DOI: https://doi.org/10.18485/philologia.2024.22.22.1
UDC: 811.163.41'373.421
811.163.41'37

DESINONIMIZACIJA KROZ PRIZMU KORPUSNE I VEKTORSKE ANALIZE KONTEKSTUALNIH PREFERENCI LEKSEMA *KOMPJUTER* I *RAČUNAR*

Matija Nešović[1][2]
Institut za srpski jezik SANU
Beograd, Srbija
https://orcid.org/0009-0004-5822-6675

Koristeći elektronski korpus *Serbian Web Corpus PDRS 1.0*, kao i računarski alat za pretvaranje reči-tokena u vektore *word2vec*, u ovom radu smo analizirali primere upotrebe leksema *kompjuter* i *računar* kako bismo utvrdili stepen njihove semantičke bliskosti. Dobijeni rezultati pokazuju da pomenute imenice preferiraju načelno različita kontekstualna okruženja, zbog čega zastupamo tezu da ih ne treba smatrati istoznačnicama u savremenom srpskom jeziku. Sprovedeno istraživanje ilustruje mogućnost primene raznovrsnih kompjuterskih resursa prilikom rešavanja kako praktičnih, tako i teorijskih lingvističkih pitanja.

Ključne reči: *kompjuter*, *računar*, sinonimi, desinonimizacija, distribuciona semantika, kontekst, korpus, word2vec.

1. UVODNA RAZMATRANJA

Interesovanje za formalne i sadržinske odnose među rečima (uključujući i sinonimiju) postojalo je i mnogo pre formiranja lingvistike kao naučne discipline. Ovim pitanjima bavili su se, između ostalog, i antički filozofi, uglavnom nastojeći da preko jezičkih reprezentacija prodru u suštinu odnosa između jezika i stvarnosti (v. Ivić 1970: 12–14; Kotzia/Chriti 2013: 127–128). Osim toga, leksičke relacije čine okosnicu velikog broja retorskih figura (poput tautologije, antiteze itd.), zbog čega je njihov praktični potencijal bio prepoznat od samih početaka besedništva (isp. Herrick 1996: 41).

Razvoj leksikografske prakse, koji je posebno obeležio epohu prosvećenosti, nametao je potrebu za izučavanjem značenjskih odnosa među leksemama (v. Sudilovskaya 2018: 5). Ipak, data pitanja počinju da se razmatraju na čvrstim naučnim temeljima tek sa pojavom strukturalizma. Jedan od pionira leksičke paradigmatike bio je nemački naučnik Jost Trir, koji je u lingvistiku uveo pojam leksičkog polja. U pitanju

1 Kontakt podaci (E-mail): matija.nesovic@isj.sanu.ac.rs
2 Ovaj rad finansiralo je Ministarstvo nauke, tehnološkog razvoja i inovacija Republike Srbije prema Ugovoru broj 451-03-47/2023-01 od 17. 1. 2023. godine, koji je sklopljen sa Institutom za srpski jezik SANU.

je apstraktna struktura koja objedinjuje različite reči u okviru istog pojmovnog domena. U skladu sa strukturalističkom paradigmom, elementi datog skupa nisu samostalni, već su uzajamno zavisni; promene na nivou pojedinačnih entiteta neizostavno utiču na sistem u celini (isp. Kleparski/Rusinek 2007: 188–190). Evidentno je da je ovakav teorijski konstrukt (koji „okuplja" značenjski povezane leksičke jedinice) u tesnoj vezi sa semantičkim odnosima, i to naročito sa sinonimijom i hiponimijom. Razvojem učenja o semskoj strukturi i polisemiji istraživanja u ovom pravcu dobila su novu dimenziju; veliki doprinos na tom polju pružili su ruski semantičari (za kratak pregled v. Sudilovskaya 2018: 6–8), a kod nas eminentni lingvisti poput Milke Ivić, Irene Grickat Radulović, Darinke Gortan Premk, Stane Ristić, Rajne Dragićević i dr., zahvaljujući čijim zalaganjima je pod okriljem tzv. Beogradske leksikografske škole formiran tim vrsnih stručnjaka koji su na polju teorijske i praktične leksikologije postigli (i nastavljaju da postižu) rezultate vredne pažnje (v. Dragićević 2017). Koristeći se poglavito komponencijalnom analizom kao metodom, istraživači su sada na naučnoj osnovi mogli da ispituju tanane značenjske nijanse po kojima se, na primer, razlikuju relativni sinonimi.

Pored ove tradicionalne linije fundamentalnih istraživanja, na Zapadu se od sredine XX veka intenzivno razvija i formalna semantika, posebna interdisciplinarna oblast koja, ukrštajući temeljne lingvističke fenomene (poput ambigviteta, referencije i sl.) sa strogim metodološkim aparatom formalne logike, problematiku značenja prenosi sa verbalnog na rečenični plan, čime se fokus premešta sa paradigmatskog na sintagmatski nivo. Pored toga, kognitivna lingvistika, kao jedan od vodećih pravaca ne samo u lokalnim (američkim) već i u globalnim okvirima, iako okrenuta „atomizovanim" konceptima kao diskretnim strukturama (često obeležavanim pojedinačnim rečima) iza kojih se krije naša predstava o određenim fragmentima stvarnosti, počiva na obuhvatnom metodu semantičke analize. Tako, na primer, prilikom razmatranja koncepta LJUBAV istraživač mora, osim stožerne lekseme, uzeti u obzir i širi kontekst u kojem se ona upotrebljava (npr. *Leteo je na krilima ljubavi*), druge reči iz istog semantičkog polja (npr. *voleti*) i njihove sintagmatske realizacije, kao i primere u kojima centralna leksema izostaje, ali tematski korespondira sa rečeničnom situacijom (npr. *Razišli su im se putevi*). Takođe, pojam leksičkog značenja dovodi se u vezu i sa enciklopedijskim znanjem, a prilikom semantičkog opisa koriste se širi strukturni modeli (poput scenarija i frejmova) (v. Dragićević 2007: 92–93).

Intenzivan razvoj korpusne lingvistike postavio je nove standarde istraživanja na polju leksičke semantike. Danas je u svetskoj nauci gotovo nezamislivo da se o značenjskim fenomenima govori bez pokrića u obimnoj i kvantitativno obrađenoj građi. Zaključci se donose na osnovu praćenja leksičke jedinice „u pogonu", a ne na osnovu njenih sistemskih obeležja i relacija. Uopšte uzev, može se reći da savremenu lingvističku misao odlikuje izrazit empirizam i usmerenost na jezičku upotrebu. S tim u vezi, sve više su u fokusu kontekstualna svojstva leksema, a sve manje njihove apstraktne semantičke crte i paradigmatski odnosi.

1.1. RAĐANJE DISTRIBUCIONE SEMANTIKE: ŠTA JE (I GDE JE) ZNAČENJE?

Kao što smo iz navedenog kratkog prikaza videli, stručnjaci su se kroz istoriju bavili različitim aspektima semantičke problematike. Ipak, nezavisno od konkretnog istraživačkog cilja (besednička praksa, pisanje rečnika ili pak teorijsko utemeljenje

koncepta semantičkog polja, te razmatranje logičke strukture iskaza), sve predstavljene pravce objedinjuje neizostavna usmerenost na značenjske odnose među leksemama na paradigmatskoj i (posebno u novije vreme) sintagmatskoj ravni. Naravno, takav pristup nameće i potrebu za omeđivanjem samog pojma leksičkog značenja.

Značenje, kao centralno pitanje leksikologije, tradicionalno se određuje kao odnos između jezika, mišljenja i stvarnosti (v. Dragićević 2007: 54). Ovakav model, prema kojem je prostor značenja, metaforički rečeno, „oivičen" (i određen) stranicama čuvenog Ogden–Ričardsovog trougla, omogućava da se efikasno sagledaju neke od ključnih distinkcija u okviru leksičke semantike, poput one između denotacije i designacije. Međutim, iako je dati pristup čvrsto teorijski utemeljen, njegov praktični potencijal je upitan.[3]

Kroz istoriju lingvistike u više navrata su (de)fokusirani neki od ponuđenih triju parametara; tako, na primer, De Sosir naglašava da „označeno" predstavlja konceptualnu projekciju stvarnosti koja je u našem mentalnom prostoru povezana sa psihičkim otiskom zvuka (isp. De Saussure 2011 [1916]: 12), dok se za formalne semantičare značenje nalazi u spoljnom svetu (v. Portner 2005: 11–12).

Treća mogućnost, prema kojoj bi se leksičko značenje crpilo samo iz jezika, postulirana je, između ostalog, u okviru tzv. distribucione semantike. Naime, ova disciplina, koja se naročito intenzivno razvija od početka XXI veka, počiva na pretpostavci da je značenje reči moguće odrediti prema tome u kakvom se kontekstu ona javlja. Ova hipoteza može se slikovito parafrazirati naširoko citiranim rečima britanskog lingviste Dž. R. Ferta (Firth 1962 [1957]: 11): „You shall know a word by the company it keeps!" Korene ovakvog pristupa treba tražiti u američkom (harisovskom) distribucionalizmu, koji u izvesnom smislu predstavlja radikalnu nadgradnju strukturalizma – u njemu se, naime, raskida svaka veza sa mentalizmom, a kako bi se leksema valjano semantički analizirala, treba razmotriti njene (empirijski dostupne i proverljive) kontekstualne preference. Distribucioni semantičari odlaze korak dalje i, osim metodološke, predlažu i eksplanatornu (kognitivnu) dimenziju date postavke: linearna kombinatorika leksičke jedinice počinje da se tumači i kao srž njene semantike u svesti govornikâ; drugim rečima, njena sintagmatika *postaje* i njena paradigmatika (naročito u „jačoj" verziji ove teorije) (isp. Lenci 2008: 4–7). Naravno, ovako smela teza dočekana je u naučnim krugovima sa podozrenjem, budući da, pre svega, delimično ili potpuno zanemaruje čulno i motoričko iskustvo kao konstitutivne činioce značenja. Ipak, nezavisno od toga da li se okruženje u kojem se određena leksema javlja posmatra kao suština ili tek posledica njenog značenja (u vezi sa ovim pitanjem v. Lenci 2008), treba istaći da su distribucioni semantičari umnogome zaslužni za afirmaciju konteksta kao ključnog pojma moderne nauke o jeziku. Upravo su temeljne ideje razvijene u okviru ove lingvističke škole dale snažan zamah empirijski orijentisanim istraživanjima, kao i razvoju korpusne i računarske lingvistike (uključujući i vektorske modele, koji su predstavljeni u ovom radu).

1.2. KONTEKST I KOLOKATIVNOST KAO INDIKATORI LEKSIČKOG ZNAČENJA

S obzirom na to da metodološki imperativ distribucione semantike čini promatranje reči u kontekstu, i ne čudi što se sam termin *kolokacija* tradicionalno vezuje za ime Dž. R.

3 Ovo se naročito odnosi na primenjivost datog obrasca u računarstvu. Kako bi kompjuter uspeo da „dopre" do značenja, neophodno je da ono bude „opipljivo" (a ne internalizovano), numerički merljivo (a ne deskriptivno) i uporedivo sa ostalim članovima iste klase (a ne izolovano).

Ferta (v. Dražić 2014: 28). Kao što je poznato, ovaj pojam je u literaturi definisan na različite načine. To je i razumljivo ako se uzme u obzir potreba da se on razgraniči prema srodnim i isto tako nedovoljno omeđenim lingvističkim konceptima poput slobodnih leksičkih veza i idioma. T. Prćić (2016: 147–148) određuje kolokacije kao spojeve koje odlikuje moguća ili minimalna zamenljivost, relativno slaba postojanost (što podrazumeva izvesnu gramatičku „fleksibilnost") i (semantička) prozirnost. Na ovaj način se (nasuprot frazemima) potcrtava kompozicionalni karakter kolokacija, ali se istovremeno ističe i njihova delimična ustaljenost (naspram slobodnih leksičkih kombinacija).

Iako je, kao što je već predočeno, naučna relevantnost sintagmatskog jezičkog nivoa dovođena u pitanje u okviru određenih lingvističkih pravaca, to ne poništava činjenicu da kontekst (u najmanju ruku) stoji u očiglednoj korelaciji sa značenjskim svojstvima reči (isp. npr. Shmelev 1973: 160; Gortan-Premk 1997: 49; Dragićević 2007: 222–223). S tim u vezi, kolokaciona analiza predstavlja vrlo koristan metod za utvrđivanje i poređenje semantičkog sadržaja leksema. Ipak, kako primećuje Lj. Gotštajn (1986: 41), leksička kompatibilnost *in potentia* ne mora se uvek realizovati u vidu odgovarajuće kolokacije. Drukčije rečeno, „kolokacione sile" mogu određene spojeve preferirati, a druge (uprkos ispunjenosti semantičkih uslova) – odbacivati. Ovo bi, bar teoretski, moglo predstavljati metodološki problem prilikom semantičke analize, budući da, prema ovakvoj postavci, sintagmatika jedinice ne mora sasvim precizno odražavati njenu paradigmatiku. Na primer, iako su imenice *učenik* i *đak* nesumnjivo sinonimi, naporedna konstrukcija *učenik i učitelj* u građi je potvrđena daleko većim brojem primera od istoznačne konstrukcije *đak i učitelj*. Međutim, ako je korpus dovoljno reprezentativan, ovakva (ređa) razilaženja ne mogu značajno uticati na dobijene rezultate.

Takođe, treba istaći da se termin *kolokat* u ovom radu koristi u širem značenju nego što je to iznad precizirano, što je upravo u skladu sa ranije predstavljenom distribucionalističkom doktrinom. Naime, vektorski modeli (iz t. 3) obično su zasnovani na obradi znatno šireg konteksta (npr. u rasponu od -5 do +5), te „kolokatom" centralne lekseme nazivamo svaku reč sa kojom se ona udružuje u proizvoljno zadatom opsegu. Kada je reč o korpusnoj analizi (iz t. 2), dato određenje je u većini slučajeva opravdano i objektivnim svojstvima razmatranih spojeva: najrelevantniji među njima (npr. *laptop računar*, *tablet računar*, *prenosivi računar* i sl.) poseduju glavne karakteristike kolokacija (isp. Dražić 2014: 80).

1.3. SINONIMIJA I DESINONIMIZACIJA

U literaturi se, prema kriterijumu semantičke sličnosti, najčešće polazi od podele na dve osnovne vrste sinonima: apsolutne (istoznačnice) i relativne (bliskoznačnice) (isp. Dragićević 2007: 245). Istoznačnice se obično definišu kao reči koje su međusobno zamenljive u svim kontekstima (v. Cruse 1986: 268). U skladu s tim, test supstitucije pominje se kao jedan od glavnih metoda za utvrđivanje potpune značenjske podudarnosti među analiziranim leksemama. Ako pak navedeno određenje prevedemo sa sintagmatske na paradigmatsku ravan, možemo zaključiti da su posredi lekseme čije se polisemične (ili češće – monosemične) strukture u potpunosti poklapaju. S druge strane, bliskoznačnice se mogu razlikovati kako po opsegu polisemične strukture (*kuća* i *dom* dele samo neka značenja), tako i po određenoj komponenti konkretnog leksičkog značenja (designaciji,

konotaciji, upotrebnoj vrednosti i sl. – isp. graduelne sinonime *topao* i *vreo*) (up. Šipka 1998: 45–46). Nezavisno od toga koji je model relativne sinonimije posredi, njene posledice su uvek iste – zamenljivost zahvaćenih leksema u kontekstu nije potpuna, već ograničena.

U nauci vlada konsenzus da su apsolutni sinonimi u praksi izuzetno retki, budući da se njihovo postojanje protivi zakonu jezičke ekonomije. Osim toga, ako se u određenom sinhronom preseku i može ustanoviti potpuna značenjska ekvivalencija dveju (ili više) leksema, obično se smatra da će ona biti kratkog veka, pošto istoznačnice teže desinonimizaciji, tj. semantičkoj diferencijaciji. U pitanju je proces koji tipično podrazumeva sužavanje (specijalizaciju) značenja jednog od sinonima (up. Dragićević 2007: 248–249), usled čega se novouspostavljeni semantički odnos neretko transformiše u hiponimiju.

Ako prihvatimo ovde iznetu tezu o važnosti konteksta u proceni sinonimičnosti (uz nužne ograde koje su iznete u prethodnom poglavlju) i pokušamo na taj način da ustanovimo stepen značenjske sličnosti među određenim rečima, smatramo da je najbolje koristiti elektronske korpuse i druge računarske alate koji omogućavaju automatsku i obuhvatnu analizu distribucionih karakteristika leksema. Ukoliko je pažnja istraživača usmerena na semantičke promene koje su reči pretrpele u određenom vremenskom okviru (kao što je to delom slučaj u ovom radu), naročito su korisni korpusi koji dopuštaju filtriranje rezultata pretrage prema godini u kojoj su zabeleženi primeri.

Naposletku, kako uopšte nastaju sinonimi? Kao jedan od glavnih izvora leksičke sinonimije (osim stvaranja ekspresivne leksike, polisemije i derivacije) izdvaja se upravo pozajmljivanje – naporedo sa preuzetom leksemom u jeziku-primaocu obično već postoji ili se naknadno formira i domaća reč sa istim značenjem (v. Dragićević 2007: 249–250).

1.4. ZNAČENJSKI ODNOS ANGLICIZMA I ODGOVARAJUĆE DOMAĆE LEKSEME

Kako u ovom radu poredimo značenja lekseme preuzete iz engleskog jezika (*kompjuter*) i njenog prevodnog ekvivalenta (*računar*), potrebno je osvrnuti se i na problematiku semantičke adaptacije anglicizama u srpskom jeziku i njihovog odnosa prema odgovarajućim domaćim rečima.

Budući da nove predmete ili koncepte načelno usvajamo zasebno („jedan po jedan") i uniformno (sa jedinstvenim kriterijalnim obeležjima), anglicizme (i pozajmljenice uopšte) obično preuzimamo u svega jednom ili malom broju značenja (isp. Filipović 1986: 65).[4] Taj proces u jeziku-primaocu neretko prati i stvaranje značenjskog ekvivalenta od domaćeg leksičkog i tvorbenog materijala. Formirani leksemski par u početku može predstavljati apsolutne sinonime (ili tzv. trenutne „leksičke dublete"), ali, kao što smo istakli u prethodnom potpoglavlju, s vremenom među njima mora nastupiti semantička diferencijacija na denotativnom i(li) konotativnom (u najširem smislu) planu, što je mehanizam koji se tipično „prelama" na samoj pozajmljenici (isp. Milić 2013: 110–111).

Za predmet ovog rada naročito su važna razmatranja T. Prćića (2019: 148) o tzv. hiposinonimiji, koja podrazumeva hibridni semantički odnos između anglicizma i domaće lekseme: prema autoru, ovakve reči „istovremeno se nalaze u dvostrukom

4 Naravno, pozajmljenica u jeziku-primaocu sekundarno može razviti dodatna značenja, strana izvornoj leksemi (isp. Filipović 1986: 66; Nešović 2023).

odnosu: i sinonimije (zbog istovetnog osnovnog značenja) i hiponimije (zbog dodatnih obeležja anglicizma), pri čemu funkciju hiperonima vrši semantički opštija postojeća reč, dok funkciju hiponima vrši semantički specifičniji anglicizam (očigledni ili sirovi)". Prćić kao primere navodi parove *đus* prema *voćni sok*, *hamburger* prema *pljeskavica*, *dil* prema *dogovor* itd. Navedeno određenje hiposinonimije nije najjasnije, budući da nije precizirano šta se podrazumeva pod „istovetnim osnovnim značenjem": ta istovetnost primarnih značenja (barem u datim primerima) može figurirati samo u međujezičkoj relaciji izvorne engleske lekseme (npr. *deal*) i srpskog ekvivalenta (*dogovor*), ali se gubi u okvirima jezika-primaoca (*đus* označava samo sok od narandže). Ipak, čini se da je autor na pravom tragu kada sinonimiju i hiponimiju razmatra kao srodne, ispripletene i temeljne značenjske odnose u koje (simultano ili sukcesivno) stupaju pozajmljenica i odgovarajuća domaća leksema.

Najzad, od toga kako ćemo odrediti relaciju između anglicizma i njegovog srpskog ekvivalenta (kao apsolutnu sinonimiju, relativnu sinonimiju ili pak hiponimiju) umnogome zavisi i odgovor na pitanje da li nam je određeni anglicizam u jeziku uopšte potreban (up. Prćić 2019: 129–134).

1.5. PREDMET, METODOLOGIJA I CILJEVI RADA

Na tragu ideja distribucionih semantičara, u ovom članku želimo da pokažemo da se određene računarske metode mogu koristiti za procenu sinonimičnosti dveju ili više leksema. Centralni deo rada čine dva poglavlja: prvo ćemo, koristeći elektronski korpus (*PDRS 1.0*), sprovesti kolokacionu analizu reči *kompjuter* i *računar* (t. 2), a zatim ćemo pomoću odgovarajućeg modela za obradu teksta (*word2vec*) predstaviti i uporediti vektorske reprezentacije navedenih leksema (t. 3). Na kraju ćemo dobijene rezultate razmotriti iz teorijske vizure i izneti glavne zaključke istraživanja (t. 4).

Neposredni povod za pisanje ovog rada predstavlja sledeći pasaž iz udžbenika *Leksikologija srpskog jezika* Rajne Dragićević (2007: 248): „У садашњем часу, на пример, лексеме *компјутер* и *рачунар* представљају истозначнице у савременом српском језику [i RSJ ih tretira na isti način – M. N.]. То, међутим, вероватно, неће дуго трајати. Једна од њих ће или нестати или ће спецификовати значење." Želeli smo da utvrdimo da li se autorkina pretpostavka u međuvremenu obistinila, tj. da li su ove dve reči podlegle procesu desinonimizacije, što bi bilo u skladu sa uobičajenim shvatanjem sinonimije kao „jezičkog trenutka" i fenomena koji se suprotstavlja načelu jezičke ekonomije.

Dakle, cilj rada jeste da se ustanovi da li su lekseme *kompjuter* i *računar* istoznačnice. Još važnije od toga, pokušaćemo da skrenemo pažnju na neke od prednosti računarske obrade leksičkih odnosa.

2. KORPUSNA ANALIZA

Za potrebe ovog dela istraživanja korišćen je veb-korpus *Serbian Web Corpus PDRS 1.0* (izrađen u okviru projekta *Javni diskurs u Republici Srbiji*), koji sadrži više od 700 miliona tokena. Građa za ovaj korpus ekscerpirana je sa internet sajtova sa domenom *.rs*.

Pomoću jednostavnih upita ([lemma = 'računar'] i [lemma = 'kompjuter']) i opcije za izdvajanje najčešćih kolokata u opsegu -1, došli smo do leksema koje obično

prethode rečima *računar* i *kompjuter* (tim redom). Rezultati su predstavljeni u sledećim tabelama[5].

Collocate	Freq	Coll. freq.	logDice
desktop	755	3618	8.7738
PC	758	6104	8.7174
laptop	814	10524	8.716
personalan	716	4880	8.6654
tablet	562	5093	8.3107
prenosan	507	6678	8.1232
tvoj	1064	75327	8.0869
notebook	299	1016	7.5054
vaš	2334	507312	7.093
kvantni	234	3551	7.0855
prenosiv	188	5593	6.7185
disk	177	14857	6.4194
memorija	166	15264	6.3182
stoni	128	4902	6.181
kućni	191	40440	6.0671

Collocate	Freq	Coll. freq.	logDice
kvantni	80	3551	7.3157
personalan	66	4880	6.9261
board	43	539	6.7117
voice-input	31	31	6.2952
dt	27	445	6.0504
adresa	194	104100	5.7637
ekran	76	36459	5.6568
bord	21	1003	5.6288
ronilački	21	1334	5.5949
desktop	20	3618	5.3098
ispred	116	85024	5.2786
putni	53	32002	5.2733
pomoću	61	39585	5.251
tastatura	21	5658	5.2123
trip	15	1555	5.0872

Kao što se iz priloženih tabela vidi, minimalni levi kontekst lekseme *računar* (ako zanemarimo zamenice) uglavnom sačinjavaju reči koje preciznije određuju vrstu uređaja – prema frekvenciji dominiraju *laptop* (pored ređeg *notebook*), *PC*, *desktop* (uz znatno ređe *stoni*) i *personalan*, a odmah iza njih su *tablet* i *prenosan* (uz *prenosiv*). Dakle, ispostavlja se da leksema *računar* primarno označava bilo koji veći („ne-džepni") elektronski uređaj koji ima ekran i odašilje neki vizuelni signal, interaguje sa korisnikom i ispunjava određene „računarske" operacije (otvara, izvršava i zatvara programe i sl.).

Što se tiče lekseme *kompjuter*, njen najčešći levi kolokat predstavlja reč *adresa* (*Freq*: 194). Ređe se javljaju lekseme koje se odnose na vrstu kompjutera – *kvantni*, *personalan* (kalk prema engl. *personal computer*), *putni* / *(on) board* (oba predstavljaju kompjuterske sisteme u automobilima), te (u malom broju slučajeva) *ronilački*, *desktop*. Pored toga, činjenica je da većina navedenih atributa upućuje na gabaritne ili ugrađene (nemobilne) uređaje, što bi moglo da govori u prilog tvrdnji da leksema *kompjuter* tipično označava veliki, neprenosivi („stoni") računar koji se sastoji od monitora i kućišta i mora biti priključen u struju da bi radio. O tome posredno svedoči i izostanak kolokata kao što su *laptop*, *prenosni*, *tablet* i sl., kao i relativno mali ukupan broj anteponiranih leksema kvalifikativnog tipa, što se može objasniti tezom da je reč *kompjuter* značenjski već dovoljno specifikovana.

5 U svakoj koloni predstavljen je jedan parametar analize: apsolutna frekvencija pojavljivanja datog kolokata na prvoj poziciji s leve strane (*Freq*), apsolutna frekvencija pojavljivanja datog kolokata na bilo kojoj poziciji (ne samo prvoj levoj) u okviru istog primera (*Coll. Freq.*), kolokacioni „potencijal" (*logDice*), koji uzima u obzir ne samo apsolutnu frekvenciju već i učestalost pojavljivanja primarnog i sekundarnog kolokata van konteksta zadate kolokacije. Ovim parametrom (kojem su, inače, tvorci korpusa s razlogom pridali najveći značaj u kolokacionoj analizi) meri se koliko su analizirane reči „zavisne" jedna od druge. Kao što se vidi, ukupno je prikazano po 15 kolokata za svaku reč, koji su rangirani prema pomenutoj vrednosti *logDice*.

3. VEKTORSKA ANALIZA

Pored elektronskih korpusa, postoji još nekoliko računarskih metoda pomoću kojih se mogu analizirati semantički odnosi među leksemama. To su, između ostalog, modeli zasnovani na neuronskim mrežama, koji spadaju u domen veštačke inteligencije (AI). Oni pokazuju visoku efikasnost kada je reč o obradi prirodnog jezika (engl. *NLP – natural language processing*), zahvaljujući čemu su implementirani u mnoge alate koji se koriste za računarsku obradu i produkciju teksta – npr. za mašinsko prevođenje, čet-botove, optimizaciju pretraživača, analizu sentimenta, parafraziranje, pretvaranje govora u zapis itd. Međutim, ovi resursi zasad se vrlo malo koriste u teorijski orijentisanim istraživanjima. Indikativan je podatak da u srpskoj lingvistici, koliko je nama poznato, nije objavljen nijedan rad u kojem bi se autor služio nekim od dostupnih AI-alata pri razmatranju određenog fundamentalnog pitanja.[6]

Jedan od takvih resursa jeste tzv. *word2vec*, koji je uobličen pre deset godina (Mikolov *et al.* 2013). Uprošćeno rečeno, on funkcioniše tako što se isprva unosi velika količina građe (obično u milionima reči). Nakon što model identifikuje i obradi kolokate svake reči u zadatom kolokacionom opsegu, svim rečima se dodeljuje odgovarajući multidimenzionalni vektor (obično od sto ili više dimenzija). Na ovaj način se prvobitni tekstualni input pretvara u brojnu vrednost, koja se potom može podvrgavati različitim matematičkim operacijama – vektori se mogu upoređivati (može se meriti stepen njihove sličnosti), pa čak i sabirati i oduzimati (tako bi, na primer, u dobrom korpusu rezultat izraza *kralj - muško + žensko* trebalo da bude – *kraljica* (v. Mikolov *et al.* 2013: 2)).

Trebalo bi objasniti kako se ovaj metod razlikuje od korpusnog. Naime, i u jednom i u drugom slučaju uzimaju se u obzir kolokati određene lekseme, a kolokacioni opseg se može ciljano podešavati (iako smo se u ovde sprovedenoj korpusnoj analizi ograničili isključivo na poziciju -1, tu vrednost je moguće promeniti). Ipak, razlika je u tome što *word2vec* automatski poredi vektore stotina hiljada reči. S tim u vezi, zahvaljujući mašinski izmerenoj kosinusnoj sličnosti među njima, možemo za svaku reč iz građe utvrditi, na primer, još deset (ili više) reči koje su joj najsličnije po svom kontekstualnom ponašanju. Nasuprot tome, javno dostupni korpusi daju uvid u najčešće kolokate tražene lekseme (ili, eventualno, većeg broja njih), ali ne omogućavaju automatsku identifikaciju značenjski bliskih reči iz čitavog korpusa. Drugim rečima, korpusnim metodom možemo pouzdano ustanoviti sintagmatiku određene jedinice, a vektorskim – i pomoću nje „obračunatu" paradigmatiku.

Mi smo za potrebe ovog istraživanja napravili sopstveni neanotirani korpus. On je znatno manjeg obima nego što je to preporučljivo kada se koristi vektorski metod – sadrži svega 16007 primera i 30110 unikatnih reči. Međutim, ovaj nedostatak pokušali smo da premostimo pažljivim odabirom građe, koju smo direktno preuzimali iz već pomenutog javno dostupnog korpusa *PDRS 1.0*, ali ne nasumično, već ciljano kopirajući isključivo one primere u kojima je upotrebljena neka od sledećih leksema: *kompjuter*, *računar*, *laptop*, *tablet*, *telefon*, *televizor*, *uređaj*, *aparat*. Na taj način smo, po ugledu na neka ranija istraživanja (Dusserre/Padró 2017), fokus modela preusmerili ka onim rečima

6 Nasuprot tome, teorijski potencijal ovakvih alata bolje je prepoznat u inostranstvu, te su oni korišćeni u nezanemarljivom broju radova posvećenih različitim lingvističkim pitanjima (v. npr. Antipenko/Mitrofanova 2019; Savytska *et al.* 2021).

koje će nam biti potrebne u krajnjoj analizi (a čija je frekvencija sasvim zadovoljavajuća). Pošto naš korpus nije lematizovan, sve morfološke realizacije gorenavedenih leksema sveli smo pomoću regularnih izraza na njihov osnovni oblik (kako bi odgovarajući vektor bio povezan sa čitavom paradigmom, a ne sa svakom formom ponaosob).

Za implementaciju datog modela korišćen je programski jezik *Python*. Upotreba metoda *word2vec* omogućena je posredstvom besplatne biblioteke *Gensim*. Osim toga, odabran je standardni algoritam *CBOW*. Za svaku reč određen je 100-dimenzionalni vektor, a u obzir je uziman kontekst u opsegu -5 do +5 (tj. postavljen je hiperparametar *window = 5*)[7].

U nastavku ćemo navesti po pet reči čiji je vektorski profil najsličniji u odnosu na vektore leksema *računar* i *kompjuter*.[8]

računar	***kompjuter***
[('uredjaj', 0.8961764574050903), ('tv', 0.8953549861907959), ('aparat', 0.8803419470787048), ('wifi', 0.880127489566803), ('mejla', 0.8756557106971741)]	[('ovde', 0.9074090719223022), ('ekran', 0.9051103591918945), ('drugi', 0.903938889503479), ('svoj', 0.9033220410346985), ('pomoc', 0.898833708877563)]

Iako je očigledno da predstavljeni vektori, usled određenih nedostataka samog korpusa (manji broj primera i tokena, većinska nelematizovanost, prisustvo gramatičkih reči), nisu idealni, ipak se mogu zapaziti određene tendencije. Naime, leksema *računar* bliska je rečima kao što su *uređaj* i *aparat*, koje imaju krajnje uopšteno značenje. Ove imenice (uključujući i *tv*) često su postponirane u odnosu na neki atribut koji specifikuje njihovo značenje (isp. spojeve tipa *klima-uređaj*, *smart tv* i sl.). U prošlom poglavlju pokazali smo da se i leksema *računar* ponaša na sličan način (npr. *desktop računar*, *tablet računar*, *laptop računar*), tako da je ovakva veza očekivana. Kao što se može videti, leksema *kompjuter* tipično ne pokazuje takva kontekstualna svojstva.

Do sličnih zaključaka možemo doći i ako razmotrimo vektorske mreže reči *uređaj* i *tablet*.

uređaj	***tablet***
[('ajped', 0.9112445712089539), ('aparat', 0.9073323011398315), ('internetu', 0.9047685861587524), ('racunar', 0.8922036290168762), ('tv', 0.8921940922737122)]	[('mobilni', 0.9259875416755676), ('desktop', 0.8949421048164368), ('preko', 0.8930915594100952), ('televizor', 0.8896180391311646), ('vasem', 0.8843291997909546)]

7 Ova odluka doneta je u skladu sa tezom da se semantički bliske reči najbolje određuju preko relativno užeg konteksta, dok se širi kontekst (npr. -10 : +10) može koristiti za dobijanje reči koje pripadaju istom pojmovnom domenu, ali nisu nužno sličnog značenja, tj. ne moraju deliti istu arhisemu (isp. Levy/Goldberg 2014).

8 Inače, ovi rezultati mogu se i vizuelno predstaviti u koordinatnom prostoru (preko biblioteke *Matplotlib*) nakon što se vektori svedu na dve dimenzije (što je izvodljivo pomoću statističke funkcije *t-SNE* u okviru dostupne programske biblioteke *scikit-learn*).

Kao što smo i očekivali, leksema *uređaj* uporediva je sa imenicama šireg značenja, uključujući *računar*. S druge strane, reč *tablet* upotrebljava se u sličnom (adnominalnom) kontekstu kao i *mobilni*, *desktop* – isp. spojeve *tablet / desktop računar*, *mobilni telefon*.[9]

4. DISKUSIJA I ZAKLJUČAK

Kada je reč o teorijskom nivou istraživanja, sprovedena analiza pokazala je da lekseme *računar* i *kompjuter* ne predstavljaju apsolutne sinonime u savremenom srpskom jeziku, budući da nisu zamenljivi u svim kontekstima (**laptop kompjuter*, **tablet kompjuter*, [?]*desktop kompjuter* itd.). Ipak, nije lako precizno utvrditi kako je tekao proces njihove desinonimizacije. Možemo pretpostaviti da su ove dve reči u početku (sa pojavom kompjutera) zaista i predstavljale istoznačnice (što je i tipično za parove stranih i domaćih termina (v. Dragićević 2007: 245)). Međutim, razvoj i ekspanzija novih elektronskih uređaja koji obavljaju iste osnovne operacije kao i kompjuteri, ali se od njih po određenim inherentnim svojstvima razlikuju (npr. po obeležju prenosivosti) – poput laptopova i tableta – doveo je, kako nam se čini, do širenja semantike lekseme *računar*, te se ona počela koristiti kao opšti naziv za različite uređaje koji su funkcionalno bliski. Ovakav mehanizam u suprotnosti je sa značenjskom specijalizacijom kao tipičnim mehanizmom kojim se razrešava sinonimija (isp. Dragićević 2007: 248–249). Ipak, iako se nameće zaključak da domaća leksema *računar* danas ima nešto drugačije značenje od pozajmljenice *kompjuter*, iz sprovedene analize nije najjasnije kako bi trebalo predstaviti njenu polisemičnu strukturu. Naime, ne treba zanemariti pretpostavku (koju bi valjalo potvrditi psiholingvističkim eksperimentima) da i danas leksema *računar* najpre pobuđuje mentalnu predstavu stonog računara (tj. kompjutera). Šta nam to govori o njenom (tradicionalno shvaćenom) leksičkom značenju? Naime, mogli bismo reći da je ova leksema monosemična. U tom slučaju njena designacija (kao apstraktan skup obaveznih svojstava) bila bi znatno šira (i „objedinjavala" bi kako prenosive, tako i neprenosive uređaje), ali njena denotacija (kao realan, prototipičan predstavnik objekta (isp. Dragićević 2007: 58)) i dalje bi bila čvrsto spregnuta sa slikom desktop računara.[10] Ovakva postavka može se dovesti u vezu sa semantičkom nespecifikovanošću kao jednim od koncepata koji se podrobnije razrađuju u modernoj lingvistici (v. Murphy 2010: 84, gde se kao primer navodi reč *clock*, koja može označavati različite tipove satova: digitalne, analogne, budilnike i sl.), ali i sa similisemijom (v. Gortan-Premk 1997: 59–67; Dragićević 2007: 136–137), koja podrazumeva suptilnu hijerarhiju između srodnih značenjskih subrealizacija (npr. a. *glava* [čoveka], b. *glava* [životinje]). Zapravo, čini se da je značenjski „preobražaj" lekseme *računar* još uvek u povoju; on se, figurativno rečeno, odigrava pred našim očima. Data leksema još uvek nije raskinula čvrste semantičke „niti" koje je vezuju za izvorni koncept, ali istovremeno pokazuje jasnu tendenciju ka širenju

9 U datoj tabeli interesantan je i odnos između reči *tv* i *televizor*, od kojih se prva kontekstualno ponaša kao imenica opštije, a druga – specifičnije semantike. Zašto je to tako? Pretpostavljamo da odgovor leži u činjenici da ćemo skraćenicu *TV* mnogo češće čuti u preuzetom izrazu *smart TV* , dok će se imenica *televizor* obično koristiti bez atributskog determinatora. Ovaj primer nas istovremeno podseća da nas kompjuterski obrađena građa ponekad može navesti na pogrešne zaključke.

10 Fazičnu značenjsku strukturu lekseme *računar* bilo bi sasvim prikladno analizirati i iz perspektive teorije prototipa: u njenom centru bio bi kompjuter, bližu periferiju činili bi prenosivi računari poput laptopova, a dalju – tableti i slični uređaji.

svog polaznog značenja; dakle, ona prema leksemi *kompjuter* zauzima ambivalentan odnos – može se tumačiti kao njen relativni sinonim ili pak hiperonim (povodom datog teorijskog problema v. Dragićević 2007: 296–297). S obzirom na sve što je iznad rečeno, ovde je umesno poslužiti se Prćićevim (2019: 148) hibridnim terminom *hiposinonimija*, budući da on odlično odražava fazičnu i kompleksnu prirodu semantičkih relacija u koje stupaju analizirane lekseme. Najzad, s obzirom na hiperprodukciju novih računarskih sistema i ubrzano zastarevanje postojećih (uključujući klasične kompjutere), možemo očekivati dalju semantičku generalizaciju lekseme *računar* i potencijalni razvoj pravog opšteg značenja (v. Dragićević 2007: 135–136), koje je apstraktno i „nevidljivo", ali natkriljuje sve pojedinačne i ravnopravne semanteme bez jasne hijerarhije među njima (kao što, na primer, leksema *kora* označava tvrdi spoljni omotač različitih entiteta: drveta, plodova, hleba itd.). Naravno, ovaj proces pratiće i stabilizacija uskog, specijalnog značenja anglicizma *kompjuter* ('stoni računar').

Iz svega navedenog proističe zaključak da su obe reči našle svoje mesto u leksičkom sistemu, kao i da su potrebne srpskom jeziku. S tim u vezi, možemo reći da su neopravdani normativistički napori da se jedna od ovih dveju leksema odbaci (*kompjuter* jer je u pitanju pozajmljenica, odn. *računar* jer nije razvio razgranatu mrežu derivata). Ovaj slučaj još jednom potvrđuje tezu da sinonimija podleže spontanoj samoregulaciji, zbog čega spoljne intervencije najčešće i nisu neophodne.

Smatramo da je najveći doprinos ovog rada metodološke prirode. Naime, pokušali smo da pokažemo da se teorijskim pitanjima leksičke semantike može pristupati i uz oslanjanje na najnovija dostignuća računarske lingvistike i informatike. Duboko smo uvereni da se o značenju reči najobjektivnije može suditi upravo na osnovu kvantitativne analize njenih kontekstualnih preferenci. Pošto je takav istraživački postupak složen (naročito kada se operiše većim brojem jedinica), naučnici se mogu poslužiti brojnim tehnološkim alatima koji su im na raspolaganju, poput ovde pomenutih elektronskih korpusa i vektorskih semantičkih modela.

Najzad, procena sinonimičnosti (i drugih semantičkih parametara) može, naravno, imati i primenu u leksikografskoj praksi. Informacije o značenjskim vezama među leksemama dragocene su za različite vrste opštih i specijalnih rečnika (među njima, podrazumeva se, i za sinonimske). Kako bi takvi važni podaci bili što pouzdaniji, poželjno je sakupiti obimnu i raznovrsnu građu, koju bi zatim trebalo statistički obraditi. Osim toga, kvantitativna semantika mogla bi predstavljati i okosnicu za osmišljavanje raznovrsnih digitalnih rečnika i drugih elektronskih jezičkih resursa.

LITERATURA

Antipenko, A. A. i O. A. Mitrofanova. 2019. Исследование ассоциативных связей слов в корпусе социальных сетей с помощью дистрибутивносемантических моделей. *Компьютерная лингвистика и вычислительные онтологии 3*, 77–91.

Cruse, A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Dragićević, R. 2007. *Leksikologija srpskog jezika*. Beograd: Zavod za udžbenike.

Dragićević, R. 2017. Srpska leksikologija danas: sadašnje stanje i perspektive. *Južnoslovenski filolog LXXIII(3–4)*, 259–290.

Dražić, J. 2014. *Leksičke i gramatičke kolokacije u srpskom jeziku*. Novi Sad: Filozofski fakultet.

Dusserre, E. and M. Padró. 2017. Bigger does not mean better! [Internet]. Dostupno na: https://aclanthology.org/W17-6908.pdf [1. 12. 2023].

Filipović, R. 1986. *Teorija jezika u kontaktu : uvod u lingvistiku jezičnih dodira*. Zagreb: Jugoslavenska akademija znanosti i umjetnosti: Školska knjiga.

Firth, J. R. 1962. A Synopsis of Linguistic Theory, 1930-55. In J. R. Firth (ed.) *Studies in Linguistic Analysis. Special Volume of the Philological Society*. Oxford: Blackwell, 1–31.

Gortan-Premk, D. 1997. *Polisemija i organizacija leksičkog sistema u srpskome jeziku*. Beograd: Zavod za udžbenike i nastavna sredstva.

Gotštajn, Lj. 1986. *Sinonimija u kolokacijama (sa primerima engleskog jezika naučne argumentacije)*. Novi Sad: Filozofski fakultet.

Herrick, J. 2005. *The History and Theory of Rhetoric: An Introduction*. 3rd edition. Boston: Allyn and Beacon.

Ivić, M. 1970. *Pravci u lingvistici*. Ljubljana: Državna založba Slovenije.

Kleparski, G. and A. Rusinek. 2007. The tradition of field theory and the study of lexical semantic change. *Studia Anglica Resoviensia 4*, 188–205.

Kotzia, P. and M. Chriti. 2014. Ancient Philosophers on Language. In G. Giannakis *et al*. (eds.) *The Encyclopedia of Ancient Greek Language and Linguistics*. Leiden and Boston: Brill, 124–133.

Lenci, A. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica 20(1)*, 1–31.

Levy, O. and Y. Goldberg. 2014. Dependency-Based Word Embeddings. [Internet]. Dostupno na: https://aclanthology.org/P14-2050.pdf [1. 12. 2023].

Mikolov, T. *et al*. Efficient Estimation of Word Representations in Vector Space. [Internet]. Dostupno na: https://arxiv.org/pdf/1301.3781 [1. 12. 2023].

Milić, M. Anglicizmi kao sinonimi u srpskom jeziku. [Internet]. Dostupno na: https://digitalna.ff.uns.ac.rs/sadrzaj/2013/978-86-6065-166-4 [1. 12. 2023].

Murphy, M. L. 2010. *Lexical Meaning*. Cambridge: Cambridge University Press.

Nešović, M. 2023. Anglicizmi u srpskom i ruskom omladinskom žargonu (morfosintaksički i semantički aspekt). *Slavistika XXVII(1)*, 180–197.

Portner, P. 2005. *What is Meaning? Fundamentals of Formal Semantics*. Oxford: Blackwell.

Prćić, T. 2016. Semantika i pragmatika reči. [Internet]. Dostupno na: https://digitalna.ff.uns.ac.rs/sadrzaj/2016/978-86-6065-356-9 [1. 12. 2023].

Prćić, T. 2019. Engleski u srpskom, treće izdanje. [Internet]. Dostupno na: https://digitalna.ff.uns.ac.rs/sadrzaj/2019/978-86-6065-512-9 [1. 12. 2023].

Savytska, N. V. *et al*. 2021. Using Word2vec Technique to Determine Semantic and Morphologic Similarity in Embedded Words of the Ukrainian Language. [Internet]. Dostupno na: https://ceur-ws.org/Vol-2870/paper21.pdf [1. 12. 2023].

de Saussure, F. 2011. *Course in General Linguistics* (eds. P. Meisel and H. Saussy, trans. W. Baskin). New York: Columbia University Press.

Shmelev, D. N. 1973. *Проблемы семантического анализа лексики (на материале русского языка)*. Москва: Издательство „Наука“.

Sudilovskaya, V. G. 2018. *Введение в лексикологию: учебное пособие*. Санкт-Петербург: Балтийский государственный технический университет «Военмех».

Šipka, D. 1998. *Osnovi leksikologije i srodnih disciplina*. Novi Sad: Matica srpska.

SUMMARY

DESYNONYMIZATION THROUGH THE LENS OF CORPUS AND VECTOR ANALYSIS OF THE CONTEXTUAL PROPERTIES OF THE WORDS *KOMPJUTER* AND *RAČUNAR*

In this paper, we discuss the meanings of the Serbian words *kompjuter* and *računar* ('computer') in terms of their semantic proximity. Using web corpus data, as well as *word2vec* method for measuring the cosine similarity between their vector representations (based on their contextual preferences), we conclude that, contrary to popular belief, these words should not be considered absolute synonyms. Specifically, we propose that the loanword *kompjuter* has a narrower sense ('PC', 'desktop computer'), whereas its loan-translation counterpart *računar* carries a broader meaning ('any larger computational device'). This conclusion is based on the fact that the two lexemes in question do not share distributional patterns: while *kompjuter* is typically used as a syntactically free expression, *računar* is often preceded by various attributes that specify the meaning of the entire nominal phrase (e.g. *laptop računar*, *tablet računar*, *iPad računar*). Consequently, these words are not always contextually interchangeable. Additionally, we propose that computational resources should be utilized when addressing various practical and theoretical linguistic problems.

KEYWORDS: *kompjuter*, *računar*, synonyms, desynonymization, distributional semantics, context, corpus, *word2vec*.

PODACI O ČLANKU:
Originalni naučni rad
Primljen: 2. oktobra 2023.
Ispravljen: 12. decembra 2023.
Prihvaćen: 12. decembra 2023.