Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

UDK 81'322:811.16(497)
Nikola Dobrić
Odsjek za anglistiku i amerikanistiku, Alpen-Adria Univerzitet u Klagenfurtu

SAVREMENI JEZIČKI KORPUSI NA ZAPADNOM BALKANU – ISTORIJAT, TRENUTNO STANJE I BUDUĆNOST

Zapadni Balkan ima bogatu istoriju konstrukcije jezičkih korpusa. Prvi elektronski korpus u regionu je konstruisan samo nekoliko godina posle prvog elektronskog korpusa u svetu, dok se ideja razvitka elektronskih jezičkih resursa razvila na ovim prostorima još ranije. Ovakav rani razvitak obrade prirodnog jezika je donekle usporen (negde i skoro zaustavljen) nesretnim događajima devedesetih godina prošlog veka. Na sreću, protekle dve dekade bile su obeležene značajnim napretkom u razvoju korpusa zapadno-balkanskih jezika. Ovaj članak prvo daje istorijski pregled razvitka jezičkih korpusa i korpusne lingvistike u regionu u periodu između 1950. i 1990. godine, kao i trenutno stanje i buduću perspektivu.

Ključne reči: korpusi, Zapadni Balkan, istorijat, pregled, jezički resursi, obrada prirodnog jezika

The West Balkans have had a rich history in developing language corpora. The first electronic corpus in the region was created only a few years after the very first one in the world, while the idea of developing electronic language resources dates even further back. This early development of natural language processing was somewhat hampered by the unfortunate events of the 1990s, but in the last two decades there has been some substantial improvement in the development of the West Balkan language corpora. The paper presents a historical overview of the language corpora development in the region in the period from 1950 to 1990 as well as its current state and future prospects.

Keywords: corpora, West Balkans, history, overview, language resources, natural language processing

1 Jezički korpusi na Zapadnom Balkanu do 1990. godine[1]

Iako je važnost jezičkih korpusa[2] u savremenoj lingvisitici danas opšte poznata, samo je mali broj lingvista u svetu prepoznao tu važnost tako rano kao lingvistička zajednica Zapadnog Balkana. Prateći razvoj mašinskog prevođenja četrdesetih i pedesetih godina prošlog veka, prvenstveno u Sjedinjenim Američkim Državama, prvi korpus u slične svrhe je takođe započet u to vreme i na ovim prostorima.

[1] Autor bi hteo posebno da se zahvali Tomažu Erjavecu, Simonu Kreku, Volfgangu Tojbertu (Wolfgang Teubert), Špeli Vintar, Dušku Vitasu i ponajviše Primožu Jakopinu na njihovom uvidu u istorijat razvitka jezičkih korpusa na Zapadnom Balkanu.

[2] Članak podrazumeva svaku veću isključivo elektronsku bazu tekstova, imenovanu korpusom bez obzira na njen nivo anotacije i obrade teksta (Meyer 2002). Članak se takođe bavi jednojezičkim korpusima, mada pokriva paralelne ili višejezičke korpuse.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

Započeo ga je psiholog Đorđe Kostić 1957. godine u Beogradu sa ciljem razvitka jezičkih tehnologija za prepoznavanje govora i mašinsko prevođenje sa tadašnjeg srpsko-hrvatskog jezika. Projekat je trajao do 1962. godine (Kostić 2003: 261), ali korpus tada ipak nije elektronski obrađen. Idući u korak ne samo sa teoretskim nego i sa tehnološkim inovacijama u području obrade prirodnog jezika, prvi elektronski korpus na Zapadnom Balkanu napravljen je u Zagrebu već 1967. godine, samo tri godine posle pojavljivanja prvog elektronskog korpusa na svetu, Brown Corpus korpusa. Bio je to elektronski obrađeni ep *Osman* Ivana Gundulića koji je pripremio Željko Bujas. Pojava ovog korpusa je pokrenula lavinu interesovanja za stvaranje elektronskih korpusa i već 1968. imamo još jedan korpus konstruisan u Zagrebu, Jezik Marka Marulića, koji je pripremio Milan Moguš (Tadić 1997: 388) i koji je dalje proširen sedamdesetih i osamdesetih godina prošlog veka[3] i rezultirao Jednomilijunskim korpusom hrvatskog književnog jezika (iliti takozvanim Moguševim Korpusom). Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu je takođe bio dom većem korpusnom projektu između 1972. i 1975. godine pod naslovom Englesko-hrvatski leksikografski korpus koji je važno pomenuti kao još jedan primer ranog razvitka elektronskih korpusa u regionu iako je kao paralelni korpus izvan opsega ovog rada. Godine 1971. Denis Poniž je na Univerzitetu u Ljubljani prekucao dve kutije 80-kolonskih papirnih kartica teksta koje su zajedno sadržavale nekih 4.000 redova iliti 320.000 karaktera uzetih iz molitvenog opusa Janeza (Krstnika) Svetokriškog, dok je Tomo Pisanski isprogramirao frekvencijski brojač slova i još neke oblike računarske analize. Rezultati ovog poduhvata su objavljeni 1974. godine u knjizi *Slovenski jezik, literatura, računalniki* (podnaslovljenom *numerično-statistično raziskovanje konstantnih in spremenljivih količin v slovenskem jeziku, prozi in poeziji*). Iste godine simpozijum *Informatica 74* na Bledu je ukazao na mogućnosti i potrebe računarske obrade teksta u regionu (Tancig i Tancig 1974). Tri godine kasnije, 1977., Peter Šerber (Peter Scherber) je na Univerzitetu u Getingenu objavio *Slovar Prešernovega pesniškega jezika* (prvi lematizovani konkordancijski rečnik nekog slovenskog jezika) zasnovan na njegovom malom elektronskom korpusu dela Franca Prešerna. 1980. godine na Univerzitetu u Ljubljani Primož Jakopin je, uz pomoć Melite Ambrožič i Jure Dimca, elektronski obradio 400.000 reči iz dela Cirila Kosmača (ovaj korpus je još uvek dostupan na internet stranici Slovarske in besedilne zbirke). Per Jakobsen (Per Jacobsen) je 1980. godine u Danskoj objavio *Kvantitativnu analizu Balada Petrice Kerempuha*, studiju zasnovanu na elektronskoj korpusnoj konkordanciji toga dela.[4] Matematički institut na Univerzitetu u Beogradu je u to vreme, 1981. godine, takođe započeo veliki projekat pod nazivom Matematička i računarska lingvistika sa ciljem konstruisanja elektronskog korpusa savremenog srpskog jezika. Peter Tancig i Tomaž Erjavec su 1989. godine konstruisali korpus Verbalni napadi na JNA[5] koji se sastoji

[3] Ovaj projekat se u tom periodu mogao naći pod više različitih imena, kao na primer Kompjutorska analiza tekstova stare hrvatske književnosti ili Korpus suvremenog hrvatskog književnog jezika.

[4] Ovakvih korpusnih studija koje su podrazumevale digitalizovanje pojedine knjige ili više knjiga jednog pisca kako bi se mogle proučiti njihove konkordancije u ovom periodu je bilo više, tako da rad pominje samo nekoliko obimnijih analiza.

[5] Jugoslovenska narodna armija.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

od 259.217 reči uzetih iz novinskih članaka iz perioda april–avgust iste godine i koji su za temu imali JNA.

Osim ovih individualnih i nezavisnih projekata u Beogradu, Ljubljani i Zagrebu ovaj period su takođe obeležili i zajednički korpusni projekti. Rudolf Filipović je 1968. godine u Zagrebu započeo konstrukciju višejezičnog korpusa u okviru projekta Yugoslav Serbo-Croatian-English Contrastive Project. Dati korpus je bio zasnovan na prevođenju već pomenutog Brown Corpus korpusa što je rezultiralo prvim elektronskim paralelnim korpusom u svetu i prvom upotrebom računara u kontrastivnoj lingvistici (Tadić 1997: 388). Projekat je trajao do 1971. godine i doveo do dalje popularizacije korpusne lingvistike u regionu. Još jedan važan zajednički projekat u kome su učestvovali univerziteti u Beogradu, Ljubljani i Zagrebu je započet 1988. godine i bio je zasnovan na zajedničkom učestvovanju u međunarodnom projektu pod nazivom Jezičke industrije (Language Industries). Zavod za lingvistiku Sveučilišta u Zagrebu, kao jedan od pokretača korpusne lingvistike na Zapadnom Balkanu, je koordinirao projekat koji je bio važan i zbog toga što je omogućio uspostavljanje međunarodnih veza sa raznim svetskim univerzitetima i saradnje sa velikim evropskim korpusnim projektima i centrima. Ovaj period je takođe proizveo još dva paralelna korpusa koje je važno spomenuti: srpsko-slovenački korpus jezika uputstava za lekove i srpsko-hrvatsko-slovenački korpus saveznih zakona.

Sav ovaj zajednički rad na konstrukciji korpusa i popularizaciji korpusne lingvistike je doveo do formiranja ideje stvaranja velikog korpusa svih zapadno-balkanskih (tada jugoslovenskih) jezika koja se prvi put pojavila još 1978. godine na prvoj ROJP[6] (Računarska obrada jezičkih podataka/Računalniška obdelava jezikovnih podatkov) konferenciji koju su originalno pokrenuli Peter Tancig i Milan Šipka (ideja koja je donekle proistekla i iz računarski orijentisanih Informatica konferencija). Prvi koraci ka konstrukciji ovakvog korpusa bili su obeleženi sve većom saradnjom Grupe za jezičke tehnologije iz Beograda, Instituta Jozef Stefan iz Ljubljane i Sveučilišnog računskog centra (SRCE) iz Zagreba. Nažalost, ova plodna i perspektivna saradnja zapadno-balkanskih univerziteta među sobom i sa svetskim univerzitetima, projektima i centrima je grubo prekinuta nestretnim događajima tokom devedesetih godina prošlog veka. Njihovi putevi razvitka korpusne lingvistike nastavili su se potpuno razdvojeni, a tek poslednjih nekoliko godina možemo videti nove, iako skromne, početke nove saradnje.

2 Jezički korpusi na Zapadnom Balkanu posle 1990. godine

Ovaj period u regionu bio je prvenstveno obeležen obnovom učešća odnosno prisustva Zapadnog Balkana u međunarodnom naučnom krugu i filološkim proučavanjima, ali i fokusiranjem pažnje na konstrukcije velikih nacionalnih korusa u zapadno-balkanskim zemljama. Međunarodna saradnja, koja je bila ključni preduslov za stvaranje jezičkih tehnologija neophodnih za obradu jezika regiona, je perso-

[6] Od kojih su ROJP 3, održan 1985. godine na Bledu, i ROJP 4, održan 1988. u Portorožu, zbog svojih zaključaka od posebne važnosti za korpusnu lingvistiku u regionu.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

nifikovana kroz dva značajna naučna projekta započeta sredinom devedesetih godina prošlog veka – TELRI i MULTEXT-East.

Trans-European Language Resources Infrastructure ili TELRI (TELRI I i TELRI II) je bio projekat koji je izveden u dve faze, finansiran od strane Evropske komisije i predvođen Volfgangom Tojbertom čiji cilj je bio povezivanje svih evropskih centara za jezičke tehnologije i kroz tako stvorenu saradnju konstruisanje jednojezičnih i višejezičnih (paralalenih) korpusa (kao i elektronskih rečnika, leksičkih baza podataka i računarskih programa neophodnih za obradu različitih jezika obuhvaćenih projektom). TELRI I, koji je trajao od 1995. do 1998. godine, na početku je uključivao samo slovenački jezik (predstavljen učešćem Tomaža Erjaveca i Instituta Jozef Stefan) dok su se ostali zapadno-balkanski jezici pridružili projektu kasnije ili na početku TELRI II projekta 1998. godine (uklučujući tadašnji Matematički fakultet Univerziteta u Beogradu, Filozofski fakultet Sveučilišta u Zagrebu i Filološki fakultet Univerziteta Sv. Kiril i Metodije u Skoplju). TELRI II je trajao do 2002. godine a ukupan obimni sadržaj projekta je još uvek dostupan u istraživačke svrhe na njegovoj intenet stranici. MULTEXT-East je proizišao iz MULTEXT[7] korpusnog projekta koji je između 1995. i 1997. godine obuhvatao bamanankan, bugarski, katalonski, češki, holandski, engleski, estonski, francuski, nemački, mađarski, italijanski, kikongo, oskitanski, rumunski, slovenački, španski, švedski i svahili jezike (svi podaci su takođe dostupni u naučne svrhe na internet stranici MULTEXT projekta). Prvi rezultat MULTEXT-East projekta, objavljen 1998. godine, je bio paralelni korpus šest jezika (osim engleskog bili su tu bugarski, češki, estonski, mađarski, rumunski i slovenački) i koji se sastojao od morfosintaksički anotiranog teksta knjige *1984*. Džordža Orvela (Erjavec 2010). Rezultati objavljeni 2001., 2004. i 2010. godine su uključili još 10 dodatnih, mahom zapadno-balkanskih jezika: hrvatski, litvanski, makedonski, persijski, rozajski, ruski, srpski, slovački i ukrajinski. Projekat je još uvek u toku a više podataka se može pronaći na njegovoj internet stranici.[8]

Ova dva projekta su dala novi elan razvoju korpusa u regionu i obezbedili su računarske alate neophodne za takve poduhvate. U kombinaciji sa novonastalim nacionalnim motivima, ovi projekti su neposredno omogućili postojanje većine savremenih jezičkih korpusa na Zapadnom Balkanu. Posmatrajući bogati stariji korpusni opus stvoren pre 1990. godine i međunarodna iskustva stečena kroz pomenute projekte u toku devedesetih godina dvadesetog veka, može se reći da su neke zemlje Zapadnog Balkana nastavile sigurnim koracima ka kompleksnom razvoju korpusne lingvistike na ovim prostorima, neke zemlje su napravile tek nekoliko novih koraka dok ostale tek treba da se upuste u ovu lingvističku avanturu.[9]

---

[7] Multilingual Text.

[8] Osim učestvovanja u ova dva velika međunarodna projekta, zapadno-balkanske zemlje se mogu naći kao učesnici i u drugim internacionalnim korpusnim poduhvatima, na primer učešće Hrvatske i Slovenije u stvaranju korpusa dečjeg jezika CHILDES (Child Language Data Exchange System).

[9] Pregled postojećih i dostupnih korpusa zapadno-evropskih jezika koji sledi dat je abecednim redom.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

2.1 Korpusi bosanskog jezika

Kada se pogleda bosanski jezik, činjenica je da praktično nema elektronskih baza tekstova koje bi pokrivale ovaj jezik, a kamoli anotiranih i obrađenih korpusa. Stanje je takvo, čini se, zbog lingvističkog i političkog statusa bosanskog jezika u prošlosti (Baotić 2004). U suštini jedini samodefinisani dostupan korpus bosanskog jezika se može naći na Univerzitetu u Oslu. The Oslo Corpus of Bosnian Texts je rezultat projekta koji teče na Odseku za orijentalne studije na Univerzitetu u Oslu[10] i u njihovoj Labratoriji za tekst[11] pod vođstvom Džejn Bondi Johanesen (Janne Bondi Johannessen). Korpus sadrži oko 1,5 miliona reči različitih žanrova (književnost, eseji, dečje pripovetke, narodne pripovetke, islamski religijski tekstovi, pravni tekstovi i novinski članci) uglavnom iz devedesetih godina dvadesetog veka.[12]

2.2 Korpusi crnogorskog i makedonskog jezika

Makedonski je jedan od zapadno-balkanskih jezika koji nije bio jako prisutan u razvitku korpusa u periodu pre 1990. godine i koji nažalost ni sada nije dovoljno pokriven jezičkim korpusima. Kako su makedonski lingvisti učestvovali u TELRI i MULTEXT-East projektima, tehnoligije za obradu i anotiranje makedonskog jezika su dovoljno razvijene (Vojnovski, Džeroski i Erjavec 2005), i to je ipak dovelo do konstrukcije nekoliko, iako još nereprezentativnih, korpusa.

Jedan od njih je Makedonski elektronski korpus, konstruisan od strane Georgea Mitrevskog i Instituta za makedonski jezik Univerziteta u Skoplju, koji je još uvek u izgradnji (iako je korpus i sada besplatno dostupan). Tekstovi su tokenizirani, a dalja obrada i anotiranje je u planu. Jedini drugi dostupan korpus makedonskog jezika, osim tekstova u okviru MULTEXT-East i TELRI korpusa, je Jednojezički i višejezički Gralis korpus makedonskog jezika (Monolinguale und multilinguale Gralis-Korpus der Mazedonische Sprache). To je mali korpus koji je Branko Tošović pokrenuo 2008. godine na Karl-Francens Univerzitetu u Gracu i predstavlja deo većeg i već pomenutog Gralis-Korpus projekta. Trenutno broji 47.000 reči i biće potpuno završen 2016. godine.

Kako je status crnogorskog jezika još uvek, ako ne politički i institucionalno onda bar lingvistički neodređen (Greenberg 2004), trenutno nema dostupnih jednojezičnih korpusa ovog jezika, projekata u toku niti najava konstrukcije istih. Jedini dostupan samodeklarativni korpus crnogorskog jezika je paralelni Montecorpus korpus započet 2009. godine. Korpus trenutno uključuje 3.316.152 reči uzetih iz prevoda tekstova koje obrađuje crnogorsko ministarstvo za pridruživanje Evropskoj

[10] Department for East European and Oriental Studies, University of Oslo.

[11] Text Laboratory.

[12] Dobri izvori za bosanski jezik iliti za BHS (bosanski/hrvatski/srpski) su takođe različiti korpusni projekti koje vodi Branko Tošović na Univerzitetu u Gracu, uključujući Gralis BHS korpus (Gralis-Korpus); projekat Ivo Andrić u evropskom kontekstu (2007.–2015.) koji podrazumeva konstrukciju paralelnog BHS-nemačkog korpusa njegovih dela; i Lirski, humoristički i satirički svet Branka Ćopića (2011.–2016.), projekat koji za cilj ima konstrukciju elektronskog korpusa njegovih dela; kao i književno-korpusne inicijative, na primer pan-balkanska kolekcija elektronskih tekstova koja se može naći u projektu Rastko.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

uniji. Lingvistički značaj ovog projekta kao i nivo obrade i anotiranja tekstova su još uvek nepoznati.

2.3 Korpusi hrvatskog jezika

Potpuno opravdavši tradiciju obrade prirodnog teksta na Sveučilištu u Zagrebu dugu pola veka hrvatski lingvisti su možda postigli više na polju pokrivenosti svoga jezika koropusima nego većina njihovih kolega na Zapadnom Balkanu. Hrvatski se jezik može pohvaliti i velikim, dobro obrađenim opštim (nacionalnim) korpusom koji se može posmatrati kao odličan primer ostalim zemljama u regionu (iako je slovenački jezik takođe odlično pokriven korpusima).

Hrvatski nacionalni korpus (HNK) se dakle rodio kao ideja još ranih devedesetih godina prošlog veka (Tadić 1990) dok je rad na njemu zapravo započeo 1996. godine u okviru projekta Računalna obrada hrvatskoga jezika pod vođstvom Vesne Muhvić-Dimanovski. Kao osnovni cilj projekta zamišljena je konstrukcija višemilionskog korpusa savremenog hrvatskog jezika, čija je struktura, kao i ime, definisana po ugledu na Britanski nacionalni korpus (British National Corpus (BNC)). Kroz ubrzani razvoj ovaj korpus danas istraživačima nudi 101,3 miliona reči anotiranih prema MULTEXT-East standardima (Agić, Tadić i Dovedan 2009) uzetih iz godina 1996.–2004. Korpus je još uvek u razvitku: cilj njegove izgradnje je dalja optimizacija, bolja reprezentativnost i veći stepen obrade i anotiranja. Korpus je otvoren za pretragu i besplatno je dostupan na njegovoj internet stranici.[13]

Još jedan od važnih korpusa hrvatskog jezika je Hrvatska jezična mrežna riznica. Korpus je deo tekućeg projekta započetog 2005. godine koji za cilj ima sakupljanje javno dostupnih tekstova na hrvatskom jeziku u što većem broju (uključujući književne tekstove, rečnike i drugi javno dosupne izvore počevši od 19. veka). Ostali korpusi hrvatskog jezika koje treba spomenuti su sledeći[14]:

- Hrvatski jezični korpus: konstruisan je kao potkorpus Hrvatske jezične mrežne riznice i predstavlja elektronsku kolekciju važnijih dela hrvatske književnosti (uključujući romane, pripovetke, drame i poeziju), eseja, naučnih publikacija, udžbenika, elektronskih publikacija i novina;
- Hrvatski mofološki leksikon i lematizacijski poslužitelj: to je leksička baza podataka koja obuhvata oko 100.000 lema opšteg jezika, ličnih muških i ženskih imena i prezimena (Tadić i Flugosi 2003);
- Hrvatska ovisnosna banka stabala: to je takođe tekući projekat Zavoda za lingvistiku sa Sveučilišta u Zagrebu, čija svrha je konstruisanje potpuno sintaksički anotiranog hrvatskog korpusa od bar 100.000 reči (Berović, Agić i Tadić 2012);
- Intratext zbirka vjerskih tekstova na hrvatskom jeziku: elektronski tekstovi koji su dostupni uključuju Bibliju, Katekizam rimokatoličke Crkve, itd.;

[13] Više informacija o alatima i servisima za obradu hrvatskog jezika se može pronaći na internet stranici Jezične tehnologije za hrvatski jezik.

[14] Neki izvori, kao na primer internet stranica Jezične tehnologije za hrvatski jezik navode još neke korpuse hrvatskog jezika (na primer Korpus tekstova Starih pisaca hrvatskih; Korpus tekstova udžbenika za osnovne i srednje škole u Republici Hrvatskoj; ili *Kur'an* u elektronskom obliku), ali pošto ih nije bilo moguće pronaći niti pronaći više informacija o njima, oni nisu uključeni u datu listu.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

- Hrvatska biskupska konferencija: Stari Zavet i Novi Zavet dostupni u elektronskom obliku; i
- Korpus Silvija Strahimira Kranjčevića: elektronska kolekcija njegovih dela.

2.4 Korpusi slovenačkog jezika

Pošto su temelje obrade prirodnog teksta u Sloveniji postavili Primož Jakopin 1980. godine i Tomaž Erjavec i Peter Tancig 1989., prvi sledeći važan korak u razvoju korpusa slovenačkog jezika napravljen je 1995. godine od strane Mirana Hladnika na Filozofskom fakultetu Univerziteta u Ljubljani. Miran Hladnik je te godine počeo objavljivati elektronsku zbirku slovenačkih književnih dela, mahom starijeg datuma, kojima je istekla zaštita autorskih prava.[15] Koristeći tekstove iz ove zbirke Primož Jakopin je 1999. godine, kao deo rada na svojoj doktorskoj disertaciji (pod naslovom *Zgornja meja entropije pri leposlovnih besedilih v slovenskem jeziku*), konstruisao korpus od 3 miliona reči iz kojega je nastao CORTES korpus.

CORTES (Corpus of Texts in Slovene) korpus je bio još jedan važan korak za slovenačku korpusnu lingvistiku, a sastojao se od isključivo književnih dela, uključujući 112 proznih dela iz pera 41 pisca iz perioda 1858.–1998. (Grzybek 2007: 172). Korpus je sa svojih 3 miliona reči svoj dom našao na Filozofskom fakultetu Univerziteta u Ljubljani dok je dalji ubrzani razvoj počeo kada je 2000. godine prebačen na Institut za slovenski jezik Frana Ramovša u sklopu SAZU. Kroz nekoliko nadogradnji i transkripte u narednih pet godina (uključujući tekstove iz dnevnih novina DELO, transkripte rasprava u slovenačkom parlamentu u periodu 1996.–2004., i dodatne književne tekstove) korpus, koji se sada zove Slovarske in besedilne zbirke, danas se može pohvaliti sa 318 miliona reči i nekoliko potkorpusa (kao što su Beseda i Nova Beseda, Poizvedbe po označenih besedilih i Ciril Kosmač korpus) i ostalih jezičkih resursa (na primer rečnika). Dati potkorpusi su različito obrađeni i anotirani, od toga da su očišćeni od šumova (grešaka i sl.) pa sve do anotiranja prema vrsti reči i lematizacije.

Najprominentniji, a svakako i najreferentniji korpus slovenačkog jezika je FIDA Plus korpus (naslednik FIDA korpusa) (Krek 2012). FIDA korpusni projekat je započeo 1997.[16] godine i u vreme svog završetka 2000. godine se sastojao od nekih 100 miliona reči. Kako je od samog početka planiran da bude veoma reprezentativan, korpus za izvore ima velik broj jezičkih varijanti i registara iz perioda 1950.–2000. (iako je većina tekstova zapravo iz devedesetih godina prošlog veka), a uglavnom se radi o pisanom jeziku uz manji broj izvora govornog jezika (uglavnom transkripata diskusija iz slovenačkog parlamenta). Glavni nedostaci originalnog FIDA korpusa su bili nedovoljna dostupnost u nekomercijalne svrhe (pošto je korpus bio finansiran od strane komercijalnih ulagača) i tadašnja nedovoljna razvijenost računarskih alata

[15] Zbirka je objavljena na internet stranici današnje Zbirke slovenskih leposlovnih besedil.

[16] Originalna inicijativa za stvaranje reprezentativnog opšteg (nacionalnog) korpusa slovenačkog jezika je došla sa Filozofskog fakulteta, Fakulteta za društvene nauke Univerziteta u Ljubljani i sa Instituta Jozef Stefan a korpus je u početku bio finansiran od državne Javne agencije za raziskovalno dejavnost Republike Slovenije i ko-finansiran od strane dva komercijalna partnera – DZS izdavačke kuće i Amebis računarske kompanije.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

za obradu slovenačkog jezika (Arhar et al. 2007). FIDA Plus, pod vođstvom Marka Stabeja, za cilj ima potpuno besplatnu dostupnost, povećanje obima korpusa i bolji nivo anotacije (Grzybek 2007: 173). Novi korpus se sastoji od šireg izbora izvora (iako još uvek nema dovoljan udeo govornog jezika) i uglavnom sadrži primere relativno savremenog slovenačkog jezika iz perioda 1990.–2000. Trenutno ima 621.150.000 reči koje su anotirane prema MUTEXT-East modelu anotacije (Erjavec 1998).

Ostali jednojezični korpusi slovenačkog jezika koje treba navesti su[17]:

- Sloleks leksikalna baza: je tekući projekat započet 2008. godine (sa planiranim završetkom 2013.) koji nastoji da konstruiše korpus slovenačkog jezika od milijardu reči koji bi se koristio primarno u leksikografske svrhe (GANTAR i KREK 2011);
- JOS korpus (JOS100k i JOS1M): ovaj korpus trenutno sadrži milion reči uzetih iz FIDA Plus korpusa koje su sve potpuno ili delimično (ručno) lematizivane i anotirane za morfosintaksičke karakteristike, sintačke odnose među njima i WordNet sinsetove za pojedine imenice (ERJAVEC et al 2010);
- Zbirka slovenskih leposlovnih besedil: to je ogromna zbirka raznih slovenačkih književnih tekstova (poezije i proze) koji su dostupni u elektronskom obliku;
- iKorpus: korpus od 14 miliona reči koji sadrži tekstove vezane za informacione tehnologije i računarstvo;
- KoRP korpus PR tekstova: korpus javno dostupnih reklamnih tekstova koji broji 18 miliona reči (Žgank et al. 2006);
- Slovene Dependency Treebank (SDT): mali sintaksički anotiran korpus tekstova na slovenačkom jeziku koji za osnovu ima 30.000 reči uzetih kao uzorak iz slovenačke komponente paralelnog MULTEXT-East korpusa (Džeroski et al. ALL 2006); [18]
- Korpus govorjene slovenščine (GOS): izgrađuje se od 2007. godine, a za 2013. je planirano da sadrži milion reči isključivo govornog slovenačkog jezika;
- Učni korpus govorjene slovenščine: je mali korpus govornog slovenačkog jezika koji je sakupila Jana Zemljarič-Miklavčič i koji se sastoji od 15.000 ručno anotiranih reči spontanog slovenačkog kao drugog odnosno stranog jezika (Miklavčič 2006);
- Jezikovni viri starejše slovenščine IMP: elektronska zbirka od preko 150 knjiga i novina od 16. veka sve do dvadesetih godina dvadesetog veka; i

[17] Neki izvori (kao na primer Logar 2000) navode još neke korpuse slovenačkog jezika (kao na primer BNSI: Broadcast News korpus koji se sastoji od transkripata dnevnih informativnih emisija nacionalne televizije Slovenije u periodu 1999.–2003.; SiBN korpus sa 2,3 miliona reči koji se takođe sastoji od transkripta dnevnih informativnih emisija nacionalne slovenačke televizije, u periodu 2003.–2004.; SloParl korpus sa 23 miliona reči iz transkripata rasprava slovenačkog parlamenta (više o ovom projektu može se naći na internet stranici DSPLAB Universiteta u Mariboru); i Korpus vojaških besedil zajedno sa Korpusom besedil odnosov z javnostmi čija internet stranica (na http://www.korp.fdv.uni-lj.si/), ne funkcioniše, i o kojima nije moguće pronaći više informacija. Zbog nedostupnosti i nedostatka informacija ili publikacija o ovim korpusima oni nisu mogli biti uključeni u datu listu korpusa slovenačkog jezika.

[18] Dobar izvor informacija o tehnologijama i servisima za obradu slovenačkog jezika je internet stranica Slovene Natural Language Server.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

- Referenčni korpus starejše slovenščine goo300k: korpus starog slovenačkog jezika koji je zapravo deo Jezikovni viri starejše slovenščine IMP korpusa. Razlika je u tome što su 300.000 reči ovog korpusa ručno obrađene i anotirane (Erjavec 2012).

2.5 Korpusi srpskog jezika

Iako je veoma aktivno učestvovala u ranim regionalnim i međunarodnim korpusnim projektima u periodu od 1950.–1990. godine, i iako se relativno uspešno vratila u međunarodne lingvističke tokove posle loše političke i ekonomske situacije sredinom devedesetih godina prošlog veka, čini se da, nažalost, srpska lingvistička zajednica nije uradila dovoljno kada je u pitanju konstruisanje korpusa. Dostupno je samo nekoliko korpusa od kojih ni jedan ne zadovoljava ni svetske ni regionalne standarde reprezentativnosti, veličine ili nivoa obrade (Dobrić 2009).

Prvi od dva prominentnija korpusa srpskog jezika[19] predstavlja konačnu fazu razvitka prethodno spomenutog korpusa Đorđa Kostića započetog 1957. godine. Njegov sin, Aeksandar Kostić, je 1996. Godine, posle četiri decenije razvitka korpusa, pretvorio sav materijal u elektronski oblik. Nazvan Korpus srpskog jezika, korpus se sastoji od 11 miliona reči i može se pohvaliti dobro konstruisanom dijahronom dimenzijom (jer uključuje i stare tekstove iz 12. veka) i veoma detaljnom, ručno izvedenom anotacijom, gde je svaka reč anotirana za njen gramatički status, broj grafema, slogova i za njenu fonološku strukturu (Kostić 2003). S druge strane, u korpusu uopšte nema govornog jezika, nema dovoljno tekstova savremenog jezika i u njemu nije dovoljno lingvistički razgraničeno šta u dijahronoj dimenziji predstavlja srpski jezik a šta tadašnji srpsko-hrvatski[20].

Drugi poznati korpus srpskog jezika je svakako SrpKorp korpus savremenog srpskog jezika[21]. Konstrukcija ovog korpusa je započeta još 1981. godine kao deo već pomenutog projekta Matematička i računarska lingvistika (Krstev et al. 2003). Korpus je nastavio sa daljim razvitkom sve do danas Duško Vitas, zajedno sa Cvetanom Krstev i ostalim saradnicima sa Grupe za jezičke tehnologije na Matematičkom fakultetu Univerziteta u Beogradu. Korpus danas broji 113 miliona reči isključivo iz pisanih izvora i automatski je anotiran za vrste reči (Utvić 2011; Popović 2010). Korpusi srpskog jezika koje takođe vredi pomenuti uključuju[22]:

[19] Pojedini usmeni izvori spominju i korpus SANU (Srpske akademije nauka i umetnosti) koji je korišćen za konstrukciju njihovih rečnika srpskog jezika kao i rečnika Matice srpske i koji se stvaraju od 19. veka. Ovaj pregled ne uključuje dati korpus iz nekoliko razloga: korpus je konstruisan uglavnom u neelektronskom obliku (iako od skora ima dosta pomaka ka elektronskoj obradi); korpus je namenjen isključivo u leksikografske svrhe i ne prati savremene standarde konstrukcije korpusa (uglavnom se fokusira na kljiževne izvore); i na kraju, korpus je nedostupan, netransparentan i kao takav neupotrebljiv za širu lingvističku zajednicu.

[20] Kako korpus sadrži brojne tekstove iz druge polovine dvadesetog veka, na primer govore Josipa Broza Tita, nije lako razgraničiti gde po savremenim lingvističkim kriterijumima prestaje hrvatski a počinje srpski jezik.

[21] Prethodno nazivan Korpus savremenog srpskog jezika.

[22] Pregled postojećih jezičkih tehnologija, alata i servisa, razvijenih za obradu srpskog jezika, može se pronaći na internet stranici Jezičke tehnologije – resursi i alati.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

- Antologija srpske književnosti: je projekat pokrenut na Učiteljskom Fakultetu u Beogradu i sadrži preko 130 dela narodne, stare i moderne književnosti u elektronskom (isključivo ćiriličnom) obliku;
- Rastko projekat: osim sinhronog i dijahronog pregleda pan-balkanske književnosti ovaj korpus, započet 1997. godine, uključuje i poseban potkorpus tekstova isključivo na srpskom jeziku, počevši od srednjevekovnih tekstova sve do savremenih izvora; i
- Elektronski korpus dela Laze Kostića: to je projekat započet 2009. g. pod pokroviteljstvom Matice srpske koji podrazumeva stvaranje korpusa njegovih dela. Završetak i dostupnost korpusa još nisu poznati[23].

3 Budućnost

Ako se pogleda ovaj sažet i iscrpan pregled dostupnih korpusa zapadno-balkanskih jezika[24] mora se priznati da je veoma pohvalno što su neke zemlje iz regiona, kao na primer Hrvatska i Slovenija, uspele da opravdaju sav nagovešteni potencijal i lingvističko nasleđe veoma ranog razvoja korpusne lingvistike na Zapadnom Balkanu. Nije pak pohvalno to što neke druge zemlje, a to važi za Srbiju i još više za Crnu Goru, Bosnu, Makedoniju, još uvek nisu ostvarile zadovoljavajuće rezultate u konstrukciji svojih jezičkih korpusa.

I zaista, budućnost hrvatskog i slovenačkog jezika, što se tiče korpusne pokrivenosti, izgleda veoma blistava. Osim konstantnih napora ka poboljšanju jezičkih tehnologija koje bi rezultirale boljom obradom ovih jezika, obe zemlje su u poslednjih nekoliko decenija uspešno iskoristile i domaće i međunarodne izvore naučnog finansiranja i tako znatno proširile svoj korpusni opus. Perspektiva srpskog jezika takođe nije tako loša – jezičke tehnologije i alati za obradu srpskog jezika su prilično dobro formirani. Najveći problem, čini se, pretstavlja nedostatak političke i institucionalne podrške (koja je prisutna u Hrvatskoj i Sloveniji) i nedostatak aktivnosti na međunarodnom planu. Ako bi se ta dva faktora pokrenula, postoji više sposobnih institucija u Beogradu koje su stručne i voljne da sustignu korpusno naprednije susede u regionu.

Situacija je nažalost mnogo gora kada su u pitanju ostala tri jezika Zapadnog Balkana. Dok makedonski ima neke jezičke tehnologije razvijene i upotrebljive za sopstvenu obradu (nastale, između ostalog u okviru MULTEXT-East projekta) trenutno ipak nema većih aktivnih projekata niti posebnih instituta koji bi radili na razvoju naučno ambicioznijih korpusa makedonskog jezika, a takođe se može primetiti i značajno odsustvo kako finansijske, tako i političke podrške za razvitak korpusne lingvistike u Makedoniji. Bosanski i crnogorski jezici, zbog raznih problema vezanih za njihov lingvistički i politički status (za bosanski ranije, a crnogorski sada) nažalost nemaju dovoljno korpusnih tehnologija posebno razvijenih za njih.

[23] U najavi je takođe i projekat konstrukcije korpusa savremenog srpskog jezika u saradnji Matice srpske i Odseka za srpski jezik i lingvistiku Filozofskog fakulteta Univerziteta u Novom Sadu, iako detalji još nisu poznati.

[24] Celokupna lista sa aktivnim linkovima se može pronaći na internet stranici UniKlu West Balkan Corpora Page.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

Iako mogu značajno koristiti jezičke tehnologije namenjene obradi lingvistički srodnih jezika kao što su hrvatski ili srpski, za bilo kakav obimniji razvoj korpusa (koji danas skoro da i ne postoje) potrebno je mnogo domaće institucijalne i finansijske podrške. Iz navedenoga sledi da četiri poslednja pomenuta jezika moraju osigurati mnogo jaču podršku za popularizaciju korpusa u svojim zemljama i obezbediti bolje izvore finansiranja, jer samo uz ta dva faktora može se pokrenuti razvoj korpusne lingvistike, a sustizanje svetskih i regionalnih standarda će onda biti samo pitanje vremena.

Važno je takođe razumeti da rani razvoj korpusa na Zapadnom Balkanu nije obavezao regionalne lingvističke institucije samo na dalji konstantan razvoj. Velika imena lingvistike iz ovog regiona su svojim primerom obavezala buduće generacije i na konstantnu međusobnu saradnju. Zajednički projekti zapadno-balkanskih zemalja, koji su bili tako važni između 1950. i 1990. godine, kasnije su (osim zajedničkog rada na međunarodnim projektima) skoro sasvim zamrli. Osim manjih projekata, kao što su recimo hrvatsko-slovenački paralelni korpusi[25], inicijative za razvitak višejezičnih korpusa zapadno-balkanskih jezika ili za zajednički razvoj jezičkih tehnologija za obradu ovih srodnih jezika nema ni približno u dovoljnoj meri. Korak napred bi svakako bio rad na razvitku paralelnih korpusa, razvoj BHS dijahronih korpusa ili možda čak obnavljanje ideje sveobuhvatnog južnoslovenskog korpusa (uključujući, odnosno zajedno sa bugarskim jezikom). Ljudi, stručnost i tehnološki resursi postoje, ali još uvek nedostaju volje i želje.

Na kraju, možemo se samo nadati da će sve zemlje Zapadnog Balkana nastaviti da grade nove i da unapređuju postojeće korpuse svojih jezika (neke po mogućstvu sa više elana nego do sada) jer su jezički resursi jednog (posebno nemnogoljudnog) naroda i jedne nacije neprocenjivo jezičko i kulturno blago, a korpus je prava riznica starog i savremenog jezika, nezaobilazna i neophodna za nacionalni opstanak u globalnom svetu današnjice.

Viri i literatura

Željko Agić, Marko Tadić i Zdravko Dovedan, 2009: Tagset reductions in morphosyntactic tagging of Croatian texts. *Proceedings of the INFuture 2009 digital resources and knowledge sharing conference*. Ur. H. Stančić et al. 289–298.

Špela Arhar i Vojko Gorjanc, 2007: *Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa*. Ljubljana: Slavistično društvo Slovenije.

Josip Baotić, 2004: The Language Situation in Bosnia and Herzegovina. *Language in the former Yugoslav lands*. Ur. R. Bugarski, C. Hawkesworth. Bloomington: Slavica Publisher. 117–125.

Daša Berović, Željko Agić i Marko Tadić, 2012: Croatian dependency treebank: Recent development and initial experiments. *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Ur. N. Calzolari et al. 1902–1906.

[25] Kao na primer Hrvatsko-slovenski paralelni korpus.

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

Peter Grzybek, 2007: *Contributions to the science of text and language: Word length studies and related issues*. Berlin: Springer Verlag.

Nikola Dobrić, 2009: Corpus linguistics as the new paradigm of language research. *Philologia* 7. 47–57.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdeněk Žabokrtsky i Andreja Žele, 2006: Towards a Slovene dependency treebank. *Proceedings of the 5th international conference on language resources and evaluation (LREC'06)*. Ur. European Language Resources Association (ELRA). 1388–1391.

Tomaž Erjavec, 1998: The MULTEXT Slovene lexicon. *Proceedings of the 7th electrotechnical conference (ERK)*. 189–192.

Tomaž Erjavec i Simon Krek, 2008: The JOS Morphosyntactically tagged corpus of Slovene. *Proceedings of the 6th international conference on language resources and evaluation (LREC'08)*. Ur. European Language Resources Association (ELRA). 322–326.

--, 2010: MULTEXT-East Version 4: Multilingual morphosyntactic specifications, lexicons and corpora. *Proceedings of the seventh conference on international language resources and evaluation (LREC'10)*. Ur. N. Calzolari et al. 1535–1538.

--, 2012: The goo300k corpus of historical Slovene. *Proceedings of the 8th international conference on language resources and evaluation (LREC'12)*. Ur. N. Calzolari et al. 225–260.

Tomaž Erjavec, Darja Fišer, Simon Krek i Nina LEDINEK, 2010: The JOS linguistically tagged corpus of Slovene. *Proceedings of the seventh conference on international language resources and evaluation (LREC'10)*. Ur. N. Calzolari et al. 1806–1809.

Polona Gantar i Simon Krek, 2011: Slovene Lexical Database. *Proceedings of the natural language processing, multilinguality: 6th international conference*. Ur. D. Majchráková, R. Garabík. 72–80.

Robert Greenberg, 2004: From Serbo-Croatian to Montenegrin? Politics of language in Montenegro. *Language in the former Yugoslav lands*. Ur. R. Bugarski, C. Hawkesworth. Bloomington: Slavica Publisher. 53–64.

Per Jakobsen, 1980: Kvantitativna analiza balada Petrice Kerempuha. Københavns: Slavisk boghandel.

Aleksandar Kostić, 2003: Elektronski kopus srpskog jezika Đorđa Kostića. *Zbornik Matice srpske za slavistiku* 64. 260–264.

Simon Krek, 2012: *Slovene language in a digital age*. Berlin: Springer Verlag.

Cveta Krstev, Gordana Pavlović-Lažetić, Ivan Obradović i Duško Vitas, 2003: Corpora Issues in Validation of Serbian Wordnet. *Proceedings of the 6th international conference TSD 2003: Text, speech and dialogue*. Ur. V. Matoušek, P. Mautner. Berlin: Springer Verlag.132–137.

Nataša Logar, 2008: Pregled korpusov za slovenščino. Prezentacija na 3. posvetu

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

Slovenskega društva za jezikovne tehnologije.

Charles Meyer, 2002: *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Zoran Popović, 2010: Taggers applied on texts in Serbian. *Proceedings of the INFOtheca'10 conference*. 21a–38a.

Marko Tadić, 1990: Zašto nam je potreban višemilijunski referentni korpus? *Informatička teorija u primenjenoj lingvistici*. 95–98.

--, 1997: Računalna obradba hrvatskih korpusa: Povijest, stanje i perspektive. *Proceedings of the XII. international Slavic congress*. 387–394.

Marko Tadić i Sanja Fulgosi, 2003: Building the Croatian morphological lexicon. *Proceedings of the EACL2003 workshop on morphological šrocessing of Slavic languages*. 41–46.

Peter Tancig i Simona Tancig, 1974: Uporaba računalnika pri konstrukciji testov znanja in pri obdelavi rezultatov. *Zbornik Informatica '74*. Ljubljana: IJS.

Miloš Utvić, 2011: Annotating the corpus of contemporary Serbian. *Proceedings of the INFOtheca'12 conference*. 36a–47a.

Viktor Vojnovski, Sašo Džeroski i Tomaž Erjavec, 2005: Learning POS tagging from a tagged Macedonian text corpus. *Proceedings of the 8th international multiconference Information society (IS)*. 199–202.

Arno Wonisch, 2012: *Das Pronominalsystem des Bosnischen/ Bosniakischen, Kroatischen und Serbischen*. Beč: Lit Verlag.

Jana Zemljarič Miklavčič, 2006: Korpus govorjene slovenščine. *Proceedings of the 5th Slovenian and 1st international conference language technologies (IS–LTC'06)*. Ur. T. Erjavec, J. Gros. 124−127.

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj i Zdravko Kačič, 2006: Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. *Proceedings of the 5th Slovenian and 1st international conference language technologies (IS–LTC'06)*. Ur. T. Erjavec, J. Gros. 115–119.

Antologija srpske književnosti (http://www.antologijasrpskeknjizevnosti.rs/....aspx).

Beseda korpus (http://bos.zrc-sazu.si/beseda.html).

British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/)

Brown Corpus (http://www.essex.ac.uk/linguistics/clmt/... brown/brown.html).

Child Language Data Exchange System (CHILDES) (http://childes.psy.cmu.edu/).

Ciril Kosmač korpus (http://bos.zrc-sazu.si/ckb_en.html).

DASPLAB (Laboratorija za digitalno procesiranje signalov) (http://www.dsplab.uni-mb.si/Dsplab/index_eng.php).

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

FIDA (http://www.tei-c.org/Activities/Projects/fi01.xml).

FIDA Plus (http://www.fidaplus.net/).

Gralis korpus (Gralis-Korpus) http://www-gewi.uni-graz.at/gralis/korpusarium/gralis_korpus.html

Hrvatska biskupska konferencija: Stari Zavet i Novi Zavet (http://www.hbk.hr/biblija/nz/index.html & http://www.hbk.hr/biblija/sz/index.html).

Hrvatska jezična mrežna riznica (http://riznica.ihjj.hr/).

Hrvatska ovisnosna banka stabala (http://hobs.ffzg.hr/default_en.html).

Hrvatski jezični korpus (http://riznica.ihjj.hr/).

Hrvatski mofološki leksikon i lematizacijski poslužitelj (http://hml.ffzg.hr/hml/info.php?show=hml).

Hrvatski nacionalni korpus (HNK) (http://www.hnk.ffzg.hr/).

Hrvatsko-slovenski paralelni korpus (http://www.hnk.ffzg.hr/hr-si_pcorp/).

Ikorpus (http://nl2.ijs.si/dsi.html & http://www.islovar.org/iskanje_enostavno.asp).

Institut Fran Ramovš (http://isjfr.zrc-sazu.si/en#v).

Institut Jozef Stefan (http://www.ijs.si/).

Institut za makedonski jazik Krste Misirkov (http://www.ukim.edu.mk/en_struktura_contact.php?inst=34).

Intratext zbirka vjerskih tekstova na hrvatskome (http://www.intratext.com/SCR/).

Ivo Andrić u evropskom kontekstu (http://www-gewi.uni-graz.at/gralis/projektarium/Andric/index.html).

Jednojezički and višejezički Gralis korpus makedonskog jezika (Monolinguale and multilinguale Gralis-Korpus der Mazedonische Sprache) (http://www-gewi.uni-graz.at/gralis/projektarium/Mak-Korpus/index.html).

Jezičke tehnologije – resursi i alati (http://poincare.matf.bg.ac.rs/~cvetana/LT-pregled.html).

Jezične tehnologije za hrvatski jezik (http://jthj.ffzg.hr/default_english.htm).

JOS korpus (http://nl.ijs.si/jos/index-en.html).

KoRP korpus PR tekstova (http://www.korp.fdv.uni-lj.si).

Korpus govorjene slovenščine (GOS)(http://www.korpus-gos.net/Support/About).

Korpus srpskog jezika (http://www.serbian-corpus.edu.rs/indexns.htm).

Lirski, humoristički i satirički svet Branka Ćopića (http://www-gewi.uni-graz.at/gralis/projektarium/Copic/index.html).

Makedonski elektronski korpus (http://www.tekstlab.uio.no/glossa/html/index_dev.php?corpus=mak).

Montekorpus (http://www.eiprevod.gov.me/korpus/).

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

MULTEXT (http://aune.lpl.univ-aix.fr/projects/multext/).

MULTEXT-East (http://nl.ijs.si/ME/).

Nova Beseda korpus (http://bos.zrc-sazu.si/a_beseda.html).

Poizvedbe po označenih besedilih (http://bos.zrc-sazu.si/ckb_en.html).

Rastko projekt (http://www.rastko.rs/).

Referenčni korpus starejše slovenščine goo300k (http://nl.ijs.si/imp-cuwi/imp-goo).

Silvije Strahimir Kranjčević lorpus (http://www.sskranjcevic.hr/uvod.ASP?PisID=1).

Sloleks leksikalna baza (http://www.slovenscina.eu/).

Slovarske in besedilne zbirke (http://bos.zrc-sazu.si/index_en.html).

Slovene Dependency Treebank (SDT) (http://nl.ijs.si/sdt/).

Slovene Natural Language Server (http://nl.ijs.si/).

Slovensko društvo za jezikovne tehnologije (http://www.sdjt.si/viri.html).

SRCE institut (http://www.srce.unizg.hr/homepage/).

SrpKorp Korpus savremenog srpskog jezika (http://korpus.matf.bg.ac.rs/prezentacija/korpus.html & http://www.korpus.matf.bg.ac.rs/SrpLemKor/).

TELRI (http://telri.nytud.hu/).

TELRI I (http://telri.nytud.hu/start.html).

TELRI II (http://telri.nytud.hu/telri2/intro.html).

The Oslo Corpus of Bosnian Texts (http://www.tekstlab.uio.no/Bosnian/Corpus.html#cont).

Učni korpus govorjene slovenščine (http://torvald.aksis.uib.no/talem/jana/s9.html).

UniKlu West Balkan Corpora Page (http://www.uni-klu.ac.at/iaa/inhalt/2525.htm).

Verbalni napadi na JNA (http://nl.ijs.si/tei/teiHeaders/VAYNA-header-en.html).

Zavod za lingvistiku Filozofskog fakulteta Univerziteta u Zagrebu (http://www.ffzg.unizg.hr/zzl/).

Zbirka slovenskih leposlovnih besedil (http://lit.ijs.si/leposl.html).

Summary

The paper looks at the current available corpora of the West Balkan languages and what they have to offer to researchers. The current state is put into focus by a detailed outline of the history of the development of corpora in this region, starting with the very first electronic corpora in the 1960s and following their common development until 1990 (being that all of the languages and countries understood as West Balkans belonged to the same country in this period). The paper also follows the beginnings of their individual development in the last two decades and emphasizes

Slavistična revija (https://srl.si) je ponujena pod licenco
Creative Commons, priznanje avtorstva 4.0 international.
URL https://srl.si/sql_pdf/SRL_2012_4_07.pdf | DOST. 18/02/26 9.30

the importance of the international projects that helped to reform the technological resources necessary for the construction of contemporary corpora. The conclusions the author arrives at speak volumes about the amount of work some of the countries involved still need to invest in order to reach both world and regional standards in the construction of corpora. They also point out the need for a renewal of regional cooperation that was so fruitful in the early years of corpus linguistics in the West Balkans.