Речник САНУ као база терминолошких речника (на примеру речника кулинарства)

Рада Стијовић, Олга Сабо, Ранка Станковић

Рада Стијовић, Олга Сабо, Ранка Станковић

РЕЧНИК САНУ КАО БАЗА ТЕРМИНОЛОШКИХ РЕЧНИКА (НА ПРИМЕРУ РЕЧНИКА КУЛИНАРСТВА)

Циљ овога рада јесте да покаже како богата и разноврсна лексичка грађа Речника САНУ може послужити за израду разноврсних терминолошких речника (овде на примеру речника кулинарства) и како се процесом дигитализације Речник може обогатити, осавременити и како се рад на њему може убрзати.

Кључне речи: Речник САНУ, терминолошки речници, кулинарска лексика, дигитализација

Уводне напомене

Речник српскохрватског књижевног и народног језика треба да садржи, када буде завршен[1], целокупну лексику[2] савременог српског језика[3], како књижевног језика тако и српских народних говора[4]. Све одредничке речи илустроване су примерима из писаних извора и са најширег српског дијалекатског подручја, презентованим тако да показују временску и просторну заступљеност речи[5]. Овако богат лексички материјал омогућује разноврсна лингвистичка и ванлингвистичка истраживања. Могуће је на основу њега сагледати материјалну и духовну култура српског народа, његов начин живота, начине размишљања, етичка начела, обичаје[6], занимања, контакте и међусобне односе са другим народима[7]. Могу се изводити разноврсна статистичка истраживања[8], формирати синонимски низови итд. [9]

[1] До сада је изашло 19 томова, 20. је у припреми за штампу.

[2] Ова констатација је условна, јер по концепцији речника страна и ускоспецијализована лексика се уноси у оној мери у којој је она позната једном образованом средњошколцу.

[3] Мисли се на књижевни језик вуковског типа, дакле од Вука, па и Доситеја до данас.

[4] Примери књижевног језика ексцерпирани су из дела свих области живота и стваралаштва, а дијалекатска грађа је прикупљана на широком простору српских говора – од Книна, Баније, Лике, преко Босне и Херцеговине, Славоније, Барање, до Војводине, Ресаве, Призрена, Тимока, Врања, Црне Горе, а укључени су и српски говори у Македонији.

[5] Свако значење речи је илустровано примерима из писаног извора или са терена, и то тако да се настоји показати да ли реч живи од најстаријег времена које је обухваћено овим речником до данас, као и на ком терену је потврђена. Број примера зависи од учесталости употребе речи, али их је највише шест.

[6] У Речнику, нпр., налазимо податак да се у неким крајевима Босанске крајине, као и међу Србима у Хрватској крајини орачима кад заору прву бразду износи погача која се зове *брaзданица, браздењача* или *браздионица*, у зависности од краја, да се у Ресави за Божић прави колач који се зове *виноград* због украса који се стављају на њега а који симболизују чокоте винове лозе, сазнајемо и шта је *василица*, када се праве *младенчићи,* а када чорба *киселица* итд.). Бројне су фразеолошке јединице које садрже ову лексику, а које, такође, бележи Речник САНУ (прост као пасуљ, прогутати кнедлу, бити у свакој чорби мирођија, солити памет). Формирње синонимских низова – дулечара, бундевара, тиквењак, дулечњек..., па качамак, палента, пура, мамаљуга, макарон, жгањци.

[7] У Речнику се налазе и романизам *пашада* као ознака прибора за јело и германизмом *есцајг*, али потврђено у различитим говорима. Осим тога, наћи ће се турска чорба, немачка супа и словенска јуха.

База Речника може послужити и као грађа за израду бројних других речника – књижевног језика, дијалеката, терминолошких речника итд. Ми смо за предмет овога рада узели Речник као извор за терминолошке речнике. Како он садржи терминолошку лексику из 78 области, определили смо се за кулинарску, која је у нашој свакодневици несумњиво најзаступљенија и могло се очекивати и да ње има највише у Речнику и да је она у великој мери обухваћена Речником. Претраживати стотине хиљада одредница и милионе примера у папирној верзији Речника тешко је изводљиво. Зато смо желели да покажемо како дигитална форма овог дела тај поступак чини лако остваривим.

Да би се дошло до релевантних података, било је потребно наћи метод екстраховања ове лексике, с обзиром на то да она (за разлику од већине других терминолошких области) није системски обележавана одговарајућим квалификатором (*кув., кул.*), нарочито када је лексема део општег лексичког фонда (*зготовити, месити*). Помоћ су нам пружиле обавезне кључне речи у дефиницијама којима се описује значење и употреба неке лексеме. При дефинисању неке зачинске биљке, поред осталог увек стоји „која се употребљава као зачин“, „која служи као зачин“ и сл. (в. *мајоран, мирођија*). Посуђе или прибор за јело дефинишу се са „део стоног прибора“ (*кашика, пирун, ложица*), „суд/посуда за (кување, пржење и др.)“ (*лонац, тигањ*). Послужили су нам и кувари у електронском облику, пре свега Катарине Поповић-Мицине, који се налази у изворима за Речник, као и морфолошки и електронски речници и граматике развијене у оквиру Групе за језичке технологије Универзитета у Београду. Они су нам омогућили да дођемо не само до примера који су уједно и одреднице, већ и до лексема које су садржане у целокупном речничком корпусу, а које по концепцији речника не могу стајати као одредничке речи (бројни двочлани и вишечлани термини) или још нису обрађене у досадашњим томовима.

Критеријуми за израду терминолошког речника

У уводном делу рада желели смо да прикажемо структуру Речника САНУ, његово лексичко обиље, велики потенцијал који пружа као могућа терминолошка база и огроман значај процеса дигитализације као једног од битних фактора који омогућава обогаћивање и осавремењивање постојећег Речника, и убрзавање процеса његове израде.

Да бисмо искористили Речник САНУ као терминолошку базу било је потребно успоставити одређене критеријуме који би одредили поступак избора речи, одређивање модела и примену одговарајућих процедура за обраду текста. О самом процесу припреме и обраде корпуса и могућим допунама Речника САНУ биће више речи у завршном делу рада.

Први корак је био издвајање, из свих деветнаест томова Речника САНУ, речи означених са кув./кул.. Затим је извршена анализа структуре одредница и дефиниција

[8] Нпр., о заступљености или учесталости неке лексике у одређеном времену или код одређеног писца.

[9] Такви су примери који се налазе у Речнику: бундевара, дулечара, дулечњак, тиквењак или: качамак, палента, пура, мамаљуга, макарон, жгањци.

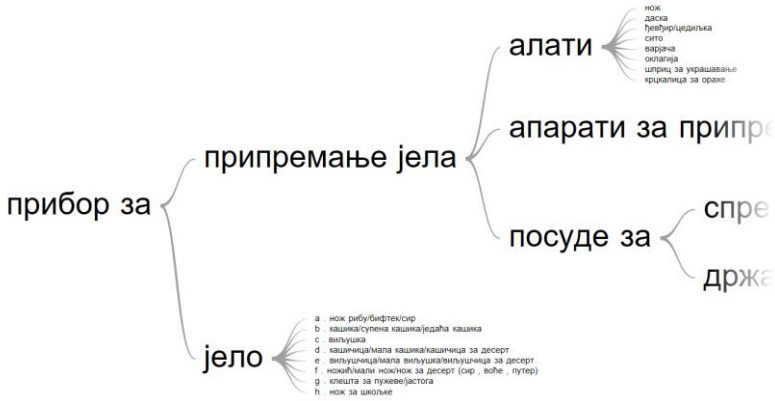
са циљем да се утврде одговарајући модели. Да бисмо могли прећи на наредне кораке било је потребно пронаћи неке одговоре и установити неке критеријуме:

1. Које се стране речи и термини уносе у РСАНУ?

Ово питање односно критеријум на основу кога се уносе речи је веома битан јер нас упућује на степен сложености термина из појединих области. Када говоримо о страним речима и терминима из одређених области, који се уносе у Речнику САНУ подразумева се ниво знања опште културе једног образованог човека. Полазећи од тога спонтано се наметнуло питање у којој мери се променио ниво знања током деценија, и колико је општа култура порасла што се свакако морало одразити на усвајање већег броја страних речи или термина. Данас, речи као што су *блендирати, брускети*, *тостирати, шејк* (неке ће се појавити у наредним томовима, док би другима било место у почетним) сигурно представљају део лексике једног образованог човека док неке нису ни постојале у време настанка првих томова. И, што је још значајније, током процеса дигитализације ови термини би се лако могли унети и тиме допринети осавремењивању лексичког фонда.

2. Које ћемо речи унети у Терминолошки речник кулинарства?

Да бисмо могли да одредимо које речи би обухватао ТЕРМИНОЛОШКИ РЕЧНИК КУЛИНАРСТВА било је потребно установити како би изгледала структура једног РЕЧНИКА КУЛИНАРСТВА, који има много шири домен у односу на терминолошки речник, и на основу тога издвојити оне области које би ушле у терминолошки речник.



Слика 1. Илустрација једног сегмента структуре Речника кулинарства

3. Шта је кулинарски термин, а шта реч која припада кулинарској лексици?

Најтежи део у овом послу је, свакако, одредити шта је кулинарски термин, а шта реч која припада кулинарској лексици, или општем лексичком фонду.

Шта урадити са речима као што су вода, тањир: плитки тањир, нож: нож за рибу, кашика у значењу мере, или прибора за јело, виљушка: виљушка на трактору и сл.?

За потребе овога рада ми смо се определили за кулинарску лексику да бисмо боље представили поступак проналажења и обележавања речи.

4. Да ли унети, и на који начин, квалификатор *кув.* код речи које нису обележене?

Паралелно са процесом одређивања однос. проналажења речи у РСАНУ, појавило се и питање обележавања речи однос. додељивање квалификатора *кув.*, током процеса дигитализације. Ово је веома значајно питање, јер би уношењем квалификатора који се односи на неку терминолошку област, речничка терминолошка база однс. корпус био знатно проширен.

5. Терминолошка мултидисциплинарност: веће могућности за истраживање

Приликом дефинисања области које би биле део једног Речника кулинарства имали смо у виду и упитнике које се користе за дијалекатска испитивања. Ово је пример како би се две истраживачке области (дијалекатска истраживања омогућавају прикупљања обредне и обичајне лексике која је у Речнику САНУ обично обележена квалификатором *етн*.) могле повезати што би се као принцип могло применити и на остале терминолошке области..

УПИТНИК ЗА КУЛИНАРСКУ ЛЕКСИКУ

...

Припремање хране

...

26. Како се зове кад се јело кува у води (*барити, кувати, варити*)

_____

27 Како се зове кад се храна прелије врелом водом (*попарити*)

_____

28. Како се зове кад се храна стави на ватру да се кува (*приставити*)

_____

29. Како се зове кад се заврши кување, припремање хране (*зготовити*)

_____

30. Како се зове кад се храна припрема на врелој масноћи (*пржити, пригати, цварити*)

_____

Примери питања из упитника за дијалекатска истраживања

ораница ж **… 2.** етн. покр. *колач са избразданом горњом кором који се меси уочи Божића и ломи на Бадње вече.* — На Бадњак ујутро жене испеку посебне хљебове и колаче, који се зову посебним именима као Зорњача, Даница, Ораница (Влаш. 1, 35).

василица ж *колач који се по обичају меси на Мали Божић и на Нову годину, вар.васиљица, васуљица;* исп. чесница.

Примери обредне лексиле у Речнику САНУ

бразданица ж *погача за ораче кад заору прву бразду*, вар. бразденица.

окопаница ж етн. покр. *пита коју домаћица изнесе пред раднике кад се заврши копање, окопавање кукуруза*.

Примери обичајне лексике у Речнику САНУ

Пре него што пређемо на обраду корпуса потребно је поменути који су били наши циљеви.

Желели смо да утврдимо колико је речи обележено квалификатором *кув./кул.* Да покушамо да одредимо **критеријум** на основу кога можемо разграничити лексику која се користи у кулинарству од кулинарске терминологије, изаберемо **метод екстраховања ове лексике**, с обзиром на то да речи из кулинарства (за разлику од већине других терминолошких области) нису системски обележаване одговарајућим квалификаторима (нарочито када је лексема део општег лексичког фонда па се подударају) терминолошко и опште значење. **применимо семантичке маркере** који ће омогућити обележавање и проналажење ових речи, и у процесу дигитализације Речника видљиво обележити ову лексику да постане лако претражива, и на тај начин допринети да Речник буде редовно допуњаван новом грађом и на тај начин стално осавремењиван.

палента ж (ген. мн. паленти) (тал. polenta) кув. *јело од укуваног кукурузног брашна, качамак, пура.*

ћувеч м (тур. güveç, земљани суд) 1. а. *дубљи суд за печење (обично земљани, овалног облика).* — Влада мишљење, да су најбољи судови — ћувечи грнчарски (Влад. Т. 1, 73)... б. покр. *велики плићи земљани суд за разливање млека.* — Разладила сам три ћувеча млеко (КМ, Ел. Г. 3). (Приштина, КЕМ).

2. *јело од разног поврћа и пиринча (обично с месом).* — Кад месо буде кувано ... а кромпири постану меки, — онда је ћувеч готов (Влад. Т. 1, 73). Ћувеч се највише готови у јесен од дебелог овчег меса са патлиџанима и пиринчем (Ник. В. 2, 109).

Примери обележених и необележених речи из Речника САНУ

Процес припреме и обраде корпуса

Процес екстракције и лематизације термина отпочео је прикупљањем више кувара у пдф облику, после чега је уследило оптичко препознавање карактера, аутоматско пресловљавање и корекција хифенације, а потом и обрада електронским речницима за српски језик (Крстев и други, 2003). Треба напоменути да се електронски речници разликују од машински читљивих речника, јер су електронски речници намењени рачунарској обради текста (софтверу), за разлику од машински читљивог речника који је намењен човеку као кориснику. Човек као корисник углавном у

речнику тражи значење неке речи или њен превод, док се електронски речници користе у истраживањима језика и креирању језичких алата. Морфолошке речнике српског језика развили су проф. др Цветана Крстев и проф. др Душко Витас уз помоћ Групе за језичке технологије Универзитета у Београду.

Анализа обрађеног корпуса обухватила је екстракцију речи и фраза засновану на доменским и семантичким категоријама, потом анализу препознатих и непрепознатих речи са делимичним поређењем електронских речника са Речником САНУ. Поређење је делимично зато што је истраживање базирано на доступном подскупу Речника САНУ (чија је дигитализација у току), а истраживање се такође ограничава само на лексику текстова кулинарског домена.

Имајући у виду да је корпус текстова произведен аутоматски из пдф датотека, у тексту је велики број речи био преломљен (растављен), те се јавила потреба за аутоматском корекцијом хифенације. За то су искоришћене каскаде графова које консултују електронске речнике, тако да је аутоматски кориговано 350 речи. Ево неколико примера речи код којих је успешно уклоњена хифенација: *конзумира-ња, уп-рошћавате, миро-ђију, мли-ну, бун-девиним, кори-јандера, индиј-ског, пра-зилук*... Систем је пројектован тако да оставља цртицу тамо где је она саставни део сложене фразе, тако да су следећи примери остали непромењени: *пола-пола, медено-слатку, карамел-бомбона, златно-смеђу, дан-два, зелено-љубичасте*… Анализом резултата нису пронађени примери погрешне корекције хифенације. Одређени број прелома није могуће аутоматски разрешити и у том случају је систем оставио текст без промена. На пример у речи: *кафе-на*, *кафе* и *на* могу бити засебне речи, те тако систем није вршио замену са *кафена*. Слично, није замењено ни: *за-чин, маши-ни, неко-лико, на-прави, посоли-ти, нај-мање, не-природан, рен-де, посоли-ти, су-сам* ... Креирани корпус се састоји од 260.000 речи, односно 1.360.000 карактера. Овде је неопходна ручна корекција текста.

Лингвистичка обрада је урађена применом електронских речника, након чега је креиране „врећа речи“ са фреквенцијама појављивања и рачунањем релативних фреквенција „на милион речи“. За поређење релативне фреквенције кулинарских термина са референтним корпусом коришћен је Корпус савременог српског језика (korpus.matf.bg.ac.rs) од двадесет два милиона речи, (Утвић, 2011). Кључност термина (енг. keyness) се рачуна као однос релативне фреквенције (на милион) у кулинарском корпусу и у општем корпусу.

Коришћењем електронских речника у креираном кулинарском корпусу препознате су семантичке категорије карактеристичне за лексику кулинарских текстова (Крстев, Лазић, 2015):

Табела: Примери из корпуса који су препознати семантичким маркерима за текстове кулинарског домена

Маркер	Опис маркера	Примери из корпуса
+DOM=Culinary	кулинарски домен	павлака, путер, побиберити, алева паприка, динстати, пропржити, нарендати плех, надев, сос, пармезан, сенф тиквица, беланац, презла целер, рендан, рерна, нарендан,

+Food	општи маркер за храну	бибер, торта, глазура, шећер, надев, уље
+Alim	храна која се конзумира скоро природна	бадем, целер, босиљак, першун, рузмарин
+Drink	пиће	јогурт, прошек, млаћеница, црно вино, воћни сок, пиво
+Prod	производ	брашно, ванилин шећер, марципан, џем, чоколада
+Course	готово јело	савијача, куглоф, пире, десерт, качамак, сарма
+Ing	додатак	со, лимунтус, конзерванс, каротен, заслађивач
+Meal	оброк	предјело, роштиљ, закуска, ручак, пикник, чај, гозба
+MesApp	приближна мера	прстохват, кесица, комадић, штанглица, чаша
+Cont	посуде	кашичица, кесица, шоља, тањир, кутија, шака, кеса, пакет
+Por	порција	комадић, колутић, штанглица, режањ, кап, врх, крак
+Uten	справа	калуп, плех, кашика, посуда, модлица, шерпа
+WoP	начин припреме јела	кухање, желирање, пропржити, развљати
+Part	део	јуфкица, ребро, зрно, средина, лист
+Wh	целина	коцка, штапић, ролна, векна
+Set	скуп	веза, прстохват

Могућности допуне и осавремењивања Речника САНУ

Процењује се да је у Речнику САНУ квалификатором *кув.* означено око сто речничких одредница из куварског домена. У првом тому су означен: *аспик, бајцовати и баница*, у петом*: завара, заваривати, заварити и завезача*, у десетом: *котлет, крокет, кромпирача, крофна, крумпир-пирјан*, у седамнаестом: *одморити*, док у деветнаестом тому има четрдесет једна одредница са *кув*. и једна (пендевиш) са *кул*.

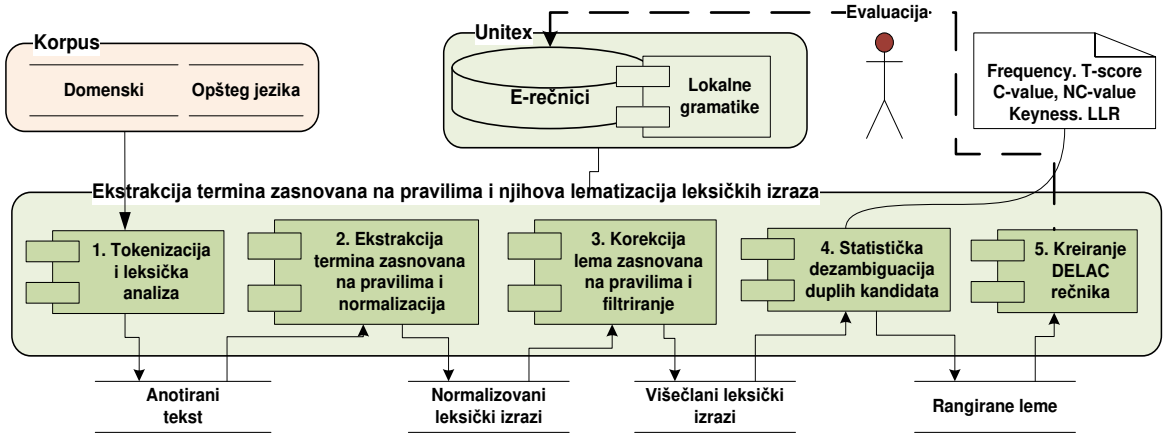
Процењује се да се још око 630 одредница из Речника САНУ може на основу електронског речника означити да су из куварског домена, на пример: *вечера, месо, млеко, кафа, вино, воће, јело, јаје, намирница, брашно, месни, колач, жито, алкохол*, *желе, маст*. Осим означавања домена могу се увести и нови квалификатори којима се могу и прецизније обележити речи. На пример, ако се уведе квалификатор за категорију *оброк* било би обухваћено више од десет одредница (*вечера, оброк, доручак, гозба, банкет, закуска, мезе, зајутрак* итд.). Или, квалификатором за *готово јело* било би обухваћено више од сто речи (*жито, кобасица, бурек, капама, гулаш, кајгана, качамак, вешалица*), за *пића* око седамдесет (*млеко, кафа, вино, лоза, виски, напитак, јогурт, вотка, ликер*), за категорију *начин припреме* преко тридесет (*кувати, барити, конзервисати, одмрзнути, отклопити, маринирати*).

Постоје такође речи које постоје у електронским речницима које су из домена кулинарства, а које се не појављују у Речнику САНУ и процењује се да је око триста таквих одредница, на пример:

- авокадо, ајдамер, ајсберг, амарето, армањак, ароматизирати, аронија, афусал;
- багет, балзамико, бананица, бешамел, бланширање, бланширати, блендирати, бренди, бри, броколи, брускета, бурбон, бургер, бурито;
- ванилица, вафл, вегета, винобран, вонгола;
- гаспаћо, гаспачо, гауда, гвакамола, гирос, граделе, гранцла, гратинирати, грејпфрут, грилијаж, грилијаш, гриловати, грицкалице;
- дебрецина, дехидратор, динстати, добиберити, догрејати, дресинг.

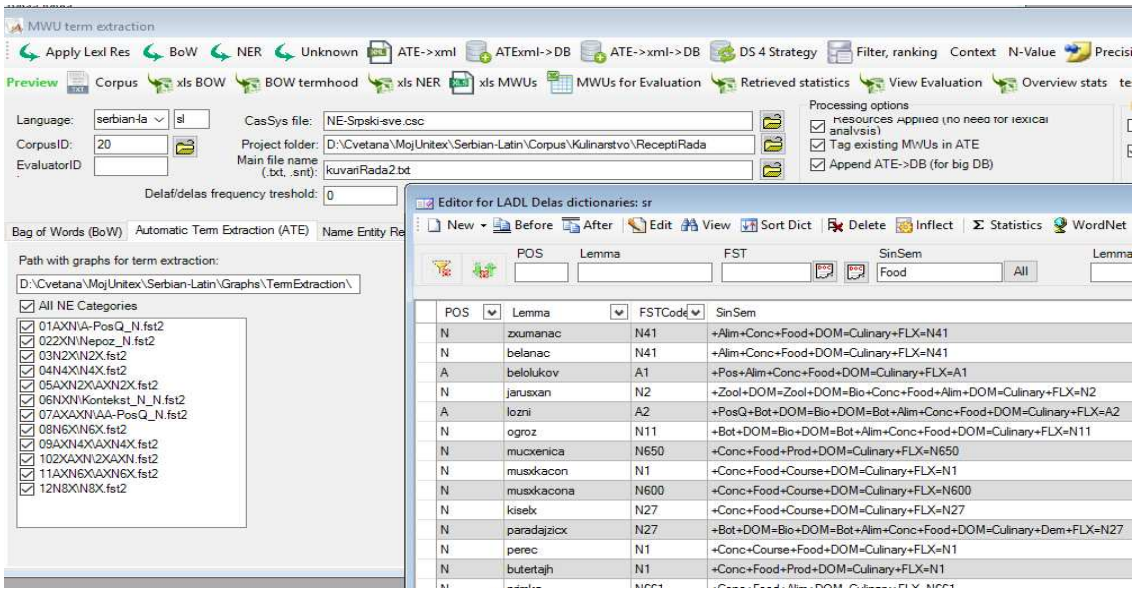
Екстракција термина из текста

За решавање проблема екстракције и лематизације вишечланих термина из текста, одабран је приступ заснован на правилима која се ослањају на систем језичких ресурса, као што су морфолошки и електронски речници, и граматике развијене у оквиру Групе за језичке технологије Универзитета у Београду. Аутоматска процедура за екстракцију вишечланих речи (синтагми и фраза) као кандидата за термине и њихова лематизацију остварена је коришћењем система Unitex за обраду корпуса и LeXimir-a, вишенаменског софтверског алата за управљање лексичким ресурсима. Целокупан процес је аутоматизован почевши од токенизације и лексичке анализе текста до производње одредница за речнике, а потребна је само незнатна ручна интервенција (слика 2), (Станковић и други, 2016).



Слика 2. Процес аутоматске екстракција и лематизације термина

Систем се ослања на карактеристике Unitex-a за анализу текста и израду коначних трансдуктора, као и на бројне функционалности LeXimir-a у производњи лема у стандардним форматима (Станковић и други, 2011). Сви резултати и припадајући метаподаци похрањени су у бази података на SQL серверу. Скуп лема произведених у наведеном поступку филтрирања се даље процесира увођењем мера које комбинују језичке и статистичке податке. Наиме, свакој леми, осим фреквенције, придружују се и основне мере (C -Value T-Score, LLR, Keyness) (Frantzi, 2000; Dunning, 1993; Kilgarriff, 2014.) тако да се за одабрану меру и одговарајући праг, генерише скуп лема за евалуацију. На слици 3 је приказан панел за примену електронских речника и локалних граматика, екстракцију према предефинисаним синтаксичким обрасцима, лематизацију и управљање евалуацијом.



Слика 3. Радно окружење за екстракцију термина

Описаћемо неке од синтаксичких образаца за екстракцију:

AXN – вишечлана реч се састоји од придева и именице и придев и именица се слажу у погледу све четири граматичке категорије (род, број, падеж, аниматност); пример свињско месо, кукурузно брашно, алуминијумска фолија, индијски орах;

2XN – испред именице се налази реч која не подлеже флексији; то је најчешће префикс или прилог изведен од придева, а сепаратор је најчешће цртица; пример пек-папир, сардел-паста;

N2X – иза именице се налази реч која не подлеже флексији; то је најчешће именица у генитиву или инструменталу; пример зрно бибера, корен целера, главица купуса, лист ловора, млевење меса.

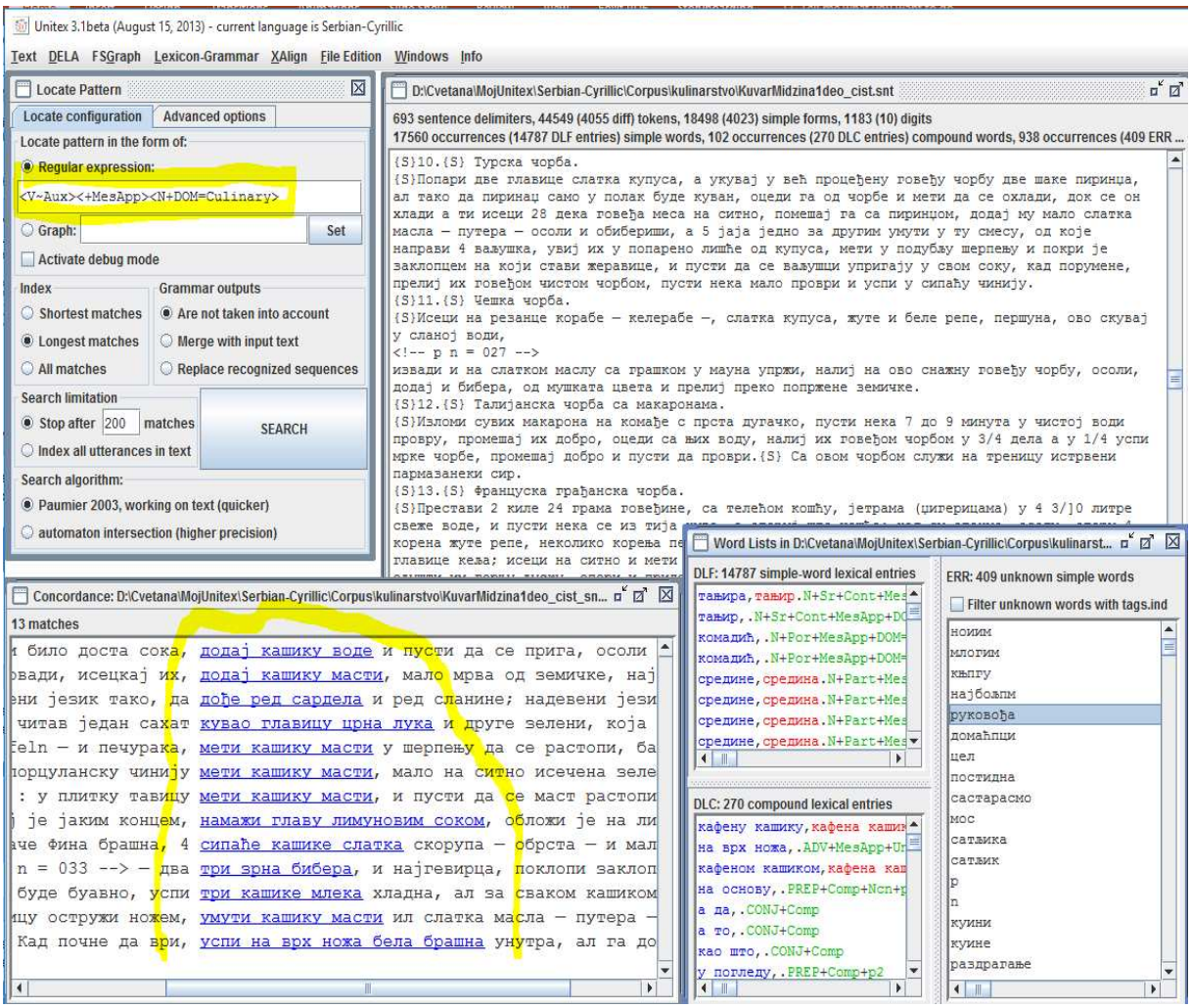
Екстракција полилексичких јединица је урађена из текста Великог српског кувара Катарине Поповић-Миџине, у ком се могу препознати специфичности лексике у односу на савремене куваре. Тако су препознате следеће именице: *шерпења, повлака, умокац, исек, милерам, кастрола, дека, говеђина, комаће, масло, треница, заклопац, питија, лукац, аспик, пржолица, капар, сахат, аван, смрчак* итд. Такође, појављују се и карактеристични глаголи: *подавати, поруменити, трти, поспати, поруменети, исећи, размутити, попржити, упирјанити, пирјати, метути, окиселити, мести, калаисати, динстовати, подати, усхтети, зготовити, наденути, опенити, откапљати, надева* и др. и придеви: *сипаћи, топал, подубок, утучен, мушкатов, истрвен, размућен, добивен, истучен, тучан, зготовљен, бојадисан* итд.

Постоји одређени скуп речи које електронски речници не препознају, а постоје у Речнику САНУ, као на пример: беланцад, бишкота, буаван, додавајући, гркљанчићи, кожура, куина, лаворика, обиберисати и др., тако да се унапређење ресурса може посматрати у два смера.

Претраживање по обрасцима

Један од водећих истраживача у области обраде природних језика, Дан Цурафски, у књизи „Језик хране“ (Jurafsky, 2014), приказује еволуцију обичаја, назива, рецепата који прате говор о храни. Ту се може видети лексичка и концептуална сложеност језика хране. Покушали смо да извучемо карактеристичне изразе који претходе, или следе, тражене термине, или се налазе у дефиницији. У раду је представљено неколико карактеристичних образаца.

Претраживање кулинарског корпуса, осим по облицима и лемама, могуће је коришћењем колмплекснијих упита, представљених регуларним изразима или графовима. Образац облика ***<V~Aux><+MessApp><N+DOM=Culinary>*** проналази глагол (који није помоћни), за којим следи реч која означава кулинарску меру и потом именица из кулинарског домена. На слици је приказано радно окружење за задавање обрасца упита и конкорданце добијене као одзив за постављени упит: *додај кашику воде, мети кашику масти, намажи главу лимуновим соком и сл.,* али препознаје и *три зрна бибера,* јер је *три* облик глагола *трти* (императив другог лица једнине).



Слика 4. Образац за претрагу и одговарајуће конкорданце у окружењу Unitex

Терминологију је могуће пронаћи (и обележити) и на основу карактеристичних фраза у дефиницијама. За зачинске биљке ће увек стајати: *„која се употребљава као*

зачин“, или *„која се због ароматичног мириса и укуса употребљава“, „лековита биљка која се“.* Креиран је регуларни израз који почиње са било којим обликом односне заменице ***који, -а, -е,*** потом следи више речи, а на крају је ***као зачин***: ***<који><МОТ>*<као><зачин>*** на основу ког се добија 28 конкорданци, од којих су неке:

мајоран	„**која** расте по Средоземљу а употребљава се **као зачин** и у медицини“
мирођија	„**која** се због ароматичног мириса и укуса употребљава **као зачин“**
ким	„**која** се и код нас гаји и употребљава **као зачин** и у лекарству“
забела	„**која** се додаје **као зачин** за чорбу“
зелен	„**које** се употребљавају **као зачин“**
орашчић	„**који** служи **као зачин** и у апотекарству“
зачинаст	„**који** је **као зачин**, ароматичан“.

За посуђе и прибор за се често у дефиницији користи „део стоног прибора“, „прибор за јело“ или „суд/посуда за“ (очекује се: кување, пржење, печење, разливање, цеђење) тако да је могуће креирати различите регуларне изразе, а овом приликом представљамо један који проналази стони прибор/ прибор за јело или посуда односно суд за којим следи „за“ и нека реч из домена кулинарства.

Напоменимо да <МОТ>* означава да се између речи могу појавити још неке речи, али њихово појављивање није обавезно.

Образац:
(<стони><прибор>)+(<прибор><за><МОТ>*<јело>)+((<посуда>+<суд>)<за><DOM= Culinary>) је дао 170 кандидата. Ево неколико примера:

ментруга	посуда за брашно
илика	посуда за вино
мерница	посуда за жито
јушник	посуда за јуху
голица	посуда за млијеко
кртољ	посуда за кашике
есцајг	прибор за јело
једало	прибор за јело
набадаљка	део прибора за јело
карафиндл	део стоног прибора
кашика	део стоног прибора за јело
биберњача	суд за бибер; исп. биберница, биберњак.
бакра	суд за вариво
кондир	суд за вино
котрва	суд за воду
ешаљка	суд за храну;

За јела, пите, колаче и слаткише је креиран образац ***(<јело>+<пита>+<колач>+<слаткиш>) од***, који је пронашао 308 одговарајућих конкорданци, од којих наводимо неке:

ражњић	врста јела од мањих комада меса
јајченик	јело од куваног млека и јаја.
брокета	врста јела од брашна и меса

цушпајз јело од поврћа, вариво
ораховача колач од ораха
куглоф врста колача од брашна, јаја, млека
зељаница слана пита од зеља
патишпаљ, патишпан, патишпањ врста колача, слаткиша од брашна, шећера и јаја.

Закључна разматрања

Представљени резултати упућују на то да је повезивање различитих језичких ресурса и алата могуће и да доприноси обогаћењу и унапређењу терминолошког аспекта Речника САНУ, као и да пружа нову, савремену грађу која тек треба да буде обрађена у наредним томовима. Поступак представљен у овом раду није карактеристичан специјално за кулинарски домен, он се може применити и за друге домене (нпр. за рударство и геологију, где постоје развијени електронски речници и корпуси, а потом за математику, физику, етнологију, право итд.). Када је у питању конкретно лексика кулинарских текстова, даљи кораци би свакако обухватили проширење корпуса и дигитализацију осталих кувара у библиотеци Института САНУ. Свакако треба радити и на набавци нових кувара, пожељно у е-облику. Потребно је, такође, радити на даљој аутоматизацији која би убрзала: генерисање листе кандидата, процедуре уноса нових термина и допуњавање постојећих. План даљих истраживања обухвата и генерисање листе нових речи и израза, статистичке обраде корпуса, израду сервиса за екстракцију и лематизацију, као и имплементацију синтаксичких шема описаних у [Крстев, Лазић, 2015].

Оно што је потребно нарочито нагласити јесте да се процесом дигитализације чини видљивом и доступном лексика из целокупног корпуса Речника потврђена провереним и нарочитим скраћеницама обележеним примерима, што уз остале поступке, убрзава и рад на папирној верзији Речника.

Извори

Поповић-Миџина Катарина, *Велики српски кувар.* Нови Сад, 1904.
Речник српскохрватског књижевног и народног језика. Београд, 1959–2014, I–XIX.

Литература

1. Krstev, Cvetana, Duško Vitas and Gordana Pavlović-Lažetić. „Resources and methods in the morphosyntactic processing of Serbo-Croatian.” In Gerhild Zybatow et al. (eds.) Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003, pp. 3-17. Frankfurt am Main.
2. Vitas, D., Popović, Lj., Krstev, C., Obradović, I., Pavlović-Lažetić, G. and Stanojević, M. (2012). The Serbian Language in the Digital Age. Berlin; Springer-Verlag.
3. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography 1(1), pp.7–36.
4. Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp.61 –74.

5. Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:.the C-value/NC-value method. International Journal on Digital Libraries, 3(2), pp. 115-130.
6. Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. „Rule based automatic multi-word term extraction and lemmatization." In: 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož.
7. Ranka Stanković, Ivan Obradović, Cvetana Krstev, Duško Vitas, "Production of morphological dictionaries of multi-word units using a multipurpose tool", In: Proceedings of the Computational Linguistics-Applications Conference, 2011.
8. Милош Утвић, „Анотација Корпуса савременог српског језика". ИНФОтека 12, бр. 2 (децембар 2011): 39-51.
9. Staša Vujičić Stanković, Cvetana Krstev, Duško Vitas, "Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain", In The Proceedings of Seventh Global WordNet Conference 2014.
10. Jurafsky, D. (2014). The Language of Food: A Linguist Reads the Menu. New York City, NY, USA: W.W. Norton & Company
11. Cvetana Krstev, Biljana Lazić, "Glagoli u kuhinji i za stolom", Naučni sastanak slavista u Vukove dane - Srpski jezik i njegovi resursi: teorija, opis i primene, Vol. 44/3, pp. 117-136, 2015, Međunarodni slavistički centar, Beograd
12. Александра Тртовац, Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама, докторска дисертација (у процедури за одбрану), Ментор: проф. Цветана Крстев

Рада Стијовић, Олга Сабо, Ранка Станковић

Резиме

РЕЧНИК САНУ КАО БАЗА ТЕРМИНОЛОШКИХ РЕЧНИКА

(НА ПРИМЕРУ РЕЧНИКА КУЛИНАРСТВА)

У раду се разматра могућност израде речника кулинарства на основу кулинарске лексике садржане у тезаурусном Речнику српскохрватског књижевног и народног језика САНУ.

Да би се дошло до релевантних података, било је потребно наћи метод екстраховања ове лексике, с обзиром на то да она (за разлику од већине других терминолошких области) није системски обележавана одговарајућим квалификаторима.

Као полазиште користили смо дигитализовану верзију Великог српског кувара Катарине Поповић Миџине (обрађену алатом Unitex уз примену морфолошких речника српског језика у DELA формату). После овога применили смо алат Leximir за екстраховање списка лема са фреквенцијама у кувару, који смо, потом, филтрирали семантичким маркерима.

Тако добијена листа упоређена је са одредницама Речника; одвојене су одреднице које се налазе у Речнику, а немају информације о кулинарском домену (треба их допунити), и речи које не постоје у Речнику (треба их унети).

Један део рада посветили смо коришћењу синтаксичких графова за екстракцију вишечланих лексичких јединица које се према фреквенцији издвајају као кандидати за термини јер је њихова учесталост у кулинарском тексту значајно већа од учесталости у корпусу савременог српског језика.

На овај начин успели смо да идентификујемо изузетно богат фонд кулинарске лексике садржане у Речнику и покажемо како традиционални речник током процеса дигитализације постаје база терминолошких речника која омогућава разноврсну употребу. Осим тога, процес дигитализације и осавремењивање рада на Речнику омогућавају да се Речник обогати новом, савременом лексиком и новим примерима, а рад на Речнику убрза.

SASA DICTIONARY AS A BASE FOR TERMINOLOGICAL DICTIONARIES

(ON THE EXAMPLE OF CULINARY VOCABULARY)

In this paper is discussed the possibility of creating culinary vocabulary based on the culinary lexica contained in the Dictionary of Serbo-Croatian Literary and folk Language SASA, that is built as a thesaurus.

To obtain the relevant data, it was necessary to find a method of extracting this lexicon, given that it (unlike most other areas of terminology) is not systematically labelled with appropriate qualifiers.

As a starting point, we used a digitized version of the Great Serbian chef Katarina Popovic Midžine (processed using Unitex tool and morphological dictionaries for Serbian language in DELA format). After this we applied Leximir tool to extract a list of the lemma frequencies in the cookbook, which are then filtered and classified by applying semantic markers. The obtained list was compared to a list of the lexical entry of the SASA dictionary; extracted the entries from the Dictionary that have no information about the culinary domain (should be amended), and words that do not exist in the Dictionary (should be entered).

In this research are also used syntactic graphs for the extraction of multiword lexical units that are allocated to the frequency of terms that are significantly more frequent in a culinary text than in the corpus of contemporary Serbian language.

Using this approach, we were able to identify extremely rich term collection for culinary lexicon contained in the SASA Dictionary and show how traditional vocabulary during the digitization process becomes a base for terminology dictionaries that allows different uses. In addition, the process of digitalization and modernization of work on the SASA Dictionary provide its enrichment with the new, modern lexica and new examples, and to accelerate work on the SASA Dictionary.