Рада Р. Стијовић[*]
Институт за српски језик САНУ[**]
Београд, Србија

Ранка М. Станковић
Рударски-геолошки факултет[***]
Београд, Србија

ДИГИТАЛНО ИЗДАЊЕ РЕЧНИКА САНУ: ФОРМАЛНИ ОПИС МИКРОСТРУКТУРЕ РЕЧНИКА САНУ

У Упутству намењеном обрађивачима који раде на изради Речника српскохрватског књижевног и народног језика дате су детаљне и прецизне смернице које описују микроструктуру речничког чланка, укључујући и разрешење могућих недоумица. Упутство је намењено састављању штампане верзије Речника па се поједине смернице односе на графичко обликовање одреднице (употреба одређеног типа слова и сл.). У припреми дигиталног издања Речника, микроструктура речничког чланка мора се описати на начин који ће омогућити различита претраживања текста Речника (изван уобичајеног редоследа одређеног азбучним поретком одредница). У том светлу у раду се описује начин разлагања речничког чланка у дигиталном облику на поља обележена XML-етикетама, а сагласно међународним стандардима (LMF, TEI). Микроструктура чланка описана на овај начин не зависи више од могуће графичке презентације, а омогућава вишеслојну претрагу текста Речника.

Кључне речи: рачунарска лексикографија, лексикографска радна станица, језички ресурси, речнички чланак, српски језик, Речник САНУ.

1. Увод

Речник српскохрватског књижевног и народног језика САНУ (у даљем тексту РСАНУ), као што је познато, јесте једнојезични дескриптивни речник академијског типа. Досадашњих 20 томова[1] представља нешто више од половине предвиђеног речничког опуса. Речник треба да садржи целокупну лексику српског књижевног језика и српских народних говора забележену током протекла два века. У њега, поред књижевних, улазе и покрајинске речи, мање познате или непознате књижевном језику, као и речи које у књижевном језику нису пожељне, које представљају архаизме, варваризме, арго или уобичајене неправилности изражавања. Али се у њега уноси и моменат нормативности тиме што се читалац нарочитим ознакама упозорава на то да је

[*] rada.stijovic@isj.sanu.ac.rs

[**] 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.

[***] 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] Двадесети том је у припреми за штампу и верујемо да ће до изласка овог рада бити одштампан.

реч покрајинска, неправилна и сл. (Упутства). Све одредничке речи дате су у акценатском и морфолошком лику који одговара књижевнојезичком стандарду. Акценатска и обличка одступања од савремене норме наводе се у заградама на одговарајућем месту.[2] Примери којима се илуструју одреднице ексцерпирани су из писаних извора[3] и прикупљени на најширем српском дијалекатском подручју. Исписани су на преко пет милиона картица, а у речнику презентовани тако да показују временску и просторну заступљеност речи. Овако богат лексички материјал омогућује разноврсна лингвистичка и ванлингвистичка истраживања.

Речничка грађа се повремено допуњава новим речима и новим изворима, али потпуну савременост при оваквом темпу израде (један том у три године) није могуће постићи[4]. То се нарочито односи на грађу коју би требало унети у већ објављене томове, што је могуће урадити тек по завршетку израде речника и његовом поновном издању. Треба ипак нагласити да РСАНУ, чак и ако изгуби нешто од своје практичне актуелности, никада неће изгубити од своје научне актуелности. Међутим, мора се рећи и то да би примена савремених технологија омогућила да он не изгуби ни своју практичну актуелност. Наиме, дигитализацијом речника било би омогућено његово константно допуњавање и осавремењавање, као и праћење промена у језику. Осим тога, дигитализовани речник био би изванредан корпус за израду наредних томова, као и за разноврсне претраге у научне и друге сврхе. Дигитализовањем речника, картица са примерима и библиотечких извора, као и осавремењавањем начина израде речника знатно би се убрзала његова израда.

У Институту за српски језик почело је 2016. године скенирање листића са речничком грађом (планирано је да се средином 2018. заврши, чиме би се, поред осталог, ово драгоцено културно благо спасило од пропадања). Ове године је, у сарадњи са информатичким тимом са Рударско-геолошког факултета у Београду, почело и анотирање скенираних листића (у првој фази само одедницом, а у наредној свим подацима који би омогућили потпуну претраживост текста).[5] Исти информатичко-лингвистички тим[6] отпочео је 2016. године припрему за дигитализовану верзију РСАНУ. Засада су креиране процедуре за анализу дигитализованог текста 1. и 19. тома, сегментирани речнички чланци и информационе целине у оквиру речничког чланка према моделу који ће бити описан у одељку 2.

Део резултата ове активности доносимо у овом раду.

1.1 Формализације речничког чланка

Традиционално обележавање елемената речничког чланка постиже се визуелно, углавном стилом фонта, на основу чега читалац може препознати компоненте речничког чланка: одредницу, квалификаторе, дефиницију, примере, изворе. Рачунарски програм из овако обележеног текста не може да екстрахује потребне

[2] Дијалекатска грађа којом се илуструју одреднице доноси се у оригиналном виду, онако како је записана на терену.

[3] Осим грађе на овим картицама, у РСАНУ се уносе примери из бројних речника (једнојезичних, двојезичних, дијалекатских, књижевног језика, терминолошких и др.), енциклопедија, лексикона, именика итд.

[4] Истраживања су показала да је „темпо израде Речника САНУ веома … сличан темпу израде речника овога типа у европским језицима који су израђени без информатичке подршке“ (Ивановић 2016:152).

[5] Материјалну подршку овом послу Институту је пружило Министарство културе и информисања Републике Србије.

[6] На челу информатичког тима стоје Душко Витас и Ранка Станковић (Универзитет у Београду), а на челу лингвистичког Рада Стијовић, а за речник и Олга Сабо (Институт за српаки језик САНУ).

информације, тако да, ако желимо не само машински читљив већ машински разумљив текст чланка, морамо спровести другачије обележавање речничког чланка. Припрема дигиталног издања РСАНУ захтева формално описивање микроструктуре речничког чланка како би рачунар могао да препозна и разуме унутрашње целине речничког чланка.

У (Витас, Сабо 1988) размотрен је оквир за осавремењивање рада на РСАНУ, предложено је конципирање лексикографске базе као средства за експлицирање и формализовање лексикографског знања и похрањивање резултата лексикографског рада. На основу садржаја овакве базе било би могуће да се, поред унапређења и убрзавања рада на РСАНУ, генеришу и други лексикографски ресурси (Витас, Крстев 2014).

Утврђивање структуре речничког чланка, повезивање његових целина са дозвољеним садржајем може у знатној мери олакшати процес уређивања и аутоматизовати провере формалне исправности речничког чланка (употреба скраћеница, структура дефиниција, редослед одредница, сортирање примера и сл.). Припрема текста за публиковање на различитим медијумима, као нпр. за веб портал, мобилни телефон, таблет, могли би се аутоматски произвести, уз, наравно, традиционални папирни формат (Витас, Сабо 1988). Структурирање речничког чланка омогућава приказивање садржаја лексичке базе са различитим нивоима детаљности, за различите намене, нивое привилегија и профиле корисника.

Развоју лексикографске базе претходи развој модела података за разлагање речничког чланка у дигиталном облику, обично на поља обележена XML-етикетама, сагласно неком од модела за обележавање лексикографских ресурса, на пример Оквир за лексичко обележавање LMF (Lexical Markup Framework)[7], Иницијатива за обележавање текста TEI (Text Encoding Initiative)[8] и Лексички модел за онтологије lemon[9] (Lexicon Model for Ontologies).

Н. Иде и колеге (Иде 1993) показују да модели релационих података, укључујући ненормализоване моделе који омогућавају угнежђене релације, не могу у потпуности да ухвате структурне особине лексичких информација и предлажу модел заснован на структурама својстава које се потом мапира у објектно оријентисан модел података и имплементира у објектно-оријентисан систем за управљање базама података. Касније се ови модели преводе у LMF оквир за лексичко обележавање, као апстрактан метамодел који обезбеђује заједнички, стандардизовани оквир за израду лексичких база, које обезбеђује бележење лингвистичких информација и њихово коришћење у различитим апликацијама и за различите задатке (Калцолари 2013). LMF пружа заједничку и дељену репрезентацију лексичких објеката која укључује морфолошке, синтаксичке и семантичке аспекте, коришћењем јединственог модела електронских језичких ресурса од малих до веома великих.

TEI даје препоруке за обележавање и размену електронских текстова, а један од модула препорука намењен је кодирању лексичких ресурса свих врста, па тако и једнојезичних и вишејезичних речника. Све су бројнији штампани речници који се преводе у електронски облик.[10]

[7] https://www.iso.org/standard/37327.html

[8] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DIEN

[9] http://www.w3.org/2016/05/ontolex/

[10] И. Бјелаковић наводи податак, позивајући се на текст С. Гренџера о електронској лексикографији, да је 2010. године најављено објављивање искључиво електронског издања *Oxford English Dictionary* зато што је потражња за овим типом издања далеко већа од потражње за штампаним издањем (Бјелаковић 2016: 172). Ауторка износи и став да се данас у многим земљама назив лексикографија изједначава са називом електронска лексикографија (Исто).

Циљ лексичког модела за онтологије јесте да обезбеди богату лексичку и лингвистичку основу за онтологије, што обухвата представљање морфолошких и синтаксичких особина лексичких уноса, као и синтаксичко-семантичке информације, односно значења ових лексичких уноса (одредница) у односу на онтологију или речник (Мек Кре 2011).

Речнички чланак обележен према неком од стандарда (LMF, TEI, lemon) не зависи од могуће графичке презентације, а омогућава вишеслојну претрагу, анализу и презентацију текста речника.

1.2 Однос микро- и макроструктуре речника

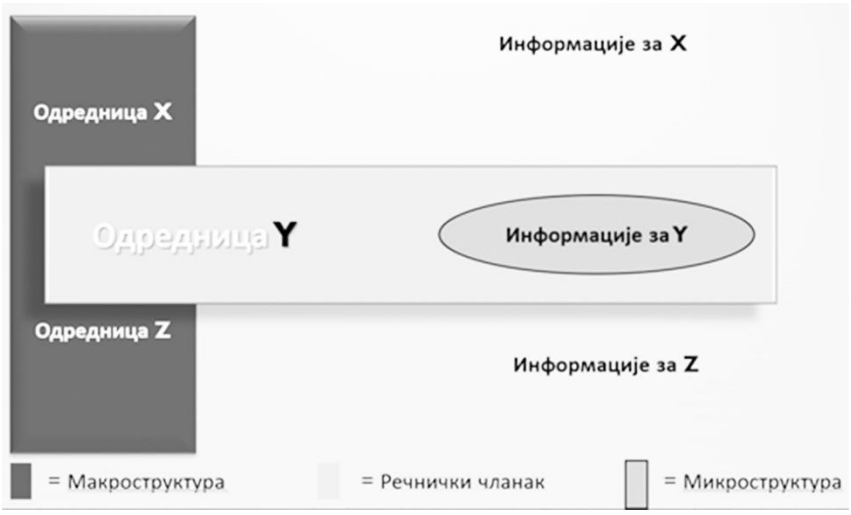
Макроструктура речника односи се на спољни изглед, облик и величину речника, тип речника и број лексичких уноса у речнику. Може се рећи да макроструктура одговара колекцији речничких чланака (РЧ) који се обрађују и може се представити као:

$$М = РЧ_1, РЧ_2, РЧ_3 \ldots РЧ_{n,} \text{ где је } РЧ_1 = (ИК^{i.1}, ИК^{i.2}, ИК^{i.3} \ldots ИК^{i.\,n}),$$

где је i информациона целина, а ИК информациона класа, која може бити: лема (одредница), фонологија, врста речи, флексијска морфема, дериват, синтакса, семантика… Сваки речнички чланак састоји се од n информационих целина. Потребно је раздвојити информационе целине тако да рачунар може да успостави везу између предефинисаних информационих класа, одреднице, флексије, дефиниције, примера, извора.

У заглављу речничког чланка РСАНУ бележе се одредница и показатељи граматичке категорије, за чим следи етимологија па квалификатори којим се обележава употребна и стилска вредност одреднице, потом значења са ознаком, дефиницијом (испред које се у неким случајевима наводе одређени синтаксички подаци), примерима и изворима. На крају су вишечлане лексичке јединице: изрази (терминолошке синтагме и фразеологизми) и пословице.

Торстен (Торстен 2006) помиње приступе који издвајају и мезоструктуру са дефиницијама међусобне повезаности речничких одредница и других информација које нису директно кодиране у самим речничким чланцима, на пример граматичка правила.



Слика 1. Микро- и макроструктура речника

1.3 Систематизација скраћеница

Да бисмо препознали информационе целине описане моделом микроструктуре, било је неопходно, поред осталог, систематизовати скраћенице, која су у папирној верзије РСАНУ дата као јединствена целина азбучним редом. Доносимо овде део те систематизације:

временски маркирана лексика: арх., заст., ист., рсл., ссл., стсл., сткњ., цсл.
стилски маркирана лексика: вулг., деч., ђач., експр., еуф., ир., индив., књ.(ишки), ков., необ., неодом., нераспр., неуоб., неол., песн., пеј./погрд., подр., презр., разг., фам., фиг., хип., шаљ., шатр.
нестандардна лексика: варв., дијал., нар. некњ., непр., погр.
парадигматски односи: вар., син., супр.
изговор: ек., ијек./ј., ик.
наречја: кајк., чак., шток.
територијално маркирана лексика: зап. кр., ист. кр., југозап. кр., локал., покр.
књижевни жанрови: НЗаг, НП, НПосл, НПр
етимологија: алб., грч., лат., мађ./мађ., нем., рум., рус., тал., тур., фр.
граматичке скраћенице: за врсту речи: везн., зам., им., предл., прид., прил., узв.; за грам. род: м, ж, с; за гл. вид: несвр., свр., трен., уч.; гл. род: непрел., повр., уз. повр., прел.; гл. времена: аор., импф., през., перф.; прид. вид: одр., неодр.; бројивост (граматичка): јд., мн., дв.; (лексичка): зб., зб. им, супл. мн.
терминолошке скраћенице: агр., адм., анат., биол., бот., вој., геогр., грађ., грам., екон., зоол., информ., ист., кув./кул., мат., мед., мит., муз., поз., правн., рлг., руд., спорт., трг., фарм., филм., шах., шум.
географске скраћенице 1: азбук., бањал., бач., ваљ., власот., врањ., дубр., дурм., зај., knин., књаж., колуб., кос., краг., круш., леск., морав., нишав., новопаз., пирот., подрин., подун., призр., ресав., рудн., сирин., смед., срем., таков., црмн.
географске скраћенице 2: БиХ, Војв., Далм., КМ, Скоп. ЦГ, Слав., Срб., Херц., Хрв., ЦГ, Шум.
део назива: Б., В., Г., Д., Ист., Ј., З., М., Н., С., Срп.
техничке скраћенице 1: в., исп., И., изр./Изр., Ред., ф.(амилија), пор.(одица).
техничке скраћенице 2: г., др., изд., књ., сл., тзв., тј., ур.

2. Микроструктура чланка

Како је већ поменуто, потребно је формално описати структуру речничког чланка тако да се омогући препознавање речничких чланака различитог обима и структуре, и оне сасвим кратке, са једним значењем и без примера, као и оне са бројним значењима, примерима, потврђеним и у изразима и пословицама.

У овом раду ће се анализирати подела речничког чланка РСАНУ на следеће целине: 1) одредница, 2) граматички подаци, 3) етимологија[11], 4) терминолошка област, 5) језичка и стилска вредност одреднице, 6) испорећене речи за целу одредницу, 7) дефиниција (описна, синонимска или упућивачка), 8) синоним, који се наводи иза скраћенице *син*., антоним иза ; и скраћенице *супр*., упућенице иза скраћеница *вар./исп*., 9) примери, који следе иза . — , 10) извори, који се налазе у загради, 11) изрази

[11] Речник САНУ није етимолошки речник. Подаци о пореклу неких страних речи (не дају се за оне које су стекле карактер домаћих речи – школа, боја и сл.) имају улогу нормативне и стилистичке информације, а доносе се тако што се наводи име народа од кога смо реч примили без улажења дубље у порекло, а за интернационалну лексику указује се на матични језик. В. о томе М. Пешикан, *Наш књижевни језик на сто година послије Вука*, Београд: Друштво за српскохрватски језик и књижевност СР Србије, 1970, 164–166.

(синтагматски и фразеолошки), који се наводе у посебном одељку иза скраћенице Изр. и 12) пословице иза скраћенице НПосл.

У следећем тексту приказан је пример дигиталне верзије чланка обележене TEI[12] XML етикетама, које се потом могу форматирати коришћењем каскадних стилских листова CSS[13] и трансформисати на различите начине, пригодно за штампу или приказ на вебу. Када је речнички чланак структуриран и јасно подељен на логичке целине XML етикетама, могуће је из истог текста генерисати различите визуелне репрезентације рецимо одреднице масним словима, дефиниције писане плавим курзивом, примери зеленом бојом, изворе магента, квалификаторе засебно… Но предуслов за исправно структурирање и дељење на логичке целине јесте да се направи формалан опис речничког чланка. Једноставан пример речничког чланка:

омӯхāч, -а́ча м агр. покр. *врста грожђа* (Крижевац, Борј.).

који је преведен у TEI запис има следећи облик:

```
<entry n="1">
 <form type="lemma"><orth>омӯхāч</orth>,</form>
 <form type="inflected">-а́ча</form>
 <gramGrp><gen>м</gen></gramGrp>
 <sense><usg type="dom">агр.</usg>
     <usg type="geo">покр.</usg>
     <def>врста грожђа</def>
     <cit><bibl>(<placeName>Крижевац</placeName>, <author>Борј.</author>).</bibl></cit>
  </sense>
  </entry>.
```

Речнички чланак подељен је на информационе целине, које припадају одређеним информационим класама, у XML запису обележене етикетама и атрибутима. На пример етикета *<gen>* означава граматички род, док *<usg type="dom">* упућује на терминолошку област, <def> на дефиницију. Угнежђивање XML етикета омогућава и да се библиографски извор обележи са <cit><bibl>, а онда у оквиру њега да се даље структурирају информације, на пример прецизира место коришћењем етикете *<placeName>* или аутор етикетом *<author>*. У оквиру овог истраживања урађено је делимично поравнање скраћеница које се користе у РСАНУ са етикетама стандарда TEI и LMF, колико је било могуће у овој фази дигитализације.

Опис микроструктуре почећемо општим целинама које се у даљој разради описују са специфичностима за поједине врсте речи. Слика 2 приказује основну структуру речничког чланка, где у заглављу имамо одредницу, затим показатеље њене граматичке категорије[14] и етимологију. За овим следе терминолошке скраћенице да означе којој области припада дефинисана реч (бот., зоол. мат. итд.), па квалификатори језичке и стилске вредности, који се налазе пре дефиниције (покр., заст, арх., вулг. итд.)[15]. Систем дефиниција обухвата најчешће описну дефиницију и синоним(е)[16], некад само једно од то двоје, а некад је дефиниција упућивачка (в.), којом се читалац усмерава на правилнију, обичнију или, код једнаких по вредности, прву по реду

[12] http://www.tei-c.org/

[13] http://www.w3.org/Style/CSS/

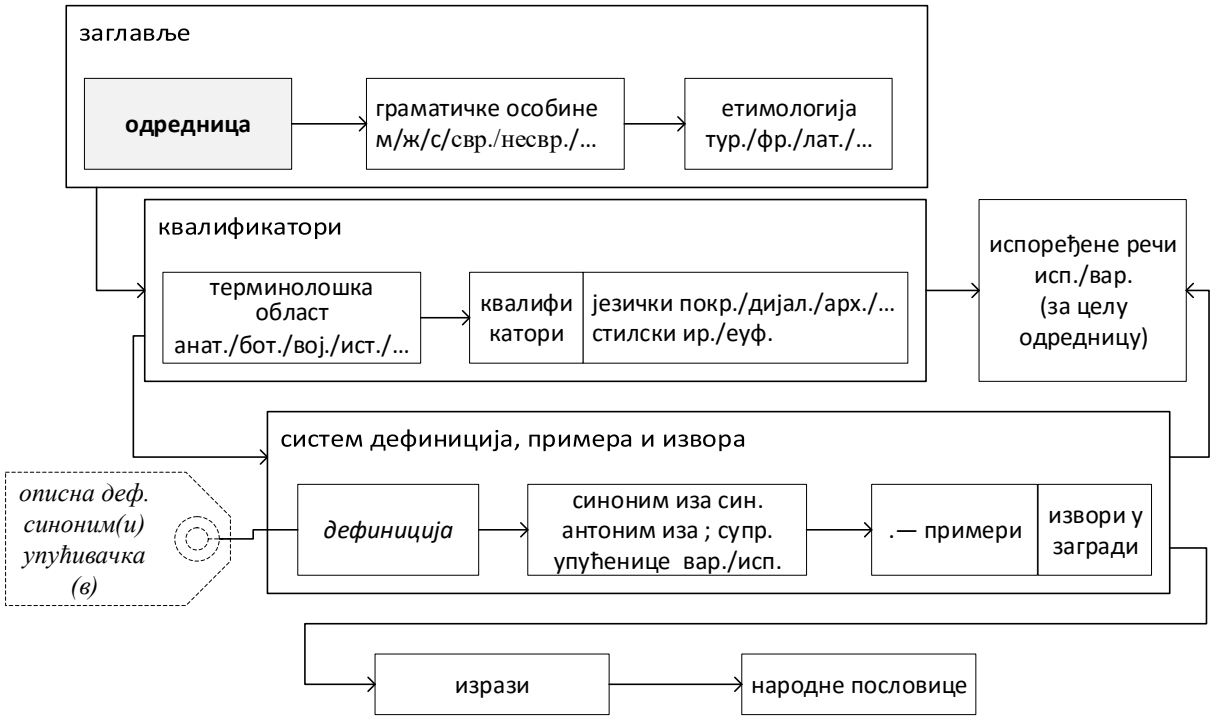
[14] У случајевима акценатских и морфолошких одступања неких облика (падежа и броја код именица, броја, падежа и вида код придева, конјугације код глагола, компаратива код придева и неких прилога) наводе се у загради, после ознака граматичке категорије, и ови облици.

[15] Могу се односити на целу реч или на појединачно значење када долази непосредно иза броја који означава значење.

[16] Обично није реч о потпуним синонимима већ о речима које својим блиским значењем употпуњују описни део дефиниције.

лексему (анатемник *в. безбожник*, оцеан *в. океан* и др.)[17]; иза тачке запете долазе фонетске варијанте и антоними ако их има[18], као и испорећења, која су углавном извесно поткрепљивање дефиниције и која могу припадати и другим граматичким врстама, али која на неки начин доприносе разумевању значења одреднице (бесан … 1. а. *који је оболео од беснила, побеснео*; исп бес[1] (2а)). За дефиницијама, одвојени тачком и цртом, следе примери, који потврђују и илуструју одређено значење дате речи, иза којих се у загради наводи прецизан извор одакле пример потиче; евентуално се даје само извор у коме је реч потврђена без примера (в. ниже).

Појединачна значења вишезначних речи обележена су арапским бројевима одвојеним у посебне пасусе или, ако су значења блиска, словима азбуке (у оквиру истог пасуса), док се преливи једног значења, његове најблаже нијансе наводе у оквиру исте дефиниције, раздвојени само тачком и запетом; ређе се јавља још један ниво хијерархије обележен са 1), 2), 3). Нерефлексивна и рефлексивна форма глагола обележавају се римским бројевима I. и II. Пример разгранатих значења може се видети у речничком чланку одреднице *очистити*, где имамо дубину четири кода: I.-7-а.-1). После свих значења, у посебном пасусу, доносе се изрази (синтагматски и фразеолошки) са ознаком Изр. За обележавање различитих значења израза користе се арапски бројеви са заградама. Речнички чланак завршава се народним пословицама означеним са НПосл и датим у посебном пасусу.



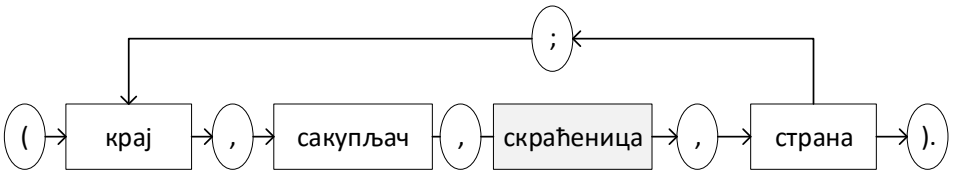
Слика 2. Општа микроструктура речничког чланка

Сваки од сегмената речничког чланка даље се рашчлањује, а саму нотацију описаћемо на примеру извора, који се наводе у загради. Одредница се најчешће илуструје примером, иза кога се наводи извор. У неким случајевима се не наводе примери већ се одредница потврђује само изворoм или са више њих и тада се у загради, без тачке и црте после дефиниције, наводи тај извор или листа извора раздвојених

[17] Иза упућивачке дефиниције следи описна у случајевима када се упућује на правилнију реч која ће у Речнику тек бити обрађена у неком од наредних томова: аналој … *в. налоњ, покретни сточић, сталак на склапање, пулт за књигу или икону*.

[18] Неки од поступака, попут навођења антонима, примењени су само у првим томовима речника.

тачком и запетом.[19] Сваки извор има своју скраћеницу, која се налази у каталогу скраћеница.[20] За илустрацију ево неколико примера са различитим нивоима детаљности, односно са различитом структуром: (Ков. Душ. 1, 197), (Степановић К., Пол. 1957, 16033/21), (Дворска, Јадар, Нен. Н.), (Бачка, Ост. Т. 1), (Срб., Поп. П. М.), (Лика, Богд. Л.), (Дубица, Хрв., Хрваћ.), као и унакрсно референцирање на друге изворе: (БВ 1908, 64; ДК 1912, 119; Сп. ЧГ, 272; Кар. 1903, 47; Сп. СЂБ, 631). Слика 3 приказује формални опис извора којим је могуће препознати све наведене типове примера. У засивљеним правоугаоницима налази се обавезан део, док су остали опциони. Скраћеница треба да се сравни са каталогом скраћеница који се формира на основу списка скраћеница наведених у РСАНУ.



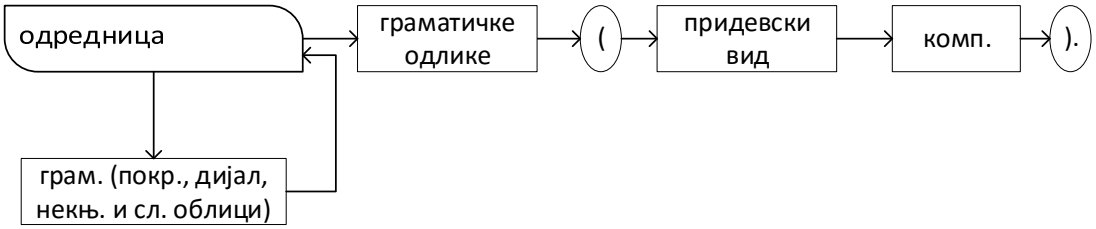
Слика 3. Формални опис извора

За препознавање врста речи креирана су правила, јер у речничком чланку не постоји експлицитно забележена информација о свим врстама речи.[21] Индиректни показатељ граматичке категорије именица јесте ознака њиховог рода (м/ж/с), а глагола ознака глаголског вида (свр./несвр.). Придеви се доносе у сва три рода, што такође посредно указује на врсту речи. Слика 4 илуструје пример карактеристичних елемената речничког чланка придева. У заглављу речничког чланка наводи се одредница у мушком роду, после чега долазе наставци за женски и средњи род: пепељаст, -а, -о; ако има гласовних промена, наставци су другачији: активан, -вна, -вно; згодан, -дна, -дно; бео, бела, бело (у овој групи придева требало је сачинити коначан списак могућих наставака). У загради следе граматичке одлике придева: одређени вид (одр. или обично одр.), па компаратив (комп.); у новој загради дају се дијалекатски и некњижевни облици придева: оштар … (комп. оштрији, -а, -е) (дијал. остар, -тра, -тро), осим ако имамо само неки сегмент некњижеван, онда он иде у прву заграду, мада има и ретких другачијих решења, као нпр.: благ, блага, благо (одр. благи, -а, -о; комп. блажи, -а, -е, непр. блажији; заст. мн. блази).

[19] То је углавном случај са збиркама речи из народних говора у којима су скупљачи навели само реч и крај у коме се она употребљава, као и поједини речници или енциклопедије.

[20] Ако је реч о збирци из народних говора, наводи се место или крај (или обоје, ако је место мало или није нимало познато) где је реч записана па, иза запете, име скупљача. Ако се ради о писаном извору, даје се скраћеница аутора или дела ако нема аутора (обично непотписани чланци у новинама, енциклопедијама, разна документа и сл. грађа) и страна на којој се налази пример.

[21] За неке одреднице постоје директни показатељи граматичке категорије, као, нпр. за прилоге, предлоге, везнике, узвике и речце. Уз њих се наводе скраћенице прил., предл., везн., узв. (осим ако је узвик део дефиниције: пам и пам … *узвик за подражавање звука при ударању, лупању и сл.*), а за речцу цела реч. Информација о класи заменица даје се квалификатором зам. или у оквиру дефиниције, док је податак о класи бројева увек део дефиниције. О директним и индиректним показатељима граматичке категорије в. у Гортан–Премк 1980: 108–109. Исп. и Марковић 2014: 72–82.



Слика 4. Пример детаља за придеве

Остатак речничког чланка исти је као код других врста речи и обухвата етимологију, квалификаторе употребне и стилске вредности, дефиницију италиком за којом следи тачка и црта па примери са извором у загради (или само са извором/изворима без примера). Једини обавезни делови речничког чланка јесу: придев у сва три рода (или назнака да је придев непроменљив, нпр.: бадемли прид. непром.), дефиниција и извор.[22] На слици 5 дат је пример поделе речничког чланка једног придева на информационе целине.

Слика 5. Пример рашчлањавања речничког чланка

3. Поређење микроструктуре првог и последњег тома Речника САНУ

Систематизоване скраћенице примењене су за препознавање информационих целина описаних моделом микроструктуре у претходном одељку на примеру првог и деветнаестог тома. У овом раду дајемо кратак преглед урађеног истраживања, које је започето поређењем са резултатима описаним у (Ђинђић 2014). Аутоматском процедуром је препознато и обележено у првом тому 15.988 речничких чланака док је у деветнаестом тај број 11.200, што је блиско налазу М. Ђинђић са 11.511 одредница. Аутоматска процедура није евалуирана ручно, тако да је разлика вероватно последица неодговарајућег препознавања аутоматском процедуром или типографском грешком у дигиталној верзији изворног текстуалног документа.

Урађена је анализа према врстама речи и поређене су апсолутне и релативне фреквенције, што је приказано на слици 6 за најучесталије врсте речи. Када су именице у питању, препознато је 10.633 (66.2%) у првом тому и 7.808 (69.7%) у деветнаестом тому, потом глагола 1.364 (8.5%) односно 1.192 (10.6%), придева 2.654 (16,5%) односно 1.391 (12,4%). У првом тому за 607 (3.8%) одредница није одређена врста речи, а у деветнаестом за 521 (4.7%). Анализом непрепознатих врста речи уочено је да су у питању углавном придеви. Допуном правила у наредном периоду и они се могу препознати, као и, истина ретки, речнички чланци који се састоје само од одреднице и дефиниције, нпр.: „**пето-** *први део сложеница којим се означава да је оно што значи други део сложенице састављено, сложено…*”

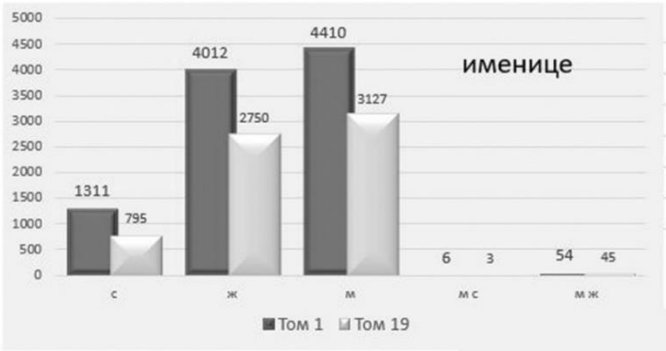
[22] У придевској служби јављају се и неке именице. У том случају се код именице то експлицитно каже, нпр.: алем м … 2. (у прид. служби) непром.

pos	Том 1	Том 19	Том 1 %	Том 19 %
?	607	521	3.8	4.7
им.	10633	7808	66.2	69.7
гл.	1364	1192	8.5	10.6
прид.	2645	1391	16.5	12.4
прил.	670	259	4.2	2.3
речца	11	12	0.1	0.1
узв.	140	17	0.9	0.2

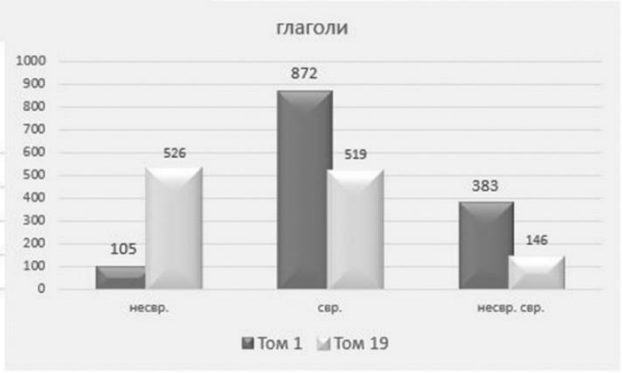
Слика 6. Анализа фреквенција одредница према врстама речи у првом и деветнаестом тому

Даље су анализиране именице према роду и глаголи према виду. Систем је аутоматски препознао у првом тому 1.311 именица средњег рода, 4.012 женског, 4.410 мушког граматичког рода, док је 6 имало мушки и средњи, а 54 мушки и женски. У деветнаестом тому је 795 средњег, 2.750 женског, 3.127 мушког, док је мушки и средњи род препознат за 3 одреднице и 45 мушки и женски. Ове резултате треба посматрати као прелиминарне, све док не буде урађена ручна евалуација и, по потреби, корекција аутоматски препознатих информационих целина.

Што се тиче глагола, у првом тому је 105 несвршених, 872 свршена и 383 двовидска, док је у деветнаестом тому 526 несвршених, 519 свршених и 146 двовидских.



	Том 1	Том 19
с	1311	795
ж	4012	2750
м	4410	3127
м с	6	3
м ж	54	45



	Том 1	Том 19
несвр.	105	526
свр.	872	519
несвр. свр.	383	146

Слика 7. Анализа именица према роду и глагола према виду

4. Закључак

Истраживање је показало да је формални опис структуре речничког чланка РСАНУ могућ. Експерименти са дигитализацијом 1. и 19. тома РСАНУ показали су да је скенирани и кориговани текст речника могуће превести у структурирани документ у стандардизованим форматима. Тако обликован речник је у формату који се непосредно може унети у базу података, потом визуелизовати и претраживати по разноврсним сценаријима. У раду су представљени различити примери формалног описа сагласни са међународним, општеприхваћеним стандардима. Изазови у превођењу неструктуриране дигиталне форме у презентовани модел илустровани су на примерима. Приказани примери представљају фреквенцијску анализу одредница у речнику на основу препознатих граматичких обележја, на примеру поређења првог и деветнаестог тома на којима је тестиран развијени модел. Даља истраживања треба усмерити на обраду осталих томова, унапређење и имплементацију модела, као и евалуацију решења кроз развој система за управљање лексичком базом.

Литература

Бјелаковић 2016: И. Бјелаковић, Електронско издање *Речника славеносрпског језика* (предности, проблеми, могућности), у: С. Ристић и др. (ур.), *Лексикологија и лексикографија у светлу савремених приступа,* Београд: Институт за српски језик САНУ, 169–180.

Витас 1979: Д. Витас, Приказ једног система за аутоматску обраду текста, Симпозијум *INFORMATICA '79*, Блед, 710.

Витас 1980: Д. Витас, Генерисање именичких облика у српскохрватском, *Informatica* 80(3), Љубљана: Словеначко друштво за информатику, 49–55.

Витас 1982: Д. Витас, Приказ једног система за аутоматску обраду текста, Зборник са II научног скупа *"Рачунарска обрада лингвистичких података"*, Блед: Институт Јожеф Стефан, 457–465.

Витас, Крстев 2015: Д. Витас, Ц. Крстев, Нацрт за информатизовани речник српског језика, *Научни састанак слависта у Вукове дане*, 44/3, 105–116.

Гортан-Премк 1980: Д. Гортан-Премк, О граматичкој информацији и семантичкој идентификацији у великом описном речнику, *Наш језик* XXIV/3, 107–114.

Драгићевић 2007: Р. Драгићевић, О Једнотомнику и поводом њега, Речник српског језика, *Књижевност и језик*, LIV 3–4, Нови Сад: Матица српска, 407–412.

Ђинђић 2014: М. Ђинђић: Деветнаести том Речника САНУ, *Наш језик* XLV/3−4, 115–119.

Ивановић и др. 2016: Н. Ивановић, М. Јакић, С. Ристић, Грађа Речника САНУ – потребе и могућности дигитализације у светлу свременихг приступа, *у:* С. Ристић и др. (ред.), *Лексикологија и лексикографија у светлу савремених приступа,* Београд: Институт за српски језик САНУ, 133–154.

Иде и др. 1993: N. Ide, J. L. Maitre, J. Véronis, Outline of a model for lexical databases, In *Information Processing and Management*, 29/2, 159–186.

Калцолари и др. 2013: N. Calzolari, M. Monachini, C. Soria, LMF – Historical Context and Perspectives, in: *LMF Lexical Markup Framework*, Eds: G. Francopoulo, P. Paroubek, John Wiley & Sons, Inc.

Марковић 2014: А. М. Марковић, Граматика у српским речницима, *у:* Р. Драгићевић (ред.), *Савремемна српска лексикографија у теорији пракси*; Београд: Филолошки факултет, 69–91.

Мек Кре и др. 2011: J. McCrae, D. Spohr, P. Cimiano. Linking Lexical Resources and

Ontologies on the Semantic Web with Lemon, in: *Extended Semantic Web Conference* 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg.

РСАНУ 1959–2014: *Речник српскохрватског књижевног и народног језика САНУ*, I–XIX, Београд: САНУ и Институт за српски језик САНУ.

Сабо, Витас 1998: О. Сабо, Д. Витас, Могућност осавремењивања израде речника на примеру Речника српскохрватског књижевног и народног језика САНУ и Института за српскохрватски језик, IV међународни научни скуп „*Рачунарска обрада језичких података*", Порторож: Институт Јожеф Стефан, 375–384.

Трипел 2006: T. Trippel, The Lexicon Graph Model: *A generic model for multimodal lexicon development*, AQ-Verlag, Saarbrücken, Germany.

Упутство: Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017.

Rada Stijović, Ranka Stanković

DIGITAL EDITION OF THE SASA DICTIONARY: FORMAL DESCRIPTION OF SASA DICTIONARY MICROSTRUCTURE

Summary

In the Guidelines for Dictionary preparation and writing of the Serbo-Croatian Literary and Vernacular Language (SASA Dictionary), detailed and precise procedure and rules are presented and describe the microstructure of the lexical entry, including the resolution of possible perplexity. These Guidelines are intended for composition of the printed version of the SASA Dictionary, so certain guidelines refer to the graphical formatting of the lexical entry (use of a certain type of letter, etc.).

In the preparation of the digital edition of the Dictionary, the microstructure of the dictionary article must be described in a way that will allow various types of searching possibilities of the Word text (outside the usual order determined by the alphabet order of the lexical entries). In this context, the paper describes the way in which the dictionary article units are recognized, marked with XML tags, in accordance with international standards (LMF, TEI). The microstructure of the dictionary article described in this way is not dependent from the possible graphic presentation, but it allows a multi-layered search of the Dictionary text.

Key words: computer lexicography, lexicographic workstation, language resources, lexical entry, Serbian language, SASA Dictionary.