

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ
<p>1. Датум и орган који је именовео комисију 20. IV 2022. Наставно-научно веће Филолошког факултета</p> <p>2. Састав комисије са знаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и називом факултета, установе у којој је члан комисије запослен:</p> <ol style="list-style-type: none">1. др Цветана Крстев, редовни професор, библиотечка информатика, 20. V 2014, Филолошки факултет Универзитета у Београду2. др Милош Утвић, доцент, библиотечка информатика, 17. XII 2019, Филолошки факултет Универзитета у Београду3. др Јован Чудомировић, доцент, српски језик, 6. VII 2021, Филолошки факултет Универзитета у Београду4. др Ранка Станковић, ванредни професор, математика и информатика, 8. VI 2020, Рударско-геолошки факултет Универзитета у Београду5. др Александра Марковић, научни сарадник, језик и књижевност – филологија, 31. I 2021, Институт за српски језик САНУ
II ПОДАЦИ О КАНДИДАТУ
<ol style="list-style-type: none">1. Име, име једног родитеља, презиме: Биљана (Ђорђе) Рујевић (девојачко Лазић)2. Датум рођења, општина, република: 30. VI 1988. Београд, Србија3. Датум одбране, место и назив мастер рада: 9. X 2012. Београд. „Архиви медијских кућа у свету“4. Научна област из које је стечено академско звање мастера: Филолошке науке, мастер професор језика и књижевности
III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ: Речници у дигиталном добу – информатичка подршка за српски језик
IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ: Навести кратак садржај са знаком броја страна поглавља, слика, шема, графика и сл.
<p>Докторска дисертација мастера Биљане Рујевић бави се развојем модела лексикографске базе података, као и израдом апликације за управљање лексичким ресурсима (речницима) заснованим на предложеном моделу и за њихов развој. Истраживање посебно обухвата:</p> <ul style="list-style-type: none">• преглед савремених, стандардизованих модела за развој речника, посебно за развој речника за потребе обраде природних језика;• развој модела лексикографске базе података заснованог на постојећим стандардима и формирање лексикографске базе електронских морфолошких речника српског језика засноване на овом моделу;• развој апликације за одржавање и развој електронских морфолошких речника српског језика смештених у лексикографску базу;

- евалуација остварених резултата на корпусу, лексици и терминологији из домена геологије.

У истраживању се користе методе моделирања података, као и експерименталне методе попут корпусних метода претраживања и екстракције информација. За процену квалитета остварених резултата коришћена је евалуација од стране два стручњака.

Дисертација обухвата 208 страна, а у оквиру тога 8 поглавља (140 страна), списак коришћене литературе (12 страна, 174 библиографске јединице), 3 прилога (40 страна), уводни и завршни материјал (насловне стране на српском, енглеском и руском, апстракт на српском и енглеском, списак чланова комисије, захвалност, садржај, списак слика, списак табела, биографија кандидата и три изјаве, 16 страна). У дисертацији укупно има 42 слике (и 28 у прилогу) и 15 табела. Поглавља дисертације су:

1. Увод (2 стране).
2. Поглед на традиционалне и електронске речнике (27 страна);
3. Језички ресурси за српски (7 страна);
4. Модел лексикографских података и развој модела базе *Лексимирка* (61 страна);
5. Формирање лексикографске базе српског језика у *Лексимирки* и успостављање унутрашњих и спољашњих веза (20 страна);
6. Постављање презентације на веб (9 страна);
7. Евалуација резултата на лексици (терминологији) из геолошког домена (12 страна);
8. Закључак (2 стране).

Прилози дисертације су:

1. Поређење категорија података из Морфолошког речника српског језика у односу на скраћенице из Речника САНУ и концепте из онтологије SUMO (5 страна);
2. Упутство за коришћење апликације *Лексимирка* (26 страна);
3. Списак успостављених релација са бројем правила, бројем успостављених веза и примерима (9 страна).

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дисертације кандидаткиња Биљана Рујевић представља предмет и циљ својих истраживања, даје њихов кратак преглед по поглављима и укратко приказује остварене резултате.

У другом поглављу – „Поглед на традиционалне и електронске речнике“ – кандидаткиња пре свега уводи појам *дигиталног речника* (или речника у дигиталном облику) и даје узорне примере првих таквих речника који су настали у другој половини XX века. Кандидаткиња потом уводи појмове *дигитализовани речник* и *речник настао у дигиталном окружењу*. Дигитализовани речници настају превођењем постојећих речника из традиционалног, папирног облика у дигитални облик, и они су најчешће и даље намењени људима. Речници настали у дигиталном окружењу могу такође бити намењени људима, али њихова првенствена намена могу бити и рачунарске апликације које се ослањају на обраду природних језика. Како ови речници најчешће настају применом метода електронске лексикографије, кандидаткиња укратко представља основне фазе њиховог развоја, а то су фаза припреме, фаза прикупљања извора (то јест, формирања корпуса), фаза аутоматизације (то јест, аутоматске језичке обраде корпуса), фаза обраде која подразумева одабир кандидата за речник, фаза аутоматске компилације прикупљених података и фаза припреме за објављивање (на пример, на мрежи). Имајући у виду специфичну форму дигиталних речника, њихова класификација се у великој мери разликује од класификације традиционалних речника, тако да се они разврставају у зависности од њиховог изворног облика, довршености, хипертекстуалности, интерактивности с корисницима, мултимедијалности и приступа садржају. У даљем тексту кандидаткиња даје преглед неких репрезентативних дигиталних речника доступних на вебу за француски, немачки, енглески, руски, хрватски, словеначки и српски језик и указује на информације које су корисницима доступне преко ових лексичких портала. Како израда нових речника у

дигиталном облику није могућа без коришћења лексикографских база података и система за писање речника, кандидаткиња пре свега уводи појам лексикографске базе података као колекције структурираних лексикографских података и указује на предности коришћења оваквих база у изради речника. Кандидаткиња потом представља два позната система за израду речника, TLex и Lexonotu, који се називају и лексикографске радне станице, и приказује могућности које они нуде корисницима.

У трећем поглављу – „Језички ресурси за српски“ – кандидаткиња Биљана Рујевић укратко представља три језичка ресурса на којима се заснива њен рад на овој дисертацији. Први од њих су Морфолошки речници српског језика (SrpMD) који се развијају већ дуги низ година и користе за интелигентно претраживање и друге сврхе у рачунарским апликацијама. Ови речници покривају општу лексику и властита имена, као монолексемске и полилексемске јединице. Други ресурс је семантичка мрежа ворднет за српски језик, који се такође развија већ дуго времена, а користи се за семантичко проширивање упита и друге примене у апликацијама које се заснивају на обради српског језика. Трећи ресурс су корпуси српског језика, од којих су за кандидаткињин рад најзначајнији. Корпус савременог српског језика SrpKor2013, као *de facto* референтни корпус српског језика, као и доменски корпус из области рударства, који је кандидаткињи послужио за проверу остварења постављених циљева и евалуацију добијених резултата.

У четвртном поглављу – „Модел лексикографских података и развој модела лексикографске базе *Лексимирка*“ – кандидаткиња представља теоријске основе свога рада. Оно је посвећено формалним моделима лексикографских података. Конкретно, у првом делу овог поглавља кандидаткиња детаљно представља три најшире прихваћена стандардизована модела. Први од њих је развијен као модул *Речници* (Dictionaries) у оквиру смерница за кодирање текста TEI. Кандидаткиња описује како се у овом моделу представљају макро- и микро-структура речника, а затим како се, користећи XML као формат, кодира речник као целина, а како речнички чланци и који се све елементи могу користити за описивање делова речничких чланака. Други представљени модел је *Оквир за лексичко обележавање*, *LMF*, који се користи да би се стандардизовали лексички подаци и тиме омогућило да се једном развијени лексички ресурс може примењивати у разне сврхе, било за развој машински читљивих речника било за потребе апликација за обраду природних језика. Модел *LMF* поред основног пакета садржи низ додатних пакета који су прилагођени специфичним потребама, нпр. машинскичитљивим речницима или специфичним подацима, какви су полилексемске јединице у апликацијама обраде природног језика. Кандидаткиња приказује основни и додатне пакете модела *LMF* користећи UML дијаграме, а за сваки од њих даје примере употребе у српском језику користећи XML формат или поједностављене UML записе (налик формату JSON). Као трећи, кандидаткиња представља модел *lemon* који спаја онтологије са лингвистичким и лексичким ресурсима и заснива се на многим постојећим стандардима за онтологије на вебу и лексичке ресурсе, укључујући и *LMF*. И овај модел се састоји од основног модула *ontolex* и више придружених модула за специфичне потребе. Све ове модуле кандидаткиња представља дијаграмима и даје илустративне примере за њих на српском језику у виду поједностављене RDF синтаксе. На крају овог одељка кандидаткиња пореди могућности ова три модела, даје примере њихове употребе у пракси и указује на домене у којима сваки од њих даје најбоље резултате.

У другом делу истог поглавља кандидаткиња представља модел лексикографске базе података, развијен у оквиру рада на дисертацији, који се ослања на претходно описане моделе *LMF* и *lemon*, као и на међународни стандард за Регистар категорија података (DCR). Развијени модел података подржава складиштење лексичких јединица из Морфолошких електронских речника српског језика, и то из речника монолексемских и полилексемских лема, као и речника облика. Осим тога, модел омогућава раздвајање и описивање значења, повезивање полилексемских јединица са саставним деловима, као и повезивање лексичких уноса по разним критеријумима. Да би се приказала општост модела, кандидаткиња приказује како би се у истом моделу приказали подаци из Морфолошког електронског речника француског језика. Како је контрола унетих података један од циљева изградње и увођења лексикографске базе података, кандидаткиња је при конципирању базе посебно водила рачуна о развоју модела

ознака које се користе у Морфолошким електронским речницима српског, као и њиховом повезивању са сличним ознакама које се користе у Речнику САНУ и вршној онтологији SUMO. Успостављање ових веза олакшава контролу података и омогућава потенцијално преливање података из сличних ресурса.

У петом поглављу – „Формирање лексикографске базе српског језика у *Лексимирки* и успостављање унутрашњих и спољашњих веза“ – кандидаткиња описује реализацију лексикографске базе *Лексимирка* засноване на моделу описаном у претходном поглављу. Посебну пажњу кандидаткиња посвећује успостављању веза између лексичких уноса у бази, пре свега повезивању одредница из електронског речника српског језика који је базом подржан. Пре формирања лексикографске базе одреднице у овом речнику нису биле повезане, премда су многе биле обележене маркерима који указују на постојање одређених веза између одредница различитог изговора, варијантних облика или одредница повезаних процесом деривације. У новој лексикографској бази ове одреднице су аутоматски повезане, а добијени резултати су евалуирани на узорку. Осим тога, одреднице електронског речника су повезане са синсетовима Српског ворднета, чиме је омогућено њихово узајамно богаћење и унапређивање. Лексикографска база *Лексимирка* омогућава, осим тога, повезивање и са спољашњим ресурсима, пре свега са корпусима српског језика, омогућавајући њихову напредну претрагу, чиме се знатно олакшава одржавање електронских речника и унапређивање њиховог садржаја. Коначно, уноси у лексикографској бази *Лексимирка* су преко истоимене апликације повезани и са спољашњим базама података: узорцима речника или комплетним речницима, уколико су они слободно доступни, као и ресурсима на вебу, од којих је најзначајнији Викиречник, а потом и са вишејезичним речницима и семантичким мрежама BabelNet, Glosbe и Termi, што све олакшава развој електронских речника и његово коришћење, како у хуманом тако и машинском окружењу.

У шестом поглављу – „Постављање презентације на веб“ – које је логични наставак претходног поглавља, кандидаткиња укратко представља могућности веб апликације *Лексимирка*, која подржава и омогућава приступ описаној лексикографској бази. Кандидаткиња представља корисничко сучеље (интерфејс), привилегије за приступ бази, могућности њеног претраживања, као и формате у којима се садржај базе или неки њен део може извести ради коришћења у другим апликацијама. На крају кандидаткиња представља неке веб-сервисе који се заснивају на лексичким ресурсима из базе *Лексимирка*. Више детаља о начину коришћења апликације *Лексимирка* кандидаткиња је дала у Прилогу 2 дисертације.

У седмом поглављу – „Евалуација резултата на лексици (терминологији) из геолошког домена“ – кандидаткиња Биљана Рујевић демонстрира применљивост лексикографске базе и апликације *Лексимирка* на обради текстова и лексике из геолошког домена. Кандидаткиња је прво припремила скуп текстова из овог домена, од њих оформила корпус, који је потом обрадила у *Лексимирки* морфолошким речницима записаним у њеној бази, што је омогућило прво допуну речника недостајућом (моноксемском) лексиком, а потом и екстракцију полилексемских термина. Развијеним моделом за повезивање лексичких уноса новододати уноси су повезани на одговарајући начин. С обзиром на то да се у текстовима из геолошког домена користи велики број придева за означавање боја, кандидаткиња илуструје примену морфолошких речника на припремљени корпус излистивањем конкорданци оваквих придева и формирањем фреквенцијске листе.

У завршном поглављу – „Закључак“ – кандидаткиња резимира доприносе ове дисертације и даје смернице за будући рад. Допринос дисертације је развој модела лексикографске базе и софтверског решења апликације *Лексимирка* са бројним функцијама за одржавање и претраживање речника међусобно повезаних моноксемских и полилексемских јединица, његово повезивање са другим ресурсима и издавање произвољног скупа јединица у различитим форматима. Будући рад предвиђа даље усавршавање апликације *Лексимирка* додавањем нових функција, као што су аутоматски избор добрих примера из корпуса, генерисање речничких дефиниција, израда профила речи и сл.

На крају дисертације мастера Биљане Рујевић налазе се 3 прилога:

1. У прилогу 1 дато је поређење категорија података (маркера) који се користе у Морфолошким речницима српског језика са ознакама, односно, скраћеницама из

Речника САНУ, као и са концептима из вршне онтологије SUMO, тамо где је то применљиво.

- У прилогу 2 дато је детаљно упутство за коришћење апликације *Лексимирка* проласком кроз све његове меније и табеле.
- У прилогу 3 дат је списак свих успостављених релација између лексичких јединица у бази *Лексимирка*, са бројем примењених правила и уз навођење илустративних примера.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

Обрадовић, Иван, Александра Томашевић, Ранка Станковић, и Биљана Лазих.

2017. „Увођење доменских и семантичких маркера за област рударства у српске електронске речнике“. У: Научни састанак слависта у Вукове дане, уредили Рајна Драгићевић и Александар Милановић, 46/3: 147–58. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду.

<https://doi.org/10.18485/msc.2017.46.3.ch10>.

Утвић, Милош, Ранка Станковић, Александра Томашевић, Михаило Шкорић, и

Биљана Лазих. 2019. „Претрага корпуса заснована на употреби екстерних лексичких ресурса путем веб-сервиса“. У: Научни састанак слависта у Вукове дане, уредили Божо Ђорић 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>.

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>.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је кандидаткиња мастер Биљана Рујевић остварила циљеве зацртане у пријави дисертације. Кандидаткиња је дала преглед савремених, стандардизованих модела речника за потребе обраде природних језика, развила модел заснован на постојећим стандардима и на основу њега формирала лексикографску базу електронских морфолошких речника српског језика, што је омогућило развој апликације за њихово одржавање и коришћење, коју је кандидаткиња користила за евалуацију остварених резултата на корпусу, лексички и терминологији из домена геологије.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је мастер Биљана Рујевић користила релевантну и савремену литературу, те да је постављене проблеме обрадила детаљно и сагледавајући их из разних углова. Овим радом мастер Биљана Рујевић је изградила модел и лексикографску базу података који обухватају речнике за обраду српског језика (а могу садржати речнике за потребе обраде и других природних језика), и израдила концепт и омогућила развој лексикографске апликације за креирање, одржавање, унапређивање и коришћење ових речника, који тиме постају доступни ширем кругу истраживача.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА

Комисија сматра да је кандидаткиња Биљана Рујевић у својој дисертацији *Речници у дигиталном добу – информатичка подршка за српски језик* успешно обрадила значајну тему коришћењем савремених метода, да је текст дисертације урађен према одобреној пријави дисертације и да је реч о раду који представља оригинално и самостално научно дело.

IX ПРЕДЛОГ:

На основу укупне оцене дисертације, Комисија предлаже Наставно-научном већу Филолошког факултета Универзитета у Београду да прихвати извештај о дисертацији *Речници у дигиталном добу – информатичка подршка за српски језик* кандидаткиње мастера Биљане Рујевић и упути га Већу научних области друштвено-хуманистичких наука Универзитета у Београду, како би кандидаткиња била позвана на усмену одбрану рада.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

1. др Цветана Крстев, редовни професор
Филолошки факултет, Универзитет у Београду
2. др Милош Утвић, доцент
Филолошки факултет, Универзитет у Београду
3. др Јован Чудомировић, доцент
Филолошки факултет, Универзитет у Београду
4. др Ранка Станковић, ванредни професор
Рударско-геолошки факултет, Универзитет у Београду
5. др Александра Марковић, научни сарадник
Институт за српски језик, САНУ