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
25. IX 2019. 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 Милош Утвић, доцент, библиотечка информатика, 18. XI 2014, Филолошки факултет Универзитета у Београду
3. др Јелена Костић Томовић, редовни професор, немачки језик, 16. маја 2018, Филолошки факултет Универзитета у Београду
4. др Гордана Павловић-Лажетић, редовни професор, рачунарство и информатика, 22. I 2009, Математички факултет Универзитета у Београду
5. др Ранка Станковић, ванредни професор, математика и информатика, 19.V 2015, Рударско-геолошки факултет Универзитета у Београду

II ПОДАЦИ О КАНДИДАТУ

1. Име, име једног родитеља, презиме:
Јелена Слободан Андоновски

2. Датум рођења, општина, република:
21. IX 1987. Лесковац, Србија

3. Датум одбране, место и назив мастер рада:
30. IX 2011. Београд. „Европеана – врата до европског културног наслеђа“

4. Научна област из које је стечено академско звање мастера:

Библиотекарство и информатика

III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

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

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Навести кратак садржај са назнаком броја страна поглавља, слика, шема, графикона и сл.

Докторска дисертација Јелене Андоновски бави се изградњом српско-немачког паралелног литерарног корпуса кога као значајан језички ресурс могу да користе истраживачи у својим лингвистичким и филолошким истраживањима и рачунарске апликације које се заснивају на обради природно-језичких података. Истраживање посебно обухвата:

- основне појмове о рачунарским корпусима, једнојезичним и вишејезичним, начинима њихове изградње и могућностима њиховог коришћења, с посебним освртом на корпусе који садрже текстове на српском језику;
- преглед постојећих стандарда метаподатака који омогућавају ефикасно проналажење информација у корпусима, посебно у паралелним корпусима;
- представљање идеје мреже „отворених повезаних података“ и њеног значаја за семантички веб;
- српско-немачки паралелни литерарни корпус кроз све фазе његове изградње, од одабира текстова до постављања корпуса на веб, уз примере његовог коришћења и укључивања у мрежу отворених повезаних лингвистичких података.

У истраживању се користе експерименталне методе, као што су корпусне методе претраживања, и методе семантичког веба. За процену квалитета остварених резултата коришћен је једноставан поступак евалуације.

Дисертација обухвата 346 страна, а у оквиру тога 7 поглавља (255 страна), списак коришћене литературе (34 стране, 282 библиографске јединице), 15 прилога (34 стране), уводни и завршни материјал (насловне стране, апстракт на српском, енглеском и руском, садржај, списак табела, списак слика, списак скраћеница, биографија кандидата, 23 стране). У дисертацији укупно има 70 слика и 15 табела. Поглавља дисертације су:

1. Увод (4 стране).
2. Рачунарски језички корпуси (66 страна).
3. Израда паралелних корпуса у Србији (26 страна).
4. Стварање предуслова за ефикасно проналажење информација у паралелним корпусима (29 страна).
5. Семантички веб и мрежа отворених повезаних података (46 страна);
6. Српско-немачки паралелни корпус – СрпНемКор (79 страна);
7. Постигнути резултати и будући рад (5 стране).

Прилози дисертације су:

А. Библиографија (34 стране).
Б. Додаци (15 прилога, 34 стране)

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дисертације кандидаткиња Јелена Андоновски представља предмет

својих истраживања, даје њихов кратак преглед по поглављима и укратко приказује остварене резултате.

У другом поглављу „Рачунарски језички корпуси“ кандидаткиња, полазећи од појма *корпусна лингвистика*, представља језичке корпусе као значајне ресурсе у различитим областима истраживања и проучавања језика са посебним освртом на паралелне корпусе као једну од врста језичких корпуса која је последњих деценија постала изузетно значајна за двојезичну и вишејезичну лексикографију, учење страних језика, превођење и машинско превођење, истраживање терминологије, лингвистичка истраживања, упоредна изучавања два и више језика и друго. У овом поглављу су представљени и неки примери добре праксе паралелних корпуса у свету: корпуси некњижевних текстова (Acquis Communautaire, EuroParl, SETimes, OpenSubtitles, BabelNet) и корпуси књижевних текстова, којих је знатно мање (Платонова *Република*, и Орвелова *1984*). Кандидаткиња даје затим преглед развоја корпусне лингвистике у Србији, посебно се осврћући на напоре и истраживања који су довели до развоја Корпуса савременог српског језика (СрпКор), а затим и до развоја два паралелна корпуса: српско-француски корпус и српско-енглески корпус. Оба ова корпуса садрже и књижевне и некњижевне текстове, као и текстове оригинално написане на српском језику и текстове преведене на српски са француског, односно енглеског језика. Представљена су и два паралелна вишејезична текста, Орвелова *1984* и Вернов роман *Пут око света за 80 дана* који је у потпуности развила Група за језичке ресурсе и технологије Универзитета у Београду и који обухвата текстове романа на 20 језика. Кандидаткиња представља са више детаља два ресурса исте Групе, дигиталну библиотеку паралелних текстова и систем за њихово претраживање и коришћење, обједињене под називом *Библиша*. Резултати истраживања кандидаткиње такође су постали део *Библише*.

Треће поглавље „Израда паралелних корпуса у Србији“, детаљно објашњава поступак израде једног паралелног корпуса. Процес паралелизације подразумева да су текстови у електронском облику коректно припремљени и упарени, а резултати представљени у одговарајућем стандардном формату. Упаривање подразумева повезивање одговарајућих сегмената текстова – „преводних еквивалената“ – као што су пасуси, или реченице, односно делови реченица. Сам поступак припреме текстова и креирање паралелног корпуса пролази кроз неколико фаза, као што су прикупљање и дигитализација текстова, примена алата и језичких ресурса за обраду текстова пре паралелизације, анотација корпусног материјала (снабдевање текстова изванјезичким и језичким анотацијама), поступак паралелизације уз помоћ одговарајућег софтвера, и коначно производња корпуса у једном или више одабраних излазних формата. Поред објашњења поступка, у поглављу су представљени језички ресурси (електронски морфолошки речници српског језика, српски ворднет) и софтвер (IMS OCWB, Unitex, XAlign, ACIDE) који се у Србији користе за припрему и израду паралелних корпуса, као и начини претраге паралелних корпуса који су до сада развијени у Србији. Како се у корпусу који се анализира у дисертацији ради о делима савремених писаца, односно писаца који су писали у другој половини 20. века, посебно питање које се у овом поглављу разматра је питање поштовања ауторских права током приступа и коришћења корпусног материјала, као и поштовање приватности у складу са законима и важећим прописима.

У четвртом поглављу „Стварање предуслова за ефикасно проналажење информација у паралелним корпусима“ кандидаткиња се бави описом и означавањем докумената са становишта њихових формалних и садржинских особина како би се она могла претраживати и проналазити у базама података. Највећи део поглавља посвећен је формалном опису дигиталних докумената и објеката, односно додељивању метаподатака. Кандидаткиња на почетку уводи појам *метаподатак*, а затим анализира врсте метаподатака, истиче њихов значај и улогу у дигиталним репозиторијумима, базама података, електронским каталозима и другим савременим колекторима података. Кандидаткиња потом представља неке од међународних стандарда за израду метаподатака који су у широкој употреби у свету и у Србији: Даблинско језгро, METS (Стандард за кодирање и пренос података), MODS (шема за опис метаподатака), заглавље TEI (Иницијатива за кодирање текста). Поред додељивања формалног описа, дигитална документа се данас и садржински индексирају применом различитих техника које спадају у домен обраде природних језика. Све то обезбеђује да комплетни текстови докумената

похрањених у дигиталним колекцијама буду претраживи. Кандидаткиња у овом поглављу представља неке од техника припреме и додатне обраде текстова — технологију оптичког препознавања карактера, аутоматско индексирање садржаја докумената и препознавање именованих ентитета, односно назива (*named entities*) – које су примењене и на корпус паралелних текстова као предмет дисертације.

Пето поглавље „Семантички веб и мрежа отворених повезаних података“ дефинише појам *семантичког веба* као мрежу података на вебу (уместо мреже докумената на вебу) са придруженим значењем и међусобно повезаних тако да омогуће ефикасније проналажење информација. Кандидаткиња представља кључни стандард за описивање значења података и за њихово повезивање на нивоу значења – RDF (*Resource Description Framework*) – који је омогућио стварање мреже отворених повезаних података LOD (*Linked Open Data*). У овом поглављу су представљени и неки ресурси који су део семантичког веба (ДБпедија, Википодаци) и омогућавају његову реализацију (отворени повезани подаци, системи за организацију знања). Посебна пажња је посвећена значају семантичког веба за библиотеке и њиховом узајамном односу у коме свака страна значајно доприноси развоју друге стране. На крају поглавља приказане су неке иницијативе и пројекти који се заснивају на семантичком вебу (Европеана, Дигитална народна библиотека Америке, Оквир за библиографски опис), њихови модели, начини претраге и приступа информацијама, а указано је и на предности овакве структуре за различите заједнице, а пре свега за библиотеке у најширем смислу.

Главни резултати дисертације представљени су у шестом поглављу „Српско-немачки паралелни корпус – СрпНемКор“. Кандидаткиња прво представља идеје које су довеле до креирања овог литерарног корпуса, а потом и критеријуме којима се руководила приликом одабира романа уврштених у корпус. Иницијални садржај корпуса састоји се од седам савремених романа написаних на српском језику и њихових превода на немачки језик, те седам савремених романа написаних на немачком језику и преведених на српски језик, што значи да се корпус састоји од укупно 28 текстова, то јест, 14 паралелних текстова. Овако формиран корпус има 1.657.329 текућих речи, при чему српски, односно, немачки део корпуса имају приближно исти број речи (48,87%, односно 51,13% величине корпуса). Први изазов с којим се кандидаткиња сусрела био је проналажење самих текстова, у дигиталном или папирном облику. Како је само мали број одабраних текстова био доступан у дигиталном облику, кандидаткиња је у библиотекама у земљи и иностранству, као и у приватним колекцијама пронашла штампана дела која су потом прошла све фазе дигитализације и обраде описане у претходним поглављима дисертације: сканирање, оптичко препознавање карактера, кориговање, структурну и морфолошку анотацију и коначно паралелизацију, односно формирање паралелних текстова. За овако формиране паралелне текстове припремљени су метаподаци што је омогућило и њихово укључивање у дигиталну библиотеку паралелних колекција *Библиша*. Међу метаподацима сваког од текстова се, поред уобичајених података (име аутора, назив дела, итд.), налази и УДК број као и идентификациони бројеви одговарајућих записа који се тичу текста у нормативним базама података VIAF (Virtual International Authority File), GND (Gemeinsame Normdatei или Universal Authority File) и LCNAF (Library of Congress Name Authority File), као и у бази података Википодаци (VikiData). С обзиром да дигитална библиотека *Библиша* омогућава ефикасну претрагу колекција, било преко метаподатака било преко садржаја докумената, кандидаткиња је представила могућности претраге СрпНемКор, као и постојеће лексичке ресурсе који подржавају напредне технике претраживања, укључујући и двојезично претраживање. Посебну пажњу кандидаткиња посвећује унапређивању постојећих лексичких ресурса и њиховој допуни за немачки језик. Тако су постојећи електронски речници српског језика допуњени са више од 2.500 нових одредница. За двојезично српско-немачко претраживање формиран је потпуно нови ресурс коришћењем система BilTE (Bilingual Terminology Extraction) који су развили сарадници Групе за језичке ресурсе и технологије Универзитета у Београду. Тако је добијен ресурс који садржи 3.984 српско-немачка преводна пара. Након идентификовања синонима и њиховог повезивања, овај ресурс је укључен у базу *Терми*, један од ресурса који подржавају двојезичну претрагу у *Библиши*. Кандидаткиња је анализирала могућност аутоматског анотирања текстова корпуса именованим ентитетима (имена људи, геополитичка имена, имена организација) расположивим

алатима за српски и немачки језик, јер би такве анотације значајно обогатиле могућности претраживања. Међутим, како су тестирани алати произведени махом за потребе анотирања новинских текстова, њихова примена на литерарне текстове није била превише успешна, тако је кандидаткиња одустала од укључивања ових напредних анотација у развијени корпус.

У завршном делу овог поглавља, кандидаткиња приказује како су технике и стандарди семантичког веба примењени на СрпНемКор. Један део се односи на повезивање библиографских метаподатака текстова из корпуса са релевантним ресурсима на вебу (VIAF, GND, LCNAF и Википодаци), док други део описује припрему скупа података припремљених према принципима семантичког веба и њихово укључивање у облак „Отворени повезани подаци“. За ове потребе је на основу паралелне колекције генерисан узорак двојезичног речника општег типа. Овај ресурс се заснива на раније успостављеним преводним паровима обогаћеним везама синонимије. Да би се могао укључити у отворене повезане податке, овај ресурс је трансформисан у RDF формат коришћењем два модела података: Оквир за означавање лексике (Lexical Markup Framework – LMF) и Модел лексикона за онтологије (Lexicon Model for Ontologies – lemon). Представљањем ресурса у овим форматима, уз припремљено заглавље метаподатака, створени су услови за објављивање ресурса као скупа отворених повезаних података.

У седмом поглављу „Постигнути резултати и будући рад“ кандидаткиња Јелена Андоновски даје сажет приказ свог рада и постигнутих резултата – формирање српско-немачког литерарног корпуса СрпНемКор, његово укључивање у дигиталну библиотеку паралелних колекција *Библиша*, изградња двојезичног лексичког ресурса који подржава напредну двојезичну претрагу и коначно његова трансформација у формат који омогућава објављивање као скуп отворених података. Кандидаткиња анализира и представља више могућности за проширивање и унапређивање постигнутих резултата. Предвиђено је увећање корпуса новим текстовима (литерарног или неког другог типа) и његово укључивање у корпусе српског језика претраживе помоћу софтвера IMS OCWB, којима припадају и други паралелни корпуси српског језика (српско-енглески и српско-немачки). Кандидаткиња даље сматра да би за унапређивање претраге системом *Библиша* било корисно укључивање семантичке мреже ворднет за немачки језик (таква мрежа за српски језик је већ укључена), на пример Open-de-WordNet. Друго значајно проширење била би могућност морфолошког проширивања упита за немачки језик (таква могућност за српски већ постоји). Кандидаткиња планира да анализира стање потребних алата за немачки језик у циљу проналажења алата који је истовремено у отвореном приступу и даје задовољавајуће резултате.

На крају дисертације мастер Јелена Андоновски је приложила 3 додатка и 15 прилога:

1. У додатку 1 дат је списак слика у дисертацији.
2. У додатку 2 дат је списак табела у дисертацији.
3. У додатку 3 дат је списак скраћеница које су коришћене у дисертацији, одвојено ћириличне и латиничне.
4. У прилозима 1-8 дати су примери записа метаподатака исте библиографске јединице, превода дела Томаса Бернхарда *Моје награде* на српски језик, у различитим стандардима и форматима: COMARC/B/ISO2709, MARC21/ISO2709, Даблинско језгро/XML, METS/XML, MODS/XML, MARK21/XML, COMARC/XML, TEI заглавље/XML
5. У прилогу 9 је дат табеларни упоредни приказ записа метаподатака у различитим стандардима и форматима.
6. У прилозима 10-15 дати су примери записа истог дела Томаса Бернхарда у референтним базама података: GND/корисничко окружење, GND/RDF/Turtle, VIAF/HTML, LCNAF/корисничко окружење, LCNAF/RDF/XML, Википодаци/корисничко окружење, EDM/корисничко окружење, BIBFRAME/корисничко окружење.

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.

Andonovski, Jelena, Branislava Šandrih i Olivera Kitanović. „Bilingual lexical extraction based on word alignment for improving corpus search“. *The Electronic Library* Vol. 37, No. 4(2019): 722-739, https://doi.org/10.1108/EL-03-2019-0056. Dostupno na: https://www.emerald.com/insight/content/doi/10.1108/EL-03-2019-0056/full/html
Impakt factor časopisa za 2018: 0.886
Petogodišnji impact factor u 2018: 1.119
Oblast: Library and Information Studies

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је кандидаткиња мастер Јелена Андоновски остварила циљеве зацртане у пријави дисертације. Кандидаткиња је дала прецизан опис изградње паралелних корпуса, посебно се осврћући на значај метаподатака, детаљно описала поступак изградње литерарног паралелног српско-немачког корпуса, од почетне фазе одабира дела за корпус до завршне фазе његовог објављивања на вебу у дигиталној библиотеци паралелних колекција, развила ресурсе за унапређивање претраге корпуса које је потом припремила у формату потребном за њихово објављивање као отворених повезаних података.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је мастер Јелена Андоновски користила релевантну и савремену литературу, те да је постављене проблеме обрадила детаљно и сагледавајући их из разних углова. Овим радом мастер Јелена Андоновски је изградила један значајан нови информатички ресурс за српски (и немачки) језик који је постао доступан широком кругу истраживача, поставила је основе за његово унапређивање и развила методологију за изградњу сличних ресурса у будућности.

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 Андоновски у својој дисертацији *Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса* 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 *Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса* кандидаткиње Јелене Андоновски 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 Милош Утвић, доцент
   Филолошки факултет Универзитета у Београду

3. др Јелена Костић Томовић, редовни професор
   Филолошки факултет Универзитета у Београду

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

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