

Образац за пријаву техничког решења¹

Назив техничког решења	Детектор штампарских и правописних грешака за српски језик (<i>anSpellChecker</i>)
Аутори техничког решења	Милана Бојанић, Стеван Острогонац, Милан Сечујски, Наташа Вујновић Седлар, Синиша Сузић
Категорија техничког решења	Ново техничко решење – прототип (M85)

За кога је рађено техничко решење и у оквиру ког пројекта МПНТР:

Ово техничко решење урађено је на технолошком пројекту “Развој дијалошких система за српски и друге јужнословенске језике” (ТР32035, 2011-14) на Факултету техничких наука и у предузећу АлфаНум у Новом Саду.

Ко користи техничко решење:

Предузеће АлфаНум у Новом Саду као партиципант на пројекту ТР32035.

Година када је техничко решење урађено:

Развој техничког решења је започет у јануару, а довршен крајем марта 2012. године. Примена техничког решења је почела од 01.04.2012. године у предузећу АлфаНум.

Ко је прихватио-примењује техничко решење:

Ово техничко решење иницијално је реализовано за потребе развоја говорних и језичких ресурса у предузећу АлфаНум, користи се и у развојном тиму на Факултету техничких наука, а предузеће АлфаНум планира да га, по завршетку тестирања прототипа (до краја 2012. године) презентује као нови производ за потребе новинара, публициста и свих који пишу много текстова – да им уштеди време за тражење штампарских грешака у текстовима које су написали.

¹ У складу са одредбама *Правилника о поступку и начину вредновања, и квантитавном исказивању научноистраживачких резултата истраживача*, који је 21.03.2008. године донео Национални савет за научни и технолошки развој Републике Србије («Службени гласник РС», бр. 38/2008).

Како су резултати верификовани (од стране кога тела):

- 1) Техничко решење је реализовано у Лабораторији за акустику и говорне технологије на Факултету техничких наука, имплементирано и испитано на развојним системима у предузећу АлфаНум, а сада се налази у свакодневној експлоатацији у предузећу АлфаНум у Новом Саду као прво такво техничко решење у Србији и/или ширем региону.
- 2) Техничко решење базирано је на више научних радова пријављених као резултати на пројектима технолошког развоја Министарства у претходном периоду, између осталог:
 - Milan Sečujski: "Development of language resources for the Serbian language required for part-of-speech tagging", Chapter in book: „Speech and Language: Interdisciplinary Research III“, Eds.: S. T. Jovičić, M. Sovilj, 2009, LAAC and IEPPS, Belgrade, ISBN 978-86-81879-27-6, pp. 125-139.
- 3) Дато је писано мишљење два рецензента-експерта из области техничког решења:
 - a. Проф. др Снежана Гудурић, Филозофски факултет у Новом Саду,
 - b. Доц. др Александра Јовановић, Електронски факултет у Нишу.
- 4) Наставно-научно веће Департмана за енергетику, електронику и телекомуникације и Факултета техничких наука, на основу мишљења рецензената и приложених доказа, издали су Уверење о признавању техничког решења које потврђује да оно испуњава све услове да буде признато као прототип (M85), у складу са Правилником Министарства.

На који начин се користи (кратак опис):

Апликација *anSpellChecker* представља техничко решење које омогућава детекцију штампарских грешака у текстуалним садржајима на српском језику. Ослањајући се на морфолошки речник српског језика који садржи преко четири милиона изведених облика речи, *anSpellChecker* одређује речи у тексту које највероватније садрже грешку.

Корисници прослеђују апликацији текстуални садржај у ANSI, UTF8 или UNICODE формату. Затим се аутоматски креира фајл са садржајем у коме су обележене речи које нису пронађене у морфолошком речнику. Корисници тада могу на једноставан начин прегледати обележене речи и исправити евентуалне грешке. Наравно, међу обележеним речима могу се наћи и оне које су исправно написане, али не постоје у речнику. То су углавном страна имена или кованице попут речи „шездесетседмогодишњакиња“.

Ова апликација је иницијално креирана за потребе сарадника предузећа АлфаНум који су задужени за писање транскрипција за аудио-фајлове из говорне базе која се користи за обуку система за препознавање говора. С обзиром да овакве транскрипције садрже посебне ознаке за различите врсте акустичких или лингвистичких неправилности у

материјалу (нпр. за речи које су изговорене разумљиво али са изостављеним гласом попут „ноги“ уместо „многи“), *anSpellChecker* је прилагођен оваквим садржајима у смислу тумачења тих ознака у зависности од којих се поједине речи изузимају из контроле исправности записа, јер су свакако обележене као неправилне.

Опис техничког решења:

Детектор штампарских и правописних грешака за српски језик (*anSpellChecker*)

Детектор штампарских грешака *anSpellChecker* представља прво напредније решење проблема отклањања грешака у текстовима на екавском наречју српског језика. Ово техничко решење је настало, између осталог, захваљујући резултатима постигнутим на претходним пројектима Министарства, током којих је развијен акценатско-морфолошки речник српског језика као језички ресурс од ширег значаја.

Обим и свеобухватност садржаја овог речника даје кључну предност апликацији *anSpellChecker* у детекцији грешака у текстовима у односу на сличне постојеће апликације, јер је експлицитно решен проблем постојања великог броја изведених облика речи у српском језику – речник садржи променљиве врсте речи у свим њиховим појавним облицима. Такође, ова апликација подржава у извесној мери и ијекавско наречје српског језика, јер поседује базу најчешћих речи на ијекавици које се појављују у текстовима на екавском наречју. Ово важи и за често употребљаване речи енглеског језика, попут „Facebook“, „system“, „Windows“ итд.

Постојање веома обимне базе речи оправдава концепт детекције грешака простом претрагом и поређењем са уносима речника. Иако још увек није развијен модул који ће грешке детектовати уз додатну анализу граматичке структуре реченица, већ ова основна верзија апликације даје веома добре резултате у детекцији штампарских грешака и ретко даје „лажне узбуне“.

Област на коју се техничко решење односи:

Техничко решење припада области обраде природног језика и представља спој лингвистичких знања и информационо-комуникационих технологија.

Проблем који се техничким решењем решава:

Висок степен инфлективности српског језика онемогућавао је у протеклом периоду развој интелигентнијих детектора штампарских и правописних грешака у текстуалним садржајима. Све сличне апликације које су развијене за српски језик базирале су се на стандардним речницима који садрже само каноничке (основне) облике речи. Ове апликације могле су понудити кориснику да изабере једну од понуђених речи на

основу неколико почетних слова, али су све понуђене речи биле лишене информација о морфолошким категоријама попут рода, броја, лица или падежа. Услед тога, њихов потенцијал у олакшавању детекције грешака при куцању текста био је веома ограничен.

Са увођењем информација о изведеним облицима речи у речник (нпр. поред речи „огањ“ у речнику постоје и облици „огња“, „огњу“, „огњем“) значајно је проширен скуп секвенци слова које се сматрају речима српског језика. Секвенце слова које се не налазе у овом скупу највероватније су речи које садрже грешку. Другим речима, док би се код ранијих верзија детектора штампарских и правописних грешака све нпр. именице које нису записане у основном облику третирале као погрешно написане (нпр. „девојкама“), код *anSpellChecker*-а би то биле само речи које се, из било ког разлога не поклапају ни са једним изведеним обликом неке конкретне именице (нпр. „девојкама“ или „девојкма“).

Стање решености тог проблема у свету:

Концепт детекције штампарских грешака на основу речника даје веома добре резултате за језике мањег степена инфлективности, попут енглеског. За језике са веома комплексном морфологијом потребно је користити речнике који садрже све изведене облике речи. Ипак, мали је број језика, попут нпр. руског, за које адекватни речници постоје. За детаљнију анализу граматичких конструкција при детекцији штампарских грешака користе се углавном концепти коначних аутомата или модела језика заснованих на *N*-грамима. Ове методе се, међутим, углавном користе код језика са једноставнијом морфологијом, због рачунске захтевности алгоритама претраге.

Објашњење суштине техничког решења и детаљан опис са карактеристикама, укључујући и пратеће илустрације и техничке цртеже (техничке карактеристике):

Програм *anSpellChecker* креиран је као самостална (енг. *standalone*) апликација за детекцију штампарских грешака. Тренутно је подржана његова примена на тзв. *plain text* формату. Основну функционалност овог програма приказује блок шема на сл. 1.

Први блок, који служи за анализу текста, има две основне функционалности. Прва функционалност односи се на издвајање речи из текстуалног садржаја који садржи и знаке интерпункције, као и цифарске записе и друге специјалне карактере који не треба да буду укључени у проверу исправности записа. Друга функционалност, која се активира по потреби, односи се на одлучивање о томе које ће речи бити подвргнуте детекцији грешака. С обзиром на то да је *anSpellChecker* иницијално био намењен за потребе писања транскрипција за аудио фајлове који сачињавају говорну базу за обуку препознавача говора компаније АлфаНум, у оквиру блока за анализу текста имплементирана је и интерпретација ознака које се у оваквим транскрипцијама користе. Неке од ових

ознака односе се на акустичка оштећења говорног сигнала, а друге на „речи“ које не садрже никакве акустичке сметње, али не спадају у скуп валидних секвенци слова (односно секвенци са семантичким значењем) у српском језику.



Слика 1 – Блок шема детектора штампарских грешака

Ради илустрације, у наставку је дат изглед записа једне реченице из текстуалног документа који садржи транскрипције за аудио фајлове:

12_10_2012_15_30.wav Јуче сам, шетајући са ЕС ДЕ (С. Д.) [сватио] (схватио) да се не <желим> (желим) вратити кући.

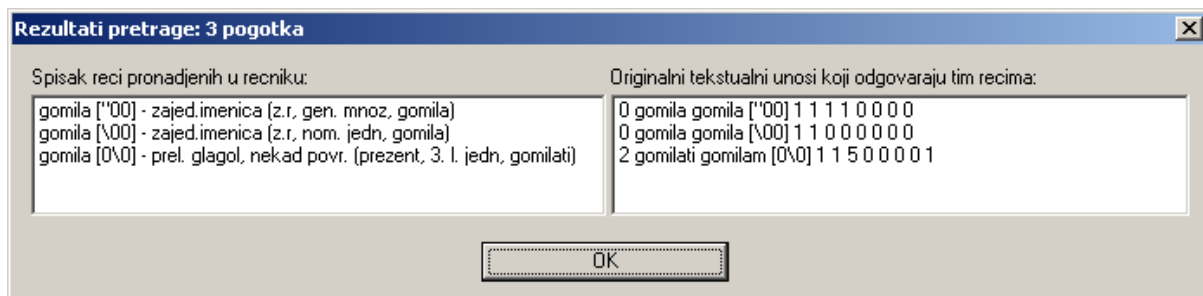
Задатак модула за анализу текста јесте да на основу посебних ознака у транскрипцији донесе одлуку о томе које ће речи бити прескочене у процесу детекције штампарских грешака. На пример, секвенце писане искључиво великим словима попут „ЕС“ и „ДЕ“ не прослеђују се модулу за детекцију грешке, јер су то транскрипције које се користе искључиво за обуку препознавача говора, а иза њих следи у обичним заградама запис у форми у којој би требао да постоји у тексту. Угласе заграде („[]“) у транскрипцијама означавају ситуације када је реч изговорена без присуства било каквих акустичких сметњи, при чему је сваки глас правилно артикулисан, али сама реч не постоји у стандардном српском језику. Запис између ознака „<>“ служи за лабелирање лоше артикулације гласова или шума у сигналу. Након посебно означених записа (са заградама „<>“ и „[]“), у „обичним“ заградама дат је запис речи које би требало да се налазе на местима оштећења и ти записи се прослеђују модулу за детекцију грешака.

Назив аудио фајла се, наравно, изузима из детекције грешака. Дакле, секвенца која би у наведеном примеру била прослеђена наредном модулу гласила би:

Јуче сам шетајући са СД схватио да се не желим вратити кући

Ово су само неки од примера ознака које се користе у писању транскрипција. Све ознаке су, наравно, недвосмислене и имају прецизно дефинисано значење у датом контексту. Уколико се, међутим, *anSpellChecker* користи за проверу исправности текстова који нису везани за транскрипције (што се лако детектује изостанком назива аудио фајла на почетку реченице), заграде и сличне ознаке се третирају као стандардни интерпункцијски симболи.

Модул за детекцију грешака имплементира брзу претрагу акценатско-морфолошког речника. При учитавању речника, сваком уносу се додељује идентификациони број и креира се табела пресликавања између речи и идентификационих бројева, односно табела парова реч-идентификациони број. Затим се, за сваку прослеђену реч, претражује табела и, уколико се нађе бар једно поклапање, шаље се информација да је реч исправна. У супротном, ако се дође до краја табеле, датој речи се додељује „невалидан“ идентификациони број. За једну реч је могуће пронаћи и више поклапања, што је илустровано на слици 2.



Слика 2 – Приказ резултата претраге речника за реч „гомила“

Списак оригиналних уноса у речник који је дат на слици 2 са десне стране приказује информациони садржај акценатско-морфолошког речника. На слици 3 је приказан кратак извод из овог речника.

Прва колона, која је попуњена цифрама (од 0 до 9) означава врсту речи на коју се односи дати унос (у конкретном примеру су са 0 кодоване именице, а са 2 глаголи). Затим следи колона у којој су дати записи речи у каноничким формама. Ове форме се још називају и леме. Након њих следе речи у изведеним облицима (које су од интереса за детектор штампарских грешака). Следи део који се односи на акценатске информације (у угластим заградама). Након тога, сваки унос садржи по осам цифара које означавају вредности морфолошких категорија речи попут рода, броја, лица, падежа итд.

```
2 mesti meti [/0] 1 0 3 1 0 0 0 1
0 metak metka ["0] 0 1 3 0 0 0 0 0
0 metak metka ["0] 0 1 1 0 0 0 0 0
0 metak metke ["0] 0 1 3 1 0 0 0 0
0 metak metkom ["0] 0 1 5 0 0 0 0 0
0 metak metku ["0] 0 1 4 0 0 0 0 0
0 metak metku ["0] 0 1 2 0 0 0 0 0
0 metla metla [\0] 1 1 0 0 0 0 0 0
0 metla metlama [\00] 1 1 5 1 0 0 0 0
0 metla metlama [\00] 1 1 2 1 0 0 0 0
0 metla metle [\0] 1 1 4 1 0 0 0 0
0 metla metle [\0] 1 1 3 1 0 0 0 0
0 metla metle [\0] 1 1 0 1 0 0 0 0
0 metla metle [\0] 1 1 1 0 0 0 0 0
0 metla metli [\0] 1 1 1 1 0 0 0 0
0 metla metli [\0] 1 1 2 0 0 0 0 0
```

Слика 3 – Извод из акценатско-морфолошког речника српског језика

Наравно, пошто различите врсте речи имају и различите скупове морфолошких категорија, нису увек сви елементи уноса у употреби, али увек постоје подразумеване вредности како би формат записа био јединствен. Ово омогућава ефикасну компресију података помоћу тзв. *run length* алгоритама. *Run Length Encoding (RLE)* алгоритам² врши компресију података без губитака тако што низове истих вредности (нпр. карактера у тексту) унутар неког документа замењује са два податка: први представља број понављања, а други представља саму вредност (карактер) која се понавља. Висок степен компресије се постиже уколико документ садржи веома дуге низове идентичних вредности (нпр. текстуални документ са поновљеним карактерима или слика која се састоји од већих једнобојних површина).

Све речи за које се установи да не постоје у речнику бивају обележене додавањем префикса „#“. На тај начин корисник може једноставном претрагом да прегледа текст и направи евентуалне измене.

Опција која је креирана са циљем да се омогући унапређење апликације допуњавањем садржаја речника јесте снимање списка различитих речи које су обележене као погрешно написане у засебан документ. Овакав извештај служи у каснијој ручној претрази чији је циљ да се прикупе речи које су исправне, али још увек не постоје у речнику те су означене као грешке.

Како је реализован и где се примењује, односно које су могућности примене (техничке могућности):

У актуелној верзији, *anSpellChecker* је дизајниран као апликација веома једноставна за коришћење. Позива се из командне линије навођењем јединог обавезног параметра –

² Held G., *Data Compression: Techniques and Applications, Hardware and Software Considerations*, second edition, John Wiley & Sons, New York, NY, 1987.

назива документа са текстуалним садржајем који је потребно преконтролисати. Други (опциони) параметар односи се на извештај о списку речи које су означене као грешке. Као резултат контроле грешака креира се нов текстуални документ са називом који се састоји од назива оригиналног документа и додатог суфикса „_fixed“. Осим из командне линије, детектор грешака се може активирати простим превлачењем иконице текстуалног документа на иконицу апликације. Тада не постоји могућност да се испише извештај о грешкама.

Што се тиче писма, сама претрага морфолошког речника захтева латинични текст, али се писмо детектује аутоматски (на основу текста) и врши се конверзија, уколико је потребна.

Програм *anSpellChecker* спада у групу алата који значајно могу олакшати рад када год је потребно вршити контролу исправности записа. Успешност детекције грешака код ове апликације је значајно боља у односу на алате са сличном функционалношћу због сасвим другачијег приступа проблему који је био главни мотив реализације овог техничког решења – новог производа предузећа АлфаНум.

Докази (прилози):

- Писано мишљење два рецензента-експерта из области техничког решења.
- Потврда предузећа АлфаНум.
- Уверење о признавању техничког решења Наставно-научног већа Факултета техничких наука, на основу мишљења рецензената и приложених доказа, издато након одговарајуће процедуре на Департману за енергетику, електронику и телекомуникације.

Нови Сад, децембар 2012. године.

Подносилац пријаве:

Проф. др Владо Делић
Руководилац пројекта TP32035

POTVRDA

Ovim potvrđujemo da je 01.04.2012. godine u preduzeću AlfaNum d.o.o. pušten u rad detektor štamarskih i pravopisnih grešaka za srpski jezik(*anSpellChecker*) koji smo sami razvili za potrebe razvoja govornih i jezičkih resursa. Rešenje je inicijalno instalirano za potrebe testiranja, ali ga preduzeće AlfaNum koristi i za sopstvene potrebe.

U Novom Sadu
18. januar 2013


Darko Pekar, dipl.ing., Direktor



РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	Детектор штампарских и правописних грешака за српски језик (anSpellChecker)
Аутори техничког решења:	Милана Бојанић, Стеван Острогонац, Милан Сечујски, Наташа Вујновић Седлар, Синиша Сузић
Реализатори:	Факултет техничких наука и АлфаНум у Новом Саду
Пројекти на којим је развијено:	“Развој дијалогских система за српски и друге јужно-словенске језике”, ТР32035 код МПНТР (2011-2014)
Област на коју се односи:	Електроника, телекомуникације и информационе технологије
Корисници:	Предузеће АлфаНум у Новом Саду
Категорија техничког решења:	Ново техничко решење – прототип (M85)

Подаци о рецензенту:

Име, презиме и звање:	Др Снежана Гудурић, редовни професор
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабрана у звање редовног професора 21.07.2009. год. на Филозофском факултету у Новом Саду, за у.н.о. Филолошке науке
Установа где је запослен:	Филозофски факултет Универзитета у Новом Саду

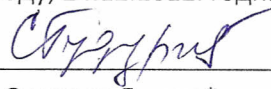
Стручно мишљење рецензента:

Резултат научно-истраживачког рада “Детектор штампарских и правописних грешака за српски језик (anSpellChecker)” испуњава услове за признање својства техничког решења и то као **ново техничко решење – прототип (M85)**, све у смислу одредби које се односе на техничка решења у Правилнику који је 21.03.2008. године донео Национални савет за научни и технолошки развој Републике Србије.

Образложење за техничко решење (ТР):

- Ово ТР представља програм за аутоматску детекцију штампарских и правописних грешака у текстовима на српском језику. Ослања се на морфолошки речник који омогућава решавање проблема постојања великог броја изведених облика речи у српском језику.
- Апликацији се даје назив текстуалног документа, а као резултат настаје нови документ у коме су обележене речи које садрже грешку. Осим екавског наречја, у извесној мери по-држано је и ијекавско наречје српског језика, као и најчешћи изрази на енглеском језику. Детекцију грешака могуће је вршити текстовима писаним и ћирилицом и латиницом.
- ТР је развијено коришћењем научних метода на пројектима МПНТР, претежно на сопственој опреми и без коришћења туђе патентне/лиценцене документације. Као резултат развојних активности овог ТР произашао је и одређен број научних радова.
- ТР се већ примењује код горе наведених корисника, дакле у земљи и иностранству, и представља прво решење овакве врсте (генерисање новог документа са означеним грешкама) које подржава и ијекавско наречје српског језика и које решава проблем често употребљаваних речи енглеског језика, а што није био случај са на изглед сличним, али ипак битно различитим, решењем које је тренутно у употреби за коректуру српских текстова (РАС).

У Новом Саду, 24.12.2012. године.


Др Снежана Гудурић

РЕЦЕНЗИЈА ТЕХНИЧКОГ РЕШЕЊА

Подаци о техничком решењу:

Назив техничког решења:	Детектор штампарских и правописних грешака за српски језик (anSpellChecker)
Аутори техничког решења:	Милана Бојанић, Стеван Острогонац, Милан Сечујски, Наташа Вујновић Седлар, Синиша Сузић
Реализатори:	Факултет техничких наука и АлфаНум у Новом Саду
Пројекти на којим је развијено:	“Развој дијалошких система за српски и друге јужно-словенске језике”, ТР32035 код МПНТР (2011-2014)
Област на коју се односи:	Електроника, телекомуникације и информационе технологије
Корисници:	Предузеће АлфаНум у Новом Саду
Категорија техничког решења:	Ново техничко решење – прототип (М85)

Подаци о рецензенту:

Име, презиме и звање:	Др Александра Јовановић, доцент
Ужа научна област за коју је изабран у звање, датум избора у звање и назив факултета:	Изабрана у звање доцента 15.03.2011. год. на Електронском факултету у Нишу, за у.н.о. Телекомуникације
Установа где је запослен:	Електронски факултет Универзитета у Нишу

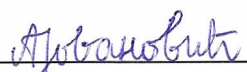
Стручно мишљење рецензента:

Резултат научно-истраживачког рада “Детектор штампарских и правописних грешака за српски језик (anSpellChecker)” испуњава услове за признање својства техничког решења и то као **ново техничко решење – прототип (М85)**, све у смислу одредби које се односе на техничка решења у Правилнику који је 21.03.2008. године донео Национални савет за научни и технолошки развој Републике Србије.

Образложење за техничко решење (ТР):

- *AnSpellChecker* омогућује аутоматску детекцију грешака у текстовима писаним на српском језику. Коришћењем морфолошког речника у великој мери је превазиђен проблем комплексне морфологије и високог степена инфлективности српског језика.
- Програму се прослеђује текстуални запис у коме апликација проналази и обележава грешке поређењем секвенци слова са уносима у речник. Секвенце које не постоје у речнику сматрају се грешкама. Подржано је екавско наречје српског језика, делимично и ијекавско, као и одређени број честих речи и фраза на енглеском језику.
- ТР је развијено коришћењем научних метода на пројектима МПНТР, претежно на сопственој опреми и без коришћења туђе патентне/лиценценце документације. Као резултат развојних активности овог ТР произашао је и одређен број научних радова.
- ТР се већ примењује код горе наведених корисника, дакле у земљи и иностранству, и представља прво такво решење проблема у региону.

У Нишу, 28.12.2012. године.


Др Александра Јовановић



УНИВЕРЗИТЕТ
У НОВОМ САДУ



ФАКУЛТЕТ
ТЕХНИЧКИХ НАУКА

Трг Доситеја Обрадовића 6, 21000 Нови Сад, Република Србија
Деканат: 021 6350-413; 021 450-810; Централa: 021 485 2000
Рачуноводство: 021 458-220; Студентска служба: 021 6350-763
Телефакс: 021 458-133; e-mail: ftndean@uns.ac.rs

ИНТЕГРИСАНИ
СИСТЕМ
МЕНАџМЕНТА
СЕРТИФИКОВАН ОД:



Наш број: 01.сл

Ваш број:

Датум: 2013-01-30

ИЗВОД ИЗ ЗАПИСНИКА

Наставно-научног већа Факултета техничких наука у Новом Саду, на 4. редовној седници одржаној дана 30.01.2013. године, донело је следећу одлуку:

-непотребно изостављено-

Тачка 15.1.40.: Питања научноистраживачког рада и међународне сарадње / верификација нових техничких решења

Одлука

На основу позитивног извештаја рецензената верификује се
техничко решење (M85) под називом:

ДЕТЕКТОР ШТАМПАРСКИХ И ПРАВОПИСНИХ ГРЕШАКА ЗА СРПСКИ ЈЕЗИК (ANSPELLCHECKER)

Аутори техничког решења: Милана Бојанић, Стеван Острогонац, Милан Сечујски, Наташа Вујновић Седлар, Синиша Сузић.

-непотребно изостављено-

Записник водила:

Јасмина Димић, дипл. правник

Тачност података оверава:
Секретар

Иван Нешковић, дипл. правник

Декан



Проф. др Раде Дорословачки