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. VI 2014. 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 Божо Ћорић, редовни професор, српски језик, 18. X 1985, Филолошки факултет Универзитета у Београду 4. др Добросав Миловановић, ванредни професор, управно правна научна област, 5.VI 2012, Правни факултет Универзитета у Београду 5. др Душко Витас, ванредни професор, рачунарство и информатика, 8. VII 2011, Математички факултет Универзитета у Београду
II ПОДАЦИ О КАНДИДАТУ
1. Име, име једног родитеља, презиме: Небојша М. Васиљевић 2. Датум рођења, општина, република: 16. XII 1968. Београд, Србија 3. Датум одбране, место и назив магистарске тезе: 15. IV 1998. Београд. Формалне методе и модели података 4. Научна област из које је стечено академско звање магистра наука: Рачунарство
III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
Аутоматска обрада правних текстова на српском језику
IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ: **Навести кратак садржај са назнаком броја страна поглавља, слика, шема, графикона и сл.**
Докторска дисертација проучава посебна језичка правила у текстовима прописа на српском језику која се могу изразити методама рачунарске лингвистике, са циљем да се на теоријском нивоу створи основа за развој софтверских алата за обраду текстова прописа и да се испитају конкретне могућности имплементације. Истраживање посебно обухвата: • опис општих језичких особености правног језика на основу припремљеног корпуса и расположивих језичких ресурса; • прецизан опис методама рачунарске лингвистике посебних језичких правила која се користе у прописима; • имплементацију синтаксне анализе формалне структуре текста прописа, као почетне фазе обраде;

- израду експерименталних софтверских алата, као и могућности примене таквих и сличних решења.

У истраживању се користе методе засноване на правилима, насупрот статистичким методама које се такође користе у обради природних језика. Коришћена је методологија локалних граматика представљених коначним трансдукторима у форми синтаксних графова и уз подршку електронских речника.

Као технолошка основа за подршку наведеној методологији коришћен је систем *Unitex* који је развиен у *LADL*-у (*Laboratoire d'Automatique Documantaire et Linguistique*), у оквиру Универзитета Марн-ла-Вале, Француска, као и електронски речници развијени у оквиру Групе за језичке технологије Универзитата у Београду.

Формалним методама су посебно детаљно описане упућујуће фразе и структура правног акта. На основу тога је изграђен софтверски алат који за дати чисти текст закона формира приказ текста тако да се прегледно изражава препозната структура и мрежа упућивања, омогућава навигација кроз упућивања и приказују се идентификовани пропусти у структури и упућивањима.

Дисертација обухвата 214 страна, а у оквиру тога 9 поглавља (152 стране), списак коришћене литературе (8 страна, 76 библиографских јединица) и 4 прилога (54 стране). У дисертацији укупно има 105 слика, 24 дијаграма и 10 табела. Поглавља дисертације су:

1. Увод (7 страна).
2. Правни текст и природно-језичка обрада (14 страна, 1 слика).
3. Корпус (37 страна, 15 слика, 6 табела).
4. Упућујуће фразе (34 стране, 4 слике, 10 дијаграма, 1 табела).
5. Логичка структура (12 страна, 2 слике, 12 дијаграма, 1 табела).
6. Синтаксна анализа (8 страна, 6 слика, 1 табела).
7. Контролни документ (7 страна, 9 слика).
8. Израђени софтверски алати (26 страна, 12 слика, 2 дијаграма).
9. Закључак и даљи рад (7 страна).

Прилози дисертације су:

А. Синтаксни графови (15 страна, 56 слика).
Б. CasSys формат датотеке (1 страна).
В. XSLT specifikacija (2 стране).
Г. Списак закона у корпусу (36 страна, 1 табела).

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дисертације мр Небојше Васиљевића представљени су предмет, циљ и методологија истраживања. Предмет истраживања ове дисертације су прописи (закони и подзаконска акта) и њихов језик, који представља специфичан подјезик у оквиру административног функционалног стила, за који се обично сматра да представља формални сегмент језика са врло ограниченим репертоаром лексичких и синтаксичких структура. Језик прописа одликује, с једне стране, прецизно утврђена логичка структура јер је текст закона подељен у чланове, ставове и тачке по правилима која су строго утврђена. С друге стране, и поред тога што су језичке конструкције у овом подјезику понекад сличне формалним језицима, оне су прожете слободом природно-језичког изражавања, тако да оно што на први поглед представља поједностављену (формалну) структуру, заправо може бити и сложеније јер задржава наслеђену сложеност природног језика и истовремено уводи специфична методолошка правила за израду прописа. Кандидат мр Небојша Васиљевић је поставио следеће циљеве истраживања: (а) да на основу припремљеног корпуса опише језичке особености правних прописа на српском језику; (б) да на формалан начин опише посебна језичка правила која се користе у прописима; (в) да реализује у виду почетне фазе обраде текста прописа синтаксну анализу њихове формалне структуре; (г) да изгради експериментални софтвер за припрему текстова прописа и навигацију кроз њих. Кандидат се у свом раду ослања пре свега

на методе корпусне анализе које користи за опис структуре текстова прописа на српском језику. Осим тога, он у овом уводном поглављу образлаже зашто се у свом раду опредељује за методе засноване на правилима, конкретно за методе засноване на примени коначних аутомата и електронских речника, насупрот врло присутним статистичким методама обраде природних језика. Наиме, с обзиром на постављене циљеве истраживања, који обухватају анализу специфичних језичких конструкција у правном језику, методе засноване на правилима су далеко примерније.

У другом поглављу „Правни текст и природно-језичка обрада“, кандидат даје преглед главних задатака и досадашњих резултата у области природно-језичке обраде правних текстова, укључујући истраживања, развијене системе и софтверске алате који су на располагању. Кандидат, пре свега, показује да су кључне подобласти обраде природног језика и вештачке интелигенције, као што су: претрага текста, екстракција информација, аутоматска коректура, аутоматско превођење и аутоматско резоновање, присутне и у обради правних текстова ради, на пример, њихове претраге у циљу проналажења релевантних одредби прописа, подршке њиховој изради у складу са обавезним и уобичајеним правилима за одговарајућу врсту аката, анализе и конструкције правне аргументације, и сл. Кандидат затим прецизира структуру сваког задатка природно-језичке обраде у којој препознаје аналитичку фазу у којој је обрада усредсређена на извођење информација из улазног текста (анотирање улазног текста) и синтетичку фазу у којој се формира излазни резултат на основу претходно изведених информација. На почетку аналитичке фазе обраде се обично врши природно-језичка анализа, да би се наставило са (а) непосредном синтезом крајњег резултата; (б) додатном анализом без напредне семантичке обраде или (в) додатном напредном семантичком обрадом. Као резултат овог разматрања кандидат позиционира своје истраживање у сложеној мрежи разних видова природнојезичке обраде. Кандидат затим даје крађи преглед досадашњих истраживања у области природно-језичке обраде правних текстова истичући да је потреба за таквом обрадом нарочито изражена у прецедентним правним системима попут америчког, канадског и британског, где су извор права и све претходне одлуке судова. Посебна пажња се посвећује неколицини радова чија је тема ближе повезана с кандидатовим истраживањем – анализом и детекцијом упућујућих фраза – у сличном (правна акта) или различитом контексту (судске одлуке). На крају овог поглавља кандидат даје преглед најпознатијих општих алата за обраду природних језика који се могу мање или више успешно уклапати у разноврсне задатке, те даје крађи приказ алата *Unitex* одабраног за потребе свог истраживања.

У трећем поглављу „Корпус“ кандидат мр Небојша Васиљевић је, пре свега, описао корпус закона који је коришћен у истраживању и који се састоји од 681 законског текста који су донети у периоду од почетка 2005. до маја 2014. године. Сви ови текстови су преузети са званичног веб-сајта Народне скупштине Републике Србије. Списак свих законских текстова из корпуса је наведен у прилогу Г, заједно са неким основним бројчаним показатељима. Кандидат, даље, описује које методе и процедуре су коришћене за трансформацију текстова из оригиналног облика (погодног за штампање) у облик погодан за аутоматску обраду (на начин описан у претходном поглављу). На основу овако припремљеног корпуса, кандидат је израчунао и представио основне статистичке показатеље корпуса законских текстова и упоредио их са одговарајућим показатељима Корпуса савременог српског језика (СрпКор). Ово поређење је показало у којој мери законски текстови прате лексичке карактеристике општег језика, а у којим сегментима се појављују специфичности (на пример, висока учесталост групе облика речи „овог”, „члан”, „члана”, „закона” и „става” у односу на њихово појављивање у општем језику). Кандидат затим описује формат и садржај електронских речника српског језика и њихову улогу у обради корпуса. Детаљно је обрађен проблем непрепознатих речи и улога електронских речника за његово превазилажење. На крају мр Небојша Васиљевић представља анализу препознавања именованих ентитета (имена особа, геополитичких назива, организација, временских и бројчаних израза) у корпусу законских текстова која је подржана електронским речницима и локалним граматикама, а која илуструје специфичност законских текстова у односу на општи језик (нпр. новинске текстове).

Најважнији резултати дисертације мр Небојше Васиљевића изложени су у поглављима 4-7. У четвртом поглављу „Упућујуће фразе“ кандидат је детаљно истражио и прецизно описао

стуктуру фраза којима се у прописима упућује на делове истог или другог прописа. Због учесталости коришћења упућивања у прописима, као и важности упућивања за разумевање правних норми, кроз праксу израде прописа су се и код нас и у другим земљама успоставили стандарди у начину изражавања упућивања. Ови стандарди покривају већину форми упућивања, али не увек довољно детаљно, а такође се у пракси припреме нових прописа ови стандарди углавном поштују, иако не увек. Стога је кандидат, пре него што је приступио формализацији синтаксних правила за упућујуће фразе у прописима Републике Србије, утврдио основна начела којих ће се придржавати: (а) формално установљена правила; (б) преовлађујућа пракса у савременим законима (на основу израђеног корпуса прописа); (в) општа језичка правила; (г) логичност правила; (д) разумљивост и једноставнос правила; и (ђ) једнозначност правила. За прецизан опис структуре упућујућих фраза кандидат је осмислио дијаграме који, с једне стране, олакшавају праћење изградње овог описа, а с друге стране, олакшавају превођење формалног описа упућујућих фраза у синтаксне графове чија је улога препознавање упућујућих фраза у анализираном тексту прописа. Кандидат је поступно описао структуру упућујућих фраза, полазећи од оних једноставнијих („елементарне упућујуће фразе“) и водећи рачуна о дозвољеним варијацијама – упућивање на акт, упућивање на делове акта, специфичност нумерације код измена и допуна прописа, повратно упућивање (упућивање на класификациону јединицу у којој се налази сама упућујућа фраза) и просто једноструко и вишеструко упућивање (код кога се наводи више нумерација). За дефинисање сложенијих упућујућих фраза, кандидат уводи појмове затворених упућујућих фраза код којих је упућивање потпуно одређено и отворених упућујућих фраза код којих упућивање није до краја прецизирано. У прописима се, по правилу, користе затворене упућујуће фразе али је ово разликовање битно ради прецизирања сложених форми упућивања које се формирају слагањем једноставнијих упућујућих фраза, и то на два начина: хоризонталним слагањем (набрајањем) и вертикалним слагањем (додатним одређивањем контекста отворене упућујуће фразе). На основу овог формалног описа упућујућих фраза кандидат је израдио колекцију синтаксних графова применљивих у оквиру изабраног програмског система *Unitex* и подржаних електронским речницима (сви синтаксни графови ове колекције садржани су у прилозима А.1 и А.2). На крају овог поглавља кандидат приказује најједноставнију примену израђених синтаксиних графова, израду конкорданци упућујућих фраза. Кандидат илуструје као посебну погодност синтаксних графова њихову прилагодљивост, јер се они могу подесити да се више или мање строго придржавају постављених правила, према потреби предвиђене примене.

У петом поглављу „Логичка структура“ кандидат мр Небојша Васиљевић је сличним методама као за упућујуће фразе прецизно описао логичку структуру текста прописа. Под логичком структуром акта се подразумева његова подела на класификационе јединице (логичка структура текста је превод устаљеног енглеског термина *logical layout*). По правилу, основна класификациона јединица прописа је члан, док обимнији прописи (најчешће они који имају више од 20 чланова) могу имати и шире класификационе јединице, и то: део, главу, одељак и пододељак. При решавању овог задатка, кандидат је имао у виду и крајњу примену препознавања логичке структуре прописа коришћењем одабраног програмсог система *Unitex*. Наиме, овај систем није намењен пуној синтаксној анализи текста, али је зато врло успешан у препознавању мањих структура у локалном окружењу (такозваним *локалним граматикама*). У конкретном случају препознавања логичке структуре, то значи да су целине текста које представљају део, главу и члан сувише крупне да би биле јединице препознавања, па је кандидат стога дао нешто другачији поглед на структуру прописа, у коме се текст прописа види као секвенца мањих целина које ће се касније имплементирати као засебне јединице препознавања. При формализацији оваквог погледа на логичку структуру прописа кандидат је водио рачуна да у што већој мери смањи грешке при препознавању, које су неминовне код препознавања на локалном нивоу. Користећи формализам дијаграма уведен у претходном поглављу кандидат је затим описао две основне (локалне) јединице препознавања – структуру блока изнад члана и структуру члана. У структури блока изнад члана је обавезно појављивање ознаке члана, чему претходе опциона појављивања ознаке дела, ознаке главе, ознаке одељка и ознаке пододељка са својим правилима нумерисања. Структура члана се састоји од ставова, тачка, подтачака и алинеја са својим правилима нумерисања и разграничавања. Као и у случају

упућујућих фраза, кандидат је на основу овог формалног описа логичке структуре прописа израдио колекцију синтаксних графова применљивих у оквиру програмског система *Unitex*, а сви синтаксни графови ове колекције садржани су у прилогу А.4.

Шесто поглавље „Синтаксна анализа“ је праткичан наставак претходног поглавља и у њему кандидат образлаже како се у датом тексту може ефективно одредити његова структура која произилази из претходно дефинисаних синтаксних правила коришћењем програмског система *Unitex*. Описана је методологија синтаксне анализе логичке структуре и упућујућих фраза, до нивоа имплементације у конкретним алатима, као и форма резултата синтаксне анализе. Конкретно, синтаксна анализа се обавља у две фазе. У првој фази се у тексту обележавају лексичке етикете коришћењем каскаде трансдуктора, који представља колекцију трансдуктора који се примењују редом један за другим, при чему сваки до њих користи резултате претходних трансдуктора. Лексичке етикете омогућавају да се неки део текста експлицитно обележи као недељива јединица обраде, или токен, и да му се доделе по вољи синтаксна, семантичка и флективна својства; на пример, неке лексичке етикете за синтаксну анализу текста прописа су *tagNumeracija* и *tagUpucivanjePNStav*. У другој фази се лексичке етикете, које су погодне за рад у првој фази, замењују уобичајеним *XML* етикетама које се могу диркетно применити у разним другим апликацијама. У овој фази кандидат је користио уобичајену *XSLT* технологију за трансформацију резултата синтаксне анализе *Unitex*-ом у циљни *XML* формат. Опис трансформације у *XSLT* нотацији кандидат је дао у прилогу В.

У седмом поглављу „Контролни документ“ мр Небојша Васиљевић је описао алат за формирање контролног документа на основу чистог текста прописа. Алат представља прототип софтверског решења применљивог у пракси и сублимира претходно описане резултате истраживања. Основна сврха контролног документа је да на прегледан начин прикаже који елементи логичке структуре документа су препознати и на који начин, да прикаже детаље мреже упућивања, као и да помогне у уочавању техничких пропуста и методолошких несагласности при изради прописа. Израда контролног документа заснива се на резултатима синтаксне анализе описане у претходном поглављу, а кандидат се определио за *XML* формат (са неким додатим *XHTML* елементима) који омогућава лако приказивање у веб прегледачима. У овој фази рада кандидат је користио *CSS* спецификацију, програмски језик *Java* и стандардну подршку *Java* платформе за обраду *XML* докумената. Генерисање контролног документа представља практичан алат који може помоћи при изради и редакцији закона, као и за анализу раније израђених закона али се може користити и као подршка анотацији текста приликом припреме за унос у базе података које садрже текстове прописа.

У осмом поглављу „Израђени софтверски алати“ кандидат је описао више софтверских алата које је развио за потребе овог истраживања, како алата који решавају практичне потребе истраживања, тако и алата који имплементирају резултате истраживања. Треба посебно истаћи да се приликом развоја ових алата кандидат руководио начелом шире употребљивости, то јест могућности каснијег коришћења и изван овог истраживања, па се у том смислу алати које је кандидат произвео могу поделити у три групе: (а) без шире употребљивости — алати који су по својој природи везани за специфичан случај коришћења; (б) потенцијална шира употребљивост — алати који су израђени тако да су употребљиви и изван истраживања, али није развијен готов производ путем кога би тај алат био ефективно доступан ширем кругу корисника; (в) ефективна шира употребљивост — алати који су ефективно доступни ширем кругу корисника путем одговарајућег готовог производа. Већина алата које је кандидат развио спадају у групу (б) што значи да се њихова употребљивост не завршава с овим истраживањем, већ ће се они уз извесну дораду моћи користити и за друге задатке. У овом поглављу кандидат је описао и поједине техничке детаље имплементације.

У деветом поглављу „Закључак и даљи рад“ мр Небојша Васиљевић је истакао главне резултате истраживања и приказао више могућих праваца даљег рада. Као главне резултате истраживања кандидат је истакао формалан опис структуре текста прописа, прецизан опис језичких правила за упућујуће фразе и софтверски алат за формирање контролног документа на основу чистог текста прописа. Кандидат истиче да ово истраживање отвара многобројне могућности за развој обраде правних текстова на српском језику. Неки од непосредних задатака који се јављају на истраживачком нивоу су проширивање анализе упућујућих фраза у

делу упућивања на друге законе (која обухвата и препознавање њихових имена), формална репрезентација знања које је исказано одредбама прописа коришћењем одговарајућих формалних језика и проширивање анализе логичке структуре правног теста на препознавање цитата, што је пре свега значајно код измена и допуна те код аутоматског формирања пречишћеног текста на основу одредаба о изменама и допунама. Такође важан задатак за будући рад је израда и одржавање разноврсних језичких ресурса (корпуса и раченика) за правне текстове. На нивоу имплементације постоје бројне могућности за надоградњу контролног документа и његово прерастање у у алат за проверу текста прописа који се израђује и као такав може бити проширен са правописном провером, као и механизмом за сугерисање исправки.

На крају дисертације мр Небојша Васиљевић је приложио четири додатка:
А. Синтаксни графови (15 страна, 56 слика). Приказани су ситаксни графови за систем *Unitex* у којима на основу резултата из поглавља 4 и 5.

Б. CasSys формат датотеке (1 страна). Дата је спецификација трансдукторских каскада које су описане у поглављу 6 у CasSys формату датотеке.

В. XSLT спецификација (2 стране). Дата је XSLT спецификација путем које се резултат примена каскада из прилога Б преводи у циљни XML формат синтаксног дрвета

Г. Списак закона у корпусу (36 страна, 1 табела). Приказана је табела са 681 текстом закона и основним статистичким подацима за сваки текст.

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.

Д. Витас, Н. Васиљевић и Ц. Крстев, „Информатички поглед на корпус закона републике Србије”, Српски језик – студије српске и словенске, св. 19, стр. 377-394, Београд, 2014. УДК 911.163.41`322.2 811.163.41`373.611

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

Nebojša Vasiljević, “Multidocument concordances in Unitex”, In *Natural Language Processing for Serbian – Resources and Applications,* eds. G. Pavlović-Lažetić, C. Krstev, I. Obradović, D. Vitas, Faculty of Mathematics, University of Belgrade, Belgrade, 2014.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је кандидат мр Небојша Васиљевић остварио циљеве зацртане у пријави дисертације. Кандидат је припремио обиман корпус законских текстова на српском језику који је користио у свим фазама истраживања и који остаје као вредан ресурс за даља истраживања у области аутоматске обраде правних текстова. Кандидат је детаљно и на формалан начин описао један сегмент правних текстова – њихову логичку структуру и синтаксну структуру упућујућих фраза – и израдио алате који на основу овог формалног описа врше синтаксну анализу. На основу резултата овог истраживања кандидат је изградио експериментални софтвер за контролу постојећих и припрему нових текстова прописа који је прилагођен потребама потенцијалних корисника. Све детаље изграђених алата и софтвера кандидат је у дисертацији ставио на располагање чиме је омогућио репродукцију резултата истраживања као и будуће надградње.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је мр

Небојша Васиљевић користио релевантну и савремену литературу, те да је постављене проблеме обрадио детаљно и сагледавајући их из разних углова. Овим радом Небојша Васиљевић је отворио једно ново поље истраживања у области обраде српског језика а будућим истраживачим ставио на располагање изузетно значајне ресурсе и алате за даљи рад.

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 мр Небојша Васиљевић у својој дисертацији *Аутоматска обрада правних текстова на српском језику* 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 *Аутоматска обрада правних текстова на српском језику* кандидата мр Небојше Васиљевића и упути га Већу за друштвено-хуманистичке науке Универзитета у Београду, како би кандидат био позван 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 Божо Ћорић, редовни професор
   Филолошки факултет Универзитета у Београду

4. др Добросав Миловановић, ванредни професор
   Правни факултет Универзитета у Београду

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