

Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica

Apolonija Gantar

 <https://orcid.org/0000-0001-5822-6414>

Filozofska fakulteta Univerze v Ljubljani (Slovenija)

apolonija.gantar@ff.uni-lj.si

Sažetak

U radu se razmatraju mogućnosti korišćenja veštačke inteligencije (VI) u leksikografske svrhe, a posebno se ispituje uspešnost generativnih metoda u formiranju rečničkih definicija i dobijanju dobrih rečničkih primera za semantički i strukturno složene frazeološke jedinice (FJ). Analiza generisanih rečničkih definicija za skup od 35 slovenačkih frazeoloških jedinica pokazala je dobre preliminarne rezultate u odgovarajućem prepoznavanju značenja FJ, čak i u više od 60% primera. Veštačka inteligencija je lošije prošla prilikom kreiranja rečničkih primera upotrebe, koji su u većini slučajeva veštački i neintuitivni na prvi pogled. Istraživanje, između ostalog, pokazuje i nezavisnost uspešnosti generisanih definicija od nivoa idiomatičnosti FJ i potrebu upoznavanja veštačke inteligencije s negativnim odnosno bukvalno upotrebljenim FJ, ako one postoje u jeziku.

Ključne reči: leksikografija, veštačka inteligencija, frazeološke jedinice, rečničke definicije, slovenački jezik.

Abstract

In the paper, we discuss some possibilities of using artificial intelligence for lexicographical purposes, in particular we examine the success of generative methods in forming dictionary definitions and obtaining good dictionary examples for semantically and structurally complex phraseological units. The analysis of the generated dictionary definitions for 35 Slovenian phraseological units showed good preliminary results in appropriately recognising the meaning of PU, even in more than 60% of examples. Artificial intelligence fared worse when creating dictionary examples, often artificial and unintuitive at first glance. The research, among other things, also shows the independence of the success of the generated definitions from the level of idiomativeness of PU and the need to acquaint artificial intelligence with negative or literal examples of PU if they exist in the language.

Keywords: lexicography, artificial intelligence, phraseological units, definitions, Slovene.

1. Uvod

Nakon što je u novembru 2022. ChatGPT predstavljen javnosti, očekivanja leksikografske zajednice bila su razumljivo velika budući da njeni članovi već deceniju nastoje da što više automatizuju leksikografske postupke, naročito one koji zahtevaju dugotrajan ručni rad.

U oblasti leksikografije u poslednjih su 40, a naročito 20 godina razvijene jezičke tehnologije, resursi i alati, kao što su npr. anotirani korpusi i alati za njihovu analizu. To je omogućilo brže uređivanje velike količine autentičnih jezičkih podataka i, tokom godina, da se leksikografski proces postepeno približava sve većoj automatizaciji. Zbog toga su se s dolaskom veštačke inteligencije (VI) leksikografima i korpusnim lingvistima razumljivo javi-

la pitanja, kao što su npr.: da li se i — štaviše — kako se veštačka inteligencija može upotrebiti pri izradi rečnika, možemo li jednostavno zaboraviti na sve tehnologije koje su u oblasti leksikografije razvijene u poslednjih 20 ili čak 40 godina? Prema našem mišljenju, jedno od ključnih pitanja koje mi leksikografi moramo postaviti sebi, ukoliko želimo realan suživot u svetu veštačke inteligencije, jeste kako postojeće izvore i semantičko znanje koje sadrže postojeći rečnici upotrebiti za unapređenje rada veštačke inteligencije pri izradi rečnika i jezičkih modela uopšte? A u širem smislu utvrditi kako sami ljudi obrađuju jezik, naročito semantičke informacije, da bismo zapravo mogli da razumemo kako veštačka inteligencija obrađuje semantičke informacije. Kada to shvatimo, možda ćemo moći da unapredimo rad ve-

štačke inteligencije, uključujući i one zadatke koji su do sada bili isključivo u domenu leksikografa — tj. razlučivanje značenja višeznačnih reči (eng. *word sense disambiguation*), pisanje rečničkih definicija i navođenje dobrih rečničkih primera.

Polazeći od istraživanja koja su nastala u prvih šest meseci nakon puštanja u rad ChatGPT-a (DE SHRYVER 2023; BARRETT 2023; LEW 2023; McKEAN – FITZGERALD 2023; RUNDELL 2023; JAKUBIČEK – RUNDELL 2023), u vezi s ulogom VI pri izradi rečnika izvedeni su sledeći zaključci. Kao prvo, za izradu velikih jezičkih modela potrebna je velika količina pouzdanih podataka, naročito semantičkih. Zbog toga su rečnici, koji sadrže te podatke u strukturiranom obliku veoma tražen i važan semantički izvor. Što se tiče korišćenja metoda veštačke inteligencije pri izradi rečnika, odnosno rečničkih članaka, istraživanja kažu da ona iznenađujuće dobro radi pri formulisanju definicija — čak u više od 60 procenata. Ipak, razlučivanje značenja višeznačnih reči i za ChatGPT predstavlja veliki izazov, jer daje lošije rezultate pri prepoznavanju višeznačnih reči. Kako navodi Randel (RUNDELL 2023: 6), neka značenja su udvojena, druga izmišljena, a neka su važna značenja čak i izostavljena. Ujedno je utvrđeno da se automatski generisane definicije mogu značajno unaprediti pomoću jezičkih upita (engl. *prompts*) i da ChatGPT ima najlošije rezultate pri navođenju dobrih, autentičnih primera. Naime, primeri ChatGPT-a u većini su slučajeva, kako u svojoj studiji navodi Randel (RUNDELL 2023: 7), neuverljivi i često na prvi pogled izmišljeni i neintuitivni.

U nastavku rada predstavljamo nalaze kratkog istraživanja, u kome smo proveravali razumevanje značenja i upotrebe frazeoloških jedinica (FJ) od strane veštačke inteligencije, odnosno proverom adekvatnosti veštački generisane rečničke definicije. U radu najpre opisujemo ciljeve i metode istraživanja, kao i pretpostavke istraživanja. U nastavku predstavljamo skup frazeoloških jedinica i jezički upit. Centralni deo rada namenjen je leksikografskoj analizi automatski generisanih definicija u odnosu na one ručno rađene u postojećim rečnicima. U posebnom odeljku prvo utvrđujemo sposobnost veštačke inteligencije da prepozna značenje frazeološke jedinice, zatim analiziramo nalaze sa stanovišta semantičkih i kulturnih svojstava frazeoloških jedinica. Članak završavamo nalazima koje je istraživanje pružilo za dalji rad.

2. Formulisanje rečničkih definicija pomoću veštačke inteligencije — na primeru slovenačkih frazeoloških jedinica

Istraživanje koje je predstavljeno u nastavku ovog rada spada u širi okvir interdisciplinarnih istraživanja koje u Centru za jezičke resurse i tehnologije Univerziteta u Ljubljani sprovodimo zajedno s kolegama sa Fakulteta za računarstvo i informatiku i s Filozofskog fakulteta

Univerziteta u Ljubljani, a koje su usmerene ka poboljšanju metoda za automatsko prikupljanje semantičkih podataka u leksikografske i jezikotehnoške svrhe (npr. KOSEM i dr. 2024; GAPSA i dr. 2024).

2.1. Ciljevi i metode

Cilj je našeg istraživanja testirati metodu po kojoj pomoću jezičkih upita dobijamo automatski generisane rečničke definicije za frazeološke jedinice. Pri tome, frazeološke jedinice razumemo kao leksičke jedinice od više reči koje pokazuju određen stepen strukturne ustaljenosti i semantičke netransparentnosti, a koje imaju izraženu ekspresivnu ili pragmatičnu ulogu u jeziku (GANTAR i dr. 2021). Dalje, želimo da izvršimo jezičku analizu rezultata u kojoj ćemo proveriti generisane definicije za unapred određeni skup frazeoloških jedinica, i to s aspekta adekvatno, odnosno neadekvatno prepoznatog značenja, odnosno definicije i višeznačnosti s obzirom na stepen idiomatičnosti, kao i s aspekta navođenja adekvatnih rečničkih primera upotrebe. Rezultate ChatGPT-a želimo da uporedimo s postojećim definicijama frazeoloških jedinica u nekom od slovenačkih rečnika, npr. u opštem rečniku SSKJ2, *Rečniku slovenačkih frazema* (KEBER 2015) i u *Velikom slovenačko-mađarskom rečniku* (VSMS).

Pre nego što se posvetimo opisu istraživanja, potrebno je pomenuti i njegova ograničenja, koja utiču na donošenje konačnih zaključaka. Prvi nedostatak ogleda se u veoma malom skupu frazeoloških jedinica, koji ne dozvoljava uopštavanje niti pouzdane kvantitativne analize. Druga metodološka slabost leži u činjenici da istraživanje nije moguće ponoviti jer ChatGPT pri ponavljanju istog jezičkog upita daje različite odgovore. Upravo zato smo odlučili da testiramo rad ChatGPT-a jednokratnim, ali unapred promišljenim jezičkim upitom. Pored toga, u istraživanju smo koristili besplatni, javno dostupni ChatGPT v3.5, pri čemu nova verzija v4 već provereno bolje radi, što znači da su neki naši zaključci možda već prevaziđeni, odnosno zastareli. I najzad, utvrđivanje stepena idiomatičnosti, koju smo u istraživanju upotreбили kao varijablu, izrazito je subjektivan postupak, a ujedno i jedan od osnovnih leksikografskih zadataka pri leksikografskoj obradi frazeoloških jedinica.

Ovim istraživanjem stoga smo prvenstveno želeli da dobijemo odgovore kuda usmeriti naše istraživanje za optimalno korišćenje veštačke inteligencije — naravno, pod pretpostavkom da rečnik sadrži legitimne i proverene podatke koji odražavaju jezik jezičke zajednice (a ne jezik veštačke inteligencije).

2.2. Pretpostavke

S obzirom na ranije pomenute opšte tvrdnje, naše su pretpostavke (1) da se metode veštačke inteligencije

moгу uspešno upotrebiti pri formulisanju definicija i za frazeološke jedinice; te (2) da je ChatGPT sposoban da prepozna različita značenja višeznačnih frazeoloških jedinica, iako je uspešnost modela pri tome povezana sa značenjskom specifičnošću konkretne frazeološke jedinice i stepenom njene idiomatičnosti; pored toga, pretpostavljamo (3) da veliki jezički modeli, koji ne sadrže specifične semantičke podatke za određeni jezik — u našem primeru slovenački — ne dostižu visok nivo pri prepoznavanju značenja frazeoloških jedinica i njihove tipične upotrebe.

2.3. Skup frazeoloških jedinica

Spisak frazeoloških jedinica koje smo uključili u istraživanje sadrži 35 jedinica. Njih smo odabrali na osnovu sledećih kriterijuma: (1) frazeološka jedinica morala je biti evidentirana i leksikografski obrađena u nekom od postojećih, javno dostupnih rečnika. Ovaj kriterijum omogućio nam je objektivno poređenje automatski generisanih definicija s leksikografskim definicijama. Zatim, (2) frazeološka jedinica morala je imati najmanje

10 potvrda u referentnom korpusu standardnog slovenačkog jezika *Gigafida* v2.0 (KREK i dr. 2019) (u daljem tekstu: GFv2). Takođe, (3) uzeli smo u obzir zastupljenost najčešćih sintaksičkih struktura, pri čemu prevladavaju glagolske i imeničke frazeološke jedinice, koje su i inače česte u jeziku. Dalje, (4) određen deo frazeoloških jedinica sa spiska morao je u rečniku pokazivati višeznačnost. Isto tako, (5) određen deo frazeoloških jedinica morao je sadržati elemente specifične za jezik i kulturu. (6) Uzeli smo u obzir zastupljenost frazeoloških jedinica s različitim stepenom idiomatičnosti po unapred utvrđenim kriterijumima, koje predstavljamo u odeljku 2.5.2.

Spisak frazeoloških jedinica koje su obuhvaćene istraživanjem prikazan je u tabeli 1. Simbolom ~ uводимo ekvivalent FJ u srpskom jeziku, pod navodnicima ‘’ označavamo značenje FJ odnosno njen prevod, i u [] doslovni prevod komponenti FJ, gde to ima smisla za razumevanje stepena idiomatičnosti.

1	<i>prepeljati koga žejnega čez vodo</i> ~ prevesti nekog žednog preko vode	2	<i>podarjenemu konju se ne gleda v zobe</i> ~ poklonu se u zube ne gleda	3	<i>dati roko v ogenj</i> ~ staviti ruku u vatru	4	<i>na prostem</i> ~ na otvorenom
5	<i>iti rakom žvižgat</i> [žviždit na rakove] ~ baciti kašiku	6	<i>živeti iz rok v usta</i> ~ živeti od danas do sutra	7	<i>hoditi narobe</i> ‘izazivati teškoće’ [ići naopako]	8	<i>naleteti na gluha ušesa</i> ~ govoriti gluvim ušima
9	<i>pomesti pod preprogo</i> ~ gurnuti pod tepih	10	<i>imeti poln kufer</i> ~ imati pun kufer	11	<i>kapo dol</i> ~ kapa dole	12	<i>ne tič ne miš</i> ~ nit smrdi nit miriše
13	<i>pasti v vodo</i> ~ pasti u vodu	14	<i>pasti dol</i> ~ pasti na nos	15	<i>nositi hlače</i> ~ nositi pantalone	16	<i>na štiri oči</i> ~ u četiri oka
17	<i>suh kot poper</i> ~ suv kao barut	18	<i>biti kvit</i> ~ biti kvit	19	<i>biti za</i> ~ biti za	20	<i>zamrznjeni predsednik</i> [zamrznuti predsednik]
21	<i>rdeče številke</i> ~ crvene brojke	22	<i>bolj ali manj</i> ~ više ili manje	23	<i>na vse kriplje</i> ~ iz petnih žila	24	<i>dati vse od sebe</i> ~ dati sve od sebe
25	<i>na vse ali nič</i> ~ na sve ili ništa	26	<i>stric iz ozadja</i> ~ vladar iz senke	27	<i>trinajsto prase</i> ~ trinaesto prase	28	<i>gordijski voz</i> ~ Gordijev čvor
29	<i>slamnata vdova</i> ~ bela udovica	30	<i>španska vas</i> ~ špansko selo	31	<i>mala malica</i> ‘što je jednostavno i nekomplikovano’	32	<i>priti do</i> ~ doći do
33	<i>domaća naloga</i> ~ domaći zadatak	34	<i>imeti velike oči</i> ~ imati velike oči	35	<i>biti bolj papeški od papeža</i> ~ biti veći katolik od pape		

Tabela 1. Spisak frazeoloških jedinica uključenih u jezički upit.

2.4. Jezički upit

Zadatak smo ChatGPT-u postavili u formi jednokratnog jezičkog upita, koji je glasilo: Kakvo je značenje, ili tačnije rečnička definicija slovenačkih frazeoloških jedinica s donjeg spiska? Navedi rečničku definiciju na gramatički ispravnom slovenačkom jeziku i po dve rečenice kao primere tipične upotrebe frazeološke jedinice u kontekstu. Neka definicije budu kratke i jezgrovite. Ako postoji više od jednog značenja određene frazeološke jedinice, navedi definicije za sva značenja i primere upotrebe za svako značenje. Značenja numeriši.

2.5 Leksikografska analiza

Na osnovu ovako postavljenog jezičkog upita generisano je 35 rečničkih definicija za isti broj značenja i 56 primera upotrebe. Rezultati analize na osnovu postavljenih pitanja dati su u nastavku rada.

2.5.1. Prepoznavanje značenja. Najpre smo proverili adekvatnost automatski generisanih rečničkih definicija tako što smo utvrdili u koliko je primera ChatGPT adekvatno prepoznao barem jedno od nedoslovnih značenja frazeološke jedinice koja su zabeležena u nekom od rečnika. Upotrebili smo trostepenu lestvicu, pri čemu nam DA znači da je definicija adekvatna, tj. da razumljivo objašnjava značenje frazeološke jedinice koje je zabeleženo u nekom od rečnika; NE znači da ChatGPT nije adekvatno prepoznao značenje, što dalje znači da je generisana rečnička definicija neadekvatna; konačno, DELIMIČNO znači da je ChatGPT značenje frazeološke jedinice formulisao drugačije od rečnika, ali je prepoznao smisao njenog osnovnog značenja. Rezultati su predstavljeni u tabeli 2:

Adekvatnost ChatGPT definicije	Br. FJ	U %
DA	21	60
DELIMIČNO	6	17
NE	8	23
Ukupno	35	100

Tabela 2. Numerički prikaz adekvatno generisanih rečničkih definicija.

Kako pokazuju podaci iz tabele 2, ChatGPT je u više od polovine primera (60%) adekvatno prepoznao i definisao barem jedno od značenja frazeološke jedinice. To se vidi u primeru (1) i (2):

(1) biti kvit ~ biti kvit

SSKJ: brez dolga, brez obveznosti, na čistem 'bez dugova, bez obaveza'.

ChatGPT: biti izenačen ali poravnati dolgove 'biti izjednačen ili izmiriti dugove'.

(2) podarjenemu konju se ne gleda v zobe ~ poklonu se u zube ne gleda

SSKJ: pri podarjeni stvari se ne smejo iskati napake 'poklonjenosti stvari se ne smeju tražiti greške'.

ChatGPT: ne kritizirati ali dvomiti v kvaliteto nečesa, kar je bilo podarjeno 'ne kritikovati niti dovoditi u pitanje kvalitet nečega što je poklonjeno'.

Dalje, kao što se vidi iz primera (3) i (4), kod kojih smo naveli delimično adekvatno generisanu definiciju, u pitanju je frazeološka jedinica gde se rečnička, odnosno leksikografska i generisana definicija razlikuju, ali se može reći da veštački generisana definicija daje osnovni smisao frazema, te da tako i te primere realno možemo uvrstiti u grupu primera s adekvatno prepoznatim značenjem. To dodatno povećava uspešnost ChatGPT-a:

(3) naleteti na gluha ušesa ~ govoriti gluvim ušima

SSKJ: naleteti na ljudi, ki se ne dajo prepričati 'naiči na ljude koji se ne mogu ubediti'.

ChatGPT: nekomu se pritoževati ali prositi za pomoč, vendar ne biti uslišan ali razumljen 'žaliti se nekome ili ga moliti za pomoč, ali ostati bez razumevanja ili uslišenja molbe'.

(4) ne tič ne miš ~ nit smrdi nit miriše

SSKJ: ki je brez izrazitih lastnosti, značilnosti 'koji nema izrazite osobine, karakteristike'.

ChatGPT: oseba ali entiteta, ki ni jasno opredeljena ali uvrščena v nobeno določeno kategorijo ali skupino 'osoba ili entitet koji nije jasno definisan ili uvršten u neku određenu kategoriju ili grupu'.

Frazeoloških jedinica kojima ChatGPT nije adekvatno prepoznao značenje, kao što se vidi iz primera (5) i (6), ima relativno malo — od 35 obrađenih frazeoloških jedinica, takvih ima svega 8, što čini približno četvrtinu građe:

(5) slamnata vdova ~ bela udovica

SSKJ: žena, katere mož je dalj časa odsoten 'žena čiji je muž duže vreme odsutan'.

ChatGPT: Oseba, ki prejema pokojnino ali druge ugodnosti po smrti partnerja 'Osoba koja prima penziju i druge pogodnosti nakon smrti partnera'.

(6) prepeljati koga žejnega čez vodo ~ prevesti nekog žednog preko vode

SSKJ: z zvitim govorjenjem, ravnanjem izmakniti se izpolnitvi zahtev koga 'uvijenim govorenjem, ponašanjem izbeći ispunjavanje nećijih zahteva'.

ChatGPT: 1. Uspeti izvesti nekaj zahtevnega ali tveganega; 2. Spretno iskoristiti priliku ali rešiti težavu 1. 'Uspeti izvesti nešto zahtevno ili rizično'; 2. 'Spretno iskoristiti priliku ili rešiti teškoću'.

Među frazeološkim jedinicama koje je VI neadekvatno razumela nalaze se naročito višeznačne frazeološke jedinice, kod kojih ChatGPT prepoznaje samo doslovno, ali ne i frazeološko značenje.

2.5.2. Stepen idiomatičnosti. U nastavku istraživanja želeli smo utvrditi da li je i u kojoj meri prepoznavanje adekvatnog značenja frazeološke jedinice od strane VI uslovljeno stepenom njene idiomatičnosti, pri čemu idiomatičnost razumemo kao jednu od ključnih definicijskih karakteristika frazeoloških jedinica.

Stepenom 1 označavali smo skoro potpuno semantički transparentne veze. S obzirom na definicijske karakteristike frazeoloških jedinica, koje pored semantičke netransparentnosti zahtevaju i metaforičnost, na našem spisku su takve samo 2 jedinice, za koje je karakteristična pre svega strukturna ustaljenost, npr. *bolj ili manj* ~ više ili manje; *na prostem* ~ na otvorenom.

Stepenom 2 beležili smo delimično semantički transparentne frazeološke jedinice, kod kojih je značenje moguće razumeti na osnovu određenih elemenata, ponajpre na osnovu metaforičkih prenosa i asocijativnosti ustaljenih u jeziku, npr. *rdeče številke* ~ crvene brojke; *na vse ili nič* ~ na sve ili ništa; *živeti iz rok v usta* ~ živeti od danas do sutra. Stepenom 2 označili smo i frazalne glagole tipa: *priti do* ~ doći do; *biti za* ~ biti za, te frazeološke jedinice sa zajedničkom kulturnom, istorijskom ili političko-geografskom pozadinom, npr. *gordijski voz* ~ Gordijev čvor.

Stepenom 3 označavali smo frazeološke jedinice s najvišim stepenom idiomatičnosti, npr. *biti kvit* ~ biti kvit; *na vse kriplje* ~ iz petnih žila; *iti rakom žvižgat* ~ baciti kašiku, *hoditi narobe* 'izazivati teškoće, nevolju' i sl. Rezultate analize prikazuje tabela 3.

Stepen idiomatičnosti	Značenjska adekvatnost						Ukupno br. FJ
	DA br. FJ	%	NE br. FJ	%	DELI-MIČNO br. FJ	%	
1	2	100	0	0	0	0	2
2	8	61,5	4	31	1	7	13
3	11	55	4	20	5	25	20
Ukupno	21	60	8	23	6	17	35

Tabela 3. Značenjska adekvatnost s obzirom na stepen idiomatičnosti FJ.

S obzirom na visok procenat frazeoloških jedinica s adekvatno generisanim rečničkim definicijama (čak 60%), može se očekivati da će visok procenat značenjski adekvatnih frazeoloških jedinica biti i kod onih koje imaju najviši stepen semantičke netransparentnosti — takvih je zapravo više od polovine frazeoloških jedinica s našeg spiska (55%). I kod frazeoloških jedinica sa srednjim stepenom idiomatičnosti ChatGPT adekvatno je prepoznao značenje i ponudio adekvatnu rečničku definiciju — u više od 60 % primera. S druge strane, među frazeološkim jedinicama s najvišim i srednjim stepenom idiomatičnosti isto je tako visok procenat frazeoloških jedinica kod kojih ChatGPT nije adekvatno prepoznao

značenje — reč je o preostaloj polovini frazeoloških jedinica s našeg spiska: ukupno 51 %.

Iako je naš skup primera suviše mali za donošenje opštih zaključaka, u daljim istraživanjima možemo ispitati novu hipotezu — naime, da stepen idiomatičnosti frazeoloških jedinica ne utiče suštinski na prepoznavanje FJ od strane ChatGPT i na generisanje adekvatnih rečničkih definicija. To ujedno stavlja pod znak pitanja našu polaznu pretpostavku — da će ChatGPT imati lošije rezultate pri razumevanju frazeoloških jedinica u slučaju semantički netransparentnih i jezički i kulturno specifičnih frazeoloških jedinica. S obzirom na rezultate istraživanja, još uvek nije moguće potpuno oboriti tu pretpostavku, jer među neadekvatno definisanim frazeološkim jedinicama s najvišim stepenom idiomatičnosti nalazimo naročito one koje uključuju kulturno specifične elemente (*iti rakom žvižgat* [zviždati na rakove] ~ baciti kašiku) ili koje su jezički, pre svega sintaksički, specifične (*hoditi narobe* ~ izazivati teškoće, nevolju), naročito ako kod njih postoji i doslovno značenje, kao u prikazanim primeru (7):

(7) *hoditi narobe* 'izazivati teškoće, nevolju'

GFv2: ne uzdržati, povzročiti težave.

ChatGPT: gibati se v nasprotni ali nepravilni smeri glede na običajen način hoje. 'Kretati se u suprotnom ili nepravilnom smeru u odnosu na uobičajen način hodanja'.

2.5.3. Rečnički primeri upotrebe. Treba još pogledati kako se ChatGPT snašao pri navođenju rečničkih primera upotrebe. Naime, navođenje adekvatnih i ilustrativnih primera upotrebe u rečnicima je kod značenjski kompleksnih frazeoloških jedinica utoliko važnije pošto bi trebalo da primeri odražavaju tipičnu upotrebu, smetanje u kontekst prema sintaksičkim karakteristikama, govorne situacije, obeležnosti frazeološke jedinice i sl. Upravo zbog toga je uloga leksikografa u tom zadatku od ključnog značaja.

Kako je već utvrđeno u prethodnim istraživanjima, ChatGPT u ovom zadatku ne postiže dobre rezultate, što potvrđuje i naše istraživanje. Samo u 10% slučajeva generisane primere mogli bismo da ocenimo kao rečnički adekvatne. Na primer, u slučaju relativno semantički transparentne frazeološke jedinice *na prostem* (primer 8) u značenju 'na otvorenom, izvan prostorija' ChatGPT je, po našem mišljenju, naveo dobar primer, iako možda lošiji od ručno izabranog iz korpusa GFv2, ali svakako bolji od navedenog u SSKJ2:

(8) *na prostem* ~ na otvorenom

SSKJ2: ležati na prostem 'ležati na otvorenom'.

GFv2: Zaradi suše je kurjenje na prostem prepovedano, zato bodo policisti proti kršiteljem ukrepali. 'Zbog suše je zabra-

njeno paljenje na otvorenom, pa će policija preduzeti mere protiv prekršilaca’.

ChatGPT: Poleti radi obedujemo na prostem. ‘Leti volimo da jedemo na otvorenom’.

Kod značenjski, gramatički ili stilski neadekvatnih primera koje navodi ChatGPT, može se primetiti ilustrovanje doslovnog značenja naročito kod onih frazeoloških jedinica kojima je ChatGPT i inače pripisao pogrešnu definiciju (v. primer 9). Isto važi za primere kod kojih je ChatGPT inače adekvatno prepoznao značenje (v. primer 10). U svim ostalim primerima, bez obzira na to je li ChatGPT adekvatno prepoznao značenje ili nije, rečnički primeri na prvi su pogled neintuitivni i veštački, ili čak pogrešno predstavljaju značenje ili kontekst upotrebe:

(9) *pasti dol* ~ *pasti na nos*

SSKJ2: /

GFv2: Ko sem bila še aktivna mama, delavka, kmetica itd. je bil najsrečnejši trenutek dneva, ko sem pozno zvečer “padla dol” in brez vsake misli zaspala! ‘Kad sam bila još aktivna mama, radnica, seljanka itd... najsrečniji trenutak dana bio je kada bih kasno uveče pala na nos i bez ijedne misli zaspala!’

ChatGPT: Ker ni bil dovolj pripravljen, je na izpitu padel dol. ‘Pošto nije bio dovoljno pripremljen, na ispitu je pao na nos’.

(10) *hoditi narobe* ‘izazivati teškoće, nevolju’

SSKJ2: vse mi hodi narobe ‘sve mi ide naopako’.

GFv2: Nikoli mi ni hodilo narobe, da prihajam iz pravega podeželja. ‘Nikada mi nije pošlo naopako što dolazim iz pravog sela’.

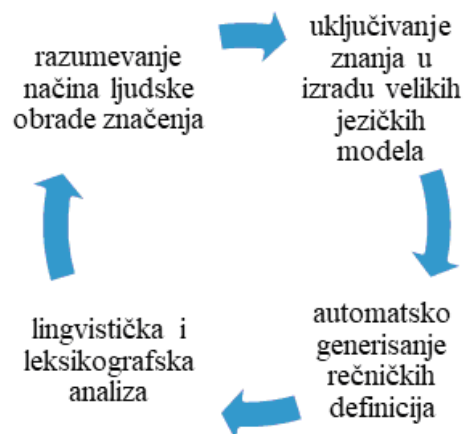
ChatGPT: Nogavice su mi drsele in sem hodil narobe, dokler nisem opazil, da sem jih obrnil na pravilen način. ‘Čarape su mi klizile i hodao sam naopako dok nisam primetio i okrenuo ih na pravu stranu’.

3. Zaključci i dalji rad

Kao što smo već pomenuli (v. odeljak 2.1), naše istraživanje ima nekoliko ograničenja, koja nam onemogućavaju donošenje opštijih zaključaka. Uprkos tome, ono je ukazalo na neke karakteristike ponašanja veštačke inteligencije pri adekvatnom prepoznavanju značenja, oblikovanju rečničkih objašnjenja i navođenju rečničkih primera, koje, ako ništa drugo, zahtevaju dodatna razmišljanja i proveravanje pretpostavki. S obzirom na prethodna istraživanja u ovoj oblasti (v. odeljak 1), i naše je istraživanje potvrdilo upotrebljivost metoda veštačke inteligencije pri prepoznavanju značenja i rečničkog opisa, i to na kompleksnim jezičkim jedinicama kao što su frazeološke jedinice. S obzirom na visok procenat semantički adekvatnih rečničkih opisa, može se zaključiti da je količina podataka — između ostalog i rečničkih — koje jezički model ima na raspolaganju ključna za njegovo uspešno funkcionisanje. Naša pretpostavka da će stepen idiomatičnosti frazeoloških jedinica i prisu-

stvo elemenata specifičnih za jezik imati suštinski uticaj na uspešnost rada ChatGPT samo je delimično potvrđena, jer ChatGPT ima dobre rezultate pri prepoznavanju značenja, kako kod semantički transparentnih tako i kod potpuno netransparentnih frazeoloških jedinica. Pri tome je zanimljiv zaključak da u slučaju pogrešnog razumevanja značenja frazeoloških jedinica ChatGPT obično nudi doslovno značenje. Zato ima smisla da, kada želimo da jezički model nauči ljudski jezik, skup primera za učenje opremimo primerima i s doslovnim i s frazeološkim značenjem. Što se tiče leksikografskog posla, upotreba veštačke inteligencije pri navođenju dobrih rečničkih primera ostaje najveći izazov. Naime, tu veštačka inteligencija daje loše rezultate čak i onda kada ispravno razume značenje frazeološke jedinice. U tom smislu, generisani primer ne dostiže kompleksne karakteristike dobrog rečničkog primera, kakav može ponuditi leksikograf — i to leksikograf (ili alatka) koji izabere realan, neizmišljen primer iz korpusa.

Šta, dakle, ostaje leksikografima? O budućnosti je inače veoma nezahvalno govoriti, ali ako pretpostavimo da će savremenom društvu još uvek biti potreban legitiman i pouzdan rečnički opis (na isti način kao što su nam sve potrebne legitimne, proverene informacije u svetu uopšte), smatramo da će uloga leksikografa s jedne strane biti usmerena na istraživanje čovekovog razumevanja značenja — mnogo više nego što je to bila do sad, a s druge strane na što više automatizovano generisanje rečničkih opisa koji će, pod nadzorom leksikografa, dostizati zadovoljavajuće rezultate (up. Prikaz 1).



Prikaz 1. Uključivanje VI u leksikografski proces.

Na osnovu leksikografske analize automatski generisanih rečničkih opisa, zaključci će moći da se upotrebe u razumevanju načina čovekovog procesuiranja značenja i da se uključe nazad u unapređenje jezičkih modela, čime se zatvara začarani krug (Prikaz 1).

Napomena

Istraživanje upotrebe veštačke inteligencije u leksikografske svrhe sprovedeno je u okviru programa istraživanja br. P6-0215 (*Slovenski jezik — osnovna, kontrastivna i primenjena istraživanja*), koje iz državnog budžeta sufinansira Javna agencija za naučna istraživanja i inovacije Republike Slovenije.

Literatura

- BARRETT, Grant (2023). Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content. Paper presented at the 24th Biennial Conference of the Dictionary Society of North America. Boulder, CO, USA, 31 May – 3 June 2023.
- DE SCHRYVER, Gilles-Maurice (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387.
- GANTAR, Polona, KREK, Simon, KOSEM, Iztok (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V: Iztok Kosem (ur.), *Kolokacije v slovenščini*. 1. izd. (str. 15–41). Ljubljana: Znanstvena založba Filozofske fakultete. <<https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6969-1>>. [5. 9. 2024]
- GAPSA, Magdalena, ARHAR HOLDT Špela, KOSEM, Iztok (2024). Kako dober je chat GPT pri umeščanju sopomenk pod pomene. V: Špela Arhar Holdt, Tomaž Erjavec (ur.), *Konferenca Jezikovne tehnologije in digitalna humanistika* (str. 144–162). Ljubljana: Inštitut za novejšo zgodovino.
- JAKUBIČEK, Miloš, RUNDELL, Michael (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography?. V: Marek Medved' i dr. (eds), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 508–523). Brno: Lexical Computing.
- KOSEM, Iztok i dr. (2024). Examining the potential of AI in the annotation of corpus examples for language learning. *15th International Corpus Linguistics Conference* (pp. 93–95). Las Palmas de Gran

Canaria, Spain 22–24 May 2024; [book of abstracts]. <<https://drive.google.com/file/d/1rHS4OwztEPvYOPwHE5Mxn-InK2ErbiHV/view>>. [5. 9. 2024]

- LEW, Robert (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), 704.
- MCKEAN, Erin, FITZGERALD, Will (2023). The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 10–20). Seoul: Yonsei University.
- RUNDELL, Michael (2023). Automating the Creation of Dictionaries: Are We Nearly There?. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 1–9). Seoul: Yonsei University.

Izvori

- KEBER, Janez (2015). *Slovar slovenskih frazemov*. Elektronska objava. Spletna izdaja na www.fran.si. Ljubljana 2015, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. <www.fran.si/192/janez-keber-frazeoloski-slovar-slovenskega-jezika>. [5. 9. 2024]
- KREK, Simon, ERJAVEC, Tomaž, REPAR, Andraž, ČIBEJ, Jaka, ARHAR HOLDT, Špela, GANTAR, Polona, KOSEM, Iztok, ROBNIK ŠIKONJA, Marko, LJUBEŠIČ, Nikola, DOBROVOLJC, Kaja, LASKOWSKI, Cyprian Adam, GRČAR, Miha, HOLOZAN, Peter, ŠUSTER, Simon, GORJANC, Vojko, STABEJ, Marko, LOGAR, Nataša (2019). *Corpus of written standard Slovene Gigafida 2.0*. Ljubljana: Centre for Language Resources and Technologies, University of Ljubljana. CLARIN. SI data & tools. <<http://hdl.handle.net/11356/1320>>. [5. 9. 2024]
- SSKJ2: *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja. Spletna izdaja na <www.fran.si>. [5. 9. 2024]
- VSMS: *Veliki slovensko-madžarski slovar*, <<https://viri.cjvt.si/slovensko-madzarski/slv/>>. [5. 9. 2024]

Apolonija Gantar

Generating Dictionary Definitions Using Artificial Intelligence on the Example of Slovenian Phraseological Units

(Summary)

In this paper, we discuss the possibility of using artificial intelligence to test the performance of generative methods in creating dictionary definitions and obtaining good dictionary examples for semantically and structurally complex phraseological units (PU), which are defined as multi-word lexical units that express a certain degree of structural stability and semantic opacity and have a noticeable expressive or pragmatic role in the language (GANTAR et al. 2021). The test study included 35 structurally, semantically (polysemous, varying degrees of idiomaticity) and culturally different phraseological units. We then asked ChatGPT to produce dictionary definitions for each PU and to illustrate each PU meaning with a dictionary example. The answers obtained were analyzed regarding the adequacy of the meaning recognition based on the generated definition, and the results were compared with the definitions in existing dictionaries. We independently analyzed the performance of ChatGPT in generating good dictionary examples. The analysis showed ChatGPT's very good performance in generating dictionary definitions. In as many as 60% of the cases, ChatGPT correctly identified at least one PU meaning and provided a good dictionary definition. Compared to dictionary definitions in existing dictionaries, the automatically generated definitions were even more precise and meaningful in some cases. Starting from the assumption that the degree of idiomaticity and cultural and linguistic specificity of PU has a crucial influence on the performance of ChatGPT in generating dictionary definitions, we compared the performance of ChatGPT with the degree of idiomaticity of PUs. The results showed that ChatGPT performs equally well in meaning recognition for linguistically and culturally specific PUs as for semantically relatively transparent ones. Given that ChatGPT defines literal usage when meaning is inadequately recognized, it is reasonable to consider literal examples when learning language models. ChatGPT performs worse when providing dictionary examples, even in cases where the meaning of the PU is correctly identified. From the point of view of use in the lexicographic process, tools for extracting dictionary examples from the corpus, such as the Gdex function within SketchEngine, remain more valuable. We conclude the paper with the view that using generative intelligence methods when describing PU in a dictionary makes sense and that more targeted linguistic prompts could improve the generated definitions. Preliminary research has further shown that the degree of idiomaticity does not significantly impact the performance of the AI, and we assume that the quality of the results depends mainly on the amount of texts containing different types of metaphorical language use available to the language model.