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
17. XII 2013. 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, ванредни професор, библиотечка информатика, 9. III 2009. године, Филолошки факултет Универзитета у Београду
2. др Божо Ђорић, редовни професор, 18. X 1985, српски језик, Филолошки факултет Универзитета у Београду
3. др Весна Половина, редовни професор, 18. V 2001, општа лингвистика, Филолошки факултет Универзитета у Београду
4. др Гордана Павловић-Лажетић, редовни професор, 22. I 2009, рачунарство и информатика, Математички факултет Универзитета у Београду
5. др Душко Витас, ванредни професор, 25. V 2006, рачунарство и информатика, Математички факултет Универзитета у Београду

II ПОДАЦИ О КАНДИДАТУ

1. Име, име једног родитеља, презиме:
Милош Војислав Утвић
2. Датум рођења, општина, република:
29. I 1976. Прокупље, Србија
3. Датум одбране, место и назив магистарске тезе:
18. II 2008. Београд. Коначни аутомати у регуларној именској деривацији
4. Научна област из које је стечено академско звање магистра наука:
Рачунарство

III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Изградња референтног корпуса савременог српског језика

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

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

У докторској дисертацији се разматра проблем метода и алата за конструкцију корпуса савременог српског језика као референтног језичког ресурса. У првом делу рада се разматрају општа питања која се односе на дефиницију, историјат и класификацију корпуса, као и на корпусну лингвистику као методологију истраживања језика. Уводи се скуп параметара за процену и поређење корпуса. Посебна пажња је посвећена репрезентативности и балансираности корпуса као узрока језика. Истражен је и утицај Интернета, а посебно веба на критичко преиспитивање дефиниције корпуса. Детаљно су представљени национални корпуси словенских језика и историјат развоја српске корпусне лингвистике. У другом делу рада описују се опште методе и фазе у оквиру претходне обраде и анализе корпуса. Посебно су представљени репрезентативни интегрисани корпусни алати и извршена је њихова упоредна

анализа коришћењем уведеног скупа параметара поређења. У трећем делу рада је описан процес конструкције Корпуса савременог српског језика (СрпКор) с посебним освртом на актуелну верзију СрпКор2013. Изложен је историјат развоја овог корпуса и анализирани су његови параметри на основу чега је омогућено поређење са сличним корпусима других језика. Детаљно су описане све фазе укључене у изградњу корпуса као и могућности претраге. У овом делу дисертације је представљена и изградња паралелних корпуса у којима је изворни или циљни језик српски.

Докторска дисертација обухвата 8 поглавља текста (317 страна, 36 табела, 54 слике), библиографију (36 страна, 332 јединице) и додатке (36 страна), укупно 396 страна. Поглавља дисертације су:

1. Корпусна лингвистика као методологија (56 страна, 6 табела)
2. Претходна обрада корпуса (101 страна, 9 табела, 27 слика),
3. Претрага и анализа корпуса (38 страна, 6 табела, 6 слика),
4. Преглед постојећих алата за рад на корпусу (45 страна, 3 табеле, 4 слике),
5. Корпус савременог српског језика (СрпКор) (13 страна, 3 табеле),
6. Фазе у креирању корпуса СрпКор (43 стране, 5 табела, 14 слика),
7. Претрага (9 страна, 2 табеле, 3 слике) и
8. Закључак и даљи рад (3 стране).

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном делу дисертације Милош Утвић уводи појам корпусне лингвистике и прати њен развој у четири фазе које обележавају: преелектронски корпуси, критика Чомског, електронски корпуси прве генерације и електронски корпуси друге генерације, тј. савремени корпуси. С обзиром да се велики број актуелних дефиниција корпуса слаже да је у питању „репрезентативни узорак појединачног језика или језичког варијетета“, кандидат анализира појам „репрезентативности“ и даје критеријуме за оцену репрезентативности корпуса. Премда многи аутори сматрају да је репрезентативност корпуса циљ који се не може постићи, понуђени критеријуми могу водити ауторе корпуса ка приближавању том циљу. Милош Утвић даље уводи параметре за класификацију корпуса: носач, обим (величина), домен, намена, период, извор, начин анотације, вишејезичност. Ове параметре ће кандидат користити при представљању и поређењу националних корпуса словенских језика: белоруског, бугарског, чешког, хрватског, пољског, руског, словачког, словеначког и украјинског, а они ће бити коришћени и у завршном делу рада за сагледавање позиције Корпуса савременог српског језика према корпусима других словенских језика. Коначно, аутор представља нови феномен „веб као корпус“ као потенцијални корпус треће генерације и указује на његове предности и недостатке, на које указују његови заговорници, односно критичари.

У другом делу дисертације, насловљеном „Креирање и анализа корпуса“, Милош Утвић се бави општим питањима везаним за компилацију и претрагу корпуса. Овај део дисертације подељен је у три поглавља. У поглављу „Претходна обрада корпуса“ кандидат се бави неизбежним корацима које сви аутори корпуса морају предузети пре него што приступе формирању самог корпуса. Овде ћемо поменути само неке од најважнијих задатака којима је и кандидат посветио нужну пажњу. Први од њих тиче се ауторских права и заштите приватности. По самој природи ствари, у корпус махом улазе текстови који су заштићени законом о ауторским и сродним правима. Према томе, аутор корпуса би требало да добије сагласност носилаца ауторских права за све текстове које планира да уврсти у корпус. Тај циљ је био тешко остварив и у време првих електронских корпуса са релативно малим бројем узорака, док је данас, када је број узорака неупоредиво већи и када је за многе текстове, посебно оне преузете с Интернета, веома тешко утврдити ко је носилац ауторских права, овај циљ практично недостижан. С друге стране, и сам корпус је ауторско дело заштићено ауторским правима. Да би се превазишла ова два повезана проблема, појам „поштене употребе“ којим се суспендују искључива ауторска права у ограниченим ситуацијама, уврштен је у законе о ауторским правима. Корпусна сучеља, која кориснику омогућавају да кроз произведене конкорданце види само сегменте узорака текстова обезбеђују поштовање „поштене употребе“.

Други важан корак који претходи формирању корпуса је припрема свих узорака текстова у електронском облику и то на једнообразан начин. То, пре свега, значи да сви текстови треба да буду кодирани коришћењем истог кодног распореда, те је потребно извршити потребна подешавања. Овај проблем је посебно истакнут у случају корпуса српског језика у које улазе подједнако узорци текстова записани ћириличним и латиничним писмом. Како писмо које је оригинално коришћено за запис узорка текста није од значаја за већину лингвистичких и литерарних корпусних истраживања, сви узорци текстова треба да буду „неутрални“ у односу на писмо. Милош Утвић се у овом поглављу детаљно бавио овим питањем, представио решења коришћена у прошлости и предложио адекватно решење које је примењено у случају Корпуса савременог српског језика. Следе виши нивои припреме електронских текстова, пре свега токенизација – утврђивање шта су корпусне речи текста – и одређивање краја реченице. Коректно извршење ових задатака је пресудно да би претраживање корпуса било у складу с корисниковим очекивањима. Кандидат указује на проблеме који се јављају при решавању ових задатака и представља основне методе за њихово решавање: методе засноване на ручно конструисаним правилима и методе машинског учења. На крају овог одељка Милош Утвић се бави највишим нивоом припреме електронских текстова, а то је њихова анотација као поступак придруживања додатних информација узорцима текстова, реченицама, деловима реченица или токенима. Премда се одређени број корпусних лингвиста противи анотацији јер сматрају да ниједно истраживање засновано на анотираном корпусу није непристрасно, тј. да у великој мери одражава погледе онога ко је корпус анотирао (о чему Милош Утвић говори у првом делу рада), чињеница је да је већина корпуса данас анотирана и да многи истраживачи очекују да раде управо с оваквим корпусима. Кандидат, стога, представља предности и мане ручног према аутоматском анотирању корпуса, уграђених анотација према независним анотацијама. Милош Утвић је посебно детаљно обрадио најзначајније стандарде за анотацију корпуса, како оне за структурну анотацију – XML и на њему засновани формати, независни формати („хоризонтални“ и „вертикални“ формат текста), TEI CES/XCES и ISO LAF, тако и оне за морфосинтаксичку анотацију – LADL/DELA и Multext-East.

Друго поглавље „Претрага и анализа корпуса“ Милош Утвић је посветио, пре свега, основним методама за претрагу корпуса, међу којима најистакнутије место заузимају регуларни изрази (појам из теорије формалних језика), а затим конкорданцама, као најпопуларнијем и најинформативнијем формату за приказ резултата претраге. Милош Утвић представља и основне методе статистичке анализе корпуса, као што су израда листи учесталости и утврђивање колокација које се данас сматрају стандардним методама за анализу корпуса.

У трећем поглављу „Преглед постојећих алата за рад са корпусом“ Милош Утвић представља на систематичан начин најпознатије корпусне алате. С обзиром да је понуда оваквих алата данас веома богата, Милош Утвић је понудио скуп параметара који се могу користити за њихову процену и који могу да буду посебно корисни онима који се тек одлучују за набавку и коришћење неког од њих. Тај скуп параметара укључује следеће: лиценца софтвера, платформа (хардверски и софтверски предуслови), клијент-сервер архитектура и веб, ажурност и подршка, проширивост. На процену неког корпусног алата утиче и пристуство, односно одсуство, неког од додатних модула: претходна обрада, језички ресурси, типови претраге (прилагођавање конкорданци) и расположиве статистичке функције.

Трећи део дисертације Милоша Утвића „Корпус савременог српског језика“ је њен најзначајнији део јер се у њему приказују резултати оригиналног истраживачког рада чији је резултат израда Корпуса савременог српског језика (СрпКор). Овај део је подељен у четири поглавља. У поглављу „Корпус савременог српског језика (СрпКор)“ кандидат даје краћи историјски преглед настанка овог корпуса, навођењем листе бројних домаћих и међународних пројеката који су непосредно, али чешће само посредно допринели његовој изградњи. Пратећи историјски развој овог корпуса, кандидат поименце наводи многобројне истраживаче и студенте који су својим, често волонтерским, радом допринели његовој изградњи. Користећи параметре за класификацију корпуса уведене у првом делу рада, Милош Утвић представља

СрпКор. Овај корпус сада има 122 милиона корпусних речи које потичу из 4.890 узорака текстова који су настали у периоду од 1920. године до данас. Корпус чине искључиво писани текстови. Већину текстова у корпусу чине новински текстови, али значајан је и број административних и књижевних текстова. Окосницу корпуса чине текстови оргинално написани на српском језику; уз њих, корпус садржи и известан број превода. Корпус је анотиран на три нивоа: библиографским информацијама о корпусним текстовима, структурним информацијама и морфосинтаксичким информацијама (врста речи и лема).

Наредно поглавље „Фазе у креирању корпуса СрпКор“ сведочи о изузетној приљежности кандидата Милоша Утвића и о његовом дугогодишњем самопрегорном раду на обављању низа комплексних и деликатних задатака, што се односи не само на послове прикупљања текстова и њихову конверзију у униформни формат електронског текста већ пре свега на послове анотације. Придруживање библиографских информација је пажљиво урађено за све узорке текстова у корпусу, било претрагом база COBISS.SR и COBIB.SR било консултовањем изворних папирних издања, а скуп придружених информација садржи осим традиционалних библиографских података (наслов, аутор, година издања, и сл.) и податке о функционалном стилу и УДК класификацији. Све ове информације омогућавају корисницима корпуса да постављају сложене упите које укључују све библиографске информације (нпр., ако корисник жели само примере из дела Иве Андрића). Структурна анотација је урађена применом спецификације TEI-LITE за један део узорака текстова. Најсложенији је био задатак морфолошке анотације, за чије решење је било потребно, пре свега, прецизно дефинисати скуп морфолошких дескриптора (етикета) који се придружују појединачним токенима, изабрати алат за аутоматско етикетирање, а потом и припремити помоћне ресурсе на основу којих се изабрани алат обучава за рад. Аутор се определо за TreeTagger који сваком токену придружује не само информацију о врсти речи већ и лему, који је обучен на колекцији полуаутоматски анотираних текстова (припремљеној у оквиру делатности Групе за језичке технологије при Математичком факултету Универзитета у Београду).

Премда је изградња Корпуса савременог српског језика у фокусу дисертације Милоша Утвића, на крају поглавља „Фазе у креирању корпуса СрпКор“ кандидат се бави и корацима у изградњи вишејезичних, пре свега, двојезичних корпуса у којима је један од језика српски. Припрема текстова за ову врсту корпуса је посебно захтевна јер они морају да буду брижљиво структурно анотирани. Премда се поравнавање текстова на нивоу изабране структурне целине обавља аутоматски, потребна је контрола поравнатих јединица. Аутор представља алат за поравнавање коришћен за двојезичне текстове код којих је један језик српски (Xalign), као и сопствени алат за корекцију (ACIDE) и израду излаза у различитим форматима (HTML и формат преводилачких меморија TMX).

У поглављу „Претрага“ кандидат Милош Утвић представља сопстево веб-сучеље које омогућава једноставну и напредну претрагу СрпКор-а. Оно се заснива на CQP упитном језику као саставном делу корпусног алата IMS OCWB који подржава рад са СрпКор-ом на вебу. Једноставна претрага омогућава претрагу по корпусним речима (без анотација) задавањем сложених упита регуларним изразима. Излазне конкорданце могу бити у два формата: „кључна реч у контексту“ и „кључна реч у пасусу“, а сортирање редова конкорданци је могуће и по левом и по десном контексту. Сваком реду конкорданци је придружена веза ка пуном библиографском опису извора. Корисник може да селектује редове конкорданци који се могу запамти за локалну употребу. Напредна претрага омогућава претрагу по морфолошкој анотацији корпуса (по врсти речи и леми).

На крају дисертације Милош Утвић је приложио шест додатака. У додацима B-E су дати примери који читаоцу олакшавају разумевање појединих делова текста. У додатку A наведена је листа донатора текстова за корпус која укључује осим организација (Медијска кућа *Ебарт*, новинска кућа *Политика* и издавач *Службени гласник*) и више истакнутих појединаца из научног и културног живота. У додатку F је дат терминолошки енглеско-српски речник који би

могао да послужи као основа за успостављање стабилне терминологије у области корпусне лингвистике и, шире, информатичке обраде језика.

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.

1. Miloš Utvić, Ranka Stanković, Ivan Obradović, „Integrisano okruženje za pripremu paralelizovanog korpusa, Sonderdruck aus: Die Unterschide zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen (Razlike izmedju bosanskog/bošnjačkog, hrvatskog i srpskog jezika), ed. Branko Tošović, pp. 563-578, Münster, Germany, LITVerlag, 2008.
2. Милош Утвић. „Анотација Корпуса савременог српског језика". У ИНФОтека : часопис за информатику и библиотекарство, год. 12, бр. 2, стр. 39-51, Београд, 2011. УДК 004.9:811.163.41'374

У случају радова прихваћених за објављивање, таксативно навести називе радова, где и када ће бити објављени и приложити потврду о томе.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, Duško Vitas, "A system for named entity recognition based on local grammars", J Logic Computation first published online February 19, 2013, Oxford Journals, doi:10.1093/logcom/exs079, http://logcom.oxfordjournals.org/content/early/recent (рад је прихваћен за објављивање у броју за април 2014, а сада се налази у такозваном раном приступу и доступан је на вебу).

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је остварен добар део циљева зацртаних у пријави дисертације. Успостављени су критеријуми за поређење корпуса и алата за њихову анализу. Изграђен је СрпКор, електронски, синхрони и динамички корпус савременог српског језика, величине 122 милиона речи. Корпус је анотиран на три нивоа: корпусним текстовима су придружене библиографске информације, један део текстова је структурно анотиран, док су токени корпусних текстова аутоматски анотирани морфолошким информацијама о врсти речи и леми. Балансираност корпуса још увек није постигнута, али с обзиром да је корпус динамички и да се непрестано ажурира, овом циљу ће се тежити у будућности. Изграђено веб-сучеље које омогућује једноставну и напредну претрагу је у великој мери прилагођено потребама разноврсних корисника. Кандидат је доградио и израдио алате за припрему паралелизованих текстова за двојезичне корпусе који су у изградњи.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је Милош Утвић користио релевантну и савремену литературу, те да је постављене проблеме обрадио детаљно и сагледавајући их из разних углова. Овим радом Милош Утвић је ставио на располагање широкој научној јавности изузетно значајне ресурсе за разноврсна научна истраживања. Рад на изградњи корпуса српског језика није завршен: истраживање и дисертација Милоша Утвића заснивају се на раду његових претходника, а његово истраживање ће послужити као основа за даља истраживања у области корпусне лингвистике и информатичке обраде српског језика.

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

Комисија сматра да је кандидат Милош Утвић у својој дисертацији *Изградња референтног корпуса савременог српског језика* успешно обрадио ову комплексну и изузетно значајну тему, 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 *Изградња референтног корпуса савременог српског језика* кандидата мр Милоша Утвића и упути га Већу за друштвено-хуманистичке науке Универзитета у Београду, како би кандидат био позван на усмену одбрану рада.

Кликните да бисте почели унос текста.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

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

2. др Божо Ђорић, редовни професор
   Филолошки факултет Универзитета у Београду

3. др Весна Половина, редовни професор
   Филолошки факултет Универзитета у Београду

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

5. др Душко Витас, ванредни професор
   Математички факултет Универзитета у Београду