О фамилији корпуса савременог српског језика СрпКор

Научни рад DOI: 10.18485/judig.2025.1.ch1

Душко Витас, 0000-0003-4194-692X
Ранка Станковић, 0000-0001-5123-6273
Цветана Крстев[1] *0000-0003-3328-9392*

Апстракт

Рад приказује мотиве за развој и еволуцију СрпКор-а, описујући промене у коришћеним алатима, у димензијама корпуса, временској покривености, као и нивоима анотације. Поред грађе са веба, у овим корпусима је присутна и грађа која није доступна на мрежи, нарочито из научног и књижевног домена, али и она која због технолошких промена, није била архивирана на изворним сајтовима. Биће описан и његов даљи развој који укључује обогаћивање метаподатака, допуну анотација и унос нових садржаја осим материјала прикупљеног са веба. Актуелна верзија корпуса је повезана са другим језичким, посебно лексичким ресурсима, кроз систем Лексимирка што омогућава значајно унапређење лексикографског рада, али и оригинално интегрисано окружење за језичке експерименте.

Кључне речи: СрпКор, корпуси, српски језик, лематизација, Лексимирка

1. Увод

Акроним СрпКор означава фамилију електронских корпуса савременог српског језика чија је изградња почела крајем седамдесетих година прошлога века, а која је постала шире видљива заинтересованој истраживачкој заједници објављивањем његове прве верзије на вебу 2003. године. У овом дугом периоду, посебно пре појаве корисних текстуелних ресурса на вебу, развој корпуса се састојао у прикупљању и

[1] Друштво за језичке реурсе и технологије JePTex, {vitas|cvetana|ranka}@jerteh.rs

обради грађе као и у развоју метода обраде корпуса. Наиме, електронски корпус није само колекција текстова у дигиталном облику (како се то, на пример, наводи у (Добрић 2012)), већ подразумева више компонената које ће заједно овакву колекцију учинити корисном у језичким и другим истраживањима. Ове компоненте, поред самих текстова, чине, пре свега, софтверска подршка организацији и експлоатацији колекције текстова и средства за различите нивое анотације текстова који ће се наћи у корпусу (Витас 2023).

СрпКор је, водећи рачуна о овим компонентама, током своје изградње прошао различите метаморфозе које пружају слику, како о еволуцији софтверске подршке за конструкцију и експлоатацију корпуса, тако и о развоју система анотација на различитим нивоима (метаподаци, морфолошко обележавање, лематизација, именовани ентитети, итд).

Крајње скромни услови (у поређењу са другим срединама, како у броју истраживача укључених у изградњу корпуса, додељеним финансијским средствима из различитих извора, расположивом опремом) су наметнули стратегију поступног развоја корпуса која је подразумевала да ће се нове верзије корпуса ослањати на материјал припремљен и употребљен у оним верзијама које су јој претходиле.

У раду ће бити илустрована еволуција у развоју СрпКор-а почев од његове прве верзије до данас пратећи упливе различитих средстава која су коришћена у изградњи појединачних верзија, као и промене димензија и система анотације текстова. Посебно ће бити описана структура појединачних верзија корпуса, њихове димензије, обухваћени временски период и ниво анотације.

Основне замисли приликом конципирања веб-верзије корпуса су прво изложене у (Витас & Поповић 2003), а затим у (Utvić 2013) где су описани бројни детаљи за верзију СрКор-а из 2013. године. Интеракције корпуса са речницима су разматране у (Krstev & Vitas 2005, Vitas & Krstev 2012).

У одељку 2 су описани први кораци обраде природних језика у Србији. Одељак 3 је посвећен првим корпусима на вебу, док се и одељку 4 говори о даљем развоју корпуса и њиховим верзијама.

2. Први кораци

Интересовање за (рачунарску) обраду природних језика је у Србији почело почетком осамдесетих година прошлог века, знатно касније него у другим европским срединама, па чак и оним које су и језички и географски биле блиске у то доба. Могући узроци за овакво

занемаривање једне важне језичке технологије данас су тешко разумљиви. Могуће их је тражити, на пример, у традиционалном одсуству математичког и информатичког образовања филолога у Србији, чије се последице и даље живо осећају. Овоме је могао посебно допринети и одјек Пирсовог извештаја (Pierce и други 1966.) који је негативно оценио развој џорџтаунског пројекта аутоматског превођења и обуставио његово финансирање[2]. Како је део овог пројекта био финансиран у Југославији (стр. 41 и даље), обустављање овог пројекта је морало изазвати код традиционално образованих лингвиста оправдано зазирање од рачунарских технологија, (видети, на пример, (Ивић, 2001, стр. 61 и даље), (Rajić, 1981)). Зазирање од информатичких поступака у обради језичких података се види и кроз читав низ различитих квантитативних језичких истраживања која су до осамдесетих година прошлог века обављена ручним пребројавањем и уазбучавањем одређеног језичког материјала (погледати (Витас, 2023)).

С друге стране, информатички оријентисани математичари у то исто доба се упознају са теоријом формалних језика и, посебно, са теоријом формалних граматика и језика Ноама Чомског и њиховом применом у компилацији програмских језика (Витас, 2018). Ова теорија природно буди и занимање за своју примену на природне језике, али су за такве експерименте недостајали неопходни ресурси – језичка грађа и програми за њену обраду. Ово је у извесном смислу последица одсуства интересовања филолога за обраду језика која је у другим срединама била покретач развоја потребних информатичких средстава.

Крајем 1977. године је у Сарајеву, у организацији Института за језик и књижевност, одржан скуп под насловом „Компјутерска обрада лингвистичких података“ (Šipka, 1978). Једини рад у коме се рачунар примењивао на материјала српског (српскохрватског) језика је приказивао неке од расподела графема у тексту (Tomić, 1978). У другим срединама ондашње Југославије била су увелико у току истраживања везана за корпусе (пре свега у Загребу), као и различити програмски експерименти са словеначким језиком.

Подстакнуте овим скупом, започете су и у Београду активности на подручју обраде српског језика. Први кораци су се састојали у развоју неопходне програмске опреме. Тако је развијен властити систем за обраду корпуса, Аурора, који је улазни текст обрађивао у две фазе (Vitas 1979). У првој је формирана интерна репрезентација улаза у облику инвертованог текста, који није зависио од улазне азбуке[3]. У

[2] https://en.wikipedia.org/wiki/Georgetown%E2%80%93IBM_experiment

[3] Текстови у систему АУРОРА су били кодирани користећи ознаке CX, CY,... за Ћ, Ч,… или Ć, Č… што је омогућило обраду која не зависи од азбуке изворног текста.

другој фази су у почетку из интерне репрезентације генерисане конкорданце целог текста (или према задатом кључу) (Слика 1), као и други индекси (фреквенцијар, атерго, итд.), а касније и друге апликације (попут откривања типографских грешака, проналажење информација или генерисања битекста).

Слика 1. Први е-текстови и конкорданце

Непосредно за овим кораком се приступило сложенијем питању: формализацији именске морфологије, а резултат је био генератор именичке и придевске деклинације на основу морфографемске дефиниције леме (слика 2). Овај корак је омогућио израду парцијално лематизираних конкорданци текста, а представљао је и почетак рада на пројекту израде морфолошког електронског речника српског језика (Vitas 1980, Витас 2023).

Слика 2. Први пример лематизације и обележавања у српскохрватском језику

У исто време, током 1978, формиран је Семинар за математичку и рачунарску лингвистику у Математичком институту[4], а нешто касније, почев од 1981. године, и научно-истраживачки пројекат „Математичка и рачунарска лингвистика“ где су се окупили, с једне стране, математичари и информатичари, а са друге, лингвисти. Занимљиво је да су у том раном периоду живо интересовање за обраду језика показивали лингвисти коју се бавили другим језицима, а не српским (англисти, германисти, романисти и други). Крајем 1981. године, Математички институт је на Филолошком факултету организовао тродневни семинар „О репрезентативним узорцима природних језика (теорија корпуса)“ на коме је др Волфганг Тојберт са Института за немачки језик у Манхајму излагао немачка искуства у изградњи и експлоатацији корпуса.

Из ових почетних активности у Србији се да видети да је развој метода обраде српског језика почео међу информатичарима за разлику од других средина у којима су лингвисти били носиоци таквих активности. Природно је да су информатичари били заинтересовани, пре свега, за информатичке методе које омогућавају обраду једног језика, док су питања која се постављају једном лингвисти била у другом плану – информатичке компетенције их не могу формулисати.

Могућности масивних обрада текста су биле у оно доба вишеструко ограничене. С једне стране су то биле расположиве информатичке технологија где је унос текста вршен преко бушених картица, са врло ограниченим меморијским капацитетима (унутрашња меморија од 128KB) и користећи неприкладни програмски језик (Fortran IV). С друге стране, текстови у е-облику нису били доступни премда је већина великих штампарија у Београду већ тада користила затворене рачунарске системе са излазом на фотослогу. Из оваквих система није било могуће добити потребан излаз на магнетној траци или неком другом носачу који би омогућио преузимање већ унетих текстова у друге рачунарске системе. Такође, језици за обележавање текста, типа SGML и XML, који омогућавају преносивост, вишеструко коришћење и одрживост, још нису били развијени.

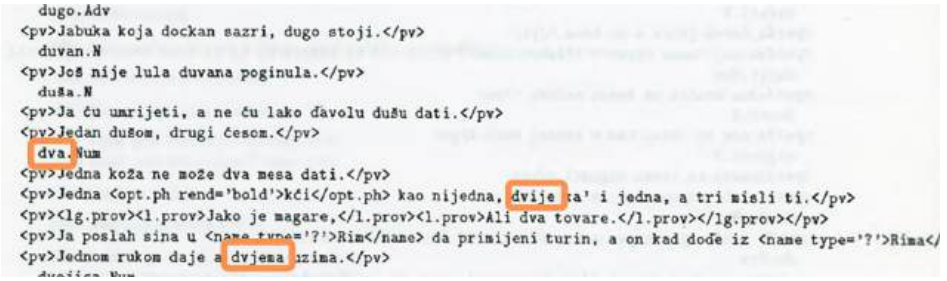
Упркос овим ограничењима, током прве деценије обраде српског језика било је прикупљено текстова у укупној дужини од око милион речи. Ова колекција се састојала од текстова уџбеника, литерарних и законских текстова и није имала посебно издиференцирану структуру. Значајан део литерарних текстова стигао је захваљујући Марк Хенингу

[4] Далеки наследник овог семинара је данас семинар који се одржава у организацији Друштва за језичке ресурсе и технологије – JePTex.

са Универзитета у Орхусу који је сканирао известан број романа на српскохрватском језику[5].

Напредак у обради српског језика се огледа и у три конференције са међународним учешћем које је, као наставак сарајевске конференције, организовао институт Јожеф Штефан из Љубљане 1982, 1985. и 1988. године. Преглед ових зборника показује убрзани раст броја учесника из Србије, као и разноврсност тема које су обрађиване. Занимљиво је да се у опсежној библиографији радова из опште лингвистике за 1988. годину (Филолог 46[6]) забележен само један од радова са конференције одржане 1988. године: остали радови на материјалу српског језика нису забележени јер су долазили из информатичких средина!

Прва верзија морфолошког речника и идеја протографеме помоћу система Аурора је имплементирана у оквиру докторске дисертације на тексту Вукових пословица (Krstev 1996, Krstev 1997). Протографема је коришћена на нивоу интерне репрезентације речника да омогући екстракцију различитих дијалекатских облика одреднице у аутоматској лематизацији (Слика 3).



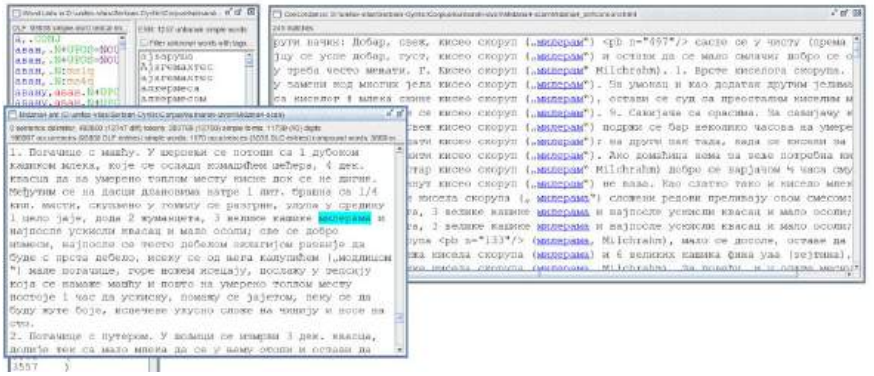
Слика 3. Лематизиране конкорданце Вукових пословица

Од краја осамдесетих година прошлог века, истраживања српског језика су била укључена прво у кооперативни пројекат „Језичке индустрије“, а затим и у пројекат Telri[7] (Trans-European Language Resources Infrastructure) средином деведесетих година двадесетог века упркос санкцијама Уједињених нација које су истраживачима отежавале међународну сарадњу. Ови пројекти су омогућили да се унапреде методе обраде српског језика кроз сарадњу са истраживачким групама других европских језика у овој области. Посебно значајни резултати су били морфосинтаксички опис српског језика у формализму Multext-East

[5] http://www.yurope.com/books/yu/
[6] https://dais.sanu.ac.rs/bitstream/id/4224/bitstream_4224.pdf
[7] http://telri.nytud.hu/start.html i http://telri.nytud.hu/telriass/index.html

(Krstev и други 2004) и примена овог описа у обележавању и лематизацији Орвеловог романа 1984 (Krstev и други 2011) као и паралелизацији овог текста са енглеским оригиналом.

Поред ових пројеката, посебно подршку развоју обраде српског је пружила париска лабораторија LADL (Laboratoire d'Automatique Documentaire et Linguistique) коју је водио проф. Морис Грос[8]. Из ове сарадње је потекла пуна формализација флективне морфологије српског језика и започет је развој морфолошког електронског речника српског на истим принципима који су примењени у изради оваквих речника за француски, енглески и бројне друге језике[9] (Krstev 2008). Теоријску основу ових речника представља теорија коначних аутомата и трансдуктора (Gross, 1989). У оквиру ове сарадње, београдска група за језичке технологије је овладала програмским алатом за обраду корпуса методом лексичког препознавања Intex (данас Unitex/GramLab)[10] (Слика 4). Овај систем гради интерну репрезентацију текста у облику коначног аутомата, а претрага се врши комплексним упитима који се формулишу на основу садржаја електронских речника у облику коначних трансдуктора, чиме је омогућена не само претрага, већ и модификација корпуса. Из исте школе је преузет и систем CorpusWeb, саставни део програма GlossaNet за надгледање веба (Fairon, 1998), а уз помоћ овог система, био је прикупљен значајан број текстова са раног српског веба, а затим и обрађен помоћу е-речника. У овом окружењу су настали и први већи електронски корпуси српског језика који су обрађивани и експлоатисани локално.



Слика 4. Интерфејс алата Unitex/GramLab са конкорданцама и речником текста

[8] https://fr.wikipedia.org/wiki/Maurice_Gross
[9] https://infolingu.univ-mlv.fr/MenuPrincipal.html#
[10] https://unitexgramlab.org/

3. Први корпус српског језика на вебу

Са продором персоналних рачунара и развојем веба, и међу српским лингвистичким круговима се почело развијати интересовање за поступке аутоматске обраде српског језика. Резултат ове нове атмосфере је, поред осталог било и формирање научно-истраживачког пројекта „Интеракција између текста и речника“[11]. Један од циљева пројекта је био изградња и постављање на веб корпуса савременог српског језика, пре свега захваљујући настојањима и подршци проф. Љубомира Поповића и проф. Љиљане Суботић. Како је пројекат био теоријске природе, „практична“ решења нису била прихватана као резултат рада на пројекту тако да је ова прва верзија корпуса узгредни резултат овог пројекта без одговарајуће валоризације. Са скромним људским и технолошким ресурсима, током 2003. године корпус је компилиран под системом IMS CWB[12] (Utvić, стр. 249) са веб-интерфејсом састављеним на Математичком факултету (Слика 5). За приступ корпусу било је неопходно да се добије корисничка идентификација. Током 20 година постојања, за приступ овој верзији корпуса[13] је регистровано око 900 корисника из земље и света који су поставили више од милион упита.

Ова верзија данас носи ознаку СрпКор-2 или СрпКор2003. Његова основна намена је била да пружи истраживачима савременог српског језика експерименталну грађу за текућа истраживања, али и да развије културу употребе електронских језичких ресурса. У том светлу, претраживање је у првој верзији корпуса било ограничено на проширене регуларне изразе без других могућности које пружа систем IMS CWB.

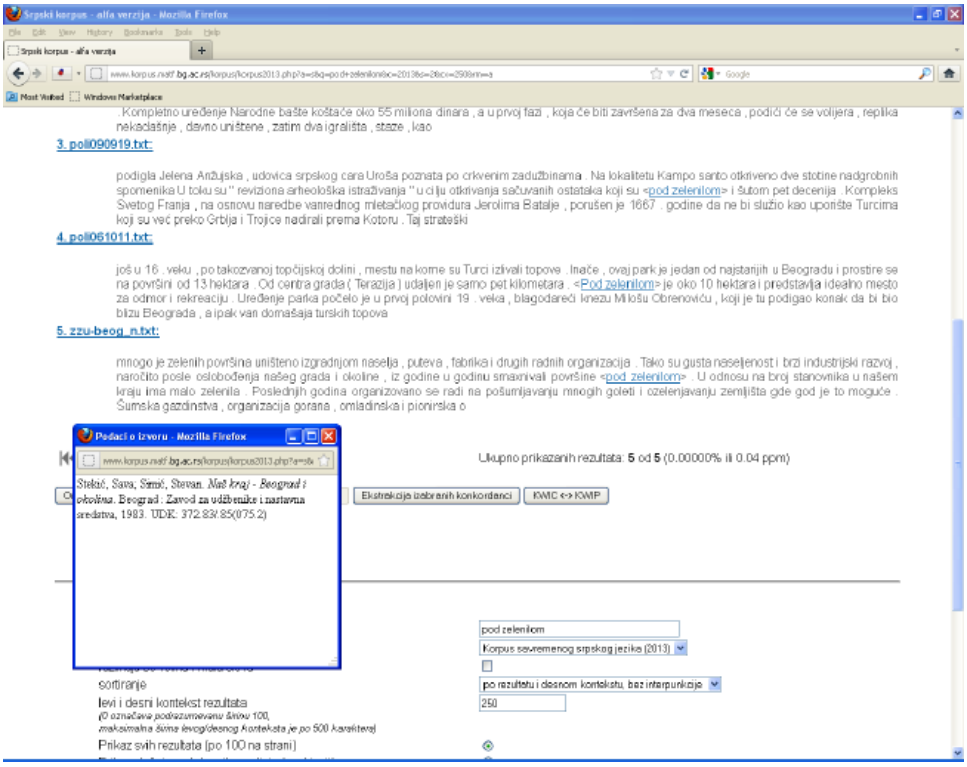
Корпус је садржавао око 23 милиона речи сировог текста без икаквих анотација. Грађа која је унета у корпус, поред различитих извора преузетих са веба, садржала је и материјал који је прикупљан од раних осамдесетих година 20. века, као и известан број литерарних текстова које су донирали њихови аутори. Међу овим текстовима посебно место је чинила преводна књижевност. У корпус је била укључена и уџбеничка литература и мањи број научних монографија из различитих области. Покушај балансирања садржаја корпуса је морао

[11] Пројекат 101743 „Интеракција између текста и речника“ је финансирало Министарство науке и технолошког развоја у периоду 2002-2006. година као пројекат основних истраживања у области језика, а реализатори су били Математички и Филолошки факултети Универзитета у Београду и Филозофски факултет Универзитета у Новом Саду.

[12] https://cwb.sourceforge.io/index.php

[13] Ова верзија је и даље доступна на адреси http://www.korpus.matf.bg.ac.rs/korpus/login.php (приступљено 22.04.2025.)

бити напуштен пред дилемом да ли саставити балансиран или велики корпус. Наиме, једномесечна продукција једних дневних новина, нпр. „Политике“, је премашивала у броју речи укупну годишњу домаћу литерарну продукцију.



Слика 5. Веб корпус на Математичком факултету

Проблеми који су се појавили приликом састављања корпуса су обухватали, пре свега, проблем конверзије на јединствену кодну шему ћириличних и латиничних текстова који су могли бити и у различитим кодним шемама. Значајан део грађе, прикупљене са веба, већ следеће године је био постао недоступан преласком сајтова са ISO-889x-серије на Unicode. Такође, садржај корпуса је ограничен на српски језик екавског изговора. За грађу на српском језику ијекавског изговора је планирано да се формира одвојен корпус када се за то стекну услови[14].

[14] Као почетни корак ка формирању корпуса српског језика ијекавског изговора започет је пројекат дигитализације описан у (Шуђур & Марковић, 2025).

Даље модификације корпуса су донеле верзију са метаподацима о изворима и пуни репертоар функција у претраживању које пружа систем IMS CWB.

Наставак рада на корпусима савременог српског језика је 2013. донео нову верзију овог корпуса која носи ознаку СрпКор3 или СрпКор2013. Ова верзија, иницијално доступна кроз исти интерфејс као и СрпКор2, је његов прави надскуп проширен текстовима прикупљеним после 2003. године. Овај корпус, који садржи 122 милиона речи, је у технолошком смислу пратио она решења која су била примењена у изградњи СрпКор3, али је значајно обогаћен како у погледу разноврсности текстова, тако и расположивих функција. Овај корпус је лематизиран са обележеном врстом речи и основним метаподацима о изворима. Његов детаљан опис и структура су дати у (Утвић 2013).

4. Даљи развој корпуса и верзије СрпКор-а

Даљи развој пројекта прикупљања текстова и обраде корпуса био је у великој мери вођен еволуцијом корисничких захтева и све израженијим информатичким потребама циљних група корисника корпуса – информатичар с једне стране, а са друге лингвиста, истраживача друштвених наука, педагога и шире јавности заинтересоване за анализу језика. Иницијални захтеви за једноставну претрагу по речима и фразама постепено су прерасли у потребу за комплекснијим упитима који укључују морфолошке, синтаксичке и семантичке критеријуме, као и за напредним статистичким алатима за анализу података из корпуса (нпр. извлачење колокација, стварање фреквенцијских листа, анализа дистрибуције по регистрима, ауторима и сл.).

Паралелно са овим, развој информатичких технологија и појава нових софтверских решења за рад са великим текстуалним подацима учинили су доступним моћније платформе за управљање корпусима. Узимајући у обзир како растуће корисничке захтеве, тако и потребе за ефикасним управљањем значајном количином анотираног текста, донета је одлука о миграцији или имплементацији на платформу која може да пружи напредније функционалности у односу на ранија решења.

Након анализе доступних опција, изабрана је платформа NoSketchEngine (Kilgarriff и други 2014). Ова платформа је препозната као робусно и флексибилно решење које пружа широк спектар алата за корпусну лингвистику, укључујући подршку за комплексне упите користећи језик CQL (Corpus Query Language), могућност креирања

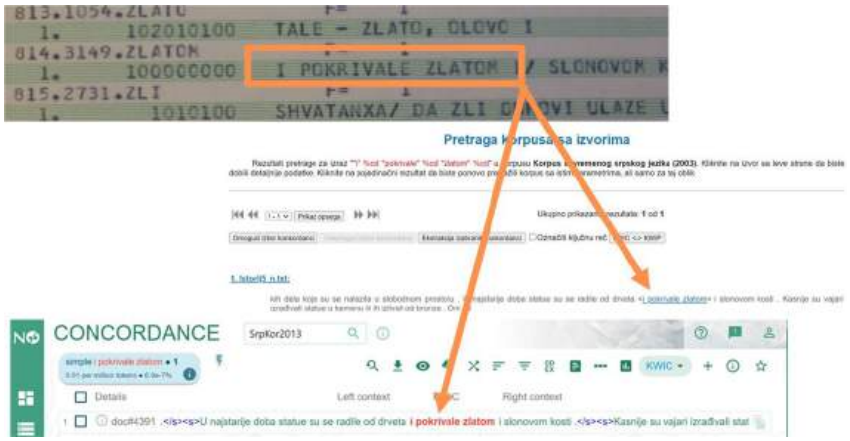
различитих врста индекса и анотација, као и напредне алате за анализу. Имплементацијом корпуса на платформи NoSketchEngine, омогућено је не само испуњење тренутних, већ и предвиђање будућих корисничких захтева, значајно унапређујући могућности истраживања и анализе доступних текстуалних података.

Sketch Engine (Kilgarriff et al., 2004) је софтвер за корпусну лингвистику који обрађује унете корпусе и граматичке шаблоне, генеришући детаљне информације о употреби речи ("word sketches"), тезаурусе и поређења блиских синонима ("sketch differences"). То је комерцијални алат који захтева претплату, док NoSketch Engine представља верзију отвореног кода Sketch Engine-a, али са значајним ограничењима у функционалности. За разлику од комерцијалне верзије, NoSketch Engine не долази са готовим корпусима – корисник мора сам да припреми своје корпусе и поседује техничко знање за њихову обраду и форматирање. Недостају многи напредни алати за анализу (као што су "word sketches", тезаурус, н-грами, трендови), као и алати за аутоматску изградњу корпуса (токенизација, лематизација, тагирање) и графички интерфејс за управљање корпусима. Корисник је у потпуности одговоран за преузимање, инсталацију, хостовање и администрацију NoSketch Engine-a. Подршка је доступна преко Google групе, а постоји и демо инсталација за пробу функционалности.

Одржавање и развој корпуса је од 2018. године преузело Друштва за језичке ресурсе и технологије – JePTex[15], које је урадило миграцију корпуса на нову платформу[16], а потом наставило са даљим унапређењима и проширењима текстуалних ресурса. Када је реч о доступним корпусима, са циљем да се у одређеној мери балансирају критеријуми представљања општег језика, доступна су два корпуса. Први од њих је корпус савременог српског језика СрпКор2013 који је преузет са сајта Математичког факултета Универзитета у Београду[17] (поменут у претходном одељку). Током миграције се прешло на латиницу (UNICODE) тако да је замењено ASCII кодирање осмишљено у првом систему за обраду текста АУРОРА. Крајем 2021. године објављена је нови корпус који је допуна претходном и не садржи текстове корпуса СрпКор2013. У поређењу са претходном верзијом (СрпКор2013), СрпКор2021 обухвата више од 600 милиона речи, што га чини једним од највећих корпуса савременог српског језика који има донекле балансиран садржај. Корпус је комплетно анотиран, где је сваки токен обележен припадајућом врстом речи и

[15] http://jerteh.rs
[16] https://noske.jerteh.rs/
[17] http://www.korpus.matf.bg.ac.rs/korpus/login.php

лемом (основним обликом) (Stanković и други 2020). Увођење новог, робуснијег система за управљање корпусним текстовима омогућава његово перманентно унапређивање кроз лако додавање нових корпуса и ажурирање постојећих података. Треба напоменути да се током свих ових миграција водило рачуна да не дође до губитка информација, чему сведочи слика 6, где видимо примере исте конкорданце генерисане системом АУРОРА из прошлог века, а затим у системима СрпКор2 (под IMS CQP) и СрпКор3 (NoSketch).



Слика 6. Животни пут текста

Разноврсност садржаја представља кључну карактеристику фамилије корпуса СрпКор. Поред великог броја новинских текстова, корпус укључује текстове са Википедије, књижевне текстове, уџбенике из широког спектра научних и образовних области, докторске дисертације, као и друге врсте текстова из различитих домена. Ова разноликост обезбеђује бољу репрезентативност различитих регистара и стилова савременог српског језика. У СрпКор унети су и текстови на српском језику из паралелизованих корпуса који су настајали упоредо са СрпКор-ом[18]. На овај начин је делимично компензован утицај веб-садржаја на састав корпуса. С друге стране, овакви текстови који су, по правилу, изузетно значајни у културном смислу, јер не само да нису присутни у грађи са веба, већ обично не улазе ни у традиционалне лексикографске корпусе. Њих чине одабрани преводи на српски језик научних, књижевних, филозофских, антрополошких, историјских и сличних текстови преузети из едиција угледних издавачких кућа.

[18] Посебно богата колекција књижевних текстова је преузета из италијанко-српског корпуса.

Корпуси су доступни за претрагу регистрованим корисницима платформе NoSketch на сајту Друштва JePTex[19]. Поред СрпКор2021, NoSketch платформа на сајту JePTex подржава и домаће доменске корпусе из специфичних области (математика, рударство, геологија, итд.) демонстрирајући флексибилност и моћ платформе за разноврсна истраживања. Табела 1 приказује квантитативне карактеристике изабраних корпуса. У табели су зеленом бојом засенчени корпуси који се доступни уз бесплатну регистрацију (текстови са Википедије, SrpELTeC потколекција (Krstev и други 2021), паралелни Српско-италијански корпус (Moderc и други 2023)) чиме се подстиче шира употреба платформе, док је за коришћење корпуса СрпКор и осталих корпуса потребна симболична претплата у виду чланарине у Друштву.

Табела 1. Статистика изабраних корпуса на платформи noske.jerteh.rs

Корпуси	Токени	Речи	Документи	Реченице
BiKes_sr	1.848.802	1.547.103	336	82.459
BiKes_en	2.072.074	1.766.421	336	82.459
GeoSrpKor	1.316.646	1.067.583	69	
ItSrNER_sr	314.619	254.828	4	10000
ItSrNER_it	331.375	272.026	4	10000
Matematika	1.932.271	1.341.242	89	
RudKorp	3.542.016	2.713.594	172	
SerbItaCor3_sr	12.306.854	10.225.632	267	
SerbItaCor3_it	13.166.565	11.051.704	267	
SkolKor	4.736.884	3.650.849	82	
SrpELTeC	5.923.024	4.766.056	108	144.387
SrFudKo	9.948.301	8.318.399	37	487
SrpKor2013	145.275.324	121.142.927	5.038	5.178.311
SrpKor2021	716.878.652	606.683.682	1.916	29.718.407
WikiKorpus	81.267.249	64.510.812	609.380	609.380
ZlatniKorpus	1.333.804	1.104.864	8	
Укупно	**1.002.194.460**	**840.417.722**	**618.113**	**35.835.890**

Међу књижевним текстовима посебно се издваја српска потколекција Европске колекције књижевних текстова (ELTeC), која обухвата у основној верзији 100 романа, припремљених у оквиру међународне COST акције CA16204 *Distant Reading for European Literary*

[19] https://noske.jerteh.rs

History[20], чиме се омогућава компаративна анализа са корпусима других европских језика. „Плус“, односно проширена, верзија SrpELTeC+ представља проширење по различитим димензијама, величини, периоду, варијететима текста. Актуелна верзија SrpELTeC+ тренутно садржи више од 190 дела, међу којима су осим романа, путописи и приповетке, а на корпусу се интензивно и даље ради. Важан ресурс свакако представљају резултати прве фазе пројекта “Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920)” који реализује Центар за дигиталну хуманистику Филозофског факултета Пале, а подржава Министарство за научнотехнолошки развој и високо образовање Републике Српске. Изградња Корпуса српског књижевног наслеђа ијекавског изговора обухватила је 46 публикација, и то приповетке, путописе и романе (Шуњур & Марковић 2025).

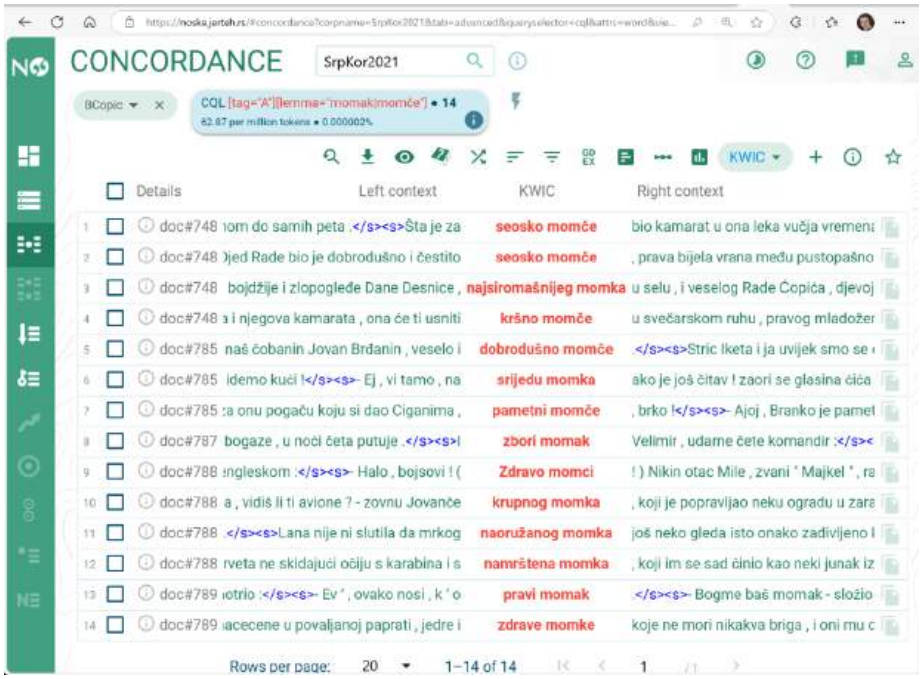
Ниво метаподатака придружених корпусима се разликује, али оно што је заједничко је да су забележени подаци о аутору, години издања, наслову, и регистру текста. Дистрибуција текстова по регистрима у корпусима српског језика из табеле 1 (обједињено), рачунато према броју речи, показује да административни текстови чине 1,3%, књижевност 8,7%, наука 3,2%, новински текстови 67,6%, разговорни 10,4%, уџбеници 0,5%, вики 7,3%, и друго (некласификовано) 1,1%.

На основу метаподатака могуће је креирати подкорпусе и над њима постављати упите, тако да можемо истраживати рецимо само административни језик или, на пример, само језик једног писца. Слика 7 приказује упит на покорпусу BCopic (издвојена дела Бранка Ћопића) корпуса SrpKor2021 којим се тражи одговор на питање „Који придеви прате именицу *момак*?“, што преведено на језик CQL гласи: [tag="A"][lemma="momak|momče"]. Резултат је приказан у облику конкорданци из којих корисник може да види из ког Ћопићевог дела долазе примери и њихов контекст.

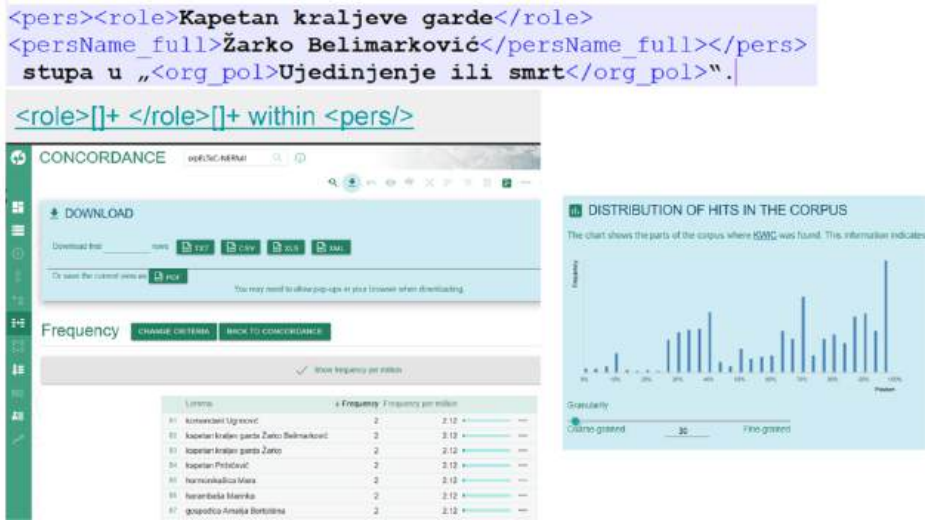
Осим анотација на нивоу речи (лема и врста речи) које су биле поменуте, одређени скуп текстова има и структурне анотације – обележени су одељци, пасуси и реченице, а у неким текстовима и именовани ентитети. Највећи део текстова је обележен са 7 категорија именованих етикета (Stanković и други 2021) одабраних за SrpELTeC корпус (Krstev и други 2021), али један део текстова је обележен и много богатијим скупом етикета коришћењем система за препознавање именованих етикета заснованог на морфолошким речницима и аутоматима (Krstev и други 2019). Слика 8 приказује пример коришћења структурног обележавања у упитима. Упит тражи

[20] https://www.distant-reading.net/

појављивање етикете <role> (занимања, титуле, позиције) унутар етикете <pers> (особа).



Слика 7. Пример упита на подкорпусу



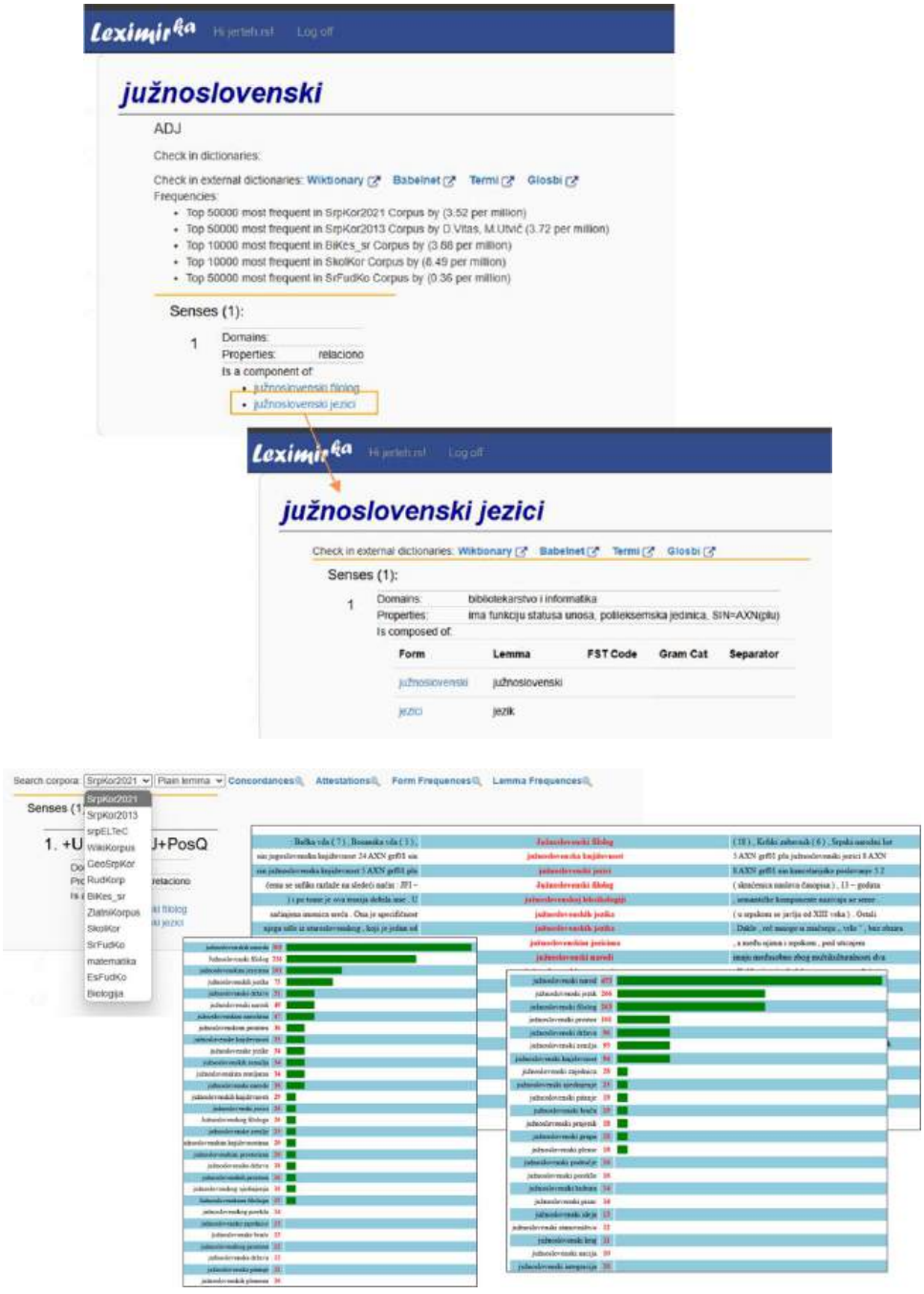
Слика 8. Пример структурно обележеног текста и упита који укључује структурне етикете

5. Интеракције СрпКор-а са другим ресурсима

Интеракција корпуса савременог српског језика (СрпКор) са другим језичким ресурсима представља кључни аспект модерне лингвистике и обраде природних језика. Ова спрега, првенствено са речницима који пружају структурирано лексичко знање и језичким моделима који обухватају дистрибутивне и семантичке односе, омогућава обострано обогаћивање. Корпус пружа емпиријску основу за валидацију и проширење речника, док речници и модели унапређују могућности корпусних анализа и анотације.

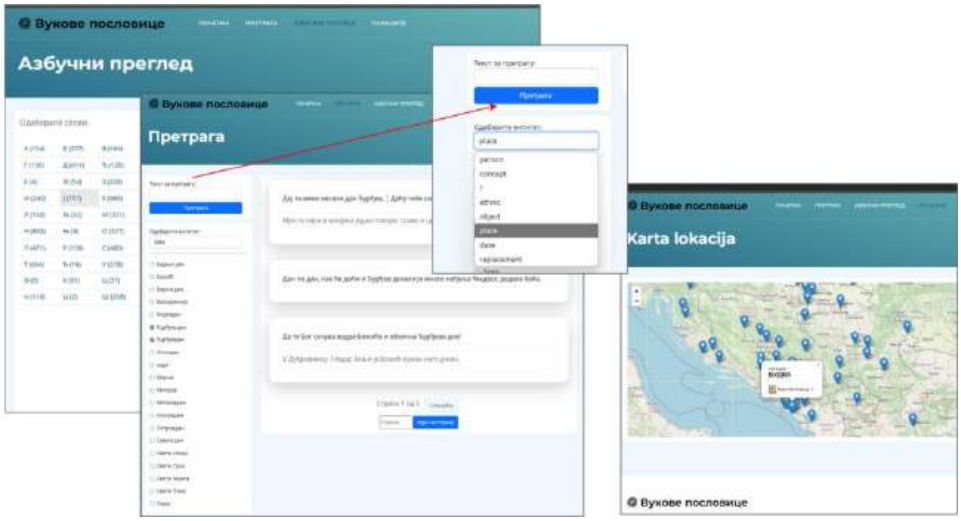
Успостављена је снажна двосмерна спрега између лексичке базе Лексимирка и фамилије корпуса СрпКор (Lazić, Škorić, 2020). Напоменимо да је база Лексимирка настала трансформацијом електронских речника за српски језик (Krstev 2008) у релациону базу (Stanković и други 2018) прилагођену за даље серијализације у екосистем повезаних лингвистичких података. Ова интеракција омогућава обострано обогаћивање ресурса. Са једне стране, кроз специјализовани интерфејс лексичке базе Лексимирка, корисници имају директан и интерактиван увид у примере употребе речи и њихових различитих облика у аутентичном корпусном контексту (слика 9). Такође је могуће истраживати појављивање речи у специфичним синтаксичким обрасцима или колокацијама детектованим на корпусу. Са друге стране, лексичка база и придружени морфолошки ресурси имају кључну улогу у унапређењу система за лематизацију корпуса. Систем за лематизацију се континуирано унапређује из верзије у верзију корпуса, а његова прецизност и обухватност директно зависе од квалитета и обухвата електронских морфолошких речника српског језика. Ови речници, коришћењем моћног система за обраду текста, попут Unitex-a, омогућавају не само похрањивање лема и њихових основних морфолошких информација, већ и прецизно генерисање свих валидних флективних облика сваке леме. Ова свеобухватна морфолошка информација је затим интегрисана у процес анотације корпуса, обезбеђујући да се свака појава речи у тексту исправно повеже са својом основном лемом.

Речнички чланци базе Лексимирка повезани су и са Ворднетом (Krstev и други 2004, Stanković и други 2018), терминолошком базом Терми (Kitanović и други 2021), одређеним бројем дигитализованих традиционалних речника и речницима са веба.



Слика 9. Интерфејс лексикона и корпуса на примеру Лексимирке и СрпКор-а: проналажење разноврсних информација полазећи од упита “јужнословенски” у систему Лексимирка.

Напоменимо још један пример текста који је прошао вишедеценијску трансформацију и добио ново „рухо“. Дигитализација Вукових пословица и израда њиховог индекса део су резултата докторске дисертације “Један прилаз информатичком моделирању текста и алгоритми његове трансформације” (1997) проф. др Цветане Крстев (Krstev 1996, Krstev 1997). Из изворног SGML формата који је пратио смернице TEI, генерисан је XML облик на основу ког су подаци преточени у PosgreSQL релациону базу. Развијена је веб апликација НарПос[21] (слика 10), уз помоћ које се могу прегледати пословице уређене у азбучном поретку, а могу се и, претраживати преко индекса, филтрирати по темама али и проналазити на карти, на основу информације која се налази у самом тексту пословице или у објашњењу које је прати).



Слика 10. Веб апликација Народне пословице

Развој савремених језичких модела, укључујући оне за обраду и разумевање српског језика, фундаментално зависи од доступности и квалитета великих текстуалних корпуса. Иако веб-корпуси нуде неупоредиву количину података и ефикасно одражавају тренутну, често неформалну, употребу језика на интернету, њихова нехомогеност, присуство грешака, сленга, нестанардних варијанти и недостатак конзистентне, дубинске лингвистичке анотације могу представљати ограничење за тренирање модела који захтевају разумевање структурираног и формалнијег језика. Због тога су

[21] https://narpos.jerteh.rs/

пажљиво прикупљени, балансирани и анотирани корпуси, чији садржај не потиче искључиво са веба (већ укључује и књижевна дела, научне радове, званичне документе, транскрибовани говор, итд.), незаменљиви за изградњу робусних језичких модела способних да обрађују различите регистре и стилове. Ови "чистији" и структурираније корпуси пружају поузданију основу за учење сложених граматичких образаца, нијанси семантике и специфичности различитих домена, што је од пресудног значаја за развој модела високе прецизности и поузданости, посебно за задатке који превазилазе површинску обраду текста. Подразумева се да је припрема оваквих корпуса далеко захтевнија како у фази прикупљања текстова тако и у фази њихове обраде која претходи укључивању у корпус.

Пројекат ТЕСЛА[22] посебно се бави векторизацијом текста – представљањем речи, фраза или читавих докумената као нумеричких вектора, што је темељна техника у савременој обради природних језика. Овим се отварају врата за развој низа напредних апликација за српски језик – од анализе осећања и ставова и класификације текста до машинског превођења и система за одговарање на питања. Успешност ових апликација директно корелира са доступношћу и квалитетом великих колекција текстуалних података, а посебно анотираних корпуса који пружају лингвистички обогаћене информације. У том контексту, планови за будући развој корпуса СрпКор усмерени су ка стварању још већег и детаљније анотираног ресурса.

6. Закључак

У овом раду пружамо свеобухватан преглед развојног пута корпуса савременог српског језика. Полазећи од почетних фаза конципирања и прикупљања података, пратимо стварање првог корпуса српског језика доступног путем веб-технологија, што је представљало значајан корак у дигитализацији и доступности језичких ресурса за српски. Детаљно елаборирамо даљи развој и различите итерације корпуса, које су донеле значајна проширења у обиму, разноврсности садржаја и дубини лингвистичке анотације, одражавајући еволуцију истраживачких потреба и технолошких могућности. Коначно, представљамо актуелну фазу развоја, фокусирану на интеграцију и синергију корпуса са другим значајним језичким ресурсима, као што су лексичке базе и језички модели, што

[22] https://tesla.rgf.bg.ac.rs/

омогућава обострано обогаћивање и отвара нове могућности за напредне језичке анализе и апликације. Овај преглед документује трансформацију корпуса од пионирског подухвата до савременог, интероперабилног ресурса, илуструјући изазове и решења током више декада развоја.

Даљи развој корпуса савременог српског језика, усмерен ка стварању робустне и детаљне лингвистичке основе за напредне апликације попут векторизације текста, планиран је кроз више паралелних активности. Крајем 2025. године планирано је објављивање нове верзије корпуса на платформи NoSketch, која ће обухватити значајно проширење садржаја. Ово проширење укључиће новинске текстове из периода 2022-2025, као и нове типове текстова попут путописа, мемоара и историјских новина, проширујући временску димензију и доприносећи културно-историјској вредности корпуса, уз континуирано допуњавање одабраним веб-садржајима. Паралелно са повећањем обима, фокус је стављен на детаљније обогаћивање метаподатака. Ово ће омогућити флексибилније креирање подкорпуса према различитим димензијама, укључујући период, домен, изговор, ауторе или специфичне типове текста, чиме ће се омогућити прецизнија циљана истраживања. Квалитет корпуса ће бити додатно унапређен кроз допуну и обогаћивање лингвистичких анотација. У плану је комплетирање поделе на реченице, препознавање именованих ентитета, те додавање детаљнијих граматичких информација и повезивање са базама знања. У том циљу, активно се ради на развоју и унапређењу модела и алата за аутоматску анотацију који ће подржати, не само стандардно означавање врста речи (PoS) и лематизацију, већ и сложеније задатке попут препознавања именованих ентитета, додељивања значења речима и креирања веза ка базама знања. Сви ови напори усмерени су ка изградњи корпуса који ће пружити незаменљив ресурс за лингвистичка истраживања и развој напредних алата за обраду српског језика.

Захвалница: Ово истраживање је подржао Фонд за науку Републике Србије, #7276, Text Embeddings - Serbian Language Applications - TESLA.

Литература

[1] Fairon, C. (1998). GlossaNet: Parsing a web site as a corpus. *Lingvisticæ Investigationes*, 22(1-2), 327–340.

[2] Gross, M. (1989). The Use of Finite Automata in the Lexical Representation of Natural Language; in: Gross, M., Perrin, D. (Eds.). *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science. Berlin: Springer Verlag.

[3] Dobrić, N. (2012) "Savremeni jezički korpusi na Zapadnom Balkanu–Istorijat, trenutno stanje i budućnost." *Slavistična revija* 60: 677-692.

[4] Ivić, M. (2001). *Pravci u lingvistici* (Vol. 2, 9. izd.). Čigoja štampa, Biblioteka XX vek, Beograd.

[5] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The sketch engine. *Lexicography*, 1(1), 7-36.

[6] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress* (стр. 105-116).

[7] Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I., & Kolonja, L. (2021). A data-driven approach for raw material terminology. Applied Sciences, 11(7), 2892.

[8] Krstev, C. (1996) [računarska obrada leksike], *Vukove narodne poslovice s registrom ključnih reči*, ed. Slobodan Đorđević, Nolit, Beograd, 1996.

[9] Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije* (doktorska disertacija, Univerzitet u Beogradu, Matematički fakultet).

[10] Krstev, Cvetana. *Processing of Serbian. Automata, texts, and electronic dictionaries*. Faculty of Philology of the University of Belgrade, 2008.

[11] Krstev, C. Pavlović-Lažetić, G., Vitas, D., & Obradović, I. (2004) Using Textual and Lexical Resources in Developing Serbian Wordnet, *Romanian Journal of Information Science and Technology*, Romanian Academy, Publishing House of the Romanian Academy, 7(1-2), 147-161.

[12] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2), 473-489.

[13] Krstev, C., Vitas, D., & Erjavec, T. (2004). MULTEXT-East Resources for Serbian. У T. Erjavec & J. Zganec Gros (Ур.), *Zbornik 7. mednarodne multikonference "Informacijska druzba IS 2004", Jezikovne tehnologije*. Institut "Jožef Stefan".

[14] Krstev, C., & Vitas, D. (2005) “Corpus and Lexicon - Mutual Incompleteness”, *Proceedings of the Corpus Linguistics Conference*, 14-17 July 2005,

Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/.

[15] Krstev, C., Vitas, D., & Trtovac, A. (2011). Orwell's 1984 – the Case of Serbian Revisited. У Z. Vetulani (Ур..), *Proceedings of 5th Language & Technology Conference* (стр. 570–574). Fundacja Uniwersytetu im. A. Mickiewicza.

[16] Krstev, C., & Stanković, R. (2022). *Deliverable D1.35 Report on the Serbian Language. European Language Equality (ELE)*; EU project no. LC-01641480–101018166. https://europ ean-language-equality.eu/reports/language-report-serbian.pdf.

[17] Krstev, C., & Stanković, R. (2023). Language Report Serbian. In: Rehm, G., Way, A. (eds) *European Language Equality*. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_32

[18] Lazić, B., & Škorić, M. (2020). From DELA based dictionary to Leximirka lexical database. *Infotheca - Journal For Digital Humanities*, 19(2), 81-98. doi:10.18485/infotheca.2019.19.2.4

[19] Moderc, S., Stanković, R., Tomašević, A., & Škorić, M. (2023). An italian-serbian sentence aligned parallel literary corpus. *Review of the National Center for Digitization*, 43 (2023), 78–91.

[20] Pierce, J. R., & Carroll, J. B. (1966). *Language and machines—Computers in translation and linguistics* [ALPAC report]. National Academy of Sciences, National Research Council, Washington, DC.

[21] Rajić, Lj. (1981). *Teorija i poetika prevođenja*. Prosveta, Beograd.

[22] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). Annotation of the serbian eltec collection. *Infotheca–Journal for Digital Humanities*, 21(2), 43-59.

[23] Stanković, R., Krstev, C., Lazić, B., & Škorić, M. (2018). Electronic Dictionaries–from File System to Lemon-based Lexical Database. In the *6th Workshop on Linked Data in Linguistic (LDL-2018),* Towards Linguistic Data Science.

[24] Stanković, R., Mladenović, M., Obradović, I., Vitas, M., & Krstev, C. (2018). Resource-based WordNet augmentation and enrichment. У *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)* (стр. 104-114).

[25] Stanković, R., Sandrih, B., Krstev, C., Utvić, M., & Škorić, M. (2020). Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In LREC 2020-12th International Conference on Language Resources and Evaluation, Conference Proceedings (pp. 3954-3962). European Language Resources Association (ELRA).

[26] Šipka, M. (Ур.). (1978). *Kompjuterska obrada lingvističkih podataka*, Posebna izdanja 4. Institut za književnost i jezik, Sarajevo.

[27] Tomić, T. (1978). Statistička analiza srpskohrvatskog teksta pomoću računara. У M. Šipka (Ур.), *Kompjuterska obrada lingvističkih podataka* (стр. 221–236). Institut za književnost i jezik.

[28] Utvić, M. (2013). *Izgradnja referentnog korpusa savremenog srpskog jezika.* Doktorska disertacija. Univerzitet u Beogradu, Filološki fakultet. https://nardus.mpn.gov.rs/handle/123456789/4091

[29] Vitas, D. & Krstev C. (2012) "Tvorbeni obrasci u elektronskom rečniku srpskog jezika", Međunarodni komitet slavista. Komisija za tvorbu reči. *Međunarodna naučna konferencija Tvorba reči i njeni resursi u slovenskim jezicima* (14), pp. 515-525, Filološki fakultet Univerziteta u Beogradu, Beograd, ISBN 978-86-6153-116-3

[30] Vitas, D. (1979) Prikaz jednog sistema za automatsku obradu teksta, INFORMATICA'79, Bled, (стр. 7 10)

[31] Vitas, D. (1980) Generisanje imeničkih oblika u srpskohrvatskom, *Informatica* 80(3) Slovenačko društvo za informatiku, Ljubljana, 49-55.

[32] Витас, Д. (2023) Белешке о ручној и аутоматској обради српског језика, *Језик Данас*, бр. 22, 2023, Матица Српска, Нови Сад.

[33] Витас, Д. (2018) О развоју информатике међу математичарима, *Инфотека* 18 (1), 2018,

[34] Витас, Д., Поповић, Љ. (2003) „Конспект за изградњу референтног корпуса српског стандардног језика", *Научни састанак слависта у Вукове дане* 31/1 - МСЦ, Београд, стр. 221 - 227.

[35] Шуњур, С., Марковић, Ј. (2025). Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза), Зборник радова конференције Јужнословенски језици у дигиталном окружењу ЈуДиг 2025 (у овом броју), Универзитет у Београду, Филолошки факултет.

About the Family of the Corpus of the Modern Serbian Language SrpKor

Duško Vitas, Ranka Stanković, Cvetana Krstev

Summary

The SrpKor family comprises electronic corpora of contemporary Serbian, initiated in the late 1970s and first made publicly available online in 2003. Its development, especially before the advent of abundant web-based resources, involved systematic text collection, corpus processing tool creation, and annotation methodology design. SrpKor is not merely a digital text collection but an integrated resource with software infrastructure, linguistic annotations (metadata, morphological tagging, lemmatization, named entities), and search capabilities, evolving through multiple versions under constrained resources.

Early Serbian NLP efforts in the 1980s lagged behind other European contexts, partly due to limited computational literacy among philologists and skepticism toward language technology. Initial advances came from computer scientists, leading to the Aurora corpus processing system and early morphological analyzers. International collaborations in the 1990s (e.g., Telri, LADL) introduced formal morphological modeling and tools such as Unitex/GramLab, facilitating large-scale corpus compilation.

The first web corpus (SrpKor2003) contained ~23 million unannotated words; SrpKor2013 expanded to 122 million words, with lemmatization and part-of-speech tagging. Responding to growing research needs, the corpora migrated to the NoSketchEngine platform, culminating in SrpKor2021—a 600M-word, fully annotated balanced corpus. These corpora include diverse genres: news, literature, Wikipedia, textbooks, academic texts, parallel translations, and culturally significant non-web materials. Metadata enables subcorpus creation by domain, author, or register.

SrpKor interacts closely with lexical databases such as Leximirka, enhancing both lexicon and corpus quality. Structural and named entity annotations support complex queries. Additional domain-specific corpora (e.g., mining, geology, mathematics) demonstrate the platform's versatility.

Future development focuses on expanding temporal coverage (including 2022–2025 texts), enriching metadata, refining linguistic annotation (sentence segmentation, detailed grammar, named entities), and

linking to knowledge bases. These efforts aim to provide a robust, richly annotated resource supporting advanced applications like text vectorization, sentiment analysis, and machine translation, ensuring SrpKor remains a cornerstone for Serbian language research and technology development.

Keywords: SrpKor, corpora, Serbian language, lemmatization, Leximirka