Аутоматска екстракција дефиниција – допринос убрзању израде речника

Рада Стијовић, Цветана Крстев, Ранка Станковић

ISBN 978-86-82873-82-2, – (2021), стр. 113–137
UDK: 811.163.41'374
371.671
COBISS.SR-ID: 54614281

Рада Р. СТИЈОВИЋ
Институт за српски језик САНУ
rada.stijovic@isj.sanu.ac.rs

Цветана Ј. КРСТЕВ
Универзитет у Београду, Филолошки факултет
cvetana@matf.bg.ac.rs

Ранка М. СТАНКОВИЋ
Универзитет у Београду, Рударско-геолошки факултет
ranka.stankovic@rgf.bg.ac.rs

АУТОМАТСКА ЕКСТРАКЦИЈА ДЕФИНИЦИЈА – ДОПРИНОС УБРЗАЊУ ИЗРАДЕ РЕЧНИКА

У раду су представљени резултати прелиминарних испитивања аутоматске екстракције кандидата за речничке дефиниције из неструктурираних текстова на српском језику, са циљем убрзања израде речника. Коришћењем постојећих речника српског језика анализирани су различити начини моделирања и екстракције дефиниција. Анализа је показала да велики број дефиниција има структуру коју је могуће моделирати, чему сведоче статистика дефиниција груписаних по типовима, пример модела и евалуација на уџбеницима биологије.

Кључне речи: дескриптивни речници, метаанализа лексикографских дефиниција, аутоматска екстракција дефиниција, електронски речници, српски језик.

1. Увод

У доба електронске лексикографије у коме се за производњу речника користе информатичка средства, посебна пажња посвећује се аутоматском или полуатоматском обављању неких задатака. Тако се приступило и аутоматизацији неких од задатака укључених у израду речника српског језика. Истраживања везана за екстракцију пригодних речничких при-

мера (GDEX – good dictionary examples) показала су да се екстракција информација из корпуса грађе за речник може користити за убрзање рада на Речнику српскохрватскога књижевног и народног језика Српске академије наука и уметности (РСАНУ), али и других речника (Станковић et al. 2019), тако да се кренуло и у правцу екстракције реченица које могу послужити као речничке дефиниције. Уз то, корист од екстракције свакако представља и препознавање парадигматских лексичких односа, што подразумева проналажење синонима, антонима, хиперонима, хипонима. Коришћење електронских речника може убрзати рад на речнику и аутоматским придруживањем граматичких информација: врста речи, облика речи, различитих типова маркера.

Проблем аутоматске екстракције дефиниција из текста до сада није темељно истраживан за српски језик. Претпоставка је да би његово решавање могло знатно да допринесе убрзању израде речника.

При истраживању представљеном у овом раду кренуло се од детаљне анализе различитих лексичких и синтаксичких карактеристика дефиниција у Речнику САНУ и од прегледа структуре појединих типова дефиниција који су у кратким цртама описани у Упутствима за израду Речника (в. Упутства РСАНУ 1959 [2017]). Једна од постављених хипотеза истраживања јесте да велики број именичких дефиниција има структуру коју је могуће моделирати.

Ослањајући се на постојеће дескриптивне и морфолошке речнике српског језика, анализирани су различити начини моделирања и екстракције дефиниција. Истраживачки корпус подскупа дефиниција именица из пет томова Речника САНУ (1, 2, 18, 19, 20) анализиран је уз помоћ морфолошких електронских речника српског језика и локалних граматика (Крстев 2008). Први, непосредни циљ истраживања јесте систематизација речничких дефиниција која треба да омогући откривање оних дефиниција које одступају од уобичајених или усвојених модела, на пример при изради речника. Дугорочни циљ је изградња система који би из наменског корпуса за задату реч издвајао реченице које одговарају некој од дефинисаних структура дефиниција.

У другом одељку овог рада говори се о проблему моделирања речничке дефиниције и методама за аутоматску екстракцију кандидата за речничке дефиниције. Трећи одељак даје анализу лексичких и синтаксичких карактеристика појединих типова дефиниција засновану на корпусу дефиниција именица из пет дигитализованих томова РСАНУ, статистику дефиниција груписаних по типовима и укратко представља морфолошке електронске речнике српског језика (е-речнике) и локалне граматике у алату Unitex, на којима се предложени приступ заснива. Прелиминарни модел речничких дефиниција именица евалуиран је на изабраним уџбеницима из биологије, а остварени резултати приказани

су у четвртом одељку, уз примере интеграције дефиниција у различите речнике. Остварени резултати и планови за даље истраживање изложени су на крају рада.

2. Аутоматска екстракција дефиниција

2.1. Појам дефиниције у лексикографији

Према стандарду „Терминолошки рад – Вокабулар – Део 1: Теорија и примена“ (SRPS ISO 1087-1:2003) дефиниција је „представљање неког појма описним исказом који омогућава његово разликовање од сродних појмова“. Овај стандард разликује дефиниције по интензији (садржају) и по екстензији (обиму). Дефиниције неког појма по интензији наводе надређени појам и дистинктивне карактеристике; на пример, за појам „сијалица са усијаним влакном“ оваква дефиниција је *електрична сијалица чије се влакно загрева електричном струјом на такав начин да сијалица емитује светлост*. Дефиниција по екстензији (обиму) даје опис појма набрајањем свих подређених појмова према једном критеријуму потподеле, на пример, оваква дефиниција 18. колоне Периодног система елемената била би *хелијум, неон, аргон, криптон, ксенон и радон*, док би дефиниција ретких гасова била *хелијум, неон, аргон, криптон, ксенон* и *радон*.

Лексикографска дефиниција је идентификација семантичког садржаја одређене лексеме, с релевантним елементима реализације. Семантички садржај се састоји из архисеме, која носи информацију о припадности лексеме некој широј лексичко-семантичкој групи, и сема нижег ранга, које носе информације о индивидуалним карактеристикама лексема, на основу којих се једна лексема разликује од друге у истој лексичко-семантичкој групи. Основни типови лексикографске дефиниције јесу описна, синонимска, комбинована – описна и синонимска. Описна дефиниција без синонима употребљава се код дефинисања основних значења простих, неизведених речи (које немају синониме), а синонимска, коју чине речи блиске или исте семантичке структуре, када постоји довољан број одредница синонимних са том лексемом или када је та лексема маркирана у погледу ширине употребе (Гортан-Премк 2014: 131–132). Комбиноване дефиниције карактеристичне су за велике описне речнике – описним делом се идентификује појам директно и наводе елементи значења, а синонимом се индиректно упућује на семантички садржај (Гортан-Премк 1982: 50–51; Гортан-Премк 1983).

Barnbrook (2002) у делу „Defining Language, A local grammar of definition sentences“ истиче да је дефинисање основна активност језика, која је од посебне важности за лингвисте. Он представља подскуп општег језика енглеског језика који се користи у дефиницијским реченицама, развија таксономију типова дефиниција, као и граматику дефиницијских

реченица. На основу ових резултата развијен је и софтвер за парсирање текста писаног на енглеском који може да издвоји функционалне компоненте. Приступ који се представља у овом раду делом је инспирисан управо Барнбруковом таксономијом дефиниција (структурним обрасцима) и граматиком језика дефиниција.

2.2. Приступи проблему аутоматске екстракције дефиниција

Проблем аутоматске екстракције дефиниција јесте релевантан задатак у различитим областима обраде природних језика: у системима за одговарање на питања (eng. Question Answering – QA) који на основу једног или мноштва извора синтетизују одговор на питање типа „Шта је“, за писање речника и развој онтологија (Navigli & Velardi 2010). Приступи решавању овог проблема често се заснивају на изради и примени лексичко-синтаксичких образаца. На пример, у енглеском језику се за дефиниције често срећу следећи обрасци (Jin et al. 2013):

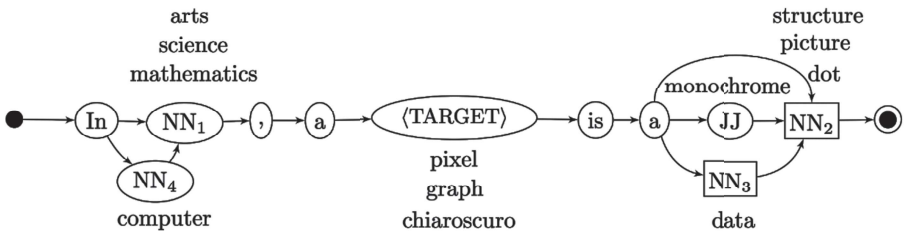
<term> defined (as|by) <definition>
define(s)? <term> as <definition>
definition of <term> <definition>
<term> a measure of <definition>
<term> is DT <definition> (that|which|where)
<term> comprise(s)? <definition>
<term> consist(s)? of <definition>
<term> denote(s)? <definition>
<term> designate(s)? <definition>
<definition> (is|are|also) called <term>
<definition> (is|are|also) known as <term>
“part of/in/on/…”, “type of”, “ is <hyperonym> …”,…

Међутим, аутори Navigli и Velardi истичу да примена оваквих образаца може дати слабији одзив (не препознају довољно дефиниција) и мању прецизност (препознају реченице које нису дефиниције), због врло променљивих синтаксичких структура реченица којима се појмови дефинишу. Стога они предлажу да се за моделирање текстуалних дефиниција користе WordClass Lattices (WCLs), генерализација решетки речи. Метода решетке класа речи је алтернатива лексичко-синтаксичким обрасцима мотивисана побољшањем прецизности и одзива. Решетка (енгл. lattice) је усмерени ациклични граф, поткласа коначних аутомата. Решеткаста структура има за циљ очување главних разлика међу различитим секвенцама, при чему се истовремено уклањају сувишне информације. Сравњивање образаца заснива се на употреби звездице (џокерски знак *), којом се олакшава кластеризација реченица. Свака класа реченица се затим генералише решетком класа речи (свака класа је или нека фреквентна реч или врста речи). Кључна карактеристика приступа јесте

способност идентификације дефиниција и издвајања хиперонима. У следећим примерима[1] дефиниција појмови који се дефинишу означени су подебљаним словима, а њихови надређени појмови курзивом. Слика 1 приказује решетку класе речи за наведене реченице, где се потпора сваке класе речи наводи поред одговарајућег чвора.

[In arts, a **chiaroscuro**]$_{DF}$ [is]$_{VF}$ [a monochrome *picture*]$_{GF}$.
[In mathematics, a **graph**]$_{DF}$ [is]$_{VF}$ [a *data structure*]$_{GF}$ [that consists of . . .]$_{REST}$.

[In computer science, a **pixel**]$_{DF}$ [is]$_{VF}$ [a *dot*]$_{GF}$ [that is part of a computer image]$_{REST}$.



Слика 1. Пример решетке класе речи са потпором класе речи поред одговарајућег чвора

Задатак аутоматске екстракције дефиниција може се формализовати на различите начине. Једна од могућности јесте да се посматра као проблем класификације реченица на оне који су потенцијални кандидати за дефиницију неког појма и на оне које то нису, односно као проблем утврђивања да ли реченица садржи пар појам–дефиниција или не. Аутоматску екстракцију дефиниција можемо поставити и као задатак анотације где се захтева идентификовање граница појмова и њихових дефиниција. На пример, у реченици *„Имунитет је отпорност организама на болести…“* одредница (или термин у терминолошком речнику) и њена дефиниција могли би бити обележени XML етикетама на следећи начин:

<odr>Имунитет</odr> <rel>је</rel> <def>отпорност организама на болести</def>…

У литератури се могу срести различити приступи екстракцији дефиниција, а најчешће коришћени су они који се заснивају на лингвистичким правилима и шаблонима који обухватају обрасце за изражавање релације појма и дефиниције; потом статистичким моделима машинског учења са синтаксичким и семантичким карактеристикама (фичерима, енг. features), а у новије време и на дубоком учењу, користећи угнежђене репрезентације речи (енг. word embeddings) путем вишеслојних неуронских мрежа (Veyseh et al. 2019).

[1] Примери су наведени према (Navigli & Velardi 2010).

Морфосинтаксички обрасци примењени на доменском корпусу из области рачунарске лингвистике користе се за екстракцију кандидата за дефиниције за словеначки и енглески језик (Polak et al. 2012), аутоматско препознавање терминологије и семантичко означавање ворднет (енг. WordNet) значењима. Алат развијен на основу резултата овог рада[2] користи обрасце облика „NP is NP“, „NP refers to NP“, „NP denotes NP“, при чему се NP односи на именичку фразу (енг. noun phrase). За детектовање дефиниција аутори користе и словеначки ворднет: реченице које потенцијално садрже дефиниције почињу речју из ворднета, а даље садрже бар још једну реч из истог ланца релација надређени или подређени. Систему се параметарски прослеђује информација о језику текста у коме се детектују дефиниције (EN, SL), као и положај појма у словеначком (да ли појам мора бити на почетку реченице, након унапред дефинисаних почетних образаца, што је подразумеван избор, или може бити било где у реченици).

Spala и др. (2019) уз корпус дају детаљнију шему анотације како би се истражиле разнолике структуре дефиниција термина у слободном и полуструктурираном тексту. Осим основног појма (Term) који се дефинише и главне дефиниције (Definition), анотирају се и сегменти реченице који садрже псеудониме или додатне називе (Alias Term), који се повезују са основним термином, потом именичке фразе (Referential Term) које се реферишу на претходно обележени термин, секундарне дефиниције (Secondary Definition) са додатним информацијама, квалификаторе (Qualifier) који спецификују евентуалне услове под којим дефиниција важи и слично. Обележени сегменти се повезују одговарајућим релацијама.

Анализом вертикалног уланчавања концепата присутних у њиховим лексикографским дефиницијама баве се Ристић и др. (2018) на примеру тематског поља ‘кућа, грађевински објекат’, указујући на низ од највиших нивоа појмовне категоризације, сложених и простих примитива, према нижим хиперонимским нивоима (МЕСТО > ПРОСТОР [ГДЕ], ПРОСТОР > ПРОСТРАНСТВО [НЕШТО ШТО се (безгранично) простире на све стране]). Наводе такође да би овај тип истраживања могао да допринесе унакрсном претраживању лексема и другим напредним претрагама у електронском издању дескриптивног речника савременог српског језика, што је значајно како лексикографима тако и корисницима речника.

Контекст у ком се у нашем раду информатички моделирају дефиниције ради аутоматске екстракција јесте помоћ при изради речника, односно скицирање речника (eng. Dictionary drafting) (Kilgarriff & Rychlý 2010), које подразумева развој заснован на корпусу. Резултати су постиг-

[2] Алат је доступан онлајн на адреси http://clowdflows.org/workflow/8165/

нути коришћењем електронског речника српског језика и локалних граматика развијених на основу резултата претходних сличних истраживања (Barnbrook 2002), анализе дефиниција у постојећим речницима и наших претходних истраживања (Крстев et al. 2015).

3. Анализа и препознавање дефиниција у Речнику САНУ

Анализа лексичких и синтаксичких карактеристика дефиниција почела је истраживањем дефиниција у Речнику САНУ (РСАНУ). Упутство за обраду Речника даје неформалан опис структура дефиниција, где се наводи да прво долази описна дефиниција, а затим дефиниција путем замене сродним речима. На пример: *ажурност ... особина, стање онога што је ажурно; тачност, уредност.* Најближа значења могу се одвојити само ';' и тако представљају једну дефиницију; на пример: *бајаџо ... циркуски лакрдијаш, кловн; човек који је маскиран као кловн*.

Илуструјмо пар модела датих у смерницама: 1) ако именицу треба дефинисати релативном реченицом, дефиниција треба да започне речима „онај који …“, за разлику од придева где дефиниција почиње са „који …“ ; 2) за апстрактне именице које се завршавају са -ост, -ство, -ота, -оћа, -ина итд., прво се даје општа дефиниција: особина (и/или стање, или својство) онога који је … онога што је … (после чега следи придев који је у основи те апстрактне именице), и евентуално 2–3 синонима за допуну дефиниције.

За изградњу система за аутоматску екстракцију неопходно је формализовати моделе типских дефиниција, за шта само Упутство није било довољно, тако да се јавила потреба за креирањем корпуса за анализу структуре дефиниција. Као најбоље решење показао се РСАНУ, у коме су лексикографски поступци, утемељени на Упутствима, током израде Речника усавршавани, у складу са најновијим лингвистичким сазнањима, без нарушавања основне концепције. Сарадници на изради Речника, бавећи се у свом научном раду семантиком и простих, неизведених речи и деривата, као и организацијом лексичког система, уграђивали су своје теоријске лексиколошке синтезе у лексикографски рад. У складу са лингвистичко-лексиколошким начелима наше науке о језику уобличена је у Речнику једна у великој мери уједначена синтаксичка и лексичка структура дефиниција.[3] У Речнику се тежи идентификационим тумачењима, што подразумева да је дефиниција у граматич-

[3] Бројни лексикографски и технички поступци разасути по томовима Речника, познати су само лексикографима који израђују Речник, али се све више актуелизују као предмет појединачних истраживања у светлу савремених приступа различитих усмерења. Осим наведених аутора, в. и: Ристић 2006, 2009 и 2015; Ристић, Лазић Коњик 2020; Новокмет 2020 и др.

ком и у значењском смислу што верниja замена одреднице.[4] Осим тога, настоји се да дефиниције буду што сажетије, али да притом исказују све што је релевантно, и да буду максимално унифициране (Грицкат 1981: 5, Грицкат-Радуловић 1988: 36, Фекете 1993: 44).[5] Приликом дефинисања користи се хијерархијска организација лексичког система, тј. наводи се најпре хипероним речи која се дефинише, а потом се доносе њене специфичне карактеристике, критеријска обележја појма који се дефинише, на основу којих се успостављају сличности и разлике у односу на друге појмове, чији називи чине уређен низ хипонима истог надређеног појма, хиперонима. На тај начин се, осим хијерархијске уређености лексичког система, успостављају и системски односи између његових јединица – лексема у виду граматичко-семантичких класа, које захтевају и унифицирани начин дефинисања.

Инспирисани подухватима великих лексикографских кућа и истраживачких центара који су препознали могућности које пружају лексикографске базе, посебно у подржавању савремених начина на који се корисник служи речницима (Rundell 2014), кренуло се у формализацију микроструктуре речничког чланка РСАНУ (Стијовић и Станковић 2017).

---

[4] Потпуна идентификација дефиниције и одреднице подразумева да се речи одређене граматичке врсте или граматичког облика дефинишу помоћу речи те исте врсте и облика (придев придевом, инфинитив инфинитивом итд.) или помоћу синтагми које функционишу као јединице датог граматичког типа (бео = који је боје снега и сл.) (Грицкат 1981: 4–5).

[5] Као пример уједначености могу да послуже дефиниције ботаничких и зоолошких, али и неких других термина и група речи. Када је реч о терминима, одражава се хијерерхијска структура у таксономији (врста–род–породица/фамилија итд.): **брмељ** … *бодљикава зељаста биљка* Carthamus (Kentrophyllum) lanatus *из ф.[амилије]* Compositae, **бросква** … *зељаста биљка* Brassica napus rapifera (Napobrassica) *из ф.* Cruciferae, *која се гаји због меснато задебљалог корена за јело;* **калош** … *украсна биљка* Anemone coronaria *из. ф.* Ranunculaceae; **бргљун**… *врста сарделе дуге до 20 см* Engraulis encrasicholus *из пор.[ородице]* Engraulidae, **видра** … *дивља животиња рђастосмеђе боје крзна* Lutra lutra *из пор. куна* Mustelidae, *која живи поред вода и храни се рибом,* **оштригар** … *врста птице вивчарице* Haematopus ostralegus *из пор.* Charadriadae, *која живи на обалама мора и река.*

Називи месеци су потпуно унифицирани по моделу: **јануар** … први месец у години, сијечањ; **март** … трећи месец у години, ожујак, **мај** … пети месец у години, свибањ итд.

У народном именовању животиња по некој особини користи се формула *назив за + особина + и име таквој животињи*: **бањов** … *назив за бањаста риба и име таквој животињи,* **баран** … *назив за црно-бела јарца и име таквој животињи,* **баца** … *назив за овцу без белеге на глави и име таквој животињи.*

У меронимији: *део тела, орган, део биљке*: **глава** … *горњи део човечијег, односно предњи део животињског тела у коме се налази мозак и главна чула,* **врат** … *део тела који везује главу и труп,* **димњаци** … ***а.*** *део човечјег тела између кукова и пупка; препоне.* ***б.*** *бочни део тела између грудног коша и кука, слабине,* **врећица** *део тела у облику кесе* **а.** *трбушна марамица* mesenterium. **б** *лежиште неког болесног израштаја.*

Лексичка база података је пројектована тако да може да складишти структурирани запис речничког чланка у релациону структуру, а софтверско решење омогућава трансформацију неструктуираног текста Word документа у релациону базу (Станковић *et al.* 2018).

За потребе ових истраживања сачињен је корпус састављен од дефиниција екстрахованих из пет томова Речника САНУ. Из овог корпуса издвојено је и анализирано 62.710 дефиниција за именице с циљем моделирања дефиниција за разне класе именица.

Дефиниције именица из РСАНУ анализиране су помоћу српских морфолошких електронских речника и локалних граматика (Крстев 2008) реализованих помоћу коначних трансдуктора имплементираних у алату Unitex[6]. Електронски морфолошки речници српског језика намењени искључиво аутоматској обради развијају се већ дуги низ година, те њихова данашња величина и садржај обезбеђују успешно коришћење у реалним апликацијама обраде српског језика. Ови речници садрже аутоматски генерисане облике речи више од 150.000 лема, а њихов садржај покрива како општу лексику тако и властита имена – лична, геополитичка, имена организација и сл. Такође, речник садржи и полилексемске јединице (енг. *multi-word units*), које се у традиционалним речницима бележе као синтагме или фраземе. Основна јединица ових речника јесте облик речи коме је придружена лема (најчешће одредница традиционалног речника), врста речи, маркери који описују њена синтаксичка, семантичка, употребна, дијалектска, доменска својства, а потом и кодови морфосинтаксичке реализације.

Коначни трансдуктори су апстрактне математичке конструкције које омогућавају моделирање локалних граматика за описивање неких језичких конструкција, на пример, именичких фраза. Коначан трансдуктор у свом раду „пролази“ кроз текст који анализира у покушају да сравни део текста с моделом који представља. У случају успешног препознавања, коначни трансдуктор производи неки резултат, што може да буде модификација полазног текста у виду додавања ознака врсте препознатих речи, или тип препознате синтаксне структуре (Vitas & Krstev 2012). Коначни трансдуктори се ради лакше изградње и коришћења визуелизују преко графова, а један пример је дат у овом одељку.

Прво су анализирани ен-грами, ниске од n суседних речи које се јављају у дефиницијама са придруженом фреквенцијом појављивања. За наше потребе издвајали смо ниске са почетка дефиниција од три до највише пет речи које почињу са једним од два карактеристична „окидача“ (енг. trigger) „онај који“ и „која“. У табели 1 дати су обрасци који се

[6] Unitex је софтверски пакет отвореног кода за обраду корпуса који је слободан за коришћење (https://unitexgramlab.org)

појављују више од петнаест пута; може се видети да непосредно иза ова два окидача обично следи глагол.

Табела 1. Фреквенције карактеристичних ен-грама који почињу са „онај који“ и „која“

Израз	Учесталост	Израз	Учесталост
онај који је	449	која служи за	106
онај који се	243	која је пореклом	92
онај који је пореклом	118	која је пореклом из	86
онај који је пореклом из	111	која се бави	77
онај који се бави	79	која се употребљава	48
онај који има	58	која живи у	46
онај који прави	46	која расте на	37
онај који носи	38	која се користи	36
онај који нешто	35	која служи као	28
онај који не	22	која се у	27
онај који нема	22	која се бави проучавањем	20
онај који је без	21	која се добија	20
онај који воли	18	која се ставља	19
онај који даје	18	која се употребљава за	17
онај који много	18	која се гаји	16
онај који некога	18	која живи на	15
онај који ради	17	која расте у	15
онај који врши	16	која се употребљава у	15

На левој страни табеле 2 приказани су примери окидача „врста“, док се у десном делу табеле приказују примери за окидач „назив за“ са примерима образаца који се појављују више од десет пута, а може се видети да иза ова два окидача обично следи именица. Трећу групу примера чине они који се односе на особину или стање, где уочавамо појављивање различитог редоследа речи „особина“ и „стање“, иза којих следи заменица „онога“.

Табела 2. Фреквенције ен-грама за окидаче „врста“, „назив за“ и „особина/стање“

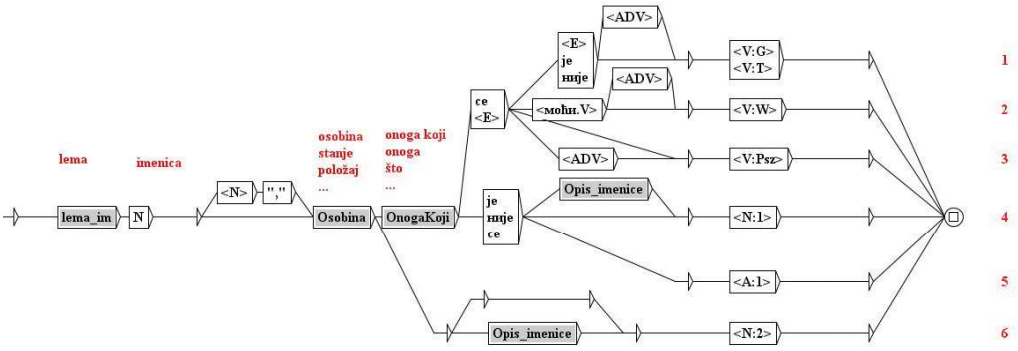
Израз	Учесталост
врста винове лозе	60
врста морске рибе	55
врста народног кола	54
врста дечје игре	22
врста народне игре	21
врста друштвене игре	16
врста јела од	16
врста народног веза	15
врста дивље патке	12
врста рибарске мреже	12
врста слатководне рибе	12
врста речне рибе	11
врста белог грожђа	10
врста шаре у	10

Израз	Учесталост
назив за разне	31
назив за разне врсте	19
назив за неке	18
назив за род/врсте	15
назив за краву/вола	13
назив за више	11
назив за домаће	11
назив за домаће животиње	11
назив за јарца/козу	11
назив за поједине	10

особина онога који	190
особина онога што	182
особина онога који је	180
особина онога што је	175
стање онога који	111
стање онога који је	107
стање онога што	60
стање онога што је	59
особина или стање	18
особина и стање	14
особина и стање онога	13
особина или стање онога	13
стање и особина	13
стање и особина онога	13

Ови подаци помогли су моделирање најчешће коришћених шема дефиниција именица у локалним граматикама. Локална граматика која моделира (и препознаје) дефиниције именица које представљају особину, стање или положај некога или нечега дата је графом који треба овако тумачити: приликом препознавања трансдуктор треба да стигне из почетног стања (стрелица са леве стране) до завршног стања (квадратић у кружићу с десне стране) по неком од путева. У графу са слике 2 препознаје се прво лема, ознака врсте речи (именица), потом окидачи (*особина*, *стање*, *положај* итд., а потом *онога који*, *онога што* итд.), иза чега следи наставак дефиниције описан са шест главних путања (обележених бројевима 1–6), које се ослањају на препознавање придева, именица или глагола у одговарајућим облицима. На пример, у путањи 5, ознака <A:1>

значи да иза окидача и помоћног глагола *јесам* до препознавања води придев у номинативу. Неки карактеристични примери препознавања за сваку од ових 6 путања које моделирају карактеристичне дефинице ове класе именица јесу:



Слика 2. Граф за препознавање дефиниција именица које представљају особину, стање или положај некога или нечега

1. банкротство N стање онога који је банкротирао
   оронулост N стање и особина онога који је оронуо
   везаност N стање, особина онога који је везан
   варљивост N својство онога што се лако вари
   осамостаљеност N особина и стање, положај онога који се осамосталио
2. оповргљивост N особина онога што се може оповргнути
3. бебунство N својство онога који се бебуни
   белина N особина онога што светли
4. боемство N особина онога који је боем
   паланчанство N стање, особина онога који је паланчанин
5. бесрдачност N особина онога који није срдачан
   оријентисаност N положај, особина онога ко је оријентисан
   пластичност N особина, својство, стање онога што је пластично
   перманентност N особина или стање онога што је перманентно
6. беснило N стање душевне поремећености код алкохоличара
   базичност N особина неке базне супстанце

Сличне локалне граматике развијене су и за препознавање дефиниција биљака, животиња и биљних органа. Неки примери препознавања овим локалним граматикама у корпусу РСАНУ јесу:

биљке

памук N род грмоликих (ређе дрволиких) и култивисаних, претежно једногодишњих биљака
палма N општи назив за разне родове већином дрвенастих биљака из ф.
плавина N сорта винове лозе тамноплавог, ситног грожђа

животиње
антилопа N ма која од животиња из ове фамилије
барна N назив за домаћу животињу мрке длаке и име таквој животињи
була N име разним домаћим животињама (кобили, крави, овци, кози, керуши и др.)
део/орган
архегонија N женски расплодни орган код биљака…
палетак N ситан, закржљао или недозрео клип кукуруза…
вампир N неправилно развијени плод шљиве

Анализа дефиниција показала је да велики број именица има веома једноставне дефиниције које једном речју објашњавају лему, на пример за презимена, или само упућују на неку другу лему, на пример *види.* До сада је графовима моделирано 17 типова именичких дефиниција који препознају скоро 50% дефиниција именица у корпусу који чине пет томова РСАНУ. У табели 3 дат је преглед моделираних дефиниција, бројеви препознатих дефиниција за сваку од њих, као и карактеристични примери у којима су окидачи дати у курзиву, именица која се дефинише у малој капитали, а именица са којом се повезује задебљаним словима.

Табела 3. Преглед типова именица чије су дефиниције моделиране локалним граматикама

Тип	**Број препознатих дефиниција**	**Пример (тригер италиком)**
види	15367	ПИРИТ N *в.* **биотит**
презиме	3681	**Витас** N *презиме*
глаголска именица	2337	АДАПТИРАЊЕ N *гл. им. од* **адаптирати**
врста	2111	ПАЛИЈАТА N *врста* староримске **комедије** …
деминутив	1919	АНЂЕЛЧЕ N *дем. и хип. од* **анђео**
име или надимак	913	**Палић** N породични *надимак*
женска особа	743	БАЛАВИЦА N млада незрела **жена** или **девојка**
аугментатив	532	БУРЖУЈЧИНА N *аугм. и пеј. од* **буржуј**
особина или стање	710	БЕЗНАЧАЈНОСТ N *особина онога који је* **безначајан**
географско име	474	**Арабија** N велико *полуострво* у југозападној **Азији**
становник	270	**Бокељка** N *становница* **Боке Которске**
народ	14	**Асири** N ратнички *народ*

остала властита имена	98	**Водолија** N једанаесто по реду зодијачко *сазвежђе*
збирне именице	159	**пашчад** N *зб. им.* од **пашче**
биљке	487	**бесцветница** N систематски *назив* за такве *биљке*
животиње	1254	**бикан** N *назив* за домаћу *животињу* (мужјака) и име таквој *животињи*
део биљака/животиња	94	**палетак** N ситан, закржљао или недозрео *клип кукуруза*
Укупно	31171 (49.7%)	

Локалне граматике које моделирају дефиниције именица (и других врста речи) могу на различите начине допринети изради речника и других лексичких ресурса. Пре свега, приликом израде речника, оне могу помоћи у провери усаглашености дефиниција са усвојеним обрасцима. Осим тога, оне омогућавају аутоматско повезивање речничких чланака у бази, што се може видети и у примерима из табеле 3. Шире, оне омогућавају обогаћивање и повезивање речника са другим лексичким изворима за српски, а овде се у првом реду мисли на морфолошке електронске речнике и семантичку мрежу речи ворднет. Коначно, а што ће бити приказано у наредном одељку, локалне граматике дефиниција могу се користити и за екстракцију дефиниција из доменских корпуса.

4. Анализа на корпусу

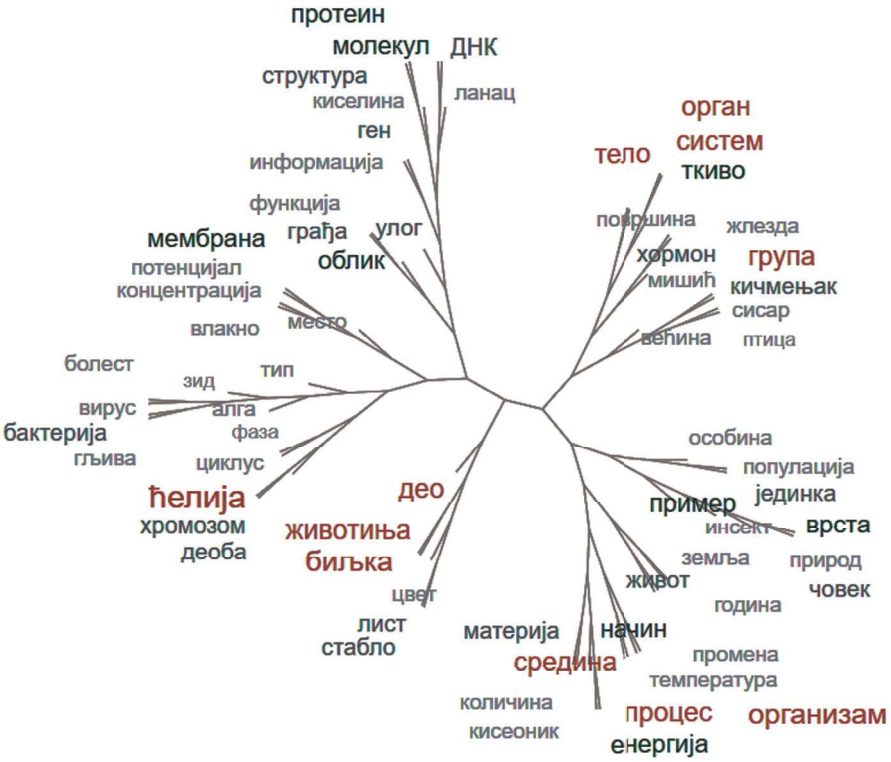
4.1. Креирање корпуса уџбеника

Ради провере хипотезе да је могуће аутоматски екстраховати дефиниције из доменских корпуса локалним граматикама које моделирају дефиниције именица формиран је корпус из домена биологије. Овај корпус текстова креиран је од 6 уџбеника за основну и средњу школу. Уџбеници су сканирани, обављено је оптичко препознавање карактера, а потом су грешке препознавања ручно исправљане. Корпус се састоји од 19.446 реченица, 698.210 (37.345 различитих) токена, 303.829 (37.278 различитих) речи.[7] Корпус је обрађен електронским речницима српског језика који су препознали приближно 238.000 облика речи, 5.000 облика синтагми док је ~5.700 речи остало непрепознато. Међу препознатим синтагмама су *жута грозница*, *аеробне бактерије*, *агрегатно стање*, *биљка месождерка, бела крвна зрнца, вода за пиће, запаљење плућа, угљен-диоксид* и

[7] Под „речима“ се овде подразумевају ниске састављене само од алфабетских карактера, док међу „токене“ осим речи спадају и бројеви исписани цифрама, специјални и интерпункцијски знаци.

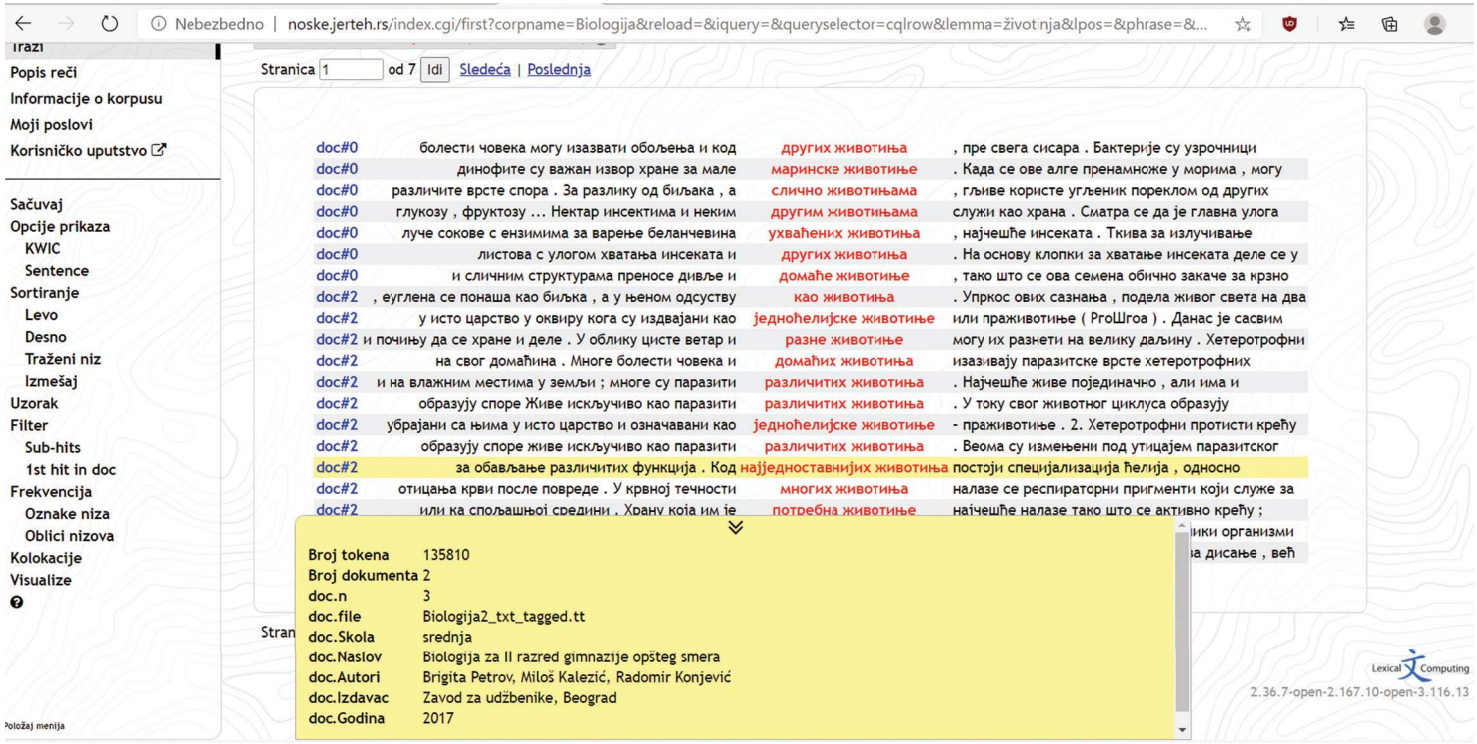
сл. Међу непрепознатим речима су, пре свега, заостале грешке оптичког препознавања карактера, а затим страна имена и скраћенице (DNK, RNK, ATR, HIV). Одређен број речи које спадају у терминологију биологије, као на пример *органела*, *цитоплазма*, *зигот*, *ретикулум*, *хлоропласт* и сл., још увек се не налазе у електронском речнику па су остале непрепознате.

Брзи преглед садржаја текста облаком стабала (TreeCloud) приказан је на слици 3. Попут облака речи, облак стабала приказује најчешће речи у тексту, при чему величина фонта којим су те речи записане одражава њихову учесталост у тексту, док распоред речи на дрвету одражава њихову семантичку удаљеност[8] (Gambette 2009) израчунату на основу текста. Такви облаци стабала помажу у идентификовању главних тема документа, а могу се користити и за анализу текста.



Слика 3. Облак стабала (TreeCloud) корпуса уџбеника биологије

[8] Семантичка удаљеност између две речи у тексту се изводи из њиховог заједничког појављивања (енг. cooccurrence) у тексту (реченици, параграфу, одељку, или некој другој целини). Не постоји јединствена формула за израчунавање семантичке удаљености. За различите примене користе се различите формуле, а више од 20 њих је систематизовано и дефинисано у (Evert 2005).



Слика 4. Пример конкорданци произведених упитним обрасцем <A>+<животиња> над тагираним корпусом

Корпус уџбеника биологије је аутоматски лематизован и тагиран врстом речи и скраћеним граматичким категоријама (Stanković et al. 2020), а потом публикован на платформи за претрагу корпуса Друштва за језичке ресурсе и технологије Јертех.[9] Систем за претрагу једнојезичких и вишејезичких корпуса заснован је на платформи NoSketch Engine (Rychly 2007), која је софтвер отвореног кода, као скраћена верзија комерцијалног алата Sketch Engine.[10] На слици 4 приказано је претраживање лематизованог и тагираног корпуса обрасцем <A>+<животиња> (именица *животиња* у ма ком облику којој претходи придев) и произведене конкорданце. Види се, такође, да је за сваки пример конкорданци могуће утврдити у ком је документу (уџбенику) корпуса пронађен.

4.2. Препознавање кандидата у корпусу уџбеника

Да бисмо утврдили да ли је у тексту доменског корпуса могуће препознати дефиниције појмова специфичних за домен, применили смо на корпус шест уџбеника биологије три локалне граматике представљене у одељку 3, и то оне за дефинисање биљних и животињских врста и делова, односно органа, биљака и животиња. У табели 4 приказани су примери успешног препознавања дефиниција. Треба напоменути да су у неким случајевима препознати само почетни делови дефиниција, на пример, за последњи пример у табели 4 потпуна дефиниција гласи ***Изданак*** *је вегетативни орган биљке који гради стабло са листовима* (подвучени део није препознат). Аутоматско препознавање потпуних дефиниција сложен је задатак, јер захтева граматике за потпуно парсирање реченица. Ипак, и делимично препознавање може бити корисно за указивање на реченице које потенцијално садрже дефиниције.

успешно препознавање	
биљке	**Европски ариш** је брзорастућа врста квалитетног и трајног дрвета
	Орхидеје су вишегодишње зељасте биљке
	Голосеменице су дрвенасте биљке
животиње	**Пауколики зглавкари** су најчешће грабљиве животиње
	Термити су друштвени (социјални) инсекти
	Опнокрилци су већином друштвени инсекти
	Sequoiadendron giganteum је једна од најстаријих и највиших живих биљака
органи	**Корен** је карактеристичан вегетативни орган биљака
	Изданак је вегетативни орган биљке

Табела 4. Примери препознавања дефиниција из корпуса уџбеника биологије

[9] Претрага корпуса на адреси: http://noske.jerteh.rs/; приступ корпусу захтева аутентификацију и ауторизацију корисника.

[10] https://www.sketchengine.eu/

Дефиниције препознате у тексту су у неким случајевима сличне дифиницијама из РСАНУ, нпр., *орхидеја фамилија вишегодишњих зељастих украсних биљака*, док се понекад разликују, на пример, *бубашваба* је у РСАНУ дефинисана са *в. црни жохар*[11] док се у тексту корпуса јавља дефиниција ***Бубашвабе*** *су ноћне животиње, честе по људским насељима* (подвучени део није препознат). Локалне граматике описане у одељку 3 концентришу се на моделирање и препознавање лексикографских дефиниција. У РСАНУ дефиниције понекад могу садржати и додатна објашњења која више наликују енциклопедијском објашњењу, а њих је тешко обухватити локалним граматикама. Дефиниције у слободном тексту обично садрже и више слободно формулисаних описа и објашњења. Да би се могло обавити веродостојно поређење дефиниција из речника (конкретно РСАНУ) и доменског корпуса било би потребно допунити корпус разноврснијим изворима, на пример високошколским уџбеницима, научним радовима и сл.

Резултате добијене овим експериментом можемо да посматрамо као прелиминарне и обећавајуће. Да би се са сигурношћу могла утврдити њихова успешност потребна је, с једне стране, дорада локалних граматика, а с друге стране допуна електронских речника доменским терминима.

4.3. Предлог речничких чланака

Једна од могућности будућег лексикографског система јесте аутоматска провера фонда речи у дефиницијама, на пример, да ли одговара узрасту, да ли се у дефиниције користе општепознате речи, да ли садржи пригодну синтаксичку структуру и слично.

Структура речничког чланка зависи од врсте и намене речника. Ниже се доносе примери речничког чланка лексеме *дрво* из два описна речника – Речника за основце[12], који садржи лексику са којом се ученици сусрећу током основношколског образовања и једнотомног Речника српског језика Матице српске, у коме је лексема *дрво* презентована тако да доноси релевантне податке којима би требало да влада један средњошколац.[13]

---

[11] С обзиром на то да се у РСАНУ доносе не само књижевне речи већ и дијалектизми, покрајинске речи, речи различите употребне вредности итд., то се у поступку дефинисања прибегава и упућивање на другу реч (*в. то и то*) – на фонетски правилнију, односно уобичајенију варијанту или на реч истог значења која је правилнија, односно обичнија. Овим поступком се даје нормативна препорука и економише простором.

[12] У питању је Речник за основце Виолете Бабић у издању Креативног центра, који је у припреми за штампу.

[13] Најкомплекснији речнички чланци садржани су у РСАНУ. Они нуде обиље информација о фонетским, морфолошким, синтаксичким, семантичким, прагматичким,

У оквиру речничког чланка речника за основну школу на левој страни је одредница, која својом бојом показује врсту речи, а на десној значења са којима се ученик сусреће током основне школе. Чланку су придружени изрази и пословице, као и деривационо гнездо лексеме.

дрво

1. дрвенаста биљка чије је стабло причвршћено кореном за земљу, а при врху има гране и крошњу.
2. комад стабла који се употребљава за потпалу, огрев.
3. део стабла и грана који се користи за прављење намештаја, у занатству и грађевинарству.

□ ***божићно дрво*** јелка која се кити за Божић;
пуца дрво и камен изузетно је хладно;
седи дрво на дрво говорило се ђаку који покушава да каже или уради нешто, а није способан за то;
дрва у шуму носити радити непотребан посао;
не видети шуму од дрвећа в. шума (изр.);
дрво се савија док је младо човек се учи, навикава на нешто одмалена.

□ дрвар, дрвара, дрвен, дрвенарија, дрвенаст, дрвенасто, дрвено, дрвеће, дрвље, дрвни, дрвод(ј)еља, дрвод(ј)ељница, дрвод(ј)ељски, дрвојед, дрвокрадица, дрволик, дрвопрерађивач, дрворед, дрворез, дрворезац, дрвос(ј)еча, дрвостругар, дрвце, одрвенети.

Пример речничког чланка за исту одредницу *дрво*, али за средњу школу, садржи акцентовану одредницу, граматичке податке, скраћеницу за терминолошку област, систем дефиниција с примерима (у виду синтагме или реченице) и синтагматске и фразеолошке изразе, међу којима је и једна пословица.

дрво, дрвета им. с

1. (мн. дрвета, ген. дрвéтā, дат. инстр. и лок. дрветима; зб. им. дрвеће с) бот. *вишегодишња дрвенаста биљка чије је стабло причвршћено кореном за земљу, у доњем и средњем делу је голо, а у горњем делу се грана и ствара крошњу*: листопадно дрво, зимзелено дрво, Дрвеће још није озеленело.
2. (мн. дрва, ген. дрвā; зб. им. дрвље с) *исечени комад стабла или гране које се употребљава за огрев или као грађа*: огревно дрво, цепати дрва, обрада дрвета, трговац дрвима.
3. *најчвршћи део стабла и грана, тврда материја између сржи и лике која се употребљава у индустрији, занатству, грађевинарству и др.*: Кад се стабло посече на њему се виде срж, дрво, лика и кора.

когнитивним, комуникативним и другим карактеристикама српског језика, а „по уређености и устаљеном начину представљања подсећају на зоне системске лексикографије“ и представљају солидну основу за развој савремене српске лексикографије и лингвистике уопште (Ристић 2003).

□ ***божићно дрво*** младица јеле која се за Божић кити разним украсима, јелка; ***генеалошко (родословно, породично) дрво*** схема у облику дрвета којом се представља гранање неке породице, животињске класе и сл., породично стабло;

дрво живота рлг. дрво које је, према Библији, расло у рају и чији плодови дају вечни живот, бесмртност;

***дрво познања (сазнања, спознаје добра и зла** и сл.)* рлг. дрво које је, према Библији, расло у рају и чији су плод окусили Адам и Ева (иако им је то Бог био забранио) и због тога били истерани из раја, дрво чији плодови дају знање о добру и злу;

мускатово дрво бот. *зимзелено дрво* Myristica fragrans, чији се плод употребљава као зачин и за справљање мирисних есенција, орашчић;

човек од дрвета неосетљив, себичан човек;

дрва у шуму носити радити непотребан и узалудан посао;

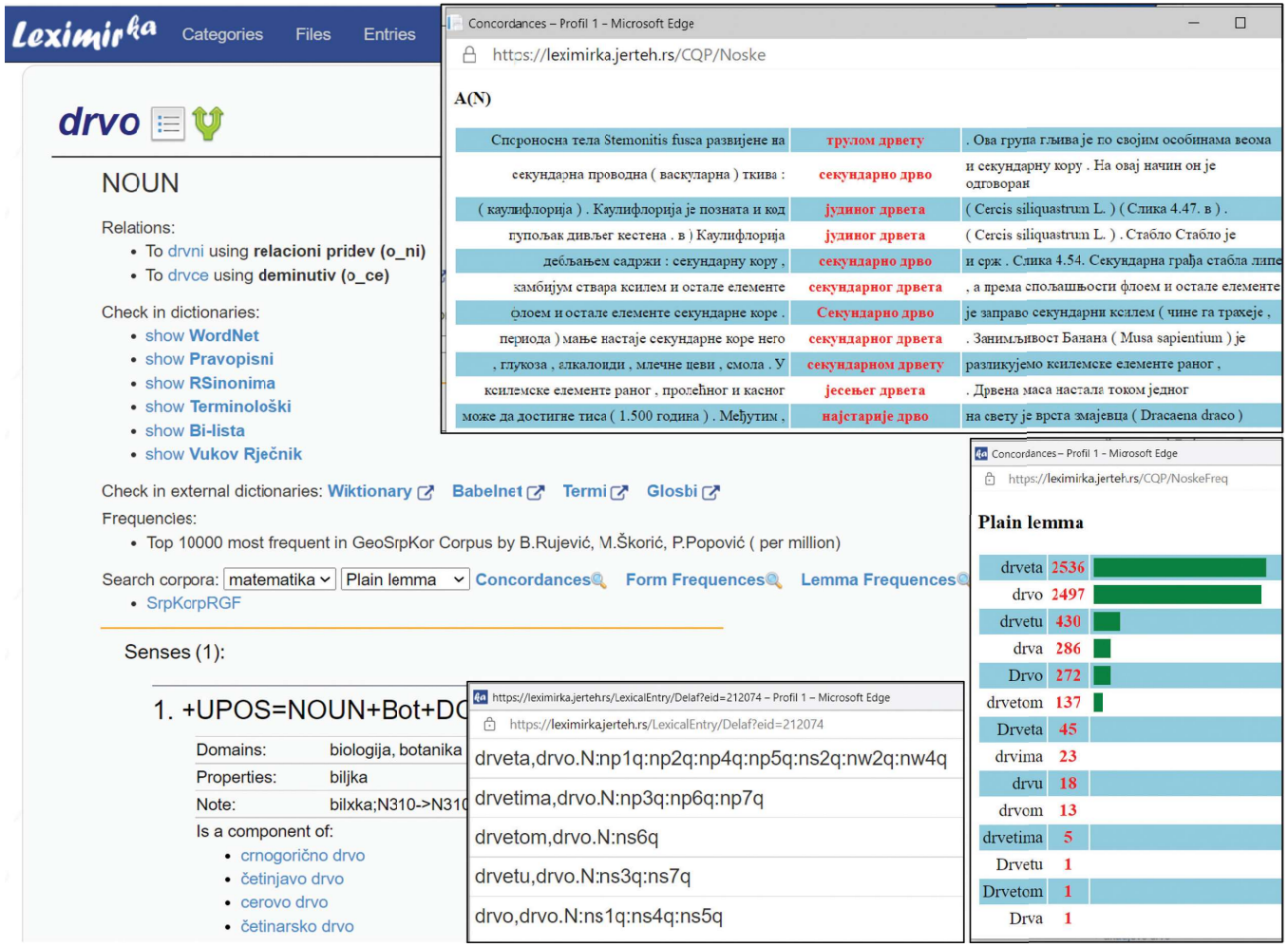
дрвље и камење бацати оштро напасти, нападати, критиковати;

дрво се савија док је младо одмалена се човек учи, навикава на нешто;

не видети шуму од дрвећа од споредних, небитних ствари не видети оно што је битно, важно;

пуца дрво и камен изузетно је хладно.

Како истиче Витас (2015), информатизовано лексикографско окружење у нашим условима морало би да испуни сложеније задатке од непосредног рада на производњи једног (или више) традиционалних речника отиснутих на папиру. Пре свега, такво окружење би могло и морало, с једне стране, да омогући задовољење различитих истраживачких потреба, а са друге, да обезбеди акумулирање истраживачких резултата. С тим циљем приступило се изградњи лексичке базе и лексикографске станице *Лексимирка*, у којој су похрањени електронски речници српског језика и други лексички ресурси (на пример, српски ворднет). На слици 5 приказана је одредница **дрво** из електронског речника, са неким од придружених информација које корисник по избору може видети; на слици су, на пример, наведени сви флективни облици ове одреднице са придруженим граматичким информацијама (кодираним према електронском речнику), као и повезане синтагме (*црногорично дрво*, *четинарско дрво*, итд.). Како је у *Лексимирки* имплементирана тесна веза са више корпуса српског језика, могуће је приказивање хистограма учесталости облика, као и различитих конкорданци (за задату лему или за синтаксне обрасце у којима је она једна од компоненти), што је такође приказано на слици). Са панела одреднице може се, у зависности од привилегија које су кориснику додељене, прићи другим речницима (нпр. српском ворднету) и ресурсима на интернету (нпр. Викиречнику) и погледати повезане речничке одреднице (алтернативни изговор, деривацијом изведени облици, дублети, и сл.).



Слика 5. Један панел лексикографске станице *Лексимирка*

5. Закључак

Даља истраживања ће бити усмерена на развој модула за систематизацију дефиниција и откривање оних дефиниције које одступају од уобичајених модела и за друге врсте речи. Дугорочни циљ остаје изградња система који ће из текста књига и уџбеника за задату реч издвојити реченице које одговарају некој од дефинисаних структура дефиниција. У овом тренутку је разматрана екстракција дефиниција само за појединачне речи, али се планира проширење и на полилексемске јединице. Прелиминарна евалуација на корпусу уџбеника биологије треба да се прошири на друге предмете и друге врсте текста.

Примена додатних техника (екстракције фичера, вектора речи, статистичке анализе, машинског и дубоког учења) верујемо да ће унапредити аутоматску екстракцију и омогућити класификацију примера и дефиниција према значењу (sense clustering).

6. Литература

Витас, Душко, Цветана Крстев. „Нацрт за информатизовани речник српског језика“. У: *Научни састанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене*, 44/3, Београд: Међународни славистички центар. Филолошки факултет, 2015, 105–116.

Гортан-Премк, Даринка. „О семантичком садржају лексикографске дефиниције“. У: Д. Ћупић (ур.), *Лексикографија и лексикологија*, Београд – Нови Сад, 1982, 49–51.

Гортан-Премк, Даринка. „Синонимски низ у лексикографској дефиницији“. У: *Научни састанак слависта у Вукове дане*, 12/1, 1983, 45–50.

Гортан-Премк, Даринка. „Дефинисање у српској лексикографији“. У: Р. Драгићевић (ур.), *Савремена српска лексикографија у теорији и пракси*, Београд: Филолошки факултет, 2014, 131–139.

Грицкат, Ирена. „Лексикографски поступак у речницима САНУ и ЈАЗУ – показан на глаголима са префиксом за-“. *Наш језик* XXV/1–2, 1981, 3–24.

Грицкат-Радуловић, Ирена. „Речник српске академије наука и уметности – почеци, лик, перспективе“. *Глас САНУ*, књ. 352, Одељење језика и књижевности, књ. 13, 1988, 25–44.

Новокмет, Слободан. *Називи животиња у српском језику – семантичка и лингвокултуролошка анализа*, Монографије 30. Београд: Институт за српски језик САНУ, 2020.

Ристић, Стана, Ивана Лазић Коњик, Ненад Ивановић. „Метајезик лексикографске дефиниције у дескриптивном речнику (на материјалу речника српског језика)“. *Лужнословенски филолог* 74/1, 2018, 81–96.

Ристић, Стана, Ивана Лазић Коњик. *Когнитивни правац у српској етнолингвистици – почеци развоја и актуелни проблеми*, Монографије 29. Београд: Институт за српски језик САНУ, 2020.

Ристић, Стана. „Лексикографски метајезик и српска дескриптивна лексикографија“. У: *Научни састанак слависта у Вукове дане* 31/1, Београд: Међународни славистички центар. Филолошки факултет, 2003, 119–130.

Ристић, Стана. *Раслојеност лексике српског језика и лексичка норма*, Монографије 3. Београд: Институт за српски језик САНУ, 2006.

Ристић, Стана. *Модификација значења и лексички модификатори у српском језику*, Монографије 10. Београд: Институт за српски језик САНУ, 2009.

Ристић, Стана. *Граматички и когнитивни аспекти лексичког значења*, Монографије 22. Београд: Институт за српски језик САНУ, 2015.

SRPS ISO 1087-1:2003, Терминолошки рад – Вокабулар – Део 1: Теорија и примена, Институт за стандардизацију Србије, 2003.

Стијовић, Рада, Ранка Станковић. „Дигитално издање *Речника САНУ*: формални опис микроструктуре *Речника САНУ*“. У: *Научни састанак слависта у Вукове дане* 47/1, Београд: Међународни славистички центар. Филолошки факултет, 2018, 427–440.

РСАНУ: *Речник српскохрватског књижевног и народног језика* САНУ, I–XXI (1959–2019), Београд: Институт за српски језик САНУ и САНУ.

Упутства за израду *Речника САНУ* 1959 [2017]. Београд: Институт за српски језик САНУ (рукопис).

Фекете, Егон. „О *Речнику српскохрватског књижевног и народног језика САНУ*“. У: *Сто година лексикографског рада у САНУ*, Београд 1993, 21–49.

*

Barnbrook Geoff. *Defining Language, A local grammar of definition sentences. Studies in Corpus Linguistics*, John Benjamins Publishing Company, 2002.

Evert, Stefan. *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Phd Thesis, 75–91, 2005.

Gambette, Philippe, Jean Véronis. „Visualising a Text with a Tree Cloud“. In: Locarek-Junge H. and Weihs C., editors, *Classification as a Tool of Research*, Proc. of IFCS'09 (11th Conference of the International Federation of Classification Societies) (supplementary material), 2009, https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643.

Jin, Yiping *et al.* „Mining scientific terms and their definitions: A study of the ACL anthology“. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

Navigli, Roberto, Paola Velardi. „Learning Word-Class Lattices for Definition and Hypernym Extraction“. In: *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, 2010, 1318–1327.

Kilgarriff, Adam, Pavel Rychlý. „Semi-Automatic Dictionary Drafting“, In: *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Uganda: Menha Publishers Ltd., 2010, 299–312.

Krstev, Cvetana. *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade, Belgrade, 2008.

Krstev, Cvetana, Duško Vitas, Ranka Stanković. „A Lexical Approach to Acronyms and their Definitions“. In: *Proceedings of 7th Language & Technology Conference*, November 27–29, 2015, Poznań, Poland, eds. Zygmunt Vetulani & Joseph Mariani, 219–223, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, 2015.

Pollak, Senja, Anže Vavpetic, Janez Kranjc, Nada Lavrac, Špela Vintar. „NLP workflow for on-line definition extraction from English and Slovene text corpora“. In: *Proceedings of KONVENS* 2012, Vienna, September 19, 2012, 53–60.

Rundell, Michael. „Macmillan English Dictionary: The End of Print?“ In: *Slovenščina 2.0: empirical, applied and interdisciplinary research 2.2*, 2014, 1–14.

Rychly, Pavel. „Manatee/bonito – a modular corpus manager“. In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 2007, 65–70.

Spala, Sasha, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, Carl Dockhorn. „DEFT: A corpus for definition extraction in free-and semi-structured text“. In: *Proceedings of the 13th Linguistic Annotation Workshop*, 2019, 124–131.

Stanković, Ranka, Rada Stijović, Duško Vitas, Cvetana Krstev, Olga Sabo. „The Dictionary of the Serbian Academy: from the Text to the Lexical Database“. In: *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana: Ljubljana University Press, Faculty of Arts. Čibej, Jaka, Gorjanc, Vojko, Kosem, Iztok & Krek, Simon (eds.), 2018, 941–949.

Stanković, Ranka, Branislava Šandrih, Rada Stijović, Cvetana Krstev, Duško Vitas, Aleksandra Marković. „SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian“. In: *Electronic lexicography in the 21st century*. Proceedings of the eLex 2019 conference, 1–3 October 2019, Sintra, Portugal, eds. I. Kosem et als., 2019, 248–269.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, Mihailo Škorić. Machine „Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian“. In: *Proceedings of the 12th International Conference on Language Resources and Evaluation*, LREC eds. Nicoletta Calzolari *et al.*, 2020, 3947–3955, 2020.

Pouran Ben Veyseh, Amir, Franck Dernoncourt, Dejing Dou, Thien Huu Nguyen. „A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency“. In: *AAAI*, 2020, 9098–9105.

Vitas, Duško, Cvetana Krstev. „Processing of Corpora of Serbian Using Electronic Dictionaries“. In: *Prace Filologiczne*, vol. LXIII, Warszawa, 2012, 279–292.

Уџбеници:

Берић, Тања, Гордана Субаков-Симић, Пеђа Јанаћковић. *Уџбеник биологије за ӣрви разред гимназије*, Београд: Нови Логос, 2016.

Кризманић, Имре, Зорица Лазић, Албина Холод. *Уџбеник биологије за седми разред основне школе*, Београд: Нови Логос, 2019

Петров, Бригита, Милош Калезић, Радомир Коњевић. *Биологија за II разред гимназије оӣшӣег смера*, Београд: Завод за уџбенике, 2017.

Стилиновић, Слободан, Иван Јаковљевић. *Шумске кулӣуре и ӣланӣаже за III и IV разред шумарске школе*, Београд: Завод за уџбенике, 2003.

Цвијић, Гордана, Јелена Ђорђевић, Надежда Недељковић. *Биологија за III разред гимназије општег смера*, Београд: Завод за уџбенике, 2019.
Цветковић, Драгана, Дмитар Лакушић, Гордана Матић, Александра Кораћ, Слободан Јовановић. *Биологија за IV разред гимназије општег смера*, Београд: Завод за уџбенике, 2019.

Rada R. Stijović, Cvetana J. Krstev, Ranka M. Stanković

AUTOMATIC EXTRACTION OF DEFINITIONS – A CONTRIBUTION TO ACCELERATING THE PROCESS OF COMPILING DICTIONARIES

S u m m a r y

The paper presents the results of preliminary tests of automatic extraction of candidates for dictionary definitions from unstructured texts in the Serbian language, with the aim of accelerating the production of the dictionary. Using the existing dictionaries of the Serbian language, different ways of modelling and extracting definitions were analyzed. The analysis showed that a large number of definitions have a structure that can be modelled, as evidenced by the statistics of definitions grouped by type, example models and evaluations in Biology textbooks.

Keywords: descriptive dictionaries, meta-analysis of lexicographic definitions, automatic definition extraction, electronic dictionaries, Serbian language.