Оригиналан научни рад
UDK 027.7:004.4(497.11)
811.163.41'33
001.103.2
https://doi.org/10.18485/bibliotekar.2021.63.1.3

Јелена С. Андоновски
Универзитетска библиотека
„Светозар Марковић", Београд
andonovski@unilib.rs

ПАРАЛЕЛНИ КОРПУСИ У СРБИЈИ[1] – МОГУЋНОСТИ ЗА ПАРАЛЕЛНО ПРОНАЛАЖЕЊЕ ИНФОРМАЦИЈА НА ДВА ИЛИ ВИШЕ ЈЕЗИКА

Сажетак: Паралелни корпуси представљају врсту вишејезичних корпуса који су последњих деценија постали изузетно значајни у области обраде природних језика (енгл. Natural Language Processing – NLP) и један од важнијих ресурса за истраживаче у различитим областима лингвистике и сродним језичким дисциплинама. Под паралелним корпусима подразумевају се језички корпуси који садрже један текст или више оригиналних текстова и њихове преводе на један језик или више језика, поравнате на једном нивоу или више структурних нивоа текста (на пример, на нивоу реченице, пасуса и одељка). Они су најчешће двојезични, али није ретко да постоје и на једном језику што подразумева да корпусни садржај чине различита издања истог текста на одабраном језику. Паралелне корпусе који обухватају српски језик у Србији развија Група за језичке технологије која је у међувремену прерасла у Друштво за језичке ресурсе и технологије – JePTex. До данас су развијени следећи корпуси: два већа корпуса, српско-француски (СрпФранКор) и српско-енглески (СрпЕнгКор) корпус, затим, дигитална библиотека Библиша која садржи више паралелних двојезичних колекција и вишејезична колекција *Вишејезични Верн*. Поред ових корпуса текстови на српском језику део су и вишејезичних корпуса Платонова *Република* и Орвелова *1984* који су развијени у оквиру међународних пројеката, али и неких корпуса који се тренутно развијају у региону и свету. У раду ће бити приказани корпуси које развија Друштво за језичке ресурсе и технологије, њихова структура и намена, као и могућности за проналажење информација у њима.

[1] Рад представља прерађени текст о паралелним корпусима у Србији који је детаљније разрађен у докторској дисертацији: Ј. Андоновски, „Мрежа повезаних отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса" (докторска дис., Филолошки факултет, Београд, 2019), http://phaidrabg.bg.ac.rs/o:22874

Кључне речи: корпусна лингвистика, језички корпуси, паралелни корпуси, обрада природних језика, проналажење информација.

Увод

Језички корпуси представљају један од важнијих ресурса за истраживања у области лингвистике и сродним језичким дисциплинама. У ужем смислу дефинишу се као велика колекција текстова,[2] док у најширем смислу представљају емпиријски материјал намењен истраживању језика.[3] Они се користе у свим лингвистичким дисциплинама као помоћно средство за анализу књижевно-уметничких дела или текстова који припадају неком другом посебном функционалном стилу (новински, административни, разговорни, научни и научно-популарни), при чему посебну улогу заузимају у статистичкој анализи језика. Неке од области лингвистике у оквиру којих се истраживања заснивају на корпусима су: лексикографија, социолингвистика, анализа дискурса, морфологија, фонологија, семантика, синтакса, компаративна и контрастивна лингвистика, методика наставе, когнитивна лингвистика.[4]

У зависности од нивоа анотације, корпуси омогућавају истраживачима да уоче различите примере употребе језика, да приликом претраге сагледају фреквентност појављивања одређене речи или фразе постављене кроз упите за претрагу, у којим се све облицима и варијантама задата реч или фраза појављује, као и у каквој је семантичкој корелацији са другим речима и фразама и њиховим облицима у датом корпусу. Такође, са новим технологијама повезивања садржаја на вебу (Linked Open Data[5]) могуће је сагледати и фреквентност појављивања речи или фразе у разним базама

---

2 Tony McEnery and Andrew Wilson, *Corpora and translation: uses and future prospects*, 1993, http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf (преузето 1. 3. 2021).

3 Душко Витас и Љубомир Поповић, „Конспект за изградњу референтног корпуса српског стандардног језика", у *Научни састанак слависта у Вукове дане* (Београд: МСЦ, 2003), 221–227.

4 Nikola Dobrić, „Corpus linguistics – the basic form of linguistic analysis", *Philologiano* 7 (2009): 40, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309960 (преузето 1. 3. 2021).

5 Jonathan Blaney, „Introduction to the Principles of Linked Open Data", *The Programming Historian* (12. 5. 2020), https://programminghistorian.org/en/lessons/intro-to-linked-data (преузето 1. 3. 2021), https://doi.org/10.46430/phen0068.

података, репозиторијумима, онтологијама,[6] контролисаним речницима[7] и сличним изворима.

Приликом креирања корпуса група истраживача и техничких стручњака одређују његов садржај, методологију рада, језик на коме ће бити заступљени текстови, период који ће бити обухваћен, што све одређује врсту корпуса. Класификација корпуса може се извршити на основу више параметара од којих су најважнији носач, обим (величина), домен, намена, период, извор, начин анотације и број укључених језика те се тако корпуси могу поделити на: пределектронске и електронске, опште и специјализоване, статичке и динамичке, синхроне и дијахроне, говорне или текстуалне, неетикетиране и етикетиране (анотиране), једнојезичне и вишејезичне, при чему се вишејезични даље могу поделити на паралелне и упоредне.[8]

Под паралелним корпусима подразумевају се корпуси који садрже један текст или више оригиналних текстова и њихове преводе на један језик или више језика поравнате на једном ниову или више структурних нивоа текста (на пример, на нивоу реченице, пасуса и одељка).[9] То су најчешће двојезични паралелни корпуси, али није ретко да постоје и корпуси са паралелним текстовима на једном језику који подразумевају да корпусни садржај чине различита издања истог текста на одабраном језику било да су у питању различита издања текста који је оригинално настао на том језику или су то различита издања превода неког текста. Њихова предност је што садрже преводе истог текста на два или више језика, што даље омогућава истраживачима да сагледају све предности корпуса, претходно наведене у овом поглављу, упоредо на два или више језика. Поред лингвистичких области, постали су значајни и у практичним језичким областима као што су превођење, терминолошка екстракција или производња преводилачких

---

[6] Онтологија је машински читљив речник у одговарајућем формату чији елементи (класе и својства) имају јасно дефинисане типове и двосмерне логичке везе које омогућавају повезивање са другим елементима (класама и својствима). Термин је преузет из филозофије и представља формалну репрезентацију знања у некој области.

[7] Контролисани речници (нормативне датотеке, речници, тезауруси и др.) развијени су за различите научне области и области људског знања. Садрже структуриране податке, а у библиотечкој заједници се вековима развијају како би се направио неки унифицирани систем за индексирања људског знања које би било претраживо са једног места.

[8] Miloš Utvić, „Izgradnja referenctnog korpusa savremenog srpskog jezika" (doktorska dis., Filološki fakultet, Beograd, 2013), 24–31, http://phaidrabg.bg.ac.rs/o:10061 (преузето 1. 3. 2021).

[9] Luzius Töny, *Corpora als Ressourcen für die maschinelle Übersetzung* (Seminar Maschinelle *Übersetzung*) 11, http://www.swanrad.ch/downloads/mt_1.pdf (преузето 1. 3. 2021).

меморија,[10] а последњих деценија доста је рађено и на развоју система за машинско превођење који користе управо упарене јединице[11] паралелних корпуса за изградњу модела превођења.

Развој корпусне лингвистике у Србији почиње већ средином прошлог века, а њен интензиван развој наставља се седамдесетих и осамдесетих година када је на Математичком институту Српске академије наука и уметности[12] започео са радом стални Семинар за математичку и рачунарску лингвистику, чији је оснивач и руководилац др Душко Витас, професор Математичког факултета Универзитета у Београду. У оквиру ове школе настаје и Група за језичке технологије која окупља истраживаче са Универзитета у Београду[13] и која је своју делатност усмерила на развој језичких ресурса и алата за аутоматску обраду српског језика, поред осталог, језичких корпуса (једнојезичних и вишејезичних) и електронских речника.[14] Група за језичке технологије Универзитета у Београду у међувремену је прерасла у Друштво за језичке ресурсе и технологије (у даљем тексту: JePTex).

У раду ће бити приказани паралелни корпуси које развија JePTex последњих деценија, њихове карактеристике и структура, као и могућности проналажења информација у њима. Важно је напоменути да припрема паралелних корпуса захтева много времена како у погледу одабира и прикупљања материјала, тако и у погледу техничке припреме, обраде и паралелизације одабраних текстова. Студенти се на Катедри за библиотекарство и информатику на Филолошком факултету у Београду већ током студија срећу са паралелним корпусима у оквиру информатичких

---

[10] Eric Laporte, Duško Vitas and Cvetana Krstev, „Preparation and exploitation of Bilingual Texts", *Lux Coreana* No. 1(2006): 110, http://poincare.matf.bg.ac.rs/~cvetana/biblio/VKL.pdf (преузето 1. 3. 2021).

[11] Упарене јединице представљају семантички еквивалентне верзије истог текста на два или више језика (текст и његов превод). То су структурни елементи текста (пасуси, реченице или речи) који су анотирани на одговарајући начин и између њих је успостављена експлицитна веза применом одговарајућег софтвера.

[12] Математички институт Српске академије наука и уметности, http://www.mi.sanu.ac.rs/

[13] Група за језичке технологије, Математички факултет, Катедра за рачунарство и информатику, http://www.racunarstvo.matf.bg.ac.rs/?content=nauka_jezicke_tehnologije

[14] О развоју корпусне лингвистике у Србији више у: Ј. Андоновски, „Мрежа повезаних отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса" (докторска дис., Филолошки факултет, Београд, 2019), 42-47, http://phaidrabg.bg.ac.rs/o:22874 (преузето 1. 3. 2021); M. Utvić, „Izgradnja referenctnog korpusa savremenog srpskog jezika" (doktorska dis., Filološki fakultet, Beograd, 2013), 52–55, http://phaidrabg.bg.ac.rs/o:10061 (преузето 1. 3. 2021).

предмета. Такође, настава информатике омогућава студентима да савладају и технике електронске обраде текстова и различите врсте анотација, како логичких структура, тако и садржаја електронског текста, као и да се упознају са структуром система за проналажење информација. Ово им касније, у пракси, омогућава да учествују у изради паралелних корпуса и система за њихову претрагу.

Паралелни корпуси у Србији

Развој паралелних корпуса у Србији започет је учешћем Групе за језичке технологије у пројекту TELRI (Trans-European Language Resources Infrastructure).[15] Резултат рада на пројекту је CD „East Meets West – A Compendium of Multilingual Resources"[16] у два тома. Садржај првог тома настао је у потпуности у оквиру акције TELRI када је припремљен вишејезични паралелни корпус дела *Рейублика* грчког филозофа Платона. Текст *Рейублике* је паралелизован на 21 европски језик, а у оквиру пројекта урађена је и паралелизација са српским преводом овог дела.[17] Други том добијеног CD-а садржи резултате европског пројекта MULTEXT-East[18] у оквиру кога је развијен вишејезични паралелни корпус Орвелова 1984[19] и разноврсни језички алати за језике који су били део пројекта међу којима

---

[15] Trans-European Language Resources Infrastructure, http://telri.nytud.hu/

[16] Tomaž Erjavec, Ann Lawson, and Laurent Romary, eds., *East meets West: A Compendium of Multilingual Resources* (Mannheim: TELRI Association, e.V., Institut für deutsche Sprache, 1998).

[17] Duško Vitas, Goran Nenadić and Cvetana Krstev, „[Electronic edition of Serbian translation of Plato's Republic aligned with 17 languages by Duško Vitas, Goran Nenadić, Cvetana Krstev]", in *East meets West – A compendium of Multilingual Resources*, eds. Tomaž Erjavec, Ann Lawson, Laurent Romary (Mannheim: TELRI Association e.V., Institut für deutsche Sprache, 1998).

[18] Ludmila Dimitrova, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaaler and Dan Tufis, „Multext-east: Parallel and comparable corpora and lexicons for six central and Eastern European languages", in *Volume 1 Proceedings of the 17th international conference on Computational linguistics* (Association for Computational Linguistics, 1998), http://www.aclweb.org/anthology/P98-1050 (преузето 2. 3. 2021).

[19] Tomaž Erjavec and Nancy Ide, „The MULTEXT-East Corpus", in *Proceeding of First International Conference on Language Resources & Evaluation, Granada, Spain, 28–30 May* (1998), 971–974. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.5846&rep=rep1&type=pdf; (преузето 2. 3. 2021)

су посебно значајни морфосинтаксички речници.[20] Српски језик постао је део корпуса реализацијом пројекта TELRI, односно у оквиру треће верзије пројекта MULTEXT-East која је изашла 2004. године[21] када је урађен структурно анотирани корпус српске верзије романа *1984* и паралелизација са осталим језицима[22] и када су припремљени морфосинтаксички речници за српски језик.[23]

Након учешћа у овим пројектима и користећи стечена искуства JeP-Tex је наставио да ради на развоју паралелних корпуса који обухватају српски језик тако да данас постоје два већа корпуса, српско-француски (СрпФранКор) и српско-енглески (СрпЕнгКор), затим вишејезична колекција Вишејезични Верн, дигитална библиотека Библиша која садржи више паралелних двојезичних колекција, као и српско-српски/хрватски паралелни корпус који је још у развоју. Ови паралелни корпуси примарно су развијени за потребе лингвистичких и лексикографских истраживања и у њихов садржај је укључен и знатан број књижевних текстова.

Српско-француски корпус (СрпФранКор)

Српско-француски корпус – СрпФранКор, први је паралелни двојезични корпус на коме је започет рад у Србији после учешћа у међународним пројектима TELRI и MULTEXT-East. Највећи део корпуса чине класична дела француске књижевности настала од краја 18. века па до данас, затим новински текстови (важан део чине текстови преузети из

---

[20] Tomaž Erjavec, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić and Duško Vitas, „The MULTEXT-East Morphosyntactic Specifications for Slavic Languages", in *Proceedings of the Workshop on Morphological Processing of Slavic Languages: 10th Conference of the European Chapter, EACL,* eds. Tomaž Erjavec and Duško Vitas (Budapest, 2003), 25–32, http://poincare.matf.bg.ac.rs/~cvetana/biblio/04erjavec.pdf (преузето 2. 3. 2021).

[21] Tomaž Erjavec, „MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora", in *LREC* (2004), 2544–2547, http://nl.ijs.si/et/teach/jsi07-hlt/Bib/Multext_LREC04.pdf (преузето 2. 3. 2021); Multext-East Resources, Version 3, http://nl.ijs.si/ME/Vault/V3/

[22] Duško Vitas, Cvetana Krstev, „[Electronic edition of Serbian translation of Orwell's 1884 aligned with 7 languages by Duško Vitas, Cvetana Krstev]", in *East meets West – A compendium of Multilingual Resources*, eds. Tomaž Erjavec, Ann Lawson, Laurent Romary (Mannheim: TELRI Association e.V., Institut für deutsche Sprache, 1998).

[23] Cvetana Krstev, Duško Vitas and Tomaž Erjavec, „Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian", *Informatica* Vol. 28, No. 4 (2004): 431–436, http://poincare.matf.bg.ac.rs/~cvetana/biblio/mtesr-inform04.pdf (преузето 2. 3. 2021).

француског часописа „Le Monde Diplomatique"[24]), али и неки савремени текстови из области филозофије, социологије, етнологије и науке.[25] Према последњим подацима корпус садржи 31 књижевни текст од чега је 28 текстова оригинално написано на француском језику и преведено на српски (један са два превода), два текста оригинално написана на српском језику која су преведена на француски и један енглески роман који је преведен на француски и српски језик.[26] Прве анализе корпусног садржаја показале су да се у добијеном корпусу могу пронаћи решења за многе преводе која не постоје у двојезичним француско-српским речницима. Такође, корпус даје анализу стратегије превођења која даље омогућава решавање лексичког јаза или двосмислености у оригиналном тексту, као и проналажење недоследности у преводу.

Поред овог корпуса важно је поменути и француско-српско-енглески корпус (ParColLab[27]), настао као резултат вишегодишње сарадње лингвиста и информатичара, као и стручњака за рачунарску обраду природних језика из Француске и Србије. Намењен је истраживачима за проучавање различитих области лингвистике ових трију језика, наставним и педагошким курикулумима (употреба у оквиру наставе и учења ових језика као страних, школовању преводилаца, припреми наставног материјала) и бесплатно је доступан преко интернета уз отворен кориснички налог.

Српско-енглески корпус (СрпЕнгКор)

Након рада на корпусу СрпФранКор, Група за језичке технологије започела је рад и на паралелном српско-енглеском корпусу, *СрпЕнгКор*.

---

[24] Le Monde Diplomatique, http://www.monde-diplomatique.fr/

[25] Duško Vitas and Cvetana Krstev, "Literature and Aligned Texts", in *Readings in Multilinguality*, eds. Milena Slavcheva, Galia Angelova and Kiril Simov (Sofia: Institute for Parallel Processing, Bulgarian Academy of Sciences, 2006), 148–155, http://poincare.matf.bg.ac.rs/~cvetana/biblio/CvDv-Paskaleva.pdf (преузето 2. 3. 2021).

[26] Подаци су преузети са приступне странице СрпФранКор-а. http://www.korpus.matf.bg.ac.rs/SrpFranKor/ (преузето 2. 3. 2021).

[27] ParColLab, http://parcolab.univ-tlse2.fr/en/; Antonio Balvet, Dejan Stosic and Aleksandra Miletic, "TALC-Sef a Manually-revised POS-Tagged Literary Corpus in Serbian, English and French", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (Reykjavik: European Language Resources Association, 2014), 26–31, http://www.lrec-conf.org/proceedings/lrec2014/index.html (преузето 2. 3. 2021).

Први текст у овом корпусу био је роман *1984* Џорџа Орвела, преузет из истоименог вишејезичног корпуса, а рад је настављен паралелизацијом дела ауторке Џејн Остин (Jane Austen).[28] У току даљег рада корпус је допуњен и другим делима класичне енглеске књижевности – Томаса Хардија (Thomas Hardy) и Ернеста Хемингвеја (Ernest Hemingway), али и делима савремених писаца као што су Ден Браун (Dan Brown), Џ. К. Роулинг (J. K. Rowling) и други. Поред дела на енглеском језику која су преведена на српски, део корпуса постала су и дела наших писаца преведених на енглески – Данила Киша, Драгана Великића, Светислава Басаре и других.

Поред књижевних дела саставни део корпуса постали су и новински чланци из корпуса SETimes[29] при чему је формиран поткорпус BALKANTIMES, који садржи вести из Југоисточне Европе на десет језика, а у оквиру пројекта Intera (Integrated European Language data Repository Area)[30] СрпЕнгКор допуњен је паралелним текстовима из области права, пословања, образовања и здравствене заштите при чему је формиран поткорпус SELFEH (Serbian-English Law Finance Education and Health).[31] [32]

Може се закључити да СрпЕнгКор садржи текстове оригинално написане на енглеском језику који су преведени на српски, текстове оригинално написане на српском језику који су преведени на енглески, као и поравнате енглеске и српске преводе текстова који су оригинално написани на француском језику.

---

[28] Cvetana Krstev and Duško Vitas, „Analigned English-Serbian corpus", in *Volume I ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Belgrade, 4–6 December 2009*, eds. N. Tomović & J. Vujić (Belgrade: Faculty of Philology, University of Belgrade, 2011), 495–508, http://poincare.matf.bg.ac.rs/~cvetana/biblio/AlignedCorpus-full-final.pdf (2. 3. 2021).

[29] Francis M. Tyers and Murat Serdar Alperen, „South-east European times: A parallel corpus of Balkan languages", in *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages* (2010), 49–53.

[30] Maria Gavrilidou, Peny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos and Stelios Piperidis, „Building parallel corpora for eContent professionals", in *MLR '04 Proceedings of the Workshop on Multilingual Linguistic Ressources* (Stroudsburg: Association for Computational Linguistics, 2004), 97–100, https://www.aclweb.org/anthology/W04-2213.pdf (преузето 2. 3. 2021).

[31] Serbian-English Law Finance Education and Health, http://www.korpus.matf.bg.ac.rs/prezentacija/selfeh.html

[32] Miloš Utvić, „Izgradnja referenctnog korpusa savremenog srpskog jezika" (doktorska dis., Filološki fakultet, Beograd, 2013), 252–253, http://phaidrabg.bg.ac.rs/o:10061 (преузето 2. 3. 2021).

Вишејезични Верн

Колекција Вишејезични Верн представља паралелни вишејезични корпус романа Жила Верна (Jules Verne) *Пут око света за 80 дана*, изграђен у оквиру самосталног пројекта који је реализован на Математичком и Филолошком факултету Универзитета у Београду. Овај роман одабран је зато што је Жил Верн најпревођенији француски аутор и други најпревођенији аутор на свету,[33] те су његова дела доступна у електронском облику на многим језицима.[34] Са друге стране, специфичан садржај самог текста погодан је за различите врсте анализа, посебно за поступак препознавања и обраде именованих ентитета.[35] [36] Корпус је иницијално сачињен од паралелних превода романа на шеснаест језика (бугарски, хрватски, енглески, француски, немачки, грчки, мађарски, италијански, македонски, пољски, португалски, румунски, руски, српски, словеначки, шпански), а данас садржи двадесет превода поравнатих са оригиналном француском верзијом текста, а у припреми су и преводи на холандском и кинеском језику, као и још по једна верзија превода на енглеском и немачком.

---

[33] Статистички подаци о фреквентности превођења аутора у свету доступни су на: http://www.unesco.org/xtrans/bsstatexp.aspx?crit1L=5&nTyp=min&topN=50

[34] Duško Vitas and Cvetana Krstev, „Construction and Exploitation of X-Serbian Bitexts", in *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation,* eds. Cristina Vertan and Walther v. Hahn (Cambridge: Cambridge Scholars Publishing, 2012), 207–227, http://poincare.matf.bg.ac.rs/~cvetana/biblio/DvCv-CambridgeS-2012.pdf (преузето 2. 3. 2021).

[35] Именовани ентитети представљају властита имена (имена особа, имена организација и локација), затим, изразе којима се описују датуми и време на часовнику и бројчане изразе (процентуалних и новчаних израза) који се могу пронаћи у тексту. Данас постоје развијени алати за многе језике који аутоматски врше екстракцију ових израза из одабраних текстова.

[36] Duško Vitas, Svetla Koeva, Cvetana Krstev and Ivan Obradović, „Tour du monde through the dictionaries", in *Actes du 27eme Colloque International sur le Lexique et la Gammaire,* eds. M. Constant, T. Nakamura, M. De Gioia, S. Vecchiato (Paris: Universite Paris-Est, Institut Gaspard-Monge, 2008), 249–256, http://poincare.matf.bg.ac.rs/~cvetana/biblio/akvila-en-fin.pdf (преузето 2. 3. 2021).

Дигитална библиотека Библиша

Библиша[37] је веб-апликација коју је развила Група за језичке технологије Универзитета у Београду ради унапређења могућности претраживања вишејезичних дигиталних библиотека електронских часописа.[38] Дигитална библиотека Библише тренутно садржи једанаест текстуалних колекција упарених докумената: пет колекција поравнатих чланака из домаћих научних часописа који излазе на српском и енглеском језику (*Инфотека*,[39] *Подземни радови*,[40] *Архитектура и урбанизам*,[41] *Стоматолошки гласник Србије*,[42] *Менаџмент*); две колекције поравнатих техничких извештаја са пројеката (BEAKTEL TEMPUS, CESAR); три паралелна доменска корпуса (Intera, EIEner – паралелни корпус текстова из области енергетике и Mining – паралелни корпус текстова из области рударства), и један корпус књижевних текстова (СрпНемКор – паралелни корпус књижевних текстова на српском и немачком језику).

Начин припреме текстова за Библишу, односно поступак паралелизације, не разликује се од поступка паралелизације текстова за поменуте корпусе. Користе се исте методе за обраду и анотацију текстова и исти софтвер за паралелизацију. Међутим, структура алата Библиша другачија је у односу на претходно наведене корпусе, сложена је и састоји се из неколико компонената: лексички ресурси (интегрисани за проширење корисничких упита за претрагу), текстуалне колекције (документи на два језика поравната до нивоа реченице), веб-сервиси (интегрисани за приступ лексичким ресурсима) и веб-сумеђа (енгл. *web interface*) (развијена

---

37 Biblisha, http://jerteh.rs/biblisha/

38 Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić, „A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals", in *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 23–25 May 2012,* eds. Nicoletta Calzolari et al. (Istanbul: European Language Resources Association, 2012), 1710, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf (преузето 3. 3. 2021).

39 *Инфотека*, http://infoteka.bg.ac.rs/index.php/sr

40 *Подземни радови*, http://www.rgf.rs/publikacije/PodzemniRadovi/?lang=sr

41 *Архитектура и урбанизам*, http://www.iaus.ac.rs/code/navigate.aspx?Id=

42 *Стоматолошки гласник*, http://www.stomglas.org.rs/

за потребе корисника).[43] За разлику од корпуса наведених у претходним одељцима, у Библиши је развијена одређена структура метаподатака за опис дигиталних објекта. Како је она иницијално креирана као дигитална библиотека електронских часописа, тако је и структура метаподатака првобитно била осмишљена да опише сам часопис, сваки број појединачно и сваки чланак појединачно. С временом структура је прилагођена и опису друге врсте материјала.

Поред различите структуре, дигитална библиотека Библиша разликује се и у могућностима претраге у односу на наведене корпусе о чему ће бити речи у наредном одељку.

Проналажење информација у паралелним корпусима

Као што је речено у уводу, језички корпуси представљају један од важнијих ресурса за истраживања у области лингвистике и сродним језичким дисциплинама. Корпуси поменути у овом раду могу се претраживати на различите начине и преко различитих параметара. Јако је важно нагласити да је за њихову претрагу неопходна ауторизација, тј. отварање корисничког налога од стране одговорног лица. Заинтересовани истраживачи на захтев и уз образложење добијају корисничко име и шифру за приступ корпусу. Такође, важно је напоменути да приликом претраге корпуса корисници ни у једном моменту не добијају приступ пуном садржају текстова[44] у корпусу, већ се кроз постављање упита за претрагу добијају резултати у виду конкорданци[45] које представљају листу појављи-

[43] Ranka Stanković, Cvetana Krstev, Ivan Obradović and Olivera Kitanović,„Indexing of Textual Databases Based on Lexical Resources: - A Case Study for Serbian", in *Semantic Keyword-Based Search on Structured Data Sources - First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, September 8–9* (Springer, 2015), 167–181, 2015. DOI 10.1007/978-3-319-27932-9_15; Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović, „Keyword-Based Search on Bilingual Digital Libraries", in *Semantic Keyword-Based Search on Structured Data Sources – Second COST Action IC1302 International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9,* eds. A. Calì, D. Gorgan and M. Ugarte, LNCS 10151 (Springer, 2017), 112–123.

[44] Изузетак могу бити текстови који се већ налазе у отвореном приступу. На пример, чланци у часописима који су у отвреном приступу, а у овом случају такве су двојезичне текстуалне колекције у дигиталној библиотеци Библиша, на пример, чланци часописа *Инфотека*.

[45] Конкорданца је метод за визуелизацију података из корпуса који одговарају корисниковом упиту.

вања токена[46] који су задовољили упит за претрагу, са позицијама у тексту и цитатима или изводима из контекста. Добијене конкорданце корисник може да прегледа страну по страну и приступи свакој страни у резултатима претраге, а може и да изабере оне које жели да задржи у приказу, као и да сачува издвојене конкорданце.

Начин претраге садржаја поменутих корпуса се разликује. Претрага корпуса који су део Корпуса савременог српског језика[47] (СрпФранКор и СрпЕнгКор) врши се преко регуларних израза.[48] Претрага преко регуларних израза биће илустрована на корпусу СрпЕнгКор примером „bibliotek[a-z]*", односно „librar[a-z]*". Применом овог регуларног израза на српском језику као резултат добијено је 3.616 погодака, док је применом регуларног израза на енглеском добијено 4.567 погодака. Уз све добијене конкорданце као резултат претраге на српском стоји паралелна конкорданца на енглеском и обрнуто. Токени који су задовољили упит за претрагу назначени су у сваком добијеном резултату (Слике 1 и 2).

[46] Токен представља појединачно појављивање неке језичке јединице чија фреквентност спада у поље квантитативне анализе.

[47] Miloš Utvić, „Izgradnja referenctnog korpusa savremenog srpskog jezika" (doktorska dis., Filološki fakultet, Beograd, 2013), http://phaidrabg.bg.ac.rs/o:10061 (преузето 1. 3. 2021).

[48] У области рачунарства и информатике регуларни изрази се дефинишу као ниска која описује, мења или упарује скуп ниски према одређеним синтаксичким правилима. У случају да корисници не знају да користе регуларне изразе за претрагу уз формулар за претрагу су наведени примери, као и веза ка упутству за коришћење регуларних израза.

1034239: Projekat Evropske Unije Tempus UMI _ JEP 16059 usmeren je na prevazilaženje navedenih problema , pružanjem finansijske pomoći za grupne posete srpskih <bibliotekara> evropskim univerzitetima .

EN: EU Tempus project UMI _ JEP 16059 is aiming to overcome stated problems by funding group visits to European universities for Serbian librarians

1034329: Grupu su činila dva <bibliotekara> gđa . Maja Đorđević iz Univerzitetske biblioteke u Beogradu i gđa . Snežana Jančić iz Univerzitetske biblioteke u Nišu, i dva sistem adminsitratora, g . Adam Sofronijević iz Univerzitetske biblioteke u Beogradu i gđa. Vesna Abadić iz Univerzitetske biblioteke u Kragujevcu .

EN: The group consisted of two librarians , Mrs . Maja Djordjevic from the University Library in Belgrade and Mrs . Snezana Jancic from the University Library in Nis , and two system administrators , Mr . Adam Sofronijevic from the University Library in Belgrade and Mrs . Vesna Abadic from the University Library in Kragujevac

1034401: Gđa . Sju Harli srdačno je dočekala <bibliotekare> , dok su g . Alan Hopkinson i njegove kolege gđa . Sju Blek , g . Aleks Burkhal i g . Endi Mekgregor bili strpljivi i stručni domaćini sistem administratorima.

EN: Ms . Sue Hurley warmly welcomed librarians while system administrators found a competent and patient guide in Mr . Alan Hopkinson , and his colleagues Mrs . Sue Black , Mr . Alex Burchall and Mr . Andy McGregor

1034248: Imajući u vidu da mladi <bibliotekari> nisu do sada imali priliku da razmene profesionalna iskustva sa svojim kolegama iz Evrope , lako je zaključiti koliko novostečeno iskustvo može postati važno i koliko ono može da doprinese profesionalnom razvoju svakog pojednica .

EN: Having in mind that young librarians have never had the chance to exchange professional experience with their counterparts in Europe , it is easy to draw a conclusion how significant this newly gained personal experience will become and how much it can contribute to personal professional development

Слика 1. Пример резултата претраге корпуса СрпЕнгКор упитом „bibliotek[a-z]*“:

EN: The Health Sciences Library , the Kathrine R . Everett Law Library , and several independent libraries , including the Carolina Population Center Library , Grant Source Library , Highway Safety Research Center Library , Odum Insistute Library , Park Library of Journalism and Mass Communication , Occupational Information Library , and Sociology and Political Science Reading Rooms , complete the campus network
2541341: Biblioteka zdravstvenih nauka , zatim Katherine R . Everett Law Library , i nekoliko nezavisnih biblioteka , koje uključuju Carolina Population Center Library , Grant Source <Library> , Highway Safety Research Center Library , Odum Insistute Library , Park Library of Journalism and Mass Communication , Occupational Information Library , and Sociology and Political Science Reading Rooms , upotpunjuju mrežu kampusa .
EN: The Health Sciences Library , the Kathrine R . Everett Law Library , and several independent libraries , including the Carolina Population Center Library , Grant Source Library , Highway Safety Research Center Library , Odum Insistute Library , Park Library of Journalism and Mass Communication , Occupational Information Library , and Sociology and Political Science Reading Rooms , complete the campus network
2621988: Međutim , prva biblioteka na aerodromu otvorena je na aerodromu u Nešvilu , u SAD , 1962 . godine (Nashville public library - <library> history) , u saradnji javne biblioteke u Nešvilu i uprave aerodroma u Nešvilu .

Слика 2. Пример резултата претраге корпуса СрпНемКор упитом „librar[a-z]*“:

Претрага дигиталне библиотеке Библиша разликује се од претраге Корпуса савременог српског језика. И Библиша нуди могућност регистрације корисника, с тим што Библишу могу претраживати и нерегистровани корисници, али они виде мањи број поравнатих сегмената у поређењу са регистрованим корисницима. Детаљни метаподаци докумената и преглед у формату PDF такође им је доступан.

Библиша омогућава претрагу комплетних поравнатих текстова корпуса са сумеђом која је потпуно прилагођена корисницима (енгл. *user-friendly*) и то на два начина: претрагу преко метаподатака и претрагу пуног текста. Претрага преко метаподатака је једнојезична и омогућава корисницима да комбинују више поља за претрагу (речи из наслова, име аутора, кључне речи, речи из сажетка, речи из текста) при чему корисници дефинишу језик претраге из листе која се налази на почетку претраживача

и бирају да ли желе да претражују све расположиве колекције или неку одређену. Као резултат добија се листа докумената која одговара постављеном упиту.

Слика 3 приказује неке примере резултата претраге преко метаподатака, тј. преко кључне речи „biblioteka". Добијено је 16 резултата, а у сваком резултату приказан је идентификациони број документа, краћи преглед метаподатака и дата је могућност кориснику да прегледа комплетне метаподатаке који описују документ (могућност *detaljnije*), да приступе тексту у формату TMX[49] (могућност *tmx*), да приступе комплетном тексту или делу текста у формату PDF где год је то могуће (могућност *pdf*) и да погледају „врећу речи" ако она постоји за дати документ (могућност *bow*[50]).

Broj pogodaka: 16

Document	About
1.2011.1.1	**Naslov**: Delatnost Karnegijevih zadužbina na Balkanu posle Prvog svetskog rata: Univerzitetska biblioteka u Beogradu, 1919-1926 **Autori**: Nadine Akhund **Ključne reči**: Karnegijeva fondacija za mir, Univerzitetska **biblioteka** u Beogradu, izgradnja [detaljnije] [tmx] [pdf] [bow]
1.2011.1.10	**Naslov**: ACCESSIT (Accelerate the circulation of culture through exchange of skills in information technology) **Autori**: Predrag Đukić **Ključne reči**: AccessIT, digitalizacija, digitalne biblioteke, **Biblioteka** grada Beograda, dLibra [detaljnije] [tmx] [pdf] [bow]
1.2011.1.5	**Naslov**: Približiti biblioteku korisnicima: biblioteke u alternativnim prostorima **Autori**: Adam Sofronijević, Jelena Andonovski **Ključne reči**: Biblioteke, alternativni prostori, **biblioteka** na aerodromu, **biblioteka** u metrou, **biblioteka** na brodu, bibliotekarstvo, unapređenje usluga, iPAD [detaljnije] [tmx] [pdf] [bow]

Слика 3. Примери резултата претраге у Библиши преко метаподатка „кључна реч" (keywords) „biblioteka"

[49] Translation Memory eXchange (Преводилачка меморија за размену) јесте ISO стандард (ISO 24616:2018) за складиштење такозваних преводилачких меморија (Translation Memories) (збирке одредница у којима је текст на једном језику повезан са еквивалентним преводом текста на другом језику) и њихову размену између различитих софтверских преводилачких алата, као и између различитих фирми које се баве одржавањем преводилачких меморија.

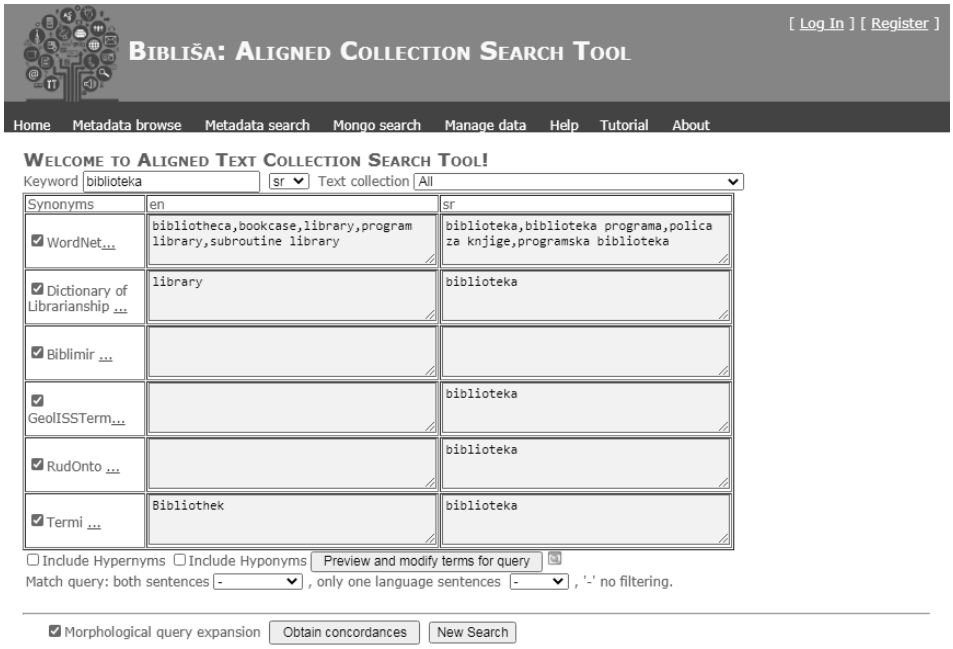
[50] Bag of Words је векторска репрезентација докумената за потребе њиховог рангирања.

Поред једнојезичне претраге преко метаподатака, Библиша омогућава и двојезичну претрагу комплетног текста колекција уз могућност морфолошког и семантичког проширење упита позивањем различитих лексичких и терминолошких ресурса које развија JePTex. Под морфолошким проширењем упита подразумева се позивање ресурса који омогућавају генерисање свих флективних облика кључних речи датих у упиту. У случају Библише морфолошко проширење упита се, тренутно, врши само за српски језик и за то се користе електронски морфолошки речници српског језика.[51] Под семантичким проширењем упита подразумева се позивање ресурса за генерисање свих семантички еквивалентних лексема и термина на српском и њихови преводи на енглеском, односно немачком језику који су у Библиши заступљени.

Корисници дефинишу упит за претрагу уношењем кључних речи у поље за претрагу и, као и код претраге преко метаподатака, бирају језик претраге и колекцију коју желе да претражују. Као резултат добијају се листе кључних речи распоређене према одговарајућим лексичким ресурсима на српском и енглеском, односно немачком језику, а корисници могу да их користе онакве какве јесу или да их уређују брисањем или додавањем нових, а од добијених резултата генеришу се конкорданце.

На пример, ако се као упит за претрагу постави „biblioteka" на српском језику над свим понуђеним колекцијама и означе се сви расположиви лексички ресурси за семантичко проширење, добијају се резултати које илуструје Слика 4.

[51] Duško Vitas and Cvetana Krstev, „Processing of Corpora of Serbian Using Electronic Dictionaries", *Prace Filologiczne* Vol. LXIII (2012): 279–292, http://poincare.matf.bg.ac.rs/~cvetana/biblio/22_Vitas_Krstev.pdf (преузето 9. 3. 2021); Cvetana Krstev, Duško Vitas and Agata Savary, „Prerequisites for a Comprehensive Dictionary of Serbian", in *Proceedings of the 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August, 2006*, eds. Tapio Salakoski, Filip Ginter, Sampo Pyysalo, Tapio Pahikkala. Serija Lecture Notes in Artificial Intelligence: Subseries of Lecture Notes in Computer Science, eds. J.G. Carbonell, J. Siekmann, (Heidelberg, Berlin: Springer, 2006), 552–564; Cvetana Krstev and Duško Vitas, „An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds", in *Arena Romanistica*, eds. B. Lamiroy, E. Laporte, T. Kyriakopoulou (Bergen: University of Bergen, Department of Foreign Languages, 2009), 204–212, http://poincare.matf.bg.ac.rs/~cvetana/biblio/Krstev-Vitas-LGC09.pdf (преузето 9. 3. 2021).



Слика 4. Окружење у Библиши за претрагу комплетног текста свих колекција преко упита „biblioteka" са листом семантички еквивалентних лексема на српском, енглеском и немачком језику

Добијени резултати претраге могу се проширити надређеним и подређеним појмовима, а понуђена опција „Obtain concordances" генерише конкорданце на основу добијених резултата претраге који су приказани на Слици 5. Свака произведена конкорданца садржи информацију о изворном документу и везу ка метаподацима који додатно описују овај документ, а кључна реч постављена кроз упит за претрагу (у овом случају „biblioteka") додатно је означена у резултатима претраге.

	Number of concordances (en/de/fr): 2601	Broj konkordansi (sr): 2601
Zaman Shuvaa et al., 2011, vol. XII:1, ID: 1.2011.1.3 **metadata**	Bangladesh National **Library** has 5 staff employed for digitization, National **Library** of the Republic of Indonesia has 12 staff, and the National **Library** of the Philippines has 26 staff employed for digitization.	U Nacionalna **biblioteci** Bangladeša petoro ljudi je angažovano na digitalizaciji, u Nacionalnoj **biblioteci** Republike Indonezije dvanaest, u Nacionalnoj **biblioteci** Filipina dvadeset i šest.

Jordan, 2011, XII:2, ID: 1.2011.2.2 metadata	There are CBS installations at the Royal **Library** of the Netherlands, the German National **Library** and the National **Library** of Australia, and at ABES (Agence bibliographique de l'enseignement superieur) in France, to name a few.	Instalacije CBS sistema postoje na primer u Kraljevskoj **biblioteci** Holandije, Nacionalnoj **biblioteci** Nemačke, Nacionalnoj **biblioteci** Australije i u ABES (Agence bibliographique de l'enseignement superieur) u Francuskoj.
Authority Control in Serbia / Ana Savić = Normativna kontrola u Srbiji / Ana Savić metadata	**n76** As the highly complex task was not progressing as planned, the team size gradually increased, firstly the Matica Srpska **Library** engaged more cataloguers, then the University **Library** "Svetozar Marković" and finally the National **Library** of Serbia.	**n76** Izuzetno složen posao redakcije normativnih zapisa nije se odvijao zadovoljavajućom brzinom i postepeno se redaktorski tim širio, najpre u **Biblioteci** Matice srpske, zatim Univerzitetskoj **biblioteci** „Svetozar Marković" i, na kraju, u Narodnoj **biblioteci** Srbije.
Komo / Srdxan Valxarevicx = Como / Srdxan Valxarevicx, ID: 11.2.006 metadata	n3270 Er hatte sie irgendwo in einer **Bibliothek** für mich ausgegraben.	n3270 Iskopao ga je u nekoj **biblioteci**, za mene.

Слика 5. Примери произведених конкорданци за упит „biblioteka"

Конкорданце могу приказивати одговор на постављени упит на више начина. Први начин је њихово генерисање без било каквих филтера, као што је приказано на Слици 5. Приказују се сви резултати који садрже реч из постављеног упита („*biblioteka*") у оба језика (српски и енглески, односно немачки). Други начин омогућава да се сагледају све конкорданце у којима се јавља одговор на упит на два одабрана језика, на пример "EN&SR". Ово значи да се добијају конкорданце у којима се одговор на постављени упит јавља у једном језику заједно са његовим еквивалентима у другом језику, односно као резултат генеришу се конкорданце које садрже одговор на упит на један или на оба одабрана језика једне паралелне колекције. Трећи начин је да се сагледају све конкорданце у којима се одговор на упит јавља у текстовима на једном од језика одабране паралелне колекције, али не и на другом, на пример на "EN" (Слика 6).

Number of concordances (en/de/fr): 271		Broj konkordansi (sr): 271
Hopkinson, 2007, vol. VIII:1/2, ID: 1.2007.1/2.3 metadata	I also saw **Bibliotheca** exhibiting at the conference and was impressed.	Video sam tada sistem Bibliotheca koji je bio izložen na konferenciji i bio sam zadivljen.
Hopkinson, 2007, vol. VIII:1/2, ID: 1.2007.1/2.3 metadata	So if there this standard had already become available, we would not have had to engage in so much discussion with **Bibliotheca.**	Tako da su ti standardi bili na raspolaganju, mi ne bi morali toliko da se angažujemo u raspravama sa provajderom Bibliotheca.
Hopkinson, 2007, vol. VIII:1/2, ID: 1.2007.1/2.3 metadata	When we discussed with **Bibliotheca** how we would implement RFID, they wanted our assistance in developing the 'data model' for our implementation.	Kada smo diskutovali sa dobavljačem Bibliotheca o tome kako bi mogli da uvedemo RFID, oni su tražili našu pomoć u razvoju modela podataka za naš sistem.

Слика 6. Примери произведених конкорданци за упит „biblioteka" са опцијом приказа резултата "EN"

Закључак

Паралелни корпуси су последњих деценија постали изузетно значајни за различите области истраживања језика, као што су двојезична или вишејезична лексикографија, учење страних језика, процеси превођења и машинског превођења, истраживање терминологије, лингвистичка истраживања, упоредна изучавања језика итд. У овом раду приказани су паралелни корпуси које последњих деценија JePTex, поред других језичких алата, интензивно развија. До данас, а посебно са развојем дигиталне библиотеке Библиша, припремљен је велики број паралелних вишејезичних колекција (углавном двојезичних). Истраживачи имају могућност да своја истраживања врше на корпусним колекцијама како књижевних тако и стручних текстова из различитих научних области (библиотекарство и информатика, рударство и енергетика, архитектура, стоматологија, економија, филозофија, и сл.). Захваљујући развијеним системима за претрагу, посебно комплексном систему алата Библиша, могућности за проналажењем информација у доступним корпусима су велике, а приказ добијених резултата истраживачима омогућава да сагледају све неопходне информације. Такође, могућност морфолошког и семантичког проширења упита омогућава истраживачима са сагледају све флективне облике одређене кључне речи на српском језику и њене еквивалнете у другим

језицима, док семантичко проширење упита на више језика омогућава да се сагледају семантички еквиваленти једне речи у датом језику и њихови преводи на друге језике, односно варијанте превода на другим језицима. Осим тога, истраживачи могу утврдити и фреквентност појављивања речи задате у упиту у више језика, као и многе друге могућности које су већ у раду поменуте.

Важно је истаћи да доступни ресурси нису коначни већ су подложни променама и допунама те истраживачи имају могућност да кроз своја истраживања утврде грешке које се јављају и које треба исправити, затим, укажу на нове термине које је пожељно додати, или пак утврде да се у резултатима претраге појављују жељени термини који нису означени, односно нису део одговарајућег лексичког ресурса. Све измене истраживачи могу сами унети у већ постојеће ресурсе ако за то имају привилегију. У плану је допуна ресурса за морфолошко и семантичко проширење упита на страним језицима (посебно немачком) за шта је потребно утврдити расположивост језичких ресурса који се у ове сврхе могу користити, да ли су они у слободном приступу или је потребно тражити дозволу за њихово коришћење.

Literatura:

1. Andonovski, Jelena. „Mreža povezanih otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa". Doktorska dis., Filološki fakultet, Beograd, 2019, http://phaidrabg.bg.ac.rs/o:22874 (na ćirilici)
2. Balvet, Antonio, Dejan Stosic and Aleksandra Miletic. „TALC-Sef a Manually-revised POS-Tagged Literary Corpus in Serbian, English and French". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 26–31. Reykjavik: European Language Resources Association, 2014, http://www.lrec-conf.org/proceedings/lrec2014/index.html (preuzeto 2. 3. 2021).
3. Blaney, Jonathan. „Introduction to the Principles of Linked Open Data". *The Programming Historian* (12. 5. 2020), https://programminghistorian.org/en/lessons/intro-to-linked-data (preuzeto 1. 3. 2021). https://doi.org/10.46430/phen0068.
4. Dimitrova, Ludmila, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaaler and Dan Tufis. „MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages". In *Volume 1 Proceedings of the 17th international conference on Computational linguistics* (Association for

Computational Linguistics, 1998), http://www.aclweb.org/anthology/P98-1050 (preuzeto 2. 3. 2021).

5. Dobrić, Nikola. „Corpus Linguistics – the Basic Form of Linguistic Analysis". *Philologiano* 7 (2009): 31–41, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309960 (preuzeto 1. 3. 2021).
6. Erjavec, Tomaž. „MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora". In *LREC*, 2544–2547. 2004, http://nl.ijs.si/et/teach/jsi07-hlt/Bib/Multext_LREC04.pdf (preuzeto 2. 3. 2021).
7. Erjavec, Tomaž and Nancy Ide. „The MULTEXT-East Corpus". In *Proceeding of First International Conference on Language Resources & Evaluation, Granada, Spain, 28–30 May,* 971–974. 1998, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.5846&rep=rep1&type=pdf, (preuzeto 2. 3. 2021).
8. Erjavec, Tomaž, Ann Lawson and Laurent Romary. eds. *East Meets West: A Compendium of Multilingual Resources*. Mannheim: TELRI Association, e.V., Institut für deutsche Sprache, 1998.
9. Erjavec, Tomaž, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić and Duško Vitas. „The MULTEXT-East Morphosyntactic Specifications for Slavic Languages". In *Proceedings of the Workshop on Morphological Processing of Slavic Languages: 10th Conference of the European Chapter, EACL,* eds. Tomaž Erjavec and Duško Vitas, 25–32. Budapest, 2003, http://poincare.matf.bg.ac.rs/~cvetana/biblio/04erjavec.pdf (preuzeto 2. 3. 2021).
10. Gavrilidou, Maria, Peny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos and Stelios Piperidis. „Building Parallel Corpora for eContent Professionals", in *MLR '04 Proceedings of the Workshop on Multilingual Linguistic Ressources* (Stroudsburg: Association for Computational Linguistics, 2004), 97–100, https://www.aclweb.org/anthology/W04-2213.pdf (preuzeto 2. 3. 2021).
11. Krstev, Cvetana and Duško Vitas. „An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds". In *Arena Romanistica*, eds. B. Lamiroy, E. Laporte, T. Kyriakopoulou, 204–212. Bergen: University of Bergen, Department of Foreign Languages, 2009, http://poincare.matf.bg.ac.rs/~cvetana/biblio/Krstev-Vitas-LGC09.pdf (preuzeto 9. 3. 2021).
12. Krstev, Cvetana and Duško Vitas. „Analigned English-Serbian Corpus". In *Volume I ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, *Belgrade, 4–6 December 2009*, eds. N. Tomović & J. Vujić, 495–508. Belgrade: Faculty of Philology, University of Belgrade, 2011, http://poincare.matf.bg.ac.rs/~cvetana/biblio/AlignedCorpus-full-final.pdf (preuzeto 2. 3. 2021).
13. Krstev, Cvetana, Duško Vitas and Agata Savary. „Prerequisites for a Comprehensive Dictionary of Serbian". In *Proceedings of the 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August, 2006,* eds. Tapio Salakoski, Filip Ginter,

Sampo Pyysalo, Tapio Pahikkala. Serija *Lecture Notes in Artificial Intelligence: Subseries of Lecture Notes in Computer Science*, eds. J.G. Carbonell, J. Siekmann, 552–564. Heidelberg, Berlin: Springer, 2006.

14. Krstev, Cvetana, Duško Vitas and Tomaž Erjavec. „Morpho-Syntactic Descriptions in MULTEXT-East – the Case of Serbian". *Informatica* Vol. 28, No. 4 (2004): 431–436, http://poincare.matf.bg.ac.rs/~cvetana/biblio/mtesr-inform04.pdf (preuzeto 2. 3. 2021).
15. Laporte, Eric, Duško Vitas and Cvetana Krstev. „Preparation and Exploitation of Bilingual Texts". *Lux Coreana* No. 1 (2006): 110–132, http://poincare.matf.bg.ac.rs/~cvetana/biblio/VKL.pdf (preuzeto 1. 3. 2021).
16. McEnery, Tony and Andrew Wilson. *Corpora and Translation: Uses and Future Prospects*, 1993, http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf (preuzeto 1. 3. 2021).
17. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. „A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals". In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 23–25 May 2012*, eds. Nicoletta Calzolari et al., 1710–1717. Istanbul: European Language Resources Association, 2012, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf (преузето 3. 3. 2021).
18. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Olivera Kitanović. „Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian". In *Semantic Keyword-Based Search on Structured Data Sources – First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, September 8–9,* 167–181. Springer, 2015. DOI 10.1007/978-3-319-27932-9_15
19. Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović and Olivera Kitanović. „Keyword-Based Search on Bilingual Digital Libraries". In *Semantic Keyword-Based Search on Structured Data Sources – Second COST Action IC1302 International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9*, eds. A. Calì, D. Gorgan and M. Ugarte, LNCS 10151, 112–123. Springer, 2017.
20. Tyers, Francis M. and Murat Serdar Alperen. „South-East European Times: A Parallel Corpus of Balkan Languages". In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, 2010, 49–53.
21. Utvić, Miloš. „Izgradnja referentnog korpusa savremenog srpskog jezika". Doktorska dis., Filološki fakultet, Beograd, 2013, http://phaidrabg.bg.ac.rs/o:10061 (preuzeto 1. 3. 2021).
22. Vitas, Duško and Cvetana Krstev. „[Electronic Edition of Serbian Translation of Orwell's 1884 aligned with 7 languages by Duško Vitas, Cvetana Krstev]". In *East Meets West – A compendium of Multilingual Resources*, eds. Tomaž Erjavec, Ann

Lawson, Laurent Romary. Mannheim: TELRI Association e.V., Institut für deutsche Sprache, 1998.

23. Vitas, Duško and Cvetana Krstev. „Literature and Aligned Texts". In *Readings in Multilinguality*, eds. Milena Slavcheva, Galia Angelova and Kiril Simov, 148–155. Sofia: Institute for Parallel Processing, Bulgarian Academy of Sciences, 2006, http://poincare.matf.bg.ac.rs/~cvetana/biblio/CvDv-Paskaleva.pdf (preuzeto 2. 3. 2021).
24. Vitas, Duško and Cvetana Krstev. „Construction and Exploitation of X-Serbian Bitexts". In *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation*, eds. Cristina Vertan and Walther v. Hahn, 207–227. Cambridge: Cambridge Scholars Publishing, 2012, http://poincare.matf.bg.ac.rs/~cvetana/biblio/DvCv-CambridgeS-2012.pdf (preuzeto 2. 3. 2021).
25. Vitas, Duško i Ljubomir Popović. „Konspekt za izgradnju referentnog korpusa srpskog standardnog jezika". U *Naučni sastanak slavista u Vukove dane,* 31(1): 221–227. Beograd: MSC, 2003. (na ćirilici)
26. Vitas, Duško, Goran Nenadić and Cvetana Krstev. „[Electronic Edition of Serbian Translation of Plato's Republic Aligned with 17 Languages by Duško Vitas, Goran Nenadić, Cvetana Krstev]". In *East meets West – A Compendium of Multilingual Resources*, eds. Tomaž Erjavec, Ann Lawson, Laurent Romary. Mannheim: TELRI Association e.V., Institut für deutsche Sprache, 1998.
27. Vitas, Duško, Svetla Koeva, Cvetana Krstev and Ivan Obradović. „Tour du monde through the dictionaries". In *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, eds. M. Constant, T. Nakamura, M. De Gioia, S. Vecchiato, 249–256. Paris: Universite Paris-Est, Institut Gaspard-Monge, 2008, http://poincare.matf.bg.ac.rs/~cvetana/biblio/akvila-en-fin.pdf (preuzeto 2. 3. 2021).
28. Vitas, Duško and Cvetana Krstev. „Processing of Corpora of Serbian Using Electronic Dictionaries". *Prace Filologiczne* Vol. LXIII (2012): 279–292, http://poincare.matf.bg.ac.rs/~cvetana/biblio/22_Vitas_Krstev.pdf (preuzeto 9. 3. 2021);

Jelena S. Andonovski
University Library "Svetozar Marković", Belgrade
andonovski@unilib.rs

PARALLEL CORPORA IN SERBIA – POSSIBILITIES FOR SIMULTANEOUS INFORMATION RETRIEVAL IN TWO OR MORE LANGUAGES

Summary: Aligned multilingual corpora have become essential resources in multilingual Natural Language Processing (NLP) in the last decades, as well as one of the major resources for researchers in various areas of linguistics and related language disciplines. Parallel corpora are language corpora that contain a collection of one or more original texts in one language and their translations into one or more other languages. Original texts and their translations are aligned at some level of text divisions (e.g. sentence, paragraph, and chapter level). In most cases, parallel corpora contain texts in only two languages but also there are examples of one-language parallel corpora containing a collection of different editions of the same text in one language. In Serbia, JeRTeh, Language Resources and Technologies Society (former Group for Language Technologies) has been developing parallel corpora containing Serbian texts for decades. Until today, JeRTeh has developed: Serbian-French aligned corpus (SrpFranKor) and Serbian-English aligned corpus (SrpEngKor), digital library Biblisha with several parallel collections, and multilingual edition *Multilingual Vern*. In addition, corpora texts in the Serbian language are part of multilingual parallel corpora *Plato's Republic* and *Orwell's 1984* developed during the international projects, as well as part of some corpora developing now in the region and the world. This paper presents corpora developed by Group for Language Technologies, their structure and purpose, as well as possibilities for information retrieval in them.

Keywords: corpus linguistics, language corpora, parallel corpora, natural language processing, information retrieval.

Примљено: 9. марта 2021.
Исправке: 8. јуна 2021.
Прихваћено: 14. јуна 2021.