UDK 811.163.41'374
811.163.41'322
Stručni rad
Rukopis primljen 9. X. 2019.
Prihvaćen za tisak 23. I. 2020.
doi.org/10.31724/rihjj.46.2.32

Рада Стијовић[1]
Институт за српски језик САНУ
Кнеза Михаила 36/И, RS-11000 Београд
rada.stijovic@isj.sanu.ac.rs
Ранка Станковић[2]
Универзитет у Београду, Рударско-геолошки факултет
Ђушина 7, RS-11000 Београд
ranka.stankovic@rgf.bg.ac.rs
Михаило Шкорић[2]
Универзитет у Београду, Рударско-геолошки факултет
Ђушина 7, RS-11000 Београд
mihailo.skoric@rgf.bg.ac.rs

ВЕБ-АЛАТ ЗА УПРАВЉАЊЕ ГРАЂОМ РЕЧНИКА САНУ И АНОТАЦИЈА ЛИСТИЋА

Грађа на основу које се израђује *Речник српскохрватског књижевног и народног језика САНУ*, а која садржи материјал из преко 4.500 писаних извора и 300 рукописних збирки речи са подручја народних говора штокавског наречја, забележена је на око 5.000.000 листића. Богат лексички материјал, који обухвата књижевни и народни језик у протекла два века и на основу кога треба да се напише још најмање 15 томова Речника, пружа могућност и за разноврсна лингвистичка и ванлингвистичка истраживања. Из тог разлога се приступило дигитализацији грађе и њеној организацији у дигитални репозиторијум, који ће омогућити да се ови у највећем броју врло трошни листићи сачувају, да се учине подесним за аутоматизацију рада на Речнику, што води знатном убрзању његове израде, као и да се грађа записана на њима учини доступном научној, стручној и широј културној јавности. У раду ће бити представљена веб-апликација која је креирана за управљање електронским верзијама листића и резултати досадашњег рада.

[1] Овај рад је настао у оквиру пројекта *Лингвистичка истраживања савременог српског књижевног језика и израда Речника српскохрватског књижевног и народног језика САНУ* (178009), који финансира Министарство просвете, науке и технолошког развоја Републике Србије.
[2] Овај рад је настао у оквиру пројекта *Инфраструктура за технолошки потпомогнуто учење у Србији* (ИИИ 47003), који финансира Министарство просвете, науке и технолошког развоја Републике Србије.

1. О грађи за Речник САНУ

1.1. *Речник српскохрватског књижевног и народног језика САНУ* (у даљем тексту Речник САНУ; SANU) јесте једнојезични дескриптивни речник академијског типа. Он треба, када буде завршен, да обухвати целокупну лексику[3] српско(хрватско)г књижевног језика у последња два века (од Доситеја и Вука Караџића до данас) и српских народних говора. Досадашњих 20 томова представља нешто више од половине предвиђеног речничког опуса.[4] Грађа на основу које се израђује Речник САНУ ексцерпирана је из преко 4.500 писаних извора из свих области живота и рада[5] и сакупљана током 19. и 20. столећа у говорима штокавског наречја (најстарија збирка речи је из 1948. год.). Забележена је на око пет милиона листића, уазбучена и смештена у одговарајуће кутије.

1.2. Значај грађе је многострук. На основу ње треба да се напише још најмање 15 томова Речника САНУ. Она је добра основа за разна фонетска, морфолошка и синтаксичка истраживања. На основу ње може се пратити развој језика током протекла два века, могу се сазнати бројне дијалекатске црте из времена када су збирке настале (већина примера једини су подаци о говору неког краја тога времена), могу се правити етимолошке студије[6], могу се на основу ње изводити и разноврсна статистичка истраживања, формирати синонимски низови итд. Грађа пружа податке и за ванјезичка истраживања – о материјалној и духовној култури током протекла два века, о народним обичајима, веровањима, занимањима и др. Она има и културолошку вредност – на њој су радила многа знаменита имена наше културне историје – Јован Јовановић Змај, Јован Скерлић, Радоје Домановић, Исидора Секулић, Милан Решетар и др. (о грађи више у: Ivanović 2013 и Pavković 2014: 55–59).

[3] Исказ „*целокупну лексику*" је услован, јер по концепцији речника страна и ускоспецијализована лексика уноси се у оној мери у којој она треба да буде позната једној образованој особи.

[4] Први том је објављен 1959, двадесети 2017, а двадесет први је у завршној фази израде и, према плану, треба да буде окончан до краја ове године.

[5] Напоменули бисмо да се под једним изворомподразумевају сва годишта неког часописа (нпр., 20 годишта Бранковог кола), целокупно издање периодичних публикација, речника, енциклопедија (нпр., 10 књига Војне енциклопедије или 4 тома Речника наших старих мера) итд.

[6] Она је важан извор за *Етимолошки речник српског језика*, који се израђује у Институту за српски језик САНУ (објављена и необјављена грађа) и *Етимолошки речник словенских језика* Руске академије наука (грађа објављена у Речнику).

2. Дигитализација грађе

2.1. Године 2016. започета је у Институту за српски језик дигитализација речничке грађе. Основни циљ је био да се ови врло стари и махом трошни листићи сачувају, као и да се учине подесним за аутоматизацију рада на Речнику САНУ, што би, уз дигитализацију и осталих речничких ресурса (досадашњих томова Речника и библиотеке из које је ексцерпирана грађа), убрзало рад на изради даљих томова Речника (в. Stijović i Stanković 2018: 427–440). Циљ је био и створити могућност да се грађа записана на листићима, а која се само у мањој мери доноси у Речнику, учини доступном научној, стручној и широј културној јавности.

2.2. У периоду 2016–2018. целокупна грађа је скенирана. За управљање електронским верзијама листића осмишљено је 2017. године, заједничким радом информатичког тима, корпусних лингвиста и лексикографа, софтверско решење у виду веб-апликације, а према спецификацији и потребама запослених у Институту за српски језик САНУ који у свакодневном раду користе листиће.[7]

2.3. Апликацијом је предвиђено анотирање листића основним метаподацима и праћење напредовања анотација путем коришћења различитих статистичких извештаја. Апликација је заснована на комбинацији PHP и MySQL технологија и омогућује везу између корисника различитих профила, апликативног сервера, репозиторијума скенираних листића[8] и базе података у коју се анотације похрањују. Овлашћени корисници могу јој приступити путем веб-читача, анотатирати листиће коришћењем форми апликације, чиме се допуњује база података.

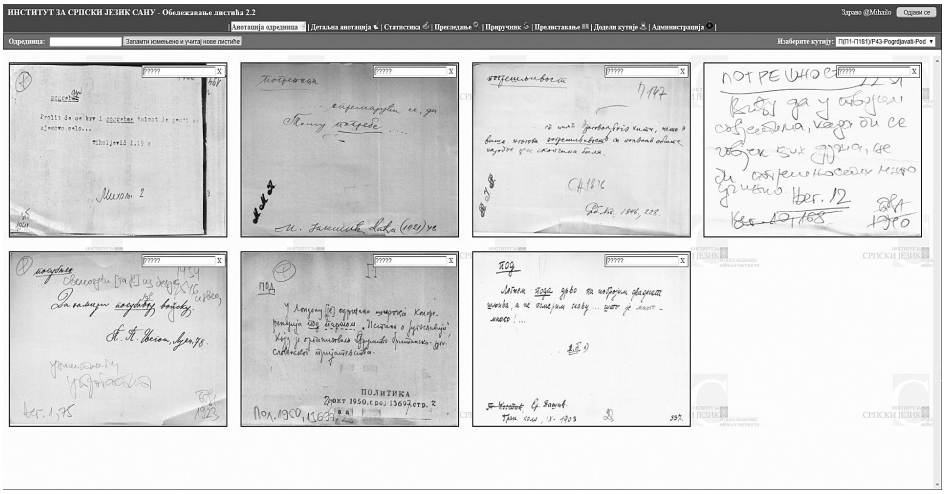
3. Анотација

3.1. Предвиђено је да се у првој фази рада листићи анотирају само одредницом, и то у оном облику у коме је одредница записана у заглављу листића,

[7] Концепт решења дали су Рада Стијовић и Душко Витас, а на развоју и одржавању софтвера раде Ранка Станковић и Михаило Шкорић.

[8] Сви скенирани листићи налазе се у облику jpg датотека организованих у каталоге који одговарају кутијама, односно секцијама листића. Метаподаци о сликама листића, подељене по каталозима према кутији у којој се налазе њихове физичке копије аутоматски су пресликане у базу података, чиме је значајно убрзана анотација листића..

како би се сачувала аутентичност грађе. Почело се са анотирањем речи на слово П (које се обрађују од 18. тома и биће обухваћене са још најмање пет томова) па се завршило са речима на слово Ш.

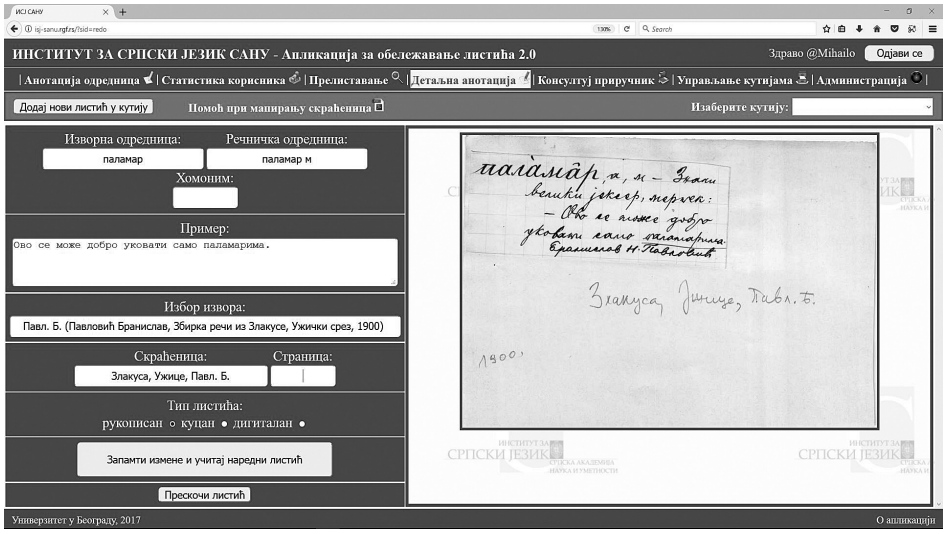


Илустрација 1: Фаза прве анотације

Анотација листића обавља се у три корака: (1) одредница исписана у заглављу листића укуца се у обележено текстуално поље; (2) кликом миша на сличице листића, обележе се сви листићи (један по један), анотацију листића могуће је поништити кликом на X и на листић унети исправну одредницу; (3) након што су сви жељени листићи обележени, промене се чувају у бази кликом на дугме „Запамти и учитај нове листиће". Уколико анотатори нису у могућности да унесу одредницу, уместо ње се у предвиђено поље уписују цифре од 1 до 6, у зависности од проблема на који су наишли: цифром „1". уколико не могу да прочитају одредницу[9], цифром „2" уколико на листићу постоји више једнаких одредница или је нема у заглављу, цифром „4" уколико је листић празан, цифром „5" уколико се одредница не види због грешке при скенирању и цифром „6" уколико је листић окренут наопако (цифра „3" се не користи због сличности са словом З). По завршетку анотације успешно обележени листићи ће притиском на дугме нестати из тела апликације (и исправка њихове анотације биће онемогућена у тој фази рада), а уместо њих ће се учитати нови листићи, необележени и спремни за даље анотирање.

[9] Анотатори су углавном студенти Филолошког факултета из Београда, ненавикли на различите рукописе на листићима.

3.2. За другу фазу рада предвиђена је пуна анотација и претраживост листића по различитим основама. Пуна анотација подразумева унос речничке одреднице тј. одреднице у облику који ће имати у Речнику (што значи њен књижевни лик, без обзира на то у ком је лику забележена у извору – дијалекатском, некњижевном и сл.), затим ознаке за хомоним, пуног текста примера, извора из кога је пример узет са скраћеницом под којом се доноси у Речнику, као и типа листића (рукопис, куцан текст) и додатних напомена. Софтверско решење подразумева и супервизора корисника, затим, прелиставање, што значи и корекције претходно анотираних листића (од стране корисника који их је анотирао и његових супервизора), креирање поља за уношење одреднице у облику какав ће бити у финалној верзији речника САНУ, преглед статистике анотације одредница, као и унетих одредница (од стране корисника који их је анотирао и његових супервизора), могућност додавање нових листића из нове грађе; импортовање у базу дигитално ексцерпираних листића (из ексела).

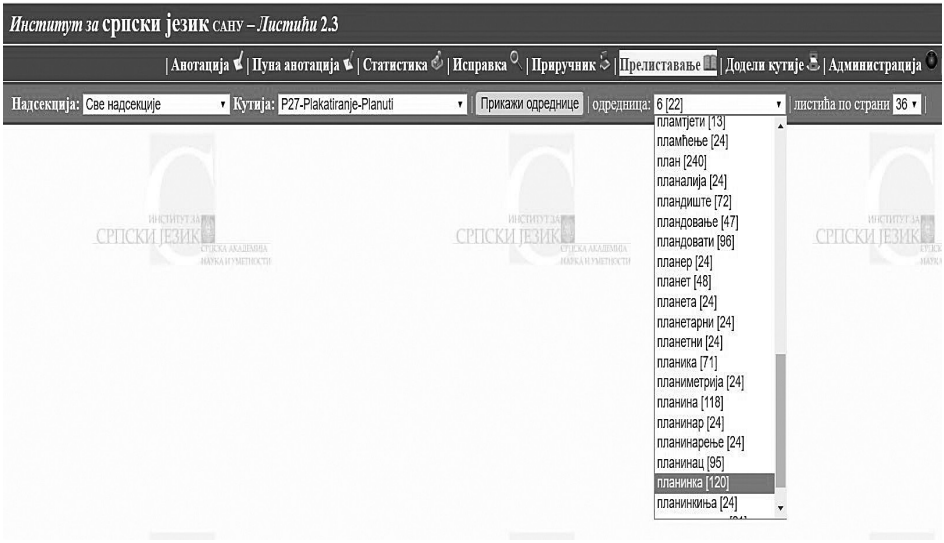


Илустрација 2: Фаза друге анотације

3.3. Припремљена су и упутства за рад (којима се може приступити кроз апликацију) и видео-материјал са примерима употребе софтверског решења, а у више наврата су организоване образовне радионице за кориснике апликације.

Различити статистички прикази помажу у праћењу процеса анотације листића: број ангажованих анотатора у одређеним периодима, учинак њиховог рада по данима, сатима, броју анотираних листића, броју различитих одредница итд. Тако су, нпр., добијени подаци да је на анотацији до сада повремено радило 12 различитих анотатора, али у сваком тренутку њих 3–5 анотира истовремено са различитим интензитетом рада. Укупно је анотирано 2.035.265 листића, смештених у 813 секција, са 215.400 различитих одредница.[10]

3.4. Резултати анотације омогућују оквирни попис одредница до краја Речника[11], као и процену броја потврда за одрђене лексеме (нпр. 9110 потврда за предлог *у* или 7503 за именицу *рука*[12]).



Илустрација 3: Претраживање по одредницама

3.5. У даљем процесу рада предвиђено је унапређење софтверског решења, које подразумева: уазбучавање листића (уношење нових секција са новим разграничењима и прерасподелу анотираних листића по новим секцијама), исправљање неисправно оријентисаних листића, допуну базе листићима

[10] Ово су подаци сачињени за излагање рада у Загребу маја 2019.

[11] Говоримо о оквирном попису одредница, јер у Речник САНУ, поред грађе записане на листићима, улазе и примери из енциклопедија, речника, лексикона и сл. који нису ексцерпирани, али их сарадници на Речнику обавезно консултују.

[12] Ове податке треба схватити условно пошто још није извршено уазбучавање целокупне грађе, још постоје „кутије" са неразврстаном грађом, са новом грађом, са дијалекатским збиркама итд.

од А до О, детаљну анотацију одабраних листића, интеграцију са дигитално ексцерпираним листићима (у екселу), интеграцију ознака на листићима са изворима, модул за управљање одредницама

4. Закључак

С обзиром на значај грађе на основу које се израђује једно од најважнијих пројеката савремене српске науке и културе – Речник САНУ, информатичко-лингвистички тим је приступио дигитализацији листића на којима је грађа записана.

Истовремено се приступило скенирању листића и њиховом превођењу у електронски облик и изради дигиталног каталога листића, којим је обезбеђено организовање ове обимне грађе и њена лакша доступност и прегледност. Израђена је апликација за опремање листића основним метаподацима. За прву фазу рада предвиђена је анотација листића само одредницом, а за другу фазу пуна анотација и претраживост листића по различитим основама – одреднице у облику који ће имати у Речнику, ознаке за хомоним, пуног текста примера, извора из кога је пример узет са скраћеницом под којом се доноси у Речнику, типа листића (рукопис, куцан текст) и додатних напомена.

Софтверско решење подразумева и корисника супервизора, затим, прелиставање, што значи и корекцију претходно анотираних листића, преглед статистике анотације одредница, као и унетих одредница (од стране корисника који их је анотирао и његових супервизора), могућност додавања листића из нове грађе, импортовање у базу дигитално ексцерпираних листића (из ексела) и др.

Резултати анотације омогућили су оквирни попис одредница до краја Речника, као и процену броја потврда за одрђене лексеме.

У даљем процесу рада предвиђено је унапређење софтверског решења, које подразумева: уазбучавање листића (уношење нових секција са новим разграничењима и прерасподелу анотираних листића по новим секцијама), исправљање неисправно оријентисаних листића, интеграцију са дигитално ексцерпираним листићима (у екселу), статистику претраживања

пуних анотација, интеграцију ознака на листићима са изворима и модул за управљање одредницама.

Сви ови поступци омогућили су да се драгоцени листићи сачувају, а уз предстојеће, испланиране послове, омогућиће и да се аутоматизацијом рада на Речнику САНУ убрза израда овог дугорочног пројекта.

Литература

Ivanović 2013. = Ивановић, Ненад. 2013. *Речник САНУ и његова улога у лексичкој стандардизацији српског језика (са историјског и лексикографског аспекта)*. Докторски рад. Филолошки факултет у Београду. Београд. 344 стр. [Ivanović, Nenad. 2013. *Rečnik SANU i njegova uloga u leksičkoj standardizaciji srpskog jezika (sa istorijskog i leksikografskog aspekta)*. Doktorski rad. Filološki fakultet u Beogradu. Beograd. 344 str.]

Ivanović i dr. 2016. = Ивановић, Ненад и др. 2016. Грађа Речника САНУ – потребе и могућности дигитализације у светлу савремених приступа. *Лексикологија и лексикографија у светлу савремених приступа*. Ур. Ристић, Стана. Институт за српски језик САНУ. Београд. 133–154. doi.org/10.5281/zenodo.437537. [Ivanović, Nenad i dr. 2016. Građa Rečnika SANU – potrebe i mogućnosti digitalizacije u svetlu savremenih pristupa. *Leksikologija i leksikografija u svetlu savremenih pristupa*. Ur. Ristić, Stana. Institut za srpski jezik SANU. Beograd. 133–154. doi.org/10.5281/zenodo.437537.]

Pavković 2014. = Павковић, Васа. 2014. Грађа за Речник САНУ. *Савремена српска лексикографија у теорији и пракси*. Ур. Драгићевић, Рајна. Филолошки факултет. Београд. 55–59. [Pavković, Vasa. 2014. Građa za Rečnik SANU. *Savremena srpska leksikografija u teoriji i praksi*. Ur. Dragićević, Rajna. Filološki fakultet. Beograd. 55–59.]

Sabo, Olga; Vitas, Duško. 1988. Mogućnosti osavremenjivanja izrade rečnika na primeru Rečnika srpskohrvatskog književnog i narodnog jezika. *Zbornik sa 4. naučnog skupa Računalniška obdelava jezikovnih podatkov*. Institut Jožef Stefan. Portorož. 375–384.

SANU = *Речник српскохрватског књижевног и народног језика САНУ, I–XX*. 1959. – 2017. Институт за српски језик САНУ. Београд. [*Rečnik srpskohrvatskog književnog i narodnog jezika SANU*, I–XX. 1959. – 2017. Institut za srpski jezik SANU. Beograd.]

Stijović 2017. = Стијовић, Рада. 2017. Грађа Речника САНУ – благо које треба сачувати. *Наш језик* 48/3–4. 201–207. [Stijović, Rada. 2017. Građa Rečnika SANU – blago koje treba sačuvati. *Naš jezik* 48/3–4. 201–207.]

Stijović i Stanković 2018. = Стијовић, Рада; Станковић, Ранка. 2018. Дигитално издање Речника САНУ: формални опис микроструктуре Речника САНУ. *Научни састанак слависта у Вукове дане* 47/1. 427–440. doi.org/10.18485/msc.2018.47.1.ch40. [Stijović, Rada; Stanković, Ranka. 2018. Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU. *Naučni sastanak slavista u Vukove dane* 47/1. 427–440. doi.org/10.18485/msc.2018.47.1.ch40.]

Vitas i Krstev 2015. = Витас, Душко; Крстев, Цветана. 2015. Нацрт за информатизовани речник српског језика: Међународни научни састанак слависта у Вукове дане. *Српски језик и његови ресурси: теорија, опис и примене* 44/3. 105–116. [Vitas, Duško; Krstev, Cvetana. 2015. Nacrt za informatizovani rečnik srpskog jezika: Međunarodni naučni sastanak slavista u Vukove dane. *Srpski jezik i njegovi resursi: teorija, opis i primen*e 44/3. 105–116.]

Rada Stijović
Ranka Stanković
Mihailo Škorić

A Web Tool for Managing Material for SASA Dictionary and the Annotation of Lexicographic Card Files

Abstract

The material for the development of the Dictionary of the Serbo-Croatian Standard and Vernacular Language was collected across 160 years and is recorded on roughly 5,000,000 lexicographic citation cards. It was manually excerpted from over 4,500 written sources and collected in the field in all pronunciations of the Štokavian dialect. At least 15 new volumes of the dictionary are planned based on these card files. They can also serve as the basis for various phonetic, morphological, and syntactic research, as well as for analysing language development over the past two centuries, dialectal specifics from the time when the collections were created (often the only data on the speech of a region at the time of collection), and etymological studies. Its cultural value is also exceptional, as it includes contributions from many illustrious names in Serbian cultural history – Jovan Jovanović Zmaj, Jovan Skerlić, Radoje Domanović, Isidora Sekulić, Milan Rešetar, etc.

This precious, delicate material was scanned from 2016-2018, and in 2017, a web application was developed to efficiently annotate the electronic cards. It was further enhanced based on user needs, enabling (in addition to constricted annotation, where only card headwords were marked) a more detailed annotation including dictionary

entry form, homonym tag, attestation and bibliographic reference, abbreviation in the dictionary, and card type (handwritten, typed).
This paper will present the web tool and annotation results from the letter P to Š. So far, 12 different annotators have been working on the annotations, 3-5 annotating simultaneously at any moment with varying intensity. Of the 813 sections with 2,010,508 paper slips, 795 sections with 1,934,583 paper slips were processed with constricted annotation, including 201,487 different headwords. Annotation results will offer an estimation of the remaining number of words for the dictionary SASA with headword list.

Mrežni alat za upravljanje građom *Rečnika SANU* i anotacija listića

Sažetak

Građa koja je poslužila za izradu *Rečnika srpskohrvatskoga književnog i narodnog jezika SANU*, a koja sadržava materijal iz više od 4500 pisanih izvora i 300 rukopisnih zbirka riječi s područja narodnih govora štokavskoga narječja, zabilježena je na otprilike 5 000 000 listića. Bogat leksički materijal, na kojemu se nalazi književni i narodni jezik posljednjih dvaju stoljeća i na temelju kojega bi trebalo napisati još najmanje 15 svezaka rječnika, pruža mogućnost i za raznovrsna lingvistička i izvanlingvistička istraživanja. Zbog toga se pristupilo digitalizaciji građe i njezinoj organizaciji u digitalni repozitorij, koji će omogućiti to da se vrlo trošni listići u najvećemu broju sačuvaju i da se učine prikladnim za automatizaciju rada na rječniku. To vodi znatnomu ubrzanju njegove izrade, kao i tomu da se zapisana građa učini dostupnom stručnoj i široj kulturnoj javnosti. U radu je predstavljena mrežna aplikacija koja je izrađena za upravljanje elektroničkim verzijama listića te su prikazani rezultati dosadašnjega rada.

Кључне речи: лексикографска грађа, листићи, лексикографски алат, дигитализација, анотација
Keywords: lexicographic material, lexicographic tool, card files, digitization, annotation
Ključne riječi: leksikografska građa, leksikografski alat, listići, digitalizacija, anotacija