

СрпНемКор и отворени повезани подаци

УДК 811.163.41'322.2:811.112.2'322.2]:004.822

САЖЕТАК: У раду ће бити приказане могућности примене принципа и смерница семантичког веба на паралелном корпусу романа *СрпНемКор*. Са једне стране, анализирано је увезивање неких библиографских метаподатака о романима (име аутора и наслов) са записима у референтним ресурсима у семантичким базама података, док је, са друге стране, на основу садржаја корпуса генерисан двојезични речник општег типа, *SrpNemLex*, као скуп отворених повезаних података. На крају је илустрована репрезентација романа из корпуса Википодацима и претраживање SPARQL упитима са различитим могућностима за визуализацију добијених резултата.

КЉУЧНЕ РЕЧИ: СрпНемКор, метаподаци, семантички веб, отворени повезани подаци, Linked Open Data, RDF, *SrpNemLex*, Википодаци, SPARQL

РАД ПРИМЉЕН: 30. март 2023.

РАД ПРИХВАЋЕН: 9. мај 2023.

Јелена Андоновски

andonovski@unilib.rs

Универзитетска библиотека

„Светозар Марковић“

Београд, Србија

1. Увод

Семантички веб или Веб 3.0 дефинисан је као механизам за структурисање и увезивање података са прецизно дефинисаним значењем од чега се формирају међусобно повезани скупови података (енгл. *data sets*) који се могу претраживати преко система обједињене претраге. Тако је од „веба докумената“ (енгл. *Web of Documents*) створен „веб података“ (енгл. *Web of Data*)¹ стварајући од веба глобалну базу знања. У овом концепту садржај података исказан на природном језику експлицитно се анотира тако да га машине могу прочитати и коректно интерпретирати (разумети) (Rurarz et al. 2021).

1. [Linked Data. W3C](#), приступљено 10.3.2023.

Кључна парадигма семантичког веба су „повезани подаци“ (Linked Data - LD) који омогућавају да принципи и технологије семантичког веба (Janik, Scherp, and Staab 2011) постану доступни и примењиви за различите апликације, а скупови података међусобно увезани и доступни у отвореном приступу, Linked Open Data (LOD) (Radulovic et al. 2015). Концепт семантичког веба и технологије отворених повезаних података проширују традиционални веб употребом стандардног језика за означавање и сродних алата за обраду, где оквир за описивање ресурса на вебу (Resource Description Framework – RDF)² има велику улогу и омогућава ефикаснија решења за проналажење информација (Shah et al. 2002; Stanković and Davidović 2021).

Иницијатива „отворени повезани подаци“ развија се више од деценије и обједињује разноврсне ресурсе и врсте података у технологији облака. Највећи део облака чине базе знања од којих је једна Википодаци (Wikidata)³. Википодаци је вишејезична база знања општег типа коју је развила Задужбина Викимедија (Wikimedia Foundation) са циљем да омогући преузимање и чување структурираних података на више језика са Википедије у интероперабилном машински читљивом формату (Vrandečić and Krötzsch 2014). Она се користи не само за структурисање података из различитих Викимедијиних пројеката, већ и у бројним апликацијама заснованим на семантичком вебу, а и ми смо је користили у нашем истраживању. Облак LOD је подељен у девет подоблака⁴ који су даље подељени на поткатегорије. Облак LOD је подељен у девет подоблака *лингвистика*⁵ који су даље подељени на поткатегорије. Један од подоблака који је од посебног значаја за наше истраживање је *лингвистика*⁶ који обједињује или описује лингвистичке ресурсе који се могу користити за потребе истраживања у различитим областима лингвистике или у оквиру области обраде природних језика (McCrae 2016; Cimiano et al., 2020a). Подоблак *лингвистика* (Chiarcos and Pareja-Lora 2019; Chiarcos et al. 2020; Cimiano et al., 2020b) подељен је на седам категорија: корпуси; лексикони и речници; терминологија, тезауруси и базе знања;

2. [Resource Description Framework](#), приступљено 10.3.2023.

3. [Wikidata](#), приступљено 10.3.2023.

4. [The Linked Open Data Cloud](#), приступљено 20.4.2023.

5. [The Linked Open Data Cloud. Subclouds by domain. Linguistics](#), приступљено 20.4.2023.

6. [The Linked Open Data Cloud](#), приступљено 20.4.2023.

метаподаци о лингвистичким ресурсима; категорије лингвистичких података; типолошке базе података; друго.

У раду смо анализирали могућности примене принципа и смерница семантичког веба на садржај паралелног корпуса СрпНемКор. Прво, повезали смо библиографске метаподатке о романима (име аутора и наслов) са одговарајућим записима у референтним базама знања, друго, на основу садржаја корпуса припремљен је двојезични речник општег типа, *SrpNemLex*, као скуп отворених повезаних података и, на крају, анализирали смо пример Википодатака за један од романа из корпуса и могућности претраживања постављањем SPARQL упита.

2. Паралелни корпус *СрпНемКор*

Паралелни српско-немачки корпус књижевних текстова, СрпНемКор, настао је у оквиру рада на докторској дисертацији „Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса“ (Андоновски 2019). Корпус садржи укупно 14 паралелизованих романа, од чега седам српских писаца који су преведени на немачки језици и седам писаца немачког говорног подручја (у овом случају се мисли на писце са територије Немачке и Аустрије) који су преведени на српски језик. Након урађене паралелизације корпус је смештен у дигиталну библиотеку Библиша⁷ као посебна колекција СрпНемКор⁸ са две потколекције: Романи оригинално написани на немачком језику и Романи оригинално написани на српском језику.

Библиша је дигитална библиотека коју је развила Група за језичке технологије Универзитета у Београду, а коју данас одржава Друштво за језичке ресурсе и технологије – ЈеРТех⁹. Креирана је са циљем унапређења могућности претраживања вишејезичних дигиталних библиотека електронских часописа и других паралелизованих текстова (Stanković et al. 2012). Библиша омогућава вишејезичну претрагу комплетног текста паралелних колекција уз могућност морфолошког и семантичког проширења упита позивањем различитих вишејезичних лексичких и термилошких ресурса (Stanković et

7. Библиша, приступљено 20.4.2023.

8. СрпНемКор у дигиталној библиотеци Библиша, приступљено 20.4.2023.

9. Друштво за језичке ресурсе и технологије - ЈеРТех, приступљено 20.4.2023.

al. 2015; Stanković et al. 2016; Stanković et al. 2017). До постављања колекције СрпНемКор, библиотека Библиша је садржала паралелне српско-енглеске колекције текстова па је и претраживање, уз могућност проширења упита, било могуће само на ова два језика. Како би се омогућила претрага колекције на немачком језику и користиле могућности алата за проширење упита, допуњен је један од термилошких ресурса, вишејезична база Терми,¹⁰ листом преводних немачко-српских еквивалената лексичких јединица са синонимима (детаљније описано у следећем поглављу) која је истовремено послужила и за генерисање двојезичног речника општег типа као скуп „отворени повезани подаци“ (детаљније описано у одељку 4.2).¹¹

3. Листа преводних немачко-српских еквивалената

Листа преводних немачко-српских еквивалената лексичких јединица настала је на основу садржаја корпуса *СрпНемКор* (Andonovski, Šandrih, and Kitanović 2019). За ове потребе користили смо алат BilTE (Bilingual Terminology Extraction)¹² (Krstev et al. 2018) који су развили чланови друштва ЈеРТех, а који омогућава екстракцију двојезичне терминологије на основу паралелних двојезичних текстуалних колекција, термилошких листи изворног језика и система за екстракцију полилексемских термина циљног језика. У нашем случају за изворни језик одређен је немачки, а за циљни српски. За ове потребе користили смо следеће језичке ресурсе:

1. поравнати корпус књижевних текстова СрпНемКор са 48.004 упарених сегмената (Андоновски 2021);
2. листу лексичких јединица за немачки језик као изворни која је припремљена коришћењем два извора. Као први извор коришћен је немачки Ворднет, Open-de-WordNet¹³ (*Open DE WordNet Initiative*

10. Терми је термилошка вишејезична база која подржава развој термилошких речника из различитих домена (математика, рачунарство, рударство, библиотекарство, рачунарска лингвистика и многи други) и омогућава семантичко проширење упита у *Bibliši*, приступљено 20.4.2023.

11. Двојезични речник *SrpNemLex*, приступљено 20.4.2023.

12. BilTE – екстракција двојезичне терминологије, приступљено 20.4.2023.

13. Иницијатива Open-de-WordNet покренута је са идејом да се направи Ворднет на немачком језику који ће постати део ширег вишејезичног Ворднет окружења, бити у потпуно отвореном приступу и моћи да се користи у оквиру

2023), из кога је добијено приближно 120.000 литерала. Као други извор коришћена је листа најфреквентнијих речи на немачком језику која је преузета из Викиречника,¹⁴ а која приближно броји око 10.000 речи. Ове две листе спојене су у једну и добијена је прелиминарна листа лексичких јединица за немачки језик. Како би се постигли бољи резултати урађена је лематизација речи из добијене листе за шта смо користили модел доступан на spaCy¹⁵ који омогућава лематизацију користећи табелу од 355.354 монолексемских јединица. После лематизације и елиминације дупликата добијена је листа од 27.638 различитих лексичких јединица на немачком језику;

- екстрактор полилексемске терминологије (MWT) за српски језик као циљни. За екстракцију полилексемских термина на српском делу корпуса коришћен је алат *LeXimir* заснован на електронским морфолошким речницима и локалним граматикама за екстракцију полилексемских јединица за српски језик (Stanković et al. 2016). Како се систем користи само за екстракцију полилексемских јединица произвели смо додатно и „врећу речи“¹⁶ из српских текстова користећи систем Unitex.¹⁷ Произведене листе садржале су 94.802 монолексемских и 48.159 полилексемских јединица. Након тога, урадили смо лематизацију применом електронских морфолошких речника за српски језик (Krstev 2008; Крстев 2019; Stanković et al. 2018) после чега су елиминисани дупликати и добијена је листа од 77.297 различитих лексичких јединица на српском језику.

платформе Алата за обраду природних језика ([Natural Language Toolkit](#)). Прва верзија [Open-de-WordNet](#) урађена је на основу лексикона синонима на немачком језику у отвореном приступу [OpenThesaurus](#) ([OpenThesaurus German synonym lexicon](#)) и [енглеског Ворднета](#) и објављена у пролеће 2017. године. Више о иницијативи на [Open-de-WordNet](#), а сам тезаурус је доступан на [Open-de-WordNet](#) и може се користити под лиценцом [Creative Commons Attribution Share Alike 4.0 International](#) која дозвољава комерцијалну употребу, измене, дистрибуцију и приватну употребу.

14. [Wiktionary:Frequency lists](#), приступљено 20.4.2023.

15. [spaCy](#) је бесплатна библиотека отвореног кода писана у програмским језицима Python и Cython која нуди моделе за означавање именованих ентитета у енглеском, немачком, шпанском, португалском, француском, италијанском, холандском и у вишејезичним текстовима.

16. Врећа речи (енгл. *Bag of Words*) представља скуп речи у једном тексту са израчунатом фреквенцијом њиховог појављивања у том тексту независно од њихових граматичких облика и реда речи у реченици.

17. Комплет алата за обраду корпуса [Unitex](#), приступљено 20.4.2023.

Применом алата за машинско превођење GIZA++¹⁸ (Och and Ney 2003; Koehn, Och, and Marc 2003) произведена је листа преводних немачко-српских еквивалената полилексемских јединица која је након филтрације садржала 14.142 потенцијална кандидата. Након ручне евалуације добијена је коначна листа од 3.984 исправна преводна еквивалента која је заједно са произведеном листом синонима за оба језика (864 синонима на српском и 791 синоним на немачком језику) интегрисана у базу *Терми*. У табели 1 приказани су неки примери преводних еквивалената, док су у табели 2 приказани неки примери синонима преводних еквивалената за оба језика.

Изворни језик (немачки)	Циљни језик (српски)	Енглески
Haltestelle	аутобуска станица	bus station
Hoffnungslos	безнадежан случај	hopeless case
Allee	алеја	alley
alt Frau	старица	old woman

Табела 1. Немачко-српски преводни еквиваленати лексичких јединица.

4. *СрпНемКор* и отворени повезани подаци

У овом одељку анализирали смо могућност примене принципа и смерница семантичког веба. У одељку 4.1 приказали смо повезивање библиографских метаподатка са релевантним записима у референтним базама података, а у одељку 4.2 двојезични речник општег типа као скуп повезаних отворених података.

4.1 Увезивање библиографских метаподатака

Колекција *СрпНемКор* је у дигиталној библиотеци *Библиша* описана одговарајућим метаподацима што је детаљније објашњено у (Андоновски 2019, 2021), а ми ћемо овде анализирати само повезивање метаподатака „име аутора“ и „наслов дела“ са записима у три нормативне

18. GIZA++ – алат за тренирање статистичких модела за превођење, приступљено 20.4.2023.

лексичке јединице на српском	Синоними на немачком
потпун	absolut, gänzlich, total, völig, vollkommen, vollständig
базен	Bassin, Becken, Schwimmbad, Schwimmbecken
мајка	Mama, Mutter, Mutti
Лексичке јединице на немачком	Синоними на српском
Augenblick	трен, тренутак момент, час
fremd	туђ, туђински, непознат
grenzenlos	безграничан, бескрајан

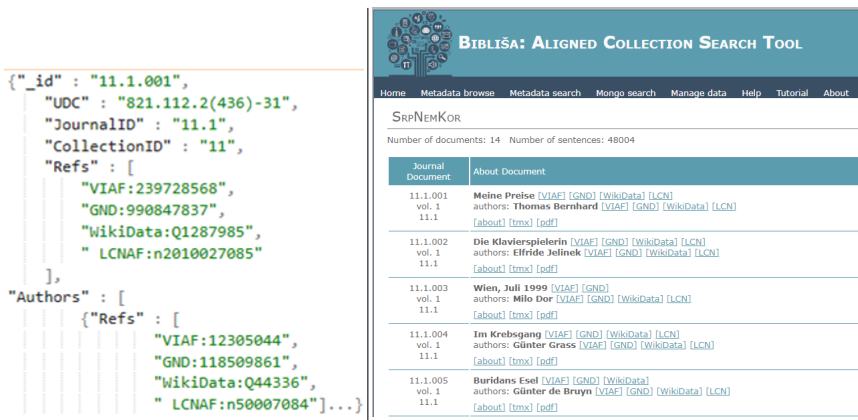
Табела 2. Синоними преводних немачко-српских еквивалената лексичких јединица

датотеке и једној општој бази знања које су део облака LOD, а то су: Међународна виртуелна нормативна датотека (Virtual International Authority File – VIAF),¹⁹ Нормативна датотека Националне библиотеке Немачке (Gemeinsame Normdatei – GND)²⁰ Нормативна датотека имена Конгресне библиотеке (Library of Congress Name Authority File – LCNAF)²¹ и база знања Википодаци (Wikidata). У свакој од поменутих база записи имају своје јединствене идентификаторе (ID) преко којих је урађено повезивање. Они су наведени у одговарајућем делу шеме за метаподатке у *Библиши* и груписани заједно у зависности од тога да ли се односе на име аутора или наслов дела. На слици 1 (лево) приказана је структура метаподатака у Библиши за роман „Моје награде“ (Meine Preise) аутора Томаса Бернхарда (Thomas Bernhard) где видимо да елемент “Refs” групише ID бројеве записа у поменутих базама на начин и у оном редоследу како се после појављују у корисничком окружењу. На слици 1 (десно) видимо панел *Библише* где уз име аутора односно наслов дела стоје линкови ка путањама записа у поменутих базама.

19. Virtual International Authority File – VIAF

20. Gemeinsame Normdatei – GND

21. Library of Congress Name Authority File – LCNAF



Слика 1. Повезивање метаподатака у *БиблИши* са записима у одабраним базама (лево); Приказ веза у *БиблИши* са одабраним базама (десно).

4.2 Двојезични речник као скуп отворених повезаних података

Немачко-српски електронски речник општег типа, *SrpNemLex*, који ми овде представљамо као скуп отворених повезаних података, генерисан је извозом и трансформацијом листе немачко-српских преводних еквивалената са синонимима (описано у одељку 3.) из базе *Терми*. За генерисање речника као скуп отворених повезаних података коришћен је Модел лексикона за онтологије (*Lexicon Model for Ontologies – OntoLex*) (Cimiano, McCrae, and Buitelaar 2016; McCrae 2016; Bosque-Gil and Gracia 2019) који је развила Онтолекс група (*Ontology-Lexicon community group – OntoLex*).²² Модел је заснован на формату RDF и намењен за опис речника и других лингвистичких ресурса који садрже лексикографске податке, у форми информатичке онтологије на вебу (Buitelaar et al. 2011). Модел следи принцип „семантика кроз референце” (*semantics by reference*) (McCrae et al. 2012) што значи да се додатне лингвистичке информације о лексичкој јединици исказују реферисањем на релевантне онтологије (Bosque-Gil and Gracia 2019). *OntoLex* модел је надоградња моделу *lemon* (Gracia and Vila-Suero 2015; Gracia et al. 2018), који је коришћен као претходна верзија и доста

22. *Ontology-Lexica Community group*

класа и својстава за опис одредица и преводних еквивалената је из њега задржана.

Целокупна листа припремљених немачко-српских преводних еквивалената извезена је из базе *Терми* у модел OntoLex за шта је коришћена програмска библиотека dotNetRDF,²³ а за приказ добијених резултата користи се синтакса Turtle/RDF.²⁴ За разлику од претходне верзије модела *lemon* када су формиране три одвојене датотеке (датотека речника изворног језика (у нашем случају немачки), датотека речника циљног језика (у нашем случају српски) и датотека преводних еквивалената) (Андоновски 2019), сада је генерисана једна датотека која садржи заглавље са метаподацима и три скупа података: скуп података на немачком језику (одреднице у немачком делу речника), скуп података на српском језику (одреднице у српском делу речника) и скуп преводних еквивалената.

Одреднице на оба језика груписане су преко својства *lime:entry* у одвојене скупове података. Све одреднице у скуповима података су поређане у абecedном редоследу и имају словни идентификатор који се састоји из писаног облика одреднице, податка о врсти речи и податка о језику. На пример, :AUGENBLICK-N-DE или :TREN-N-SR. После сваког скупа података, у нашем примеру на српском и немачком језику, свака одредница додатно је описана на следећи начин: врста речи дефинисана преко својства *lexinfo:partOfSpeech*, идентификатор одреднице дефинисан преко својства *ontolex:Form* и писани облик одреднице дефинисан преко својства *ontolex:writtenRep* са податаком о језику који се наводи уз ознаку @, док се преко својства *ontolex:LexicalEntry* указује да је реч о одредници која припада одређеном скупу података. Када је реч о словима са дијакритичким ознакама у оба језика (у српском *č, ć, š, ž, đ*, у немачком *ö, ä, ü*) у речнику су она изражена кодовима (кодирање за веб), на пример, %C4%8DAS – ČAS, OBU%C4%87AR – OBUĆAR, %C3%96FFENTLICHKEIT – ÖFFENTLICHKEIT. У табели 3 дали смо примере одредница у речнику SrpNemLex у формату RDF/Turtle.

Приликом формирања преводних еквивалената свака одредница у скуповима података добија алфанумерички идентификатор од чега се формира идентификатор преводног еквивалента, на пример, :112383-SR-SENSE-112383-DE-SENSE-TRANS. Преводни еквиваленти се, затим, групишу у посебан скуп података преко класе *vartrans:Translation* после

23. dotNetRDF

24. Turtle

Одредница у скупу података на немачком језику	Објашњење
:Augenblick-n-de <i>lexinfo:partOfSpeech</i> lexinfo:noun; <i>a ontolox:LexicalEntry</i> ; <i>ontolox:Form</i> :Augenblick-n-de-form. :Augenblick-n-de-form <i>ontolox:writtenRep</i> "Augenblick"@de.	врста речи - вредност из Lexinfo онтологије одредница идентификатор одреднице писани облик одреднице
Одредница у скупу података на српском језику	Објашњење
:%C4%8Das-n-sr <i>lexinfo:partOfSpeech</i> lexinfo:noun; <i>a ontolox:LexicalEntry</i> ; <i>ontolox:Form</i> :%C4%8Das-n-sr-form. :%C4%8Das-n-sr-form <i>ontolox:writtenRep</i> "čas"@sr.	врста речи - вредност из Lexinfo онтологије одредница идентификатор одреднице писани облик одреднице

Табела 3. Примери одредница у речнику SrpNemLex у формату RDF/Turtle.

чега је сваки преводни еквивалент појединачно описан тако да су дефинисане све одреднице које га чине, у нашем примеру одреднице на немачком и српском језику. Свака одредница у преводном еквиваленту појединачно је описана преко класе *ontolox:LexicalSense* на следећи начин: преко својства *skos:definition* приказује се значење одреднице у форми текстуалне дефиниције са податком о језику, ако постоји, преко својства *ontolox:isLexicalizedSenseOf* наводи се одредница у формату како је наведена у припадајућем скупу података језика, а својствима *vartrans:source* и *vartrans:target* указује се на скупове података изворног и циљног језика односно скупа података којима одреднице из преводног еквивалента припадају. У табели 4 дали смо пример једног преводног еквивалента из речника SrpNemLex, док је на слици 2 дат графички приказ истог преводног еквивалента.

Поред онтологија које су моделом OntoLex предвиђене за генерисање и структурисање речника (@prefix *ontolox:*, @prefix *vartrans:*, @prefix *lime:*) могуће је и коришћење додатних онтологија како би се одредницама у речнику и преводним еквивалентима које формирају додала додатна лексикографска објашњења као што су врсте речи, дефиниције и слично. У нашем примеру ми смо користили онтологије

Пример преводног еквивалента	Објашњење
<pre>:112383-de-sense a ontolex:LexicalSense; skos:definition "bestimmte Zeitspanne von kurzer Dauer, Moment, Augenblick"@de; ontolex:isLexicalizedSenseOf :Augenblick-n-de, :Moment-n-de, :Zeitpunkt-n- de :112383-sr-sense a ontolex:LexicalSense; skos:definition "vrlo kratki deo vremena, tj. vreme potrebno da se trepne okom"@sr; ontolex:isLexicalizedSenseOf :%20tren-n-sr, :%20trenetak-n-sr, :%C4%8Das-n-sr, :moment-n-sr. :112383-sr-sense-112383-de-sense-trans a vartrans:Translation; vartrans:source :112383-sr-sense; vartrans:target :112383-de-sense.</pre>	<p>одредница из скупа података на нем. језику дефиниција одреднице</p> <p>формат одреднице на нем. језику</p> <p>одредница из скупа података на срп. језику дефиниција одреднице</p> <p>формат одреднице на срп. језику</p> <p>преводни еквивалент – класа <i>vartrans:Translation</i> скуп података изворног језика (немачки) скуп података циљног језика (српски)</p>

Табела 4. Пример преводних еквивалента у речнику SrpNemLex у RDF/Turtle формату.

Lexinfo²⁵ и SKOS²⁶ (@prefix lexinfo: и @prefix skos:). Lexinfo (Buiteelaar et al. 2009) је онтологија која омогућава анотацију лингвистичких својстава одредница. Делимично је изведена из регистра DatCatInfo,²⁷ Репозиторијум категорије података (Data Category Repository – DCR), који је развијен према стандарду ISO 12620:2019²⁸ као замена за регистар метаподатака ISOcat²⁹ (Kemps-Snijders et al. 2008). У нашем примеру речника, Lexinfo смо користили за дефинисање врсте речи одреднице. У претходно наведеним примерима преко својства *lexinfo:partOfSpeech* дефинисана је врста речи дате одреднице и његова вредност је LEX-INFO:NOUN на основу чега видимо да је реч о именицама.

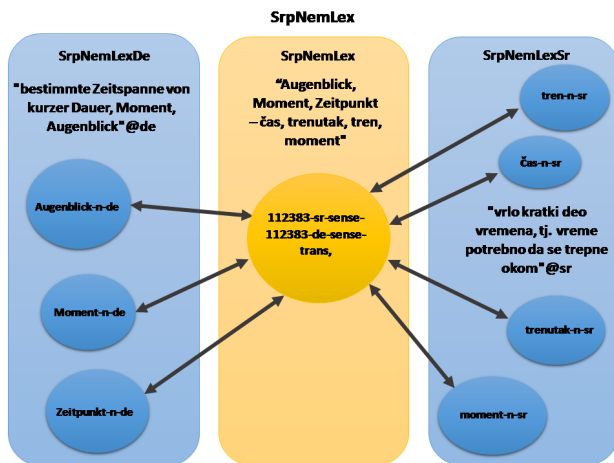
25. Lexinfo

26. Simple Knowledge Organization System

27. DatCatInfo

28. ISO 12620:2019

29. Званичан регистар информација о лингвистичким категоријама података на вебу и значајан ресурс за истраживања и развој различитих лингвистичких дисциплина.



Слика 2. Графички приказ примера из речника SrpNemLex.

Друга онтологија коју смо користили је SKOS (Isaac and Summers 2009) за исказивање значења одреднице у форми дефиниције. У претходно наведеним примерима преводних еквивалената значење одредница наведено је преко својства *skos:definition* чија је вредност ниска карактера уз коју стоји код за језик на коме је дефиниција. У нашем примеру из табеле 4 имамо пример дефиниција и на немачком и на српском језику:

- *skos:definition* "bestimmte Zeitspanne von kurzer Dauer, Moment, Augenblick"@de, односно
- *skos:definition* "vrlo kratki deo vremena, tj. vreme potrebno da se trepne okom"@sr;

Да би структура речника била потпуна потребно је креирати одговарајуће метаподатке који дају све потребне информације о ресурсу. За потребе овог рада креирали смо заглавље са основним метаподацима (кратак опис речника – *dc:description*, метаподатак о томе где је речник тренутно складиштен – *dc:source*, наслов речника – *dc:title* и аутори односно сарадници који су на њему радили – *dct:contributor*), док ће потпуни метаподаци бити креирани у каснијем истраживању применом модела LexMeta³⁰ за опис лексичких ресурса (Lindemann, Labropoulou,

30. LexMeta, приступљено 9.5.2023.

and Klaes 2022). Пример тренутног заглавља са метаподацима за речник *SrpNemLex*:

```
:SrpNemLexTrans dc:description "This is the lemon version of the
bilingual dictionary for German and Serbian languages. Bilingual
lemon dictionary contains lexical units with synonyms and was
generated from German-Serbian parallel corpus. Corpus is the subject
of the doctoral dissertation Linked Open Data and Language Resoruces
in preparing Serbian-German Literary Corpus. For German-Serbian
lemon bilingual correspondence, the corresponding source and target
monolingual entries (LexicalEntry) were generated in addition to the
bilingual correspondence (SenseAxis) element."@en;
dc:source <http://llod.jerteh.rs/id/>;
dc:title "German-Serbian lemon Bilingual dictionary";
dct:contributor "Branislava Šandrih", "Cvetana Krstev", "Jelena
Anonovski", "Olivera Kitanović", "Ranka Stanković";
```

5. СрпНемКор у Википодацима

У овом одељку илустровали смо репрезентацију једног од романа из корпуса СрпНемКор Википодацима и претраживање помоћу SPARQL упита са различитим могућностима за визуализацију добијених резултата. У тренутку припреме корпуса СрпНемКор постојали су Википодаци за све ауторе који су заступљени у корпусу и за шест наслова романа као књижевно дело. Овим истраживањем започели смо припрему Википодатака за остале романе из корпуса према принципу који је коришћен у пројекту WikiELTeC³¹ (Ikonić Nešić, Stanković, and Rujević 2021) реализован на корпусу српских романа srpELTeC³² (Trtovac, Milnović, and Krstev 2022) који је настао у оквиру COST акције CA16204 (2017-2021) „Удаљено читање за европску историју књижевности“.³³

31. WikiELTeC; Википедија:Википројекат WikiELTeC, приступљено 21.4.2023.

32. srpELTeC: Колекција српских књижевних текстова (ELTeC)

33. CA16204 – Distant Reading for European Literary History (DISTANT-READING)

5.1 Википодаци

Википодаци су оријентисани на документе, усредсређени на ставке, које представљају теме, концепте или објекте. Уз поседовање личног налога могу се креирати нове или уређивати постојеће ставке и то на језику који се одабере на почетку, чиме се дефинише подразумевани језик. Свака ставка у Википодацима има јединствени идентификатор, такозвани трајни идентификатор (енгл. *permanent identifier* – QID), који се додељује аутоматски приликом креирања записа и не може се касније мењати (Ikonić Nešić, Stanković, and Rujević 2021). Идентификатор је у форми позитивног целог броја коме претходи велико слово Q, што омогућава људима и машинама да разумеју разлике између ставки са истим или сличним именом али различитим значењем. Идентификатор ставке повезан је са насловом и описом, може имати више псеудонима и одређен број изјава којима се представљају својства и вредности ставке која се описује (Станковић and Давидовић 2021). Изјаве су у форми уређене тројке, тзв. RDF тројке, како је и прописано смерницама и принципима семантичког веба: субјекат (ставка) – предикат (релација) – објекат (вредност) (Schreiber and Raimond 2014).³⁴ Субјекат (почиње ознаком Q) је било која тема (особа, предмет, место, концепт), на пример, Q571 (књига); предикат (почиње ознаком P) је карактеристика релевантна за субјекат, на пример P123 (издавач); објекат је додатна информација о субјекту и може бити „литерал“ односно ниска карактера, на пример, „1. новембар 1926“, затим, неки други ентитет, на пример, Q3711 (Београд) или URL адреса (Hernández, Hogan, and Krötzsch 2015; Erxleben et al. 2014).

5.2 „Буриданов магарац“ у Википодацима

За илустрацију примера из нашег корпуса у Википодацима одабрали смо роман „Буриданов магарац“ аутора Гинтера де Бројна. На почетку истраживања утврдили смо да у Википодацима постоје следеће ставке везане за овај роман: књижевно дело „Буриданов магарац : роман“ (Q110637016)³⁵ са којим је већ успостављена веза у *Библиши* приликом повезивања метаподатака (одељак 4.1), израз „Буриданов магарац“ (Q340003)³⁶ као филозофска параболоа, филм из 1980. године који је

34. RDF 1.1 primer 2014

35. Буриданов магарац; роман

36. Буриданов магарац

урађен по мотивима романа Буриданов магарац „Glück im Hinterhaus (engl. Backhouse Bliss)“ (Q1532735)³⁷ и аутор романа „Гинтер де Бројн“ (Q62753)³⁸ са којим је већ успостављена веза у Библиши приликом повезивања метаподатака (одељак 4.1). Према искуству из пројекта WikiELTeC неопходно је да за сваки роман постоји ставка описана изјавом „је (P31) књижевно дело (Q7725634)“ и ставка (или ставке) описане изјавом „је (P31) издање (Q3331189)“ које међусобно морају бити повезане преко својстава „издање“ (P747) односно „издање од“ (P629).

Као што је већ поменуто, ставка „књижевно дело“ је за наш пример већ постојала у Википодацима и била је описана следећим својствима:

- „је“ (P31),
- „наслов“ (P1476),
- „изведено дело“ (P4969),
- „облик стваралачког дела“ (P7937),
- „аутор“ (P50),
- „језик дела или имена“ (P407),
- „датум издавања“ (P577),
- „GND ID“ (P227).

За потребе нашег истраживања додали смо вредности за својства:

- „издање“ (P747),
- „место радње“ (P840),
- „радња се дешава у времену“ (P2408) и
- „ликови“ (P674).

У табели 5 дали смо упоредни приказ изјава у Википодацима за ставке „Буриданов магарац: роман“ (Q110637016), књижевно дело, и „Гинтер де Бројн“ (Q62753), аутор романа Буриданов магарац, на природном језику, њихову анотацију ID идентификаторима и у облику изјава. У обе ставке субјекат се не понавља, изјаве истог субјекта су одвојене знаком тачка-зарез („;“), док је иза последње изјаве тачка („.“). Такође, видимо да су вредности објеката најчешће друге ставке које имају свој QID, осим својстава „датум издавања“ (P577), „датум рођења“ (P569) и „датум смрти“ (P570) чије су вредности литерали (редом: „1968.“, „1. новембра 1926“ и „4. октобра 2020“) и својства P227 чија је вредност ID број ентитета у нормативној датотеци GND и представља јединствени идентификатор, „4575022-1“ односно „118516388“.

37. [Glück im Hinterhaus](#)

38. [Гинтер де Бројн](#)

Буриданов магарац је књижевно дело аутора Гинтера де Бројна објављено 1968. године на немачком језику. GND ID број је 4575022-1.

Буриданов магарац (Q110637016) је (P31) књижевно дело (Q7725634). аутора (P50) Гинтера де Бројна (Q62753). објављено (P577) „1968“. на језику дела (P407) немачки (Q108). GND ID (P227) је „4575022-1“.

Q110637016

P31 Q7725634;
P50 Q62753;
P577 „1968“;
P407 Q108;
P227 „4575022-1“.

Гинтер де Бројн је рођен 1. новембра 1926. године у Берлину, а умро 4. октобра 2020. године у Бад Зарову. Био је писац и библиотекар, члан Берлинске академије уметности и Немачке академије језика и поезије. Његов GND ID је „118516388“.

Гинтер де Бројн (Q62753) је рођен (P569) „1. новембра 1926“. у (P19) Берлину (Q64). а умро (P570) „4. октобра 2020“. у (P20) Бад Зарову (Q49421). Гинтер де Бројн (Q62753) је био (P106) писац (Q36180) и библиотекар (Q182436), члан (P463) Берлинске академије уметности (Q414110). и Немачке академије језика и поезије (Q459620). Његов (Q62753) GND ID (P227) је „118516388“.

Q62753

P569 „1. новембра 1926“;
P19 Q64;
P570 „4. октобра 2020“;
P20 Q49421;
P106 Q36180;
Q182436;
P463 Q414110;
Q459620;
P227 „118516388“.

Табела 5. Упоредни приказ исказа на природном језику, коришћењем идентификатора и у облику Wikidata изјава.

Поред ставке „књижевно дело“ креирали смо и ставке „издање“. У нашем истраживању реч је о немачко-српском паралелном корпусу који је настао паралелизацијом текстова на немачком и српском језику. У нашем примеру, корпусно издање романа „Буриданов магарац“ настало је на основу издања на немачком (у нашем примеру из 1992. године) и издања на српском језику (у нашем примеру из 2007. године). Из тог разлога урађена је репрезентација Википодатака за три издања ставке (Q110637016): Буриданов магарац; роман: СрпНемКор издање (Q117220537),³⁹ Буриданов магарац; роман (издање из 1992. године) (Q117221046)⁴⁰ и Буриданов магарац; роман (издање из 2007. године) (Q117220339).⁴¹

За сваку ставку „издање“ користили смо следећа својства за опис:

- „је“ (P31),
- „наслов“ (P1476),

39. Буриданов магарац; роман: СрпНемКор издање

40. Буриданов магарац; роман (издање из 1992. године)

41. Буриданов магарац; роман (издање из 2007. године)

Буриданов магарац : роман (Q110637016)

роман немачког писца Гинтера де Бројна уреди

↳ На другим језицима

језик	Назив	Опис	Псеудоними
српски / srpski	Буриданов магарац : роман	роман немачког писца Гинтера де Бројна	
енглески	Buridan's ass	1968 novel by Günter de Bruyn	
немачки	Buridans Esel	Roman von Günter de Bruyn	
хрватски	Ознака није дефинисана	Опис није дефинисан	

издање	Буриданов магарац : роман	СрпНемКор издање	уреди
језик дела или имена		српски језик	
језик		немачки језик	
датум издавања		2019	
место издавања		Београд	
↳ 0 референце			+ додај референцу
Буриданов магарац : роман (издање из 1992. године)		уреди	
језик дела или имена		немачки језик	
језик		немачки језик	
датум издавања		1992	
место издавања		Франкфурт на Мајни	
↳ 0 референце			+ додај референцу
Буриданов магарац : роман (издање из 2007. године)		уреди	
језик дела или имена		српски језик	
језик		српски језик	
датум издавања		2007	
место издавања		Београд	
↳ 0 референце			+ додај референцу
			+ додај вредност

Слика 3. Својство „издање“ (P747) у ставци „Буриданов магарац: роман“ (Q110637016).

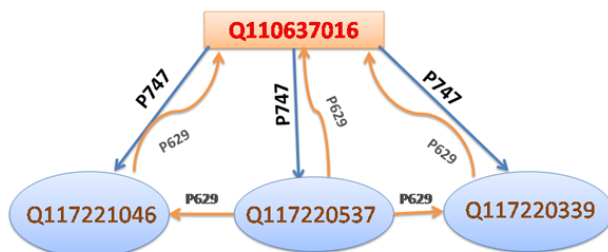
- „облик стваралачког дела“ (P7937),
- „издање од“ (P629),
- „аутор“ (P50),
- „издавач“ (P123), (за ставке Q117221046 и Q117221046)
- „место издавања“ (P291),
- „језик дела или имена“ (P407),
- „датум издавања“ (P577),
- „радња се дешава у времену“ (P2408),
- „место радње“ (P840),
- „број страна“ (P1104),
- „објављено у“ (P1433).

СрпНемКор издање је додатно описано и својствима „пуно дело доступно на“ (P953) и „број речи“ (P6570).

Ставка Q110637016 је преко својства „издање“ (P747) повезана са ставкама Q117220537, Q117220537 и Q117220339 (слика 3), а оне су преко

својства „издање од“ (P629) повезана са ставком Q110637016. Ставка Q117220537 (СрпНемКор издање) је додатно преко својства „издање од“ (P629) повезан и са издањима из 1992. (Q117221046) и 2007. (Q117220339) од којих је и настало (слика 4).

Уз вредности објеката могу стајати додатно квалификатори и референце којима се дају додатне информације о самој вредности објекта. Они су, такође, у форми уређене RDF тројке. На пример, на слици 3 видимо квалификаторе уз вредности својства „издање“ (P747) и то су: „језик дела или имена“, „датум издавања“ и „место издавања“. У овом примеру поменути квалификатори дају корисницима детаљније информације о издањима одабраног књижевног дела.



Слика 4. Својства „издање“ (P747) и „издање од“ (P629) у примеру Википодатака за роман „Буриданов магарац“.

Референце могу имати облик класичног библиографског цитирања или могу бити веб адресе и скупови података на које се реферише од којих сваки може бити ставка у Википодацима. У нашем примеру референце ћемо показати у ставци „Гинтер де Бројн“ (Q62753) и то већ у првој изјави где су наведене следеће референце: једна указује на извор преузимања Википедију на руском језику, а друга указује на Међународну нормативну датотеку Француске (BnF authorities)⁴² уз коју стоји и датум преузимања „10. октобар 2015” и URL адреса “<http://data.bnf.fr/ark:/12148/cb118858431>” (слика 5).

42. BnF authorities



Слика 5. Пример референци у ставци „Гинтер де Бројн“ (Q62753).

5.3 Постављање SPARQL упита

База Википодаци омогућава претрагу метаподатака постављањем SPARQL⁴³ упита у корисничком окружењу,⁴⁴ а као помоћ корисницима су на располагању бројни примери упита који се могу искористити тако да корисници који и не познају SPARQL језик могу да задовоље своје информационе потребе. Резултати постављених упита могу се приказати на више начина: табела (*table*), слика (*image grid*), графикон (*graph builder*), карта (*map*), линијски графикон (*line chart*), стубичасти дијаграм (*bar chart*), расути дијаграм (*scatter chart*), графикон површи (*area chart*), мехурасти дијаграм (*bubble chart*), мапа дрвета (*tree map*), дрво (*tree*), временска скала (*timeline*), димензионе коцке (*dimensions*) или као интерактивни граф (*graph*). SPARQL упите илустровали смо на два примера. Први пример излистава сва издања романа Буриданов магарац са подацима о аутору, месту издавања, датуму издавања и облику стваралачког дела.⁴⁵ Резултате смо приказали на два начина: као табелу (слика 6)⁴⁶ и интерактивни граф (слика 7)⁴⁷.

43. Simple Protocol and RDF Query Language

44. Wikidata Query Service

45. SPARQL упит, први пример

46. Визуализација првог SPARQL упита као табела

47. Визуализација првог SPARQL упита као граф

Пример 1:

```
#defaultView:Table #defaultView:Graph
SELECT ?izdanja ?izdanjaLabel ?autor ?autorLabel ?mestoIzdavanja
?mestoIzdavanjaLabel ?datumIzdavanja ?oblik ?oblikLabel
WHERE { #izdanja (P629) romana Buridanov magarac (Q110637016)
  ?izdanja wdt:P629 wd:Q110637016;
  #autor (P50) Ginter de Brojn(Q62753)
  wdt:P50 ?autor;
  #mesto izdavanja
  wdt:P291 ?mestoIzdavanja;
  #datum izdavanja
  wdt:P577 ?datumIzdavanja;
  #oblik stvaralачkog dela
  wdt:P7937 ?oblik.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],sr".}
LIMIT 100
```

izdanja	izdanjaLabel	autor	autorLabel	mestoIzdavanja	mestoIzdavanjaLabel	datumIzdavanja	oblik	oblikLabel
Q wd:Q11722033 9	Буриданов магарач : роман (издање из 2007. године)	Q wd:Q62753	Гинтер де Бројн	Q wd:Q3711	Београд	Januau 1, 2007	Q wd:Q8261	роман
Q wd:Q11722053 7	Буриданов магарач : роман : СрпНемКор издање	Q wd:Q62753	Гинтер де Бројн	Q wd:Q3711	Београд	Januau 1, 2019	Q wd:Q461183	издуск текстова
Q wd:Q11722104 6	Буриданов магарач : роман (издање из 1992. године)	Q wd:Q62753	Гинтер де Бројн	Q wd:Q1794	Франкфурт на Мајни	Januau 1, 1992	Q wd:Q8261	роман

Слика 6. Визуализација првог SPARQL упита у облику табеле.



Слика 7. Визуализација првог SPARQL упита у облику интерактивног графа.

Други пример излистава све главне ликове који се појављују у роману Буриданов магарац са подацима о наслову романа и аутору.⁴⁸ Резултате смо приказали као дрво (слика 8).⁴⁹

Визуализација SPARQL упита у облику интерактивног графа омогућава не само да се виде изјаве (ставка-својство-вредност) које су резултат постављеног упита, већ је могуће сагледати све скупове изјава и њихове везе које у произведеном резултату постоје. На слици 9 показали смо интерактивни граф романа Буриданов магарац који је у раду анализиран.

Пример 2:

```
#defaultView:Tree
SELECT ?likovi ?likoviLabel ?pol ?polLabel ?roman ?romanLabel ?tvorac
?tvoracLabel
WHERE {
  #likovi (P1441) romana Buridanov mgarac (Q110637016)
  ?likovi wdt:P1441 wd:Q110637016;
  #pol (P21)
  wdt:P21 ?pol;
  #pojavljuje se u romanu Buridanov mgarac
  wdt:P1441 ?roman;
  #tvorac (P170) Ginter de Brojn
  wdt:P170 ?tvorac.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],sr".}
LIMIT 100
```

6. Закључак и будући рад

У раду смо показали како поједини принципи и смернице семантичког веба могу бити примењени на један паралелни корпус књижевних текстова који садржи и српски језик. Увезивање метаподатака из корпуса са записима у релевантним семантичким базама подата из облака LOD или генерисање двојезничног речника као скупа повезаних отворених података на основу садржаја корпуса или илустрација метаподатака из корпуса Википодацима само су неки од начина да се садржај једног паралелног корпуса прикаже у широкој мрежи повезаних отворених података.

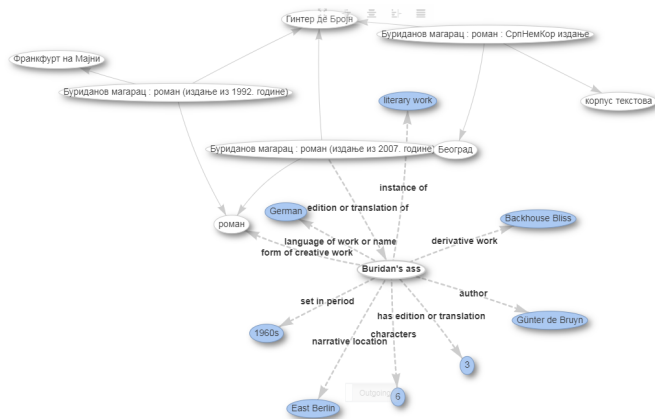
Поред резултата истраживања приказаних раду, постоји још много могућности примене семантичког веба на оваквом лингвистичком ресурсу. У будућој фази истраживања наставићемо рад на

48. SPARQL упит, други пример

49. Визуализација другог SPARQL упита као дрво

- ▲ |||| Карл Ерп
 - ▲ |||| мушки
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн
- ▲ |||| Елизабет Ерп
 - ▲ |||| женски
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн
- ▲ |||| Петер Ерп
 - ▲ |||| мушки
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн
- ▲ |||| Катарина Ерп
 - ▲ |||| женски
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн
- ▲ |||| Госпођица Бродер
 - ▲ |||| женски
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн
- ▲ |||| Тео Хаслер
 - ▲ |||| мушки
 - ▲ |||| Буриданов магарац : роман
 - |||| Гинтер де Бројн

Слика 8. Визуализација другог SPARQL упита у облику дрвета.



Слика 9. Интерактивни граф ставке „Буриданов магарац: роман“ (Q110637016).

припремљеном речнику *SrpNemLex*, пре свега на изради детаљнијег заглавља са метаподацима о речнику и објављивање у облаку LOD, али и истраживање и анализу о фреквентности појављивања одредница из речника у корпусу *SrpNemKor* и Корпусу савременог српског језика.⁵⁰

Поред тога, наставићемо креирање Википодатака за преостале романе из колекције *SrpNemKor* по узору на роман „Буриданов магарац“, али примењујући аутоматски начин учитавања метаподатака који је био примењен на романе из корпуса *srpELTeC*. Поред тога, биће анализирана анотација именованих ентитета у корпусу и могућности увезивања са одговарајућим базама знања како би садржај корпуса постао део ширег окружења које омогућава лакше проналажење информација и различите начине приказа и исписивање садржаја. Могућност да се из ширег облака LOD дође до садржаја на српском језику допринеће бољој заступљености српског језика на вебу.

Литература

Andonovski, Jelena, Šandrih, and Olivera Kitanović. 2019. “Bilingual lexical extraction based on word alignment for improving corpus search.” *The Electronic Library* 37 (4): 722–739.

50. Корпус савременог српског језика

- Bosque-Gil, Julia, and Jorge et al. Gracia, eds. 2019. *The OntoLex Lemon Lexicography Module: Final Community Group Report 13 September 2019*. Преузето 13.3.2023, <https://www.w3.org/2019/09/lexicog/>. W3C.
- Buitelaar, Paul, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. “Towards linguistically grounded ontologies.” In *ESWC2009, LNCS 5554, Towards linguistically grounded ontologies*, edited by L. Aroyo et al., 111–125. Преузето 14.03.2019, https://link.springer.com/content/pdf/10.1007/978-3-642-02121-3_12.pdf. Berlin, Heidelberg: Springer.
- Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. “Ontology lexicalisation: The lemon perspective.” In *WS 2 Workshop Extended Abstracts, 9th Int. Conf. on Terminology and Artificial Intelligence, TIA 2011, Paris, 10 November 2011*, 33–36. http://oa.upm.es/9772/1/Ontology_Lexicalisation.pdf.
- Chiarcos, Christian, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. “On the Linguistic Linked Open Data Infrastructure.” In *Proc. of the 1st International Workshop on Language Technology Platforms*, 8–15. Marseille, France: ELRA. ISBN: 979-10-95546-64-1. <https://aclanthology.org/2020.iwltpl-1.2>.
- Chiarcos, Christian, and Antonio Pareja-Lora. 2019. “Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction.” In *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, 1–17. The MIT Press.
- Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020b. *Linguistic Linked Data*. Springer International Publishing.
- Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020a. “Linguistic Linked Open Data Cloud.” In *Linguistic Linked Data: Representation, Generation and Applications*, 29–41. Springer, Cham. https://doi.org/10.1007/978-3-030-30225-2_3.
- Cimiano, Philipp, John P. McCrae, and Paul Buitelaar. 2016. *Lexicon Model for Ontologies: Community Report, 10 May 2016: Final Community Group Report 10 May 2016*. Преузето 13.3.2023, <https://www.w3.org/2016/05/ontolex/>. W3C.

- Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. “Introducing Wikidata to the Linked Data Web.” In *The Int. Semantic Web Conf.– ISWC 2014*, edited by Peter et al. Mika, 50–65. Cham: Springer International Publishing. ISBN: 978-3-319-11964-9.
- Gracia, Jorge, Marta Villegas, Asunción Gómez-Pérez, and Núria Bel. 2018. “The apertium bilingual dictionaries on the web of data.” *Semantic Web* 9 (2): 231–240.
- Gracia, Jorge, and Daniel Vila-Suero. 2015. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries. Final Community Group Report 29 September 2015*. Преузето 13.3.2019, <https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/>. W3C.
- Hernández, Daniel, Aidan Hogan, and Markus Krötzsch. 2015. “Reifying RDF: What Works Well With Wikidata?” In *11th Int. Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2015) At the 14th Int. Semantic Web Conf. (ISWC2015)*, 32–47.
- Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. “Serbian ELTeC Sub-Collection in Wikidata.” *Infotheca* 21 (2): 60–87. <https://doi.org/10.18485/infotheca.2021.21.2.4>.
- Isaac, Antoine, and Ed Summers. 2009. *Simple Knowledge Organization System Primer*. Working Group Note, W3C. Преузето 14.3.2023, <https://travesia.mcu.es/server/api/core/bitstreams/3f55d0de-b342-4d5d-8016-013b52e81707/content>.
- Janik, Maciej, Ansgar Scherp, and Steffen Staab. 2011. “The Semantic Web: Collective Intelligence on the Web.” Преузето 10.3.2023, <http://www.ansgarscherp.net/publications/pdf/J07-JanikScherpStaab-TheSemanticWeb-Preprint.pdf>, *Informatik-Spektrum* 34 (5): 469–483.
- Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. “ISocat: Corralling data categories in the wild.” In *Proc. of the 6th Int. Conf. LREC'08*, 887–891. Преузето 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf. ELRA.

- Koehn, Philipp, Franz Josef Och, and Daniel Marc. 2003. “Statistical phrase-based translation.” In *Proc. of the 2003 Conf. of the North American Chapter of the ACL on Human Language Technology-Volume 1*, 1:48–54. Association for Computational Linguistics.
- Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Belgrade: Faculty of Philology, University of Belgrade,
- Krstev, Cvetana, Branislava Sandrih, Ranka Stanković, and Miljana Mladenović. 2018. “Using English Baits to Catch Serbian Multi-Word Terminology.” In *Proc. of the 11th Int. Conf. LREC 2018*, edited by Nicoletta Calzolari et al., 2487–2494. Preuzeto 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf. ELRA.
- Lindemann, David, Penny Labropoulou, and Christiane Klaes. 2022. “Introducing LexMeta: a metadata model for lexical resources.” In *Dictionaries and Society. Proc. of the XX EURALEX Int. Congress, 12–16 July 2022, Mannheim, Germany*, 310–320. IDS-Verlag.
- McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, and Jorge et al. Gracia. 2012. “Interchanging lexical resources on the Semantic Web.” Preuzeto 14.03.2023, <https://link.springer.com/content/pdf/10.1007%2Fs10579-012-9182-3.pdf>, *Language Resources and Evaluation* 46 (4): 701–7019.
- McCrae, John Philip et al. 2016. “The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud.” In *Proc. of the 10th Int. Conf. LREC’16*, 2435–2441. ELRA. <https://doi.org/10.5167/uzh-132353>.
- Och, Franz Josef, and Hermann Ney. 2003. “A Systematic Comparison of Various Statistical Alignment Models.” *Computational linguistics* 29 (1): 19–51.
- Open DE WordNet Initiative*. 2023. Pristupljeno 13.3.2023, <https://ikum.medien-campus.h-da.de/projekt/open-de-wordnet-initiative/>.
- Radulovic, Filip, María Poveda-Villalón, Daniel Vila-Suero, Víctor Rodríguez-Doncel, Raúl García-Castro, and Asunción Gómez-Pérez. 2015. “Guidelines for Linked Data generation and publication: An example in building energy consumption.” *Automation in Construction*, no. 57, 178–187.

- Rurarz, Daniel, Aneta Poniszewska-Maranda, Iryna Ivanochko, and Oksana Urikova. 2021. "Use of Semantic Networks to Search for Information." In *Developments in Information & Knowledge Management for Business Applications*, 1:93–199. Springer, Cham. https://doi.org/10.1007/978-3-030-62151-3_3.
- Schreiber, Guus, and Yves Raimond. 2014. *RDF 1.1 Primer*. W3C Note. <https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>. W3C, June.
- Shah, Urvi, Tim Finin, Anupam Joshi, R Scott Cost, and James Matfield. 2002. "Information Retrieval on the Semantic Web." In *Proceedings of the eleventh international conference on Information and knowledge management*, 1:461–468. <https://doi.org/10.1145/584792.584868>.
- Stanković, Ranka, and Lazar Davidović. 2021. "Infotheca (Q25460443) in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. ISSN: 2217-9461. <https://doi.org/10.18485/infotheca.2021.21.1.5>. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.1.5_en.
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries – from File System to *lemon* Based Lexical Database." In *Proc. of the 11th Int. Conf. LREC 2018 - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), Miyazaki, Japan, May 7-12, 2018*, edited by John P. et al McCrae. ELRA. http://lrec-conf.org/workshops/lrec2018/W23/pdf/3_W23.pdf.
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. "A bilingual digital library for academic and entrepreneurial knowledge management." In *Proc. of 10th International Forum on Knowledge Asset Dynamics-IFKAD*, 1764–1777. Preuzeto 13.03.2023, http://poincare.matf.bg.ac.rs/~\unhbox\voidb@x\protect\penalty\@M\cvetana/biblio/IFKAD_RS_CK_BL_DV-fin.pdf.
- Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. "Rule-based Automatic Multi-word Term Extraction and Lemmatization." In *Proc. of the 10th Int. Conf. LREC'16*, 507–514. ELRA. <https://aclanthology.org/L16-1081>.

- Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. “A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals.” In *Proc. of the 8th Int. Conf. LREC 2012*, edited by Nicoletta et al. Calzolari, 1710–1717. Preuzeto 16.04.2016, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf. Istanbul: ELRA.
- Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2017. “Keyword-Based Search on Bilingual Digital Libraries.” In *Semantic Keyword-Based Search on Structured Data Sources*, edited by Andrea Cali, Dorian Gorgan, and Martín Ugarte, 112–123. Springer International Publishing. ISBN: 978-3-319-53640-8.
- Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2022. “The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection.” *Infotheca* 21 (2): 7–25. <https://doi.org/10.18485/infotheca.2021.21.2.1>.
- Vrandečić, Denny, and Markus Krötzsch. 2014. “Wikidata: a free collaborative knowledgebase.” *Communications of the ACM* 57 (10): 78–85. <https://doi.org/10.1145/2629489>.
- Андоновски, Јелена С. 2019. “Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса.” Докторска дисертација, Универзитет у Београду, Филолошки факултет.
- Андоновски, Јелена С. 2021. *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семантички веб*. Београд: Универзитетска библиотека „Светозар Марковић“.
- Крстев, Цветана. 2019. “О одабиру одредница за електронски речник српског језика и њиховом повезивању.” In *Научни састанак слависта у Вукове дане*, 48:133–147. Београд : Филолошки факултет, Међународни славистички центар. <https://doi.org/10.18485/msc.2019.48.3.ch7>.
- Станковић, Ранка, and Лазар Давидовић. 2021. “Инфотека (Q25460443) у Википодацима.” Преузето 17.3.2023, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.1.5_sr, *Инфотека* 21 (1): 93–104.