УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЛОШКИ ФАКУЛТЕТ

УПУТСТВО ЗА ПИСАЊЕ ИЗВЕШТАЈА О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ
1. Датум и орган који је именовао комисију 24. I 2018. Научно-наставно веће Филолошког факултета 2. Састав комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен: 1. др Цветана Крстев, редовни професор, библиотечка информатика, 20. V 2014, Филолошки факултет Универзитета у Београду 2. др Александра Вранеш, редовни професор, библиотекарство и информатика, 14. XII 2004, Филолошки факултет Универзитета у Београду 3. др Ранка Станковић, ванредни професор, математика и информатика, 19.V 2015, Рударско-геолошки факултет Универзитета у Београду 4. 5.
II ПОДАЦИ О КАНДИДАТУ
1. Име, име једног родитеља, презиме: Јелена Д. Митровић 2. Датум рођења, општина, република: 25. XI 1980. Ужице, Србија 3. Датум одбране, место и назив мастер рада: 25. XI 2010. Београд. Пројекат „Гугл књиге“ 4. Научна област из које је стечено академско звање мастера: Библиотекарство и информатика
III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада
IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ: **Навести кратак садржај са назнаком броја страна поглавља, слика, шема, графикона и сл.**
Докторска дисертација се бави лексичким ресурсима за српски језик намењеним апликацијама за обраду природних језика и њиховом надоградњом коришћењем метода групне расподеле рада. Истраживање посебно обухвата: • преглед постојећих лексичких ресурса за српски језик и анализу могућности њихове надоградње коришћењем групне расподеле рада; • анализу постојећих приступа групној расподели рада и њихове применљивости на развој лексичких ресурса; • решавање једног проблема допуне Српског ворднета коришћењем за ту намену припремљеног система групне расподеле рада; • коришћење допуњеног Српског ворднета за решавање конкретног задатка

препознавања реторичких фигура ироније и сарказма у текстовима на српском језику. У истраживању се користе експерименталне методе, као што су корпусне методе претраживања и методе групне расподеле рада. За процену квалитета остварених резултата користе се статистичке методе (Студентов t-тест, Крипендорфов алфа коефицијент).

Дисертација обухвата 220 страна, а у оквиру тога 6 поглавља (142 стране), списак коришћене литературе (17 страна, 213 библиографских јединица), 6 прилога (25 страна), уводни и завршни материјал (насловне стране, апстракт на српском, енглеском и руском, садржај, списак табела, списак слика, биографија кандидата, 26 страна). У дисертацији укупно има 34 слика и 39 табела. Поглавља дисертације су:

1. Увод (4 стране).
2. Језички ресурси и алати у обради природног језика (58 страна).
3. Групна расподела рада (43 стране).
4. Нове семантичке релације у Српском ворднету на основу реторичке фигуре поређење (23 стране).
5. Примена унапређеног Српског ворднета (11 страна).
6. Закључак и будући рад (3 стране).

Прилози дисертације су:

А. Литература (17 страна).
Б. Прилози (6 прилога, 25 страна)

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дисертације кандидаткиња Јелена Митровић смешта своје истраживање на међу теоријске области рачунарске лингвистике као гране лингвистике и примењене области обраде природних језика настале у оквиру рачунарства, стављајући већи нагласак на примену. У овом уводном поглављу кандидаткиња поставља основне *циљеве* свог истраживања, а то је, пре свега, развој језичких ресурса и алата за српски језик неопходних за решавање многих задатака који се заснивају на обради природних језика, *методе* рада које се превасходно заснивају на моделу групне расподеле рада и *резултате* свог истраживања, а то су систем за примену модела групне расподеле рада на прикупљање језичких података и унапређени језички ресурси за српски језик.

У другом поглављу „Језички ресурси и алати у обради природног језика“, кандидаткиња представља језичке ресурсе и базичне алате који се користе за обраду природних језика. Кроз историјски преглед развоја обраде природних језика кандидаткиња указује на велике међународне пројекте и асоцијације у оквиру којих су развијани неки од најзначајнијих језичких ресурса и алата. Посебну пажњу кандидаткиња посвећује оним међународним пројектима у оквиру којих су настали неки од најважнијих ресурса за обраду српског језика (TELRI, BalkaNet, CESAR и Parseme – COST Action као најзначајнији). Кандидаткиња даље представља два важна ресурса за српски језик које је користила у раду на изради докторске дисертације. Корпус савременог српског језика (СрпКор), чија је прва верзија постала доступна на вебу 2003. године, а нова, анотирана верзија 2013. године, представља веома значајан ресурс за разне врсте лингвистичких и филолошких истраживања. Солидност резултата добијених претраживањем СрпКор-а обезбеђена је разноврсношћу текстова у њему (новински, административни, књижевно-уметнички, научни и научно-популарни), премда пуна репрезентативност корпуса још увек није достигнута. Други важан ресурс су електронски морфолошки речници српског језика (е-СМР) који су намењени обради текстова на српском језику. Ови речници су настали по узору на сличне речнике за француски језик које је осмислио и у највећој мери изградио француски лингвиста проф. Морис Грос. Електронски морфолошки речници српског језика који садрже како монолексемске тако и полилексемске јединице (термини настали по узору на устаљене термине из области обраде природних језика на енглеском, *simple words* и *multi-word units*) имају велику покривеност, али је због сталног

развоја језика и ове речнике потребно одржавати и допуњавати. Кандидаткиња затим представља ресурсе чијој изради и унапређивању је допринела радом на овој докторској дисертацији. Формална, или информатичка, онтологија је онтологија дата на неком формалном језику, а њена основна улога је дељење и вишеструка употреба знања од стране различитих интелигентних агената и апликација. Онтологије представљају основу семантичког веба који подразумева да се на веб постављају документа чије је значење описано на начин који рачунарске апликације разумеју те могу корисницима да пруже квалитетније одговоре на постављене упите. Кандидаткиња са више детаља представља формалну онтологију реторичких фигура за српски језик *РетФиг* чијој изради је допринела у току рада на овој докторској дисертацији. Последњи део овог поглавља кандидаткиња посвећује Ворднету, информатичкој лексичко-семантичкој мрежи, која представља лингвистички ресурс који је нашао вишеструку примену у обради природног језика и као такав је постао *de facto* стандард у тој области. Кандидаткиња пре свега представља изворни Ворднет, данас познат као Принстонски ворднет, чији развој је 1985. године отпочела група лингвиста и психолингвиста на челу са проф. Џорџом Милером. Ворднет се састоји из синсетова, то јест скупова речи синонимног значења, који су међусобно повезани основним семантичким релацијама као што су хипонимија и меронимија, али и многим другим чије карактеристике кандидаткиња детаљно описује. Као важан информатички ресурс Ворднет је проширен многим додатним структурама, а једна од првих је доменска хијерархија која се састоји од 164 хијерархијски организованих ознака домена. Друго значајно проширење Ворднета јесте његово повезивање са SUMO онтологијом, која представља формалну онтологију највишег нивоа. Под утицајем, и на основу, Принстонског ворднета настали су ворднетови многих других језика, између осталог и Српски ворднет. Развој Српског ворднета отпочео је у оквиру европског пројекта Балканет, а настављен по окончању пројекта у оквиру разних волонтерских активности. Српски ворднет тренутно (децембар 2017, датум окончања текста докторске дисертације) има приближно 22.500 синсетова и још увек по обухвату значајно заостаје за Принстонским ворднетом, посебно за неке врсте речи, као што су придеви и глаголи. Кандидаткиња, с тога, у наредним поглављима представља методологију за унапређење рада на Српском ворднету чија примена је и резултовала додавањем више од 300 придевских синсетова. И Српски ворднет је обогаћен новим структурама, као што су доменска хијерархија и SUMO онтологија, потом SentiWordNet, ресурс који сваком синсету додељује три ознаке интензитета и поларитета осећања – позитиван, негативан и објективан став или осећање, а остварена је и веза са српским морфолошким е-речницима. На крају овог поглавља кандидаткиња представља и основне алате у обради природног језика, посебно оне који се користе за развој српских ресурса, те оних који користе те исте ресурсе у свом раду. Посебну пажњу кандидаткиња посвећује веб апликацији Serbian WordNet Editing (SWNE) који се користи за изградњу, одржавање и унапређивање Српског ворднета, а у чијем конципирању је кандидаткиња активно учествовала.

Најважнији резултати истраживања изложени су у поглављима 3, 4 и 5. У трећем поглављу „Групна расподела рада“ кандидаткиња представља један нови метод обављања задатака и управљања који се све више примењује у науци и култури (енглески термин је *crowdsourcing*). Овај метод се заснива на чињеници да су неки задаци још увек сложени (или недостижни) за рачунар али једноставни за људе, посебно ако их обавља велики број људи. Ова група људи се, с тога, некад назива *људска процесорска јединица* (енг. Human Processing Unit по угледу на Computer Processing Unit). Кандидаткиња истиче да је овакав начин рада постојао и у прошлости те указује на неке изузетне примере (на пример, рад на Оксфордском речнику енглеског језика или рад на изради логаритамских таблица) али је добио нови замах и потенцијал развојем интернета. Кандидаткиња потом истражује који мотиви покрећу људе да се укључују у пројекте групне расподеле рада и истиче да они могу бити: лични – жеља за учествовањем у пројекту и стицањем нових знања, друштвени – остваривање престижа, рецимо у односу на друге учеснике у пројекту и финансијска. Основна подела метода групне расподеле рада односи се на врсту задатака које учесници обављају, а то могу бити микрозадаци и макрозадаци. У случају микрозадатака сваки учесник у пројекту треба да уради неки мали, добро дефинисан задатак, а на крају се сва решења „склапају“ у коначно решење до кога треба доћи у датом пројекту. Макрозадаци се односе на пружање читавог задатка или проблема на

увид групи људи, како би свако могао да одреди који и колики део ће решити и како ће пројекту допринети у складу са својим знањима и компетенцијама. У оквиру методе групне расподеле рада развијено је више жанрова, а најпознатији су групна мудрост, механизовани рад и игре са сврхом. Групна мудрост односи се на појаву да у неком пројекту група која се састоји од великог броја учесника може бити много успешнија од неколико стручњака, а као успешан пример овакве расподеле рада се често истиче Википедија. Механизовани рад односи се на решавање задатака који обично захтевају веома мало времена и напора, а за шта „радници“ добијају неку малу новчану надокнаду. Најпознатија платформа за овако организовану групну расподелу рада је Amazon Mechanical Turk. Игре са сврхом се разликују од осталих видова групне расподеле рада пре свега по мотивацији учесника, а то је исти мотив који их подстиче да играју игрице – жеља за забавом. Овако организована групна расподела рада разликује се од уобичајених игрица највише по томе што играјући их учесници, мање или више неприметно, решавају и неке задатке. Групна расподела рада је коришћена у низу друштвено корисних задатака, те кандидаткиња истиче неке од најубедљивијих примера, као што је пројекат исправљања грешака насталих оптичким препознавањем карактера у старим новинским текстовима који је водила Народна библиотека Финске и пројекат Duolingo учења језика кроз превођење. Групна расподела рада, а посебно игре са сврхом, је коришћена и у многим пројектима везаним за обраду природних језика за обављање задатака као што су анотација, валидација или евалуација лингвистичких података (нпр. анотација корпуса, од обележавања врсте речи до обележавања њиховог значења, парафразирање делова реченица, успостављање веза између речи, обележавање поларитета осећања израженог неким речима и сл.). За све пројекте групне расподеле рада, па и оне везане за обраду природних језика, значајно је да се успостави систем провере квалитета обављеног посла. Када су у питању микрозадаци, често исте задатке решава већи број учесника, па се као исправно решење узима оно које су понудили сви (или већина) учесника. Друга могућност је постављање „златних задатака“ на које су одговори унапред познати, па се на основу успешности решавања тих задатака може проценити општа успешност учесника. Осим тога, поузданост одговора или решења учесника може се мерити разним статистичким методама. Кандидаткиња представља неке од највише коришћених статистичких показатеља, задржавајући се највише на такозваном Крипендорфовом алфа коефицијенту који је и сама користила за процену поузданости одговора које је понудило више учесника.

Кандидаткиња Јелена Митровић је посветила четврто поглавље „Нове семантичке релације у Српском ворднету на основу реторичке фигуре поређење“ пројекту групне расподеле рада који је она осмислила, спровела и оценила. Циљ овог пројекта је увођење у Српски ворднет нове релације специфичанЗа/спецификованСа (SpecificOf/SpecifiedBy) која повезује именичке и придевске синсетове на основу реторичке фигуре поређења. Реторичка фигура поређења повезује уобичајене, устаљене концепте са изузетним, необичним, често измишљеним концептима и тако омогућава узајамно деловање норме и креативности. Ова реторичка фигура се у српском језику најчешће изражава у облику „*придев* као *именица*“ (нпр. „снажан ако медвед“). Пројекат кандидаткиње је, с тога, отпочео претрагом анотираног СрпКор-а у коме су тражене конструкције овог облика. У првом експерименту кандидаткиња је из добијене листе примера, после филтрирања, одабрала подскуп оних за које је на основу своје језичке компетенције сматрала да представљају добре кандидате. Они су уврштени у упитнике који су дељени преко друштвених мрежа и у којима су учесници одговорима „да/не“ означавали понуђене кандидате као нешто што се користи у свакодневном говору или не. Квалитет рада добровољних учесника је мерен на основу једног „златног питања“ (питања на које се једино очекивао одговор „не“), а потом су као релевантни учесници одабрани они код којих није било значајне разлике између аритметичких средина њихових одговора за шта је коришћен *Студентов t-тест*. Коначно су као релеватни одговори одабрани они код којих је Крипендорфов алфа коефицијент показао задовољавајућу сагласност између учесника. На тај начин се дошло до скупа од 53 конструкција „*придев* као *именица*“ које функционишу као реторичка фигура поређења. Кандидаткиња је потом показала да уколико би се од целокупног филтрираног излаза претраге СрпКор-а одабрао само подскуп оних примера који имају фреквенцију појављивања већу од три, 84% израза у њему садржаних били би изрази које су

учесници експеримента групне расподеле рада потврдили. Други пројекат групне расподеле рада је кандидаткиња спровела на исти начин, само што су после филтрирања резултата претраге СрпКор-а кандидати за упитнике бирани насумично. Иако је у овом случају било више скупова релевантних одговора (према Крипендорфовом алфа коефицијенту), број потврђених израза „*придев* као *именица*“ је, управо због насумичног одабира кандидата, био знатно мањи, само 21. У овом случају се фреквенција појављивања израза у СрпКор-у од пет или више показала као добар праг, јер је у том скупу 86% израза потврђено и експертизом учесника у другој групној расподели рада. На основу ових резултата кандидаткиња је аутоматски или полуаутоматски допунила Српски ворднет са 225 релација типа специфичанЗа/специфицированСа.

Пето поглавље „Примена унапређеног Српског ворднета“ кандидаткиња Јелена Митровић је посветила првим успешним применама унапређеног Српског ворднета, конкретно применама релације специфичанЗа/специфицированСа. Једна од таквих примена је аутоматско проналажење реторичких фигура ироније и сарказма у кратким порукама на друштвеним мрежама (Твитер) на српском језику. За овај задатак Српски ворднет је трансформисан у формалну онтологију над којом је могуће закључивање, чиме је омогућено повезивање релацијом антонимије удаљених синсетова, то јест, синсетова који нису директни антоними. Новододате семантичке релације у Српском ворднету, специфичанЗа/специфицированСа, су суштински важне за препознавање ових реторичких фигура: на пример, ако су придеви *спор* и *пуж* повезани паром ових инверзних релација, а релација антонимије постоји између придева *брз* и *спор*, правилима логичног закључивања спроведеним над онтологијом Српског ворднета може се аутоматски закључити да израз *брз као пуж* представља иронично тврђење. Слично истраживање кандидаткиња је спровела и за грчки језик, чиме је потврђена универзалност примењених метода.

У шестом поглављу „Закључак и будући рад“ кандидаткиња Јелена Митровић даје сажет приказ свог рада и постигнутих резултата – допуњавање Српског вордента придевским синсетовима, додавање нових релација у Српски ворднет као и додавање примера поређења који су прикупљени пројектима групне расподеле рада у формалну онтологију реторичких фигура *РетФиг*. Кандидаткиња планира слично истраживање за конструкције типа „*глагол* као *именица*“, нпр. „поцрвенети као булка“, у коме би се учеснички упитници разликовали јер би се уместо „да/не“ очекивали конкретни одговори, нпр. поцрвенети као *<допуни именицом>*. Кандидаткиња предвиђа даљи кооперативни рад на изградњи Српског ворднета уз коришћење новоизграђених алата и метода групне расподеле рада. Осим тога, кандидаткиња планира рад на проширењу лексичко-семантичке мреже типа ворднет за немачки језик – GermaNet, што ће захтевати прилагођавање методологије додавања нових семантичких веза, узимајући у обзир посебну структуру лексичко-семантичке мреже GermaNet, у којој су синсетови придева организовани другачије него у Ворднету, уз коришћење хијерархијског приступа.

На крају дисертације мастер Јелена Митровић је приложила шест додатака:
1. Листа 25 именичких синсетова који представљају коренове хијерархијских дрвета Принстонског ворднета (па самим тим и Српског ворднета).
2. Основни подаци о Српским морфолошким е-речницима, број одредница у речницима монолексемских и полилексемских јединица, одабрани семантички и други маркери, као и неки илустративни примери.
3. Основни подаци о Српском ворднету (број синсетова према врсти речи, број литерала по синсетовима, расподела значења према врсти речи, расподела ознака позитивног и негативном осећања по синсетовима).
4. Подаци о везама међу синсетовима у Српском ворднету (врста и број лексемско-семантичких релација), листа аутоматски додатих веза у Српски ворднет типа специфичанЗа/специфицированСа и листа кандидата за ручно повезивање.
5. Уводни текст упитника у оба пројекта групне расподеле рада.
6. Речник термина и скраћеница који су коришћени у тексту дисертације.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ, уз напомену: Навести називе радова, где и када су објављени.

J. Mitrović, C. O'Reilly, M. Mladenović, S. Handschuh, "Ontological Representations of Rhetorical Figures for Argumentation Mining", *Argument & Computation Journal*, Vol 8, Issue 2, 2017. DOI: 10.3233/AAC-170027

M. Mladenović and **J. Mitrović**, "Ontology of Rhetorical Figures for Serbian", *Lecture Notes in Computer Science and Artificial Intelligence* 8082/Springer-Verlag Berlin Heilderberg, 2013, pp. 386-393. DOI: 10.1007/978-3-642-40585-3_49

J. Mitrović, "Crowdsourcing and its application," *INFOtheca*, vol. 14, pp. 37-46, Jun 2013.

M. Mladenović, **J. Mitrović**, "Semantic Networks for Serbian – New Tools for Developing and Maintaining a WordNet", *Proceedings of the 35th Anniversary of Computational Linguistics in Serbia*, 1-11, Belgrade, November 2013.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је кандидаткиња мастер Јелена Митровић остварила циљеве зацртане у пријави дисертације. Кандидаткиња је дала детаљан преглед постојећих лексичких ресурса за српски језик и анализу могућности њихове надоградње коришћењем групне расподеле рада, обавила анализу постојећих приступа групној расподели рада и њихове применљивости на развој лексичких ресурса, понудила решење једног проблема допуне Српског ворднета коришћењем за ту намену припремљене апликације групне расподеле рада и решила један конкретан задатак препознавања реторичких фигура ироније и сарказма у текстовима на српском језик коришћењем тако допуњеног Српског ворднета.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је мастер Јелена Митровић користила релевантну и савремену литературу, те да је постављене проблеме обрадила детаљно и сагледавајући их из разних углова. Овим радом Јелена Митровић је увела нове методе у област рачунарске обраде српског језика а дограђене лексичке ресурсе ставила на располагање будућим истраживачима.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА
НАПОМЕНА: Навести позитивну или негативну оцену начина приказа и тумачења резултата истраживања.

Комисија сматра да је кандидаткиња Јелена Митровић у својој дисертацији *Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада* успешно обрадила значајну тему коришћењем нових метода, да је текст дисертације урађен према одобреној пријави дисертације, и да је реч о раду који представља оригинално и самостално научно дело.

X ПРЕДЛОГ:

На основу укупне оцене дисертације, комисија предлаже Научно-наставном већу Филолошког факултета Универзитета у Београду да прихвати извештај о дисертацији *Електронски језички*

ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада кандидаткиње Јелене Митровић и упути га Већу за друштвено-хуманистичке науке Универзитета у Београду, како би кандидаткиња била позвана на усмену одбрану рада.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

1. др Цветана Крстев, редовни професор
   Филолошки факултет Универзитета у Београду

2. др Александра Вранеш, редовни професор
   Филолошки факултет Универзитета у Београду

3. др Ранка Станковић, ванредни професор
   Рударско-геолошки факултет Универзитета у Београду