Mojca Kompara Lukančič

Prepoznavanje krajšav in razvezav v angleških besedilih s področja varstvoslovja z algoritmom Krajšavar ter orodjema ChatGPT in Perplexity

Cobiss: 1.01
https://doi.org/10.3986/JZ.31.2.05

V prispevku se osredotočamo na uporabo orodij umetne inteligence, in sicer ChatGPT ter Perplexity, v procesu prepoznavanja krajšav in razvezav v angleških besedilih stroke, in sicer na primeru varstvoslovnih besedil, ter izsledke primerjamo s filtracijo besedil z algoritmom Krajšavar. Angleška besedila, ki smo jih uporabili pri filtraciji z orodjema umetne inteligence in algoritmom, smo ročno zbrali na podlagi tipološke klasifikacije angleških varstvoslovnih besedil. V prispevku predstavimo značilnosti algoritma Krajšavar, opišemo njegov razvoj in delovanje ter v nadaljevanju orišemo zbiranje besedil in pripravo gradiva za filtracijo. Osredinimo se na uporabo orodij ChatGPT in Perplexity pri samodejnem prepoznavanju krajšavno-razvezavnih parov v angleških varstvoslovnih besedilih, preverimo izsledke filtracije ter jih primerjamo z izsledki, pridobljenimi z algoritmom Krajšavar, in z rezultati ročnega pregleda.
Ključne besede: krajšave, angleščina, slovar, algoritem, umetna inteligenca

Recognizing Abbreviations and Their Expansions in English Criminal Justice and Security Texts Using the Krajšavar Algorithm, ChatGPT, and Perplexity

This article examines the application of artificial intelligence tools, specifically ChatGPT and Perplexity, for recognizing abbreviations and their corresponding expansions in English professional texts dealing with criminal justice and security. The results are compared with those obtained through text filtering using the Krajšavar algorithm. The English texts used for filtering with both AI tools and the algorithm were manually collected based on a typological classification of English criminal justice and security texts. This article presents the main features of the Krajšavar algorithm, outlines its development and functioning, and describes the text collection and material preparation for filtering. The analysis focuses on using ChatGPT and Perplexity for the automatic recognition of abbreviation–expansion pairs in English criminal justice and security texts. The filtering results are evaluated and subsequently compared with those obtained through the Krajšavar algorithm and manual verification.
Keywords: abbreviations, English, dictionary, algorithm, artificial intelligence

Mojca Kompara Lukančič ▪ Univerza v Mariboru, Fakulteta za turizem – Fakulteta za varnostne vede ▪ mojca.kompara@um.si ▪ https://orcid.org/0000-0003-2368-4161

Prispevek temelji na raziskovalnih podatkih, ki se hranijo v Digitalni knjižnici Univerze v Mariboru in so javno dostopni na povezavi https://dk.um.si/IzpisGradiva.php?lang=slv&id=95507. Predstavljena raziskava je bila podprta z nepovratnimi sredstvi programa Erasmus+, št, projekta 2024-1-RO01-KA220-HED-000248845, ki jih financira EU. Izražena stališča in mnenja so izključno stališča avtorice in ne odražajo nujno stališč EU ali ANPCDEFP. Niti EU niti ANPCDEFP ne moreta biti odgovorni zanje.

1 Uvod

Položaj krajšav v slovenskem prostoru je bil podrobneje predstavljen v delih, ki so se osredinjala na njihovo prisotnost v normativnih priročnikih (gl. Kompara Lukančič 2018; Verovnik 2023) in slovarjih (Kompara Lukančič 2010). Že Gabrovšek (1994: 164) je izpostavil, da so s krajšavami »križi in težave«, da nastajajo hitro in da je pomembno, da so slovarji krajšav čim bolj novi. Hitro nastajanje krajšav, njihovo vključevanje v pravopise, splošne in specializirane slovarje ter priprava samostojnih krajšavnih slovarskih zbirk so privedli do uporabe računalniških orodij, priprave algoritmov za samodejno prepoznavanje krajšav in razvezav v besedilih (njihov pregled sledi v nadaljevanju; gl. Kompara Lukančič 2018) ter uporabe umetne inteligence pri prepoznavanju krajšav in razvezav.

Uporabo računalniških orodij v slovenskem leksikografskem prostoru zaznamo že pri izdaji *Velikega nemško-slovenskega slovarja* (Debenjak – Debenjak – Debenjak 1992). Izpostaviti je treba programa STEVE in EVE oz. EVA, ki ju je razvil Primož Jakopin in s pomočjo katerih so bili izdani *Odzadnji slovar slovenskega knjižnega jezika po Slovarju slovenskega knjižnega jezika* (Hajnšek-Holz – Jakopin 1996), *Slovar govorov Zadrečke doline med Gornjim Gradom in Nazarjami (A–H)* (Weiss 1998), *Etimološki slovar slovenskega jezika* (Bezlaj 1995), *Slovenski etimološki slovar* (Snoj 1997) in *Slovenski pravopis* (2001). Uporabo jezikovnih tehnologij pri sestavi slovarjev so izpostavili številni avtorji; gre za avtomatizirane procese, ki so omogočili premik od klasičnega ročnega slovaropisja k delno avtomatiziranemu (Weiss 1991; Humar 2004; Košmrlj-Levačič – Seliškar 2004; Kosem – Gantar – Krek 2013; Rundell 2023). Uporaba jezikovnih tehnologij je bila ključna pri zasnovi najsodobnejših slovenskih digitalnih leksikografskih virov, in sicer *Slovarja sopomenk sodobne slovenščine* (2017), *Kolokacijskega slovarja sodobne slovenščine* (2018), *Velikega slovensko-madžarskega slovarja* (Kosem 2024), *Digitalne slovarske baze za slovenščino* (2023) itn. O vključevanju jezikovnih tehnologij v slovaropisje in samodejni pripravi geselskih člankov pišeta tudi Kompara Lukančič in Holozan (2011), ki izpostavljata samodejni pristop pri pripravi slovarjev krajšav. Gre za pripravo in uporabo algoritma Krajšavar, ki omogoča samodejno prepoznavanje krajšav in razvezav v elektronskih besedilih (za podroben pregled algoritmov, ki so bili pripravljeni za filtracijo angleških besedil, gl. Kompara Lukančič 2018); z razvojem tega algoritma se avtorica tega prispevka ukvarja že več let (Kompara Lukančič 2010; 2011; 2018). Pojav umetne inteligence v leksikografiji je sprva zamajal leksikografsko skupnost (de Schryver 2023; Jakubíček – Rundell 2023; Vossen 2022; Lew 2023), a kmalu privedel do uporabe orodij umetne inteligence pri sestavi geselskih člankov (Lew 2023). V tem prispevku se ukvarjamo z njeno uporabo pri prepoznavanju krajšav in razvezav. Namen prispevka je na podlagi posodobljenega algoritma za prepoznavanje

krajšav in razvezav (Krajšavar) iz angleških besedil, ki sodijo v področje varstvoslovja, na podlagi besedilne tipologije za varstvoslovna besedila pridobiti krajšavno-razvezavne pare in primerjati pridobljene izsledke z izsledki orodij umetne inteligence (ChatGPT in Perplexity).[1]

2 Krajšavar – algoritem za samodejno prepoznavanje krajšav in razvezav v elektronskih besedilih

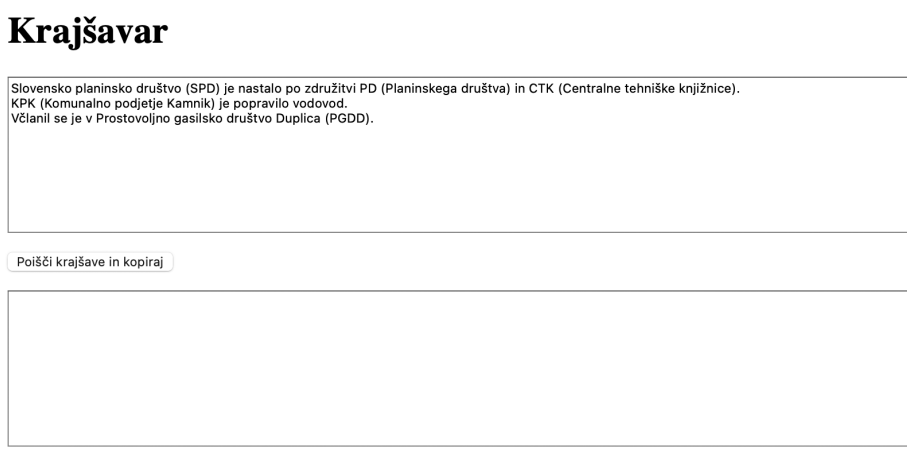
Za krajšave se je družba zanimala že v času Cicera (Kompara Lukančič 2018). Morda je njihova tipološka značilnost pripomogla k temu, da so se z njihovim samodejnim zbiranjem že pred več kot dvema desetletjema začeli ukvarjati predvsem v angleškem prostoru, in sicer s pojavom algoritmov za prepoznavanje krajšav in razvezav oz. pomenov v elektronskih besedilih. V Kompara Lukančič 2018 so podrobneje predstavljene značilnosti posameznih algoritmov, in sicer se delijo glede na značilnosti prepoznavanja krajšav in razvezav. Taghva in Gilbreth (1999) prepoznavata akronime, ki so zapisani z velikimi tiskanimi črkami in obsegajo od tri do deset znakov, razvezave pa črpata iz sobesedila, pri čemer imajo ključno vlogo začetne črke. Yeates (1999) prav tako prepoznava akronime, zapisane z velikimi črkami, in razvezave prepozna iz sobesedila, pri čemer so omejene na začetne tri črke. Larkey idr. (2000) prepoznavajo akronime, zapisane z velikimi črkami, a dopuščajo tudi nabor izjem v smislu ostalih krajšavnih tipov, in sicer do največ devet znakov; razvezave prepoznava po vzorcu akronim (razvezava) ali razvezava (akronim). Byrd in Park (2001) prepoznavata akronime, zapisane z velikimi tiskanimi črkami, pri čemer mora biti velika tiskana vsaj ena črka; vključujeta tudi številke in akronime, dolge od dva do deset znakov. Razvezave iščeta po vzorcu akronim (razvezava) ali razvezava (akronim). Schwartz in Hearst (2003) prepoznavata akronime, zapisane v oklepaju ali zunaj njega; ti vsebujejo od dva do deset znakov; razvezave iščeta po vzorcu akronim (razvezava) ali razvezava (akronim). Zahariev (2004) prepoznava krajšave v oklepaju ali zunaj njega, razvezave pa po vzorcu akronim (razvezava) ali razvezava (akronim). Jun Xu Yalou in Huang (2005) prepoznavata akronime, zapisane z velikimi tiskanimi črkami, v sestavi od dveh do desetih znakov, razvezave pa po vzorcu akronim (razvezava) ali razvezava (akronim). Zhou, Torvik in Smalheiser (2006) prepoznavajo akronime, zapisane z velikimi tiskanimi črkami v oklepajih ali zunaj njih in z nekaj izjemami krajšavnih tipov; razvezave prepoznavajo po vzorcu akronim (razvezava) ali razvezava (akronim). Šateva in Nikolov (2008) prepoznavata akronime, zapisane z velikimi tiskanimi črkami z do petimi znaki; razvezave prepoznavata iz sobesedila

1 Raziskovalni podatki (Kompara Lukančič 2025b) so na voljo na naslednji povezavi: https://dk.um.si/IzpisGradiva.php?lang=slv&id=95507.

po vzorcu akronim (razvezava) ali razvezava (akronim). Kuo idr. (2009) krajšave prepoznavajo tudi v oglatih oklepajih, pri čemer je lahko v oklepaju krajšava ali pa razvezava; uvajajo tudi uporabo vejice, pri čemer sta krajšava in razvezava ločeni z vejico. Gelernter in Balaji (2013) se usmerjata v mikrobesedila, v katerih so tudi krajšave, ki se pojavljajo v imenih krajev, ulic in cest, in sicer v kratkih tekstovnih sporočilih. Wu idr. (2015) pri prepoznavanju krajšav preučujejo uporabo nevronskih vektorskih predstavitev besed za razločevanje krajšav v kliničnem kontekstu v okviru treh metod, in sicer vektorske značilnosti na podlagi okoliškega besedila, leve in desne vektorske značilnosti na podlagi okoliškega sobesedila ter maksimalne vektorske značilnosti na podlagi okoliškega sobesedila. Liu idr. (2017) se ukvarjajo s pridobivanjem razvezav na podlagi označevanja zaporedij v smislu pogojnih naključnih polj. Montalvo idr. (2018) predlagajo pet sistemov za prepoznavanje krajšavno-razvezavnih parov in tri sisteme za prepoznavanje krajšav brez razvezav v sobesedilu. Veyseh idr. (2020; 2022) prepoznavajo krajšave v znanstvenih besedilih, in sicer z metodo identifikacije in razločevanja. Osredotočajo se predvsem na dvoje: ekstrakcijo in razločevanje krajšav v nizu tujih jezikov, in sicer na primeru pravnih in znanstvenih besedil, ob pomoči označevanja in ob uporabi nevronskih mrež. Huang idr. (2022) se usmerjajo k prepoznavanju krajšav in razvezav na podlagi začetnih črk, ob predpostavki, da razvezave stojijo v neposredni bližini krajšave.

Vsem algoritmom je skupno, da krajšave in razvezave prepoznavajo v angleških elektronskih besedilih, z izjemo algoritma, ki ga je pripravila avtorica tega prispevka (Kompara Lukančič 2011; 2018), saj gre za prvi algoritem, ki krajšave in razvezave prepoznava v slovenskih elektronskih besedilih. Tu omenimo, da je bil algoritem iz leta 2011 pripravljen za filtracijo slovenskih besedil, algoritem Krajšavar pa je bil prirejen za filtracijo angleških besedil. Priprava slovenskega algoritma za prepoznavanje krajšav in razvezav se je pričela leta 2009 (Kompara Lukančič 2009), in sicer je bil v prvi fazi namenjen filtraciji slovenskih elektronskih besedil. V naslednjih letih se je avtorica z razvojem algoritma usmerila tudi v zametke filtracije tujih besedil, in sicer na primeru angleških in italijanskih (Kompara Lukančič 2011). Algoritem je bil zasnovan tako, da je prepoznaval samo krajšavno-razvezavne pare, saj je bila končni cilj algoritma sestava geslovnika za pripravo slovarja krajšav. Torej, če krajšava v besedilu ni imela razvezave, je algoritem ni prepoznal, prav tako velja za razvezavo, ki v besedilu ni imela krajšave. Skladno s pravopisnimi pravili, da se krajšava ob prvi pojavitvi razveže, smo se odločili, da algoritem deluje tako, da najde tiste krajšavno-razvezavne pare, ki stojijo pred ali za krajšavo in so v oklepaju ali pa je v oklepaju krajšava. Algoritem preskoči vse krajšave, ki nimajo razvezave, prav tako preskoči krajšave, zapisane s posebnimi znaki, npr. pomišljajem ipd. Algoritem je bil ponovno posodobljen leta 2024 ob pomoči informatika dr. Petra Holozana. Ta je razvil posodobljeno

različico algoritma (gl. sliko 1), ki omogoča filtracijo obsežnejše količine besedil ter ekstrakcijo krajšav in razvezav v angleških besedilih. Algoritem je bil pripravljen na podlagi tipoloških značilnosti angleških krajšav (gl. Kompara Lukančič 2023a).



Slika 1: Krajšavar – algoritem za samodejno prepoznavanje krajšav in razvezav v elektronskih besedilih

Končna priprava algoritma za samodejno prepoznavanje krajšav in razvezav v angleških besedilih zajema nabor korakov, ki se začne s pripravo osnovnih pravil – gre predvsem za upoštevanje tipoloških značilnosti angleških krajšav (gl. Kompara Lukančič 2011), ki so ključne pri prepoznavanju krajšav in v nadaljevanju razvezav. Sledita priprava pravil za prepoznavanje razvezav in končna implementacija v uporabniku prijazni digitalni obliki. Algoritem Krajšavar je bil sprva pripravljen za interno rabo in testiranje besedil. Kot je razvidno s slike 1, ga sestavljata dve okni: v prvo se vnese besedilo za filtracijo, v drugem oknu pa se nato pojavijo krajšavno-razvezavni pari. V ozadju preprostega uporabniškega vmesnika poteka kompleksno delovanje algoritma, ki je opisano v nadaljevanju.

V prvi fazi algoritem razdeli besedilo po posameznih besedah, pri čemer tudi ločila obravnava kot posamezne besede, in sicer zgolj zaradi postopka ločevanja. Tako kot predhodno opisani algoritmi krajšave in razvezave išče v nizu krajšava (razvezava), (krajšava) razvezava, razvezava (krajšava), (razvezava) krajšava. Krajšave in/ali razvezave torej išče levo ali desno od oklepaja. Algoritem filtrira besede iz besedila in išče uklepaj. Ko ga najde, nadaljuje in išče prvi zaklepaj, ki mu sledi. Če se med uklepajem in zaklepajem pojavi zgolj ena beseda, ki je zapisana z veliko začetnico ali je v celoti zapisana z velikimi tiskanimi črkami

in je sestavljena iz vsaj dveh črk, algoritem predpostavlja, da gre za krajšavo. Algoritem išče v poljubnem nadaljevanju besedila in ni zamejen na isto poved. Pri iskanju razvezav začne z iskanjem besed, ki stojijo pred ali za oklepajem, in poskuša s sovpadanjem do desetih besed, ki stojijo pred ali za oklepajem. V nadaljevanju išče prvo besedo, ki sovpada s prvo črko v krajšavi, ki sledi uklepaju in je zapisana z veliko začetnico. Algoritem išče sovpadanje besed, ki stojijo levo ali desno od oklepaja, in sicer do deset besed, pri čemer preveri sovpadanje od prve do desete besede, saj utegnejo biti vmes besede, ki se ne krajšajo in kot take niso prisotne v krajšavi, npr. predlogi, vezniki. Tu je treba podariti, da ta postopek velja, če je v oklepaju krajšava. Če je v oklepaju razvezava, algoritem išče krajšavo, ki stoji levo ali desno od oklepaja. Algoritem kot rezultat vrne besedilo od najdene ustrezne prve besede do vključno zaklepaja ne glede na druge besede v tem nizu, zapis začetnic teh besed in neprve črke v krajšavi. Na ta način najde tudi razvezave, v katerih nastopajo dodatne besede, npr. vezniki, predlogi – *Univerza v Ljubljani (UL)*, in krajšave, ki vsebujejo črke, ki niso začetnice besed v razvezavi, npr. *Andragoški center Republike Slovenije (ACS)*. Omeniti velja še tipološke značilnosti angleških krajšav, ki so drugačne od slovenskih (gl. Kompara Lukančič 2011), npr. da z veliko začetnico zapisujejo naslove.

3 Metodologija

V prispevku se osredotočamo na uporabo algoritma Krajšavar pri filtraciji angleških besedil s področja varstvoslovja in izsledke primerjamo z izsledki filtracije, pridobljene z orodjema ChatGPT in Perplexity. Področje varstvoslovja je z vidika jezikovnega raziskovanja, razvoja terminologije, prevajanja ipd. v slovenskem prostoru močno podhranjeno (Kompara Lukančič 2023b), primanjkuje referenčnih gradiv, ki so po večini zastarela, pa tudi slovarjev in glosarjev. Podrobno se je z varstvoslovjem ukvarjala Kompara Lukančič, ki je obenem opozorila na vrzel, ki jo je treba zapolniti, saj je razvoj jezika stroke ključen za obstoj jezika. Avtorica se je z jezikom stroke dotaknila tudi tipološke klasifikacije varstvoslovnih besedil (gl. Kompara Lukančič 2023; 2025), ki je po njenem mnenju nujna za sistematično zbiranje in nadaljnjo analizo varstvoslovnih besedil ter poznavanje značilnosti varstvoslovnega strokovnega jezika. Avtorica varstvoslovna besedila deli glede na tipološko klasifikacijo, ki jo je pripravila med letoma 2023 in 2025 in ki temelji na klasifikaciji turističnih besedil, ki jo je razvila Mikolič (2007). Kompara Lukančič (2025) je tudi mnenja, da je klasifikacija, ki jo je pripravila Mikolič (2007), univerzalna, torej uporabna tudi za druga področja strokovnega jezika in druge jezike. Kompara Lukančič (Kompara Lukančič – Smajla 2025) varstvoslovna besedila v prvi vrsti deli na tista, ki so namenjena javnosti, in tista, ki vsebujejo tajne podatke

in so zato interne narave. Besedilna tipologija s področja varstvoslovja je podrobneje opisana v Kompara Lukančič – Smajla 2025, v katerem se avtorica dotakne pomena tipologije, predvsem v smislu sistematičnega zbiranja besedil, opazovanja značilnosti posameznih besedil in pridobivanja nabora besedil za potrebe njihove filtracije z algoritmom Krajšavar. Kompara Lukančič (Kompara Lukančič – Smajla 2025) pri svoji tipološki klasifikaciji področje varstvoslovja najprej razdeli na podpodročja varnosti, pravosodja, policije, kriminalistike, prava, zakonodaje in vojske. Kategorizacijo tipologije varstvoslovnih besedil razdeli glede na namen, referenco in medij ter zaradi preglednosti podpodročja združi v varnost, policijo, pravo in vojsko (Kompara Lukančič – Smajla 2025). Glede na tipološko kategorizacijo varstvoslovnih besedil besedila razdeli na (1) pravna besedila s področja varnosti, policije, prava in vojske, (2) znanstvena besedila s področja varnosti, policije, prava in vojske, (3) strokovna in poljudnoznanstvena besedila s področja varnosti, policije, prava in vojske, (4) publicistična besedila s področja varnosti, policije, prava in vojske, (5) splošna besedila s področja varnosti, policije, prava in vojske ter (6) promocijska besedila s področja varnosti, policije, prava in vojske. Klasifikacija varstvoslovnih besedil (gl. Kompara Lukančič – Smajla 2025) nam je omogočila sistematično zbiranje besedil po posameznih kategorijah, in sicer smo za potrebe naše analize zbrali angleška besedila, ki sodijo v podpodročja varstvoslovja, ki jih omenja Kompara Lukančič 2023b.

Besedila smo pridobili ročno s spleta, in sicer z vnosom angleških ključnih besed s področja varstvoslovja, tj. *police* 'policija', *crime* 'kriminal', *criminal justice* 'varstvoslovje', *army* 'vojska' in '*security* 'varnost'. Besedila smo iskali skladno s tipološkimi značilnostmi, in sicer na svetovnem spletu v iskalniku Google ter Google Učenjak; slednjega smo uporabili predvsem za pridobivanje znanstvenih besedil. Zbrana besedila smo nato filtrirali z algoritmom za samodejno prepoznavanje krajšav in razvezav Krajšavar ter pridobili angleške krajšavno-razvezavne pare s področja in podpodročij varstvoslovja. V prispevku se osredotočamo na primerjavo izsledkov filtracije besedil z algoritmom Krajšavar ter z orodjema ChatGPT in Perplexity. Za potrebe raziskave smo se želeli osrediniti na besedila, ki sodijo v enotno tipološko kategorijo, zato smo se odločili za znanstvene prispevke, tj. znanstvena besedila s področja varnosti, policije, prava in vojske (kategorija 2), ker so ta praviloma podobne oz. omejene dolžine in niso predolga oz. prekratka za ročno analizo vsebnosti krajšavno-razvezavnih parov. Skupno smo filtrirali 21 prispevkov.[2] SNamen raziskave je prikazati, kako orodji umetne inteligence ChatGPT in Perplexity delujeta pri prepoznavanju krajšav in razvezav. Uporabili smo prosto dostopno, neplačljivo različico orodja ChatGPT, ki je bila na voljo marca 2025, in ob vnosu napotkov naložili besedila oz. jih prekopirali v

2 Seznam prispevkov je dostopen v okviru raziskovalnih podatkov (Kompara Lukančič 2025), gl. op. 1.

iskalnik ter preverili točnost pri prepoznavanju krajšav in razvezav v angleških besedilih. Da smo primerjali rezultate, pridobljene z orodjem ChatGPT, smo uporabili še plačljivo različico orodja Perplexity, ki je bila na voljo marca 2025. Pri obeh orodjih smo uporabili enak napotek, in sicer niz v angleškem jeziku »In the attached text find abbreviations and their meanings/expansions« (sl. V priloženem besedilu poišči vse krajšave in njihove pomene). Nato smo primerjali izsledke, ki smo jih pridobili s filtracijo besedil z algoritmom Krajšavar ter orodjema ChatGPT in Perplexity.

4 Izsledki raziskave

Raziskava je temeljila na filtriranju besedil iz druge tipološke kategorije (Kompara Lukančič – Smajla 2025), tj. znanstvena besedila s področja varnosti, policije, prava in vojske. Za potrebe raziskave smo podrobneje analizirali 21 besedil, kar predstavlja 10 % vseh besedil, ki so podrobneje predstavljena v Kompara Lukančič 2025). V teh 21 besedilih je bilo po ročnem pregledu najdenih 99 krajšavno-razvezavnih parov, od katerih je Krajšavar prepoznal 82 parov, ChatGPT 71, Perplexity pa 63. Omeniti velja, da gre za eno pojavitev, tj. en krajšavno-razvezavni par, in ne za ponovitve para. Po podrobni analizi filtriranih besedil z algoritmom Krajšavar in po ročnem preverjanju krajšavno-razvezavnih parov v besedilih smo ugotovili, da je priklic 91 %, natančnost algoritma pa 83 %, pri čemer smo odstotke zaokrožili (F1 znaša 0,87). Krajšavar ni prepoznal 13 krajšav, v 19 primerih pa je prepoznal krajšave, ki to niso. Za krajšavno-razvezavne pare, ki jih algoritem Krajšavar ni prepoznal, navajamo primere v preglednici 1.

Preglednica 1: Krajšavno-razvezavni pari, ki jih Krajšavar ne prepozna

Pojavitev samostalnika poleg krajšave	Posebni znaki v razvezavi ali male črke v krajšavi	Posebne krajšave	Zapis razvezav z malimi črkami
International Ship and Port Facilities Security Code (ISPS Code)	science and technology studies (STS)	Pamphlet (PAM)	field manual (FM)
	blood alcohol concentration (BAC)	tactical standard operating procedure (TACSOP)	international political sociology (IPS)
	standard operating procedures (SOPs)	ACC accidents	sociology of scientific knowledge (SSK)
	road traffic injuries (RTIs)	SPER suspicious person	portable document format (PDF)

Primere krajšavno-razvezavnih parov, ki jih Krajšavar ne prepozna, smo v preglednici 1 razdelili po načinu zapisa: pri kategoriji 1 imajo nizi poleg krajšave v oklepaju zapisan še samostalnik, ki ga algoritem Krajšavar ne more prepoznati. Pri kategoriji 2 sledijo zbrani pari, ki so sestavljeni iz posebnih znakov, in sicer vezaja v razvezavi ter malih črk v krajšavi. Tudi teh algoritem Krajšavar ne prepozna. Pri kategoriji 3 sledijo krajšave, ki so okrajšane na poseben način, npr. *ACC – accidents*. Sledi kategorija 4 – zapis razvezav z malimi črkami, kjer utegnemo kot vzrok za neprepoznavanje navesti zapis razvezav z malimi začetnimi črkami, npr. *field manual (FM)*.

Preglednica 2: Krajšavno-razvezavni pari, ki to niso in jih Krajšavar prepozna (nekaj primerov)

Imena, zapisana z veliko začetnico	Tuji jezik	Zapis z velikimi tiskanimi črkami
Schneider : (Socialism)	(Nazionale) National Diaristic Archive Foundation	(S) SPONSORING / MONITORING AGENCY NAME
Hellwig, first issued in (Hellwig)		(S) SPONSORING / MONITORING AGENCY NAME
Bourdieu : scientific capital (Bourdieu)		
Gieryn has called boundary work (Gieryn)		(S) S ACRONYM
William Stern, Paul Plaut, and Albert Hellwig (Wolffram)		(S) S REPORT NUMBER

V preglednici 2 so prikazani krajšavno-razvezavni pari, ki to niso, a jih algoritem Krajšavar prepozna kot take, in sicer gre za **(1)** imena, ki so zapisana z veliko začetnico, **(2)** zapise v tujem jeziku, ki ni angleški, ki se v besedilu pojavijo v oklepaju, in **(3)** zapise z velikimi tiskanimi črkami. Tako Krajšavar kot orodji ChatGPT in Perplexity ob filtraciji krajšavno-razvezavne pare prikažejo v tolikšni pojavitvi, kot se pojavijo v besedilu. Za potrebe prikaza izsledkov smo uporabili samo eno pojavitev krajšavno-razvezavnega para, in sicer tako pri Krajšavarju kot pri obeh orodjih umetne inteligence. To pomeni, da pri orodjih ChatGPT in Perplexity ni bilo izrecnih navodil, da se zapiše samo prva pojavitev para, je pa bilo uporabljeno navodilo, da izpiše par tako, kot se pojavi. V ta namen podrobneje analiziramo besedilo št. 13. Izsledki so predstavljeni v preglednici 3.

Preglednica 3: Filtracija besedila z algoritmom Krajšavar ter orodjema ChatGPT in Perplexity

Št. besedila	Avtor	Naslov	Vir
13.	Shawn Neely, Chris M. Anson	The Army and the Academy as Textual Communities: Exploring Mismatches in the Concepts of Attribution, Appropriation, and Shared Goals	https://apps.dtic.mil/sti/pdfs/ADA535143.pdf

Krajšavar	ChatGPT	Perplexity
SAR Same as Report	SOPs - Standard Operating Procedures	SOP: Standard Operating Procedure
TRADOC Training and Doctrine Command	TRADOC - Training and Doctrine Command	TRADOC: Training and Doctrine Command
AR Army Regulation	USMA - United States Military Academy	USMA: United States Military Academy
AKO Army Knowledge Online	U.S. - United States	AR: Army Regulation
CALL Center for Army Lessons Learned	U.S. Air Force Academy - United States Air Force Academy	FM: Field Manual
TTPs Tactics and Procedures	U.S. Army - United States Army	DA PAM: Department of the Army Pamphlet
BAS Battalion Aid Station	KS - Kansas	UCMJ: Uniform Code of Military Justice
AMEDD Army Medical Department	OMB - Office of Management and Budget	AKO: Army Knowledge Online
DA Department of the Army	SAR - Standard Annual Report	CALL: Center for Army Lessons Learned
Krajšavar ne najde:	SOP - Standard Operating Procedure	TACSOP: Tactical Standard Operating Procedure
standard operating procedures (SOPs)	AKO - Army Knowledge Online	DMOC: Division Medical Operations Center
tactical standard operating procedure (TACSOP)	CALL - Center for Army Lessons Learned	MAJ: Major
Pamphlet (PAM)	TTPs - Techniques, Tactics, and Procedures	1LT: First Lieutenant
field manual (FM)	MAJ - Major	AMEDD: Army Medical Department
	1LT - First Lieutenant	RGL: Reading Grade Level
	DA PAM 25-40 - Department of the Army Pamphlet 25-40: Army Publishing, Action Officer's Guide	VMI: Virginia Military Institute
	DA PAM 25-30 - Department of the Army Pamphlet 25-30: Army Publishing Program	USNA: United States Naval Academy
	RGL - Reading Grade Level	CGSC: Command and General Staff College
	LTG - Lieutenant General	BAS: Battalion Aid Station
	FM 3-24 Field Manual 3-24	PR: Public Relations
	PR Public Relations	
	e.g. exempli gratia	
	etc. et cetera	

Kot je razvidno iz preglednice 3, je po številu pridobljenih krajšavno-razvezavnih parov najuspešnejše orodje ChatGPT s 23 pridobljenimi pari, sledita Perplexity z 20 in Krajšavar z devetimi. Po ročnem pregledu vsebnosti krajšavno-razvezavnih parov v besedilu jih je bilo najdenih 13. Krajšavar je tako spregledal štiri pare, in sicer *standard operating procedures (SOPs)*, *tactical standard operating procedure (TACSOP)*, *Pamphlet (PAM)* in *field manual (FM)*. Razlogi, zakaj navedeni primeri niso bili prepoznani, so razloženi v preglednici 1. Opazimo, da Krajšavar sicer delno pravilno prepozna par *Techniques, Tactics and Procedures (TTPs)*, pri katerem je sicer pozabil navesti prvo besedo, tj. *Techniques*, a je par vseeno prepoznal. ChatGPT par *Techniques, Tactics and Procedures (TTPs)* prepozna, Perplexity pa ne.

Orodje ChatGPT sicer prepozna največje število krajšavno-razvezavnih parov, 23, a niso vsi prisotni v besedilu. Pravilno prepoznanih je šest parov: *SOPs – Standard Operating Procedures*, *TRADOC – Training and Doctrine Command*, *SAR – Standard Annual Report*, *AKO – Army Knowledge Online*, *CALL – Center for Army Lessons Learned*, *TTPs – Techniques, Tactics, and Procedures*. Ostali prepoznani pari se v besedilu ne pojavijo, npr. *USMA – United States Military Academy*, *KS – Kansas*, *OMB – Office of Management and Budget* itn. Izpostavimo še par *SOP – Standard Operating Procedure*, ta se v besedilu pojavi v paru *standard operating procedures (SOPs)* ter v zapisu *tactical standard operating procedure (TACSOP)*. Opazimo torej nepravilno prepoznavanja krajšavno-razvezavnega para *SOP – Standard Operating Procedure*. Orodje ChatGPT prepozna tudi nekaj okrajšav, npr. *e.g.* in *etc.*, ki v besedilu nimata svoje razvezave.

Orodje Perplexity je uspešnejše, saj prepozna 8 parov, tj. AMED - Army Medical Department, BAS - Battalion Aid Station, AKO - Army Knowledge Online, CALL - Center for Army Lessons Learned, TACSOP - Tactical Standard Operating Procedure itn. Tako kot ChatGPT prepozna par SOP - Standard Operating Procedure, ne pa para standard operating procedures - SOPs. Med krajšavno--razvezavnimi pari, ki se v besedilu ne pojavijo, so USMA - United States Military Academy, DA PAM - Department of the Army Pamphlet, UCMJ - Uniform Code of Military Justice, DMOC - Division Medical Operations Center, MAJ – Major in 1LT - First Lieutenant. Kot zanimivost izpostavimo, da orodje Perplexity prepozna par RGL [Reading Grade Level], ki ga zaradi oglatih oklepajev Krajšavar ne prepozna, prav tako ta krajšavno-razvezavni par prepozna orodje ChatGPT. Za razliko od orodja ChatGPT pa Perplexity ne prepozna okrajšav tipa e.g. ali etc.

V nadaljevanju v preglednici 4 številčno povzamemo, koliko dobljenih, pravih in haluciniranih krajšavno-razvezavnih parov je bilo prepoznanih z algoritmom Krajšavar ter z orodjema ChatGPT in Perplexity, dodani pa so tudi izsledki ročnega pregleda.

Preglednica 4: Število dobljenih, pravih in haluciniranih krajšavno-razvezavnih parov s Krajšavarjem ter orodjema ChatGPT in Perplexity po posameznem besedilu

Št. besedila	Število parov – ročni pregled	Dobljeni pari s Krajšavarjem/%	Dobljeni pari s ChatGPT/%	Dobljeni pravi pari s ChatGPT/%	Halucinirani s ChatGPT/%	Dobljeni pari s Perplexity/%	Dobljeni pravi pari s Perplexity/%	Halucinirani s Perplexity/%
1.	2	2 (100 %)	2 (100 %)	2 (100 %)	5 (250 %)	2 (100 %)	2 (100 %)	3 (150 %)
2.	6	5 (83 %)	5 (83 %)	5 (83 %)	14 (233 %)	0 (0 %)	0 (0 %)	0 (0 %)
3.	4	4 (100 %)	4 (100 %)	4 (100 %)	4 (100 %)	0 (0 %)	0 (0 %)	0 (0 %)
4.	4	4 (100 %)	4 (100 %)	4 (100 %)	4 (100 %)	4 (100 %)	4 (100 %)	2 (50 %)
5.	3	1 (33 %)	3 (100 %)	3 (100 %)	13 (433 %)	3 (100 %)	3 (100 %)	6 (200 %)
6.	12	9 (75 %)	9 (75 %)	9 (75 %)	15 (125 %)	4 (33 %)	4 (33 %)	4 (33 %)
7.	5	2 (40 %)	6 (120 %)	4 (80 %)	4 (80 %)	5 (100 %)	4 (80 %)	2 (40 %)
8.	2	2 (100 %)	2 (100 %)	2 (100 %)	7 (350 %)	2 (100 %)	2 (100 %)	7 (350 %)
9.	2	2 (100 %)	2 (100 %)	2 (100 %)	6 (300 %)	2 (100 %)	2 (100 %)	4 (200 %)
10	1	1 (100 %)	1 (100 %)	1 (100 %)	2 (200 %)	1 (100 %)	1 (100 %)	6 (600 %)
11.	2	2 (100 %)	0 (0 %)	0 (0 %)	7 (350 %)	2 (100 %)	2 (100 %)	10 (500 %)
12.	3	2 (66 %)	3 (100 %)	3 (100 %)	3 (100 %)	3 (100 %)	3 (100 %)	13 (433 %)
13.	13	9 (69 %)	6 (46 %)	4 (30 %)	17 (130 %)	7 (53 %)	5 (38 %)	13 (100 %)
14.	6	5 (83 %)	6 (100 %)	5 (83 %)	9 (150 %)	4 (66 %)	5 (83 %)	16 (266 %)
15.	1	1 (100 %)	0 (0 %)	0 (0 %)	6 (600 %)	0 (0 %)	0 (0 %)	0 (0 %)
16.	1	1 (100 %)	0 (0 %)	0 (0 %)	2 (200 %)	1 (100 %)	1 (100 %)	4 (400 %)
17.	10	8 (80 %)	6 (60 %)	6 (60 %)	3 (30 %)	6 (60 %)	6 (60 %)	8 (80 %)
18.	3	3 (100 %)	2 (66 %)	2 (66 %)	4 (133 %)	2 (66 %)	2 (66 %)	7 (233 %)
19.	11	11 (100 %)	7 (63 %)	7 (63 %)	4 (36 %)	7 (63 %)	7 (63 %)	7 (63 %)
20.	3	3 (100 %)	0 (0 %)	0 (0 %)	2 (66 %)	3 (100 %)	3 (100 %)	6 (200 %)
21.	5	5 (100 %)	3 (60 %)	3 (60 %)	4 (80 %)	5 (100 %)	5 (100 %)	7 (140 %)
Skupaj	99	82	71	63	135	63	61	125

Kot je razvidno iz preglednice 4, algoritem Krajšavar prepozna krajšavno-razvezavne pare pri vseh besedilih, ChatGPT pa ne najde pravilnih krajšavno-razvezavnih parov pri besedilih 11, 15, 16, in 20. Orodje Perplexity ne najde pravih krajšavno-razvezavnih parov pri besedilih 2, 3 in 15, pri slednjih orodje dejansko ne najde nobenega krajšavno-razvezavnega para. Krajšavar največ krajšavno-razvezavnih parov najde v besedilih 6, 13 in 19, ChatGPT pri besedilu 6, Perplexity pa pri besedilih 17 in 19. V preglednici 4 izpostavimo še krajšavno-razvezavne pare, ki to niso, in sicer primerjamo izsledke, pridobljene z orodjema umetne inteligence ChatGPT in Perplexity. Podrobnejši izsledki za algoritem Krajšavar so predstavljeni v preglednici 2; skupno je bilo prepoznanih 19 parov. Iz preglednice 4 je razvidno, da so halucinacije pri orodju ChatGPT prisotne pri 135 parih, pri orodju Perplexity pa pri 125 parih. Gre za pojav krajšavno-razvezavnih parov, ki jih v besedilu ni; med slednjimi izpostavimo krajšave, ki so v besedilih prisotne, npr. krajšave univerz ali organizacij. Razvezav v besedilih ni. Med primeri smo zaznali tudi prisotnost okrajšav. Podrobnejši pregled haluciniranih parov je viden v prilogi in v preglednici 5.

Preglednica 5: Halucinirani pari, ki se ne pojavijo v besedilu 13

ChatGPT	Perplexity
USMA - United States Military Academy	SOP: Standard Operating Procedure
U.S. - United States	USMA: United States Military Academy
U.S. Air Force Academy - United States Air Force Academy	DA PAM: Department of the Army Pamphlet
U.S. Army - United States Army	UCMJ: Uniform Code of Military Justice
KS - Kansas	DMOC: Division Medical Operations Center
OMB - Office of Management and Budget	MAJ: Major
SOP - Standard Operating Procedure	1LT: First Lieutenant
MAJ - Major	RGL: Reading Grade Level
1LT - First Lieutenant	VMI: Virginia Military Institute
DA PAM 25-40 - Department of the Army Pamphlet 25-40: Army Publishing, Action Officer's Guide	USNA: United States Naval Academy
DA PAM 25-30 - Department of the Army Pamphlet 25-30: Army Publishing Program	CGSC: Command and General Staff College
RGL - Reading Grade Level	PR: Public Relations
LTG - Lieutenant General	
FM 3-24 Field Manual 3-24	
PR Public Relations	
e.g. exempli gratia	
etc. et cetera	

V preglednici 5 so razvidni pari, ki se po ročnem pregledu v besedilu 13 ne pojavijo. Pari iz besedila 13 so izpostavljeni, ker je bilo v tem besedilu največ halucinaranih. Po pregledu smo ugotovili, da se po večini pojavijo samo krajšave, npr. USMA, LTG, etc., ali pa samo razvezave, npr. Kansas. Opazimo še, da se med pari pojavi tudi niz RGL - Reading Grade Level, ki ga prepoznata obe orodji. Razveza-va se pri slednjem v besedilu pojavi v oglatih oklepajih, zato Krajšavar tega para ne prepozna. Pri parih iz preglednice 5 lahko govorimo o dveh vrstah halucinacije. Poudariti velja, da orodje ChatGPT prepozna krajšave, ki se pojavijo v besedilu, npr. USMA in LTG, čeprav v besedilu ni njihovih razvezav, prepozna pa tudi pare, ki v besedilu sploh niso prisotni, npr. KS in CGSC. Tu bi želeli poudariti, da je bil cilj raziskave v naboru izbranih besedil iskati krajšavno-razvezavne pare. V ta namen je bilo tudi navodilo za filtracijo z orodjema ChatGPT in Perplexity usmerjeno v iskanje krajšavno-razvezavnih parov v priloženem besedilu, saj samo na tak način pridobimo primerljive izsledke z algoritmom Krajšavar, ki krajšavno--razvezavne pare išče izključno v izbranem besedilu. Tu velja izpostaviti, da tako ChatGPT kot Perplexity prepoznata nekaj več parov, kot jih je ročno najdenih v besedilu. To je seveda pozitivno, če želimo pridobiti čim več parov, negativna plat pa je, da so vmes tudi halucinacije in da utegnejo biti rezultati manj zanesljivi.

5 Sklep

Prispevek se osredinja na primerjavo izsledkov filtracije angleških varstvoslovnih besedil z algoritmom Krajšavar in orodjema umetne inteligence (ChatGPT in Perplexity) v procesu prepoznavanja krajšavno-razvezavnih parov. Filtrirali smo 21 besedil, ki smo jih zbrali na podlagi tipološke klasifikacije varstvoslovnih besedil, in sicer smo zaradi usklajene dolžine in s tem lažjega ročnega pregledovanja kot kategorijo izbrali znanstvena besedila s področja varnosti, policije, prava in vojske (2). Ročno smo pregledali nabor krajšavno-razvezavnih parov v filtriranih besedilih ter izsledke primerjali v smislu točnosti in natančnosti pridobljenih krajšavno-razvezavnih parov z algoritmom Krajšavar ter z orodjema ChatGPT in Perplexity. Iz pridobljenih izsledkov lahko povemo, da je pri prepoznavanju krajšavno-razvezavnih parov najbolj točen algoritem Krajšavar, ki pridobi 82 % točnih parov, sledita ChatGPT z 71 % in Perplexity s 63 %. Haluciniranih krajšavno-razvezavnih parov, pridobljenih z orodjem ChatGPT, je za 214 % več kot pravih parov, z orodjem Perplexity pa za 204 % več. Tu velja omeniti, da so med haluciniranimi pari tudi tisti, ki v besedilu niso sestavljeni iz krajšave in razvezave; praviloma se pojavi samo krajšava. Smiselno je ločevati med pari, ki so v besedilu dejansko prisotni v obliki krajšave ali razvezave, ter pari, ki v besedilu sploh niso pristoni. V besedilu 13 orodje ChatGPT dopolni vse najdene krajšave,

orodje Perplexity pa dopolni razvezavo *Virginia Military Institute* s krajšavo VMI in razvezavo *Uniform Code of Military Justice* s krajšavo UCMJ. Orodje dopolni tudi preostale krajšave z izjemo krajšavno-razvezavnega para CGSC - *Command and General Staff College*, ki se v besedilu ne pojavi, torej orodje par povsem halucinira. Prednost orodij ChatGPT in Perplexity je nedvomno v tem, da najdeta pare krajšav in razvezav, tudi če se slednje ne pojavijo v besedilu. Te zmožnosti Krajšavar nima, saj išče samo krajšavno-razvezavne pare, ki se pojavijo v besedilu. Cilj raziskave je bil vsekakor iskati izključno krajšavno-razvezavne pare, ki se pojavijo v besedilu, saj lahko na tak način primerjamo točnost algoritma in orodij umetne inteligence. Raziskava je pokazala visoko točnost algoritma Krajšavar, ki pa bo vsekakor dodatno izboljšan, npr. možnosti nalaganja dokumenta, kot to omogočata ChatGPT in Perplexity, saj se bo tako poenostavila filtracija. Prednost orodij umetne inteligence vidimo predvsem v zmožnosti pridobivanja krajšav in razvezav tudi izven filtriranega besedila, kar je seveda pozitivno pri izgradnji potencialnih podatkovnih baz krajšavno-razvezavnih parov, ki utegnejo služiti kot gradivo za pripravo glosarjev ali slovarjev. Vsekakor bi bilo treba pri pogovornih botih raziskati, ali ti utegnejo delovati bolje, če se v navodila dodajo natančnejši opisi oz. če je treba take opise ponoviti pri vsaki posamezni filtraciji besedila.

Literatura

Bezlaj 1995 = France Bezlaj, *Etimološki slovar slovenskega jezika* 3: *P–S,* dopolnila in uredila Metka Furlan – Marko Snoj, Ljubljana: Mladinska knjiga, 1995.

Byrd – Park 2011 = Roy J. Byrd – Youngja Park, Hybrid TextMining for Finding Abbreviations and Their Definitions, *IMB Thomas J. Watson Research Center* (2011), 167–170.

ChatGPT, marec 2025, https://chatgpt.com.

Debenjak – Debenjak – Debenjak 1992 = Doris Debenjak – Božidar Debenjak – Primož Debenjak, *Veliki nemško-slovenski slovar = Grosses deutsch-slowenisches Wörterbuch*, Ljubljana: Državna založba Slovenije, 1992.

de Schryver 2023 = Gilles-Maurice de Schryver, Generative AI and Lexicography: the Current State of the Art Using ChatGPT, *International Journal of Lexicography* 36.4 (2023), 355–387, DOI: https://doi.org/10.1093/ijl/ecad021.

***Digitalna slovarska baza za slovenščino*,** 2023–, https://www.cjvt.si/blog/oznaka/digitalna-slovarska-baza-za-slovenscino/.

Gabrovšek 1994 = Dušan Gabrovšek, Kodifikacija angleškega jezika v specializiranih enojezičnih slovarjih: too much of everything?, *Vestnik* 28.1–2 (1994), 150–180.

Gelernter – Judith 2013 = Judith Gelernter – Shilpa Balaji, An algorithm for local geoparsing of microtext, *Geoinformatica* 17 (2013), 635–667.

Hajnšek-Holz – Jakopin 1996 = Milena Hajnšek-Holz – Primož Jakopin, *Odzadnji slovar slovenskega jezika po Slovarju slovenskega knjižnega jezika*, Ljubljana: Založba ZRC, ZRC SAZU, 1996.

Huang idr. 2022 = Xiusheng Huang – Bin Li – Fei Xia – Yixuan Weng, A novel initial reminder framework for acronym extraction, v: *SDU@AAAI-22,* 2022, https://ceur-ws.org/Vol-3164/paper29.pdf.

Humar 2004 = Marjeta Humar (ur.), *Terminologija v času globalizacije: zbornik prispevkov s simpozija Terminologija v času globalizacije, Ljubljana, 5.-6. junij 2003 = Terminology at the time of globalization*, Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC = Scientific Research Centre SASA, ZRC Publishing, 2004.

Jakubíček – Rundell 2023 = Miloš Jakubíček – Michael Rundell, The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography. v: *Electronic lexicography in the 21st century: invisible lexicography*, Brno: Lexical Computing, 2023, 518–533, https://elex.link/elex2023/wp-content/uploads/102.pdf.

Kolokacijski slovar sodobne slovenščine, 2018–, https://viri.cjvt.si/kolokacije/slv/#

Kompara Lukančič 2009 = Mojca Kompara Lukančič, Prepoznavanje krajšav v besedilih, *Jezikoslovni zapiski* 15.1–2 (2009), 95–112.

Kompara Lukančič 2010 = Mojca Kompara Lukančič, Krajšavni slovarji, *Jezikoslovni zapiski* 16.2 (2010), 111–129.

Kompara Lukančič 2011 = Mojca Kompara Lukančič, Razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih, *Jezikoslovni zapiski* 17.2 (2011), 107–122.

Kompara Lukančič 2018 = Mojca Kompara Lukančič, *Sinhrono-diahroni pregled krajšav v slovenskem prostoru in sestava slovarja krajšav*, Maribor: Univerza v Mariboru, Univerzitetna založba, 2018.

Kompara Lukančič 2023a = Mojca Kompara Lukančič, Compilation of English entries in the contemporary Slovene dictionary of abbreviations, *International Journal of Lexicography* 36.2 (2023), 195–210.

Kompara Lukančič 2023b = Mojca Kompara Lukančič, *English for specific purposes: selected readings from the field of English for criminal justice and security*, Maribor: Univerza v Mariboru, Univerzitetna založba, 2023.

Kompara Lukančič 2025 = Mojca Kompara Lukančič, *Prepoznavanje krajšav in razvezav v angleških besedilih s področja varstvoslovja z algoritmom Krajšavar ter orodjema ChatGPT in Perplexity* [zaključena zbirka raziskovalnih podatkov], 2025, https://dk.um.si/IzpisGradiva.php?lang=slv&id=95507.

Kompara Lukančič – Holozan 2011 = Mojca Kompara Lukančič – Peter Holozan, What is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using Termania website, v: *Electronic lexicography in the 21st century: new applications for new users*, ur. Iztok Kosem – Karmen Kosem, Ljubljana: Trojina, 2011, 140–146.

Kompara Lukančič – Smajla 2025 = Mojca Kompara Lukančič – Tilen Smajla, Krajšavar—an algorithm for recognizing English abbreviations in texts related to criminal justice and security, *International Journal of Lexicography* 38.3 (2025), 237–269, DOI: https://doi.org/10.1093/ijl/ecaf012.

Kosem 2024 = Iztok Kosem, *Veliki slovensko-madžarski slovar,* različica 2.0, rastoči slovar, Založba ZRC SAZU, 2024–, https://franja.si/slovar/sl-ma.

Kosem – Gantar – Krek 2013 = Iztok Kosem – Polona Gantar – Simon Krek, Avtomatizacija leksikografskih postopkov, *Slovenščina 2.0* 1.2 (2013), 139–164.

Košmrlj-Levačič – Seliškar 2004 = Borislava Košmrlj-Levačič – Tomaž Seliškar, Uporabniški računalniški program *SlovarRed 2.0*, v: *Terminologija v času globalizacije*, ur. Marjeta Humar, Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC, 2004, 179–199.

Kuo idr. 2009 = Cheng-Ju Kuo – Maurice HT Ling – Kuan-Ting Lin – Chun-Nan Hsu, BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature, v: *Eight International Conference on Bioinformatics* 10 (2009), S7, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S15-S7.

Larkey idr. 2000 = Leah S. Larkey – Paul Ogilvie – M. Andrew Price – Brenden Tamilio, Acrophile: an Automated Acronym Extractor and Server in Digital Libraries, v: *Proceedings of the Fifth ACM Conference on Digital Libraries*, ur. Peter J. Nürnberg – David L. Hicks – Richard Futura, New York: Association for Computing Machinery, 2000, 205–214.

Lew 2023 = Robert Lew, ChatGPT as a COBUILD lexicographer, *Humanit Soc Sci Commun* 10 (2023), 704, DOI: https://doi.org/10.1057/s41599-023-02119-6.

Liu – Liu – Huang 2017 = Jie Liu – Caihua Liu – Yalou Huang, Multi-granularity sequence labeling model for acronym expansion identification, *Information Sciences* 378 (2017), 462–474.

Mikolič 2007 = Vesna Mikolič, Tipologija turističnih besedil s poudarkom na turističnooglaševalskih besedilih, *Jezik in slovstvo* 52.3–4 (2007), 107–116.

Montalvo idr. 2018 = Soto Montalvo – Raquel Martínez – Mario Almagro – Susana Lorenzo, MAM-TRA-MED at Biomedical Abbreviation Recognition and Resolution - IberEval 2018, v: *CEUR Workshop Proceedings*, ur. María Teresa Martín-Valdivia – María Dolores Molina-González – Salud María Jiménez-Zafra, 2018, 1–8, https://ceur-ws.org/Vol-2150/BARR2_paper1.pdf.

Perplexity = *Perplexity, marec 2025,* https://www.perplexity.ai

Rundell 2023 = Michael Rundell, Automating the creation of dictionaries: are we nearly there, v: *Asialex 2023 Proceedings, Lexicography, Artificial Intelligence and Dictionary Users*, Seoul: Yonsey University, 2023, 9.

Schwartz – Hearst 2003 = Ariel S. Schwartz – Marti A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical texts, v: *Proceedings of the Pacific Symposium on Biocomputing*, ur. Russ B. Altman – A. Keith Dunken – Lawrence Hunter – Tiffany A. Jung – Teri E. Klein, Kauai: Indiana University School of Medicine, 2003, 451–462.

Slovar sopomenk sodobne slovenščine = *Slovar sopomenk sodobne slovenščine*, 2017–, http://viri.cjvt.si/sopomenke/slv/

Snoj 1998 = Marko Snoj, *Slovenski etimološki slovar*, Ljubljana: Založba ZRC SAZU, 1998.

SP 2001 = *Slovenski pravopis*, 2014, www.fran.si.

Šatev – Nikolov 2008 = Vesna Šatev – Nicolas Nikolov, Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language, v: *Jezikovne tehnologije: zbornik 11. mednarodne multikonference Informacijska družba – IS*, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Institut Jožef Stefan, 2008, 75–79.

Taghva – Gilbreth 1999 = Kazem Taghva – Jeff Gilbreth, Recognizing acronyms and their definitions, *International Journal on Document Analysis and Recognition* 1.4 (1999), 191–198.

Veyseh idr. 2020 = Amir Pouran Ben Veyseh – Franck Dernoncourt – Thein Huu Nguyen – Walter Chang – Leo Anthony Celi, Acronym identification and disambiguation shared tasks for scientific document understanding. *arXiv preprint arXiv:2012.11760*, https://ceur-ws.org/Vol-2831/paper33.pdf

Veyseh idr. 2022 = Amir Pouran Ben Veyseh – Nicole Meister – Franck Dernoncourt – Thein Huu Nguyen, Acronym extraction and acronym disambiguation shared tasks at the Scientific Document Understanding Workshop 2022, v: *Proceedings of the Scientific Document Understanding Workshop 2022*, ur. Amir Pouran Ben Veyseh idr., 2022, https://ceur-ws.org/Vol-3164/.

Vossen 2022 = Piek Vossen, ChatGPT Is a Waste of Time, *VU Magazine*, 2022, https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=en.

Verovnik 2023 = Tina Verovnik, Pomen javne razprave za prenovo pravopisnih pravil, *Škrabčevi dnevi* 12, ur. Danila Zuljan Kumar – Helena Dobrovoljc, Nova Gorica: Založba Univerze, 2023, 35–43.

Weiss 1998 = Peter Weiss, *Slovar govorov Zadrečke doline: med Gornjim Gradom in Nazarjami: poskusni zvezek (A–H)*, Ljubljana: Založba ZRC SAZU, 1998.

Weiss 1991 = Peter Weiss, Zasnova novega odzadnjega slovarja slovenskega jezika, *Jezikoslovni zapiski* 1.1 (1991), 121–139.

Wu idr. 2015 = Yonghui Wu – Jun Xu – Yaoyun Zhang – Hua Xu, Clinical abbreviation disambiguation using neural word embeddings, v: *Proceedings of BioNLP 15*, ur. Kevin Bretonnel Cohen idr., Beijing: Association for Computational Linguistics, 2015, 171–176, DOI: 10.18653/v1/W15-38.

Xu – Huang 2005 = Jun Xu – Ya-Lou Huang, A machine learning approach to recognizing acronyms and their expansion. v: *2005 International Conference on Machine Learning and Cybernetics* 4, ur. Daniel S. Yeung – Zhi-Qiang Liu, Guangzhou, China: Springer-Verlag, 2005, 2313–2319.

Yeates 1999 = Stuart Yeates, Automatic extraction of acronyms from text, v: *Proceedings of the Third New Zealand Computer Science Research Students,* ur. David Bainbridge – Stuart A. Yeast, 1999, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=31fcd3c1ac62b3612c071adc13c3b041644a2222.

Zahariev 2004 = Manuel Zahariev, *A (Acronyms)*, doktorska disertacija, Simon Fraser University, School of Computing Science, 2004.

Zhou – Torvik – Smalheiser 2006 = Wei Zhou – Vetle I. Torvik – Neil R. Smalheiser, ADAM: another database of abbreviations in MEDLINE, *Bioinformatics* 22 (2006), 2813–2818.

Summary

Recognizing Abbreviations and Their Expansions in English Criminal Justice and Security Texts Using the Krajšavar Algorithm, ChatGPT, and Perplexity

The integration of artificial intelligence (AI) tools into linguistic analysis has introduced novel methodologies for recognizing abbreviation–expansion pairs in domain-specific texts. This study evaluates the performance of ChatGPT and Perplexity in comparison with the Krajšavar algorithm, focusing on English-language texts within the field of criminal justice and security studies. A manually prepared corpus of 21 scientific texts—spanning law enforcement, military, legal, and security domains—was subjected to automated filtering and manual validation.

The Krajšavar algorithm, designed to detect abbreviation–expansion pairs explicitly present in text, demonstrated superior accuracy (82%) compared with ChatGPT (71%) and Perplexity (63%). However, both AI tools exhibited a high incidence of hallucinated pairs, with ChatGPT and Perplexity generating 214% and 204% more false positives than valid matches, respectively. Despite this limitation, the AI tools showed an enhanced capacity to infer expansions not explicitly stated, suggesting potential utility in constructing comprehensive abbreviation databases.

The study underscores the importance of aligning tool capabilities with research objectives, particularly when precision in textual analysis is paramount. Future improvements to the Krajšavar algorithm—such as the addition of document upload functionality—may enhance its usability and integration. Overall, the findings highlight the complementary strengths and limitations of AI-driven and algorithmic approaches to abbreviation recognition within specialized corpora.