SPLOŠNE INFORMACIJE

Naslov: Primerjalna semantično-pragmatična analiza intenzitetnih označevalcev v avtorskih in umetno generiranih presojevalnih besedilih : raziskovalni podatki

Avtorica: Diana Košir

ORCID: https://orcid.org/0009-0009-4428-9698

Ustanova: Znanstveno-raziskovalno središče Koper, Inštitut za jezikoslovje

E-mail: diana.kosir@zrs-kp.si

Trajanje raziskave: 2026

Financiranje: ARIS P5-0409 Razsežnosti slovenstva med lokalnim in globalnim v začetku tretjega tisočletja

MOŽNOSTI DOSTOPA

Izvleček:

Pilotna primerjalna raziskava obravnava razlike v rabi nekaterih intenzitetnih označevalcev (izbranih prislovov in členkov v vlogi modifikacije intenzitete sporočila v smeri krepitve oz. šibitve) v avtorskih presojevalnih besedilih ter besedilih, ustvarjenih z modeloma ChatGPT-5 (brezplačna verzija) in GaMS-12B-Instruct. Analiza delovnega korpusa (36 kolumn in blogov) je pokazala, da je za avtorska besedila značilno bolj raznoliko besedišče (označevalci intenzitete), zlasti izrazi, ki izražajo subjektivno vrednotenje tvorca sporočila. Za ChatGPT je značilen previdnejši in bolj zadržan slog z izrazi, kot so *morda, verjetno, lahko* in *pogosto*, medtem ko GaMS pogosteje uporablja pomenske presežnike, npr. *vedno*/*stalno*, *najbolj*, in manj šibilcev, zato deluje bolj odločno, okrepljeno in posplošujoče. Rezultati kažejo, da se oba modela v rabi intenzitetnih označevalcev razlikujeta od avtorskih besedil, pa tudi drug od drugega.

Ključne besede: veliki jezikovni modeli (VJM), ChatGPT, GaMS, kolumna in blog, intenziteta jezika, pragmatika in semantika, korpusna analiza

Opis raziskovalnih podatkov in metodologija:

Raziskovalni podatki so delovni korpus »Korpus avtorskih in umetno generiranih presojevalnih besedil«, ki vsebuje 36 dokumentov oz. 28.908 besed in je bil izdelan z orodjem Sketch Engine. Nabor korpusnega gradiva je potekal v obdobju od januarja do maja 2026.

Najprej je bilo v obdobju januar–marec 2026 izbranih 12 avtorskih kolumn oz. blogov, po 6 na sorodno temo, in sicer (1) kakovost življenja v sodobni družbi in (2) svoboda govora. Besedila obsegajo med 600 in 1000 besed, gre pa za kolumne avtorjev Kozme Ahačiča, Grege Repovža, Petre Božič Blagajac, Andreja Velkavrha, Matjaža Figlja, Luke Martina Tomažiča, Marka Crnkoviča in Tomaža Štiha ter dva bloga Matica Munca (natančni bibliografski podatki so navedeni v Excelovi datoteki »Korpus_metapodatki«).

Sledilo je generiranje besedil z UI (ChatGPT, GaMS), upoštevaje ugotovljene lastnosti avtorskih besedil (tj. žanr, obseg, tema, stil). V obdobju april–maj 2026 je bilo skupno generiranih 24 besedil, po 12 z vsakim modelom.

Najprej je bil uporabljen poziv A in ugotovili smo, da so rezultat besedila brez humornih insertov in slovenskih družbeno-kulturnih referenc, ki so bili sicer prepoznani v avtorskih besedilih, zato sta bili vpeljani še variaciji poziva B in C. Različni pozivi (angl. *prompts*) so navedeni v spodnji tabeli 1.

Vrsta poziva	**Poziv**
Nevtralen (A)	**Tema 1**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o kakovosti življenja v današnji družbi. Besedilo naj obsega 600 do 1000 besed. **Tema 2**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o svobodi govora. Besedilo naj obsega 600 do 1000 besed.
Družbeni kontekst (B)	**Tema 1**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o kakovosti življenja v današnji družbi. Vključi slovenski družbeno-kulturni kontekst. Besedilo naj obsega 600 do 1000 besed. **Tema 2**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o svobodi govora. Vključi slovenski družbeno-kulturni kontekst. Besedilo naj obsega 600 do 1000 besed.
Humoren slog (C)	**Tema 1**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o kakovosti življenja v današnji družbi. Besedilo naj bo napisano v humornem slogu. Besedilo naj obsega 600 do 1000 besed. **Tema 2**: Napiši kolumno oz. blog, ki je vrednotenjsko besedilo, o svobodi govora. Besedilo naj bo napisano v humornem slogu. Besedilo naj obsega 600 do 1000 besed.

Tabela 1: Različni pozivi

Na isto temo (kakovost življenja v sodobni družbi; svoboda govora) je bilo z vsakim modelom generiranih šest besedil, z vsakim pozivom po dve. Strukturo delovnega korpusa prikazuje tabela 2.

Vir	Tema	Poziv	Število ponovitev	Skupaj besedil (N)
Jezikovni model (ChatGPT; GaMS)	Tema 1: *Kakovost življenja v sodobni družbi*	Poziv A – brez usmeritve o slogu	2	24
		Poziv B – kulturno občutljiv		
		Poziv C – humorno obarvan slog		
	Tema 2: *Svoboda govora na družbenih omrežjih*	Poziv A		
		Poziv B		
		Poziv C		
Spletni/tiskani medij	Tema 1	/	/	12
	Tema 2			

Tabela 2: Struktura delovnega korpusa.

Metapodatki o podatkovnem vnosu

Splošne informacije:

- **Jezik**: slovenščina
- **Tip besedila**: presojevalna besedila (kolumna, blog)
- **Vir podatkov**: glej tabelo Excel
- **Seznam datotek**:

- **PreberiMe.pdf**: datoteka s splošnimi ter metodološkimi informacijami in opisom podatkov
- **Korpus avtorskih in UI besedil.zip**: komprimirana datoteka, ki vsebuje tri mape z različnimi podkorpusi besedil glede na avtorstvo oziroma izvor: »Avtorska besedila«, »Besedila ChatGPT« in »Besedila GaMS« (v vsaki mapi je 12 besedilnih datotek, ki so poimenovane glede na avtorstvo, temo in zaporedno številko ponovitve pri generiranih besedilih, npr. *Avtor_Kakovost_življenja_4*, *ChatGPT_Kakovost_življenja_5*, *GaMS_Svoboda_govora_1*, pri čemer pri UI besedilih

zaporedne številke 1–2 pomenijo generiranje besedil s pozivi A, 3–4 uporabo pozivov B s slovenskim družbeno-kulturnim kontekstom, 5–6 pa uporabo pozivov C s humornim slogom)

- **Korpus_metapodatki.xlsx**: Excelov dokument z metapodatki o besedilih, vključenih v analizo
- **Frekvencna lista_clenek.pdf**: seznam členkov (35) z absolutnimi in relativnimi frekvencami v korpusu
- **Frekvencna lista_prislov.pdf**: seznam 50 najpogostejših prislovov z absolutnimi in relativnimi frekvencami v korpusu
- **Način zbiranja in urejanja podatkov**: ročni, strojni (uporaba generativne UI); frekvenčne liste (Sketch Engine), ročna selekcija in analiza konkordanc

Statistika korpusa

- **Število vseh znakov**: 34.394
- **Število besed**: 28.908
- **Število stavkov**: 2.050
- **Število dokumentov**: 36

Leksična analiza:

- **Velikost leksikona**:
  - **Besede**: 7.302
  - **Leme**: 4.007
  - **POS-tagi (označene besedne vrste)**: 666

Tagging in označevanje morfoloških lastnosti:

POS-tagi vključujejo slovnične oznake, kot so:

- noun S.*
- verb G.*
- adjective P.*
- adverb R.*
- pronoun Z.*

- adposition D.*
- conjunction V.*
- particle L.*
- interjection M.*

Povzetek:

Podatki so rezultat pilotne primerjalne raziskave avtorskih in s pomočjo UI generiranih presojevalnih besedil, da bi ugotovili, ali se UI besedila, ki so po lastnostih sorodna izbranim avtorskim presojevalnim besedilom (kolumna, blog), razlikujejo z vidika intenzitete jezika, kar se odraža tudi v rabi besednih intenzitetnih označevalcev (po modelu Mikolič 2020). Namen primerjalne korpusne analize (sicer manjšega delovnega korpusa) avtorskih in umetno generiranih besedil z jezikovnima modeloma ChatGPT-5 (brezplačna verzija) in GaMS-12B-Instruct je ugotoviti, ali je slovenski jezikovni model glede na raznolikost in semantično ustreznost rabe intenzitetnih označevalcev (že) bliže slovenskim avtorskim besedilom v primerjavi z globalnim modelom ChatGPT, kar bi lahko bil eden od kazalnikov jezikovne kompetence.

Za potrebe raziskave je s pomočjo orodja Sketch Engine nastal manjši delovni korpus »Korpus avtorskih in umetno generiranih presojevalnih besedil«, ki je sestavljen iz treh podkorpusov: (1) 12 avtorskih kolumn oz. blogov, (2) 12 besedil, generiranih z jezikovnim modelom ChatGPT in (3) 12 besedil, generiranih s slovenskim modelom GaMS. S pomočjo orodja Sketch Engine so bili identificirani najpogostejši besedni izrazi, ki se v besedilih nastopajo v vlogi intenzitetnih označevalcev (krepilci, šibilci). V analizi smo se omejili na najpogostejše prislove in členke v tej vlogi med prvimi 50 najpogostejšimi prislovi oz. členki na frekvenčni listi korpusnih lem.

Analiza je podala podobne rezultate, kot nekatere študije na anglosaškem gradivu, da obstajajo razlike v rabi intenzitetnih označevalcev med avtorskimi in s ChatGPT-jem generiranimi besedili (pri nas tudi z GaMS-om). Pri avtorskih besedilih se je pokazala večja raznolikost rabljenih izrazov v tej vlogi, zlasti so za avtorska besedila značilni izrazi, ki odražajo subjektivno vrednotenje tvorca, ki pa niso značilni za naša UI generirana besedila. Razlika v pogostosti rabe označevalcev intenzitete se je pokazala tudi pri primerjavi dveh jezikovnih modelov. V stilu besedil ChatGPT izstopa prisotnost izrazov *morda*, *verjetno*, *lahko* in *pogosto*, ki kažejo na težnjo k previdnemu, zadržanemu in posplošujočemu izražanju (izogibanje premočnim sodbam), kar sicer ni značilnost izbrane vrste presojevalnih besedil (kolumna, blog). Obratno se je pri besedilih modela GaMS pokazala visoka prisotnost izrazov *vedno*/*stalno*, *najbolj*, ki so pomenski presežniki, izrazito malo pa je šibilcev, kar kaže na odločen, okrepljen, a tudi posplošujoč stil (skrajnostno prikazovanje resničnosti).