

Изградња речника фудбалске терминологије вођена подацима и OntoLex моделом

УДК 811.163.41'322.2

САЖЕТАК: Рад истражује иновативни приступ развоју речника фудбалске терминологије применом повезаних података (енгл. *Linked Data*). Традиционални приступи изради речника често укључују секвенцијално писање речничких чланака и ручну анализу текстуалних корпуса, што може бити споро и захтевно. Наш приступ користи технике рачунарске лингвистике и аутоматизацију како би се убрзао и побољшао процес креирања речника. Дигитални речник фудбалске терминологије који смо развили намењен је и људима и софтверским апликацијама, пружајући корисницима брз приступ, могућност сталног ажурирања и интеграцију са другим дигиталним ресурсима. Овај рад позива се на српски део српско-шпанског речника фудбалских термина, који се развија у оквиру докторске дисертације Јелене Лазаревић.

КЉУЧНЕ РЕЧИ: речник, фудбал, корпус, повезани подаци, рачунарска лингвистика.

РАД ПРИМЉЕН: 10. јун 2024.

РАД ПРИХВАЋЕН: 27. јун 2024.

Јелена Лазаревић

jelazarevic1@gmail.com

ORCID: 0009-0004-4481-2729

докторанд

Универзитет у Београду

Филолошки факултет

Београд, Србија

Оливера Китановић

olivera.kitanovic@rgf.bg.ac.rs

ORCID: 0000-0002-7571-2729

Универзитет у Београду

Рударско-геолошки факултет

Београд, Србија

1. Увод

Интеграција језичких ресурса једна је од главних тема у савременим истраживањима у вези са језичким технологијама. У порасту је примењивање приступа вођеног подацима (енгл. *data-driven approach*) ради побољшаних могућности аутоматизације обраде података. Приступ вођен подацима је методологија која се ослања на анализу и тумачење података као основе за доношење одлука и решавање

проблема. Уместо интуиције, искуства или претходних теоријских знања, приступ вођен подацима користи конкретне, објективне и мерљиве информације добијене из података. Овај приступ је изузетно популаран у научним дисциплинама, попут лингвистике, медицине, различитих информатичких технологија (Kitanović et al. 2021). У овом раду представит ćemo истраживање могућности убрзања и побољшања традиционалног приступа изради речника кроз пример развоја двојезичног речника фудбалских термина. Традиционални приступ креирања речничких чланака, тачније писања једног по једног, најчешће у алфаветском или азбучном редоследу, ослања се на анализу грађе у виду корпуса или ексцерпираних примера уз коришћење референтних речника за сваки појединачни термин. Помоћу технологија рачунарске лингвистике процес се убрзава и аутоматизује. Наиме, чланци се не обрађују секвенцијално, већ је примењен приступ вођен подацима и паралелном обрадом већег скупа података одједном.

Берг и Оландер (Bergh and Ohlander 2019) уочили су да је у XX веку речи фудбалске терминологије постале присутније, а неки термини су постали уобичајени у говору навијача. Језик фудбала је у сталној промени, прилагођава се догађајима у самој игри и око ње. Због свог значаја и велике медијске присутности, фудбал као „народна игра“ утиче на то да енглески речници опште намене укључују фудбалске термине као део општег језика.

Дигитални речници имају бројне предности у односу на традиционалне, штампане речнике. Фуертес-Оливера и Тарп (Fuentes-Olivera and Tarp 2014) сматрају да је идеално решење за изазове новог доба управо онлајн – дигитално издање речника. Прва и кључна предност речника доступног на интернету је брзина претраге речи или фраза, будући да се у штампаним речницима странице морају ручно прелиставати. Друга важна предност јесте могућност сталног ажурирања, насупротив штампаним речницима који могу застарети током времена. Дигитални речници омогућавају прилагођавање врсте фонта, величине слова и других параметара како би задовољили потребе корисника, а могу садржати и додатне информације, попут детаљно објашњене етимологије и изговора речи. Доступност путем мобилних апликација или рачунарских платформи омогућава корисницима приступ речнику било када и било где. Интерактивни елементи као што су хиперлинкови и мултимедијални садржаји чине искуство учења језика динамичнијим. Такође, дигитални речници често садрже директну претрагу унутар текста речничког чланка, што је корисно за

брзо проналажење значења или информација о одређеној речи док се чита неки текст.

Интегрисање речника у дигиталне платформе за учење језика омогућава праћење напредовања у знању језика кроз прављење личних листа речи и чувања бележака, као креирање картица за подсећање и других алата којима се олакшава учење језика. Интегрисање речника са корпусима текстова је веома значајно и стога што корисници могу да уче из конкретних примера њихове употребе – било да су у питању појединачне речи, фразе или конструкције.

Насупрот светској пракси, у Србији штампани речници и даље имају већи степен употребе односу на дигиталне, услед осећаја аутентичности искуства или навика употребе. У савременом добу, дигитални речници постали су неопходан алат за оне који уче и користе стране језике. Не треба заборавити да је из дигиталног облика речника могуће припремити за штампу више различитих облика и обима речника, у различитим форматима, са разним нивоима детаљности и већим бројем речничких чланака.

Речник фудбалске терминологије који представљамо у раду је дигитални речник намењен и људима и машинама; тачније, софтверским апликацијама и део је ширег истраживања и изградње првог дигиталног фудбалског речника на српском језику, са преводима на шпански језик. У контексту истраживања насталог као део докторске дисертације Јелене Лазаревић под називом: „Језичке одлике дискурса нових медија о фудбалу: контрастивна анализа на корпусу српског и шпанског језика.“ Нагласак је на спрези техника рачунарске лингвистике, као и примени технологија семантичке мреже, како би се задовољиле потребе појединца и корисника, али и машине, односно софтверских алата као алтернативних корисника речника.

Речник је намењен спортским радницима: спортистима, тренерима, судијама, професорима, студентима, новинарима, као и свима онима који желе да овладају прецизнијом интерпретацијом фудбалских термина. Овај рад се посебно фокусира на српски део двојезичког српско-шпанског речника фудбалских термина.

У одељку 2. овог рада биће сагледана два начина издвајања информација потребних за изградњу речника: аутоматска ексцерпција из корпуса *срФудКо* кандидата за термине, фреквенција и примера употребе (Krstev et al. 2015; Ivanović et al. 2022) и екстракција информација из постојећих речника и глосара доступних онлајн (Kitanović et al. 2021).

2. Припрема података за изградњу речника

Корпус *срФудКо* (Lazarević et al. 2023) изграђен је од новинских чланака о фудбалу на српском језику који су прикупљени са пет српских дигиталних новинских сајтова: *Б92*, *Блиц*, *Мондо*, *Политика* и *Спорт Клуб*. Чланци су аутоматски преузети коришћењем различитих техника за прикупљање података са интернета, након чега је уследила дедупликација текстова, елиминисање кратких чланака (краћих од 3.000 карактера), реченица на другим језицима и табела које садрже само нумеричке резултате. Након филтрирања чланака и чишћења текста, креиран је корпус *срФудКо* који садржи 10.100.553 токена, од којих су 8.618.426 речи, док остатак чине знакови интерпункције. Садржај корпуса је аотиран врстама речи и лематизован коришћењем тагера за српски језик *SrpKor4 Tagging-TreeTagger* (Stanković et al. 2020; Stanković, Škorić, and Šandrih Todorović 2022).

Корпус *срФудКо* је затим коришћен за ексерпцију информација. Први резултат аутоматске ексерпције из корпуса је листа кандидата речи, праћена фреквенцијама појављивања у корпусу. Затим су издвојени илустративни примери употребе појединих термина. У првој фази су екстраховани, евалуирани и обележени једночлани термини, а уследила је примењена екстракција вишечланих термина (Krstev et al. 2015) најчешћих синтаксичких образаца, заснована на морфолошким речницима српског језика и локалним граматика (Krstev 2008). Осим фреквенција, за једночлане термине је рачуната и кључност¹ (енгл. *keyness*) термина, где су поређене фреквенције у корпусу са фреквенцијама у корпусу савременог српског језика *СрпKop2013* (Утвић 2011). Када су у питању вишечлани термини коришћене су комплексније асоцијативне мере *T-Score*, *CValue*, и друге (Vu, Aw, and Zhang 2008). Детаљи ове фазе развоја речника су описани у раду: *Football terminology: compilation and transformation into OntoLex-Lemon resource* (Lazarević et al. 2023).

Када је реч о екстракцији информација из постојећих речника и глосара доступних на интернету, преузети су речници различитог типа и нивоа детаљности. Уместо да се за сваки чланак појединачно користи речник, планирано је укрштање, усклађивање и обрада свих извора како би се олакшала и убрзала изградња савременог фудбалског речника богатог информацијама.

Интеграција података се ослањала на више извора, али главну улогу су имали: 1) четворојезични речник, српско-енглеско-

1. <https://www.sketchengine.eu/documentation/simple-maths/>

француско-шпанска листа термина (Mihajlović 2003) и 2) *Kicktionary*² (Schmidt 2009) заснован на концепту семантике оквира по узору на *FrameNet*³ (Fillmore et al. 2003). Семантика оквира је теорија значења која се фокусира на концептуалне структуре (оквири) које представљају позадину потребну за разумевање речи и израза у различитим контекстима. Ови оквири укључују повезане елементе и односе који се активирају приликом употребе неке речи или израза, омогућавајући боље разумевање њиховог значења и контекста. Оквири и лексичке јединице на енглеском су укрштене са енглеском страном четворојезичног речника како би се обогатила српска страна речника семантичким оквирима.

С обзиром на чињеницу да ниједан српски речник фудбалске терминологије нема описне дефиниције појмова, а имајући у виду напредак машинског превођења, зарад брже израде речника, који је предмет истраживања, користили смо глосаре фудбалске терминологије на енглеском и шпанском језику и то полуаутоматском обрадом. Осим речника *Kicktionary*, који такође садржи одређене дефиниције, коришћени су и речници доступни у тренутку обраде. Неопходно је напоменути да су неки од речника у међувремену постали недоступни на раније познатим адресама:

- Глосар *Field*⁴ са 900 најчешће коришћених термина организован је кроз илустроване сценарије. Изворни језик је португалски а садржи еквиваленте на енглеском и шпанском, заједно са примерима употребе. Поглед на речник омогућен је кроз листу фудбалских термина и кроз сценарије. Хомонимија и полисемија је имплементирана кроз одвојене речничке чланке и генерално је структура инспирисана концептима семантике оквира. Анализа полисемичних речи и колокација у корпусу укључила је, поред осталих фаза, компилацију трију упоредивих корпуса за шпански, португалски и енглески.

2. <http://www.kicktionary.de/> вишејезички лексикон који обухвата појмове и њихове дефиниције везане за фудбалске утакмице, тактике, играче, и опрему, са циљем да олакша учење и разумевање фудбалске терминологије кроз визуелизације и објашњења.

3. *FrameNet* је лексичка база података која документује семантичке оквири и како они повезују речи са њиховим значењима у различитим контекстима.

4. раније доступан на <http://dicionariofield.com.br/>, видети <https://www.facebook.com/dicionariofield>

- Фудбалски глосар *UEFA*⁵ који смо користили има укупно 8.086 речничких чланака, од тога на немачком 1.893, на енглеском 2.306 и на шпанском 1.887.
- Фудбалски глосар *UsingEnglish*⁶ као део речника који садржи различите домене има свега 74 речничка чланка за фреквентне одреднице.
- Деветојезични речник⁷ упарених фудбалских термина има 77 термина са преводним еквивалентима на енглеском, немачком, француском, шпанском, италијанском, португалском, пољском, руском и украјинском.

Речник *ФудЛе* настао на основу докторске дисертације Јелене Лазаревић садржи сада 2.648 српско-шпанских преводних еквивалената којима су придружене врсте речи и фреквенције у корпусу *срФудКо*, при чему 93 речничка чланка имају и придружену дефиницију. У српском делу речника једночланих термина има 990, двочланих има 966, са три компоненте 369, са четири 200, а са пет и више има 123. Повезивање са семантичким оквирима речника *Kicktionary* је урађено за 142 лексичке јединице. Након екстракције терминолошких фраза из корпуса *срФудКо*, ручне евалуације и обележавања доменских маркера: да ли је у питању спортски термин или је у питању термин карактеристичан за фудбалски домен, добијена је листа од 1.943 термина којима су придружени код синтаксичке групе, фреквенција и мере асоцијације.

3. Компилација и трансформација речника у OntoLex модел

Употреба модела лексикона за онтологије *OntoLex-Lemon*,⁸ краће *OntoLex-a* (*LExicon Model for ONtologies*) (McCrae et al. 2017), у порасту је када је у питању имплементација лексичких ресурса као скупа повезаних података на интернету. Одреднице, без обира на то да ли се односе на једночлане или вишечлане термине из домена фудбала, заједно са резултатима екстракције из корпуса *срФудКо* (фреквенције и

5. раније доступан на <https://www.uefa.com/insideuefa/dictionary/>

6. <https://www.usingenglish.com/glossary/football-vocabulary/>

7. https://www.uefa.com/MultimediaFiles/Download/competitions/General/01/80/75/52/1807552_DOWNLOAD.pdf

8. <https://www.w3.org/2016/05/ontolex/>, <https://jograncia.github.io/ontolex-lexicog/>

примери), обogaћени информацијама преузетим и обрађеним из речника и глосара, представљени су коришћењем модела *OntoLex*.

Поред превода, речнички чланак указује на врсту речи појединачне одреднице, и потенцијално граматичка својства. Додатне информације о термину у виду дефиниције значења, везе ка концептима у другим ресурсима, попут база знања, онтологија или лексичких база значајне су кориснику ради бољег разумевања конкретног термина. Контекст употребе кроз изабране примере екстраховане из корпуса уз фреквенције појављивања у корпусу треба да пружи додатне информације и употпуну терминолошку слику.

Треба напоменути да су прве употребе *Lemon*, а потом и *OntoLex-Lemon* модела за српски језик повезане са прављењем лексичке базе Лексимирка⁹ (Stanković et al. 2018) и превођење система електронских речника за српски (Krstev 2008) у релациону базу (Пујевић 2022). У раду се користи модел *OntoLex* за моделирање RDF¹⁰ (енгл. *Resource Description Framework*), односно представљања лингвистичких описа лексичких јединица из генерисаног речника уз проширење лексичког слоја, како би се обезбедило њихово повезивање и машинска разумљивост на интернету. *OntoLex* модули су уведени кроз примере имплементације, почевши са основним (*OntoLex*) модулом који дефинише лексичке јединице, облике речи и лексичка значења. Модул *vartrans*¹¹ има две важне класе: *Translation* и *TranslationSet*, где превод представља везу између двају лексичких значења, повезаних са лексичким јединицама на различитим језицима. *LexInfo*¹² онтологија типова, вредности и својстава користи се као каталог категорија података (нпр. за означавање граматичког рода, броја, врсте речи итд.).

Категорије превода су представљене упућивањем на екстерни каталог *OEG Translation Categories*.¹³ Други често коришћени речници, као што је *Dublin Core*, *DCMI Metadata Terms*,¹⁴ користе се за метаподатке о изворима, ауторству, верзији или лиценци. За моделирање следиле су се препоруке и приступи дати у смерницама *Guidelines for*

9. <https://leximirka.jerteh.rs/>

10. <https://www.w3.org/RDF/>

11. <https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans>

12. <https://lexinfo.net/>, <https://github.com/ontolex/lexinfo>

13. <http://purl.org/net/translation-categories>

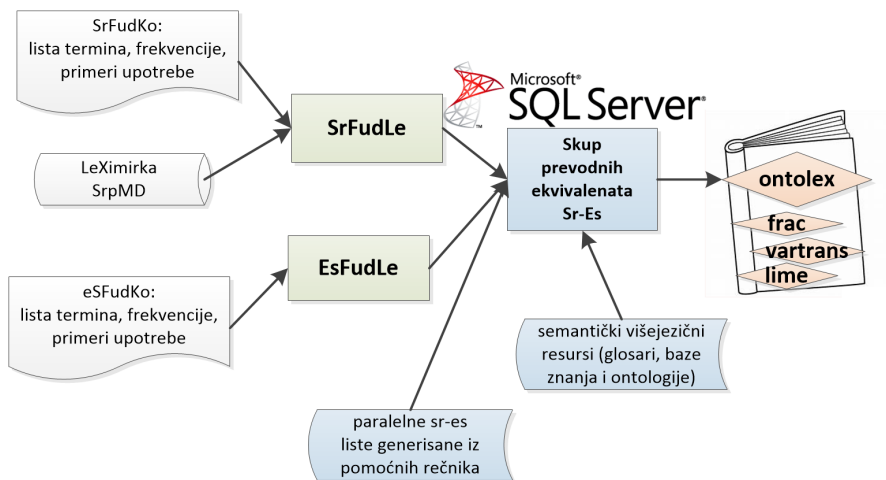
14. <http://purl.org/dc/elements/1.1/>

*Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0)*¹⁵ (Río Gracia et al. 2023), и низу смерница „Guidelines and best practices for LLOD“.

Према наведеним препорукама, двојезични речник се организује тако да су појединачни језици и одвојени RDF графови, који се потом повезују додатним графом који садржи везе између преводних еквивалената. Генерисани речник се састоји из четири графа, односно скупа података:

- Изворни лексикон, у овом случају српски,
- Циљни лексикони, у овом случају шпански,
- Скуп преводних еквивалената (енгл. *TranslationSet*).

Приступ коришћења ресурса припремљених у претходним фазама (леви и доњи део слике) и модели коришћени за трансформисање у RDF облик (десни део слике) илустровани на слици 1, тако да се на излазу добија српско-шпански двојезични речник *ФудЛе* чији су речнички чланци обogaћени бројним информацијама.



Слика 1. Графички приказ интеграције језичких података из различитих ресурса у речник *ФудЛе*.

15. https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

OntoLex-FrAC модул (Chiarcos et al. 2020) коришћен је да у речник фудбалске терминологије укључи информације о фреквенцијама појављивања у корпусима, примере употребе термина у контексту и колокацијама екстрахованим из корпуса *срФудКо*. Аутоматски екстраховани термини, који су затим ручно евалуирани и обележени етикетама за домен спорт или фудбал, допуњени су флективним облицима који се даље трансформишу у модел *OntoLex*. Даље су термини повезани са примерима из корпуса који су одабрани коришћењем GDEX¹⁶ (Good Dictionary Examples) алгоритама (Kilgarrieff et al. 2008).

Морфолошки речник вишечланих термина је направљен коришћењем алата LeXimir (Stanković et al. 2011), а затим је трансформисан апликацијом која прати *OntoLex* спецификације и објављене примере (Chiarcos, Fäth, and Ionov 2022). Граматичке информације, морфосинтаксичка својства облика речи дате су у складу са LexInfo вокабуларом. У наредним одељцима представљамо употребу основног модула модела *OntoLex* и модула за фреквенције *OntoLex-FrAC*, примере употребе и остале информације засноване на корпусу (Chiarcos et al. 2022).

3.1 Језгро речника *ФудЛе*

Корпус *срФудКо* је латинични, стога је и речник *ФудЛе* написан латиничним писмом. Пример термина: „фудбалска утакмица“ је приказан у раду (Lazarević et al. 2023), док у овом раду дајемо пример приказа термина „жути картон“ у моделу *OntoLex*, који је проширен додатним информацијама о значењу и релацијама ка другим ресурсима и језицима. Жути картон се дефинише као упозорење које судија упућује играчу због неспортског понашања. Судија жутим картоном санкционише играча за фаул у случају када понашање играча не приказује тежину прекршаја који заслужује црвени картон, као што је туча или фаул са намером да повреди противника. Овај термин постоји и у српском морфолошком речнику СрпMD сложених речи (Krstev and Vitas 2009) у облику DELAC записа (Savary, Krstev, and Vitas 2007), а његов изворни запис је:

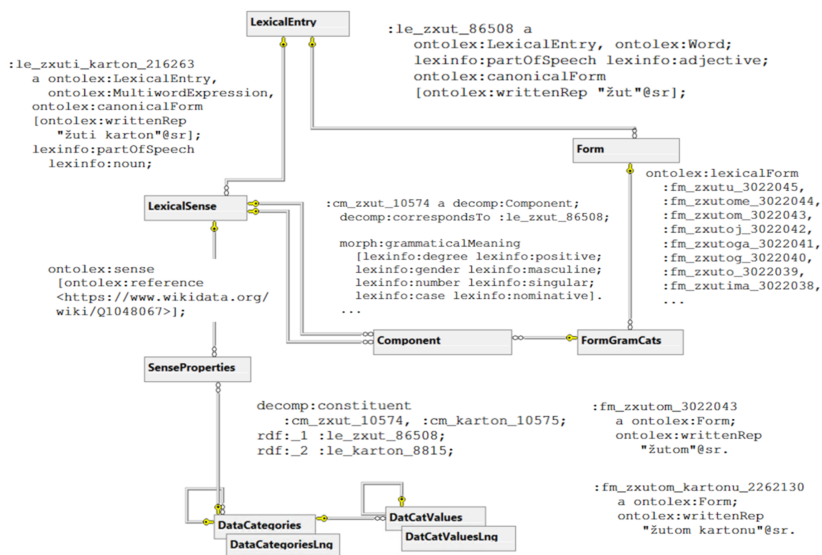
žuti(žut.A8:adms1g) karton(karton.N1:ms1g),
NC_AXN+DOM=Sport+Comp+Conc

16. <https://www.sketchengine.eu/documentation/manual-for-gdex/>

Коначни трансдуктор NC_AXN генерише флективне облике за морфолошке електронске речнике сложених речи, где NC означава именичку колокацију, а AXN приказује колокацију облика придев-именица, где се придев слаже са именицом по свом граматичком броју, роду, падежу и аниматности. За компоненте које трансдуктор укључује, захтевају се информације о леми (жути и картон), трансдуктору (A8 и N1) за појединачне компонентне колокације и вредности за граматичка својства (adms1g и ms1q). Граматичка својства су следећа: a - позитиван степен, d - одређени (дуги) облик, m - граматички мушки род, s - једнина, l - номинатив, g - аниматност није од значаја, q - неживо. Већина граматичких својстава се лако пресликава у онтологију LexInfo, али опција када аниматност није од значаја није предвиђена. Имајући у виду да се термин: „жути картон“ може наћи у неколико терминолошких речника, коришћењем модела *OntoLex*, овој лексикализованом колокацији додељујемо и њено специфично значење.

Део модела података базе података система *LeXimirka*, имплементиране у MS SQL Серверу (Stanković et al. 2018; Рујевић 2022), приказан је на слици 2 која приказује табеле за лексичке јединице (*LexicalEntry*) и флективне облике (*Form*), као и компоненте за вишечлане речи (*Component*). Једна од улога система *LeXimirka* је генерисање флективних облика и њихових граматичких информација, за једночлане и вишечлане термине речника *ФудЛе*. Граматичке информације су повезане са флективним облицима кроз категорије података (*DataCategories*) и њихове вредности (*DataValues*). Систем је опремљен метаподацима који се односе на повезане информације између категорија података у српским морфолошким речницима и речнику *LexInfo*. Једна лексичка јединица може имати више значења, сачувана у табели *LexicalSense*, која појединачне категорије прикупља преко функције *SenseProperties* из каталога *DataCategories*.

Запис на слици 2 приказује колокацију „жути картон“, где се назив лексичке јединице *le_zxuti_karton_216263* састоји од префикса *le* који потиче од *LexicalEntry*, самог термина, односно колокације: „жути картон“ и примарног кључа 216263 у табели *LexicalEntry* базе података *LeXimirka*. Када су у питању компоненте, слично се бира префикс *sm* да обележи записе који долазе из табеле *Component*, док префикс *fm* означава записе из табеле *Form*. Напоменимо да се за поједина слова мењају према кодирању Аурора (Vitas 1979) који је осмислио проф. др Душко Витас како би се обезбедило једнозначно ASCII кодирање које је



Слика 2. Дијаграм базе података у MS SQL Серверу са повезаним табелама и примером експортираних података у *OntoLex*.

погодно за називе објеката скупа података, док се литерали, тј. писани облици, кодирају коришћењем кодне шеме UTF-8.

За лексичку јединицу `:le_zxuti_karton_216263` видимо да припада класи *ontolex:LexicalEntry* и да је у питању вишечлани термин (*ontolex:MultiwordExpression*), коју наводимо у канонском облику, односно као лему (*ontolex:canonicalForm*). Затим се наводи писани облик леме, уз податак да је у питању именица. Када је значење у питању, упућује се на два екстерна ресурса. Први је база знања Википодаци (Wikidata) која има одговарајући ставку „жути картон“ чији је код Q1048067 и целокупна URL адреса: <https://www.wikidata.org/wiki/Q1048067>. Има преводне еквиваленте на 38 језика, па осим лексикализације у српском, такође има и шпански превод: “tarjeta amarilla”.

```

:le_zxuti_karton_216263 a ontolex:LexicalEntry,
  ontolex:MultiwordExpression, ontolex:canonicalForm
  [ontolex:writtenRep "žuti karton"@sr];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>.
  
```

```

decomp:constituent :cm_zxuti_10574, :cm_karton_10575;
rdf:_1 :le_zxut_86508;
rdf:_2 :le_karton_8815.

# komponente kanonskog oblika
:cm_zxuti_10574 a decomp:Component;
decomp:correspondsTo :le_zxut_86508;

morph:grammaticalMeaning [
  lexinfo:degree lexinfo:positive;
  lexinfo:gender lexinfo:masculine;
  lexinfo:number lexinfo:singular;
  lexinfo:case lexinfo:nominative].
...

```

Припадајући део на шпанском језику можемо видети у наставку:

```

:le_tarjeta_amarilla_10001 a ontolex:LexicalEntry,
  ontolex:MultiwordExpression, ontolex:canonicalForm
[ontolex:writtenRep "tarjeta amarilla"@es];
lexinfo:partOfSpeech lexinfo:noun;
ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>;

decomp:constituent :cm_tarjeta_1, :cm_amarilla_2;
rdf:_1 :le_tarjeta_1;
rdf:_2 :le_amarillo_2.

```

3.2 Дефинисање значења у речнику *ФудЛе*

Класа *ontolex:LexicalSense* дефинише лексичко значење лексичке јединице када се тумачи као да се односи на одговарајући онтолошки елемент. Лексичко значење представља реификацију или опредмећење упаривања лексичке јединице и онтолошке целине на коју се односи. Веза између лексичке јединице и онтолошког ентитета преко објекта *ontolex:LexicalSense* имплицира да се лексичка јединица може користити за упућивање на задати онтолошки ентитет.

Идеја прецизнијег повезивања лексикона *FrameNet* и *OntoLex* модела приказана је у (Robin, Kulkarni, and Buitelaar 2023). Код дефинисања лексичких јединица показано је да се могу користити за означавање одређеног онтолошког предиката, коришћењем својства: *ontolex:denotes*. Дакле, лексичка јединица означава предметну класу или онтолошки елемент. Уз то је могуће представити евокацију одређеног менталног концепта од задате лексичке јединице, која се не односи на класу са формалном интерпретацијом у неком моделу. Класа лексички концепт (*ontolex:LexicalConcept*) представља менталну апстракцију, концепт или јединицу мишљења која се може лексикализовати, тако да класа *ontolex:LexicalConcept* представља поткласу елемента *skos:Concept*.

Дефиниције се додају лексичком концепту као глоса и то коришћењем својства *skos:definition*.

Истражујући изградњу семантичког речника фудбала, ослањајући се на корпус из фудбалског домена, Wu i Li (Wu and Li 2017) користили су Word2Vec (Mikolov et al. 2013) ради обучавања векторског модела за речи у корпусу. Комбинујући статистички алгоритам TF-IDF и векторе речи, одређивали су концепте карактеристичне за фудбалски домен. За листу термина из области фудбала израчунали су растојања од речи у векторском моделу и издвојили три најближе речи, то јест, термина са вишом вредношћу косинуса као синонине за улазну реч за изградњу семантичког речника. Показали су да постојећи семантички речници, као што је WordNet, не одговарају потребама појединих домена зато што је семантичка грануларност речи мања. Дакле, потребно је направити семантички речник који за један појам повезује синонине, као и подређене и надређене појмове и дефиниције на различитим језицима. Тиме се добија семантички речник којим се имплементира семантичко претраживање текстова из фудбалског домена.

Креирали смо сопствени идентификатор концепта на основу прикупљених ресурса. Речник *UEFA*, са терминима и дефиницијама је узет као узор за дефинисање концепата. Конкретно, пример је *:ct_yellow_card_1*, где је *ct* скраћено од “concept”, потом је узет енглески термин и додат идентификатор (ради једнозначности) и уместо размака се користи доња црта “_”.

```
:conceptLexicon a ontolex:ConceptSet .
:le_zxuti_karton_216263 ontolex:evokes :ct_yellow_card_1.
:le_tarjeta_amarilla_10001 ontolex:evokes :ct_yellow_card_1.

:ct_yellow_card_1 a ontolex:LexicalConcept ;
  ontolex:LexicalisedSense :sn_zxuti_karton_216263, :sn_tarjeta_amarilla_10001;
  ontolex:isConceptOf <https://www.wikidata.org/wiki/Q1048067> ;
  skos:definition "Disciplinska sankcija koju sudija izriče tokom utakmice
    protiv igrača zbog kršenja Pravila igre, posebno zbog lošeg ponašanja u
    sportu ili odlaganja ponovnog početka igre."@sr;
  skos:definición "Sanción disciplinaria impuesta durante un partido por el
    árbitro a un jugador por infracción de las Reglas de Juego, en particular
    por mala conducta deportiva o por retrasar la reanudación del juego."@es;

ontolex:isEvokedBy :le_zxuti_karton_216263, sn_tarjeta_amarilla_10001;
skos:inScheme :FudLe .
```

Имајући у виду да се за *ontolex:LexicalSense* може узети глоса која представља специфичности коришћења употребом својства *ontolex:usage*. Наиме, интерпретација лексичке јединице у односу на значење дефинисано у датој онтологији је често прилагођено условима

употребе или прагматичним импликацијама, посебно због регистра, конотација или варијација у значењу. Ове нијансе значења могу се спецификовати коришћењем својства, што омогућава прикупљање информација о условима употребе и прагматичким импликацијама, под којима се лексички унос упућује на онтолошко значење.

Поменути услови употребе не уводе се уместо формално дефинисаног значења, већ допуњују одговарајуће значење додатним информацијама које описују употребу лексичке јединице. У случају речника *ФудЛе* додатне дефиниције прикупљене из различитих извора, екстраховане из корпуса или директно написане, уписиване су на овај начин.

```
:sn_zxuti_karton_216263 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html>;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_zxuti_karton_216263 ;
  ontolex:usage [ rdf:value "Žuti karton koji sudija vadi iz svog džepa da bi pokazao
    igraču koji je uradio nešto prilično loše, kao što je loš faul ili igranje rukom,
    ali ne tako loš kao prekršaj crvenog kartona kao što je tuča. Žuti karton često
    ide zajedno sa drugom kaznom kao što je slobodan udarac. Dva žuta kartona na jednom
    meču znače crveni karton, a dva žuta u određenom periodu suspenziju."@sr ] .

:sn_tarjeta_amarilla_10001 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html> ;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_tarjeta_amarilla_10001 ;
  ontolex:usage [ rdf:value "La tarjeta amarilla se muestra a un jugador por acciones
    como faltas fuertes o tocar el balón con la mano. Puede ir acompañada de sanciones
    como tiros libres. Dos tarjetas amarillas en un partido significan una roja, y dos
    en un período llevan a suspensión."@es ] .
```

3.3 Имплементација фреквенција и примера у речнику *ФудЛе*

Документација за модул *OntoLex* и информације о фреквенцијама, примерима употребе у контексту и другим корпусним информацијама “Frequency, Attestations and Corpus data” (FrAC) налази се на [линку](#) (Chiarcos et al. 2022; Chiarcos et al. 2020). Модул FrAC је усмерен на допуну речника и других лингвистичких ресурса који садрже лексикографске податке моделом за репрезентацију статистике изведене из корпуса: информације о учесталости и појављивању, колокације, показивачима из лексичких ресурса на корпусе и друге текстуалне колекције: потврде употребе у корпусној грађи, обележавање корпуса и других језичких извора лексичким информацијама. лематизацију према речнику, и дистрибутивну семантику, односно векторе колокације, ембединге речи, значења и појмова.

Модул се бави случајевима коришћења у лексикографији заснованој на корпусу, корпусној лингвистици и обради природног језика, у комбинацији са основним и другим *OntoLex* модулима.

Помоћна класа *:SrFudKo* дефинисана је да обезбеди лакше руковање и краћу анотацију. Упућивање на фреквенције су коришћене за корпус *срФудКо*, објављен на инстанци “noSketch” (Kilgarriff et al. 2014) коју одржава Друштво за језичке ресурсе и технологије JePTeX.¹⁷

Уведена је помоћна класа *:SrFudKo_lemma_freq* за фреквенцију свих облика конкретног термина, уз напомену да се термин може састојати од једне или више компоненти. Мотив за увђење наведене класе је компактније кодирање јер се смањује понављање сегмената кода. Поређење фреквенција са референтним општим корпусом српског језика СрпКор2021 (Krstev and Stanković 2023; Škorić and Janković 2024) обезбеђено је кроз евидентирање фреквенција у том корпусу коришћењем објекта *:SrpKor2021*.

```
# korpus fudbala - srpski
:SrFudKo a owl:Class;
  rdfs:subClassOf [a owl:Restriction;
    owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrFudKo>] .
:SrFudKo_lemma_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
  [a owl:Restriction; owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# opšti korpus savremenog srpskog jezika
:SrpKor2021 a owl:Class;
  rdfs:subClassOf [a owl:Restriction; owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrpKor2021>] .
:SrpKor2021_lemma_freq rdfs:subClassOf
  frac:Frequency, :SrpKor2021, [a owl:Restriction;
    owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# korpus fudbala - španski
:EsFudKo a owl:Class;
  rdfs:subClassOf [a owl:Restriction;
    owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=EsFudKo>] .
:EsFudKo_lemma_freq rdfs:subClassOf
  frac:Frequency, :EsFudKo, [a owl:Restriction;
    owl:onProperty dct:description; owl:hasValue "lemma frequency"].
```

У наставку се илуструју фреквенције лема у корпусу фудбала и општег језика за српски језик. Напоменимо да је један од фактора одређивања термина (енгл. *termhood*) управо претходно поменути појам кључности (енгл. *keyness*) што се рачуна на основу фреквенција у ова два корпуса. Дата су два могућа начина записивања фреквенција.

17. JePTeX

```
# frekvencije na nivou leme (prvi način)
:le_zxut_86508 frac:frequency
  [a :SrFudKo_token_freq; rdf:value "2313" ] .
:le_zxut_86508 frac:frequency
  [a :SrpKor2021_token_freq; rdf:value "23377" ] .

# frekvencije na nivou leme (drugi način)
:le_zxut_86508 frac:frequency
  :freq_SrFudKo_le_zxut_86508,
  :freq_SrpKor2021_le_zxut_86508.
:freq_SrFudKo_le_zxut_86508
  a :SrFudKo_token_freq; rdf:value "2313".
:freq_SrpKor2021_le_zxut_86508
  a :SrpKor2021_token_freq;
  rdf:value "23377".

# frekvencije na nivou leme za višechlanu reč
:le_zxuti_karton_216263 frac:frequency
  :freq_SrFudKo_le_zxuti_karton_216263,
  :freq_SrpKor2021_le_zxuti_karton_216263.
```

За шпански језик дајемо примере фреквенција на корпусу *esФудКо*. За CQL (Corpus Query Language) упит [lemma = "amarillo"] добија се 1.517 појављивања, при чему се следеће фреквенције јављају по појединачним облицима: *amarilla* (839), *amarillas* (244), *amarillo* (221), *amarillos* (85), *Amarillas* (55), *Amarilla* (41), *AMARILLAS* (21), *Amarillo* (10) и *AMARILLA* (1). У речнику *ФудЛе* приказују се обједињено фреквенције, без обзира на мала и велика слова, дакле *amarillas* ће имати $244+55+21=320$

```
# frekvencije na nivou leme (prvi način) španski
:le_amarillo_2 frac:frequency
  [a :EsFudKo_token_freq; rdf:value "1517" ] .

# frekvencije na nivou leme (drugi način)
:le_amarillo_2 frac:frequency :freq_ESFudKo_le_amarillo_2.
:freq_ErFudKo_le_amarillo_2
  a :EsFudKo_token_freq;
  rdf:value "1517".
```

Апсолутне фреквенције добијају се коришћењем CQL израза, док се релативне фреквенције рачунају на милион речи – дељењем са укупним бројем речи у корпусу и множе са милион. Када су у питању фреквенције вишечланих елемената у српском, односно шпанском језику, једно од могућих решења својстава је следеће (Lazarević et al. 2023):

```
:SrFudKo_mwe_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
  [owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "mwe frequency"].
```

```
:EsFudKo_mwe_freq rdfs:subClassOf
  frac:Frequency, :EsFudKo,
  [owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "mwe frequency"].
```

У наставку су дати примери за фреквенције вишечлане леме „жути картон“. На примеру корпуса *срФудКо* дато је појашњење и упити којима се долази до фреквенција.

Фреквенције лема се добијају упитом „[lemma="žut"]|[lemma="karton"]“, где осим претходно наведених облика имамо и: „žuti kartoni“ (124), „žuta kartona“ (95), „žutih kartona“, „žutog kartona“ (84), „žute kartone“ (62), „žutim kartonom“ (51), „žuti kartoni“ (15),...

```
# mwe frekvencija lema
:le_zxuti_karton_216263
  frac:frequency [a :SrFudKo_mwe_freq; rdf:value "1996"];
  frac:frequency [a :SrpKor2021_mwe_freq; rdf:value "3192"];
  frac:head :le_karton_8815 .
```

На шпанском језику дајемо пример за претрагу по лемима CQL упитом [lemma = "tarjeta"]|[lemma = "amarillo"] који даје 312 погодака: *tarjeta amarilla* (214), *tarjetas amarillas* (75), *Tarjetas amarillas* (12), *Tarjeta amarilla* (10), *tarjeta amarillas* (1). Када дајемо фреквенције рачунамо неосетљивост на мала и велика слова.

```
# mwe lemma frequency za primer na španskom
:fm_tarjeta_amarilla_10001
  frac:frequency [a :EsFudKo_mwe_freq; rdf:value "224"];
  frac:head :le_tarjeta_1 .
:le_tarjetas_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq; rdf:value "87"];
  frac:head :le_tarjeta_1 .
:le_tarjeta_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq; rdf:value "1"];
  frac:head :le_tarjeta_1 .
```

Треба напоменути да је могуће увођење појединачног објекта за фреквенције, са којим се затим повезују подаци:

```
# frekvencije za španski
:le_tarjeta_amarilla_10001 frac:frequency :freq_le_tarjeta_amarilla_10001.
```

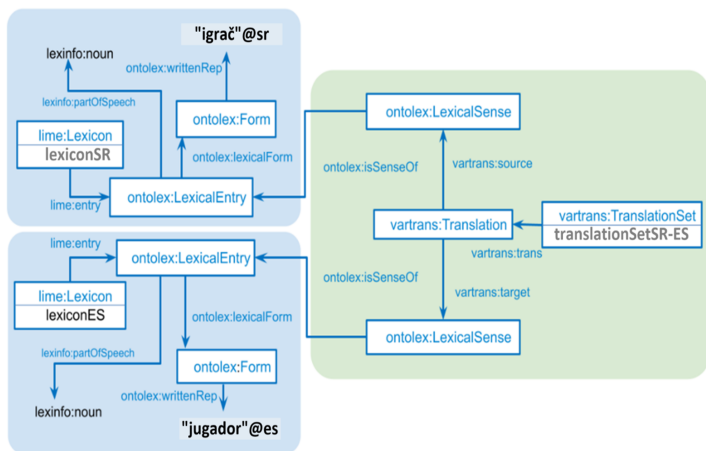
4. Преводни еквиваленти у речнику *ФудЛе*

Као што је претходно поменуто, Речник фудбалске терминологије *ФудЛе* састоји се од више различитих компонената, међу којима овде

представљамо: изворни, српски део лексикона, циљни, шпански део и скуп превода (енгл. *TranslationSet*).

Оваква подела се природније уклапа у основну шему *OntoLex* и модул *vartrans*. Као резултат, два независна једнојезична лексикона на српском и на шпанском језику објављују се као повезани скупови података, заједно са скупом превода који их повезује.

Објављивање речника и за друге језике је могућ по истој шеми. Слика 3 илуструје шему представљања која се користи за превод лексичке јединице „играч“ са српског на шпански у „jugador“, у контексту основног простора имена речника *ontoLex* и *vartrans*.



Слика 3. Моделирање превода коришћењем *OntoLex* и *vartrans*.

Лексичка јединица (*ontolex:LexicalEntry*) и њена придружена својства се користе за додавање лексичких информација, док их *vartrans:Translation* класа повезује путем лексичког значења *ontolex:LexicalSense*. На пример: својство писана форма (*ontolex:writtenRep*) повезује субјекат класе облик (*ontolex:Form*) и објекат „играч“ на српском језику или „jugador“ на шпанском.

Постоје и једноставније могућности када се лексичке јединице повезују, без дефинисања посредничког значења. Тада се користи предикат *vartrans:translatableAs*. Међутим, преводни еквиваленти се

успостављају између специфичних значења речи и коришћење класе *vartrans:Translation*, као повезнице између два значења лексичких јединица преко својства *ontolex:LexicalSense*, омогућава да јасно представимо ову чињеницу.

За публикување речника одабран је следећи образац URI идентификатора: `http://{domain}/{type}/{concept}/{reference}`, где `{type}` декларише тип ресурса који се идентификује. На пример: `'id'` или `'item'` за објекте из реалног света; `'doc'` за документе који описују те објекте; `'def'` за концепте; `'set'` за скупове података; или ниска специфична за контекст, као што је `'authority'` или `'dc-terms'` (Archer, Goedertier, and Loutas 2012). Сличан приступ је коришћен за публикување Српско-немачког речника *SrpNemLex* (An-donovski 2023).

У случају *ФудЛе* су пројектована три скупа података:

- Скуп података изворног језика (српски):
<https://lloj.jerteh.rs/id/fudle/lexiconSR>
- Скуп података циљног језика (шпански):
<https://lloj.jerteh.rs/id/fudle/lexiconES>
- Скуп података преводних еквивалената (српско-шпански преводни):
<https://lloj.jerteh.rs/id/fudle/tranSetSR-ES>

Следи пример за превод одредница `"играч"@sr` и `"jugador"@es` где су имена објеката прилагођена лакшем разумевању креираних веза између преводних еквивалената на нивоу значења речи.

```
# :lexiconSR a lime:Lexicon .
...
:lexiconSR lime:entry :lexiconSR/igracy-n-sr .
:lexiconSR/igracy-n-sr a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconSR/igracy-n-sr-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconSR/igracy-n-sr-form a ontolex:Form ;
  ontolex:writtenRep "играч"@sr .
:lexiconES a lime:Lexicon .
...
:lexiconES lime:entry :lexiconES/jugador-n-es .
:lexiconES/jugador-n-es a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconES/jugador-n-es-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconES/jugador-n-es-form a ontolex:Form ;
  ontolex:writtenRep "jugador"@es .
...
:tranSetSR-ES a vartrans:TranslationSet ;
...
:tranSetSR-ES vartrans:trans
  :tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans.
:tranSetSR-ES/igracy_jugador-n-sr-sense a ontolex:LexicalSense ;
```

```
ontolex:isSenseOf :lexiconSR/igracy-n-sr .  
:tranSetSR-ES/jugador_igracy-n-es-sense a ontolex:LexicalSense ;  
  ontolex:isSenseOf :lexiconES/jugador-n-es .  
:tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans  
  a vartrans:Translation ;  
  vartrans:source :tranSetSR-ES/igracy_jugador-n-sr-sense ;  
  vartrans:target :tranSetSR-ES/jugador_igracy-n-es-sense .
```

Модул FrAC омогућава да се прикажу и колокације и ембединзи (Chiarcos et al. 2022; Chiarcos et al. 2020), односно различита векторска представљања.

5. Закључак

Развој фудбалског речника путем методологије повезаних података показао је значајне предности у односу на традиционалне приступе. Аутоматизација процеса, коришћење корпуса и интеграција различитих дигиталних ресурса омогућили су бржу и ефикаснију израду речника. Дигитални речници нуде многе предности, као што су брзина претраге, стална ажурирања и прилагодљивост корисницима, што их чини незамењивим алатом у савременом добу. Примена техника рачунарске лингвистике и технологија семантичке мреже омогућила је стварање речника који задовољава потребе људи и машина. У будућности очекујемо да се овакви приступи примене на израду речника у другим областима, чиме се додатно унапређује област лексикографије и језичких технологија.

Захвалница

Истраживање је спроведено уз подршку Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

Литература

Andonovski, Jelena. 2023. “SrpNemKor and Linked Open Data.” *Infotheca - Journal for Digital Humanities* 23 (1): 33–60. ISSN: 2217-9461. <https://doi.org/10.18485/infotheca.2023.23.1.2>.

- Archer, Ph., S. Goedertier, and N. Loutas. 2012. *Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC*. Technical report. European Commission.
- Bergh, Gunnar, and Sölve Ohlander. 2019. “A hundred years of football English: A dictionary study on the relationship of a special language to general language.”
- Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. “Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC.” In *Proceedings of the 29th International Conference on Computational Linguistics*, 4018–4027.
- Chiarcos, Christian, Christian Fäth, and Maxim Ionov. 2022. “Unifying morphology resources with OntoLex-morph. a case study in German.” In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4842–4850.
- Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. “Modelling frequency and attestations for ontollex-lemon.” In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, 1–9.
- Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. “FrameNet in action: The case of attaching.” *International journal of lexicography* 16 (3): 297–332.
- Fuertes-Olivera, Pedro A, and Sven Tarp. 2014. *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. Vol. 146. Walter de Gruyter GmbH & Co KG.
- Ivanović, Tanja, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. “Corpus-based bilingual terminology extraction in the power engineering domain.” *Terminology* 28 (2): 228–263.
- Kilgarrieff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine: ten years on.” *Lexicography* 1 (1): 7–36.
- Kilgarrieff, Adam, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. “GDEX: Automatically finding good dictionary examples in a corpus.” In *Proceedings of the XIII EURALEX international congress*, 1:425–432. Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra ...

- Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. “A data driven approach for raw material terminology.” *Applied Sciences* 11 (7): 2892.
- Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.
- Krstev, Cvetana, and Ranka Stanković. 2023. “Language Report Serbian.” In *European Language Equality: A Strategic Agenda for Digital Language Equality*, edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.
- Krstev, Cvetana, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. “Terminology Acquisition and Description Using Lexical Resources and Local Grammars.” In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence*. Granada, Spain: LexiCon (Universidad de Granada).
- Krstev, Cvetana, and Duško Vitas. 2009. “An Effective Method for Developing a Comprehensive Morphological E-Dictionary of Compounds.” In *Proceedings of the 28th Conference on Lexis and Grammar*, edited by B. Lamiroy, E. Laporte, and T. Kyriakopoulou, 204–212. 29th September – 3rd October 2009, in Arena Romanistica. Bergen: University of Bergen, Department of Foreign Languages.
- Lazarević, Jelena, Ranka Stanković, Mihailo Škorić, and Biljana Rujević. 2023. “Football terminology: compilation and transformation into OntoLex-Lemon resource.” In *Proceedings of LDK 2023 – 4th Conference on Language, Data and Knowledge*. 12–15 September, Vienna, Austria. Lisbon: NOVA FCSH - CLUNL. <https://doi.org/10.34619/srmk-injj>.
- McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. “The Ontolex-Lemon model: Development and applications.” In *Proceedings of eLex 2017 conference*, 19–21.
- Mihajlović, Aleksandar. 2003. *Fudbalski rečnik: srpsko-englesko-francusko-španski*. Beograd : A. Mihajlović, 2003 (Beograd : Solidarnost).
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. “Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems* 26.

- Río Gracia, Jorge del, et al. 2023. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0) Draft Community Group Report*. Technical report. 03 October 2023. October. https://bpmlod.github.io/Bilingual_Dictionaries_Report/.
- Robin, Cécile, Atharva Kulkarni, and Paul Buitelaar. 2023. "Identifying FrameNet Lexical Semantic Structures for Knowledge Graph Extraction from Financial Customer Interactions." In *Proceedings of the 12th Global Wordnet Conference*, edited by German Rigau, Francis Bond, and Alexandre Rademaker, 91–100. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, January. <https://aclanthology.org/2023.gwc-1.11/>.
- Savary, Agata, Cvetana Krstev, and Duško Vitas. 2007. "Inflectional non compositionality and variation of compounds in French, Polish and Serbian, and their automatic processing." In *Bulag – Bulletin de Linguistique Appliquée et Générale: Les langues slaves et le français : approches formelles dans les études contrastives*, edited by Aleksandra Dziadkiewicz and Izabella Thomas, 73–94. 32. Besançon: Presses Universitaires de Franche Comté.
- Schmidt, Thomas. 2009. "The Kicktionary—A multilingual lexical resource of football language." *Multilingual FrameNets in computational lexicography. Methods and applications*, 101–134.
- Škorić, Mihailo, and Nikola Janković. 2024. "New Textual Corpora for Serbian Language Modeling." Faculty of Philology, University of Belgrade, *Infoteka* 24 (1).
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries - from File System to lemon Based Lexical Database." In *Proceedings of the 11th LREC – W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018)*. Miyazaki, Japan: ELRA.
- Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference*, 77–84.

- Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. “Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian.” In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3954–3962. Marseille, France: ELRA.
- Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. “Parallel bidirectionally pretrained taggers as feature generators.” *Applied Sciences* 12 (10): 5028.
- Vitas, Duško. 1979. “Prikaz jednog sistema za automatsku obradu teksta.” *Zbornik radova XIV jugoslovenskog međunarodnog simpozijuma o obradi podataka Informatica* 79:7–101.
- Vu, Thuy, Aiti Aw, and Min Zhang. 2008. “Term extraction through unit-hood and termhood unification.” In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.
- Wu, Jiguang, and Ying Li. 2017. “Research on construction of semantic dictionary in the football field.” In *Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, 303–306. IEEE Xplore Full-Text PDF available. IEEE.
- Рујевић, Биљана. 2022. “Речници у дигиталном добу – информатичка подршка за српски језик.” Докторска дисертација, Универзитет у Београду, Филолошки факултет.
- Утвић, Милош. 2011. “Анотација Корпуса савременог српског језика.” *ИНФОтека* 12, no. 2 (December): 39–51.