су у речницима.

представљање података из речника у виду лексикографске базе, на пример, извоз речника у различитим форматима, веб-сервиси који омогућавају да речнике користе друге апликације.

Евалуација резултата је спроведена на корпусу из области геологије, развијеном за потребе истраживања ове дисертације. На основу овог корпуса *Морфолошки речници* су допуњени новим лексичким записима али и новим маркерима и релацијама. Претходно је кроз сегмент апликације *Лексимирка* намењен евалуацији извршена евалуација аутоматски екстраховане терминологије из поменутог геолошког корпуса.

Очекивани резултати истраживања ове докторске дисертације јесу:

- модел лексикографске базе података,
- лексикографска база података за потребе складиштења информација из *Морфолошког речника за српски језик*,
- веб-апликација за потребе администрације, управљања и проширења садашњег електронског *Морфолошког речника* и презентацију лексичких записа ширем кругу корисника.[2]

Очекује се да ће апликација за управљање речником на вебу допринети удобнијем раду на процесима доградње и управљања *Морфолошким речницима српског језика*, као и налажењу нове примене овог ресурса, било у погледу новог круга корисника, било за развој апликација са компонентама обраде природних језика.

[2] У поглављу 9 се налази библиографија коришћених и цитираних радова уређена према *Приручнику Чикаго стила* – 17. издање (енг. *The Chicago Manual of Style, 17th. ed*).

2. Поглед на традиционалне и електронске речнике

2.1 Од традиционалних ка дигиталним речницима

Историја практичне лексикографије, односно лексикографије која се бави практичним пословима на изради речника, сеже у далеку прошлост, за разлику од теоријске лексикографије која је млађа дисциплина, делом због тога што се бави вредновањем производа практичне лексикографије.

Најстарији познати речници потичу из око 2300. године пре нове ере. Њихова постојбина је град Елба у Месопотамији. Били су представљени у форми глинених таблица и бавили су се преводима сумерских речи на акадски језик. Сматра се да је први тематски тезаурус настао у старом Египту 1750. године пре нове ере, а да су први глосари настали у Индији у II и III веку пре нове ере. Они су разјашњавали теже разумљиве изразе из индијских веда. Први кинески речник потиче из III века пре нове ере и садржи 3.500 речи пореклом из старих текстова подељених у 19 поглавља. Грчки речник из I века пре нове ере представља Хомеров лексикон и верује се да га је написао граматичар Аполон (Haslam 1994). Први речник персијске цивилизације који није сачуван настао је између IX и X века нове ере, док из XV века потиче најстарији персијско-турски речник (Драгићевић 2014). Развој европске лексикографије везује се за постанак националних књижевних језика. Први двојезични речници настају средином XV века за потребе учења латинског језика. Јужни Словени развијају лексикографију у XV веку на подручју Хрватске. Од почетка XVII века оснивају се националне академије које се баве састављањем речника. У овом покрету предњаче Италија (1612), Француска (1694), Шпанија (1726-1739) и Русија (1789-1794), док лексикографски рад у европским земљама доживљава процват у XIX веку.

Почетак дигиталних речника своје наговештаје даје шездесетих година XX века. Тада су Олни (Olney) и Ревард (Revard) *Вебстеров речник* (енг. *Webster's Seventh New Collegiate Dictionary*) са штампане папирне верзије техником прекуцавања перфорирањем пренели на бушене картице. Циљ овог подухвата била је рачунарска анализа речника (De Schryver 2003). Рачунари су у почетној фази коришћени у процесу стварања речника само као терминали за унос података. Тако је коришћење магнетних трака за производњу штампаних издања речника водило ка настанку машински читљивих речника. Ипак, прекретница која је подстакла настанак електронских речника јесте појава језичких корпуса. Ладислав Згуста у предговору књиге *Приручник лексикографије*, написаном 1968. године, наводи да је за подухват машинског превођења од прворазредног значаја међусобна зависност лексикона и граматике (Zgusta 1991). Дакле, Згуста још крајем 60-их година XX века препознаје и осмишљава појам електронског речника и могућности његове примене.

Први машински читљив речник претворен је у лексикографску базу 1978. године. Ради се о *Лонгмановом речнику савременог енглеског језика* (енг. *Longman Dictionary of Contemporary English, LDOCE*) (Granger и Paquot 2012). Наредна битна тачка у развоју електронских речника била је 1987. година када је објављен *Collins COBUILD речник енглеског језика* (енг. *Collins COBUILD English Language Dictionary*). Револуционарност овог речника огледала се у томе што је био заснован на анализи електронског корпуса савременог енглеског језика.

Сасвим природно и у складу са еволуцијом носача информација, велика револуција за кориснике речника наступила је појавом речника на CD-ROM носачима информација. Ово је донело новине у начину приступа информацијама у речницима који се више нису морали прегледати линеарно, попут папирних верзија речника, већ су омогућавали једноставну навигацију која се састојала у тражењу речи путем претраге кључним речима и праћењу хипервеза за упутнице. Деведесетих година XX века долази

до брзе појаве различитих електронских носача информација па самим тим и различитих начина приступа речницима. Сасвим сигурно највећу револуцију у погледу приступа електронским речницима представља појава интернета, па савремена лексикографија у свету данас најчешће подразумева само дигиталне речнике (De Schryver 2003). Једна потврда овог тврђења је одлука издавачке куће Оксфордског универзитета (енг. *Oxford University Press*), донета 2010. године, да свој чувени речник са дугом традицијом *Оксфордски речник енглеског језика* (енг. *Oxford English Dictionary*) убудуће објављује само као онлајн издање. У међувремену је велики број речника постао доступан онлајн а за њихову израду се користе бројни ресурси и алати развијени за обраду природних језика који лексикографима олакшавају посао. Међу њима су најзначајнији електронски корпуси из којих се врши ексцерпција одредница, примера и синтаксичких модела, морфолошки речници, семантичке мреже које ближе одређују значења и дефиниције, алати који одређују да ли је реч термин у некој области, итд. Ови и слични ресурси и алати доприносе и развоју низа могућности које речник као ресурс пружа корисницима. Тако данас уз помоћ једног речника можемо за изабрану одредницу видети све облике леме, најчешће колокације, превод на друге језике, илустрацију или фотографију, изговор у облику звучног записа, значењски повезане речи и низ других информација.

2.2 Основа за развој дигиталних речника српског језика

Српска лексикографија као и лексикографије других језика почиње да се развија из практичних побуда. Након 1690. године која је за српску историју протекла у знаку Велике сеобе народа под Арсенијем Чарнојевићем, писмени слојеви становништва у новом, католичком и германском окружењу, осетили су потребу за учењем латинског и немачког језика. Према подацима из *Српске библиографије за новију књижевност 1741-1867* библиографа Стојана Новаковића, објављено је седам двојезичних речника који укључују српски и друге језике (немачки, латински, грчки) (Драгићевић 2014). *Латинско-славеносрпски речник* чији је аутор највероватније Захарије Орфелин сматра се првим српским речником. Објављен је 1766. године као део граматике латинског језика. Затим следи још једна граматика опремљена речником, највероватније истог аутора. И први немачко-славеносрпски речник који потиче из 1774. године био је део граматике. Најважнији представник класичног преводног речника, за разлику од претходних који су били тематски, јесте *Немецки и сербски словар* непознатог аутора штампан у Бечу у штампарији Ј. Курцбека у два издања (1790. и 1791. године) (Ружин Ивановић 2018). Захарије Папа Георгије је 1803. објавио први српско-грчки речник за потребе трговинске размене са Грцима. Од половине XIX века настају и двојезични речници који укључују друге европске језике. Свакако да је XIX век у оквирима српске лексикографије обележен радом Вука Стефановића Караџића. *Српски рјечник истолкован њемачким и латинским ријечима* изашао је 1818. године, а његово друго издање штампано је 1852. године. Неизбежно је поменути и лексикографски рад Ђуре Даничића чији резултат је пре свега *Рјечник из књижевних старина српских* у три свеске, издат 1863. и 1864. године.

Институционални рад у српској лексикографији се реализује средином XIX века. Почетком тог века започета је израда *Рјечника хрватскога или српскога језика ЈАЗУ* (Југославенска академија знаности и умјетности). Овај речник је објављиван скоро сто година, од 1882. до 1976. године, у 97 свезака повезаних у 23 тома и „представља најобимније завршено дело српске и хрватске лексикографије“ (Мацановић 2018).

Теоријско-методолошке темеље јединственог лексикографског програма за израду дескриптивног речника српског језика поставили су у континуитету чланови Друштва српске словесности, Српског ученог друштва, потом и Српске краљевске академије. Јован Стејић, као члан Друштва српске словесности, 1853. у 5. свесци *Гласника*

објављује *Предлог за српски речник и српску граматику*. Стојан Новаковић, као члан Српске краљевске академије, у посланици поводом стогодишњице рођења Вука Караџића (1887), *Српска краљевска академија и неговање језика српског*, даје упутство шта би све требало да уђе у речник и којом методологијом би требало да се ради.

Први том *Речника српскохрватског књижевног и народног језика САНУ* (Српска академија наука и уметности) (скр. *Речник САНУ*), објављен је 1959. године. Објављивању првог тома речника претходили су издавање огледних свезака (од 1913. до 1953. године)[3] и *Упутства за израду речника САНУ,* која су израђена и умножена 1959. године као интерни рукопис Института за српски језик САНУ (Гортан-Премк 2010). Ова *Упутства* су касније допуњавана и према њима се и данас израђује *Речник САНУ*. Треба напоменути да су целокупан начин рада и искуства стечена на изради *Речника САНУ* резултовали формирањем тзв. Београдске лексикографске школе (Ристић 2014). Неизбежно је поменути значај рада Александра Белића који се огледа у *Уводу* првом тому *Речника САНУ* у коме он даје теоријске, концепцијске и методолошке поставке Београдске лексикографске школе. Рад на овом речнику и даље траје, а тренутно последњи том, књига XXI, која обрађује лексику од „погдекад(а)“ до „покупити“, изашао је из штампе 2019. године.

У XX веку светлост дана је, захваљујући тиму лексикографа који је лексикографска искуства стекао на изради *Речника српскохрватског књижевног и народног језика САНУ*, на темељима Београдске лексикографске школе, угледао и *Речник српскохрватскога књижевног језика* Матице српске у шест томова, са преко 150.000 одредница и њиховим речничким чланцима, објављиван од 1967. до 1976. године. Матица српска је 2007. године објавила и једнотомни *Речник српскога језика*. Овај речник је доживео више издања, друго из 2011. и треће 2018. године.

Почетак организованог рада у области терминологије у Србији везује се за Друштво српске словесности. Чланови Друштва су средином XIX века издвојили и предложили групу речи која до тада није била у широј употреби. Неке од тих речи су тада предложене, неке су преузете из народног говора а некима је проширено значење (В. Јовановић 2018). Групи тада предложених речи припадају: *правобранитељ*, *сазвежђе*, *прибор*, *попис*, *сукоб*, *саобразност*, итд. Са развојем многих области науке, војске, администрације, уметности и културе у првој половини XX века долази до главног развоја терминолошке лексикографије. Године 1938. излази двосмерни *Француско-српскохрватски војни речник* О. Обрадовића који у уводу износи детаљно разрађене методолошке и језичке проблеме терминолошке лексикографије који су присутни и данас. Од 1990. године до данас број терминолошких речника на српском језику прелази 200 речника и то је најпродуктивнији период српске терминолошке лексикографије (В. Јовановић 2018). Двојезични и вишејезични терминолошки речници настају из потребе да се преводиоцима олакша превођење термина из стручних области знања са једног језика на други. Услед неконзистентности у библиографским описима и недостатка описа старијих издања узајамна база COBISS[4] (слов. *Kooperativni online bibliografski sistem in servisi* - Кооперативни онлајн библиографски систем и сервиси) не даје прецизне

[3] Прва свеска под називом *Огледно издање бр. 1: Српски речник књижевнога и народнога језика* изашла је 1913. године. Друга свеска *Речник српскога књижевног и народног језика, Огледна свеска* објављена је 1944. године, док је трећа, под називом *Огледна свеска Речника српскохрватског књижевног језика,* објављена 1953. године.

[4] http://sr.cobiss.net/ (приступљено 01.09.2021)

резултате[5]. Према подацима из тезе (Begenešić 2010) само српско-немачких и немачко-српских терминолошких речника издатих у интервалу од 1945. до 2000. године има 130 док узајамна база садржи опис незнатно већег броја терминолошких речника.

[5] Претрага која би требало да пронађе терминолошке речнике даје свега око 150 речника. На листи није познати *Њемачко-српски рударски речник* Драгутина Степановића из 1923. године јер није обрађен. Такође недостаје и петојезични *Рударски речник* из 1970. године који јесте обрађен у систему COBISS али није адекватно библиографски обрађен.

2.3 Врсте речника

Према једној дефиницији, речник је „публикација која систематски региструје и тумачи већи или мањи, општији или посебан део укупног фонда речи једног или више језика“ (Bugarski 1995). Састављањем речника и теоријским принципима њиховог настајања бави се лексикографија која је под утицајем дигиталне револуције доживела велике промене. У складу са тим, иако се речници могу поделити на основу више наменских, садржинских и других критеријума, наш поглед на речнике захтева поделу првенствено на традиционалне и електронске речнике.

2.3.1 Традиционални речници

Под традиционалним речницима подразумевају се речници намењени људима и штампани на папиру. Они се, пре свега, деле на енциклопедијске и језичке речнике. Енциклопедијски речници или енциклопедије са језичким речницима имају заједнички готово само начин представљања података - по алфабетском поретку речи.

Језички речници се могу поделити на више начина имајући у виду различите аспекте њихове намене, употребе, обухвата и сл. Ранко Бугарски наводи поделу речника (Bugarski 1995) према: броју језика чију лексику обрађује, на једнојезичне, двојезичне и вишејезичне; намени, на школске, приручне, академске и др.; временском обухвату, на етимолошке, историјске и савремене; обухвату, на опште и специјализоване; слоју језика који захватају, на ортографске, акцентолошке, граматичке, терминолошке, фразеолошке и др. Данко Шипка (Šipka 2006) речнике дели према сфери интересовања, на лингвистичке и енциклопедијске, према броју језика, на једнојезичке, двојезичке и вишејезичке, према обухвату лексике, на опште и посебне, према начину представљања лексике, на генералне и специјалне. Ладислав Згуста, језичке речнике класификује (Zgusta 1991) према: броју језика које обухватају, на једнојезичне, двојезичне и вишејезичне; намени, на педагошке, ортографске, научне, дескриптивне и сл.; временском обухвату, на дијахронијске и синхронијске (дијахронијске речнике разврстава на историјске и етимолошке); обухвату, на опште (стандарднодескриптивне и општедескриптивне речнике) и ограничене (посебне) речнике; обиму, на речнике малог обима, средњег обима и сл. Згуста издваја и категорију речника традиционалних типова и под њима подразумева тезаурусе, као речнике који теже исцрпности, и посебне речнике које потом дели на: вокабуларе, као двојезичне преводне листе речи; и глосаре, као листе речи са објашњењем оних речи за које се сматра да је читаоцу тешко да их разуме, иако припадају његовом говорном језику.

На основу наведеног, можемо закључити да аутори класификују речнике на сличан начин, уз извесне терминолошке разлике.

Оно што је свакако главна одлика традиционалних речника јесте традиционални приступ представљању података у овом типу речника. Речнички чланци су представљени линеарно на папиру, у виду надовезивања различитих лексикографских елемената које су у складу са типом речника. Још једна од одлика традиционалних речника која се полако губи јесте и традиционални приступ изради што се огледа у примени методологије класичне лексикографије. Ипак, са развојем електронске лексикографије, и при изради традиционалних речника почеле су да се користе методе електронске лексикографије. Корпус за ексцерпцију постао је углавном електронски, користе се специјализовани програми за писање речника, итд.

Као последица информационе револуције, појавила су се питања везана за опстанак традиционалних речника. Искристалисане су главне предности традиционалног речника у односу на електронски: традиционални речник има симболичну вредност као физички објекат, опипљив је и прилично једноставан за

коришћење (прелиставањем и прегледањем страна), не умара очи, трајан у смислу конзистентности функција које нуди и, при том, за коришћење није потребан рачунар нити било који други уређај (De Schryver 2003). Пракса показује да ово ипак није довољно па би се рекло да примат преузимају речници у дигиталном облику.

2.3.2 Речници у дигиталном облику

Речници у дигиталном облику могу бити дигитализовани штампани речници (традиционални речници) или речници који су настали као дигитални (енг. *born digital*) а који могу бити намењени људима или рачунарским програмима.

Дигитализовани речници

Сканирање је широко прихваћен начин за конверзију речника из аналогног, папирног облика, у дигитални облик. Сканиране слике страница речника нипошто не задовољавају потребе корисника јер у смислу претраживања не пружају много више од традиционалних речника, осим што су физички „преносивије“. Оно што већина корисника речника сматра основном предношћу електронског речника јесте напредан систем претраге. Како би се добио сасвим функционалан речник потребно је после сканирања урадити још неколико захтевних корака. Како би се омогућило индексирање текста, односно претрага садржаја, првобитно сканирани речник се обрађује помоћу софтвера који омогућава оптичко препознавање карактера – OCR (енг. *Optical Character Recognition*). Како резултат софтвера за оптичко препознавање карактера није савршен, добијен текст је потребно преконтролисати и кориговати. На овај начин се долази до претраживог текста али ни то није последња тачка обраде јер речници, кроз употребу различитих квалификатора, симбола и стилова у форматирању текста, користе језик који пружа додатне информације о садржају речника. Ове информације такође треба искористити и укључити их у претраживање. Наредни корак који то омогућава јесте парсирање текста (енг. *parsing*). Парсирани текст је могуће сместити у лексикографску базу која доприноси практичнијем одржавању и унапређивању речника, премда се дигитални речник може одражавати без употребе базе података. Постоје различити облици дигиталних речника који се чувају у XML (енг. *eXtensible Markup Language*) формату[6], другим текстуалним форматима, нпр. DELA (више о овом формату у поглављу 3.1), или у виду HTML (енг. *HyperText Markup Language*) презентације[7]. И ови речници могу бити намењени људима и рачунарским програмима али се пуна функционалност и контрола речника постиже применом лексикографске базе. Алтернативу парсирању текста пружају програми засновани на машинском учењу који препознају делове речника и обележавају их. Пример таквог програма јесте *GROBID-Dictionaries* (Khemakhem, Foppiano, и Romary 2017) који издваја речнички чланак и његове делове из PDF (енг. *Portable Document Format*) формата[8], над којим је извршено оптичко препознавање карактера, и обележава их у складу са TEI смерницама (више о овом у поглављу 4.1.1). Треба напоменути да програм *GROBID-Dictionaries* најбоље ради над речницима који садрже структурално унифициране речничке чланке код којих нема значајних одступања у дужини или структури речничког чланка. Ово углавном

[6]XML – Широко заступљени проширив језик за означавање који дефинише скуп правила за означавање докумената који је разумљив и људима и рачунарима. Више прочитати на: https://www.w3.org/XML/ (приступљено 15.10.2018)

[7] https://www.w3.org/html/ (приступљено 15.10.2018)

[8] https://sr.wikipedia.org/wiki/PDF (приступљено 4.1.2019)

представља проблем је се речнички чланци, нарочито у описним речницима, разликују по дужини и структури.

Поред сканирања, као начин за конверзију речника из штампаног у дигитални облик користи се и ручно прекуцавање текта. Управо на овај начин је настала машински читљива верзија *Оксфордског речника енглеског језика* (енг. *Oxford English Dictionary*, *OED*) (Berg и остали 1988). Речник је прекуцан у оквиру пројекта који је започео 1984. године и трајао пет година. Коштао је преко 13 милиона долара и укључивао 120 људи који су прекуцавали текст као и 50 људи који су читали и проверавали прекуцано. Текст речника је том приликом обележен језиком SGML (енг. *Standard Generalized Markup Language*), претечи раније поменутог језика XML. Језик SGML је у том тренутку представљао сасвим нову схему за обележавање типографских специфичности које одговарају деловима речничког чланка. Други пример конверзије текста речника из штампане у дигиталну верзију прекуцавањем јесте *Речник немачког језика* (нем. Deutsches Wörterbuch) браће Грим који представља један од најбитнијих докумената у историји немачке културе. Како би настала дигитална верзија, 331.056 речничких чланака, тј. 300 милиона карактера овог речника је чак два пута прекуцано у Кини (Christmann 2001). Прекуцавање је извршено независно два пута, како би се елиминисале грешке. Дигитална верзија речника је објављена 2004. године.

Речници настали у дигиталном окружењу

Речници настали у дигиталном окружењу, с друге стране, могу од самог почетка бити креирани применом електронске лексикографије, односно њених процеса. Теоретски постоје три фазе у стварању речника и то: планирање, писање и производња. Практично, рад на стварању онлајн речника који је заснован на корпусу, како је детаљно изложено у раду Анете Клоза (Annette Klosa), може се поделити у шест фаза (Klosa 2013): припрема, прикупљање података, аутоматизација, обрада података, анализа података и фаза припреме за објаву на мрежи.

Фаза припреме састоји се из писања скице иза чега следи планирање и израда концептуалног плана. Планирање подразумева финансијски и организациони план посла, као и планирање људских ресурса. Концептуални план је упутство за лексикографа које садржи опис корпуса, опис речничких записа, означавање делова информација у речнику, итд. Планирање људских ресурса мора да обухвати, осим лексикографа, и корпусне лингвисте, информатичаре, графичке дизајнере, техничку подршку и друге стручњаке јер свако од њих има своју улогу у производњи речника. Корпусни лингвиста ради на дизајну корпуса заједно са лексикографима. Графички дизајнер, у договору са лексикографом и информатичарем, скицира различите приказе који се прилагођавају потребама различитих група корисника којима је речник намењен. У овој фази се израђује и пилот-студија која садржи прелиминарни скуп лексикографских информација, листу одредница, план објављивања речника, итд. Оно што је посебна погодност ове врсте речника јесте могућност објављивања „незавршене верзије речника“, односно, речнику се накнадно или сукцесивно могу додавати нови речнички записи или нове врсте информација које не морају бити искључиво лексичке природе (модули) што омогућава да се речник брзо и често ажурира без пуно техничких потешкоћа која се могу јавити код традиционалних речника (штампа комплетне нове верзије речника који може имати и више томова).

Фаза прикупљања података подразумева прикупљање примарних, секундарних и других извора, звучних и филмских записа, илустрација, итд. Код речника на мрежи, примарни извор је електронски корпус за који текстове прикупља корпусни

лингвиста, секундарни извори су други речници, били они папирни или електронски, граматике и друге лексичке базе података.

Фаза аутоматизације подразумева анотацију, морфосинтаксичко обележавање и лематизацију корпуса, постављање система за одржавање и претрагу корпуса, инсталацију система за писање речника (енг. *dictionary writing system*) и друге сродне задатке.

Фаза обраде података подразумева ексцерпцију кандидата за речничке записе, идентификацију фреквенција у корпусу, анализу колокација, одређивање структуре лексикографске базе, означавање звучних и филмских записа, као и многе друге задатке. Овом фазом се махом баве рачунарски и корпусни лингвисти који ове информације добијају из корпуса.

Код фазе анализе података ради се на аутоматској компилацији информација, писању речничких записа, повезивању записа хипервезама, укључивању илустрација, филмских записа, итд. Корпусни и рачунарски лингвисти и лексикографи заједно раде на аутоматској компилацији информација о фреквенцији, колокацијама, парадигматским паровима, аутоматској екстракцији примера из корпуса и другим функцијама. Лексикографи се у овој фази баве истим пословима као и приликом креирања традиционалног речника, али уз то повезују речничке записе са екстерним изворима као што су веб-стране, енциклопедије и речници, илустрације, прикази, звучни или филмски записи.

Фаза припреме за објављивање на мрежи укључује редакцију, рецензију, тестирање хипервеза, мултимедијалних елемената, тестирање презентације на мрежи, креирање упутстава за коришћење, итд.

Клоса напомиње да наведене фазе не морају пратити једна другу нити садржати све подзадатке (Klosa 2013), што све зависи од типа речника и организације посла. Она такође наглашава да рачунари играју основну улогу у свим фазама развоја и да је рад на отвореном електронском речнику посао без окончања који се своди на организацију и контролу процеса.

У раду (Granger и Paquot 2012) дат је приказ најзначајнијих новина које доносе електронски речници у односу на традиционалне а то су: интеграција корпуса, провера квалитета података, ефикасност приступа, прилагодљивост и укључивање корисника у писање речника удруженим радом.

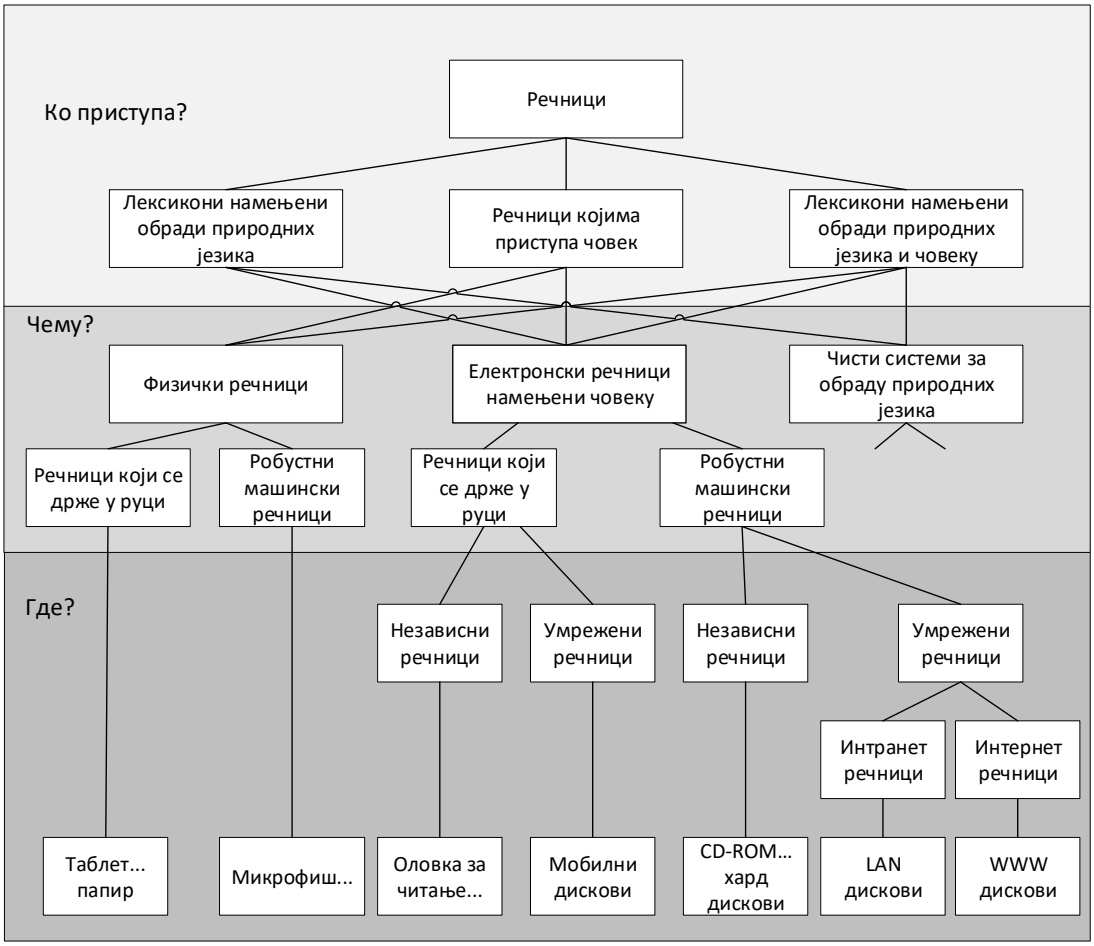
Као резултат производње речника у дигиталном окружењу су настали речнички портали који повезују више речника доступних на мрежи кроз заједничку инфраструктуру. Међу речничким порталима постоје две подгрупе, речничке мреже и речничке збирке. У речничким мрежама су речници повезани путем заједничке микроструктуре па је због захтевности преклапања број речника у њима мањи. Типичан пример речничке мреже јесте речнички портал *Онлајн лексички информациони систем за немачки језик* - OWID[9] (нем. *Online-Wortschatz-Informationssystem Deutsch*). Речничке збирке, с друге стране, садрже већи број речника јер је међусобно повезивање мање захтевно (Štrkalj Despot и Möhrs 2015). Пример речничке збирке јесте портал *Твој речник*[10] (енг. *Your Dictionary)*.

Оно што је неопходно разјаснити пре разматрања класификације електронских речника јесте сам термин „електронски речник“. У свом зачетку, средином

[9] https://www.owid.de/ (приступљено 20.02.2022)

[10] https://www.yourdictionary.com/ (приступљено 20.02.2022)

XX века, електронски речници су носили назив „машински читљиви речници“. По једној од дефиниција (Jacquet-Pfau 2002) у којој се прави дистинкција између електронских и машински читљивих речника „електронски речник је намењен рачунарској обради текста-корпуса за разлику од машински читљивог речника намењеног кориснику рачунара“. Исти став дели Морис Грос (Gross 1987) с обзиром на то да он електронским речницима сматра речнике DELA (фр. *Dictionnaires électroniques du LADL*), намењене употреби од стране рачунара. Такав став је присутан и у раду (Крстев 2000) где су машински речници дефинисани као дигитализована копија папирних издања, док су електронски речници „намењени аутоматској трансформацији текста, а не људском кориснику“ (Крстев 2000). Стога ће се у овом раду користити исти став, дакле, електронским речником ће се сматрати речник у дигиталном формату намењен рачунарским програмима за потребе различитих апликација. Ово је неопходно нагласити јер у данашње време постоје и друга становишта по којима се дигитални речници намењени човеку називају електронским, док се речници намењени рачунарској обради називају машински читљивим („European Dictionary Portal: Criteria for Inclusion“ 2017). Такав приступ је заступљен и у две често цитиране класификације које ће бити представљене у наставку.



Слика 1 Таксономија речника по Де Шриверу (De Schryver 2003)

Широко прихваћена класификација речника из 2003. године (De Schryver 2003) настала је анализом и комбинацијом више постојећих типологија речника. Она

комбинује Мартинову (Martin) типологију[11] из 1992. године, типологију Ајдове[12] (Ide) из 1993. године, Шарпову[13] (Sharpe) из 1995. године, типологију Лејрове[14] (Lehr) из 1996. године и типологију Несијеве[15] (Nesi) из 2000. године са текућим правцима развоја (у погледу медијума) што јој даје додатну вредност. Де Шриверова (De Schryver) типологија се декларише по етапама, кроз одговор на три питања:

- Ко приступа речницима?
- Чему се приступа?
- Где се приступа речничким подацима?

На слици 1 представљена су три слоја, обојена различитим нијансама, која дају одговор на претходно постављена питања.

Одговор на прво питање јесте да речнику приступају човек или машине. Рачунари приступају лексиконима намењеним обради природних језика, док човек приступа њему читљивим речницима. С друге стране, и човек и машине могу користити речнике намењене обради природних језика.

Друго питање тиче се речничког медијума. Де Шривер овде прво уводи поделу на електронске и физичке медијуме, које потом дели на ручне уређаје (мање, преносиве уређаје за читање) и робусне машине. Аутор под робусним машинама подразумева лаптоп и стоне рачунаре. Електронски речници намењени човеку могу бити базе података, лексикони за обраду природних језика (нпр. лексикон за проверу типографских грешака) или базе података дизајниране за обраду природних језика али које могу користити и људи (нпр. Ворднет[16] (енг. *WordNet*) и Фрејмнет[17] (енг. *FrameNet*)). Физички медијум подразумева штампану верзију речника намењеног човеку или штампану верзију лексикона истовремено предвиђеног за обраду природних језика и људској употреби.

Одговор на треће питање односи се на врсту складишта, у рачунарском смислу, или хардвера на коме се налази речник. У разматрању физичког медијума, Де Шривер дијахронијски набраја глинене плочице, воштане дрвене даске, папирус, свитак, кодекс (рукопис на прегаменту или папиру), све до штампаних страна. Ту су и записи конзервираних речника са претходно набројаних специфичних носача смештени на

[11] Мартинова типологија препознаје: (1) речнике за људске кориснике, (2) рачунарски засноване речнике, (3) машински читљиве речнике, (4) лексичке/терминолошке банке, (5) машинске речнике, (6) лексичке базе података и (7) лексиконе вештачке интелигенције.

[12] Типологију Ајдове препознаје: (1) специфичне електронске двојезичне речнике, (2) електронске бележнице, (3) електронске двојезичне речнике на CD-ROM-у и (4) софтвер за електронски речник.

[13] Шарпова типологија је на типологију Ајдове додала још два типа: (5) електронске двојезичне речнике на дискети и (6) електронске двојезичне речнике са сканерима за оптичко препознавање карактера.

[14] Типологија Лејрове у првој фази речнике дели на офлајн и онлајн речнике. Офлајн речници се деле на џепне електронске речнике и речнике за персонални рачунар са каснијим подврстама на речнике на CD-ROM-у, дискети и друге.

[15] Ова типологија констатује да су разлике између подела добијених претходним типологијама замагљене. Тако се нпр. онлајн материјали комерцијализују и продају на CD-ROM издању, с друге стране садржаји CD-ROM речника се постављају на интернет итд.

[16] Ворднет је вишејезична семантичка мрежа која представља скуп синсетова и релација између њих (Fellbaum 1998). Више о Ворднету може се прочитати у делу 3.1.

[17] Фрејмнет је лексичка база података енглеског језика заснована на анотацији примера како су речи коришћене у текућем тексту. Ова база је машински читљива али је истовремено намењена и људима. Више о пројекту: https://framenet.icsi.berkeley.edu/fndrupal/about (приступљено 18.07.2021)

микрофиш. У овом случају је неопходна робусна машина која чита оригиналне слике са микрофиша[18]. Електронски речници, било они који се држе у руци, било робустни машински речници могу се свести на независне речнике и речнике који су умрежени. Аутор као пример независних ручних речника наводи речнике на преносивим електронским уређајима (енг. *Portable electronic devices, PED*), оловкама за читање[19] (енг. *reading pens*) или лексиконе који служе као помоћ при слању порука путем мобилних телефона. Ручним умреженим речницима Де Шривер сматра онлајн речнике који су доступни преко уређаја који су врста мобилних електронских роковника (врло ретко се срећу у свакодневном животу).

Де Шривер типичним примерима робустних независних електронских речника сматра речнике складиштене на дисковима, који могу бити CD-ROM, DVD, тврди дискови (енг. *hard disk*), итд. Тврди дискови се најчешће користе за складиштење умрежених речника.

Електронски робустни умрежени речници могу бити доступни путем интранета или интернета. Речници представљени путем интранета доступни су корисницима локалне мреже (енг. *Local Area Network, LAN*) најчешће унутар организација библиотека, универзитета, итд., док су робустни умрежени електронски речници доступни путем интернета и складиште се на серверима.

Према широко утемељеној класификацији речника Ангелике Шторер (Angelika Storrer) (Klosa 2013), која је сачинила поделу електронских речника доступних преко интернета, они се у зависности од начина производње, доступности, довршености, хипертекстуалности, интеракције са корисницима, мултимедијалности и приступа могу класификовати на следећи начин:

- Изворни облик:
  - изворно објављен као штампани речник;
  - изворно објављен као електронски речник али за коришћење на персоналном рачунару или офлајн;
  - изворно објављен у електронском формату (енг. *born digital dictionaries*);
- Довршеност:
  - статички речник;
  - динамички речник;
- Хипертекстуалност:
  - хипертекстуални речник;
  - речник без хипертекстуалних могућности;
- Интеракција са корисницима:
  - речник са могућношћу интеракције;
  - речник без могућности интеракције;
- Мултимедијалност:
  - речник са текстом, илустрацијама и дијаграмима;
  - речник са текстом и звучним записима;
  - речник са текстом, звучним записима и видео записима;
  - речник без мултимедијалних садржаја;
- Приступ речнику:
  - речник са приступом у виду померања горе-доле речничких записа;
  - речник са приступом преко листе одредница које су у облику хипервеза;

[18] Микрофиш је облик микрофилма 105x149mm.

[19] Уређај који помаже особама са дислексијом при читању.

- речник коме се приступа путем поља за претрагу;
- речник са комбинованим приступом.

Ми ћемо се у наредним поглављима бавити првенствено електронским речницима намењеним рачунарским програмима и апликацијама доступним путем интернета. Имајући у виду прву класификацију по Де Шриверу *Морфолошки речник српског језика* којим ћемо се ми бавити према одговору на питање ко приступа речнику спада у лексиконе намењене обради природних језика и човеку. Према одговору на питање чему се приступа наш речник спада у електронске речнике па потом робустне машинске речнике. Према одговору на питање где се приступа речничким подацима, наш речник спада у умрежене интернет речнике. Пратећи класификацију Ангелике Штрорер *Морфолошки речник српског језика* према изворном облику спада у речник изворно објављен у електронском формату. Према довршености спада у динамички речник, док по хипертекстуалности спада у хипертекстуалне. *Морфолошки речник српског језика* омогућава интеракцију са корисницима али према мултимедијалности спада у речнике без мултимедијалних садржаја. Према приступу речнику спада у речник са комбинованим приступом (приступ одредницама у виду померања листе горе-доле, приступ одредницама у облику хипервеза и путем поља за претрагу).

Методе електронске лексикографије

Имајући у виду да се лексикографија данас ближи крају прелаза од традиционалне ка електронској, теоретичари лексикографије се баве питањем положаја и односа лексикографије са другим наукама у научној класификацији. Како је теоријска лексикографија наука о речницима, првенствено традиционалним, поставља се питање постојања потребе за новом теоријом која ће укључивати контекст који је производ новоформираног дигиталног окружења речника (Tarp 2012), односно теоријом која ће обухватити методе електронске лексикографије.

Свен Тарп (Sven Tarp), који се бави функционалном теоријом лексикографије, закључује да „лексикографија неће престати да буде независна дисциплина са сопственим тематским пољем, са сопственом теоријом и праксом али ће тежити повезивању и интеракцији са сродним дисциплинама у широкој области информационих наука“ (Tarp 2012). У раду (Bergenholtz и Bothma 2011) износи се став да је лексикограф посебна врста информатичког експерта те да лексикограф не мора бити делом лингвиста, осим у случају када се бави лексикографијом за потребе споразумевања. Иако постоји тежња да лексикографија са преласком из традиционалне у електронску пређе из поља лингвистике у поље информационих наука, реална ситуација је да је она у пресеку ова два скупа, како се сликовито изразио Адам Килгариф (Adam Kilgarriff): „Лексикограф је подељена душа, делом научник, делом градитељ алата. Научник је лингвиста који жели да опише језик. Градитељ алата, служећи се информационим наукама, жели да помогне кориснику да пронађе информацију коју тражи“[20] (Kilgarriff 2012).

У складу са претходно изнетим закључцима, у зависности од типа речника који се креира, лексикографија користи различите методе ослањајући се на друге науке. Методе које лексикографија у основи користи јесу лингвистичке, социолошке, филолошке, као и интерпретативне. У наставку ће поред ових метода бити наведене и методе практичне лексикографије, изложене у (Schierholz 2015), које се користе током фаза у стварању раније помињаних електронског речника на мрежи.

[20] Изворно: „A lexicographer is a divided soul, part scientist, part tool-builder. The scientist is a linguist, wanting to describe the language. The tool-builder wants to help the user find the information they want.“

Током припремне фазе користе се лингвистичке методе, као и методе пословног управљања како би се регулисали финансијски планови, људски ресурси, распоред посла, итд. Како би се посао ефикасно обављао, често је потребно прихватати савете и процене искусног лексикографа, нарочито по питању распореда посла. Наиме, у прошлости је било уобичајено да лексикографи који се баве писањем речничких чланака то раде по алфабетском поретку па се дешавало да су чланци с почетка речника (са почетним словима алфабета), рађени на почетку пројекта, исцрпнији, дужи и бројнији, док су они с краја речника, рађени при истеку временских и осталих ресурса, краћи и бројно мање заступљени. План којим би се избегавале ове потенцијалне опасности јесте да се лексикографи који су стручњаци за одређену област знања баве речничким чланцима који покривају ову област, или да се изврши расподела посла па да се сваки лексикограф бави само одређеним сегментима речничког чланка. Приликом одређивања распореда речничких чланака, при алфабетском поретку користи се метода алфабетизације, док се код поретка према тематским областима користе системи група концепата и потконцепата, као и погодна метода означавања како би се леме и чланци систематично означили у систему. Морају бити развијени алати за лексикографски рад и показни чланци са најбољим примерима, како би се користили уједначени принципи за све аспекте речничког чланка. У фази прикупљања података користе се методе корпусне лингвистике како би се дошло до репрезентативног и избалансираног корпуса који одговара намени речника. Након тога се користе и методе обраде природних језика како би се извршили задаци обраде корпуса, као што су токенизација, вертикализација, лематизација и морфосинтаксичка анотација. У фази обраде података примењује се метода одабира лема у складу са дистрибуцијом у корпусу имајући у виду тип речника, одлике других речника из исте области и намењених истој циљној групи. У фази припреме за објаву на мрежи следе се унапред постављене процедуре за уредничку контролу и тестирање хипервеза и мултимедијалних сегмената.

Преглед угледних речника доступних у дигиталном окружењу

Згуста у првој монографији посвећеној лексикографији, *Приручник лексикографије*, наводи да се главне информације о методама лексикографије могу добити пажљивим проучавањем добрих речника који су у датом тренутку доступни (Zgusta 1991). Ми ћемо ово правило класичне лексикографије применити кроз анализу електронских речника и портала доступних на вебу, како би детаљније размотрили могућности које електронска лексикографија пружа.

Информатизовани Трезор француског језика (фр. *Trésor de la langue Française informatisé, TLFi*), који је слободно доступан на вебу за прелиставање и претраживање[21], садржи француску лексику XIX и XX века у 16 томова и обрађује 100.000 лексичких јединица са 270.000 дефиниција и 430.000 примера који потичу из речника *Трезор француског језика* (фр. *Trésor de la langue Française*). Речи је могуће прелиставати по абецедном устројству али је такође могућа претрага путем поља за претрагу. Поред писаног, могуће је претраживати и фонетски облик одреднице. На располагању су два приказа резултата претраге: први нуди глобални приказ свих одредница речничких чланака који одговарају постављеном упиту док други даје детаљни приказ сваког речничког чланка који одговара упиту са могућношћу навигације на следећи или претходни речнички чланак. Информатизовани речник пружа и разноврсне могућности приказа речи: приказ флексије тражене речи, изговор, бојење у специфичну боју делова речничког чланка који се односе на извор (цео извор, наслов извора, аутор, датум), врсту

[21] http://atilf.atilf.fr/tlf.htm (приступљено 18.07.2021)

речи (граматички код), напомене о контексту коришћења, дефиниције, домен употребе, синониме, антониме, синтагме у чијем грађењу учествује одредница. Када се у горњем левом углу поред одреднице појави одређени знак (слово H) добија се могућност претраге речи у пет различитих речника француског језика. На слици 2 дат је приказ речничког чланка за лему „*bonheur*“ (срећа), уз коришћење опције бојења његових делова. Црвеном бојом обојен је део који се односи на граматичке категорије, зеленом дефиниције, љубичастом примери употребе, жутом датум издавања извора из кога је пример, наранџастом аутор извора примера и тиркизном бојом синтагме које садрже лему.

Слика 2 Приказ дела речничког чланка за лему „*bonheur*“ у *Информатизованом Трезору француског језика*

Оно на шта ћемо се вратити јесу различите могућности за претрагу, видљиве као картице у заглављу стране приказане на слици 2. Поред поменутог начина претраге путем постављања упита уносом једне речи, пружају се додатне три могућности – претрага уз асистенцију, комплексна претрага и листе речи. Претрага уз асистенцију омогућава да се одаберу квалификатори уз облик који се претражује што омогућава да се резултати претраге за задату реч сузе ограничењима попут врсте речи, домена у коме се реч користи, типа значења, у смислу да ли је значење иронично, метонимично и др., а да се потом одаберу и део структуре речника који се претражује (нпр. дефиниције, етимологија, примери). Овом врстом претраге се могу и елиминисати појављивања неке

друге речи. Комплексна претрага омогућава одабир дела речничког чланка за претрагу, дефинисање да ли одабрани део речничког чланка садржи реч за претрагу, као и одабир саме речи или граматичке категорије за претрагу (уз реч је могуће користити и регуларне изразе). Могуће је комбиновати и више упита постављених уз помоћ ова три поља а такође је уз сам упит могуће написати и коментар који ће заједно са упитом остати сачуван у историји претраживања. Трећа врста претраге на основу упита, који се може поставити двојако, формира листе речи. Могуће је претраживати речи у потрази за флективним облицима[22] и тада постављањем упита уносом, на пример, инфинитива глагола добијамо листу речи која садржи све облике тог глагола, а уносом облика мушког рода једнине именице[23] добијамо листу која се састоји од женског облика једнине, мушког множине итд. Други начин за креирање листе речи јесте креирање подлисте *Трезора француског језика*. У овом случају се упити постављају коришћењем регуларних израза и резултат, односно листа речи, садржи све речи које одговарају упиту а део су *Трезора француског језика*. У оба случаја, и при претрази флективним облицима, и подскупа *Трезора*, добијена листа се може сачувати а њу корисник може поново да прегледа, па чак и да поставља упите над њеним садржајем.

Узевши у обзир све што смо навели до сада, можемо закључити да је *Информатизовани Трезор француског језика* много више од речника јер пружа доста корисних могућности у виду алата за претрагу који посматрају речнике као корпус и могу бити од велике користи истраживачима, пре свега лингвистима.

Мрежа речника за немачки језик (нем. *Wörterbuchnetz*)[24] развијена је на Универзитету у Триру и омогућава истовремено претраживање 28 речника немачког језика. Она пружа могућност приказивања тражене речи или њеног објашњења у умреженим речницима. Треба напоменути да ниједан од ових речника није настао као дигиталан, сви су ретродигитализовани у склопу различитих пројеката па се ниво приказа и претраге разликује од речника до речника (Moulin и Nyhan 2014). Сви речници се могу прелиставати на исти начин, путем алфабетски устројених одредница у менију са леве стране. Неки речници пружају и могућност приказа детаљнијих библиографских информација о извору из кога потиче пример преласком миша преко тог дела речничког чланка. Уграђена је и могућност преузимања директне хипервезе до датог речничког чланка. Поједини речници омогућавају прегледање и преузимање скениране стране (слике) из оригиналног издања речника на којој се налази дати речнички чланак.

Дигитални речник немачког језика (нем. *Das Digitale Wörterbuch der deutschen Sprache*, *DWDS*) доступан је кроз *Информациони систем речи немачког језика из прошлости и садашњости* (нем. *Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*)[25], који почива на три сегмента, речницима, корпусима и делу посвећеном статистичким евалуацијама. Речнички сегмент садржи 6 речника са укупно преко 584.000 одредница. Корпусни сегмент најбрже расте и тренутно садржи 34 корпуса сачињених од преко 70 милиона докумената.

Сегмент посвећен речницима омогућава приказ: граматичких информација, звучних записа изговора, сложених речи у којима се тражена реч налази као

[22] Овај термин, који би могао деловати као плеоназам, се користи како би се тај облик разликовао од деривата, који се у информатици такође сматрају облицима.

[23] У француском језику род није класификациона категорија именица.

[24] http://www.woerterbuchnetz.de (приступљено 18.07.2021)

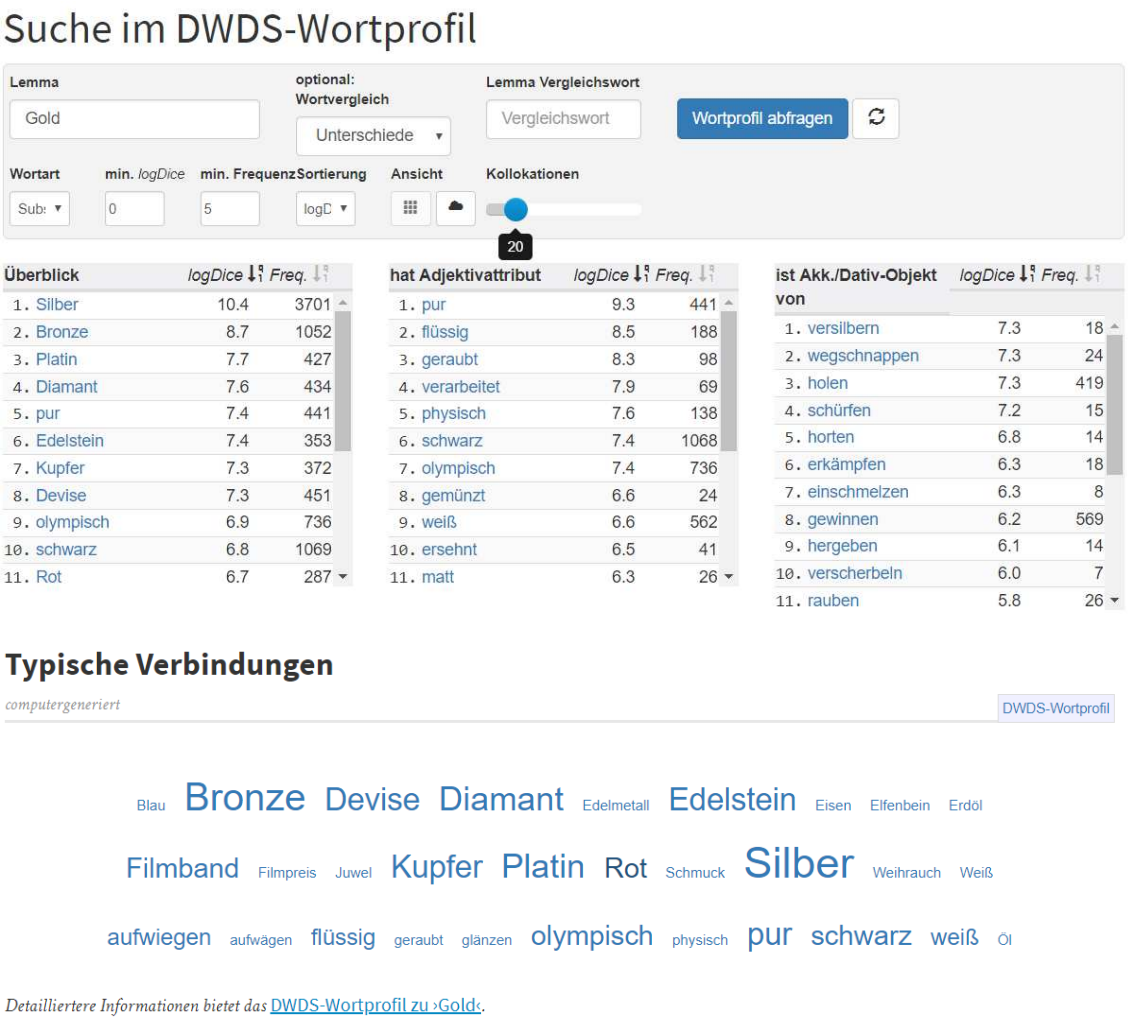
[25] https://www.dwds.de/ (приступљено 18.07.2021)

конституент, информација из речника (дефиниција, примери и извори) са ознаком речника из кога речнички чланак потиче и етимологије. Омогућен је и приказ информација из *Отвореног тезауруса*[26] (енг. *Open thesaurus*), рачунарски генерисаних најчешћих колокација са профилом речи, примера из корпуса у виду конкорданци, хипервезе до речничког чланка речи које су алфабетски испред и иза текуће речи, фреквенције речи и графика хронолошке фреквентности речи. Као новину ћемо посебно издвојити опцију профила речи који се приказује кроз табелу или облак речи. Слика 3 илуструје ове опције на примеру речи „*Gold*“ (злато). Информације у профилу речи се формирају на основу статистичких показатеља појављивања речи у корпусима. Дати су односи речи са другим речима које се најчешће јављају у њеној околини праћени информацијама о фреквенцији и мери logDice, статистичкој информацији која показује колико је дата реч добар кандидат за колокацију, уз могућност сортирања пратећих речи, било преко фреквенције било према logDice. Кликом на сваку колокацију добијају се конкорданце из корпуса са примерима дате колокације. Још једна од напредних опција које овај речник пружа јесте могућност извоза конкорданци, као и профила речи у csv[27] i RTF[28] форматима погодним за даље коришћење. Ова опција више није доступна нерегистрованим корисницима.

[26] www.openthesaurus.de (приступљено 18.07.2021)

[27] CSV (comma-separated values) је формат текстуалне датотеке у којој су приказане вредности раздвојене зарезом.

[28] RTF (Rich Text Format) је формат који садржи информације о фонтовима, форматирању и слично. Развијен је од стране Мајкрософта за потребе размене сопствених производа али је постао широко прихваћен.



Слика 3 Приказ профила речи „*Gold*“ из *Дигиталног речника немачког језика* у виду табеле (горе) и облака речи (доле)

Оксфордски речници енглеског језика (енг. *Oxford English Dictionaries*, *OED*)[29] са граматикама производ су комерцијалне издавачке куће Оксфордског универзитета (енг. *Oxford University Press*). Они пружају информације о изговору (у виду звучног записа, као и у писаном облику коришћењем IPA алфабета), врсти речи, дефиницијама са примерима у облику реченица, уз могућност приказа додатних примера који илуструју употребу, синонимима, пореклу речи, поново са могућношћу прегледања додатних примера, листи речи које се римују са траженом речи, подели на слогове, а на располагању је и листа енглеских речника у којима се може потражити задата реч. Дате су и листе сличних и повезаних речи (речи у којима је тражена реч део вишечлане лексичке јединице). *Оксфордски речници* пружају корисницима који се баве живим језиком актуелне информације о томе које нове речи су ушле у речник, које су најпопуларније речи у различитим енглеским говорним подручјима и сл. Оно што је можда главна одлика *Оксфордског речника* јесте актуелност (динамичност) речника па је сасвим оправдана нова маркетиншка измена назива у *Оксфордски енглески „живи“ речници* (енг. *Oxford English Living Dictionaries*). Тако је, на пример, у *Оксфордски енглески речник* ушла реч *babymoon* са значењем опуштајућег одмора будућих родитеља пре рођења бебе и периода зближавања родитеља са бебом након њеног рођења (слика 4), док је реч

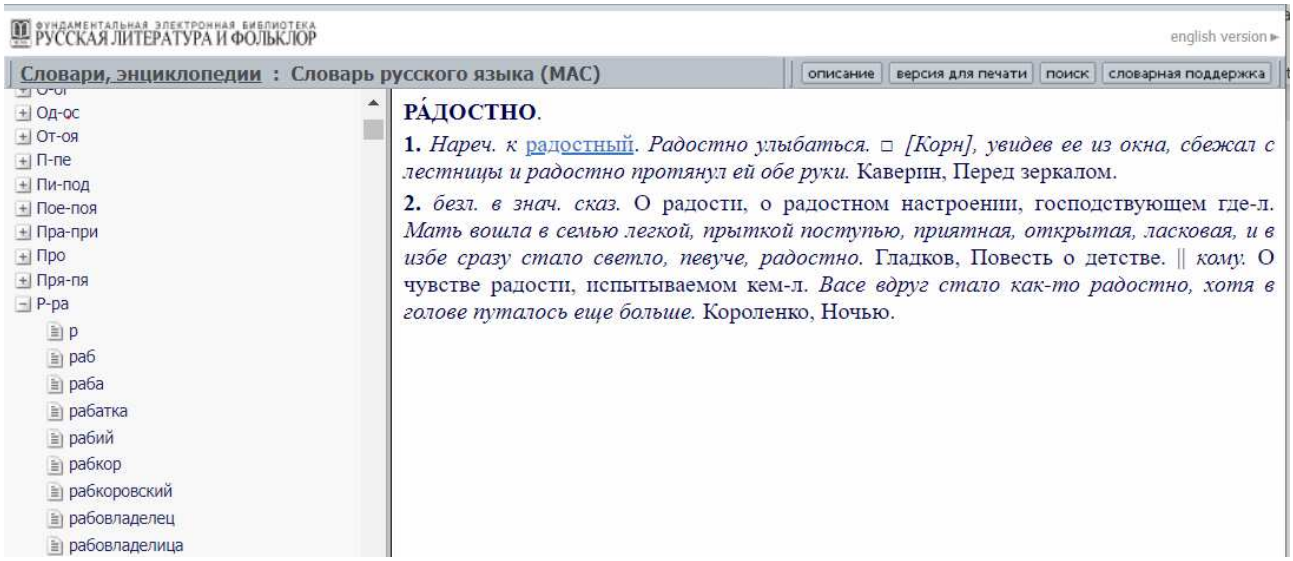
[29] https://en.oxforddictionaries.com/ (приступљено 18.07.2021)

influencer добила ново значење везано за маркетинг као особа која има способност утицаја на потенцијалне купце да купе производ или услугу промовисањем путем друштвених мрежа. Од 2014. године, издавач *Оксфордских речника* покренуо је нови пројекат *Оксфордски глобални језици* (енг. *Oxford Global Languages, OGL*) којим планира развој речника за мање језике и то по угледу на речник енглеског језика, коришћењем развијене лексикографске инфраструктуре.

Слика 4 Део речничког чланка из *Оксфордског речника енглеског језика* који описује нову реч „*babymoon*“

Путем портала *Основна електронска библиотека* (рус. *Фундаментальная электронная библиотека*), доступно је више изворно штампаних речника руског језика. Један од језичких речника је и *Речник руског језика у 4 тома*[30] (рус. *Словарь русского языка в 4 т*) или *Мали академски речник* (рус. *Малый академический словарь*) који садржи више од 80.000 речи. Сваком речнику појединачно је могуће приступити путем интерфејса за прелиставање одредница (слика 5) или путем претраге одредница (опција „поиск“ на слици 5). У опису електронског издања је наведено да су одреднице представљене у алфабетском поретку, независно од тома штампане верзије речника. Сама ознака тома у оквиру кога је штампана реч, као и пагинација у оквиру тома се приказују у прозору за приказ описа одреднице (опција „описание“), као и у верзији речничког чланка за штампу (опција „версия для печати“). Овај речник омогућава праћење хипервеза до упућеница.

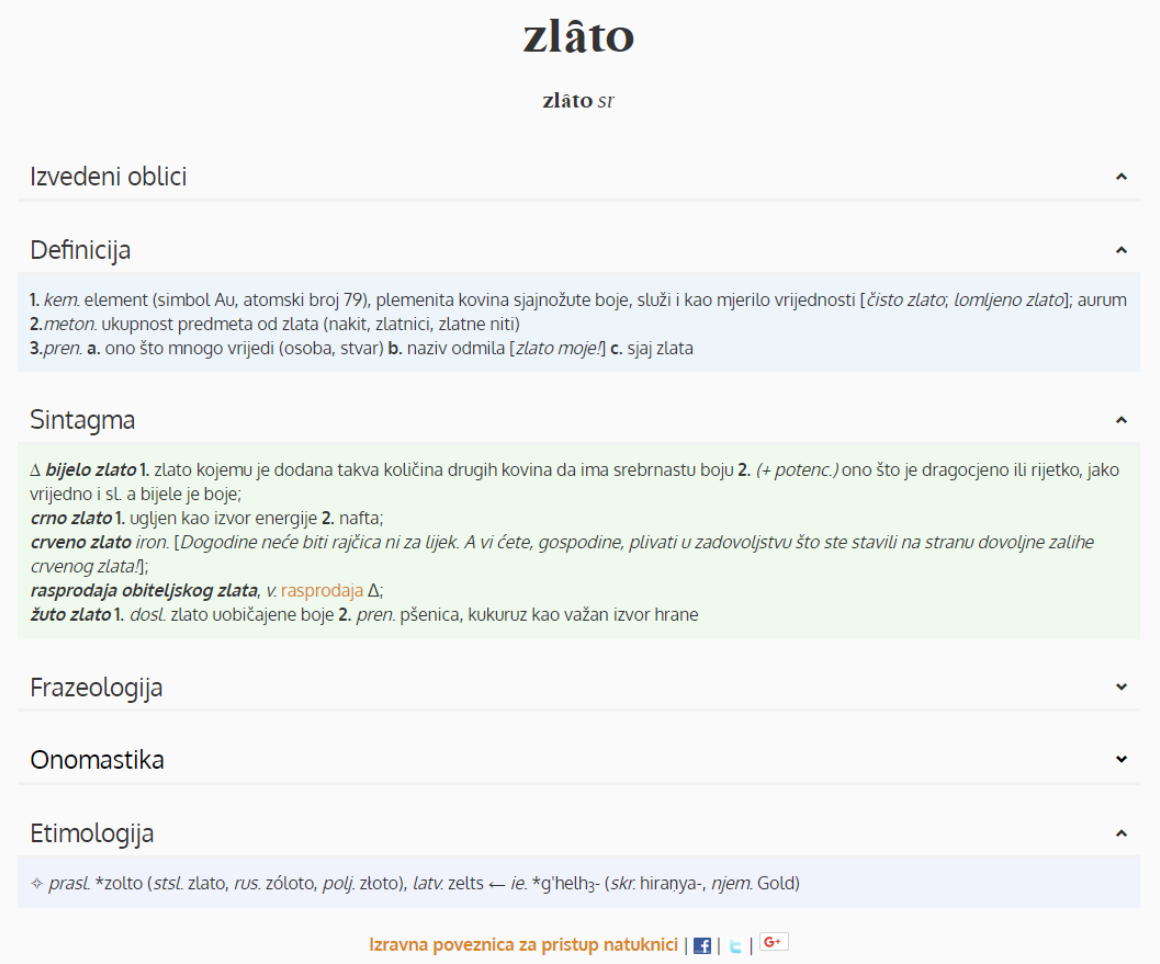
[30] http://www.feb-web.ru/feb/mas/mas-abc/default.asp (приступљено 18.07.2021)



Слика 5 Речнички чланак “ра́достно” из *Речника руског језика*

Хрватски језични портал, *ХЈП*[31] постоји од 2006. године и развијен је заједничким пројектом издавачке куће Знање и Свеучилишног рачунског центра Срце. ХЈП садржи око 117.000 одредница, које су преузете из 6 различитих речника хрватског језика. Овај портал приказује информације из речничког чланка: одредницу и врсту речи којој припада, граматичке информације, домен употребе речи, дефиниција, примери, синоними, синтагме, фразеологија, упутнице, ономастика и етимологија. Делови речничког чланка су у приказу обојени различитим бојама што их чини уочљивијим (слика 6). Коришћењем посебног дугмета у HTML презентацији могуће је уклонити сваки појединачни сегмент из визуелног приказа. Корисницима је на располагању и испис свих флективних облика за око 73.000 одредница.

[31] http://hjp.znanje.hr/ (приступљено 18.07.2021)

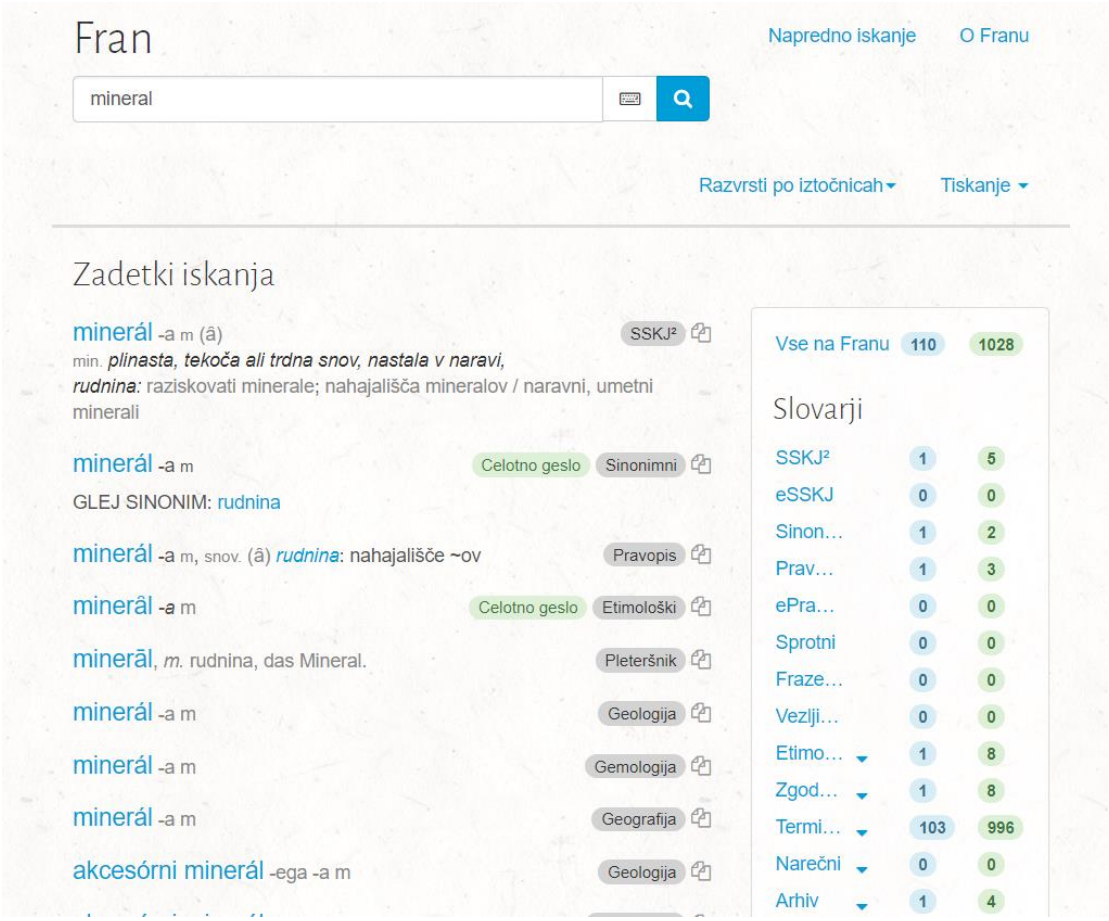


Слика 6 Приказ речничког чланка за реч „*zlato*“ на *Хрватском језичном порталу*

Након дигиталне верзије *Речника словеначког књижевног језика* (слов. *Slovar slovenskega knjižnega jezika*) Института за словеначки језик Франа Рамовша (слов. *Inštitut za slovenski jezik Frana Ramovša*), који је давао могућност претраге речника укрштањем различитих критеријума и употребом релационих оператора („==“ једнако или „!=“ различито) и семантичких ознака, настао је нови речнички портал *Фран* (слов. *Fran*)[32]. Овај портал на једном месту омогућава приступ речницима које је издао или и даље издаје Институт за словеначки језик Фран Рамовш. Доступно је преко 30 речника, неколико језичких атласа и водича са укупно преко 590.000 речничких чланака (Ledinek, Ahačič, и Perdih 2015). Ови извори су формално подељени у 7 група (општи, етимолошки, историјски, терминолошки и дијалекатски речници, корпуси и као једна група граматички и правописни приручници) што није згодно за претрагу јер корисник, уколико није другачије дефинисао, као резултате добија одговарајуће записе из свих извора. У менију датом са десне стране екрана за претрагу скраћено су дати сви извори и број појављивања тражене речи у њима. Резултати се приказују у облику речничког чланка из традиционалног речника уз информацију из ког тачно извора чланак потиче и могућности извоза цитата самог извора (слика 7). Постоји и дугме које покреће копирање речничког чланка у виду текста, а корисник може и да унесе предлоге за корекцију. Поред доброг дизајна и могућности које портал *Фран* пружа, треба напоменути да он, за разлику од већ поменутих речника и портала, омогућава

[32] https://fran.si/ (приступљено 18.07.2021)

комуникацију са корисницима, пошто корисници могу да додају коментаре који садрже предлоге корекција. Осим тога, на порталу су и детаљна упутства за употребу која су креирана према корисничкој циљној групи. Како на порталу постоји предлог вежби за укључивање портала *Фран* у наставу, може се закључити да постоји јака стратегија за промоцију портала, језичких и лексикографских ресурса, а самим тим и рада Института за словеначки језик, међу децом школског узраста и њиховим наставницима.



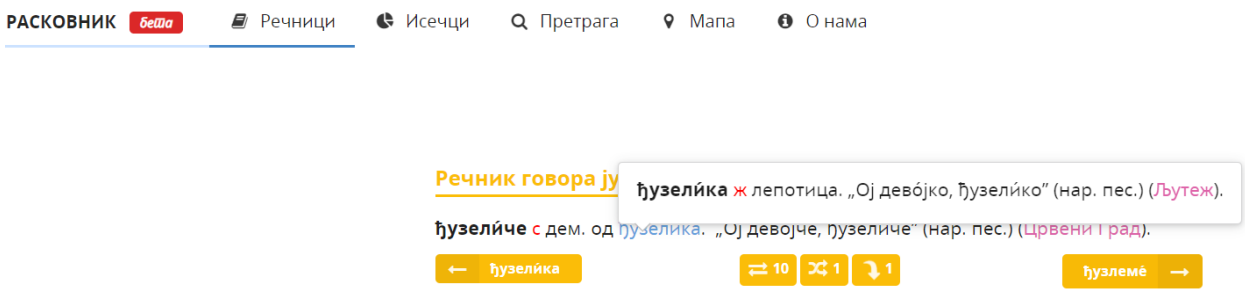
Слика 7 Приказ речничког чланка за реч „mineral“на порталу *Фран*

Расковник, српски лексикографски портал и платформа за упоредна истраживања српске лексике[33] развио је Институт за српски језик САНУ у сарадњи са Центром за дигиталне хуманистичке науке, а уз помоћ Министарства културе и информисања Републике Србије. Ова платформа пружа приступ дигитализованим верзијама више речника. У тренутку писања ове дисертације је доступно 5 речника српског језика: *Српски рјечник* Вука Стефановића Караџића (46.910 одредница), *Речник косовско-метохиског дијалекта* Глише Елезовића (13.556 одредница), *Речник говора јужне Србије* Момчила Златановића (9.178 одредница), *Рјечник говора Прошћења* Милоша Вујичића (5.213 одредница) и *Рјечник дубровачког говора* Михаила Бојанића и Растиславе Тривунац (9.940 одредница). Претрага је омогућена обједињено или по појединачним речницима и то као инстант претрага по лемама у виду приказа понуђених резултата током куцања задате ниске у поље за претрагу, претрага леме, дефиниције или примера коришћењем специјалних карактера за потребе проналаска различитих облика речи. На слици 8 је дат приказ речничког чланка „ђузели́че“ из *Речника говора јужне Србије*. На слици је у облику искачућег облака приказан инстант приказ речничког чланка на који

[33] http://raskovnik.org/ (приступљено 18.07.2021)

се упућује (ђузели́ка). Називи места (Љутеж и Црвени Град) који у датом примеру говоре о местима у којима су забележене речи дати су заградама и воде до ознаке места обележене на мапи. На мапи је могуће видети све речи прикупљене у датом месту.

Речник говора јужне Србије Момчила Златановића је од 2013. године био доступан и путем веб-апликације *Дигитални речник говора југа Србије*[34] која поред лингвистичких информација пружа и изговор одредница, примере употребе речи, графичке приказе локација уз помоћ сервиса *GoogleMaps* и *Geocoding*, претрагу по различитим деловима речничког чланка, итд. (Младеновић 2014).



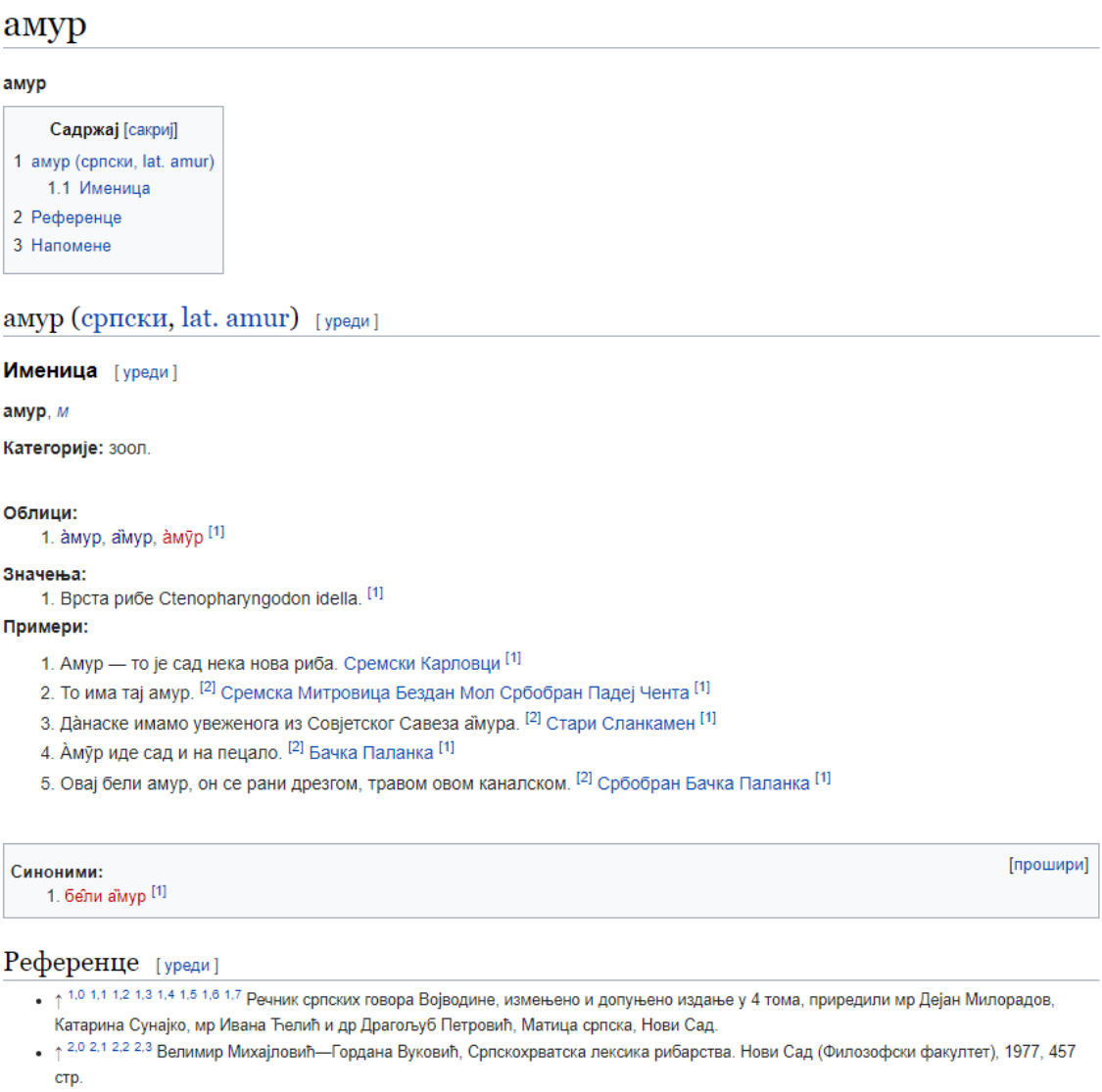
Слика 8 Приказ речничког чланка за реч „ђузели́че"на платформи *Расковник*

Платформа *Расковник* у секцији „Исечци" пружа приступ тематски издвојеним речничким чланцима. На пример, исечак *Врањанске пословице* обједињује приступ речничким чланцима који садрже пословице у *Речнику говора јужне Србије*.

Викиречник[35] (енг. *Wiktionary*) је пројекат Фондације Викимедија (енг. *Wikimedia Foundation*) покренут 2002. године доступан за преко 150 језика и представља речник на вебу. Сасвим очекивано, језик са највише речничких чланака је енглески језик, потом следе француски, малгашки, руски, кинески, немачки језик, итд. Као и већина онлајн речника, *Викиречник* омогућава претрагу одредница путем поља за претрагу али и прелиставање речника по алфабетском распореду. Речнички чланци, осим уобичајених текстуалних информација, могу садржати и звучни и графички садржај. Могуће је приступити речничким чланцима за преводе на друге језике, уколико они постоје. Оно што је карактеристично за овај речник јесте волонтерски рад корисника платформе на његовој доградњи. У оквиру речничких чланака су најчешће представљени подаци пореклом из других традиционалних речника. На слици 9 је представљен речнички чланак за реч „амур" из *Викиречника* српског језика. Уз значење и примере употребе речи стоје ознаке бројева референци (речника и речничке грађе) из којих су они преузети. Делови речничког чланка приказани у виду хипервеза представљају упућенице ка другим речничким чланцима из речника. Оно што јесте мана овог речника јесте неуједначеност структуре речничких чланака.

[34] http://www.vranje.co.rs/Default.aspx#sthash.ls8dR3nw.dpbs (приступљено 18.07.2021)

[35] https://www.wiktionary.org/ (приступљено 20.07.2021)



амур

амур

Садржај [сакриј]
1 амур (српски, lat. amur)
1.1 Именица
2 Референце
3 Напомене

амур (српски, lat. amur) [уреди]

Именица [уреди]

амур, *м*

Категорије: зоол.

Облици:
1. а̀мур, а̏мур, а̀мӯр [1]

Значења:
1. Врста рибе Ctenopharyngodon idella. [1]

Примери:
1. Амур — то је сад нека нова риба. Сремски Карловци [1]
2. То има тај амур. [2] Сремска Митровица Бездан Мол Србобран Падеј Чента [1]
3. Дà нaске имамо увеженога из Совјетског Савеза а̏мура. [2] Стари Сланкамен [1]
4. Àмӯр иде сад и на пецало. [2] Бачка Паланка [1]
5. Овај бели амур, он се рани дрезгом, травом овом каналском. [2] Србобран Бачка Паланка [1]

Синоними: [прошири]
1. бе̏ли а̏мур [1]

Референце [уреди]

- ↑ 1,0 1,1 1,2 1,3 1,4 1,5 1,6 1,7 Речник српских говора Војводине, измењено и допуњено издање у 4 тома, приредили мр Дејан Милорадов, Катарина Сунајко, мр Ивана Ћелић и др Драгољуб Петровић, Матица српска, Нови Сад.
- ↑ 2,0 2,1 2,2 2,3 Велимир Михајловић—Гордана Вуковић, Српскохрватска лексика рибарства. Нови Сад (Филозофски факултет), 1977, 457 стр.

Слика 9 Приказ речничког чланка за реч „амур“ у *Викиречнику* за српски језик

Већина речника и портала које смо анализирали су постали део *Портала за речнике* (енг. *Dictionary portal*), доступног путем адресе http://www.dictionaryportal.eu/en/. Овај *Портал* је плод рада COST[36] акције Европска мрежа за е-лексикографију (енг. *European Network of e-Lexicography, ENeL*). Битно је напоменути да су сви представљени речници и портали намењени људима, али већина њих је заснована на алатима и ресурсима за обраду природних језика, рачунарске и корпусне лингвистике.

2.3.3 Информатичка подршка изради речника

Лексикографска база

База података представља колекцију међусобно повезаних података који су организовани на униформан начин што олакшава да их користе различити програми. Један од начина организације података јесу табеле (код релационих база података). Табела може да има више колона где свака колона представља неку особину или атрибут. Једну базу података може користити низ различитих програма писаних у

[36] European Cooperation in Science and Technology

различитим програмским језицима. Базе података као највећу предност доносе интегритет података који се односи на прецизност, пуноважност и коректност података у бази (Mitić 2021). Поред интегритета коришћење базе података доноси и олакшано претраживање, могућност извоза података у различитим форматима према потребама корисника (било да су они апликације или људи), лакша проширења базе новим подацима, итд. Један од примера употребе база података је организовање и употреба лексикографских података и такву базу ћемо звати лексикографска база података.

Лексикографска база је структурирана колекција лексикографских података (нпр. лема, домен припадности, ознака флективне класе, URL адреса, итд.). Она најчешће има више поља у односу на број поља који лексикограф предвиди јер се различите лексикографске информације у пракси могу репрезентовати уз коришћење више потпоља. На пример, лексикограф може да захтева поље за веб-адресу које ће у бази података бити представљено кроз два потпоља, једно за краћи назив који ће бити приказан у речничком чланку као хипервеза за праву веб-адресу, а друго потпоље за целовиту адресу (Bergenholtz и Nielsen 2013).

Лексикографска база пружа низ нових могућности као што су успостављање и одржавање нових релација међу лексичким јединицама, допуна постојећег речника новим информацијама и новим лексичким јединицама, постављање различитих сложених упита над базом, а самим тим и екстракције жељених информација из базе за потребе производње различитих облика речника. Лексикографска база омогућава и поновну употребу лексикографских података што доприноси убрзаном раду на речничким пројектима.

Подаци из лексикографске базе могу се користити као материјал за креирање речника намењених човеку, било штампаних, било дигиталних, али и за креирање речника намењених за употребу од стране рачунара, односно речника за потребе обраде природних језика (енг. *Natural language processing, NLP*).

Системи за писање речника

Оно што је готово незаобилазан део настајања електронског речника изворно објављеног у дигиталном формату јесу програми који пружају подршку у развоју речника - системи за писање речника (енг. *Dictionary writing system - DWS*), лексикографске радне станице (енг. *lexicographic workbench*), програми за компилацију речника или системи за уређење речника (De Schryver 2011, Pala и Horák 2006). Овакви системи су намењени тимовима који производе речнике, издавачким кућама које издају речнике најразличитијих типова, али и истраживачима. Упитник реализован у оквиру ENeL Радне групе за иновативне електронске речнике, са циљем испитивања система за писање речника и постављање упита над корпусима, а који су испуниле институције које се баве лексикографијом из 25 земаља Европе, даје преглед типова и карактеристика система који се користе у ове сврхе (Krek, Abel, and Tiberius 2014). Упитник показује да 70% институција користи неки систем за писање речника. Пракса показује да иако постоји доста развијеног софтвера, од којих ће неки бити овде представљени, нови пројекат најчешће изискује нови прилагођени систем.

Неке од предности коришћења система за писање речника су убрзање процеса обраде лексикографских података, олакшан процес редакције речника, подршка тимском раду и практична контрола структуре речника (De Schryver 2011). Како бисмо илустровали могућности система за писање речника, настојаћемо да опишемо неке од функција два система: TLex и Lexonomy.

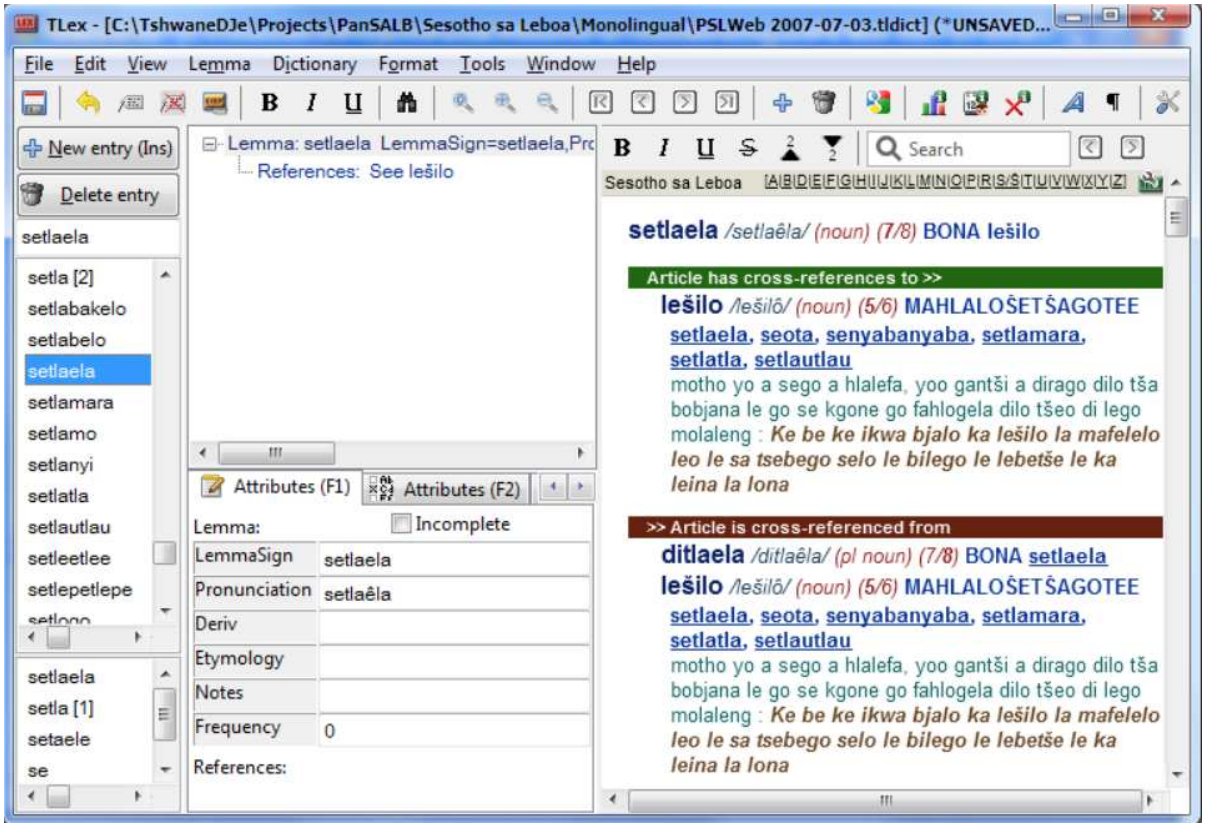
Систем за писање речника TLex[37], познат и под именом TshwaneLex[38] [изговор: чванелекс], је комерцијални софтвер који се развија од 2002. године и служи за писање једнојезичних, двојезичних и вишејезичних речника. Настао је у Јужноафричкој републици па се прво користио за речнике различитих афричких језика и енглеског језика (махом за двојезичне речнике). Подржава Unicode стандард те се може користити за све светске језике („TshwaneDJe Software: TLex Lexicography, Terminology and Corpus Software“ 2018). Потпун пакет система TLex укључује четири повезане апликације: систем за компилацију речника TLex (енг. *TLex Dictionary Compilation Software*), систем за руковање базом термина tlTerm (енг. *tlTerm Professional Termbase Software*), софтвер за израду конкорданци tlCorpus (енг. *tlCorpus Concordance Software*), као и апликацију за визуелни приказ речника писаних помоћу ових апликација - tlReader (енг. *tlReader*). Овај програм је доступан у виду апликације која се инсталира на стоне персоналне рачунаре.

TLex омогућава кориснику који креира речник да у реалном времену, у прозору који даје приказ речника намењен крајњем кориснику речника, види историја свих промена (унос нових података и измене) које прави при писању речника. Омогућено је аутоматско повезивање упутница као и приказ упућеница из речника, односно у претходно поменутом прозору за приказ промена се могу видети лексички запис који упућује и лексички запис на који се упућује (слика 10). Могућ је и приказ упутнице и на појединачно значење, уколико се упутница односи на посебно значење. Уз то треба напоменути да се хомографи нумеришу аутоматски и да је омогућено прилагођавање означавања различитих делова речничког чланка. TLex пружа могућности додавања илустрација, звука и видео записа у речник. Такође је омогућено поређење база података различитих верзија речника. Уз помоћ едитора за DTD (енг. *Document Type Definition*)[39] TLex пружа могућност дефинисања елемената који ће се користити у сваком појединачном речничком пројекту као и начина њиховог коришћења. Захваљујући употреби базе података, омогућен је истовремени рад више корисника без бојазни да ће оно што је један корисник направио бити обрисано. Верзија програма TLex објављена 2018. године омогућава корисницима да, поред уметања примера употребе уз помоћ апликације tlCorpus, додају примере и из Excel документа. На слици 10 је приказан пример из речника језика северни сото. Лема на којој корисник ради је „setlaela“ у чланку те одреднице се упућује на „lešilo“ а из чланка „ditlaela“ се упућује на текућу одредницу „setlaela“.

[37] https://tshwanedje.com (приступљено 10.12.2018)

[38] Чване (Tshwane) је традиционални афрички назив за град Преторија, родни град Дејвида Џофија (David Joffe), коаутора система Tlex.

[39] DTD је скуп декларација за обележавање које дефинишу тип документа. Он прецизно декларише који елементи и референце могу да се појаве у документу, и на ком месту, као и који су садржај и атрибути ових елемената. Користи се за дефинисање XML докумената.



Слика 10 Приказ упутница у склопу TLex-а (Tshwanedje 2020)

Апликација tlCorpus омогућава да примери добијени претрагом корпуса буду аутоматски увезени у лексички запис који се уређује кроз апликацију TLex, а подржано је и приказивање колокација речи. Сам корпус може бити у различитим форматима: txt, HTML, doc, docx, RTF (енг. *Rich Text Format*), PDF документа па чак и у форматима за електронске књиге, као што су MOBI, EPUB и CHM, с тим што у том случају на рачунару мора бити инсталиран програм отвореног кода Calibre[40]. Ова апликација подржава документа величине до 4GB. Оно што је потребно напоменути јесте да tlCorpus омогућава аутоматско означавање врста речи из корпуса применом техника обраде природних језика јер су препознате предности коришћења анотираног корпуса у односу на сирови текст (De Schryver и De Pauw 2007).

Сегмент tlReader (енг. *tlReader*) омогућава визуелизацију речника креираних уз помоћ система TLex. Прикази се могу користити за израду штампаних речника, електронских речника доступних на вебу, као и електронских речника доступних на компакт диску.

Систем за писање речника TLex дозвољава извоз лексичких података у форматима HTML, CSV (енг. *Comma Separated Values*), RTF, XML, txt и кроз интерфејс за приступ бази података ODBC (енг. *Open Database Connectivity*).

Систем за писање речника Lexonomy[41] (Měchura 2017) је за разлику од претходно представљеног система TLex програм отвореног кода који је доступан као веб-апликација. Аутор напомиње да је систем Lexonomy погодан за писање малих

[40]Calibre је програм отвореног кода који служи за управљање колекцијом електронских књига, што подразумева и пребацивање електронских књига на уређај за читање, као и конверзију електронске књиге у неки други формат.

[41] https://www.lexonomy.eu (приступљено 10.12.2018)

речника и речника средње величине. Вишекориснички рад се остварује тако што један корисник креира речник да би потом са одређеним корисницима поделио приступ уређивању. Аутор речника има могућност дефинисања подешавања на нивоу самог речника (енг. *Dictionary settings*), лексичког записа (енг. *Entry settings*) као и подешавања верзије речника за објављивање (енг. *Publishing*). На нивоу подешавања речника, аутор може мењати назив и опис речника, управљати корисничким привилегијама, којима се одређује који корисник може да уређује речник, конфигурише, извози или увози лексичке записе у речник, или управља едитором лексичких записа, у смислу одабира лаичког (енг. *laic mode*) или „напредног" режима (енг. *nerd mode*) рада. У лаичком режиму се етикете приказују као графичке опције, док се у напредном режиму етикете и структура лексичких записа приказују као XML етикете.

На нивоу управљања лексичким записом, корисник може да прави структуру лексичког записа дефинисањем елемената и атрибута који се могу користити унутар лексичког записа. Садржај елемената могу да буду други елементи, текст, текст са уметнутим елементима, вредност из задате листе, али садржај елемената може бити и празан што аутору речника даје велику слободу у осмишљавању структуре речника. У делу управљања лексичким записом је могуће дефинисати делове записа који ће бити такозвани подзаписи (енг. *subentry*). Ова опција је посебно згодна за дељење исте дефиниције у више лексичких записа или представљање компонената полилексемских израза (Měchura 2018). Такође се може подесити који елементи лексичког записа ће бити претраживи кроз поље за претрагу.

У подешавању верзије речника за објављивање могуће је форматирати сваки елемент речничког чланка, односно подесити његов визуелни приказ. Поред тога, могуће је одабрати једну од две опције за објављивање: јавно доступан или заштићен. Поред три поменуте врсте подешавања, могуће је извршити подешавања спољашњих извора података (енг. *external data sources*). Ово се првенствено односи на означавање делова лексичког записа који се могу уметати из спољашњих извора, какви су примери колокације или дефиниције. Lexonomy предвиђа коришћење корпуса који су доступни путем алата Sketch Engine па је предефинисана стаза за повезивање са спољашњим извором адреса сучеља за програмирање апликација[42] (енг. *application programming interface, API*) алата Sketch Engine. Из Sketch Engine-а је, с друге стране, могуће директно у Lexonomy уметнути дефиницију или колокације уз коришћење опције „добри речнички примери" (енг. *good dictionary examples, GDEX*)[43]. Lexonomy чува лексичке податке у XML формату а нуди могућност извоза у истом формату. Систем за писање речника Lexonomy и алат за управљање корпусима Sketch Engine заједно чине инфраструктуру за лексикографију коју промовише пројекат Европска лексикографска инфраструктура (енг. *European Lexicographic Infrastructure, ELEXIS*) (Jakubíček и остали 2018).

[42] Сучеље за програмирање апликација, апликациони програмски интерфејс или, како је још заступљено у терминологији на српском језику, програмски интерфејс апликације јесте интерфејс који дефинише на који начин програми могу да захтевају услуге других програма или оперативних система. Више се може прочитати на: https://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975-e-160?rskey=o43LEG&result=201 (приступљено 10.04.2020)

[43] Опција GDEX представља систем за евалуацију реченица у смислу њихове адекватности за употребу у својству значењских потврда у речнику. Неки од критеријума који се узимају у обзир приликом проналажења примера таквих реченица јесу дужина реченице, комплексност вокабулара, контекст, итд. (Kilgarriff и остали 2008). У раду (Stanković и остали 2019) је дата анализа примера из Речника САНУ у циљу израде модела за GDEX за српски језик.

3. Језички ресурси за српски језик

3.1 Морфолошки речници

Морфолошки речници српског језика (Krstev 2008) су електронски речници намењени превасходно употреби у рачунарским апликацијама које се баве обрадом природних језика. Они су значајан ресурс за језике са богатом флексијом, какав је и српски. Почивају на формату DELA (фр. *Dictionnaires électroniques du LADL*) који је развијен у лабораторији LADL (*Laboratoire d'Automatique Documentaire et Linguistique)* под руководством професора Мориса Гроса (Maurice Gross). Први речник овог типа развијен је за француски језик. Речници у формату DELA развијени су у оквиру мреже RELEX[44] и за друге језике од којих су неки енглески, немачки, грчки, италијански, латински, шпански итд. Од словенских језика, ови речници, осим за српски језик, постоје и за бугарски, пољски и руски језик (Krstev и остали 2009). Рад на утемељењу, развоју и примени Морфолошких речника за српски језик започет је научно-истраживачким радом професора Душка Витаса (Vitas 1993) и професорке Цветане Крстев (Krstev 1997), који и данас руководе развојем и допуном речника.

Морфолошки речници DELA се састоје од речника монолексемских јединица (енг. *simple words*) и речника полилексемских јединица (сложених речи – енг. *compounds* или вишечланих јединица или израза - енг. *multiword units, MWU* или енг. *multiword expressions, MWE*) (Paumier 2016). С обзиром на то да систем *Морфолошких речника за српски језик* почива на теорији коначних аутомата (Vitas 2006), он је заснован на морфолошким и локалним граматикама представљеним коначним трансдукторима (енг. *Finite State Transducer, FST*) којима се дефинишу и генеришу сви флективни облици у речницима.

Основне компоненте система морфолошких речника монолексемских речи су DELAS (фр. *DELA de formes simple*) и DELAF (фр. *DELA de formes Fléchies*)(Krstev 1997). Компонента DELAS се састоји од леме речи и флективног, семантичког и синтаксичког обележја распоређених према следећој структури:

```
lema,VrstaReči#fst[+Marker]*
```

Облик леме је најчешће у складу са лексикографском праксом, што значи да је за именице лема облик номинатива једнине, за придеве номинатива једнине мушког рода, неодређеног вида, за глаголе инфинитив и за непроменљиве врсте речи сама реч. Облик леме не прати увек слепо лексикографску праксу. На пример, у Предговору шестотомном *Речнику српскохрватскога књижевног језика* Матице српске наведено је да су демоними давани у облику множине (*Речник српскохрватскога књижевног језика, Књ.1, А-Е* 1967). Облик једнине је обично наведен у оквиру речничког чланка одреднице која се води под множину (*Немци, Срби , Албанци, Французи)*. У пракси се понекад јављају речнички чланци са одредницом само у једнини (*Италијан, Италијанка* или *Шпанац*, *Шпањолац, Шпанкиња, Шпањолка*) али и речнички чланци, који дефинишу припаднике истог народа, са одредницама и у множини и у једнини (*Швеђани*, *Швеђанка* или *Турци*, *Турчин*, *Туркиња*). Демоними су у *Морфолошким речницима* представљени обликом номинатива једнине јер за ову врсту речника није од значаја како би људи тражили неку реч у речнику (Крстев 2019). Речи су распоређене у класе према традиционалној подели у граматикама српског језика (Станојчић и Поповић 2020). Врста речи може имати једну од следећих вредности: N (енг. *noun*) за именицу, PREP (енг. *preposition*) за предлог, CONJ

[44] Међународну мрежу RELEX чини више лабораторија које се баве рачунарском лингвистиком. Ова мрежа је настала на иницијативу Мориса Гроса и његових сарадника из лабораторије LADL. Више о овој Мрежи може се прочитати на: http://infolingu.univ-mlv.fr/english/Relex/Relex.html (приступљено 10.09.2018)

(енг. *conjunction*) за везник, PAR (енг. *particle*) за речцу, INT (енг. *interjection*) за узвик, NUM (енг. *numeral*) за број, PRO (енг. *pronoun*) за заменицу, A (енг. *adjective*) за придев, ADV (енг. *adverb*) за прилог, V (енг. *verb*) за глагол. Ознака за врсту речи праћена је ознаком FST, односно трансдуктора коначних стања који дефинише флективну класу. Ознака флективне класе даје информацију о флективној парадигми према којој се мења дата реч. Свака флективна парадигма је описана на недвосмислен начин придруживањем ознаке могућој комбинацији скупа флективних наставака. Потом следе најразличитији маркери који се могу сврстати у доменске (нпр. DOM=Geol за геологију), семантичке (нпр. Prof за занимање), синтаксичке (нпр. p7 за предлог који захтева локатив), деривационе (нпр. VN који означава глаголску именицу), изговорне (нпр. Ijk за ијекавски изговор), као и ознаке различитих информација као што су ознаке земаља (нпр. CC2=BR за Бразил), мерних јединица (нпр. SI=cm за центиметар), итд. Маркери се у лексичком запису могу јавити опционо или више пута и набрајају се коришћењем сепаратора „+“.

Следи пример записа из речника монолексемских јединица:

```
volga,N600+Conc+Vehicle+Erg+DOM=Transport
```

Облик „volga“ представља лему, потом иза запете следе ознака флективне класе N600 што значи да се ради о именици треће врсте са немаркираним наставцима. Сематички маркер Conc означава да се ради о конкретној именици, маркер Vehicle означава да се ради о возилу, Erg да се ради о производу – заштићеном имену (ергониму), док домен DOM=Transport означава да термин припада домену саобраћаја. Дакле, ради се о запису који описује возило волга. Запис који описује истоимену руску реку изгледа овако:

```
Volga,N623+NProp+Top+Hyd+River+CC2=RU
```

Дакле, „Volga“ је лема, N623 флективна класа ове именице (мења се исто као volga али нема множину), NProp маркер властитог имена, Top је општи маркер геополитичког имена, Hyd ознака за водену површину, River ознака за реку и најзад ознака `CC2=RU` ознака за Русију.

Леме са придруженим флективним кодом, т.ј. флективним трансдуктором из речника DELAS, омогућавају да се аутоматски генеришу одреднице у речнику DELAF. Записи речника DELAF састоје се од облика речи, леме, ознаке врсте речи и граматичких категорија у следећем формату:

```
oblikReči,lema.VrstaReči:[gramatička_kategorija]*
```

Граматичке категорије се јављају код променљивих врста речи те знак „*“ показује да су ознаке граматичких категорија опционе или се јављају више пута. Свака граматичка категорија представљена је једним јединственим карактером без коришћења међусобних сепаратора у синтакси а редослед њиховог навођења није од значаја. Као пример записа из речника DELAF послужиће нам један флективни облик речи волга:

```
volgu,volga.N+Conc+Vehicle+Erg+DOM=Transport:fs4q
```

Дакле, запис описује облик „volgu“, чија је лема „volga“ и врста речи именица означена кодом N. Потом следе маркери из речника DELAS. Иза знака „:“ следе граматичке категорије f за женски род, s за једнину као ознаку броја, 4 као ознака за падеж акузатив и q као ознака анимантности која показује да је реч неаниматна. Речник DELAF се генерише аутоматски из речника DELAS и флективних аутомата.

Основне компоненте система морфолошких речника полилексемских речи су DELAC (фр. *DELA de formes composés*) и DELACF (фр. *DELA de formes Composées Fléchies*). Следи запис за полилексемску јединицу „златна грозница“ из речника DELAC:

```
zlatna(zlatan.A18:aefs1g) groznica(groznica.N650:fs1q),NC_AXN3+Fig+Comp
```

Ова лема се састоји из две компоненте „zlatna“ и „groznica“. У пратећим заградама су дати описи флективног облика који формирају ову полилексемску јединицу у датом канонском облику (номинатив једине). Дакле, облик „zlatna“ део је флективне парадигме леме придева „zlatan“ који припада флективној класи A18 и представља позитив (a), придевског вида који у овој флективној поарадигми није од значаја (e), женског рода (f), једнине (s), у номинативу (1) и без обележја аниматности (g). Компонента „groznica“ представља именицу „groznica“ која се мења у складу са флективном парадигмом N650 и представља женски род (f) једнине (s) номинатива (1) без обележја аниматности (q). Иза леме праћене знаком „ ,“следи ознака флективне класе полилексемског израза NC_AXN3. Ова флективна класа означава да се полилексемска реч састоји од именице којој претходи придев који се слаже са именицом у роду, броју, падежу и обележју аниматности, док се граматички број полилексемске речи не мења, односно остаје исти као у леми. Потом следе два семантичка маркера која показују да лексички запис има фигуративно значење (Fig) и да је у питању полилексемска јединица (Comp).

У наставку следи запис из речника DELACF који описује један флективни облик за полилексемску јединицу „златна грозница“ која је претходно представљена у речнику DELAC:

```
zlatnu groznicu,zlatna(zlatan.A18:aefs1g) groznica(groznica.N650:fs1q).NC:fs4q
```

Ради се о облику полилексемске именице (NC) „zlatnu groznicu“ који је женског рода (f) у једнини (s), акузативу (4) и неаниматан (q). На овом примеру се види да је полилексемска јединица попримила флективно понашање друге компоненте, именице „groznica.“ Записи речника DELACF се генеришу аутоматски уз помоћ речника DELAC и флективних аутомата монолексемских и полилексемских јединица.

Према подацима из јула 2021. године, речник монолексемских јединица се састоји од 205.003 леме, са 207.312 значења, док је број полилексемских јединица обухваћених речником 22.865, са 23.014 значења. Најзаступљеније врсте речи су именице (116.192 лема), придеви (64.274 лема) и глаголи (21.159 лема).

Одржавање морфолошких речника је првобитно реализовано кроз подсистем радне станице за управљање лексичким ресурсима WS4LR (енг. *Work Station for Lexical Resources*) (Krstev и остали 2006), развијене у оквиру докторске дисертације професорке Ранке Станковић (Stanković 2009), која је касније прерасла у софтверски алат LeXimir. Могућности које LeXimir пружа (Stanković и остали 2011) су дистрибуција речника у више датотека, претраживање и издвајање подскупова лема на основу различитих критеријума који су саставни део DELA формата. Омогућена је веза са коначним трансдукторима и регуларним изразима који описују флексију дате леме. Ово је корисно из два разлога. Први је прегледање и кориговање флективних трансдуктора, уколико за тиме има потребе. Други је могућност генерисања свих облика нове леме што је значајно за проверу одабира кода флективне класе. LeXimir има и специјалне могућности за рад са речницима у формату DELAC, а пре свега формирање речника DELAC на основу листе полилексемских јединица, што је омогућено моделом који предвиђа исправну флективну класу полилексемске јединице као и облике њених компонената (Krstev и Vitas 2009).

Примена морфолошких речника српског језика је вишеструка, почев од основних задатака обраде текста коришћењем система Unitex, кроз постављање различитих сложених упита регуларним изразима или графовима како би се из текста екстраховали различити подаци или да би се обавила нека сложена трансформација текста. Речници се примењују и при различитим задацима од који су неки аутоматско препознавање термина у различитим доменима (Krstev и остали 2015), препознавање именованих ентитета (Krstev и остали 2014), аутоматско препознавање временских израза (Jaćimović 2016), аутоматска обрада правних текстова и аутоматско успостављање упутнице (Васиљевић 2014), екстракција информација из доменских текстова (Vujičić Stanković 2016), претрага информација у дигиталним библиотекама (Тртовац 2016) (Томашевић 2018), корекције текста враћањем дијакритичких знакова (Krstev, Stanković, и Vitas 2018), итд.

3.2 Ворднет

Праву револуцију у свету моделирања лексикографије изазвала је појава ворднета (Miller и остали 1990). Вишејезична семантичка мрежа *Ворднет* представља скуп синсетова и релација између њих (Fellbaum 1998). Синсет је скуп речи које у неком контексту имају исто или приближно значење. Релације које је могуће изразити преко ворднета су бројне. Неке од њих су: подређеност и надређеност, антонимија, синонимија, „бити у стању нечега“ (битна за повезивање именичких и придевских синсетова). Први и највећи ворднет је *Принстонски ворднет* (енг. *Princeton Wordnet of English*, *PWN*) за енглески језик који је послужио као полазна тачка за развој сличних ресурса, било појединачно, било кроз пројекте попут *EuroWordNet*-а[45] или *BalkaNet*-а[46]. Пројекат *EuroWordNet* започет 1996. године резултовао је вишејезичком лексичком базом ворднетова за осам европских земаља.

Изградња ворднета за српски језик започета је у склопу пројекта *BalkaNet*. По завршетку пројекта, 2004. године, база података садржала је 7.000 синсетова повезаних са одговарајућим синсетовима балканских језика и синсетовима из Принстонског ворднета. То су синсетови који репрезентују исте концепте на разним језицима. Од 2006. године отпочело се са кооперативним волонтерским радом на доградњи Српског ворднета синсетовима из различитих домена. Том приликом додати су синсетови из домена биологије, биомедицине, религије, права, лингвистике, библиотекарства, књижевности, рачунарства и кулинарства, као и синсетови којима се изражавају осећања (Крстев и остали 2008). Српски ворднет је у годинама које су уследиле доживео бројна побољшања. Кроз CESAR[47] (енг. *Central and South-East European Resources*) пројекат је описан метаподацима и постао доступан путем META-SHARE[48] репозиторијума под условима CC-BY-NC лиценце. Потом је сачињено стабилно окружење за руковање и доградњу. Са првобитно коришћеног алата за одржавање, VisDic[49], се прешло на апликацију развијену за рад на вебу. Ова апликација је надоместила недостатке који су били приметни током одржавања алатом VisDic. Омогућени су истовремени рад више корисника, напредна претрага, као и потпуна XML подршка. Такође је развијено и сучеље за корисничку претрагу и доградњу ворднета за српски језик (Mladenović, Mitrović, и

[45] http://projects.illc.uva.nl/EuroWordNet/ (приступљено 10.07.2018)

[46] http://www.dblab.upatras.gr/balkanet/ (приступљено 10.07.2018)

[47] http://www.meta-net.eu/projects/cesar/ (приступљено 10.07.2018)

[48] http://www.meta-share.org/ (приступљено 10.07.2018)

[49] https://nlp.fi.muni.cz/projects/visdic/ (приступљено 10.10.2018)

Krstev 2014). *Српски Ворднет* је доступан за преглед преко апликације за веб BulNet[50] и, према подацима из септембра 2021. године, садржи 22.571 синсет, од којих 18.276 синсета представља именице, 2.249 глаголе, 1.920 придеве и 126 прилоге. (Stanković, Mladenović, и остали 2018). Укупан број литерала је 37.458 а различитих литерала 33.035.

3.3 Корпуси

Корпус је лексички ресурс који подразумева колекцију текстова. У данашње време се првенствено мисли на текстове у електронској форми који представљају репрезентативни узорак појединачног језика или језичког варијетета (Xiao 2010) према критеријумима као што су: тема, величина, функционални стил и временски оквир настанка (Томашевић 2018). Ипак, корпуси нису увек били електронски. Ако се ограничимо на српски језик, први неелектронски корпус српског језика, дијахронијски корпус српског језика Ђорђа Костића, развијан је у периоду од 1957. до 1962. године и садржи око 11 милиона речи из текстова насталих од XII века до 1962. године (Utvić 2013). Овај корпус је дигитализован и данас се користи у сврхе квантитативног описа српског језика (Đorđević 2017).

Корпус савременог српског језика (скр. *СрпКор*) почиње да се развија од 1981. године иницијативом професора Душка Витаса (Vitas 1981). Детаљан приказ развоја *СрпКор*-а дат је у докторској дисертацији Милоша Утвића (Utvić 2013). Прва верзија овог корпуса, публикована на вебу 2003. године (Vitas и Krstev 2012), чију је садржину и структуру одредио професор Витас, названа је *Неетикетирани корпус српског језика* (скр. НЕТК) или *СрпКор2003* и садржала је 22,2 милиона речи. Потом су уследиле још три верзије *СрпКор*-а, *СрпКор2011* (113 милиона речи), *СрпКор2012* (118 милиона речи) и *СрпКор2013* (122 милиона речи), које су аутоматски анотиране (врста речи и лема), допуњене библиографским информацијама о изворима и информацијама о функционалном стилу, као и информацији да ли је текст изворно настао на српском језику или је преведен (Utvić 2013). Верзија *СрпКор2013* обухвата текстове настале у временском интервалу од 1920. до 2013. године. Обухваћено је 4.890 текстова који припадају следећим функционалним стиловима: књижевно-уметничком, научном и научно-популарном, новинском, административном. Текстови који се нису могли сврстати ни у један од ових стилова смештени су у групу текстова са неодређеним функционалним стилом. Корпус је структурно анотиран у односу на смернице TEI-Lite P5[51] о којима ће бити више речи у одељку 4.1.1. Аутоматска морфолошка анотација Корпуса извршена је алатом TreeTagger[52] коришћењем ознака које се користе у Морфолошким речницима српског језика (Утвић 2011).

СрпКор је доступан за претрагу регистрованим корисницима путем адресе: http://www.korpus.matf.bg.ac.rs/korpus/login.php. *СрпКор* дозвољава једноставну претрагу, основним пољима за претрагу, и напредну претрагу која укључује могућности упитног језика CQP (енг. *Corpus Query Language*). Основна претрага даје могућност просте претраге текста корпуса без анотације и то коришћењем кода Аурора[53], као и задавање упита регуларним изразима. Резултати претраге се исписују у конкорданцама праћеним библиографском информацијом, и то 100 конкорданци по страни. Постоје могућности

[50] http://dcl.bas.bg/bulnet/ (приступљено 10.10.2018)

[51] http://www.tei-c.org/guidelines/customization/lite/ (приступљено 10.10.2018)

[52] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (приступљено 10.10.2018)

[53] Више о коду Аурора: http://www.korpus.matf.bg.ac.rs/prezentacija/uputstvo.html (приступљено 12.10.2018)

сортирања конкорданци по левом и десном контексту или подешавања дужине контекста. Могуће је и управљање конкорданцама у смислу одабира за приказ и њиховог чувања. Напредна претрага, уз могућности основне претраге, омогућава претрагу по морфолошкој анотацији, а то значи према леми или врсти речи.

У поређењу са другим корпусима српског језика *СрпКор* јесте на известан начин референтни корпус српског језика. У раду који се бави поређењем корпуса српског језика наведено је да *СрпКор* „располаже највећим бројем оних карактеристика које одликују данас модерне корпусе у свету. Доступан је преко интернета, омогућена су претраживања у поткорпусима, састоји се од релативно прихватљивог броја речи, располаже и етикетираним поткорпусима." (Urkom 2010). *СрпКор* је коришћен при бројним језичким истраживањима која су имала различите циљеве, између осталих и испитивање језичких промена у српском и хрватском језику насталих утицајем глобалних језика (Madsen 2017), испитивање начина богаћења српског језика (Драгићевић 2018), испитивање просторних фрагментизатора (Шапић 2018) или придевских атрибута уз поједине именице (Пантић 2017) или за разматрање грађе за морфолошко истраживање (Алексић 2017). Корпус је такође коришћен за потребе кодирања евиденцијалности у српском новинском дискурсу (Aleksić 2019), контрастивне анализе израза (Rajić 2015), испитивање метафоре (Jovanović 2019) или футура (Radovanovic 2017) у српском језику. *СрпКор* и његови различити поткорпуси су коришћени за развој и тестирање других језичких ресурса и алата за српски језик развијених у оквиру Друштва за језичке ресурсе и технологије. Поткорпуси се користе као референтни корпуси приликом одређивања кључних речи у специфичним доменским корпусима. Такође се користе и за обележавање именованих ентитета. *Морфолошки речници за српски језик*, детаљније описани у овом поглављу, делом су допуњени ексцерпцијом из *СрпКор*-а.

Развијено је више доменских корпуса српског језика (библиотекарство и информатика, кулинарство, итд.) а овде ћемо се посебно посветити корпусу рударских текстова. Старија верзија рударског корпуса названа *MineCorp* (Lazić и Stanković 2015) настала је 2015. године за потребе тестирања система за екстракцију термина (Stanković, Krstev, Obradović, Lazić, и остали 2015) и садржала је нешто више од 600.000 речи. Проширени *Рударски корпус - РудКор* формиран је као компонента система за управљање рударском пројектном документацијом (Tomašević и остали 2019) али и ради спровођења терминолошких истраживања. Овај корпус припада доменима рударства и заштите животне и радне средине а чине га докумената која су део пројектне документације, законска регулатива, текстови докторских дисертација и одабрана научна и стручна литература. Рударски корпус је величине 150.365 реченица, односно 2.719.086 облика речи (100.414 лема). Приликом обраде овог корпуса системом за корпусну анализу Unitex пронађено је око 1.900 речи кандидата за *Морфолошке речнике српског језика* (Томашевић и остали 2017).

Рударски корпус је креиран помоћу алата отвореног кода CWB[54] (енг. *IMS Open Corpus Workbench*), који омогућава индексирање и анотацију корпуса, и апликације за постављање упита CQP, која омогућава и приказ конкорданци које одговарају постављеном упиту. Корпус је анотиран (на нивоу врсте речи) и лематизован програмом TreeTagger и претражив преко платформе отвореног кода намењене управљању корпусима текста *NoSketchEngine*[55].

[54] http://cwb.sourceforge.net/ (приступљено 8.01.2020)

[55] https://noske.jerteh.rs/ (приступљено 13.01.2022)

Континуирани развој Рударског корпуса омогућава допуну речника полилексемским рударским терминима применом система за аутоматску екстракцију вишелексемских термина (Stanković, Krstev, Obradović, Lazić, и остали 2015). За допуну речника је од изузетног значаја проширење *Корпуса* новим текстовима који се додају у дигиталну библиотеку *Ромека* (Томашевић и остали 2017), развијену на Рударско-геолошком факултету током рада на докторској дисертацији (Томашевић 2018).

Поред до сада представљених једнојезичних корпуса развијени су и паралелни (или поравнати) корпуси, односно двојезични и вишејезични корпуси који укључују и српски језик. Под окриљем групе оснивача Друштва за језичке ресурсе и технологије развијени су двојезични *Енглеско-српски корпус* (Krstev и Vitas 2011) и *Француско-српски корпус* (Vitas, Krstev, и Laporte 2006) (Vitas и Krstev 2006). Осим двојезичних, постоје и вишејезични корпуси *Орвелова 1984.*, паралелни корпус превода на 12 језика, као и корпус романа Жила Верна *Пут око света за 80 дана* паралелно на 16 језика (Vitas и остали 2008). Паралелни корпуси се користе за различите потребе код обраде природних језика као што су машинско превођење, развој преводилачких меморија (енг. *translation memory*), екстракција двојезичне и вишејезичне терминологије(Krstev и остали 2018) или допуна ворднета дефиницијама (Krstev и Vitas 2011).

4. Модели лексикографских података и развој модела лексикографске базе *Лексимирка*

4.1 Стандардизовани модели за постављање лексикографске базе

Упоредо са првим иницијативама везаним за машински читљиве речнике (осамдесетих година) препознат је и значај стандардизације на пољу језичких ресурса (Calzolari, Monachini, и Soria 2013). Стандардизација је процес уједначавања који доприноси да завршни производ задовољава критеријуме квалитета, односно, норме предвиђене стандардом[56]. Крајњи циљ стандардизације јесте осигурање квалитета, безбедности и ефикасности. Стандарди у пољу језичких ресурса, пре свега теже уједначавању представљања језичких ресурса како би се могли користити за различите примене у оквиру разноврсних апликација. Стандардима у овом пољу кроз конзорцијуме и стандардизационе институције баве се појединци који се баве језиком и језичким ресурсима.

Технички комитет 37 - Језик и терминологија (енг. *Technical committee TC 37 Language and terminology*) (ISO/TC 37/SC 2008) Међународне организације за стандардизацију (енг. *International Organization for Standardization - ISO*) („ISO - International Organization for Standardization“ 2018) бави се стандардизацијом описа, ресурса, технологија и сервиса везаних за терминологију, превођење и друге језичке активности у вишејезичном међународном информационом друштву. Комисија за терминологију А037 („А037 Терминологија“ 2018), Института за стандардизацију Србије – ИСС („Институт за стандардизацију Србије“ 2018), јединог националног тела за стандардизацију, бави се усклађивањем са ISO стандардима из области језика у Србији.

Поред организације ISO, стандардизацијом у области језичких технологија на међународном нивоу баве се и конзорцијуми Иницијатива за обележавање текста (енг. *Text Encoding Initiative, TEI*) и Конзорцијум W3C (енг. *World Wide Web Consortium*). Треба напоменути да ово нису једини спроводиоци идеје стандардизације у лексичким ресурсима, постоје и друге иницијативе, али су претходно поменути широко заступљени и утицајни.

У наставку ће бити представљена три широко заступљена модела за представљање лексикографских података, настала као плод рада претходно наведених организација: Иницијатива за обележавање текста - TEI, Оквир за лексичко обележавање - LMF и *Lemon* модел који је развијен као стандард за дељење лексичких информација на семантичком вебу (McCrae, Aguado-de-Cea, и остали 2012).

4.1.1 Иницијатива за обележавање текста - TEI

Иницијатива за обележавање текста (енг. *Text Encoding Initiative, TEI*) истовремено представља конзорцијум који се бави успостављањем правила за обележавање разнородних текстуалних докумената на уједначен начин ради обележавања и размене података, скуп јавно доступних смерница за обележавање текста

[56] Стандарди се према начини настанка могу поделити на дефакто и дејуре стандарде. Први тип стандарда – дефакто јесу они стандарди који су настали на основу широко прихваћене употребе или неформалне конвенције. Пример оваквог стандарда јесте Microsoft Word DOC (document) рачунарски формат, који је постао дефакто стандард за размену текстуалних датотека захваљујући доминацији програма Word, односно компаније Мајкрософт на тржишту, и кога је подржавају канцеларијски програми других произвођача. Дејуре стандарди су плод правних споразума, уговора или регулатива. Познати пример овог стандарда јесте рачунарски формат HTML (енг. *Hypertext Markup Language, HTML*, језик за означавање хипертекста) који је првобитно био дефакто стандард, све до 1995. године, када је Радна група за интернет инжењеринг (енг. *Internet Engineering Task Force, IETF*) сачинила HTML 2.0 стандард („A history of HTML“ 1998).

у формату XML и заједница пројеката и појединаца који користе TEI смернице (Erjavec 2010).

Почеци TEI-ја сежу у крај осамдесетих година XX века, у време када су готово сви академски хуманистички пројекти који би требало да имају користи од дигиталних технологија у смислу умрежавања, компатибилности и размене података у пракси изгледали сасвим супротно. Део проблема је био у томе што су формати који су коришћени били део пословне стратегије самих фирми које су производиле софтвер (или се бавиле електронских издаваштвом), док је део проблема имао узроке у недостатку комуникације и успостављања стратегије. Имајући у виду текуће проблеме, у новембру 1987. године током састанка на Универзитету Васар, коме су присуствовале академске групе из различитих области, као и представници пројеката и професионалних удружења библиотекара и архивиста из Европе и Северне Америке и Азије, уобличена је идеја о Иницијативи за обележавање текста („History – TEI: Text Encoding Initiative“ 2018). Организација рада на TEI смерницама поверена је спонзорским организацијама: Удружењу за рачунаре у хуманистици (енг. *The Association for Computers in the Humanities, ACH*), Удружењу за рачунарство у књижевности и лингвистици (енг. *Association for Literary and Linguistic Computing, ALLC*), и Удружењу за рачунарску лингвистику (енг. *Association for Computational Linguistics, ACL*). Као резултат рада, у јуну 1990. године, настала је прва скица Смерница TEI, позната као P1. После два круга корекција, проширења и додавања, у мају 1994. године, прва званична верзија Смерница (P3) угледала је светлост дана. Конзорцијум TEI настаје 1999. године када су Универзитет Вирџиније и Универзитет Бергена дали предлог Извршном комитету да се формира организација која ће на међународном нивоу одржавати, промовисати и развијати TEI. Убрзо по усвајању предлога још два универзитета, Универзитет Браун и Универзитет Оксфорд, постала су домаћини TEI-ја. Потом је у јуну 2002. године објављена верзија Смерница P4 чија револуционарност се огледа у употреби XML-а. Актуелна верзија P5 објављена је 2007. године, након темељне ревизије и јавног позива за побољшања и развој нових делова попут, представљања карактера, графике, описа манускрипта и других врста текста („History – TEI: Text Encoding Initiative“ 2018). Ова верзија има 23 поглавља која се баве скуповима елемената за обележавање различитих типова докумената.

TEI смернице за обележавање и размену електронског текста (енг. *Guidelines for Electronic Text Encoding and Interchange*) дефинишу језик за представљање својстава текста (нпр. структуре, концепта, врсте информација). Крајњи циљ јесте пригодно обележен текст за потребе истраживања у различитим научним областима, а пре свега у хуманистичким наукама. Рад са дигиталним, обележеним текстом даје бројне могућности у истраживању садржине текста из најразличитијих аспеката.

Смернице су представљене као модуларна и проширива XML схема. Детаљи схеме описани су богатом документацијом доступном на званичној страни TEI иницијативе[57]. Документи обележени у складу са TEI смерницама су заправо XML документи па је стога неопходно да документи буду синтаксички коректни али и валидни у односу на TEI XML схему. За прецизирање валидности TEI документа најчешће се користе DTD језик (енг. *Document Type Definition*), језик XML схеме[58] (енг. *XML Schema language*), као и RELAX NG[59] језик. Данас је метод који је највише у употреби RELAX NG.

[57] http://www.tei-c.org/index.xml (приступљено 15.08.2021)

[58] https://www.w3.org/XML/Core/ (приступљено 20.07.2018)

[59] http://relaxng.org/ (приступљено 20.07.2018)

TEI иницијатива нуди, сходно потребама, могућност одабира прилагођеног пакета (енг. *customization*) TEI смерница, како би се избегло гомилање елемената који неће бити употребљени. Неки од постојећих пакета су Bare (схема са минимумом елемената), Corpus (схема са елементима за обележавање информација у корпусима), MS (схема са елементима за обележавање манускрипта). Свакако најпопуларнији пакет представља Lite. Овај подскуп TEI елемената омогућава основно обележавање једноставних докумената који је за велики број примена довољан. Врло корисну помоћ при креирању корисничких прилагођених пакета представља веб-алат Roma[60]. Он служи за одабир елемената који ће бити присутни у схеми, за преузимање на тај начин прилагођене схеме, као и за преузимање унапред дефинисаних пакета.

Етикете заглавља и основне структуре TEI документа су универзалне. Сваки TEI документ почива на коришћењу етикета и опционих парова атрибут=вредност. Елемент је компонента структуре која представља текстуалну јединицу. Сваки елемент је означен одговарајућом етикетом, отвореном нпр. `<s>` и затвореном нпр. `</s>`. Садржај који се налази између отворене и затворене етикете назива се садржајем елемента. Пример `<s>`**Данас је био кишан дан.**`</s>` илуструје елемент реченицу означену ознаком `<s/>`, док је садржај елемента „Данас је био кишан дан.“. Када је елемент празан почетна и завршна етикета се могу изразити једном етикетом, нпр. `<s/>.` Атрибути се користе да описивање информација које су опис датог појављивања елемента али нису део садржаја. Атрибутима се придружује њихова вредност па тако настају парови атрибут=вредност. Вредност атрибута се наводи под знацима наводника. Код примера `<s n=”1”>`**Данас је био кишан дан.**`</s>` атрибут n има вредност 1 и показује нам да је дата реченица прва по реду.

Заглавље TEI документа приказује податке намењене корисницима обележеног текста, софтверу који га обрађује и каталогизаторима који га обрађују. У заглављу се описују дигитални текст, извор, његово означавање, као и свака измена текућег TEI документа. Овај скуп података ставља се под етикете `<teiHeader>`. Заглавље садржи пет главних делова, а то су: (1) опис датотеке, `<fileDesc>`, у коме се наводе библиографски подаци саме датотеке, односно дигиталног документа као и изворног документа, ако постоји, (2) опис обележавања, `<encodingDesc>`, који приказује везу између изворног текста и електронске верзије, (3) профил текста, `<profileDesc>`, који описује небиблиографски аспект текста, попут језика, класификације, (4) елемент за складиштење - `<xenoData>` који омогућава уметање метаподатака из схеме која није TEI (нпр. библиографски подаци у MARC формату) и (5) историја промена - `<revisionDesc>` која омогућава бележење свих промена на текућем документу.

У наставку рада ће бити описано поглавље 9 Смерница TEI P5 посвећено речницима, и то верзија 3.2.0 ажурирана 10. јула 2017. године[61] (ревизија 0fcf651). Обележавање овим моделом предвиђено је за све типове речника, почев од класичних (и штампаних и електронских намењених човеку), једнојезичних и вишејезичних, до оних који су намењени употреби од стране рачунара.

Еквивалент речнику представља елемент **body** (тело) који је корени елемент сваког TEI документа (без заглавља).

Више речничких записа могуће је груписати коришћењем етикете **div** (одељак). Примена овог елемента може бити од користи при обради лексике која припада једном слову, или код вишејезичних речника, приликом означавања дела

[60] http://www.tei-c.org/Roma/ (приступљено 15.08.2017)

[61] Верзија TEI смерница 3.2.0 била је доступна у току рада на одабиру модела лексикографске базе. Поглавље 9 посвећено речницима је касније доживело значајније измене и објављено под називом TEI Lex 0. Коришћена верзија је доступна путем платформе Sourceforge: https://sourceforge.net/projects/tei/

речника посвећеног једном од језика. Тада се сви речнички записи који припадају слову или језику стављају под једну етикету **div**.

У зависности од начина на који се обележава структура речнички чланак је могуће означити на више начина. Етикета **entry** представља еквивалент речничком чланку. Њом се обухватају све устаљене информације које се сматрају делом једног речничког записа.

Следи илустрација упрошћене структуре имагинарног речника који се зове *Речник са три лексичка записа*. У њему су коришћењем елемента **div** записи груписани на основу почетног слова.

```
<body>
    <div>
        <head>Речник са три лексичка записа</head>
        <!-- први лексички запис -->
        <entry>
            <form>
                <orth>кућа</orth>
            </form>
        </entry>
        <!-- други лексички запис-->
        <entry>
            <form>
                <orth>кључ</orth>
            </form>
        </entry>
    </div>
    <div>
        <!-- трећи лексички запис-->
        <entry>
            <form>
                <orth>пас</orth>
            </form>
        </entry>
    </div>
</body>
```

С друге стране, етикета **entryFree** користи се за обележавање неструктурираног записа. Атрибут **type** се уз ове етикете може користити да би се означио тип записа (главни, повезани, упутница и сл.). Етикетом **superEntry** окружује се више записа који заједно формирају неку врсту јединице као што су на пример хомографи.

Елемент **hom** групише хомографне записе који се разликују по врсти речи. Значење речничког записа (дефиниција, примера, преводни еквиваленти) окружује се етикетама **sense**. Уколико постоји потреба, више прецизнијих значења се може понављати у оквиру једног значења. Овај случај се јавља код приказивања специфичнијег значења у односу на шира. Тада се етикетама додељује атрибут **n** чија вредност носи ознаку хијерархије значења.

Претходно набројани елементи, **entry, entryFree, hom** и **sense** могу обухватати елементе који служе да опишу облик речи, граматичке податке, изговор, употребу, итд. Такви елементи су такозвани чиниоци лексичког записа на највишем нивоу (следећем највишем у односу на претходне елементе). Они су представљени у наставку рада.

1. Елемент **form** групише податке о свим облицима одреднице/леме. Употребом вредности атрибута **type** ближе се одређује тип одреднице. Предложене вредности су: simple (монолексемска реч), lemma (лема), variant (варијанта), compound (сложена), derivative (изведена), inflected (флективна) и phrase (израз). У оквиру елемента **form** могуће је навођење других писаних и изговорних облика одреднице. Етикетом **orth** означава се писани облик одреднице. Њом се засебно може обележити више флективних облика речи. Атрибутом **type** дефинише се врста облика леме. Тако би, на пример, одредница записа *hidrometeorolog* из *Српског морфолошког речника* била описана као:

```
<form type="lemma">
    <orth>hidrometeorolog</orth>
</form>
```

С друге стране, један флективни облик, похрањен у облику DELAF речника био би представљен као:

```
<form type="inflected">
    <orth>hidrometeorolozima</orth>
</form>
```

Атрибутом **extent** наводи се да ли се елемент односи на целу реч, префикс, суфикс, инфикс или њен део. Овај атрибут се у истом смислу користи код елемента **pron** који садржи информацију о начину изговора одреднице. Елемент **hyph** (hyphenation) садржи одредницу у облику који указује како реч треба поделити када се јави на крају ретка. Етикетом **syll** (syllabification) означава се одредница подељена на слогове. Етикетом **stress** означава се акцентовани облик одреднице. На овом степену је могуће и коришћење етикете **lbl** за потребе означавања синтаксног израза који на неки начин одређују одредницу, нпр. превод, приближно, синоним... Поред претходно набројаних елемената, у оквиру **form** елемента могу се наћи етикета **gram** (grammatical information) и друге етикете које се користе за означавање морфолошких категорија, рода, броја, врсте речи и сл. али ће оне касније бити описане.

2. Елемент **gramGrp** групише све врсте граматичких информација. Може бити коришћен у оквиру елемента **sense**. Он може садржати све врсте морфолошких информација. Елемент **case** означава падеж у коме је дати облик. Елемент **colloc** садржи скуп речи које се често јављају уз одредницу. Податак о граматичком роду обележава се етикетама **gen**. Елементом **iType** означава се флективна класа речничког записа**.** Етикетом **per** означава се граматичко лице флективног облика датог у речнику. Етикетом **number** означава се граматички број облика речи. Врста речи се означава етикетом **pos** и она се придружује одредници лексичког записа. Етикетом **subc** обележавају се информације о поткатегоријама, као што су прелазност, бројивост, итд. За ове поткатегорије не постоје друге предефинисане етикете већ се оне бележе у оквиру етикета **subc**. Граматичко време се означава етикетом **tns.** Етикета **mood** служи за означавање глаголског начина, нпр. императив, потенцијал, футур II.

Елемент **gram,** који може бити и потомак елемента **form,** садржи поделементе који указују на језик из кога одредница потиче (етимологија) **- lang**, повезивање са ортографским обликом одреднице (нпр. ћириличком или латиничком верзијом одреднице или местом где се у склопу полилексемске јединице јавља одредница) **- oRef** и повезивање са изговорним обликом одреднице - **pRef.** С друге стране, елемент **gram** се као поделемент **gramGrp,** уз коришћење атрибута **type** и одговарајућих вредности које представљају граматичке категорије, нпр. **per**, **number**, **subc** и слично може користити за означавање граматичких категорија. На пример, наш претходни пример флективног облика *hidrometeorolozima* проширен граматичким информацијама (граматички род – мушки, број – множина, падеж – локатив и анимантност - анимантан) би изгледао:

```
<form type="inflected">
    <orth>hidrometeorolozima</orth>
    <gramGrp>
        <gram type="rod">m</gram>
        <gram type="broj">p</gram>
        <gram type="padez">7</gram>
        <gram type="animatnost">v</gram>
    </gramGrp>
</form>
```

Исти пример би коришћењем предефинисаних етикета могао бити обележен и на следећи начин:

```
<form type="inflected">
    <orth>hidrometeorolozima</orth>
    <gramGrp>
        <gen>m</gen>
        <number>p</number>
        <case>7</case>
        <animate>v</animate>
    </gramGrp>
</form>
```

Препорука стандарда је да се сви елементи за опис граматичких категорија групишу као деца елемента **gramGrp**.

3. Етикетама **def** окружују се дефиниције одреднице. Дефиниција описује значење речничког записа. Она се може јавити директно унутар записа. Када постоји више дефиниција, оне припадају различитим значењима. За разлику од претходно приказаних чиниоца лексичког записа, елемент **def** не служи искључиво као кровни елемент за групу по хијерархији нижих елемената.

4. Етикетама **cit** окружују се делови текста преузети из другог документа намењени илустрацији употребе речничког записа. У наведеном делу текста се мора јавити реч из одреднице јер се на овај начин дефинише њена употреба. У вишејезичним речницима ова етикета користи се за ознаку превода. Ово се постиже додавањем атрибута **type** чијом вредношћу се указује на превод (нпр. `<cit type="translation">`). Овај елемент се може користити и за навођење библиографских информација о извору примера употребе код једнојезичних речника и тада вредност атрибута **type** елемента **cit** указује да се ради о примеру (нпр. `<cit type="example">`). Ове вредности атрибута **type** нису унапред предефинисане TEI смерницама. Преузети део се наводи у оквиру поделемента **quote**, док се део који указује на библиографију наводи у облику поделемента **bibl** који може садржати низ других поделемената - **author**, **editor**, **title**, **pubPlace** и друге који су релевантни за идентификацију извора. Ове етикете нису специфичне за модул посвећен речницима већ се могу користити за све врсте текста.

5. Елементом **usg** дефинишу се подаци о употреби. Они обично упућују на употребу речи из речничког записа - географску, временску, доменску, стилску, фреквенцијску, прихватљиву или граматичку. Ови аспекти употребе се прецизирају употребом атрибута **type** коме се додељују предефинисане вредности које осликавају претходно наведене типове употребе. Код вишејезичних речника елемент **usg** се са атрибутом **type** користи за семантичке смернице које помажу кориснику речника да боље схвати значење речи. Тако је могуће навести синонимију (нпр. `<usg type="syn">`), хиперонимију (нпр. `<usg type="hyper">`), колокацију (нпр. `<usg type="colloc">`) итд.

6. Елементи за међусобно упућивање на друге речничке записе могу се јавити у неколико облика. Разликовне нијансе у значењима су изузетно мале. Елемент **xr** садржи реч или израз који упућује корисника речника на друго место у истом речнику

али и у другом документу. Етикета **ref** дефинише упутницу на друго место са могућношћу укључивања неког пропратног текста. Елемент **ptr** пребацује курсор на друго место коришћењем атрибута којим се одређује врста „прелаза“. Курсор се може пребацити навођењем хипервезе за пребацивање (нпр. `<ptr target="http://jerteh.rs"/>`), дефинисане тачке на екрану (нпр. `<ptr target="#p143 #p144"/>`), ознаке поглавља (нпр. `<ptr cRef="1.5.1"/>`) и сл. Овој групи елемената се може прикључити и елемент **lbl** којим се означавају облик, пример, превод или нека друга ознака.

7. Етикетама **note** означава се забелешка, најчешће, лексикографа. Забелешке могу указивати на објашњења по питању употребе, граматике, изузетака и сл. Оне се често јављају као посебан елемент на крају речничког записа.

8. Елемент **etym** додаје речничком запису етимолошку информацију. Кроз различите врсте речника описи етимолошких информација варирају, од краћих до сложених и од структурираних до неструктурираних описа. Зато смо овде издвојили само неколико најбитнијих елемената за обележавање информација релевантних за етимологију. Етикатама **lang** обележава се језик порекла. Етикетама **date** означавају се век, датум, или година који су наведени у речнику и односе се на етимологију. Елемент **mentioned** означава пре свега речи или изразе из другог језика поменуте у етимолошком делу. Елемент **gloss** означава израз или реч која је употребљена ради дефинисања друге речи или израза. Елементом **pron** обележавају се изговори речи које су наведене у овом блоку.

7. Етикетама **re** (related entry) обележава се запис који се из неког разлога јавља у другом запису. Такви су примери код сложених речи, израза и колокација. Елемент **re** може садржати исте елементе као и **entry,** с тим што не сме садржати угнежђен исти елемент.

Уз наведене елементе, дефинисано је коришћење најразличитијих атрибута, од оних чије вредности служе за бројчано означавање редоследа, до оних који ближе одређују типове елемената. Преглед дозвољених атрибута могуће је проверити у оквиру описа сваког појединачног елемента. Атрибути се такође користе за навођење информација које нису примарне или које неће бити експлицитно наведене у самом речнику. Тако су, на пример, у сету атрибута **Lexicographic**, који обухвата сет атрибута уобичајен за све елементе у модулу за речнике, и атрибути **norm** чија вредност даје нормализовану вредност ознаке која се означава (нпр. ако је у речнику ознака за глагол г, вредност атрибута **norm** ће бити „глагол“) и **expand** који се дефинише истом аналогијом за ознаку проширеног назива.

Пошто је један од циљева стандардног етикетирања речника могућност размене података, елементи који се користе у више речника се на неки начин морају једнозначно повезати. Нпр. информације о врсти речи се не бележе на исти начин. Неке од варијанти за придев су „п“, „прид.“, „придев“, „ADJ“, „adj.“, „A“, итд. Како би се постигла једнозначност означавања, користе се категорије које прописује ISO у документу (ISO/TC 37/SC 3 Management of terminology resources 2009)- Регистар категорија података (енг. *Data Category Registry, DCR*) у коме је јединствена ознака за придев: http://www.isocat.org/rest/dc/1230. Следи један упрошћен пример лексичког записа *hidrogeološki* у коме је на овакав начин означена врста речи:

```
<entry>
    <form>
        <orth>hidrogeološki</orth>
    </form>
    <gramGrp>
        <pos dcr:datcat=http://www.isocat.org/datcat/DC-1345
             dcr:valueDatcat="http://www.isocat.org/datcat/DC-1230">adj</pos>
```

```
    </gramGrp>
</entry>
```

TEI смернице се највише користе за традиционалне речнике, било изворно дигиталне, било дигитализоване. Један од интересантних примера употребе смерница јесте за потребе описа речника угрожених језика као што је „*Nxaʔamxcín*[62]“ (Czaykowska-Higgins, Holmes, и Kell 2014) или „*Mixtepec-Mixtec*“[63] (J. Bowers и Romary 2018). Објављени су и радови који говоре о употреби ових смерница за потребе обележавања великих академијских речника попут *Речника савременог португалског језика* (пор. *Dicionárioda Língua Portuguesa Contemporânea*) (Salgado, Costa, и Tasovac 2019) или речника афро-азијске породице језика (Mörth и остали 2014).

4.1.2 Оквир за лексичко обележавање - LMF

Оквир за лексичко обележавање (енг. *Lexical Marckup Framework, LMF*) настао је као плод петогодишњег рада групе од 60 истраживача са искуством у развоју лексикона за обраду природних језика и машински читљивих речника. Током 2003. године су начињени први кораци у развоју овог модела удруженим радом америчке и француске делегације на формирању стандарда за представљање речника. Током 2004. године Комитет за управљање језичким ресурсима (енг. *Language resource management*) организације ISO - ISO-TC37/SC4 покреће заједнички пројекат.

LMF прописује стандардизован оквир за бележење лингвистичких информација приликом креирања рачунарских лексикона. Циљеви LMF-а су коришћење истих речника за различите апликације да би се обавили различити задаци (Francopoulo и George 2013), као и изградња модела за стварање и употребу лексичких ресурса и могућност спајања индивидуалних електронских ресурса у један глобални (Francopoulo и остали 2006).

Категорије које користи LMF дефинисане су стандардом ISO 24613:2008 Управљање језичким ресурсима – Оквир за лексичко обележавање (енг. *ISO 24613:2008 Language resource management - Lexical markup framework (LMF)*) (ISO/TC 37 2008, 2008) Комитета за управљање језичким ресурсима ISO-TC37/SC4. Структуру LMF модела треба допуњавати лингвистичим константама преузетим из Регистра категорија података DCR и ISO/IEC 11179-3:2013 (енг. *Information technology - Metadata registries, MDR*) (ISO/IEC JTC 1/SC 32 2013). Посебно треба повести рачуна о тумачењу датих дефиниција како не би дошло до размимоилажења приликом имплементације модела (Francopoulo и George 2013).

Спецификација LMF модела представљена је кроз подскуп UML језика (енг. *Unified Modeling Language*) са идејом приказа лингвистичког описа. Представљене су класе, везе међу класама, као и скуп категорија података који функционишу као парови UML атрибута и вредности. UML спецификација је усклађена са XML серијализацијом која је дата у информативном делу стандарда у виду примера с потпуним DTD документом.

LMF модел заснива се на основном пакету и пакетима проширења. Основни пакет представља костур хијерархије информација у лексичком ресурсу (Francopoulo и остали 2007). Пакети проширења користе класе из основног пакета за проширење додатним лексемским информацијама (морфолошким, семантичким, синтактичким и др.). На слици 11 дат је приказ основног пакета LMF модела у виду UML дијаграма. У кућицама су приказане класе. Везе међу кућицама (класама) су означене бројевима. Ознака 1 значи да се веза мора успоставити тачно једном. Ознака 1..* указује да се веза

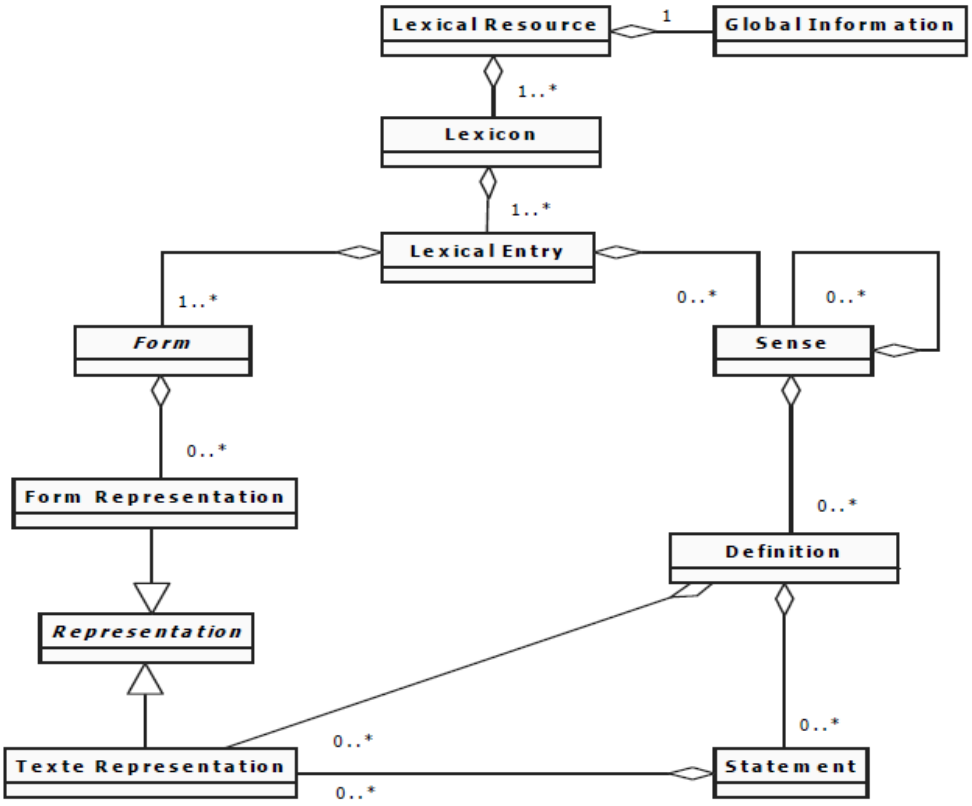
[62] Језик индијанског народа Салиш који се говори у држави Вашингтон (САД).

[63] Језик са 9-10.000 говорника у мексичкој држави Оахака.

може јавити једном или више пута, док ознака 0..* указује на необавезност везе или појављивање више пута. Уколико се на вези налази ознака празног дијаманта класа уз коју је ознака садржи класу са којом је у вези. Празна стрелица показује да је оно на шта се њом упућује специфичнији тип класе са које се упућује.

Lexical Resource (лексички ресурс) је једночлана класа која представља комплетан ресурс, а која може садржати један или више речника (1..*). Класа **Global Information** (опште информације) намењена је административним подацима и другим општим атрибутима који важе за језички ресурс у целости. Такви атрибути се могу односити на језик ресурса, писмо или кодирање. Ова класа мора садржати бар један атрибут /language coding/ - кодирање језика. Језички ресурс мора садржати један административни податак (ознака 1 уз везу). Класа **Lexicon** (лексикон) садржи све лексичке записе (одреднице) за дати језик. Лексикон мора имати најмање један запис.

Класа **Lexical Entry** (лексички запис) представља лексему одређеног језика. Она служи за повезивање и управљање класама **Form** (облик) и **Sense** (значење). Мора садржати једну **Form** класу, док појављивање **Sense** класе није обавезно (0..*).



Слика 11 Дијаграм основног LMF пакета - класе из основног пакета (Francopoulo и George 2013)

Form (облик) је класа која представља лексему, морфолошку варијанту лексеме или морфему[64]. Ова класа дозвољава коришћење поткласа. Опциона класа **Form Representation** (репрезентације облика) представља једну ортографску варијанту класе **Form**, за случај да постоји потреба за том информацијом. Ова класа се означава уколико

[64]„Најмања семантичка јединица која има значење јесте морфема. Морфеме су несамосталне јединице и због тога не улазе у лексикон“ (Драгићевић 2010)

постоји канонски облик, транслитерација или изговор. **Representation** (репрезентација) је додатна специфичнија класа која се може исказати Unicode ниском или јединственим атрибут-вредност паровима, који описују језик, писмо или ортографију.

Sense (значење) је класа која представља значење лексеме. Она дозвољава хијерархијска значења како би се показало да је једно значење специфичније од другог значења текуће одреднице. Ова могућност је на дијаграму приказана петљом означене са 0..* која води из кућице **Sense** и враћа ка њој. **Definition** (дефиниција) је класа која представља слободан текстуални опис значења намењен човеку који користи речник. **Sense** не мора садржати ниједну дефиницију, а може их имати и више. Свака дефиниција се може удружити са више **Text Representation** (репрезентација текста) инстанци да би се управљало дефиницијама на више језика. **Statement** (исказ) је класа која представља слободан текстуални опис који профињује и допуњује дефиницију.

У наставку је у виду XML-а представљен пример лексичког записа на именици *„лежиште“*:

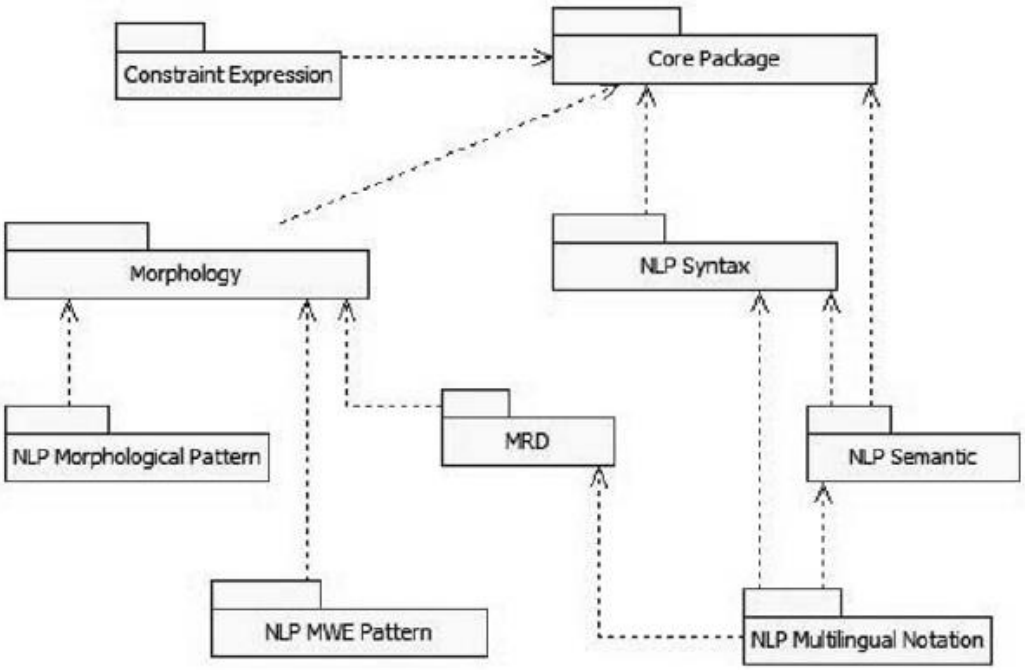
```
<LexicalResource dtdVersion="16">
    <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
    </GlobalInformation>
    <Lexicon>
        <feat att="language" val="srp"/>
        <LexicalEntry>
            <feat att="partOfSpeech" val="N"/>
            <Lemma>
                <feat att="writtenForm" val="ležište"/>
            </Lemma>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

Техничке информације о лексичком ресурсу – верзија dtd-ја (16) и коришћење ознаке језика према трокарактерском стандарду ISO 639-3.

Речник се односи на српски језик чија је ознака према претходно наведеном стандарду „srp.“

Лексички запис се састоји од информација о врсти речи која је именица *N* и леми *лежиште*.

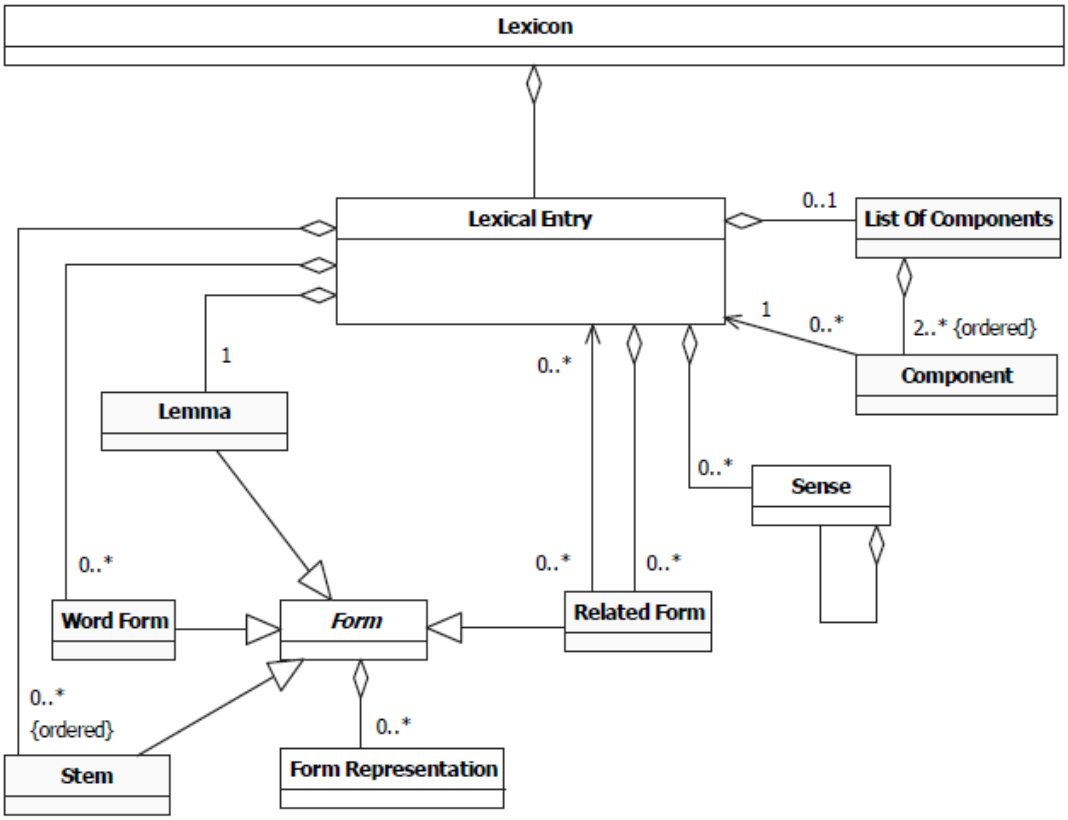
Представљено језгро LMF модела је обавезно. Поред њега доступно је више пакета проширења који се користе да би се описале друге лингвистичке информације. Ови пакети су зависни од основног пакета док су међусобно ослањају један на други у зависности од лингвистичких података заступљених у сваком појединачном пакету. На слици 12 представљене су међузависности пакета LMF модела. Испрекиданим стрелицама су приказане зависности између пакета. Стандардом је предвиђен пакет за бележење морфолошких информација (енг. *Morphology Extension*), као и низ пакета за обраду природних језика: обрасци вишелексемских језичких израза (енг. *NLP Multiword Expression Pattern*), машински читљиви речници (енг. *Machine Readable Dictionary - MRD*), синтакса (енг. *NLP syntax*), семантика (енг. *NLP Semantic Extension*) и вишејезичне нотације (енг. *NLP Multilingual Notations*). Пакет за машински читљиве речнике се не везује директно за основни пакет већ је зависан од морфолошког пакета јер он садржи информације о морфологији лексичких записа. С друге стране пакет за вишејезичну нотацију је зависан од пакета за морфолошки читљиве речнике јер пакет за машински читљиве речнике садржи информације о преводилачким еквивалентима.



Слика 12 Међусобне зависности пакета у LMF-у (Francopoulo и George 2013)

Пакет за опис морфолошких информација

Пакет за опис морфолошких информација (енг. *Morphology extension*) служи за развој речника који садрже опис морфологије лексичких записа. Ова могућност је неопходна за језике који имају богату флексију. Ово практично значи да ће сви облици једне речи бити наведени у оквиру једног речничког записа. На слици 13 приказан је модел пакета који ближе одређује морфолошке обрасце.



Слика 13 Модел пакета за морфолошке информације (Francopoulo и George 2013)

Lemma (лема) је потклаcа класе **Form**. Она представља облик речи који је конвенцијом одабран да представља речнички запис. Често се ради о једном од флективних облика, корену речи или сложеној фрази. **Word Form** (облик речи), потклаcа класе **Form,** илуструје реч у облику у коме се појављује када се користи у реченици или фрази. **Stem** (корен) је потклаcа која представља морфему. **Related Form** је потклаcа класе **Form** која представља облик речи или морфеме које су у вези са речничким записом. За илустрацију овог односа, као инстанца класе **Form** може послужити именица *„рудар“* са потклаcом **Related Form** чија инстанца може бити изведени релациони придев *„рударски“*.

Полилексемске јединице се представљају применом класе **List Of Components** (листа компонената). Ова класа служи за агрегацију компоненти полилексемске јединице. Лексички запис може садржати ниједну или једну листу компоненти. Класа **Component** (компонента) дефинише појединачне компоненте из листе компонената. Листа компоненти мора садржати најмање две компоненте у одређеном редоследу (2..*{ordered}). Како саме компоненте полилексемских јединица и саме могу бити полилексемске јединице, механизам се може употребити и повратно па отуда класа за компоненту упућује на лексички запис везом означеном са (0..*).

У табели 1 приказани су атрибути који се могу користити за опис класа приказаних у оквиру помоћног пакета за опис морфолошких информација.

Табела 1 Атрибути чија је употреба могућа у оквиру пакета за опис морфолошких информација (ISO/TC 37/SC 2008)

Име класе	Пример атрибута	Напомена
Lema	*writtenForm* *phoneticForm* *geographicVariant* *scheme*	Вредности атрибута *writtenForm* и *phoneticForm* су *Unicode* ниске
Word Form	*writtenForm* *phoneticForm* *hyphenation* *gramaticalNumber* *gramaticalGender* *gramaticalTense* *person* *case*	Када је вредност атрибута *writtenForm* „рударство“ вредност атрибута *hyphenation* је „ру-дар-ство“
Related Form	*writtenForm* *phoneticForm* *type*	

У наставку је у виду XML-а представљен један пример лексичког записа са морфолошким информацијама. Представљена је именица *„лежиште“* у облицима једнине и множине номинатива:

```
<LexicalResource dtdVersion="16">
    <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
    </GlobalInformation>
    <Lexicon>
        <feat att="language" val="srp"/>
        <LexicalEntry>
            <feat att="partOfSpeech" val="N"/>
            <Lemma>
                <feat att="writtenForm" val="ležište"/>
            </Lemma>
            <WordForm>
                <feat att="writtenForm" val="ležište"/>
                <feat att="case" val="1"/>
                <feat att="grammaticalNumber" val="s"/>
            </WordForm>
            <WordForm>
                <feat att="writtenForm" val="ležišta"/>
                <feat att="case" val="1"/>
                <feat att="grammaticalNumber" val="p"/>
            </WordForm>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

Техничке информације о лексичком ресурсу – верзија dtd-ја (16) и коришћење ознаке језика према трокарактерском стандарду ISO 639-3.

Речник се односи на српски језик чија је ознака према претходно наведеном стандарду „srp.“

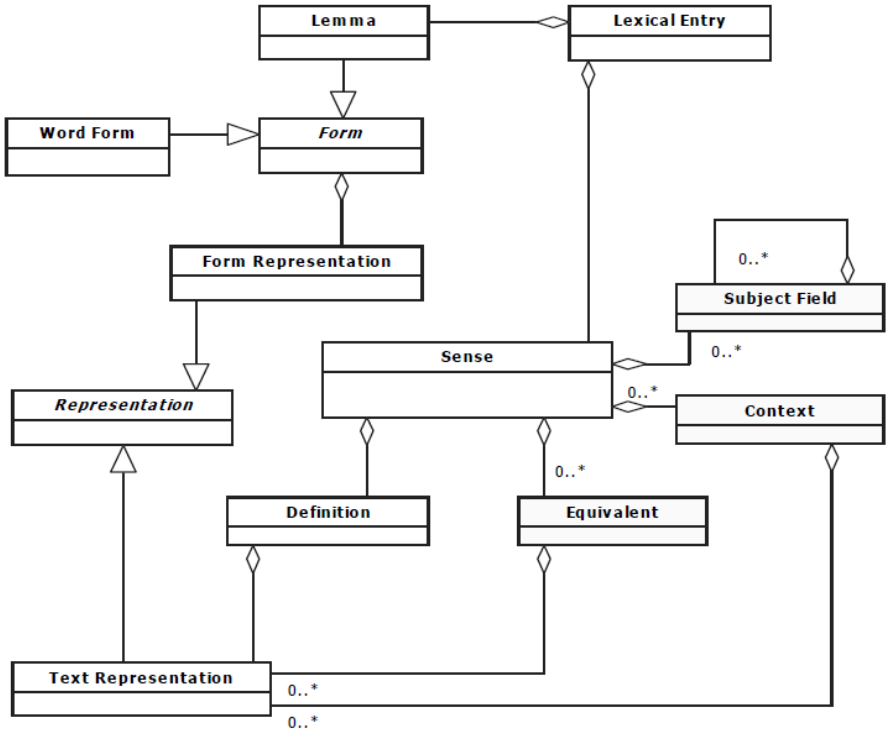
Лексички запис се састоји од информација о врсти речи која је именица *N*, леми *лежиште*, и два облика - *лежиште* који представља номинатив (1) једнине (s) и *лежишта* који представља номинатив (1) множине (p).

Пакет проширења за машински читљиве речнике

Пакет проширења за машински читљиве речнике (енг. *Machine Readable Dictionary extension*) ослања се на пакет за морфологију. Он је намењен кодирању једнојезичних и вишејезичних речника које користе преводиоци, као и кодирању

лексичких података намењених системима за обраду природних језика (ISO/TC 37/SC 2008). Модел овог пакета приказан је на слици 14.

Класа **Equivalent** (еквивалент) представља преводилачки еквивалент речи која је описана класом **Lemma**. **Context** (контекст) је класа представљена у виду текстуалне ниске којом се описује контекст употребе дате речи. Класа **Subject Field** (тематско поље) представља текстуалну ниску која даје информацију о домену. Инстанца класе **Text Representation,** која је потклаcа класе **Form Representation**, садржи посебну ортографију и једну или више категорија података које описују атрибуте те ортографије. Ова инстанца представља различита писма и језике.



Слика 14 Модел пакета за машински читљиве речнике (Francopoulo и George 2013)

Следи пример који уводи значење именице „*лежиште*“:

```
<LexicalResource dtdVersion="16">
    <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
    </GlobalInformation>
    <Lexicon>
        <feat att="language" val="srp"/>
        <LexicalEntry>
            <feat att="partOfSpeech" val="N"/>
            <Lemma>
                <feat att="writtenForm" val="ležište"/>
            </Lemma>
            <Sense>
                <SubjectField>
                   <feat att="label" val="geologija"/>
                </SubjectField>
                <Context>
                   <TextRepresentation>
                     <feat att="language" val="srp"/>
                     <feat att="writtenForm" val="ležište
zlata"/>
                   </TextRepresentation>
                </Context>
            </Sense>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

Техничке информације о лексичком ресурсу – верзија dtd-ја (16) и коришћење ознаке језика према трокарактерском стандарду ISO 639-3.

Речник се односи на српски језик чија је ознака према претходно наведеном стандарду „srp.“

Лексички запис се састоји од информација о врсти речи која је именица *N* и леми *лежиште*.

Значење је описано тематским пољем са вредношћу домена *геологија* и контекстом употребе *„лежиште злата“*.

Проширење за морфолошке обрасце за потребе обраде природних језика

Проширење за морфолошке обрасце за потребе обраде природних језика (енг. *NLP morphological pattern extension*) доступно је за потребе описа морфологије датог језика. Циљ овог пакета је подршка организацији и складиштењу лексичких информација од значаја за анализу и продукцију морфолошких облика, били они флективни, аглутинативни[65], изведени или сложени.

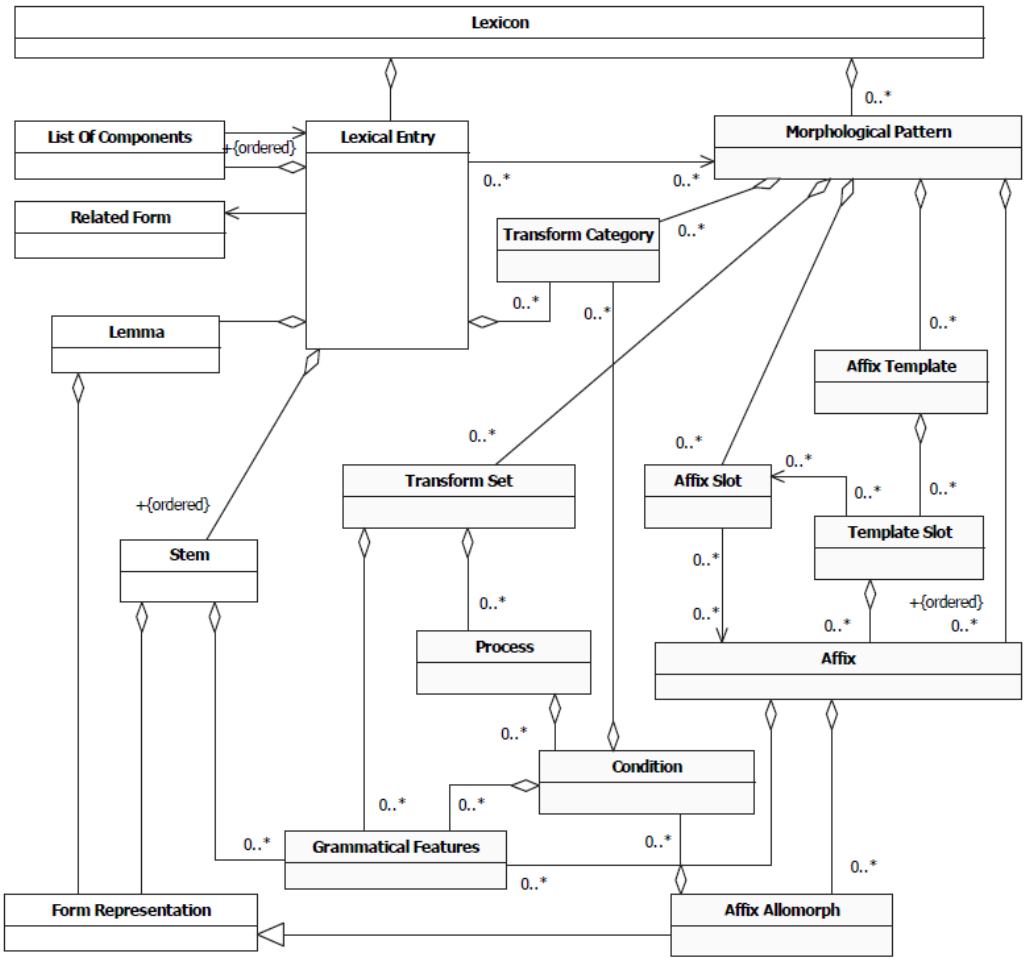
Класа **Morphological Pattern** је заједничка за све наведене морфолошке облике. Засебни облици који се могу јавити су корен, морфема или аломорф[66]. Такође се могу јавити заједнички облици у виду афикса, као и придружена правила.

Потребно је напоменути да су класе **Lexical Entry** и **Morphological Pattern** међусобно здружене у односу на надређену класу **Lexicon**. **Lexical Entry** класа управља облицима речи и морфемама јединственим за један речнички запис, док **Morphological Pattern** дефинише класе које формирају схему која описује неколико речничких записа.

[65]Аглутинација – „творба речи и облика додавањем афикса с одговарајућим значењем корену или основи речи“ (*Речник српскога језика* 2011).

[66] Фоно-морфолошка варијанта морфеме.

Слика 15 илуструје модел проширења за морфолошке обрасце за потребе обраде природних језика.



Слика 15 Проширење за морфолошке обрасце за потребе обраде природних језика (Francopoulo и George 2013)

Појављивање класа **Transform Set** у оквиру класе **Morphological Pattern** није обавезно, али се она може појавити и више пута. Класа **Transform Set** удружује класе **Process** и **Grammatical Features** које ограничавају опсег обрасца који се описује. Класа **Process** представља правила за лингвистичке процесе који се примењују на само један облик речи, афикс или морфему или њихове комбинације док класа **Grammatical Features** представља неуређену комбинацију граматичких својстава. **Condition** представља услове који одређују или ограничавају употребу класа **Process** или **Affix Allomorph.** Класа **Affix Allomorph**, која је уопштење **Form Representation** класе, представља аломорфе канонских афикса у свим писмима и појављивањима. Она је такође удружена са инстанцама класе **Condition** које описују фонолошко окружење или друге услове који утичу на продукцију аломорфа (нпр. где је граница корена аломорфа).

Affix је класа која представља афикс, реч или морфему која је неопходна граматичким својствима да би анализирали или генерисали облике речи. Ова класа управља једним или неколиким аломорфима афикса удружених кроз класу **Affix Allomorph**. **Affix Template** управља обрасцима према афиксима који су према морфологији груписани на флективне, деривационе и аглутинативне кроз **Template Slot** класу. Атрибути који се користе уз **Affix Template** описују позицију афикса, њихов број у уређеном низу и друге специјалне услове који се могу применити на образац афикса.

Template Slot представља скуп афикса који могу бити додати уређеном низу у **Affix Template** класи**.**

Класа **Affix Slot** повезује скуп афикса који се истом положају морфеме додају кроз класу **Template Slot**. Овај скуп афикса представља подскуп афикса одређених кроз **Morphological Paradigm**. Један афикс може бити упућен једним или више објеката **Affix Slot** класе. Атрибути класе **Affix Slot** описују врсту афикса (суфикс, инфикс, префикс)**,** ранг афикса у уређеном скупу, број афикса у уређеном скупу или било који специјални услов примењив на афикс.

У табели 2 дата је листа атрибута које је могуће користити за дефинисање класа из описаног пакета.

Табела 2 Атрибути пакета проширење за морфолошке обрасце за потребе обраде природних језика (ISO/TC 37/SC 2008)

Име класе	Пример атрибута	Напомена
Morphological Pattern	*id* *comment* *example* *partOfSpeech* *patternType*	Занимљиво је да се инстанца *Morphological Pattern* користи за дељење. Једна инстанца *Morphological Pattern ce* не може користити за две различите врсте речи.
Transform Set	*comment*	Ова класа повезује инстанце и служи за писање напомена.
Process	*operator* *affixRank* *componentRank* *stemRank* *rule* *stringValue*	Вредности *operator* могу бити, /addLemma/, /addAffix/, или /addComponentStem/. Вредности *rule* су ниске вредности које представљају скуп лингвистичких правила, нпр. обрасце као што су /CVx/ или формализме као /[X]n -> [1 *ut*]v/.
Condition	*id* *location* *agreement* *affix* *transformType*	
Affix	*writtenForm* *type*	Атрибут *type* може бити одређен вредностима попут /prefix/ или /suffix/.
Affix Template	*type*	Атрибут *type* може бити одређен вредностима попут /prefix/ или /suffix/.
Affix Slot	*id*	
Template Slot	*label* *position* *required*	Атрибут /position/ одређује позицију афикса унутар облика речи.
Affix Allomorph	*writtenForm*	
Transform Category	*id* *comment*	

У наставку следи опис морфолошког обрасца „muški_nepostojanoA“ из *Морфолошких речника српског језика* за именице мушког рода које у флексији имају гласовну промену непостојано *а*. Пример је дат у поједностављеном опису UML модела и односи се на лему „*метар*“. У наставку следе објашњења линија из модела LMF:

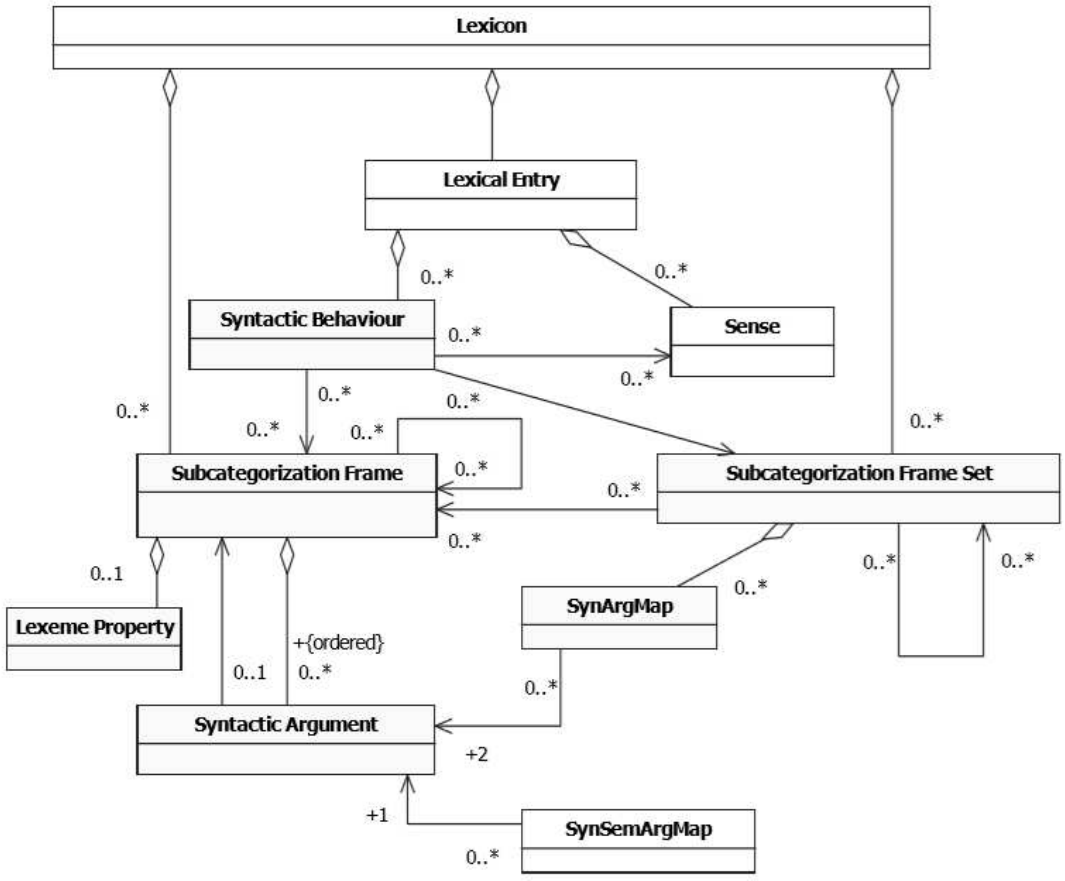
```
:Lexicon          Language="srp"
:LexicalEntry              partOfSpeech="noun"
:Lemma                     writtenForm="metar"
:MorphologicalPattern      id="muški_nepostojanoA"
                  partOfSpeech="noun"
                  comment="flektivni obrazac za
imenice muškog roda s nepostojanim a"
                  example="litar"
```

```
речник је на срп. језику
врста речи је именица
облик леме је „metar“
идентиф. морф. обрасца -
„muški_nepostojanoA“ и важи за
именице
у коментару је опис флект.
обрасца
као пример је дата и именица
„litar“
```

:TransformSet :GrammaticalFeatures grammaticalNumber="singular" grammaticalCase="nominative" :Process operator="addLema" :TransformSet :GrammaticalFeatures grammaticalNumber="singular" grammaticalCase="genitive" :Process operator="addLemma" :Process operator="removeAfter" stringValue="2" :Process operator="addAfter" stringValue="ra" :TransformSet :GrammaticalFeatures grammaticalNumber="plural" grammaticalCase="genitive" :Process operator="addLemma" :Process operator="addAfter" stringValue="a"	класа TransformSet уводи грам. својства битна за овај морф. образац код граматичког броја једнине и облика у номинативу означава лему код грам. облика једнине у генитиву врши се замена тако што се операцијом уклањања са краја леме ниске дужине 2 и додавањем на крај ниске вредности „ra" код грам. облика множине у генитиву врши се трансформација додавањем на лему ниске чија је вредност „a"

Пакет проширења за синтаксу у обради природних језика

Пакет проширења за синтаксу у обради природних језика (енг. *NLP syntax extension*), приказан на слици 16, има за циљ опис својства лексеме када је у комбинацији са другим лексемама у реченици. У табели 3 дати су атрибути које је могуће користити уз класе пакета.



Слика 16 Синтакса у обради природних језика (Francopoulo и George 2013)

Класа **Syntactic Behaviour** (синтаксичко понашање) представља једно од могућих понашања лексеме и опционо је прикључена инстанци класе **Lexical Entry**

(лексички запис) и инстанци класе **Sense**. Класа **Subcategorization Frame** (оквир поткатегорије), чију инстанцу деле све инстанце класе **Lexical Entry** које имају исто синтаксичко понашање у једном језику, представља једну синтаксичку структуру. Њен централни чвор је класа **Lexeme Property** (лексичко својство), док је класа **Syntactic Argument** (синтаксички аргумент) њен аргумент. Класа **Subcategorization Frame Set** (скуп оквира поткатегорија) представља скуп синтаксних структура и могућу везу између њих. Класа **SynArgMap** (мапа синтаксичких аргумената) представља везу која спаја више инстанци класе **Syntactic Argument** исте инстанце **Subcategorization Frame Set**.

Табела 3 Атрибути пакета проширења за синтаксу у обради природних језика (ISO/TC 37/SC 2008)

Име класе	Пример атрибута	Напомена
Syntactic Behaviour	*id* *label*	
Subcategorization Frame	*id* *label* *comment*	
Lexeme Property	*partOfSpeech* *mood* *voice* *auxiliary* *position*	Категорија података */position/* наводи релативну позицију лексеме у реченици имајући у виду синтаксички аргумент.
Syntactic Argument	*syntacticFunction* *syntacticConstituent* *introducer* *label* *restriction* *example*	Категорија података /syntacticFunction/ може имати вредности /subject/, /object/ или нешто друго. Ннеке од могућих вредности /syntacticConstituent/ могу бити /NP/ или /PP/ за именску фразу или предлошку фразу. Категорија података/introducer/ одређује предлог који је потребан за увођење категорије /syntacticConstituent/.
Subcategorization Frame Set	*id* *label* *comment*	
SynArgMap	*comment*	

Следећи пример илуструје приказ два синтаксичка понашања глагола *поломити*: као прелазни глагол с објектом и као повратни глагол кроз модел LMF.

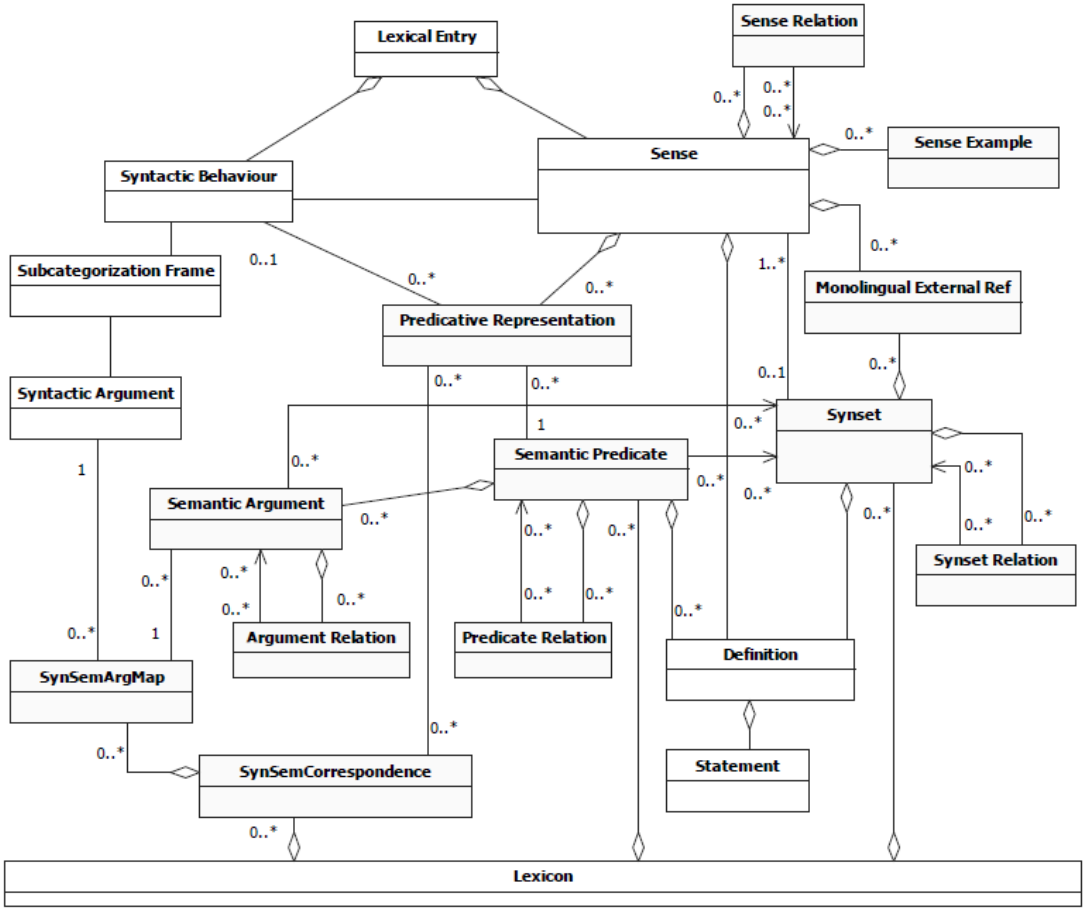
:LexicalEntry partOfSpeech=”glagol” :Lemma writtenForm=”polomiti” :Subcategorization Frame Set Id=”praviPovratniGlagolTip1” :Subcategorization Frame Id=”pravilanSVO”	Скуп оквира поткатегорија чији је идентификациони атрибут „praviPovratniGlagolTip1“ састоји се из два оквира дефинисана аргументима. Први оквир чији је id „pravilanSVO“ дефинисан је аргументима id „synArgX“ субјекат и конституент „NP“ и „synArgY“

:SyntacticArgument :syntacticFunction=”subjekat” :syntacticConstituent=”NP” Id=”synArgX” :SyntacticArgument :syntacticFunction=”objekat” :syntacticConstituent=”NP” Id=”synArgY” :Subcategorization Frame Id=”pravilanSV” :SyntacticArgument :syntacticFunction=”subjekat” :syntacticConstituent=”NP” Id=”synArgZ” :SyntacticArgument :syntacticFunction=”objekat” :syntacticConstituent=”se” Id=”Refse” :SynArgMap (spaja čvorove :SyntacticArgument Id=”synArgY” i :SyntacticArgument Id=”synArgZ”) :comment=”objekat prelaznog glagola je subjekat pravog povratnog glagola”	објекат и „NP“. Овај синтаксички оквир илуструје пример „Она је поломила чашу“. Други синтаксни оквир „pravilanSV“дефинисан је аргументима „synArgZ“ који наводи синтаксну функцију објекта и конституент „NP“ и „Refse“ дефинисан као синтаксна функција објекта са синтаксичким конституентом „se“. Пример за дефинисано понашање глагола би био „Чаша се поломила“ Класом SynArgMap су спојени синтаксни аргументи понашања у виду коментара је дат опис „objekat prelaznog glagola je subjekat pravog povratnog glagola“.

Пакет проширења за семантику у обради природних језика

Пакет проширења за семантику у обради природних језика (енг. *NLP semantic extension*) има за циљ да опише једно значење и његов однос са другим значењима која припадају једном језику. С обзиром на повезаност семантике и синтаксе, овај пакет, приказан на слици 17, описује везу са синтаксом па су стога на слици приказане и класе које су описане у претходном одељку.

Овај пакет користи терминологију Ворднета као ресурса који илуструје семантичку мрежу. **Synset** (синсет) је класа која представља једно значење у оквиру једног језика. **Synset** повезује синониме у виду скупа синонима од инстанци класе **Lexical Entry** унутар исте врсте речи (нпр. синсет: рачунар, компјутер). Класа **Synset Relation** представља усмерену везу између примерака **Synset** класе. Веза је усмерена јер нису све релације симетричне. Код односа синонимије није од значаја које од два повезана значења је у односу са којим, док код, на пример, хиперонимије то јесте од значаја (дигитални рачунар је хипероним у односу на рачунар док обрнуто не важи). Класа **Sense Example** (пример значења) служи да прикаже пример специфичног значења **Sense** инстанце.



Слика 17 Семантичко проширење за обраду природних језика (Francopoulo и George 2013)

Класа **Semantic Predicate** (семантички предикат) представља сажето значење заједно са својом класом **Semantic Argument** (семантички аргумент) која је аргумент семантичког предиката. Инстанце претходно наведене две класе могу бити повезане са једном или више инстанци класе **Synset**. Инстанца класе **Semantic Predicate** односи се на инстанцу класе **Lexicon**. У основи инстанца класе **Semantic Predicate** може представљати заједнички смисао различитих значења која нису прави синоними. Ова значења могу бити везана за инстанце класе **Lexical Entry** чије су врсте речи различите. Тако би на пример за инстанцу класе **Lexical Entry** „читати“ у значењу „распознавати слова у писаном или штампаном тексту“, инстанца класе **Semantic Predicate** могла бити дефинисана са два семантичка аргумента: једним за особу која чита и другим за оно што се чита. Друге инстанце класе **Lexical Entry** повезане са истим предикатом могле би бити „читалац“, „читатељ“ или „читач“.

Класа **Predicative Representation** (представа предиката) представља везу између класа **Sense** и **Semantic Predicate.** Тако би код претходног примера веза између значења глагола и предиката могла бити означена као „master“, док би веза значења именице и предиката („читати“ и „читалац“) била означена као „agentiveNominalization“ (вршилац радње настао деривацијом од друге врсте речи додавањем афикса).

Класа **Argument Relation** (аргумент везе) је веза између инстанци класе **Semantic Argument** исте инстанце класе **Semantic Predicate**. **SynSemArgMap** је веза између семантичког и синтаксичког аргумента (описаног у пакету за синтаксу).

SynSemCorrespondence представља скуп инстанци **SynSemArgMap** за дату инстанцу **Subcategorization Frame**.

Класа **Predicate Relation** представља директну везу између инстанци класе **Semantic Predicate**. **Monolingual External Ref** је класа која представља везу између инстанце класе **Sense** или инстанце **Synset** и неког спољашњег система.

У табели 4 дати су атрибути које је могуће користити уз класе пакета за семантику у обради природних језика.

Табела 4 Атрибути пакета проширења за семантику у обради природних језика (ISO/TC 37/SC 2008)

Име класе	Пример атрибута	Напомена
Sense	*id* *dating* *style* *frequency* *animacy*	
Sense Relation	*label*	Ова класа има више улога – може да представља антонимију, меронимију, итд.
Sense Example	*text* *source* *language*	
Semantic Predicate	*label* *definition*	
Predicative Representation	*type* *comment*	Нпр. семантичка деривација значења именице и значења глагола може бити повезана истим предикатом. У том случају ова класа код значења именице може да буде */verbNominalization/*.
Semantic Argument	*semanticRole* *restriction*	
Argument Relation		
Semantic Type		
SynSemArgMap		
SynSemCorrespondence		
Predicate Relation	*label* *type*	
Synset	*label* *source*	
Synset Relation	*label* *type*	
Monolingual External Ref	*externalSystem* *externalReference*	

У наставку је представљен пример односа значења речи *рачунар* и *компјутер*, као и односа хипонимије са значењем речи *дигитални рачунар* и односа хиперонимије са значењем речи *машина*. Однос и ознаке синсета су преузете уз ворднета.

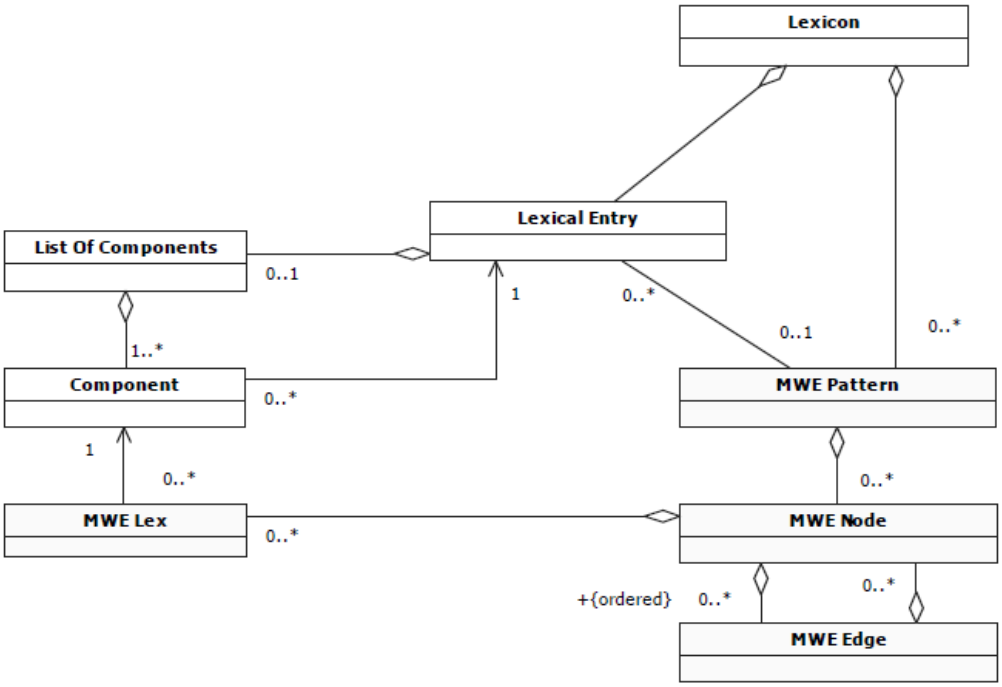
:LexicalEntry partOfSpeech=”imenica” :Lemma writtenForm=”računar” partOfSpeech=”imenica” :Sense id=”računar1” synset=“eng-30-03082979-n“	именица računar са значењем синсета ознаке eng-30-03082979-n
:LexicalEntry partOfSpeech=”imenica” :Lemma writtenForm=”kompjuter” :Sense id=”kompjuter1” synset=“eng-30-03082979-n“	именица kompjuter са значењем синсета ознаке eng-30-03082979-n
:LexicalEntry partOfSpeech=”imenica” :Lemma writtenForm=”digitalni računar” :Sense id=”digrač1” synset=“eng-30-03196324-n“	именица digitalni računar са значењем синсета ознаке eng-30-03196324-n
:LexicalEntry partOfSpeech=”imenica” :Lemma writtenForm=”mašina” :Sense id=”mašina1” synset=“eng-30-03699975-n“	именица mašina са значењем синсета ознаке eng-30-03699975-n
:Synset id=“eng-30-03082979-n“ :SynsetRelation target=“eng-30-03196324-n“ label=”hiponim”	Синсет ознаке eng-30-03082979-n је у вези хипонимије са синсетом eng-30-03196324-n (рачунар и компјутер су хипоними речи дигитални рачунар)
:Synset id=“eng-30-03082979-n“ :SynsetRelation target=“eng-30-03699975-n“ label=”hiperonim”	Синсет ознаке eng-30-03082979-n је у вези хиперонимије са синсетом eng-30-03699975-n (рачунар и компјутер су хипероним речи машина)

Пакет за вишечлане лексичке изразе у обради природних језика

Пакет за полилексемске изразе у обради природних језика (енг. *NLP multiword expression pattern extension*) има за циљ представљање структуре полилексемских израза (колокације, синтагме и сл.) у датом језику. За неке једноставније конструкције полилексемских израза могуће је коришћење пакета за морфолошке обрасце али у случају конструкција које немају фиксне елементе овај пакет представља адекватан избор. На слици 18 дат је приказ модела пакета за вишечлане лексичке изразе у обради природних језика.

MWE Pattern (образац полилексемског израза) је класа која представља тачан тип појаве комбиновања речи. Примерак **Lexical Entry** класе мора бити усклађен са примерком **List of Components** класе тј. листом компонената полилексемског израза (детаљније у пакету за морфолошке информације). Класа **MWE Node** представља детаље

о структури полилексемског израза. **MWE Edge** класа представља мањи део информације о **MWE Node** класи. У пракси он прецизира да је неки део обрасца полилексемског израза индиректни објекат и слично. Инстанца класе **MWE Edge** је рекурзивно повезана са инстанцом класе **MWE Node**. Класа **MWE Lex** је веза са лексичком компонентом која је одређена редоследом представљеним у инстанци класе **List Of Components**.



Слика 18 Пакет за вишечлане лексичке изразе у обради природних језика (Francopoulo и George 2013)

У табели 5 представљени су атрибути класа пакета за полилексемске изразе у обради природних језика.

Табела 5 Атрибути пакета за вишечлане лексичке изразе у обради природних језика (ISO/TC 37/SC 2008)

Име класе	Пример атрибута	Напомена
MWE Pattern	*id* *comment*	Сврха примерка *MWE Pattern* класе је да обједини све полилексемске изразе који имају исту структуру. Сваки примерак има свој *id* тако да могу бити повезани међувезама.
MWE Node	*syntacticConstituent* *semanticRestriction* *grammaticalNumber* *grammaticalGender* *grammaticalCase*	
MWE Edge	*Function*	
MWE Lex	*structureHead* *rank* *graphicalSeparator*	

Следи илустрација употребе пакета на примеру *кнедла са шљивама*.

:Lemma writtenForm=”knedla sa šljivama” :LexicalEntry partOfSpeech=”imenica” :ListOfComponents :Component id=”1” :LexicalEntry partOfSpeech=”imenica” :Lemma writtenForm=”knedla” :Component id=”2” :LexicalEntry :Lemma writtenForm=”sa” :Component id=”3” :LexicalEntry :Lemma writtenForm=”šljiva”	Запис ”knedla sa šljivama” се састоји од компоненти именице ”knedla” облика ”sa” и облика речи ”šljiva”
:MWEPattern id=“NC_PrepN“ comment=“izraz se sastoji od imenice koja se menja praćene predlogom i imenicom u odgovarajućem padežu“ :MWENode syntacticConstituent=”NC” :MWELex rank=“1“	овај израз припада морфолошком обрасцу “NC_PrepN“ чији је опис дат у коментару овај образац се састоји од компоненте ”NC” која на правом месту означава именицу која се мења кроз падеже и бројеве

graphicalSeparator="razmak" structureHead="nosilac fleksije" :MWENode syntacticConstituent="PrepN" :MWEEdge function="imenica se slaže sa predlogom" :MWELex rank="2" graphicalSeparator="razmak" partOfSpeech="predlog" :MWELex rank="3" graphicalSeparator="razmak" partOfSpeech="imenica" grammaticalNumber="množina"	потом од компоненте "PrepN" која означава две фиксне компоненте уз ограничење да се именица слаже са предлогом предлог именица у облику множине

Пакет проширења за обележавање вишејезичности

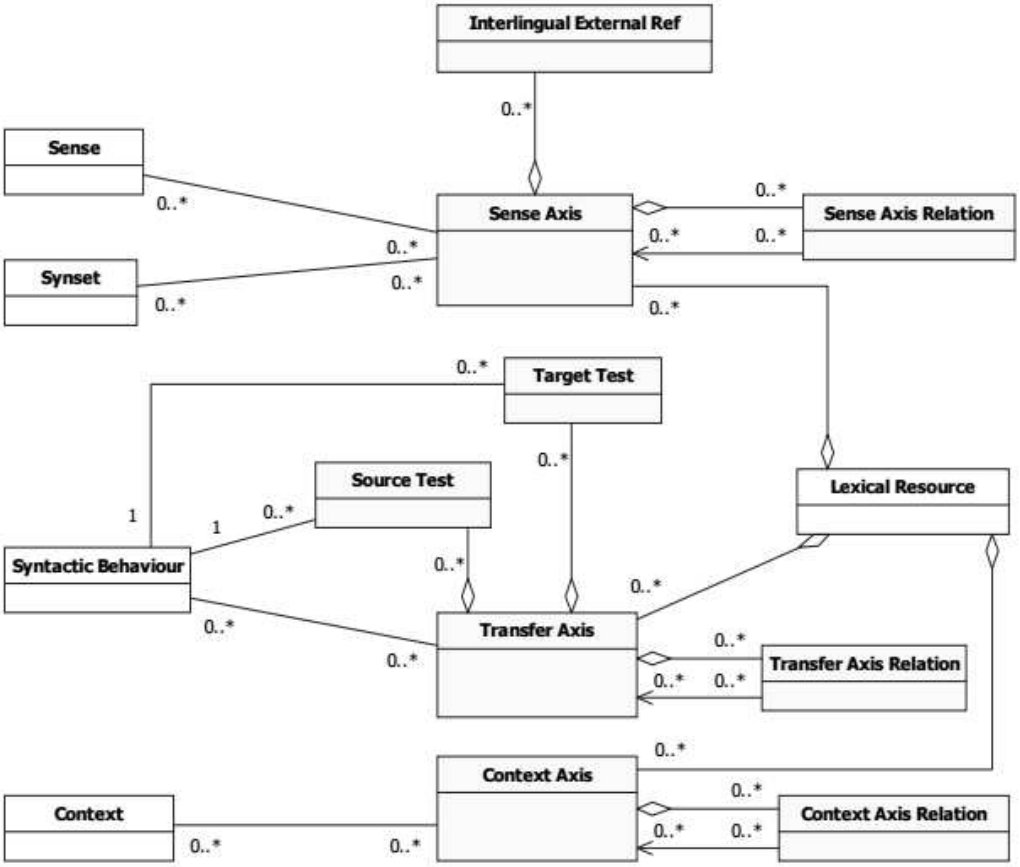
Пакет проширења за обележавање вишејезичности (енг. *Multilingual notation extension*) служи за превођење значења и синтаксних понашања међу различитим језицима. Другим речима овај пакет описује представљање истоветних примерака класа **Sense** или **Syntactic Behaviour** унутар два или више језика. Структура пакета приказана на слици 19 представља двојезични лексикон где је свака веза употребљена да представи еквиваленте значења из различитих језика.

Sense Axis (оса значења) јесте класа која представља везу између блиско повезаних значења у различитим језицима и примењује приступ заснован на међујезичком пивоту[67]. Класа **Sense Axis Relation** (веза оса значења) представља везу међу примерцима **Sense Axis** класе.

Transfer Axis (оса превода) је класа која представља међујезичку везу између неколико примерака **Syntactic Behaviour** класе пореклом из различитих језика. **Transfer Axis Relation** (веза осе превода) је класа која представља везу међу примерцима класе **Transfer Axis**. **Source Test** (изворни тест) је класа која представља услов који утиче на превод с обзиром на употребу текста у изворном језику. **Target Test** (циљни тест) је класа која представља услов који утиче на превод узимајући у обзир употребу у циљном језику. Ове две класе су повезане класом **Syntactic Behaviour**.

Context Axis Relation (контекст везе оса) је класа која представља везу између две инстанце класе **Context Axis**.

[67]Међујезички пивот (енг. *interlingual pivot*) заснива се на машинском парсирању синтаксе изворног језика и семантичке анализе.



Слика 19 Проширење за обележавање вишејезичности (Francopoulo и George 2013)

Следећи пример илуструје употребу овог пакета за потребе представљања превода српске речи *обала* енглеским речима *riverbank* и *seaside.*

:Sense label="sr:obala" :Sense Axis :Sense Axis Relation comment="uzan pojas kopna u kontaktu sa vodenom površinom" label="preciznije" :Sense Axis :Sense label="en:riverbank" :Sense label="en:seaside"	Значење речи *обала* у српском језику које се односи на узан појас копна уз водену површину у енглеском језику је исказано кроз прецизнија значења речи riverbank (обала реке) и seaside (обала мора)

Оквир за лексичко обележавање LMF је успешно примењен на многе лексичке ресурсе. Неки од њих *су Холандски електронски лексикон полилексемских израза - DUELME* (енг. *Dutch Electronic Lexicon of Multiword Expressions*) (Odijk 2013), табеле Лексикон-граматике глагола из француског језика (Laporte, Tolone, и Matthieu 2013), Лексикон Арапског језика (Elleuch, Gargouri, и Ben Hamadou 2021), ворднет (Hayashi и остали 2012).

4.1.3 Lemon

Модел *Lemon* (енг. *lexicon model for ontologies*) повезује светове који су пре његовог настанка били неспојиви, с једне стране, онтологије (скупови појмова и њихових релација на вебу – више о онтологијама у информатици у одељку 4.2.2) и, с

друге стране, лексичке и лингвистичке ресурсе. Сврха модела јесте подршка језичкој заснованости онтологије – додавањем информације о томе како су елементи онтологије лексикализовани у природном језику. Глави подстицај за развој овог модела био је немогућност постојећих стандарда за онтологије, попут *Језика за онтологије на вебу* OWL (енг. *Web Ontology Language*) (Horridge и остали 2004), да опишу морфологију, граматичке категорије, употребу и остала сродна својства онтолошких јединица. *Lemon* модел заснован је на стандарду RDF (енг. *Resource Description Framework*) („RDF Model and Syntax“ 2018) који омогућава прожимање онтологија и лексике на вебу (McCrae, Guadalupe Aguado-de-Cea, и остали 2012).

За опис лингвистичких својстава *lemon* модел препоручује екстерне речнике и онтологије као што је *LexInfo*[68]. *LexInfo* је OWL онтологија која омогућава придруживање детаљних лингвистичких информација елементима онтологије (Cimiano и остали 2011). У њему су дефинисана морфосинтаксичка својства као што су анимнатност, падеж, глаголски вид, лице, глаголско време, итд.

Сам модел је заснован на онтологијама као што су SKOS (енг. *Simple Knowledge Organization System*), LIR (енг. *Linguistic Information Repository*), LexInfo и моделу LMF па је самим тим врло компатибилан са њима. Поглавље 4.1.2 посвећено је детаљном опису LMF модела. SKOS се бави успостављањем стандарда у области система који се баве организацијом знања (тезаурусима, класификационим схемама, тематским системима одредница и таксономијама) у оквиру семантичког веба („Introduction to SKOS - SKOS Simple Knowledge Organization System“ 2018). LIR је објављен у облику онтологије OWL која покрива подскуп елемената за лексички и терминолошки опис за потребе лингвистичке реализације доменских онтологија на природним језицима („LIR - Linguistic Information Repository“ 2018).

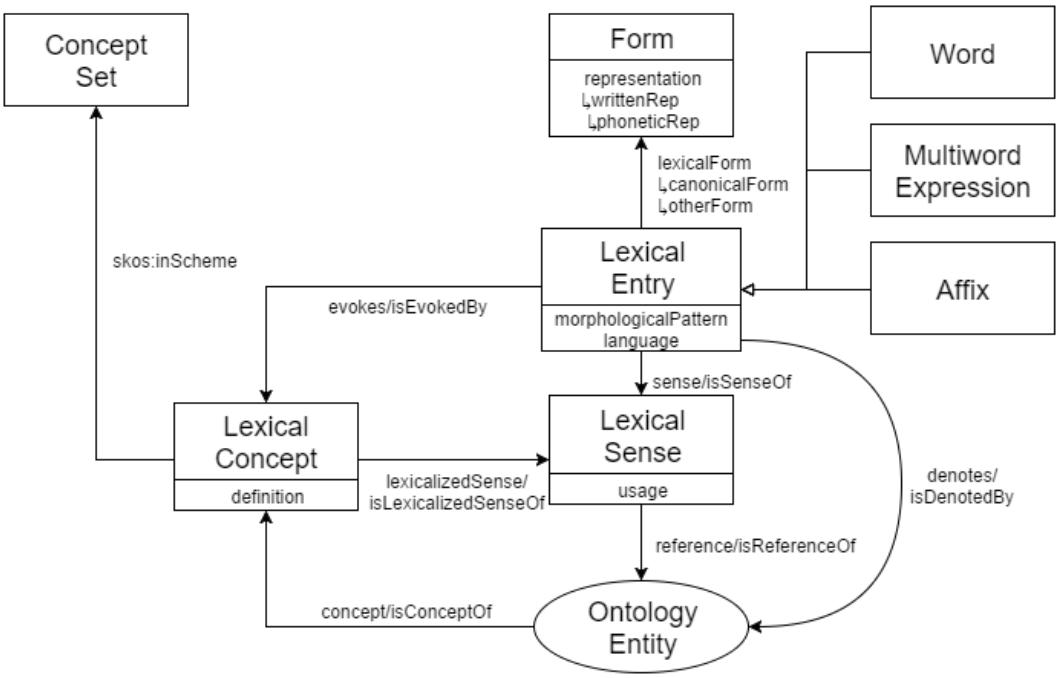
Четири главне карактеристике *lemon* модела су: сажетост, дескриптивност, модуларност и заснованост на RDF синтакси (McCrae, Aguado-de-Cea, и остали 2012). Овај модел је већим делом развијен 2010. године кроз пројекат *Monnet* али се он развија и даље захваљујући групи *Ontology-lexica community* тако да су ажурне информације увек доступне на веб-страни[69]. Опис *lemon* модела у овом раду заснива се на верзији из 2016. године приказаној у извештају Групе („Lexicon Model for Ontologies“ 2016).

И *lemon* модел је модуларно заснован, и чине га пет основних модула:

- повезивање онтологије и лексикона (Ontology-lexicon interface – *ontolex*)
- синтакса и семантика (Syntax and Semantics – *synsem*)
- декомпозиција – (Decomposition – *decomp*)
- варијације и превод (Variation and Translation – *vartrans*)
- лингвистички метаподаци (Linguistic Metadata – *lime*)

[68] http://lexinfo.net/ontology/2.0/lexinfo (приступљено 23.08.2021)

[69] https://www.w3.org/ (приступљено 23.08.2021)



Слика 20 Модел основног модула *ontolex* („Lexicon Model for Ontologies“ 2016)

Модул ontolex, главни модул *lemon* модела, представљен је на слици 20. У уоквиреним деловима дате су класе. Уколико је уоквирени део састављен од два дела у горњем делу се налази назив класе док су у доњем делу приказани њени атрибути (својства) (нпр. класи **Lexical Entry** су додељени атрибути morphologicalPattern и language). Стрелице испуњених глава указују на својства објекта (нпр. **Form** је класа којој припада својство lexicalForm објекта класе **Lexical Entry**), док стрелице са празним главама указују на класу чија је изворишна класа потклаcа (нпр. **Word**, **Multiword Expression** и **Affix** су поткласе класе **Lexical Entry**). Код стрелица праћених ознакама X/Y, X је име својства објекта док је Y назив инвертног својства (нпр. concept/isconceptOf: атрибут објекта Ontology Entity је концепт а инвертни атрибут од **Lexical Concept** је isConceptOf).

Главна класа језгра ontolex модела јесте **Lexical Entry**. Лексички запис је реч, вишечлани језички израз или афикс са јединственом врстом речи, морфолошким обрасцем и другим својствима и представља основну јединицу анализе речника. Самим тим, како је представљено на слици, поткласе **Lexical Entry** могу бити **Word**, **Multiword Expression** или **Affix**. Потклаcа **Word** представља један токен. Потклаcа **Multiword Expression** састоји се из две или више лексема, док потклаcа **Affix** представља морфему.

У наставку следи пример више лексичких записа који припадају представљеним потклаcама. Описани су реч *звезда* (линија 1), полилексемски израз *пуно радно време* (линија 2) и афикс *не* (линија 3). Ови записи су представљени коришћењем текстуалне синтаксе RDF корњача која је тако названа према буквалном преводу енглеског назива *turtle*. Ова синтакса омогућава запису у облику RDF да буде исписан у облику природног текста.

```
1   :zvezda a ontolex:Word
2   :puno_radno_vreme a ontolex:MultiwordExpression
3   :ne- a ontolex:Affix
```

Сва три записа пре првог члана имају ознаку : што значи да тај члан дефинише корисник. Запис приказан у линији 1 показује да реч *звезда* припада потклаcи **Word**, тј. да је монолексемска реч. Део „ontolex:Word“ значи да **Word** потиче из ontolex модела. Запис

приказан линијом 2 показује да *пуно радно време* припада поткласи полилексемских израза (**Multiword Expression**) модела ontolex. Према истој аналогији, афикс *не-* припада поткласи афикса (**Affix**).

Класа **Form** представља један граматички облик лексичког записа. Лексички запис може бити придружен једном од својих облика коришћењем једног од понуђених својстава. Својство **Lexical Form** повезује лексички запис са граматичким обликом лексичког записа.

Препоручено је коришћење својства типа података (енг. *datatype property*) **Representation** које садржи два подсвојства, **Written Representation** и **Phonetic Representation**. Корисницима је омогућено да додају друга подсвојства, према потребама. Писани облик леме описује се својством **Written Representation**. Лексички запис може имати једну или више писаних верзија. Следи пример описа једнине и множине именице *мачка*:

```
1   :lex_mačka a ontolex:LexicalEntry;
2   ontolex:lexicalForm :form_mačka_singular,
3   :form_mačke_plural.
4
5   :form_mačka_singular a ontolex:Form ;
6   ontolex:writtenRep "mačka"@sr .
7
8   :form_mačke_plural a ontolex:Form ;
9   ontolex:writtenRep "mačke"@sr .
```

Првом линијом је исказано да реч *мачка* припада класи **Lexical Entry** модела ontolex. Линија 2 показује да облик *мачка* представља облик једнине, док трећа линија показује да облик *мачке* представља облик множине. Линијом 5 је исказано да облик једнине *мачка* припада класи **Form**, док је линијом 6 исказано да облик *мачка* у српском језику представља подсвојство **Written Representation**. Линијама 8 и 9 је истом аналогијом представљен облик множине *мачке*.

Ортографске варијанте и изговорне варијанте у *lemon* моделу представљају се као различити облици класе **Form** користећи својство **Written Representation**. Следећи пример илуструје део записа за лему *дете*. Описани су ијекавски и екавски облик именице у српском језику. Линијом 5 је приказано да облик *дијете* представља писану верзију у ијекавском изговору српског језика, док облик *дете* представља писану верзију речи дете у екавском изговору српског језика.

```
1   :lex_dete a ontolex:LexicalEntry;
2   ontolex:form :form_dete.
3
4   :form_dete a ontolex:Form;
5   ontolex:writtenRep "dijete"@sr-IJK, "dete"@sr-EK.
```

Phonetic Representation је својство којим се представља фонетски изговор облика коришћењем *међународног фонетског алфабета IPA* (енг. *International Phonetic Alphabet*).

Препорука је да се за повезивање лексичког записа са неким од облика користе подсвојства својства **LexicalForm** - **Canonical Form** и **Other Form**. Подсвојство **Canonical Form** повезује лексички запис са канонским обликом који је обично лема, чији облик зависи од језичких конвенција за дати језик. Следи пример навођења канонског облика за реч *мачка* у српском језику:

```
1   :lex_mačka a ontolex:LexicalEntry, ontolex:Word;
2   ontolex:canonicalForm :form_mačka;
3   rdfs:label "mačka"@sr .
```

Пример се ослања на претходно коришћени запис којим се илуструје навођење једнине и множине именице *мачка* па отуда запис „:form_mačka“ у линији 2. У линији 3 се налази ознака „rdfs:label“ чији префикс „rdfs“ означава да је ознака „label“ пореклом из речника RDFS који представља упрошћен скуп ознака из стандарда RDF (енг. *simple RDF*). Сама ознака „rdfs:label“ означава верзију речи читљиву говорнику српског језика.

Поред канонског је могуће навести друге неканонске[70] облике који се подводе под назив „други облици“ и дефинишу се подсвојством **Other Form**.

Својство **Morphological Pattern** служи за навођење морфолошке класе како би се избегло навођење листе облика код речи са правилним флективним облицима. Овим својством се дефинишу специфични морфолошки обрасци за флексију монолексемских речи и полилексемских израза. Укључивање оваквих образаца се описује другим прилагођеним речницима попут речника *LIAM* (енг. *The Lemon Inflectional and Agglutinative Morphology Module for OntoLex*).

Значење лексичког записа се у оквиру модела *lemon* одређује повезивањем са онтолошким концептом који обухвата или представља његово значење. Својство **Denotes** повезује лексички запис са предикатом у датој онтологији. Овај предикат представља денотационо значење лексичког записа.

Пример који следи повезује запис *nac* са значењем у онтологији *DBpedia*:

```
1   :lex_pas a ontolex:LexicalEntry;
2   ontolex:canonicalForm :form_pas;
3   ontolex:denotes<http://dbpedia.org/page/Dog >.
```

Линијом 3 је приказано да реч *nac* има значење које одговара запису у онтологији *DBpedia* представљеним у виду URI-ја.

Следећи пример показује како можемо обележити реч са више значења:

```
1   :jezik a ontolex:LexicalEntry ;
2   ontolex:denotes<http://dbpedia.org/page/Language> ;
3   ontolex:denotes<http://dbpedia.org/page/Beef_tongue> .
```

Линијама 2 и 3 се указује на URI-је записа из онтологије *DBpedia* на које реч *језик* према значењу може указивати. Дакле, линијом 2 се указује на језик као средство комуникације, док се линијом 3 указује на значење говеђег језика који се користи као састојак јела у кулинарском домену.

Мета повезивања не мора увек да буде примерак у онтологији већ се може упућивати на класу, својство или тип података.

Својства у моделу за повезивање са онтологијама имају инверзно својство „is x-ed by“ где је x оригинално име својства које омогућава лексикону да буде дефинисан на онтолошки начин. У случају својства **Denotes** ово својство би било **isDenotedBy**.

Некада значење лексичког записа није изричито дато у онтологији. Тада *lemon* модел предвиђа креирање нове класе између лексикона и онтологије употребом недељивих (енг. *atomic*) онтолошких ентитета дефинисаних у онтологији. Као илустрација употребе дат је пример, за изражавања значења придева „female“, тј. женски у енглеском језику. Значење придева женски је исказано уз помоћ концепата онтологије *DBpedia*. Креирана је анонимна рестрикциона класа на сучељу између лексикона и онтологије. Линијом 9 је исказан почетак рестрикције значења комбинацијом својства „gender“ онтологије *DBpedia* показаног линијом 10 и вредности „Female“ из онтологије

[70] При том се мисли на све друге облике који нису канонски, нпр. у српском све облике именица осим оног који је канонски, тј. облик номинатива једине.

DBpedia приказане линијом 11. Дакле, исказано је да придев *женски* представља члана нове класе (линија 12) настале спајањем ознаке пола и вредности „женски“.

```
1   :female a ontolex:LexicalEntry;
2   lexinfo:partOfSpeech lexinfo:adjective;
3   ontolex:canonicalForm :female_canonical_form;
4   ontolex:sense :female_sense.
5
6   :female_canonical_form ontolex:writtenRep "female"@en.
7
8   :female_sense ontolex:reference [
9   a owl:Restriction;
10  owl:onProperty<http://dbpedia.org/ontology/gender> ;
11  owl:hasValue<http://dbpedia.org/resource/Female> ] ;
12  synsem:isA :female_arg .
```

Имајући у виду многе практичне ситуације приликом моделовања, претходно представљено својство **Denotes** није довољно да обухвати прецизну везу између лексичког записа и његовог значења у односу на онтологију. Из тог разлога *lemon* модел уводи посреднички елемент назван **Lexical sense.**

Класа **Lexical sense** представља значење лексичког записа интерпретираног кроз позив на одговарајући онтолошки елемент. Ова класа је пар јединствено одређеног лексичких записа и јединствено одређеног ентитета онтологије на који се реферише. Путем објекта **Lexical sense** могу се додати својства (контекст, регистар, домен, итд.) пару који чине лексички запис и онтолошки предикат.

Lexical entry је својством **Sense** повезан са **Lexical sense**, док је **Lexical sense** повезан са онтологијом својством **Reference**. Својство **Sense** повезује лексички запис са једним од његових лексичких значења, док својство **Reference** повезује лексичко значење са онтолошким предикатом који представља ознаку одговарајућег лексичког записа. Овај ланац назван **Sense - Reference** је заправо еквивалент својству **Denotes** које је претходно описано.

Интерпретација речи поштујући њено значење дефинисано у датој онтологији је често условљена употребом услова или прагматичких импликација везаних за регистар, контекст или нијансе значења речи. Својство **Usage** описује те услове при повезивању лексичког записа са онтологијом. Илустрација употребе својства **Usage** у документацији описа модела дата је на примеру француских речи *rivière* и *fleuve* које означавају реку, с тим што се првом инменицом означава она река која се улива у реку, док се другом означава река која се улива у море. Ми ћемо дати пример на српским речима за опис родбинске везе брата једног од родитеља, *ујак* и *стриц*. Обе речи означавају брата родитеља с тим што прва означава брата мајке, а друга брата оца. Самим тим ће њихове значењске нијансе бити забележене коришћењем својства **Usage**. На овај начин постављени услови употребе не користе се уместо формалног значења већ га допуњују додатном информацијом о употреби.

```
1   :ujak a ontolex:LexicalEntry ;
2    ontolex:sense :ujak_sense .
3
4   :stric a ontolex:LexicalEntry ;
5    ontolex:sense :stric_sense .
6
7   :ujak_sense ontolex:reference <https://dbpedia.org/ontology/Uncle>;
8     ontolex:usage [
9       rdf:value "Brat majke"@sr
10    ] .
11
12   :stric_sense ontolex:reference <https://dbpedia.org/ontology/Uncle>;
```

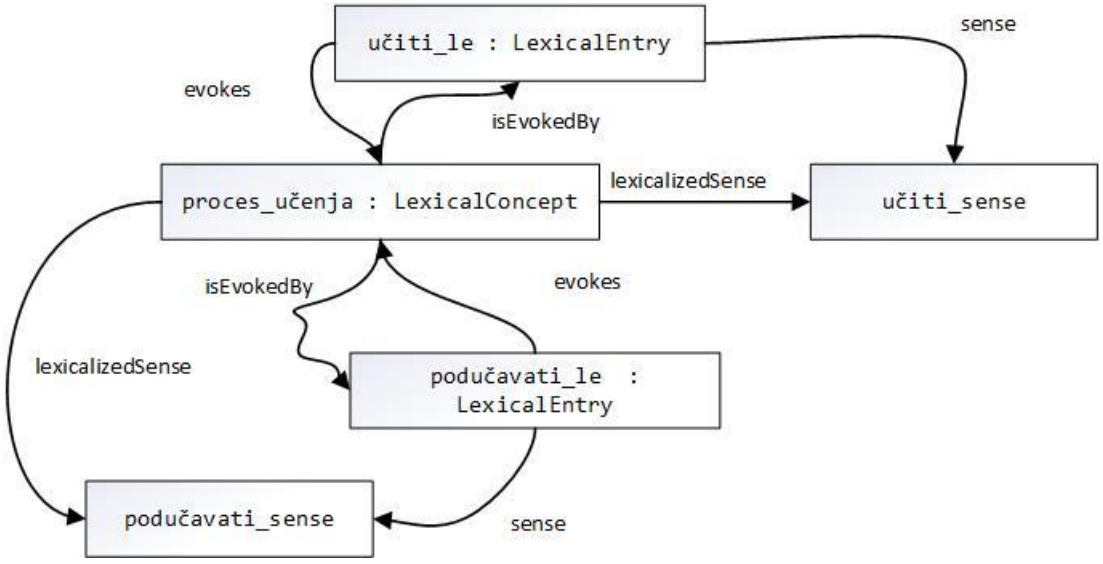
```
13     ontolex:usage [
14       rdf:value "Brat oca"@sr
15     ] .
```

За изражавање чињенице да лексички запис призива посебан ментални концепт, уместо да као што је претходно описано реферише на класу из формалне интерпретације неког модела, користи се класа **Lexical Concept**. Ова класа је поткласа класе **skos:Concept**, класе из шире онтологије *SKOS*. Она представља менталну апстракцију, концепт, јединицу или мисао која може бити лексикализована датом колекцијом значења.

Лексички запис „призива“ лексички концепт, слично томе како лексички запис означава онтолошки запис. Својство **Evokes** повезује лексички запис са лексичким концептом који призива, односно са менталним концептом на који говорника датог језика дата реч асоцира. У документацији *lemon* модела дат је пример реченице *„John F. Kennedy died in 1963.“* где се глагол "*die (in)*" може искористити за генерисање URI-ја **deathDate** унутар **SPARQL** упита. Ми ћемо пример илустровати на реченици „*Nikola Tesla је рођен 1856. године.“.* Глагол „родити (се)“ се може искористити за генерисање URI-ја **birthDate** (линија 2 у примеру). Истовремено глагол „родити (се)“ призива концепт рођења (Birth), што је исказано линијом 3 у следећем примеру.

```
1   :roditi a ontolex:Word ;
2   ontolex:denotes<http://dbpedia.org/ontology/birthDate > ;
3   ontolex:evokes  :Birth .
```

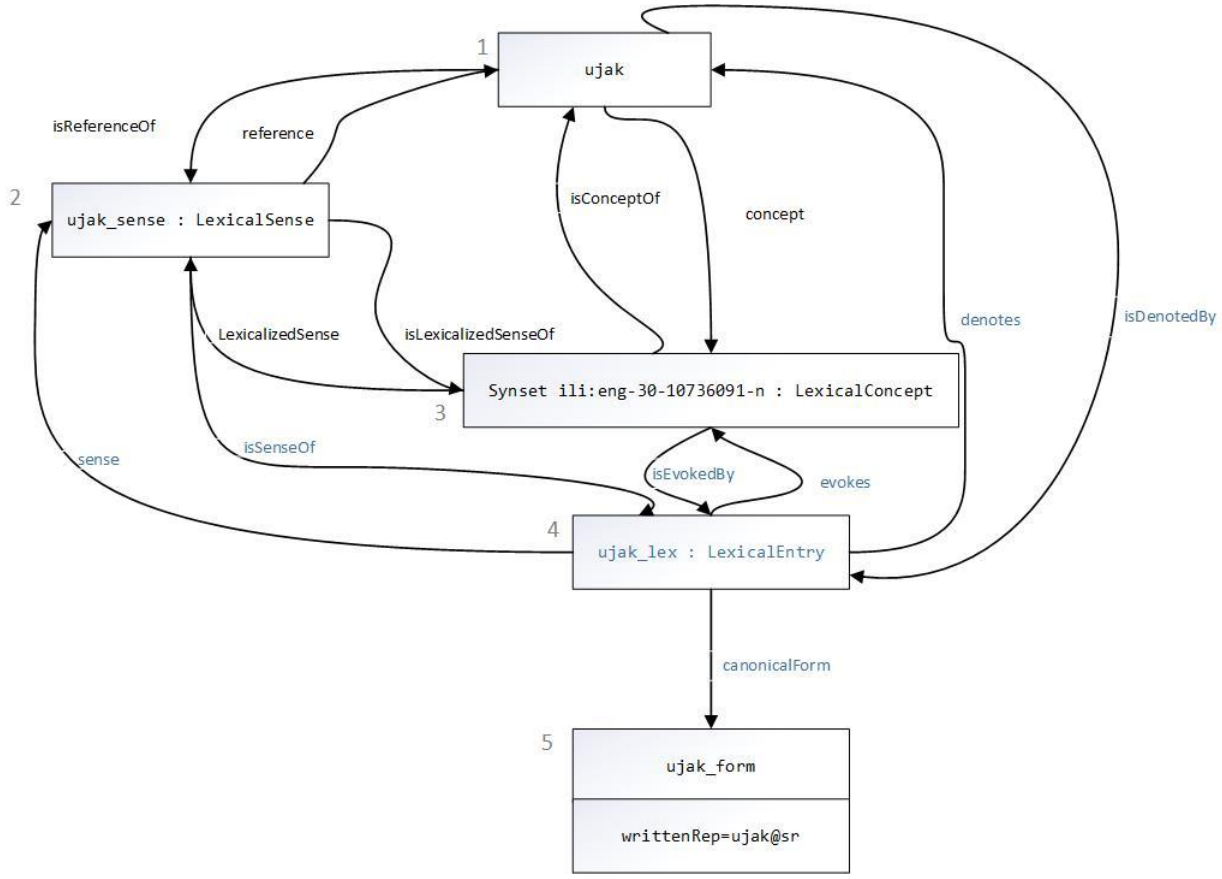
Својство **Lexicalized Sense** повезује лексички концепт, који представља менталну апстракцију, концепт или јединицу мисли, са одговарајућим лексичким значењем које лексикализује концепт. Слика 21 илуструје лексички концепт на примеру односа глагола *учити* и *подучавати*. Лексички концепт за процес учења је лексикализован значењима *учити* и *подучавати*. Лексички запис *учити,* са значењем (својство **sense**) *учити,* с друге стране призива (својство **evokes**) концепт процеса учења који је призван (својство **isEvokedBy**) датим лексичким записом. Исти принцип важи и за лексички запис *подучавати*.



Слика 21 Илустрација употребе концепта Lexical Concept

Слично се може повезати лексички концепт са записом у онтологији користећи својство **Concept**. Пример дат на слици 22 преузет из *Српског ворднета*

илуструје употребу својстава **Denotes, Sense, Evokes, Concept** и **Lexicalized sense** како би се исказао синсет означен међујезичким индексом *eng-30-10736091-n*. Међујезички индекс је индекс скупа значења једне речи у различитим језицима.



Слика 22 Пример представљања синсета из *Ворднета за српски језик - ујак*

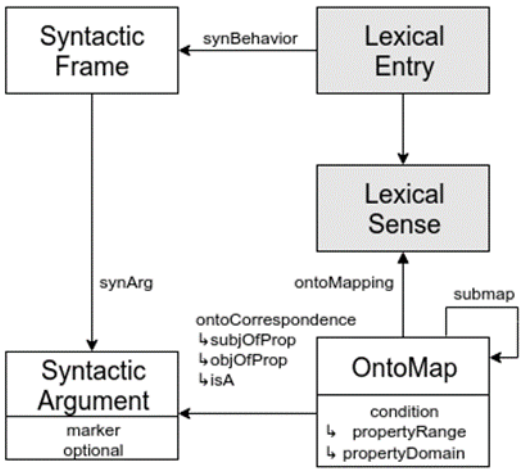
У наставку следи исти пример исказан у нотацији RDF.

```
:ujak_lex a ontolex:LexicalEntry ;
  ontolex:canonicalForm :ujak_form ;
  ontolex:sense :ujak_sense ;
  ontolex:denotes <https://dbpedia.org/ontology/uncle> ;
  ontolex:evokes ili:eng-30-10736091-n .

:ujak_form ontolex:writtenRep "ujak"@sr .

:ujak_sense a ontolex:LexicalSense ;
   ontolex:reference <https://dbpedia.org/ontology/uncle> ;
   ontolex:isLexicalizedSenseOf ili:eng-30-10736091-n ;
   ontolex:isSenseOf :ujak_lex .

 <https://dbpedia.org/ontology/Uncle>
   ontolex:concept pwn:30-10736091-n ;
   ontolex:isReferenceOf :ujak_sense ;
   ontolex:isDenotedBy :ujak_lex .

 ili:eng-30-10736091-n a ontolex:LexicalConcept;
   ontolex:isEvokedBy :ujak_lex ;
   ontolex:lexicalizedSense :ujak_sense ;
   ontolex:isConceptOf <http://dbpedia.org/resource/Cat> .
```

Линијом 1 је исказан лексички запис *ујак*, потом следи канонски облик из кућице која је на слици означена бројем 4. У линији 3 је исказано значење које упућује на кућицу `ujak_sense` (број 2 слици). Линија 4 указује на везу са кућицом број 1 повезану својством `denotes`, док линија 5 својством `evokes` призива синсет у Ворднету. Линија 7 приказује кућицу број 5 у којој је писани облик у српском језику. Линије 9-12 описују кућицу број 2, тј. класу **LexicalSense** и то својства исказана везама одозго на доле – својство `reference`, `isLexicalizedSenseOf` са синсетом и `isSenseOf` са лексичким записом. Линије 14-17 исказују кућицу број 1 (концепт у онтологији *DBpedia*) и њене везе са синсетом, значењем, и лексичким записом. Линијама 19-22 је исказан однос класе **LexicalConcept** (синсет из Ворднета) (кућица 3) са лексичким записом којим је призван, значењем које лексикализује и концептом из онтологије.

Синтакса и семантика (synsem)

Речи се углавном употребљавају у контексту, а глагол у функцији предиката захтева одређени број аргумената. Прелазни глаголи, на пример, поред субјекатског аргумента захтевају и објекатски. На слици 23 илустрован је *lemon* модул за синтаксу и семантику (*synsem*).



Слика 23 Модул за синтаксу и семантику („Lexicon Model for Ontologies“ 2016)

Класа **Syntactic Frame** представља синтаксно понашање речи у смислу синтаксних аргумената који се захтевају. Она у основи описује поткатегорије структуре речи као и синтаксне аргументе који се захтевају. Својство **Syntactic Behavior** (synBehavior) повезује лексички запис са једним од његових синтаксних понашања како је забележено у синтаксном оквиру.

Следећи пример описује глагол *читати* као прелазни глагол:

```
1   :čitati_lex a ontolex:LexicalEntry ;
2   ontolex:canonicalForm :čitati_form ;
3   synsem:synBehavior :čitati_frame_transitive .
4
5   :čitati_frame_transitive a synsem:SyntacticFrame, lexinfo:TransitiveFrame.
6
7   :čitati_form ontolex:writtenRep "čitati"@sr .
```

Линијом 1 је описана класа лексичког записа лексеме *читати.* Друга линија даје канонски облик овог записа преко својства канонског облика, док је трећом линијом својством за синтаксичко понашање `synBehavior` из пакета `synsem` исказан оквир за прелазни глагол. Потом су у линији 5 оквиру прелазног глагола додељене класе синтаксног оквира и оквира прелазног глагола из онтологије *LexInfo*. Линијом 7 је исказан писани облик леме коришћењем модула *ontolex* у српском језику.

Класа **Syntactic Argument** представља место које треба попунити како би синтаксни оквир био потпун. Овом класом се најчешће изражавају граматичке функције попут субјекта, директног или индиректног објекта итд. Својство **marker** служи за навођење маркера синтаксног аргумента. Маркер може бити маркер падежа или лексички запис попут предлога или партикуле.

Својство **SynArg** повезује синтаксни оквир са једним од његових синтаксних аргумената. Следи проширени пример глагола *читати*:

```
1   :čitati_lex a ontolex:LexicalEntry ;
2   ontolex:canonicalForm :čitati_form ;
3   synsem:synBehavior :čitati_frame_transitive .
4
5   :čitati_form ontolex:writtenRep "čitati"@sr.
6
7   :čitati_frame_transitive a synsem:SyntacticFrame, lexinfo:TransitiveFrame;
8          lexinfo:subject :čitati_frame_subj;
9          lexinfo:directObject :čitati_frame_obj.
```

Наведени пример је проширен у односу на претходни линијама 8 и 9 којима се наводе подсвојства својства **SynArg –** синтаксички аргументи `subject` и `directObject` пореклом из онтологије *LexInfo*, тј. прецизира синтаксичко понашање глагола *читати*.

Као што је раније наведено у опису *lemon* модела, синтаксни оквири треба да буду мапирани или везани са онтолошким структурама које представљају њихово значење. На исти начин на који лексичко значење повезује лексички запис са онтолошким ентитетом, класа **OntoMap** (онтолошко мапирање) повезује синтаксни оквир са онтолошким ентитетом.

Класа **OntoMap** прецизира како се синтаксни оквир и његови синтаксни аргументи повезују са скупом концепата и својстава у онтологији која одређују значење синтаксног оквира. Својство **ontoMapping** је предвиђено за повезивање односом 1:1 класа **OntoMap** и **Lexical Sense**. С обзиром на то, препоручено је да уколико лексикон подразумева истовремено класе **OntoMap** и **Lexical Sense**, ове две класе буду дефинисане коришћењем истог URI-ја, јер нема техничких разлога за њихово разликовање пошто имају сличне функције.

Модул **synsem** уводи својство **ontoCorrespondence**. Ово својство, које повезује аргумент предиката дефинисан у онтологији са синтаксним аргументом, који представља аргумент предиката у синтаксичком оквиру, има три подсвојства.

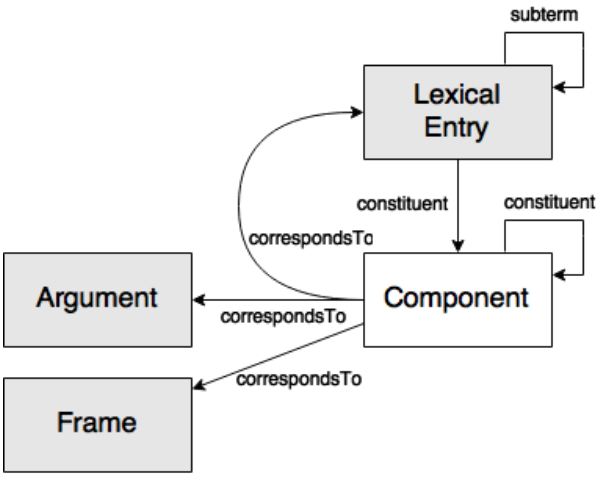
Подсвојство својства **ontoCorrespondence**, **isA**, представља појединачни аргумент класе или унарног предиката. У складу са RDF/OWL терминологијом, први аргумент својства је његов субјекат (**subject**) док је други аргумент објекат (**object**). Стога подсвојство **subjOfProp** (пун назив - **Subject of Property**) представља први аргумент субјекта бинарног предиката (својства) у онтологији, док **objOfProp** (пун назив - **Object of Property**) представља његов други аргумент. Пример употребе дат је на глаголу *читати* који је повезан са онтологијом *DBpedia* (линија 4), прецизирањем везе

између аргумената својства „читач“ (reader) и аргумената који их синтаксно сагледавају. Успостављање везе између субјекта и објекта исказано је линијама 12 - 14.

```
1   :čitati_lex a ontolex:LexicalEntry ;
2   ontolex:canonicalForm :čitati_form ;
3   synsem:synBehavior :čitati_frame_transitive ;
4   ontolex:denotes <http://dbpedia.org/page/Reader> .
5
6   :čitati_form ontolex:writtenRep "čitati"@sr.
7
8   :čitati_frame_transitive a synsem:SyntacticFrame, lexinfo:TransitiveFrame;
9          lexinfo:subject :čitati_frame_subj;
10          lexinfo:directObject :čitati_frame_obj.
11
12   :čitati_ontomap a synsem:OntoMap;
13            synsem:subjOfProp :čitati_obj;
14            synsem:objOfProp :čitati_subj.
```

Модул декомпозиција (decomp)

Декомпозиција је процес растављања вишечланог или сложеног лексичког записа на елементе који га формирају. На слици 24 дат је приказ модула за декомпозицију.



Слика 24 Модул за декомпозицију (*decomp*) („Lexicon Model for Ontologies“ 2016)

Најједноставнији начин разрешавања декомпозиције је коришћење **subterm** својства које указује да ли је лексички запис део другог записа. Следи пример сложеног термина *пуно радно време* (линија 1) који у себи садржи и термин *радно време* (линија 2):

```
1   :PunoRadnoVreme a ontolex:LexicalEntry ;
2      decomp:subterm :RadnoVreme .
```

Ово својство може бити коришћено да се испише садржај сливеница. Следећи пример приказује растављање сливенице *чуваркућа* (линија 1) на делове *чувар* (линија 2) и *кућа* (линија 3).

```
1   :čuvarkuća a ontolex:LexicalEntry ;
```

```
2     decomp:subterm :čuvar_lex;
3     decomp:subterm :kuća_lex .
```

Својство **Subterm** повезује сложени лексички запис са записима који га формирају. Наглашено је да ово својство не наводи посебан флективни облик лексичког записа у сложеници (сливеници) или вишечланом изразу или позицију на којој се садржани лексички запис јавља. Следећи пример илуструје коришћење својства **subterm** за опис сложенице *западноевропски*. Управо на овом примеру ћемо приметити да се као **subterm** наводи канонски облик дела сложенице *западни* (линија 2) уместо флективног облика *западно*.

```
1   :Zapadnoevropski a ontolex:LexicalEntry ;
2     decomp:subterm :Zapadni_lex;
3     decomp:subterm :Evropski_lex .
```

За навођење унутрашње структуре сложенице или полилексемског израза треба описати све његове компоненте. За то се користи класа **Component** која представља реализацију лексичког записа који чини део сложеног лексичког записа. Својство **Constituent** (линије 2 и 5) повезује лексички запис или компоненту са компонентом од које су сачињени. Раније наведени пример *пуно радно време* био би представљен на следећи начин:

```
1   :PunoRadnoVreme a ontolex:LexicalEntry ;
2     decomp:constituent :Puno_comp , :RadnoVreme_comp ;
3         :Puno_comp a decomp:Component .
4         :RadnoVreme a ontolex:LexicalEntry ;
5           decomp:constituent :Radno_comp , :Vreme_comp .
```

Како компонента представља посебан облик лексичког записа који је део сложеног лексичког записа, потребно је компоненту повезати са лексичким записом чији је она облик, а зато се користи својство **correspondsTo**. Могуће је додавати и граматичке категорије како би се јасно означио облик лексичког записа. Следи пример описивања лексичког записа *антилопске ципеле* из *Морфолошких речника за српски језик*. За додавање граматичких категорија коришћена је онтологија *LexInfo*.

```
1   :antilopske_cipele_lex a ontolex:LexicalEntry ;
2     decomp:constituent :antilopske_component;
3     decomp:constituent :cipele_component .
4
5   :antilopske_component a decomp:Component;
6     decomp:correspondsTo :antilopski_lex;
7     lexinfo:gender lexinfo:feminine;
8     lexinfo:number lexinfo:plural.
9
10   :cipele_component a decomp:Component;
11     decomp:correspondsTo :cipela_lex;
12     lexinfo:gender lexinfo:feminine;
13     lexinfo:number lexinfo:plural.
```

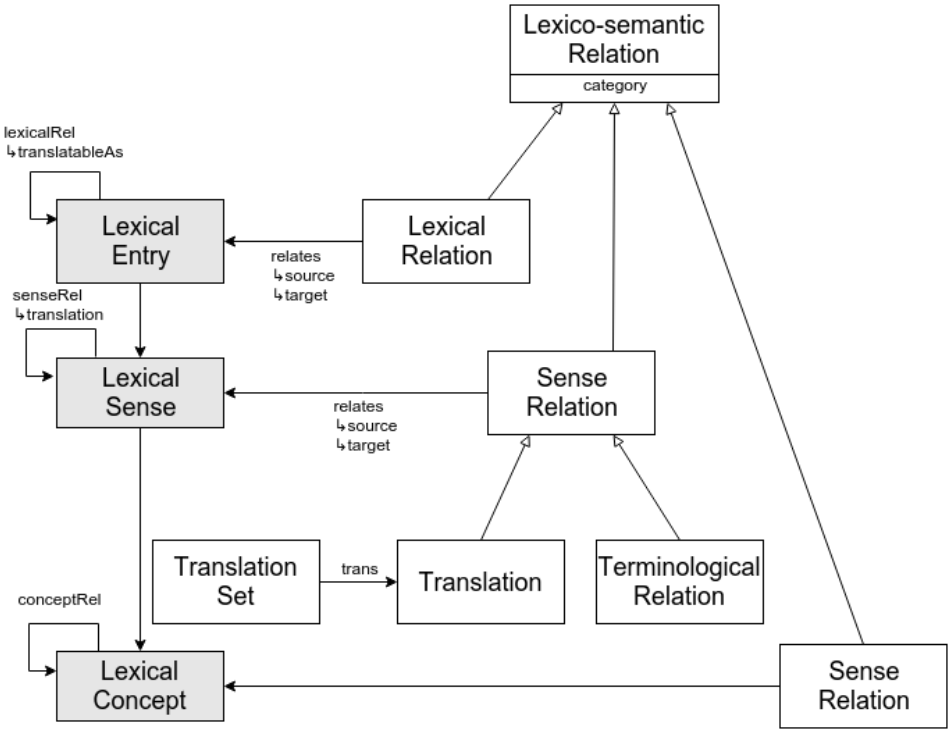
Линијама 6-8 и 11-13 коришћењем својства `correspondsTo` из модула `decomp` наведене су леме лексичких записа чији облици се користе у формирању полилексемског израза. У оба случаја се користе облици множине женског рода. Уколико постоји потреба за навођењем непроменљивог редоследа компоненти, то је могуће постићи коришћењем RDF својстава `rdf:_1, rdf:_2, rdf:_3` итд.

Такође је могуће спецификовати структуру фразе коришћењем синтаксних категорија. Уз чвор (енг. *node*) који описује чиниоце полилексемског израза се додају својство **olia:hasTag** из онтологије лингвистичких ознака OliA (енг. *Ontologies of Linguistic*

Annotation)[71] и адекватна ознака из Пенсилванијске банке дрвета **Penn TreeBank**[72] (енг. *Pennsilvanya Tree Bank*).

Модул варијације и превод - vartrans

Модул за варијације и превод (енг. *Variation & Translation, vartrans*) уводи речник за представљање међусобних веза варијантних лексичких записа и лексичких значења. Графичка илустрација модела дата је на слици 25.



Слика 25 Модул за варијације и превод („Lexicon Model for Ontologies“ 2016)

Својства **lexicalRel** и **senseRel** омогућавају успостављање везе између два записа или значења. Док **lexicalRel** повезује два лексичка записа која су у лексичкој релацији, својство **senseRel,** према истој аналогији, повезује два лексичка значења која су у некој семантичкој релацији. Њих не треба користити директно већ треба увести додатна подсвојства. Следећи пример односа полилексемског израза *геолошки чекић* и речи *чекић* илуструје увођење подсвојства **hypernym** за коришћење својства **senseRel**.

```
1   :geološki_čekić lexinfo:hypernym :čekić.
2     lexinfo:hypernym rdfs:subProperty vartrans:senseRel.
```

Класа **Lexical Relation** представља везу између два записа која су повезана граматички, деривационо или на неки други лингвистички оправдан начин. Деривационим релацијама повезују се облици речи који имају исту основу а могу представљати другу врсту речи као што је, на пример, веза придева и прилога *јак* и *јако* или између глагола *листати* и именице *лист*. Морфосинтаксичка релација повезује

[71] http://nachhalt.sfb632.uni-potsdam.de/owl/ (приступљено 23.08.2021)

[72] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (приступљено 23.08.2021)

облике попут *електронска књига* и *е-књига*. Својством **category** је могуће прецизирати врсту релације.

Класа **Sense Relation** служи за представљање семантичке везе два повезана лексичка значења. Типови семантичких релација међу значењима могу бити хиперонимија и хипонимија, синонимија, антонимија, превод и друго.

Класа **Terminological Relation,** као поткласа **Sense Relation**, је семантичка веза која спаја два значења термина који су семантички повезани у смислу да се у употреби могу заменити. Таква значења могу да се односе на хронолошку, географску, стилску или неку другу варијанту речи. У таквом односу су, на пример, термини *ручник* и *пешкир*, *крух* и *хлеб*, итд.

Лексичко-семантичка класа **Lexico-Semantic Relation,** као наткласа **Lexical Relation** и **Sense Relation**, представља везу између лексичких записа или значења која су повезана неком лексичком или семантичком везом. Својство **relates** служи за повезивање класе **Lexico-Semantic Relation** и лексичког записа. Његово подсвојство **source** означава да је лексичко значење или лексички запис укључен у релацију као извор, док **target** наводи да су они укључени као мета.

Следећим примером ћемо илустровати коришћење лексичке релације „акроним“ на лексичким записима за акроним *UNESCO* и пун назив организације „*United Nations Educational, Scientific and Cultural Organization*“.

```
1   :UNESCO a ontolex:LexicalEntry ;
2        ontolex:sense :UNESCO_sense;
3        ontolex:lexicalForm :UNESCO_form.
4
5   :UNESCO_sense ontolex:reference <https://dbpedia.org/page/UNESCO> .
6
7   :United_Nations_Educational_Scientific_and_Cultural_Organisation a ontolex:LexicalEntry;
8        ontolex:sense : United_Nations_Educational_Scientific_and_Cultural_Organisation
_sense ;
9        ontolex:lexicalForm :
United_Nations_Educational_Scientific_and_Cultural_Organisation_form.
10
11   :United_Nations_Educational_Scientific_and_Cultural_Organisation_sense ontolex:reference
<https://dbpedia.org/page/UNESCO> .
12
13   :UNESCO_form ontolex:writtenRep "UNESCO"@en .
14   : United_Nations_Educational_Scientific_and_Cultural_Organisation_form
ontolex:writtenRep" United_Nations_Educational_Scientific_and_Cultural_Organisation"@en .
15
16   :UNESCO_acronym a vartrans:LexicalRelation ;
17         vartrans:source : United_Nations_Educational_Scientific_and_Cultural_Organisation ;
18         vartrans:target :UNESCO ;
19         vartrans:category :acronym.
```

Линијом 1 је описан лексички запис *UNESCO* са значењем (линија 2) и лексичким обликом (линија 3). Као значење за *UNESCO* дата је веза ка онтологији *DBpedia* (линија 5). И за лексички запис који представља пун назив организације је дато значење повезивањем са онтологијом *DBpedia* (линије 7-12). Линијама 13-14 су представљени писани облици обе верзије назива. Линијом 16 је уведена класа `LexicalRelation` за акроним док су својствима `source` за изворни облик (линија 17) и `target` за циљни скраћени назив (линија 18) наведена оба облика назива организације.

Поред набројаних веза, овим модулом је могуће направити везу између концепата што се може користити за потребе описивања веза међу синсетима у

Ворднету. Својство **conceptRel** повезује два лексичка концепта која су у значењском односу.

Превод повезује два лексичка записа која припадају различитим језицима у случају да је њихово значење еквивалентно. Постоје три нивоа за исказивање једнакости превода. Код првог начина сва значења записа на различитим језицима су повезана са једним концептом онтологије коришћењем класе **Sense Relation**. Код другог нивоа исказивања једнакости лексички записи не указују на тачно исти концепт али се може рећи да су значења еквивалентна ако се поклапају у многим контекстима. У овом случају се користи класа **Translation** (потпкласа класе **Sense Relation**) која изражава семантичку везу између два значења лексичких записа на различитим. За ове потребе може се користи и својство **translation** (**Lexical Sense**, слика 25) за повезивање лексичких значења записа који су у преводној релацији.

Код трећег нивоа изједначавања превода се прецизира контекст у ком се значења лексичких записа на различитим језицима поклапају. Ово се описује навођењем тачног контекста и услова под којим је поклапање могуће. Својством **translatableAs** (**Lexical Entry**, слика 25) повезује се лексички запис на једном језику са лексичким записом на другом језику у одређеном контексту.

Преводи се, у зависности од језика, извора или неког другог критеријума, могу груписати коришћењем класе **Translation Set**. Како би се скуп превода повезао са неким преводом који садржи, користи се својство **trans**.

Модул лингвистичких метаподатака - *lime*

Пети модул *lime* односи се на лингвистичке метаподатке (енг. ***LI****nguistic* ***ME****tadata*). Он омогућава опис метаподацима на нивоу повезивања лексикона и онтологије, а замишљен је да буде компатибилан и комплементаран са схемама метаподатака *Даблинско језгро*[73] (енг. *Dublin Core*), онтологијом *PROV*[74], *DCAT*[75] или *VoID*[76].

Три главна ентитета које *lime* препознаје јесу: скуп референтних података (енг. *reference dataset*), лексикон (енг. *lexicon*) и скуп концепата (енг. *concept set*).

Lime разликује три главна типа скупова којима се додају метаподаци:

- Скуп лексикализација (енг. *set of lexicalizations*) који садржи повезивања између онтолошких логичких предиката и лексичких записа из лексикона;

[73] Даблинско језгро: http://dublincore.org/ (приступљено 23.08.2021)

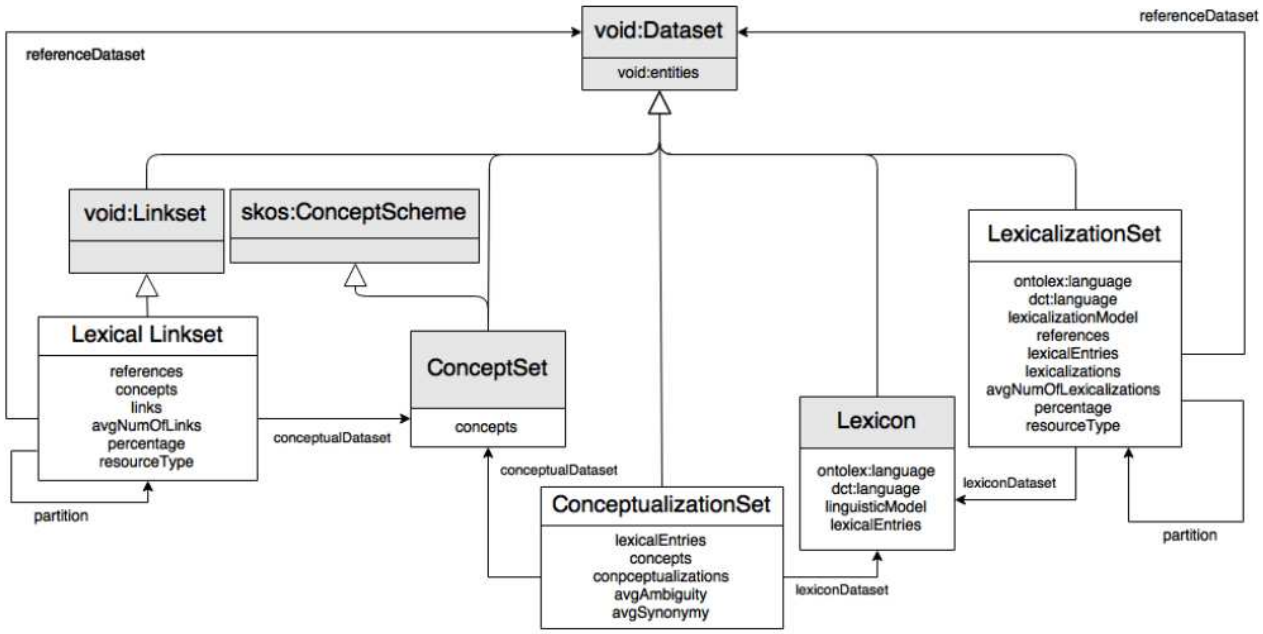
[74] Онтологија *PROV* (енг. *Provenance*) обезбеђује скуп класа, својстава и рестрикција које се користе да прикажу информације о пореклу потребне за размену међу различитим системима и различитим контекстима. https://www.w3.org/TR/prov-o/ (приступљено 23.08.2021)

[75] Речник DCAT (енг. *Data Catalog Vocabulary*) је у серијализацији RDF и намењен је олакшавању размене података међу каталозима података објављеним на вебу: https://www.w3.org/TR/vocab-dcat/ (приступљено 23.08.2021)

[76] *VoID (енг. Vocabulary Identification)* је речник за метаподатке о скуповима RDF података. https://www.w3.org/TR/void/ (приступљено 23.08.2021)

- Скуп концептуализација (енг. *set of conceptualizations*) који садржи повезивања лексичких концепата и скупа концепата и записа у лексикону;
- Скуп лексичких веза (енг. *set of lexical links*) који повезује лексички концепт из скупа концепата са везама у онтологији.

Приказ модула *lime*, заједно са ентитетима и скуповима, дат је на слици 26.



Слика 26 Модул за лингвистичке метаподатке превод („Lexicon Model for Ontologies“ 2016)

Главни ентитет који окупља метаподатке је објекат који представља речник, односно лексикон, стога ћемо опису овог дела посветити посебан простор. Класа **Lexicon** представља скуп лексичких записа за одређени језик или домен и мора садржати

најмање један лексички запис. Својство које повезује речник са лексичким записом јесте **Entry**. Како сви записи у једном речнику морају бити на истом језику, за додавање ознаке језика речнику, његовом запису, концепту или лексикализованом скупу, користи се својство **Language**. За ознаку језика је препоручена употреба својства **language** из Даблинског језгра са везом ка коду језика из онтологије *Lexvo.org*[77] која пружа информације о језицима и писмима или *Речника Конгресне библиотеке*[78] (енг. *The Library of Congress Vocabulary*) тј. кодовима за језике према стандарду *ISO*. Својство **Lexical Entries** наводи бројем колико има речничких записа у речнику, лексикализованом или концептуализованом скупу.

Битно је напоменути да се својство **Linguistic Catalog** користи за навођење каталога или речника лингвистичких категорија који је одабран за анотацију граматичких својстава записа у речнику. Следи пример речника који садржи три записа на српском језику а користи *LexInfo* за дефинисање граматичких категорија, у овом примеру, врсте речи.

```
1   :lexicon a lime:Lexicon;
2     lime:language "sr";
3  dct:language <http://id.loc.gov/vocabulary/iso639-1/sr>,
<http://lexvo.org/id/iso639-3/srp> ;
4     lime:lexicalEntries "3"^^xsd:integer;
5     lime:linguisticCatalog <http://www.lexinfo.net/ontologies/2.0/lexinfo> ;
6     dct:description "Ovo je primer rečnika"@sr;
7     dct:description "This is an example lexicon"@en;
8     dct:creator <Biljana>;
9     lime:entry :lex_mačka;
10     lime:entry :lex_pisati;
11      lime:entry :lex_obojen .
12
13   :lex_mačka a ontolex:LexicalEntry, lexinfo:Noun;
14      ontolex:canonicalForm :form_mačka.
15   :form_mačka ontolex:writtenRep "mačka"@sr.
16
17   :lex_pisati a ontolex:LexicalEntry, lexinfo:Verb;
18      ontolex:canonicalForm :form_pisati.
19   :form_pisati ontolex:writtenRep "pisati"@sr .
20
21   :lex_obojen a ontolex:LexicalEntry, lexinfo:Adjective;
22      ontolex:canonicalForm :form_obojen.
23   :form_obojen ontolex:writtenRep "obojen"@sr .
```

У примеру је описан речник (линија 1) на српском језику (линија 2), при чему се језик речника спецификује коришћењем етикета из Даблинског језгра (префикс `dct`) и ознакама из *Речника Конгресне библиотеке* и онтологије *Lexvo.org* (линија 3). Линијом 4 је исказано да речник садржи 3 лексичка записа, док је линијом 5 својством `linguisticCatalog` наведено да се онтологија *LexInfo* користи за лингвистичке податке. Линијама 6 и 7 је дат опис речника на српском и енглеском језику. У линији 8 је наведен аутор речника, а потом су наведени и лексички записи. Од линије 13 до 23 су наведени описи лексичких записа.

Уз све набројано *lime* модул дефинише и сценарија за објављивање података из *lemon* модела. У самом моделу су описана три типа ентитета:

[77] http://www.lexvo.org/ (приступљено 23.08.2021)

[78] http://id.loc.gov/vocabulary/iso639-1.html (приступљено 23.08.2021)

- Лексикони - *lime:Lexicon*
- Лексикализације - *lime:LexicalizationSet*
- Референтни скуп података или онтолошки речник - *VoID Dataset* или *VOAF Vocabulary*.

Сваки од ових типова може бити објављен као засебан извор података, а могу бити објављени и у комбинацији или као обједињен ресурс.

Аутори *lemon* модела наводе четири уобичајена начина за објављивање:

- Независни ресурси – У овом случају се сва три ентитета објављују као независни извори података. У пракси је могуће да референтни скуп података и лексикон постоје независно или су производи различитих аутора које потом трећа страна повезује скупом лексикализација заснованим на лексичким записима постојећег лексикона.
- Повезивање са лексиконима треће стране – Лексикон опште намене је објављен као независан ресурс. Потом, током развоја референтног скупа података или онтологије аутори одлуче да га објаве заједно са скупом лексикализација заснованим на лексичким записима из постојећег лексикона.
- Повезивање са онтологијом треће стране – Лексикон се кроји по постојећем референтном скупу података који је објављен са скупом лексикализација. Сада је референтни скуп или онтолошки речник старији од лексикона који је направљен за њега.
- Интегрисано – У овом случају се сва три ентитета објављују заједнички и чине један извор података. Најчешће се сва три ентитета развијају у једној средини.

Многи савремени ресурси користе *lemon* модел. *PAROLE/SIMPLE* представља лексиконе мапиране са *lemon* моделом и *LexInfo* онтологијом (Villegas и Bel 2015).

DBnary је скуп вишејезичких лексичких података настао из Wiktionary издања за различите језике, а заснован је на *lemon* лексичком моделу (Sérasset 2015). *DBpedia Wiktionary* представља симбиозу онтологије *DBpedia*, усклађене са *lemon* моделом, и Wiktionary речника коме *DBpedia* омогућава да буде машински читљив речник (Brekle 2012). *LemonUby* представља *UBY* ресурсе[79] (енг. *Large-Scale Unified Lexical-Semantic Resource*) у *lemon* облику (UKP 2018). *Дељени ресурси за анализу осећања Eurosentiment* (Sánchez-Rada, Iglesias, и Gil 2015) такође су усклађени са *lemon* моделом, као и *PanLex* (Kamholz, Pool, и Colowick 2014), преводилачки ресурс који комбинује различите преводилачке ресурсе и више од 2.500 речника тежећи да успостави преводне еквиваленте међу свим језицима (за сада је укључено више од 5.000 језика) на нивоу лексема.

4.1.4 Поређење представљених модела

С обзиром на то да је *lemon* модел проистекао из LMF модела, јасно је да постоји доста сличности између њих али су аутори *lemon* модела настојали да превазиђу неке недостатке LMF-а. Оба модела су модуларна, што се огледа у могућности опционог коришћења модула у *lemon*-у, односно пакета у LMF-у.

[79] У ове ресурсе се убрајају *FrameNet*, *OmegaWiki* за енглески језик, *OmegaWiki* за немачки језик, *VerbNet*, *Wiktionary* за енглески језик, *Wiktionary* за немачки језик и *Принстонски WordNet* верзија 3.0.

Lemon модел има мање елемената од LMF-а. Истовремено су називи неких елемената у *lemon* моделу преузети из LMF-а и модификовани како би се имена скратила и у исто време смањила сличност у именовању класа и њихових својстава. Тако је **SubcategorizationFrame** постало **Frame** у *lemon*-у, **Sense** је **LexicalSense**, **MWENode** је **Node**, **MonolingualExternalRef** је **OntologyReference**, а **SyntacticBehaviour** је постало **synBehavior** (McCrae, Aguado-de-Cea, и остали 2012).

За разлику од LMF-а *lemon* тежи коришћењу спољашњих већ постојећих лексикона за различите потребе што је добро јер се тиме избегава стварање нових система без потребе. Он изоставља морфолошки опис јер предвиђа коришћење спољашњих лексикона за бележење морфолошких информација, а такође дефинише синсетове кроз везивање за онтологије. Такође нуди опис синтаксичких оквира користећи се структуром фразе и прагматичким контекстом (McCrae, Aguado-de-Cea, и остали 2012).

LMF као XML серијализација не предвиђа објављивање лексикона на вебу у виду повезаних података. Он се такође не бави односом лексикона и онтологија. Из претходно изложеног је јасно да *lemon* у овом смислу има предност над LMF-ом.

Једна од главних и за нас пресудних разлика јесте приступ семантици. *Lemon* тежи онтолошком приступу дефинисању значења јер он нуди више могућности за изражавање нијанси и специфичности значења у појединим језицима у односу на модел заснован на лексикону значења. Осим тога на проширењу модела *lemon* се интензивно ради те се појављују нови модули. Током 2019. године су се појавили модули за лексикографију (енг. *The OntoLex Lemon Lexicography Module - lexicog*) и модул за информације о фреквенцијама, потврдама и корпусима (енг. *Module for frequency, attestation and corpus information - FrAC*) (Chiarcos и остали 2020). Модел за лексикографију је настао из потребе да се постојећи лексички ресурси објављују у виду отворених повезаних података (Bosque-Gil и остали 2019) што и јесте главни правац у развоју савремене лексикографије.

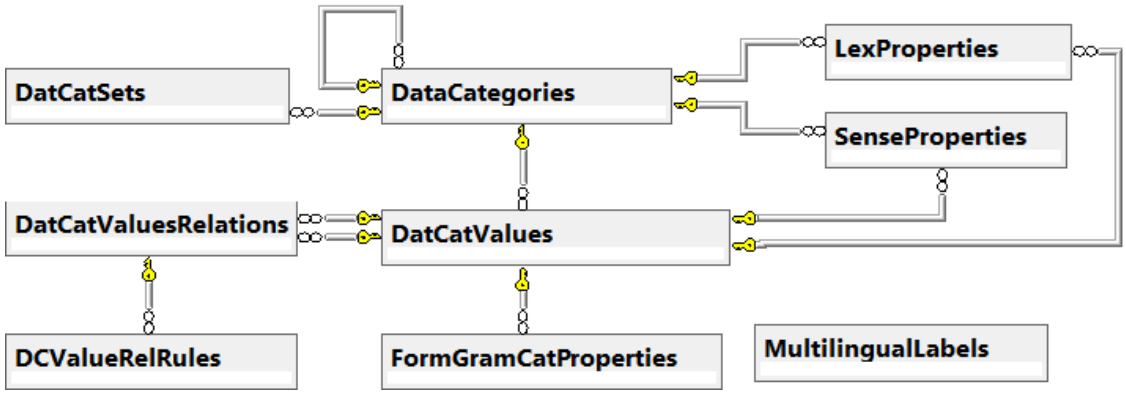
Када је у питању TEI модел, у више извора је наведено да је он више прилагођен за коришћење при опису традиционалних речника за људску употребу (Bański, Bowers, и Erjavec 2017) (Khemakhem, Foppiano, и Romary 2017). С друге стране, у оквиру TEI заједнице постоји велико интересовање за повезивање обележених података са онтологијама и за придруживање Отвореним повезаним подацима (енг. *Linked Open Data, LOD*) (J. T. Bowers и Declerck 2016). Како би се превазишао проблем XML серијализације у контексту повезивања података у пракси постоје примери мапирања TEI речника са SKOS системом који је заснован на RDF синтакси (Declerck, Mörth, и Wand-Vogt 2014) или *lemon* моделом (McCrae, Guadalupe Aguado-de-Cea, и остали 2012). RDF нотација је незаменљива због једноставности изражавања и могућности ефективне интеграције података из различитих извора (Berners-Lee 1998). Као таква представља основу за идеју семантичког веба тј. отворених повезаних података.

4.2 Одабир адекватног модела лексикографске базе и њен развој

4.2.1 Модел развијене лексикографске базе

Да би се обезбедили представљање, одржавање и развој података *Морфолошких речника за српски језик* (поглавље 3.1), у склопу рада на овој дисертацији, развијен је модел лексикографске базе података инспирисан претходно представљеним моделима *lemon* (поглавље 4.1.3) и LMF (поглавње 4.1.2), као и *Регистром категорија података DCR*. Модел лексикографске базе обезбеђује адекватно место за складиштење лексикографских информација у одговарајуће табеле и њихово повезивање са другим лексикографским информацијама уз помоћ релација.

На слици 27 приказан је модел лексикографске базе која складишти граматичке, опште, деривационе, изговорне, варијацијске, синтаксичке, доменске и семантичке маркере коришћене у морфолошким речницима (Stanković, Krstev, и остали 2018). Табела **DataCategories** похрањује информације о категоријама маркера, односно информацију да ли је маркер нпр. граматички, синтаксички, доменски, семантички итд. Табела **DataCategories** је повезана сама са собом што осликава могућност постојања хијерархије категорија. На пример, категорију „sintaksički markeri“ чине категорије „glagolski vid“, „glagolski rod“, „marker rečca“, „marker vrsta i funkcija zamenica“ и други. Записима табеле **DatCatSets** дефинише се опсег примене одређене категорије маркера, тј. на које врсте речи се примењује категорија маркера. На пример, у морфолошким речницима постоји категорија маркера која се односи само на заменице. Пример таквог маркера јесте „+PrsJB“ који означава личну заменицу у једнини без ознаке рода што би биле заменице *mu* и *ja*. Опсег примене маркера може бити и неколико врста речи. Уколико се категорија односи на све врсте речи, запис у табели **DatCatSets** има вредност „MOT“[80]. Табела **DatCatSets** је са табелом **DataCategories** повезана везом више према један јер се првом дефинише врста речи на које се примењује једна категорија из табеле **DataCategories** а једна категорија се може односити на више врста речи. На пример, падеж се може односити на именице, придеве, заменице и бројеве.



Слика 27 Модел лексикографске базе која складишти информације о категоријама

Конкретна вредност маркера, у облику у коме се јавља у речнику, представљена је као запис у табели **DatCatValues**. Више вредности маркера из исте категорије чини једну категорију која је запис у табели **DataCategories**. У складу са тим маркери заменица „+PrsJB“ (лична у једнини без рода), „+PrsMB“ (лична у множини без рода) и „+PrsJG“ (лична у једнини са родом) припадају категорији „marker vrsta i funkcija zamenica“. Оно што је потребно напоменути јесте да је хијерархија, осим на нивоу категорија, могућа и на нивоу вредности категорија (маркера). Ово је посебно битно код доменских маркера. Да би се остварили релације међу вредностима уведена је табела **DatCatValuesRelations.** Она се користи и код повезивања маркера што је од примарних значаја за успостављање веза између одредница, детаљно описаних у одељку 5.2.1. Пример овакве релације била би релација „EkIjk“ која представља везу између маркера „+Ek“ и „+Ijk“, односно везу између екавског и ијекавског облика речи. У табели **DCValueRelRules** описана су појединачно јединствена правила за повезивање која чине један тип релације. Једно од правила за релацију „EkIjk“ прецизира да једна од повезаних речи мора садржати подниску „je“ и маркер „+Ijk“, док друга мора садржати подниску „e“ и маркер „+Ek“. Помоћу овог правила су, на пример, повезане речи „bezbjednost“ и

[80] Од француског *mot* (реч) која се користи у систему Unitex да означи било коју реч.

„bezbednost". Вредности маркера којима се обележавају делови лексичког записа на нивоу значења (табела **LexicalSense** на слици 28) налазе се као записи у табели **SenseProperties**. Овде ће се наћи семантички маркери и маркери домена (нпр. +Hum за људе и +DOM=Geol за домен геологије). Табела **SenseProperties** је веза модела дела базе који складишти информације о категоријама (слика 28) са табелом модела лексикографске базе за речнике DELAS (слика 28) и DELAC (слика 29). Вредности маркера којима се обележавају делови лексичког записа на нивоу **LexicalEntry**, тј. на граматичком нивоу, јесу записи у табели **LexProperties**. Пример таквих маркера је +Ek за екавски изговор. Табела **LexProperties** је такође веза са другим табелама – моделима лексикографске базе за речнике DELAS (слика 28) и DELAC (слика 29). Табелом **FormGramCatProperties** представљене су вредности граматичких категорија које се јављају уз облике речи, односно у DELAF речнику. Примери таквих ознака граматичких категорија јесу „1" за номинатив, „2" за генитив, „m" за мушки род, „f" за женски род итд.

Табела **MultilingualLabels** је креирана са идејом да се метајезик речника представи вишејезично. Тренутно је језик за опис категорија само српски али би се у овој табели могли дефинисати називи категорија на другим језицима, било са циљем описа морфолошког речника неког другог језика или представљања информација из Српског морфолошког речника на другим језицима.

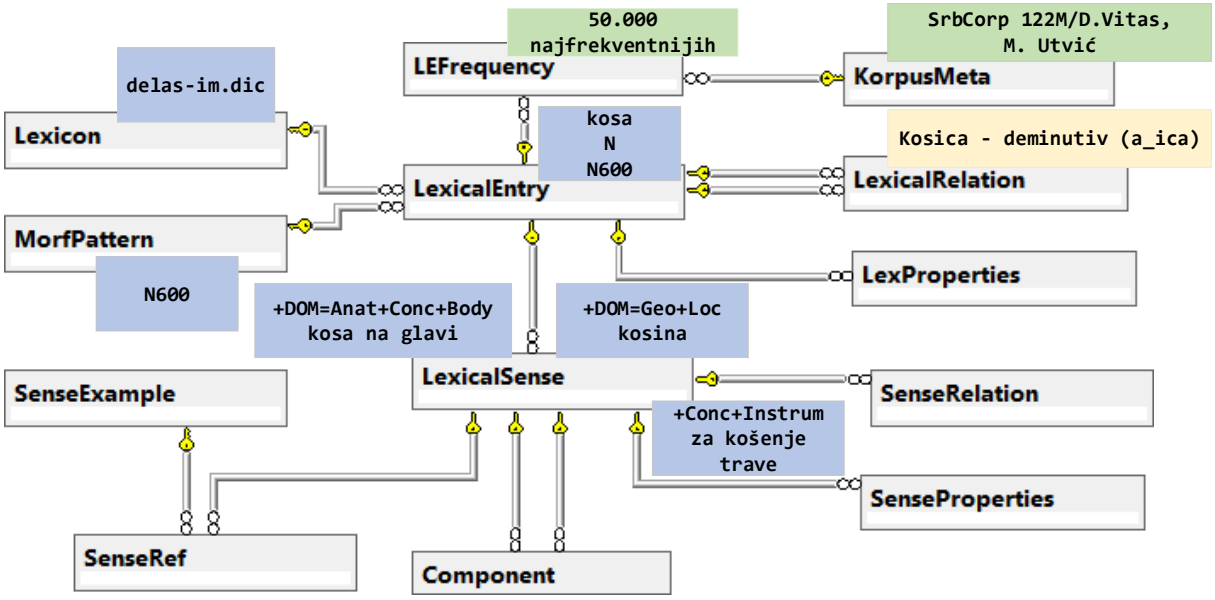
У наставку ће бити описан модел за представљање информација морфолошких речника DELA, као и модел за представљање речника облика DELAF.

На слици 28 представљен је приказ модела лексикографске базе морфолошких речника – DELA, детаљније описаних у поглављу 3.1, заједно са илустрацијом смештаја лексичких информација на примеру лексичког записа *коса,* којим су представљене три хомографне именице[81], који у формату DELA изгледају овако:

```
kosa,N600+DOM=Anat+Conc+Body//kosa na glavi
kosa,N600+DOM=Geo+Loc//kosina
kosa,N600+Conc+Instrum//za košenje trave
```

Табеле из модела су представљене сивом бојом. У правоугаоницима обојеним плавом бојом су представљене лексичке информације из речника DELA, у зеленим правоугаоницима дате су информације везане за корпус, док је у жутом правоугаонику дата информација о вези са другим лексичким записом.

[81] У шестотомном *Речнику Матице српске* овај пример је представљен у виду три лексичка записа за хомограф *кòса.* У Српском морфолошком речнику се акценти не бележе.



Слика 28 Приказ модела лексикографске базе са примером из речника монолексемских речи – DELAS

У табелу **Lexicon** смештају се подаци који описују један речник. Ови подаци се илуструју атрибутима: идентификациони број речника, назив, језик, тип и опис речника. За изабрани пример, атрибут назива речника, има вредност „delas-im.dic". Са табелом **Lexicon** је повезана табела **LexicalEntry** која складишти већи део информација из DELAS речника, односно већи део информација из лексичког записа. За једну репрезентацију табеле **Lexicon,** што је један речник, везује се један или више лексичких записа (репрезентације табеле **LexicalEntry**)**.** У ову табелу се смештају информације које се односе на идентификациони број лексичког записа, лему, канонски облик, тип записа, врсту речи, ознаку морфолошке класе, као и администраторске информације попут белешке, статуса за објављивање лексичког записа, језика или идентификационог броја речника, дакле добар део информација из лексичког записа у формату DELA. Имајући у виду наш пример, овде ће бити смештене лема „kosa", ознака врсте речи „N", ознака морфолошке класе „N600", као и ознака за српски језик „sr". У речнику монолексемских речи ће тип лексичког записа имати вредност „**s**" (енг. *simple*), код полилексемских израза је то вредност „**c**" (енг. *compound*). Табела **LexicalEntry** је везана за табелу **MorfPattern** јер се путем атрибута ове табеле дефинишу карактеристике појединачних морфолошких класа. Сви лексички записи са истом ознаком морфолошке класе су везани за једну репрезентацију табеле **MorfPattern,** што практично значи да су за морфолошку класу „N600" везани сви записи који се флективно мењају по овом обрасцу. Тако ће за ову морфолошку класу бити везани и лексички записи „frankofonija", „fantazmagorija", „odiseja" итд.

Табела **LexicalRelation** представља везе између лексичких записа који су повезани. У повезане записе спадају облици добијени деривацијом, као и варијанте из екавског и ијекавског изговора. Пример „kosa" је у лексичкој релацији деминутив са варијантом „kosica" која се остварује правилом замене наставка „**a**" наставком „**ica**", односно правилом "**a_ica**". Више детаља о лексичким релацијама биће у одељку 5.2.

Табела **LexicalSense** представља везу између лексичког записа и онтологија, што су у нашем примеру доменске ознаке и маркери и илуструје синтаксичко-семантичка значења лексичког записа. Један запис из табеле **LexicalEntry** (лексички запис) може имати више значења, односно бити везан за више записа табеле **LexicalSense**. Наш пример управо илуструје сценарио по коме лексички запис има три

значења. Графемски облик „kosa“ има три различита значења. Прво значење, представљено записом табеле **LexicalSense** је дефинисано ознакама доменског маркера „+DOM=Anat“ за анатомију и семантичких маркера „+Conc“ за конкретну именицу, и „+Body“ за тело што указује да се ради о делу људског или животињског тела. Ниска „kosa na glavi“ је део напомене лексикографа и такође је део табеле **LexicalSense**. Друго значење представљено у табели јесте дефинисано ознакама за доменски маркер географија, означен са „+DOM=Geo“, и семантичким својством са вредношћу „+Loc“ за место, уз напомену лексикографа „kosina“ што сугерише да се ради о значењу косине као географског појма. Треће значење је дефинисано семантичким маркерима за конкретну именицу „+Conc“ и „+Instrum“ за опрему и делове опреме, што сугерише да се ради о значењу алатке за кошење, или како је у напомени аутора записа уписано „za košenje trave“. Сваки појединачни коришћени маркер представља запис у табели **SenseProperties**. Табела **SenseRelation** омогућава повезивање појединачног значења лексичког записа са његовим синонимима али такође омогућава везу са ворднетом.

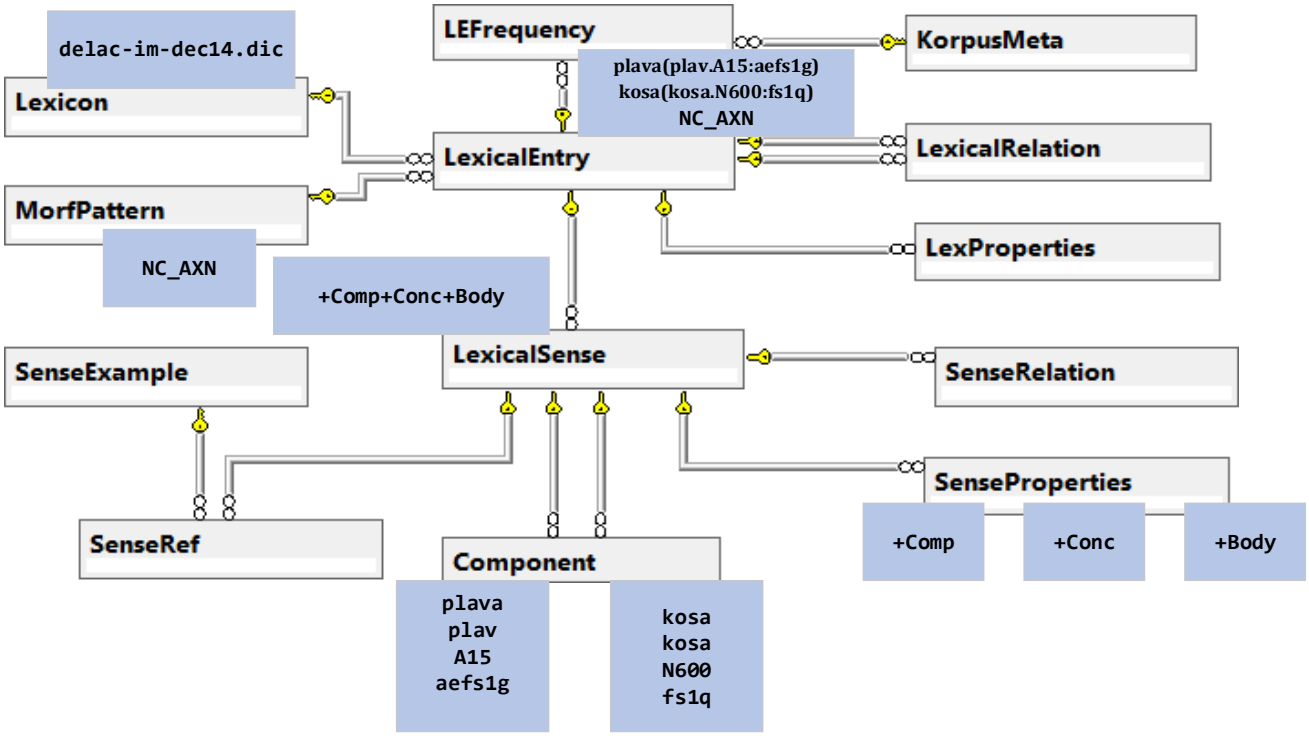
Табела **SenseRef** повезује лексичко значење са примером употребе речи у одређеном значењу, односно табелом **SenseExample**. Практично, у табелу **SenseRef** се смештају информације о библиографском извору из кога потиче пример, док се сам пример налази у табели **SenseExample**.

Табела **LEFrequency** представља везу лексичког записа са табелом **KorpusMeta** и даје информацију о фреквенцији употребе леме записа у одређеном корпусу описаном табелом **KorpusMeta** којом су дефинисани метаподаци корпуса. Како је илустровано на слици 28, лема с графемским обликом *коса* спада у 50.000 најфреквентнијих речи у Корпусу савременог српског језика од 122 милиона речи аутора Душка Витаса и Милоша Утвића (Utvić 2013). Податак о фреквенцији се односи на сва три могућа значења.

Табела **Component** се користи код полилексемских израза за дефинисање компоненти које тај израз формирају, а илустрација употребе ове табеле дата је на слици 29. Дакле, полилексемски израз „plava kosa“ састоји се од две компоненте „plava“ и „kosa“. Овај лексички запис је у речнику DELAC приказан на следећи начин:

```
plava(plav.A15:aefs1g) kosa(kosa.N600:fs1q),NC_AXN+Comp+Conc+Body
```

Атрибутима табеле **Component** описују се појединачне компоненте овог полилексемског израза. Овај запис говори да се ова полилексемска јединица мења по обрасцу „NC_AXN“ што значи да се полилексемска јединица мења по броју и падежу и састоји од именице којој претходи придев који се слаже са именицом у роду, броју и падежу. У табели **Component** се дефинише да је „plava“ облик придева „plav“ који се мења по флективној класи „A15“, и то његов облик дефинисан граматичким категоријама „aefs1g“, односно позитив женског рода једнине у номинативу, а да је „kosa“ именица флективне класе „N600“ која се у леми полилексемске јединице користи у облику чије су граматичке категорије „fs1q“, односно женски род једнине номинатива. Један запис у табели **Component** је једна компонента. Дакле, у табели **Component** складиште се понаособ информације о облицима који чине компоненту.



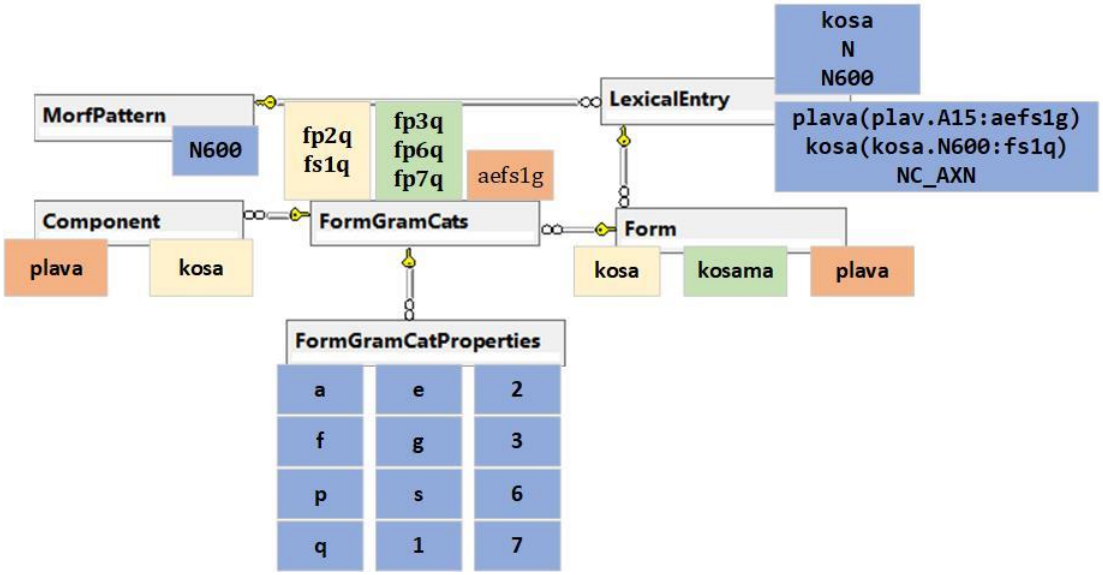
Слика 29 Приказ модела лексикографске базе са примером из речника полилексемских израза – DELAC

Слика 29 приказује модел лексикографске базе са примером смештаја лексичких информација из речника полилексемских израза – DELAC на примеру претходно поменутог лексичког записа *плава коса*. Дакле, информације о речнику чији је део лексикографски запис се налазе у табели **Lexicon**. У овом случају је назив речника „`delac-im-dec14.dic`“. Лема “`plava(plav.A15:aefs1g) kosa(kosa.N600:fs1q)`” је запис табеле **LexicalEntry**, као и информације о врсти лексичког записа „c“ за полилексемски израз и флективном коду „`NC_AXN`“. У табели **LexicalSense** се налази значење полилексемског израза представљено маркерима за вишечлану реч „`+Comp`“, конкретну именицу „`+Conc`“ и тело „`+Body`“. Сваки појединачни маркер је представљен табелом **SenseProperties**.

Модел лексикографске базе података за представљање речника флективних облика DELAF приказан је на слици 30. На слици се налази и пример употребе модела илустрован са два флективна облика монолексемске јединице *коса* (жута и зелена позадина) која су у формату DELAF представљена на следећи начин:

```
kosa,kosa.N:fp2q:fs1q
kosama,kosa.N:fp3q:fp6q:fp7q
```

Први облик „kosa“ одговара групама граматичких категорија „fp2q“ – женском роду множине генитива и „fs1q“ – женском роду једнине номинатива. Други облик „kosama“ дефинисан је категоријама „fp3q“ – женским родом множине датива, „fp6q“ женским родом множине инструментала и „fp7q“ женским родом множине локатива.



Слика 30 Приказ модела лексикографске базе речника DELAF

Табела **LexicalEntry** и табела **MorfPattern** су готово истоветне као у моделу лексикографске базе за DELA речнике. Табела **LexicalEntry** повезана је са табелом **MorfPattern** чији записи представљају различите морфолошке класе, тј. обрасце флективног понашања. Сви лексички записи који се флективно мењају према једној морфолошкој класи су придружени једном запису у табели **MorfPattern**. Табела **Form** је везана за табелу **LexicalEntry** односом више према један јер се у њој налазе сви облици јединице из табеле **LexicalEntry**. Јединица „kosa“ из табеле **LexicalEntry** представљена је на слици са два записа у табели **Form**, облицима „kosa“ и „kosama“. Записи из табеле **FormGramCats**, која садржи могуће комбинације груписаних граматичких категорија које описују тачне облике речи, повезани су са једним записом из табеле **Form,** односно једним обликом речи везом много–један. Дакле, групе граматичких категорија „fp3q“ (женски род множине датива), „fp6q“ (женски род множине инструментала) и „fp7q“ (женски род множине локатива) описују облик „kosama“, што је на слици приказано у зеленим оквирима. Облик „kosa“ и пратеће групе граматичких категорија на слици су приказани жутом бојом. У табели **FormGramCatProperties** похрањује се понаособ свака граматичка категорија. Класе табеле **FormGramCatProperties** су појединачно „f“ – женски род, „1“ – номинатив, „2“ – датив и тако даље. Табела **Component** се користи код одређивања облика полилексемских израза. У случају полилексемског израза *плава коса* табела **Component** складишти информације о компонентама које формирају израз - *плава* и *коса*. Облик *плава* (наранџаста позадина на слици) је дефинисан граматичким категоријама „aefs1g“, односно да се ради о позитиву женског рода једнине у номинативу придева *плав* (флективна класа A15), док је облик *коса* женски род једнине у номинативу именице *коса* (флективна класа N600) дефинисан граматичким категоријама „`fs1q`“.

Примена лексикографске базе на француски језик

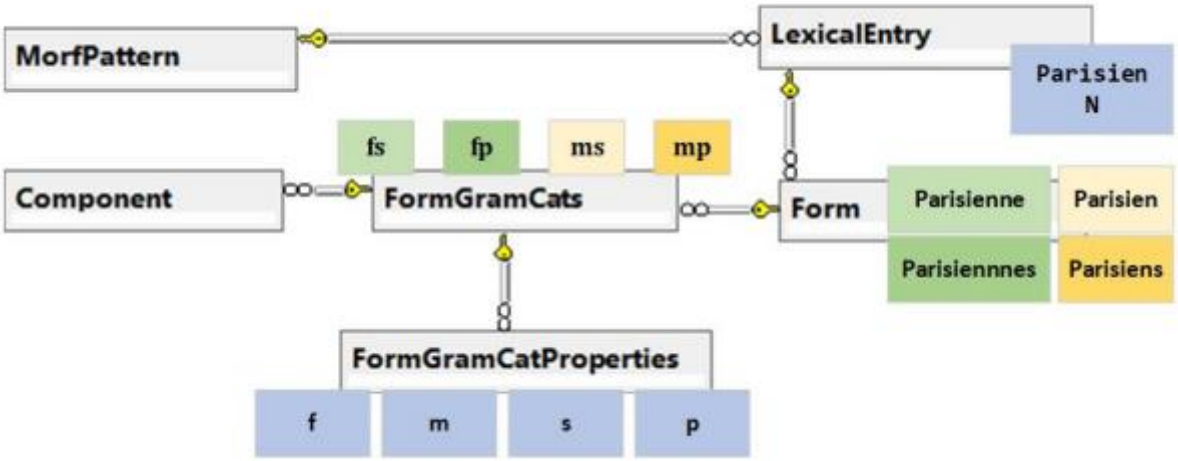
Како бисмо показали независност предложеног модела од језика, илустроваћемо примену лексикографске базе *Лексимирка* на 5 лексичких записа из

Морфолошких речника за француски језик – конкретно из речника чији је назив „*Prolex-Unitex.dic*“[82] (Maurel 2008).

```
Paris,Paris.N+PR+Toponyme+Ville:ms:fs
Parisien,Parisien.N+PR+Hum+Toponyme+Ville:ms
Parisienne,Parisien.N+PR+Hum+Toponyme+Ville:fs[83]
Parisiennes,Parisien.N+Hum+Toponyme+Ville:fp
Parisiens,Parisien.N+Hum+Toponyme+Ville:mp
```

Назив речника ће се наћи у табели **Lexicon**. У табели **LexicalSense** ће се наћи све информације у виду маркера, дакле, маркер за властито име + PR, топоним +Toponyme, град +Ville и човека +Hum (слика 28). Информација о лемама „Paris“ и „Parisien“ и врсти речи смештају се у табелу **LexicalEntry**.

На слици 31 је приказан смештај информација из 4 лексичка записа која представљају флективне облике леме “Parisien”. Лема и врста ређи ће се наћи у табели **LexicalEntry** док ће се сви облици (Parisien , Parisienne, Parisiennes, Parisiens) наћи у табели **Form**. Све граматичке информације које су придружене облицима су похрањене у табели **FormGramCats**, док су појединачне граматичке категорије похрањене у табели **FormGramCatProperties**. На слици су истим бојама обележени облици и њихове пратеће групе граматичких категорија. На пример, жутом бојом су означени облик „Parisien“ и пратеће категорије ms које означавају да се ради о мушком роду једнине.



Слика 31 Приказ смештаја информација *из Морфолошког речника за француски језик* у моделу лексикографске базе (Lazić и Škorić 2019)

С обзиром на то да први лексички запис представља град а наредна 4 означена маркером +Hum представљају становнике тог града, могуће је успоставити релацију деривације. Правило које би повезало први и други запис би било засновано на маркерима +Toponyme+Ville из првог записа и суфикса „ien“ и маркера +Hum из другог записа. Правило које би повезало први и трећи запис би захтевало суфикс „ienne“ код другог записа док би захтевани маркери остали исти као и у претходном правилу. Истом аналогијом би са првим записом била повезана и преостала два записа. Истом релацијом би били повезани лексички записи за градове ”Péone” и ”Plélauff” и лексички записи за њихове становнике ”Péonien” и ”Plélauffien”.

[82] Речник је доступан на следећој адреси: https://github.com/UnitexGramLab/unitex-lingua/tree/master/fr/Dela (приступљено 26.08.2021)

[83] У француском језику се род традиционално сматра флективном категоријом именица које означавају жива бића те је онда *Parisienne* један флективни облик именице *Parisien*.

Наведеним примерима смо показали да је лексикографску базу могуће користити независно од језика речника.

4.2.2 Доменске онтологије

Појам онтологија има двојако значење. Прво и изворно је пореклом из филозофије и представља грану метафизике која се бави постојањем бића. Друго значење се односи на информатички термин и према широко прихваћеној дефиницији Тома Грубера (Tom Gruber) представља „експлицитну спецификацију концептуализације“[84](Gruber 2009). Он под концептуализацијом подразумева: „предмете, појмове и друге ентитете за које се претпоставља да постоје у некој области интересовања и везе које постоје међу њима“[85] (Gruber 1995). Онтологија се најчешће састоји од: класа или концепата (енг. *classes, concepts*), инстанци (индивидуа или примерака) класа, релација (енг. *relations*), својстава или атрибута (енг. *properties, attributes*) и формалних правила (енг. *axioms*). Релације могу постојати између класа, индивидуа или међусобно. Формалним правилима се исказују знања која нису дата експлицитно, тј. релације међу концептима заснована на знању о стварности (Митровић 2018).

Основни разлози за креирање информатичких онтологија данас јесу дељење и поновна употреба знања од стране машина и апликација на вебу. Ако заједнице које се баве истом облашћу користе исти модел података, њихова размена је једноставнија и ефикаснија (Sen и Duffy 2005). Како би подаци били машински читљиви и размењиви на семантичком нивоу између апликација на вебу, неопходно је да информације буду представљене у виду онтологија. Довољно је да на вебу постоји основно доменско знање представљено онтологијом како би се стекао услов за развој интелигентних апликација везаних за специфичан домен. Присутан је став да „онтологије пружају инфраструктуру за трансформацију *веба информација и података* у *веб знања – семантички веб*“[86] (Gašević и остали 2006).

Развој информатичких онтологија подстакнут је потребама вештачке интелигенције и семантичког веба да би се оне данас користиле у многим областима, попут медицине, рударства, библиотекарства, итд. Онтологије се користе за спецификацију речника за размену података међу системима за потребе семантичког веба, као саставни део стандарда W3C. Стандард W3C за представљање онтологија на семантичком вебу препоручује коришћење Језика за онтологије на вебу – OWL (енг. *Web Ontology Language*). RDF је још један од стандарда семантичког веба који се бави концептуалним описом информација и семантичких веза међу електронским изворима. Заснива се на постојању „уређених тројки“ (енг. *triples*): субјекат, предикат и објекат.

Потребно је напоменути да се за убрзање развоја онтологија користе алати или окружења за онтолошко учење (енг. *ontology learning environment*) који се ослањају на области обраде природних језика и обраде текста. Најуопштеније речено, методама поменутих области се врши екстракција информација које се користе у онтологији.

Сходно томе који ниво знања представљају, онтологије могу бити: горње онтологије (енг. *upper ontologies*), доменске онтологије (енг. *domain ontologies*) и

[84]Изворно: „explicit specification of a conceptualization“.

[85]Изворно: „the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them“.

[86] Изворно: „Ontologies provide an infrastructure for transforming the Web of information and data into the Web of knowledge – the Semantic Web“.

онтологије задатака (енг. *task ontologies*). Горње онтологије садрже основни речник који садржи опште концепте који се користе кроз велики број различитих домена. Примери такве онтологије су онтологија SUMO (енг. *Suggested Upper Merged Ontology*), о којој ће бити више речи у наставку, као и помињано Даблинско језгро. Доменске онтологије садрже концепте који припадају једној области знања и описују значења примењена на тај специфичан домен. Онтологије задатака садрже формална правила потребна за извршавање задатака. Ми ћемо се у даљем раду највише бавити доменским онтологијама из области рударства и геологије али ћемо се прво позабавити онтологијом SUMO.

Горња онтологија SUMO је настала 2000. године спајањем више јавно доступних онтолошких садржаја у једну структуру. Ови онтолошки садржаји су укључивали постојеће онтологије: онтологије доступне на серверу Ontolingua, горњу онтологију Џона Сове (John Sowa) и онтологије које је развио *Институт за биомедицинске технологије Националног савета за истраживања у Риму* (Institute for Biomedical Technologies - Consiglio Nazionale delle Ricerche, ITBM-CNR). Осим тога укључено је и више мереотополошких теорија[87] (Niles и Pease 2001). Након одабира релевантног садржаја из свих придружених извора, извршено је њихово повезивање са веб-страном радне групе SUO (Standard Upper Ontology Working Group), као и превођење на језик SUO-KIF (Knowledge Interchange Format) путем синтаксичког обједињавања.

Последња верзија базе података онтологије SUMO (верзија 3.0 из 2021. године) доступне преко алата/веб-стране сигма (Sigma) садржи 13.552 термина, 193.208 аксиома и 6.231 правило. Онтологија је осим у KIF језику, доступна још и кроз језике OWL, TPTP (енг. *Thousands of Problems for Theorem Provers*) и traditionalLogic.

Велике предности онтологије SUMO јесу њена величина (највећа горња онтологија у отвореном приступу), висок број инкорпорираних формалних правила, као и инстанци повучених из *DBPedia*-е. Онтологија SUMO је такође поравната са Принстонским ворднетом а преко њега и са другим ворднетима, између осталог и са Ворднетом за српски језик. Све наведено онтологију SUMO чини погодном за истраживање, задатке аутоматског резоновања, примену при семантичком проширењу упита, као и у лингвистици.

Доменска онтологија из области рударства - *РудОнто*[88] је развијана под окриљем Рударско-геолошког факултета на темељу терминолошких ресурса који су настали током рада на различитим рударским пројектима. Такви ресурси су таксономија роторног багера, таксономије везане за рударску опрему као и портал посвећен геолошкој терминологији и номенклатури развијен у оквиру пројекта *Геолошког информационог система Србије - GeolISS*. Главни циљ онтологије *РудОнто* јесте структурирање и ефикасна употреба знања у области рударства. Радом на тези која се бавила креирањем система пословне интелигенције за управљање заштитом на раду у рударству онтологија је проширена делом који се бави заштитом на раду у рударству (Kolonja 2016). Тако проширена онтологија *РудОнто* садржи близу 7.000 термина из рударства на српском језику који илуструју најчешће коришћене концепте. Присутно је и око 1.200 еквивалената термина на енглеском језику, као и незнатан број на француском, руском и другим језицима. Најчешће коришћена лексичка релација јесте синонимија којом су повезани термини који чине један концепт. С друге стране, најчешће

[87] Мереотопологија је теорија првог реда која утеловљује мереолошке и тополошке односе међу целинама, деловима и границама између делова.

[88] Онтологија *РудОнто* је доступна за прегледање путем: http://rudonto.rgf.bg.ac.rs/ (приступљено 10.10.2019)

употребљавана семантичка релација јесте „хипоним/хипероним“ којом се обележавају надређени и подређени тј. специфични концепти. Онтологија је намењена проналажењу информација и контроли података који се користе у оквиру система пословне интелигенције (Kolonja и остали 2016).

Језик за обележавање у геонаукама - *GeoSciML* (енг. *Geoscience Markup Language*) представља модел података и стандард за трансфер података који се користе у геонаукама. Настао је на иницијативу Британског геолошког завода (British Geological Survey - BGS) на темељима постојећег *Географског језика за обележавање - GML* (енг. *Geography Mark-up Language*) који садржи компоненте за геометријски координатни систем, тополошке и временске информације, дефиниције и речнике, јединице и мере и постојеће апликације XMML (енг. *eXploration and Mining MarkupLanguage*) која је била намењена потребама истраживања лежишта и рударској индустрији (Sen и Duffy 2005). *GeoSciML* је заснован на језику XML. Тренутно актуелна верзија језика *GeoSciML* 4.1 из 2016. године покрива домен геологије – геолошке јединице, стратиграфију, геолошке структуре и времена, геоморфологију, геохемију и материјале и карактеристике које се појављују при узорковању у геонаукама – подаци везани за бушотине, геолошке узорке и лабораторијско испитивање (Boisvert, Raymond, и Sen 2016). Језик *GeoSciML* користе бројни геолошки пројекти, између осталих и пројекат *OneGeology Europe* који има за циљ да се геолошки просторни подаци које поседују европски геолошки заводи усагласе и постану дељиви изван националне геолошке заједнице (Благојевић и остали 2014). Кључна улога језика *GeoSciML* јесте у томе што власници геолошких скупова података, ради унификације мера и термина, не морају да мењају своје податке како би их хармонизовали са осталима, већ их преводе у структуру *GeoSciML*-а. Употреба стандардизованих речника из области геонаука доприноси конзистентности у семантичком језику који се користи међу скуповима података.

Језик за обележавање ресурса Земље - *EarthResourceML* (енг. *Earth Resource Markup Language*) је такође стандард за размену података заснован на језику XML. Служи за размену информација о својствима која се користе у областима минералних ресурса, рудника и рудничких активности. *EarthResourceML* је заснован на претходно описаном језику и стандарду *GeoSciML*, тј. представља његову надградњу намењену опису геолошких материјала са минералним лежиштима. *EarthResourceML* различити пројекти користе као стандард за размену података широм света. Глобални пројекат *OneGeology*[89] га користи за потребе портала за приказ геолошких мапа и минералних ресурса. Пројекти *AuScope*[90] и *AUSGIN*[91] га користе за испоруку података о минералним ресурсима на државном и територијалном нивоу у оквиру Аустралије. Пројекат *European INSPIRE Directive*[92] користи *EarthResourceML* као стандард за размену информација о минералним ресурсима међу земљама Европске уније.

[89] http://www.onegeology.org/ (приступљено 15.08.2019)

[90] https://www.auscope.org.au/ (приступљено 15.08.2019)

[91] http://portal.geoscience.gov.au/ (приступљено 15.08.2019)

[92] https://inspire.ec.europa.eu/ (приступљено 15.08.2019)

Мапирање категорија података из *Морфолошких речника српског језика*

Током рада на овој дисертацији смо, како би се обезбедила контрола података и омогућила допуна *Морфолошких речника српског језика* (*МРС*), поредили категорије које се у њему користе са концептима онтологије SUMO и скраћеницама које су у употреби у *Речнику САНУ*.

У прилогу 1 је приказано поређење (мапирање) односа категорија података које се користе у оквиру *МРС* у односу на скраћенице које се користе или су коришћене у *Речнику САНУ*[93], као и однос категорија података *МРС* у односу на концепте онтологије SUMO. Однос ознака категорија података *МРС* (колоне „МРС“) и квалификатора коришћених у *Речнику САНУ* (колона „РСАНУ“) приказан је ознакама из колоне „однос са РСАНУ“. Универзална значења коришћених ознака односа су:

- = - једнака употреба,
- [- поређени појам је шири у односу на појам са којим се пореди и
-] - поређени појам је ужи у односу на појам са којим се пореди.

Две колоне „МРС“ постоје у случају када су две ознаке категорија података *МРС* истим односом упоређене са једним квалификатором из *Речника САНУ* или концептом онтологије SUMO. У колони „МРС“ је поред ознаке категорије података у угластим заградама дато додатно појашњење уз вредност „А“ јер ова ознака може означавати граматичку категорију (gramCat) - аорист, како је приказано у 8. реду, и врсту речи (POS) - придев, како је приказано у 10. реду Прилога 1. Граматичке категорије падежа у *МРС* јесу у односу једнакости са квалификаторима који означавају падеже у *РСАНУ*, нпр. ознака „З“ из МРС је једнака квалификатору „дат.“ за датив из РСАНУ. Флективни облици у речнику DELAF у *МРС* су означени ознаком за одговарајући падеж, док се у *РСАНУ* квалификатори којима се означавају падежи користе у одређеним случајевима (за морфолошка одступања у парадигми или када је потребно назначити рекцију одреднице). У оба речника се једнако користе и ознаке за домене ботанике, грађевинарства, геологије, медицине, итд. С друге стране у РСАНУ се користи ознака за народну медицину (нар. мед.) која се у МРС не користи али је одређена као ужи појам у односу на ознаку домена медицине „DOM=Med“ која се у МРС користи. Пример категорије података из *МРС* која је ужа у односу на скраћеницу из *РСАНУ* јесте „HumColl“ (људи збирно) у односу на скраћеницу „зб. им.“ (збирна именица).

Поређење категорија МРС са концептима онтологије SUMO показује да су нпр. у једнаком односу категорија семантичког маркера за научну област „FoS“ и истоименог концепта из онтологије SUMO „FieldOfStudy“. Једнаки су и семантички општи маркер за водену површину „Hyd“ и концепт „WaterArea“ (водена површина) из онтологије SUMO.

Смисао поређења ознака из речника *МРС* са ознакама из *Речника САНУ* и онтологије *SUMO* јесте у контроли података и потенцијалној аутоматској допуни речника. Ово поређење може допринети преиспитивању коришћених ознака, у смислу да ли су оне адекватне, да ли је скуп ознака потпун. Ако би РСАНУ био у дигиталном формату, била би могућа аутоматска анализа означених лексичких чланака и евентуална допуна на основу лексичких записа из МРС. Поређење са концептима из онтологије SUMO могло би допринети повезивању једнако обележених лексичких записа из *Морфолошких речника српског језика* са инстанцама класа из онтологије SUMO.

[93]Списак квалификатора коришћених у Речнику САНУ може се пронаћи у раду: (Стијовић и Станковић 2018).

Други пример мапирања које ћемо приказати јесте мапирање категорија које се користе у *МРС* са категоријама које се користе за врсте речи у оквиру *Универзалних зависности* (енг. *Universal Dependencies, UD*). *Универзалне зависности* представљају оквир за универзално означавање граматичких категорија у преко 100 различитих језика настао са циљем да се омогући парсирање у различитим језицима. У табели 6 је приказано повезивање ознака категорија које се користе у МРС са ознакама из скупа за српски језик наведеним у UD[94]. Прва колна представља ознаке из МРС, док друга колона представља ознаку из сета за српски језик UD. Трећом колоном је дат опис датих категорија. Приметићемо да се доста често за једну ознаку из UD користи комбинација више ознака из *МРС*. Тако ће се за помоћни глагол у *МРС* користити две ознаке: ознака за врсту речи глагол V у комбинацији са маркером глагола +Aux чије присуство у лексичком запису показује да је глагол помоћни. С друге стране ће ознака за глагол који није помоћни VERB из UD свој еквивалент у *МРС* имати у ознаци за глагол без маркера +Aux. Ознака за заједничке именице из UD би означавала записе означене са N али без маркера +NProp за властите именице у МРС, док ознака за властите именице PROPN представља еквивалент комбинацији ознаке за именицу N и маркера за властиту именицу +NProp у МРС. Ознака за скраћенице, префиксе, стране речи и друго X из UD спаја ознаке за две постојеће врсте речи из МРС – ABB и PREF. Ознаке за знак интерпункције PUNCT и симбол SYM из UD немају еквивалент у ознакама из МРС пошто се у речнику налазе само речи (у смислу ниски алфабетских симбола).

Табела 6 Мапирање ознака из МРС са ознакама за врсту речи из UD

Ознаке из МРС	**Ознака из сета за срп. јез. UD**	**Опис**
A	ADJ	придев
PREP	ADP	предлог
ADV	ADV	прилог
V+Aux	AUX	помоћни глагол
CONJ+Cconj	CCONJ	напоредни везници
PRO+ProA	DET	придевска заменица
INT	INTJ	узвик
N без +NProp	NOUN	заједничка именице
NUM	NUM	број
PAR	PART	речца
PRO без +ProA	PRON	заменица
N+NProp	PROPN	властита именица
	PUNCT	знак интерпункције
CONJ+Sconj	SCONJ	зависни везници (субординативни везници)
	SYM	симбол
V без +Aux	VERB	глагол који није помоћни

[94] https://universaldependencies.org/treebanks/sr_set/index.html (приступљено 13.03.2021)

ABB, PREF	X	скраћенице, префикси, стране речи и друго

Наведене ознаке из UD су увезене у лексикографску базу па су, на основу мапирања, лексичким записима придружене одговарајуће вредности ознака које су видљиве кроз приказ лексичког записа. Претрага лексикографске базе помоћу врсте речи из UD је могућа коришћењем префикса +UPOS=OznakaIzUD. Тако бисмо, на пример, глаголе који нису помоћни излистали претрагом својства „+UPOS=VERB“ (више о претрази лексикографске базе следи у одељку 6.2).

Мапирање значења једнојезичних речника

У склопу пројекта ELEXIS, пре свега за потребе радионице GLOBALEX, развијен је ресурс који повезује значења речи у различитим једнојезичним речницима (Ahmadi и остали 2020). Овакав ресурс проналази примену у тренирању статистичких модела и евалуацији резултата у обради природних језика за потребе поравнања значења речи и откривања семантичких веза. Ресурс обухвата 15 језика и 17 скупова података. И српски језик је, захваљујући раду чланова Друштва за језичке ресурсе и технологије (делом и аутора ове дисертације), заступљен као скуп података и то у виду поравнања значења између *Ворднета за српски језик* и *Речника српскохрватског књижевног језика I-VI Матице српске*. За повезивање два значења су коришћени односи једнакости, ширег, ужег, повезаног и неповезаног. Повезано је 599 значења из *Речника Матице српске* са 1.768 значења из *Ворднета.* Ови подаци су сада доступни у формату JSON (енг. *JavaScript Object Notation*). Следи пример поравнања значења речи *рука.* Линијом 2 је исказано да се ради о именици. Од линије 6 креће навођење значења из *Речника Матице српске.* Ознаком „`#text`“ су дате дефиниције значења (нпр. линија 8) праћене екстерном идентификационом ознаком (линија 9) и ознаком значења у речнику (линија 10). Од линије 63 креће опис значења из *Ворднета за српски језик*. Од линије 73 се наводи поравнање значења. Поравнање значења је исказано на нивоу описа значења уместо коришћења ознака значења (external_ID) јер је то био избор руководилаца пројекта. Ознаком „`sense_source`“ означена су значења из *Речника Матице српске* док су ознаком „`sense_target`“ означена значења из *Ворднета за српски језик.* Значење из *Речника Матице српске „један од горњих удова човечјег тела, екстремитета од рамена до врхова прстију“* је релацијом једнакости (`"semantic_relationship": "exact"`– линија 77) повезано са значењем из Ворднета: „*Човеков горњи уд који почиње од рамена*.“. Друго поравнање се односи на уже значење (`"semantic_relationship": "narrower"` – линија 82) између значења из *Речника Матице српске* "*крајњи део тога екстремитета, од зглоба или шаке до врхова прстију*.“ (линија 80) и значења "*Екстремитет горњих удова*.", из ворднета.

```
1{
2            "lemma": "рука",
3            "part-of-speech_tag": "noun",
4            "gender": "",
5            "meta_ID": "",
6            "resource_1_senses": [
7                {
8                    "#text": "један од горњих удова човечјег тела, екстремитета од
рамена до врхова прстију.",
9                    "external_ID": "ruka.N.1a.",
10                    "RMS_label": "1a."
11                },
12                {
13                    "#text": "крајњи део тога екстремитета, од зглоба или шаке до
врхова прстију.",
14                    "external_ID": "ruka.N.16.",
15                    "RMS_label": "16."
```

```
                },
                {
                    "#text": "извршилац каквог посла, радник.",
                    "external_ID": "ruka.N.2a.",
                    "RMS_label": "2a."
                },
                {
                    "#text": "посао, рад, делатност.",
                    "external_ID": "ruka.N.2б.",
                    "RMS_label": "2б."
                },
                {
                    "#text": "власт, моћ, надлежност, својина",
                    "external_ID": "ruka.N.4",
                    "RMS_label": "4"
                },
                {
                    "#text": "начин, метод (рада, понашања и сл.).",
                    "external_ID": "ruka.N.5б.",
                    "RMS_label": "5б."
                },
                {
                    "#text": "друштвени слој, сталеж, ранг.",
                    "external_ID": "ruka.N.5в.",
                    "RMS_label": "5в."
                },
                {
                    "#text": "извор, порекло.",
                    "external_ID": "ruka.N.6",
                    "RMS_label": "6"
                },
                {
                    "#text": "страна (лева или десна).",
                    "external_ID": "ruka.N.7",
                    "RMS_label": "7"
                },
                {
                    "#text": "пристанак на брак.",
                    "external_ID": "ruka.N.8",
                    "RMS_label": "8"
                },
                {
                    "#text": "слобода рада, деловања, поступања, иницијатива",
                    "external_ID": "ruka.N.9",
                    "RMS_label": "9"
                }
            ],
            "resource_2_senses": [
                {
                    "#text": "Човеков горњи уд који почиње од рамена.",
                    "external_ID": "ENG30-05563770-n"
                },
                {
                    "#text": "Екстремитет горњих удова.",
                    "external_ID": "ENG30-05564590-n"
                }
            ],
            "alignment": [
                {
                    "sense_source": "један од горњих удова човечјег тела,
екстремитета од рамена до врхова прстију.",
                    "sense_target": "Човеков горњи уд који почиње од рамена.",
```

```
                    "semantic_relationship": "exact"
                },
                {
                    "sense_source": "крајњи део тога екстремитета, од зглоба или шаке до врхова прстију.",
                    "sense_target": "Екстремитет горњих удова.",
                    "semantic_relationship": "narrower"
                }
            ]
        }
```

5. Формирање лексикографске базе српског језика у *Лексимирки* и успостављање унутрашњих и спољашњих веза

5.1 Формирање лексикографске базе

Након успостављања модела лексикографске базе *Лексимирка*, у склопу рада на овој дисертацији, развијене су процедуре за миграцију података из речника у новонасталу базу података. Ове процедуре су укључене у постојећу апликацију *Лексимир* (Krstev и остали 2013) како би у време тестирања *Лексимирке* био подржан истовремени развој *Морфолошких речника за српски језик* кроз обе апликације. Старију апликацију *Лексимир* одликују рад над системом датотека уместо рада над базом података који постоји код *Лексимирке* (Stanković, Krstev, и остали 2018). Самим тим стара апликација нема референцијални интегритет података. Стара апликација је десктоп апликација која није дозвољавала истовремени вишекорисnички рад. Један од разлога за паралелни развој обе апликације јесте и лакши прелазак на нову развојну средину. Процедуре за миграцију су аутоматске и укључују структурирање информација о лемама и облицима како би се подаци мапирали са пољима у лексикографској бази података.

Записи из речника DELAS су мапирани са табелама **LexicalEntry** и **LexicalSense**. У поља табеле **LexicalEntry** су смештене информације о леми и флективној класи. Део записа који се односи на све врсте маркера смештен је у поља табеле **LexicalSense**. Лексички записи који су хомографи и имају исту флективну класу су у лексикографској бази представљени као један лексички запис у табели **LexicalEntry**. Маркери који означавају њихова значења су, с друге стране, представљени засебно кроз табелу **LexicalSense**. Раније поменути пример хомографне именице „kosa“ (поглавље 4.2.1) илуструје овај пример који је приказан на слици 28. Лема „kosa“ заједно са ознакама врсте речи (N) и флективне класе (N600) се налази у табели **LexicalEntry**, док су маркери који дефинишу појединачна значења која представљају косу на глави (+DOM=Anat+Conc+Body), косину (+DOM=Geo+Loc) и инструмент за кошење траве (+Conc+Instrum) представљени као засебни записи у табели **LexicalSense**. Лексички записи монолексемских речи који су компонента полилексемске јединице су повезани са том полилексемском јединицом путем табеле **LexicalSense** (слика 29) коришћењем табеле **Componet.**

Записи речника DELAF мапирани су са моделом лексикографске базе приказаним на слици 30. Ако се задржимо на примеру записа „kosa“, сви његови флективни облици („kosa“, „kosama“, „kose“, „kosi“, „koso“ и „kosom“) ће бити смештени у табели **Forms.** Скуп граматичких категорија које описују један облик биће дат табелом **FormGramCats**. У пракси се дешава да један облик речи може бити описан уз помоћ више скупова граматичких категорија сачуваних у табели **FormGramCats**, као што је случај код облика „kosi“ који је описан скуповима категорија „fs3q“ и „fs7q“ тј. представља женски облик једнине датива али и локатива. Појединачне категорије су мапиране са табелом **FormGramCatProperties**.

За потребе представљања категорија података развијен је модел базе представљен сликом 27. Нова лексикографска база садржи све категорије података које су до сада биле у употреби у речницима (различите врсте маркера, информација и домена) али су сада категорије представљене у хијерархијском односу како би била омогућена њихова контролисана употреба. Ово представља новину у односу на ранији линеарни начин представљања категорија. Хијерархијска структура значи да је, на пример, ознака за граматички број једине „s“ представљена као потомак категорије за граматички број која је опет једна од врста граматичких категорија. Детаљи о начину смештаја категорија у базу дати су у делу 4.2.1.

5.2 Развој модела повезивања речи у лексикографској бази и креирање механизама за полуаутоматско хармонизовање и усклађивање

5.2.1 Повезивање одредница у е-речнику – врсте веза

Новонастала апликација *Лексимирка,* ослањајући се на лексикографску базу, пружа опцију повезивања одредница, односно лексичких записа, што је новина у *Морфолошким речницима српског језика*.

Постоји више типова релација између одредница *у Морфолошким речницима српског језика*. Оне се начелно према типу могу поделити на варијационе и деривационе релације. У посебну врсту релације се може сврстати изговорна релација „Ek-Ijk", односно релација која повезује лексичке записе екавског и ијекавског изговора, нпр. *бео* и *бијел*, *снешко* и *сњешко*, итд.

Ове везе су у начелу успостављене и у традиционалним речницима српског језика. На пример, за случај екавског, односно, ијекавског изговора шестотомни речник *Речник Матице српске*[95] изузетно даје двострану релацију између два речничка чланка која дефинишу екавски и ијекавски облик речи (*бео* и *бијел*). У већини случајева се у екавском речничком чланку само указује на постојање ијекавских облика (*бедник* и *биједник*).

бедник, ијек. биједник, м [...]

бео, бела, бело, ијек. бијел и био [...]

бијел-, ек. бел-.

бијел, -ела, -ело, ек. бео.

И у *Речнику САНУ* се поступа на овај начин, с тим што се скраћеницом „ијек.", односно у првом тому, скраћеницом „ј." (јужни изговор) уводи ијекавски изговор који је равноправан екавском изговору. У предговору првом тому *Речника* наведено је да су речи јужног изговора унете у *Речник* као засебне одреднице али без дефиниција, примера и других података. Са тих речи се упућује на исте речи екавског (источног) изговора, где се ијекавске варијанте обрађују као врста дублета. Уз екавске варијанте се дају сви подаци са дефиницијама и примерима (*бео* и *бијел*) (*Речник српскохрватског књижевног и народног језика. Књ. 1, А-Богољуб* 1959). Ради уштеде простора, ијекавске варијанте које припадају већој породици речи са истим кореном се не уносе као одреднице, већ се као одредница наводи корен са цртицом који упућује на исти екавски корен (*бијед-, бијел*-).

бедник ј. **биједник** м [дефиниције и примери]

бео, бела, бело ј. **бијел** (покр. био), бијела, бијело (дијал. бел, бела, бело ј. бијел, бјела, бјело; одр. бели, -а, -о ј. бијели, -а, -о; комп. бељи, дијал. белији, ј. бјељи, дијал. ј. бјелији) [дефиниције и примери]

бијел[2] в. бео.

бијел- в. бел-

биједˉ в. бед-

У једнотомном *Речнику Матице српске* наводе се екавски облици праћени (и)јекавским изговором уведеним скраћеницом „јек.".

бедник јек. **биједник** [...]

[95] Видети предговор *Речника*.

бео, бела, -о јек. **бијел**, бијела, -о (одр. бели јек. бијели; комп. б(ј)ељи) [...]

Варијантни облици, тј. дублети са гласовним разликама или различитим префиксима или суфиксима дати су у шестотомном *Речнику Матице српске* са дефиницијом у уобичајеном облику, док се сви други облици упућују ка том облику (*кафа* и *кава* и *моделирати*, *моделисати* и *моделовати*), или су дати као упоредне одреднице у истом чланку (*адресирати* и *адресовати*).

кава ж тур. = кафа **1.** бот. *тропска грмолика биљка зрнаста плода* Coffea arabica; *њезин плод*. **2.** *напитак приређен од пржена кавина плода или његова сурогата*. — Имамо каве и шећера. *Дом*. **3.** покр. *в. кавана*. — Силни Турци из каве иђаху. *НП Вук*. У исто доба навраћали [су] у Перину каву. *ЛМС 1951*.

кафа ж тур. = кава. — У господским кућама има једна соба гдје се пече кафа. *Маж. М*. Синоћ ... мимо кафу ја сам турску прошла. *НПХ*

моделирати, -елирам сврш. и несврш. *обликовати, уобличити, уобличавати*; *вајати*. — Његове слике ... пластично моделирају оно што се не види. *Бен.*

моделисати, -ишем сврш. и несврш. *моделирати.* — Неће никад моћи да женско тело моделишу као што може да га моделише модерна хаљина. *Цар М.*

моделовати, -лујем сврш и несврш. *моделирати.* — Странац ће штошта моделовати по своме ћефу. *Цар Е.*

адресирати, адресирам и **адресовати,**-сујем сврш. и несврш. **1.** *(на)писати адресу*. **2.** *упутити, упућивати*. – Видео је на столу лист ... адресован на газдино име. *Ђон*

И у једнотомном *Речнику Матице српске* варијантним облицима се приступа на истоветан начин, с тим што облик *моделисати* није обухваћен због обима речника. Следе речнички чланци из овог речника:

кава ж в. *кафа.*

кафа ж тур. **а.** бот. *назив за зимзелене биљке Coffea из ф. Rubiaceae од којих се највише гаје C. arabica и C. liberica*; *семе ове биљке*. **б.** *напитак справљен од прженог и самлевеног семена ове биљке*. **в.** разг. *количина кафе (б) која запрема одређену посуду (шољи(иц)у или сл.)*: попити две кафе. [...]

моделирати, -елирам и **моделовати,** -лујем свр. и несвр. **а.** ум. *уобличити, уобличавати, обликовати ликовно дело (најчешће вајарско)*. **б.** *израдити, израђивати модел.* **в.** фиг. *да(ва)ти некоме или нечему изразит облик, рељефно приказ(ив)ати; уопште уобличити, уобличавати, обликовати*.

адресирати, -есирам (адресовати, -сујем) свр. и несвр. **1.** *(на)писати адресу на писму или другој пошиљци; упутити, упућивати на нечију адресу*. **2.** (некоме нешто) *наменити, упутити некоме*. – Те увреде су биле адресиране нама.

У *Речнику САНУ* су ови типови дублета повезани тако што су дефиниција и примери дати код стандардног облика, док су код облика који се упућује на стандардни дати примери његове употребе (*кафа* и *кава*, *моделовати* и *моделирати* и *моделисати*, *адресовати* и *адресирати*). За упућивање се врши скраћеница „в“ (види).

кава ж *в. кафа*. [примери]

кафа ж (ген. мн. кафа) (тур. kahve); исп. кава [дефиниције и примери]

моделирати (се), -елирам (се) (-ирају) свр. и несвр. в. *моделовати (се).* [примери]

моделисати (се), - ишем (се) свр. и несвр. в. *моделовати (се).* [примери]

мòделовати, -лујём свр. и несвр. [дефиниција и примери]

адресирати (се), адресирам (се) (-ирају) свр. и несвр. (нем. adressieren) в. *адресовати (се).* [примери]

адресовати, -сујем (аор. адресова) свр. и несвр. исп. адресирати (се), атресирати (се) [дефиниције и примери]

Одреднице деривационих облика се у шестотомном *Речнику Матице српске* повезују преко дефиниција које садрже основне облике (*радник*, *радница* и *раднички*)(Стијовић, Крстев, и Станковић 2021).

радник м **1. а.** *онај који ради, обавља одређени посао, онај који се бави одређеном делатношћу (било умном, било физичком).* — Умјетник је радник, а сваки радник је умјетник у неку руку. *Матош*. Око Лицеја окупља [се] група културних и јавних радника. *Милис*. **б.** *онај који се претежно бави физичким радом, који обавља физичке послове у предузећу, у индустрији и сл.* — Томе каменоломном раднику експлодирале [су] у џепу динамитне патроне. *Јонке*. **в.** *човек који воли да ради, онај који добро ради, радиша. И-Б Рј*. **2.** *мрав који ради (добавља храну, одржава гнездо, негује младе).* — У друштву термита извршена је подела рада ... Радници одржавају гнездо, добављају храну ... негују младе. *Станк. С*.

радница ж 1. *жена радник.* — Удавила [се] у Дунаву једна радница. *Уск*. **2.** (обично с додатком: пчела) зоол. в. *радилица* (*3*).

раднички -а, -о *који се односи на раднике, који је у вези с радницима.* ~ класа, ~ самоуправљање, ~ савет, ~ покрет.

И кроз једнотомни *Речник Матице српске* се деривационе везе успостављају преко дефиниција које садрже основне облике. Следе изводи речничких чланака за одреднице *радник, радница и раднички*.

радник м **1. а.** *онај који ради, који обавља одређени посао, онај који се бави одређеном делатношћу, занимањем, делатник на одређеном подручју, у одређеној области друштвеног живота*; […] **б.** *онај који се претежно бави физичким радом, особа која обавља физичке послове у индустрији, занатству, трговини и која за утрошену енергију прима плату; у ширем смислу – свака запослена особа (првенствено припадник радништва као класе)*: индустријски ~, металуршки ~, занатлијски ~, ~ у трговини, угоститељски ~. **в.** *особа која воли да ради и која добро ради, вредна, марљива особа, радиша*. — Радник је то, нема му премца. 2. *мрав који ради (добавља храну, одржава гнездо и негује младе)*. […]

радница ж 1. *женска особа радник*.

раднички прил. *на начин радника, као радник, радници*: ~ обучен.

Речник САНУ такође деривационе везе представља махом путем дефиниција што ћемо илустровати кроз речничке чланке *витез* и *витешкиња*.

витез м (вок. витеже; мн. витезови и витези, ретко витежеви) [дефиниције и примери]

витешкиња ж *жена витез (припадница витешког племићког сталежа; јунакиња)*; исп. витезица. [примери]

Кроз апликацију *Лексимирка* везе између екавског и ијекавског облика речи успостављају се на основу маркера за изговор - екавски „+Ek“ и ијекавски „+Ijk“ којима су обележени лексички записи. Врста речи није од значаја за ову релацију. Следи пример лексичких записа из *Морфолошких речника српског језика* повезана релацијом „Ek-Ijk“:

```
beo,A38+Col+Ek
bijel,A14+Col+Ijk
bednik,N10+Hum+Ek
bjednik,N10+Hum+Ijk
```

Правило којим су повезана прва два лексичка записа дефинише да један лексички запис мора садржати подниску *„ео“* док други мора садржати подниску *„ијел“*. Овим правилом је успостављено 7 релација. Неки од повезаних парова су: *цео* и *цијел*, *сребрнобео* и *сребрнобијел*, као и *полубео* и *полубијел*. У другом случају, правило је да један лексички запис садржи подниску *„е“* док други мора да садржи подниску *„ије“*. На овај начин је успостављена 101 веза а неки од повезаних парова су: *осменути* и *осмијенути*, *петао* и *пијетао*, *препис* и *пријепис*.

У варијације спадају релације којима се повезују два лексичка записа која представљају језичке варијанте речи истог значења нпр. *дедуцирати* и *дедуковати*, *кафа* и *кава*, *хладан* и *ладан* итд. Како би се успоставила ова веза између два лексичка записа потребно је да они буду обележени адекватним маркерима. Следе лексички записи за леме *кафа* и *кава* повезани варијантном везом ф_в, као и лексичи записи за леме *адресирати* и *адресовати* повезани варијантном везом ирати_овати:

```
kafa,N600+DOM=Culinary+DER=FV+Conc+Drink+Food+Prod
kava,N600+DOM=Culinary+DER=VF+Conc+Drink+Food+Prod
adresirati,V1+Imperf+Perf+Tr+Iref+Ref+DER=IratiOvati
adresovati,V18+Imperf+Perf+Tr+Iref+Ref+DER=OvatiIrati
```

Први лексички запис је обележен варијационим маркером +DER=FV који означава да у лексичком запису подниска „ф“ може бити замењена подниском „в“. Други лексички запис је означен варијационим маркером +DER=VF који је знак да подниска „в“ може бити замењена подниском „ф“. Врста речи није од значаја ни за ову релацију. На основу овог правила успоставља се релација именована као „varijanta(f_v)“. Овим правилом је успостављено 53 везе међу лексичким записима а неки од парова су: *кефтати* и *кевтати*, *салфета* и *салвета*, *кафопија* и *кавопија*, као и *куглоф* и *куглов*. У случају релације између глагола *адресирати* и *адресовати*, правилом је дефинисано да један лексички запис садржи маркер +DER=IratiOvati што значи да садржи подниску *„ирати“* која у другом лексичком запису може бити замењена подниском *„овати“*. Други лексички запис за повезивање означен је маркером +DER=OvatiIrati што значи да подниска *„овати“* може бити замењена подниском *„ирати“*. На основу овог правила које повезује 11 парова речи успоставља се релација названа „varijanta(irati_ovati)“. Неки од парова речи су: *рапортирати* и *рапортовати*, *аванзирати* и *аванзовати*, *фаворизирати* и *фаворизовати*.

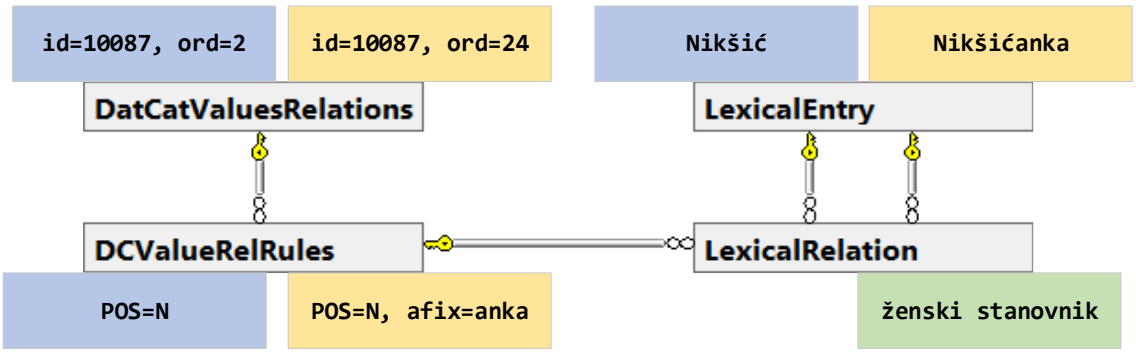
У деривационе релације спадају оне које повезују лексичке записе који су повезани на основу деривационих правила. У ову врсту релација спадају везе: *витез* и *витешки* (релација – релациони придев), *радник* и *радница* (релација – моција рода), *радник* и *раднички* (релација – релациони придев). Ове релације су карактеристичне за одређене врсте речи. Следе лексички записи из *Речника* који илуструју наведене релације:

```
vitez,N120+Hum
viteški,A2+PosQ
```

```
radnik,N10+Hum
radnički,A2+PosQ
radnica,N651+Hum+GM
```

Правило за успостављање релације *релациони придев* заснива се на повезивању именице и придева и постојању маркера +PosQ на придеву. Правилом је дефинисано да лексички запис који представља именицу треба да садржи подниску *„з“* док лексички запис који означава придев треба да садржи подниску *„шки“*. Овим правилом су успостављене 52 релације међу лексичким записима. Тако је повезан пар: *Париз* и *паришки*. Истом релацијом али другим правилом које захтева подниску *„к“* код именице и подниску „чки“ код придева повезан је пар *радник* и *раднички*. Овим правилом су повезана 52 пара речи а неке од њих су: *записник* и *записнички*, *дневник* и *дневнички*, *речник* и *речнички*. Релација моције рода између записа *радник* и *радница* остварује се на основу критеријума да обе речи морају бити именице и да једна реч садржи маркер +GM који означава моцију рода. Прва реч треба да садржи подниску *„к“* док друга реч треба да садржи подниску *„ца“*. Овим правилом је успостављено 59 релација међу паровима речи од којих су неки: *срећник* и *срећница*, *вереник* и *вереница*, *љубљеник* и *љубљеница*.

На слици 32 дат је приказ модела успостављених веза међу лексичким записима. У табели **LexicalEntry** налазе се информације о појединачним лексичким записима међу којима се успоставља релација. Ова табела је са две везе (енг. *Relationship*) повезана са табелом **LexicalRelation**, односно два лексичка записа (полазни и крајњи) су повезани релацијама које их спајају.



Слика 32 Приказ модела за успостављање релација

Табела **LexicalRelation** је повезана са табелом **DCValueRelRules** у којој се налазе појединачна правила за успостављање релација. Овим правилима се дефинишу критеријуми за успостављање релација, као што су захтевана врста речи, флективна класа, афикс или маркер којим је означен лексички запис. Полазна и циљана лексичка јединица треба да задовоље критеријуме постављене у табели **DCValueRelRules**. На пример, једно правило за успостављање релације „женски становник“ између лексичког записа који означава геополитичко име и лексичког записа који означава женског становника јесте: полазна реч треба да буде именица (POS = N) док остали критеријуми из табеле **DCValueRelRules** за њу нису од значаја, док циљана лексичка јединица треба такође да буде именица (POS = N) и да садржи подниску *„анка“*. Овим правилом за релацију су повезани лексички записи: *Никшић* и *Никшићанка*, *Перу* и *Перуанка*, *Тибет* и *Тибетанка*, *Чиле* и *Чилеанка*, итд. Али ово није довољно за коректно успостављање релације. За потпуну дефиницију релације неопходна је и табела **DatCatValuesRelations** која на основу кода категорије (све категорије су описане у сегменту апликације „Categories“) дефинише везу са маркером којим треба да буду означене полазна, односно циљана лексичка јединица. Тако је за претходни пример од значаја да полазни лексички запис садржи семантички маркер „+Top“ (општи маркер геополитичког имена) чија је ознака категорије „id=10087, ord=2“ похрањена у табели **DatCatValuesRelations** заједно

са ознаком категорије семантичког маркера „+Inh“ (становник геополитичке локације) којим је означен крајњи лексички запис „id=10087, ord=24“. Више правила за успостављање релација из табеле **DCValueRelRules** може бити везано за једно правило из табеле **DatCatValuesRelations**. За постојеће правило из табеле **DatCatValuesRelations** којим се повезује назив геополитичког имена са женским становником локације на коју се геополитичко име односи (полазни лексички запис садржи маркер „+Top“ док крајњи лексички запис садржи семантички маркер „+Inh“), поред поменутог правила које повезује пар *Никшић* и *Никшићанка*, дефинисаног табелом **DCValueRelRules** постоје и друга правила дефинисана истом табелом: полазни и крајњи лексички запис садрже ознаку врсте речи именица (POS = N) и нпр. полазни поднсику „*д*“ и крајњи “*ђанка*” (пар *Београд* и *Београђанка*), полазни подниску „*дија*“ и крајњи “*ђанка*” (пар *Холандија* и *Холанђанка*), полазни подниску „*а*“ и крајњи „*ињa*” (пар *Енглеска* и *Енглескињa*), итд.

Неке од потешкоћа на које се наилази приликом успостављања релација међу лексичким записима јесу ситуације када лексички записи нису обележени одговарајућим маркерима који су предуслов за успостављање релације, у речнику недостају лексички записи за упаривање или када услед широко постављених правила долази до погрешних повезивања лексичких записа.

Када неки лексички запис није обележен маркером постоји могућност његове допуне адекватним маркером и ручног повезивања са одговарајућим записом (опција Add Relation) или поновног покретања аутоматског успостављања релације на одговарајућем правилу (кроз сегмент за успостављање релација), што је згодно у случају допуне више лексичких записа.

Пример погрешног повезивања лексичких записа би био пар *Вран – Врањанка. Вран* представља планину у Босни док *Врањанка* означава становницу Врања. Овај пар је погрешно повезан правилом које спаја речи са подниском „*н*“ мењајући је у „*њанка*“ релације „ženski stanovnik“ која спаја топониме (маркер +Top) са становницама (+Inh). У пракси се погрешно успостављене везе ручно раскидају.

У јануару 2022. године кроз апликацију *Лексимирка* је успостављено укупно 103.589 повезивања лексичких записа. Од тог броја кроз деривационе релације остварено је 94.732 повезивања, кроз изговорну релацију EkIjk 5.456 повезивања, док је кроз варијациону релацију остварено 3.401 повезивање. Највећи број релација је успостављен аутоматски – 103.472, ручно је успостављено 46 релација а 71 је означена као погрешна.

У Прилогу 3 дат је списак свих успостављених релација у виду табеле којом су приказани идентификациони број релације, врста релације, ознака релације, полазна вредност релације, крајња вредност релације, број правила на којима се заснива релација, број успостављених веза, као и примери повезивања том релацијом.

5.2.2 Повезивање са *Ворднетом*

Како би развој и коришћење лексичких ресурса били што ефикаснији корисно је један ресурс обогаћивати информацијама из других ресурса. Један од примера јесте обогаћивање *Ворднета за српски језик* информацијама из *Морфолошког речника српског језика.*

Опште правило ворднета јесте да све речи које припадају једном синсету припадају истој врсти речи. Следећи синсетови илуструју то правило:

(1) utvrđenje:1; tvrđava:1; grad:A1a [Odbrambena struktura koja se sastoji od zidova ili rovova izgrađenih oko uporišta da bi se ono ojačalo.]
(2) grad:A2; [Administrativna podela okruga.]
(3) grad:B1; [Mera jačine alkoholnih pića izražena kao ceo broj.]

(4)grad:C1; [Padavina koju čine komadići leda.]

Наведена су 4 синсета који за литерале имају именице. Литерали су ниске знакова које представљају концепт синсета. Дефиниције придружене синсетовима су дате у угластим заградама. Приметићемо да се у свим синсетовима налази именица *„град“* која је употребљена у различитим значењима. Међутим, ради се о 3 различите именице: две акцентоване *„гра̑д“* и једна *„гра̀д“*. Али, како се у српском ворднету као и у *Морфолошком речнику српског језика* акценти не бележе, у свим синсетовима се појављује исти литерал *„град“.* Како се ради о 3 различите именице њима су придружене 3 флективне класе: у прва два случаја то је класа N81 (именице мушког рода с проширеном множином), у трећем је то класа N1 (именице мушког рода с немаркираном променом) а у четвртом је то класа N1001 (именица мушког рода без множине). Записи ове 3 именице и 4 значења у *Морфолошким речницима српског језика* су:

```
grad,N81+Conc+Facility
grad,N81+Top+Gr
grad,N1+Mes
grad,N1001+Phenom
```

Као што се види различита значења су означена различитим маркерима. Пренос ових информација из *Морфолошких речника српског језика* у српски ворднет може допринети побољшаној ефикасности коришћења *Ворднета за српски језик* у претраживањима јер се отклања вишезначност а претражује се само одговарајућим облицима. Информације пренесене из *Морфолошких речника српског језика* бележе се на литералу у пољу „lnote“.

Следећи пример показује скраћени приказ синсета (2) у приказу на језику XML:

```
<SRPWN>
<SYNSET>
<ID>ENG30-08672199-n</ID>
<POS>n</POS>
<SYNONYM>
<LITERAL>grad<SENSE>A1a</SENSE><LNOTE>N1+Top+Gr</LNOTE></LITERAL>
</SYNONYM>
<DEF> Administrativna podela okruga.</DEF>
</SYNSET>
</SRPWN>
```

Црвеном бојом и коришћењем етикета <LNOTE> означено је поменуто поље „lnote“ у коме се налазе морфолошка класа и семантички маркери пренети из *Морфолошких речника српског језика*. Податак „A1a“ окружен етикетом <SENSE> представља ознаку значења преузету из *Речника српскохрватскога књижевног језика* Матице српске. А означава прву одредницу од 3 које одговарају литералу *„град“* а 2 је ознака значења у оквиру те одреднице.

Скраћени приказ синсета (1) са литералима *„утврђење“ „тврђава“*и *„град“* изгледа овако:

```
<SRPWN>
<SYNSET>
<ID>ENG30-03385557-n</ID>
<POS>n</POS>
<SYNONYM>
<LITERAL>utvrđenje<SENSE>1</SENSE><LNOTE>N300+Conc+Facility</LNOTE></LITERAL>
<LITERAL>tvrđava<SENSE>1</SENSE><LNOTE>N600+Conc+Facility</LITERAL>
<LITERAL>grad<SENSE>A1a</SENSE><LNOTE>N81+Conc+Facility</LITERAL>
</SYNONYM>
<DEF>Odbrambena struktura koja se sastoji od zidova ili rovova izgrađenih oko
uporišta da bi se ono ojačalo.</DEF>
```

```
</SRPWN>
```

Уз све литерале стоје информације преузете из *Морфолошких речника српског језика* и *Речника српскохрватскога књижевног језика* Матице српске.

Други пример обогаћивања *Ворднета* јесте пренос деривационих информација из *Морфолошких речника српског језика*. У претходном одељку 5.2.1 је описан принцип повезивања лексичких записа у морфолошком речнику, између осталих и успостављање деривационих релација. У ворднету се користи релација „XPoS“ како би се надоместило то што се у једном синсету налази само једна врста речи. Информације о деривационим релацијама које постоје у *Морфолошким речницима* се могу користити за додавање релација „derived“, као и за креирање нових синсетова од изведених речи.

Следе примери два деривационо повезана лексичка записа из *Морфолошких речника српског језика*:

```
Beč,N1001+NProp+Top+Gr+CC2=AT+Val=Wie
bečki,A2+PosQ+NProp+Top+Gr+CC2=AT   relacioni pridev (_ki)
```

Запис „*бечки*“ је деривационом релацијом „relacioni pridev (_ki)“ повезан са записом „*Беч*“. Информација о релацији се може искористити за повезивање синсетова уз помоћ релације „derived“. Следи скраћени пример синсета „*бечки*“, у коме је црвеном бојом у оквиру етикета <ILR> (енг. *Interlingua relation*) означено која је врста релације у питању („derived“), и с којим синсетом се успоставља веза - „Beč“ – „ENG30-08846324-n“:

```
<SRPWN>
<SYNSET>
<ID>ENG30-02971192-a</ID>
<POS>a</POS>
<SYNONYM>
<LITERAL>bečki<SENSE>1</SENSE></LITERAL>
</SYNONYM>
<DEF>koji se odnosi na Beč i njegove stanovnike</DEF>
<BCS></BCS>
<ILR>ENG30-08846324-n<TYPE>derived</TYPE></ILR>
</SYNSET>
</SRPWN>
```

Међу следећа два синсета је могуће успоставити релацију стања „be_in_state“ на основу врсте речи и суфикса (информације из *MPC*):

```
siguran:6; uveren:1 [Kod koga nema osećanja nesigurnosti ili kolebanja; pouzdan i
ubeđen]
sigurnost:10; uverenost:1 [Stanje u kome je neko siguran, uveren u nešto.]
```

За успостављање ове релације је потребно да полазна реч буде придев а циљна именица са суфиксом „*ост*“. Следи скраћени приказ првог синсета где је црвеном бојом означена релација „be_in_state“ заједно са ознаком другог синсета „ENG-30-05697135-n“ са којим је остварена ова релација.

```
<SRPWN>
<SYNSET>
<ID>eng-30-00336831-a</ID>
<POS>a</POS>
<SYNONYM>
<LITERAL>siguran<SENSE>6</SENSE><LNOTE>>+UPOS=ADJ</LNOTE></LITERAL>
<LITERAL>uveren<SENSE>1</SENSE><LNOTE>+UPOS=ADJ+PP+Ek </LNOTE></LITERAL>
</SYNONYM>
<DEF>Kod koga nema osećanja nesigurnosti ili kolebanja; pouzdan i ubeđen </DEF>
<BCS></BCS>
<ILR>ENG-30-05697135-n<TYPE>be_in_state</TYPE></ILR>
<ILR>ENG-30-05697363-n<TYPE>be_in_state</TYPE></ILR>
```

```
<ILR>ENG-30-00337404-a<TYPE>near_antonym</TYPE></ILR>
<ILR>ENG-30-08540903-n<TYPE>specificof</TYPE></ILR>
<ILR>ENG-30-11444117-n <TYPE>specificof</TYPE></ILR>
</SYNSET>
</SRPWN>
```

Поред обогаћивања *Ворднета за српски језик* трансфером информација из *Морфолошких речника*, могућ је и обрнут след. Ворднет својом хијерархијском структуром која је успостављена релацијама надређености и подређености може обогатити *Морфолошки речник српског језика* увођењем нових семантичких ознака. Томе могу допринети и везе ворднета са онтологијом SUMO и таксономијом домена.

5.2.3 Повезивање са корпусом

Приказ конкорданци

Лексички записи приказани кроз апликацију *Лексимирка* омогућавају претрагу различитих корпуса и то на два начина. Више детаља о претрази корпуса може се видети у у Прилогу 2.

Један начин интеракције између апликације и корпуса јесте основна претрага лемом. Тада се над корпусом поставља упит који претражује све флективне облике леме лексичког записа у коме се корисник налази. На овај начин су претраживи *Корпус геолошких текстова - ГеоСрпКор* чија је величина 1.078.435 монолексемских облика речи (више детаља о обухвату корпуса дато је у поглављу 7) и корпус рударских текстова – *РудКор* од 2.719.086 облика речи (детаљније описан у поглављу 3.3).

Други начин повезивања апликације и корпуса јесте напреднија претрага корпуса путем одабира унапред постављеног обрасца. *Корпус ГеоСрпКор* допушта овакав вид претраге јер је претходно аутоматски обележен проширеним анотацијама приказаним у табели 7. Овај корпус је доступан преко платформе отвореног кода *NoSketchEngine* која омогућава претрагу помоћу простог упита, леме, фразе, речи, карактера али и коришћењем језика за постављање упита над корпусима – CQL (енг. *Corpus Query Language*).

Табела 7 Проширени тагови коришћени у корпусу *ГеоСрпКор*

Категорије		**Таг у корп.**
N (Noun) именица		N
род	f – женски	N:f
	m – мушки	N:m
	n – средњи	N:n
A (Adjective) придев		A
род, степен компарације	f – женски, a - позитив	A:af
	m - мушки, a - позитив	A:am
	n - средњи, a - позитив	A:an
	f - женски, b - компаратив	A:bf
	m - мушки, b - компаратив	A:bm
	n - средњи, b - компаратив	A:bn
	f - женски, c - суперлатив	A:cf
	m - мушки, c - суперлатив	A:cm
	n - средњи, c - суперлатив	A:cn
V (Verb) глагол		V
род	f – женски	V:f
	m – мушки	V:m
	n – средњи	V:n
PRO (Pronoun) заменица		PRO

NUM (Number) број		NUM
PREP (Preposition) предлог		PREP
CONJ (Conjunction) везник		CONJ
INT (Interjection) узвик		INT
PAR (Particle) речца		PAR
ADV (Adverb) прилог		ADV
PREF (Prefix) префикс		PREF
ABB (Abbreviation) скраћеница		ABB
RN (Roman numeral) римски број		RN
PUNCT (Punctuation) знак интерпункције		PUNCT
SENT (Sentence and marker) реченица и маркер		SENT
? (non-Serbian words or suffixes in compounds) речи или суфикси у сложеницама несрпског порекла		X

На пример, постављањем упита `[tag="A.*"] [lemma="peskovi"]` на језику CQL, добићемо конкорданце које садрже све облике леме *пескови* којима претходи придев. Лема *пескови* је карактеристична за област геологије као термин јер се увек користи у множини. Овим упитом се добија 959 линија конкорданци од којих су неке приказане у табели 8. У првој колони су дате ознаке докумената (тумача) из којих су конкорданце извучене, следећа колона представља леви контекст, у последњој колони се налази десни контекст док су црвеном бојом означене комбинације речи које задовољавају постављени упит. Неки од пронађених облика су: *палудинским песковима, средњезрних пескова, неогених пескова, лискуновитим песковима, алевритски пескови*.

Табела 8 Извод из конкорданци које одговарају упиту `[tag="A.*"] [lemma="peskovi"]` добијених на корпусу *ГеоСрпКор*

doc#62	**tri bušotine kojima su kaptirani izdani u**	paludinskim peskovima	**i to : Ki-2/H (Kikinda) , VS-l/H i VS-2/H (**
doc#61	slojevi su izgrađeni od sitnozrnih do	**srednjezrnih peskova**	(0,06 - 0,40 mm) . Ovaj nivo hipotermalnih voda
doc#42	su glinovitim i peskovitim laporima ,	**glinovitim peskovima**	i šljunkovitim glinama . makrofauna ukazuje na
doc#13	sedimentima . naizmenično smenjivanje	**neogenih peskova**	, šljunkova i peščara sa vulkanskim
doc#65	do 445 m dubine) . pojave peska veće akumulacije	**fluvijalnih peskova**	koji leže neposredno ispod kulturnog sloja
doc#5	i sitnozrnim subarkozama , koji se smenjuju sa	**alevritskim peskovima**	, alevritskim glinama i peskovitim laporcima .
doc#34	i sivo plavičastim srednjezrnim do sitnozrnim	**liskunovitim peskovima**	. sortiranost zrna je odlična (so = 1,2 - 1,35) ,
doc#65	Dunava tekla od šljunkovito rečnog dna pa preko	**srednjozrnih peskova**	, eolskih i povodanskih akumulacija do
doc#61	- Kikinda pretpostavljamo 60 - 75m . Peskovi ,	**alevritski peskovi**	i šljunkovito - peskoviti alevriti (aj - g)
doc#19	je predstavljena peskovitim šljunkovima ,	**šljunkovitim peskovima**	i alevritskim peskovima . Ređe se sreću i

Упит `[flemma="ugalj"]` ће вратити конкорданце које садрже морфолошки проширене облике леме *угаљ* на латиници и ћирилици (*ugalj, угаљ, ugalja, угаља, uglja, угља, uglje, угље, ugljem, угљем, uglji, угљи, ugljima, угљима, uglju, угљу*). Упитом [synlemma="voda"] добићемо конкорданце које садрже све облике леме *вода* и њених синонима забележених у оквиру речника DELA (*suza, voda, vodosnabdevanje, суза, вода, водоснабдевање*). Оваква претрага је омогућена захваљујући проширењу упита

коришћењем ресурса за проширење упита и скупа веб-сервиса *ВебРан* (више о веб-сервисима може се прочитати у делу 6.3.2).

Када корисник у оквиру прегледа монолексемског лексичког записа одабере опцију за приказ конкорданци, одабиром тачног корпуса, страна са конкорданцама које садрже лему лексичког записа ће се појавити у новом табу у оквиру веб-прелистача (слика 33). Подразумевано је излиставање по 20 линија конкорданци које садрже тражени појам.

Коришћењем напредних могућности за претрагу корпуса *ГеоСрпКор* употребом унапред дефинисаног образаца и коришћењем дугмета за претрагу (🔍) излиставају се конкорданце које садрже конструкцију која одговара задатом обрасцу. На пример, употребом обрасца „A<N>PrepNp“ добићемо конкорданце које садрже конструкцију у којој је на првом месту придев праћен именицом из лексичког записа а потом и предлошко-падежна конструкција[96]. Уколико из лексичког записа *глина* позовемо овај упит, добићемо конкорданце приказане на слици 33. Међу излистаним линијама конкорданци су оне које садрже конструкције: *песковите глине са шљунковима*, *бетонитске глине у повлати*, *лапоровите глине са остракодама*, *масне глине са сочивима*, итд.

више смењују са пешчарима , преко којих леже	peskovite gline sa konkrecijama	karbonata i ugljenisanim biljnim ostacima .
sloja , debljine oko 40 cm (D. Škerl , 1962) .	Sive gline sa ostacima	fosilnih biljaka određenih kao
ugljonosne serije . To su žute do crvenomrke	peskovite gline sa šljunkovima	i konglomeratima . deo ovih naslaga svakako
sedimenti . To su crvene i mrke peskovite i	šljunkovite gline sa slojevima	peska , šljunka i konglomerata . sličnim
sedimenti Prebreze pripadaju sarmatu . pojave	bentonitskih glina u povlati	ugljenog sloja rudnika Jankova Klisura
ili preko starijih stena . To su crvene i mrke	peskovito-šljunkovite gline sa slojevima	peska , šljunkova i konglomerata , dok su u
sedimenata , a čine ih sive i žutomrke do crvene	peskovite gline sa proslojcima	peska i ređe šljunka . Ovim sedimentima je na
basena javljaju se žutomrke do crvenkaste	peskovite gline sa proslojcima	peska , interpretirane kao jezerski sedimenti
debljine nekoliko santimetara ili proslojci	ugljevitih glina u peskovima	i glinama donjeg dela miocenske serije . nemaju
i šljunkova , dok se u povlati uglja nalaze	sive gline sa ostacima	slatkovodne faune (kongerije , viviparusi , i
, idući odozdo pa više : nečisti lapori i	laporovite gline sa Candona	veljae , C. hvosnoica ; tanak sloj šljunka ;
, viviparusima i ostrakodama ; lapori i	laporovite gline sa ostrakodama	, kongerijama i viviparusima . ukupna debljina
. predstavljen je šljunkovima , peskovima i	nečistim glinama sa konkrecijama	od CaCO3 . fosilni ostaci nisu nađeni . ugljeni
i suvog Lukovca . konkordantno preko sivih	laporovitih glina sa faunom	iz donjeg pliocena (kongersije i dr .) leže
peskovite i listaste gline . U listastim	tamnosivim glinama na listu	kosovska Mitrovica nađeni su ostaci
: alevrolitski peskovi , alevrolitske gline i	peskovite gline sa proslojcima	laporaca i slabo vezanih peščara . Jedan deo
pri erupcijama na površinu zemlje ili u vodu . U	peskovitim glinama oko sela	Kruševice (Vlasotince 51) nađeni su riblji
miocen se završava peskovito-laporovitim	masnim glinama sa sočivima	i proslojcima žutih , kvarcnih , slabo vezanih
mnogo biljnog detritusa i proslojke uglja .	peskovite gline sa šljunkovima	predstavljaju završni deo . požeški basen
slojevi uglja . U višim horizontima nalaze se	konglomeratične gline sa proslojcima	konglomerata , glina i peskova . U jugozapadnom

Слика 33 Линије конкорданци за упит „ANPrepNp“ у оквиру лексичког записа *глина*

Предефинисани обрасци су одабрани из скупа образаца за екстракцију терминологије (Krstev и остали 2015) пошто је то оно што корисник углавном тражи. Универзално правило је да ће реч из тренутног лексичког записа у обрасцу бити на позицији која је уоквирена изломљеним заградама. Неки други доступни обрасци за претрагу именица у корпусу *ГеоСрпКор* су:

– A<N> – проналази изразе који се састоје од придева праћеног именицом која је у лексичком запису. За лексички запис *глина*, у корпусу

[96] У српском језику предлози не чине синтагму.

ГеоСрпКор ћемо пронаћи конкорданце које садрже појављивања: *песковита глина*, *црвена глина*, *угљевита глина*, *серицитисана глина,* итд.

- AA<N> - проналази изразе који се састоје од два придева праћена именицом из тренутног лексичког записа. Неки од погодака за лексички запис *глина* су: *бетонитска сивозелена глина, мрка песковита глина, зелена масна глина, сива лапоровита глина*, итд.
- N<Ngi> – проналази именицу праћену именицом из лексичког записа (*глина*) у генитиву или инструменталу. Примери који се добијају јесу: *подина глине, тип глине, сочиво глине, минерал глине, анализа глине, фација глина*, итд.
- <N>PrepNp - проналази именицу из лексичког записа иза које следи предлошко-падежна конструкција. У корпусу *ГеоСрпКор* неки од погодака су: *глина са угљем, глина у песковима, глина са прослојцима, глина са шљунковима, смењивање са глинама*, итд.
- AN<Ngi> – проналази придев праћен двема именицама од којих је друга одређена лексичким записом и у облику генитива или инструментала. Овим обрасцем проналазимо: *монморионитска врста глине, мали садржај глине, ситне грудве глине, минералошко испитивање глине, брахихалинска фација глине*, итд.
- N<Ngi>PrepNp – проналази именицу праћену именицом одређеном лексичким записом у генитиву или инструменталу иза које следи предлошко-падежна конструкција. На овај начин долазимо до поготка: *алверит глина са лапоровитом.*
- <N>PrepNpNgi - проналази именицу из лексичког записа иза које следи предлошко-падежна конструкција па потом именица у генитиву или инструменталу. На овај начин долазимо до следећих комбинација: *глина са конкрецијама карбоната*, *кречњаци са прослојцима глина, глине са појавама угља, глина у локалитету Беочина, глина око села Крушевице*, итд.
- <N>VN – проналази именицу одређену лексичким записом праћену глаголом и именицом. Овим обрасцем проналазимо конструкције: *глине су наслаге, глине имају честице*, итд.

Кориснику ће се, у зависности од тога којој врсти речи припада текући лексички запис, у падајућем менију појавити обрасци који садрже ту врсту речи. Тако ће на пример кроз лексички запис придева *еолски* за претрагу бити доступни обрасци који проналазе конкорданце које садрже придев: „<A>N“ (*еолска акумулација, еолска прашина*), „A<A>N“ (*млађи еолски пескови, лесоидни еолски пескови*), „<A>NNgi“ (*еолски транспорт материјала*, *еолске творевине Баната*), „<A>NPrepNp“ (*еолска акумулација у вирму, еолски песак у комплексу*). У случају да су у обрасцу придеви на две позиције, придев из лексичког записа се подразумева на позицији окруженој изломљеним заградама. Листа понуђених образаца је флексибилна јер се обрасци додају преко базе података па је могуће додавање нових образаца без измена апликације.

Фреквенције речи у различитим доменима

Свака област науке и струке користи специфичне термине који имају одређено значење у том домену. Како би се дошло до информације о најспецифичнијим речима (терминима) за одређени доменски корпус у апликацији *Лексимирка* је развијен сегмент који даје информације о кључности лексичког записа у том доменском корпусу.

Како би се одредила кључност речи у доменском корпусу користи се методологија која ће бити приказана у наставку.

У приказу лексичког записа могуће је добити информацију о фреквенцији појављивања одређене речи у неком од корпуса доступних путем сегмента за корпусе (опција Corpora). За потребе одређивања фреквенције кључних термина у одређеном доменском корпусу у односу на други референтни корпус користи се мера „једноставна математика“ (енг. *simple maths*)(„Statistics used in the Sketch Engine“ 2015) (Kilgarriff 2009). Овом мером се укључују променљиве које дозвољавају кориснику да се усредсреди на високофреквентне или нискофреквентне речи из корпуса. Више вредности ове мере (нпр. 100 или 1000) се односе на фреквентније речи док се ниже вредности (нпр. 1 или 0,1) односе на мање фреквентне речи.

Статистика која се користи при одређивању специфичне фреквентности за доменски корпус је: „реч *P је толико и толико пута фреквентнија у корпусу X него у корпусу У.*“ Циљ ове мере је да одреди да ли се одређена реч појављује значајно чешће у једном корпусу у односу на њено појављивање у другом. Скор кључности (енг. *keyness*) речи за одређени корпус у односу на други се рачуна на основу следеће формуле:

$$\frac{fpm_{focus} + N}{fpm_{ref} + N}$$

Где је fpm_{focus} нормализована фреквенција (на милион) појављивања речи у корпусу у коме се истражује док је fpm_{ref} нормализована фреквенција (на милион) појављивања речи у референтном корпусу. N је параметар за уједначавање чија је подразумевана вредност 1.

Нормализована фреквенција појављивања у појединачним корпусима се рачуна на следећи начин:

$$fpm_{focus} \text{или } fpm_{ref} = \frac{\text{број појављивања у корпусу x 1.000.000}}{\text{величина корпуса}}$$

Следи рачун за кључност речи *шкриљац* у корпусу *ГеоСрпКор чија је величина* 1.028.962 речи. За референтни корпус узет је подскуп *Корпуса савременог српског језика* од 117.077.960 речи назван *СрпКор122М* (Stankovic и остали 2020).

Нормализована фреквенција појављивања речи *шкриљац* у корпусу *ГеоСрпКор 3.552,12* која је сврстава у 100 најфреквентнијих речи у овом корпусу добијена је на следећи начин:

$$fpm_{focus} = \frac{3.655 \text{ x } 1.000.000}{1.028.962} = 3.552{,}12$$

Нормализована фреквенција појављивања речи *шкриљац* у референтном корпусу *СрпКор122М* 1,57, која је сврстава у групу од 50.000 најфреквентнијих речи у овом корпусу, добијена је на следећи начин:

$$fpm_{ref} = \frac{184 \text{ x } 1.000.000}{117.077.960} = \frac{184.000.000}{117.077.960} = 1{,}57$$

Скор кључности речи *шкриљац* у корпусу *ГеоСрпКор* у односу на референтни корпус *СрпКор122М* се рачуна на следећи начин:

$$\frac{fpm_{focus}+N}{fpm_{ref}+N} = \frac{3{,}552.12+1}{1.57+1} = \frac{3553.12}{2.57} = 1.382$$

На основу добијене вредности долази се до закључка да је реч *шкриљац* 1.382 пута фреквентнија у корпусу *ГеоСрпКор* него у корпусу *СрпКор122М.*

Следи иста рачуница за реч *гвожђе*:

$$fpm_{focus} = \frac{298 \text{ x } 1.000.000}{1.028.962} = 289{,}61$$

$$fpm_{ref} = \frac{2.464 \text{ x } 1.000.000}{117.077.960} = 21{,}04$$

$$\frac{fpm_{focus}+N}{fpm_{ref}+N} = \frac{289{,}61+1}{21{,}04+1} = \frac{290{,}61}{22{,}04} = 12{,}97$$

Закључак до кога се долази јесте да је реч *гвожђе* 12,97 пута фреквентнија у корпусу *ГеоСрпКор* него у референтном корпусу *СрпКор122М.* Према нормализованој фреквенцији у доменском корпусу *ГеоСрпКор* ова реч спада у категорију од 100 најфреквентнијих речи док у корпусу *СрпКор122М* спада у категорију од 5.000 најфреквентнијих речи.

Следи рачун за реч *седимент* која у корпусу *ГеоСрпКор* има највећи број појављивања, чак 9.455.

Нормализована фреквенција појављивања ове речи која је сврстава у 100 најфреквентнијих речи у корпусу *ГеоСрпКор* је:

$$fpm_{focus} = \frac{9.455 \text{ x } 1.000.000}{1.028.962} = 9.188{,}87$$

С друге стране реч *седимент* се не појављује у корпусу *СрпКор122М* па рачун за нормализовани фреквенцију у овом корпусу изгледа овако:

$$fpm_{ref} = \frac{0 \text{ x } 1.000.000}{117.077.960} = \frac{0 \text{ x } 1.000.000}{117.077.960} = 0$$

Аутор методологије по којој се израчунава скор кључности наводи да је случај када се реч не појављује у референтном корпусу спорна и предлаже да се у том случају свим фреквенцијама (и фреквенцији у доменском корпусу) дода 1 или неки други број како би се избегло дељење са 0 (Kilgarriff 2009). У том случају би овај рачун изгледао овако:

$$fpm_{ref} = \frac{(0+1) \text{ x } 1.000.000}{117.077.960} = \frac{1.000.000}{117.077.960} = 0{,}0085$$

Нови рачун нормализоване фреквенције у доменском корпусу би био:

$$fpm_{focus} = \frac{(9.455 + 1) \text{ x } 1.000.000}{1.028.962} = 9.189{,}84$$

На основу ових вредности следи израчунавање скора кључности:

$$\frac{fpm_{focus}+N}{fpm_{ref}+N} = \frac{9.189{,}84+1}{0{,}0085+1} = \frac{9.190{,}84}{1{,}0085} = 9.113{,}38$$

Закључак је да је реч *седимент* 9.113,38 пута фреквентнија у доменском корпусу него у референтном.

На претходна три примера смо видели да што је већи скор кључности, то је термин специфичнији у домену корпуса. Реч *гвожђе* није толико често коришћен термин

у геологији док се у *СрпКор122М* користи доста често, термин *шкриљац* је доста заступљенији у домену геологије (а у СрпКор се користи ређе), док термин *седимент* представља најзаступљенији термин у овом домену, а у СрпКор се не појављује уопште.

Поред претходно наведене статистике корисник има могућност да излиста фреквенције свих облика одређене речи (дугме Form Frequencies) и леме (дугме Lemma Frequencies) у изабраном корпусу. На пример, можемо да видимо да се лема *шкриљац* у корпусу *ГеоСрпКор* појављује 3.226 пута, док се у корпусу *РудКор* појављује 155 пута. У табели 9 је дат приказ појављивања облика речи *шкриљац* у геолошком корпусу *ГеоСрпКор* и рударском корпусу *РудКор.* Ове фреквенције појављивања нам не значе пуно за упоредно поређење појављивања речи у датим корпусима с обзиром на то да корпуси нису исте величине па су нам за тако нешто потребне нормализоване фреквенције чији је поступак добијања описан у претходном пасусу.

Табела 9 Фреквенције појављивања облика речи *шкриљац*

Облик речи	*ГеоСрпКор*	*РудКор*
шкриљаца	1444	89
шкриљци	964	50
шкриљцима	574	9
шкриљац	22	3
шкриљца	14	2
шкриљцу	0	1
шкриљце	208	1

5.2.4 Профил речи

Један од циљева апликације *Лексимирка* јесте да, на основу података добијених из лексикографске базе и екстерних извора, структурирано прикаже информације о лексичком запису у облику који бисмо могли назвати профил речи. На овај начин ће бити приказане и информације које нису биле садржане у лексичким записима *Морфолошких речника за српски језик*. Овај профил речи може да садржи информације о:

- леми,
- флективним облицима,
- могућим деривационим облицима,
- речима са којима реч дели флективну парадигму,
- врсти речи,
- повезаности са другим речима,
- лексичким записима актуелне речи у другим речницима,
- фреквенцијама појављивања речи у језичким корпусима,
- конкорданцама из појединих корпуса,
- фреквенцији облика и лема у појединачним корпусима,
- значењима речи,
- полилексемским изразима чији је конституент актуелна реч.

Побројане информације су доступне кориснику кроз корисничко сучеље апликације које ће бити описано у одељку 6.1 и Прилогу 2.

5.2.5 Повезивање са екстерним ресурсима

Како би се кориснику који има потребу за консултовањем других лексичких извора олакшао рад и уштедело време за претрагу, у апликацију *Лексимирка* је интегрисан систем претраге лексичких ресурса који могу бити похрањени у локалној бази података или на вебу. Повезивање лексичких записа из апликације *Лексимирка* са другим ресурсима омогућено је уз помоћ табеле **LinkedLexicon** која похрањује податке о екстерним ресурсима. Сваком ресурсу са којим се повезују записи из *Лексимирке* су додељени идентификациони број (поље **LexiconID**), кратак назив (поље **Lexicon**), као и информација да ли је је екстерни ресурс на вебу или је у питању локална база података (поље **LinkDb**). Поље **LinkParam** говори о томе да ли се повезивање врши преко леме или коришћењем кључа (идентификатора). Поље **BaseURL** у комбинацији са податком из поља **LinkParam**, тј. лемом или кључем, даје конкретан URL који се позива у новој картици. На пример, URL за претрагу ресурса *Wiktionary* лемом *глина* ће имати облик „https://en.wiktionary.org/wiki/глина“. Поље **BaseSQL** ће садржати упит на језику SQL (енг. *Structured Query Language*), док ће поље **Description** садржати кратак опис и библиографско реферирање на ресурс. Поље **BaseURL** се користи код претраге ресурса на вебу, док је поље **BaseSQL** намењено претрази локалних база података.

Повезане базе података

Апликација *Лексимирка* одређеном кругу корисника који имају право на претрагу других лексичких ресурса нуди могућност приказа речничких чланака из неколико постојећих речника чији су делови рашчитани[97] и увезени у локалну базу података или су доступни преко веб-сервиса о којима ће бити речи у одељку 6.3.2. Под овим се подразумева приказ речничких чланака из неколико речника српског језика:

- *Речник српскога језика* Матице српске[98] (*Речник српскога језика* 2011) – обухваћени су илустративни примери из 5% дигитализованих рашчитаних речничких чланака из издања из 2011. године,
- *Речник српскохрватског књижевног и народног језика САНУ* – обухваћени су илустративни примери из 5 рашчитаних томова (Stanković и остали 2019), и то првог тома (*Речник српскохрватског књижевног и народног језика. Књ. 1, А-Богољуб* 1959), другог тома (*Речник српскохрватског књижевног и народног језика. Књ. 2, Богољуб-Вражогрнци* 1962), осамнаестог тома (*Речник српскохрватског књижевног и народног језика. Књ. 18, оповргавање - оцарити* 2010), деветнаестог тома (*Речник српскохрватског књижевног и народног језика. Књ. 19, оцат - петогласник* 2014) и двадесетог тома (*Речник српскохрватског књижевног и народног језика. Књ. 20, петогодан - погдегод* 2017),
- *Вуков рјечник* (Караџић 1818) – рашчитана верзија,
- *Систематски речник српскохрватског језика*[99] Ранка Јовановића (Р. Јовановић 1938),

[97] Рашчитавање текста овде подразумева оптичко препознавање карактера и рашчлањење делова речничког чланка.

[98] Сканирани Речник је доступан на интернету путем хипервезе: https://archive.org/details/recnik-srpskoga-jezika-2011/mode/2up (приступљено: 18.09.2021)

[99] Доступан у целости на интернету путем Викиизворника: https://sr.wikisource.org/wiki/Систематски_речник_српскохрватског_језика_(1936) (приступљено: 18.09.2021.)

– *Лексикон страних речи и израза*[100] Милана Вујаклије (Вујаклија 2014),
– *Речник синонима*[101] Павла Ћосића (Ћосић 2008) – обухваћено је 5% дигитализованих рашчитаних речничких чланака,
– *Ворднет за српски језик* (Stanković, Mladenović, и остали 2018).

Из традиционалних речника српског језика додат је део речничких чланака како би се приказале могућности интеграције. Нови речници се могу накнадно додавати у базу сходно њиховој доступности. На слици 34 дат је приказ лексичког записа *глина* са приказаним повезивањима са речничким чланцима из *Речника српскога језика* Матице српске и *Речника српскохрватског књижевног и народног језика САНУ.*

Слика 34 Речнички чланак *глина* са приказом повезаних чланака из других речника

Ресурси на вебу

На описани начин апликација *Лексимирка* омогућава повезивање речничких чланака са различитим лексичким изворима на вебу (опција *Check in external dictionaries* на слици 34). Повезивање речничких чланака је извршено са: семантичком мрежом

[100] Доступан у целости на интернету путем следећег линка: https://books.google.rs/books?id=jSzNAQAAQBAJ&printsec=frontcover&hl=sr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false (приступљено: 18.09.2021)

[101] Доступан у целости на интернету путем следећег линка: https://pdfcoffee.com/qdownload/pavle-osi-renik-sinonima-2008pdf-pdf-free.html (приступљено: 18.09.2021)

BabelNet[102], *Викиречник*[103], базом *Википодаци*[104], терминолошком базом *Терми*[105] и вишејезичким порталом *Glosbe*[106]. Кликом на назив одабраног спољњег ресурса из приказа лексичког чланка у апликацији *Лексимирка* у новом прозору се започиње претрага лемом тренутног чланка.

Упит лемом *глина* постављен речнику *Викиречник*[107] приказује речничке чланке на бугарском, македонском, руском, српском, српскохрватском и украјинском језику. База *Википодаци* враћа информације о запису *clay (глина)*. Погоци који се добијају постављањем истог упита над *BabelNet*-ом[108] (слика 35) јесу концепт *глина* на српском језику, именовани ентитети *Глина* за град и реку у Хрватској, као и за насељено место у Румунији. Портал *Glosbe*[109] постављањем истог упита, поред превода на енглески језик, нуди приказ конкорданци са употребом речи *глина* пореклом из преводилачких меморија.

[102] *BabelNet* је вишејезична лексичка семантичка мрежа настала повезивањем Википедије са ворднетом. https://babelnet.org/ (приступљено 8.09.2021)

[103] *Викиречник* (енг. *Wiktionary)* представља вишејезични речник слободног садржаја на вебу настао удруженим радом корисника. https://www.wiktionary.org/ (приступљено 8.09.2021)

[104] *Википодаци* (енг. *Wikidata*) - база знања која је извор података за друге пројекте Викимедије уређивана од стране корисника Викимедије. https://www.wikidata.org/wiki/Wikidata:Main_Page (приступљено 8.09.2021)

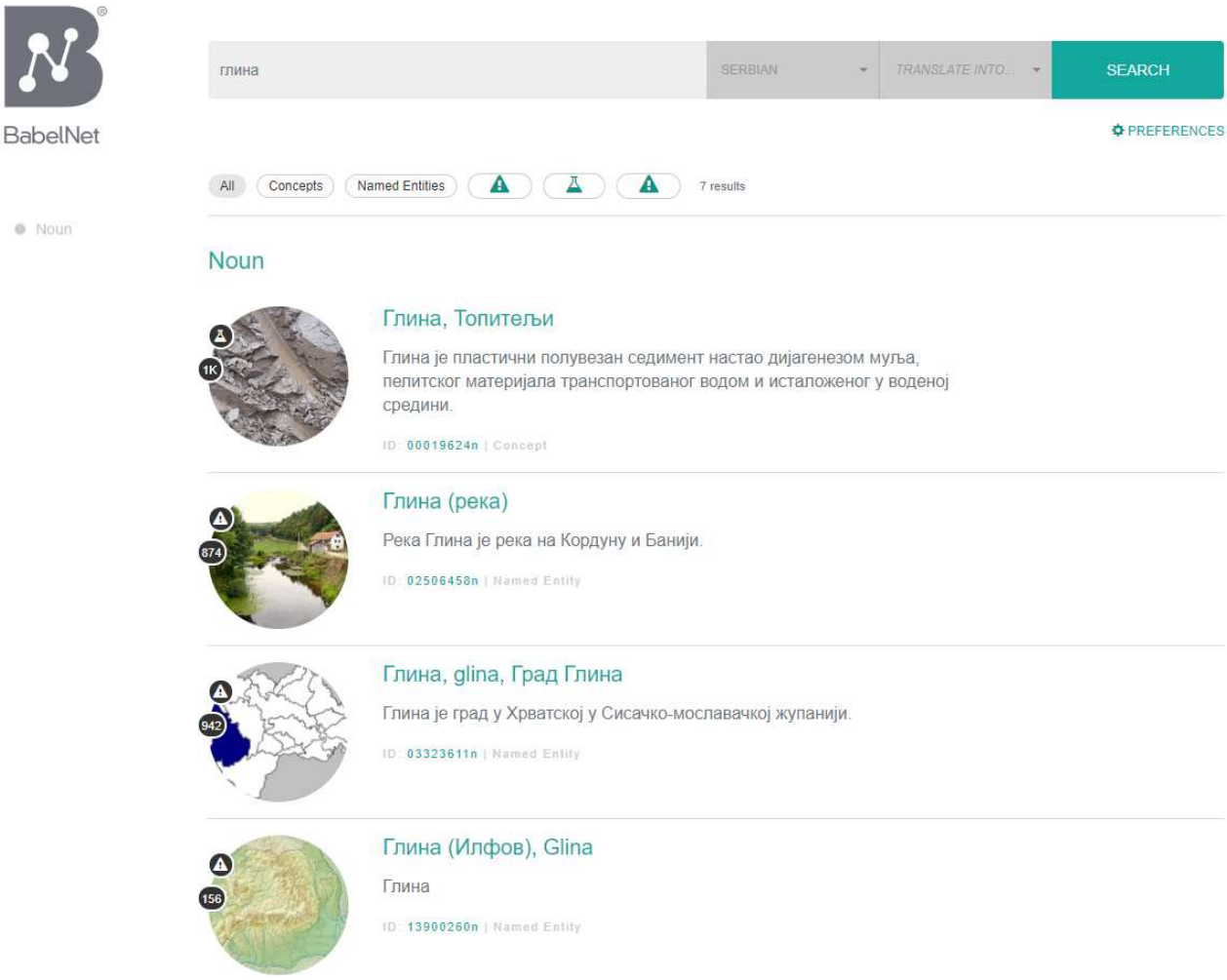
[105] Терми је апликација која подржава развој терминолошких речника и пружа могућност извоза података у различитим форматима (нпр. ексел, TBX, csv). http://termi.rgf.bg.ac.rs/ (приступљено 8.01.2020)

[106] *Glosbe* је вишејезични онлајн речнички портал који пружа велики број међујезичких превода са линијама конкорданци из преводилачких меморија. https://sr.glosbe.com/ (приступљено 8.09.2021)

[107] https://en.wiktionary.org/wiki/глина (приступљено 8.09.2021)

[108] http://live.babelnet.org/search?lang=SR&word=глина (приступљено 8.09.2021)

[109] https://sr.glosbe.com/sr/en/ глина (приступљено 8.09.2021)



Слика 35 Приказ резултата упита *глина* у BabelNet-у

6. Постављање презентације на веб

6.1 Опис корисничког сучеља

Корисничко сучеље развијене апликације *Лексимирка* има двоструку намену. Прва намена је за потребе претраживања и прегледа ограниченог скупа података из *Морфолошких речника српског језика*, док се друга намена односи на уређивање и развој истог речника.

Регистровани корисник без додељених додатних привилегија може да претражује лексичке записе из подскупа *Морфолошких речника српског језика* који подразумева 100.000 најфреквентнијих у корпусу *СрпКор* и том приликом види: приказ леме, везе са другим лексичким записима, везе са лексичким ресурсима који су на вебу и ограниченим бројем лексичких ресурса из повезане базе података, фреквенције коришћења леме у различитим корпусима и основно значење што подразумева да су маркери из *Речника* исказани речима.

Наредни ниво корисника са одређеним привилегијама може да претражује све лексичке записе из базе са ограничењем од 12.000 карактера за претрагу. Овај корисник у односу на корисника претходног нивоа има приступ већем броју повезаних речника у повезаној бази података. На располагању му је и део везан за претрагу корпуса и приказ информација о фреквенцији леме и појединачних облика одреднице из лексичког записа. Овај корисник може да учита запис из речника флективних облика, као и списак свих лексичких записа који деле исту флективну парадигму. Овај корисник има и могућност уређивања података у лексичком запису.

Корисник са додељеним администраторским привилегијама има могућност приступа апликацији за управљање и развој *Морфолошких речника српског језика.* Ово подразумева уређивање и прегледање различитих сегмената апликације *Лексимирка*: лексичких категорија (опција „Categories“), датотека речника (опција „Files“), лексичких записа (опција „Entries“), коришћених корпуса (опција Corpora), евалуације кандидата за речник полилексемских израза (опција „Evaluation“), лексичких релација (опција „Relations“), и флективних образаца (опција „Morphology“).

Кроз први сегмент корисник приступа панелу за преглед и управљање свим категоријама података који се користе у *Речнику.* Панел се у складу са жељом корисника приказује двојако, у облику дрвета или у облику табеле. Корисник може кориговати метаподатке о категоријама али може и креирати нове. Више детаља о овом сегменту може се прочитати у Прилогу 2.

Кроз сегмент за датотеке *Речника* могућ је преглед постојећих метаподатака о датотекама из којих су увезени лексички записи, као и њихово уређивање. Уз сваку датотеку је могућ приказ свих лексичких записа који су у њеном оквиру. Такође је могуће додавање нове датотеке.

Сегмент апликације *Лексимирка* за лексичке записе омогућава преглед, претрагу и управљање лексичким записима из базе. Овом сегменту је могуће приступити избором једног од два панела: панела за прелиставање и приказ лексичких записа („Lexical Entries“) или панела пуне табеле („Full Table“). Главна разлика између ова два панела јесте главно поље за претрагу. На првом панелу се кроз поље за претрагу претражују канонски облици лема док се на другом панелу претражује према категоријама које су придружене лексичким записима. Друга разлика између ова два панела јесте приказ излистаних лексичких записа. Кроз колону панела пуне табеле су излистане све ознаке значења у лексичком запису па она пружа више могућности за филтрирање према категоријама из те колоне. Путем два представљена панела је могуће приступити пуном приказу или уређивању постојећег лексичког записа. Подаци о

лексичком запису подразумевају лему и њене флективне облике, као и могуће деривационе облике, приказ информације о датотеци којој запис припада, као и приступ лемама са којима дели флективну класу. Доступан је и приказ веза са другим лексичким записима. У приказу лексичког записа је дат преглед леме у речницима доступним у локалној бази података, као и у ресурсима на вебу (више детаља било је у поглављу 5.2.5). Потом следе информације о нормализованим фреквенцијама у оквиру појединих корпуса, као и опције за претраживање доступних корпуса према предефинисаним обрасцима. Резултати се приказују у виду линија конкорданци из корпуса или приказа фреквенција облика или леме, у зависности од тога шта корисник одабере (више детаља било је у поглављу 5.2.3). Након ових података следе информације о значењима лексичког записа исказане кроз називе маркера на српском језику али и кроз саме ознаке маркера и домена. Уз одговарајућа значења су повезане и везе са лексичким записима полилексемских израза у чијем формирању учествује лексички запис (нпр. у оквиру једног значења записа „*лопта*“ је веза са лексичким записима „*тенис-лопта*“, „*Земљина лопта*“, „*брејк-лопта*“, итд.). Уређивање постојећег лексичког записа подразумева уређивање свих набројаних сегмената осим оних који се односе на корпусе и приказ записа у речницима доступним у локалној бази података и на вебу. Кроз сегмент за лексичке записе могуће је директно креирање новог лексичког записа (монолексемског или полилексемског израза). Детаљна упутства о коришћењу могуће је прочитати у Прилогу 2.

Сегментом за корпусе могућ је приступ метаподацима о корпусима који се користе у оквиру система, као и њиховом уређивању. Могуће је додавање фреквенција лексичких записа у текућем корпусу према методологији која је описана у поглављу 5.2.3. Детаљне инструкције о коришћењу овог сегмента апликације следе у Прилогу 2.

Сегмент за евалуацију кандидата за речник полилексемских израза служи као алат за потребе вишеевалуаторског рада на процени кандидата за *Речник*. Више информација о практичној примени биће у поглављу 7.1 док је упутство за употребу у Прилогу 2.

Сегмент за лексичке релације пружа приступ скуповима правила за успостављање постојећих веза међу лексичким записима (према методологији изложеној у поглављу 5.2.1). Овим сегментом се приступа креирању нових релација, њихових правила, као и евентуалној корекцији или брисању постојећих правила. Детаљно упутство за употребу је дато у Прилогу 2.

Кроз сегмент за приказ флективних образаца су могући приказ и уређивање метаподатака о свим флективним класама које се користе у оквиру *Речника*. Могуће је додавање новог флективног обрасца као и евентуално брисање постојећег. Више детаља о овом сегменту могуће је прочитати у Прилогу 2.

6.2 Претраживање лексикографске базе

Претраживање лексикографске базе је омогућено са више приступних тачака, у зависности од информационих потреба корисника. За претрагу лексичких записа су доступне три приступне тачке („Leximirka“, „Entries“ и „Full Table“), док је саме категорије могуће претраживати и уређивати путем панела „Categories“.

Уколико корисник жели да пронађе лексичке записе на основу леме и погледа њихов садржај (опис), најефикаснији начин да дође до описа је да користи основну претрагу на почетној страни апликације *Лексимирка*. Ову претрагу је могуће задати латиничним или ћириличним словима или коришћењем словних кодова из Ауроре. Упити задати латиницом „srećnik“ или ћирилицом „срећник“ су еквиваленти упиту „srecxnik“ задатим кодом Аурора и вратиће исти резултат. Када корисник жели да

ограничи број резултата на тачну ниску слова, може то учинити коришћењем помоћне опције „Exact match“. Уколико корисник жели да претражује по свим флективним облицима речи, може користити опцију „Search all forms“. Као илустрација ове претраге ће нам послужити претрага ниском „најсрећнији“ која уз примену опције „Search all forms“ враћа лексички запис „srećan“ будући да је задата ниска његов флективни облик.

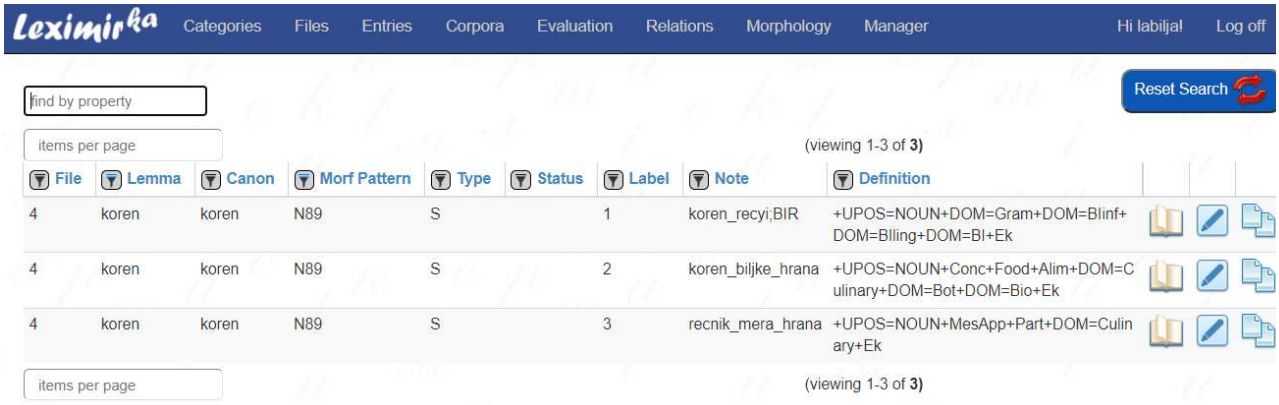
Преостале две приступне тачке за претрагу лексикографске базе су предвиђене за развојно окружење и служе за претрагу по свим елементима лексичког записа. Претрага се у оба случаја врши задавањем упита коришћењем искључиво кода Аурора. Сама претрага у начелу почива на употреби филтера уз доступне колоне у комбинацији са додатним главним пољем за претрагу. Код панела за претрагу лексичких записа је то поље „find entry“, док је код панела за претрагу пуне табеле то поље „find by property“. Следи више информација о овим начинима претраге.

Код панела за претрагу лексичких записа[110] претрага по пољу „find entry“ се састоји од задавања ниске која је садржана у леми записа и аутоматски започиње са уносом првог карактера. Дакле, уколико унесемо само слово „s“, у табели ће бити излистани сви записи који почињу овим словом. Претрага путем овог поља не разликује употребу малих и великих слова. Корисник је у могућности да сузи резултате претраге придруживањем додатних филтера за лексичке категорије предвиђене овим панелом. Панелом су предвиђена: поља за идентификациони број речника, лема, канонски облик леме, ознака за врсту речи, тип речи (монолексемска или полилексемска јединица), морфолошки образац, статус, језик и белешка уз лексички запис. Тако на пример, уз задату ниску „srecx“, која враћа 17 лексичких записа, можемо сузити резултат филтером за врсту речи додавањем услова да она почиње са „N“ (именицом) (POS StartsWith N). Сада као резултат добијамо 12 лексичких записа. Поред лексичких записа који представљају монолексемске речи које су именице („srecxkovicx“, „srecxa“, „Srecxko“ и друге) међу резултатима се налази и полилексемски израз „srecxan brak“ чија је ознака врсте речи „NC“.

Претрага путем панела који приказује пуну табелу „Full Table“[111] омогућава претрагу на нивоу свих излистаних лексичких информација. У овој табели су редови представљени на нивоу дефиниције значења лексичког записа. На пример, уколико лема има 3 значења, биће представљена са 3 реда у табели. Такав случај је са лексичким записом „koren“ који има 3 значења. Прво значење се односи на корен речи и дефинисано је ознакама „+DOM=Gram+DOM=Biinf+DOM=Biling+DOM=BI+Ek“, дакле припада доменима граматике, информатике, лингвистике и библиотекарства и информатике и представља екавски изговор. Друго значење које се односи на корен биљке који се може користити и за исхрану дефинисано је категоријама „+Conc+Food+Alim+DOM=Culinary+DOM=Bot+DOM=Bio+Ek“ које редом означавају конкретну именицу, храну, храну која се користи у кулинарству необрађена. Домени употребе су кулинарски, ботанички и биолошки, а изговор екавски. Треће значење се односи на меру у кулинарству и означено је категоријама „+MesApp+Part+DOM=Culinary+Ek“ које означавају приближну меру у кулинарству, као део веће целине, кулинарски домен и екавски облик. Дакле, овај лексички запис је у табели панела представљен са 3 реда и доступан за претрагу по све три дефиниције понаособ (слика 36).

[110] http://leximirka.jerteh.rs/LexicalEntry/LexicalEntry (приступљено 4.6.2020)

[111] http://leximirka.jerteh.rs/LexicalEntry/FullTable (приступљено 4.06.2020)



Слика 36 Приказ лексичког записа „koren“ кроз панел „Full Table“

Главно поље за претрагу „find by property“ је предвиђено за претрагу ознака категорија које се налазе у колони за дефиницију значења (колона Definition). У ово поље је могуће додавати више ознака као упит, једну по једну, чиме се надомешћује ограничење постављања једног филтера по колони. С друге стране је могуће остале колоне филтрирати по неком критеријуму.

У будућем раду би требало омогућити претрагу елиминацијом неке од ознака. Пример такве претраге би био излиставање лексичких записа који припадају домену геологије а нису пореклом из датотеке речника која садржи монолексемске јединице из области рударства и геологије “delas-mining.dic” (ид 25) или из датотеке полилексемских јединица „delac-geols.dic“ (ид 117) у којима се махом налазе речи из геолошког домена. Овакву претрагу бисмо започели преко панела који приказује пуну табелу (слика 36) претрагом својства “DOM=Geol” или филтрирањем по колони за дефиницију са „Contains“ “+DOM=Geol”. Добили бисмо 4.308 значења која задовољавају овај упит али ћемо приметити да су записи пореклом из различитих датотека речника. Моћи ћемо рецимо да филтрирамо лексичке записе који су увезени из датотеке са ид-ом 25 (филтрирање по колони “File” “Equals” 25) и да видимо да је у њој 2.338 значења али нећемо моћи да је искључимо како бисмо могли да видимо колико је значења пореклом из других датотека. У могућности смо да видимо значења лексичких записа пореклом из датотека чији је ид мањи од 25 или већи од 25 али то нема баш пуно смисла. Такође нисмо у могућности да излистамо значења записа пореклом из две датотеке истовремено. Нпр. из датотеке речника чији је ид 25 и датотеке чији је ид 117. Из истог разлога не бисмо могли да искључимо, на пример придеве који означавају да је нешто од неког материјала (маркер „MatA“).

Било би корисно уколико би лексикографска база омогућавала приступ корпусно заснованим информацијама о лексичким записима (речима) попут колокација које пружа опција скице речи[112] (енг. *Word sketch*) из алата за управљање корпусима Sketch Engine или уколико би била омогућена визуелизација података из лексикографске базе, нпр. остварених лексичких релација неке речи.

112 https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ (приступљено 12.02.2019)

6.3 Функционалности

6.3.1 Извоз речника у формату према потреби корисника

Апликација *Лексимирка* омогућава извоз различитих врста речника генерисаних на основу лексикографских података из базе. У пракси се више датотека речника монолексемских речи (DELAS) спаја у јединствен излазни речник ради компактнијег коришћења и компилације која је неопходна како би се могао користити у систему *Unitex*. У лексикографској бази се уз сваки назив датотеке речника монолексемских јединица (поље File) наводи име датотеке које ће датотека имати када се компилује у компоненту DELAF. Уколико је поље за име излазне датотеке празно, датотека речника се не спаја и користи се изворно име.

Систем омогућава да се приликом ове производње новог речника изоставе лексички записи према неком критеријуму. На пример, могу се изоставити записи који су означени маркером +Arch (архаично/покрајински) уколико речником обрађујемо савремене текстове. Такође је могуће ограничити маркере који се придружују речи тако да се остави само наведена листа, нпр. „+Conc+Hum+Bot+Zool+VN+PosQ+Comp“.

Развијене су процедуре које потпуно аутоматски креирају речнике према одређеним параметрима. Ово практично значи да се речници добијају на дугме, и то следећи често коришћени типови речника:

- Изворна и пуна верзија речника монолексемских јединица DELAS за јавну дистрибуцију
- Речници DELAF од пуне верзије речника монолексемских јединица
- Латинички и ћирилички речници DELAF од речника DELAF у Аурора кодирању
- Изведени инвертовани речници DELAF – разликују се од уобичајених речника по томе што се лема налази на првом месту у запису а облик речи на другом месту и практично служе за индексирање лема и друге примене.
- Компиловани речник DELAF
- Речник за рестаурацију дијакритика
- Речник за пребацивање екавских облика у ијекавске и обрнуто

Поред наведених могућности за извоз речника, у плану је прилагођавање апликације *Лексимирка* додавању у лексикографску базу нових комплетних речника (опција Upload Lexicon). Такође је у плану формирање „контејнера“ за новонастале речнике предвиђеног за складиштење лексичких записа креираних директно кроз апликацију од стране корисника до момента кад корисник са вишом привилегијом не потврди њихову ваљаност.

6.3.2 Веб-сервиси засновани на речнику

Према дефиницији Конзорцијума W3C, „веб-сервис је софтверски систем дизајниран да подржи несметану интеракцију између машина путем мреже. Он садржи сучеље описано у машински обрадивом формату. Други системи ступају у контакт са веб-сервисом на начин прописан описом сервиса коришћењем порука у формату SOAP[113], обично преведених коришћењем HTTP[114] протокола у облик XML серијализације у

[113] Simple Object Access Protocol

[114] Hypertext Transfer Protocol

складу са другим стандардима за веб[115]“ (Haas и Brown 2004). Сучеље које се користи за комуникацију између машина (програма) се назива програмским сучељем апликације – API. Једна од архитектура која се користи за веб-сервисе јесте REST[116]. У овој архитектури сервер пружа приступ ресурсима који су у формату попут текста, XML-а или JSON-а, док им клијент приступа и користи их према потреби.

Како би се омогућила употреба лексичких ресурса од стране апликација за обраду природних језика, развијају се веб-сервиси који овај задатак чине могућим. Конкретно скуп веб-сервиса *Вебран* омогућава проширење упита засновано на *Морфолошким речницима за српски језик* и другим лексичким ресурсима. Сервис *Вебран* је настао као надградња претходно коришћеног система названог *ewsQueryExpand* (Stanković 2009). Осим *Морфолошких речника за српски језик*, кроз *Вебран* је путем морфолошког и семантичког проширења упита могуће користити информације из *Ворднета за српски језик* (приказан у поглављу 2.33.1), терминолошког портала *Терми* (Kitanović и остали 2021), онтологије *РудОнто* (описана у одељку 4.2.2), геолошког терминолошког речника *ГеолИСС терм*[117], као и *Речника библиотекарства и информатике*[118]. Само проширење упита засновано на *Вебрану* подразумева проналажење синонима речи и њихово укључивање у упит, проналажење семантички повезаних речи као што су антоними, мерономи, хипоними и хипероними, као и проналажење свих флективних облика речи (Stanković и Utvić 2019). *Вебран* је заснован на Мајкрософтовој технологији *RESTful API* (*ASP.NET Web API*) за развој веб-сервиса коју одликује једноставност и флексибилност формата генерисаних података (Stanković, Krstev, Obradović, и Kitanović 2015).

Коришћење веб-сервиса *Вебран* за проширење упита спроводи се кроз три корака: апликација која користи услуге веб-сервиса путем URL адресе контактира веб сервис *Вебран* слањем упита који је корисник задао након чега *Вебран* претражује интегрисане лексичке ресурсе (између осталих и *Морфолошке речнике за српски језик*) и шаље резултате апликацији. Апликација потом, на основу одговора *Вебрана*, генерише крајњи и проширени упит за претрагу коју потом спроводи над ресурсом и приказује резултате (Утвић и остали 2019).

6.3.3 Примери имплементације веб-сервиса

Сервис *Вебран* се користи за проширење упита при претрази текстуалних ресурса у које спадају корпуси, дигиталне библиотеке и репозиторијуми. У наставку су дати примери примене на *Корпусу савременог српског језика – СрпКор* и доменском корпусу *РудКор,* дигиталној библиотеци *Библиша*[119] и *Дигиталном репозиторијуму Рударско-геолошког факултета - ДрРГФ*[120] (Popović, Škorić, и Rujević 2020).

---

[115] Изворно: „A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.“

[116] REpresentational State Transfer

[117] http://geoliss.mre.gov.rs/recnik/ (приступљено 8.01.2020)

[118] http://rbi.nb.rs/sr/home.html (приступљено 8.01.2020)

[119] http://jerteh.rs/biblisha/ (приступљено 8.01.2020)

[120] http://dr.rgf.bg.ac.rs/s/repo/indeks?q= (приступљено 4.04.2021)

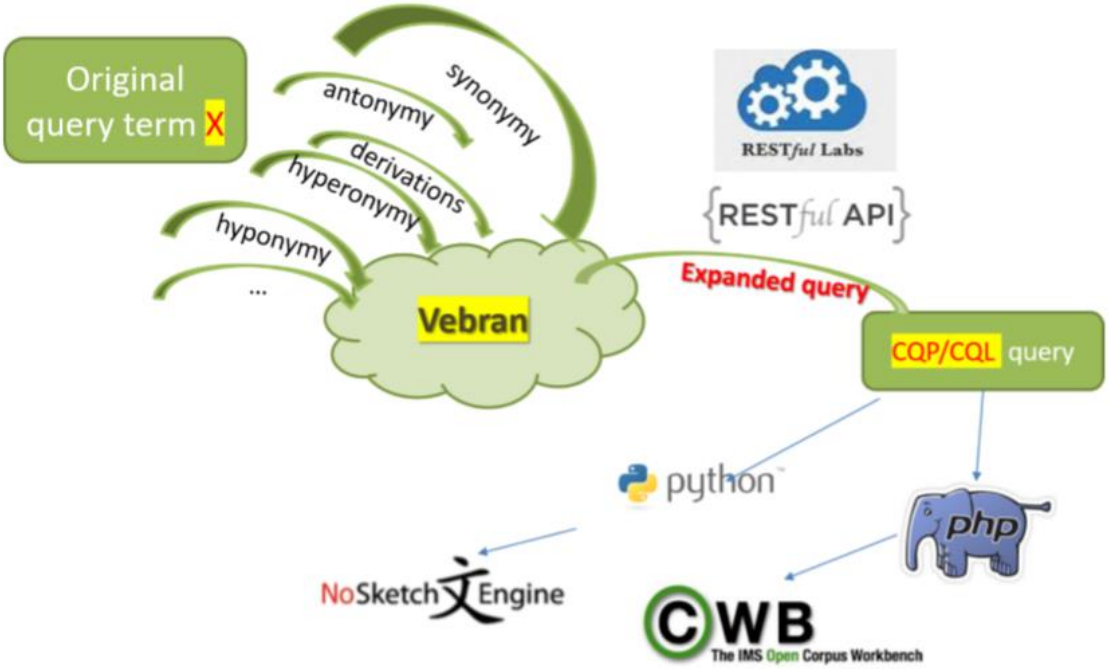
На слици 37 приказана је архитектура система за проширење упита заснована на веб-сервису *Вебран* и примењива на корпусе *СрпКор* и *РудКор*. Детаљан опис семантичког проширења упита дат је у раду (Утвић и остали 2019). На слици је са X означена полазна лексема коју корисник поставља као упит. Лексема може бити проширена, тј. повезана са осталим терминима семантичком релацијом (на слици синонимијом, деривацијом, антонимијом, хипонимијом, хиперонимијом и др.) или флективним облицима. Посредством технологије *RESTful API* се успоставља веза са *Вебраном*. Сам проширени упит се прослеђује алатима за претрагу корпуса *NoSketch Engine*[121] и *CWB*. Ова два алата су отвореног кода па је њихово прилагођавање за проширење упита начињено у програмским језицима *PHP (*CWB) и *Python* (*NoSketch Engine*).

Са̂мо морфолошко проширење упита при претрази корпуса се користи како би се побољшали резултати претраге јер на тај начин расте одзив. Али може бити проблема са падом прецизности резултата због потенцијалних проблема насталих коришћењем алата *TreeTagger* за обележавање врста речи. Један од могућих проблема би био фаворизовање одређених лема у односу на друге када више лема садржи хомографне флективне облике. Морфолошко проширење упита се заснива на замени лексеме X унијом њених флективних облика који су дати на писму претраживаног текстуалног ресурса. За морфолошко проширење се користе четири функције развијене у веб-сервису *Вебран*: *delaf*, *obliciZaCQP*, *delafs* и *MWUzaCQP*.

Прва функција *delaf* очекује параметре за упит у формату JSON. Следи пример таквог упита за лексему *књижевност*:

```
(1) {
(2)       lema:’književnost’,
(3)       alphOut:’L’,
(4)       lngIn:’sr’,
(5)       lngOut:’sr’,
(6)       POS:’N’,
(7)       GramCats:’p’,
(8)       fleksije:false,
(9)       dlfByLemma:false
(10) }
```

[121] https://www.sketchengine.eu/nosketch-engine/ (приступљено 8.01.2020)



Слика 37 Архитектура система за проширење упита намењеног хибридној претрази корпуса коришћењем сервиса *Вебран* (Stanković и Utvić 2019)

Овим упитом се захтева проширење облицима леме *књижевност* (линија 2) на латиничном писму (линија 3), да је језик уноса српски (линија 4) и језик добијених резултата српски (линија 5), са именицом као врстом речи леме (линија 6), означеном граматичком категоријом броја са вредношћу множине (линија 7), индикатором који показује да резултат садржи само лему X или такође њене флективне облике (линија 8), индикатор који показује да ли се излазни резултат приказује у виду облика груписаних према леми или не (линија 9). Добијени резултат у облику регуларног израза `književnost(i|ima)` покрива све облике у множини.

Функција *obliciZaCQP* захтева лему као улазни параметар претраге. Врста речи је опциона и намењена примени при претрази корпуса *СрпКор* јер се за писмо резултата захтева *Ауропа.* Ово значи да линија 3 упита у формату JSON гласи: `alphOut: 'A'`. Резултат за претходни упит би био: `knxizxevnost(i|ima).`

За потребе морфолошког проширења упита у алатима за претрагу *OCWB* и *NoSketch Engine* који се користе у корпусима *СрпКор* и *РудКор* у синтаксу је уведен лажни позициони атрибут `flemma` (подразумева флективну парадигму) како би корисник могао да захтева морфолошко проширење упита. У овом случају корисник преко алата креира упит у облику `flemma=X` док веб-сервис *Вебран* враћа упит `word=flektivni_oblici` прилагођен синтакси алата за претрагу.

Трећа функција именована као *delafs* користи исте параметре упита у формату JSON (друга линија - alphOut:'CL') али производи излазни резултат у облику листе облика. Резултат нашег упита би био наведен у облику: `književnosti; književnostima; књижевности; књижевностима`. Овакав облик упита се користи у дигиталним библиотекама *Ромека* (заснована на платформи *Омека*), *Библиша* (*mongoDB*) и репозиторијуму *ДрРГФ* (платформа *Омека*) пошто њихови системи захтевају такав формат.

За потребе проширења упита постављених полилексемским јединицама посредством алата *CQP* развијена је четврта функција *MWUzaCQP*. Ова функција користи информације о полилексемским јединицама садржане у *MPC*, као и примену додатних

хеуристика за обраду полилексемских јединица које нису у речнику. Наиме, у последњем случају систем, на основу уграђеног знања, предвиђа флективне облике полилексемских јединица.

Ако се полилексемска јединица постављена као упит од стране корисника налази у речнику *MPC,* анализира се њена флективна класа и примењује се одговарајућа трансформација. На пример, за кориснички упит *игра речи* који представља полилексемски израз који се налази у речнику *MPC,* систем враћа флективне облике који су у складу са флективном класом израза - *N2X* која означава да се само прва компонента која је именица флективно мења док је друга компонента непромењива. Сервис онда генерише одговор у облику: `C:igra_L reči_W` где `C` означава сложену реч, `L` означава лему и `W` означава реч. У крајњем упиту ће `L` бити замењено атрибутом `lemma` што значи да се ниска флективно мења, док ће `W` бити замењено атрибутом `word` – непромењиви део полилексемске јединице: `[lemma="igra"][word="reči"]`.

Уколико полилексемска јединица задата као упит није у речнику *MPC,* свака компонента се анализира засебно. Уколико је у *MPC* пронађено да је компонента флективни облик који се разликује од леме, она се мења одговарајућом лемом и обележава са `L`. Ако компонента није пронађена у *MPC,* обележава се као непромењива реч `W`. Уколико корисник унесе упит *емотивна веза* који представља полилексемску јединицу која није у *MPC,* анализира се свака компонента засебно. Како су обе компоненте заступљене у речнику као монолексемске речи, сервис генерише облик `C:emotivna_L veza_L`. Крајњи упит ће изгледати: `[lemma="emotivan"][lemma="veza"]`.

У овим случајевима сервис не генерише регуларни израз који покрива све флективне облике, већ се користи позициони атрибут `lemma` и информације о анотацији које пружа *TreeTagger.*

7. Евалуација резултата на лексици (терминологији) из геолошког домена

7.1 Екстракција терминологије

Корпус текстова из геолошког домена (*ГеоСрпКор*), развијен за потребе овог истраживања, чине текстови тумача за листове основних геолошких карата (ОГК) објављени у интервалу од 1962. до 1994. године као издања Савезног геолошког завода. Карта Србије је представљена у виду 78 листова ОГК у размери 1:100.000. Сваки лист карте је праћен текстуалним тумачем који даје приказ геолошких истраживања, геолошки развој и опис терена обухваћеног појединачним листом. Изузетак су поједини тумачи који описују 2 до 3 листа. Треба напоменути да је у корпус текстова ушло 69 тумача на српском језику јер су тумачи који описују пограничне листове Качаник, Куманово, Кратово и Гостивар објављени на македонском језику. Од 69 тумача, изворно је 7 објављено на ћириличном писму, док је остатак од 62 тумача објављен латинично. Тумачи су дигитализовани а ћирилични тумачи су за потребе истраживања корпуса конвертовани на латинично писмо. Текстови су такође првобитно припремљени тако што су елиминисане илустрације, табеле, литература, сажеци на другим језицима и стране са општим подацима о публикацији.

Процесирање описаног корпуса извршено је кроз окружење за управљање лексичким ресурсима апликације *Лексимир*. Корпус је обрађен применом лексичких ресурса за српски језик након избора корпуса „tumaci“ и одабиром дугмета „Apply Lexl Res“. Овим дугметом се примењују интегрисани графови и *Морфолошки речници за српски језик*. Извршено је сегментирање текста на реченице, као и препознавање речи из постојећег *Морфолошког речника за српски језик* заједно са мерењем фреквенција појављивања датих облика речи и лема. Након процесирања је од 62 документа добијен корпус који садржи 68.677 реченица, 2.621.193 токена, 72.778 различитих токена, 1.078.435 монолексемских облика, односно 72.611 различитих монолексемских облика. Такође су идентификоване и непознате речи које је апликација *Лексимир* приказала у виду табеларног документа „tumaci_out_unkownfreq.xlsx“. Ове речи делом представљају извор за кандидате за *Морфолошки речник српског језика* док су делом резултат грешака при куцању или оптичком препознавању карактера, с обзиром на то да су текстови тумача дигитализовани. У листи речи овог документа налазе се и речи из других језика. У овом корпусу су присутне речи из латинског језика, како су називи појединих врста биљака и животиња дати на истом језику, из енглеског и руског језика с обзиром на то да су поједини краћи сегменти текста којима је дата анализа слика представљени на српском, енглеском и руском језику. Табела „tumaci_out_unknownfreq.xlsx“ садржи колоне „token“, „out“, „freq“, „len“ и „grouped“ којима се олакшава филтрација кандидата за *Морфолошки речник српског језика*. У колони „token“ се налази списак непрепознатих облика. Колона „out“ омогућава примену неких предефинисаних образаца да би се искључиле речи (токени) које нису кандидати за речник. Примери таквих образаца су „low-freq“ који представља нискофреквентан облик, односно облик који се јавља у корпусу само једанпут, обрасци „<3“ или „=3“ представљају облик чија је дужина мања од 3 или укупно 3 карактера. Група непрепознатих облика „eng“ представља облике који су вероватно пореклом из енглеског језика препознате на основу генеричке процедуре у апликацији *Лексимир*. Ова процедура је заснована на узастопном понављању два слова и настала је јер се у пракси често дешава да се у оквиру текста налазе речи пореклом из енглеског језика. Група непрепознатих облика речи означена са „wq“ и „xy“ представља облике који садрже карактер w или q и x или y што значи да ти облици нипошто нису валидни облици речи из српског језика. Неки примери искључени овим обрасцима су: „quartzite“, „schweyeri“, „vortex“, „boisayi“, итд. Филтер „ee_oo“ који препознаје облике који у себи имају „ee“ или „oo“ је занемарен јер је примећено да се њиме изузима неколико

валидних термина попут „oolitske“, „zoogeno“ или „zoogenosprudni“. Коришћењем описаних филтера из колоне „out“ број кандидата за *Морфолошки речник* је са почетних 18.873 непрепознатих облика сведен на 5.510 облика.

У колони „len“ приказан је број карактера датог облика. С обзиром на то да смо у претходној колони искључили све облике чија је дужина највише 3 карактера, јер је мало вероватно да су они кандидати за речник, најкраћи облици имају 4 карактера и њих је укупно 94. Детаљном анализом смо одлучили да елиминишемо и ове кандидате јер међу њима није било валидних. Неки примери најфреквентнијих облика дужине 4 карактера су: „apta“[122] (фрекв. 127), „alba“[123] (фрекв. 79), „arca[124]“ (фрекв. 39) и „albu“ (фрекв. 33). Облици „intrabiomikrosparitima“ (фрекв. 5) и „kvarcdioritporfiritima“ (фрекв. 4) представљају најдуже облике речи. Након изузимања облика дужине 4 карактера, преостало је 5.416 облика речи.

У колони „freq“ дате су фреквенције појављивања датог облика у анализираном корпусу. Како су применом филтера „low-freq“ изузети облици са фреквенцијом 1, распон учестаности облика је од 2 појављивања (укупно 2.135 облика) до 633 појављивања (само један облик - „dijabaz“). Како би се број кандидата свео на прихватљивији за ручну анализу искључено је и 2.276 облика који се јављају 2 пута, тако да је преостало 3.140 облика за анализу.

Преостали кандидати су потом сортирани према колони „grouped“ путем које је извршено груписање сродних облика, чиме се значајно олакшава процес описа кандидата за *Морфолошки речник српског језика* јер се више облика једне леме налази у низу. На пример, облици „alevrolitsko“ (фрекв. 20), „alevrolitskog“ (фрекв. 5), „alevrolitske“ (фрекв. 16), „alevrolitska“ (фрекв. 21) и „alevrolitskoj“ (фрекв. 7) у колони „grouped“ имају вредност „alevrolitska“ те су у низу поређани један испод другог, иако им фреквенције појављивања варирају. Сами облици лема су одређивани ручно.

Кандидати за речник су потом допуњени граматичким и семантичким ознакама. Сви кандидати су допуњени ознакама за врсту речи а именице су обележене и информацијом о роду. Од семантичких ознака су коришћене ознаке домена и семантички маркери. Након ове фазе је извршено аутоматско предвиђање флективне класе кандидата уз помоћ коначних трансдуктора чиме су добијени комплетни записи у речнику. Пример новоформираног лексичког записа за лему „*микрит*“ изгледа овако:

```
mikrit,N1001+Stena+Mat+DOM=Geol
```

из кога се види да је додељена флективна класа речи „N1001“, да се ради о стени (+Stena) и материјалу (+Mat) и да реч припада домену геологије (+DOM=Geol). Више о начину одређивања флективне класе се може прочитати у раду (Krstev и остали 2015). Морфолошки речници су на овај начин допуњени са 1.217 нових монолексемских речи.

У наредној фази обраде је извршена екстракција полилексемских израза коришћењем графова који проналазе комбинације речи одређене структуре. Ови графови су конструисани на основу анализе синтаксичке структуре полилексемских термина из неколико постојећих терминолошких извора (*Терми*, *РудОнто*, *ГеоЛИСТерм*). Пример графа описаног обрасцем „N2X“ који у тексту проналази термин који се састоји од именице праћене речју која не подлеже флексији јесте „grf03“ и он проналази

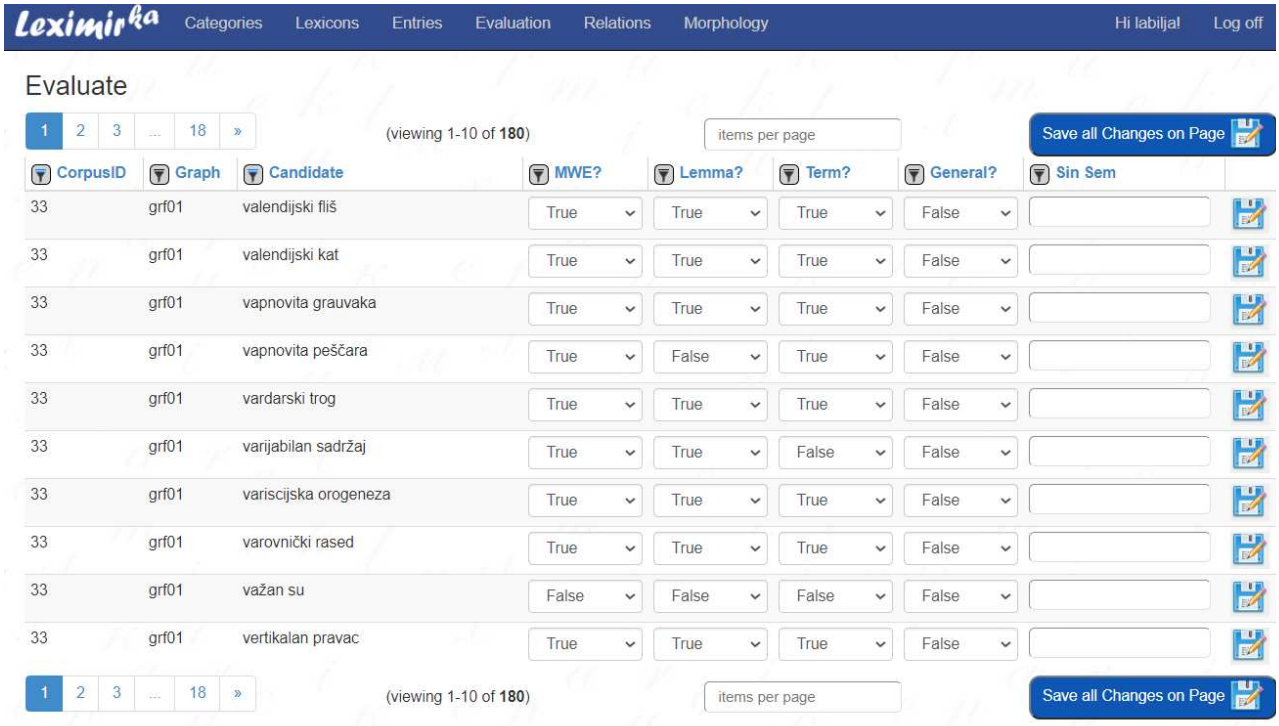
[122] Облик „*anta*“ представља флективни облик монолексемске јединице „*ant*“ – стадијум епохе доње креде у геолошкој хронологији.

[123] Облици „*алба*“ и „*албу*“ се односе на лему „*алб*“ – стадијум периода креде у геолошкој хронологији.

[124] Облик „*arca*“ је део латинских назива који су наведени у *Тумачима*.

вишечлане термине попут: „kartiranje lista“, „kolektor stena“, „ležište uglja“, итд. Сама екстракција терминологије је извршена уз помоћ апликације *Лексимир* и система *Unitex*. За екстракцију полилексемских израза из корпуса *ГеоСрпКор* је коришћено 12 графова.

Након извршене екстракције полилексемских кандидата према принципима описаним у (Krstev и остали 2015) за термине успостављена је методологија за евалуацију добијених резултата. Евалуација је извршена кроз истоимени панел (опција Evaluation) апликације *Лексимирка* приказан на слици 38. У евалуацији су учествовала два евалуатора (један је стручњак из области геологије а други аутор ове дисертације) чији је задатак био да евалуирају по 3.076 истих полилексемских кандидата. Разматрани су кандидати чија је фреквенција појављивања била 7 и више. Првобитно је на тестном узорку од 200 кандидата успостављена методологија за евалуацију у договору два евалуатора. Евалуатори су имали задатак да одреде да ли је дати кандидат (колона Candidate) полилексемски израз (колона MWE), да ли кандидат у понуђеном облику задовољава критеријум за лему (колона Lemma), да ли је термин из области геологије (колона Term) или представља термин у некој другој области (колона General). Колона Candidate се састоји од аутоматски екстрахованих термина и не пружа могућност дотеривања, нпр. исправног облика леме уколико кандидат није у облику леме. Изјашњавање се вршило одабиром позитивне опције (True) или негативне (False) из падајућег менија. Евалуатори су такође имали могућност да додају синтаксичко-семантичке ознаке кроз за то предвиђену колону SinSem. Чување измена се може вршити појединачно на нивоу кандидата (опција) или на нивоу целе стране (опција).



Слика 38 Панел за евалуацију полилексемских израза

Коенова капа (енг. *Cohen's Kappa*) (Cohen 1960) је мера која је коришћена за оцену евалуације кандидата. Коефицијент *Коенова капа* је статистичка мера за одређивање сагласности у анотацији између два евалуатора номиналном скалом мерења. Користи се када су два евалуатора независна један од другог.

Матрица сагласности у погледу евалуације тога да ли кандидат представља полилексемски израз дата је у табели 10. Оценом 1 је означен потврдан став (у евалуацији True), док је оценом 0 представљен одричан суд евалуатора (у евалуацији False) о томе да ли је кандидат полилексемски израз.

Табела 10 Матрица сагласности евалуатора 1 и 2 (полилексемски израз)

Евалуатор 1				
Евалуатор 2	Оцена	0	1	Укупно
	0	321 (a)	48 (b)	369 (a+b)
	1	91 (c)	2.616 (d)	2.707 (c+d)
	Укупно	412 (a+c)	2.664 (b+d)	3.076 (a+b+c+d)

Из табеле 10 по дијагонали означеној наранџастом бојом можемо видети да су се оба евалуатора сложила да 2.616 кандидата представља полилексемски израз, док 321 кандидат не представља полилексемски израз. По дијагонали табеле означеној жутом бојом видимо да се евалуатори нису сложили код знатно мањег броја кандидата. Евалуатор 1 је оценио да 91 кандидат не представља полилексемски израз а да 48 кандидата представља полилексемски израз, док евалуатор 2 тврди супротно.

За рачунање коефицијента *Коенова капа* (k) потребне су две мере: P_o - проценат опсервација у којима постоји слагање евалуатора и P_e – очекивана вероватноћа слагања евалуација.

Проценат опсервација у којима постоји слагање евалуатора добија се када се укупан број слагања оба евалуатора (збир по наранџастој дијагонали) подели укупним бројем оцена (укупним бројем кандидата). У нашем случају P_o износи:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{321+2616}{3076} = 0.9548$$

Очекивана вероватноћа слагања евалуација P_e представља збир вероватноћа слагања евалуатора, дакле збир вероватноће да ће оба евалуатора имати потврдан суд и вероватноће да ће оба евалуатора имати негативан суд о предмету евалуације:

$$P_e = P_1 + P_0$$

Вероватноће да ће оба евалуатора имати негативан суд - P_0 добија се формулом:

$$P_0 = \frac{a+c}{a+b+c+d} \cdot \frac{a+b}{a+b+c+d} = \frac{412}{3076} \cdot \frac{369}{3076} = 0.0161$$

Вероватноћа да ће оба евалуатора имати потврдан став - P_1 се добија формулом:

$$P_1 = \frac{b+d}{a+b+c+d} \cdot \frac{c+d}{a+b+c+d} = \frac{2664}{3076} \cdot \frac{2707}{3076} = 0.7622$$

Дакле, очекивана вероватноћа слагања евалуација P_e износи 0.7782.

И коначно, формула за израчунавање коефицијента *Коенове капе* гласи:

$$k = \frac{Po-Pe}{1-Pe} = \frac{0.9548-0.7782}{1-0.7782} = 0.7962$$

На основу представљених прорачуна према којима *Коенова капа* износи 0.7962 и увидом у табелу 11 која представља ниво сагласности евалуатора у зависности од вредности овог коефицијента, долазимо до закључка да је ниво сагласности евалуатора по питању оцене да ли је кандидат полилексемски израз или не висок (0.61 – 0.80).

Табела 11 Ниво сагласности евалуатора на основу коефицијента *Коенова капа* (k)

Коенова капа (k)	**Ниво сагласности**
< 0.00	слаб
0.00 - 0.20	низак
0.21 - 0.40	коректан
0.41 - 0.60	умерен
0.61 - 0.80	висок
0.81 - 1.00	веома висок

Матрица сагласности у погледу евалуације тога да ли флективни облик кандидата представља лему без обзира на то да ли је кандидат термин или не дата је у табели 12. Оценом 1 је означен потврдан став, док је оценом 0 представљен одричан суд евалуатора о томе да ли је облик кандидата прикладан за лему.

Табела 12 Матрица сагласности евалуатора 1 и 2 (лема)

		Евалуатор 1		
Евалуатор 2	Оцена	0	1	Укупно
	0	570 (a)	217 (b)	787 (a+b)
	1	2 (c)	2.287 (d)	2.289 (c+d)
	Укупно	572 (a+c)	2.504 (b+d)	3.076 (a+b+c+d)

Из табеле 12 по дијагонали означеној наранџастом бојом видимо да су се оба евалуатора сложила да 2.287 кандидата представља облик адекватан за лему, док њих 570 не испуњава тај критеријум. По дијагонали табеле означеној жутом бојом видимо да је евалуатор 1 оценио да 2 кандидата не представљају адекватан облик за лему док њих 217 представља исправан облик за лему. Евалуатор 2 је дао супротне оцене.

Проценат опсервација у којима постоји слагање евалуатора - P_o добија се на следећи начин:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{570+2287}{3076} = 0.9288$$

Вероватноће да ће оба евалуатора имати негативан суд - P_0 добија се формулом:

$$P_0 = \frac{a+c}{a+b+c+d} \cdot \frac{a+b}{a+b+c+d} = \frac{572}{3076} \cdot \frac{787}{3076} = 0.0476$$

Вероватноћа да ће оба евалуатора имати потврдан став - P_1 се добија формулом:

$$P_1 = \frac{b+d}{a+b+c+d} \cdot \frac{c+d}{a+b+c+d} = \frac{2504}{3076} \cdot \frac{2289}{3076} = 0.6058$$

Дакле, очекивана вероватноћа слагања евалуација P_e износиће:

$$P_e = P_1 + P_0 = 0.6058 + 0.0476 = 0.6533$$

На основу свега наведеног коефицијент *Коенова капа* ће износити:

$$k = \frac{Po-Pe}{1-Pe} = \frac{0.9288-0.6533}{1-0.6533} = 0.7946$$

Вредност коефицијента *Коенова капа* од 0.7946 нам говори да је ниво сагласности два евалуатора по питању оцене тога да ли је облик кандидата прикладан за лему висок (на основу табеле 11).

Следећа матрица сагласности, представљена табелом 13, представља оцене евалуатора да ли кандидат представља термин у домену геологије или не. Оценом 1 је означен потврдан став, док је оценом 0 представљена одрична оцена евалуатора.

Табела 13 Матрица сагласности евалуатора 1 и 2 (термин у геолошком домену)

Евалуатор 1				
Евалуатор 2	Оцена	0	1	Укупно
	0	451 (a)	128 (b)	579 (a+b)
	1	175 (c)	2.322 (d)	2.497 (c+d)
	Укупно	626 (a+c)	2.450 (b+d)	3.076 (a+b+c+d)

Из табеле 13 по дијагонали означеној наранџастом бојом видимо да су се оба евалуатора сложила да 2.322 кандидата представљају термине у области геологије, док њих 451 то нису. По дијагонали табеле означеној жутом бојом видимо да је евалуатор 2 оценио да 128 кандидата не представља геолошки термин, док њих 175 јесте термин, док евалуатор 1 износи супротне оцене.

Проценат опсервација у којима постоји слагање евалуатора - P_o израчунава се на следећи начин:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{451+2322}{3076} = 0.9015$$

Вероватноћа да ће оба евалуатора имати негативан суд - P_0 добија се формулом:

$$P_0 = \frac{a+c}{a+b+c+d} \cdot \frac{a+b}{a+b+c+d} = \frac{626}{3076} \cdot \frac{579}{3076} = 0.0383$$

Вероватноћа да ће оба евалуатора имати потврдан став - P_1 се добија формулом:

$$P_1 = \frac{b+d}{a+b+c+d} \cdot \frac{c+d}{a+b+c+d} = \frac{2450}{3076} \cdot \frac{2497}{3076} = 0.6466$$

Очекивана вероватноћа слагања евалуација P_e износиће:

$$P_e = P_1 + P_0 = 0.6466 + 0.0383 = 0.6849$$

Коефицијент *Коенова капа* ће износити:

$$k = \frac{Po-Pe}{1-Pe} = \frac{0.9015-0.6849}{1-0.6849} = 0.6874$$

Увидом у табелу 11 видимо да вредност коефицијента *Коенова капа* од 0.6874 показује да је ниво сагласности два евалуатора по питању оцене да ли је кандидат термин у области геологије поново висок.

Четврта матрица сагласности, приказана табелом 14, представља оцене евалуатора да ли кандидат представља термин у области која није геологија. Оценом 1 је означен потврдан став, док је оценом 0 представљена одрична оцена евалуатора.

Табела 14 Матрица сагласности евалуатора 1 и 2 (термин у домену који није геологија)

Евалуатор 1				
Евалуатор 2	Оцена	0	1	Укупно
	0	2.803 (a)	253 (b)	3.056 (a+b)
	1	8 (c)	12 (d)	20 (c+d)
	Укупно	2.811 (a+c)	265 (b+d)	3.076 (a+b+c+d)

По дијагонали у табели 14 означеној наранџастом бојом видимо да су се оба евалуатора сложила да свега 12 кандидата представља термине у области која није геологија, док њих чак 2.803 то нису. По дијагонали табеле означеној жутом бојом видимо да је евалуатор 2 оценио да 253 кандидата не представља термин из геологије, док њих 8 то јесте, док евалуатор 1 истовремено тврди супротно.

Проценат опсервација у којима постоји слагање евалуатора - P_o добићемо на следећи начин:

$$P_o = \frac{a+d}{a+b+c+d} = \frac{2803+12}{3076} = 0.9151$$

Вероватноћа да ће оба евалуатора имати негативан суд - P_0 добија се на следећи начин:

$$P_0 = \frac{a+c}{a+b+c+d} \cdot \frac{a+b}{a+b+c+d} = \frac{2811}{3076} \cdot \frac{3056}{3076} = 0.9079$$

Вероватноћа да ће оба евалуатора имати потврдан став - P_1 се добија формулом:

$$P_1 = \frac{b+d}{a+b+c+d} \cdot \frac{c+d}{a+b+c+d} = \frac{265}{3076} \cdot \frac{20}{3076} = 0.0006$$

Очекивана вероватноћа слагања евалуација P_e износиће:

$$P_e = P_1 + P_0 = 0.0006 + 0.9079 = 0.9085$$

Коефицијент *Коенова капа* ће износити:

$$k = \frac{Po-Pe}{1-Pe} = \frac{0.9151-0.9085}{1-0.9085} = 0.0721$$

Увидом у табелу 11 запажамо да вредност коефицијента *Коенова капа* од 0.0721 показује да је ниво сагласности два евалуатора по питању оцене да ли је кандидат термин у области која није геологија низак.

Закључак изведен на основу примене коефицијента *Коенова капа* јесте да су евалуатори имали висок ниво сагласности приликом евалуације, осим у последњем случају, те је квалитет евалуације релевантан. С обзиром на то да је последњом матрицом сагласности утврђивано да ли кандидат представља термин у области која није геологија ниво сагласности је низак због ширине питања без фокуса на једну област или унапред понуђене области.

На основу евалуације креиран је Морфолошки речник полилексемских израза *delac-geols.dic* (ид 117) који садржи 2.004 лексичка записа која припадају геолошком домену. У речник су ушли полилексемски термини за које су се оба евалуатора сложила да су полилексемски израз, да имају исправну лему и да припадају домену геологије (DOM=Geol).

7.2 Обогаћивање речника маркерима

Током процеса допуне *Морфолошких речника српског језика* новим речима ексцерпираним из корпуса *ГеоСрпКор* коришћене су постојеће ознаке категорија података али су такође уведене и нове ознаке од значаја за геолошку терминологију. Категоријама података су придружени семантички маркери који означавају минерале „+Mineral“, стене „+Stena“, палеонтологију „+Paleo“, геолошки облик „+Oblik“, као и ознака за домен палеонтологије „+DOM=Paleo“.

Семантичким маркером “+Mineral” је означено 50 записа који представљају минерале. Неки од њих су: *андензин*, *епидот*, *фелдспат*, *селадонит*, *смитсонит*, *милерит*, итд. Семантичким маркером за стене обележен је 71 лексички запис који представља врсту стене. Неки од њих су: *базалт*, *микрит*, *рожнац*, *тералит*, итд. Маркером за палеонтологију означено је 83 лексичка записа из области палеонтологије. Пример таквих записа су: *молуска*, *цијанофита*, *гастроподски*, *остракода*, итд. Маркером за геолошки облик означено је 14 записа који представљају неки физички облик појављивања материје. Тако, на пример, *факолит* представља магматско тело у облику звона утиснуто уздуж слојних плоча синклинала и антиклинала, док *дајк* представља углавном вертикалну интрузију у постојећим стенама. Доменским маркером за палеонтологију означена су 83 лексичка записа која представљају термин из домена палеонтологије, као што су, на пример, *макрофаунистички*, *нанофосил*, *спонгија*, *хипурит*, итд.

7.3 Обогаћивање речника везама

Захваљујући успостављеној лексикографској бази и моделу за успостављање релација између одредница описаном у одељку 5.2.1 новонастали речник геолошких термина је допуњен одговарајућим релацијама.

Релацијом „relacioni pridev (_ski)“ која повезује именицу и придев повезане су леме: *каолин- каолински, серпентин – серпентински*, *ацетон – ацетонски, афиолит – афиолитски*, *алб – албски*, *алотроп – алотропски*, *антимон -антимонски*, *андезит – андезитски*, *Тутин – тутински*, *Жељин – жељински*, итд. Дакле, везе су успостављене применом правила које подразумева да се повезују именица и придев који има наставак „*ски*“. Треба напоменути да је пре повезивања речник обогаћен са три лексичка записа за које је утврђено да су недостајала у речницима: *афиолит*, *алб* и *алотроп*. Релацијом „relacioni pridev (_ni)“ која повезује именицу и придев са наставком „*ни*“ повезани су записи *антрацит – антрацитни, туф – туфни, гнајс – гнајсни, кварцит – кварцитни*, итд.

Креирана је нова деривациона релација „gradivni pridev“ која повезује именицу са повезаним градивним придевом, типом речи који је доста заступљен у геологији. Ова релација је заснована на три правила. Првим правилом су повезане именице са наставком „*ар*“ и придев са суфиксом „*ровит*“. На овај начин је повезан пар *бигар – бигровит*. Другим правилом су повезане именице без специфичног наставка са придевима који се завршавају наставком „*овит*“. Овим правилом су повезани парови: *туф – туфовит*, *лапор – лапоровит*, *сумпор – сумпоровит*, итд. Трећим правилом су повезане именице које се завршавају наставком „*ак*“ и придеви који садрже наставак „*ковит*“. Ово правило повезује парове *банак – банковит*, *песак – песковит*, *шљунак – шљунковит*, итд.

Такође је креирана деривациона релација „imenica proces“ која повезује именице са одговарајућим именицама које означавају процесе с обзиром на то да је приметна честа употреба оваквих именица у геолошком домену. Овом релацијом су повезане именице са именицама које означавају процесе. Универзални критеријум за остваривање ове везе јесте да је циљана именица означена семантичким маркером именица за процес (+Process). За сада ова релација почива на једном правилу. Оно се састоји од тога да су полазни и циљани лексички запис именице и да друга именица има подниску (суфикс) „*изација*“. На овај начин је на почетку успостављено 24 релације међу лексичким записима. Међу њима су парови: *фелдспат – фелдспатизација*, *јон – јонизација*, *фосил – фосилизација*, *угљен – угљенизација*, *каолин – каолинизација*, итд. Примећено је да неки термини који означавају процесе не садрже маркер тако да су лексички записи допуњени одговарајућим маркером (+Process). Након ове допуне је укупно успостављено 33 пара лексичких записа. Неки од новоуспостављених парова су: *пелет – пелетизација*, *минерал – минерализација*, *карбонат – карбонатизација*, *калцит – калцитизација*, итд.

7.4 Обрада текстова из домена рударства и геологије са применом и без примене развијених речника

Како бисмо проверили резултат примене допуњених речника за област геологије упоредићемо резултате обраде корпуса *ГеоСрпКор* кроз систем *Unitex* пре и након допуне речника.

Увидом у датотеку која даје листу монолексемских речи у анализираном тексту (dlf) генерисану од стране система *Unitex* дошли смо до података да је пре допуне речника систем пронашао 260.587 флективних облика монолексемских речи. Након допуне речника тај број је порастао на 316.067 облика речи.

Након увида у датотеку која пружа информације о флективним облицима полилексемских израза (dlc) дошли смо до закључка да је пре допуне речника систем пронашао 3.246 флективних облика полилексемских израза. Након допуне речника тај број је порастао на 17.056 флективних облика.

Пре допуне *Мофолошких речника српског језика* систем *Unitex* је након обраде текста корпуса *ГеоСрпКор* у датотеци са непрепознатим облицима речи вратио 18.874 облика. Након обраде корпуса *ГеоСрпКор* допуњеним *Мофолошким речницима српског језика* датотека са непрепознатим речима коју производи систем *Unitex* је вратила 14.893 непрепозната облика речи. Анализом ових облика долази се до закључка да се већим делом ради о облицима који су погрешно откуцани или из других језика (нпр. енглеског или латинског).

Међу ваљане облике са ниском фреквенцијом спадају примери са фреквенцијама у загради: *зеолитисање* (1), *термометаморфизма* (1), *термометаморфозе* (1), *термометасоматске* (2), *термометасоматским* (2).

Лема *ситнозрн* пропустом није ушла у речник. Следе њени облици са апсолутним фреквенцијама: *ситнозрн* (76), *ситнозрна* (17), *ситнозрне* (165), *ситнозрни* (361), *ситнозрних* (204), *ситнозрнија* (3), *ситнозрније* (6), *ситнозрнијег* (1), *ситнозрнијем* (1), *ситнозрнији* (21), *ситнозрнијих* (6), *ситнозрнијим* (7), *ситнозрним* (123), *ситнозрно* (3), *ситнозрн* (76), *ситнозрној* (8), *ситнозрном* (14), *ситнозрну* (13).

Облици са фреквенцијама 1 свакако нису разматрани па стога у речник нису ушле леме које су валидне али су појединачни облици речи имали ниску фреквенцију:

- *мезотермални* чије фреквенције појављивања појединачних облика су ниске - *мезотермална* (1), *мезотермални* (1), *мезотермалним* (2), *мезотермалног* (2), *мезотермалној* (1), *мезотермалну* (2);
- *егзоконтактни* чије фреквенције појављивања облика су ниске: *егзоконтактне* (2), *егзоконтактни* (1), *егзоконтактних* (1), *егзоконтактно* (2), *егзоконтактној* (2).

Облици који такође нису ушли у речник јесу они који почињу речју *слаб* и вокалом *о* (*слабо-*) који се доста користи у геолошкој терминологији. Следе такви примери који су валидни али су се у корпусу појавили само једном: *слаболапоровитим*, *слабопокретљиве*, *слабосортираним*, *слабоуслојени*, *слабоуслојеним*, *слабовезним*, *слабозаобљених*.

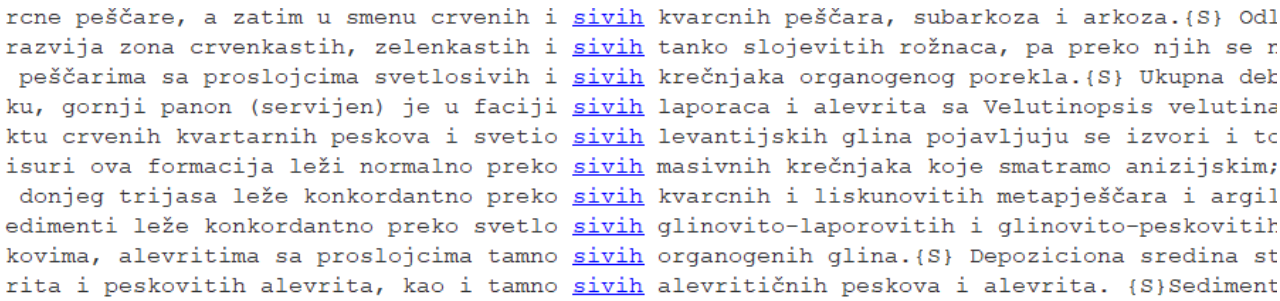
Интересантнији пример речи које нису обухваћене због ниске фреквенције појављивања јесу нијансе боја које се у геологији доста користе. У речник нису ушли следећи облици нијанси: *светлосивосмеђе* (1), *светлозеленкасте* (1), *сиворумене* (2), *сиворумени* (1), *сиворуменим* (1), *сиворуменкасте* (2), *сиворуменкасти* (1), *сивољубичасте* (1), *сивољубичастог* (1), *сиворђасти* (1), *сивкастоцрним* (1), *сивкастозелена* (1). С друге стране, додати су записи за следеће нијансе: *блиједосмеђ*, *мркозеленкаст*, *затвореносив*, *затворенозелен*, *затвореноцрвен*, *сивозеленкаст*, *сивожућкаст* и *сивоплавичаст*.

Претрагом корпуса *ГеоСрпКор* коришћењем система *Unitex* и применом допуњених *Морфолошких речника за српски језик* регуларним изразом „<A+Col>“ који проналази придеве означене семантичким маркером који означава боју добија се листа од 4.695 линија конкорданци. У табели 15 су у две колоне дате леме боја праћене фреквенцијама појављивања у корпусу *ГеоСрпКор*. На основу табеле се види да су најчешће у употреби чисте боје, и то највише сива (730 појављивања), зелена (570 појављивања), црвена (527 појављивања), црна (449 појављивања), бела (378 појављивања) и жута (193 појављивања). У табели на крају видимо и погрешно препознате облике *Морава* (275) и *био* (158). Облик *Морава* је препознат као облик леме придева *морав* који је пореклом из турског језика и чије значење је тамномодар (*Речник српскохрватскога књижевног језика. Књ. 3, К-О: (Косјерић-огранути)* 1969) а заправо се односи на реку Мораву. Облик *био* који се у пронађеним конкорданцама односи на глагол *бити* је препознат као флективни облик леме *бијел.*

Табела 15 Леме боја у корпусу *ГеоСрпКор*

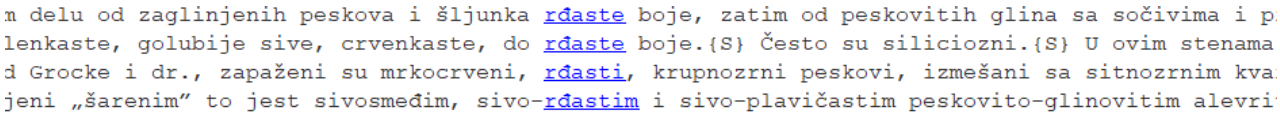
Боја	бр. појављивања	Боја	бр. појављивања
бео	378	сивозелен	53
бијел	63	сивозеленкаст	21
беличаст	41	сивожућкаст	11
беложут	5	сивожут	23
бледољубичаст	2	смеђ	124
бледосив	2	сребрн	2
бледозелен	4	светломрк	2
блиједосмеђ	3	светлосив	56
црн	449	светлосмеђ	5
црвен	527	светлозелен	22
црвенкаст	105	светложут	6
љубичаст	52	свијетлосив	6
млечнобео	1	тамноцрвен	2
мркоцрвен	35	тамнољубичаст	1
мркосив	6	тамномрк	7
мркозеленкаст	2	тамноплав	3
мркожут	17	тамносив	146
окер	5	тамносмеђ	4
окераст	6	тамнозелен	29
пепељастосив	1	затвореноцрвен	6
плав	48	затвореносив	16
плавичаст	55	затворенозелен	4
плавосив	2	затвореножут	4
плавозелен	1	зелен	570
прљавобео	2	зеленкаст	62
рђаст	4	зеленкастосив	8
румен	52	златан	1
ружичаст	19	златножут	1
сив	730	жућкаст	113
сивобео	29	жућкастосив	8
сивоцрн	4	жут	193
сивомаслинаст	4	жутомрк	7
сивомрк	15	жутосмеђ	4
сивоплав	18	жутозелен	3
сивоплавичаст	21	морава	275
сивосмеђ	31	био	158

На слици 39 је дат приказ 10 насумичних линија конкорданци за лему *сив*. Видимо у 9. и 10. линији у оквиру левог контекста да је облик *тамносивих* написан у облику дволексемског израза (*тамно сивих*) што прави проблем у лексичком препознавању па облици нису препознати као облик леме *тамносив*.



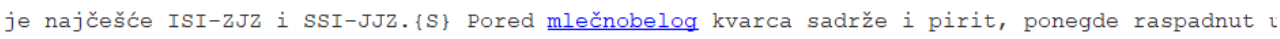
Слика 39 Линије конкорданци за лему *сив*

Интересантан пример употребе нијансе јесте рђаста боја. На слици 40 су приказане 4 линије конкорданци које садрже лему *рђаст*. Видимо да су пескови и шљункови описани том нијансом боје. У последњој линији видимо да није препозната сиворђаста нијанса јер је у тексту облик "*сиво-рђастим*" и као такав није препознат. Видимо да у самом тексту има недоследности у писању јер је у левом контексту облик "*сивосмеђим*" док је у десном контексту облик "*сиво-плавичастим*" а оба означавају нијансу боје.



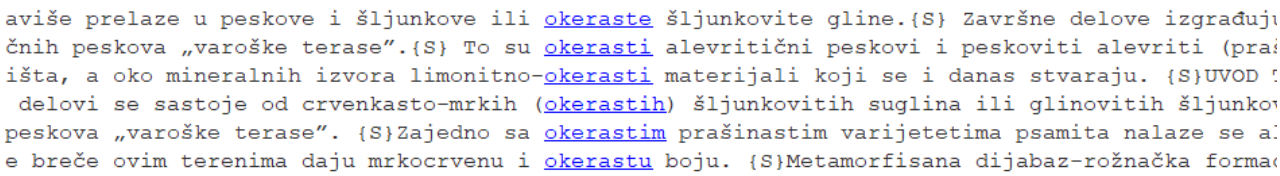
Слика 40 Линије конкорданци за лему *рђаст*

Линија конкорданце приказана на слици 41 приказује употребу млечнобеле нијансе. Видимо да се придевом *млечнобео* описује кварц.



Слика 41 Линије конкорданци за лему *млечнобео*

Окераста нијанса боје , како можемо видети на слици 42, користи се за опис глина, шљункова, пескова и стена (бреча). У 3. линији видимо непрепознат облик „*лимонитно-окерасти*" који на први поглед представља комбинацију градивног и описног придева. С обзиром на то да је лимонит по дефиницији руда гвожђа црвенкасте, смеђе или жућкасте боје рекло би се да ова комбинација придева значи да се мисли на материјале који су од црвенкастих (окерастих) лимонита.



Слика 42 Линије конкорданци за лему *окераст*

8. Закључак

8.1 Допринос

Овом докторском дисертацијом представљен је процес трансформације лексичког ресурса *Морфолошки речници за српски језик* из скупа датотека у лексикографску базу. Лексикографска база заснована на моделу који комбинује оквир за лексичко обележавање – *LMF*, *lemon* модел и каталог категорија података *DCR*, је послужила као основа за развијање веб-апликације *Лексимирка* за развој, одржавање и прегледање *Морфолошких речника за српски језик*. Новоразвијеном веб-апликацијом *Лексимирка* је омогућен истовремени рад више особа на развоју речника што није било могуће у десктоп апликацији *Лексимир* која се претходно користила за управљање речницима. Употреба апликације је такође допринела могућности новог начина коришћења *Морфолошких речника за српски језик*. Осим основне намене речника која се огледа у употреби од стране различитих апликација и програма за обраду природних језика, речник сада могу користити и људи. Апликација сада омогућава приказ хипервеза (или упутница) ка другим лексичким записима доступним у оквиру речника, преко успостављених релација са другим речима, као и преко приказа веза монолексемских речи и полилексемских израза чији су они конституент и обрнуто. Ознаке домена и семантички маркери су у оквиру прегледа лексичких записа видљиви на природном језику тако да их корисници могу читати у делу значења. С обзиром на то да су основним лексичким записима додате корпусне информације у виду фреквенција и конкорданци појављивања лема, као и везе са спољашњим лексичким ресурсима као што су *BabelNet*, *Wiktionary*, *Wikidata*, *Терми, Glosbe* апликација би могла бити коришћена приликом лексикографских истраживања. У раду је дат и пример како традиционални речници који су трансформисани у дигитални облик могу бити инкорпорирани у базу и приказани кроз сам лексички запис. Ови речници могу бити видљиви ограниченом кругу корисника имајући у виду заштиту ауторских права над речницима.

Имајући све наведено у виду, допринос дисертације се може представити као развој модела и софтверског решења лексикографске базе и апликације *Лексимирка* са бројним функцијама за одржавање и претраживање речника међусобно повезаних монолексемских и полилексемских јединица, његово повезивање са другим ресурсима и издавање произвољног скупа јединица у различитим форматима.

8.2 Правци за даљи рад

Завршетком рада на овој дисертацији сам процес усавршавања развијене базе и апликације није завршен. Како би се унапредиле могућности претраге лексикографске базе, треба радити на додавању претраге коришћењем Булових оператора или на додавању могућности за искључење неког критеријума претраге, као и за додавање више од два критеријума исте врсте.

Имајући у виду примере добре праксе у погледу опција које пружају апликације из области обраде природних језика, било би пожељно да корисник кроз апликацију *Лексимирка* може да прегледа колокације које прате одређени лексички запис.

Како би се апликација за управљање речницима користила у пуном капацитету потребно је увести финије рашчлањење корисничких налога. Потребно је увести категорије корисника одређене привилегијама за рад, нпр. лексикограф, рецензент, уредник, итд.

С обзиром на утврђену језичку независност базе података потврђену на примеру *Морфолошког речника за француски језик* у пракси треба тестирати

лексикографску базу и апликацију за речнике других језика. Неки од предлога јесу речници македонског, бугарског или латинског језика.

Користећи капацитет базе података отварају се бројне могућности за извоз речника у најразличитијим форматима сходно потребама. У будућности ће се свакако радити на извозу речника у форматима *RDF Turtle* и *RDF/XML*. С обзиром на тренд у представљању лингвистичких података у облику *Отворених повезаних података* треба размотрити опцију да један подскуп речника, нпр. геолошки речник, буде објављен у складу са тим смерницама. Како би се то постигло, било би корисно мапирати категорије података из *Морфолошких речника за српски језик* са онтологијом лингвистичких информација *LexInfo*.

9. Литература

„A037 Терминологија“. 2018. 2018. http://www.iss.rs/rs/tc/?national_committee_id=552.

Алексић, Данило. 2017. „Аутоматско прикупљање и обрада грађе за једно морфолошко истраживање“. Представљено на Семинару Друштва за језичке ресурсе и технологије, Београд, Март 30. http://jerteh.rs/wp-content/uploads/2016/03/Automatsko-prikupljanje.pdf.

Благојевић, Данка, Ранка Станковић, Петар Стејић, и Велизар Николић. 2014. „Србија у OneGeology Europe“. *Записници Српског геолошког друштва за 2013.*, 79–95.

Васиљевић, Небојша. 2014. „Аутоматска обрада правних текстова на српском језику“. Докторска дисертација, Београд: Универзитет у Београду, Филолошки факултет. https://fedorabg.bg.ac.rs/fedora/get/o:10687/bdef:Content/download.

Вујаклија, Милан. 2014. *Лексикон страних речи и израза*. Уредио Драго Ћупић. Београд; Подгорица: Мали пингвин.

Гортан-Премк, Даринка. 2010. „Ирена Грицкат-Радуловић - Велики Лексикограф и Учитељ Лексикографије“. *Јужнословенски Филолог*, изд. 66: 21–30. https://doi.org/10.2298/JFI1066021G.

Драгићевић, Рајна. 2010. *Лексикологија српског језика*. Београд: Завод за уџбенике.

———. 2014. „Развој практичне и теоријске лексикографије“. У *Савремена српска лексикографија у теорији и пракси*, уредио Рајна Драгићевић, 9–26. Београд: Филолошки факултет Универзитета у Београду.

———. 2018. *Српска лексика у прошлости и данас*. Нови Сад: Матица српска.

„Институт за стандардизацију Србије“. 2018. http://www.iss.rs/rs.

Јовановић, Владан. 2018. „Терминолошки речници“. У *Српска лексикографија од Вука до данас : каталог изложбе*, 189–201. Београд: САНУ : Савез славистичких друштава Србије.

Јовановић, Ранко. 1938. *Систематски речник српскохрватског језика*. Библиотека општег образовања. Б. м.: б. и.

Караџић, Вук Стефановић. 1818. *Српски рјечник : истолкован њемачким и латинским ријечма = Wolf Stephansohn's Serbisch-Deutsch-Lateinisches Wörterbuch = Lupi Stephani F. Lexicon Serbico-Germanico-Latinum*. У Бечу (Wien, Viennae): gedruckt bei den p.p. Armeniern. http://digital.bms.rs/pub.php?s=R19Sr-III-133.

Крстев, Цветана. 2000. „Српски језик у информатичком окружењу“. *Књижевност и језик*, изд. 1–2: 21–32.

———. 2019. „О одабиру одредница за електронски речник српског језика и њиховом повезивању“. У *Научни састанак слависта у Вукове дане*, уредио Божо Ћорић и Александар Милановић, 48/3:133–47. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. https://doi.org/10.18485/msc.2019.48.3.ch7.

Крстев, Цветана, Бојана Ђорђевић, Сања Антонић, Невена Ивковић-Берчек, Зорица Зорица, Весна Црногорац, и Љиљана Мацура. 2008. „Кооперативан рад на доградњи српског Wordneta“. *Инфотека* Год. 9 (1/2): 57–75. http://infoteka.bg.ac.rs/pdf/Srp/2008/INFOTHECA_IX_1-2_May2008_57-75.pdf.

Мацановић, Ана. 2018. „Рјечник хрватскога или српскога језика Југославенске академије знаности и умјетности“. У *Српска лексикографија од Вука до данас - каталог*

изложбе, 53–61. Београд: Српска академија наука и уметности и Савез славистичких друштава Србије.

Митровић, Јелена. 2018. „Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада“. Докторска дисертација, Београд: Универзитет у Београду, Филолошки факултет. https://fedorabg.bg.ac.rs/fedora/get/o:19057/bdef:Content/download.

Младеновић, Миљана. 2014. „Дигитални речник говора југа Србије“. *Инфотека* 15 (1): 42–55. http://infoteka.bg.ac.rs/pdf/Srp/2014/INFOTHECA_XV_1_2014_42-45.pdf.

Пантић, Марија. 2017. „Придевски атрибути уз именице ‚човек‘, ‚жена‘, ‚мушкарац‘ и ‚муж‘“. *Инфотека* Год. 17 (2): 69–98. http://infoteka.bg.ac.rs/pdf/Srp/2017-2/infoteka-2017-17-2-4.pdf.

Речник српскога језика. 2011. Нови Сад: Матица српска.

Речник српскохрватског књижевног и народног језика. Књ. 1, А-Богољуб. 1959. Београд: Институт за српскохрватски језик САНУ.

Речник српскохрватског књижевног и народног језика. Књ. 2, Богољуб-Вражогрнци. 1962. Београд: Институт за српскохрватски језик САНУ.

Речник српскохрватског књижевног и народног језика. Књ. 18, оповргавање - оцарити. 2010. Београд: Институт за српски језик САНУ.

Речник српскохрватског књижевног и народног језика. Књ. 19, оцат - петогласник. 2014. Београд: Институт за српски језик САНУ.

Речник српскохрватског књижевног и народног језика. Књ. 20, петогодан - погдегод. 2017. Београд: Институт за српски језик САНУ.

Речник српскохрватскога књижевног језика. Књ. 3, К-О: (Косјерић-огранути). 1969. Нови Сад; Загреб: Матица српска; Матица хрватска.

Речник српскохрватскога књижевног језика, Књ.1, А-Е. 1967. Нови Сад; Загреб: Матица српска; Матица хрватска.

Ристић, Стана. 2014. „Српске лексикографске институције. Београдска лексикографска школа“. У *Савремена српска лексикографија у теорији и пракси*, уредио Рајна Драгићевић, 27–54. Београд: Филолошки факултет Универзитета у Београду.

Ружин Ивановић, Татјана. 2018. „Двојезични речници“. У *Српска лексикографија од Вука до данас : каталог изложбе*, 112–27. Београд: САНУ : Савез славистичких друштава Србије.

Станојчић, Живојин, и Љубомир Поповић. 2020. *Граматика српског језика : за гимназије и средње школе*. 17. изд. Београд: Завод за уџбенике.

Стијовић, Рада, Цветана Крстев, и Ранка Станковић. 2021. „Аутоматска екстракција дефиниција – допринос убрзању израде речника“. У *Лексикографија и лексикологија у светлу актуелних проблема*, уредио Стана Ристић, Ивана Лазић Коњик, и Ненад Ивановић, 113–38. Београд: ИСЈ САНУ.

Стијовић, Рада, и Ранка Станковић. 2018. „Дигитално издање Речника САНУ: формални опис микроструктуре Речника САНУ“. Уредио Јелица Јокановић-Михајлов, Ана Кречмер, Александар Милановић, Драгана Мршевић-Радовић, Живојин Станојчић, Слободан Павловић, и Галина Тјапко. *Научни састанак слависта у Вукове дане* 47/1: 427–40. https://doi.org/10.18485/msc.2018.47.1.ch40.

Томашевић, Александра. 2018. „Развој модела за управљање рударском пројектном документацијом“. Докторска дисертација, Београд: Универзитет у Београду, Рударско-геолошки факултет. https://fedorabg.bg.ac.rs/fedora/get/o:18605/bdef:Content/download.

Томашевић, Александра, Биљана Лазић, Далибор Воркапић, Михаило Шкорић, и Љиљана Колоња. 2017. „Употреба веб платформе Омека за дигиталне библиотеке из домена рударства“. *Инфотека* Год. 17 (2): 27–51.

Тртовац, Александра. 2016. „Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама“. Докторска дисертација, Београд: Универзитет у Београду, Филолошки факултет.

Ћосић, Павле. 2008. *Речник синонима*. Едиција Вечници. Београд: Корнет.

Утвић, Милош. 2011. „Анотација Корпуса савременог српског језика“. *Инфотека* Год. 12 (2): 39–51.

Утвић, Милош, Ранка Станковић, Александра Томашевић, Михаило Шкорић, и Биљана Лазић. 2019. „Претрага корпуса заснована на употреби екстерних лексичких ресурса путем веб-сервиса“. У *Научни састанак слависта у Вукове дане*, уредио Божо Ћорић и Александар Милановић, 48/3:279–98. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. https://doi.org/10.18485/msc.2019.48.3.ch12.

Шапић, Јулија. 2018. „Просторни фрагментизатори у руском језику у поређењу са српским“. Докторска дисертација, Универзитет у Београду, Филолошки факултет. https://uvidok.rcub.bg.ac.rs/handle/123456789/2992.

„A history of HTML“. 1998. 1998. https://www.w3.org/People/Raggett/book4/ch02.html.

Ahmadi, Sina, John P. McCrae, Sanni Nimb, Fahad Khan, Monica Monachini, Bolette S. Pedersen, Thierry Declerck, и остали. 2020. „A multilingual evaluation dataset for monolingual word sense alignment“. У *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Aleksić, Marijana R. 2019. „Linguistic Coding of Evidentiality in Serbian and Spanish Journalistic Discourse“. *Филолог – часопис за језик књижевност и културу* 19 (19): 371–92. https://doi.org/10.21618/fil1919371a.

Bański, Piotr, Jack Bowers, и Tomaž Erjavec. 2017. „TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms“. У , 485–94. Brno: Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper29.pdf.

Begenešić, Dobrila. 2010. „Nemačko-srpska (srpskohrvatska) i srpsko (srpskohrvatsko)-nemačka stručna leksikografija u periodu 1945-2000“. Doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet. 10.2298/BG20101110BEGENISIC.

Berg, Donna Lee., Frank William Tompa, G. H. Gonnet, и University of Waterloo. 1988. *The New Oxford English dictionary project at the University of Waterloo*. OED ;88-01, 15 p. Waterloo, Ont.: UW Centre for the New Oxford English Dictionary. https://www.researchgate.net/publication/243451160_The_New_Oxford_English_Dictionary_Project_at_the_University_of_Waterloo.

Bergenholtz, Henning, и Theo JD Bothma. 2011. „Needs-adapted data presentation in e-information tools“. *Lexikos* 21: 53–77. https://doi.org/10.5788/21-1-37.

Bergenholtz, Henning, и Jesper Skovgård Nielsen. 2013. „What is a lexicographical database?“ *Lexikos* 23: 77–87.

Berners-Lee, Tim. 1998. „Why RDF model is different from the XML model“. *Why RDF model is different from the XML model* (blog). 1998. https://www.w3.org/DesignIssues/RDF-XML.html.

Boisvert, Eric, Ollie Raymond, и Marcus Sen, ур. 2016. „OGC Geoscience Markup Language 4.1 (GeoSciML)“. http://docs.opengeospatial.org/is/16-008/16-008.html#1.

Bosque-Gil, Julia, Jorge Gracia, John McCrae, Philipp Cimiano, Sander Stolk, Fahad Khan, Katrien Depuydt, Jesse de Does, Francesca Frontini, и Ilan Kernerman. 2019. „The OntoLex Lemon Lexicography Module“. https://www.w3.org/2019/09/lexicog/#introduction.

Bowers, Jack, и Laurent Romary. 2018. „Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec“. *Journal of the Dictionary Society of North America* 39 (2): 79–106. https://hal.inria.fr/hal-01968871/file/04_RWiP_Bowers-Romary-edited-Authors-copy.pdf.

Bowers, Jack Thompson, и Thierry Declerck. 2016. „TEI and lemon: a comparative study on the lexical encoding and interoperability“. У , 53–54. Vienna: Austrian Centre for Digital Humanities, Austrian Academy of Sciences. https://tei2016app.acdh.oeaw.ac.at/pages/show.html?document=BowersDeclerck.xml&directory=editions&stylesheet=editions.

Brekle, Jonas. 2012. „Flexible RDF data extraction from Wiktionary Leveraging the power of community build linguistic wikis“. Masterarbeit, Leipzig: Universität Leipzig.

Bugarski, Ranko. 1995. *Uvod u opštu lingvistiku*. Beograd: Čigoja.

Calzolari, Nicoletta, Monica Monachini, и Claudia Soria. 2013. „LMF – Historical Context and Perspectives“. У *LMF Lexical Markup Framework*, 1–18. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch1.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, и John Philip McCrae. 2020. „Modelling Frequency and Attestations for OntoLex-Lemon“. У *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, 1–9.

Christmann, Ruth. 2001. „Books into Bytes: Jacob and Wilhelm Grimm’s Deutsches Worterbuch on CD-ROM and on the Internet“. *Literary and Linguistic Computing* 16 (2): 121–33. https://doi.org/10.1093/llc/16.2.121.

Cimiano, P., P. Buitelaar, J. McCrae, и M. Sintek. 2011. „LexInfo: A Declarative Model for the Lexicon-Ontology Interface“. *Web Semantics: Science, Services and Agents on the World Wide Web* 9 (1). http://www.websemanticsjournal.org/index.php/ps/article/view/182.

Cohen, Jacob. 1960. „A Coefficient of Agreement for Nominal Scales“. *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Czaykowska-Higgins, Ewa, Martin Holmes, и Sarah Kell. 2014. „Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project“. *Language Documentation & Conservation* 8: 1–37.

De Schryver, Gilles-Maurice. 2003. „Lexicographers' Dreams in the Electronic-Dictionary Age". *International Journal of Lexicography* 16 (Јуни): 143–99. https://doi.org/10.1093/ijl/16.2.143.

———. 2011. „Why opting for a dedicated, professional, off-the-shelf dictionary writing system matters". У *ASIALEX Biennial International Conference*, 647–56.

De Schryver, Gilles-Maurice, и Guy De Pauw. 2007. „Dictionary Writing System (DWS)+ Corpus Query Package (CQP): The Case of" TshwaneLex"". *Lexikos* 17 (1): 226–46. https://tshwanedje.com/publications/dws+cqp.pdf.

Declerck, Thierry, Karlheinz Mörth, и Eveline Wand-Vogt. 2014. „A SKOS-Based Schema for TEI-Encoded Dictionaries". У *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 414–17. Reykjavik. https://www.researchgate.net/publication/265297624_A_SKOS-based_Schema_for_TEI-encoded_Dictionaries.

Đorđević, Bojana. 2017. „Izrada osnova formalne gramatike srpskog jezika upotrebom metagramatike". Doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet. https://fedorabg.bg.ac.rs/fedora/get/o:16929/bdef:Content/download.

Elleuch, Imen, Bilel Gargouri, и Abdelmajid Ben Hamadou. 2021. „Lexical Data Mining-based Approach for the Self-enrichment of LMF Standardized Dictionaries: Case of the Syntactico-semantic Knowledge". *Concurrency and Computation: Practice and Experience* 33 (17). https://doi.org/10.1002/cpe.6312.

Erjavec, Tomaž. 2010. „MULTEXT-East and TEI: an Investigation of a Schema for Language Engineering and Corpus Linguistics". У *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, уредио Dan Tufis и Corina Forascu. Bucuresti: Editura Academiei Romane.

„European Dictionary Portal: Criteria for Inclusion". 2017. http://www.dictionaryportal.eu/en/crit/.

Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, и Claudia Soria. 2007. „Lexical markup framework: ISO standard for semantic information in NLP lexicons". У *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*.

Francopoulo, Gil, и Monte George. 2013. „Model Description". У *LMF Lexical Markup Framework*, 19–40. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch2.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, и Claudia Soria. 2006. „Lexical Markup Framework (LMF)". У *International Conference on Language Resources and Evaluation-LREC 2006*.

Gašević, Dragan, Dragan Đurić, Vladan Devedžić, и Bran Selic. 2006. *Model Driven Architecture and Ontology Development*. Springer-Verlag.

Granger, S., и M. Paquot. 2012. *Electronic Lexicography*. OUP Oxford.

Gross, Maurice. 1987. „The Use of Finite Automata in the Lexical Representaion of Natural Language." У *Electronic Dictionaries and Automata in Computational Linguistics, LITP Spring School on Theoretical Computer Science, Saint-Pierre d'Oléron, France, May 25-29, 1987, Proceedings*, 34–50. https://doi.org/10.1007/3-540-51465-1_3.

Gruber, Thomas. 1995. „Toward principles for the design of ontologies used for knowledge sharing?“ *International journal of human-computer studies* 43 (5–6): 907–28. https://tomgruber.org/writing/onto-design.pdf.

———. 2009. „Ontology“. У *Encyclopedia of Database Systems*, уредио LING LIU и M. TAMER ÖZSU, 1963–65. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_1318.

Haas, Hugo, и Allen Brown. 2004. „Web Service“. У *Web Services Glossary*. W3C. https://www.w3.org/TR/ws-gloss/#webservice.

Haslam, Michael W. 1994. „The Homer Lexicon of Apollonius Sophista: II. Identity and Transmission“. *Classical Philology* 89 (2): 107–19. http://www.jstor.org/stable/270657.

Hayashi, Yoshihiko, Bora Savas, Monica Monachini, Claudia Soria, и Nicoletta Calzolari. 2012. „LMF-Aware Web Services for Accessing Semantic Lexicons“. *Language Resources and Evaluation* 46 (2): 253–64. https://doi.org/10.1007/s10579-012-9181-4.

„History – TEI: Text Encoding Initiative“. 2018. http://www.tei-c.org/about/history/.

Horridge, Matthew, Holger Knublauch, Alan Rector, Robert Stevens, и Chris Wroe. 2004. „A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0“. *University of Manchester*.

„Introduction to SKOS - SKOS Simple Knowledge Organization System“. 2018. https://www.w3.org/2004/02/skos/intro.

„ISO - International Organization for Standardization“. 2018. https://www.iso.org/home.html.

ISO/IEC JTC 1/SC 32. 2013. „ISO/IEC 11179-3:2013 Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes“. https://www.iso.org/standard/50340.html.

ISO/TC 37. 2008. „ISO 24613:2008 Language resource management - Lexical markup framework (LMF)“. https://www.iso.org/standard/37327.html.

ISO/TC 37/SC. 2008. „Language resource management — Lexical markup framework (LMF)“.

ISO/TC 37/SC 3 Management of terminology resources. 2009. „Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources“.

Jaćimović, Jelena. 2016. „Automatsko prepoznavanje i normalizacija vremenskih izraza u nestrukturiranim novinskim i medicinskim tekstovima na srpskom jeziku“. Doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet. https://fedorabg.bg.ac.rs/fedora/get/o:15094/bdef:Content/get.

Jacquet-Pfau, Christine. 2002. „Les dictionnaires du français sur cédérom“. *International Journal of Lexicography* 15 (Март): 89–104. https://doi.org/10.1093/ijl/15.1.89.

Jakubíček, Miloš, Iztok Kosem, Simon Krek, Sussi Olsen, Lene Offersgaard, и Bolette Sandford Pedersen. 2018. „ELEXIS-European lexicographic infrastructure“. У *CLARIN Annual Conference 2018*, 124.

Jovanović, Jovana. 2019. „Metafore istine u srpskom jeziku“. *Прилози проучавању језика*, изд. 50 (Децембар): 109–20. https://doi.org/10.19090/ppj.2019.50.109-120.

Kamholz, David, Jonathan Pool, и Susan M. Colowick. 2014. „PanLex: Building a Resource for Panlingual Lexical Translation.“ У *LREC*, 3145–50.

Khemakhem, Mohamed, Luca Foppiano, и Laurent Romary. 2017. „Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields“. У *Electronic Lexicography, ELex 2017*, 598–613:16. Brno: Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper37.pdf.

Kilgarriff, Adam. 2009. „Simple Maths for Keywords“. У *Proceedings of the Corpus Linguistics Conference*, 6. Liverpool.

———. 2012. „Pedro A. Fuertes-Olivera and Henning Bergenholtz (eds.). e-Lexicography: The Internet, Digital Initiatives and Lexicography“. *Kernerman Dictionary News*, 2012. http://kilgarriff.co.uk/Publications/kdn20_2012_pp26-29.pdf?format=raw.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, и Pavel Rychlý. 2008. „GDEX: Automatically finding good dictionary examples in a corpus“. У *Proceedings of the XIII EURALEX international congress*, 425–32. Universitat Pompeu Fabra Barcelona, Spain.

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, и Ljiljana Kolonja. 2021. „A Data Driven Approach for Raw Material Terminology“. *Applied Sciences* 11 (7): 2892. https://doi.org/10.3390/app11072892.

Klosa, Annette. 2013. „The lexicographical process (with special focus on online dictionaries)“. У *Dictionaries : an international encyclopedia of lexicography ; Suppl.*, 517–24. Berlin [u.a.]: De Gruyter Mouton.

Kolonja, Ljiljana. 2016. „Sistem poslovne inteligencije za upravljanje zaštitom na radu u rudarskoj industriji“. Beograd: Univerzitet u Beogradu, Rudarsko-geološki fakultet. http://nardus.mpn.gov.rs/bitstream/handle/123456789/7902/Disertacija.pdf?sequence=1&isAllowed=y.

Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović, и Aleksandar Cvjetić. 2016. „Development of terminological resources for expert knowledge: a case study in mining“. *Knowledge Management Research & Practice* 14 (4): 445–56. https://doi.org/10.1057/kmrp.2015.10.

Krek, Simon, Andrea Abel, и Carole Tiberius. 2014. „Survey – WG3 ENeL Dictionary Writing Systems & Corpus Query Systems“. У *Proceedings of the 3rd COST ENeL Working Group 3 („Innovative e-dictionaries“)*.

Krstev, Cvetana. 1997. „Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije“. Doktorska disertacija, Beograd: Matematički fakultet BU.

———. 2008. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, Ivan Obradović, Ranka Stanković, и Duško Vitas. 2013. „An Approach to Efficient Processing of Multi-Word Units“. У *Computational Linguistics - Applications, Studies in Computational Intelligence*, 109–29. 458.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, и Duško Vitas. 2014. „A system for named entity recognition based on local grammars“. *Journal of Logic and Computation* 24 (2): 473–89. https://doi.org/10.1093/logcom/exs079.

Krstev, Cvetana, Branislava Šandrih, Ranka Stanković, и Miljana Mladenović. 2018. „Using English Baits to Catch Serbian Multi-Word Terminology“. У *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1395.

Krstev, Cvetana, Ranka Stanković, Ivan Obradović, и Lazić, Biljana. 2015. „Terminology acquisition and description using lexical resources and local grammars“. У *Proceedings of the conference Terminology and Artificial Intelligence 2015*, 81–89. Granada, Spain. http://ceur-ws.org/Vol-1495/paper_13.pdf.

Krstev, Cvetana, Ranka Stanković, и Duško Vitas. 2018. „Knowledge and Rule-Based Diacritic Restoration in Serbian“. У *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, 41–51. Sofia: The Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences.

Krstev, Cvetana, Ranka Stanković, Duško Vitas, и Svetla Koeva. 2009. „E-Connecting Balkan Languages“. У *Proceedings of the Workshop on Multilingual resources, technologies and evaluation for Central and Eastern European Languages*, 23–29. Borovets, Bulgaria.

Krstev, Cvetana, Ranka Stanković, Duško Vitas, и Ivan Obradović. 2006. „WS4LR: A Workstation for Lexical Resources“. У *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, 1692–97. Genoa.

Krstev, Cvetana, и Duško Vitas. 2009. „An Effective Methode for Developing a Comprehensive Morphological E-Dicitonary of Compounds“. У *Arena Romanistica*, 204–12. Bergen: University of Bergen, Department of Foreign Languages.

———. 2011. „An Aligned English-Serbian Corpus“. У *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Belgrade, 4-6 December 2009*, Volume I:495–508. http://poincare.matf.bg.ac.rs/~cvetana/biblio/AlignedCorpus-full-final.pdf.

Laporte, Éric, Elsa Tolone, и Constant Matthieu. 2013. „Conversion of Lexicon-Grammar Tables to LMF: Application to French“. У *LMF Lexical Markup Framework*, 157–74. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch1.

Lazić, Biljana, и Mihailo Škorić. 2019. „From DELA Based Dictionary to Leximirka Lexical Database“. *Infotheca* 19 (2): 81–98. https://doi.org/10.18485/infotheca.2019.19.2.4.

Lazić, Biljana, и Ranka Stanković. 2015. *MineCorp, Serbian corpus from mining domain - Written Corpus*.

Ledinek, Nina, Kozma Ahačič, и Andrej Perdih. 2015. *Fran : slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. Vodnik*. Zbirka Fran. Ljubljana: Založba ZRC. http://hjp.znanje.hr/index.php?show=izvedeni.

„Lexicon Model for Ontologies“. 2016. Community Report. Ontology-Lexicon Community Group. https://www.w3.org/2016/05/ontolex/#core.

„LIR - Linguistic Information Repository“. 2018. 2018. http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/63-lir/.

Madsen, Martin Schou. 2017. „Planning Against Change: Serbian and Croatian reactions to contact-induced linguistic inovation“. University of Copenhagen.

Maurel, Denis. 2008. „Prolexbase: a Multilingual Relational Lexical Database of Proper Names“. У *Sixth language resources and evaluation conference LREC*. Marrakech, Morocco.

McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez Pérez, Jorge Gracia, и остали. 2012. „The Lemon Cookbook“. http://lemon-model.net/lemon-cookbook.pdf.

McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaa, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, и остали. 2012. „Interchanging lexical resources on the Semantic Web“. *Language Resources and Evaluation*, изд. 4: 701.

http://search.ebscohost.com.proxy.kobson.nb.rs:2048/login.aspx?direct=true&db=edsjsr&AN=edsjsr.23325377&site=eds-live.

Měchura, Michal Boleslav. 2017. „Introducing Lexonomy: an open-source dictionary writing and publishing system“. У *Proceedings of the eLex 2017 conference, 19-21 September 2017*, 662–79. Leiden: Lexical Computing.

———. 2018. „Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement“. У *The XVIII EURALEX International Congress*, 223/232.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, и Katherine J. Miller. 1990. „Introduction to WordNet: An on-line lexical database“. *International journal of lexicography* 3 (4): 235–44.

Mitić, Nenad. 2021. „Integritet u relacionim bazama podataka“. Представљено на Univerzitet u Beogradu, Matematički fakultet, Beograd. http://poincare.matf.bg.ac.rs/~nenad/rbp/6.Integritet.pdf.

Mladenović, Miljana, Jelena Mitrović, и Cvetana Krstev. 2014. „Developing and Maintaining a WordNet: Procedures and Tools“. У *The Proceedings of Seventh Global WordNet Conference 2014*, 55–62. https://aclanthology.org/W14-0108.pdf.

Mörth, Karlheinz, Laurent Romary, Gerhard Budin, и Daniel Schopper. 2014. „Modeling Frequency Data: Methodological Considerations on the Relationship between Dictionaries and Corpora“. *Journal of the Text Encoding Initiative*, изд. Issue 8 (Децембар). https://doi.org/10.4000/jtei.1356.

Moulin, Claudine, и Julianne Nyhan. 2014. „The Dynamics of Digital Publications“. У *New Publication Cultures in the Humanities*, 47–62. Exploring the Paradigm Shift. Amsterdam University Press. http://www.jstor.org/stable/j.ctt12877w9.6.

Niles, Ian, и Adam Pease. 2001. „Towards a Standard Upper Ontology“. У *Proceedings of the International Conference on Formal Ontology in Information Systems - FOIS '01*, 2001:2–9. Ogunquit, Maine, USA: ACM Press. https://doi.org/10.1145/505168.505170.

Odijk, Jan. 2013. „DUELME: Dutch Electronic Lexicon of Multiword Expressions“. У *LMF Lexical Markup Framework*, 133–43. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch1.

Pala, Karel, и Aleš Horák. 2006. „From WEB Pages to Dictionary: a Languge-independent Dictionary Writing System“. У *Proceedings of the 12th EURALEX International Congress*, 10:199–204. Turin, Italy.

Paumier, Sébastien. 2016. *Unitex User Manual*. Paris: Université Paris-Est Marne-la-Vallée.

Popović, Petar, Mihailo Škorić, и Biljana Rujević. 2020. „The Use of the Omeka Semantic Platform for the Development of the University of Belgrade, Faculty of Mining and Geology Digital Repository“. *Infotheca* 20 (1–2): 136–48. https://doi.org/10.18485/infotheca.2020.20.1_2.9.

Radovanovic, Aleksandra. 2017. „Future Tense in Serbian: A Fuzzy Linguistic Category“. *Nasleđe, Kragujevac* 14 (38): 125–39. https://doi.org/10.5937/naslKg1738125R.

Rajić, Jelena. 2015. „Análisis contrastivo de la expresión de la evidencialidad en serbio y español“. *Verba Hispanica* 23 (1): 127–48. https://doi.org/10.4312/vh.23.1.127-148.

„RDF Model and Syntax“. 2018. https://www.w3.org/TR/WD-rdf-syntax-971002/.

Salgado, Ana, Rute Costa, и Toma Tasovac. 2019. „TEI Lex-0: A Good Fit for the Encoding of the Portuguese Academy Dictionary?“, Септембар. https://doi.org/10.5281/ZENODO.3464931.

Sánchez-Rada, J. Fernando, Carlos A. Iglesias, и Ronald Gil. 2015. „A linked data model for multimodal sentiment and emotion analysis“. У *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, 11–19.

Schierholz, Stefan J. 2015. „Methods in Lexicography and Dictionary Research“. *Lexikos* 25: 323–52.

Sen, Marcus, и Tim Duffy. 2005. „GeoSciML: Development of a generic GeoScience Markup Language“. *Application of XML in the Geosciences* 31 (9): 1095–1103. https://doi.org/10.1016/j.cageo.2004.12.003.

Sérasset, Gilles. 2015. „DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF.“ *Semantic Web (1570-0844)* 6 (4): 355. http://search.ebscohost.com.proxy.kobson.nb.rs:2048/login.aspx?direct=true&db=edb&AN=110621832&site=eds-live.

Stanković, Ranka. 2009. „Modeli ekspanzije upita nad tekstuelnim resursima“. Doktorska disertacija, Beograd: Univerzitet u Beogradu, Matematički fakultet.

Stanković, Ranka, Cvetana Krstev, Lazić, Biljana, и Mihailo Škorić. 2018. „Electronic Dictionaries - from File System to Lemon Based Lexical Database“. У *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 48–56. Miyazaki, Japan: European Language Resources Association (ELRA). http://lrec-conf.org/workshops/lrec2018/W23/summaries/3_W23.html.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, и Olivera Kitanović. 2015. „Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian“. У *Semantic Keyword-based Search on Structured Data Sources*, уредио Jorge Cardoso, Francesco Guerra, Geert-Jan Houben, Alexandre Miguel Pinto, и Yannis Velegrakis, 9398:167–81. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27932-9_15.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić, и Aleksandra Trtovac. 2015. „Rule-Based Automatic Multi-Word Term Extraction and Lemmatization“. У *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23--28 May 2016*, 507–14.

Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, и Cvetana Krstev. 2018. „Resource-based WordNet Augmentation and Enrichment“. У *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, 104–14. Sofia, Bulgaria: The Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences. http://dcl.bas.bg/clib/wp-content/uploads/2018/06/CLIB_2018_Proceedings_v1_final.pdf.

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, и Duško Vitas. 2011. „Production of morphological dictionaries of multi-word units using a multipurpose tool“. У *Proceedings of the Computational Linguistics-Applications Conference*, уредио K. Jassem, P. W. Fuglewicz, M. Piasecki, и A. Przepiórkowski, 77–84. Polish Information Processing Society.

Stankovic, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, и Mihailo Skoric. 2020. „Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian“. У *Proceedings of the 12th Language Resources and*

Evaluation Conference, 3954–62. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.487.

Stanković, Ranka, Branislava Šandrih, Rada Stijović, Cvetana Krstev, Duško Vitas, и Aleksandra Marković. 2019. „SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian“. *Electronic lexicography in the 21st century: Smart lexicography*, 248–69.

Stanković, Ranka, и Miloš Utvić. 2019. „Vebran Web Services for Corpus Query Expansion“. *INFOtheca* 19 (2).

„Statistics used in the Sketch Engine“. 2015. Lexical Computing Ltd. https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf.

Šipka, Danko. 2006. *Osnovi leksikologije i srodnih disciplina*. 2., Izmijenjeno i dop. izd. Novi Sad: Matica Srpska.

Štrkalj Despot, Kristina, и Christine Möhrs. 2015. „Pogled u e-leksikografiju“. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 41: 329–53.

Tarp, Sven. 2012. „Do we need a (new) theory of lexicography?“ *Lexikos* 22: 321–32.

Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović, и Božo Kolonja. 2019. „Managing mining project documentation using human language technology“. *The Electronic Library*, 993–1009. https://doi.org/10.1108/EL-11-2017-0239.

Tshwanedje. 2020. „TshwaneDJe Software and Consulting“. Децембар 8. https://tshwanedje.com/tshwanelex/.

„TshwaneDJe Software: TLex Lexicography, Terminology and Corpus Software“. 2018. 2018. https://tshwanedje.com/tshwanelex/overview.html.

UKP. 2018. „UBY“. UKP – Technische Universität Darmstadt. 2018. https://www.informatik.tu-darmstadt.de/ukp/research_6/data/lexical_resources/uby/index.en.jsp.

Urkom, Aleksander. 2010. „Korpusna lingvistika u okruženju Srbije i zadaci razvoja srpskog korpusa“. У *Susret Kultura. Novi Sad, Srbija*, 567–75. http://szlavintezet.elte.hu/staff/urkom/anyagok/2010_Korpusna%20lingvistika%20u%20okruzenju%20Srbije%20i%20zadaci%20razvoja%20srpskog%20korpusa.pdf.

Utvić, Miloš. 2013. „Izgradnja referentnog korpusa savremenog srpskog jezika“. Doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet. https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download.

Villegas, Marta, и Núria Bel. 2015. „PAROLE/SIMPLE ‚lemon‘ ontology and lexicons.“ *Semantic Web (1570-0844)* 6 (4): 363–69. http://www.semantic-web-journal.net/system/files/swj496.pdf.

Vitas, Duško. 1981. „Generisanje imeničkih oblika u srpskohrvatskom jeziku“. *Informatica* 81 (3): 49–55.

———. 1993. „Matematički model morfologije srpskohrvatskog jezika (imenska fleksija)“. Doktorska disertacija, Beograd: Matematički fakultet BU.

———. 2006. *Prevodioci i interpretatori : (uvod u teoriju i metode kompilacije programskih jezika)*. Beograd: Matematički fakultet.

Vitas, Duško, Svetla Koeva, Cvetana Krstev, и Ivan Obradovic. 2008. „Tour du monde through the dictionaries“. У *Actes du 27eme Colloque International sur le Lexique et la Gammaire, L'Aquila, 10-13 septembre 2008*, 249–56.

Vitas, Duško, и Cvetana Krstev. 2006. „Literature and Aligned Texts“. У *Readings in Multilinguality*, 148–55. Sofia: Institute for Parallel Processing, Bulgarian Academy of Sciences.

———. 2012. „Processing of Corpora of Serbian Using Electronic Dictionaries“. *Prace Filologiczne* Vol. LXIII: 279–92. http://poincare.matf.bg.ac.rs/~cvetana/biblio/22_Vitas_Krstev.pdf.

Vitas, Duško, Cvetana Krstev, и Éric Laporte. 2006. „Preparation and exploitation of Bilingual Texts“. *Lux Coreana*, изд. No. 1: 110–32.

Vujičić Stanković, Staša. 2016. „Ekstrakcija informacija vođena ontologijama (model za srpski jezik)“. Doktorska disertacija, Beograd: Univerzitet u Beogradu, Matematički fakultet. https://fedorabg.bg.ac.rs/fedora/get/o:15083/bdef:Content/get.

Xiao, Richard. 2010. „Corpus creation“. У *Handbook of Natural Language Processing (2n Revised edition)*, 147–65. Machine Learning & Pattern Recognition Series. CRC Press, Taylor and Francis Group.

Zgusta, Ladislav. 1991. *Priručnik leksikografije*. Sarajevo: Zavod za udžbenike i nastavna sredstva.

Прилог 1. Поређење категорија података из Морфолошких речника српског језика у односу на скраћенице из Речника САНУ и концепте из онтологије SUMO

Р. бр.	МРС	МРС	однос са РСАНУ	РСАНУ	однос са SUMO	SUMO
1	1		=	ном.		
2	2		=	ген.		
3	3		=	дат.		
4	4		=	ак.		
5	5		=	вок.		
6	6		=	инстр.		
7	7		=	лок.		
8	A [gramCat]		=	аор.		
9	A [POS]		=	прид.		
10	Acr				=	acronym
11	ADV		=	прил.		
12	Art		=	ум.		
13	Aug		=	аугм.		
14	b		=	комп.		
15	Bot				=	Plant
16	c		=	суп.		
17	Cave				=	Cave
18	CC2=BA		=	БиХ		
19	CC2=BA		[	Херц.		
20	CC2=HR		[	Далм.		
21	CC2=HR		[	Слав.		
22	CC2=HR		=	Хрв.		
23	CC2=ME		=	ЦГ		
24	CC2=MK		[	Скоп. ЦГ		
25	CC2=RS		[	Војв.		
26	CC2=RS		[	КМ		
27	CC2=RS		=	Срб.		
28	CC2=RS		[	Шум.		
29	Cell				=	Cell
30	Char				=	Character
31	Coll		=	зб. им.		
32	HumColl		]	зб. им.		
33	Compound				=	CompoundSubstance
34	Conc				=	Concrete
35	CONJ		=	везн.		
36	Const		=	непром.		
37	Cur				=	Currency
38	d		=	одр.		
39	Dem		=	дем.		
40	DeoGr				=	CityDistrict
41	Desert				=	Desert

42	Dir				=	direction
43	DOM=Agro		=	агр.		
44	DOM=Anat		=	анат.		
45	DOM=Archeo		=	археол.		
46	DOM=Archi		=	архит.		
47	DOM=Art		[	вај.		
48	DOM=Art		[	слик.		
49	DOM=Art		[	фот.		
50	DOM=Astr		=	астр.		
51	DOM=BIinf		[	информ.		
52	DOM=Bio		=	биол.		
53	DOM=Bio		[	мик.		
54	DOM=Bio		[	ф.		
55	DOM=Bot		=	бот.		
56	DOM=Bot		[	фитопат.		
57	DOM=Chem		=	хем.		
58	DOM=Construct		=	грађ.		
59	DOM=Culinary		=	кув.		
60	DOM=Forest		=	шум.		
61	DOM=Geo		=	геогр.		
62	DOM=Geol		=	геол.		
63	DOM=Geol		[	мин.		
64	DOM=Geol		[	пал.		
65	DOM=Gram		=	грам.		
66	DOM=Gram		[	прав.		
67	DOM=Hist		=	ист.		
68	DOM=Ind		=	инд.		
69	DOM=Law		=	правн.		
70	DOM=Ling		=	лингв.		
71	DOM=Ling		[	сем.		
72	DOM=Ling		[	синт.		
73	DOM=Ling		[	фон.		
74	DOM=Lit		=	књиж.		
75	DOM=Lit		[	песн.		
76	DOM=Lit		[	поет.		
77	DOM=Lit		[	сткњ.		
78	DOM=Mach		=	маш.		
79	DOM=Math		=	мат.		
80	DOM=Med		=	мед.		
81	DOM=Med		[	нар. мед.		
82	DOM=Meteo		=	мет.		
83	DOM=Mining		=	руд.		
84	DOM=Mov		=	филм.		
85	DOM=Mus		=	муз.		
86	DOM=Myth		=	мит.		
87	DOM=Pharm		=	фарм.		
88	DOM=Phylos		=	фил.		

89	DOM=Phys		=	физ.		
90	DOM=Physiol		=	физиол.		
91	DOM=Pol		=	пол.		
92	DOM=Print		=	штамп.		
93	DOM=Relig		[	исл.		
94	DOM=Relig	Eth	[	јевр.		
95	DOM=Relig		[	кат.		
96	DOM=Relig		[	правосл.		
97	DOM=Relig		=	рлг.		
98	DOM=Sport		=	спорт.		
99	DOM=Stage		=	поз.		
100	DOM=Tech		=	техн.		
101	DOM=Zool		=	зоол.		
102	DOM=Zool		[	пор.		
103	Dr				=	StateOrProvince
104	DrFed				=	FederalRepublic
105	Ek		=	ек.		
106	Erg				=	Product
107	ET=AR		=	ар.		
108	ET=CSL			цсл.		
109	ET=DE	DE	=	нем.		
110	ET=EN	EN		енгл.		
111	ET=FR			фр.		
112	ET=GR			грч.		
113	ET=HU			мађ. и мац.		
114	ET=IT			тал.		
115	ET=LA			лат.		
116	ET=PERS			перс.		
117	ET=RU	RU	=	рус.		
118	ET=TR			тур.		
119	Eth					
120	f		=	ж		
121	First				=	givenName
122	Food				=	Food
123	FoS				=	FieldOfStudy
124	Furniture				=	Furniture
125	G		=	р. пр.		
126	Golf				=	Gulf
127	Gr				=	City
128	Hip		=	хип.		
129	Hum				=	Human
130	Hyd				=	WaterArea
131	i		=	енкл.		
132	I		=	импф.		
133	Ijk		=	ијек.		
134	Ik		=	ик.		
135	Imperf		=	несвр.		

136	Incorr		=	непр.		
137	Inh				=	inhabits
138	Ins				=	Island
139	Instrum				=	Device
140	INT		=	узв.		
141	It		=	непрел.		
142	Kont				=	Continent
143	Lake				=	Lake
144	Last				=	familyName
145	Law				=	Law
146	Lng				=	Language
147	m		=	м		
148	Mat				=	material
149	Meal				=	Meal
150	Mes				=	UnitOfMeasure
151	Micro				=	Microorganism
152	Mount				=	Mountain
153	Music				=	Music
154	N		=	им.		
155	n		=	с		
156	Name				=	Name
157	Neg		=	одрич.		
158	Nick				[	Name
159	NUM		=	бр.	=	Number
160	Number				=	Number
161	Onom		=	оном.		
162	Org				=	Organization
163	P		=	през.		
164	p		=	мн.		
165	PAR		=	речца		
166	Pej		=	пej.		
167	Penins				=	Peninsula
168	Perf		=	свр.		
169	Plain				=	Plain
170	POG		=	погрд.		
171	PREF		=	преф.		
172	PREP		=	предл.		
173	PRO		=	зам.		
174	Proc				=	Procedure
175	Process				=	Process
176	Prof				=	OccupationalRole
177	PsychFeature				=	PsychologicalAttribute
178	Ref		=	повр.		
179	Reg				=	Region
180	River				=	River
181	S		=	пр. сад.		
182	s		=	јд.		

183	Salute				=	Greeting
184	Sea				=	Sea
185	Sport				=	Sport
186	Strait				=	Strait
187	T		=	трп. пр.		
188	Text				=	Text
189	Time				=	time
190	Tr		=	прел.		
191	V		=	гл.		
192	Vehicle				=	Vehicle
193	Volcano				=	Volcano
194	W		=	инф.		
195	w		=	дв.		
196	X		=	пр. пр.		
197	Y		=	имп.		
198	Zgrada				=	Building
199	Zool				=	Animal

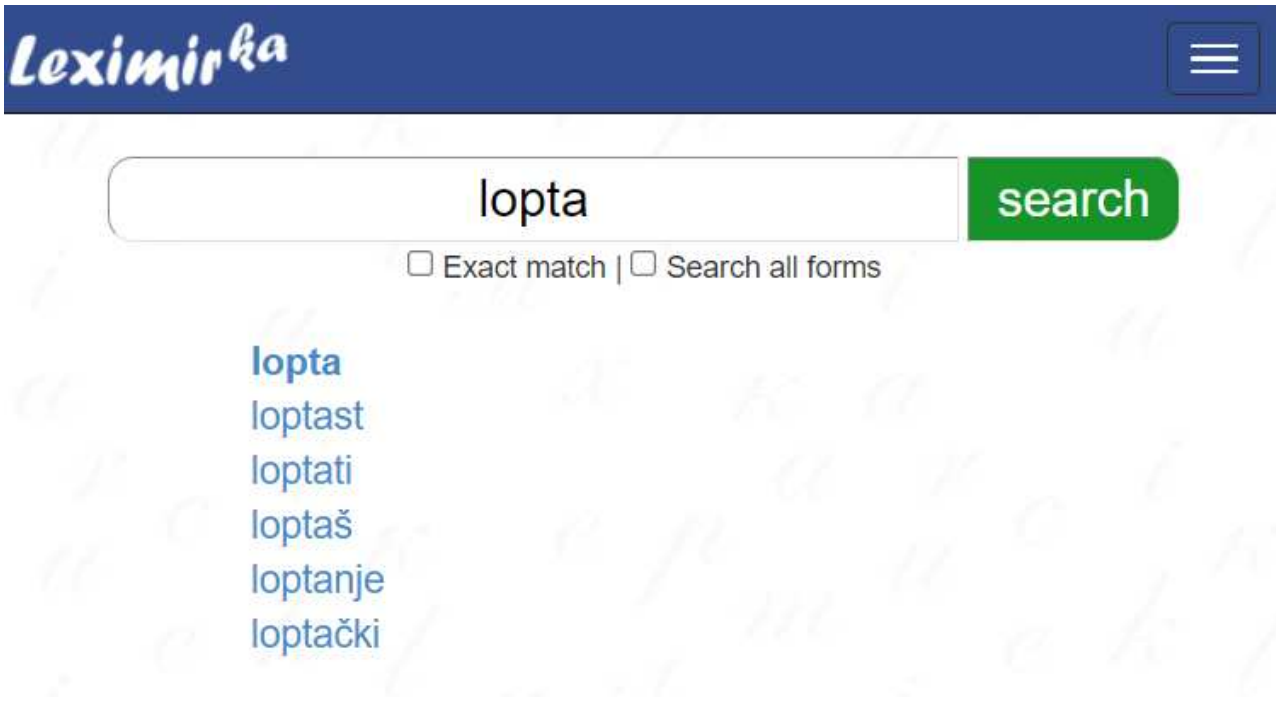
Прилог 2. Упутство за коришћење апликације *Лексимирка*

Апликација *Лексимирка* је доступна за претраживање на вебу[125].

Апликација *Лексимирка* може се користити двојако: може бити намењена широком кругу корисника регистрацијом налога или као сучеље за управљање *MPC* и његов развој уз додељивање одговарајућих привилегија корисничком налогу. Приступ апликацији је истоветан за обе врсте корисника - путем опције „Log in“, која се налази у горњем десном углу екрана, уз коришћење корисничког имена (*User name*) и лозинке (*Password*).

а) Корисничко сучеље намењено широком кругу корисника

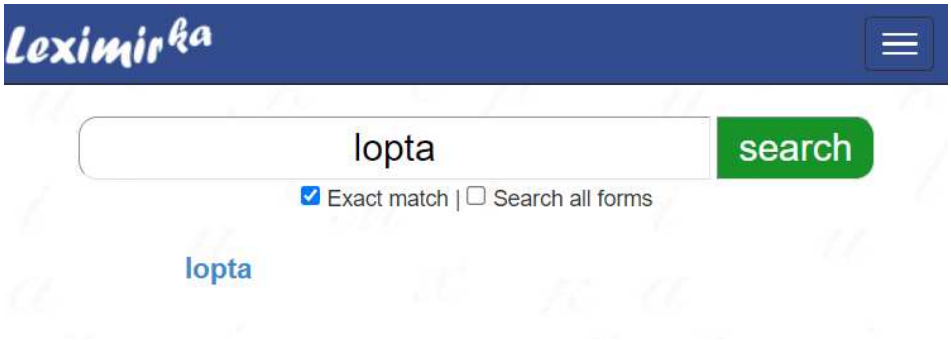
Употребом основног корисничког налога корисник је у могућности претражује лексичке записе из подскупа *MPC.* Овим корисницима су за претрагу доступне речи које спадају у 100.000 најфреквентнијих у корпусу *СрпКор.*



Слика П2.1 Резултат претраге за реч „лопта“ у јавном приказу

Упити за претрагу се постављају куцањем речи латиничним или ћириличним писмом или словима из кода Аурора док ће резултат претраге бити на латиници. За интерне записе речника се користи код Аурора. Након уноса ниске слова за претрагу, претрага се покреће притиском на тастер „*Enter*“ на тастатури или притиском на дугме „*search*“. Као резултат се добија листа речи које одговарају задатом упиту (највише 20 речи). Уколико није означена опција за потпуно поклапање (опција „*Exact match*“), листа речи ће садржати речи које садрже ниску слова која је постављена упитом (слика П2.1). На пример, за упит постављен ниском „lopta“ листа речи које одговарају упиту садржаће речи: „kloptanje“, „loptanje“, „lopta“, „polulopta“, „loptaš“, „kloptati“, „loptati“, „loptački“, „loptast“, „tenis-lopta“, итд.). Уколико се при истом упиту означи и опција за тачан облик, добија се само један погодак, реч „lopta“ (слика П2.2).

[125] URL адреса: http://leximirka.jerteh.rs/



Слика П2.2 Резултат претраге за реч „лопта“ у јавном приказу – прецизан упит

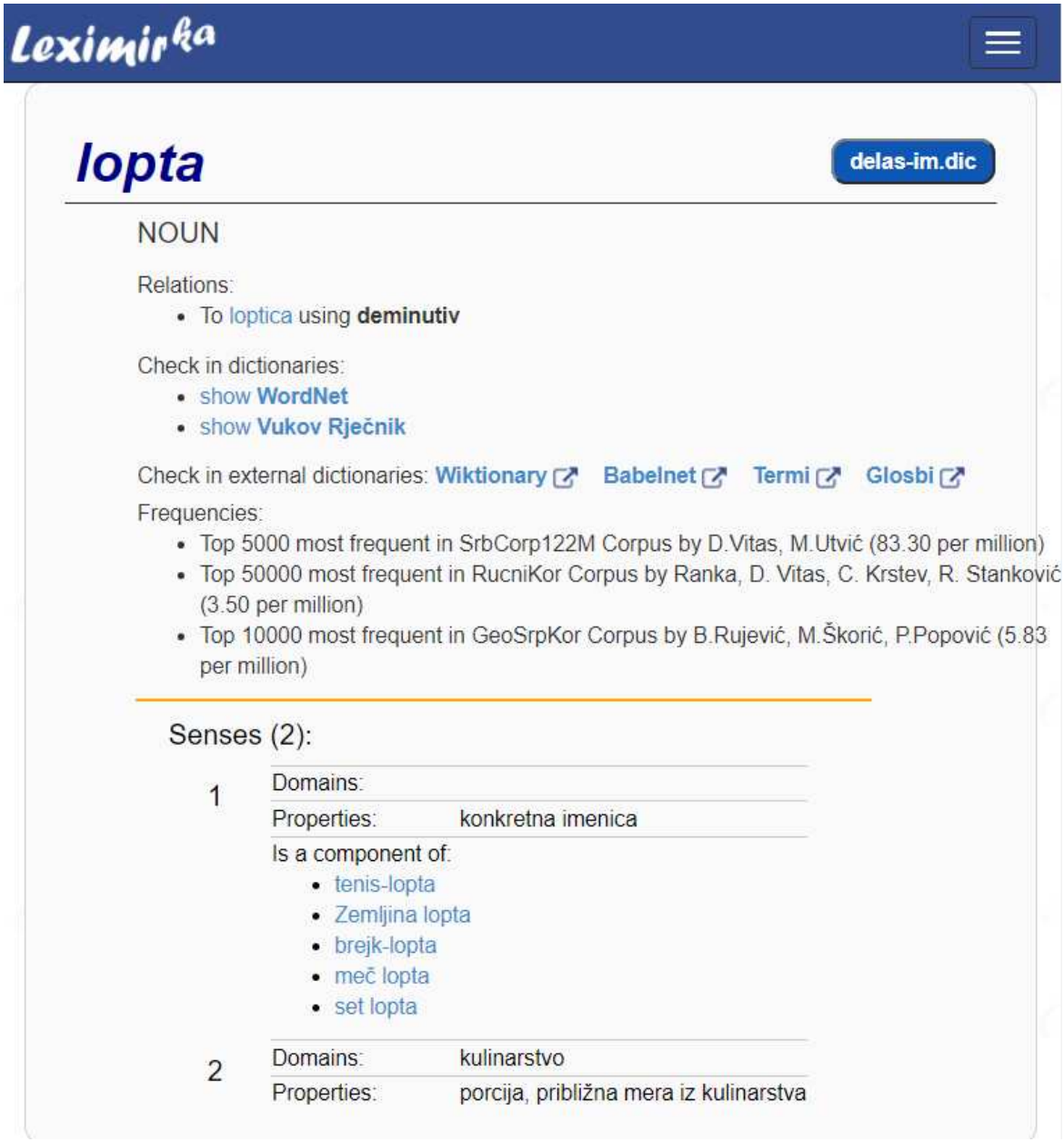
Одабиром одговарајућег поготка из понуђене листе добија се приказ лексичког записа у облику предвиђеном за јавни приказ што је приказано на слици П2.3. У горњем десном углу се налази лема што је у овом случају „lopta“. Приказ је потом подељен на четири дела: релације – „Relations“, везе са другим речницима, фреквенције –„Frequencies“ и значења – „Senses“.

У првом делу лексичког записа приказана је деминутивна релација са записом „loptica“ који је приказан у облику хипервезе чијим се одабиром долази до прегледа одговарајућег лексичког записа.

Потом следе везе са лексичким ресурсима који су на вебу (*Wiktionary*, *BabelNet*, *Termi*, *Glosbe*) и ограниченим бројем лексичких ресурса из повезане базе података (*WordNet*, *Vukov rječnik*). Веза се лексичким ресурсима на вебу се остварује отварањем новог прозора у коме је покренута претрага одабраног ресурса лемом лексичког записа. Одабиром назива речника из повезане базе података се у оквиру приказа лексичког записа у апликацији *Лексимирка* отвара картица са приказом речничког чланка пореклом из изабраног речника.

Затим следи информација о фреквентности речи „lopta“ у доступним корпусима. На пример, у *Корпусу савременог српског језика*, у овом тренутку је то верзија *СрпКор122М*, аутора Душка Витаса и Милоша Утвића – реч спада у 5.000 најфреквентнијих речи.

У трећем делу се налази приказ два значења исказана природним језиком. Првим значењем даје се информација о томе да је „lopta“ конкретна именица и за ово значење су везани полилексемски изрази у којима је „lopta“ компонента. Ови полилексемски изрази су наведени у виду упутнице ка одговарајућим записима: „tenis-lopta“, „Zemljina lopta“, „brejk-lopta“, „меč lopta“ и „set lopta“. Друго значење овог лексичког записа везано је за домен кулинарства у коме означава порцију и приближну меру из кулинарства.

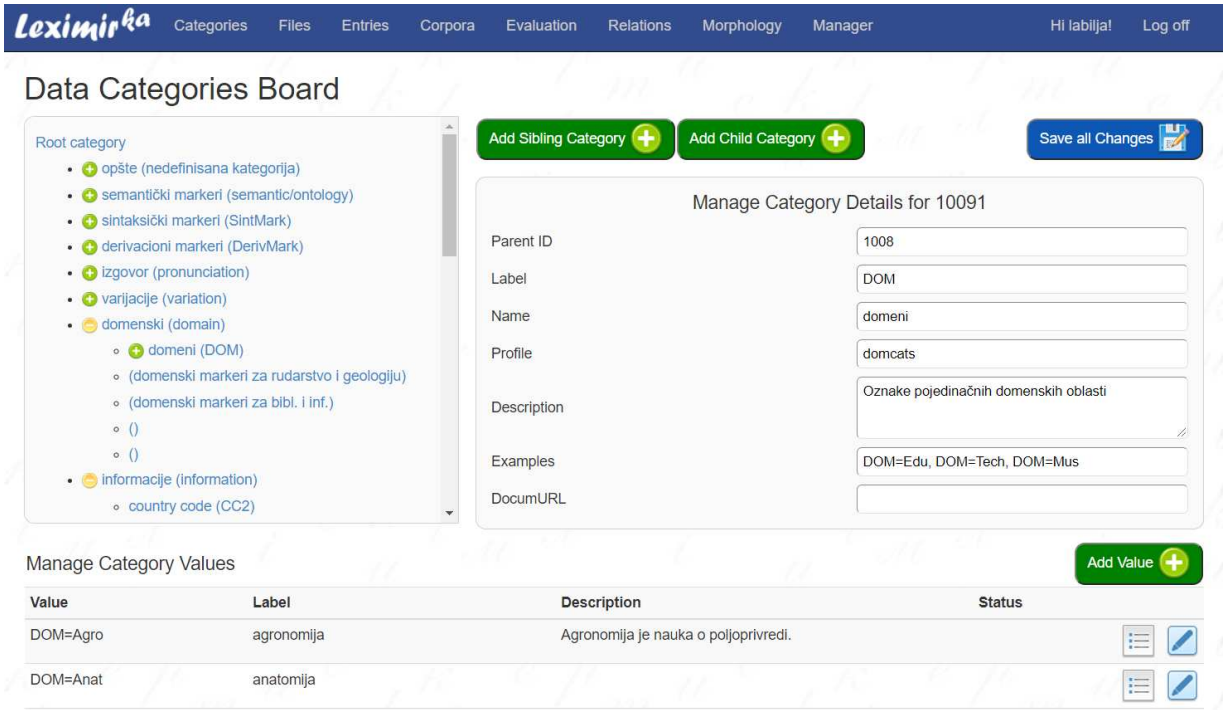


Слика П2.3 Јавни приказ лексичког записа „lopta“

Б) Корисничко сучеље за управљање и развој

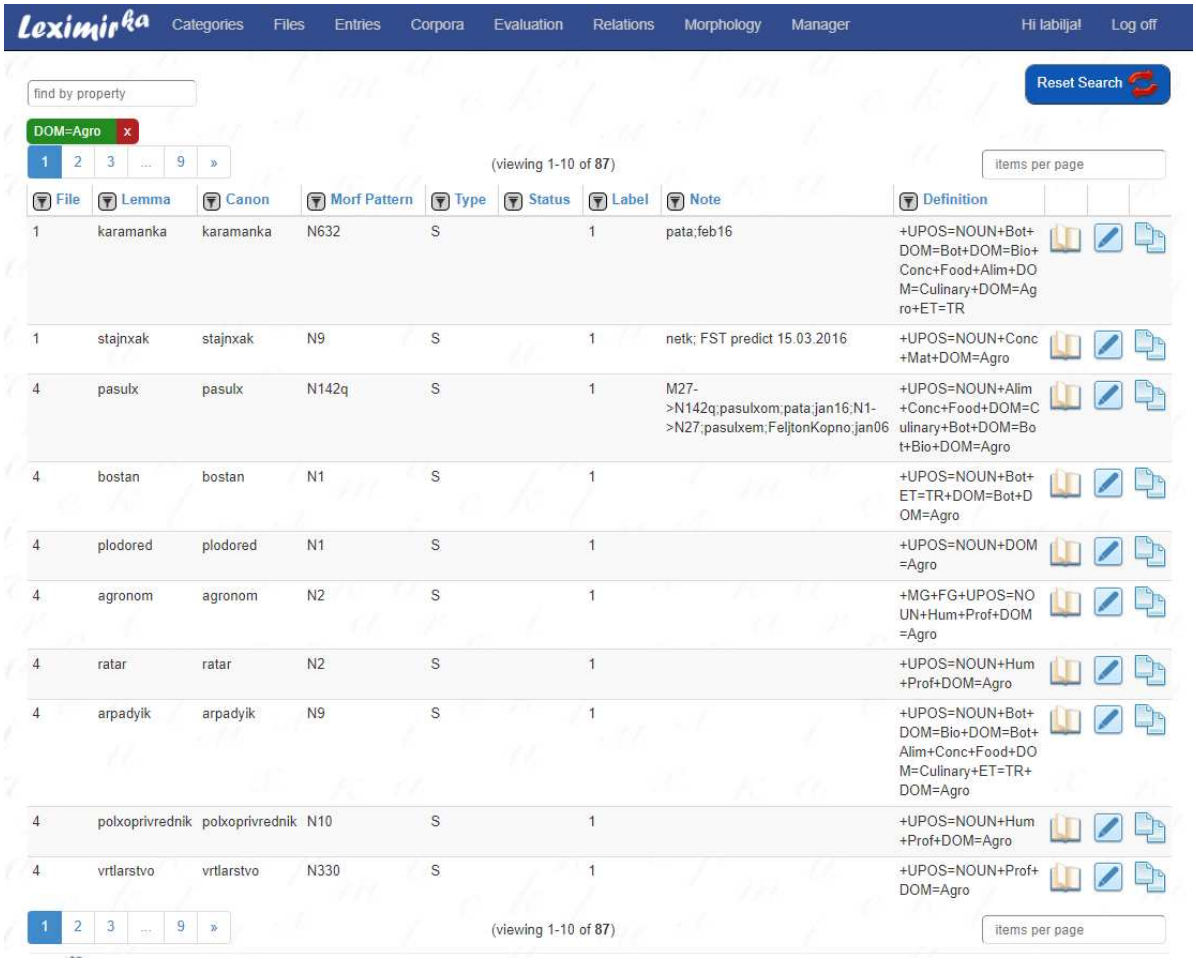
Како би се приступило корисничком сучењу за управљање и развој неопходно је да кориснику буду додељене одговарајуће привилегије од стране администратора. Корисник уз ове привилегије може да види све информације из „јавног приказа“ али и флективне облике леме, остале речи које се мењају према истој флективној класи, као и садржај речника коме припада дата реч. Поред овога регистровани корисник има могућност уређивања и прегледања различитих сегмената апликације Лексимирка: лексичких категорија (опција „Categories“), датотека речника (опција „Files“), лексичких записа (опција „Entries“), коришћених корпуса (опција Corpora), евалуације записа (опција „Evaluation“), лексичких релација (опција „Relations“), и морфолошких образаца (опција „Morphology“). Сегмент за управљање корисничким налозима (опција Manager) је намењен администраторима који додељују корисничке привилегије.

Прелазом курсора преко опције „Categories“ отвара се и опција „Tabelar View“ која омогућава табеларни приказ свих категорија које се користе у Морфолошким речницима за српски језик. Прва и предефинисана опција за приказ која се добија одабиром опције „Categories“ јесте контролна табла за управљање категоријама података - „Data Categories Board“ (Слика П2.4).



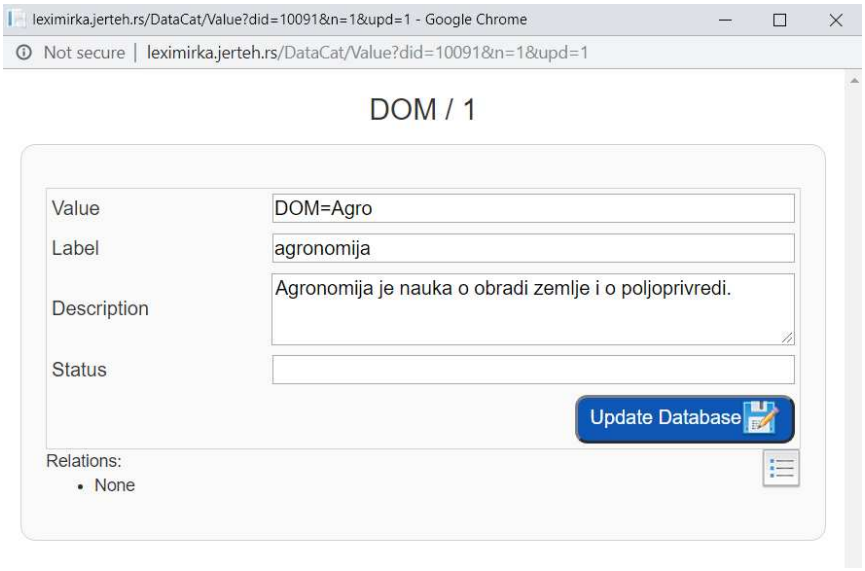
Слика П2.4 Контролна табла за управљање категоријама података

У левом делу приказа налази се дрвце свих категорија података које одражава њихов хијерархијски однос. Како се може видети на слици, основне категорије су: опште, семантички маркери, синтаксички маркери, деривациони маркери, изговор, варијације, домени, информације и граматичке категорије. У приказу на слици П2.4 је излистана категорија домена те можемо видети да су њене поткатегорије домени (без даљег разврставања), доменски маркери за рударство и доменски маркери за библиотекарство и информатику. Одабрана је категорија *домени* чији опис можемо видети са десне стране панела на слици П2.4. Опис се састоји од различитих детаља о категорији попут идентификационог броја (10091), идентификационог броја категорије која је родитељ текуће категорије (1008) (системски одређене вредности), обележја (DOM), имена (domeni), назива профила (domcats), кратког описа (Oznake pojedinačnih domenskih oblasti) и примера (DOM=Edu, DOM=Tech, DOM=Mus). У самом панелу је могуће вршити измене вредности у активним пољима и сачувати их коришћењем дугмета плаве боје у десном углу „Save All Changes“. Такође је могуће додавање нових категорија података у односу на текућу категорију података коришћењем дугмића зелене боје. Додавање категорије података истог нивоа у хијерархији се постиже коришћењем дугмета „Add Sibling Category“, док се подређена категорија додаје дугметом „Add Child Category“. У оба случаја се отвара непопуњен десни део панела са истим пољима за опис. У доњем делу панела се налазе излистане ознаке конкретних домена. На слици видимо доменске маркере за агрономију (DOM=Agro) и анатомију (DOM=Anat). Са десне стране у редовима вредности доменских маркера се налазе два дугмета: за преглед лексичких записа „View Entries“ и уређивање вредности “Edit Value”. Прво дугме води до нове стране на којој су излистани сви лексички записи означени текућом вредношћу категорије података. На слици П2.5 дат је приказ лексичких записа означених са „DOM=Agro“.



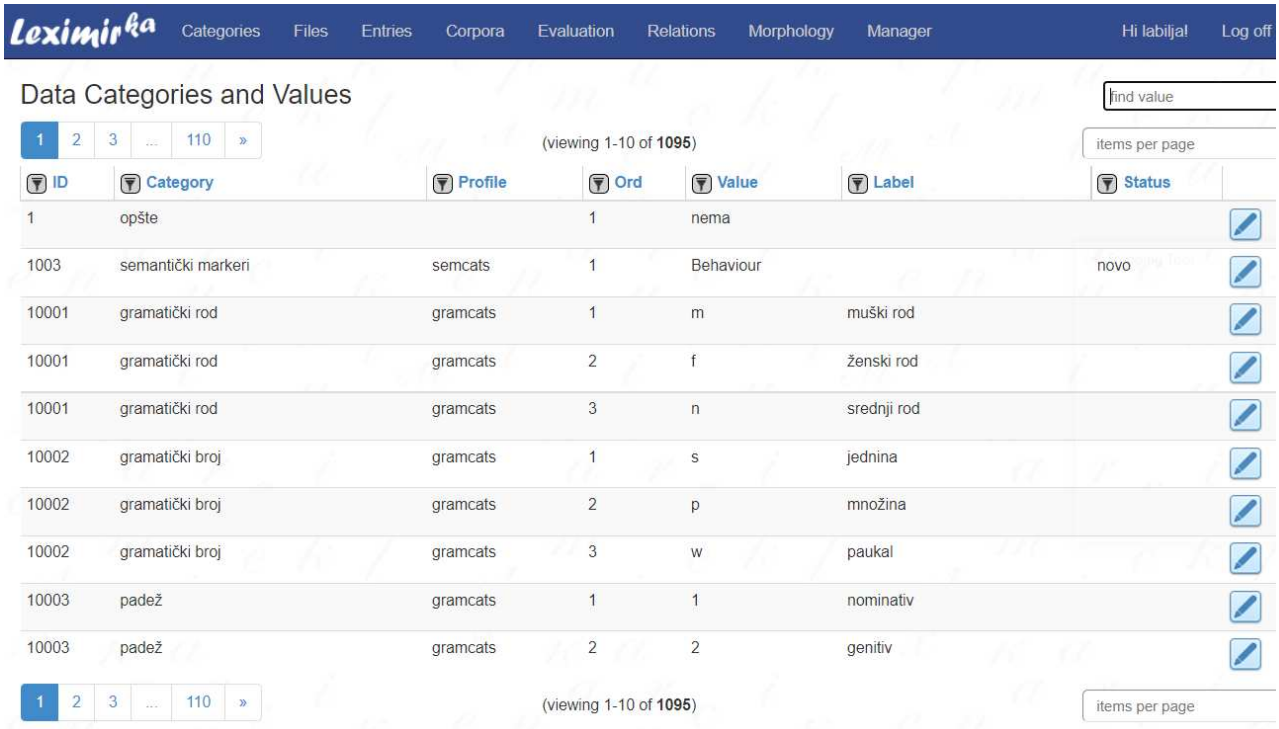
Слика П2.5 Приказ лексичких записа означених вредношћу категорије података „DOM=Agro“

Колона за дефиницију „Definition“ која се користи у панелу „Full table“ (слика П2.5). У овој колони су приказане ознаке којима је обележен лексички запис тако да је могуће додатно филтрирање подскупа лексичких записа добијеног према неким унапред задатим критеријумима путем ове колоне (преко опције уз назив колоне „Definition“). Друго дугме (приказано иконицом) води ка панелу за уређивање вредности категорије података (Слика П2.6). Дугме „View Entries List“ води ка истом панелу који је приказан на слици П2.5. Додавање нове вредности, односно доменског маркера, спроводи се уз помоћ зеленог дугмета „Add Value“ (слика П2.4) и том приликом се отвара непопуњен панел приказан на слици П2.6.



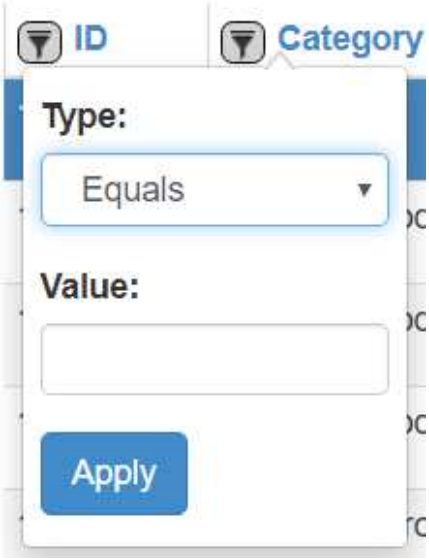
Слика П2.6 Панел за додавање и уређивање вредности категорија података

Други приступ управљању категоријама података омогућен је преко раније поменутог табеларног приказа „Tabelar View“ датог на слици П2.7. Путем горњег десног поља за претраживање могућа је претрага вредности категорија које се поново излиставају у табеларном приказу. Претрага је омогућена и преко вредности колона табеле: идентификационог броја (ID), категорије података (Category), профила (Profile), редоследа (Ord), вредности категорије података (Value), обележја (Label) и статуса (Status – празно поље предвиђено да садржи ознаку да ли је дата вредност предложена или усвојена за коришћење). Колона за редослед показује редослед дате категорије у оквиру надређене категорије. На пример, вредност колоне *редослед* за падеж у оквиру надређене категорије (grammaticalCase) је –за номинатив „1“, за генитив „2“, за датив „3“, итд. Сортирање табеле на основу вредности колоне се може обавити на уобичајен начин.



Слика П2.7 Табеларни приказ категорија података и њихових вредности

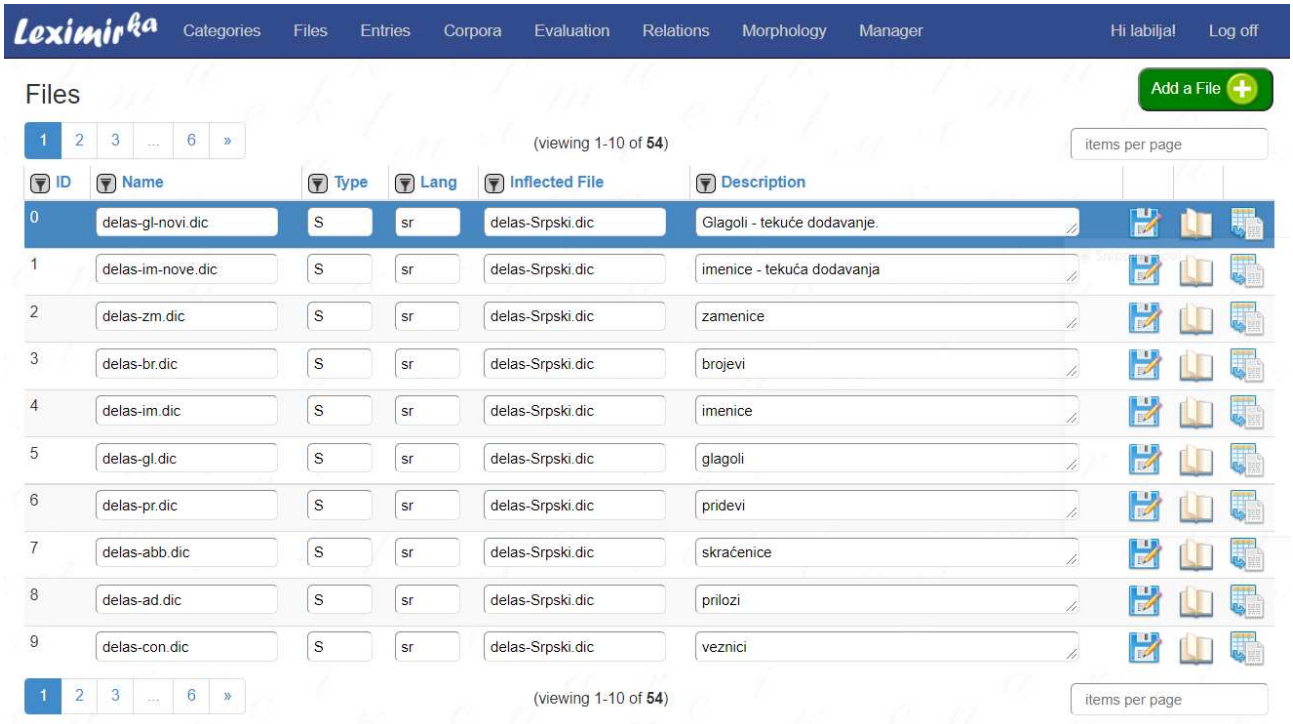
Све колоне које садрже нумерички тип података се претражују на идентичан начин. Одабиром иконице за филтрирање се отвара мали прозор у коме се бира да ли колона треба да садржи вредност која је једнака (Equals), већа (Greater than) или мања (Less than) од нумеричке вредности коју корисник уноси у пољу за претраживање (Value). Колоне које у себи садрже текстуалну вредност се филтрирају према томе да ли похрањују вредност која је иста као задата ниска (Equals), која садржи задату ниску (Contains), њом почиње (StartsWith) или се завршава (EndsWith). Приказ опције филтрирања илустрован је сликом П2.8.



Слика П2.8 Приказ опције за филтрирање
вредности категорија података

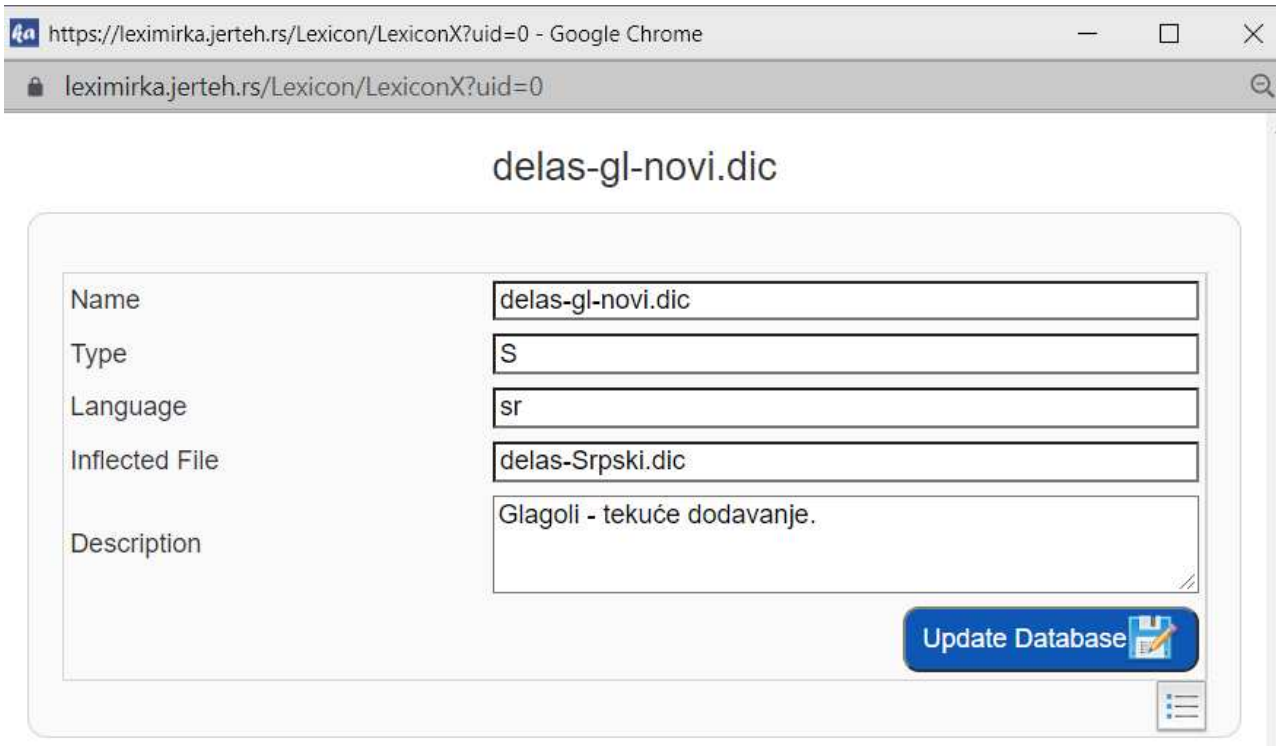
Крајње десно дугме „Edit Value“ (дато иконицом на слици П2.7) у колони приказа вредности категорије води ка панелу за уређивање вредности категорије података идентичном оном који је приказан на слици П2.6.

Илустрација панела за уређивање и прегледање датотека (опција „Files“) дата је на слици П2.9. Ове датотеке садрже део речника издвојен по неком критеријуму – нпр. глаголи или новододате речи. Овим панелом је омогућен приступ свим таквим датотекама које су похрањене у лексикографској бази. Свака датотека је представљена једним редом у табели и то идентификационим бројем (ID), називом (Name), типом (Type), језиком (Language), датотеком облика (Inflected File) и пољем за опис (Description). И овде је могуће сортирање на основу вредности колоне на уобичајен начин. Поред назива колоне на основу које је сортирана табела ће се наћи стрелица. Принцип филтрирања вредности колона је истоветан претходно описаном. Са крајње десне стране уз сваку датотеку се налазе три дугмета: чување измена (Save Changes, иконица), преглед датотеке (View, иконица), извоз датотеке у формату DELAF (Export Delaf, иконица). Више о формату DELAF може се прочитати у делу 3.1. Саме измене на постојећим подацима о датотеци се врше директно у активном панелу постављањем курсора миша на неко поље за унос података. Након измене података активацијом дугмета за чување измена (Save Changes) се подаци похрањују. Дугме за преглед речника (View) води ка прозору чији је изглед илустрован на слици П2.10. Дугме за приказ лексичких записа „Vew Entries List“ (иконица) води ка приказу лексичких записа садржаних у датој датотеци у формату за приказ који одговара сегменту апликације Лексимирка „Entries“.



Слика П2.9 Панел за уређивање и прегледање речника

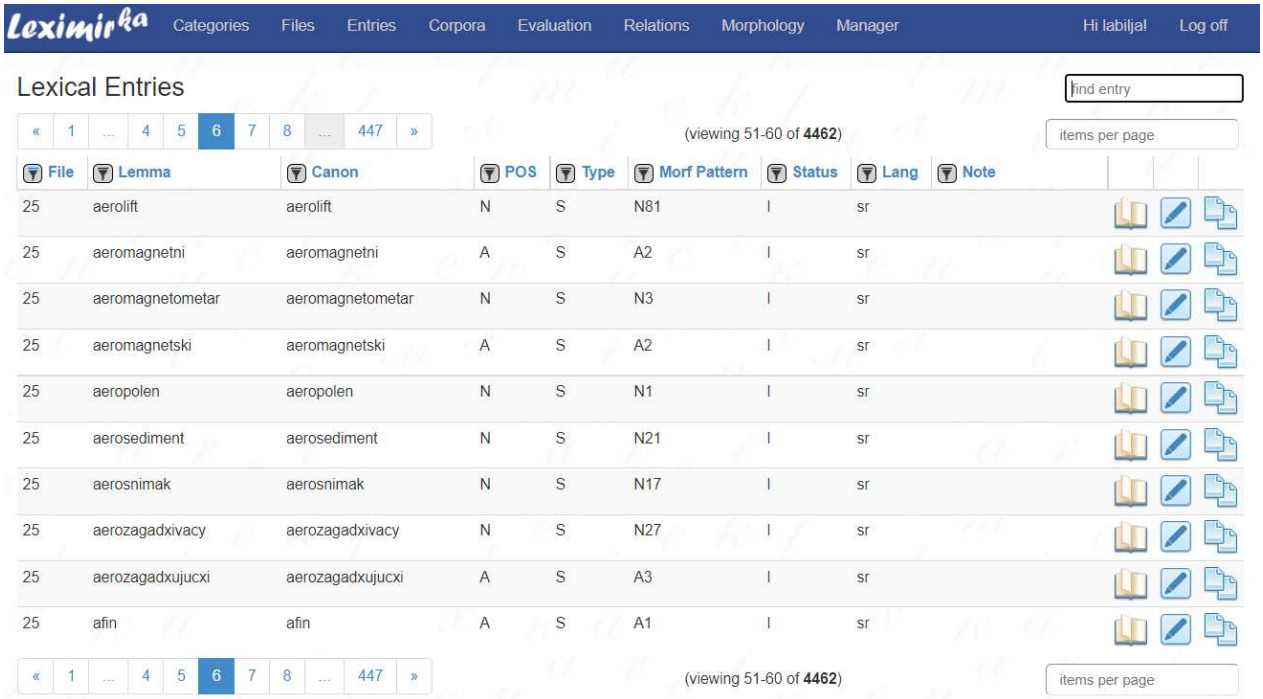
Додавање сасвим нове датотеке се започиње коришћењем зеленог дугмета „Add a File“ (слика П2.9Слика П2.9) које отвара прозор за додавање основних информација о датотеци идентичан прозору приказаном на слици П2.10.



Слика П2.10 Прозор за приказ података о датотеци и додавање нове датотеке

Уређивање, преглед и креирање нових лексичких записа омогућени су кроз сегмент „Entries“ апликације *Лексимирка*. Прелазом курсора миша преко опције “Entries” отварају се и опције за приказ пуне табеле (опција „Full table“), унос новог лексичког записа (опција „New entry“) и опција за унос новог полилексемског израза (опција „New MWU“).

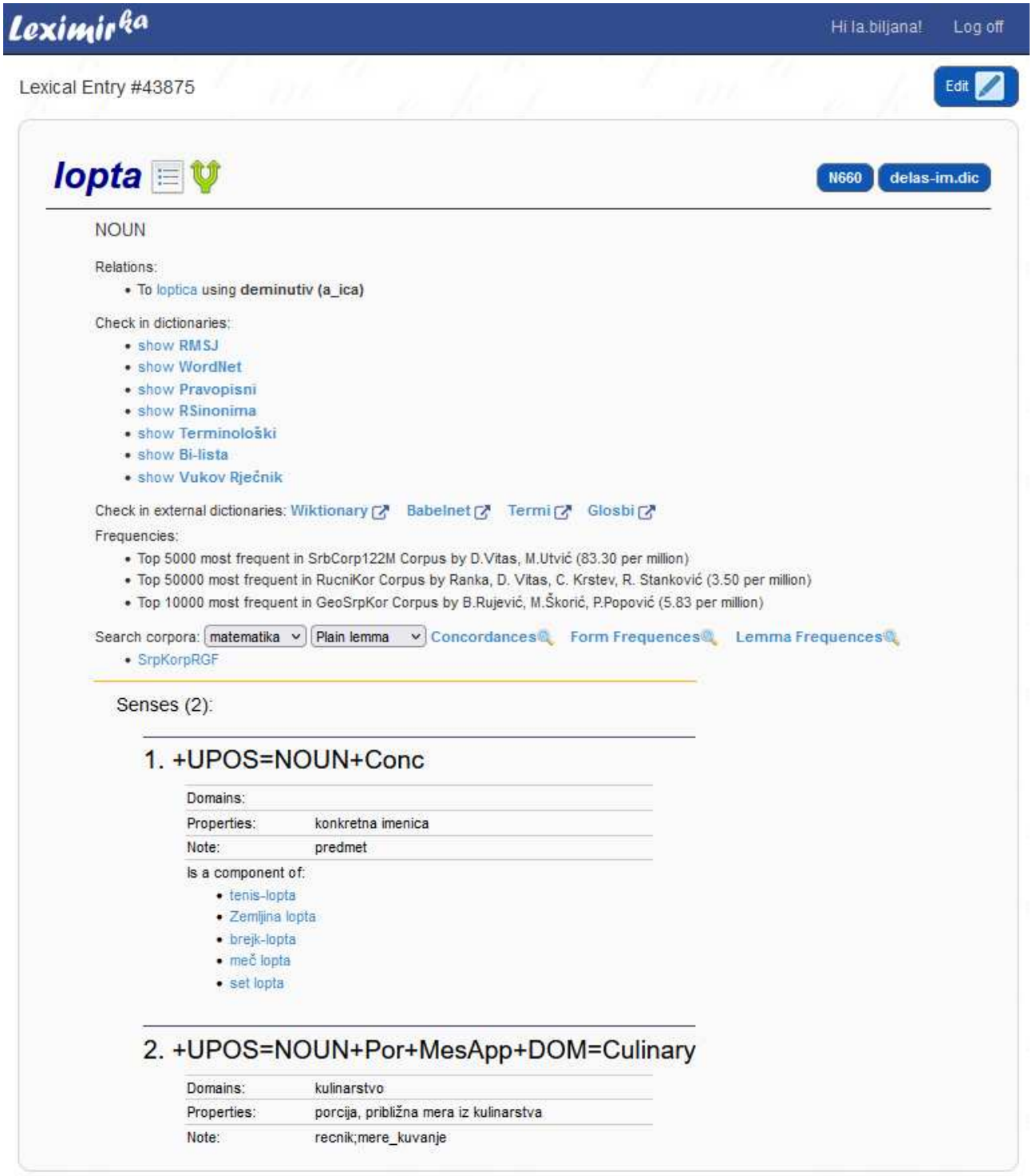
Одабиром прве и предефинисане опције „Entries“ отвара се панел за прелиставање и претрагу лексичких записа - “Lexical Entries”, приказан на слици П2.11. Сваки лексички запис је представљен једним редом у табели информацијама о идентификационом броју датотеке којој запис припада (File), леми (Lemma), канонском облику (Canon), врсти речи (POS), типу лексичког записа (Type), флективној класи (Morf Pattern), статусу лексичког записа (Status), језику (Lang) и евентуалној напомени уз лексички запис (Note). Са крајње десне стране уз сваки запис стоје три дугмета: дугме за детаљнији приказ лексичког записа (View Entry), дугме за уређивање лексичког записа (Edit Entry) и дугме за прављење копије лексичког записа (Copy Entry). На слици П2.11 приказан је део лексичких записа из датотеке речника за рударство и геологију (delas-mining.dic) који је филтриран коришћењем колоне „Lexicon“, одабиром вредности „25“ (идентификациони број датотеке delas-mining.dic). Претрага и филтрирање поља за опис лексичких записа врши се на универзалан начин описан у делу посвећеном панелу за управљање категоријама података „Data Categories and values“. Оно на шта је неопходно скренути пажњу јесте поље за претраживање „find entry“. Приликом уноса ниске за претраживање неопходно је користити код Аурора, за разлику од претраге намењене корисницима на вебу, када се може користити и латинично писмо. Подразумева се да се кроз ово поље претражују сви записи који садрже задату ниску у пољу за канонски облик леме. Уколико корисник жели да зада тачну ниску за претрагу, потребно је да изврши претрагу путем колоне за лему (Lemma), одабиром типа „Equals“.



Слика П2.11 Панел за прелиставање и претрагу лексичких записа „Lexical Entries“

Приказ панела пуне табеле (опција „Full table“) раније је приказан на слици П2.5 на којој је дат приказ лексичких записа означених вредношћу категорије података „DOM=Agro“. Основна разлика између тог панела („Full table“) и панела „Lexical Entries“ јесте главно поље за претрагу. Код тог панела (Слика П2.5) претражују се категорије којима су означени лексички записи (назив поља за претрагу је „find by property“), док се код панела за претрагу лексичких записа (Слика П2.11) претражују леме, или прецизније њихов канонски облик. Ово је потребно нагласити јер се код сложених речи (тип C, енг. *Compound*) лема и канонски облик разликују. У пољу за лему се налази облик леме у формату предвиђеном *Српским морфолошком речником* (детаљније у поглављу 3.1), док се у пољу за канонски облик налази лема исказана природним језиком. На пример, облик леме предвиђен *Српским морфолошким речником* био би „bogata(bogat.A17:aefs1g) berba(berba.N724:fs1q)“, док би канонски облик исте леме био “bogata berba”.

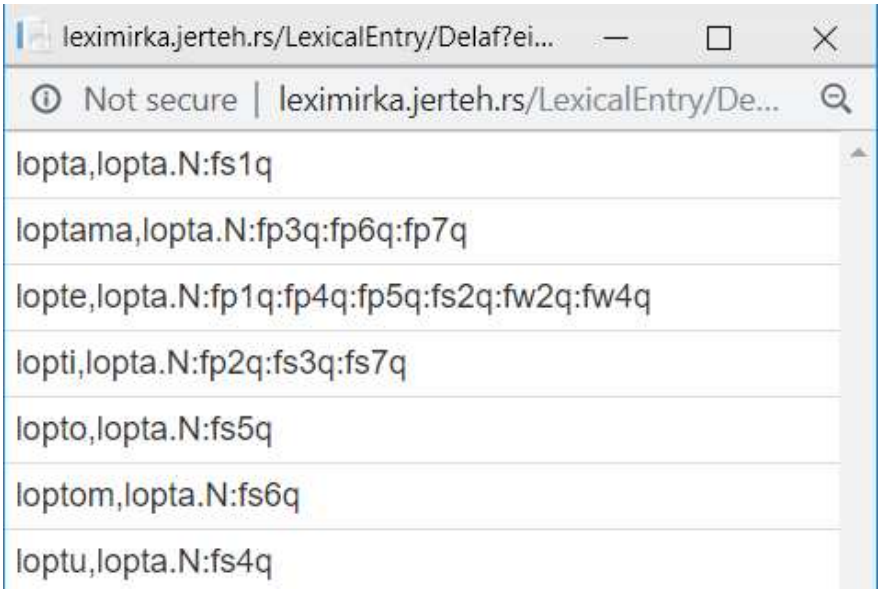
Као увод у опис опција за додавање новог лексичког записа и нове полилексемске јединице, корисно је да се позабавимо опцијама за приказ, уређивање и копирање лексичког записа који су помињани као саставни делови описа лексичког записа приликом описивања одговарајућих панела (слике П2.5 и П2.11).



Слика П2.12 Лексички запис „*lopta*“ у корисничком сучељу за управљање и развој

Панел за приказ лексичког записа (View Entry) у корисничком сучељу за управљање и развој илустрован на примеру лексичког записа „lopta“ представљен је сликом П2.12. Уколико упоредимо овај приказ са приказом истог лексичког записа у јавном приказу (слика П2.3) видећемо да је разлика у приказу додатних информација и постојању додатних могућности. Сада се у горњем десном углу приказује податак о системски одређеном идентификационом броју лексичког записа (Lexical Entry #43875). Поред леме се налази опција за приказ свих облика леме (опција List all forms) (Слика

П2.13), као и опција за приказ могућих деривација (опција List possible derivations). Са десне стране у истом реду се налазе два плава дугмета на којима су исписане флективна класа (N660) и назив датотеке у чијем саставу је лексички запис (delas-im.dic) чијим активирањем се добијају све речи које деле исту флективну класу, односно сви лексички записи датог подскупа речника. У делу Check in dictionaries су излистани други речници у којима је могуће пронаћи и консултовати лексички запис са текућом лемом (више о овоме је било речи у поглављу 5.2.5). У делу Check in external dictionaries дата је хипервеза ка екстерним изворима које је могуће консултовати датом лемом (више о овоме је било речи у поглављу 5.2.5). Доступна је и претрага различитих корпуса лемом лексичког записа (опција Concordances) или претрага фреквенција појављивања саме леме (опција Lemma Freqences) или свих њених облика (опција Form Freqences) о чему је више речи било у одељку 5.2.3. У приказу значења сада постоји и приказ садржаја поља за напомене (Note). У горњем десном углу се налази дугме за уређивање текућег лексичког записа (Edit) које води ка панелу за уређивање лексичког записа приказаном на слици П2.14.

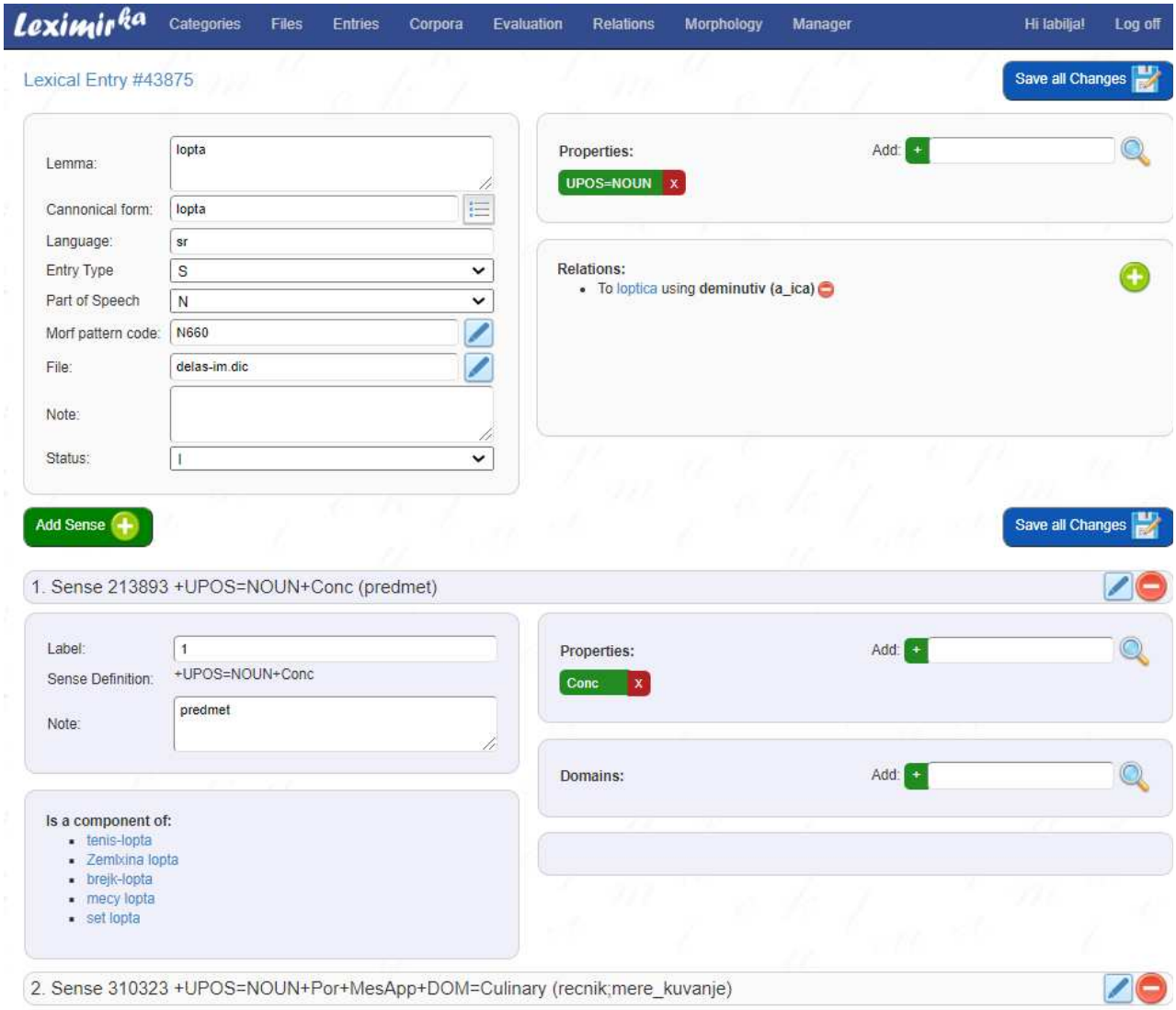


Слика П2.13 Панел са приказом свих флективних облика леме

Панелу за уређивање лексичког записа (слике Слика П2.14 и Слика П2.18) је могуће приступити са више приступних тачака путем иконице за уређивање лексичког записа (Edit entry). Панел се, условно речено,[126] састоји од две уоквирене целине. Прва се односи на опште информације о леми и лексичком запису, док се друга односи на значења. Запис приказан на слици П2.14 има два значења, с тим што је прво приказано у потпуности док је друго дато у сажетом приказу (једна од опција у *Лексимирки*) ради уштеде простора. Кроз панел је могуће кориговати све лексичке информације али на различите начине. Поља за лему (поље Lemma), канонски облик (поље Cannonical form), језик (поље Language) и напомену (поље Note) су текстуална, па се стога текст може слободно кориговати. Поља за унос типа лексичког записа (Entry Type) и врсте речи (Part of Speech) су у форми падајућег менија са понуђеним контролисаним вредностима. Поља за ознаку флективне класе (Morf pattern code) и датотеку (File) се попуњавају одабиром опције која се бира из новог прозора који се

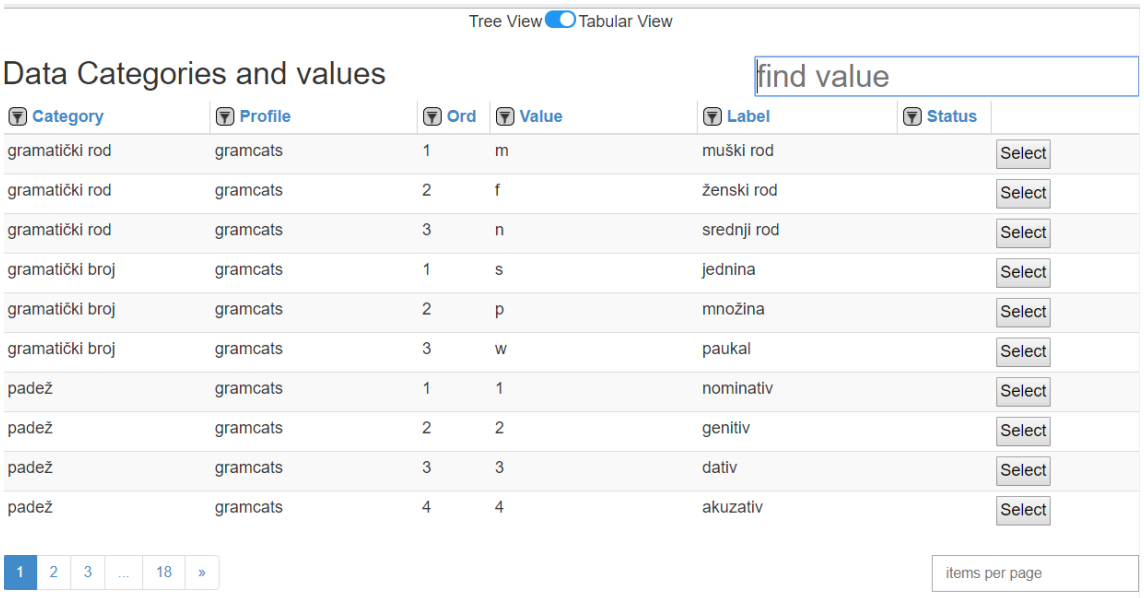
[126] Број целина које се односе на значење има онолико колико има значења.

отвара активирањем дугмета за промену вредности (Replace Value) или директним куцањем где се филтрирају понуђене вредности.



Слика П2.14 Приказ панела за уређивање лексичког записа на примеру уређивања монолексемске речи

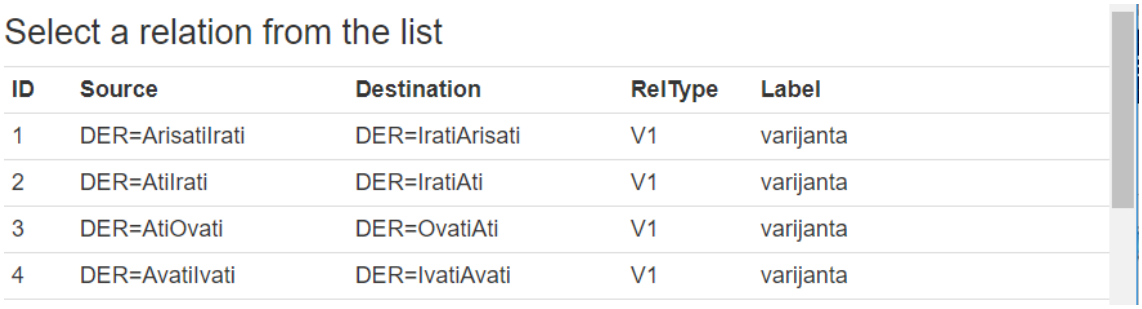
Додавање категорија података (Properties) лексичком запису на нивоу леме могуће је одабиром вредности из панела који се добија активацијом дугмета за претрагу (Search). Панел који се отвара (Слика П2.15) јесте сличан контролној табли за управљање категоријама (Слика П2.7), с тим што уз појединачну категорију уместо опција за уређивање стоји опција за избор (Select). У горњем делу се налази дугме којим се бира да ли желимо табеларни приказ категорија (Tabular View) или приказ у облику дрвцета (Tree View). Када се одаберу специфична категорија, њена вредност (Value) се појављује у пољу за својство (Properties) у панелу приказаном на слици П2.14. Како би се вредност додала потребно је потврдити избор дугметом које се налази са леве стране поља за унос. Уколико корисник зна тачну вредност категорије, може је изабрати из падајућег менија који се појављује у истом пољу када се преко њега пређе курсором миша преко поља за унос текста (Properties) и додати коришћењем истог дугмета .



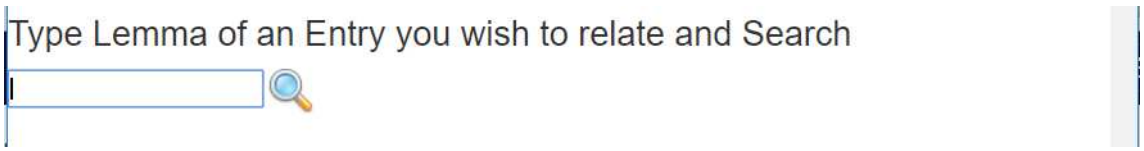
Слика П2.15 Панел за одабир и додавање категорија

Кроз панел приказан за уређивање лексичког записа (Слика П2.14) је могуће успостављати везе са другим лексичким записима кроз сегмент за релације (Relations). Уколико је успостављена нека релација, биће приказана као у текућем примеру где је лексички запис „лопта“ повезан са записом „лоптица“ релацијом деминутив (deminutiv (a_ica)). Уз успостављену релацију се налази дугме које служи за брисање релације (Delete relation ⛔). Такође је могуће додати нову релацију дугметом за додавање (Add relation ➕). Овим путем се активира панел са списком постојећих релација (Слика П2.16) са кога корисник бира одговарајућу релацију и након тога уноси лему лексичког записа (Слика П2.17) са којим жели да успостави везу. Све коначне измене начињене у лексичком запису чувају се коришћењем дугмета за чување измена (Save all changes 

) (Слика П2.14). О успостављању веза је било више речи у одељку 5.2.1.



Слика П2.16 Панел за одабир релације

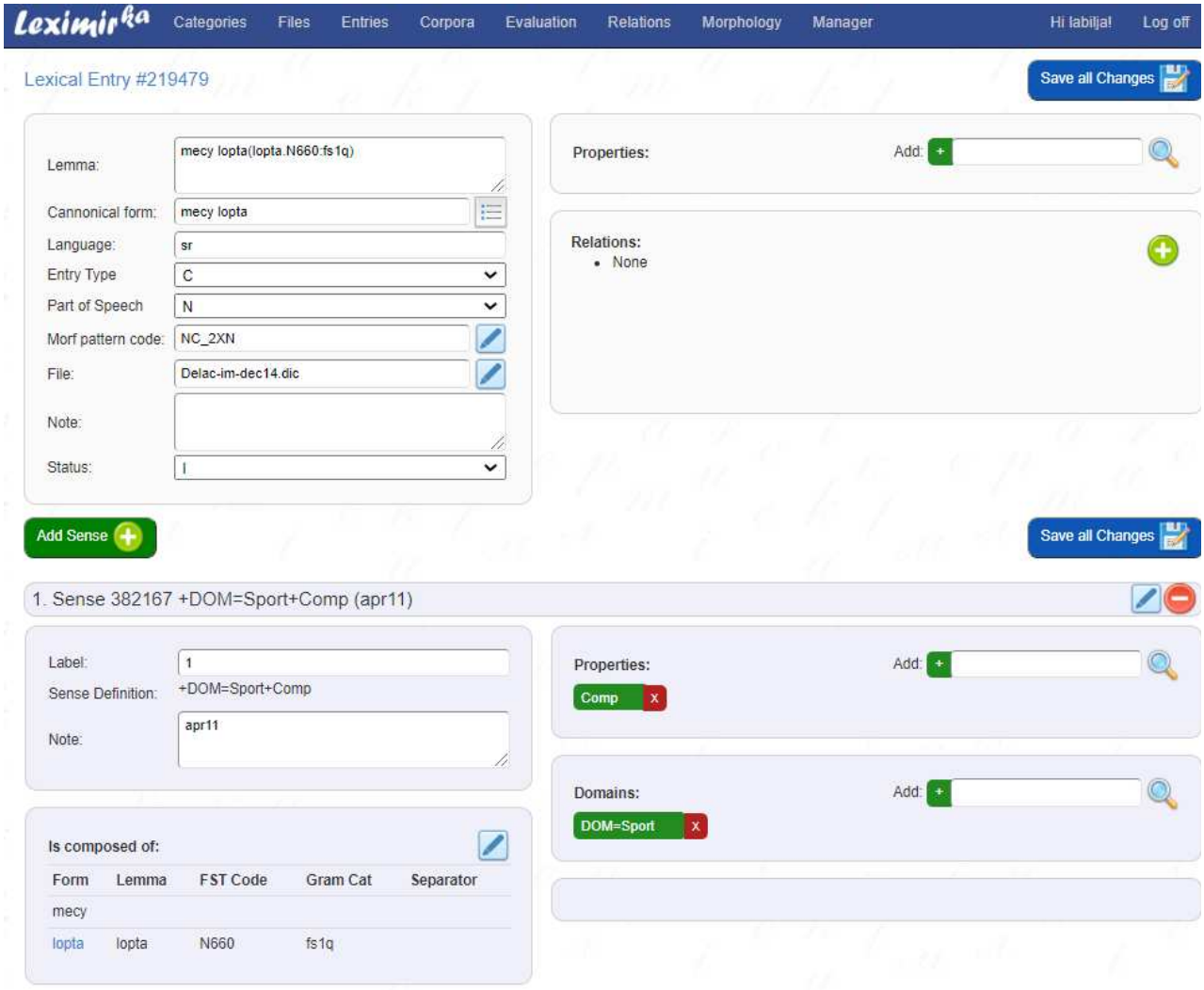


Слика П2.17 Одабир леме записа за повезивање

Значења лексичког записа приказују се на начин који је приказан за друго значење лексичког записа приказаног на слици П2.14. Потребно је одабрати дугме 

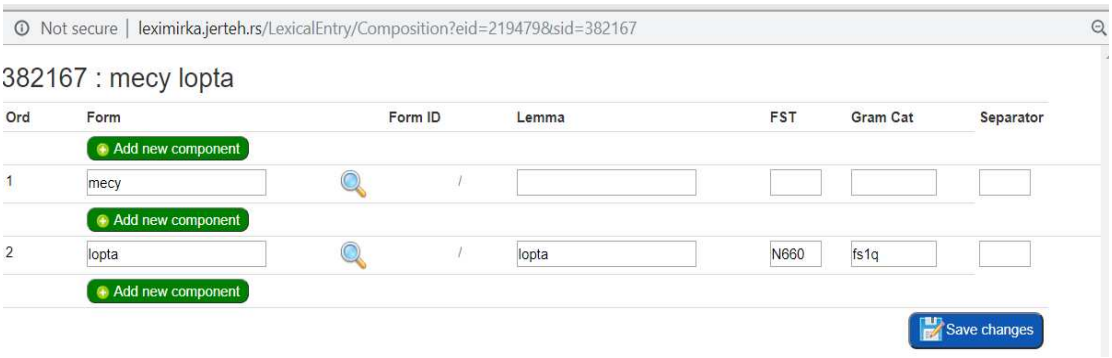
како би се приказала поља за уређивање. За обрнути поступак – уклањање поља за едитовање служи дугме ⛔ које се налази поред, уз вредности текућег значења, а исто се може урадити и простим кликом миша преко линије вредности категорија. У сегменту једног значења са леве стране су дате вредности категорија које чине једно значење. На слици П2.14 је то маркер „+Conc“. Поље ознаке (Label) се односи на текуће значење у оквиру једног лексичког записа, стога вредност 1 показује да се ради о првом значењу. У пољу за дефиницију значења (Sense Definition) приказују се вредности придружених категорија. Поље за напомене (Note) се попуњава слободним уносом текста, док се поље за дефиницију значења (Sense Definition) попуњава уз помоћ опција за додавање својстава (Properties) и домена (Domains), приказаним са десне стране. Поље за додавање својстава је раније описано. Брисање својства се врши кликом на црвени крстић у приказу додатог маркера 

. Додавање домена се врши на истоветан начин као и додавање својстава. За лексичке записе који учествују као компоненте у формирању полилексемског израза ће у делу предвиђеном за то (Is component of) бити наведене леме таквих полилексемских израза у виду хипервеза које воде до њих. Лексички запис „лопта“ је компонента полилексемских израза „тенис-лопта“, „Земљина лопта“, „брејк-лопта“, „меч лопта“ и „сет лопта“. Повезивање лексичког записа са компонентом се врши кроз панел за уређивање полилексемског израза или сложене речи, приказаном на слици П2.18.



Слика П2.18 Приказ панела за уређивање лексичког записа на примеру уређивања полилексемског израза

Панел за уређивање полилексемског израза (Слика П2.18) функционише и изгледа скоро истоветно као панел за уређивање монолексемских речи. Једина разлика је у делу који показује од којих компоненти се састоји полилексемски израз (Is composed of). У конкретном примеру видимо лексички запис „меч лопта“, који се састоји из компоненти „меч“ и „лопта“. Компонента „меч“ се не мења кроз флективне облике у овој полилексемској јединици док се компонента „лопта“ мења па је потребно остварити везу са одговарајућим лексичким записом. Панел за уређивање (Слика П2.19) се отвара дугметом за уређивање компоненти (Edit composition). Колоном за редослед је приказан редослед компоненти у полилексемском изразу. Додавање нове компоненте се постиже за то предвиђеним дугметом (Add new component). Брисање компоненте се остварује тако што се курсор миша позиционира на компоненту за брисање и потом се компонента повуче. У том тренутку се појављује иконица канте за отпатке ( ) у коју је могуће превући дату компоненту. Жељена лема се уноси у поље колоне за облик „Form“ док коришћењем дугмета за претрагу започиње претрага лексичких записа, односно њихових значења. Као резултат се отвара панел за одабир облика приказан на слици П2.20.



Слика П2.19 Панел за уређивање/додавање компоненти полилексемском изразу

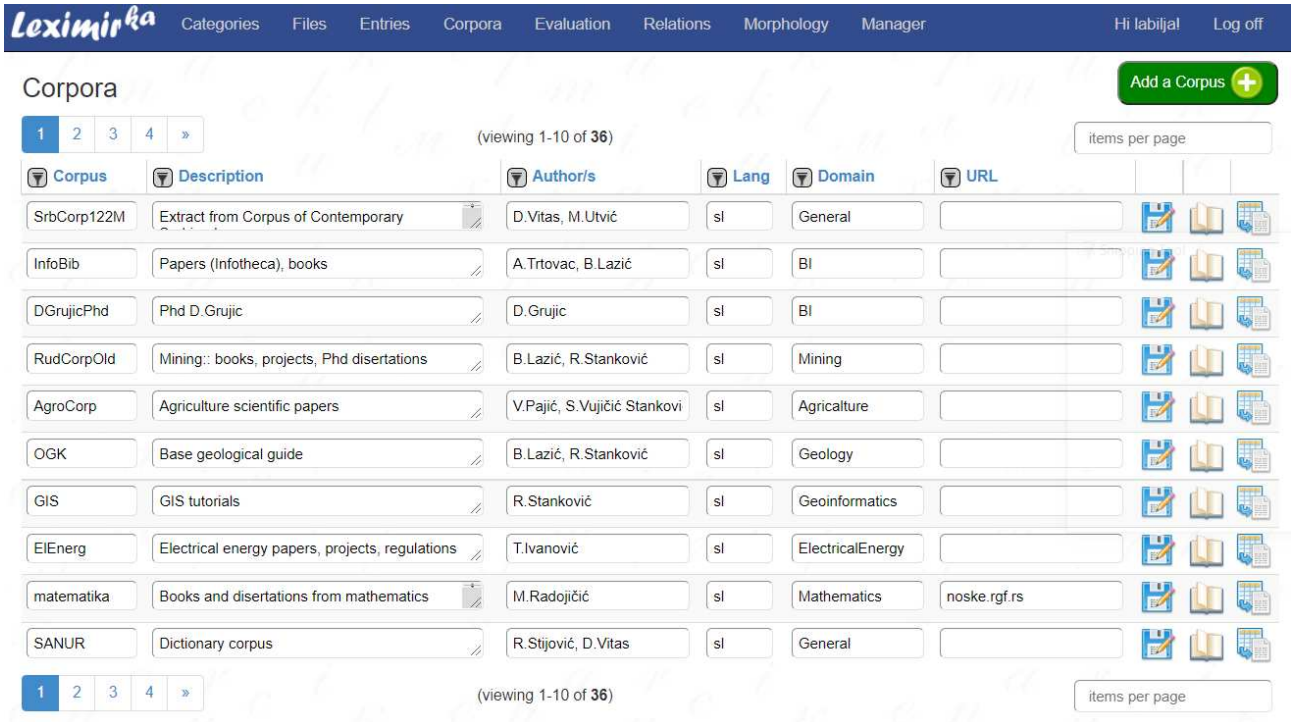
На основу значења и граматичких категорија се врши избор адекватног облика који ће бити компонента полилексемског израза. У нашем примеру је то лексички запис означен идентификационим бројем „29033“, флективну класу „N660” и то значење везано за предмет, а не за кулинарску мерну јединицу, граматичких ознака „fs1q“ које показују да се ради о облику женског рода једине номинатива (неживом). Одабир се врши кликом миша на адекватан ред. Апликација се потом враћа на панел за додавање са слике П2.19 где можемо сачувати измене дугметом Save changes. Потом је потребно сачувати све измене у панелу самог лексичког записа (Слика П2.18).

Pick Correct Entry

FormID	Lemma	GramCatsOrdNo	gramCats	Note	Definition	LexicalSenseID	Morf
29033	lopta	1	fs1q	predmet	+Conc	213893	N660
29033	lopta	1	fs1q	recnik;mere_kuvanje	+Por+MesApp+DOM=Culinary	310323	N660
1167522	loptati	1	Ays		+Imperf+It+Ref	224214	V501
1167522	loptati	2	Azs		+Imperf+It+Ref	224214	V501
1167522	loptati	3	Pzs		+Imperf+It+Ref	224214	V501

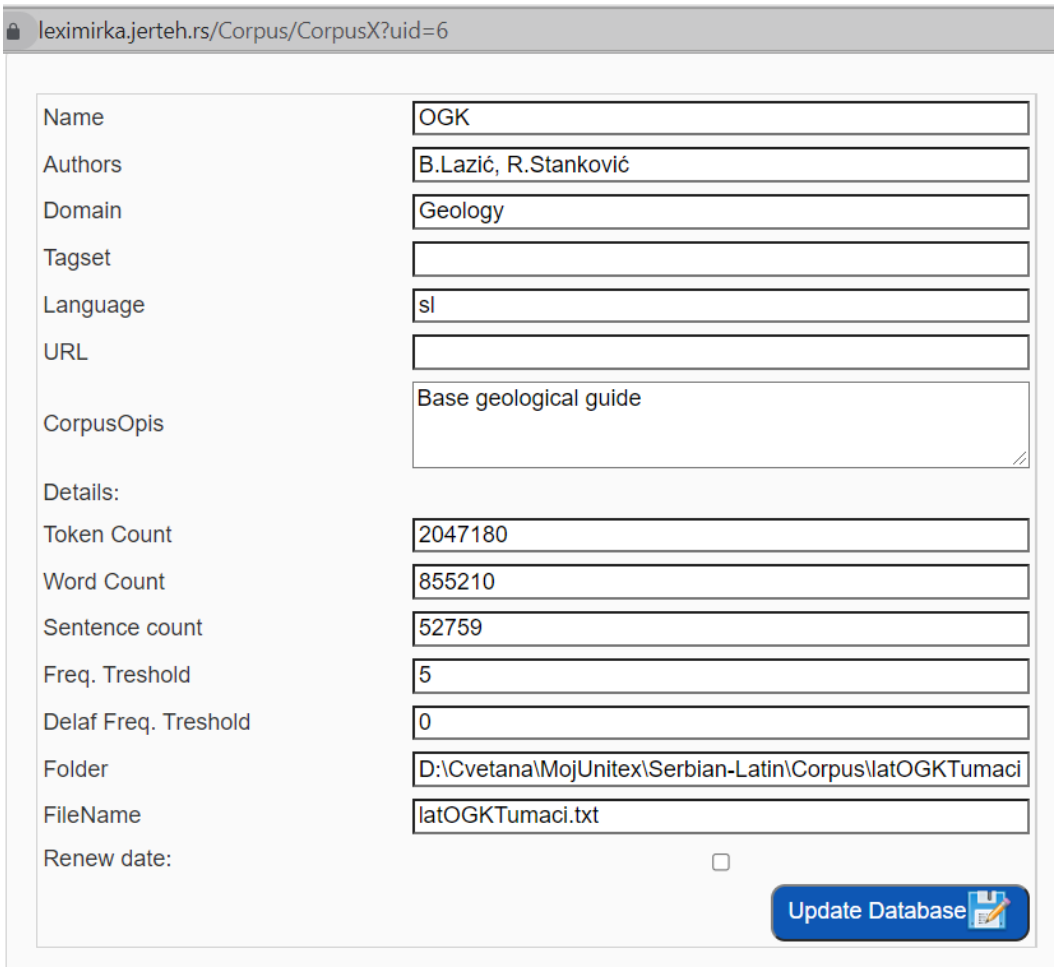
Слика П2.20 Панел за одабир облика компоненте

Сегмент за корпусе (Corpora) садржи метаподатке о корпусима коришћеним за претраживање конкорданци и израчунавање фреквенција појављивања лема и облика (Слика П2.21). Сваки корпус је приказан колонама за скраћени назив (колона Corpus), опис (колона Description), ауторе (колона Author/s), језик (колона Lang), домен коме припада (колона Domain) и хипервезу ка корпусу (опција URL). Уз метаподатке сваког појединачног корпуса доступна је позната дугмад за чување коригованих метаподатака, преглед и уређивање метаподатака и извршавање додавања фреквенција (дугме) према методологији која је описана у одељку 5.2.3.

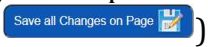


Слика П2.21 Панел за корпусе

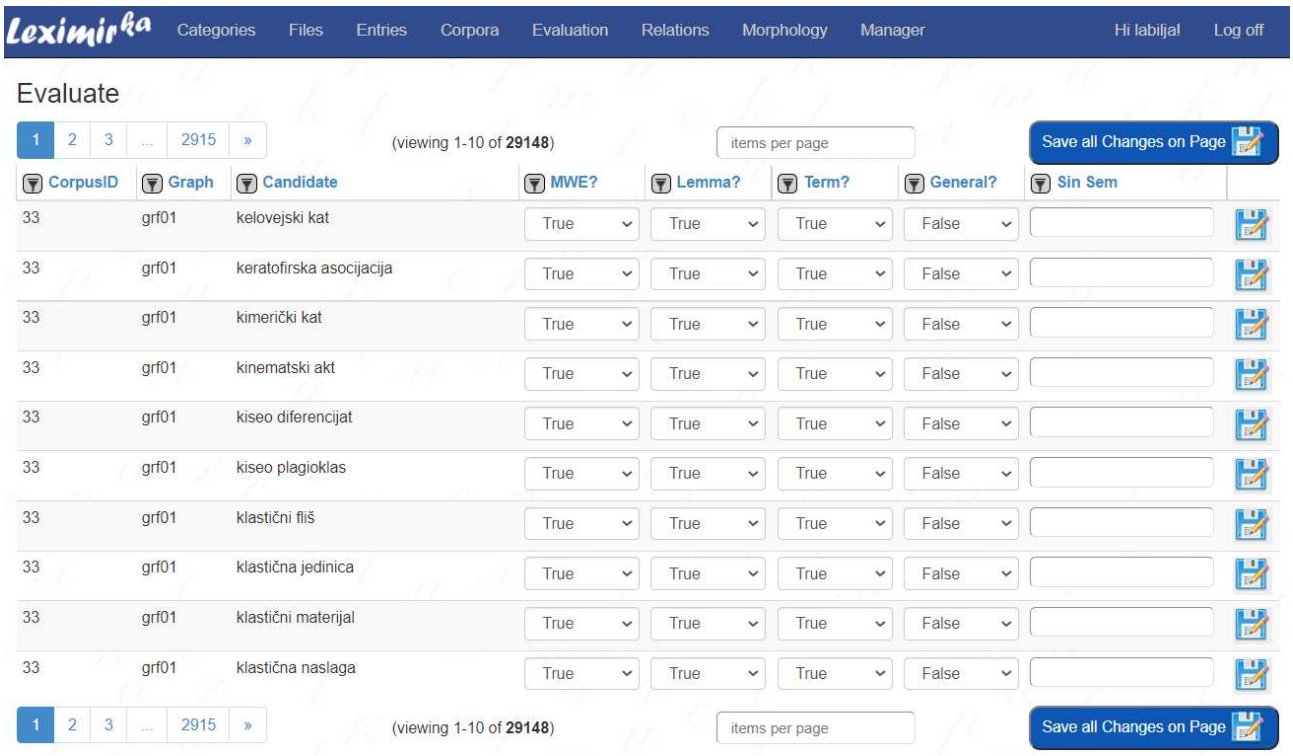
На слици П2.22 приказан је панел за преглед и уређивање детаљнијих метаподатака о корпусу. Осим података приказаних у претходном панелу, могуће је уређивати информације о сету ознака коришћених за обележавање (поље Tagset) и величини корпуса. За опис величине корпуса користе се следећи параметри: број токена (Token Count), број речи (Word Count), број реченица (Sentence count), праг који се узима за приказ фреквенција (Freq. Treshold), праг који се узима за приказ фреквенција из речника DELAF (Delaf Freq. Treshold), као и путања до датотеке (Folder) и њен назив (FileName). Могуће је и ажурирање датума додавања корпуса имајући у виду додавање новије верзије (Renew date).



Слика П2.22 Панел за уређивање метаподатака о корпусу

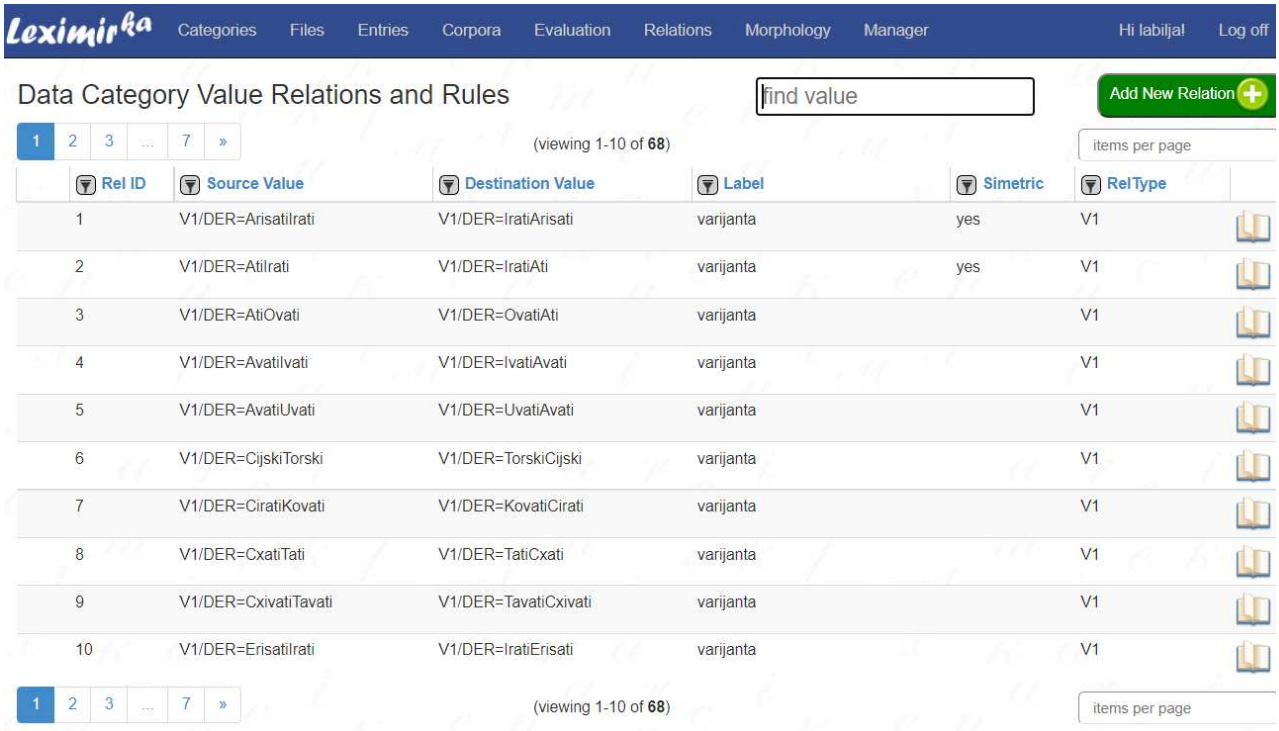
Сегмент за евалуацију записа (Evaluation) омогућава да више независних евалуатора анализира листу препознатих термина (кандидата за допуну речника). На слици П2.23 је представљен панел путем кога евалуатор оцењује да ли је екстраховани кандидат полилексемски израз (колона MWE?), да ли флективни облик речи који је екстрахован одговара леми (колона Lemma?), да ли представља термин у унапред задатом домену (колона Term?) и да ли исти кандидат представља термин у другим доменима (колона General?). Евалуација се врши избором опције истинито „True“ или погрешно „False“ у падајућем менију у оквиру колоне. Предефинисана вредност за прве три колоне је „True“, док је за четврту колону „False“. Евалуатор је такође у могућности да у колони за синтаксичке и семантичке ознаке (колона Sin Sem) дода адекватне маркере за које сматра да описују датог кандидата. Овим панелом није предвиђено кориговање неисправних лема (када је у колони Lemma вредност „False“). Прве четири колоне представљају идентификациони број корпуса из кога су екстраховани термини (колона CorpusID), ознаку графа уз помоћ кога је кандидат препознат (колона Graf), граматички број кандидата (колона Form) и самог кандидата (колона Lemma). На уобичајен начин су доступни претрага и филтрирање према колонама а чување промена је могуће на нивоу кандидата (дугме ) и на нивоу целе стране (дугме 

). Више речи о практичној примени овог панела било је у поглављу 7.1.



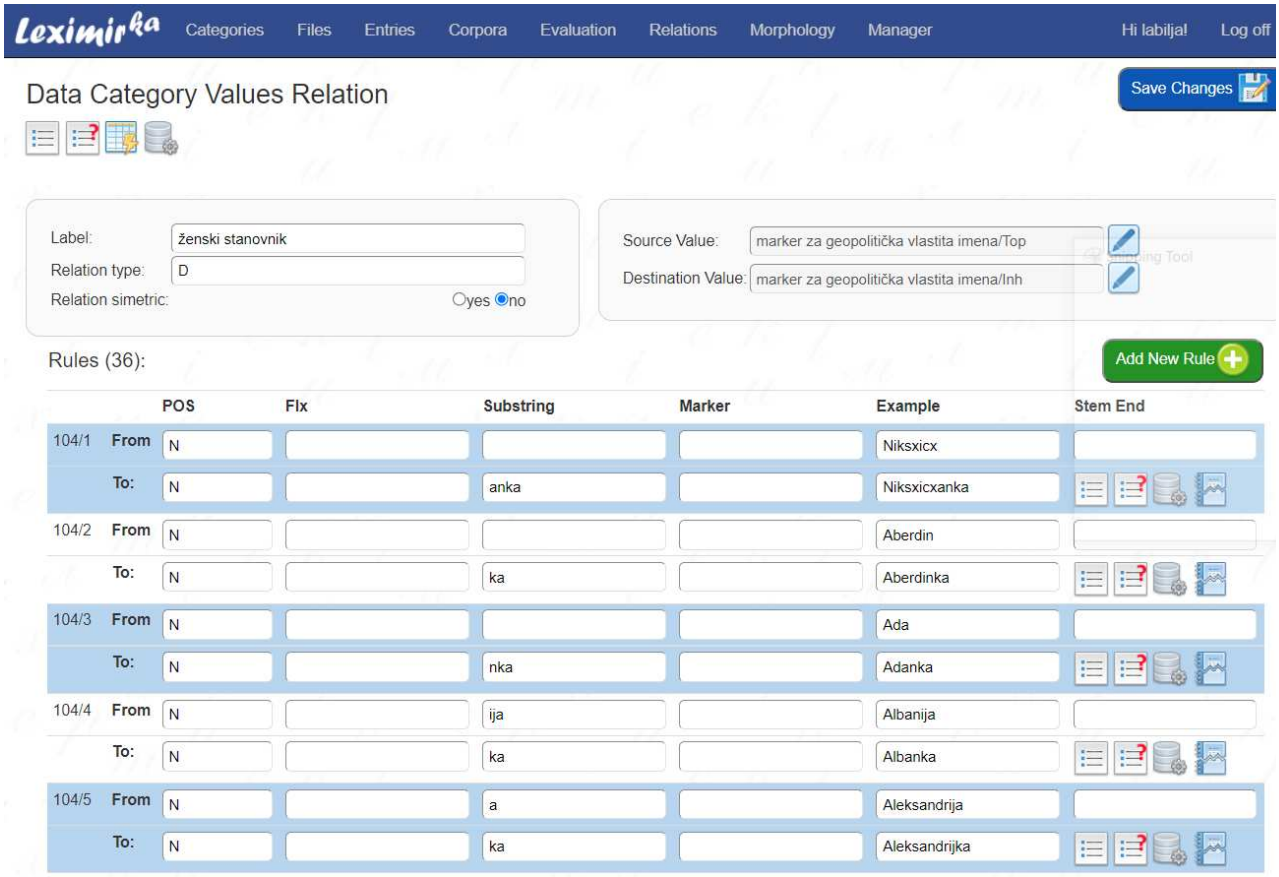
Слика П2.23 Панел за евалуацију кандидата за Речник

Сегмент „Relations“ у апликацији *Лексимирка* пружа могућност прегледа постојећих релација и управљање њима, као и креирања нових релација и правила за повезивање лексичких записа из лексикографске базе. На слици П2.24 дат је приказ почетног панела за преглед постојећих релација и управљање њима.



Слика П2.24 Панел за преглед постојећих релација и управљање њима

Свака релација је приказана једним редом у табели панела и то идентификационим бројем релације (Rel ID), ознаком релације (Source Value и Destination Value), обележјем (Label), маркером симетричности (Simetric), и типом релације (RelType). У колони за ознаку полазне вредности релације (Source Value) дата је ознака релације која се састоји од ознаке типа релације и подниске коју треба да садрже полазни и крајњи запис који се повезују (нпр. V1/DER=ErisatiIrati). Ознака дата у колони за ознаку крајње вредности релације (Destination Value) је структурно истоветна, с тим што приказује обрнути смер, подниске од крајњег ка полазном запису (нпр. V1/DER=IratiErisati). Колоном за ознаку релације (Label) дати су називи релација исказани природним језиком (нпр. елизија, негација, мушки становник, женски становник, варијанта, итд.). Колоном за симетричност (Simetric) приказује се да ли је дата релација симетрична или није. У колони за тип релације (RelType) дата је ознака типа релације (нпр. V1, El, EkIjk, итд.) (више о врстама релација било је речи у поглављу 5.2.1). Употребом дугмета за преглед правила „View Rules“ (иконица 📖) долази се до панела који садржи сва правила придружена једној релацији (Слика П2.25).

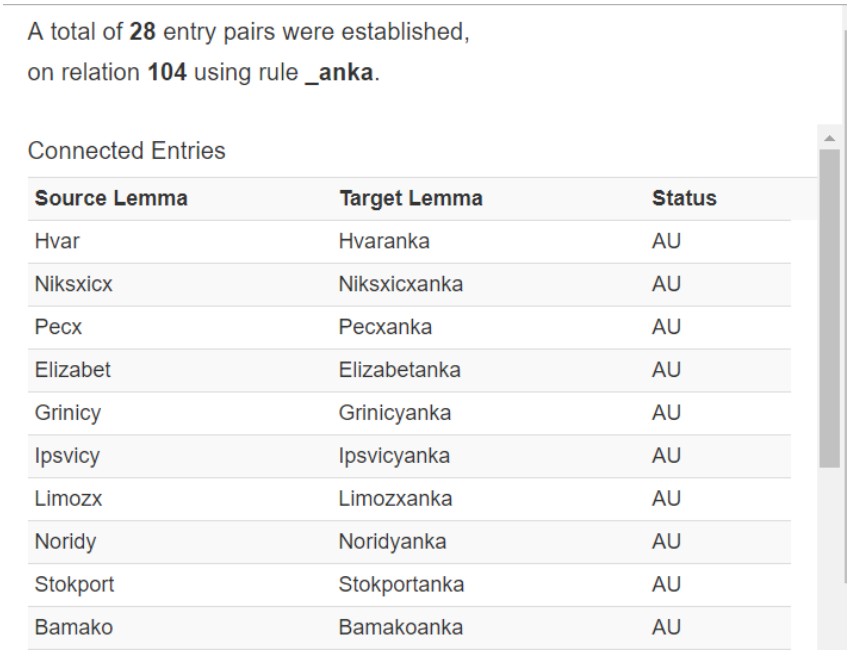


Слика П2.25 Панел за управљање правилима за успостављање релације на примеру релације *женски становник*

На слици П2.25 је приказан панел за управљање правилима за успостављање релације која је названа „женски становник“ а повезује геополитичко име са називом женског становника. У оквиреним деловима панела осенченим сивом бојом се налазе административни подаци о релацији, који се приказују као општи подаци о релацији на претходном панелу (Слика П2.24). Ове податке је могуће мењати у оквиру текућег панела. Поља за ознаку релације (Label) и тип релације (RelType) су текстуална док се симетричност (Simetric) одређује чекирањем. Поља за избор полазне и крајње вредности релације (Source Value и Destination Value) попуњавају се употребом дугмета плаве боје (иконица) које се налази уз само поље (Change Source Value и Change Destination Value). Одабиром овог дугмета искаче нови прозор који садржи панел за одабир и додавање категорија који је раније приказан на слици П2.15. У овом панелу се бира вредност категорије којом треба да буду обележени полазни и крајњи запис између којих се успоставља релација. Одабиром једне вредности се попуњавају потребна поља и тиме се дефинише први критеријум за успостављање једне релације. У примеру приказаном на слици П2.25 за полазни критеријум је постављен маркер који означава геополитичко име (+Top) а за крајњи критеријум је маркер за ознаку становника (+Inh). Све измене се чувају употребом дугмета плаве боје (иконица Save Changes) из горњег десног угла.

Сама правила за успостављање релација дата су у табели у виду редова наизменично осенчених белом и плавом бојом. Релација приказана на слици П2.25 дефинисана је уз помоћ 36 правила (информација приказана у горњем левом углу изнад табеле) од којих се на слици види 5. У првој колони се налази идентификациони број правила сачињен од идентификационог броја релације и редног броја правила у оквиру

те релације раздвојених знаком „/“, нпр. 104/1, 104/2, итд. У следећој колони се налазе информације о врсти речи (POS) полазног и крајњег лексичког записа. На слици П2.25 су приказана правила која повезују именице - N. У колони за флективну класу (Flx) могуће је задати критеријум за повезивање заснован на флективној класи полазног и крајњег лексичког записа што у овом случају није применљиво. Колоном за подниску (Substring) задају се правила која се односе на подниске које треба да садрже леме лексичких записа за повезивање. Првим правилом са слике П2.25 (104/1) дефинисано је да је за полазну лему подниска празна и да крајња лема треба да садржи подниску „anka“, док је четвртим правилом (104/4) дефинисано да полазна лема треба да садржи суфикс „ija“ док крајња лема треба да садржи суфикс „ka“. Колоном за маркер (Marker) се поставља критеријум заснован на маркеру којим треба да буду обележени полазни и крајњи запис. У овом случају је то речено на нивоу релације. У колони за пример (Example) су дати примери полазних и крајњих лема лексичких записа који су повезани на основу примене текућег правила. Примери којима се илуструје примена првог правила су *Никшић* и *Никшићанка*, док се примерима *Албанија* и *Албанка* илуструје примена четвртог правила.



A total of **28** entry pairs were established,
on relation **104** using rule **_anka**.

Connected Entries

Source Lemma	Target Lemma	Status
Hvar	Hvaranka	AU
Niksxicx	Niksxicxanka	AU
Pecx	Pecxanka	AU
Elizabet	Elizabetanka	AU
Grinicy	Grinicyanka	AU
Ipsvicy	Ipsvicyanka	AU
Limozx	Limozxanka	AU
Noridy	Noridyanka	AU
Stokport	Stokportanka	AU
Bamako	Bamakoanka	AU

Слика П2.26 Преглед повезаних лексичких записа

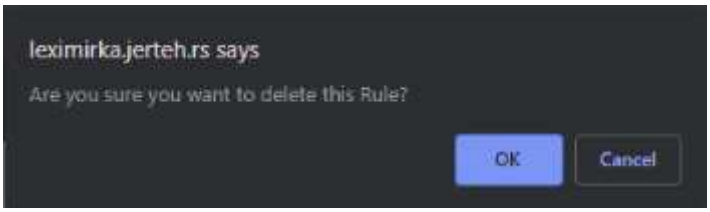
Уз свако правило за успостављање релације се налазе четири дугмета која служе за управљање правилом и његову примену. Дугметом за преглед успостављених веза „View Connections“ (иконица) искаче нови прозор у коме се налази табела са прегледом лексичких записа повезаних изабраним правилом (Слика П2.26). Путем овог приказа се добија информација о броју повезаних парова насталих применом наведеног правила одређене релације. На слици П2.26 је приказано неколико од укупно 28 парова лексичких записа повезаних на основу релације 104 правилом „_anka“, односно додавањем суфикса „anka“ на полазни лексички запис. Изворна лема (Source Lemma) је приказана првом колоном, крајња лема (Target Lemma) другом, док је трећом колоном приказан статус пара (Status), односно да је повезивање извршено аутоматски (AU).

Дугме за приказ кандидата „View Candidates“ (иконица) отвара нови прозор који приказује које лексичке записе би повезало текуће правило. Ово дугме је могуће извршити једино након креирања новог правила, односно пре извршавања

правила јер у супротном неће бити кандидата за повезивање (пошто је правило извршено, кандидати су већ повезани).

Активацијом дугмета за извршавање правила „Execute Rule“ (иконица) извршава се повезивање лексичких записа применом услова постављених правилом. Након извршеног правила се појављује прозор са повезаним лексичким записима (Слика П2.26).

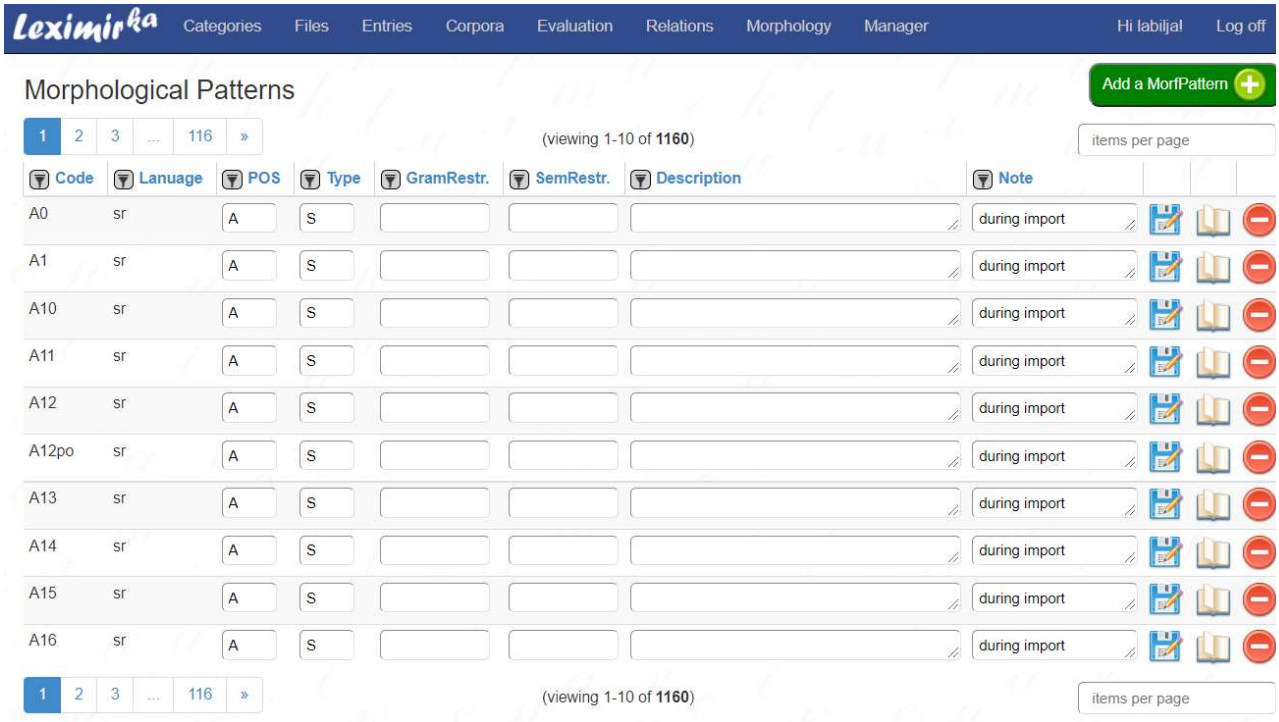
Брисање правила се постиже употребом дугмета за брисање „Delete Rule“ (иконица). Након активације дугмета се појављује прозор програма *Java* у оквиру кога се захтева потврда одлуке за брисање правила (Слика П2.27).



Слика П2.27 Брисање правила

Додавање новог правила релацији се покреће кликом на дугме за додавање „Add New Rule“ (иконица Add New Rule) чијом активацијом се на крај табеле са правилима додаје нови ред са аутоматски додељеним идентификационим бројем правила и преосталим празним пољима за попуњавање. Све измене се чувају на уобичајен начина (иконица Save Changes). У линији овог дугмета, са његове леве стране, налазе се четири дугмета која изгледају као дугмићи уз правила и имају исту улогу, али извршавају задатке на нивоу комплетне релације.

Сегмент за приказ флективних образаца (Morphology) за сада нуди приказ метаподатака о свим флективним обрасцима који се користе у *Морфолошким речницима за српски језик*. Панел морфолошки обрасци (енг. *Morphological Patterns*) приказан је на слици П2.28. Ознака флективног обрасца описана је колоном за код (Code), потом следи колона за језик на који се односи флективни образац (Language). Ознака врсте речи на коју се односи флективни образац дата је колоном за врсту речи (POS) док је тип записа, у смислу монолексемска или полилексемска реч, дат у колони за тип (Type). Колоном за граматичку рестрикцију (GramRestr.) предвиђено је бележење граматичког ограничења које се односи на актуелни образац док је колоном за семантичку рестрикцију (SemRestr) предвиђено бележење семантичког ограничења. Колоном за опис (Description) је предвиђен краћи опис морфолошког обрасца. Колона за белешке (Note) предвиђа чување напомена о датом обрасцу. Уз сваки образац су доступна дугмад за чување измена у метаподацима, детаљнији преглед метаподатака и брисање обрасца.



Слика П2.28 Панел за приказ морфолошких образаца

Прилог 3. Списак успостављених релација са бројем правила, бројем успостављених веза и примерима

ИД рел. - DCValueRelID	Врста рел. - RelType	Ознака рел. - Label	Полазна вредност рел. -SourceValue	Крајња вредност рел. - DestinationVa	Бр. правила у рел.	Бр. веза	Примери повезивања
1	V1	varijanta	DER=ArisatiIrati	DER=IratiArisati	3	6	коментарисан - коментиран; коментарисање -
2	V1	varijanta	DER=AtiIrati	DER=IratiAti	3	5	измиксан - измиксиран; миксање - миксирање;
3	V1	varijanta	DER=AtiOvati	DER=OvatiAti	3	34	шпикан - шпикован; шлепање - шлеповање;
4	V1	varijanta	DER=AvatiIvati	DER=IvatiAvati	3	678	осмишљаван - осмишљиван; укисељавање - укисељивање;
5	V1	varijanta	DER=AvatiUvati	DER=UvatiAvati	3	3	нађинђаван - нађинђуван; нађинђавање - нађинђување;
6	V1	varijanta	DER=CijskiTorski	DER=TorskiCijski	1	1	стабилизацијски - стабилизаторски
7	V1	varijanta	DER=CiratiKovati	DER=KovatiCirati	3	115	едуциран - едукован; мистифицирање -

8	V1	varijanta	DER=CxatiTati	DER=TatiCxati	3	3	схваћан - схватан; схваћање - схватање;
9	V1	varijanta	DER=CxivatiTavati	DER=TavatiCxivati	3	27	освешћиван - освештаван; раскршћивање -
10	V1	varijanta	DER=ErisatiIrati	DER=IratiErisati	3	2	хохштаплирати - хохштаплерисати;
11	V1	varijanta	DER=IratiOvati	DER=OvatiIrati	6	949	ауторизиран - ауторизован; акредитирати -
12	V1	varijanta	DER=IsatiOvati	DER=OvatiIsati	3	35	асимилисан - асимилован; оксидисање - оксидовање;
13	V1	varijanta	DER=KovatiZirati	DER=ZiratiKovati	3	3	критикован - критизиран; критиковање -
14	V1	varijanta	DER=RatiSati	DER=SatiRati	4	440	афирмиран - афирмисан; карамелирање -
15	V1	varijanta	DER=SatiTi	DER=TiSati	3	2	капарисати - капарити; сумпорисати - сумпорити

16	V1	varijanta	DER=SatiZirati	DER=ZiratiSati	4	94	симпатисан - симпатизиран; идеалисање - идеализирање
17	V1	varijanta	DER=SatiZovati	DER=ZovatiSati	3	66	популарисан - популаризован; идеалисање -
18	V1	varijanta	DER=SavatiSxavati	DER=SxavatiSavati	4	5	спасаван - спашаван; спасавајући - спашавајући;
19	V1	varijanta	DER=SivatiSxavati	DER=SxavatiSivati	3	6	надвисиван - надвишаван; надвисивање - надвишавање;
20	V1	varijanta	DER=SivatiSxivati	DER=SxivatiSivati	3	3	надвисиван - надвишиван; надвисивање - надвишивање
21	V1	varijanta	DER=ZivatiZxavati	DER=ZxavatiZivati	3	3	унизиван - унижаван; унизивање - унижавање;
22	V2	varijanta sa/bez h	DER=H0	DER=0H	2	326	хајдук - ајдук; хајвар - ајвар; хајмо - ајмо
23	V2	varijanta sa/bez i	DER=I0	DER=0I	1	1	италијаштина - талијаштина

24	V2	varijanta	DER=BV	DER=VB	1	23	барбарин - варварин; симболичан - символичан;
25	V2	varijanta	DER=CK	DER=KC	1	13	оцеан - океан; центаур - кентаур; острацизам -
26	V2	varijanta	DER=CS	DER=SC	1	6	цертификат - сертификат; суфинанциран -
27	V2	varijanta	DER=FV	DER=VF	1	91	салфета - салвета; куглоф - куглов;
28	V2	varijanta	DER=GH	DER=HG	1	1	астаган - астрахан
29	V2	varijanta	DER=GK	DER=KG	1	11	регрутовати - рекрутовати; туфегџија - туфекција;
30	V2	varijanta	DER=HJ	DER=JH	1	40	чоха - чоја; снахин - снајин; смех - смеј
31	V2	varijanta	DER=HK	DER=KH	1	128	хлор - клор; геохронологија - геокронологиј

32	V2	varijanta	DER=HV	DER=VH	1	151	кухар - кувар; ухо - уво; буха - бува
33	V2	varijanta	DER=IJ	DER=JI	1	9	ионски - јонски; ионизација - јонизација;
34	V2	varijanta	DER=JV	DER=VJ	1	2	проја - прова; силај - силав
35	V2	varijanta	DER=SSx	DER=SxS	1	25	пастицада - паштицада; стипендија - штипендија;
36	V2	varijanta	DER=SZ	DER=ZS	1	73	десерт - дезерт; циркусант - циркузант;
37	V2	varijanta	DER=ZC	DER=CZ	1	2	мезосопран - мецосопран; ензим - енцим
38	V2	varijanta	DER=VU	DER=UV	1	4	евнух - еунух; евро - еуро; евхаристија - еухаристија
39	V2	varijanta	DER=V0	DER=0V	1	5	човек - чоек; зевнути - зенути; паворски -

40	V2	varijanta	DER=SxSh	DER=ShSx	1	3	шематичан - схематичан; шизофренична -
41	V2	varijanta	DER=SxCx	DER=CxSx	1	1	зашуткан - заћуткан
42	V2	varijanta	DER=HvF	DER=FHv	1	5	хвала - фала; ухватити - уфатити; хвалити -
43	V2	varijanta	DER=HC	DER=CH	1	1	острахизам - острацизам
44	D	glagolska imenica	Imperf	VN	18	7391	блистати - блистање; бистрити - бистрење;
45	D	glagolska imenica	Perf	VN	10	1842	излечити - излечење; одоцнети - одоцњење;
48	D	prisvojni pridev - prezime	Last	Pos	24	17280	Андрић - Андрићев; Тинторето - Тинторетов;
49	D	prisvojni pridev - ime	First	Pos	1	51	Мијаило - Мијаилов; Вујо - Вујов; Јеротије -

50	D	mocija roda prezime	Last	GM	2	19764	Петерсен - Петерсенка; Мргодић - Мргодићка;
51	D	mocija roda ime	First	GM	1	0	
52	D	relacioni pridev	nema	PosQ	37	10244	Нант - нантски; виолина - виолински;
53	D	prisvojni pridev	nema	Pos	18	12183	пријатељ - пријатељев; кедар - кедров; Хенри
54	D	poimeničeni pridev	nema	FlxAdj	2	0	женски - женска; запослен - запослен
59	D	mocija roda	nema	GM	8	1370	amater - amaterka; znanac - znanica; lutkar
60	D	deminutiv	nema	Dem	43	1786	брод - бродић; стопало - стопалце; полуга -
61	D	augmentativ	nema	Aug	10	193	ров - ровина; купус - купушчина; синак -

64	Dlf	trpni pridev	nema	PP	1	10585	екранизовати - екранизован; нацврцкати - нацврцкан;
65	Dlf	radni glagolski pridev	nema	APP	1	301	оживети - оживео; смекшати - смекшао;
66	Dlf	prilog vremena sadašnjeg	nema	PGA	1	251	лелујати - лелујајући; јездити - јездећи;
67	Dlf	superlativ	nema	Sup	1	28	лагано - најлаганије; добро - најбоље;
68	Dlf	komparativ	nema	Cmp	1	73	брзо - брже; касно - касније; природно -
69	Dlf	izveden iz prideva	nema	Adj	1	1686	први - прво; заљубљен - заљубљено; згранут -
71	El	elizija	nema	El	2	114	драг - предраг; мудар - премудар;
72	El	negacija	nema	Neg	4	1676	угрејан - неугрејан; сигурно - несигурно; ко

101	EkIjk	ekavski - ijekavski	Ek	Ijk	20	5456	тетреб - тетријеб; сњешко - снешко;
102	D	Svršeni-Nesvršeni	Perf	Imperf	20	6882	избушити - бушити; нашетати - шетати;
103	D	muški stanovnik	Top	Inh	32	498	Словенија - Словенац; Рим - Римљанин; Кина - Кинез
104	D	ženski stanovnik	Top	Inh	36	427	Александрија - Александријка; Париз -
105	D	gradivni pridev	nema	nema	3	74	бигар - бигровит; лапор - лапоровит;
106	D	imenica proces	nema	Process	1	33	фелдспат - фелдспатизација; серпентин -

Списак слика

Слика 1 Таксономија речника по Де Шриверу (De Schryver 2003) 11
Слика 2 Приказ дела речничког чланка за лему „*bonheur*“ у *Информатизованом Трезору француског језика* 16
Слика 3 Приказ профила речи „*Gold*“ из *Дигиталног речника немачког језика* у виду табеле (горе) и облака речи (доле) 19
Слика 4 Део речничког чланка из *Оксфордског речника енглеског језика* који описује нову реч „*babytoon*“ 20
Слика 5 Речнички чланак “ра́достно” из *Речника руског језика* 21
Слика 6 Приказ речничког чланка за реч „*zlato*“ на *Хрватском језичном порталу* 22
Слика 7 Приказ речничког чланка за реч „mineral“на порталу *Фран* 23
Слика 8 Приказ речничког чланка за реч „ђузели́че“на платформи *Расковник* 24
Слика 9 Приказ речничког чланка за реч „амур“ у *Викиречнику* за српски језик 25
Слика 10 Приказ упутница у склопу TLex-a (Tshwanedje 2020) 28
Слика 11 Дијаграм основног LMF пакета - класе из основног пакета (Francopoulo и George 2013) 45
Слика 12 Међусобне зависности пакета у LMF-у (Francopoulo и George 2013) 47
Слика 13 Модел пакета за морфолошке информације (Francopoulo и George 2013) 48
Слика 14 Модел пакета за машински читљиве речнике (Francopoulo и George 2013) 50
Слика 15 Проширење за морфолошке обрасце за потребе обраде природних језика (Francopoulo и George 2013) 52
Слика 16 Синтакса у обради природних језика (Francopoulo и George 2013) 55
Слика 17 Семантичко проширење за обраду природних језика (Francopoulo и George 2013) 58
Слика 18 Пакет за вишечлане лексичке изразе у обради природних језика (Francopoulo и George 2013) 61
Слика 19 Проширење за обележавање вишејезичности (Francopoulo и George 2013) 64
Слика 20 Модел основног модула *ontolex* („Lexicon Model for Ontologies“ 2016) 66
Слика 21 Илустрација употребе концепта Lexical Concept 70
Слика 22 Пример представљања синсета из *Ворднета за српски језик - уjак* 71
Слика 23 Модул за синтаксу и семантику („Lexicon Model for Ontologies“ 2016) 72
Слика 24 Модул за декомпозицију (*decomp*) („Lexicon Model for Ontologies“ 2016) 74
Слика 25 Модул за варијације и превод („Lexicon Model for Ontologies“ 2016) 76
Слика 26 Модул за лингвистичке метаподатке превод („Lexicon Model for Ontologies“ 2016) 79
Слика 27 Модел лексикографске базе која складишти информације о категоријама 83
Слика 28 Приказ модела лексикографске базе са примером из речника монолексемских речи – DELAS 85
Слика 29 Приказ модела лексикографске базе са примером из речника полилексемских израза – DELAC 87
Слика 30 Приказ модела лексикографске базе речника DELAF 88
Слика 31 Приказ смештаја информација *из Морфолошког речника за француски језик* у моделу лексикографске базе (Lazić и Škorić 2019) 89
Слика 32 Приказ модела за успостављање релација 103
Слика 33 Линије конкорданци за упит „ANPrepNp“ у оквиру лексичког записа *глина* 109
Слика 34 Речнички чланак *глина* са приказом повезаних чланака из других речника 115
Слика 35 Приказ резултата упита *глина* у BabelNet-у 117
Слика 36 Приказ лексичког записа „koren“ кроз панел „Full Table“ 121
Слика 37 Архитектура система за проширење упита намењеног хибридној претрази корпуса коришћењем сервиса *Вебран* (Stanković и Utvić 2019) 125
Слика 38 Панел за евалуацију полилексемских израза 129
Слика 39 Линије конкорданци за лему *сив* 138
Слика 40 Линије конкорданци за лему *рђаст* 138
Слика 41 Линије конкорданци за лему *млечнобео* 138
Слика 42 Линије конкорданци за лему *окераст* 138

Слика П2.1 Резултат претраге за реч „лопта“ у јавном приказу vi
Слика П2.2 Резултат претраге за реч „лопта“ у јавном приказу – прецизан упит vii
Слика П2.3 Јавни приказ лексичког записа „lopta“ viii
Слика П2.4 Контролна табла за управљање категоријама података ix
Слика П2.5 Приказ лексичких записа означених вредношћу категорије података „DOM=Agro“ x
Слика П2.6 Панел за додавање и уређивање вредности категорија података xi
Слика П2.7 Табеларни приказ категорија података и њихових вредности xii
Слика П2.8 Приказ опције за филтрирање вредности категорија података xiii

Слика П2.9 Панел за уређивање и прегледање речника xiv
Слика П2.10 Прозор за приказ података о датотеци и додавање нове датотеке xiv
Слика П2.11 Панел за прелиставање и претрагу лексичких записа „Lexical Entries“ xv
Слика П2.12 Лексички запис „*lopta*“ у корисничком сучељу за управљање и развој xvii
Слика П2.13 Панел са приказом свих флективних облика леме xviii
Слика П2.14 Приказ панела за уређивање лексичког записа на примеру уређивања монолексемске речи xix
Слика П2.15 Панел за одабир и додавање категорија xx
Слика П2.16 Панел за одабир релације xx
Слика П2.17 Одабир леме записа за повезивање xx
Слика П2.18 Приказ панела за уређивање лексичког записа на примеру уређивања полилексемског израза xxii
Слика П2.19 Панел за уређивање/додавање компоненти полилексемском изразу xxiii
Слика П2.20 Панел за одабир облика компоненте xxiii
Слика П2.21 Панел за корпусе xxiv
Слика П2.22 Панел за уређивање метаподатака о корпусу xxv
Слика П2.23 Панел за евалуацију кандидата за Речник xxvi
Слика П2.24 Панел за преглед постојећих релација и управљање њима xxvii
Слика П2.25 Панел за управљање правилима за успостављање релације на примеру релације *женски становник* xxviii
Слика П2.26 Преглед повезаних лексичких записа xxix
Слика П2.27 Брисање правила xxx
Слика П2.28 Панел за приказ морфолошких образаца xxxi

Списак табела

Табела 1 Атрибути чија је употреба могућа у оквиру пакета за опис морфолошких информација (ISO/TC 37/SC 2008) 49
Табела 2 Атрибути пакета проширење за морфолошке обрасце за потребе обраде природних језика (ISO/TC 37/SC 2008) 54
Табела 3 Атрибути пакета проширења за синтаксу у обради природних језика (ISO/TC 37/SC 2008) 56
Табела 4 Атрибути пакета проширења за семантику у обради природних језика (ISO/TC 37/SC 2008) 59
Табела 5 Атрибути пакета за вишечлане лексичке изразе у обради природних језика (ISO/TC 37/SC 2008) 62
Табела 6 Мапирање ознака из MPC са ознакама за врсту речи из UD 94
Табела 7 Проширени тагови коришћени у корпусу *ГеоСрпКор* 107
Табела 8 Извод из конкорданци које одговарају упиту `[tag="A.*"] [lemma="peskovi"]` добијених на корпусу *ГеоСрпКор* 108
Табела 9 Фреквенције појављивања облика речи *шкриљац* 113
Табела 10 Матрица сагласности евалуатора 1 и 2 (полилексемски израз) 130
Табела 11 Ниво сагласности евалуатора на основу коефицијента *Коенова капа* (k) 131
Табела 12 Матрица сагласности евалуатора 1 и 2 (лема) 131
Табела 13 Матрица сагласности евалуатора 1 и 2 (термин у геолошком домену) 132
Табела 14 Матрица сагласности евалуатора 1 и 2 (термин у домену који није геологија) 133
Табела 15 Леме боја у корпусу *ГеоСрпКор* 137

Биографија аутора

Биљана Рујевић (Лазић) је рођена 30. јуна 1988. године у Београду. Дипломирала је на Катедри за библиотекарство и информатику Филолошког факултета Универзитета у Београду 2011. године. Мастер студијски програм Језик, књижевност, култура на истом факултету завршила је 2012. године одбранивши мастер рад на тему *Архиви медијских кућа у свету*. Докторске академске студије на Филолошком факултету Универзитета у Београду, модул Култура уписала је 2013. године.

Започела је радно искуство библиотекара 2011. године у Музеју науке и технике. Потом је била ангажована као библиотекар у Природњачком музеју у Београду и Музеју Николе Тесле. Од децембра 2013. године запослена је као библиотекар у Централној библиотеци Рударско-геолошког факултета Универзитета у Београду.

У периоду од 2014. до 2017. била је учесник Tempus пројекта *BAEKTEL - Blending academic and entrepreneurial knowledge in technology enhanced learning*. Од 2017. до 2018. била је учесник билатералног пројекта са Републиком Словачком - *Кванититативна анализа слогова у словенским језицима (руски, словачки, српски)*. Од 2020. године је учесник билатералног пројекта са Републиком Немачком - *Међујезичко препознавање говора мржње*.

Добитник је једне од шест стипендија *Alliance of Digital Humanities Organizations (ADHO)* за интернационалног студента учесника у организацији годишње конференције *Digital Humanities 2012*, одржане у јулу 2012. године у Хамбургу у Немачкој.

Полазник је Треће европске летње школе из области дигиталних хуманистичких наука *Култура и технологија*, одржане 2012. године на Универзитету у Лајпцигу, *Осме руске летње школе из проналажења информација* RuSSIR 2014, одржане 2014. године у Нижњем Новгороду и *Lexical Data Masterclass* одржаног 2018. године у Берлину.

Стручни испит у библиотечко-информационој делатности положила је јуна 2013. године.

Секретар је Друштва за језичке ресурсе и технологије - JePTex. Ради на развоју језичких ресурса за области рударства и геологије. Члан је Библиотекарског друштва Србије.

Члан је Комисије за стандарде *A037 – Терминологија* Института за стандардизацију Србије.

Изјава о ауторству

Име и презиме аутора Биљана Рујевић

Број индекса 13088д

Изјављујем

да је докторска дисертација под насловом

Речници у дигиталном добу – информатичка подршка за српски језик

- 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, 25.02.2022.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Биљана Рујевић

Број индекса 13088д

Студијски програм Модул Култура

Наслов рада Речници у дигиталном добу – информатичка подршка за српски језик

Ментор Проф. др Цветана Крстев, редовни професор

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, 25.02.2022.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку "Светозар Марковић" да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Речници у дигиталном добу – информатичка подршка за српски језик

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, 15.02.2022.

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.