SERBIAN LANGUAGE INSTITUTE OF SASA

LEXICOLOGY AND LEXICOGRAPHY IN THE LIGHT OF CONTEMPORARY APPROACHES

A collection of papers

Edited by:
Stana Ristić, PhD, scientific advisor
Ivana Lazić Konjik, PhD, research associate
Nenad Ivanović, PhD, research associate

Belgrade, 2016

ISBN 978-86-82873-??-?

ИНСТИТУТ ЗА СРПСКИ ЈЕЗИК САНУ

ЛЕКСИКОЛОГИЈА И ЛЕКСИКОГРАФИЈА У СВЕТЛУ САВРЕМЕНИХ ПРИСТУПА

Зборник научних радова

Уредништво:
др Стана Ристић, научни саветник
др Ивана Лазић Коњик, научни сарадник
др Ненад Ивановић, научни сарадник

Београд, 2016

УДК:

Снежана М. ПЕТРОВИЋ[1], Тома П. ТАСОВАЦ[2]
[1]Институт за српски језик САНУ
[2]Етнографски институт САНУ
snezzanaa@gmail.com, ttasovac@humanistika.org

ЛЕКСИКОГРАФСКА АНОТАЦИЈА НЕСТАНДАРДНИХ ОБЛИКА У СЛУЖБИ ПРЕТРАЖИВОСТИ ЕЛЕКТРОНСКИХ РЕЧНИКА*

У раду је представљена дигитализација рукописне збирке речи из Призрена Димитрија Чемерикића. Посебна пажња посвећена је наслављавању одредница и анотацији нестандардних облика овог дијалекатског лексикографског извора. Оне омогућавају вишенаменско претраживање речника и његово повезивање са другим електронским речницима и/ли лексикографским порталима.

Кључне речи: дигитализација, електронска лексикографија, лексикографска анотација, српски призренски говор, рукопис.

1. Увод. Дигитализација рукописне збирка речи из Призрена Димитрија Чемерикића започета је 2012. године, више од шездесет година после њене предаје Институту за српски језик САНУ. Неколико је разлога због којих је овај рукописни речник одабран да буде први лексикографски материјал, дигитализован по савременим стандардима у оквиру Института за српски језик САНУ. Најпре зато што је јединствени оригинал морао бити скениран да би се сачувао од пропадања, а за-

* Овај чланак је резултат рада на пројекту *Интердисциплинарно истраживање културног и језичког наслеђа Србије и израда мултимедијалног интернет портала „Појмовник српске културе“* (бр. 47016), који у целини финансира Министарство просвете, науке и технолошког развоја Републике Србије. Дигитализацију и напредно обележавање Збирке речи из Призрена омогућило је Министарство културе и информисања Републике Србије.

тим, да би се тако скенирани материјал учинио доступним за научни и лексикографски рад, стручној и широј јавности. Трећи, не мање важан разлог, лежао је у изазову да се рад започне управо изворoм чија ће дигитализација омогућити стицање искуства у раду са нестандардном лексичком грађом и која ће моћи да послужи као модел за пребацивање у дигитални медиј и других рукописних лексикографских картотека које се налазе у Институту за српски језик САНУ.

2.1. Збирка речи. Рукописна збирка из Призрена садржи 509 страна прекуцаног текста са речима на слова *а* до *г* и близу 17000 руком писаних листића на слова *д* до *ш*. О њеним особинама и структури, богатству лексичког материјала и културноисторијском значају написано је неколико научних студија и радова (Реметић 1996, Петровић 2010, Петровић 2012, Петровић / Тасовац 2014), а послужила је и као материјал за једну мултимедијалну изложбу (Петровић / Тасовац 2013а).

2.2. Посебну вредност ове збирке представља сам језик који рукопис документује – призренски дијалекат. Она је драгоцено сведочанство о јединственом градском говору који на различитим језичким нивоима осликава мултиетнички и мултиконфесионални карактер Призрена и који као такав данас углавном не постоји. Овом је дијалекту такође посвећено је неколико студија: Живановић 1887, Ивић 1991, Реметић 1996, Петровић 2012.

2.3. Да би се стекао увид у значај овог рукописа као лексикографског извора који је и данас актуелан, упоредили смо податке из збирке са бројем потврда из ње посведочених у РСАНУ, с обзиром на то да је она писана са намером да буде у тај речник унета. Иако се, сходно принципима рада на РСАНУ, није могло очекивати да је сваки појединачни пример у овом речнику – тезаурусу забележен, број потврда из Чемерикићеве грађе изненађујуће је мали. На слова *б*, *в*, *г* и *о* постоји известан број примера, док се у осталих 13 томова (4–16) ова збирка ниједном не помиње.[1] Изостанак овог дијалекатског материјала не значи само одсуство убикације у Призрену, па и шире, на Косову и Метохији, у РСАНУ посведочених лексема, већ неретко уопште нема ни облика ни значења речи која су код Чемерикића посведочена. Као илустрацију, навешћемо податке добијене поређењем материјала на слово *ј* из ова два лексикографска извора. У РСАНУ не постоји ниједна

[1] Слово *а* није узето у обзир, пошто је могуће да је збирка предата, тек пошто је рад на овом делу азбуке окончан.

потврда из Чемерикићеве збирке на слово *ј*.[2] У збирци из Призрена на слово *ј* има 709 одредница. Од тога 156 облика, више од једне петине, није у РСАНУ уопште забележено.

Највећи број недостајућих потврда односи се на ониме – 71 – међу којима је највише антропонима и мањи број микротопонима (називи делова Призрена, као што је *Јемиш-̄пазар* или околних села). Некада недостају основни облици, обично у дијалекатској форми, као што је *Јакоф* (поред забележеног *Јаков*), *Јанићија* (поред *Јанићије*), али и велики број изведеница: *Јажинчанин*, *Јаковче*, *Јанчет̄овица*, *Јорданов*, *Јорданов̄ица*, *Јорданов̄ичин* и сл.

Велики број Чемерикићевих потврда фонетски, односно ортографски, разликују се од оних забележених у РСАНУ[3]: јабанџилък : *јабанџи̏лук*[4], јазбъц : *јазбац*, јазбъцов : *јазбац*, јазлък : *јазлук*, јазък : *јазик* и *јазук*, јалънџија : *јаланџија*, јанглъш : *јанлиш* и *јанлаш*, јаръм-путине : *јарам*, јаслилък : *јаслилук*, јастък : *јаст̄ак* и *јаст̄ук*, јастъче : *јаст̄уче*, јасък : *јасак*, једнък : *једнак*, један : *један*, јемъц : *јемац*, јорганџилък : *јорганџилук*, јузлък : *јузлук*, јутрешњь : *јут̄рошњи*, јутрошњь : *јут̄рошњи*, јучеръшњи : *јучерашњи* и сл.

Известан број речи РСАНУ уопште не бележи, чак ни у стандардизованој варијанти:

јадничење, *јадничит̄и*, *јалънџика*, *јарадисат̄и*, *јарандисат̄и*, *језит̄ка*, *јеменџилък*, *јузбашија*, *јум-димије*, *јунашт̄ина*, *јут̄редън*. Чак и када су у питању хапакси, као што су речи *јадничење* и *јадничит̄и* ‘једноничити’ (http://www.prepis.org/items/show/13926), они су у Чемерикићевој збирци поткрепљени живописним примерима и представљају сведочанство о лексичком богатству српских народних говора, чија ризница, управо РСАНУ треба да буде[5].

---

[2] То је вероватно последица чињенице да рукописна грађа Збирке речи из Призрена никада није била спојена са остатком картотеке иако се све време налаазила у Институту и била доступна за коришћење.

[3] Овакав налаз не би се много разликовао чак и да је на ово слово Чемерикићева збирка узета као грађа, пошто су дијалекатски облици у РСАНУ мање-више доследно стандардизовани. Уп. на пример случај са грађом из *Речника косовско-мет̄охиског дијалект̄а* Г. Елезовића у Петровић 1994.

[4] Потврде из РСАНУ писане су курзивом.

[5] Није мали број одредница у РСАНУ које су илустроване примером из само једног извора – на страни 3 деветог тома има их двадесет. На пример: *јурдун* м. покр. ‘ђаво, злодух, нечастиви. – Како постаје, јурдун (бијес) га знао (БиХ, Зовко, ЗНЖ 6, 141; *јуридисат̄и* (Врање, Влај. 1); *Луреша* презиме (Марч. 1, 333), итд. Стога би методолошки било оправдано да и наизглед хапаксне речи, попут *јадничит̄и* буду уврштене у РСАНУ.

Списак недостајућих речи и варијаната није интегралан, ни исцрпно анализиран, нити је његова функција да полемише са методологијом рада на РСАНУ – он је ту да укаже на значај овог, и не само овог, дијалекатског извора за целовити увид у српско језичко наслеђе.

3.1 Приступ дигитализацији. Иако на међународном нивоу постоје различити пројекти везани за дигитализацију рукописног дијалекатског лексичког материјала, она код нас до сада није била рађена, погото-во не на начин који превазилази пуко скенирање текста.[6] Због тога смо се, радећи на дигитализацији ове збирке, по први пут суочили са одређеним бројем специфичних проблема везаних за претраживост српске дијалекатске лексике.

3.2. Са теоријско-методолошког становишта, може се рећи да постоје четири приступа дигитализацији лексичке грађе: захватање слике (image capture), захватање текста (text capture), лексикографско моделирање података (lexicographic data modeling) и лексикографско обогаћивање података (lexicographic data enrichment) (Tasovac / Petrović 2015: 386–387). Због природе извора, након фазе захватања слике, у нашем случају скенирања рукописа, прешло се на фазу *обогаћивања лексикографских података*, уз делимично *захватање текста* кроз транскрипцију.[7]

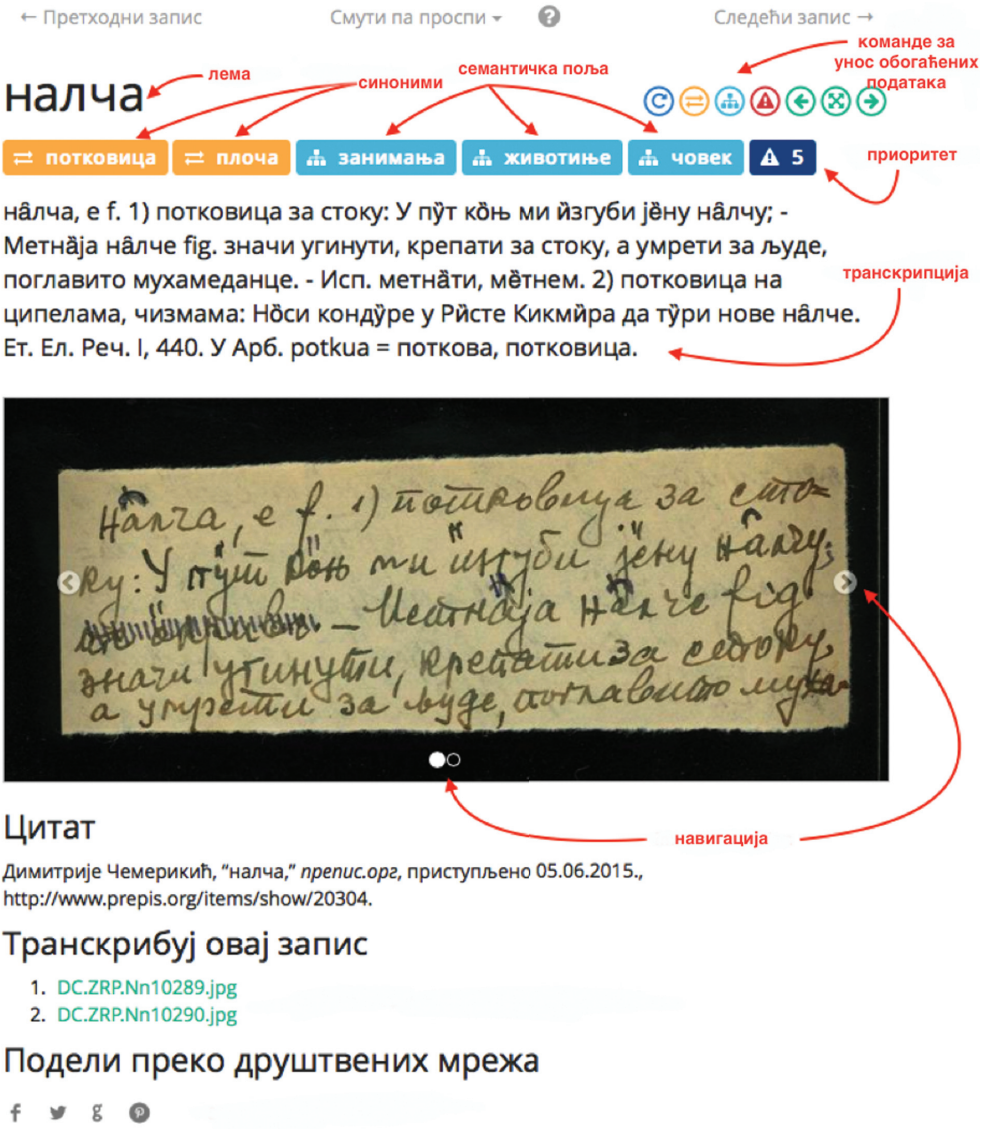
3.3. Важно је нагласити да примењени приступ дигитализацији олакшава коришћење речника на начин који Reichmann 2012: 64 одређује као употребу која превазилази границе самог текста (*texttranszendierend*). То значи да је однос према овој збирци такав да она треба да се користи не само као речник у коме се могу пронаћи информације о значењима речи, фонетским и морфолошким облицима, особинама и граматичком статусу, већ да се она треба третирати и као извор за проучавање културне и социјалне историје (Tasovac / Petrović 2015: 388; Петровић / Тасовац 2014).

4.1. Процес дигитализације. Дигитализација се одвијала у неколико фаза. Први корак је био скенирање комплетног рукописа – свих

[6] Дигитализација рукописног наслеђа текстова из периода између XII и XVIII века рађена је у оквиру пројекта *Корпус српског језика*, касније *Квантитативни опис структуре српског језика* (Kostić 2013; Костић 2012). Међутим, ради се о подухвату из области корпусне лингвистике, а не електронске лексикографије. С друге стране, дигитализација грађе РСАНУ, која је делом рукописна а делом штампана или прекуцана писаћом машином, планирана је као будућа важна активност у изради овог речника, али њена реализација још увек није започета (Ристић / Ивановић 2011).

[7] Детаљније о разлозима за овакав приступ в. Tasovac / Petrović 2015: 388.

руком писаних листића и прекуцаног материјала – чиме је добијена база од око 27000 скенова. Скенирани материјал подигнут је затим на платформу www.prepis.org (Тасовац / Петровић 2013). Даља техничка и лексикографска обрада одвијала се искључиво на самој платформи и то у шест фаза: 1) спајање скенираних листића, 2) наслављавање одредница, 3) одређивање приоритета за транскрипцију, 4) одређивање стандардизованих облика одредница, 5) одређивање синонима и 6) одређивање семантичких поља (Tasovac / Petrović 2015). Обрађена одредница на платформи изгледа овако:



4.2. У овом раду тежиште је стављено на разматрање проблема лексикографске и лингвистичке природе са којима су се аутори сусретали у различитим фазама дигитализације, као и на презентовању начина на које су ти проблеми решавани. Већина њих произилази из чињенице да се ради о дигитализацији **рукописа, а не штампаног речника** и да рукопис садржи **нестандардан језички материјал**, односно дијалекатску грађу. Ови проблеми односе се на фазе 1, 2 и 4 у лексикографској обради на платформи, док ће фазе 3, 5 и 6, због њихове специфичности и ограниченог обима овог излагања, бити обрађене засебно.

5. Рукописни материјал. Рад на фазама 1 (спајање скенираних листића), 2 (насловљавање одредница) и 3 (одређивање приоритета за транскрипцију) био је условљен тиме што је дигитализован рукописни, а не штампани речник. То значи да највећи део рукописа (слова *д* до *ш*) није прошао чак ни минималну уређивачку обраду која би претпостављала коректуру текста и одабир репрезентативних примера, лематизацију, контролу граматичких података и семантичких дефиниција. Такође, ни прекуцани материјал (слова *а* до *г*), није прошао редактуру и није до краја лексикографски обрађен, а уз то је Чемерикић накнадно исписивао руком допуне и исправке на већ прекуцаним листовима. Стога је и овај материјал, прекуцан на папиру А4 формата, морао бити прилагођен моделу лематизације на платформи. Другим речима, скенови су сечени тако да свака одредница буде засебна целина, која се по наслову може пронаћи на платформи.

6.1. Фаза 1 спајање скенираних листића: разлика између физичких и дигиталних објеката. Пошто је велики број рукописних листића исписан обострано, скенирањем једног физичког објекта добијена су два дигитална објекта. Да би електронска верзија рукописног речника била што вернија оригиналу, али и функционално прилагођена корисницима, скенови су спајани по принципу формирања минималних лексикографских гнезда.

6.2.1. Спајање скенова идентично физичком објекту. У случајевима кад је нека одредница забележена на једном листићу са обе стране, два скена спајана су у целину која је доступна у оквиру једне интернет стране, док се на појединачне стране листића прелази уз помоћ *алата за навигацију слике* (Слика 1). Код највећег броја речи овај принцип довео је до стварања одредница у електронском издању идентичних онима у рукопису.

6.2.2. Међутим, код мањег броја одредница, углавном полисемичних глагола, Чемерикић је примењивао два различита метода приликом записивања њихових значења. Први и најчешћи јесте тај да је свако значење писано на посебном листићу, који је засебно и насловљен. Такав је пример глагола *дати* који у рукопису има тридесет четири листића, односно физички независна записа. Различита значења овог глагола, његове рефлексивне облике, употребу у оквиру фразеолошких јединица и устаљених израза, аутор рукописа бележио је на посебним листићима. У лексикографским приручницима, попут РСАНУ, за чије потребе је ова збирка и писана, различита значења глагола *дати*, попут 'удати', 'донети плод', 'дозволити' и сл., налазе се у оквиру једне одреднице. Исто важи и за примере фразеолошке употребе, па су тако изрази попут *дати реч*, *дати главу*, *дати пут* у РСАНУ, за разлику од Чемерикићевог рукописа, сви смештени под одредницу *дати*. Да је рукопис приређиван за штампано издање, ова би се значења свакако нашла у оквиру једне одреднице. Међутим, код електронског издања, спајање близу шездесет скенова (највећи број листића је скениран са обе стране) довело би до стварања тешко прегледне леме, као и до немогућности за фрагментарно смештање у различита семантичка поља и одређивање појединачних синонима. Тиме би навигација и међусобно повезивање кроз речник били значајно мање ефикасни и информативни. На пример, изрази које је Чемерикић записао на посебним листићима могле су да буду напредно обележене својим синонимима и тако повезане са речима одговарајућег значења: синоним *отерати* за *дати пут* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/7957)[8], *жртвовати* за *дати главу* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/7950), *обавезати се* за *дати реч* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/7976) и сл.[9] Стога је сваки од ових листића самостално насловљен и напредно обележен и представља посебан објекат у електронској збирци.

6.3. Спајање скенова у дигиталне објекте различите од физичког објекта. У рукопису постоје случајеви у којима је Чемерикић применио други метод записивања различитих значења, облика и фразеолошке употребе речи – када се примери и значење једне лексикографске

[8] На овај начин значење израза *дати пут* повезано је са једним од значења глагола *дигнати* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/8949).

[9] Треба напоменути и да неке изразе са овим глаголом не бележи ни РСА, попут *дати на благоту* 'поверити другом стоку на исхрану и чување тако да власник добије део белог смока по договору'(Петровић / Тасовац 2013: http://www.prepis.org/items/show/7967).

целине протежу на више листића који нису засебно насловљени. У тим случајевима су различити листићи спајани у једну целину, а пример за то може се видети код глагола *изгубиҭи* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/11658#). Напредно обележавање оваквих целина веома је ограничено и овај принцип примењен је само онда када раздвајање није било могуће.

6.4. Раздвајање скенова једног физичког објеката. У извесном броју случајева Чемерикић је на двема странама једног истог листића записивао различите речи. На пример, на полеђини листића на коме се налази одредница *ћуҭинка* ‘затвор’ (Петровић / Тасовац 2013: http://www.prepis.org/items/show/6885), и који је смештен у целину речи са почетним словом *ћ*, забележио је реч *Белуша* назив цркве св. Николе, данас вероватно рушевина (Петровић / Тасовац 2013: http://www.prepis.org/items/show/6886). Ова два скена су у електронској збирци раздвојена, чиме је омогућено да буду засебно насловљена и да се, сходно свом наслову, могу претраживањем пронаћи. Раздвајање је допринело и формирању нове одреднице *Белуша* назив цркве св. Николе, с обзиром на то да се у Чемерикићевом рукопису под словом *б* реч *Белуша* помиње само као име села у околини Призрена. На тај начин је један изгубљени податак постао видљив и доступан.

7. Нестандардан језички материјал. Потешкоће приликом дигитализације овог рукописа који су последица тога што се обрађује нестандардан језички материјал, односно дијалекат, односе се углавном на фазу 2 (насловљавање листића) и фазу 4 (одређивање стандардизованих облика одредница).[10]

8.1. Фаза 2: насловљавање листића. Са лексикографске тачке гледишта, у Чемерикићевом рукопису постоје три основна типа листића: они на којима је забележен само наслов одреднице, типа *џар* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/10642), они који садрже само наслов одреднице и пример, попут *басма шиљҭе* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/29545) и они који имају, у мањој или већој мери, лексикографски обрађене речи. Листићи последњег типа су далеко најбројнији, али је ниво њихове обрађености веома различит. Многе имају облик већ прилагођен тада важећим стандардима за лексикографску обраду у РСАНУ – посебно речи на *а*, *б*, *в* и *г*, које су прекуцане – што значи да су опскрбљене основним граматичким информацијама, семантичком дефиницијом и примерима, као

[10] В. § 4.1.

што су већ прекуцана реч *бадем* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/47677), или руком писана *демирџија* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/9884).

8.2. Приликом насловљавања листића одлучено је да се за наслов, који је уједно и основ за претраживање у падајућем менију поља за претрагу, узме реч коју је сам Чемерикић на листићу означио као лему, уз минималне интервенције[11]. На пример, на једном од листића за глагол *да̄ти* (http://www.prepis.org/items/show/7976) записано је значење „обавезати се“, али се оно реализује само у синтагми *да̄ти реч*. Одлучено је да се за наслов узме само *да̄ти*, а да се ова специфична реализација нотира обележавањем синонима *обавеза̄ти се*. Аргумент у прилог оваквој одлуци лежи у чињеници да су речи у рукопису методолошки различито обрађене. У случају глагола *да̄ти* и већине речи на слово *д*, Чемерикић је различита значења писао на посебним листићима, док је у каснијим словима тај принцип напуштен. Код листића са полисемичним речима, на којима су сва значења груписана на једном месту, не би било могуће изразе стављати у наслов одреднице, већ је вишезначност решавана низањем већег броја синонима и/или смештањем у различита семантичка поља. Такав је, на пример, случај са полисемичном одредницом *чадър* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/6908), за разлику од *чадр* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/6910).

8.3. Измене у односу на оригинално дефинисане одреднице односе се на изостављање акцента у наслову електронских одредница. То је учињено како би претрага била прилагођена корисницима, с једне стране зато што се на интернету веома ретко користе акцентована слова, а с друге зато што би корисник морао знати тачно место и врсту акцента како би добио жељени податак, што би значајно смањило могућност претраживања и употребљивост електронског издања.[12] Важно је, међутим, напоменути да се приликом транскрипције самих одредница

[11] Такве минималне интервенције односе се најчешће на исправљање словних грешака као у случају речи *Циганштӣја* где је у наслову одреднице у рукопису записано *циганшиш̄тија* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/6330). У минималне интервенције спадају и писање великог уместо малог почетног слова, најчешће код етнонима, као што је *Циганштӣја*. Правописне интервенције које би разликовале фонетски лик електронске и оригиналне одреднице нису рађене.

[12] Посебан је проблем начин бележења акцента у рукопису и у каснијим студијама који исти тај материјал обрађују – у првом случају коришћен је краткосилазни акценат за означавање места акцента, а појављује се и дугосилазни, док је касније углавном краткосилазни акценат замењен ударним, а дужине су изостављене. Детаљније о акценту у овој збирци Реметић 1996: 345–355.

без изузетка транскрибују и акценти (в. нпр. Петровић / Тасовац 2013: http://www.prepis.org/items/show/32668).

8.4. Интервенције су рађене и у уједначавању ортографије, али само у случају када је сам аутор недоследно бележио један исти глас. На пример, лично име *Сака* писао је и као *Дзака*, па је у овом случају оно насловљено као *Сака*, док је варијанта *Дзака* упсана у поље *стандардизовани облик* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/10496#).

9.1. Фаза 4: одређивање стандардизованог облика одредница. Због тога што се српски призренски говор, као део призренско-тимочког дијалекта, одликује бројним архаичним и специфичним фонетским особинама, одлучено је да се дигитализовани материјал обогати и *стандаризованим обликом леме*. Доследно стандардизовање нестандардних облика је од пресудне важности за лакшу навигацију и претраживање дијалекатских речника (Landolt 2007). У нашем случају, стандардизовани облик леме не подразумева бележење само еквивалента који за дијалекатску реч постоји у стандардном српском језику, попут *дибеја* : *дебео* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/8871) – пошто такав еквивалент не би било могуће пронаћи у свим случајевима – већ и анотацију облика речи које нису забележене у стандардном језику, као што је на пример *далък* : *далак* ‘слезина’ и који бисмо могли назвати *псеудостандардним обликом*. Стандардизовање облика и увођење псеудостандардних облика се увелико користе при изради регистара за претрагу и међусобно повезивање немачких дијалекатских речника (Bickel 2013: 129). У случају облика као што су *далак* (од *далък*), *зиндан* (од *зъндан*), *јаланџија* (од *јалънџија*) и сл., ради се о речима које постоје или би по својим фонетским и морфолошким својствима могле постојати на српско-хрватском простору. Стандардни и пеудостандардни облици свима који нису говорници или познаваоци овог дијалекта олакшавају проналажење изворних облика призренске речи, али и омогућавају њихово аутоматско сравњивање са другим речницима у електронском окружењу. Обележавање ових облика речи у речницима српских дијалеката представља пример добре праксе која ће значајно олакшати индексацију и међусобно сравњивање различитих лексичких ресурса, а самим тим и допринети бољем захвату информација (information retrieval).

9.2.1. Да бисмо илустровали типове одступања од стандарда у овој збирци и начине на који су бележени стандардизовани облици, узели смо пример слова *д*. Од 1177 одредница које почињу овим словом, њих

236 има свој стандардизовани облик, што чини нешто више од петине корпуса.

9.2.2. У Чемерикићевој збирци постоје три графеме којих стандардном језику нема: *ъ*, *ӥ* и *s*. Слово *s* углавном је стандардизовано као *з* – *сејсти* : *зејсти* (Петровић / Тасовац 2013: http://www.prepis.org/items/show/10541#), док је код друга два вокала ситуација компликованија и стандардизовани облици морали су бити решавани од случаја до случаја.

9.2.3. Уместо полугласа у Чемерикићевом материјалу, у стандардном језику могу се наћи четири вокала, сви осим *е*: вокал *а* – *дънък* : *дан**а**к*[13], *дал**ъ**к* : *дал**а**к*, *дамнъшъњ* : *давн**а**шњи*, *двокаптън* : *двокапт**а**н*, *домъзлак* : *дом**а**злук*, *домъшњи* : *дом**а**ћи*, *дън* : *д**а**н*, *дънгуба* : *д**а**нгуба*, *дънък* : *д**а**н**а**к*, *дъњом* : *д**а**њу*; вокал *и* – *зъндан* : *з**и**ндан*;[14] вокал *о*: *чадър* : *чад**о**р*, вокал *у*: *домазлък* : *домазл**у**к даилък* : *дахил**у**к*, *дъгъчък*[15] : *дугачак*. Треба нагласити да је полуглас у Чемерикићевом материјалу веома очуван и да је забележен релативно мали број речи код којих је, под утицајем стандарда, полуглас замењен вокалом *а* (Реметић 1996: 363).

9.2.4. Вокал *ӥ* замењен је у стандардизованим облицима вокалима *и*: *д**ӥ**бек* : *д**и**бек* и *у*: *дибид**ӥ**з* : *дибидус*, *дӥзгӥн* : *дузгун*, *дӥзтабанлија* : *дустабанлија*, *дӥшек* : *душек*.[16]

9.2.5. Једна од типских појава у српском призренском говору јесте и губљење финалног *-т* (Реметић 1996: 442): *дебељушас* : *дебељушкаст*, *деветнајес* : *деветнаест*. Ту спадају и „скраћени инфинитивни облици“[17]: *дес : дести*, *денути*, *довес* : *довести*, *дојес* : *дојести*, *донес* : *донети*, *долес* : *долести*.

9.2.6. Испадање *-в-* у спојевима *-ств-*, *-штив-*, још једна од карактеристика призренског говора (Реметић 1996: 398), доследно је стандардизована: *девојашто* : *девојаштво*, *детињсто* : *детињство*, *доброчинсто* : *доброчинство*, *другарсто* : *другарство*, *друсто* : *друштво*.

---

[13] У овом параграфу први наведени облик је из српског призренског говора, а други је стандардизовани.

[14] За овај и следећи пример нису забележене потврде на слово *д*.

[15] У овом облику полуглас се тумачи као континуанта слоговног *л* (Реметић 1996: 374).

[16] Детаљније о овом вокалу у српском пристренском говору Реметић 1996: 366–367, Петровић 2012: 331–335.

[17] Овом приликом нећемо улазити у то да ли је проблем бележења инфинитива и у колико мери последица хиперкорекције аутора за потребе усаглашавања са принципом рада у РСАНУ.

9.2.7. У призренском говору није забележена секвенца *-иjи-* (Реметић 1996: 395–396), што потврђује и Чемерикићева грађа: *димице* : *димијице*.

9.2.8. У Чемерикићевој збирци није забележен ниједан пример за прелаз *л* > *о*, што се сматра утицајем књижевног језика. И ово је показатељ да Чемерикићев рукопис показује једно старије стање српског призренског говора. (Реметић 1996: 421): *дебеја* : *дебео*, *дебел* : *дебео*, *дибеја* : *дебео*, *деја* : *део*, *дел* : *део*, *делба* : *деоба*, *доколан* : *докон*.

9.2.9 Глас *х* не постоји у фонолошком систему овог говора (Реметић 1996: 402–406): *доодак* : *доходак*.

9.2.10. У збирци су забележени бројни нестандардни облици појединих речи, који су последица како дијалекатских особина говора, тако и утицаја старијих функционалних стилова српског језика и суседних говора: *декеври* : *децембар*, *декембар* : *децембар*, *декембер* : *децембар*, *декември* : *децембар*, *декемврије* : *децембар*, *дисембар* : *децембар*.[18]

9.2.11. У појединим случајевима застарели облици, на пример неодређени вид придева, стандардизовани су уобичајеним, одређеним видом: *десан* : *десни*.

9.2.12. Посведочен је и известан број локалних, метатезираних креација: *денеља* : *недеља*.

9.2.13. Појединим турцизмима, који у призренском говору имају облик фонетски ближи етимону, додата је стандардна варијанта: *дирӣија* : *ӣурӣија*, *дӱрӣија* : *ӣурӣија*, *дурӣија* : *ӣурӣија*.[19]

10. Закључак. У овом раду покушали смо да представимо значај и актуелност рукописне збирке речи Димитрија Чемерикића као језичког ресурса и оправданост њеног избора за огледну дигитализацију. Рад на овој збирци омогућио је ауторима бољи увид у широк спектар проблема који се јављају приликом преношења у електронски облик једног нестандардног лексикографског извора. Део тих проблема, као и низ одговарајућих решења, представљени су у овом раду са циљем да допринесу стварању што исцрпнијег списка смерница за дигитализовање и других језичких ресурса сличног типа. Приказане могућности различитог претраживања материјала добијене напредним обележавањем *ӣандардизованог облика леме* значајне су не само због тога што обогаћују материјал и пружају прилику да се овај вредан споменик

[18] Облик *децембар* није забележен у Чемерикићевој збирци.

[19] Облик *ӣурӣија* у овој збирци није посведочен.

језичке и културне историје сагледа кроз више нивоа (преко наслова одредница, стандардизованог облика, синонима и семантичких поља[20]), већ и стога што ће олакшати његово интегрисање са другим дигитализованим изворима који би у најскорије време требало да се нађу на *Речничком порталу* Института за српски језик САНУ чија је израда у току.

ЛИТЕРАТУРА

Живановић, Ј. 1887. Српски језик у околини Призренској, Пећкој, Моравској и Дибарској. У: *Сτражилово* III: 554–556, 573–574, 585–587.

Ивић Р. 1991. *Изабрани огледи III. Из срūскохрваūске дијалекūологије*. Ниш: Просвета.

Петровић, С. 1994. Значај дијалекатског материјала за проучавање турцизама у српском језику. У: *Говори ūризренско-ūимочке облаūи и суседних дијалекаūа*. Зборник радова са научног скупа. Ниш–Београд: Филозофски факултет у Нишу, 427–431.

Костић, Ђ. 2013. *Кванūиūаūивни ойис сūрукūуре срūског језика: срūски језик од XII до XVIII века. Живоūија, канон, хронике.* Београд: Службени гласник.

Петровић, С. 2010. Збирка речи из Призрена Димитрија Чемерикића као извор за проучавање језичке и културне интерференције на Косову и Метохији. У: *Косово и Меūохија у цивилизацијским ūоковима*, књ. 1, *Језик и народна ūрадиција*. Косовска Митровица: Филозофски факултет Универзитета у Приштини, 195–206.

Петровић, С. 2012. *Турцизми у срūском ūризренском говору*. Београд: Институт за српски језик САНУ.

Петровић, С., Т. Тасовац 2013 (електронски извор). *Збирка речи из Призрена Димиūрија Чемерикића.* Препис. орг: платформа за дигитална издања и транскрипцију српског рукописног наслеђа, Београд: Центар за дигиталне хуманистичке науке, Институт за српски језик САНУ, Етнографски институт САНУ. http://www.prepis.org/items/browse?collection=1

Петровић, С., Т. Тасовац 2013а. *Призрен - живоū у речима*. Београд: Институт за српски језик САНУ, Центар за дигиталне хуманистичке науке, Галерија науке и технике САНУ.

Петровић, С., Т. Тасовац 2014. Збирка речи Димитрија Чемерикића као извор за етнолингвистичка и етнолошка истраживања. *Гласник Еūнографског инсūиūуūа САНУ* 62/2, 171–180.

Реметић, С. 1996 Српски призренски говор I (гласови и облици). *Срūски дијалекūолошки зборник* 42, 319–614.

РСАНУ

Ристић, С., Н. Ивановић 2011. Предлог за модернизацију рада на Речнику САНУ. У: *Граматика и лексика у словенским језицима*. Нови Сад – Београд: Матица српска, Институт за српски језик САНУ, 529–553.

[20] Последња два поља биће детаљније обрађена на другом месту, а основне информације могу се наћи у Tasovac / Petrović 2015.

Тасовац, Т., С. Петровић (ур.) 2013 (електронски извор). *Збирка речи из Призрена Димитрија Чемерикића.* У: *Прейис. орг: йлайформа за дигийална издања и йранскрийцију срйског рукойисног наслеђа.* Београд, Центар за дигиталне хуманистичке науке, Институт за српски језик САНУ, Етнографски институт САНУ. http://www.prepis.org/

*

Bickel, H. 2013. Fortschreitende Digitalisierung: Neue Zugriffe auf das Idiotikon. In: *150 Jahre Schweizerisches Idiotikon. Beiträge zum Jubiläumskolloquium in Bern, 15. Juni 2012.* Bern, 121-36.

Kostić, A. 2013. Digitalization of Serbian written Heritage: Electronic Corpus of Serbian Language from 12th to 18th Century. In: *Speech and Language*, 4th International Conference on Fundamental and Applied Aspects of Speech and Language. Belgrade, 10–14.

Landolt, Ch. 2007. Neuere Entwicklungen in der historischen Dialektlexikographie des Deutschen. *Lexicographica* 23: 151–172.

Reichmann, O. 2012. *Historische Lexikographie: Ideen, Verwirklichungen, Reflexionen an Beispielen des Deutschen, Niederländischen und Englischen*. Berlin, Boston: De Gruyter.

Tasovac, T., S. Petrović 2015 (електронски извор). Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. In: *eLex 2015 – Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age*. https://elex.link/elex2015/conference-proceedings/paper-25/

Snežana Petrović and Toma Tasovac

LEXICOGRAPHIC ANNOTATION OF NONSTANDARD FORMS AS A SEARCHABILITY AID IN E-DICTIONARIES

S u m m a r y

The paper discusses the process of digitizing and annotating lexicographic paper slips from the manuscript collection of the Serbian dialect of Prizren, compiled by Dimitrije Čemerikić. It emphasizes two problems of the digitizing process: the determination of headwords and the standardized forms of phonetically / orthographically nonstandard lemmas. Various types of solutions are proposed as a contribution to future list guidelines for the digitization of similar lexical resources. The increased search possibilities, provided by the availability of standardized lemma forms are important not only because they enrich and offer the multi-level insight into this valuable lexicographic material, but also because they will help with the integration of various digitized resources into the Dictionary Portal of the Serbian Language Institute of SASA, which is currently under development.