Претрага корпуса заснована на употреби екстерних лексичких ресурса путем веб-сервиса

Милош Утвић, Ранка Станковић, Александра Томашевић, Михаило Шкорић, Биљана Лазић

811.163.41'322.2
811.163.41'373.46:622
https://doi.org/10.18485/msc.2019.48.3.ch12

Милош В. УТВИЋ*
Филолошки факултет
Универзитета у Београду
Ранка М. СТАНКОВИЋ**
Александра Ђ. ТОМАШЕВИЋ*
Михаило Ђ. ШКОРИЋ
Биљана Ђ. ЛАЗИЋ
Рударско-геолошки факултет
Универзитета у Београду

ПРЕТРАГА КОРПУСА ЗАСНОВАНА НА УПОТРЕБИ ЕКСТЕРНИХ ЛЕКСИЧКИХ РЕСУРСА ПУТЕМ ВЕБ-СЕРВИСА

У раду се разматра хибридни приступ претрази корпуса, илустрован на примеру алатки OCWB и NoSketch Engine, примењених на специјални корпус из области рударства (РудКор) и Корпус савременог српског језика (СрпКор). Разматрани приступ комбинује постојеће могућности алатки OCWB и NoSketch Engine, које своју претрагу заснивају на лингвистичкој анотацији корпуса, са новим могућностима претраге у виду консултовања екстерних језичких ресурса (морфолошки електронски речници српског језика и лексичка база података Српски ворднет). Хибридни приступ је реализован надоградњом веб-сучеља која поменуте алатке користе за претрагу корпуса, као и употребом веб-сервиса заснованих на технологији RESTful. Веб-сервиси омогућавају да се задати упит, у коме се појављује лексема X, по потреби прошири лексемама које се са лексемом X налазе у некој лексичкој релацији (синонимија, антонимија, хиперонимија итд.).

Кључне речи: корпус, рударство, претраживање информација, проширивање упита, лексички ресурси, лексичке релације

Увод

Софтверске алатке за претрагу корпуса заснивају свој рад у сврху екстракције лингвистичких информација из корпуса:

* misko@matf.bg.ac.rs
** {ranka.stankovic, aleksandra.tomasevic, mihailo.skoric, biljana.lazic}@rgf.bg.ac.rs

- на лингвистичкој анотацији корпусних текстова,
- на консултовању екстерних језичких ресурса током претраге и
- на комбиновању поменутих приступа.

У овом раду ће бити речи о комбинованом приступу претрази корпуса заснованој и на лингвистичкој анотацији корпусних текстова и на консултовању екстерних језичких ресурса. У питању је природан наставак претходног заједничког рада (Утвић и др. 2018) у коме је описана изградња специјализованог корпуса (под називом Рударски корпус, скр. РудКор), с једне стране ради ефикасније претраге пројектне документације из области рударства, а са друге стране као прилика да се тестирају различите софтверске алатке за претрагу корпуса и испитају услови за њихову надградњу. Овај рад разматра једну реализовану надградњу веб-сучеља (енгл. web interface) двеју алатки за претрагу корпуса, OCWB[1] (Еверт-Харди 2011) и NoSketch Engine (Рихли 2007), са циљем да се резултати искористе не само за претрагу Рударског корпуса већ и других корпуса које развијамо. Добијени резултати биће илустровани на примеру корпуса РудКор и СрпКор2013, текуће верзије Корпуса савременог српског језика (Корпус 2013), као представнике специјализованих корпуса (РудКор), односно корпуса опште намене (СрпКор), али исти хибридни приступ се може применити и на друге врсте корпуса: лексикографске, граматичке, дијалекатске, регионалне, нестандардне, корпусе језика као нематерњег, корпусе струка (енгл. *domain specific corpora*) итд.

У одељку 2 рада се, у општим цртама, описује лингвистичка анотација корпуса РудКор и СрпКор2013, као и могућности претраге тих корпуса помоћу алатки OCWB и NoSketch Engine, утемељене на описаној лингвистичкој анотацији. Одељак 3 даје преглед веб-сервиса за експанзију упита над корпусом који су засновани на лексичким и терминолошким ресурсима са посебним нагласком на рударску терминологију. Одељак 4 разматра имплементацију хибридног приступа претраге корпуса и илуструје могућности претраживања коришћењем семантичких проширења. Напослетку, одељак 5 наводи преглед закључака и смернице будућег развоја.

Претрага заснована на лингвистичкој анотацији корпуса

Алатке OCWB и NoSketch Engine подржавају изградњу и претрагу корпуса са уграђеном лингвистичком анотацијом у формату вертикал(изова)ног текста (Утвић 2014а: 129–130), као и у формату XML. Табела 1 илуструје како изгледа једна реченица из Андрићеве приповетке *Прича о везировом слону* у формату вертикалног текста (прва колона), односно у формату анотираног вертикалног текста у корпусу СрпКор2013. Поједностављено, OCWB и NoSketch Engine третирају корпус као табелу чији сваки ред одговара или

[1] OCWB је скраћеница од Open Corpus Workbench.

појединачној позицији у корпусу (токену) или XML-етикети, прва колона (која је и једина обавезна) представља конкретну вредност токена или XML-етикету, док остале колоне, ако постоје, садрже одређени тип информација придружен токену (Еверт 2019: 6). Између осталог, колоне тако моделованог корпуса се користе за придруживање лингвистичких информација токену, попут леме, врсте речи, вредности морфосинтаксичких категорија (падеж, род, број, глаголски облик и сл.), ознаке значења, ознаке примењеног правописа итд. XML-етикетама се означава почетак или крај неке структурне целине (на пример, токена, реченице, пасуса, наслова, комплетне главе романа итд.), па се и оне могу користити за придруживање лингвистичких информација, не само на нивоу токена, већ и на нивоу осталих поменутих структурних целина. У терминологији алатки OCWB и NoSketch Engine, колоне корпуса, моделованог на описани начин, се још називају **позициони атрибути**, док се различити називи употребљених XML-елемената у корпусу називају **структурни атрибути**. Анотација корпуса, који се претражују алаткама OCWB и NoSketch Engine, реализује се комбиновањем позиционих и структурних атрибута (Табела 2).

Наведени пример (Табела 1) илуструје анотацију применом позиционих атрибута pos и lemma који редом одговарају ознаци врсте речи и леми појединачног токена (позициони атрибут word). Вредности CONJ, N, V, PAR, ADV атрибута pos редом представљају ознаку за врсте речи везник, именица, глагол, партикула, прилог; вредност SENT атрибута pos, придружена последњем токену (тачка), означава да је у питању крај реченице[2].

У пододељцима 2.1 и 2.2 укратко су описане верзије корпуса СрпКор и РудКор, конструисане помоћу алатки OCWB (СрпКор и РудКор) и NoSketch Engine (РудКор). Наведени корпуси садрже анотацију метаподацима, лингвистичку анотацију и делимичну структурну анотацију, али опис у пододељцима 2.1 и 2.2 ставља нагласак на њихову лингвистичку анотацију. У пододељку 2.3 су илустроване неке могућности претраге тих корпуса засноване на њиховој лингвистичкој анотацији.

Табела 1: Реченица из Андрићеве приповетке *Прича о везировом слону* у форми вертикалног текста (прва колона), односно у форми анотираног вертикалног текста у корпусу СрпКор2013 (колоне 2–4 као позициони атрибути word, pos и lemma)[3].

[2] Комплетна листа вредности атрибута `pos`, које се користе као ознаке врсте речи у корпусу СрпКор2013, доступна је на адреси http://www.korpus.matf.bg.ac.rs/SrpLemKor/tagset.html.

[3] У корпусу СрпКор2013 се карактери *Ш* (*Š*), *ш* (*š*), *Ж* (*Ž*), *ж* (*ž*), *Ћ* (*Ć*), *ћ* (*ć*), *Ч* (*Č*), *ч* (*č*), *Ђ* (*Đ*), *ђ* (*đ*), *Љ* (*Lj*), *љ* (*lj*), *Њ* (*Nj*), *њ* (*nj*), *Џ* (*Dž*), *џ* (*dž*) редом представљају следећим записима (карактерски кôд аурора): *Sx*, *sx*, *Zx*, *zx*, *Cx*, *cx*, *Cy*, *cy*, *Dx*, *dx*, *Lx*, *lx*, *Nx*, *nx*, *Dy* и *dy*.

Реченица (P1)	**Представљање реченице (P1) у корпусу СрпКор2013**		
	Позиционi атрибут word **(вредност токена)**	**Позиционi атрибут** pos **(ознака врсте речи)**	**Позиционi атрибут** lemma **(одговарајућа лема)**
А	A	CONJ	a
везира	vezira	N	vezir
су	su	V	jesam
Травничани	Travnicyani	N	Travnicyanin
заиста	zaista	PAR	zaista
ретко	retko	ADV	retko
виђали	vidxali	V	vidxati
.	.	SENT	.

1.1 Корпус савременог српског језика (СрпКор)

СрпКор2013 је актуелна верзија Корпуса савременог српског језика (Корпус 2013)[4]. СрпКор2013 је корпус опште намене величине преко 122 милиона корпусних речи, односно преко 152 милиона токена.[5]

СрпКор2013 поседује делимичну морфолошку анотацију, у смислу да су сваком токену придружене информације о врсти речи и леми на начин описан у (Утвић 2011). Софтверска алатка TreeTagger (Schmid 1997, Schmid 1999) је употребљена за аутоматску морфолошку анотацију текстова корпуса СрпКор2013. Параметарска датотека (енгл. language parameter file) за српски језик, коју TreeTagger користи као дигитални лексички ресурс за анотацију, настала је као дериват система морфолошких електронских речника српског језика (у даљем тексту: СМР) чији су аутори Цветана Крстев и Душко Витас (Крстев 2008). Делимична морфолошка анотација у корпусу СрпКор2013 је реализована позиционим атрибутима pos (ознака врсте речи) и lemma (лема), који садрже одговарајуће информације о токену, тј. подразумеваном позиционом атрибуту word (Табела 1).

Верзија СрпКор2013 није структурно анотирана, мада појединачни корпусни текстови, од којих је верзија СрпКор2013 компилирана, поседују структурну анотацију на неким или свим структурним нивоима одељка, наслова, пасуса, реченице. Табела 2 илуструје структурну анотацију у формату вертикалног текста.

---

[4] http://www.korpus.matf.bg.ac.rs/korpus/login.php

[5] Ако скуп неалфанумеричких карактера третирамо као скуп сепаратора, корпусне речи дефинишемо као непразне ниске карактера између два узастопна сепаратора. За корпусне речи и појединачне елементе скупа сепаратора (изузимајући белине, тј. карактере размак, табулатор и знак за нови ред) користи се заједнички назив токени (Утвић 2014б: 244).

Табела 2: Реченица из Андрићеве приповетке *Прича о везировом слону* у форми анотираног вертикалног текста са структурним атрибутом **s**, односно са XML-етикетама **<s>** и **</s>** које редом означавају почетак, односно крај реченице. Позициони атрибути и њихове вредности су истоветни као у Табели 1.

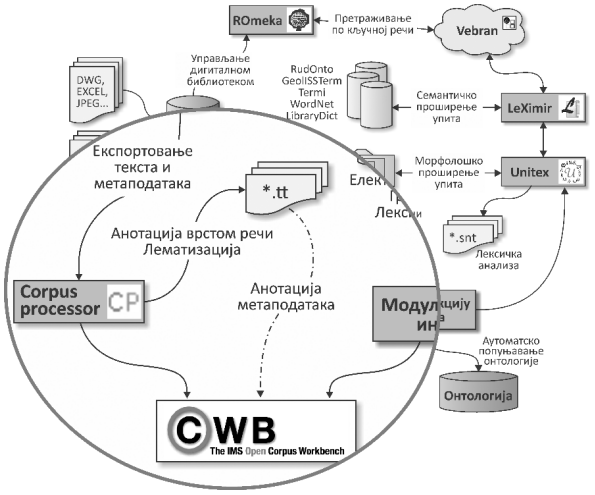
Реченица (P1)	**Представљање реченице (P1) у корпусу СрпКор2013**		
	Позициони атрибут word **(вредност токена или XML-етикета)**	**Позициони атрибут** pos **(ознака врсте речи)**	**Позициони атрибут** lemma **(одговарајућа лема)**
	<s>		
А	A	CONJ	a
везира	vezira	N	vezir
су	su	V	jesam
Травничани	Travnicyani	N	Travnicyanin
заиста	zaista	PAR	zaista
ретко	retko	ADV	retko
виђали	vidxali	V	vidxati
.	.	SENT	.
	</s>		

1.2 Специјализовани корпус из области рударства

У склопу система за управљање рударском пројектном документацијом у електронском облику (Томашевић и др. 2018; Томашевић 2018) коришћене су различите језичке технологије, међу којима издвајамо:

- дигиталну библиотеку, као централни репозиторијум за складиштење рударске пројектне документације,
- лексичке и терминолошке ресурсе и
- корпус стручних текстова на српском језику из области рударства, РудКор.

Слика 1 приказује место корпуса РудКор у систему за управљање рударском пројектном документацијом у целини, где је истакнут део система који се односи се на анотацију докумената врстом речи и лемом, снабдевање докумената метаподацима и генерисање самог корпуса.



Слика 1. Рударски корпус у систему за управљање документацијом

Изградња Рударског корпуса детаљно је описана у раду (Томашевић 2018; Утвић и др. 2018). Корпус чине 172 документа, и то: из области законске регулативе (правилници, закони, уредбе, смернице и стратегије) 94 документа (55%), литература (докторске дисертације, уџбеници и монографије) 40 докумената (23%) и пројектна документација (пројекти и студије) 38 докумената (22%). РудКор има 2.719.086 корпусних речи. РудКор садржи 100.414 корпусних типова, као елемената скупа различитих корпусних речи, док је број различитих лема 22.875.

Лингвистичка анотација актуелне верзије Рударског корпуса је аналогна лингвистичкој анотацији корпуса СрпКор2013, тј. сваки токен је анотиран на нивоу леме и врсте речи истим ресурсима, алаткама и методологијом којом је анотиран и СрпКор2013, описаним у одељку 2.1, тј. Рударски корпус користи позиционе атрибуте word (токен), lemma (лема) и tag (ознака врсте речи).[6] Тренутна верзија Рударског корпуса није структурно анотирана у целости, мада се на поткорпусу анотираних (углавном закона и правилника) тестира претрага по структурним обележјима (Утвић и др. 2018: 110–113).

1.3 Претрага корпуса СрпКор2013 и РудКор

Корпуси СрпКор2013 и РудКор доступни су за претрагу на вебу (Корпус 2013, РудКор 2018). СрпКор2013 се претражује помоћу алатке OCWB, док се за претрагу корпуса РудКор може користити или OCWB или NoSketch Engine. Информациони систем, који користи поменуте алатке за претрагу корпуса, комбинује веб-сучеље, као предњи део (енгл. front-end) система, и

[6] Комплетна листа вредности атрибута `tag` корпуса РудКор је истоветна листи вредности атрибута `pos` у корпусу СрпКор2013 на адреси http://www.korpus.matf.bg.ac.rs/SrpLemKor/tagset.html.

програмски систем за креирање, управљање и претрагу корпуса, као задњи део (енг, back-end) информационог система (Утвић и др. 2018: 112–113). Задњи део информационог система, поред осталог, садржи модул за обраду упита (у даљем тексту: процесор упита) који интерпретира задати упит, генерише резултате претраге корпуса, а веб-сучеље те резултате форматира и представља их кориснику. Према томе, веб-сучеље је заправо посредник између корисника и процесора упита, у једном смеру преноси задати упит од корисника до процесора упита, а у другом смеру преузима одговор модула за обраду упита и представља га кориснику. OCWB користи CQPweb (Харди 2012) као веб-сучеље, а IMS CWB (Еверт-Харди 2011) као задњи део информационог система, а веб-сучеље Bonito и алат за управљање корпусима Manatee чине, тим редом, предњи и задњи део информационог система NoSketch Engine (Рихли 2007). CQP (енгл. Corpus Query Processor) је назив процесора упита за OCWB (Еверт 2019).

Приликом задавања упита у оквиру напредне претраге, користи се одговарајући упитни језик заснован на регуларним изразима. Иако CQP и Manatee, као процесори упита, користе међусобно различите упитне језике — CQP-упитни језик[7] (Еверт 2019), односно Корпусни упитни језик или CQL[8] (SketchEngineDoc 2019) — формати тих упитних језика за претраживање позиционих атрибута корпуса РудКор и СрпКор2013 готово су идентични (Табела 3). Најважније разлике о којима треба водити рачуна односе се на кодирање карактера у корпусу (алфабет аурора и карактерски кôд ISO-8859-1 за СрпКор2013, односно српска латиница и карактерски кôд UTF-8 за РудКор).

Табела 3: Примери претраге корпуса РудКор и СрпКор2013 заснованих на њиховој лингвистичкој анотацији

Циљ претраге	**Корпус**	**Упит**	**Значење упита**
облици леме *ископина*	РудКор, СрпКор2013	[lemma=»iskopina»]	Сви токени којима је придружена *iskopina* као вредност атрибута lemma.
беседи као облик глагола	СрпКор2013	[word=»besedi» & pos=»V»]	Сваки токен *besedi* коме је као вредност атрибута pos придружена вредност V.
бацили као облик именице	РудКор	[word=»bacili» & tag=»N»]	Сваки токен *bacili* коме је као вредност атрибута tag придружена вредност N.

[7] У енгл. оригиналу CQP Query Language.
[8] У енгл. оригиналу Corpus Query Language, скр. CQL.

Веб-сервиси за експанзију упита над корпусом

Веб-сервиси су програмабилне целине са одређеном функционалношћу којој се може приступити из различитих система користећи интернет стандарде (HTTP/HTTPS, XML, JSON итд) и то:

- интерно, од стране једне апликације или информационог система, или
- јавно (изложени као јавни интернет сервиси), од стране више хетерогених система који евентуално могу радити заједно.

Вебран је скуп веб-сервиса направљених са циљем да омогући проширење упита, засновано на лексичким ресурсима, које упућују овлашћени корисници и апликације, а посебно је прилагођен за претрагу корпуса помоћу алатки OCWB и NoSketch Engine. Поред претраживања корпуса, Вебран се користи и за претрагу једнојезичних и вишејезичних дигиталних библиотека развијених у оквиру Групе за језичке технологије Универзитета у Београду и Друштва за језичке ресурсе и технологије Јертех. Развој Вебрана је заснован на Мајкрософт технологији RESTful API[9] (ASP.NET Web API) за развој веб-сервиса коју одликује једноставност и флексибилност формата генерисаних података.

Експанзија упита заснована на лексичким ресурсима је потребна за различите намене, од претраживања дигиталних библиотека и веба, до корпуса, при чему су поменути системи по правилу имплементирани на различитим платформама и на дистрибуираним локацијама, тако да веб-сервиси представљају одговарајућу технологију за њихову имплементацију. Вебран генерише проширење упита на основу различитих лексичких ресурса српског језика, од којих су најзначајнији:

- СМР, систем морфолошких електронских речника српског језика (Крстев 2008),
- семантичка мрежа Српски ворднет[10] (Крстев и др. 2004),
- терминолошке базе Терми[11],
- онтологија из области рударства РудОнто[12],
- Геолошки речник ГеолИСС[13],
- Речник библиотекарста и информатике[14].

Вебран-сервиси за проширење упита се користе на следећи начин (Станковић 2009; Станковић и др. 2017):

1. Апликација која прихвата кориснички упит контактира веб-сервис ради проширења упита. Контакт се остварује преко адресе веб-сер-

[9] https://dotnet.microsoft.com/apps/aspnet/apis
[10] http://www.korpus.matf.bg.ac.rs/SrpWN/
[11] http://termi.rgf.bg.ac.rs/
[12] http://rudonto.rgf.bg.ac.rs/
[13] http://geoliss.mre.gov.rs/recnik/
[14] http://rbi.nb.rs/

виса (униформни локатор ресурса, енгл. Uniform Resource Locator, скр. URL) на коју апликација шаље одговарајуће параметре (Табела 4).

2. На основу добијених параметара, веб-сервис консултује екстерне лексичке ресурсе и, на основу података о задатој леми X, генерише резултат који шаље назад апликацији.
3. Апликација на основу одговора веб-сервиса генерише проширење упита, спроводи претрагу и обавештава корисника о резултатима претраге.

У овом раду разматрамо случај када се Вебран-сервиси користе за семантичко проширење упита које корисник задаје користећи као алатку за претрагу корпуса OCWB или NoSketch Engine.

Табела 4: Параметри веб-сервиса Вебран
(у заградама су наведене конкретне вредности параметара)

Параметар	**Значење параметра**
lema	задата лема X
relType	задата лексичка релација R (в. Табелу 5, колона „Интерни назив лексичке релације”)
POS	врста речи задате леме X: именица (N), придев (A), глагол (V)
lngIn	језик задате леме X, подразумевано српски (sr)
lngOut	језик резултата, подразумевано српски (sr)
alphOut	писма резултата: ћирилица (C), латиница (L), кôд аурора (A) или нека њихова комбинација (LC, CA, LA, LCA)
fleksije	индикатор да ли резултат садржи само леме које су у лексичкој релацији R са лемом X или и све флективне облике тих лема (true, false)
dlfByLemma	индикатор да ли се у резултату облици групишу по заједничкој леми или не (true, false)

Семантичко проширење упита, упућеног информационом систему који користи OCWB или NoSketch Engine (в. одељак 2.3), реализује се на следећи начин:

(СП1) корисник преко веб-сучеља алатке за претрагу корпуса (CQPweb или Bonito) поставља упит и при томе може да захтева да резултат буде у одређеној лексичкој релацији R (синонимија, антонимија, хиперонимија итд.) са неком од лема (означимо је са X) које се појављују у упиту. На пример, корисник може тражити флективну парадигму лексеме *ископина*, (Табела 3, упит

[lemma="iskopina"]), али истовремено захтевати да се пронађу и облици свих синонима те лексеме. Овде треба приметити да процесор упита може да претражи позиционе атрибут lemma и пронађе флективну парадигму лексеме *ископина*, али не може да интерпретира семантички проширени упит јер нема информацију о томе шта су синоними лексеме *ископина*;

(СП2) ако корисник није задао захтев за семантичким проширењем упита, веб-сучеље алатке за претрагу корпуса директно прослеђује упит процесору упита и прелази се на корак (СП8);

(СП3) ако је корисник задао захтев за семантичким проширењем упита, веб-сучеље анализира упит и прослеђује веб-сервису, поред задате леме X и осталих параметара, ознаку одређене лексичке релације R (синонимија, антонимија, хиперонимија итд.) коју је корисник одабрао;

(СП4) консултовањем екстерних лексичких ресурса, веб-сервис проналази све леме које се налазе у лексичкој релацији R са лемом X (в. одељак 3);

(СП5) у зависности од захтева ког је упутило веб-сучеље алатке за претрагу корпуса, веб-сервис може резултату, добијеном у кораку (СП4), придружити и флективне парадигме пронађених лема;

(СП6) веб-сервис шаље веб-сучељу алатке за претрагу корпуса генерисани резултат;

(СП7) на основу резултата добијеног у кораку (СП6), веб-сучеље (CQPweb или Bonito) формира нови, семантички проширен упит, у складу са форматом одговарајућег упитног језика (CQP-упитни језик или CQL) и прослеђује коначни упит процесору упита;

(СП8) процесор упита (CQP или Manatee) реализује одговарајућу претрагу, саопштава резултате веб-сучељу које потом одговара кориснику на постављени упит.

Описану процедуру називамо хибридни (комбиновани) приступ претрази корпуса, с обзиром да алатке OCWB и NoSketch Engine заснивају своју претрагу на лингвистичкој анотацији, а разматрана надградња консултује екстерне језичке ресурсе током претраге. О детаљима имплементације и примерима упита који подржавају хибридну претрагу биће више речи у одељку 4.

Табела 5 илуструје неке примере проширења упита коришћењем различитих лексичких релација доступних у семантичкој мрежи Српски ворднет (Крстев и др. 2004).

Табела 5: Примери одговора веб-сервиса Вебран на задате леме и лексичке релације (у свим случајевима је претпостављено да одговор садржи само леме, не и њихове флективне парадигме).

Интерни назив лексичке релације	**Лексичка релација**	**Лема полазног упита**	**Одговор веб-сервиса (формат без груписања лема и без флективне парадигме)**
synset	синонимија	затвор	тамница; казнени завод
synset	синонимија	трачни транспортер	транспортер са траком, тракасти транспортер, транспортни систем
near_antonym	антонимија	оповргнути	доказати; показати
hyponym	хипонимија	геолошки процес	ерозија; спирање; осипање
hyponym	хипонимија	радник	намештеник; запосленик; трудбеник; волонтер; помоћник; помагач
hypernym	хиперонимија	лимонит	гвоздена руда; руда гвожђа
hypernym	хиперонимија	птица	кичмењак
holo_portion	холонимија	калцијум	гипс; кречњак; вапненац; флуорит; калцијум оксид; гашени креч; вапно

Имплементација хибридног приступа претраживању корпуса

У одељку 3 је размотрен хибридни приступ претраживању корпуса помоћу веб-сервиса и том приликом су у опису реализације наведени кораци (СП1), (СП3) и (СП7) који захтевају модификацију веб-сучеља информационог система за претрагу корпуса. CQPweb и Bonito, као веб-сучеља алатки OCWB и NoSketch Engine, отвореног су кода што омогућава њихову модификацију, тј. уграђивање описане комуникације са Вебран-сервисима у постојећи изворни кôд веб-сучеља. Модификације су имплементиране у развојним технологијама карактеристичним за њих: CQPweb — у програмском језику PHP, Bonito — у програмском језику Python. Модификације се односе на:

(М1) начин на који корисник преко веб-сучеља захтева семантичко проширење упита;

(М2) остваривање контакта веб-сучеља и веб-сервиса, укључујући и прослеђивање вредности параметара веб-сервиса (Табела 4);

(М3) генерисање семантички проширеног упита у формату који је дозвољена синтаксичка конструкција упитног језика, односно који процесор упита може да интерпретира.

Модификација (М1) се, пре свега, односи на формат (синтаксичку структуру) упита који корисник може да постави. У ту сврху се уводе квази-позиционе атрибути (у даљем тексту: КПА-и) synlemma, antlemma и hyperlemma од којих сваки тим редом одговара одређеној лексичкој релацији: синонимија, антонимија, хиперонимија (Табела 6). КПА-и се користе у упиту на исти начин као прави позиционе атрибути дефинисани у одељку 2 и илустровани у пододељку 2.3. Међутим, пошто КПА-и не постоје као стварне колоне у вертикалном формату корпусних текстова (уп. Табела 1), процесор упита не може да интерпретира упит који садржи неки КПА, те су неопходне модификације (М2) и (М3).

Модификација (М2) се састоји у примени подршке коју програмски језици PHP и Python пружају када је у питању RESTful технологија, односно контактирање веб-сервиса. Параметри веб-сервиса се прослеђују у формату JSON. При томе, највећи број вредности параметара који се прослеђују веб-сервису је фиксиран, вредност параметра *lemma* се директно преузима од веб-сучеља, једино је потребно израчунати вредност параметра *relType*, односно ознаку задате лексичке релације (синонимија, антонимија, хиперонимија), што се постиже анализом упита и препознавањем одговарајућег квази-позиционог атрибута (synlemma, antlemma, hyperlemma).

Табела 6: Лексичке релације, одговарајући квази-позициони атрибути (КПА) и примери претраге

Лексичка релација / КПА	**Упит који користи квази-позициони атрибут (КПА)**	**Значење упита који користи КПА**
синонимија / synlemma	[synlemma=»zatvor»]	синоними леме zatvor (в. Табелу 5) укључујући и њихову флективну парадигму, као и облике леме zatvor
антонимија / antlemma	[antlemma=»opovrgnuti»]	антоними леме opovrgnuti (в. Табелу 5), укључујући и њихову флективну парадигму
хиперонимија / hyperlemma	[hyperlemma=»ptica»]	хипероними леме ptica (в. Табелу 5), укључујући и њихову флективну парадигму

Модификација (М3) се односи на обраду одговора веб-сервиса који може бити у различитим форматима (XML, JSON), укључујући и замену квази-позиционих атрибута стварним позиционим атрибутима корпуса РудКор и СрпКор2013.

За разлику од правих позиционих атрибута, вредност квази-позиционих атрибута може бити и монолексемска јединица (једночлана лексема) и полилексемска јединица (вишечлана лексема, в. Табелу 5, *геолошки процес*). Такође, одговор веб-сервиса може садржати и монолексемске и полилексемске јединице (в. Табелу 5, *казнени завод* као синоним за лексему *затвор*). Пошто процесори упита алатки OCWB и NoSketch Engine подржавају искључиво монолексемске јединице као вредности (правих) позиционих атрибута, неопходно је у одговору веб-сервиса раздвојити монолексемске и полилексемске јединице пошто се оне морају обрадити на различите начине. У сврху раздвајања монолексемских и полилексемских јединица, у одговору веб-сервиса се испред монолексемске јединице наводи префикс S: (почетно слово енгл. термина *simple word* за монолексемске јединице), а испред полилексемске јединице се наводи префикс C: (почетно слово енгл. термина *compound* за полилексемске јединице).

Табела 7: Обрада одговора веб-сервиса и генерисање проширеног упита

1	**Упит који користи КПА**	[hyperlemma=»ptica»]
	Одговор веб-сервиса	S:kičmenjak
	Генерисани упит	[lemma=»kičmenjak»]
2	**Упит који користи КПА**	[antlemma=»opovrgnuti»]
	Одговор веб-сервиса	S:dokazati; S:pokazati
	Генерисани упит	[lemma=»dokazati|pokazati»]
3	**Упит који користи КПА**	[synlemma=»zatvor»]
	Одговор веб-сервиса	S:tamnica; C:kazneni zavod
	Генерисани упит	[lemma=»zatvor|tamnica»]|([lemma=»kazneni»] [lemma=»zavod»])

Генерисање проширеног упита у случају када одговор веб-сервиса садржи само леме, не и њихове флективне парадигме, реализује се у зависности од структуре одговора (Табела 7):

(Г1) Одговор веб-сервиса садржи само једну монолексемску јединицу (на пример, „S:kičmenjak”) и у том случају се упит генерише тако што се у полазном корисниковом упиту квази-позициони атрибут замењује стварним позиционим атрибутом lemma, а његова вредност одговором веб-сервиса без префикса (Табела 7, 1. упит: [lemma=”kičmenjak”]).

(Г2) Одговор веб-сервиса садржи искључиво монолексемске јединице, две или више њих. У полазном корисниковом упиту квази-позициони атрибут се такође замењује стварним позиционим атрибутом lemma, а вредност атрибута одговором веб-сервиса

без префикса у коме су монолексемске јединице међусобно раздвојене ознаком | за операцију уније (Табела 7, 2. упит: [lemma =”dokazati|pokazati”] са значењем „наћи све токене чија је лема dokazati или pokazati”).

(Г3) Одговор веб-сервиса садржи тачно једну полилексемску јединицу. Ове варијанте нема у Табели 7, али се може илустровати на примеру одговора „C:kazneni zavod”, односно „C:auto-put”. У том случају се одговор без префикса (полилексемска јединица) разбија на делове коришћењем размака или цртице као сепаратора, сваки појединачни део постаје вредност једног позиционог атрибута lemma, сваки сепаратор који није белина (на пример, цртица) постаје вредност једног позиционог атрибута word, а коначни упит је заправо низ тако генерисаних позиционих атрибута који се групише у целину помоћу заграда:
([lemma=”kazneni”] [lemma=”zavod”])
односно
([lemma=”auto”] [word =”-”] [lemma=”put”])

Значење овако добијених упита је „наћи све парове токена такве да је лема првог токена kazneni, а лема токена који за њим следи — zavod”, односно „наћи све уређене тројке токена такве да је лема првог токена auto, други токен је цртица, а лема токена који за њима следи — put”.

(Г4) Одговор веб-сервиса садржи више искључиво полилексемских јединица. Ни ова варијанта није присутна у Табели 7, али се може објаснити на примеру одговора „C:kazneni zavod; C: popravni dom”. Свака појединачна полилексемска јединица генерише по један део упита на начин описан у кораку (Г3), а коначан упит се формира спајањем тих делова ознаком | за операцију уније: ([lemma=”kazneni”][lemma=”zavod”])|([lemma=”popravni”][lemma=”dom”]). Значење овако добијеног упита је „наћи све парове токена такве да је лема првог токена kazneni, а лема токена који за њим следи — zavod, или такве да је лема првог токена popravni, а лема токена који за њим следи — dom”.

(Г5) Одговор веб-сервиса садржи и монолексемске и полилексемске јединице. У том случају монолексемске јединице генеришу један део коначног упита на начин описан у (Г1) или (Г2), полилексемске јединице генеришу други део коначног упита примењујући (Г3) или (Г4), а коначан упит се формира спајањем тих делова ознаком | за операцију уније. Управо тако је у Табели 7 на основу одговора веб-сервиса „S:tamnica; C:kazneni zavod” генерисан упит
[lemma=”zatvor|tamnica”]|([lemma=”kazneni”] [lemma=”zavod”])

са значењем „наћи све токене чија је лема zatvor или tamnica, или све парове токена такве да је лема првог токена kazneni, а лема токена који за њим следи — zavod”.

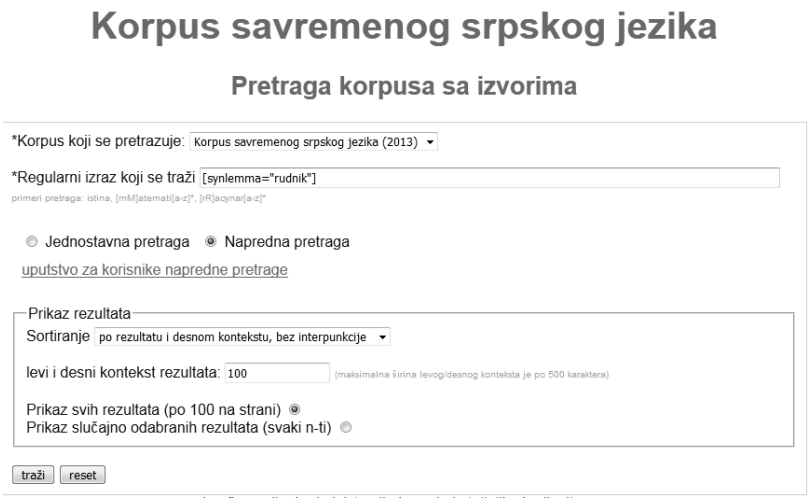
На аналоган начин се на основу одговора веб-сервиса може генерисати упит који садржи не само леме, већ и њихове флективне парадигме, а главна разлика је што се квази-позициони атрибут у корацима (Г1)–(Г5) замењује стварним позиционим атрибутом word уместо атрибутом lemma.

Закључак и будући рад

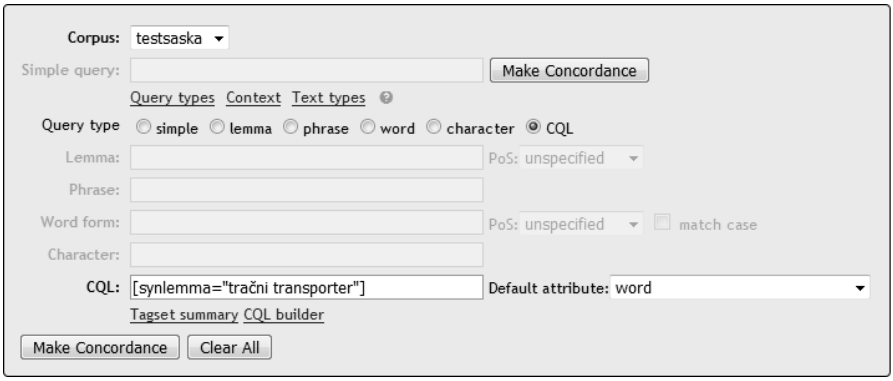
У раду је изложен приступ претрази корпуса комбиновањем лингвистичке анотације корпуса и екстерних лексичких ресурса који омогућава да резултати претраживања обухвате и лексеме које су са задатом лексемом у одабраној лексичкој релацији (синонимија, антонимија, хиперонимија). Описани хибридни приступ је успешно тестиран модификовањем веб-сучеља алатки за претрагу корпуса OCWB и NoSketch Engine, при чему је консултовање екстерних лексичких ресурса реализовано путем веб-сервиса Вебран. Вебран-сервиси су тренутно доступни само ауторизованим корисницима и апликацијама.

У овој фази развоја, упит са захтевом за семантичко проширење може се поставити једино у формату одговарајућег упитног језика (CQP-упитни језик или CQL) са квази-позиционим атрибутима (synlemma, antlemma и hyperlemma) и то:

- за корпус СрпКор2013: преко веб-сучеља Корпуса савременог српског језика у режиму напредне претраге (Слика 2);
- за корпус РудКор:
  - преко веб-сучеља Bonito (NoSketch Engine) у режиму CQL (Слика 3).
  - преко веб-сучеља CQPweb (OCWB) у режиму „CQP syntax extended (SrpWN)” (Слика 4);

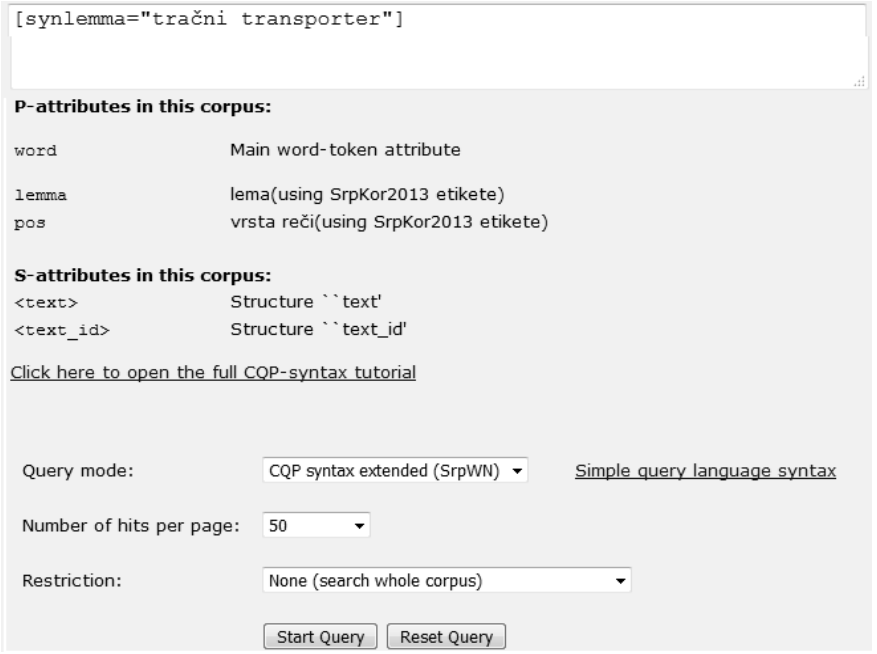


Слика 2: Веб-сучеље Корпуса савременог српског језика за претрагу корпуса СрпКор2013. Неопходно је одабрати режим напредне претраге пре задавања упита [synlemma="rudnik"] са захтевом за семантичко проширење (5481 резултат, поред лексеме *рудник*, садржи и њене синониме *окно*, *цех*, *површински коп*).



Слика 3: Веб-сучеље Bonito (NoSketch Engine) за претрагу Рударског корпуса у режиму (енгл. Query type) CQL (упит са захтевом за семантичким проширењем [synlemma="tračni transporter"] се уноси у истоимено поље CQL). Упит има 1043 резултата (*tračni transporter* 662, *transportni sistem* 381).

Иако веб-сервиси могу у свом одговору да врате и флективну парадигму лексема које су са задатом лексемом у захтеваној лексичкој релацији и да тако поправе однос пронађених резултата у односу на укупан одговарајућих јединица корпуса које су релевантне за претрагу, с обзиром да семантичко проширење упита повећава време одзива веб-сучеља (време потребно за међусобну размену података између веб-сучеља и веб-сервиса, као и за сам рад веб-сервиса, односно консултовање лексичких ресурса), у овој фази развоја смо одлучили да веб-сучеље захтева од веб-сервиса искључиво леме. Тиме се проблем проналажења флективне парадигме лексема добијених семантичким проширењем пребацује са веб-сервиса на процесор упита, односно претрагу позиционог атрибута lemma и прецизност лингвистичке анотације корпуса (в. одељак 4, кораци (Г1)–(Г5)). Непосредна последица је да проширивање упита неће функционисати на жељени начин за поједине класе полилексемских јединица, тј. произвешће одговарајући резултат само за полилексемске јединице чија је свака компонента заправо лема неке монолексемске јединице у консултованим лексичким ресурсима.



Слика 4: Веб-сучеље CQPweb за претрагу Рударског корпуса у режиму „CQP syntax extended (SrpWN)" прихвата захтев за семантичким проширењем упита (синонимима лексеме *трачни транспортер*).

На пример, ако одговор веб-сервиса садржи полилексемску јединицу *геолошки процес*, генерисани упит ([lemma="geološki"] [lemma="proces"])[15] би имао изгледа да успе јер су и *геолошки* и *процес* леме одговарајућих лексема. Међутим, ако би се у одговору веб-сервиса нашла полилексемска јединица *полицијске снаге* (као синоним лексеме *полиција*), генерисани упит [lemma="policijske"] [lemma="snage"]) не би имао никаквих изгледа да нешто пронађе зато што су у питању облици лексема *полицијски*, односно *снага*, који сами не представљају леме тих лексема.

Према томе, систем је потребно даље унапређивати. Најпре је потребно омогућити семантичко проширење упита у односу на остале лексичке релације Српског ворднета (хипонимију, холонимију, меронимију итд.). Даље, неопходно је разрешити управо описани проблем проналажења семантичког проширења за све полилексемске јединице.

Једно од питања које заслужује посебну пажњу јесте да ли треба радити на модификацији веб-сучеља тако да квази-позициони атрибути могу у упитима да се користе са свим могућностима које упитни језик дозвољава правим позиционим атрибутима. Наиме, вредност правог позиционог атри-

[15] Ово је генерисани упит у случају Рударског корпуса чији текстови користе српску латиницу. У случају корпуса СрпКор2013, генерисани упит би користио кôд аурора и имао би форму `([lemma=»geolosxki"] [lemma="proces"])`.

бута у општем случају може бити регуларни израз, док се тренутно вредност квази-позиционих атрибута интерпретира искључиво као конкретна ниска карактера која представља лему. Осим тога, у одељку 2.3 је дат пример у коме се ознаком & за логичку операцију конјукције истовремено намећу услови позиционим атрибутима истог токена (Табела 3, на пример [word="bacili" & tag="N"]). У описаној надоградњи веб-сучеља алатки OCWB и NoSketch Engine тренутно није подржан упит у коме се истовремено појављују квази-позициони атрибут и операција конјункције, попут упита [synlemma="zemlja" & lemma=".+(o|e)"] са потенцијалним значењем „облици синонима лексеме *земља* чија се лема завршава словом *о* или *е*", како би, на пример, били пронађени сви синоними средњег рода (као што су *тло*, *тле*) дате именице женског рода (у овом примеру: *земља*).

Напослетку, треба приметити да кључни фактор у реализацији идеје семантичког проширења упита представљају адекватни лексички ресурси, пре свега, адекватно попуњена лексичка база Српски ворднет, потом систем морфолошких електронских речника (СМР), а у случају РудКор-а важну улогу имају и терминолошки електронски речници специфични за област рударства. За разлику од СМР-а који је у поодмаклој фази развоја[16], Српски ворднет, иако садржи преко 22 хиљаде класа синонима[17] (Станковић и др. 2018: 106), још увек није достигао величину какву, на пример, има Енглески ворднет[18], тако да је допуна Српског ворднета свакако један од приоритета када је у питању унапређивање система за семантичко проширивање упита.

ИЗВОРИ

Корпус 2013: Душко Витас и Милош Утвић, „Корпус савременог српског језика (СрпКор), верзија СрпКор2013", Група за језичке технологије Универзитета у Београду, http://www.korpus.matf.bg.ac.rs/korpus (датум приступа: 30.6.2019).

РудКор 2018: „Рударски корпус, специјализовани корпус из области рударства", Универзитет у Београду, Рударско-геолошки факултет, Група за језичке технологије Универзитета у Београду, http://147.91.181.179/~cqp/cqpweb/ (РудКор/OCWB, датум приступа: 30.6.2019), http://147.91.181.179/~cqp/noske/ (РудКор/NoSketch Engine, датум приступа: 30.6.2019).

SketchEngineDoc 2019: "Documentation for expert users", Sketch Engine Documentation https://www.sketchengine.eu/documentation/ (датум приступа: 30.6.2019).

[16] СМР садржи преко 141 хиљаду монолексичких јединица (лема), односно преко 5 милиона њихових граматичких облика, као и више од 18 хиљада лема као полилексемских јединица (Крстев и др. 2018: 43).

[17] Класа синонима или енгл. synset.

[18] Величина Енглеског ворднета је око 5 пута већа од величине Српског ворднета.

ЛИТЕРАТУРА

Еверт-Харди 2011: Stefan Evert and Andrew Hardie, „Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium", In: *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, University of Birmingham.

Еверт 2019: Stefan Evert and The OCWB Development Team, *CQP Query Language Tutorial,* The IMS Open Corpus Workbench (CWB 3.4.16), May 2019, http://cwb.sourceforge. net/ files/ CQP_ Tutorial. pdf

Крстев и др. 2004: Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas and Ivan Obradović, "Using Textual and Lexical Resources in Developing Serbian Wordnet", *Romanian Journal of Information Science and Technology*, 7/1–2, 147–161.

Крстев 2008: Cvetana Krstev, *Processing of Serbian — Automata, Texts and Electronic Dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.

Крстев и др. 2018: Cvetana Krstev, Ranka Stanković, Duško Vitas, "Knowledge and Rule-Based Diacritic Restoration in Serbian", In: Proceedings of the Third International Conference *Computational Linguistics in Bulgaria (CLIB 2018)*, May 27-29, 2018, Sofia, Bulgaria, ISSN 2367-5675 (on-line), 41–51, http://dcl.bas.bg/clib/wp-content/uploads/2018/06/CLIB_2018_Proceedings_v1_final.pdf

Рихли 2007: Pavel Rychlý, "Manatee/Bonito — A Modular Corpus Manager", In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno : Masaryk University, 65–70.

Станковић 2009: Ранка Станковић, *Модели експанзије упита над текстуелним ресурсима* (необјављена докторска дисертација, Београд: Универзитет у Београду, Математички факултет).

Станковић и др. 2017: Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović, "Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources", In: N. Nguyen, R. Kowalczyk, A. Pinto, J. Cardoso (eds) *Transactions on Computational Collective Intelligence* XXVI, Lecture Notes in Computer Science, 10190, Cham: Springer, 162–185. DOI: 10.1007/978-3-319-59268-8_8, https:/doi.org/10.1007/978-3-319-59268-8_8.

Станковић и др. 2018: Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, Cvetana Krstev, "Resource based WordNet augmentation and enrichment", In: Proceedings of the Third International Conference *Computational Linguistics in Bulgaria (CLIB 2018)*, May 27–29, 2018, Sofia, Bulgaria, ISSN 2367-5675 (on-line), 104–114, http://dcl.bas.bg/clib/wp-content/uploads/2018/06/CLIB_2018_Proceedings_v1_final.pdf

Томашевић и др. 2018: Aleksandra Tomašević, Ranka Stanković, Miloš Utvić, Ivan Obradović and Božo Kolonja, "Managing Mining Project Documentation Using Human Language Technology", *The Electronic Library*, 36/6, 993–1009, DOI: 10.1108/EL-11-2017-0239, https:/doi.org/10.1108/EL-11-2017-0239.

Томашевић 2018: Александра Томашевић, *Развој модела за управљање рударском пројектном документацијом* (необјављена докторска дисертација, Београд: Универзитет у Београду, Рударско-геолошки факултет).

Утвић 2011: Милош Утвић, „Анотација Корпуса савременог српског језика", *ИНФОтека*, 12/2, 39–51.

Утвић 2014а: Miloš Utvić, *Izgradnja referentnog korpusa savremenog srpskog jezika* (neobjavljena doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet). https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download.

Утвић 2014б: Милош В. Утвић, „Листе учестаности Корпуса савременог српског језика", *Научни састанак слависта у Вукове дане*, 43/3, 241–262.

Утвић и др. 2018: Милош В. Утвић, Иван М. Обрадовић, Ранка М. Станковић, Александра Ђ. Томашевић, Биљана Ђ. Лазић „Изградња специјалних корпуса савременог српског језика на примеру корпуса из области рударства", *Научни састанак слависта у Вукове дане*, 47/3, 103–118. DOI: 10.18485/msc.2018.47.3.ch7, https://doi.org/10.18485/msc.2018.47.3.ch7.

Харди 2012: Andrew Hardie, "CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool", *International Journal of Corpus Linguistics*. 17/3, 380–409.

Шмид 1997: Helmut Schmid, „Probabilistic Part-of-Speech Tagging Using Decision Trees", In Jones, D. B. *et al.* (eds.) *New Methods in Language Processing*, Routledge, 154–164.

Шмид 1999: Helmut Schmid, „Improvements in Part-of-Speech Tagging with an Application to German", In: Armstrong, S. *et al.* (eds.) *Natural Language Processing Using Very Large Corpora*, Dordrecht: Springer, 13–25.

Miloš V. Utvić, Ranka M. Stanković, Aleksandra Đ. Tomašević,
Mihailo Đ. Škorić, Biljana Đ. Lazić

THE CORPUS SEARCH BASED ON USAGE OF EXTERNAL LEXICAL RESOURCES THROUGH WEB SERVICES

Summary

This paper explores a hybrid approach to corpus search illustrated by application of corpus tools OCWB and NoSketch Engine to special corpus for mining domain, RudKor, as well as to Corpus of Contemporary Serbian (SrpKor). The discussed approach combines existing functionality of tools OCWB and NoSketch Engine (a search based on embedded linguistic annotation of corpus) with newly implemented search options available in custom-upgrade of web interface. Custom-upgraded web interface of OCWB and NoSketch Engine consults external lexical resources (morphological electronic dictionaries of Serbian and lexical database Serbian Wordnet) through RESTful web services. Consequently, if necessary, user query containing lemma X can be expanded with lemmas such that specified semantic relation (synonymy, antonymy, hypernymy/ hyponymy) holds between each lemma that has been added and the lemma X.