Korpus *Srpko* kao tehnološka osnova za izradu Rečnika savremenog srpskog jezika Matice srpske

Naučni rad DOI: 10.18485/judig.2025.1.ch25

Dušanka Vujović[1], *0000-0003-2469-2702*
Branko Milosavljević[2], *0000-0003-4551-9802*

Apstrakt

Digitalizacija jezičkih resursa omogućava lak i brz pristup jezičkim podacima olakšavajući tako lingvistička istraživanja kao i analizu jezičkih fenomena. Ona je, takođe, od ključnog značaja za savremene leksikografske projekte jer obezbeđuje obilje pažljivo odabrane leksičke građe, lakšu pretragu i ekscerpciju primera kao i njihovu ugradnju u rečnik. Izgradnja elektronskog jezičkog korpusa *Srpko* predstavlja temelj za izradu novog *Rečnika savremenog srpskog jezika Matice srpske*. Elektronski jezički korpus omogućava jednostavno i brzo pretraživanje tekstualnih podataka i njihovu inkorporaciju u sadržaj rečničkog teksta. U radu se govori o principima, kao i idejnim i tehničkim rešenjima u vezi sa organizacijom korpusa. Prikazuje se njegova struktura, način prikupljanja podataka i formiranja baze podataka, budući da se u korpus uključuju podaci različitih formata, poput pisanih digitalizovanih i nedigitalizovanih tekstova, kao i audio i video zapisa. U cilju brže i lakše pretrage korpusa, ali i njegove kontrole, u softver su ugrađene različite funkcionalne pretrage i zahtevi za filtriranje rezultata što je omogućilo smanjivanje šuma, odnosno redundantnih informacija koje otežavaju pretrage. U softver je ugrađena i funkcija dobijanja različitih izveštaja na osnovu zadatih filtera. Ovako koncipiran korpus pruža osnovu za sistematsko organizovanje i

[1] Hankuk University of Foreign Studies in Seoul, Faculty of East European and Balkan Studies Univerzitet u Novom Sadu, Filozofski fakultet, dusanka.vujovic@ff.uns.ac.rs
[2] Univerzitet u Novom Sadu, Fakultet tehničkih nauka, mbranko@uns.ac.rs

predstavljanje leksičke baze podataka u cilju izrade novog *Rečnika savremenog srpskog jezika*.

Ključne reči: srpski jezik, korpus srpskog jezika, digitalizacija, rečnik

1. Uvod

Krajem 2020. godine, u Matici srpskoj intenziviran je rad na projektu izrade novog *Rečnika savremenog srpskog jezika* (RSSJ). Rečnik se koncepcijski i metodološki naslanja na ranije Matičine rečnike, šestotomni *Rečnik srpskohrvatskoga književnog jezika* (RMS) i jednotomni *Rečnik srpskoga jezika* (RSJ) kao i na *Rečnik srpskohrvatskog književnog i narodnog jezika* Instituta za jezik Srpske akademije nauka i umetnosti (RSANU). Tehnologija izrade novog Rečnika zasniva se na softverskoj platformi koju smo sami osmislili i razvili za potrebe projekta. Želeli smo da kreiramo softverska rešenja koja su prilagođena specifičnostima srpskog jezika (imajući u vidu njegovu složenu prozodijsku i morfološku strukturu) i koja u formalnom prikazu odgovaraju tradicionalnoj strukturi leksikografskog članka u dosadašnjim rečnicima srpskoga jezika. Zbog toga smo odlučili da razvijemo sopstvenu softversku platformu, umesto da adaptiramo neki od postojećih programa za izradu rečnika, čija bi već gotova rešenja trebalo dodatno prilagoditi našim potrebama.

2. O programu

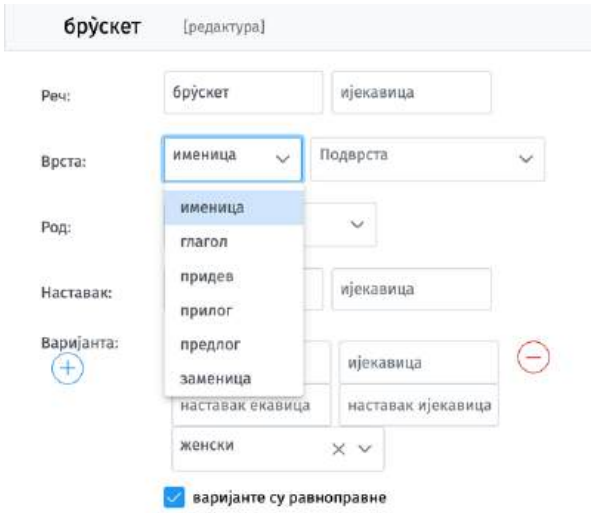
Softver za izradu Rečnika sastoji se od dva povezana programa. U cilju izrade dela programa za formiranje Rečnika prvo smo razložili elemente strukture leksikografskog članka na najmanje invarijabilne jedinice. Precizno isplanirana struktura leksikografskih elemenata, koje treba uneti u unapred definisane forme programa, veoma je važan korak u razvoju softverskih rešenja jer obezbeđuje detaljan i dosledan prikaz elemenata, njihovo efikasno kombinovanje i optimalan pristup podacima u cilju lakše ekstrakcije relevantnih informacija iz unete baze podataka (Gaus 2014: 3). Tradicionalni unos teksta na papir fokusira se na konačni izgled rečničkog članka, uz poštovanje tipografskih konvencija, dok se unos u elektronskoj formi fokusira na osnovne strukturne elemente rečničkog članka, a prikaz se formira automatski. Na taj način moguće je konzistentno menjati prikaz bez intervencija na unetom sadržaju rečničkog članka kao i formirati prikaze namenjene različitim tipovima medija kao što su štampani tekst, ekran računara ili mobilnog uređaja. Zbog kompleksnosti ukupne funkcionalnosti softvera, u ovom radu daje se prikaz samo ključnih, osnovnih funkcija.

Platforma za formiranje leksikografskog članka podeljena je na dva dela. Jedan deo predviđen je za unos leme i osnovnih gramatičkih i upotrebnih informacija vezanih za nju, odnosno, u tradicionalnoj leksikografiji tzv. leve strane rečnika. Drugi deo platforme namenjen je formiranju desne strane rečnika u koju se unosi leksikografska definicija i prateći elementi od kojih je sastavljena i koji idu uz nju[3]. Polja za formiranje leve strana rečnika (Sl. 1) predviđaju unos obaveznih i neobaveznih podataka. Neka polja imaju unapred definisanu listu mogućih vrednosti i elementi za unos dobijaju se izborom iz padajućih menija (Sl. 2), a neka omogućavaju slobodan unos teksta kao npr. polje za dodatne informacije (Sl. 1).

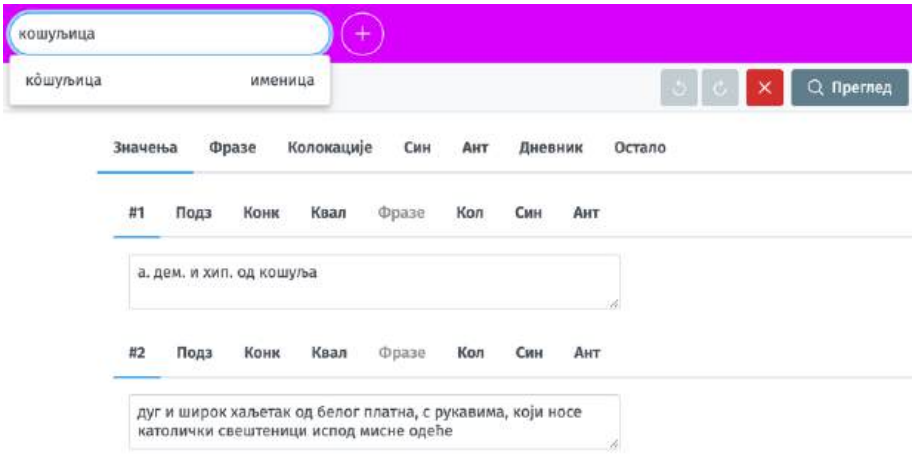
Sl. 1. Polja za unos leme i osnovnih gramatičkih i upotrebnih informacija

[3] Leva i desna strana rečnika su termini vezani za tradicionalnu leksikografiju i mi ih koristimo u radu jer je softver izrađen u cilju objavljivanja štampane verzije Rečnika u kojoj će jasno biti vidljivo šta pripada levoj, a šta desnoj strani. U elektronskoj leksikografiji pojmovi leva i desna strane rečnika nemaju više smisla jer se linearna, dvostrana struktura štampanog rečnika zamenjuje dinamičnim prikazom podataka. Organizacija rečničkih informacija više nije prostorno uslovljena, već je funkcionalna jer korisnik pristupa svim informacijama interaktivno i prema potrebi.

U polja za slobodan unos teksta omogućen je unos odredničke reči, zatim varijanata reči (bilo da je u pitanju akcenatska, morfološka ili ijekavska varijanta), različitih nastavaka reči kao i dodatnih informacija u vezi sa gramatičkim obeležjima reči, koje nismo mogli unapred da definišemo. Polja sa unapred definisanim vrednostima za unosu predviđena su za markiranje vrste reči, za gramatička obeležja kao što su rod imenica ili vid i rod glagola, za unos kvalifikatora upotrebne vrednosti reči i za obeležavanje statusa koju reč ima u trenutku obrade (npr. nova reč, upitni primer i sl).



Sl. 2. Padajući meni sa obaveznim izborom



Sl. 3. Polja za unos elemenata leksikografske definicije

Drugi deo stranice za formiranje leksikografskog članka namenjen je unosu leksikografske definicije i svih elemenata koji ulaze u njenu strukturu i dolaze uz nju. Osim polja za unos osnovnog značenja, tu su polja za unos podznačenja, konkordanci i njihovih izvora, kvalifikatora za posebna značenja, frazeologizama, kolokacija, sinonima i antonima (Sl. 3).

3. Render Rečnika iz odrednica unetih po prethodno opisanoj strukturi

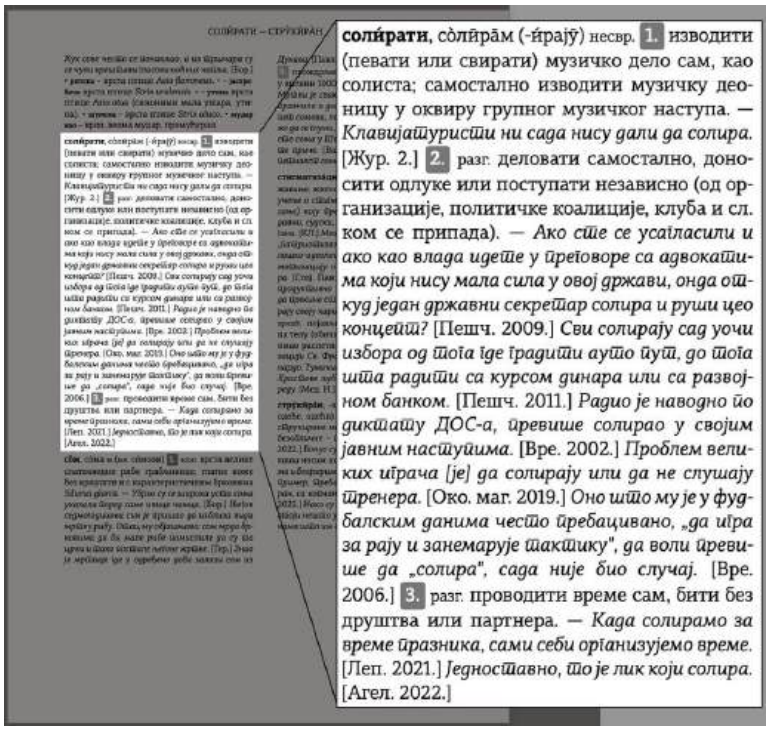
Osnovni zadatak učesnika na projektu, koji je definisala Matica srpska kao izdavač, jeste izrada Rečnika u štampanoj formi. Predviđeno je da sadrži oko 130.000 odrednica. Zbog toga je u program ugrađena funkcionalnost koja omogućava ispis i štampanje Rečnika u konačnoj, prelomljenoj formi, tako da je u svakom trenutku moguće dobiti generisani PDF prikaz (render) spreman, kako za štampu bez dodatne pripreme, tako i za prikaz na vebu, budući da je taj proces u potpunosti automatizovan. Periodično automatsko generisanje rečničkog sadržaja omogućava svakodnevni uvid u njegovu celinu u PDF formatu (Sl. 4). Elementi rečničke makrostrukture mogu da se unesu tokom rada na Rečniku, a svakako i na samom kraju izrade Rečnika i oni su sastavni deo konačnog izveštaja u PDF formatu. Unose se kao formatirani tekst ili se automatski generišu na osnovu sadržaja odrednica koje su ušle u Rečnik.

4. Baza podataka, struktura korpusa i način prikupljanja građe

Elektronski jezički korpusi neophodni su i nezaobilazni leksikografski alati jer obezbeđuju primarne izvore podataka za formiranje rečnika (RANDEL 2002: 140). Imajući to u vidu, u okviru programa za izradu Rečnika razvijena je posebna platforma za rad sa korpusom, namenjena pretragama, generisanju raznih izveštaja, ekscerpiranju primera za Rečnik i njihovom unošenju u leksikografski članak. Informacioni sistem za podršku izrade novog Rečnika, koji obuhvata i softver za rukovanje korpusom, sačinjen je kao veb aplikacija. Komponente sistema i alati za razvoj dostupni su u obliku otvorenog koda na internetu. Baza podataka kojom rukuje softver za korpus ima sledeće celine:

1. Reči čiji su gramatički oblici povezani sa njihovim osnovim oblikom po kom ih je moguće pretražiti. Trenutno se u toj bazi nalazi oko 140.000 reči (lema). Ovi podaci prikupljeni su jednim delom uz pomoć studenata Odseka za srpski jezik i lingvistiku sa novosadskog Univerziteta, koji su u bazu unosili oblike promenljivih reči, a deo podataka je automatski preuzet iz baze podataka *wikimorph-sr.* Budući da su nam ljudski i finansijski resursi veoma skromni, nismo imali mogućnosti da se bavimo anotacijom korpusa i ovo nam je bio optimalan način da, u

relativno kratkom vremenskom periodu, u programu svedemo promenljive reči na jedan oblik za pretragu i na taj način dobijemo funkcionalan korpus. Naravno, svesni smo da to nije idealno, ali dosadašnji rad na Rečniku i pripremljen Nacrt ogledne sveske, pokazali su da je dovoljno dobro za uspešnu pretragu korpusa kao i za izbor i unos odgovarajućih konkordanci.



Sl. 4. Izgled stranice i leksikografskog članka u renderu

2. Bibliografski opisi izvora koji čine korpus. Izvori su razvrstani po potkorpusima, za sada su to književni, novinski, administrativni, razgovorni i naučni potkorpus. Korpusnu građu prikupljali smo na razne načine budući da se radi o podacima dostupnim u različitim formatima kao što su pisani digitalizovani i nedigitalizovani tekstovi, audio i video zapisi. Jedan broj izvora dobili smo već pripremljene u digitalnoj formi, a druge (novinske članke, audio i video zapise) preuzimali smo sa interneta. One izvore koji su bili dostupni samo u štampanoj formi, sami smo digitalizovali tako što smo ih skenirali i OCR-ovali. Razgovorni korpus formirali smo automatskom transkripcijom izabranih razgovora u audio formatu pomoću softverskih alata kompanije OpenAI (Whisper).

3. Sadržaji izvora koji se čuvaju kao nestrukturiran tekst dostupan za pretraživanje i analizu.

4. Podaci koji predstavljaju statistiku pojavljivanja pojedinih reči u korpusu i odluke donete u vezi sa uključivanjem datih reči u Rečnik.

„Vrlo je važno dobro strukturirati tekstove u korpusu i unapred odrediti proporcionalno učešće različitih funkcionalnih stilova tako da on može, što vernije, da predstavi jedan prirodan jezik“ (VUJOVIĆ 2011: 40). S obzirom na to, u spisak dela za korpus ušla su dela uglednih i nagrađivanih srpskih pisaca sa prostora na kojima se srpski jezik govori, i to je oko 1200 monografija koje uključuju domaća i prevedena dela, zatim dnevne novine i časopisi, udžbenici za osnovnu i srednju školu kao i fakultetski udžbenici, razni priručnici, naučna i stručna literatura, drugi rečnici i dostupni pisani izvori koji predstavljaju transkripte govornog jezika a zatim i transkripti govornog jezika koje smo formirali specijalno za naš korpus. Budući da nam računarska tehnologija omogućava da u korpus unosimo celokupna dela, iskoristivost građe je maksimalna. Naravno, „veličina korpusa i njegova struktura zavisi od mogućnosti, odnosno od raspoloživih ljudskih i tehničkih potencijala, od vremena za koje treba da se sakupe, označe i analiziraju tekstovi“ (VUJOVIĆ 2011: 408).

Korpus je strukturiran po potkorpusima. Sakupljanje korpusa nije završeno jer mnoga književna dela nisu dostupna u digitalnoj formi. Proces njihove digitalizacije dugotrajan je i još uvek se radi na tim poslovima. Plan je da korpus *Srpko* obuhvati oko 4500 pisanih izvora od čega su 1454 književni izvori; 91 su rečnici; 100 su časopisi, nedeljnici i zbornici; 30 su dnevne novine iz kojih je preuzet 163.451 članak; 1900 su lektire za osnovnu školu; 725 su srednjoškolske lektire; 200 izvora su udžbenici i 153 izvora su razni administrativni tekstovi. Do sada je u korpus ušla i građa iz 5690 izvora razgovornog jezika. Tu građu smo formirali, u većini, skidanjem emisija sa *YouTube* kanala i njihovim automatskim transkribovanjem, a jednim delom i ručnim transkribovanjem domaćih filmova i snimljenih privatnih razgovora. Na slici broj 7 nalazi se trenutna statistika korpusa koja obuhvata broj izvora, broj reči i procentualnu zastupljenost u korpusu. Ti brojevi će se menjati jer se korpus još uvek dopunjava.

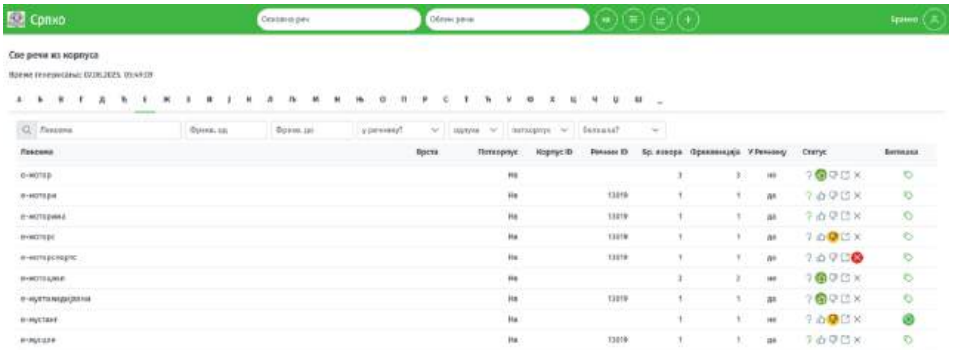
Tabela 1. Pregled statistike

Potkorpus	**Broj izvora**	**Broj reči**	**Udeo**
Književni	690	42.415.951	25%
Novinski	163.451	95.922.558	56%
Razgovorni	5.690	26.443.513	16%
Naučni	10.070	4.163.683	2%
Administrativni	153	1.065.547	1%
Ukupno	**180.054**	**170.011.252**	**100%**

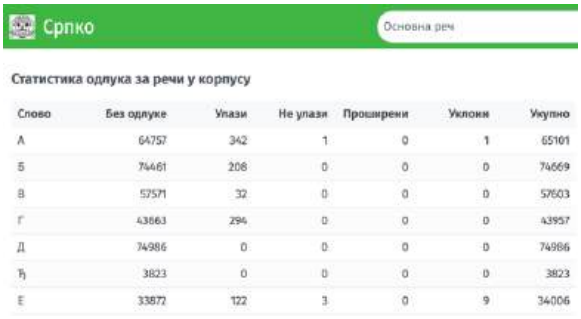
5. Izveštaji i odluke o izboru reči za Rečnik

U korpus su ugrađene različite funkcionalne pretrage i zahtevi za filtriranje rezultata kao i funkcija dobijanja različitih izveštaja na osnovu zadatih filtera (Sl. 8). To mogu biti već pripremljeni izveštaji dobijeni na osnovu unapred određenih parametara ili dinamički izveštaji koji mogu da se kreiraju po potrebi na osnovu zadatih filtera. Filter može biti, na primer, frekvencija reči, potkorpus ili odluke u vezi sa unosom zadate reči u Rečnik.

Odluke o tome da li će reč ući u Rečnik ili u da li ćemo je uopšte zadržati u nekom proširenom spisku i eventualno razmatrati kasnije ili potpuno ignorisati moguće je markirati klikom na ikonice ? 👍 👎 ↗ ✕ (Sl. 5), a detaljan izveštaj o tim odlukama dobija se klikom na poveznicu „Statistika odluka“ (Sl. 6).



Sl. 5. Filtriranje rezultata i izveštaji

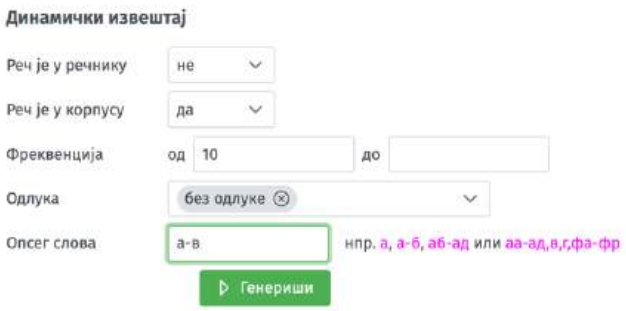


Слово	Без одлуке	Улази	Не улази	Проширени	Уклони	Укупно
А	64757	342	1	0	1	65101
Б	74461	208	0	0	0	74669
В	57571	32	0	0	0	57603
Г	43863	294	0	0	0	43957
Д	74986	0	0	0	0	74986
Ђ	3823	0	0	0	0	3823
Е	33872	122	3	0	9	34006

Sl. 6. Statistike odluka

Na stranici sa izveštajima nalaze se dve grupe izveštaja, to su statički i dinamički izveštaji. Statički izveštaji su fiksirani prema unapred predviđenim parametrima. Na primer, jedan od njih predstavlja spisak svih

reči koje se u korpusu pojavljuju više od 10 puta. U tom izveštaju tokeni kod kojih je bilo moguće identifikovati pripadnost konkretnoj lemi objedinjeni su u jednu stavku u spisku reči, a ostali tokeni se u spisku nalaze u svojim originalnim oblicima u kojima se nalaze u korpusnoj građi. Dinamičke izveštaje možemo sami da formiramo, u zavisnosti od toga šta želimo da analiziramo. Na primer, ukoliko želimo da vidimo izveštaj svih reči koje počinju na određeno slovo ili kombinaciju slova ili ukoliko hoćemo da kombinujemo neke parametre, npr. početno slovo sa određenom frekvencijom reči i sl. poslužićemo se dinamičkim izveštajem (Sl. 7 i 8).



Sl. 7. Forma za zadavanje dinamičkog izveštaja

Динамички извештај

14.11.2025. 10:28

у речнику: не

у корпусу: да

фреквенција: 10 до _

одлуке: без одлуке

слова: а-в

A

Реч	f	Реч	f	Реч	f	Реч	f
аб	1347	аврам	1018	адел	151	адрија	130
абазовић	386	аврамовић	1019	адела	40	адријан	147
абара	10	аврамовски	64	аделаида	112	адријана	191
абас	97	автобус	23	аделина	14	аер	20
абасидски	37	агалук	13	адем	147	аеробан	177
абација	31	агама	24	адеми	22	аеросолан	15
абација	36	агамемнон	193	адемовић	13	аерофлот	76
абделазиз	11	аган	10	ден	64	ажуриран	428
абдић	154	агапа	66	денски	21	ажурирање	599
абдул	139	агата	786	адео	13	азалеја	20
абдулазиз	31	агбаба	16	адет	46	азап	27
абдулах	233	агда	11	адиђар	29	азија	31
абдулхамид	28	агејев	19	адија	38	азелаичан	10
абдурахман	40	агена	17	адиктиван	29	азем	74
абед	10	агентивни	19	адикција	38	аземи	13
абер	345	агентица	23	адил	51	азер	19
абердар	41	агер	85	адина	16	азербејџан	776
абид	10	агим	135	адис	92	азербејџанац	141
аболирани	80	агитовање	48	адица	16	азербејџански	299
аболирање	15	агић	17	адиција	16	азиз	97
аборт	16	агнеза	30	адлер	127	азировић	20
абортиран	48	агнета	10	адлигат	30	азис	10

Sl. 8. Deo rezultata dinamičkog izveštaja

6. Pretraga korpusa

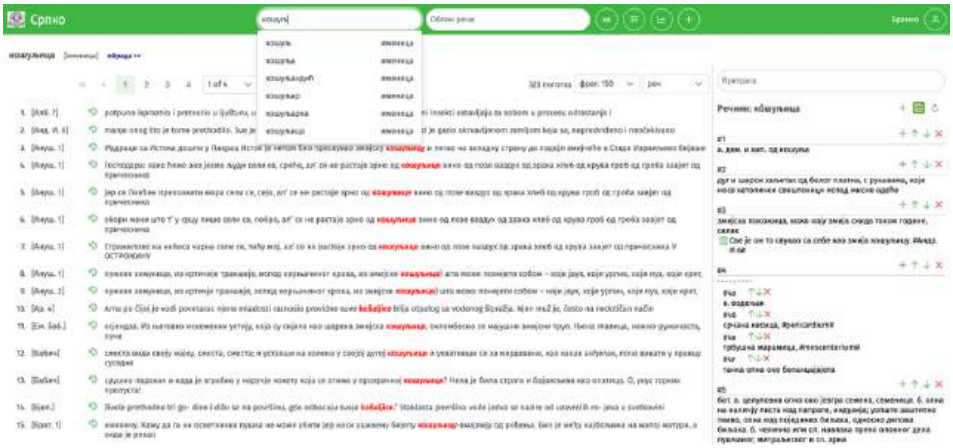
Korpus je moguće pretraživati po dva kriterijuma, po osnovnom obliku reči i po zadatom obliku reči. Rezultat pretrage po osnovnom obliku obuhvata izvore u kojima se data reč pojavljuje u bilo kom svom obliku (Sl. 9), naravno, ukoliko je u pitanju promenljiva reč. U rezultatu ove pretrage pojaviće se samo reči čiji su oblici pojavljivanja uneti u bazu podataka i povezani sa osnovnim oblikom. Njen rezultat je prikaz svih oblika pojavljivanja tražene reči. Rezultat pretrage po zadatom obliku reči omogućava pretragu određenog morfološkog oblika reči (Sl. 10) ili više reči, odnosno, konteksta (Sl. 11). Rezultat pretrage po obliku prikazuje samo taj jedan zadat oblik reči ili više reči u izvorima u kojima se nalaze. Pomoću pretrage po obliku možemo pretraživati i širi rečenični kontekst.

Sl. 9. Pretraga po osnovnom obliku

Sl. 10. Pretraga po zadatom obliku

Sl. 11. Pretraga po zadatom obliku – fraza

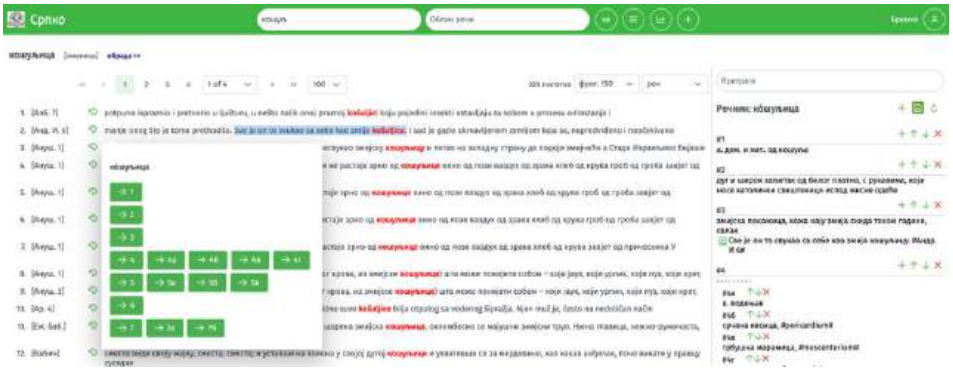
Unosom reči u polje za pretragu dobijamo je ispisanu i markiranu crvenom bojom u okviru konkordance, sa navedenim izvorom u kom se nalazi u korpusu. Širinu konteksta možemo sami da definišemo. Na desnoj strani osnovnog ekrana za rad sa korpusom pojaviće se, u svedenoj formi, ono što smo već uneli prilikom obrade zadate reči, a to znači njena značenja, podznačenja i konkordance (Sl. 12).



Sl. 12. Pregled i pretraga reči u korpusu

7. Ekscerpcija primera

Za unos novih konkordanci u rečnički članak potrebno je markirati željeni kontekst (Sl.13) i klikom na njega otvoriće se sva značenja i podznačenja unete reči.



Sl. 13. Izbor konteksta

Izborom određenog značenja unosimo izabrani kontekst u tekst članka. Pošto se u korpusu nalaze originalni tekstovi u kojima ima grešaka ili su pisani latinicom, klikom na polje pored konkordance otvara se prozor u kome možemo da intervenišemo na tekstu, bilo da ga ispravljamo ili preslovljavamo iz latinice u ćirilicu (Sl. 14).

Sl. 14. Izmena konkordance

Kada se završi unos konkordanci, sve može da se proveri na stranici za unos odrednice, klikom na dugme „pregled“ (Sl. 15). Izvor, zajedno sa konkordancom, biva automatski upisan u odgovarajuće polje.

Sl. 15. Pregled odrednice

8. Softverska arhitektura informacionog sistema

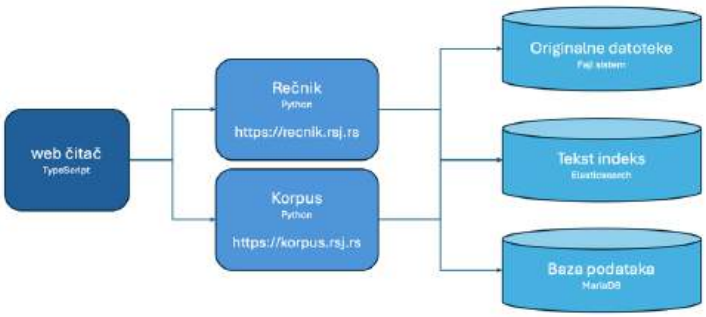
Softver korpusa i rečnika rukuje podacima u više različitih skladišta. Strukturirani podaci: rečnički članci, reči sa svim gramatičkim oblicima, bibliografski opisi izvora, statistike za izveštaje i odluke o izboru reči za Rečnik, podaci o korisnicima i pravima pristupa, čuvaju se u relacionoj bazi podataka *MariaDB*. Nestrukturirani podaci: dokumenti koji čine korpus i generisani rečnici, čuvaju se neposredno u fajl-sistemu servera na kome se izvršava program. Efikasno pretraživanje tekstualnih podataka oslanja se na poseban servis, *Elasticsearch*, koji se oslanja na fajl-sistem i koncept invertovanih datoteka.

Softver za korpus i rečnik razvijen je kroz dve odvojene aplikacije koje komuniciraju međusobno. Na taj način olakšani su nezavisan razvoj i eksploatacija ta dva osnovna dela sistema. Pored toga, pravima pristupa za korisnike korpusa i rečnika upravlja se nezavisno, što omogućava širu primenu korpusa u odnosu na rečnik, čak i izvan osnovne namene za koju je razvijen. Pristup korpusu imaju samo saradnici koji rade na izradi Rečnika, ali želja nam je da ga, ukoliko to bude moguće, kada sakupljanje građe bude završeno, učinimo i javno dostupnim.

Korpus i rečnik su implementirani kao veb aplikacije sa svojim serverskim i klijentskim slojevima. Serverski sloj je razvijen u jeziku *Python* i radnom okviru *Django*, dok je klijentski sloj razvijen u jeziku *TypeScript* i radnom okviru *Angular*. Za pristup sistemu dovoljno je koristiti neki od savremenih veb čitača. Na Sl. 16 prikazane su osnovne komponente softverske arhitekture celokupnog sistema. Aplikacije su dostupne na adresama https://korpus.rsj.rs odnosno https://recnik.rsj.rs, izvorni kod softvera je dostupan u obliku otvorenog koda na adresi https://github.com/unsftn/rsj, a osnovne informacije o projektu nalaze se na adresi https://rsj.rs.

9. Zaključak

Izgradnja elektronskog korpusa *Srpko* kao tehnološke osnovu za izradu novog *Rečnika savremenog srpskog jezika* Matice srpske predstavlja inovativan iskorak u domaćoj leksikografskog praksi. Program za korpus je, zahvaljujući pažljivo strukturiranoj arhitekturi, obezbedio prikupljanje bogate i izbalansirane leksikografske građe i omogućio njenu organizaciju u potkorpuse kao i analizu na osnovu različitih parametara. Specijalna programska rešenja, prilagođena specifičnostima srpskog jezika, omogućila su unos, pretragu (ćiriličnu i latiničnu) i ekscerpciju primera kao i automatsko generisanje finalne verzije Rečnika u različitim formatima.



Slika 16. Komponente softverske arhitekture

Uprkos ograničenim ljudskim i finansijskim resursima kao i kompromisima u vezi sa anotacijom korpusa, postignuti rezultati potvrđuju da je tako formiran korpus omogućio sistematičnu i efikasnu leksikografsku obradu. Pretrage u korpusu jednostavne su i brze, unos konkordanci i njihovih izvora u rečnički članak je automatski, generisanje izveštaja je fleksibilno, što sve skupa ubrzava i olakšava proces izrade Rečnika.

Čitav sistem otvoren je za dalji razvoj i unapređenje funkcionalnosti kao i za širenje obima kako građe tako i korisnika. Budući razvoj projekta usredsrediće se na automatsku anotaciju teksta, sistematsko bogaćenje novim izvorima, implementaciju složenijih pretraga i proširenje skupa korisnika. Ovakvim projektima unapređuje se pozicija srpskog jezika u savremenom digitalnom okruženju.

Literatura

[1] Gaus 2014: Gouws, H. R. (2014). Article structures: Moving from printed to e-dictionaries. *Lexikos, 24*(1). Preuzeto 20. jula 2025, sa https://lexikos.journals.ac.za/pub/issue/ view/80

[2] Randel 2002: Rundell, M. (2002). Good old-fashioned lexicography: Human judgment and the limits of automation. In *Lexicography and natural language processing: A Festschrift in honour of BTS Atkins*. EURALEX. Preuzeto 20. jula 2025, sa https://www.euralex.org /elx_proceedings/Lexicography%20and%20Natural%20Language%20Processing/Michael%20Rundell%20-%20Good%20Old-fashioned%20Lexicography%20Human%20 Judgment%20and%20the%20Limits%20of%20Automation.pdf

[3] Vujović 2011: Vujović, D. (2011). Internet kao elektronski korpus. *Naučni sastanak slavista u Vukove dane*, 40(1), 405–411. Beograd.

[4] Vujović, Milosavljević 2022: Вујовић, Д., Милосављевић, Б. (2022). О програму за унос и обраду речи новог вишетомника Матице српске из лексикографског угла. У *Актуелна питања лексикологије и лексикографије српскога језика* (Радови са научног скупа „Актуелна питања лексикологије и лексикографије српског језика“, одржаног у Андрићграду 9. и 10. октобра 2021), 133–149. Андрићград: Андрићев институт.

Rečnici

[1] RMS 1969–1976: *Речник српскохрватског књижевног језика.* књ. 1-3, Нови Сад – Загреб: Матица српка – Матица хрватска, књ. 4-6, Нови Сад: Матица српска.

[2] RSANU 1959 -: *Речник српскохрватског књижевног и народног језика.* Књ. 1-21. Београд: САНУ – Институт за српски језик.

[3] RSJ 2007: *Речник српскога језика*, Нови Сад: Матица српска.

The *Srpko* Corpus as a Technological Foundation for Compiling Matica Srpska's *Rečnik savremenog srpskog jezika*

Dušanka Vujović, Branko Milosavljević

Summary

The digitization of linguistic resources provides fast and convenient access to language data, thereby facilitating linguistic research and the analysis of language phenomena. For modern lexicographic projects, it is of key importance, as it ensures a rich supply of carefully selected lexical material, efficient retrieval and extraction of examples, and their seamless integration into dictionary entries. The *Srpko* electronic language corpus forms the foundation for compiling the new *Rečnik savremenog srpskog jezika* by Matica srpska. It allows for quick and precise searching of textual data and their direct incorporation into dictionary content.

This paper outlines the guiding principles, as well as the conceptual and technical solutions, applied in organizing the corpus. It describes its structure, methods of data collection, and database design for each component, given the diversity of available formats, ranging from digitized and non-digitized written texts to audio and video recordings. To enable faster and more efficient searches, as well as quality control, the software integrates a range of search functions and filtering options. It also includes a reporting tool for generating customized outputs based on selected filters. In this way, the corpus provides a systematic framework for organizing and presenting the lexical database required for dictionary compilation.

Despite limited human and financial resources, as well as necessary compromises in corpus annotation, the achieved results demonstrate that the corpus, as designed, enables systematic and efficient lexicographic processing. Corpus searches are straightforward and fast; the insertion of concordances and their sources into dictionary entries is automated; and the generation of reports is flexible, all of which significantly accelerate and facilitate the dictionary compilation process. The entire system remains open to further functional development and to expanding both the scope of the material and the range of its users. Projects of this kind strengthen the position of the Serbian language within the contemporary digital landscape.

Keywords: Serbian language, Serbian language corpus, digitization, dictionary.