
Repozitorijum simila savremenog srpskog jezika – izgradnja i evaluacija mašinski potpomognute anotacije

Milica Dinić Marinković¹[0000-0002-2641-8806],
Cvetana Krstev²[0000-0003-3328-9392],
Aleksandra Marković³[0000-0003-2985-3042],
Ranka Stanković⁴[0000-0001-5123-6273]

¹Univerzitet u Beogradu, Filološki fakultet, Srbija,
milica.dinic.marinkovic@fil.bg.ac.rs

²Društvo za jezičke resurse i tehnologije – JeRTeh, Srbija, cvetana@jerteh.rs

³Institut za srpski jezik Srpske akademije nauka i umetnosti, Srbija,
aleksandra.markovic@isj.sanu.ac.rs

⁴Univerzitet u Beogradu, Rudarsko-geološki fakultet, Srbija,
ranka.stankovic@rgf.bg.ac.rs

Scientific paper

DOI: 10.18485/judig.2026.2.ch2

Apstrakt

Cilj rada je evaluacija metodologije izgradnje *Repozitorijuma simila savremenog srpskog jezika* (SSR), razvijenog u okviru šireg istraživačkog projekta konstruisanja prvog korpusno zasnovanog jezičkog resursa za sistematsko proučavanje simila u srpskom. SSR objedinjuje korpusnu ekstrakciju, semantičku anotaciju, lingvističku validaciju i strukturirano organizovanje podataka u jedinstven, otvoren i proširiv istraživački resurs.

U radu se predstavljaju rezultati svih faza obrade – od korpusne ekstrakcije kandidata i filtriranja zasnovanog na pravilima do automatske semantičke anotacije i njene ručne validacije. Repozičtorijum se zasniva na više komplementarnih korpusa savremenog srpskog jezika. Za automatsku obradu primenjen je hibridni neuro-simbolički pristup. U ekstrakciji se kombinuju korpusni upiti i lokalne gramatike realizovane kao konačni automati u sistemu Unitex, koje koriste elektronske rečnike srpskog jezika, dok su za automatsko prepoznavanje i anotaciju kandidata izdvojenih u prethodnim koracima testirani veliki jezički modeli (dalje VJM) GPT-5.2 i Gemini.

Budući zamišljen kao dinamički resurs namenjen lingvističkim, leksikografskim i istraživanjima u okviru obrade prirodnog jezika, ali i kao osnova za dalji razvoj metoda za automatsku obradu figurativnog jezika u srpskom, u fokusu ovog rada su anotacioni model, rezultati lingvističke eva-

luacije mašinski potpomognute anotacije i implikacije dobijenih rezultata za unapređenje razvijenog sistema.

Evaluacija pokazuje da kombinacija simboličkih metoda i VJM omogućava visok stepen preciznosti u identifikaciji simila, uz zadovoljavajuću saglasnost sa stručnim anotatorima. Istovremeno, analiza upućuje da uspešna identifikacija i anotacija simila zahteva uvažavanje složenih odnosa između sintaksičke strukture, semantičke interpretacije i implicitno realizovanih elemenata poređenja, što predstavlja dodatne izazove za automatsku anotaciju. Dobijeni rezultati potvrđuju značaj hibridnog pristupa koji objedinjuje simboličke pristupe i upotrebu VJM sa lingvističkom ekspertizom ugrađenom u anotacionu shemu, evaluaciju i iterativno unapređenje modela.

Ključne reči: Simili, repozitorijum simila, srpski jezik, korpusna lingvistika, semantička anotacija, leksički resursi, figurativni jezik, Uniflex, veliki jezički modeli.

1. Uvod

Simili predstavljaju jednu od najčešće ispitivanih formi figurativnog jezika u računarskoj lingvistici. Zbog eksplicitno izraženog odnosa poređenja i relativno prepoznatljivih jezičkih obrazaca, dugo su smatrani pogodnim kandidatom za automatsku ekstrakciju i semantičku anotaciju. Ipak, početna očekivanja da će takva struktura omogućiti njihovu relativno jednostavnu obradu ustupila su mesto saznanju da identifikacija i interpretacija simila zahtevaju integraciju simboličkih i statističkih metoda, VJM i bogato anotiranih jezičkih resursa. U tom kontekstu, ovaj rad nadovezuje se na istraživanje Marković i sar. (2026), u kome je opisan postupak izgradnje *Repozitorijuma simila savremenog srpskog jezika* (SSR). I dok je prethodni rad bio usmeren na razvoj resursa i metodologiju njegove izgradnje, ovde se detaljnije razmatraju anotacioni model, rezultati evaluacije mašinski potpomognute anotacije i implikacije dobijenih rezultata za unapređenje razvijenog sistema i dalja lingvistička i računarska istraživanja simila.

1.1. Terminologija

Sam termin *simili* nije uobičajen u srpskoj lingvističkoj literaturi. Umesto njega, koriste se nazivi *poredbeni ili komparativni frazeologizam* (Mršević-Radović, 1987; Vulović, 2015) ili jednostavno *poređenje* (Simeon, 1969; Klikovac, 2004), *poredbena frazema*, *ustaljeno poređenje* (Bjelivuk, 2025), pod *poredbeni frazeologizam*.

I za komponente koje ulaze u strukturu ovih višečlanih leksičkih jedinica koristi se različita terminologija u različitim lingvističkim tradicijama.¹ Mi ćemo za subjekat poređenja (koji se u anglosaksonskoj literaturi naziva i *topic*, *tenor*, *target*), koristiti termin *komparandum*, za osnov poređenja (lat. *tertium comparationis*, koji se naziva naziva još i *ground*, *property*), korišćićemo upravo termin *osnov poređenja*, poredbenu rečcu *kao* (naziva se i *marker*), zvaćemo *komparatorom*, dok ćemo objekat poređenja (naziva se i *vehicle*, *source*), nazivati *komparatum*.

Poredbeni frazeologizmi posmatraju se i kao tematsko-rematska struktura, u kojoj temu čini osnova poređenja, a remu poredbeni deo, koji se u srpskoj frazeološkoj literaturi naziva *etalon* (Vu-[lović, 2015](#)).

U ovom radu simili se posmatraju kao šira kategorija figurativnih poređenja koja obuhvata i ustaljene poredbene frazeologizme i kreativne, kontekstualno uslovljene i nekonvencionalne poredbene konstrukcije. Poredbeni frazeologizmi, stoga, predstavljaju samo jedan podtip simila.

1.2. Prethodna istraživanja

Za naš rad relevantno je razmatranje pojava redukcije i ekspanzije frazeologizama, o kojima govori Mršević-Radović ([1987](#)). U prvom slučaju dolazi do uprošćavanja složene strukture, pa tako, na primer, poredbeni frazeologizam može izgubiti jedan ili oba konstituenta poredbene strukture (komparandum ili komparatum), čime nestaje upoređivanje, ali frazeološka jedinica može steći novo frazeološko značenje: *prati ruke kao Pilat* ← *prati ruke (od čega)*. U drugom slučaju, reč je o razvijanju već složene strukture frazeološke jedinice, koja se može proširiti novim članovima, što može biti posledica potrebe da se „oživi prvobitna motivaciona slika ... takvi naknadni eksplikativni elementi konkretizuju i oživljavaju sliku ne narušavajući pri tom frazeološko značenje“ ([Mršević-Radović, 1987](#), str. 147); primeri takve ekspanzije bili bi: *gleda kao tele* → *gleda kao tele u nova (šarena) vrata*; *smejati se kao lud* → *smejati se kao lud na brašno*. Osim toga, autorka pominje i ironične poredbene frazeologizme, tipa *pliva kao sekira* ‘ne pliva nikako, tone’, u kojima komparatum ima funkciju kvalifikatora ili kvantifikatora negativnog polariteta.²

¹ Više o toj terminologiji v. ([Marjanović, 2017](#)).

² Komparatum ovog tipa i Melvingerova ([1983](#)) svrstava u indirektnu negaciju.

Istraživanje Mpuli (2017) višestruko je značajno za naš rad. U njemu je predstavljena anotacija simila u književnim tekstovima, posebno bogatim kreativnim similiba. Autorka ukazuje na to da otvoreni simili (oni koji ne sadrže osnov poređenja) ostavljaju više prostora za imaginaciju. Simili se opisuju i anotiraju zavisno od strukture kao oni koji porede entitete (oni se javljaju uglavnom s nominalnim komparatumom) i oni koji porede procese (komparatum može biti predložka fraza, rečenica ili eliptična rečenica). Ukazuje se na probleme anotacije rečenica koje sadrže više od jednog simila i razlikuje se nekoliko tipova simila: idiomatski, perceptivni, originalni i dr.

Važna, neophodna komponenta poredbenih frazeologizama jeste osnov poređenja, zajedničko obeležje pojmova koji se porede. O tome da ono može biti izraženo eksplicitno ili implicitno prisutno govori Melvingerova (1983). Ona ističe da se pridevi i trpni pridevi, prilozi, neki glagoli i imenice kao upravne reči u poređenjima ujedno ponašaju i kao osnov poređenja (*zdrav kao bik*, *pocrveneti kao paprika*). S druge strane, uz kopulativne, semikopulativne i neke druge glagole, osnov poređenja nije upravna reč, već se on samo podrazumeva (*biti* (→ *nemiran*, *pokretljiv*) *kao živa*, *osećati se* (→ *ugodno*, *sigurno*) *kao riba u vodi*)³.

U radu Kadira i saradnika (2016) autori se bave problemom automatskog zaključivanja implicitnih osobina evociranih similiba, ističući da je velika većina simila analiziranih u tvitovima otvorena, bez eksplicitne osnove poređenja – čak 92%. Smatraju da je za obradu prirodnog jezika i za razumevanje metafore ključno zaključiti šta je osnov poređenja. Implicitna osobina mora biti semantički kompatibilna sa komponentama simila. Njihov model za zaključivanje zasniva se na nekoliko koraka: generisanju kandidata za osnov poređenja iz različitih izvora; evaluaciji tih kandidata zasnovanoj na uticaju različitih komponenta simila; zbirnom rangiranju kandidata.

Iako za naš rad nije neposredno relevantna, pomenućemo i odnos metafore i simila kao značajnu temu u lingvističkim istraživanjima. Dok neki autori smatraju da između njih ne postoji suštinska razlika (Klikovac, 2004), da su metafore zapravo implicitni simili (širu

³ Melvingerova primećuje da se poređenja s istim upravnim glagolom, za razliku od poređenja s istim upravnim pridevom, koja su po pravilu sinonimna, neretko razlikuju upravo zbog implicitne osnove poređenja: *živeti kao braća* 'živeti složno, u velikoj slozi, u ljubavi', *živeti kao mačka i pas* 'živeti u stalnom neprijateljstvu, neprestano se svađati' (Melvinger, 1983).

literaturu o ovome v. (Mathisen et al., 2025)), drugi insistiraju na razlikama, ističući da simili opisuju privremene ili slučajne sličnosti, dok metafore podrazumevaju dublju identifikaciju i trajnije osobine (Goljak, 2009). Henks (2004) naglašava da su simili češći od metafora i da se zasnivaju na kulturno uslovljenim stereotipima, kao i da se sličnima najčešće izdvaja jedna karakteristična osobina komparatuma.

U radu (Niculae and Danescu-Niculescu-Mizil, 2014) prikazana je jedna od prvih velikih korpusnih studija simila na autentičnim tekstovima, na skupu poređenja izdvojenih iz korisničkih recenzija proizvoda i ručno anotiranih prema figurativnosti. Utvrđeno je da je više od 30% analiziranih poređenja figurativno, kao i da se figurativna i doslovna poređenja mogu uspešno razlikovati na osnovu kombinacije domenski specifičnih i domenski nezavisnih obeležja. Najznačajnijim prediktorima figurativnosti pokazali su se semantička sličnost komparanduma i komparatuma, domenska specifičnost pojedinih elemenata poređenja, te psiholingvistička obeležja poput apstraktnosti i slikovitosti (*imageability*). Rezultati pokazuju da figurativna poređenja češće uključuju semantički udaljene pojmove, kao i da su zastupljenija u tekstovima sa izraženim emocionalnim stavom, nego u informativnim tekstovima.⁴

Za srpski jezik značajan doprinos predstavljaju istraživanja Mitrović i saradnika (2020), koji su analizirali ustaljena poređenja u grčkom i srpskom jeziku. Istraživanje je pokazalo visok stepen ekvivalencije između ova dva jezika, a rezultati su iskorišćeni i za proširenje srpskog vordneta (Krstev et al., 2004) novim relacijama namenjenim automatskom prepoznavanju simila.

Milošević i Nenadić (2016; 2018) napravili su korpus srpskih simila, pomoću podataka s Interneta i tehnikom prikupljanja podataka od velikog broja ljudi (*crowdsourcing*) za savremene simile, ali su u korpus ušli i simili koje je navodio Vuk Karadžić. Korpus sadrži 852 simila. Autori ukazuju na veliki broj lažno pozitivnih simila (*radi kao konj* prema *radi kao advokat*), jer nisu uspeali da pronađu leksička obeležja koja bi modelu signalizirala razliku, zbog čega su se oslonili na ručnu proveru. Zanimljivo je da presek između savremene i tradicionalne zbirke čini samo 9 simila, što ukazuje na to da se kategorija simila ne iscrpljuje u skupu konvencionalnih poredbenih

⁴ Sa sličnim rezultatom i sami smo se suočili u anotaciji srpskog proširenja korpusa ELEXIS, u kom sistem za prepoznavanje simila nije izdvojio nijedan simili u tekstu, upravo zbog prirode tekstova u ovom korpusu (Krstev et al., 2024).

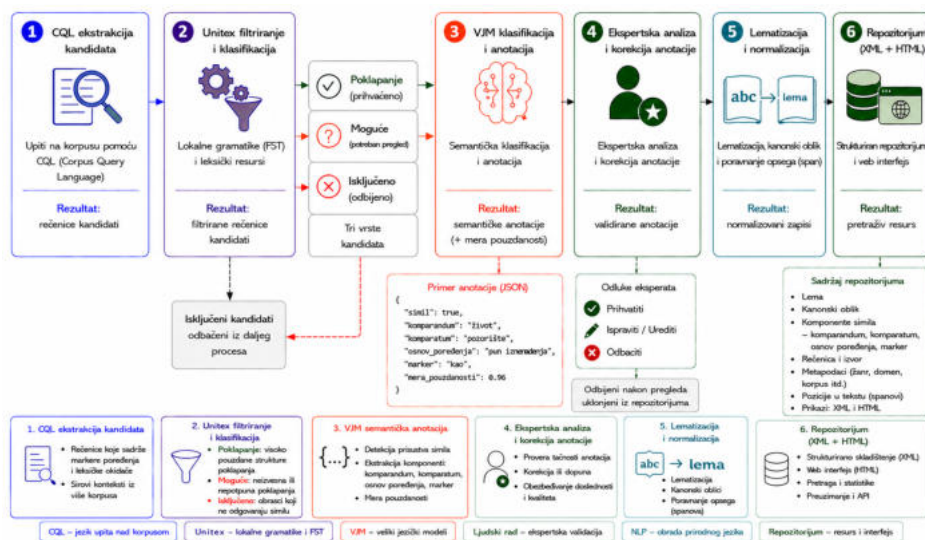
obrazaca, već predstavlja produktivno područje jezičke kreativnosti, u kome se kontinuirano uspostavljaju nove figurativne veze između domena poređenja.

Parizoska i Filipović Petrović (2017) razmatraju moguće varijante prideva u similiba s modelom pridev + *kao* + imenica. Pridevsku komponentu može zameniti sinoniman pridev (*snažan/jak kao lav*), od prideva izveden prilog (*uporno ko mazga*) ili glagol (*miran kao bubica / mirovati kao bubica*), semantički blizak (*brz kao metak / juriti/izleteti kao metak*), ali i semantički nepovezan (*bled kao duh / izgledati kao duh*). Pridevska komponenta može ostati i prazna, uz glagol *biti*, npr. *bila sam kao balon*, s implicitnom osnovom poređenja (*napuhan*). Autorke zaključuju da se uz kopulu osnov poređenja konstruiše u procesu razumevanja i da se javlja samo implicitno u značenju simila. Osnov poređenja može biti implicitan i u nekim imeničkim similiba (*bod ko kuća*, tj. velik; *obraz ko opanak*, tj. tvrd). U svim ovim slučajevima implicitna je osobina koja se obično povezuje sa komparatumom (balon – napuhan, puter – mek, med – sladak). Autorke ukazuju i na neke *kao* + imenica strukture koje se javljaju s nizom prideva i glagola koji ne moraju ukazivati na neku osobinu komparatuma; tako se može biti *umoran/ljut/ljubomoran kao pas*, ili se može *lagati/raditi/živeti kao pas*. Za razliku od slučajeva s implicitnim osnovom poređenja, ovde nema određene osobine koja se isključivo vezuje za psa, osim možda one izražene pridevom *veran*, što autorke objašnjavaju desementizacijom, zbog čega *kao pas* postaje slično intenzifikatorima. Ovo istraživanje pokazuje da simili nisu fiksne konstrukcije, već fleksibilne strukture čije varijacije zavise od semantičkih, gramatičkih i kognitivnih faktora.

U radu (Krstev, 2021) upoređuje se upotreba tradicionalnih i novih simila u korpusu SrpELTeC i savremenim srpskim romanima. Autorka pretpostavlja da savremeni pisci koriste simile ređe ili sa izmenjenim repertoarom, budući da je model mogao da prepozna samo simile koji se javljaju u korpusu starih srpskih romana, te predlaže formiranje baze srpskih simila i proširenje istraživanja na različite strukturne tipove simila. Takvi nalazi ukazuju na ograničenja pristupa zasnovanih isključivo na inventarima ustaljenih poredbenih izraza i otvaraju potrebu za izgradnjom specijalizovanog resursa koji bi omogućio sistematsko proučavanje simila različitih tipova i stepena konvencionalizovanosti. Izgradnja *Repozitorijuma simila savremenog srpskog jezika* započeta je upravo sa ciljem da odgovori na ove izazove.

2. Metodologija⁵

Metodologija ekstrakcije i anotacije simila prikazana je na Slici 1. Proces započinje izdvajanjem kandidata iz korpusa pomoću unapred definisanih CQL obrazaca, nakon čega se primenjuju lokalne gramatike i konačni automati realizovani u sistemu Unitex, uz oslanjanje na elektronske rečnike srpskog jezika (Krstev, 2008). Na osnovu pravila zasnovanih na morfosintaksičkim i leksičkim ograničenjima, kandidati se klasifikuju u tri grupe: pouzdani primeri simila (*poklapanje*), mogući primeri (*moguće*) i odbačeni primeri. Preostali kandidati se zatim prosleđuju velikim jezičkim modelima, koji određuju da li rečenica zaista sadrži simili i vrše anotaciju strukturnih elemenata simila, uključujući subjekat poređenja ili komparandum (*tenor*), osnov poređenja (*ground*), leksičku realizaciju osnova poređenja (*property*), objekat poređenja (*vehicle*) i komparator (*marker poređenja*). Nakon ručne validacije potvrđeni simili se lematizuju, povezuju sa izvornim korpusnim rečenicama i izvoze u XML format zajedno sa HTML prikazom namenjenim krajnjim korisnicima.



Slika 1: Proces ekstrakcije i anotacije simila

⁵ Odeljak 2 predstavlja proširenu i dopunjenu verziju metodološkog okvira prethodno predstavljenog u radu Marković i sar. (2026).

2.1. Ekstrakcija kandidata iz korpusa

Početna ekstrakcija kandidata zasnivala se na šest CQL obrazaca osmišljenih za prepoznavanje čestih konstrukcija simila u srpskom jeziku sa markerima poređenja: *kao*, *kô*, *k'o*, *nalik*, *poput* i *kano*, primenjenih na korpusu SrpKor2021+, koji objedinjuje korpruse SrpKor2013 i SrpKor2021 (Vitas et al., 2025). Obrasci su obuhvatali više sintaksičkih konfiguracija, uključujući strukture tipa *pridevska fraza – marker poređenja – imenička fraza* (AkaoN), *marker poređenja – imenička fraza – pridevska fraza* (kaoNA), glagole praćene predložko-padežnim konstrukcijama (VkaoPrepN), glagole sa imeničkim komparatumom (*vehicle*) (VkaoN), kao i priloške fraze praćene markerom poređenja i predložko-padežnom konstrukcijom (ADVkaoPrepN). Ovi upiti su namerno definisani široko kako bi se postigao što veći odziv (*recall*) i iz korpusa izdvojio veliki skup potencijalnih kandidata. Tabela 1 prikazuje obrasce i odgovarajuće CQL upite korišćene u prvoj fazi ekstrakcije kandidata.

2.2. Filtriranje zasnovano na pravilima i unapređenje pomoću konačnih automata

U drugom koraku lista kandidata dodatno je pročišćena korišćenjem konačnih automata i transduktora implementiranih u sistemu Unitex, uz oslanjanje na bogate elektronske rečnike srpskog jezika. Na taj način metodologija kombinuje široku ekstrakciju zasnovanu na CQL upitima sa lingvistički motivisanim unapređenjem pomoću konačnih automata (FST), čime se smanjuje količina šuma u podacima uz očuvanje visoke pokrivenosti za kasniju anotaciju. Za prvo filtriranje korišćen je postojeći inventar od preko 600 imeničkih i 300 glagolskih simila. Stavke u ovom inventaru opisuju leksičke varijacije, pa je tako u njemu (*mek+mekan+mehak*) kao (*kadiva+kadifa*) jedan simili. Osim toga, u inventaru se beleže i moguće odredbe pojedinih komponenata simila, kao, na primer, *okrugao kao* [*pun*] (*mesec+mjesec*). Na osnovu stavki iz ovog inventara kreirani su transduktori koji, vodeći računa o promenljivom redosledu komponenata simila i njihovoj udaljenosti u rečenici, sa velikom preciznoću i solidnim odzivom prepoznaju simile u obrađenom tekstu (Krstev et al., 2020). Tako je iz ekstrahovanih primera izdvojena grupa pouzdanih kandidata za simile (*poklapanje*), dok preostali kandidati iz ekstrahovanih lista sadrže kandidate koji zahtevaju dalje prečišćavanje.

Tabela 1: CQL obrasci korišćeni za ekstrakciju kandidata za simile

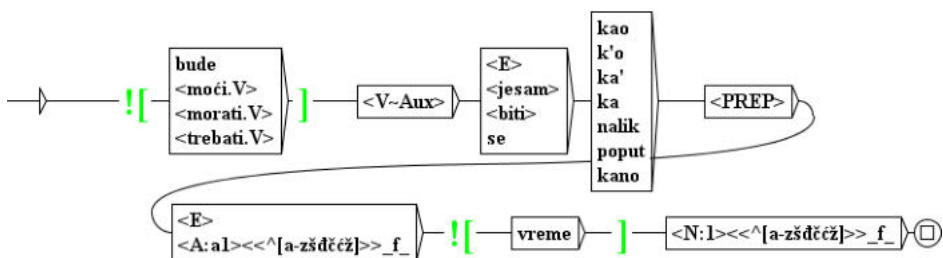
ID	Obrazac	CQL upit
1	AkaoN	[tag="A"& lemma="^[a-zčćžšđ].*"] [tag="V"& lemma="jesam biti"]? [word="kao ko k'o ka' ka nalik poput kano"] [tag="A"& lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
2	kaoNA	[word="kao ko k'o ka' ka nalik poput kano"] [tag_ud="NOUN"] [tag="V"& lemma="jesam biti"]? [tag="A"& lemma="^[a-zčćžšđ].*"]
3	VkaoPrepN	[tag="V"] [word=šetag="V"& lemma="jesam biti"]? [word="kao ko k'o ka' ka nalik poput kano"] [tag="PREP.*"] [tag="A"& lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
4	kaoPrepNV	[word="kao ko k'o ka' ka nalik poput kano"] [tag="PREP.*"] [tag="A"& lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"] [word=šetag="V"& lemma="jesam biti"]? [tag="V"]
5	VkaoN	[tag_ud="VERB"] [tag="V"& lemma="jesam bitiword=še"]? [tag_ud="NOUN"]? [word="kao ko k'o ka' nalik poput kano"] [tag="A"& lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
6	AdvkaoPrepN	[tag_ud="ADV"] [tag="V"& lemma="jesam bitiword=še"]? [word="kao ko k'o ka' ka nalik poput kano"] [tag_ud="ADP"] [tag="A"& lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]

U drugoj fazi je na kandidate neprepoznate u prvoj fazi prime-njen dodatno unapređen postupak ekstrakcije uvođenjem strožih lek-sičkih i morfosintaksičkih ograničenja na početne CQL obrasce. Kod obrasca AkaoN izostavljeno je više podvrsta prideva kako bi se sma-njio broj lažno pozitivnih kandidata: pridevi u komparativu, prisvojni pridevi, participi, pridevi izvedeni od vlastitih imena i redni brojevi. Iako ovakvo filtriranje može isključiti pojedine validne simile, poput primera *zbijeni kao sardine*, većina uklonjenih slučajeva verovatno ne

predstavlja simile. Takođe je isključena varijanta markera *ka*, jer se u velikom broju slučajeva pojavljuje kao deo predloških konstrukcija, a ne kao marker poređenja (v. Marković i saradnici 2026).

Slično unapređenje primenjeno je i na obrazac *kaoNA*. Dodatna ograničenja uvedena su zato što je početna ekstrakcija generisala veliki broj kandidata sa visokim nivoom šuma, naročito u kategoriji *moгуće*. Pridev koji predstavlja osnov poređenja (*property*) ograničen je na pozitiv, izostavljeni su uz ostale, prethodno pomenute kategorije prideva, i odnosni pridevi, a zadržani su samo pridevi koji započinju malim slovom. Ova ograničenja značajno su smanjila broj nerelevantnih podudaranja, iako istovremeno mogu dovesti do određenog smanjenja odziva (*recall*).

Kod obrasca *VkaoN* regularni izrazi nisu bili dovoljni zbog izuzetno velikog broja izdvojenih kandidata. Zbog toga je umesto njih korišćen Unitex graf. Graf je kombinovao glagole koji su već potvrđeni u prethodno identifikovanim glagolskim similima sa proizvoljnim imenicama, dok su proizvoljni glagoli kombinovani samo sa imenicama koje su već bile zabeležene u ranije prepoznatim similima. Ovaj skup dodatno je proširen uključivanjem odabranih semantičkih klasa imenica iz elektronskih rečnika srpskog jezika, uključujući nazive životinja, biljaka, prirodnih pojava, delova tela i predmeta svakodnevne upotrebe. Pošto je filtriranje bilo znatno strože nego kod pridevskih obrazaca, odbačeni primeri nisu označavani jednostavno kao *nema slaganja*, već kao *verovatno nema slaganja*, čime je ostavljena mogućnost da manji broj potencijalno validnih primera bude naknadno preispitan tokom ručne validacije.



Slika 2: Unitex graf za ekstrakciju potencijalnih kandidata simila oblika *VkaoPrepN*

Za obrasce *kaoPrepNV* i *VkaoPrepN* još uvek nisu konstruisani grafovi koji bi omogućili izdvajanje potpuno pouzdanih primera, te

je za njih izostala prva faza obrade. Za prečišćavanje rezultata korišćeni su Unitex grafovi (Slika 2) koji omogućavaju uvođenje dodatnih ograničenja: isključuju se pomoćni i modalni glagoli, a imenica i pridev koji joj eventualno prethodi treba da budu u nominativu i da ne počinju velikim slovom. I ovde je marker *ko* isključen iz obrade jer se često mešao sa upitno-odnosnom zamenicom *ko*. Obrazac *ADVka-oPrepN* obrađivan je na isti način kao i glagolsko-predložki obrasci. Međutim, početna analiza pokazuje da ova grupa sadrži veoma mali broj korisnih kandidata za simile.

Posmatrano u celini, postupak unapređenja ekstrakcije kombinuje široko korpusno pretraživanje zasnovano na CQL upitima, leksičko filtriranje, semantičke klase iz elektronskih rečnika srpskog jezika i Unitex grafove zasnovane na konačnim automatima. Na taj način izdvojeni materijal se klasifikuje u tri kategorije: pouzdane kandidate, moguće kandidate i primere koji najverovatnije ne predstavljaju simile.

2.3. Semantička anotacija uz pomoć velikih jezičkih modela

Treću fazu metodologije predstavlja automatska anotacija kandidata pomoću VJM. U tu svrhu razvijen je pristup zasnovan na upitima (promptovima), specijalizovan za prepoznavanje simila i izdvajanje i anotaciju njihovih semantičkih komponenata u srpskom jeziku. Sistemski upit definisao je zadatak kao identifikaciju figurativnih poređenja uvedenih eksplicitnim markerima poređenja, poput onih korišćenih u prethodnim fazama ekstrakcije. Posebna pažnja posvećena je razlikovanju figurativnih simila od doslovnih poređenja (*obučena kao muškarac*) i izraza koji označavaju ulogu ili zanimanje (*živeti kao podstanar, raditi kao advokat*). U upitu je naglašeno da se simili može prepoznati samo ukoliko poređenje podrazumeva prenos istaknute osobine između dva entiteta.

Za svaku rečenicu model je dobijao zadatak da utvrdi da li ona sadrži simili i da rezultat vrati u obliku strukturiranog zapisa u JSON formatu, pogodnog za dalju automatsku obradu. Ukoliko je simili prepoznat, model izdvaja osnovne semantičke komponente: *tenor* (subjekat poređenja), *vehicle* (objekat poređenja), *komparator* (marker poređenja), *property* (osnov poređenja eksplicitno izražen u tekstu) i *ground* (konceptualna osnova poređenja, tj. zajednička semantička osobina koja povezuje komparandum i komparatum, najčešće izražena deadjektivnom imenicom, npr. *belina*). Ovakva struktura

omogućava detaljan opis figurativnih poređenja, kao i njihovu kasniju pretragu i analizu u repozitorijumu (više o ovome v. u tački 4.1.). Pored toga, model generiše i kanonski oblik simila. Svaka predikcija sadrži i procenu pouzdanosti (*confidence score*), kao i kratko lingvističko obrazloženje na srpskom jeziku.

Prednost ovakvog pristupa je u tome što VJM mogu da koriste širi kontekst rečenice i implicitna semantička znanja prilikom interpretacije poređenja, čime se omogućava prepoznavanje složenijih i nekonvencionalnih simila koje je teško obuhvatiti isključivo pravilima zasnovanim na lokalnim gramatikama. Rezultat ove faze je skup automatski anotiranih kandidata, koji se zatim prosleđuju završnoj fazi ručne validacije i kontrole kvaliteta.

Rečenice su obrađivane u paketima korišćenjem dinamički generisanih korisničkih upita koji su sadržali numerisane primere. Strategijom grupne obrade smanjeno je opterećenje API servisa i omogućena anotacija velikog broja kandidata izdvojenih iz korpusa. Strukturnisani pristup zasnovan na upitima kombinuje lingvistička uputstva sa ograničenjima nad izlaznim formatom pogodnim za mašinsku obradu, čime se omogućava efikasna poluautomatska izgradnja skupa podataka o srpskim similitima pogodnog za dalju korpusnu analizu, evaluaciju i obučavanje modela.

U pripremnoj fazi evaluirani su modeli OpenAI GPT-5.4 i GPT-5.4-mini na ručno pregledanom skupu za testiranje kako bi se procenila njihova pogodnost za anotaciju simila velikih razmera. Evaluacija je bila usmerena na sposobnost modela da razlikuju figurativne simile od doslovnih poređenja, tačno identifikuju komponente simila i generišu dosledne i strukturno ispravne zapise u formatu JSON. Iako je veći model GPT-5.4 pokazao nešto stabilnije lingvističke interpretacije u dvosmislenim slučajevima, ekonomičniji GPT-5.4-mini ostvario je rezultate koji su se pokazali dovoljno pouzdanim za potrebe korpusne anotacije. Imajući u vidu izuzetno velik broj kandidatskih rečenica i povezane troškove obrade, GPT-5.4-mini izabran je za primarni model anotacije, kao kompromis između kvaliteta anotacije, brzine obrade i finansijske isplativosti.

2.4. Ručna evaluacija i validacija anotacije

Završni korak čini ručna validacija i korekcija stratifikovanog uzorka izdvojenih kandidata. Kako bi se obezbedila pouzdana evaluacija uz istovremeno održavanje prihvatljivog obima anotatorskog

rada, odabran je uzorak od 900 rečenica raspoređenih po obrascima ekstrakcije i automatski dodeljenim grupama pouzdanosti (*poklapanje* i *moгуće*). Izabrane primere nezavisno su anotirala dva eksperta (koautori ovog rada). Pored binarne klasifikacije (simili ili nije simili), anotatori su procenjivali i ispravnost izdvojenih komponenata simila, kao i lematizovanu formu.

Evaluacijom je obuhvaćeno više aspekata uspešnosti sistema. Najpre je procenjena preciznost zasebno za svaki obrazac ekstrakcije, kao i ukupna preciznost na nivou celokupnog skupa podataka. Saglasnost između anotatora izračunata je pomoću Koenovog koeficijenta κ , kako bi se procenila doslednost anotacije i stepen dvosmislenosti prisutan u interpretaciji simila. Pored toga, sprovedena je kvalitativna analiza grešaka sa ciljem identifikacije najčešćih izvora lažno pozitivnih rezultata, uključujući konstrukcije bez figurativnosti, leksikalizovane višočlane izraze i sintaksičke dvosmislenosti povezane sa similiba koji sadrže predložko-padežne konstrukcije. Performanse sistema dodatno su procenjivane u odnosu na zlatni standard⁶, korišćenjem standardnih metrika kao što su tačnost (*accuracy* – A), preciznost (*precision* – P), odziv (*recall* – R) i mera F_1 .

Evaluacioni okvir zasnovan na uzorkovanju omogućava pouzdanu procenu kvaliteta metodologije ekstrakcije uz razuman obim manualnog rada. Istovremeno, rezultati pružaju uvid u relativnu efikasnost pojedinačnih CQL obrazaca, Unitex grafova i strategija filtriranja zasnovanih na velikim jezičkim modelima, čime se obezbeđuju smernice za dalji razvoj i unapređenje sistema za automatsku ekstrakciju simila u srpskom jeziku.

2.5. Izgradnja repozitorijuma i predstavljanje podataka

Repozitorijum je konstruisan na osnovu validiranog i anotiranog skupa podataka, i organizovan oko normalizovane liste lematizovanih simila, pri čemu je svaki zapis povezan sa izvornim korpusnim primerima i njihovim semantičkim anotacijama. Podaci se čuvaju u strukturiranom XML formatu, pogodnom za dalju automatsku obradu, integraciju sa drugim jezičkim resursima i razmenu podataka između različitih sistema za obradu prirodnog jezika.

⁶ Skup od 927 rečenica nezavisno su evaluirala i korigovala dva anotatora, nakon čega su sve razlike u anotaciji usaglašene kako bi se formirao zlatni standard. Više detalja u odeljku 3.3.1.

Pored mašinski čitljive reprezentacije, razvijen je i HTML interfejs namenjen istraživačima, leksikografima i drugim korisnicima repozitorijuma. Interfejs omogućava pregled i pretragu simila prema različitim kriterijumima, uključujući lematizovanu formu, tip konstrukcije, strukturne elemente simila – komparandum, konceptualni (*ground*) i leksikalizovani (*property*) osnov poređenja, komparator i komparatum. Za svaki zapis prikazuju se originalna rečenica, anotirane komponente simila i dodatni metapodaci, što omogućava jednostavnu kvalitativnu analizu i proveru rezultata automatske anotacije.

Dvostruka reprezentacija podataka, kroz XML format i interaktivni HTML prikaz, podržava i računarsku obradu podataka i intuitivno pretraživanje i upotrebu od strane korisnika. XML zapis omogućava korišćenje repozitorijuma u korpusnim istraživanjima, zadacima obrade prirodnog jezika i razvoju novih modela za prepoznavanje figurativnog jezika, dok HTML interfejs pruža intuitivno okruženje za istraživanje i analizu simila. Na taj način repozitorijum predstavlja ne samo zbirku anotiranih primera već i održivu istraživačku infrastrukturu namenjenu proučavanju figurativnog jezika, leksikografskom radu i razvoju savremenih jezičkih tehnologija za srpski jezik.

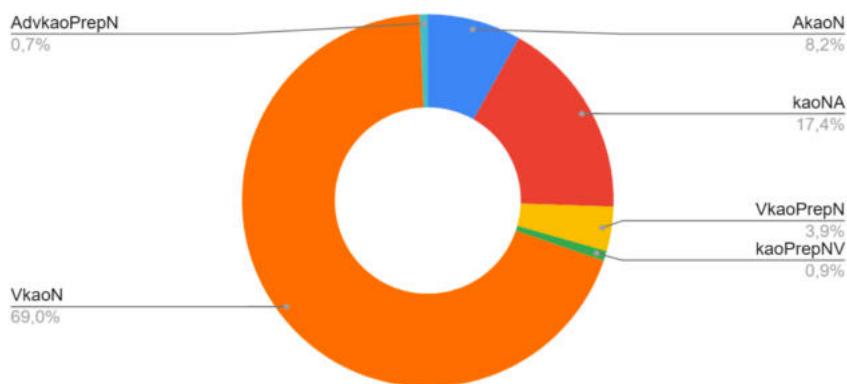
3. Rezultati⁷

3.1. Korpusna ekstrakcija i izdvajanje kandidata

Početna faza ekstrakcije zasnivala se na šest CQL obrazaca, kojima je iz korpusa izdvojeno ukupno 574.195 kandidatskih rečenica. Najproduktivniji obrazac bio je *VkaoN*, koji modeluje glagolske simile. Ovim obrascem izdvojeno je 396.319 konkordanci, 285.523 različita površinska oblika i 232.462 različite leme (Slika 3). Ovakvi rezultati potvrđuju da su glagolske poredbene konstrukcije veoma česte i leksički raznovrsne u korpusu savremenog srpskog jezika.

Drugu po veličini grupu činio je obrazac *kaoNA*, koji obuhvata konstrukcije započete markerom poređenja, sa 99.732 konkordance i 57.327 različitih lema. Sledeći po produktivnosti bio je pridevski obrazac *AkaoN*, sa 46.849 konkordanci i 32.861 različitom lemom.

⁷ Deo rezultata prikazanih u ovom odeljku prethodno je sažeto predstavljen u radu Marković i saradnici (2026). Zbog prostornih ograničenja tog rada pojedini aspekti evaluacije i predstavljanja podataka tamo nisu mogli biti detaljnije razmotreni, te su ovde prikazani u proširenom i dopunjenom obliku.



Slika 3: CQL ekstrakcija kandidata

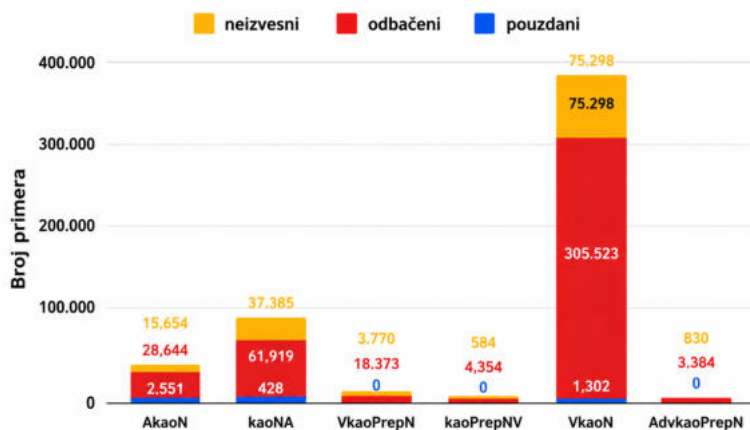
Predložke konstrukcije generisale su manje, ali i dalje veoma obimne skupove kandidata. Obrazac *VkaoPrepN* izdvojio je 22.143 konkordance i 14.547 različitih lema, dok je obrazac *kaoPrepNV* rezultovao sa 4.938 konkordanci i 3.793 leme. Najmanje produktivan bio je obrazac *ADVkaoPrepN*, sa 4.214 konkordanci i 3.452 različite leme.

Ukupan broj izdvojenih konkordanci i različitih lema ukazuje na bogatstvo i raznovrsnost poredbenih konstrukcija u srpskom jeziku. Istovremeno, rezultati potvrđuju neophodnost naknadnih faza filtriranja i klasifikacije, budući da značajan deo izdvojenih primera predstavlja doslovna poređenja, ustaljene višечlane izraze ili druge konstrukcije koje ne odgovaraju figurativnim similitudama. Zbog toga je primena lokalnih gramatika, konačnih automata i velikih jezičkih modela predstavljala ključan korak u daljem sužavanju skupa kandidata i identifikaciji stvarnih primera figurativnog poređenja.

3.2. Unitex filtriranje i klasifikacija kandidata

Slika 4 prikazuje fazu filtriranja i klasifikacije pomoću sistema Unitex, koja je značajno smanjila broj kandidatskih rečenica razdvajanjem primera na pouzdana poklapanja, odbačene slučajeve i neizvesne kandidate koji zahtevaju dodatnu proveru. Obrazac *AkaoN* dao je najveći broj pouzdanih poklapanja među pridevskim konstrukcijama, sa 2.551 potvrđenim kandidatom, uz 15.654 moguća poklapanja i 28.644 odbačena primera. Nasuprot tome, obrazac *kaoNA* rezultovao je sa svega 428 pouzdanih poklapanja, ali sa veoma velikim brojem

neizvesnih kandidata (37.385), što ukazuje na to da je ova struktura znatno podložnija šumu i dvosmislenostima.



Slika 4: Unitex filtriranje i klasifikacija

Glagolsko-predložki obrasci *VkaoPrepN* i *kaoPrepNV* u ovoj fazi nisu dali direktno potvrđene primere simila, već su uglavnom generisali odbačene ili neizvesne kandidate. Ovakav rezultat pokazuje da se ovi obrasci u ispitivanom korpusu nisu pokazali produktivnim za ekstrakciju simila i da u njima dominiraju doslovna poređenja, ustaljeni izrazi ili frazeološke konstrukcije, a ne figurativni simili (npr, *kao od bede*).

Najveći skup kandidata generisao je obrazac *VkaoN*, sa 1.302 pouzdana poklapanja, 75.298 neizvesnih kandidata i više od 305.000 odbačenih primera. Ovi rezultati istovremeno potvrđuju veliku produktivnost glagolskih poredbenih konstrukcija u srpskom jeziku, ali i visok stepen njihove semantičke dvosmislenosti. Mnogi izdvojeni primeri predstavljaju doslovna poređenja ili ustaljene izraze koji formalno odgovaraju obrascu, ali ne predstavljaju stvarne figurativne simile.

Konačno, obrazac *ADVkaoPrepN* nije proizveo nijedan pouzdano potvrđen primer u ovoj fazi, već isključivo neizvesne i odbačene kandidate.

Posmatrano u celini, faza filtriranja značajno je smanjila količinu šuma u podacima, istovremeno zadržavajući veliki broj potencijalno vrednih kandidata za naredne faze anotacije zasnovane na VJM i na ručnoj validaciji. Rezultati pokazuju da kombinacija CQL ek-

strakcije i filtriranja pomoću lokalnih gramatika i konačnih automata predstavlja efikasan mehanizam za sužavanje veoma velikog početnog skupa kandidata na broj primera pogodan za detaljniju semantičku analizu.

3.3. Validacija anotacije zasnovane na velikim jezičkim modelima

U narednoj fazi obrade analizirani su kandidati iz kategorija *poklapanje* i *moгуće*. Međutim, za tri podskupa koja su sadržala više od 15.000 primera obrađen je reprezentativni uzorak od po 3.000 kandidata. Dobijeni rezultati pokazali su veoma visok stepen saglasnosti između filtriranja zasnovanog na konačnim automatima (FST) i validacije pomoću VJM. Kod kandidata svrstanih u kategoriju *poklapanje* sva tri glavna obrasca (*AkaoN*, *kaoNA* i *VkaoN*) ostvarila su izuzetno visoke stope potvrde, u rasponu od 0,97 do 0,99. Ovi rezultati ukazuju na to da je jezički model u najvećem broju slučajeva potvrdio ispravnost kandidata koje je prethodno identifikovala faza filtriranja zasnovana na postojećem inventaru simila (vidi odeljak 2.2.), Unitex pravilima i konačnim automatima. Ukupno je 3.995 od 4.055 kandidata iz kategorije *poklapanje* potvrđeno kao stvarni simili u okviru anotacionog procesa pomoću VJM.

Nasuprot tome, kategorija *moгуće* pokazala je znatno niže stope potvrde i mnogo veće razlike između pojedinačnih obrazaca ekstrakcije. Najvišu stopu validacije među neizvesnim kandidatima ostvario je obrazac *kaoPrepNV* (0,80), dok su obrasci *VkaoPrepN* i *ADVkaoPrepN* dostigli vrednost od 0,64. Pridevski obrazac *AkaoN* zadržao je umerenu upotrebljivost i u ovoj grupi kandidata, sa stopom potvrde od 0,54. S druge strane, obrasci *kaoNA* i *VkaoN* pokazali su znatno nižu preciznost, sa stopama potvrde od svega 0,07, odnosno 0,34.

Posmatrano na nivou cele kolekcije, samo 44% kandidata označenih kao *moгуće* potvrđeno je kao stvarni simili. Ovakav rezultat potvrđuje da ova grupa sadrži znatno više šuma, dvosmislenosti i graničnih slučajeva u odnosu na kategoriju *poklapanje*. Istovremeno, rezultati pokazuju da kombinacija filtriranja zasnovanog na pravilima i validacije pomoću VJM predstavlja efikasan pristup za izdvajanje pouzdanih primera simila iz veoma velikih korpusnih skupova podataka.

3.3.1. Saglasnost između anotatora

Saglasnost između dva ljudska anotatora procenjena je korišćenjem Koenovog koeficijenta κ u dva evaluaciona režima: fleksibilnom i strogom. U fleksibilnom režimu, u kome je bilo dovoljno da oba anotatora prepoznaju prisustvo simila, postignuta je saglasnost u 848 od 926 primera. Ukupan procenat saglasnosti iznosio je 91,6%, dok je vrednost Koenovog koeficijenta κ bila 0,830, što prema uobičajenim interpretacijama predstavlja gotovo savršenu saglasnost.

U strogom režimu evaluacije, gde se pored samog prepoznavanja simila zahtevala i saglasnost u vezi sa obrascima, anotatori su se složili u 830 slučajeva. U ovom slučaju procenat saglasnosti iznosio je 89,6%, dok je Koenov koeficijent κ dostigao vrednost 0,792, što odgovara visokom nivou, odnosno značajnoj saglasnosti. Ovi rezultati ukazuju na visok stepen doslednosti među anotatorima i potvrđuju pouzdanost primenjenog postupka anotacije. Ilustrovaćemo to primerima s prepoznavanjem u strogom i fleksibilnom režimu evaluacije: što se prvog tiče, *beo kao krila serafima* evaluirano je kao simili; što se tiče drugog, *lepa kao dan božji* ispravno je prepoznato kao simili, ali oba evaluatora su se složila da je u pitanju drugi model – umesto kaoNA u pitanju je model AkaoN.

Saglasnost između ljudskih anotatora i kombinovanog sistema zasnovanog na FST pravilima i VJM bila je niža od saglasnosti postignute između samih anotatora, ali je i dalje ostala na zadovoljavajućem nivou. U fleksibilnom režimu evaluacije vrednosti Koenovog koeficijenta κ iznosile su 0,669 i 0,662 za dva anotatora, uz procenat saglasnosti veći od 83%. Ovi rezultati ukazuju na značajnu saglasnost između automatskog sistema i procena eksperata.

U strogom režimu evaluacije saglasnost se očekivano smanjila. Vrednosti Koenovog koeficijenta κ iznosile su 0,532 i 0,546, dok se ukupan procenat saglasnosti kretao oko 76%. Takvi rezultati odgovaraju umerenoj saglasnosti i pokazuju da automatski sistem uspešno prepoznaje većinu relevantnih simila, ali da razlike i dalje postoje kada se primenjuju stroži kriterijumi interpretacije i anotacije.

Posmatrano u celini, rezultati potvrđuju da kombinacija pravila zasnovanih na konačnim automatima i VJM uspešno aproksimira ljudsku anotaciju. Istovremeno, viši nivo saglasnosti među ljudskim anotatorima ukazuje da pojedini granični i dvosmisleni slučajevi i dalje zahtevaju ljudsku procenu, što potvrđuje opravdanost poluautomatskog pristupa u izgradnji repozitorijuma simila.

3.3.2. Evaluacija u odnosu na zlatni standard

Rezultati evaluacije pokazuju da oba pristupa ostvaruju veoma visoku preciznost, što potvrđuje da su identifikovani simili u najvećem broju slučajeva ispravni i da se lažno pozitivni rezultati pojavljuju relativno retko. U strogom režimu evaluacije preciznost dostiže vrednost od 0,98, što ukazuje na izuzetnu pouzdanost sistema pri potvrđivanju kandidata za simile. Međutim, ovakav pristup ostvaruje niži odziv od 0,68, što pokazuje da značajan broj stvarnih simila ostaje neprepoznat kada se primenjuju stroži kriterijumi poklapanja i anotacije.

Kao posledica toga, mera F_1 dostiže vrednost od 0,80, odražavajući kompromis između visoke preciznosti i ograničene pokrivenosti. Ovi rezultati sugerišu da strogi režim evaluacije favorizuje pouzdanost identifikovanih primera, ali po cenu propuštanja određenog broja validnih simila. Takvo ponašanje je očekivano u sistemima zasnovanim na kombinaciji pravila i automatske validacije, gde se prioritet daje minimizaciji lažno pozitivnih rezultata.

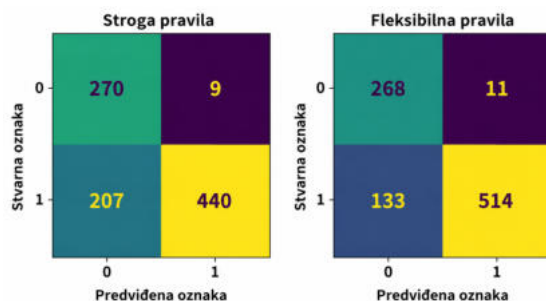
Tabela 2: Rezultati evaluacije sistema za ekstrakciju i validaciju simila u odnosu na zlatni standard

Metrika	Stroga evaluacija	Fleksibilna evaluacija
Tačnost (A)	0,7667	0,8445
Preciznost (P)	0,9800	0,9790
Odziv (R)	0,6801	0,7944
Mera F_1 (F_1)	0,8029	0,8771

Nasuprot tome, fleksibilni režim evaluacije ostvaruje znatno veći odziv od 0,79, uz gotovo identičnu preciznost od 0,98. Zahvaljujući tome postiže se značajno viša mera F_1 od 0,88, kao i bolja ukupna tačnost sistema (0,84 u poređenju sa 0,77 u strogom režimu evaluacije).

Matrice konfuzije (Slika 5) prikazuju rezultate poređenja automatski izdvojenih simila sa zlatnim standardom u dva režima evaluacije. Levi dijagram predstavlja strogi režim, u kojem se poklapanje priznaje samo ukoliko se u potpunosti podudaraju granice izdvojenog simila i njegove anotacije. Desni dijagram prikazuje fleksibilni režim, koji dozvoljava delimično preklapanje granica. Očekivano, fleksibilni režim rezultuje većim brojem tačno prepoznatih simila (514

naspram 440) i manjim brojem lažno negativnih slučajeva (133 naspram 207), dok se broj lažno pozitivnih slučajeva neznatno povećava (11 naspram 9). Rezultati potvrđuju da fleksibilni kriterijum bolje odražava praktične razlike u određivanju granica simila, posebno kod složenijih konstrukcija.

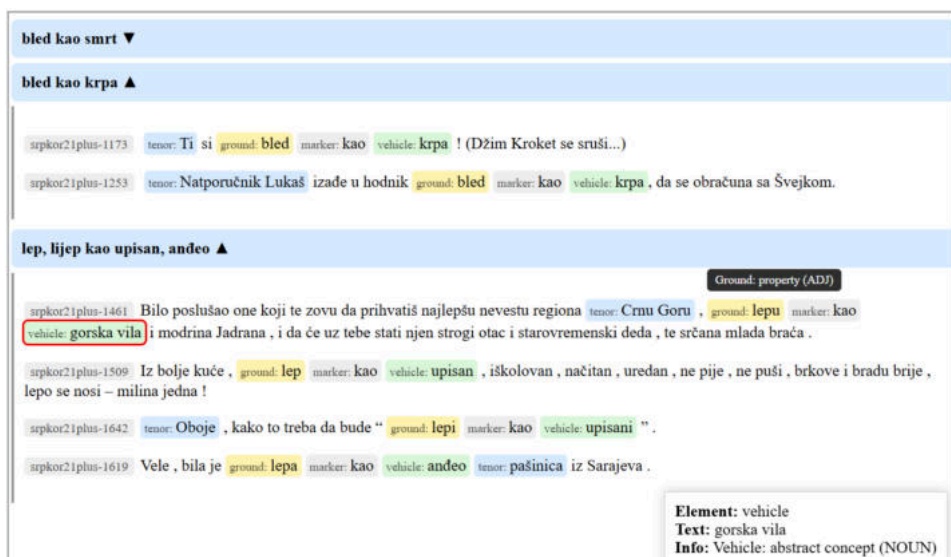


Slika 5: Matrice konfuzije dobijene evaluacijom u odnosu na zlatni standard za strogi i fleksibilni režim poređenja

Dobijeni rezultati ukazuju da je predložena metodologija ekstrakcije i validacije naročito efikasna u identifikaciji pouzdanih kandidata za simile. Razlike između strogog i fleksibilnog režima evaluacije pre svega odražavaju različit nivo detaljnosti anotacije i primenjenih kriterijuma poklapanja, a ne sistematske greške u klasifikaciji. Drugim rečima, većina neslaganja proizlazi iz različitih interpretativnih odluka i granica između kategorija, dok sposobnost sistema da razlikuje simile od lažnih simila ostaje na visokom nivou u oba evaluaciona režima.

3.4. Interfejs repozitorijuma i vizuelizacija anotacija

U Repozitorijumu simila srpskog jezika (SSR) simili su organizovani prema svojim lematizovanim kanonskim oblicima i prikazani u vidu sekcija koje se mogu proširivati i sažimati, čime se olakšava navigacija kroz velike kolekcije primera (Slika 6). Svaka sekcija sadrži korpusne potvrde upotrebe odgovarajućeg simila, zajedno sa struktuiranim semantičkim anotacijama generisanim tokom procesa ekstrakcije i validacije. Anotacioni sloj ističe osnovne komponente simila pomoću prebojenih oznaka, dok identifikatori korpusa i reference na rečenice omogućavaju praćenje porekla podataka i njihovo povezivanje sa izvornim korpusnim resursima.



Slika 6: Interfejs za pretragu repozitorijuma simila

Interfejs takođe uključuje kontekstualne opise i iskačuće informacione prozore (*tooltips*) koji pružaju dodatne lingvističke informacije o anotiranim elementima, kao što su gramatičke i semantičke kategorije, definicije komponenata simila ili dodatni metapodaci. Celokupni dizajn usmeren je ka preglednosti, jednostavnosti korišćenja i podršci istraživačkom radu. Kombinovanjem strukturiranih mašinski čitljivih anotacija i intuitivne HTML vizuelizacije, repozitorijum omogućava istovremeno korpusno istraživanje, leksikografsku obradu, razvoj jezičkih resursa i kvalitativnu validaciju figurativnih izraza.

Pored funkcije pretraživanja i pregleda, ovakav način organizacije podataka pruža osnovu za buduće proširenje repozitorijuma novim anotacionim slojevima, statističkim pregledima i alatima za analizu. Time SSR prerasta iz obične zbirke primera u istraživačku infrastrukturu namenjenu proučavanju simila, figurativnog jezika i šire semantičke organizacije savremenog srpskog jezika.

4. Lingvistička analiza mašinski potpomognute anotacije simila

U radu Marković i saradnika (2026) ustanovljeno je da obrasci ekstrakcije konstruisani za široko korpusno pretraživanje ne odgovaraju nužno različitim lingvističkim tipovima simila. Analiza je pokazala da se više obrazaca često može odnositi na iste tipove odn.

osnovne konstrukcije simila (npr, AkaoN i kaoNA, ili kaoPrepNV, VkaoPrepN i ADVkaoPrepN), kao i da se značenje poređenja često ostvaruje kroz distribuirane semantičke strukture, dok je tokom ručne evaluacije ustanovljeno da su pojedinim obrascima izdvojene vrlo heterogene grupe figurativnih konstrukcija. U istom istraživanju uočen je (ali nije dalje ispitivan) značajan broj elaboriranih, kreativnih (nekonvencionalnih), idiosinkratičnih i strukturno i semantički složenih simila, koji nisu bili neposredno anticipirani početnim dizajnom ekstrakcije. Polazeći od tih nalaza, u ovom radu usredsređujemo se upravo na navedene tipove simila, kao i na pitanja identifikacije i anotacije njihovih semantičko-sintaksičkih komponenata, sa posebnim osvrtom na granice anotacionih opsega (engl. *span*), implicitne ili neiskazane elemente, složene figurativne strukture i odnos između osnova poređenja (*ground*) i njegove jezičke realizacije (*property*). Stoga su u fokusu analize izazovi u anotaciji koji su uočeni tokom ručne validacije i korekcije automatske semantičke anotacije korpusa, kao i njihove šire metodološke i teorijske implikacije.

4.1. Osnov poređenja (*ground*) i njegova jezička realizacija (*property*)

Četvorodelna anotaciona shema strukturnih elemenata simila uvedena je, prvenstveno, iz praktičnih razloga. Analizirani korpus, kao i ranija istraživanja (Qadir et al., 2016) pokazuju da većina simila ne sadrži eksplicitno iskazan osnov poređenja. Konkretno, Kvadir i saradnici nalaze da u njihovom korpusu čak 92% primera pripada otvorenim similitima, kod kojih se osnov poređenja mora zaključivati iz drugih elemenata konstrukcije. Ukoliko bi se anotirala isključivo eksplicitno realizovana svojstva, veliki deo semantičke strukture simila ostao bi nepredstavljen. Stoga, razdvajanje osnova poređenja (*ground*) i njegove konkretne jezičke realizacije (*property*) omogućava beleženje konceptualne osnove poređenja i u slučajevima kada njena leksička realizacija izostaje.

U primenjenoj anotacionoj shemi etiketom „ground“ anotiran je konceptualni osnov poređenja, odnosno semantičko svojstvo koje omogućava uspostavljanje sličnosti između komparanduma i komparatuma (referenta i objekta poređenja), dok su ozakom „property“ anotirane njegove konkretne jezičke realizacije u okviru datog simila. Pored rešavanja ovog konkretnog praktičnog izazova u anotaciji,

uvodena distinkcija pokazala se i višestruko opravdanim analitičkim instrumentom.

- 1) Nije retka pojava da se isti osnov poređenja na konceptualnom planu realizuje različitim sinonimnim leksičkim jedinicama na jezičkom planu. Npr, u similima *jak/snažan/silan kao bik*, isto svojstvo, SNAGA, realizovano je različitim leksičkim jedinicama.
- 2) Dodatno, isti konceptualni sadržaj simila, kako su pokazale Parizoska i Filipović Petrović (2017), može se varirati kroz različite leksičke realizacije osnova poređenja, uključujući, neminovno, različite gramatičke klase reči.⁸ Npr, osnov poređenja BRZINA različito se realizuje u više simila, poput *brz/hitar/juriti/leteti/preleteti/izleteti/odleteti kao munja*.
- 3) Sa druge strane, osnovi poređenja iskazani istim leksičkim jedinicama (*property*), ne znače nužno i aktiviranje istog osnova poređenja (*ground*), zahvaljujući polisemiji i leksičkoj spojivosti. Npr., *mrtav kao fabrika u Pančevu* (NEAKTIVNOST, NEPRODUKTIVNOST, POTPUNO MIROVANJE) i *mrtva kao trula grana koja se odlomila i beskorisno crkava u travi* (BEŽIVOTNOST, IMOBILNOST, DEKOMPOZICIJA).
- 4) Na fenomen implicitnog osnova poređenja prethodno je ukazano u radu, te u tom smislu, a primenom uobičajenih postupaka začetih još u strukturalističkoj tradiciji, ovu realizaciju označavamo Ø, kao odsutstvo obeležja/svojstva, tj. formalne realizacije osnova poređenja.
- 5) Kod elaboriranih i idiosinkratičnih tipova simila osnov poređenja često nije eksplicitno iskazan, niti se može svesti na jednu leksičku jedinicu. On se u takvim slučajevima rekonstruiše iz interakcije više semantičkih elemenata raspoređenih kroz širu sintaksičku i figurativnu strukturu poređenja, zbog čega se može opisati jedino na nivou konceptualne osnove poređenja (*ground*).

Ovakvo razumevanje nije potpuno strano ni domaćoj frazeološkoj tradiciji. Mršević-Radović (1987) opisuje proces u kojem imenička komponenta poredbenog frazeologizma postepeno gubi svoje

⁸ Autorke ovaj fenomen posmatraju kroz teoriju konstruala Ronalda Lanakera. Koji god mehanizmi i procesi leže u osnovi ovog fenomena, sam fenomen je nesporan, iz ma kog teorijsko-metodološkog okvira se sagledavao. Činjenicu da se osnov poređenja može pojaviti samo kao implicitan objašnjavaju subjektivizacijom.

referencijalno značenje i svodi se na dominantnu osobinu koja motiviše poređenje. Tako se u izrazu *živeti kao grof*, imenica *grof* ne interpretira prvenstveno kao referent, već kao nosilac predstave o bogatstvu, privilegovanosti ili lagodnom načinu života. Posmatrano iz perspektive primenjene anotacione sheme, upravo ta dominantna osobina funkcionalno odgovara onome što se anotira oznakom „ground“, dok njene konkretne jezičke realizacije predstavljaju „property“.

Na kraju, imajući u vidu da je osnovni cilj konstruisanja SSR-a razvoj jezičkog resursa namenjenog različitim vrstama empirijskih istraživanja simila, razdvajanje mogućnosti pretrage prema semantičkom osnovu poređenja (*ground*) i njegovoj konkretnoj jezičkoj realizaciji (*property*) može pružiti značajne uvide u obrasce konceptualizacije i jezičke realizacije poređenja. Takva organizacija podataka omogućava istraživanje odnosa između različitih realizacija istog osnova poređenja, sinonimskih i bliskoznačnih simila, vidova grupisanja simila oko zajedničke konceptualne osnove, nezavisno od njihove konkretne leksičke realizacije, obrazaca varijabilnosti, kao i ispitivanje načina na koje se isti osnovi poređenja ostvaruju kroz različite leksičke i gramatičke obrasce, stepene konvencionalizovanosti i tipove figurativnog izražavanja.

4.2. Elaborirani, idiosinkratični i ironični simili

Tokom ručne evaluacije izdvojen je značajan broj elaboriranih/proširenih i izrazito kreativnih simila. Elaborirani simili naročito su bili zastupljeni u obrascima AkaoN i VkaoN, što je očekivano, imajući u vidu sintaksički potencijal imeničkih sintagmi za proširivanje koordinacijom, subordinacijom i različitim vrstama opisnih modifikacija.

U mnogim slučajevima komparatum prerasta u semantički bogatu i strukturno složenu figurativnu scenu. Neki od ovih elaboriranih simila zasnivaju se na relativno konvencionalnim ili univerzalno razumljivim slikama, kao što su: *crven kao karanfil na snegu*; *beo kao prvi blesak mladoga Meseca*; *svetlucati kao kroz izgažen tepih* ili *sijati kao Sunce koje nikad ne zalazi*. Drugi primeri oslanjaju se na kulturno uslovljene predstave i simboličke asocijacije, kao u izrazima: *lep kao pun Mesec četrnaeste noći*; *jarki i bleskovit kao ramazanski polumesec*; *radovati se kao dete Božiću*. Ovakvi primeri pokazuju da interpretacija simila često zavisi od kulturnih znanja i simboličkih sistema koji prevazilaze konkretna značenja pojedinačnih leksema.

U nekim kreativnim figurativnim konstrukcijama poređenje nije zasnovano na jasno izdvojenoj osobini niti na odnosu između dva eksplicitno iskazana entiteta, već na analogiji između složenih scena. U primerima poput: *kao kroz protočni bojler prošla ova diskusija* ili *esnafski seminar na kome se susreću stručnjaci da bi sa svojim kolegama razmenili iskustva, potpuno nalik na rođake koji se viđaju samo na svadbama i sahranama*, interpretacija se zasniva na rekonstrukciji šire figurativne predstave i niza inferencijalnih veza sa iskustvenim i kulturnim znanjima govornika.

Primenjena metodologija ekstrakcije i validacije pokazala se sposobnom da identifikuje i veliki broj izrazito idiosinkratičnih i autorski specifičnih simila, čija interpretacija zavisi od složenih imaginativnih asocijacija, a ne od ustaljenih obrazaca poređenja. Posebno ilustrativni primeri uključuju konstrukcije kao što su: *raznovrstan kao užasi pakla koji ga je začeo*, *bezosećajan kao komad ćeramide na kiši*, *mrtva kao trula grana koja se odlomila i beskorisno crkava u travi*, *izliti se kao potok vrijuće lave pri erupciji Vezuva* ili *zaspali kao ljubavnik koga snena i nasmejana dragana odbija*. U ovakvim slučajevima značenje poređenja ne proizlazi iz jedne dominantne zajedničke osobine, već iz čitave mreže deskriptivnih, emocionalnih i simboličkih elemenata koji zajedno grade figurativnu interpretaciju.

Ručna evaluacija pokazala je i da je model često uspešno prepoznavao veoma kreativna i strukturno složena figurativna poređenja. Posebno su zanimljivi primeri koji uključuju višestruke koordinirane i hijerarhijski ugnježdene simile, kao što su sledeći:

Mada nije imao ljudsko lice, činio mi se lep kao dva duga vlakna-sta pipka nekog insekta, ili i više, kao naglo sahranjivanje, i još više kao zakon obnavljanja osakaćenih organa, i najlepše kao tečnost prevashodno podložna truljenju.

Ona je izgledala mlitavo i blede poput raskiseljene slame, poput izbeljenih kostiju malih ptica, kao povrće koje je predugo kuvano, blede i končasto, gnjecavo – jedva nalik na samu sebe, užasno nalik na leš i gotovo potpuno lišeno vlastitog mirisa.

U primerima poput ovih koordiniraju se i subordiniraju poredbene strukture (neke i doslovne!), pri čemu se figurativna scena postepeno razvija, menja i semantički destabilizuje kroz niz uzastopnih poređenja. Značenje se ne zasniva na jednoj jasno izdvojenoj zajedničkoj osobini, već nastaje kroz progresivno nizanje različitih imaginativnih asocijacija. Ovakvi primeri dodatno potvrđuju zapažanje ([Hanks](#),

2004) da se simili često zasnivaju na neodređenim, dinamičnim i imaginativnim semantičkim asocijacijama, a ne na strogo ograničenim i jasno definisanim strukturama poređenja.

Ručnom analizom izdvojena je grupa ironičnih i evaluativno invertovanih simila. U nekim slučajevima, i to vrlo kompleksnim, model je uspeo da prepozna ironiju, kao u prethodno navedenom elaboriranom idiosinkratičkom similu: *lep kao dva duga vlaknasta pipka nekog insekta...*, što ukazuje na to da je razvijeni anotacioni sistem potpomoćan veštačkom inteligencijom sposoban ne samo da identifikuje strukturno složene simile već i da registruje određene više pragmatičke i evaluativne dimenzije figurativnog značenja. Komentar koji je model dao glasi: *Reč „lep“ se ironično i figurativno poredi sa nečim što nije doslovno lepo, uz eksplicitni marker „kao“.*

Ipak, ručnom evaluacijom ustanovljeno je da ironični i pragmatički invertovani simili i dalje predstavljaju jedan od najvećih izazova za automatsku interpretaciju. Iako je model u pojedinim slučajevima uspešno identifikovao ironičnu figurativnu nameru, čak i kod veoma složenih i semantički nestabilnih poređenja, kriterijumi na osnovu kojih dolazi do takvog prepoznavanja još uvek nisu dovoljno jasni i zahtevaju dodatna istraživanja. U većini slučajeva, međutim, ironija i evaluativna inverzija nisu uspešno prepoznate. Na primer, konstrukcije kao što su: *Virdžinijski veliki knez, lep kao studija o luku koji opisuje pas idući za svojim gospodarem, lep kao drhtanje ruku u alkoholizmu* formalno odgovaraju prototipskim pozitivno vrednovanim similiba zbog prisustva prideva „lep“, ali njihova stvarna figurativna interpretacija zasniva se na ironiji, tj. kontradiktornosti i obrnutom vrednosnom efektu. U ovim primerima pozitivna evaluacija eksplicitno iskazana na površinskom nivou dolazi u sukob sa negativnim ili apsurdnim karakteristikama komparatuma, pri čemu se željeni efekat ostvaruje upravo kroz tu kontradikciju. Takvi slučajevi pokazuju da automatsko prepoznavanje simila ostaje naročito zahtevno kada figurativno značenje ne zavisi samo od leksičkog poređenja između dva entiteta već i od šire pragmatičke, stilističke i kontekstualne interpretacije. Drugim rečima, uspešna analiza ovakvih primera zahteva modelovanje implicitnih evaluativnih stavova, kulturnih znanja, ironičnih namera i diskursnih očekivanja, što predstavlja jedan od najizazovnijih zadataka u obradi figurativnog jezika.

4.3. Semantičke komponente simila: eksplicitnost, implicitnost i distribuiranost

Prethodna analiza pokazala je da najčešće ne postoji direktno i jednoznačno prepisivanje sintaksičkih i semantičkih kategorija simila, kao i da se komponente simila često realizuju na nivou sintagme ili šire konstrukcije, a ne kroz diskretne leksičke elemente (Marković et al., 2026). Ova pojava naročito dolazi do izražaja kod elaboriranih i idiosinkratičnih simila, gde su semantičke komponente često raspoređene u složenije sintaksičke strukture.

Takođe, utvrđeno je i to da identifikacija osnova poređenja predstavlja složeniji zadatak od identifikacije komparanduma (Marković et al., 2026). Kod pridevskih simila osnov poređenja je u najvećem broju slučajeva bio pravilno prepoznat, uključujući i odgovarajući morfosintaksički oblik, naročito kada je bio eksplicitno izražen u tekstu. Međutim, model je pokazivao poteškoće u slučajevima kada je poređenje uključivalo dve koordinirane osobine. Na primer, u rečenici: *Ko iskreno i strasno ljubi Istinu, Slobodu i Otadžbinu, slobodan je i neustrašiv kao Bog, a prezren i gladan kao pas ili grof slučajno dodirnu tananu maramicu na kojoj su kapi krvi bile vlažne i crvene kao karanfil na snegu*, model je kao leksikalizovani osnov poređenja prepoznavao samo prideve koji neposredno prethode komparatoru – *neustrašiv, gladan, crvene*, dok su *slobodan, prezren, vlažne* ostali neidentifikovani.

Jedan od čestih izazova u anotaciji odnosio se na slučajeve u kojima osnov poređenja nije eksplicitno iskazan. U primeru: *predstave su se ocrtavale pred mojim očima kao žive slike*, model je kao osnov poređenja izdvojio atribut kao deo komparatuma (*žive* od *žive slike*). Osnov poređenja u navedenom primeru prisutan je samo implicitno, i jednostavno je učitati ga: prilog *jasno* ili neki njemu blizak. Isto važi i za primer: *A onda kao iz vedra neba se kaže...*, s neiskazanim osnovom poređenja, koji bi mogao biti prilog *iznenada*. U nekim slučajevima, model je kao osnov poređenja izdvajao sam poredbeni deo: *Kasnimo nekoliko meseci baš iz razloga što nismo želeli da kao na traci isporučimo zakon Narodnoj skupštini...*, dakle, model je izdvojio *kao na traci* (ovde bi osnov poredjenja bio neki od priloga *brzo, mehanički...*). Ovakvi primeri ukazuju na to da model ponekad teško razlikuje deskriptivne elemente komparatuma od osobina koje zaista predstavljaju osnov poređenja.

Dobijeni rezultati ukazuju na to da simili istovremeno funkcionišu na više reprezentacionih nivoa i da njihova anotacija ne može uvek da se zasniva na jednostavnom preslikavanju gramatičkih kategorija u semantičke komponente. U mnogim slučajevima figurativno značenje proizlazi iz interakcije više jezičkih i konceptualnih elemenata raspoređenih kroz širu sintaksičku strukturu. Ovakva zapažanja otvaraju važna pitanja za buduća istraživanja odnosa između sintaksičke organizacije, semantičke strukture i anotacije figurativnog jezika. Sa stanovišta računarske obrade jezika, rezultati pokazuju da interpretacija simila često zavisi od distribuiranog semantičkog zaključivanja i kontekstualne inferencije, a ne samo od lokalno prepoznatljivih leksičkih korespondencija između pojedinačnih elemenata poređenja.

4.4. Izazovi u identifikaciji komparanduma

Prethodna analiza (Marković et al., 2026) pokazala je da identifikacija komparanduma predstavlja jedan od složenijih zadataka u automatskoj anotaciji simila, naročito u slučajevima kada se poređenje ostvaruje kroz glagolske konstrukcije ili kada je komparandum implicitno izražen. Jedan od izazova u automatskoj anotaciji odnosio se na identifikaciju komparanduma koji nije eksplicitno izražen, naročito u slučajevima karakterističnim za srpski kao pro-drop jezik. Model je često uspešno rekonstruisao neizražene učesnike na osnovu glagolske morfologije, ali su se istovremeno javljali problemi povezani sa izborom odgovarajućeg kandidata za komparandum i određivanjem njegovih granica. Dalje je ustanovljeno da identifikacija komparanduma ne zavisi isključivo od lokalne leksičke strukture iskaza, već od interakcije glagolske morfologije, linearne organizacije rečenice i šire predikacione strukture. Kao posledica toga, granice između pojedinih komponentata simila nisu uvek bile jednoznačno određene, što je u pojedinim slučajevima dovodilo do nestabilnih granica anotacionih opsega.

Dodatni izvor nestabilnosti javljao se u slučajevima kada je model delimično rekonstruisao širu predikacionu strukturu povezanu sa poređenjem, ali pritom nije uspevao da pravilno odredi granice samog komparanduma. Posebno ilustrativan primer predstavlja rečenica: *Stvorili ste kod građana uverenje da u svojim redovima imate toliko i takve eksperte da ćemo, maltene, ako počnemo da ih izvozimo drugim državama, samo od toga živeti kao bubreg u loju*. U ovom slučaju model je kao komparandum identifikovao pomoćni glagolski oblik

ćemo. Ovakvi primeri ukazuju na to da identifikacija komparanduma nije bila zasnovana isključivo na linerarnoj blizini, već da je uključena i delimična rekonstrukcija glagolske i predikacione strukture koja okružuje izraz *simila*. Međutim, ta rekonstrukcija nije uvek bila potpuna, što je povremeno dovodilo do pogrešnog izdvajanja sintaksičkih elemenata umesto semantički relevantnih učesnika poređenja.

Ručnom analizom takođe je utvrđeno da je model često proširivao opseg komparanduma izvan samog učesnika poređenja, uključujući šire predikacione strukture. U primerima kao što su: (...) *pominjao skupe poklone koje dobijate, putovanja privatnim avionima, omogućavanja da mnogi privatnici danas žive kao bubreg u loju ili ...*, model je često kao komparandum izdvajao šire segmente poput *mnogi privatnici danas žive* ili *svi da živimo*, umesto da izoluje samo subjekat poređenja. Slično proširivanje opsega uočeno je i u primerima poput: *oni danas pište kao crv u kamenu*, gde je izdvojeni komparandum obuhvatio čitavu predikatsku rečenicu: *oni danas pište*. Model često izvrši tačnu semantičku anotaciju, ali ne pripiše odgovarajući oblik, ili proširi granice anotacionog opsega. Ovo implicira da je model uspešniji u zadacima identifikacije referenata nego u dodeljivanju odgovarajuće formalne reprezentacije u okviru anotacione sheme, što predstavlja dodatni izazov za izgradnju konzistentnih i standardizovanih anotacionih resursa.

4.5. Ostali izazovi anotacije potpomognute VJM

Tokom postupka ručne evaluacije ustanovljeno je da veliki broj formalno sličnih konstrukcija izdvojenih široko postavljenim obrascima ekstrakcije zapravo ne predstavlja simile. Među njima se nalaze diskursni markeri, doslovna poređenja, tehnička poređenja i različiti ustaljeni izrazi, kao što su: *kao na primer*, *kao u vreme*, *kao kod X*, *kao za početak*. Većina ovih instancija uspešno je prepoznata tokom procesa validacije i klasifikovana kao negativni kandidat u fazi filtriranja.

Ipak, ručnom evaluacijom utvrđeno je više ponavljajućih obrazaca u kojima je model klasifikovao doslovna ili kontekstualno motivisana poređenja kao simile. Ovi lažno pozitivni primeri najčešće su uključivali istorijska, diskursno-organizaciona ili situaciona poređenja. Na primer, konstrukcije kao što su: *pojedinci denunciraju kao u slavna doba čuvene crveno-crne koalicije*, ili *osvrtali su se na mikiensku prošlost kao na zlatno doba bogova i junaka*, sadrže eksplicitno

izraženu poredbenu strukturu, ali poređenje u njima ostaje prvenstveno referencijalno ili istorijski kontekstualizovano, a ne figurativno u smislu prototipičnog simila. Slične dvosmislenosti uočene su i u primerima: *...kada smo pozvani da nam se kao na lep način kaže da napustimo zgradu opštine* ili *...i mališan kao u odgovor zagrokta*, gde marker *kao* uvodi aproksimaciju, interpretaciju ili situaciono usklađivanje, a ne tipično figurativno poređenje. Ovi nalazi ukazuju na to da granica između simila, doslovnih poređenja, formulaičnih izraza i kontekstualnih analogija često nije oštra, već postepena, naročito u konstrukcijama koje uključuju veoma produktivne obrasce poređenja.

Posebno zanimljivu grupu predstavljaju primeri koji sadrže više od jednog simila u istoj rečenici, npr.: *(...) tvrdeći da je tobože cela njegova sudbina, kao i sudbina njegove porodice upisana u te mrlje, jasno kao na dlanu, neopozivo kao u svitak svete Tore*. U takvim slučajevima model je često uspevao da identifikuje samo prvi simili, dok su naredni ostajali neprepoznati. Na primer, u rečenici: *Bilo ih je četiri stotine mladica, sve kćeri efritske i maredske, obučene poput rascvetanih cvasti, a lake kao ptice*, sadržana su dva poređenja: *obučene poput rascvetanih cvasti* i *lake kao ptice*. Međutim, model je identifikovao samo prvo poređenje. Ovakvi slučajevi ukazuju da trenutni pristup ima ograničenja u obradi višestrukih simila unutar iste rečenice, naročito kada su oni koordinirani ili raspoređeni u složeniju sintaksičku strukturu.

5. Zaključak

U radu je predstavljena izgradnja *Repozitorijuma simila savremenog srpskog jezika* (SSR), sa posebnim fokusom na prikaz anotacionog modela i rezultate evaluacije mašinski potpomognute anotacije razvijene u svrhu konstruisanja ovog repozitorijuma.⁹ Analiza je pokazala da uspešna identifikacija i anotacija simila zahteva uvažavanje složenih odnosa između sintaksičke strukture, semantičke interpretacije i implicitno realizovanih elemenata poređenja, što predstavlja dodatne izazove za automatsku anotaciju. Dobijeni rezultati potvrđuju značaj hibridnog pristupa koji objedinjuje simboličke i statističke pristupe i primenu VJM sa lingvističkom ekspertizom ugrađenom u anotacionu shemu, evaluaciju i iterativno unapređenje modela. Oni istovremeno potvrđuju da razvoj resursa za obradu figurativnog jezika

⁹ Repoziotorijum će biti javno dostupan sa finalnom verzijom rada.

predstavlja važnu tačku susreta lingvističkih i računarskih pristupa proučavanju značenja.

Kao pravac daljeg razvoja predviđeno je proširenje metodologije na automatsko prepoznavanje drugih stilskih figura koje, za razliku od simila, nemaju jasno definisanu sintaksičku strukturu niti pokazuju visok stepen produktivnosti. Kod takvih figura biće neophodno razviti složenije postupke ekstrakcije kandidata koji će, pre primene VJM, koristiti bogate leksičko-semantičke resurse za srpski jezik. Pored elektronskih rečnika sa semantičkim markerima, Srpskog vordneta i Vikipodataka, takav pristup može uključiti ontologije i grafove znanja, korpuse sa semantičkom anotacijom, distribucione semantičke modele, kao i tehnike razrešavanja značenja reči (*word sense disambiguation*, WSD). Kombinovanjem simboličkih resursa i metoda zasnovanih na velikim jezičkim modelima moguće je značajno unaprediti izbor kandidata, smanjiti broj lažno pozitivnih rezultata i obezbediti pouzdaniju semantičku interpretaciju stilskih figura čija se identifikacija ne može zasnivati isključivo na površinskim sintaksičkim obrascima.

Zahvalnost

Ovo istraživanje je podržao Fond za nauku Republike Srbije u okviru projekta br. 7276, Text Embeddings – Serbian Language Applications“ (TESLA), COST ACTION CA21167 – UniDive i Ministarstvo nauke Republike Srbije (451-03-33/2026-03/200174).

Literatura

- Svetlana Slijepčević Bjelivuk, editor. 2025. *Rečnik novijih lingvističkih termina*. Prometej / Institut za srpski jezik SANU, Novi Sad / Beograd.
- Svetlana Goljak. 2009. Устаљена поређења у српском и белоруском језику са аспекта идиоматичности. *Научни састанак слависта у Вукове дане*, 38(1):211–220.
- Patrick Hanks. 2004. *Similes and Sets*. In Renata Blatná and Vladimír Petkevič, editors, *Jazyky a jazykověda (Languages and Linguistics: Festschrift for Professor Fr. Čermák)*, pages 1–15. Philosophy Faculty of Charles University, Prague.

- Duška Klikovac. 2004. *Metafore u mišljenju i jeziku*. Čigoja štampa, Beograd.
- Cvetana Krstev. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade, Belgrade.
- Cvetana Krstev. 2021. [White as Snow, Black as Night – Similes in Old Serbian Literary Texts](#). *Infotheca*, 21(2):119–136.
- Cvetana Krstev, Jelena Jaćimović, and Duško Vitas. 2020. [Analysis of Similes in Serbian Literary Texts \(1860–1920\) Using Computational Methods](#). In *Proceedings of CLIB 2020*.
- Cvetana Krstev, Gordana Pavlović-Lažetić, and Ivan Obradović. 2004. [Using Textual and Lexical Resources in Developing Serbian Wordnet](#). *Romanian Journal of Information Science and Technology*, 7(1-2):147–161.
- Cvetana Krstev, Sandra Ranković, Aleksandra Marković, and Tijana Mihajlov. 2024. [Towards the Semantic Annotation of SR-ELEXIS Corpus: Insights into Multiword Expressions and Named Entities](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 106–114, Torino, Italy.
- Saša P. Marjanović. 2017. *Poredbene frazeme s komponentom come/kao u francuskom i srpskom jeziku*. Doktorska disertacija, Univerzitet u Beogradu, Filološki fakultet, Beograd, Srbija.
- Aleksandra Marković, Cvetana Krstev, Milica Dinić Marinković, and Ranka Stanković. 2026. [Towards a Repository of Similes in Contemporary Serbian](#). In *Proceedings of the XXII EURALEX International Congress 2026: Lexicography in the Age of AI*, pages 295–309, Vienna. Austrian Academy of Sciences.
- Emma Krane Mathisen, Nicholas Allott, and Camilo R. Ronderos. 2025. [Cognitive Mechanisms in Simile and Metaphor Comprehension](#). *Language and Cognition*, 17:e89.
- Jasna Melvinger. 1983. [Poredbeni frazemi](#). *Jezik*, 31(5):129–135.
- Nikola Milošević and Goran Nenadić. 2016. [As Cool as a Cucumber: Towards a Corpus of Contemporary Similes in Serbian](#). *arXiv preprint*.

- Nikola Milošević and Goran Nenadić. 2018. [Creating a Contemporary Corpus of Similes in Serbian by Using Natural Language Processing](#). *arXiv preprint arXiv:1811.10422*.
- Jelena Mitrović, Stella Markantonatou, and Cvetana Krstev. 2020. [A Cross-Linguistic Study on Greek and Serbian Fixed Similes and Enrichment of Lexical Resources via Crowdsourcing](#). In Stella Markantonatou and Anastasia Christofidou, editors, *Multiword Expressions: Drawing on Data from Modern Greek and Other Languages*, pages 241–262. Research Center for Scientific Terms and Neologisms.
- Suzanne Mpouli. 2017. [Annotating Similes in Literary Texts](#). In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 26–36.
- Dragana Mršević-Radović. 1987. *Frazeološke glagolsko-imeničke sin-tagme u savremenom srpskohrvatskom jeziku*. Number 60 in *Monografije Filološkog fakulteta*. Filološki fakultet Beogradskog univerziteta, Beograd. 163 str.
- Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. [Brighter than Gold: Figurative Language in User-Generated Comparisons](#). In *Proceedings of EMNLP 2014*, pages 2008–2018, Doha, Qatar. Association for Computational Linguistics.
- Jelena Parizoska and Ivana Filipović Petrović. 2017. [Variation of Adjectival Slots in kao \('as'\) Similes in Croatian: A Cognitive Linguistic Account](#). In Ruslan Mitkov, editor, *International Conference on Computational and Corpus-Based Phraseology*, pages 348–362. Springer, Cham.
- Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2016. [Automatically Inferring Implicit Properties in Similes](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1223–1232, San Diego, California. Association for Computational Linguistics.
- Rikard Simeon. 1969. *Enciklopedijski rječnik lingvističkih naziva: na 8 jezika, hrvatsko-srpski, latinski, ruski, njemački, engleski, francuski, talijanski, španjolski*. Matica hrvatska, Zagreb.
- Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. [O familiji korpusa savremenog srpskog jezika SrpKor](#). In Jasmina Mo-

skovljević Popović and Ranka Stanković, editors, *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig*, volume 1, pages 3–27. University of Belgrade – Faculty of Philology.

Nataša Vulović. 2015. *Srpska frazeologija i religija: Lingvokulturološka istraživanja*. Institut za srpski jezik SANU, Beograd.

The Repository of Similes in Contemporary Serbian: Construction and Evaluation of Machine-Assisted Annotation

Milica Dinić Marinković, Cvetana Krstev, Aleksandra Marković, Ranka Stanković

Summary

This paper evaluates the methodology developed for constructing the *Similes Repository for Contemporary Serbian* (SSR), created within a broader research project to build the first corpus-based language resource for the systematic study of similes in Serbian. SSR integrates corpus extraction, semantic annotation, linguistic validation, and structured data organization into a unified, open, and extensible resource.

The paper presents the complete processing pipeline, from corpus-based candidate extraction and rule-based filtering to automatic semantic annotation and manual validation. SSR is based on several complementary corpora of Contemporary Serbian. Automatic processing relies on a hybrid neuro-symbolic approach that combines corpus queries and local grammars implemented as FSTs in Unitex, supported by electronic dictionaries of Serbian, with GPT-5.2 and Gemini LLMs for automatic simile identification and semantic annotation.

Conceived as a dynamic resource for linguistic, lexicographic, and NLP research, as well as a foundation for the further development of automatic figurative language processing in Serbian, the paper focuses on the annotation model, linguistic evaluation of machine-assisted annotation, and implications for further improving the proposed framework.

The evaluation demonstrates that combining symbolic methods with LLMs achieves high precision in simile identification while maintaining satisfactory agreement with expert annotators. At the same time, the analysis shows that successful simile identification and annotation requires accounting for complex interactions between syntactic structure, semantic interpretation, and implicit comparison components. The results confirm the effectiveness of a hybrid methodology that integrates symbolic methods, LLMs, and linguistic expertise throughout annotation design, evaluation, and iterative refinement of the framework.

Keywords: Similes, Simile Repository, Serbian Language, Corpus Linguistics, Semantic Annotation, Lexical Resources, Figurative Language, Unitex, Large Language Models.

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

