Дигитални речник говора југа Србије

УДК: 811.163.41'322, 811.163.41'374'282.3

Миљана Младеновић
ml.miljana@gmail.com
Универзитет у Београду, Математички факултет

АПСТРАКТ: Дигитални речник говора југа Србије представља прву целовиту реализацију дигиталне верзије једног речника дијалекта српског језика, генерисаног на основу *Речника говора јужне Србије*, аутора проф. др Момчила Златановића. Објављен је на адреси www.vranje.co.rs и иницијално је садржао 10.950 одредница. Ово је први ресурс на српском језику који, поред лингвистичких информација, обезбеђује и низ других: звучну информацију (изговор) свих одредница и примера употребе речи или фраза онако како се изговарају на дијалекту; графичке информације о географским локацијама употребе појмова коришћењем *Google Maps* и *Geocoding* сервиса; статистичке информације какве се не могу или се тешко могу добити употребом класичног речника о етимолошком пореклу одредница (колико их је из турског, персијског, латинског итд.), о појединим врстама речи, употреби речи, пореклу, значењу итд. као и повезивање речника и дељење његовог садржаја путем друштвених мрежа. Други значајан аспект речника је да ради као вики ресурс, то јест, да омогућава веб-корисницима да речник шире и допуњују на три начина: додавањем одредница које нису у речнику, увођењем примера примене и употребе постојећих појмова и коментарисањем и указивањем на нова значења, нове етимолошке предлошке и нове топонимске карактеристике везане за порекло и подручја употребе датих појмова. Развијени су алати који дају опште информације о тренду раста и употребе речника – о учесталости претраге појмова, тренутном броју одредница, броју корисника који раде на развоју речника итд.

КЉУЧНЕ РЕЧИ: призренско-тимочки дијалекат, врањски дијалекат, дигитални речник, речник дијалекта.

ДАТУМ ПРИЈЕМА РАДА: 26. фебруар 2014.
ДАТУМ ПРИХВАТАЊА РАДА: 4 мај 2014.

1. Увод

У овом раду представљен је дигитални речник варијетета српског језика који је у лингвистици познат као призренско-јужноморавски дијалекат.[1] То је, заправо, један део призренско-тимочких штокавских дијалеката српског језика, распрострањен у јужним деловима Србије.

Основ за изградњу дигиталног речника представља Треће издање речника Речник говора јужне Србије (Златановић, 2011). Сам аутор је на прикупљању појмова и примера у говору, као и самој изради папирне верзије речника, радио више од 10 година. Речи које се налазе у овом речнику аутор је највећим делом забележио радећи на терену и може се рећи да су у свакодневној употреби у Врању, Пољаници,

1 Дијалект или наречје је реч грчког порекла (διάλεκτος,, дијалектос) и односи се на варијетет једног језика који користи група људи особена по припадности одређеној географској области или социјалној, професионалној, етничкој или другој категорији. Дијалекат има свој речник који се мање или више разликује од основног, а може имати и своју граматику и фонологију.

Пчињи, Прешевској Моравици, Прешевској Црној Гори, Грделичкој клисури, Власини, Црној Трави и др. Географски посматрано, реч је о територији целог Пчињског и једног дела Топличког округа. У речнику се налазе и неке речи преузете из Путописа Хаџи-Анте Калиманца (Хаџи-Васиљевић, 1910) за који је истакнуто да је написан „правим и верним врањским говором". Известан број речи аутор је прибележио захваљујући интерним белешкама професора руског језика у Гимназији „Бора Станковић" у Врању, Тихомира Стефановића. Оне у речнику носе додатну ознаку (Т. С.).

За разлику од књижевног српског језика, дијалекат југа Србије садржи 32 гласа (фонеме). Словне ознаке за два додатна гласа избор су самог аутора. Један је „меки глас" који води порекло из старословенског и за њега је употребљена ознака (графема) **Ь**. Други је диграф дз који у овом речнику има словну ознаку **д̂з**. У говору јужне Србије у употреби су и речи страног порекла. Претрагом овог речника може се лако установити да има речи, пре свега, турског порекла, али исто тако и латинског, немачког, грчког, санскрита итд.

2. Дигитални ресурси и алати за проучавање дијалеката

Српски језик темељи се на штокавском дијасистему (језички систем изведен на основу особина језика неког макропростора) екавских и ијекавских дијалеката (Окука, 2008). Дијалекат је језички систем који има висок степен сличности са дијасистемом једног географског простора, а са друге стране има систем сопствених особина по којима се разликује од осталих дијалеката истог језика. Дијалекат могу чинити поддијалекти неког микропростора и минидијалекти (месни говори). Сваки дијалекат карактеришу географски простор на коме се користи и лингвистичке особености (фонолошке, морфолошке, лексичке и синтаксне).

Као један од дијалеката српског језика, призренско-тимочки дијалекат користи се на територији која на југозападу има границе према Албанији, на југу према Македонији, на истоку према Бугарској, а на северу се протеже до Сталаћа. Он се, због унутрашњих одступања (Ивић, 1985), дели на поддијалекте: призренско-јужноморавски, тимочко-лужнички и сврљишко-заплањски.

Са развојем информатике, проучавање дијалеката добија значајан подстицај у виду софтверских алата и дигиталних ресурса који се користе. Поред развоја дигиталних речника дијалеката (Karanikolas и др., 2013; Keymeulen и др., 2013; Ćavar и др., 2000), ради се на дигитализацији рукописних речника (Benacchio и др., 2012), на развоју алата за управљање дијалекатским речницима (Pereira и Gillier, 2012) и на софтверу за визуелизацију лингвистичких података и дијалекатских карата – атласа (Sibler и др., 2012; Petsas, 2009). Савремени географски информациони системи (ГИС) омогућују софистицирање и ефикасне анализе просторних података. Ипак, дуго времена поље лингвистике није било у фокусу истраживања географа. Када су лингвисти почели да користе ГИС технологије у изради лингвистичких атласа, дошло је до експанзије гране науке која се бави анализом географске дистрибуције и структуре језика – геолингвистике. Анализа географских информација релевантна за лингвистичка истраживања односи се на анализу и манипулацију подацима о простору, просторно-статистичку анализу и просторно моделирање (O'Sullivan и Unwin, 2010).

Циљ овога рада је да онлајн ГИС алате повежемо са дигиталним речником призренско-јужноморавског дијалекта да бисмо приказали географску распрострањеност дијалекта и како бисмо на нивоу сваке одреднице указали на географску локацију употребе исте.

3. Процес дигитализације речника

Први корак у дигитализацији *Речника говора јужне Србије* односио се на припремне радње за скенирање које представљају итеративни поступак одређивања оптималног односа величине излазне

датотеке и квалитета добијеног скенирањем. Поред оптимизације параметара, врло је важно скенирање извести професионалним високорезолуцијским скенерима. У овом поступку коришћен је *FUJITSU Image Scanner fi-Series fi-5220C* са оптичком резолуцијом до 600 *dpi*. Скенирано је укупно 550 страна у *tiff* формату. Формиране су 32 фасцикле по словима и извршена прерасподела датотека.

Други важан корак односио се на оптичко препознавање текста. У ту сврху коришћен је *Abby FineReader* 11 (Abby FineReader, 2011). Овај софтвер за оптичко препознавање карактера заснован је на техникама машинског учења. Да бисмо обезбедили ефикасну фазу учења, морали смо најпре да идентификујемо скуп писама који ће бити коришћен у процесу како учења тако и препознавања. *Abby FineReader 11* обезбеђује препознавање 168 природних језика (при чему препознаје оба српска писма: Cyrillic, Latin), четири вештачка (попут Esperanto и Interlingua језика) и седам формалних језика (C/C++, Јава, Паскал и сл.). У самом речнику срећемо: ћирилично писмо са акцентовањем (**пади́нче** – деминутив речи падина, **надлећа́ње** – од глагола надлетати, **џалдиса́** – побегнем, одјурим итд.), речи и изразе записане турским писмом (**güzel** – дотерати се, **perçem** – чуперак итд.), али и известан број речи датих латиничним писмом (најчешће су то латински називи биљака и гљива чији су локални називи дати на дијалекту: ви́лино кли́нче – ливадска печурка Marasmius oreades). Осим језика који природно користе латинично писмо, у овом речнику су и речи које потичу из грчког такође писане латинично. На пример: грч. **Faétōn** – непокривене кочије на четири точка, грч. **krommydi** – кромид, црни лук итд. Такође, јављају се и две графеме поменуте у уводном одељку: меки глас (**кла́дьнц** – извор воде, **бьбње́ње** – одјекивање и сл.) и диграф дз (**д̂зье́вни** – одјекује, **изд̂зьмбати** – појести халапљиво и сл.). Може се приметити да сва три писма садрже дијакритике. На основу добијене анализе употребе писама у самом речнику, у фази која је претходила учењу, истовремено су активирани учење и препознавање за *Serbian Cyrillic*, *Serbian Latin* и *Turkish*. С обзиром на то да основни скуп карактера ћириличног писма нема дијакритике, акцентовани знаци су уведени на два начина: проширењем *Serbian Cyrillic* скупа за слова **Á É Ó Ô á é ó ô у́** и креирањем шаблона (patterns) за слова **и́ р́ ь́** као и њихове одговарајуће *Normal*, *Bold* и *Italic* комбинације малих и великих акцентованих слова. На крају овог процеса креиран је шаблон ознаке за диграф са обрнутим бревисом **д̂з** који у врањском дијалекту представља исто што и словни знак ***s*** у македонском језику.

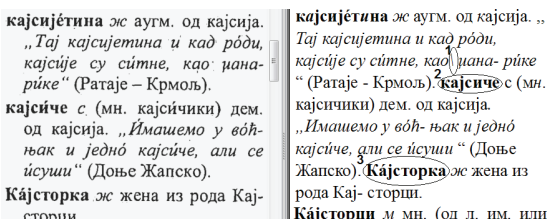
Након дефиниције скупа карактера за учење, могло се приступити самом процесу учења. Ова фаза је неопходна како би обезбедила висок степен тачности препознавања уочавањем специфичности типографије. Познато је да штампарске машине уносе известан степен дисторзије карактера тако да они никада нису истоветни са својим дигиталним матрицама. У поступку учења се препознају и отклањају нетачна препознавања карактера. Процес је итеративан и на самом је кориснику да одлучи о довољном броју итерација, као и скупу над којим ће се вршити учење. Ми смо у овом пројекту на случајан начин одабрали по једну страну за свако почетно слово и тако формирали скуп за учење који је садржао укупно 32 стране.

У следећој фази, односно у процесу оптичког препознавања текста, добијен је јединствен текстуални документ у *word* формату из скупа од 550 *tiff* докумената, где сваки документ одговара једној страни речника. Тачност препознатог текста знатно се побољшала у односу на резултате које смо добијали пре учења. Ипак, требало је решавати две основне групе проблема произашле из процеса препознавања. Прву групу проблема није било могуће отклонити аутоматским поступцима препознавања грешака, а другу је било могуће препознати, па самим тим и уклонити. У прву групу проблема спадају грешке настале услед:

1. појаве сувишних (непостојећих) карактера чији су узрок нечистоће на скенираним документима. На слици 1, дати пример означен бројем 1 односи се на ову врсту проблема. Знак „обрнута коса црта" \ појавио се услед постојања мрља у оригиналном документу. (Треба напоменути да се овај извор грешака ипак може смањити препроцесирањем *tiff* докумената употребом неког од филтера за уклањање

шумова: *Reduce noise*, *Median*, *Gaussian Blur* у *Adobe Photoshop*-у.)

2. погрешно препознатих слова, знакова интерпункције и погрешне или непрепознате акцентуације. На слици 1, дати пример означен бројем 2 односи се на ову врсту проблема. Појам **кајсиче** није препознат као акцентован. У оригиналном документу он гласи **кајси́че**.
3. неуспешне детекције краја пасуса (сваки појам мора бити у новом пасусу). На слици 1, дати пример означен бројем 3 односи се на ову врсту проблема, али и пример означен бројем 2, такође садржи ову грешку.



кајсијéтина *ж* аугм. од кајсија. *„Тај кајсијетина и кад ро́ди, кајси́је су си́тне, као џанари́ке"* (Рataje – Крмољ).
кајси́че *с* (мн. кајси́чики) дем. од кајсија. *„И́машемо у во́ћњак и једно́ кајси́че, али се и́суши"* (Доње Жапско).
Ка́јсторка *ж* жена из рода Кај-сторци.

кајсијéтина *ж* аугм. од кајсија. „ *Тај кајсијетина и кад ро́ди, кајси́је су си́тне, као џанари́ке* " (Рataje - Крмољ). **кајсиче** с (мн. кајсичики) дем. од кајсија. *„Имашемо у во́ћ- њак и једно́ кајси́че, али се и́суши* " (Доње Жапско). **Ка́јсторка** *ж* жена из рода Кај- сторци.
Ка́јстории *м* мн. (од л. им. или

СЛИКА 1. Непредвидиви извори грешака у процесу оптичког препознавања текста; лево је приказ дела оригиналног *tiff* документа који се препознаје, а десно део текста који је препознат.

Извори грешака прве групе проблема нису униформни, нису предвидиви и немају доследну логику појављивања. Из тог разлога су отклањани ручно.

У другу групу проблема спадају:

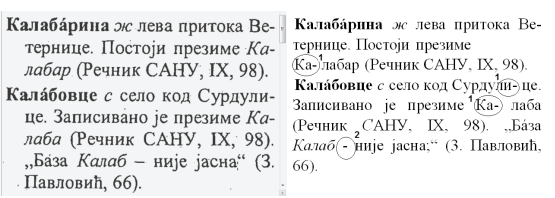
1. проблем постојања поделе речи на крају реда коју треба уклонити (на слици 2, дати примери означени бројем 1 односе се на ову врсту проблема, а пример означен бројем 2 није у тој класи)
2. препознавање цифара уместо слова и обрнуто (нпр. цифра 3 препозната је уместо слова З или цифра 0 уместо слова О)
3. неадекватно препознавање слова која се исто или слично пишу на ћирилици и латиници. С обзиром да се појмови исписују ћирилицом, неопходно је тако их и препознати. У примерима објашњења етимолошког порекла написаних изворним писмом као и у случајевима латинских назива, назива на страним језицима (француски, турски и сл.) потребно је латинично препознавање.

Погрешно могу бити препознати:

- латинично **k** и ћирилично **к**,
- латинично **u** и ћирилично искошено ***и*** (и),
- латинично **y** и ћирилично **у**,
- латинично искошено ***m*** (m) и ћирилично искошено ***т*** (т) итд.

На пример, исправно препознавање мора дати:

- реч **доли́на**, а не реч **доли́на**,
- реч ***Отиде*** (Отиде), а не реч **Отиде** (Отиде),
- реч **дува́н**, а не реч **дува́н** (где је у латинично ипсилон),
- лековита биљка „мајчина душица **Thymus serpyllum L"** – латински назив мора бити исписан латиницом и не сме се десити да слово у буде препознато као ћирилични карактер.



Калаба́рина *ж* лева притока Ветернице. Постоји презиме *Калабар* (Речник САНУ, IX, 98).
Кала́бовце *с* село код Сурдулице. Записивано је презиме *Калаба* (Речник САНУ, IX, 98). „Ба́за *Калаб* – није јасна;" (З. Павловић, 66).

Калаба́рина *ж* лева притока Ветернице. Постоји презиме (Ка-)лабар (Речник САНУ, IX, 98).
Кала́бовце *с* село код Сурдулице. Записивано је презиме (Ка-)лаба (Речник САНУ, IX, 98). „Ба́за *Калаб(-)* није јасна;" (З. Павловић, 66).

СЛИКА 2. Проблем постојања поделе речи на крају реда; лево је приказ дела оригиналног *tiff* документа који се препознаје, а десно део текста који је препознат

Да није постојала потреба истовременог препознавања латиничног и ћириличног писма, као и њихових искошених стилова, овај проблем не би постојао. Међутим, у случају овог речника, висока учесталост употребе и једног и другог писма није дала могућност искључења ма којег од њих. Овај проблем немогућности чак ни визуелног уочавања примене различитог писма унутар само једне речи (пример појма **дува́н** где се не може визуелно утврдити да ли је у питању латинично слово ипсилон) претио је да уруши систем претраге речника. Ако бисмо тражили реч **дуван** и то исписали ћирилично, а он буде у базу уписан са латиничним словом ипсилон,

претрага не би дала резултата. Проблем смо решили методом двосмерне транслитерације.

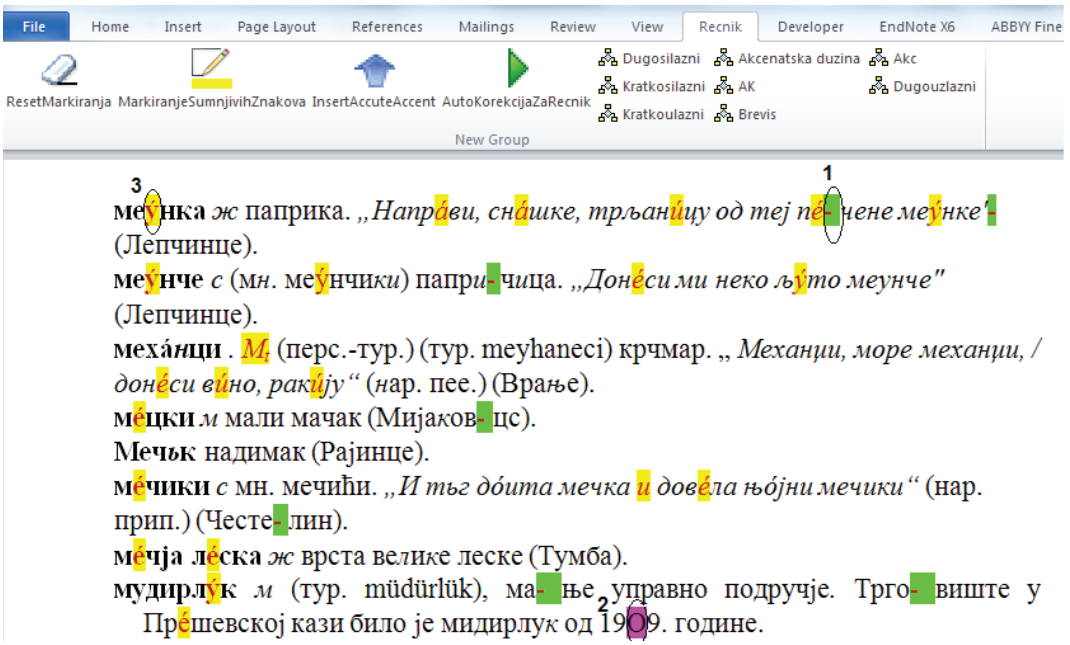
За сваку реч *word* у посматраном тексту *text* претпоставимо да је исписана ћириличним писмом. У првом кораку од дате речи методом *transliterateCtoL(word)* генеришемо исту реч исписану латиницом и именујемо је са *wordTrans*. У другом кораку од те нове речи методом супротне транслитерације *transliterateLtoC(wordTrans)* генеришемо реч *wordTransBack* исписану ћирилицом. У завршном кораку упоредимо речи *word* и *wordTransBack*, па уколико нису једнаке, реч *word* у тексту *text* означимо неисправном, тј. обележимо жутом бојом. Алгоритам двосмерне транслитерације приказан је у псеудокоду, а у табели 1 приказани примери којима се за претпостављену ћириличну реч (прва колона) генерише латинични запис те речи (друга колона), а затим од латиничне речи генерише њена ћирилична репрезентација. У случајевима неисправно написане почетне речи, садржај прве и треће колоне није идентичан.

Алгоритам 1: метод двосмерне транслитерације
Улаз: Текст од **n** необележених речи за који се жели утврдити да ли садржи речи исписане комбинацијом ћириличних и латиничних слова
Излаз: Текст од **n** речи у коме постоји **m**<=**n** обележених речи таквих да су исписане комбинацијом ћириличних и латиничних Слова
1. foreach (string word in text) 2. string wordTrans=Empty;
3. string wordTransBack=Empty; 4.Boolean wordOK=true; 5. // first transliteration step
6. wordTrans= transliterateCtoL(word); 7. // second transliteration step – transliterate back 8. wordTransBack = transliterateLtoC(wordTrans);
9. // if wordTransBack not equal to word, find mismatches 10. if (word != wordTransBack){ 11. wordOK=false; 12. Highlight(word, Color.Yellow);} 13. return text;

word	wordTrans	wordTransBack
долúна	dolúna	долýна
дувáн	dyváн	д?вáн
hymus serpyllum	x?мус серп?ллум	h?mus serp?llum

ТАБЕЛА 1. Примери детекције грешака препознавања ћириличних и латиничних слова методом двосмерне транслитерације

Методом двосмерне транслитерације налази се грешка препознавања ћириличних и латиничних слова. Сличан алгоритам креиран је за налажење грешака препознавања цифара као слова и обрнуто, док је алгоритам препознавања речи које су биле подељене на слогове заснован на примени регуларног израза облика .-\(s)* којим се детектују сва појављивања хоризонталне цртице „-“ након било ког знака. Сва три алгоритма реализована су у облику макроа у *MS Word*-у. На слици 3 дат је изглед рибона (део радне површине *MS Word*-а на коме се налазе иконе најчешће коришћених алата и опција, логички груписаних на основу задатака сличне намене) који је креиран за потребе рада на овом речнику. Функција *МаркирањеСумњивихЗнакова* садржи сва три наведена алгоритма детекције. Резултати примене ове макро-рутине дати су на слици 3 и огледају се у различито обојеним деловима текста. Зеленом бојом (означено бројем 1) су обележене појаве хифенације, ружичастом бојом (означено бројем 2) проблеми препознавања цифара и жутом бојом проблеми препознавања ћирилице и латинице (означено бројем 3). Макроом *Аутокорекција* врши се аутоматска корекција свих карактера претходно обележених макроом *МаркирањеСумњивихЗнакова*, а постоји и макро који уклања сва претходно постављена означавања у тексту (на слици 3 означен са *РесетМаркирања*). Треба напоменути да се појаве нормалног и искошеног стила, уколико је реч о истом писму не сматрају грешком. Примери дати на слици 3 за такве случајеве препознавања су речи: меу́нчики, Врање, велике, мизики итд.



СЛИКА 3. Макро-рутине у *MS Word*-у за детекцију и аутоматску корекцију проблема поделе речи на крају реда и двеју класа грешака у препознавању текста

4. Изградња базе података речника

База података коришћена у реализацији овог речника је Microsoft SQL Server 2005. Припрема пуњења базе изведена је двофазно. У првој фази је текстуални документ, добијен процесом оптичког препознавања карактера, трансформисан у *CSV* (*comma-separated values*) формат и на основу њега је креирана међуфазна база података *Recnik.mdb* у *Microsoft Access 2010*. Разлог фазног пуњења лежи у потреби моделирања и генерисања додатних група података у самом речнику, као и екстракцији одређених врста података како би се генерисале табеле шифарници. За потребе рада апликације потребни су следећи шифарници: *Словарник*, *ЕтимолошкоПорекло*, *ГеографскаЛокација*, *РодРечи*. Улога шифарника у апликацији је у поједностављењу функција претраге (њима се пуне падајуће листе у формама за претрагу) методама филтрирања група података по почетном слову, етимолошком пореклу речи, географској локацији употребе појма и роду речи појма. На пример, ако се жели направити упит који би пронашао све речи у речнику чије је етимолошко порекло француско, потребно је имати могућност селекције филтра из листе свих географских појмова који у речнику садрже дату информацију. Генерисање шифарника се могло извести на два начина. Први начин апликацију чини бржом, али мање флексибилном и реализује се екстракцијом података из речника и пуњењем фиксних, унапред креираних табела. У наведеном примеру, табела *ЕтимолошкоПорекло* садржала би појмове: албански, арапски, грчки, енглески итд. За сваки новоунети појам у речник, за који не би постојао одговарајући податак у неком од шифарника, он би се морао додатном процедуром и унети. Други начин је флексибилнији, али је систем генерисања нових података у табелама спорији. У реализацији овог речника, ми смо се одлучили за другу варијанту. Уместо фиксних табела, формиране су само упитне фразе којима се креирају сви потребни шифарници и смештају у унутрашњу меморију (тзв. „in Memory" облик) приликом активирања апликације. Тако креиране табеле се кеширају и остају непромењене у апликационом кешу све док се не појави потреба да се унесе нови податак (нпр. нови извор етимолошке одреднице)

S...	Pojam	Opis	PojamCir	PojamLat	Audio	Rod	V.	Vidi...	Primer	Poreklo	Lokacija	Status	A..	Datum
A	а́.	за истицање (ве...	а.	а.	а		0		„Гла́ва, а у гла...		Собина	активан	MZ	NULL
A	а-а	узв. за исказива...	а-а	а-а	а-а		0		„Има ли вода́?" ...		Мијовце	активан	MZ	30.06.2013 ...
A	абы́д	коровска биљка...	абад	abad	abd	м	0		„Исе́чи сас ко́су...		Ратаје - Крмољ	активан	MZ	NULL
A	абьдљи́в	-а, -о који има а...	абадљив	abadljiv	abdljiv		0		„Таj лива́да је а...		Црна Река	активан	MZ	NULL
A	Абьдљи́вица	потес	Абадљивица	Abadljivica	Abdljivica	ж	0		„Мло́го ра́сне а...		Црна Река	активан	MZ	NULL
A	а́бьче	дем. од абьд.	абаче	абаче	abce	с	0					активан	MZ	NULL
A	абаџи́ја	занатлија који и...	абаџија	абаџија	abadzija	м	0		„Милан је би́ја ...	(ар.-тур.) (тур....	Преображење	активан	MZ	NULL
A	абаџи́лък	"абаџијски зана...	абаџилак	абаџилак	abadzilak	м	0		„Де́да му се ба...	"(ар.-тур.) (тур...	Врање	активан	MZ	NULL
A	Абди́ница	њиве у Лопарди...	Абдиница	Abdinica	Abdinica	ж	0					активан	MZ	NULL
A	Абе́рка	надимак (Врање).	Аберка	Aberka	Aberka	ж	0			(ар.) (тур. habe...	Врање	активан	MZ	NULL

СЛИКА 4. Табела Речник након пуњења базе података

у неку од табела, када се она брише из кеша и креира нова. У првом случају јавља се проблем редундантности података, а у другом апликација меморијски више оптерећује веб-сервер.

У другој фази пуњења базе, датотека *Recnik.mdb* експортована је у продукциону базу *MS SQL Servera*. Након завршетка друге фазе, иницијално пуњење базе података наставља се временски најзахтевнијим делом овог пројекта: генерисањем линкова за повезивање базе података са спољним датотекама са аудио-садржајем (пододељак 3.2).

4.1 Метаподаци речника

У току прве фазе моделирања базе података речник је био описан следећим метаподацима:

- Одредница – реч или фраза дијалекта, написани са акцентом
- Опис – опис значења појма, врста речи, број, информација о деривацијама појма
- Род – род речи дефинисане појмом
- Пример – примери употребе појма у дијалекатском говору.
- Порекло – етимолошко порекло појма
- Локација – географска одредница места где је аутор забележио пример употребе појма
- Индикација унакрсног референцирања
- Унакрсна референца (над готовим речником формиране су референтне везе међу одредницама истог или сличног значења.)

Након завршетка друге фазе пуњења базе, речник је добио шест додатних врста података:

- *Појам написан ћирилицом без акцената* (за потребе претраге када корисник не користи дијакритике у формирању упита и не користи „меки глас" већ слово *а* као његову супституцију.)
- *Појам написан латиницом без акцената* (иницијално је служио да се генерише назив аудио-датотеке и добијен је аутоматском транслитерацијом појма написаног ћирилицом без акцената у појам написан латиницом, такође без акцената. У продукционој фази, овај податак се користи за претрагу када корисник упит испише латиницом.)
- *Назив аудио-датотеке* ради генерисања спољашњег линка према одговарајућем звучном запису
- *Статус* појма у речнику
- *Корисничко име* аутора записа појма
- *Датум уписа* појма у речник.

Написан је програм за генерисање садржаја додатих врста података: *Појам написан ћирилицом без акцената*, *Појам написан латиницом без акцената* и *Назив аудио-датотеке* генерисани су коришћењем података у колони *Појам*. Преостала три поља иницијално су напуњена одговарајућим константама. На слици 4 дат је приказ изгледа првих 10 записа табеле речника, након завршетка процеса пуњења базе података.

4.2. Генерисање аудио-записа

Мултимедијални запис се најчешће повезује са базом података уносом податка о релативној или апсолутној URI адреси локације на серверу

(или, у општем случају, мрежи) где се такав запис налази. У табели Речник (слика 4) колона Аудио представља податак који се односи на релативни пут, односно име, мултимедијалног записа. С обзиром да софтвер за репродукцију звука, *Adobe Shockwave Player* (Adobe, 2008), у својој верзији додатка веб-прегледачима нема подршку за *Unicode*, садржај колоне Аудио не садржи специфична слова латиничног писма са дијакритичким знаковима. Имена аудио-датотека додељивали смо сагласно садржају колоне Аудио. На пример, аудио-запис *abadzija.mp3* садржи изговор појма и примера речи **абаџи́ја**.

За потребе овог речника снимљено је 30 GB аудио-записа изговора сваког појма са коректним изговором и акцентовањем у примерима. Снимање је поверено глумицама врањског позоришта, уз консултовање проф. Златановића. Коришћен је софтвер *Audacity 2.0.2- A Free Digital Audio Editor*. Рад на обради аудио-записа односио се на рачунарску корекцију (чишћење шума, промену параметара тона) и исецање аудио-записа на појединачне датотеке. Складиштење аудио-записа смо контролисали помоћним MS Excel документом који нам је указивао на грешке у именовању аудио-датотека или на случајеве када је нека аудио-датотека из неког разлога недостајала. На слици 5 приказана је контролна форма за проверу исправности именовања аудио-записа као и података у бази о аудио-записима.

5. Изградња апликације

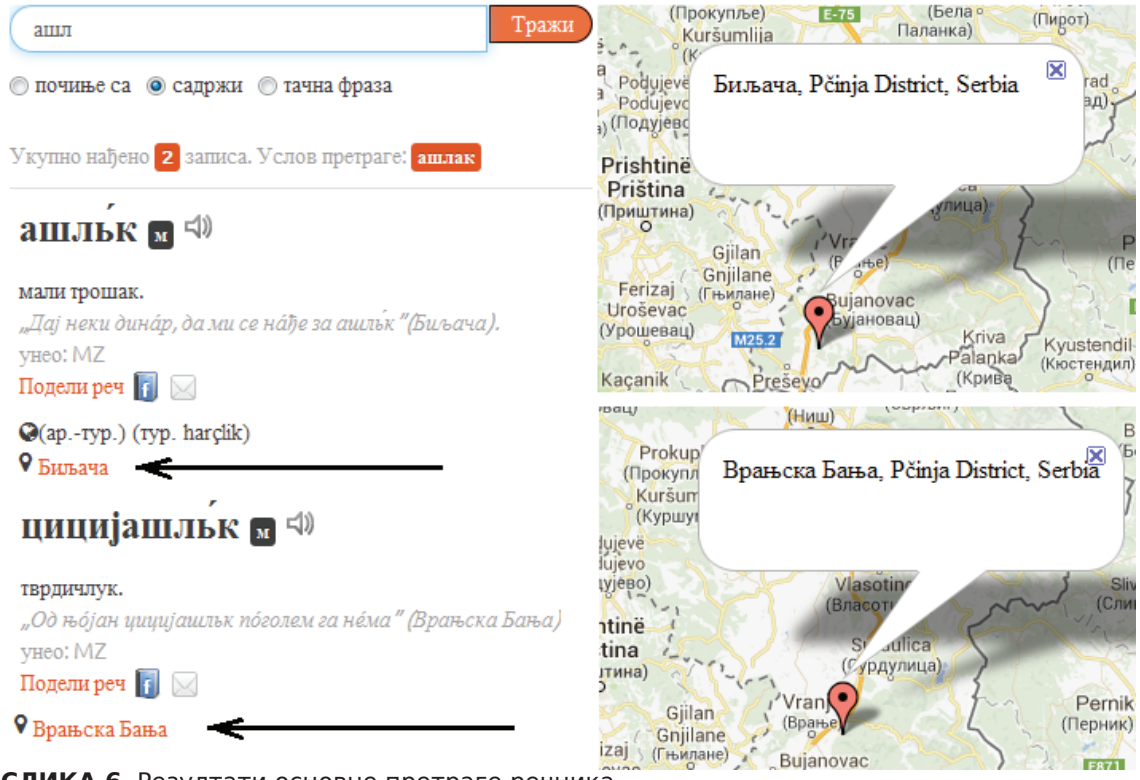
Веб-апликација за управљање речником налази се на адреси www.vranje.co.rs. Основне функционалности апликације су:

- основна претрага (претрага по појму),
- семантичка претрага,
- претрага креирањем сложених логичких упита,
- листање појмова по почетном слову,
- изговор појмова и пратећих примера,
- означавање резултата претраге на географској мапи,
- статистичке информације о речнику.

На слици 6 дат је приказ форме за основну претрагу, резултата претраге у текстуалном и графичком облику. Основна претрага подразумева могућност да корисник зада реч, део речи или фразу коју тражи. Такође, може задати и да ли ће се претрага вршити над појмовима који: почињу, садрже или су једнаки датом упиту. Резултати текстуалне претраге пре свега дају информацију о броју нађених појмова у речнику који су задовољили услов постављеног упита. Испод тога исписују се

A	F	G	H
Audio	**Folder**	**Path** ="E:\eRecnik\govor\" & F5 & "\" & A5 & ".mp3"	**FileExist**
a-a	A	E:\eRecnik\govor\A\a-a.mp3	TRUE
abadzija	A	E:\eRecnik\govor\A\abadzija.mp3	TRUE
abadzilk	A	E:\eRecnik\govor\A\abadzilk.mp3	TRUE
abcee	A	E:\eRecnik\govor\A\abcee.mp3	FALSE
abd	A	E:\eRecnik\govor\A\abd.mp3	TRUE
Abdinica	A	E:\eRecnik\govor\A\Abdinica.mp3	TRUE
abdljiv	A	E:\eRecnik\govor\A\abdljiv.mp3	TRUE
Abdljivica	A	E:\eRecnik\govor\A\Abdljivica.mp3	TRUE
Abdza	A	E:\eRecnik\govor\A\Abdza.mp3	TRUE

СЛИКА 5. Контрола исправности именовања аудио-записа



СЛИКА 6. Резултати основне претраге речника

детаљно информације о сваком од нађених појмова. Уколико неки појам садржи и податак о географској локацији употребе појма (што значи да је аутор уочио употребу тог појма на датој географској локацији), онда се дати податак софтверски трансформише у линк ка географској мапи тако да је могуће кликом добити приказ дела мапе са означеним датим подручјем. Познато је да је *Google Geocodinig web service* алат проналажења географских координата (географске ширине и дужине) из других географских података, као што су називи места, улица, адресе или поштански бројеви. Када се за жељену географску локацију (примери на слици 6. су Биљача и Врањска Бања) софтверски нађу географске координате, онда се оне параметарски проследе другом веб-сервису – Google Maps. Google Maps је алат који користи дигиталне географске карте и на основу задатих координата може означити жељене тачке и приказати их у задатом облику на мапи.

Имплементацијом техника истраживања и екстракције текста које се ослањају на регуларне изразе, генерисали смо упитне фразе које могу вршити претрагу речника на основу неког семантичког упита. Нпр. могу се тражити сви примери употребе фигуративног говора, хомоними, топоними итд. Слика 7 приказује различите могућности семантичких претрага речника. Актуелна верзија апликације нуди 11 група семантичких претрага, а на слици је приказан и први из скупа резултата на основу задатог упита: Глаголи=аорист.

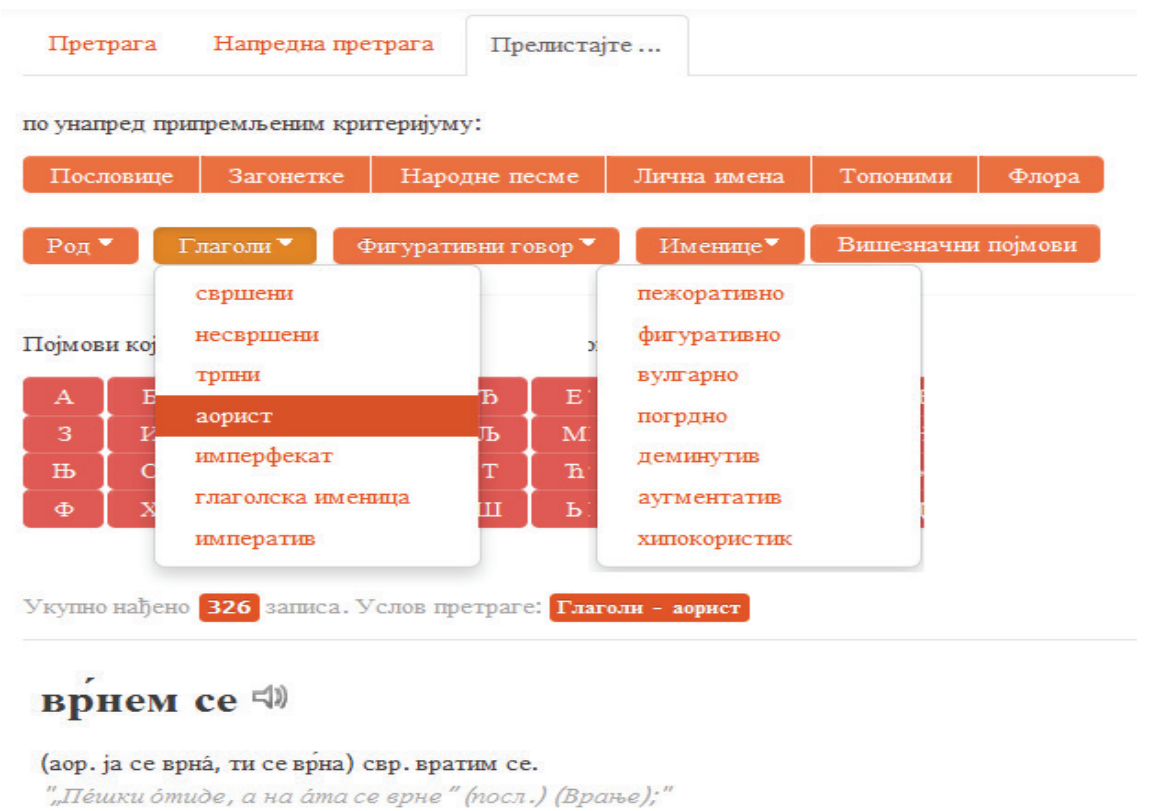
Веб-корисници унапређују речник на 3 начина: додавањем нових појмова у речник, проширењем речника увођењем примера примене и употребе постојећих појмова, коментарима (указивањем на нова значења, нове етимолошке предлошке и нове топонимске карактеристике везане за порекло и подручје употребе датих појмова). Контрола приступа садржајима обезбеђује се помоћу корисничких група и ауторства (на основу корисничког имена).

Корисничке групе дефинисане овом апликацијом су:

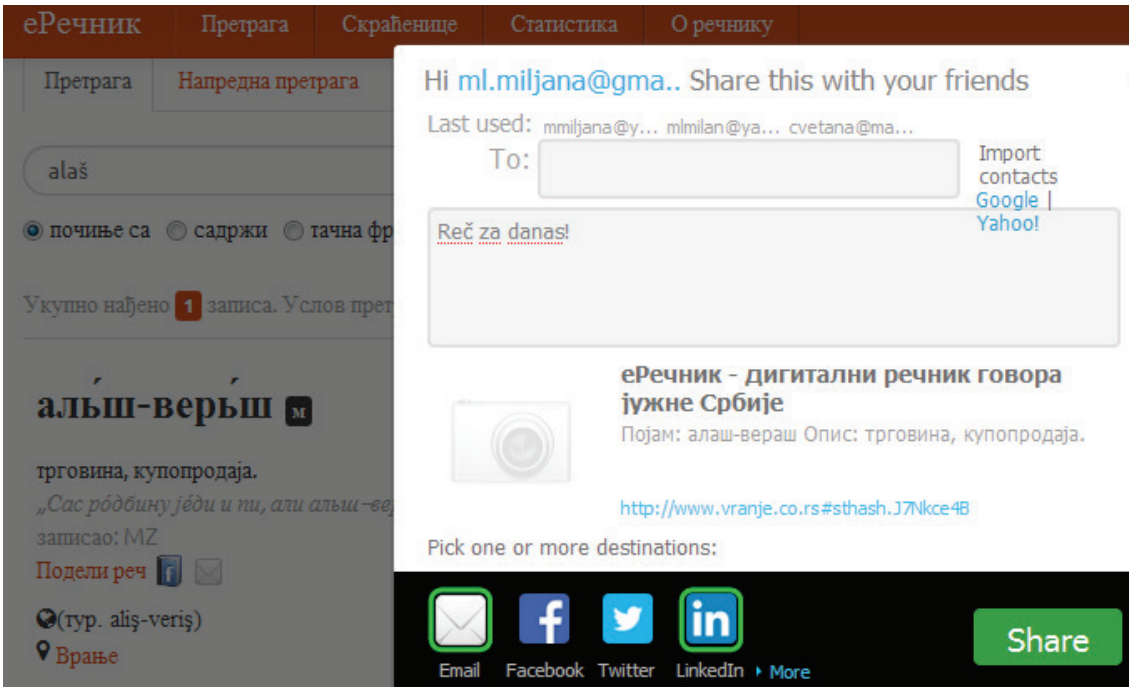
- анонимни веб-корисници – имају могућност претраге
- ауторизовани веб-корисници – имају и могућност уноса нових појмова, измене појмова које су сами унели. Појмови које су они унели имају статус „предложен".
- лексикографи – могућност уноса нових појмова, измене и брисања свих појмова. Могућност измене статуса појма са „предложен" на „активан", после чега остали корисници не могу брисати такав појам.
- администратори – управљају параметрима система и корисничким групама.

У развоју веб-апликације коришћени су следећи софтверски ресурси: релациона база *MS SQL Server 2005*, *ASP.NET Framework 4.0*, *C#*, *jQuery 1.8.9*, *AJAX*, *Twitter Bootstrap Framework 2.0.2.*, *The Google Geocoding API*, *Google Maps API v.2*, *Social Plugins za FaceBook*, *G+*, *Twitter*, *Digg*, *ShareThis Plugins*, *CSS3*.

Веб-апликација садржи имплементације веб-сервиса за повезивање са тренутно најзначајнијим друштвеним мрежама као што су: *Facebook*, *Twitter*, *Google+*, *LinkedIn* и *Digg*. Може се рећи да је ово први дигитални речник на српском језику чије речи и изразе можете размењивати и делити са пријатељима на друштвеним мрежама и путем електронске поште. На слици 8 може се видети имплементација једног од сервиса (*ShareThis*) коме се аутоматски прослеђују појам и његово значење ако се желе поделити на мрежи или електронском поштом (у овом случају је то фраза **аль́ш-верь́ш**).



СЛИКА 7. Семантичке претраге речника



СЛИКА 8. Дељење појмова из речника путем друштвених мрежа и електронске поште

6. Помоћ при изучавању особености призренско-јужноморавског дијалекта

Дигитални речник развијен у овом пројекту представља један од извора за истраживања природе јужносрпског дијалекта. Овде ћемо навести само неке специфичности до којих се може доћи применом разних врста претраге над речником.

а) Грађење времена променом акцента глагола

У описаном дијалекту, за разлику од књижевног језика, могуће је грађење прошлог времена без употребе помоћног глагола, променом акцентовања глагола. Треба напоменути, да без употребе акцената, није могуће утврдити време извршења радње јер су реченице идентичне. На пример:

Зборѝ му, зборѝ, али он ништо не прѝфати. (зборѝ – несврш. говорах, имперфекат; прѝфати – прихвати сврш. аорист)

Говорио сам му, говорио, али он ништа није прихватио.

Збо́ри му, збо́ри, али он ништо не прифа́ти. (садашње време)

Говорим му, говорим, али он ништа не прихвата.

На моти́нке наре́ђамо веша́љке и су́шимо за зиму. (садашње време)

На мотке за сушење меса наређамо каишеве меса и сушимо за зиму.

На моти́нке наређа́мо веша́љке и суши́мо за зиму. (прошло време – несврш. имперфекат)

На мотке за сушење меса наређали смо каишеве меса и сушили за зиму.

б) Глаголска анафора – (**онодӳјем, поонодӳјем, прионодӳјем, заонодӳјем, заоно́дим**) глагол који представља замену за сваки други, само уколико се нађу у истој реченици или ако се нађе у реченици иза оне која садржи глагол на који се односи. Такође, може бити замена за било који глагол за који зна и саговорник.

Пример употребе у реченици иза оне у којој је глагол у односу на који се користи глаголска анафора.

Решава́ја зада́ци цел дан ње́кња. Сви ги сам оноде́ја.

Пре неки дан, решавао је задатке целог дана. Све их је сам решио.

Пример употребе у говору када саговорници знају о којој радњи (глаголу) је реч.

„Поонóди гу још мáлко бáшчу."

Значења ове реченице могу бити различита, зависно од тога шта је саговорник претходно радио у башти. Аутор реченице сугерише да се тај посао настави још мало. На пример, може бити:

„Заливај још мало башту." или

„Окопавај још мало башту." или

„Огради још мало башту." итд.

в) Четири степена поређења придева

Сем основна три степена, постоји и четврти који се семантички налази испод позитива, тј. у себи носи мању изражајност од основне. На пример, марама може бити: **шаре́на** (позитив), **пóшарена** (компаратив), **нáјшарена** (суперлатив), али може бити и **пришаре́на** што се разуме као мање од шарена. Слично је и са већином описних придева:

приýбав (приубавúчав), ýбав, пóубав, нáјубав.

прилъýт, примáл, прилáдан, прималéчак, присúтан, прискр̀жав, прирáнке, прикиселúчав (накисео, недовољно кисео)**.**

г) Градација несвршених глагола

У књижевном српском језику имамо свршене глаголе промене стања: прогледао, пропевао, проговорио... У описаном дијалекту постоји могућност употребе истог префикса на други начин:

тéпам несвр. тучем

протепýјем несвр. тучем (помало)

лéти несвр. лети

пролúћа несвр. лети (помало)

тýрам несвр. стављам

притурýјем несвр. стављам, додајем (помало)

д) Већи број сугласника у континуитету

д̂зрд̂зóљак – човек малог раста и мршав

д̂зрд̂зóљче (мн. д̂зрд̂зóљчики) – мало и мршаво дете

Д̂зр́д̂зинци – род у Бујковцу

Д̂зр́д̂зцке – њиве у Дубници

издзьмбати – појести халапљиво

Алати за анализу овог речника који су на располагању корисницима разликују се у зависности од њихових привилегија у апликацији. Уколико су регистровани корисници, могу користити одељак напредне претраге да би добили одређене статистичке податке. На пример, уколико корисник жели да сазна колико појмова у речнику потиче из неког страног језика може креирањем упита над метаподатком *Порекло* и применом оператора „садржи" задати једну од понуђених вредности добијених екстракцијом из речника, а које се односе на етимолошко порекло појма (алб., ар., грч.,...). У том случају, упит би за одабрану вредност „нем." (немачки) изгледао:

Порекло садржи: нем.

Упит може бити и сложенији формирањем логичких AND израза. Нпр. уколико бисмо желели да сазнамо колико појмова у речнику је женског рода и немачког порекла, упит би гласио:

Порекло садржи: нем. Род једнако: ж

Добијени резултати су дати нумерички и показују број пронађених појмова као и у виду листе тих појмова. Упит којим се могу добити сви придеви који имају префикс при, а које смо описали у ставци в) овог одељка (прилýт, примáл, прилáдан,...) гласи:

Појам садржи: при Опис садржи: -а, -о

Корисници који имају привилегије лексикографа могу претраживати и лог-фајлове из којих сазнају које појмови или групе појмова веб-корисници најчешће траже. Изглед записа лог-датотеке је дат у облику:

datum:19/07/2013 vreme:16:17 pojam: ђукавац
datum:19/07/2013 vreme:16:17 pojam: уjче
datum:19/07/2013 vreme:16:18 izvor: Proverbs
datum:19/07/2013 vreme:16:19 izvor: pаламар

Из дате структуре лог-датотеке могу се вршити анализе учесталости и врста претрага над речником.

У апликацији су свима доступни и статистички подаци о броју личних имена у речнику, броју појмова чији су извори народне песме, броју појмова чији су извори пословице, броју појмова чији су извори загонетке, броју појмова који представљају фигуративну употребу језика итд. Доступна је и *Google* географска мапа на којој је означено подручје на коме се користи дијалекат.

7. Закључак

Примарни циљеви веб-апликације описане у одељку 5 су:

- изградња значајног дигиталног ресурса једног дијалекта, доступног свим интернет корисницима на нов, инвентиван начин
- подстицај интернет заједници у процесу допуне, доградње и чувања дијалекта јужне Србије
- израда дигиталног записа говора на дијалекту
- интеграција речника са геолокацијским ресурсима (*Geocoding API*, *Google Maps API v.2*)
- интеграција са друштвеним мрежама
- дељење садржаја речника електронским путем
- инспирација дигитализацији речника: Нишлијски говор и речник, Црнотравски говор, Речник пиротског говора, Народни говор и речи из власотиначког краја итд.

Апликација је осмишљена и у потпуности реализована тако да је могуће онлајн проширење постојећег речника и додавање нових речника чиме се шири дигитална база дијалеката српског језика. Даљи рад на овом речнику кретаће се у два правца: ка повећању броја појмова, примера и звучних записа, и у правцу унапређења функционалности постојеће апликације и њеног развоја за паметне телефоне и таблете.

6. Захвалност

Реализацију еРечника помогао је Регистар националног интернет домена Србије (РНИДС) у оквиру пројекта (4ПИ) у 2013. години. Велику захвалност аутори овог пројекта дугују и глумицама врањског Позоришта „Бора Станковић“, Радмили Ђорђевић и Милени Стошић.

8. Литература

Abby Fine Reader: Version 11 Users's Guide. CA: ABBYY Software Ltd, 2011.

Adobe: Adobe Director 11 User Guide. CA: Adobe Systems Incorporated, 2008.

Benacchio, Rosanna, Han Steenwijk, Željko Jozić i Nada Vajs Vinja. "Digitalna obradba rukopisnoga rječnika Vocabolario di tre nobilissimi linguaggi, italiano, illirico, e latino Ivana Tanzlinghera Zanottija (1651.—1732.)". *Filologija* No. 58 (2012): 19–38.

Златановић, Момчило. *Речник говора југа Србије*. Врање: Аурора, 2011.

Ивић, Pavle. *Dijalektologija srpsko-hrvatskog jezika*. Novi Sad: Matica srpska, 1985.

Karanikolas, Nikitas N., Eleni Galiotou, George J. Xydopoulos, Angela Ralli, Konstantinos Athanasakos i George Koronakis. "Structuring a Multimedia Tri-Dialectal Dictionary". U *Lecture Notes in Computer Science, vol. 8082*, 509–518. Berlin: Springer, 2013.

Okuka, Miloš. *Srpski dijalekti*. Zagreb: Prosvjeta, 2008.

O'Sullivan, David i David J. Unwin. *Geographic Information Analysis*. New Jersey: John Wiley & Sons Inc., 2010.

Pereira, Sandra i Raïssa Gillier. "TEDIPOR: Thesaurus of dialectal Portuguese". U *Proceedings of the 15th EURALEX International Congress*, 267-281. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 2012.

Sibler, Pius, Robert Weibel, Elvira Glaser i Gabriela Bart 2012. "Cartographic Visualization in Support of Dialectology". U *The 2012 AutoCarto International Symposium on Automated Cartography, Columbus, Ohio, USA, 16 September 2012 - 18 September 2012*. Ohio: Cartography and Geographic Information Society, 2012.

Spyridon, Petsas. "Visualising Perceptual Linguistic Data". MSc in GIS dissertation, University of Edinburgh, UK, 2009.

Van Keymeulen, Jacques i Veronique De Tier. "The Woordenbank Van De Nederlandse Dialecten." U *3rd eLex Conference. Electronic Lexicography in the 21st Century: Thinking Outside the Paper, Proceedings*, ed. Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langements, and Maria Tuulik, 261–279. Ljubljana, Slovenia: Trojina, Institute for Applied Slovene Studies, 2013.

Ćavar, Damir, Alexander Geyken i Gerald Neumann. "Digital Dictionary of the 20th Century German Language". U *Jezikoslovne Tehnologije za Slovenski Jezik: Proceedings of JS*, eds. T. Erjavec and J. Gros, 110-114. Ljubljana: Institut Jožef Stefan, 2000.

Хаџи-Васиљевић, Јован. *Путопис Хаџи-Анте Калиманца*, Београд, 1910.

Веб-адресе

Audacity. http://audacity.sourceforge.net/onlinehelp-1.2/reference.html

Shockwave. http://get.adobe.com/shockwave/

Geocoding API. https://developers.google.com/maps/documentation/geocoding/

РНИДС. http://www.rnids.rs

4ПИ. http://4pi.rs/