Korpusi za učenje srpskog jezika kao stranog u eri veštačke inteligencije

Naučni rad DOI: 10.18485/judig.2025.1.ch16

Olja Perišić[1] *0000-0002-1219-862X*

Apstrakt

U radu ćemo razmatrati izazove sa kojima se nastava srpskog jezika kao stranog zasnovana na korpusima suočava u kontekstu sve ubrzanijeg razvoja veštačke inteligencije. Dok se u oblasti Data Driven Learning-a sve više preispituje budućnost korpusnih alata (Crosthwaite, Baisa 2023, Flowerdew 2024), srpski jezik ostaje u povoju, barem što se tiče praktične upotrebe korpusa u nastavi. Naime, uočava se značajan jaz između istraživačkog delovanja stručnjaka za informatiku, posebno onih okupljenih oko udruženja JeRTeh, koji prate svetske trendove i aktivno razvijaju korpusne alate, i predavača koji bi te alate mogli da koriste u nastavi stranih jezika. Pored neophodnog afiniteta za ovu vrstu nastave, jedan od uzroka tog raskoraka je početno ulaganje u savladavanje veštine pretrage korpusima, kao i usvajanje teorijskog i metodološkog pristupa neophodnog za rad u učionici. U isto vreme iskustvo je pokazalo da studenti srpskog kao stranog relativno brzo savladavaju veštinu korišćenja korpusa i da su uglavnom kreativni u samostalnim istraživanjima, te time stiču motivaciju i sklonost ka istraživačkom radu i učenju jezika (Perišić 2023). Cilj rada je da se identifikuju i analiziraju ključni izazovi sa kojima se nastava zasnovana na korpusima u eri veštačke inteligencije suočava, sa posebnim osvrtom na generativne jezičke modele, kao i da se razmotri mogućnost njihove integracije u didaktiku srpskog jezika kao stranog.

Ključne reči: korpusi, Data Driven Learning, srpski kao strani, JeRTeh, veštačka inteligencija, generativni modeli

[1] Univerzitet u Torinu, Departman za strane jezike i književnost i moderne kulture, olja.perisic@unito.it

1. Uvod

Tokom tri i po decenije svoga postojanja pristup DDL (Data Driven Learning)[2] koji podrazumeva korišćenje korpusnih alata na svim nivoima usvajanja stranih jezika[3] prošao je kroz različite interdisciplinarne razvojne faze na različitim poljima jezičkih istraživanja. Ubrzani razvoj digitalnih tehnologija i dostupnost kompjutera i mobilne telefonije sve većem broju korisnika već od početka novog milenijuma naveli su istraživače da preispituju efikasnost korpusa u odnosu na web alate i mogućnost njihovog integrisanja u didaktiku stranih jezika (Boulton 2012, Gatto 2014, Perišić 2025).

Od 2022. g. kada je ChatGPT ušao na velika vrata u naše živote i napravio revoluciju u mnogim profesionalnim sferama, uključujući obrazovanje, naučnici koji se bave DDL pristupom našli su se pred novim izazovom i potrebom da ispitaju efikasnost i svrsishodnost korišćenja korpusa u eri generativnih modela zasnovanih na veštačkoj inteligenciji (Crosthwaite, Baisa 2023, Flowerdew 2024).

U domenu nastave srpskog jezika kao stranog, primećuje se značajan raskorak između istraživanja stručnjaka iz oblasti informacionih tehnologija, naročito onih okupljenih oko udruženja JeRTeh[4], koji aktivno prate međunarodne trendove i razvijaju korpusne alate, i predavača koji bi te alate mogli primeniti u radu sa učenicima/studentima. Pored neophodne sklonosti ka ovom pristupu, jedan od glavnih razloga tog raskoraka leži u početnom trudu potrebnom da se ovlada veštinom korpusnog pretraživanja i da se usvoje odgovarajuće teorijske i metodološke osnove koje bi omogućile njihovu uspešnu upotrebu u nastavi. Od prvog istraživanja koje se bavilo potencijalom korpusa u nastavi srpskog kao stranog jezika (Miličević, Samardžić 2010), prošla je gotovo jedna decenija do pojave novih istraživanja u oblasti hrvatskog (Posavec, 2018; Mikelić Preradović, Posavec, Unić, 2019) i srpskog jezika (Perišić Arsić, 2018), koja, nažalost, još uvek ne pokazuju značajniju tendenciju rasta[5]. S obzirom na sve prisutniju ulogu veštačke inteligencije, posebno generativnih modela koji tektonski pomeraju način na koji percipiramo didaktiku i usvajanje jezika, cilj ovog rada je da preispita potencijal generativnih modela u didaktici srpskog jezika kao stranog kroz

[2] Osnivač ovog pristupa i onaj kome dugujemo njegov naziv bio je Tim Johns koji je devedesetih godina prošlog veka izveo prve eksperimente u učionici uz pomoć štampanih konkordanci sa studentima engleskog jezika kao stranog na Univerzitetu u Birmingenu (Johns, 1990).

[3] U početku je fokus primene bio na akademskom nivou sa studentima srednjeg/višeg nivoa jezika, ali se vremenom polje njegove primene proširilo na sve nivoe jezičkog obrazovanja, ne samo na akademskom nivou, već i u srednjim i osnovnim školama (Braun 2007, Boulton 2020, Crosthwaite 2020, Hirata 2025).

[4] JeRTeh - Language Resources and Technologies Society, https://jerteh.rs/.

[5] Autorka rada se na svojim doktorskim studijama bavila ovom temom, koja je ostala u fokusu njenih istraživanja.

njihovo poređenje sa korpusnom metodologijom čija uloga u eri veštačke inteligencije mora biti preispitana, ali ne zanemarena.

2. Korpusi u eri veštačke inteligencije

Ostavljajući po strani značajna etička pitanja i pitanja regulacije upotrebe generativnih modela na različitim nivoima formalnog obrazovanja, jednostavnost ove vrste pretraga i činjenica da omogućavaju instant odgovore, kao i informativni i korektivni *feedback*, mogu biti odlučujući faktori kojima se vodi prosečan korisnik–student u odabiru jezičkih alata za učenje stranih jezika (Stewart 2025). Iz tog razloga potrebno je jasno preispitati prednosti i nedostatke ovih instrumenata u cilju njihovog boljeg razumevanja i eventualne implementacije u nastavu koja bi za osnovu imala kritičko preispitivanje, ali i eventualnu integraciju AI generativnih modela u već postojeća *corpus-based* istraživanja.

Gotovo svi autori koji su poredili generativne modele i korpuse naglasili su prednost korpusa u pogledu dostupnosti metapodataka (npr. izvora tekstova koji ulaze u njihov sastav[6]) (Crosthwaite, Baisa 2023) i autentičnosti jezičkih primera (Lin, 2023). U tom kontekstu postavlja se pitanje da li su automatski generisani tekstovi autentični s obzirom na to da nije dovoljno jasno na kojim su izvorima obučeni dok se generisanje teksta vrši na osnovu statističkih podataka i „embedding-a". Pored toga u slučaju tzv. halucinacija, tj. generisanja netačnih ili izmišljenih informacija, pitanje je da li će nematernji govornici biti u stanju da samostalno prepoznaju neautentične ili lingvistički neodgovarajuće i neprikladne sadržaje. Još jedno pitanje tiče se mogućnosti replikacije kao osnovnog načela naučnog istraživanja, s obzirom na to da generativni modeli na isto pitanje mogu pružiti različite ili različito formulisane odgovore. Ukoliko pak poredimo generativne modele sa korpusima sa metodološke tačke gledišta, kriptičnost prvih po pitanju izvora i autentičnosti podataka čini *hands-on*[7] pristup problematičnim, te je utisak da je u njihovoj trenutnoj razvojnoj fazi posredovanje nastavnika osnovni preduslov uvođenja u nastavu. Uloga predavača ne svodi se samo na tehničke aspekte nastave (npr. formulisanje promptova), već i na ukazivanje na prednosti, ali i na nedostatke ovih modela na način koji podstiče kritičko razmišljanje i promišljenu i svrsishodnu evaluaciju rezultata.

[6] Autor Kizito Tekwa (2024: 103) precizira da se radi o sledećim izvorima: „ChatGPT, for instance, uses corpus data from books, social media, Wikipedia, news articles, human speech and audio recordings, academic research papers, websites, forumes, and code repositories"

[7] Pristup *hands-on* podrazumeva direktno korišćenje korpusa ili u ovom slučaju generativnih modela, dok se pod pristupom *hands-off* podrazumeva štampanje output informacija poput konkordanci i njihova analiza i interpretacija (Boulton 2012).

U slučaju, pak, korpusne nastave učenik ne dobija gotove odgovore na istraživačka pitanja, već induktivnom metodom otkriva pravila funkcionisanja jezika ili deduktivnom metodom proverava postojeća pravila, u oba slučaja uz pomoć nastavnika, da bi vremenom sve više težio samostalnosti i autonomnom usvajanju jezika.

Pojedini autori predlažu poređenje radova pisanih od strane polaznika i automatski generisanih tekstova u cilju pronalaženja učestalih leksičkih formi koje bi ukazale na razlike u pismenoj produkciji čovek vs. mašina. Na osnovu kvalitativne analize tekstova studenti bi tokom nastave razvijali kritičko razmišljanje i imali uvid o potencijalu sopstvenih istraživanja koja mogu biti potpomognuta generativnim modelima, ali ne potpuno zamenjena njima (Zhang, Crosthwaite 2025). U tom smislu Brezina (2025) ukazuje na rizik površnog korišćenja AI instrumenata, jer imajući u vidu da se komunikacija odvija u različitim okolnostima i sa različitim komunikativnim ciljevima, podaci sami za sebe nisu dovoljni. U tom smislu autor tvrdi da je ljudski faktor i dalje nezamenjiv kada se radi o validnoj interpretaciji podataka i njihovom prilagođavanju preciznim kulturnim kontekstima i ciljevima.

Oba pristupa o kojima je reč u ovom radu (*AI-based* i *corpus-based*) u širem smislu deo su didaktičkog pristupa poznatog kao „self-directed learnig - SDL“ (Zhang, Zou, 2024) koji podrazumeva korišćenje alata u odnosu na kognitivni potencijal različitih profila polaznika. Naime, razlike među polaznicima nisu samo na nivou poznavanja jezika, već i u ličnim interesovanjima, ali i didaktičkim metodama koje su usvojili u prethodnom obrazovanju i koje nose sa sobom u vidu obrazovnog nasleđa, bez obzira na njihovu eventualnu efikasnost. S obzirom na to da su tradicionalni alati i kurikulumi standardizovani, ova vrsta personalizovanog pristupa obično je teško izvodljiva u tradicionalnoj učionici, što može ostaviti posledice kao što su nezadovoljavajući rezultati, sporo napredovanje, frustracije polaznika.

Za razliku od korpusa, generativni modeli omogućavaju korektivni feedback, mogu simulirati konverzaciju sa maternjim govornikom, drugim rečima mogu pomoći u razvijanju sposobnosti pisanja, čitanja, slušanja i sporazumevanja. Da bi ova vrsta interakcije bila efikasna potrebno je da korisnici imaju zadovoljavajući nivo AI pismenosti, odnosno da znaju na koji način da dođu do što boljih rezultata i da kritički procene efikasnost ovih alata koje ne treba prihvatati bezrezervno i bespogovorno. Interaktivno učenje koje nude ovi četbotovi nadovezuje se na konstruktivistički pristup u učenju jezika, jer omogućava da se kognitivni stil i pristup učenju svake individue valorizuje (Harel, Papert 1990). Ova vrsta konstruktivističkog pristupa učenju stranih jezika karakteristika je i

DDL-a koji podrazumeva aktivnosti „largely ustructured and open-ended“ koja zahtevaju aktivno učešće u nastavnom procesu, koji moduliraju na osnovu sopstvenih kognitivnih sposobnosti, interesovanja i metoda učenja (Flowerdew 2025).

3. Korpusi vs. generativni modeli u srpskom kao stranom jeziku

Prethodna istraživanja sprovedena na Univerzitetu u Torinu sa italofonima koji usvajaju srpski jezik kao strani, pokazala su da rad sa korpusima omogućava da se već od početnog nivoa razvija osećaj za jezik, istraživačka radoznalost i motivacija za učenje jezika (Perišić 2023). Zahvaljujući savremenim platformama za pretraživanje korpusa (Sketch Engine i NoSketch Engine[8]) studenti mogu da se obučavaju na različitim nivoima morfosintakse i leksike, objedinjenih prema principu *lexical grammar* (Sinclair 2000), istovremenog posmatranja i analize forme i značenja (*structure and meaning*). Na taj način se fokus pomera sa tradicionalnog učenja isključivo gramatičkih pravila na leksiku u kontekstu. Posebno su važna i inspirativna bila istraživanja koje su studenti radili samostalno, uglavnom na polju polisemije (razdvajanja bliskih sinonima), leksičkih praznina, kao i usvajanja glagolskog vida na višim nivoima, i sa derivacione i sa semantičke tačke gledišta. S obzirom na policentričnost nekadašnjeg srpskohrvatskog jezika, studenti su pokazali veliko interesovanje za usvajanje njegovih varijanti kroz širok spektar lingvističkih i sociolingvističkih pitanja (Perišić 2023). Ova vrsta istraživanja omogućena je pretragama tipološki srodnih korpusa poput onih zasnovanih na dokumentima preuzetim sa interneta za srpsku, hrvatsku i bosansku varijantu: SrWaC, HrWaC, BsWaC[9] (Baroni 2009, Kilgarriff et al 2010a).

Test sa kontrolnom grupom studenata koji nikada nisu koristili jezičke korpuse pokazao je da web pretrage mogu da zamene neke vrste korpusnih upita i na taj način dokažu da „googleology“ nije uvek „bad science“ (Kilgarriff 2007), posebno na srednjem i višem nivou učenja jezika (Perišić 2025). Na ovim nivoima značaj interneta dodatno se proširuje u eri veštačke inteligencije, jer je svest učenika o jezičkim

[8] Sketch Engine je trenutno najveća platforma za pretraživanje korpusa koja sadrži korpuse za preko sto različitih jezika i jezičkih varijanti i pri tom omogućava pretragu i kreiranje korpusa putem prethodne registracije. Ova platforma je korišćena za pretragu korpusa SrWaC 1.2. (Ljubešić, Klubička, 2014)
NoSketch Engine je *opensource* verzija SE-a preko koje se mogu pretraživati korpusi za srpski jezik među kojima *SerbItaCor3* (Moderc *et al*, 2023) koji je korišćen za potrebe ovoga rada, a koji sadrži paralelizovane srpske i italijanske književne tekstove.

[9] S obzirom na različit obim ovih korpusa u kvantitativnim istraživanjima studenti se obučavaju da koriste relativne frekvencije, a ne apsolutne.

strukturama autentičnog jezika već razvijena, dok je, kako je isto istraživanje pokazalo, na početnim nivoima potrebno opreznije pristupiti ovoj vrsti pretraga.

Da bismo proverili potencijal generativnih modela i uporedili ga sa korpusnim istraživanjima poći ćemo od tri već sprovedena istraživanja na polju polisemije (*kuća-dom*), leksičkih praznina (*babine*) i pozitivnih i negativnih konotacija reči tzv. semantičke prozodije (reč *žena* u kontekstu).

Praksa u učionici sa italofonima koji usvajaju srpski jezik pokazala je da postojanje dva srpska prevodna ekvivalenta italijanske lekseme *casa* (kuća i dom) dovodi do nedoumica u njihovoj upotrebi, pri čemu se često pravi paralela sa engleskim rečima *house* (građevina) i *home* (emotivni, lični prostor). Istraživanje na Univerzitetu u Torinu pokazalo je da su studenti koji su prošli obuku korišćenja korpusa u stanju da samostalno utvrde polje primene ove dve lekseme na osnovu kolokacija i tako dođu do zaključaka o njihovom značenju i adekvatnom prevodu na italijanski (Perišić 2020).

U nastojanju da ovo istraživanje proverimo preko modela ChaGPT možemo formulisati sledeći prompt:

Koji su najfrekventniji pridevi koji formiraju kolokacije reči „kuća“? Nakon toga ponovi rezultate za reč „dom“!

Radi lakšeg poređenja, rezultati istraživanja preko korpusa *SrWaC 1.2.* i *SerbItaCor3_sr* biće predstavljeni u tabeli zajedno sa rezultatima koje smo dobili preko generativnog modela ChatGPT.

Na prvom mestu uočava se da generativni modeli nisu u stanju da ponude tačne frekvencije reči s obzirom na odsustvo metapodataka vezanih za broj i vrstu tekstova na osnovu kojih je kompjuter generisao rezultate. Iz tog razloga ne možemo biti sigurni da li se zaista radi o najfrekventnijim ili najtipičnijim kolokatima. Osim toga, postavljanje istog prompta u različitim momentima daje barem delimično različite rezultate. Što se tiče sadržaja rezultata, uočava se da oba korpusna istraživanja ukazuju na značajnu komponentu institucionalne upotrebe lekseme *dom* (studentski, predstavnički, planinarski, parohijski, starački, Ruski, popravni), koje nam generativni model nudi samo u jednom slučaju (dom za stare). I u slučaju lekseme *kuća* korpusna istraživanja nude mnogobrojne rezultate ovoga tipa (izdavačka, gradska, medijska, Bela, robna, osiguravajuća, sigurna), potpuno odsutnih kod ChatGPT pretrage.

Tab. 1. Najfrekventniji pridevi lekseme kuća i dom (korpusi vs. ChatGPT).

	SrWac 1.2.[10]	SerbItaCor3_sr[11]	ChatGPT
kuća	izdavačka (12.75), gradska (6.87), medijska (6.37), Bela (6.01), robna (4.53), osiguravajuća (3.98), sigurna[12] (3.94), velika (3.59), porodična (3.53), stara (2.97)	izdavačka (9.34)[13], cela (7.56), javna (7.23), velika (5.61), seoska (5.28), stara (5.28), nova (4.47), robna (4.23), roditeljska (3.49), lepa (3.33)	velika, mala, porodična, stara, nova, lepa, seoska, napuštena, skromna, drvena
dom	studentski (3.27), predstavnički (3.04), nov (2.95), planinarski (2.13), parohijski (1.89), donji (1.55), starački (1.14), vaš (1.13), Ruski (1.03), sav (1.02)	planinarski (3.09), starački (2.93), roditeljski (1.54), nov (0.98), popravni (0.81), studentski (0.81), parohijski (0.65), star (0.57), sopstveni (0.49), pravi (0.49)	porodični, topli, miran, roditeljski, siguran, sopstveni, ljubavni, stari (u značenju „dom za stare“), zajednički, pravi, mirni

Sa stanovišta maternjeg govornika nakon prvog dela može se zaključiti da je korpusna pretraga preciznija i verodostojnija, ne samo po pitanju frekvencija, već i po pitanju razdvajanja značenja dve lekseme zahvaljujući kolokatima. Ova vrsta informacija neophodna je stranom polazniku koji je sklon pravljenju paralela sa sopstvenim maternjim jezikom ili drugim jezicima koje poznaje, čak i u slučajevima u kojima zbog jezičke asimetrije to nije moguće.

U radu u kojem smo se bavili jezičkim prazninama (Perišić Arsić 2020) jedna od analiziranih leksema bila je *babine,* izabrana s obzirom na to da se radi o kulturološkoj praznini, u ovom slučaju običaju, koji ne postoji u italijanskom jezičkom i kulturnom prostoru. Sledi pretraga preko modela ChatGPT.

Prompt: Kako objasniti reč "babine/babinje" italijanskom polazniku?

[10] Pri konsultovanju rezultata korišćena je funkcija Word Sketch.

[11] U nedostatku funkcije Word Sketch na NoSketch Engin-u korišćen je sledeći upit CQL: [tag="ADJ"][lemma="kuća|dom"]

[12] Veći broj rezultata odnosi se na instituciju pod ovim nazivom koja nudi siguran krov žrtvama nasilja.

[13] S obzirom na različitu veličinu korpusa unesene su relativne frekvencije koje su indikativne, jer funkcija Word Sketch nije potpuno ekvivalentna zadatom upitu. Na ovaj način nismo se ograničili samo na jednu vrstu pretraga, već su prikazane obe mogućnosti koje mogu biti korisne u budućim istraživanjima. Up. Kilgarriff et al (2010b).

Odgovor:

To je običajni naziv za period neposredno nakon porođaja, kada se majka oporavlja i prima posete, a gosti donose poklone za bebu i porodilju. Obično traje oko 40 dana nakon porođaja, i vezan je za verovanja, brigu o zdravlju žene i deteta, kao i izbegavanje fizičkog napora.

Rezultat je veoma precizan i sa stanovišta maternjeg govornika srpskog jezika u potpunosti zadovoljavajući tj. smatramo da bi strani student na osnovu ovog objašnjenja i mogućnosti njegovog prevoda na maternji jezik bio u stanju da u potpunosti shvati značenje ove lekseme.

Pretraga preko ChatGPT modela po pitanju definisanja jezičkih praznina pokazala se veoma uspešnom, posebno imajući u vidu mogućnost generisanja sadržaja na maternjem jeziku polaznika. Prethodno korpusno istraživanje pokazalo je da je broj izvora koji sadrže ovu reč bio nedovoljan. Samim tim ova vrsta pretrage bila je delimično problematična, jer je u nedostatku dovoljnog broja rezultata u korpusu bilo neophodno poznavanje jezika na nivou maternjeg govornika da bi se definisalo značenje na osnovu raspoloživih kolokata.

Bilo bi važno sprovesti kompleksnije istraživanje na temu kulturološki specifične leksike koje bi obuhvatilo mnogo veći broj primera, diferenciranih po preciznim kriterijumima (običaji, narodne igre i plesovi, predmeti i slično) u cilju potvrde ovog preliminarnog rezultata koji ni u kom slučaju ne može poslužiti kao potvrda efikasnosti AI modela, već samo kao podsticaj za dalja istraživanja.

Pretraga reči *žena* u srpskom korpusu tekstova preuzetih sa interneta pokazala je da se ova reč često nalazi u negativnom kontekstu, pretežno uvredljivog ili violentnog sadržaja, od načina na koji se definiše karakter žene do nasilja koje se nad njom vrši. Korpusi su takođe omogućili poređenje ove reči sa njenim muškim ekvivalentom što je potvrdilo da je kontekst reči *muškarac* pozitivnije prirode, tiče se profesionalnih aktivnosti, snage, borbenosti, ali i seksualnosti[14] (Perišić 2023). Akcenat kod žena je na reproduktivnoj ulozi i fizičkom izgledu, ali i na emocijama i profesijama kojima se bavi. Žena kao subjekt rađa, nosi dete, voli, zna, radi, živi i umire, dok se kao objekat u rečenici vezuje za glagole uglavnom negativne konotacije kao što su ubiti, voleti, silovati, tući, udati, zaposliti, varati, mrzeti.

Sledeće tabele omogućiće poređenje rezultata korpusnih istraživanja i ChatGPT pretrage:

[14] Radi se o funkciji Sketch Difference prisutnoj na platformi Sketch Engine koja omogućava poređenje dve lekseme tako što svaku označava različitom bojom (crvenom i zelenom), te se time i na vizuelnom planu može bolje uočiti koji su kolokati zajednički, a koji ekskluzivni za svaku od reči.

Koji su najfrekventniji pridevi koji formiraju kolokacije reči „žena“?

Tab.2. Najfrekventniji pridevi, kolokati reči žena (korpusi vs. ChatGPT)

	SrWac	SerbItaCor3_sr[15]	Chatgpt-based
žena	mlada (6.11), sve (4.9), mnoge (3.48), lepa (3.15), poslovna (2.56), stara (2.35), dve (1.82), uspešna (1.13), prava (1.07), zaposlena (0.88)	mlada (22.43), druga (18.77), lepa (9.83), stara (5.93), prva (4.63), sirota (3.98), jadna (3.90), udata (3.66), mrtva (3.17), dobra (3.17)	lepa, pametna, snažna, dobra, mlada, trudna, starija, udata, nezavisna, uspešna

Drugi prompt je bio: *Koji su najčešći glagoli sa reči žena kao objektom u rečenici?*

Tab.3. Najfrekventniji glagoli koji imaju kao objekat reč žena (korpusi vs. ChaGPT)

	SrWac	SerbItaCor3_sr[16]	Chatgpt-based
žena	imati (1.75), voleti (0.81), videti (0.79), ubiti (0.42), naći (0.35), ostaviti (0.3), pitati (0.25), upoznati (0.25), tražiti (0.24), dovesti (0.23)	imati (3.74), ostaviti (1.87), pogledati (0.98), ubiti (0.81), naći (0.81), posmatrati (0.65), tući (0.57), nemati (0.57), videti (0.57), ugledati (0.49)	voleti, videti, upoznati, nazvati, trebati, želeti, poštovati, brinuti, braniti, hvaliti

Na prvi pogled uočava se da rezultati obe pretrage preko modela ChatGPT imaju pozitivnu konotaciju. Na našu opasku „Primećujem da nema glagola sa negativnom konotacijom“ ChatGPT odgovara: „Mnoge publikacije i mediji (koji pune korpuse poput srWaC-a) izbegavaju direktne izraze nasilja nad ženama (npr. tući ženu, ubiti ženu, silovati ženu) — posebno u eksplicitnom obliku.“ Jasno je da se radi o halucinaciji, odnosno netačnoj informaciji, jer rezultati do kojih smo došli, ni na koji način nisu izmanipulisani, niti softver za analizu korpusa nudi mogućnost filtriranja nasilnih sadržaja.

U ovom slučaju veoma je važno uočiti značaj istraživanja zasnovanih na frekvencijama, bilo da se radi o kolokacijama koje

[15] CQL [tag="ADJ"][lemma="žena"]
[16] [tag="VERB"][word="ženu"]

razgraničavaju značenje bliskih sinonima, bilo o onima čiji je cilj donošenje zaključka o semantičkoj prozodiji određene reči tj. pozitivnom ili negativnom kontekstu. Takav kontekst ne utiče samo na semantiku same reči, već u slučaju imenice *žena* oslikava društveni položaj žene, te ove vrste pretraga mogu da pomognu u razotkrivanju stereotipa i stigmatizacije pojedinih društvenih grupa. Kada je reč o naučnim istraživanjima koja zahtevaju preciznost i potkrepljenost informacija, može se zaključiti da korpusi zadržavaju svoju vrednost i značaj čak i u eri veštačke inteligencije, bar što se tiče trenutnog stanja u ovoj oblasti.

4. Zaključak

Polazeći od obrazovnih potreba novih generacija studenata, koji aktivno integrišu savremene tehnologije u sve aspekte svoga života, uključujući tu i obrazovanje, u radu smo preispitivali efikasnost generativnih modela u odnosu na korpuse, kao i mogućnost njihove integracije u nastavu srpskog kao stranog jezika.

Pošli smo od pozitivnih rezultata korpusnih istraživanja u ovoj oblasti i činjenice da se metodologija *corpus-based* uspešno koristi i pokazuje zadovoljavajuće rezultate na svim nivoima učenja jezika, posebno na srednjem/višem kada je svest o značaju autentičnog jezika već formirana. Ovo kratko istraživanje ograničili smo preciznim jezičkim pitanjima, imajući u vidu prethodna korpusna istraživanja u nastavi srpskog kao stranog na polju polisemije (*kuća-dom*), leksičkih praznina (*babine*) i pozitivnih i negativnih konotacija reči (*žena*).

Pokazalo se da je na polju kolokacija, u slučaju razgraničavanja značenja bliskih sinonima, korpusna pretraga preciznija od AI modela, jer omogućava kvantitativnu analizu na osnovu frekvencija, što nije moguće u slučaju generativnih modela zbog odsustva metapodataka. U isto vreme, kako je prethodno citirano istraživanje pokazalo, rezultati korpusne pretrage verodostojni su i mogu dovesti do novih saznanja, tj. pružiti uvid u nova značenja reči koja nisu dovoljno izdiferencirana u postojećim, posebno bilingvalnim leksikografskim izvorima za istraživanu jezičku kombinaciju.

U definisanju leksičkih praznina generativni model pokazao se potencijalno veoma funkcionalnim na svim nivoima usvajanja jezika s obzirom na to da se ovi sadržaji mogu prevesti na maternji jezik polaznika i time biti dostupni i studentima na nižim nivoima učenja jezika. Iako je i prethodno korpusno istraživanje dalo pozitivne rezultate u slučaju istražene lekseme treba imati u vidu da se radilo o pretrazi za koju je bilo neophodno solidno tehničko i metodološko predznanje, a koja bi eventualno mogla da

se proširi konsultovanjem novih paralelnih korpusa koje imamo na raspolaganju za srpski jezik (Perišić et al 2023, Moderc et al 2023). Da bi se proverili početni pozitivni rezultati generativnih modela u definisanju ove vrste leksike bilo bi neophodno sprovesti složenije istraživanje po pitanju kulturološki specifičnih pojmova izdiferenciranih prema strogim naučnim kriterijumima.

U slučaju reči s pozitivnim ili negativnim konotacijama pretraga pomoću generativnih modela otkriva brojne problematične aspekte. Naime, već je ukazano na to da ovi modeli podležu određenom obliku cenzure, kao u slučaju sadržaja nasilnog karaktera, pa stoga „izbegavaju“ reči s negativnim konotacijama. Zbog toga istraživanja iz oblasti analize diskursa, posebno ona koja se bave govorom mržnje ili ispitivanjem nasilnih i drugih negativno obojenih sadržaja trenutno ne daju validne rezultate.

U nastavi stranih jezika u Srbiji korpusi su nedovoljno zastupljeni čak i kada je reč o većim svetskim jezicima (Vitaz, Poletanović, 2019), što onemogućava sprovođenje složenijih kvantitativnih i kvalitativnih istraživanja koja bi ukazala na potencijalne prednosti i/ili nedostatke primene generativnih modela u glotodidaktici. Kao prvi korak neophodno bi bilo osposobiti što veći broj nastavnika, na različitim nivoima obrazovanja, za korišćenje korpusa u učionici i istovremeno pratiti rezultate dobijene putem generativnih modela, u okviru neke vrste integrativne metode zasnovane na principima DDL-a.

Predlaže se, dakle, usvajanje kritičkog mišljenja koje je u osnovi DDL pristupa i koje nije isključivo ograničeno na korpuse, već kako je pokazano u prethodnim istraživanjima obuhvata i interakciju polaznika s web pretraživačima, ali i novim generativnim modelima. Ovakav integrativni okvir istraživanja može poslužiti kao osnova za definisanje didaktičkog modela koji bi se sprovodio sa studentima srpskog kao stranog jezika (i ne samo), sa ciljem razvijanja kritičkog razmišljanja na temu potencijala i efikasnosti generativnih modela na različitim nivoima usvajanja stranih jezika.

Literatura

[1] Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

[2] Boulton, A. (2012). What data for data-driven learning? EuroCALL *Review*, *20*(1), 23–27.

[3] Boulton, A. (2020). Data-driven learning for younger learners: Obstacles and optimism. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. xiv–xx). Routledge.

[4] Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, *19*(3), 307–328. https://doi.org/10.1017/S0958344007000535

[5] Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge. https://doi.org/10.4324/9780429425899

[6] Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*, https://doi.org/10.1016/j.acorp.2023.100066

[7] Flowerdew, J. (2024). Data-driven learning: From Collins Cobuild Dictionary to ChatGPT. *Language Teaching*, 1–18.

[8] Flowerdew, L. (2025). Constructivist Approaches to Data-Driven Learning. In: McCallum, L., Tafazoli, D. (eds) The Palgrave Encyclopedia of Computer-Assisted Language Learning. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_53-1

[9] Gatto, M. (2014). *The Web as Corpus: Theory and practice*. Bloomsbury Academic.

[10] Harel, I., & Papert, S. (1990). Software design as a learning environment. *Interactive Learning Environments*, *1*(1), 1–32. https://doi.org/10.1080/1049482900010102

[11] Hirata, E. (2025). Young learners and data-driven learning. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_49-1

[12] Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, *10*, 14–24.

[13] Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*, 147–151.

[14] Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010a). A corpus factory for many languages. In *LREC workshop on Web Services and Processing Pipelines*, Malta.

[15] Kilgarriff, A., Kovář, V., Krek, S., Srdanovic, I., & Tiberius, C. (2010b). A quantitative evaluation of word sketches. In *Proceedings of the 14th EURALEX International Congress* (pp. 372–379). The Netherlands.

[16] Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, *3*(3), Article 100065. https://doi.org/10.1016/j.acorp.2023.100065

[17] Ljubešić, N. & Klubička, F. (2014). {bs, hr, sr}WaC – Web corpora of Bosnian, Croatian

[18] and Serbian. In F. Bildhauer & R. Schäfer (Eds.), Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014 (pp. 29-35). Gothenburg, Sweden: Association for Computational Linguistics.

[19] Mikelić Preradović, N., Posavec, K., & Unić, D. (2019). Corpus-supported foreign language teaching of less commonly taught languages. *International Journal of Instruction*, *12*(4), 335–352.

[20] Miličević, M., & Samardžić, T. (2010). Korpusi kao sredstvo za postizanje autonomije u učenju stranog jezika. In J. Vučo & B. Milatović (Eds.), *Autonomija učenika i nastavnika u nastavi jezika i književnosti* (pp. 387–399). Filozofski fakultet.

[21] Moderc, S., Stanković, R., Tomašević, A., & Škorić, M. (2023). An Italian-Serbian sentence aligned parallel literary corpus. *Review of the NCD*, *43*, 78–91. https://www.ncd.matf.bg.ac.rs/issues/43/09.pdf

[22] Perišić, O. (2025). Corpora vs. Web nell'apprendimento del serbo come lingua straniera. In B. Avşar, F. Brachini, V. Cavalloro et al. (Eds.), *Innovamenti: Spazi e percorsi per una ricerca multidisciplinare* (pp. 341–350). Edizioni Università per Stranieri di Siena.

[23] Perišić, O. (2023). *Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico*. Edizioni dell'Orso.

[24] Perišić, O. (2020). Collocazioni nella didattica del serbo come lingua straniera. *Filolog*, *11*(22), 88–115.

[25] Perišić Arsić, O. (2018). Mogućnost primene korpusa u nastavi srpskog jezika kao stranog. In M. Kovačević (Ed.), *Savremena proučavanja jezika i književnosti* (pp. 187–199). Filološko-umetnički fakultet u Kragujevcu.

[26] Perišić Arsić, O. (2020). Translating lexical gaps: A contrastive corpus-based analysis. In M. Matešić & A. Memišević (Eds.), *Language and mind: Proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society* (pp. 93–108). Peter Lang.

[27] Perišić, O., Stanković, R., Ikonić Nešić, M., & Škorić, M. (2023). It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text. In T. Erjavec & M. Eskevich (Eds.), Selected papers from the

CLARIN Annual Conference 2022 (pp. 99–110). Linköping University Electronic Press. https://doi.org/10.3384/ecp198

[28] Posavec, K. (2018). Uporaba korpusa u poučavanju hrvatskoga kao drugoga i inoga jezika. *Studia Lexicographica*, *12*(22), 63–84.

[29] Sinclair, J. McH. (2000). Lexical grammar. *Naujoji Metodologija*, *24*, 191–203.

[30] Stewart, D. (2025). Lost in the forest: Flashpoints for language learners in Sketch Engine. *Iperstoria*, *25*, 537–554. https://doi.org/10.13136/2281-4582/2025.i25.1585

[31] Tekwa, K. (2024). Artificial Intelligence, Corpora, and Translation Studies. In D. Li & J. Corbett (Eds.), The Routledge Handbook of Corpus Translation Studies (pp. 103–118).

[32] Vitaz, M. & ¸Poletanović, M. (2019). Data-Driven Learning. The Serbian case. *EL.LE*, *8*(2), 409-422.

[33] Zhang, M. & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0196

[34] Zhang, R., & Zou, D. (2024). Self-regulated second language learning: A review of types and benefits of strategies, modes of teacher support, and pedagogical implications. *Computer Assisted Language Learning*, *37*(4), 720–765. https://doi.org/10.1080/09588221.2022.2055081

Corpora for Learning Serbian as a Foreign Language in the AI Era

Olja Perišić

Summary

This study explores the challenges of corpus-based teaching of Serbian as a Foreign Language in the context of artificial intelligence technologies. While there are ongoing global discussions about the future of DDL in the AI era, the Serbian context demonstrates a disconnect between technological advancements, particularly those developed by IT experts from JeRTeh, and their application in foreign language teaching.

The research focuses on specific linguistic areas that have been examined in previous corpus studies: polysemy (*kuća–dom*), lexical gaps (*babine*) and positive or negative word connotations (*žena*). The findings show that, in tasks such as distinguishing between closely related

synonyms, corpus searches offer more precise and reliable results than AI, due to their ability to provide frequency-based data. Generative models are useful for lexical gaps, particularly for beginner learners, though corpus-based methods yield strong results when sufficient technical knowledge is available. However, when analysing word connotations, generative models demonstrate significant limitations, frequently avoiding negative content due to embedded safety filters, rendering them unsuitable for discourse analysis.

The main issue is that corpora are not sufficiently integrated into foreign language teaching in Serbia, which limits the ability to make comprehensive comparisons with AI tools. The study recommends training more teachers in corpus-based methods and promoting integrative approaches that combine corpora and generative models. It also advocates cultivating critical thinking, which is central to DDL and involves learner engagement with search engines and AI tools, as well as corpora.

Ultimately, this research offers a model for the thoughtful incorporation of AI into language learning, particularly for less commonly taught languages such as Serbian, while preserving the strengths of corpus-based methodology.

Keywords: corpora, Data Driven Learning, Serbian as a foreign language, JeRTeh, artificial intelligence, generative models.