slovenščina 2.0, letnik 11, številka 1

Slo2.0

DIGITALNO JEZIKOSLOVJE

Digital Linguistics

2023_1

Univerza *v Ljubljani*
Filozofska fakulteta

Slovenščina 2.0
Letnik/Volume 11, Številka/Issue 1, 2023

ISSN: 2335-2736

Založila/Published by
Založba Univerze v Ljubljani

Za založbo/For the Publisher
Gregor Majdič, rektor Univerze v Ljubljani

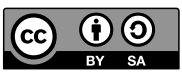
Izdala/Issued by
Znanstvena založba Filozofske fakultete Univerze v Ljubljani;
Center za jezikovne vire in tehnologije Univerze v Ljubljani

Za izdajatelja/For the Issuer
Mojca Schlamberger Brezar, dekanja Filozofske fakultete

Publikacija je brezplačna./Publication is free of charge.

Publikacija je dostopna na/Avaliable at: https://journals.uni-lj.si/slovenscina2

Revija izhaja s podporo Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije./
This journal is published with the support of the Slovenian Research and Innovation Agency (ARIS).



Kazalo / Content

UVODNIK / EDITORIAL

ČLANKI / ARTICLES

Uvodnik v tematsko številko o Digitalnem jezikoslovju

Darja FIŠER
Inštitut za novejšo zgodovino

Tomaž ERJAVEC
Institut Jožef Stefan

Pričujoča tematska številka revije *Slovenščina 2.0* se posveča digitalnemu jezikoslovju, hitro rastočemu interdisciplinarnemu področju raziskav na stičišču tradicionalnega jezikoslovja, informacijskih tehnologij in družboslovnih ved. V ospredju digitalnojezikoslovnih raziskav je ohranjanje, analiza in uporaba jezikovnih podatkov, digitalnih artefaktov z jezikom kot nosilcem medčloveškega sporazumevanja. Digitalno jezikoslovje tako pri nas kot po svetu postaja vse pomembnejše ne samo v akademskih in izobraževalnih krogih, temveč tudi v javnem in zasebnem sektorju, ki za uspešno delovanje v sodobni družbi in gospodarstvu vse bolj potrebujeta strokovnjake, vešče upravljanja z digitalnimi jezikovnimi podatki.

Tematska številka vsebuje enajst prispevkov slovenskih in tujih avtorjev, ki so bili v krajši in bolj omejeni obliki najprej predstavljeni na konferenci "Jezikovne tehnologije in digitalna humanistika" leta 2022. Prvi sklop prinaša pester nabor raziskovalnih vprašanj z različnih znanstvenih ved, od jezikoslovja in prevodoslovja pa vse do literarnih ved in zgodovinopisja, ki se jih avtorji lotevajo s korpusnim pristopom, drugi pa združuje prispevke, ki predstavljajo gradnjo eno- in večjezičnih jezikovnih virov in tehnologij za različne naloge.

Fišer, D., Erjavec, T.: Uvodnik v tematsko številko o Digitalnem jezikoslovju/ Introduction to the special issue on Digital Linguistics. Slovenščina 2.0, 11(1): 1–6.

1.20 Uvodnik / Editorial

DOI: https://doi.org/10.4312/slo2.0.2023.1.1-6

V korpusnojezikoslovnem sklopu Špela Arhar Holdt, Iztok Kosem, Eva Pori, Vojko Gorjanc, Simon Krek in Polona Gantar predstavijo inovativne rešitve za prepoznavanje in označevanje sovražnega in grobega besedišča v okviru koncepta odzivnega Slovarja sopomenk sodobne slovenščine. Špela Antloga opiše najpogostejše metode luščenja metaforičnih in metonimičnih izrazov iz jezikovnih korpusov ter na primeru korpusa g-KOMET, ki je ročno označen za metaforične izraze in metonimične prenose, ponazori poskus sistematizacije nekaterih najbolj prisotnih metonimičnih prenosov v slovenskem govorjenem jeziku. Jakob Lenardič in Kristina Pahor de Maiti proučita skladenjsko in pragmatično rabo epistemičnih in deontičnih modalnih izrazov v korpusu slovenskih družbeno sprejemljivih in nesprejemljivih komentarjev na družbenem omrežju Facebook. David Bordon poroča o raziskavi preverjanja razumljivosti nerevidiranih strojno prevedenih spletnih besedil med splošni bralci. Darja Fišer, Tjaša Konovšek in Andrej Pančur analizirajo značilnosti populističnega govora v parlamentarnih razpravah slovenskih poslancev. Andrejka Žejn in Mojca Šorli pa predstavita ročno semantično označevanje imenskih entitet glede na predlagano označevalno shemo, izdelano za korpus modernističnih literarnih besedil Maj68.

V jezikovnotehnološkem sklopu Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Taja Kuzman in Nikola Ljubešić opišejo prvi spremljevalni korpus za slovenščino Trendi ter razvoj, evalvacijo in aplikacijo algoritma za avtomatsko kategorizacijo besedil z novičarskih portalov. Kaja Dobrovoljc, Luka Terčon in Nikola Ljubešić predstavijo nove smernice, odprto dostopne ročno označene podatke ter razčlenjevalni model za označevalnik CLASSLA Staza v formalizmu Universal Dependencies za slovenščino. Petra Bago in Virna Karlić poročata o novi različici odprto dostopnega korpusa DirKorp (Korpus direktivnih govornih činova hrvatskoga jezika), namenjenega raziskavam v pragmatiki, ki vsebuje simulirano komunikacijo, implementirano v vnaprej danih pogojih stotih hrvaških govorcev. Uroš Šmajdek Matjaž Zupanič, Maj Zirkelbach in Meta Jazbinšek opišejo ter evalvirajo sistem za odgovarjanje na vprašanja v slovenskem jeziku. Gregor Donaj in Mirjam Sepesy Maučec pa predstavita uporabo podbesednih enot za nevronsko strojno prevajanje iz slovenščine v angleščino.

Posebno številko so recenzirali Zoran Bosnić, Václav Cvrček, Jaka Čibej, Helena Dobrovoljc, Kaja Dobrovoljc, Polona Gantar, Vojko

Gorjanc, Jurij Hadalin, Matej Klemen, Jakob Lenardič, Nikola Ljubešić, Matija Marolt, Maja Miličević Petrović, Matija Ogrin, Matevž Pesek, Dan Podjed, Tanja Samardžić, Marko Robnik Šikonja, Mojca Šorn, Simon Šuster, Daniel Vasić in Aleš Žagar.

Urednika posebne številke se iskreno zahvaljujeva avtorjem in recenzentom za njihovo predano delo.

V primerjavi s sorodno tematsko številko na temo jezikovnih tehnologij in digitalne humanistike iz leta 2021, kjer je bil poudarek na uvajanju naprednih tehnik in metod strojnega učenja, večjezikovnim pristopom, kritičnemu ocenjevanju obstoječih tehnologij ter razvoju storitev za končnega uporabnika, v pričujoči številki opazimo pomembno širitev korpusnega pristopa na vede izven jezikoslovja, ki pri svojem delu prav tako izhajajo iz pretežno besedilnih podatkov, ter razvoj vse bolj specializiranih jezikovnih virov in tehnologij. Oba premika nakazujeta na dozorevanje in razmah področja v Sloveniji v zadnjih nekaj letih. Z vstopom v obdobje, v katerem se metode umetne inteligence povsod po svetu intenzivno uveljavljajo v praktično vseh znanstvenih disciplinah, pa tudi v našem vsakdanjem življenju, pa bo pomen zanesljivih in visoko kakovostnih jezikovnih virov, preverljivih jezikovnih tehnologij ter čezdisciplinski prenos metodološkega znanja samo še naraščal, zato je ključno, da skupnosti zagotovimo ustrezno raziskovalnoinfrastrukturno podporo in izobraževalni okvir.

Ljubljana, julij 2023

Introduction to the special issue on Digital Linguistics

Darja FIŠER
Institute of Contemporary History

Tomaž ERJAVEC
Jožef Stefan Institute

The current special issue of the journal Slovenščina 2.0 focuses on Digital Linguistics, a growing interdisciplinary field at the crossroads of traditional linguistics, information technology and social sciences. Digital Linguistics preserves, analyses and utilises language data, i.e. digital artefacts that use language as a means of human expression. In Slovenia as well as abroad, Digital Linguistics is attracting increasing attention not only from the academic and educational communities but also from the public and private sectors, since skills in handling digital language data are considered essential in the modern economy and society.

The special issue presents eleven papers of Slovenian and international authors that were originally presented, in a shorter and more limited form, at the 2022 Language Technologies and Digital Humanities Conference. Part 1 contains a broad range of research questions from different scientific disciplines such as linguistics and translation studies but also literary studies and historiography which are approached within the corpus linguistics framework. Part 2 comprises papers that present the development of mono- and multilingual language resources and technologies for a variety of tasks.

The corpus-linguistic section opens with the paper by Špela Arhar Holdt, Iztok Kosem, Eva Pori, Vojko Gorjanc, Simon Krek and Polona Gantar who introduce innovative solutions for the recognition and mark-up of hate and offensive lexis in the framework of the Reactive Dictionary of Contemporary Slovenian Synonyms. Špela Antloga then describes the most common methods for extracting metaphorical and

metonymic expressions from language corpora and, on the case of the g-KOMET corpus, which is manually labelled for metaphoric expressions and metonymic transfers, exemplifies an attempt to systemise some of the most common metonymic transfers in the Slovenian spoken language. Jakob Lenardič and Kristina Pahor de Maiti investigate grammatical and pragmatic usage of epistemic and deontic modal expressions in the corpus of acceptable and unacceptable comments in Slovenian Facebook comments. David Bordon gives an account of an investigation to assess the comprehensibility of un-edited machine translated Web texts among general users. Darja Fišer, Tjaša Konovšek and Andrej Pančur explore the characteristics of populist discourse in parliamentary speeches of Slovenian members of parliament. Andrejka Žejn and Mojca Šorli introduce manual semantic labelling of named entities in accordance with the annotation scheme developed for the corpus of Slovenian modernist literary texts May68.

The language-technology section starts with a description of Trendi, the first monitor corpus for Slovenian by Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Taja Kuzman and Nikola Ljubešić who also describe the development, evaluation and application of their method for the automatic categorisation of texts from news portals. Kaja Dobrovoljc, Luka Terčon and Nikola Ljubešić introduce the new guidelines, openly available manually annotated datasets and the analysis model for the CLASSLA Stanza annotation tool, all for the Universal Dependencies formalism for Slovenian. Petra Bago and Virna Karlić report on the new version of the openly available DirKorp corpus, meant for research on pragmatics, which contains simulated communication implemented in pre-set conditions by one hundred Croatian speakers. Uroš Šmajdek Matjaž Zupanič, Maj Zirkelbach and Meta Jazbinšek describe and evaluate their system for question answering in Slovenian. Finally, Gregor Donaj and Mirjam Sepesy Maučec discuss and evaluate the use of sub-word units for neural machine translation from Slovenian to English.

The special issue was reviewed by Zoran Bosnić, Václav Cvrček, Jaka Čibej, Helena Dobrovoljc. Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Jurij Hadalin, Matej Klemen, Jakob Lenardič, Nikola Ljubešić, Matija Marolt, Maja Miličević Petrović, Matija Ogrin, Matevž Pesek, Dan Podjed, Tanja Samardžić, Marko Robnik Šikonja, Mojca Šorn, Simon

Šuster, Daniel Vasić in Aleš Žagar. The editors of the special issue would like to thank the authors and the reviewers for their dedicated work.

Compared to the related special issue on Language Technologies and Digital Humanities published in this journal in 2021 where the focus of research was on the implementation of state-of-the-art machine learning methods, multilingual approaches, critical evaluation of technologies, and development of services for the end user, we observe an important spread of the corpus approach to disciplines beyond linguistics which are also based on predominantly textual data, as well as the emergence of increasingly specialised language resources and technologies. This suggests that in Slovenia the field has matured and advanced significantly in the last couple of years. In the time when Artificial Intelligence is playing an increasingly important role in the methodological approaches in virtually all scientific disciplines, as well as our everyday lives all over the world, the importance of reliable and high quality language resources, verifiable language technologies and transdisciplinary knowledge transfer will only increase, which is why it is crucial that our community is equipped with an adequate research infrastructure support and a robust educational framework.

Ljubljana, July 2023

Sklop 1
Korpusnojezikoslovne raziskave

Negativno zaznamovano besedišče v *Slovarju sopomenk sodobne slovenščine 2.0*

Špela ARHAR HOLDT
Filozofska fakulteta, Univerza v Ljubljani; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Iztok KOSEM
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Eva PORI
Filozofska fakulteta, Univerza v Ljubljani

Vojko GORJANC
Filozofska fakulteta, Univerza v Ljubljani

Simon KREK
Filozofska fakulteta, Univerza v Ljubljani

Polona GANTAR
Filozofska fakulteta, Univerza v Ljubljani

V prispevku predstavljamo rešitve za prepoznavanje in označevanje zaznamovanega besedišča v okviru koncepta odzivnega Slovarja sopomenk sodobne slovenščine. Ker gre za prvi tovrstni projekt, so pripravljene rešitve v veliki meri inovativne, umeščene pa v okvir problematike avtomatske strojne izdelave slovarja, njegove odprtosti in vključenosti uporabniške skupnosti. Prispevek prikazuje postopek prepoznavanja sovražnega in grobega besedišča ter pripis

Arhar Holdt, Š., Kosem, I., Pori, E., Gorjanc, V., Krek, S., Gantar, P.: Negativno zaznamovano besedišče v Slovarju sopomenk sodobne slovenščine 2.0. *Slovenščina 2.0, 11(1): 8–32.*

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.8-32

oznak, opozorilnih ikon in daljših pojasnil. Ukvarjamo se tako s tehničnimi kot vsebinskimi vprašanji označevanja. Vsebinsko oznake temeljijo na sporočanjskem namenu in učinku, pri čemer je njihovo bistvo informacija o možnih posledicah rabe, pri tehničnih rešitvah pa veliko pozornost posvečamo digitalnemu mediju in vizualizaciji rešitev v njem. Ker je odzivnost eden ključnih konceptov slovarja, se pri rešitvah glede označevanja zavedamo pomembnosti sodelovanja z uporabniško skupnostjo, zato tudi pri dodajanju oznak predlagamo rešitve za sodelovanje s skupnostjo. Izhodiščni konferenčni prispevek je bil razširjen v vseh poglavjih, dodano pa je povsem novo poglavje o obdelavi večpomenskih iztočnic, njihovi pomenski členitvi in pomenskem opisovanju z zgledi pomenov z negativno zaznamovanostjo.

Ključne besede: slovar sopomenk, odzivni slovar, slovarske oznake, sporočanjski namen, uporabniška skupnost

1 Uvod

Slovar sopomenk sodobne slovenščine (SSSS) je oblikovan po modelu odzivnega slovarja: v prvem koraku je bil pripravljen strojno, nadaljnje urejanje podatkov pa poteka po korakih in v sodelovanju jezikoslovcev ter širše zainteresirane skupnosti (Arhar Holdt et al., 2018, str. 404). V SSSS lahko slovarski uporabniki ob strojno pripravljeno sopomensko gradivo dodajo lastne predloge sopomenk, za vse sopomenke v slovarju pa je mogoče tudi glasovati in gradivo na tak način (pomagati) urejati oz. selekcionirati.[1]

Vključevanje strojnih postopkov in predlogov uporabniške skupnosti v slovaropisne delotoke odgovarja na potrebe sodobnega časa, kot sta potreba skupnosti po odprto dostopnih jezikovnih podatkih in želja slovarskih uporabnikov po demokratičnem sodelovanju pri razvoju temeljne jezikovne infrastrukture. Na drugi strani pa ima neposredno objavljanje strojnega in uporabniško dodanega (nepregledanega) gradiva lahko tudi neželene posledice, ki jih je treba pri razvoju odzivnega modela predvideti in ustrezno obravnavati. Med prioritetami za razvoj SSSS je tako brez dvoma obravnava besedišča, ki vrednostno poimenuje posamezne družbene skupine in njihove pripadnike. Tako besedišče

1 Slovar v vmesniku je na https://viri.cjvt.si/sopomenke/slv/, kot slovarska baza pa na repozitoriju CLARIN.SI (Krek et al., 2018). Strojno pripravo slovarja opisujejo Krek et al. (2017), koncept odzivnega slovarja pa Arhar Holdt et al. (2018).

se v nepregledani različici slovarskih gesel (lahko) pojavlja na različnih mestih in na različne načine.

Namen prispevka je predstaviti obseg problematike, ki se pri odzivnem slovarju pomembno razlikuje od tradicionalnih slovaropisnih projektov, in opisati rešitve, ki so vključene v nadgradnjo SSSS iz različice 1.0 v 2.0, kot tudi naloge za nadaljnje delo. V prispevku želimo posebej izpostaviti nove načine prepoznavanja in označevanja sovražnega, grobega ter drugače negativno vrednotenega besedišča, ki so uporabne za različne sodobne jezikovne vire, ne le SSSS.

Prispevek je nadgradnja dela (Arhar Holdt et al., 2022), ki je bilo strokovni javnosti predstavljeno na znanstveni konferenci Jezikovne tehnologije in digitalna humanistika 2022. Prispevek je posodobljen v vseh poglavjih, povsem nova pa so poglavja, ki opisujejo slovaropisni pregled gradiva in končne odločitve, implementirane v SSSS 2.0.

2 Negativno zaznamovano besedišče v družbi, jeziku in slovarju

Na kratko je mogoče sovražni govor opredeliti kot "aktivno javno spodbujanje antipatije do določene, ponavadi šibke, družbene skupine" (Rebolj, 2008, str. 13), v daljši in bolj povedni obliki pa kot (Petković in Kogovšek Šalamon, 2007, str. 23):

> ustno ali pisno izražanje diskriminatornih stališč. Z njim širimo, spodbujamo, promoviramo ali opravičujemo rasno sovraštvo, ksenofobijo, homofobijo, antisemitizem, seksizem in druge oblike sovraštva, ki temeljijo na nestrpnosti. Mednje sodi tudi nestrpnost, ki se izraža z agresivnim nacionalizmom in etnocentrizmom, z diskriminacijo in sovražnostjo zoper manjšine, migrante in migrantke. Žrtve sovražnega govora praviloma niso posamezniki, pač pa ranljive družbene skupine. V osrčju sovražnega govora je prepričanje, da so nekateri ljudje manj vredni, zato je cilj sovražnega govora v razčlovečenju, ponižanju, ustrahovanju in poslabšanju družbenega položaja tistih, proti katerim je naperjen.

Motl in Bajt (2016, str. 7) ugotavljata, da je sovražni govor deležen precejšnje pozornosti v različnih vedah, od prava, sociologije in

komunikologije do psihiatrije in informatike, pridružimo pa jim lahko tudi jezikoslovje.

2.1 Kritično slovaropisje

Ameriško slovaropisje (Hughes, 2009, str. 87–105) je že pred desetletji v svoje vire načrtno vgradilo tudi občutljivost do ranljivih družbenih skupin, pri čemer ni zanemarilo nobenega od delov geselskega članka: razlag, oznak in zgledov rabe (Logar et al., 2020, str. 104). V manjši meri in pozneje, a vendarle so se opozorila o nujni tovrstni družbeni občutljivosti ter odgovornosti pojavila tudi v slovenskem prostoru (npr. Gorjanc, 2005; Kern, 2015; Logar et al., 2020, str. 91, 104), a jih kljub temu do sedaj ni polno upošteval še noben slovarski projekt.

Ni pa zgolj sovražni govor tisti, ki ga je treba v slovarjih obravnavati posebej pozorno. Kritično slovaropisje opozarja, da je treba pri slovarskih opisih izrecne (in nove) rešitve iskati pri vseh elementih, ki prinašajo vljudne in nevljudne vidike jezika, tabuiziranost, so usmerjeni v vrednotenje, konotacijo, kulturne aluzije ipd., še posebej pa je treba biti pozoren na nestabilna in spreminjajoča se poimenovanja vseh oblik drugosti (Moon, 2014, str. 85). Pri tem se sodobno slovaropisje ne more sklicevati na tradicionalne modele jezikovnega opisovanja in delovanja. Nikakor pri tem ni sprejemljivo tradicionalno razmišljanje, da "je slovar metajezikovni odsev dejanske hierarhizirane konceptualizacije sveta" (Vidovič Muha, 2013, str. 7), kar vodi v razpravljanje o resnicah v okviru slovaropisnega dela – prav nasprotno: slovaropisje mora jasno naslavljati vprašanja, ki so v svojem bistvu ideološka, saj gre za "uravnoteževanje opisa tega, kar prinašajo podatki glede pomena, s tem, na kakšen način 'naj bi bil' v postmoderni vključujoči družbi določen koncept obravnavan in predstavljen" (Moon, 2014, str. 89). Gre torej za to, da pri slovaropisnem delu končne rešitve preprosto ne morejo biti "samo jezikoslovne; neizogibno morajo biti tudi ideološke" (Moon, 2014, str. 94). Pomembno je, da se ideološkosti pri slovarskih opisih zavedamo, da odkrito in jasno povemo, da je slovaropisno delo težavno prav zato, ker je tudi ideološko (Gantar, 2015, str. 399), še posebej pri družbeno občutljivih elementih slovarja.

2.2 Od oznak k daljšim pojasnilom

Oznake in pojasnila o rabi pri slovarskih opisih naj bi bile vključene kot informacija za govorce določenega jezika predvsem kot pomoč pri odločanju o ustrezni rabi besed v določenem kontekstu (Namatende Sakwa, 2011, str. 305). Pri tem se uporablja različne sisteme oznak in pojasnil, ki so si bili v času tiskanih slovarjev v veliki meri podobni. V slovenskem okolju se pri označevanju besedišča tudi pri nastajajočih novih slovarskih opisih v veliki meri prevzema uveljavljenega s Slovarjem slovenskega knjižnega jezika, ki tako izhaja iz obdobja tiskanega slovarja, kjer se je besedišče označevalo z oznakami tipično pred pomenskim delom iztočnice, v vsebinskem smislu pa se umešča v razumevanje slovarja kot informativno-normativnega, torej predvsem v razmerju do razumevanja koncepta knjižnega jezika in označevanje besedišča, ki izhaja iz njega (Kalin Golob in Gantar, 2015, str. 452). To pri oznakah pomeni osredinjenje na pojasnila o omejitvah pri jezikovni rabi glede na rabo v knjižnem jeziku.

V sodobnem slovaropisju je tako zaradi velike količine korpusnih podatkov kot tudi novega medija slovarjev možno prikazati informacije, ki so bile prej podane s klasičnim sistemom slovarskih oznak, na različne nove načine (Kosem, 2015, str. 483), pri tem čemer je smiseln razmislek tako o tehnični izvedbi označevanja v novem slovarskem mediju kot vsebinski. V primeru negativno zaznamovanega besedišča v SSSS je pri vsebinskem označevanju potreben razmislek, kako opozoriti na okoliščine rabe in podati informacijo o pragmatičnem pomenu (Šorli, 2014, str. 480; Šorli, 2015, str. 480), in sicer na način, ki bo slovarskemu uporabniku dal jasno informacijo ne le o zaznamovanosti, ampak na podlagi analize korpusnih podatkov tudi o nameri govorcev, ko je ta ključna sestavina pomena, ko je del pragmatičnega pomena namera govorca, da npr. izraža sovraštvo ali užali. Poleg jasne kratke oznake o zaznamovanosti želimo podati tudi pojasnilo, ki uporabniku daje jasno informacijo o tem, kaj njegova jezikovna izbira povzroča.

3 Problemi SSSS 1.0

SSSS 1.0 je pripravljen strojno in povsem ročno nepregledan. V tej različici slovarja so kot iztočnice in sopomenke navedene leme (brez

besednih vrst), pomensko členitev in opis začasno nadomeščajo strojno pripravljene pomenske gruče, slovar pa tudi ne vsebuje slovarskih oznak, razen področnih.

Raziskava o odnosu uporabniške skupnosti do SSSS 1.0, v kateri je sodelovalo 671 anketirancev, je pokazala naklonjenost do večine novosti, ki jih prinaša (odzivni) slovar, npr. stalno posodabljanje, strojni postopki, digitalni format, kolokacijski podatki, povezave na korpus, uporabniško vključevanje (Arhar Holdt, 2020, str. 470). Med problematičnimi značilnostmi sta bili izpostavljeni nezanesljivost strojno pridobljenih podatkov in primanjkljaj slovarskih oznak tako pri jedrnih in bližnjih sopomenkah kot pri uporabniško dodanih. To, da ni oznak, je motilo 37 % sodelujočih (ibid., str. 472).

Pomanjkljivosti SSSS 1.0, ki jih je izpostavila uporabniška skupnost, so pri negativno zaznamovanem besedišču še posebej pereče. Na eni strani se strojno pripravljene iztočnice in sopomenski kandidati pojavljajo brez oznak ali opozoril tudi pri izrazito problematičnih primerih, kot je npr. iztočnica *buzi* s sopomenkami *peder*, *buzerant*, *toplovodar*, *homič*, *poženščen moški*. Na drugi strani je problem potencialno zavajajoča (ne)zastopanost sopomenskega gradiva, npr. vse sopomenke, ki jih najdemo pri iztočnici *zmaj – ksantipa*, *vešča*, *strupenjača*, *babura*, *coprnica*, *pošast*, *kričava ženska* – so vezane na ženski spol in imajo izrazito negativno konotacijo, čeprav se beseda rabi tudi za moške in (npr. v pomenu članov športnega kluba) brez negativne konotacije.

Tudi kolokacije in zgledi, ki so namenjeni primerjavi rabe dveh sopomenk, so iz referenčnega korpusa izvoženi strojno in so v slovarju brez oznak. Posledica je lahko sopostavitev pomensko neustreznih podatkov, npr. pri primerjavi besed *ženska – kura* najdemo prekrivne kolokacije [*stara, prava, gola*] *ženska* in [*stara, prava, gola*] *kura* ali *ženska* [*brez glave, v postelji, na odru*] in *kura* [*brez glave, v postelji, na odru*]. Korpusni zgledi načeloma pomagajo razdvoumiti problematične primere, vendar niso na voljo za vse primerjane besede, zgledi, ki so na voljo, pa niso izbrani po vsebinskih kriterijih. To je zlasti problematično pri sovražnem besedišču, npr. kolokacije [*sovražiti, tepsti, ubiti*] *pedra* ali zgledi tipa *In reskiral sem celo, da bi me imel za pedra.*

Določene težave se pojavljajo tudi pri uporabniško predlaganih sopomenkah. Tu ločujemo na eni strani zlonamerne vnose, kot je npr.

uporabniški vpis *aljaz* pri iztočnici *gej*. Za takšne primere bi bilo treba določiti natančno uredniško politiko za sprotno obravnavo na ravni vmesnika. Na drugi strani uporabniki zaznamovano besedišče dodajajo kot dejanski sopomenski predlog, npr. pri iztočnici *južnjak*, kjer so uporabniki dodali dolg niz predlogov, mdr. *jugovič*, *južni brat*, *jugič*, *trenirkar, bosanec*, *z juga*. V SSSS 1.0 so nekateri uporabniki in uporabnice oznako ali kako drugo pojasnilo v oklepaju pripisali ob svoj sopomenski predlog, npr. *bojazljivec – pezde (vulg.), Italijanka – makaronarka (slabš.),* vendar je bilo takšno označevanje sporadično in nesistematizirano. Uredniška naloga je presoditi, kateri predlogi so relevantni za vključitev v slovarsko bazo (in s katerimi slovarskimi oznakami), že uporabnikom pa omogočiti, da problematično besedišče označijo kot tako, da se torej oznaka v vmesniku prikaže hkrati z dodano sopomenko.

Različico SSSS 2.0 smo pripravili pod okriljem projekta Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL, ki ga je med leti 2021–22 financiralo Ministrstvo za kulturo Republike Slovenije. V slovarsko nadgradnjo smo vključili več ciljev, ki naslavljajo zgoraj naštete probleme: (a) razviti in preizkusiti sistem označevanja negativno zaznamovanega besedišča znotraj koncepta odzivnega slovarja; (b) identificirati besedišče, ki je negativno zaznamovano v vseh pomenih besede in ga je zato mogoče označiti na ravni leme, in ga označiti po celotnem slovarju SSSS; (c) na izbranem naboru gradiva preizkusiti označevanje negativne zaznamovanosti pri pomensko členjenih slovarskih geslih ter (d) dodati v slovarski vmesnik možnost, da uporabniki sami označijo svoje predloge. V nadaljevanju predstavimo izhodišča, metodologijo in rezultate za vsakega od ciljev, prikažemo primere implementacije in vizualizacije v prenovljenem slovarskem vmesniku, prispevek pa sklenemo z opredelitvijo načrtov in prioritet za nadaljnji razvoj slovarja.

4 Identifikacija in označevanje problematičnega besedišča

4.1 Slovaropisna izhodišča in sistem oznak

Prepoznavanje potencialnega, z vidika družbene občutljivosti problematičnega besedišča temelji na slovaropisnih izhodiščih, ki jih pripravljamo

za slovarske vire na CJVT UL, prvič pa smo jih začeli uveljavljati pri izdelavi Velikega slovensko-madžarskega slovarja (Kosem et al., 2018a). V izhodišča je vključeno prepoznavanje elementov sovražnega govora (oznaka *sovražno*), elementov nevljudnosti, žaljivosti (*grobo*) ter elementov negativnega vrednotenja ali konotacije (*izraža negativen odnos*). Omenjene vrednotenjske oznake sodijo v širši okvir t. i. sporočanjskih oznak,[2] ki opredeljujejo izraze ali pomene z vidika njihove rabe v sporočanjskem procesu in v situacijah, v katerih sporočanje poteka. V predlaganem slovaropisnem opisu so sporočanjske oznake namenjene označevanju izrazov, z izbiro katerih govorci dosegamo ali želimo doseči določen učinek pri naslovniku. Ta učinek je lahko povzročen s pozitivnim ali negativnim vrednotenjem, z uporabo v določenem govornem položaju (npr. javnem, nejavnem) ali z namenom izraziti odnos do predmetnosti ali vsebine, ki temelji na določenih družbenih normah, pričakovanjih in odstopanjih od njih. Ta sistem se od tradicionalnega označevanja besed na podlagi odnosa do knjižne norme, kot ga pozna SSKJ (t. i. stilno-zvrstni in ekspresivni kvalifikatorji), ločuje v kvalificiranju besedišča na podlagi sporočanjskega namena in učinka, pri čemer izhodišče kvalificiranja ni v opozarjanju na odstop od knjižne norme, pač pa v informiranju glede možnih posledic rabe. S takim sistemom se želimo izogniti morebitnemu kvalificiranju govorca samega, hkrati pa opozoriti na kontekst potencialno problematične rabe v informativnem smislu. To pomeni, da ne želimo uporabnikov slovarja obveščati samo o možnih učinkih rabe grobega in sovražnega besedišča, pač pa posredno pokazati tudi na okoliščine, v katerih je tako rabo mogoče prepoznati.

V slovarskem sistemu vrednotenjskih oznak s t. i. negativnega pola[3] označujemo z oznako *sovražno* izraze in pomene, ki so diskriminatorni, ksenofobični, rasistični in homofobični, ki so uperjeni proti predstavnikom skupin ali manjšin na podlagi njihove narodnosti, rase ali etničnega porekla, verskega prepričanja, spola, zdravstvenega stanja, spolne

2 Celotni sistem označevanja, ki ga razvijamo v okviru virov CJVT UL, poleg sporočanjskih oznak, ki jih notranje členimo na vrednotenjske, registrske in stilne, zajema še nabor pragmatičnih, kontekstualnih, področnih, slovničnih, časovnih in trendovskih oznak ter nabor oznak, vezanih na tuja poimenovanja in prevodne ustreznice.

3 Poleg oznak negativnega vrednotenjskega pola uporabljamo tudi pozitivni vrednotenjski oznaki “izraža pozitiven odnos” in “lahko izraža pozitiven odnos” (npr. pri *hribovka, blazen*). Oznako “lahko izraža negativen odnos” uporabljamo za omilitev negativne konotacije (npr. pri *češplja*) .

usmerjenosti, invalidnosti ter drugih lastnosti in prepričanj. Z oznako *sovražno* se torej opredeljujemo do vseh izrazov, ki spodbujajo sovraštvo, predsodke ali nestrpnost in s tem lahko predstavljajo – kot je bilo opredeljeno že v razdelku 2 – elemente sovražnega govora.

Na drugi strani z oznako *grobo* označujemo izraze ali pomene, ki so za naslovnika lahko žaljivi, z vidika družbenih in moralnih norm pa neprimerni. Tipično se nanašajo na človeško ali živalsko telo, spolnost, prehranjevanje in izločanje – zlasti torej na tabuizirano predmetnost – in so rabljeni v neformalnih govornih situacijah.

Tretji sklop predstavlja besedišče, ki izraža neodobravanje, nenaklonjenost, posmehljivost ali kritiko do lastnosti posameznikov, predmetov ali dejanj. Z oznako *izraža negativen odnos* želimo tako opozoriti na izraze z negativno konotacijo ali vrednotenjem, ki so lahko za naslovnika žaljivi ali neprijetni.

4.2 Študentska identifikacija potencialno problematičnega besedišča

Potencialno problematično besedišče v SSSS smo identificirali z ročnim pregledom iztočnic in sopomenk v slovarju. Omejili smo se na slovarske (jedrne in bližnje) sopomenke, saj pregled uporabniških predlogov zahteva dodatne uredniške premisleke in bo zato opravljen kasneje s prilagojeno metodologijo. Zaradi obilja gradiva smo delo organizirali v dva koraka: širši pregled, v katerem smo v grobem ločili potencialno problematično in neproblematično gradivo, nato pa natančnejši pregled problematičnih primerov.

Najprej smo s specializirano programsko skripto iz slovarske baze izvozili nize sopomenk, urejenih na podlagi pomenskih gruč (Krek et al., 2017), npr. *speljati se; izginiti; pobrati se; skidati se; spokati se; spizditi*, pri čemer smo strojno odstranili nize, ki so se glede nabora sopomenk podvajali, in tiste, ki so bili podmnožica kakega drugega niza. Na tak način smo pripravili 65.615 nizov različnih obsegov: od posameznih sopomenskih parov do zelo dolgih nizov, ki pa so redki: več kot 30 sopomenk vsebuje le 156 nizov, povprečje je 5 sopomenk na niz.

Čeprav strojno pomensko gručenje ni povsem natančno in se razlikuje od slovaropisne pomenske členitve, tovrstna organizacija podatkov

dobro naslovi dva pomembna problema: (a) tak pristop bistveno pohitri pregledovanje, kot bo razvidno v nadaljevanju; (b) presojanje je lahko bolj natančno, saj problematičnost posamezne leme nakazujejo ostale besede v nizu, prim. npr. *nategniti* v nizu *raztegniti; dilatirati; iztegniti; nategniti; pogrniti; razgrniti; razmakniti; razpreti; razprostreti; razviti; napeti; zavlačevati z; razpeti; prolongirati* in v nizu *pokavsati; nategniti; povaljati; porivati; pofukati; pojahati*.

Iz množice 65.615 nizov smo nato odstranili 24.945 nizov (38,0 %), pri katerih sopomenke vsebujejo področne oznake, npr. *odbojnik, deflektor, ločilnik, membrana, opna, odbojna pregrada, zvočna stena* z oznako *elektrika*. Ti podatki so terminološke narave, zato smo predvidevali zanemarljivo nizko vsebnost problematičnega besedišča. Predpostavko smo preverili s pregledom 200 naključno izbranih nizov, v katerih se problemi skladno s pričakovanji niso pojavljali. Preostalega terminološkega gradiva nismo ročno pregledovali. Ostalo je 496 nizov (0,8 %), ki vsebujejo lastnoimenske samostalnike, npr. *Antarktika, antarktično območje, južno polarno območje,* in 40.176 (61,2 %) občnoimenskih nizov, ki so bili vsi vključeni v ročni pregled.

Podatke so pregledovali študentke in študenti jezikoslovnih smeri, in sicer po trije vzporedno. Pregledovanje je potekalo v okolju *Google Sheets*. Sopomenske nize smo organizirali v vrstice tabele, kjer jim je bilo mogoče pripisati eno od naslednjih odločitev: (1) niz vsebuje sovražno ali grobo besedišče; (2) niz vsebuje besedišče, ki je drugače negativno ali (v določenem pomenu, kontekstu) izraža negativen odnos; (3) z vidika sovražnosti, grobosti, negativnosti je niz neproblematičen. Če so pregledovalci želeli, so lahko opredeliti tudi, da je (4) v nizu kako drugače zaznamovano besedišče, da (5) ne razumejo vseh besed v nizu, lahko pa so vpisali tudi dodaten komentar na svoje odločitve ali podatke.

Odločitve so študentke in študenti sprejemali na podlagi kratkih navodil za prepoznavanje potencialno negativne zaznamovanosti besed, kot so elementi sovražnega govora (rasna, etična diskriminacija, diskriminacija na podlagi spola, spolne usmerjenosti, hendikepa) negativnega vrednotenja (glede na družbeni status, gmotni položaj, obnašanje in značaj, izgled ipd.) in grobosti (glede na tabuizirano predmetnost, npr. spolnost, telesno izločanje, nasilje, in tipično neformalnost

govornega položaja). Pri mejnih primerih so izbrali tisto izmed odločitev, ki nakazuje višjo stopnjo negativnosti (npr. pri mejnih primerih med kategorijo 1 in 2 so izbirali 1).

Študentski pregled je zajemal 40.672 sopomenskih nizov, ki jih je pregledalo 6 sodelujočih, kar pomeni, da je vsak v povprečju pregledal 20.336 nizov. Priprave na označevanje (priprava podatkov in smernic, testno označevanje in uvodni sestanki za reševanje izhodiščnih dilem) so potekale v januarju in februarju, pregledovanje pa od marca do maja 2022, pri čemer je treba upoštevati, da zaradi drugih obveznosti študenti na projektih sodelujejo z omejenih naborom tedenskih ur. V povprečju so v navedenih treh mesecih porabili 72 ur na osebo, kar pomeni 12,7 sekund na sopomenski niz. Kljub veliki količini podatkov je bila naloga torej izvedljiva v relativno kratkem času, saj so študentje lahko odločitev podali takoj, ko so v nizu našli eno samo problematično besedo, natančnejše razmisleke o vrsti zaznamovanosti oz. označevanja posameznih besed pa so prepustili za drugi korak dela s podatki.

Vsak sopomenski niz so pregledali po trije študenti in študentke. Pri nalogi nas ni zanimalo ne/ujemanje med njihovimi odločitvami (čeprav posredno nakazuje težja mesta za označevanje), ampak je trojni pregled služil za zagotavljanje večje natančnosti pri ločevanju neproblematičnega gradiva od tistega, ki potrebuje nadaljnji pregled. Posamezne študentske odločitve smo pretvorili v skupne po naslednjem ključu: (1) **sovražno/grobo:** če je vsaj eden od študentov presodil, da se v nizu pojavlja sovražno ali grobo besedišče; (2) **drugače negativno:** kombinacije odločitev "druga negativnost" in "neproblematično" ali (3) **neproblematično:** če so vsi študenti presodili, da je z vidika sovražnosti, grobosti, negativnosti niz neproblematičen. Rezultate prikazuje Tabela 1.

Tabela 1: *Številčna zastopanost in delež nizov glede na skupno študentsko odločitev potencialne problematičnosti*

Kategorija končne odločitve	Število nizov v kategoriji	Delež glede na vse pregledano
Sovražno/grobo	1.810	4,5 %
Drugače negativno	12.730	31,3 %
Neproblematično	26.132	64,3 %
Skupaj	40.672	100,0 %

V Tabeli 2 navajamo nekaj nizov s po tremi sopomenkami, ki so jim študentke in študenti pripisali skladne ali različne odločitve. Kot je razvidno, lahko posamezen niz vsebuje raznoliko zaznamovano besedišče, kot tudi nezaznamovano besedišče. Sopomenske nize z najvišjo stopnjo negativne zaznamovanosti (odločitev 1) smo na projektu natančneje obravnavali, kot opisujemo v nadaljevanju. Z ročnim pregledom pa smo identificirali tudi gradivo, ki je na tak ali drugačen način relevantno za nadaljnje delo (odločitev 2), in gradivo, ki ga z vidika negativne zaznamovanosti ne bomo nadalje obravnavali (odločitev 3).

Tabela 2: *Primeri nizov s študentskimi odločitvami o nadaljnji obravnavi*

Niz sopomenk	Posamezne in skupna odločitev
fukati; porivati; natepavati	111 -> 1
skozlati; izbruhati; zbruhati	111 -> 1
pedrski; buzerantski; toplovodarski	111 -> 1
črnuhinja; zamorka; zamorklja	111 -> 1
pofukanka; prasica; zajebanka	111 -> 1
debilen; bebast; duševno zaostal	121 -> 1
kripelj; pohabljenec; pohabljenka	211 -> 1
kurnik; pajzelj; temačna luknja	222 -> 2
bedastoča; glupost; nesmisel	222 -> 2
eliminirati; likvidirati; usmrtiti	222 -> 2
izmozgano; izčrpano; mršavo	223 -> 2
imenski; nazivni; nominalni	333 -> 3
kopirni papir; indigo; karbon	333 -> 3
zaustaviti se; izklopiti se; izključiti se	333 -> 3

V "drugače negativno" so raznorodni primeri, saj so poleg zaznamovanih izrazov in pomenov (npr. *budala*, *avša*, *bedast*) študentje označevali tudi vrednotenjsko nevtralno besedišče, ki poimenuje negativne vsebine, dejanja in predmetnost. Gre zlasti za poimenovanja agresivnega obnašanja: *uničiti*, *dotolči*, nekaterih osebnih lastnosti: *pokvarjen*, *hudoben*, *ničvreden*, *grozljiv*, *grd*, *apatičnost, pokvarjenost*; videza, stanja: *neurejenost*, *razdejanje*, *zanikrnost* itd. V slovarju večina teh besed ne potrebuje oznake. Čeprav teh besed ne bomo označevali, so seznami tovrstnega besedišča pomemben rezultat ročnega pregleda, saj so koristni za različne druge namene na področju slovaropisja in strojne obdelave jezika, npr. za filtriranje gradiva z negativnim pomenom iz jezikovnih iger ali učnih gradiv, strojno pripisovanje sentimenta ipd.

4.3 Slovaropisni pregled in izbira slovarskih oznak

V 1.810 nizih z odločitvijo (1) smo določili besede in zveze, ki so relevantne za slovarsko označevanje z oznako sovražno ali grobo. Prvi izpis problematičnega gradiva iz sopomenskih nizov so pripravili študenti in študentke, nato pa je odločitve pregledala, sprejela in mestoma spremenila skupina treh jezikoslovcev. Odločanje je potekalo ob upoštevanju pojavljanja oz. rabe identificiranega besedišča v korpusih Gigafida 2.0 in Janes. Kvalitativna empirična analiza je dopolnila željo po pohitritvi prve selekcije, s čimer smo se obranili pred intuitivnim ali črno-belim presojanjem primernosti. V analizi smo identificirali tako primere, ki jih je mogoče označiti na ravni leme (npr. *črnuh, razpizditi*), kot primere, pri katerih bi bilo oznako *sovražno* ali *grobo* mogoče pripisati enemu ali več (ne pa vsem) pomenom besede (npr. *zamorec, debilen*; *nategniti, batina*).

Jezikoslovna analiza je razkrila tudi določen delež gradiva, ki je v slovar SSSS prišlo pomotoma oz. zaradi specifik v metodologiji njegove priprave iz baze Velikega angleško-slovenskega slovarja Oxford-DZS. Iz SSSS smo tako odstranili identificirane napačno izluščene primere (npr. *zamoreka, poneediti se*), množinske leme, ki se v bazi sicer pojavljajo tudi v ednini (npr. *babe, švabi*), in zveze, ki so v izvornem slovarju nastopale kot prevodne ustreznice razlagalnega tipa, v SSSS pa kot iztočnice nimajo pravega smisla (npr. *neuglajena podeželanka, bogat vulgarnež*). Rezultate kaže Tabela 3.

Tabela 3: *Številčna zastopanost in primeri iztočnic (za pripis oznak ali za pomensko členitev)*

	Primeri	**Število besed oz. zvez**
Oznaka *sovražno* na ravni leme	*črnuh, cigo, čifut, rdečuhinja, beli prasec, bela sodrga, lezba, lezbača, peder, buzerant*	27
Oznaka *grobo* na ravni leme	*podjebavati, v kurcu, zdrkati, zrajcan, pofafati ga, sranje, fentati, razpizden, sfukan, kurbarija*	140
Odstranitev iz SSSS	*zamoreka, poneediti se, babe, tipi, švabi, neuglajena podeželanka, bogat vulgarnež, črnski yuppie, bela golazen*	86
Besedišče za pomensko členjenje in (morebitno) označevanje na ravni pomenov	*baba, batina, blazen, češplja, coprnica, črv, debil, kmetica, kripelj, nabrisati*	234

Primere, kjer je problematičnost vezana na lemo ne glede na morebitno večpomenskost, smo v SSSS 2.0 označili in vizualizirali, kot prestavlja poglavje 6. Kjer se oznaka nanaša na posamezni pomen, pa smo iztočnice pomensko razčlenili in jim pripisali pomensko informacijo v obliki t. i. pomenskega indikatorja, kot predstavljamo v poglavju 5. Omeniti velja tudi, da smo pri jezikoslovni analizi veliko gradiva, ki so ga študenti umestili v kategorijo (1), premestili v kategorijo (2), kar pomeni, da ga bomo pri nadaljnjem razvoju slovarja predvidoma označili z oznako *izraža negativen odnos,* npr. *trapa, kozlarija, špeglarca, luftar.*

5 Pomenska obdelava iztočnic

Iztočnice, ki so bile s seznama relevantnih za slovarsko označevanje določene za pomensko analizo, so morale biti večpomenske, pri čemer so ob enem ali več vrednotenjsko nezaznamovanih pomenov morale nakazovati vsaj en, lahko pa tudi več negativno vrednotenih pomenov, npr. *jalov, batina, črv, gnoj, prasica* itd. Takih iztočnic je bilo na seznamu 234. Pomenska analiza je vključevala pomensko razčlenitev, ki je obsegala določitev števila in zaporedja pomenov, ter pomenski opis s pomočjo t. i. pomenskih indikatorjev.

5.1 Pomenska členitev

Znotraj pomenske obdelave iztočnic je proces pomenske členitve potekal na podlagi analize rabe besede v referenčnem korpusu standardnega jezika Gigafida 2.0 ter v korpusu Janes, za katerega velja, da se jezik v njem v marsičem razlikuje od pisnega standarda (Fišer et al., 2016, str. 68). Čeprav je pomenska členitev vključevala prepoznavanje vseh pomenov analizirane besede, torej tudi vrednotenjsko nezaznamovanih, smo bili pri pomenski analizi pozorni zlasti na pomene, ki izražajo katero od negativnih vrednotenj, še zlasti, če je bilo teh pomenov več. Večpomenskost je bilo mogoče prepoznati pri več samostojnih pomenih z negativno konotacijo, (npr. *bastard, degeneriran*) in pri stopnjevanju negativne konotacije (*cigan, debil, češplja*):

bastard

1. *izraža negativen odnos* nezakonski otrok
2. *izraža negativen odnos* izprijena, nasilna ali uporniška oseba

cigan
2. *izraža negativen odnos* podlež; lopov
4. *sovražno* pripadnik etnične skupine

debil
1. *izraža negativen odnos, lahko ironično* nepremišljena, nespametna oseba
2. *sovražno* oseba z motnjo v razvoju

češplja
2. *lahko izraža negativen odnos, neformalno* ženska ali dekle
3. *grobo, neformalno* ženski spolni organ

5.2 Pomenski opis

Za pomenski opis pomensko razčlenjenih besed smo uporabili t. i. pomenske indikatorje (Gantar, 2015, str. 164), ki so eden od treh segmentov pomenske informacije v slovarskih virih CJVT. Poleg indikatorjev (v zgornjih primerih na drugem mestu) pomenski opis tvorijo še oznaka (na prvem mestu) in razlaga, ki je v zaenkrat SSSS ne predvidevamo.

Osnovna vloga pomenskega indikatorja je na kratko in prepoznavno opredeliti pomen besede glede na njene druge pomene. Pomenski indikatorji so za to v prvi vrsti namenjeni oblikovanju t. i. pomenskega menija, ki ga poznamo iz tujejezičnih slovarjev za tujce, uvajamo pa ga tudi v slovarske vire CJVT (prim. Kolokacije 1.0, Veliki slovensko-madžarski slovar 1.0). Specifičnost posameznega vira, v našem primeru slovarja sopomenk, narekuje tudi način oblikovanja indikatorjev. Ker gre za pomene z negativno konotacijo, je bilo treba premisliti ubesedenje indikatorjev z vidika opisovanja družbeno občutljivih vsebin na eni strani ter zagotavljanja pomenske obvestilnosti in konsistentnosti na drugi. Pri oblikovanju indikatorjev smo si zato prizadevali ohranjati čim večjo nevtralnost oz. splošnost ubeseditve, zlasti pri večpomenskih besedah, ki poimenujejo posameznike, npr. "oseba" (*debil, gnoj, kača, prasica*). V primeru ženskih oblik, ki se tudi uporabljajo zgolj za ženske osebe, smo ohranili indikator "ženska" (*češplja, klošarka, debilka*). Razlog za to je odraz dejanske rabe, je pa v zvezi s temi primeri smiselno razmisliti tudi o vlogi definicije

kot tretje pomenske informacije, ki lahko prevzame pojasnjevanje specifičnih pomenskih lastnosti, indikator pa bi zato lahko ostal bolj splošen tudi pri ženskih oblikah.

Pri besedah, ki označujejo telesni ali duševni hendikep, smo skušali čim bolj konsistentno slediti uporabi nevtralnih in družbeno korektnih indikatorjev, npr. ‘oseba z motnjo v razvoju’ (sovražno: *debil, imbecil*), ‘o motnji v razvoju’ (izraža negativen odnos: *debilnost*), ‘o osebi z motnjo v razvoju’ (izraža negativen odnos: *defekten, bebast*); ‘o telesni okvari’ (izraža negativen odnos: *defekten, liliputanski*) ‘o osebi s telesno okvaro’ (izraža negativen odnos: *kruljav*), ‘oseba s telesno okvaro’ (sovražno: *kripelj, nakaza*).

Na drugi strani pa zlasti enopomenskih besed in besed z več izrazito vrednotenjskimi pomeni ni mogoče pomensko opisati s povsem nevtralnimi indikatorji. V takih primerih smo indikatorje ubesedili tako, da ustrezno ponazarjajo pomensko specifiko besede in se tako na nek način približujejo razlagam, npr.

blazen
3. *izraža poudarek, lahko izraža pozitiven odnos* o nenavadnem ali neverjetnem

psiho
1. *izraža negativen odnos, neformalno* čudaška ali nevarna oseba

Tudi pri t. i. nenevtralnih indikatorjih smo skušali slediti konsistentni ubeseditvi, če je bilo z njo mogoče zajeti sorodne pomenske lastnosti besede, npr. ‘ničvredna, izprijena oseba’ (izraža negativen odnos: *degeneriranec, podgana, pes*), ‘ničvredna oseba’ (izraža negativen odnos: *rit, ušivec*), ‘nepreudarna, nespametna ženska’ (izraža negativen odnos: *gos, goska, kura*), ‘neuglajena, nespametna oseba’ (izraža negativen odnos: *govedo, kmet*).

S problemom nevtralne ubeseditve indikatorjev smo se tokrat prvič bolj poglobljeno ukvarjali prav zaradi kvalificiranja in pomenskega opisovanja družbeno občutljivih vsebin. Menimo, da bi bilo iskanje rešitev za ustrezno ubeseditev pomenske informacije pri družbeno občutljivih vsebinah v prihodnje smiselno nadgraditi tako z jezikoslovnimi analizami besedilnega okolja kot s sociolingvističnimi in žanrskimi analizami

okoliščin rabe. Na ta način bi bilo mogoče prepoznati tudi druge, ne zgolj leksikalne možnosti vrednotenja in pridobiti vpogled v celostno pomensko sliko. Spoznanja bodo koristna tudi pri oblikovanju strategij za vključitev tovrstne informacije v celostni pomenski opis leksike v Digitalni slovarski bazi in posledično v jezikovne vire CJVT.

6 Rešitve v SSSS 2.0

V slovarskem vmesniku SSSS 2.0[4] na sovražno in grobo besedišče opozarjamo s kombinacijo opozorilne ikone in daljšega pojasnila, ki se izpiše ob kliku nanjo. V različici 2.0 slovarske oznake *izraža negativen odnos* še nismo pripisovali, bo pa na voljo uporabnikom, zato smo pripravili pojasnila za vse tri vrednotenjske oznake hkrati (Tabela 4). Pri vizualizaciji smo se namenoma odrekli pripisovanju (eno-)besednih oznak, saj bi te pri označevanju (mestoma tudi homonimnih) lem lahko vodile v napačno interpretacijo podatkov. Pri pomensko členjenih geslih so oznake pripisane posameznim pomenom, pri pomensko nečlenjenih geslih pa kombinacija ikone in pojasnila omogoči, da je problematično besedišče na prvi pregled zelo opazno, pojasnilo pa je lahko daljše in vsebuje informacije o možnem učinku na naslovnika oz. možnih posledicah rabe označene besede.

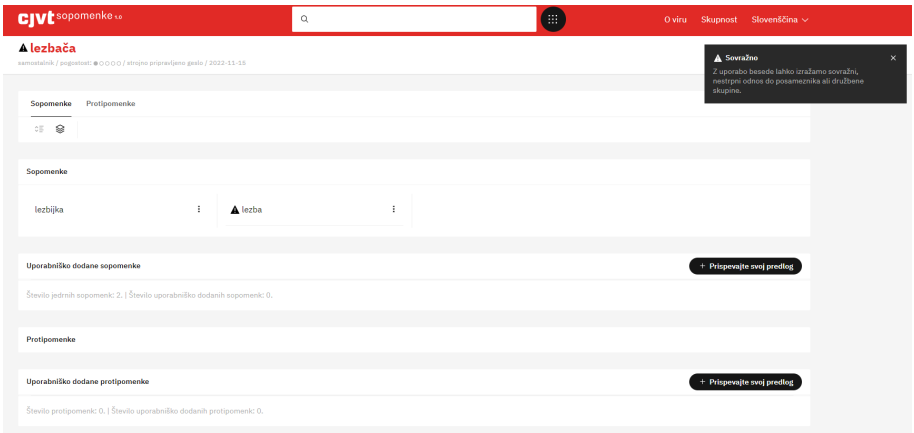
Tabela 4: *Ikone in njihova pojasnila v SSSS 2.0*

Oznaka	Ikona	Pojasnilo
Sovražno		Z uporabo besede lahko izražamo sovražni, nestrpni odnos do posameznika ali družbene skupine.
Grobo		Zaradi družbenih in moralnih norm se marsikateremu uporabniku jezika beseda lahko zdi groba ali neprimerna. Uporaba lahko povzroči nelagodje, razburi ali užali.
Izraža negativen odnos		Beseda lahko ni nevtralna. Z uporabo besede se lahko posmehujemo, izražamo neodobravanje ali kritiko do nekaterih lastnosti posameznikov, predmetov ali dejanj.

Slika 1 kaže posnetek zaslona vmesnika SSSS 2.0 Slika ponazarja, kakšna bo vizualizacija z ikonami, ki lahko stojijo pri iztočnici (*lezbača*) ali pri sopomenki (*lezba*). Klik na ikono odpre pojasnilo desno zgoraj – v

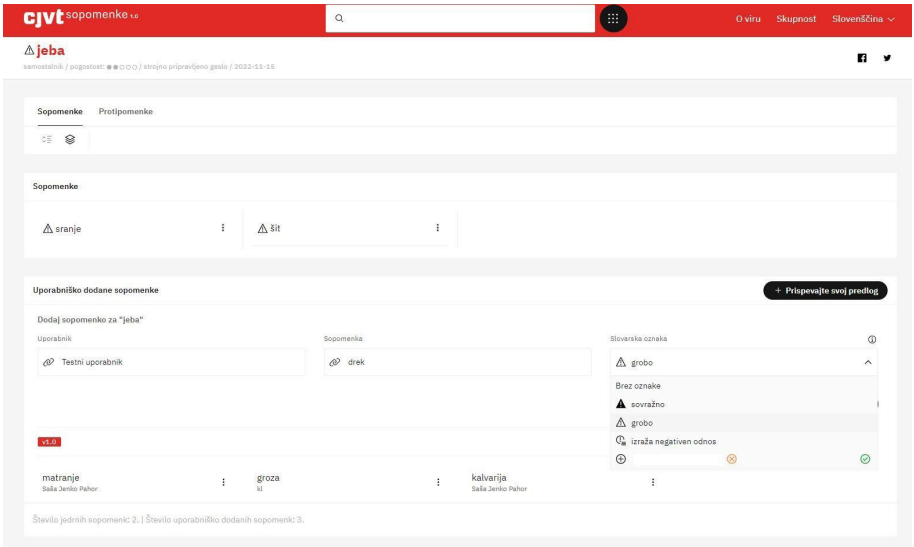
4 V času priprave prispevka je razvojno različico slovarja mogoče videti na naslovu: https://viri.cjvt.si/sopomenke-beta/slv/.

obliki pojavnega okenca, ki po nekaj sekundah samo izgine z ekrana. Razvidne so tudi nekatere druge novosti slovarja, ki se jim v tem prispevku ne posvečamo, npr. prenovljena oblikovna podoba vmesnika, informacija o besedni vrsti in pogostnosti iztočnice ter možnost dodajanja protipomenk. Ker vse težave SSSS niso enostavno in hitro rešljive, smo želeli slovarske uporabnike bolje opozoriti na trenutne omejitve. Čeprav je bila metodologija priprave SSSS 1.0 pojasnjena v razdelku *O viru*, pri samih iztočnicah ni bilo izrecnih opozoril, da je slovar pripravljen strojno, in to na vseh ravneh: sopomenke, kolokacije, korpusni zgledi, kar je lahko vodilo v napačne interpretacije slovarske vsebine. V SSSS 2.0 smo zato v zaglavje vključili tudi status izdelanosti gesla (*strojno pripravljeno geslo* na Sliki 1).



Slika 1: Iztočnica *lezbača* v SSSS 2.0: klik na opozorilno ikono odpre pojasnilo.

Protokol dodajanja sopomenk (in skladno tudi protipomenk) smo nadgradili, da bodo predlagani besedi ali zvezi uporabnice in uporabniki lahko dodali tudi slovarsko oznako oz. oznake. Privzeta izbira je, da je uporabniški predlog "brez oznake", ostale možnosti so na voljo v spustnem meniju (Slika 2). V različici SSSS 2.0 so na klik na voljo oznake *sovražno*, *grobo* in *izraža negativen odnos*, poleg tega pa okence, v katerega je mogoče vtipkati morebitno drugo oznako.



Slika 2: Iztočnica *jeba* v SSSS 2.0: uporabniško dodajanje nove sopomenke in slovarske oznake.

Pomen in raba oznak *sovražno*, *grobo* in *izraža negativen odnos* je razložena in ponazorjena s primeri, s čimer bo lahko dosežena določena stopnja enotnosti uporabniškega označevanja (informacije so na voljo na klik, gl. ikono (i) na Sliki 2). Predvideno pa je, da bodo uporabniki oznake mestoma interpretirali in uporabljali drugače, kot bi jih slovaropisci. Vse dodane oznake bodo (skupaj z dodanimi sopomenkami oz. protipomenkami) preverjene in uporabniški predlogi bodo dragoceno gradivo ne le za dopolnitev odprto dostopne slovarske baze, ampak tudi za analize širšega dojemanja označevalnega sistema ter dometa in meja oznak za negativno vrednotenje. Prav tako pomemben uvid bodo ponudile ročno vpisane oznake, ki jih bomo analizirali z vidika vsebine in pogostosti ter uporabili izsledke za nadaljnji razvoj slovarja.

7 Sklep in nadaljnje delo

Sodobno slovaropisno delo ima ob zavedanju ideološkosti, vključevanju novih pristopov, uporabi tehnologije, moči množic itd. danes veliko možnosti, da tudi vprašanja označevanja konotacije naslavlja na novo in zanj pripravlja inovativne rešitve (Gorjanc, 2017, str. 154).

V prispevku smo opisali, kako poteka obravnava sovražnega in grobega besedišča v SSSS in katere spremembe so na voljo v različici 2.0. Rešitve naslavljajo dve pomembni značilnosti SSSS: njegovo strojno izdelanost in odprtost, da pri razvoju slovarja sodeluje tudi uporabniška skupnost. V novi različici slovarja so sovražnemu in grobemu besedišču pripisane slovarske oznake oz. opozorilne ikone s pojasnili o možnih učinkih rabe in dodana je možnost, da uporabniki pripišejo oznako svojim predlogom sopomenk.

Prepoznano sovražno in grobo besedišče bo koristno tudi pri izdelavi drugih virov, kjer se za pomene izbirajo reprezentativne kolokacije in zgledi. Pri izdelavi novih gesel za Kolokacijski slovar sodobne slovenščine (Kosem et al., 2018b) npr. že zdaj pri pripravi podatkov (pred slovaropisno analizo) označujemo kolokacije, ki vsebujejo sovražno in grobo besedišče, pa tudi besedišče, ki izraža negativen odnos. Tako slovaropiske in slovaropisce opozorimo na potencialno problematične kolokacije in posledično pohitrimo delo oz. se izognemo vključevanju problematičnih vsebin. Seznami problematičnega besedišča, ki jih uporabljamo trenutno, so pripravljeni *ad hoc* iz odprto dostopnih jezikovnih virov in precej krajši od seznamov, ki bodo (lahko) nastali na osnovi predstavljenega dela.

Izražanje negativnega odnosa je pogosto vezano na posamezen pomen besede, zato bo velik del naloge izvedljiv šele ob pripravi pomensko členjenih gesel. Pri pomenski členitvi in nadaljnjem označevanju gradiva SSSS bomo uporabljali metodologijo, ki jo razvijamo pri izdelavi Velikega slovensko-madžarskega slovarja (Kosem et al., 2018a) in podatke oz. informacije, ki so na voljo v obstoječih odprto dostopnih virih za slovenščino. Preizkus prenosa metodologije smo izvedli pod okriljem projekta Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL, kjer je bila med cilje vključena tudi nadgradnja SSSS z 2.000 pomensko členjenimi gesli, ki imajo slovaropisno pregledane in razvrščene sopomenke ter kolokacije.

Nadaljnji razvoj v tej smeri omogoča predvsem projekt Nadgradnja portala jezikovnih virov CJVT, ki poteka v letih 2022–2023, financira ga Ministrstvo za kulturo. V okviru projekta se izdeluje urejevalnik za Digitalno slovarsko bazo (DSB – ta ob drugih tipih podatkov vsebuje tudi kolokacijske podatke in podatke o sopomenskosti), kar omogoča lažje

pomensko členjenje večje količine iztočnic. V okviru istega projekta v DSB integrirana večja količina pomenskih podatkov iz slovarjev, enciklopedij in podobnih virov, ki so bili odkupljeni v okviru projekt Razvoj slovenščine v digitalnem okolju in so bili dani v odprt dostop. Ob upoštevanju korpusnih podatkov, ki so že vključeni v slovaropisni proces, je s tem omogočeno tudi strojno gručenje pomenov (ang. *word sense induction*) na ravni vseh iztočnic, ki so trenutno vključene v slovarsko bazo.

Zato bi bilo v nadaljnje premisleke glede sovražnega in grobega besedišča znotraj koncepta odzivnega slovarja smiselno celoviteje vključiti tudi analizo okoliščin rabe, kar je do neke mere mogoče izvesti tudi strojno. Zanimivo bi bilo obravnavati zaznavanje in presojanje sovražnosti, grobosti v različnih tipih besedil, npr. medijskih. Ob tem se odpira tudi vprašanje formalnosti in neformalnosti položajev, na katere se ta presoja nanaša: ali posega na vse ravni izražanja ali gre zgolj za formalne, javne položaje in ali je neodvisna od generacijske ali kake druge pripadnosti presojevalca. V tej luči bi bilo lahko zanimivo sodelovanje jezikoslovk in jezikoslovcev s strokovnjakinjami in strokovnjaki s področja sociologije, etnologije, antropologije, psihologije, ki obravnavajo družbene okoliščine, generacijske kontekste, formalne in neformalne pozicije itd. Ob vsem naštetem se je treba zavedati, da slovaropisje naslavlja zlasti pojavnosti sovražnosti in grobosti, ki so v besedilnem kontekstu leksikalno izražene. Tudi s tega vidika bi bilo delo smiselno povezati s področji žanrske analize, pragmatike, kritične analize diskurza, strojne identifikacije sovražnega govora in podobnih pristopov, ki naslavljajo prikrite, implicitne, posredne načine izražanja sovražnosti oz. negativnega vrednotenja nasploh.

Zahvala

Projekt Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL je v letih 2021–2022 financiralo Ministrstvo za kulturo Republike Slovenije. Raziskovalne programe št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik), št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave) ter št. P6-0436 (Digitalna humanistika: viri, orodja in metode) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Arhar Holdt, Š. (2020). How Users Responded to a Responsive Dictionary: The Case of the Thesaurus of Modern Slovene. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, *46*(2), 465–482. doi: 10.31724/rihjj.46.2.1

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Logar, N., Gorjanc, V., & Krek, S. (2022). Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine. V D. Fišer & T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference,* 1. izdaja (str. 10–16), Ljubljana, Slovenija. Inštitut za novejšo zgodovino. Dostopno prek https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., & Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. V J. Čibej, V. Gorjanc, I. Kosem & S. Krek (ur.), *Proceedings of the 18th Euralex International Congress: Lexicography in Global Contexts* (str. 401–410). Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Fišer, D., Erjavec, T., & Ljubešić, N. (2016). JANES v0.4: korpus slovenskih spletnih uporabniških vsebin. *Računalniško posredovana komunikacija*, *4*(2), 67–99. Dostopno prek https://journals.uni-lj.si/slovenscina2/article/view/7003/6694

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/62/138/2602-1?inline=1

Gorjanc, V. (2005). Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. *Družboslovne razprave*, *21*(48), 197–209.

Gorjanc, V. (2017). *Nije rečnik za seljaka*. Biblioteka XX vek, Beograd.

Hughes, G. (2009). *Political Correctness: A History of Semantics and Culture.* 1st edition. Wiley-Blackwell, MA.

Kalin Golob, M., & Gantar, P. (2015). Stilistika in enojezični slovar: Označevanje jezikovne variantnosti. V V. Gorjanc, P. Gantar, I. Kosem & S. Krek (ur.), *Slovar sodobne slovenščine: Problemi in rešitve* (str. 446–465). Dostopno prek https://ebooks.uni-lj.si/zalozbaul//catalog/download/15/47/530-1?inline=1

Kern, B. (2015). Politična korektnost v slovaropisju. V D. Zuljan Kumar & H. Dobrovoljc (ur.), *Zbornik prispevkov s simpozija 2013* (str. 144–154). Nova Gorica: Založba Univerze.

Kosem, I. (2015). Oznake: Vrednotenjski pomen in pragmatična funkcija v slovarju. Slovarska baza in slovar. V V. Gorjanc, P. Gantar, I. Kosem & S. Krek (ur.), *Slovar sodobne slovenščine: Problemi in rešitve* (str. 482–494). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Kosem, I., Čeh Bálint, J., Gorjanc, V., Kolláth, A., Kovács, A., Krek, S., Novak-Lukanovič, S., & Rudaš, J. (2018a). *Osnutek koncepta novega velikega slovensko-madžarskega slovarja*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani. Dostopno prek https://www.cjvt.si/komass/wp-content/uploads/sites/17/2020/08/Osnutek-koncepta-VSMS-v1-1.pdf

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018b). Kolokacijski slovar sodobne slovenščine. V D. Fišer & A. Pančur (ur.), *Jezikovne tehnologije in digitalna humanistika.* Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Kosem-et-al_Kolokacijski-slovar-sodobne-slovenscine.pdf

Krek, S., Laskowski, C., & Robnik Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. V I. Kosem idr. (ur.), *Proceedings of eLex 2017: Lexicography from Scratch* (str. 93–109), Leiden, Netherlands. Dostopno prek https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf

Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B., & Dobrovoljc, K. (2018). Thesaurus of Modern Slovene 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166

Logar, N., Perger, N., Gorjanc, V., Kalin Golob, M., Kogovšek Šalamon, N., & Kosem, I. (2020). Raba slovarjev v slovenski sodni praksi. *Teorija in praksa*, 57, 89–108. Dostopno prek: https://www.fdv.uni-lj.si/docs/default-source/tip/tip_pos_2020_logar_idr.pdf?sfvrsn=0

Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. V A. Abel idr. (ur.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (str. 85–105), Bolzano/Bozen. Institute for Specialised Communication and Multilingualism. Dostopno prek https://euralex.org/elx_proceedings/Euralex2014/euralex_2014_004_p_85.pdf

Motl, A., & Bajt, V. (2016). *Sovražni govor v Republiki Sloveniji: Pregled stanja.* Mirovni inštitut, Ljubljana. Dostopno prek https://dlib.si/stream/URN:NBN:SI:DOC-F2YZP2RB/c117f4c6-8fe9-437d-8c64-5b7987a856b6/PDF

Petković, B., & Kogovšek Šalamon, N. (2007). *O diskriminaciji: Priročnik za novinarje in novinarke*. Ljubljana: Mirovni inštitut. Dostopno prek https://www.mirovni-institut.si/wp-content/uploads/2014/08/Prirocnik-o-diskriminaciji-final-all.pdf

Rebolj, D. (2008). Uporabnejša opredelitev politične korektnosti. V S. Autor & R. Kuhar (ur.), *Politična (ne)korektnost* (str. 4–15). Ljubljana: Mirovni inštitut. Dostopno prek https://www.mirovni-institut.si/wp-content/uploads/2014/08/nestrpnost-6.pdf

Sakwa, N. (2011). Problems of Usage Labelling in English Lexicography. *Lexikos, 21*, 305–315. Dostopno prek https://lexikos.journals.ac.za/pub/article/view/47

SSKJ. (2014). *Slovar slovenskega knjižnega jezika: Uvod*. Druga, dopolnjena in deloma prenovljena izdaja. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Dostopno prek https://fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

Šorli, M. (2014). Pragmatic Meaning in Lexicographical Description: Semantic Prosody on the Go. V A. Abel, C. Vettori & N. Ralli (ur.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (str. 477–491), Bolzano/Bozen. Institute for Specialised Communication and Multilingualism.

Šorli, M. (2015). Vrednotenjski pomen in pragmatična funkcija v slovarju. V V. Gorjanc, P. Gantar, I. Kosem & S. Krek (ur.), *Slovar sodobne slovenščine: Problemi in rešitve* (str. 466–480). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Vidovič Muha, A. (2013). *Moč in nemoč knjižnega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete.

Negative Vocabulary in the *Thesaurus of Modern Slovene 2.0*

The paper describes an upgraded version of the *Thesaurus of Modern Slovene 1.0*, which is currently the largest open-access collection of Slovene synonyms generated automatically. The creation of the thesaurus has introduced a new type of dictionary, referred to as a responsive dictionary, which allows the data to respond continuously to the opinions of the contributing language community. The upgrade was motivated by the results of a survey of the user community's attitudes towards the *Thesaurus of Modern Slovene*, which revealed a lack of dictionary labels, particularly for non-neutral vocabulary. As a result, the updated version of the thesaurus focuses on developing solutions for identifying and annotating extremely offensive and vulgar vocabulary. To address this, the digital medium is utilized to display information about potentially problematic vocabulary in new ways. The updated version of the thesaurus incorporates a combination of warning icons and longer explanations to provide a clear visual tag as well as an explanation about the potential consequences of word use. The identification of potentially negative words was primarily conducted manually. Synonym sets were exported from the dictionary database, ordered in semantic clusters, and reviewed by students who were provided with brief instructions to identify potentially negative words, such as elements of hate speech (discrimination based on race, ethnicity, gender, sexual orientation, or disability), negative attitudes (related to social status, wealth, behaviour and character, appearance, etc.), and vulgarity (related to taboo topics, e.g., sexuality, bodily excretions, and violence, in the typical informal speech situation). The decisions made by the students were reviewed and modified by a team of linguists, based on corpus data. As responsiveness is a key concept of the thesaurus, involving the user community in future labelling procedures is an important part of the preparation of final labelling solutions.

Keywords: thesaurus, responsive dictionary, dictionary labels, communicative purpose, user community

Grammatical and Pragmatic Aspects of Slovenian Modality in Socially Unacceptable Facebook Comments

Jakob LENARDIČ
Institute of Contemporary History

Kristina PAHOR DE MAITI
Faculty of Arts, University of Ljubljana; Institute of Contemporary History; CY Cergy Paris University

This paper investigates the grammatical and pragmatic uses of epistemic and deontic modal expressions in a corpus of Slovenian socially acceptable and unacceptable Facebook comments. We propose a set of modals that do not interpretatively vary in their modality type in order to enable robust corpus searches and reliable quantification of the results. We show that deontic, but not epistemic, modals are significantly more frequent in socially unacceptable comments, and specifically that they favour violent discourse. We complement the quantitative findings with a qualitative analysis of the discursive roles played by the modals. We explore how pragmatic communicative strategies such as hedging, boosting, and face-saving arise from the underlying syntactic and semantic properties of the modal expressions, such as the modal force and clausal syntax.

Keywords: corpus linguistics, modality, syntax, semantics, pragmatics, hate speech

Lenardič, J., Pahor de Maiti, K.: Grammatical and Pragmatic Aspects of Slovenian Modality in Socially Unacceptable Facebook Comments. Slovenščina 2.0, 11(1): 33–68.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.33-68

1 Introduction

Hate speech and other forms of socially unacceptable discourse have a negative effect on society (Delgado, 2019; Gelber and McNamara, 2016). For instance, calls to action targeting specific demographics on social media have been shown to lead to offline consequences such as real-world violence (Siegel, 2020). Linguistically, socially unacceptable attitudes are often disseminated in a dissimulated form, using pragmatic markers which superficially lessen the strength of intolerant claims or violent calls to action; nevertheless, the discursive markers of such dissimulated discourse are still not well known (Lorenzi-Bailly and Guellouz, 2019), especially outside of English social media.

In this paper, we investigate how Slovenian modal expressions contribute to the dissimulation of unacceptable discourse on social media at the interface between grammar (that is, syntax and semantics) on the one hand and discourse pragmatics on the other. We first perform a quantitative analysis, where we look at how the use of epistemic modals, which convey the speaker's truth commitment, and the use of deontic modals, which convey how the world ought to be according to a set of contextually determined circumstances, differ between unacceptable and acceptable discourse in the case of Slovenian Facebook comments obtained from the *FRENK* corpus (Ljubesić et al., 2021).

What follows is a qualitative discussion of all the observed modals.[1] We first discuss how the meaning of deontic necessity, which corresponds to some kind of obligation that needs to be fulfilled by the agent of the modalized proposition, can have a secondary pragmatic meaning that is akin to face-saving observed with epistemic modals and that arises with syntactically impersonal modals. We then discuss the only deontic likelihood modal, which is the adverb/particle *naj* ("should") co-occurring with a verb in the indicative mood, and discuss its usage

1 This paper extends our previous proceedings paper (Lenardič and Pahor de Maiti, 2022) along three dimensions. First, we make the quantitative analysis in Section 5 more precise by calculating the statistical significance for all the pairwise frequencies rather than just the overall differences. Second, in Section 6, we no longer exclusively discuss necessity modals but also the logically weaker modals – that is, those denoting likelihood or possibility. Third, we take into account an additional modal – that is, *dovoliti* –, which was omitted from the proceedings paper, while also taking into account the possible aspectual variation in verb forms.

in violent discourse from the perspective of its implicit subject-oriented semantics (Stegovec, 2019). Lastly, we discuss the deontic possibility modals denoting permission and their interaction with negation, which yields a lack of permission reading.

We then turn to the epistemic modals. We show that certainty and likelihood epistemics are primarily used to achieve a face-saving effect in pragmatics, and that they often occur in ironic contexts. We discuss how irony affects face-saving and its interaction with the underlying epistemic modality, claiming that what is being communicated with epistemics in such cases is often not truth commitment but the attenuation of controversial or impolite claims. For epistemic possibility modals, we also discuss their so-called concessive role in discourse (Palmer, 2014) in relation to the interaction between face-saving and the underlying epistemic meaning.

The paper is structured as follows. Section 2 presents the semantic and pragmatic properties of epistemic and deontic modals, while Section 3 presents some of the related corpus-linguistic work on modality in socially unacceptable discourse. Section 4 describes the make-up of the *FRENK* corpus in terms of the subtypes of socially unacceptable discourse and the criteria for the selection of the analysed modals. Section 5 presents the quantitative analysis, wherein epistemic and deontic modals are compared between the acceptable and unacceptable supersets in *FRENK*. Section 6 presents the qualitative analysis, where deontic and epistemic necessity modals are discussed in relation to the way grammar interfaces with the pragmatics. Section 7 concludes the paper.

2 Theoretical background

2.1 Semantic assumptions

Modal expressions are sentential operators that interpret a prejacent proposition within the irrealis realm of possibility (Kratzer, 2012). In terms of their lexical semantics, modal expressions are underspecified in the sense that they only encode the so-called modal force, which ranges from possibility via likelihood to necessity (Kratzer, 2012; von Fintel, 2006). The possibility, likelihood or necessity expressed by the modal is interpreted

relative to what goes on in the actual world, but this meaning component is not, in contrast to modal force, lexically encoded by the modal but rather determined by the linguistic or even extra-linguistic context.

In this paper, we look at two such contextually determined interpretations of modal expressions – the epistemic interpretation on the one hand and the deontic interpretation on the other (Coates, 1983; Kratzer, 2012; Palmer, 2014; von Fintel, 2006). Epistemic modals tie the evaluation of the possibility, likelihood, or necessity to the speaker's knowledge about the actual world. For instance, the possibility adverb *morda* in (1), taken from the *FRENK* corpus, has the reading which says that there is a possibility that the referents of the indefinite subject *nekaj jih* ("some of them") will stay in the country. This possibility reading is epistemic as it conveys that the speaker is not sure whether the possibility of their staying will actually turn out to be the case.[2]

(1) [N]ekaj jih bo *morda* ostalo v naših krajih.
"Some of them will *possibly* stay in our country."

By contrast, deontic modals do not tie the evaluation of possibility or necessity to the speaker's knowledge but to some contextually determined authority, such as a set of rules, the law, or even the speaker (Palmer, 2001, 10). An example of a deontic modal is the verb *dovoliti* in example (2), again taken from *FRENK*. This verb also denotes possibility in terms of modal force, so the deontic possibility reading roughly translates to *they should not be given the possibility* (i.e., be allowed) *to change our culture*.

(2) [S]eveda se jim ne sme *dovoliti* da bi spremenil naso *(sic)* kulturo.
"They should not be *allowed* to change our culture."

A single modal can have different readings in terms of modality type. This is, for instance, the case with the necessity modal *morati*, where the epistemic reading in (3a) conveys that the speaker is certain (i.e., epistemic necessity) that whomever they are referring to is a *bona fide* Slovenian. By contrast, the deontic reading in (3b) says that what needs to be necessarily done is preparing for the competition.

2 For ease of readability, the modal under scrutiny is typeset in italics.

(3) a. Ta *mora* biti pravi Slovenec, ni dvoma.
"He *must* be a *bona fide* Slovenian, no doubt about it."
b. Pripraviti se bodo *morali* tudi na konkurenco, ki je zdaj še nimajo.
"They *must* also prepare for the competition which they do not have."

(Roeder and Hansen, 2006, p. 163)

Lastly, note that lexical semantic notions relating to modal force, such as possibility and necessity, should not be conflated with related interpretative notions, such as uncertainty and certainty, which, however, are not lexically entailed. To exemplify, while the adverb *mogoče* and the related predicative adjective *mogoč* are both invariably possibility modals, they need not necessarily always express uncertainty under the epistemic reading. This can be seen when they are paired up with negation, which is interpreted below the modal in the case of the adverb *mogoče*, yielding the uncertainty reading, and above the modal in the case of the adjective *mogoč* (inflected for neuter *-e* because of the subjectless syntax of the matrix clause), yielding the certainty reading, as shown by the paraphrases of (4).[3]

(4) a. *Mogoče* ni bila dovolj socialna.
"It is *possible* that she was not sociable enough."
a. [N]i *mogoče*, da bi bil islam na enaki stopnji kot Zahod.
"It is not *possible* that Islam is on the same level as the West."

The fact that (4a) has a weaker interpretation than (4b) is thus a compositional effect of the different relative scopes of negation, even though at their core both modals still express possibility, as is also indicated by the paraphrases. Being lexically entailed, the force is a stable semantic property of a modal expression, which is why we will refer to a modal such as *mogoče* as a "possibility modal" rather than an "uncertainty modal" (even though it turns out that the adverbial use of *mogoče* always expresses uncertainty and never certainty in *FRENK*) through the rest of the paper.

3 While example (4a) is from *FRENK*, (4b) is from the Slovenian reference corpus *Gigafida 2.0* (Krek et al., 2019), as such negated instances of the adjectival form are not present in *FRENK*.

2.2 Modals pragmatically

Because of the intensional semantics of modals and the contextual variability of the way in which the modal force is interpreted, modals are able to play several (often overlapping) roles in discourse. In functionalist terms, they are important mainly from the perspective of the interpersonal dimension of communication (Halliday, 1970).

Interpersonally, epistemic modals are used in both positive and negative politeness strategies to satisfy the positive face needs of the speakers and the addressee, i.e., the need to be liked by the interlocutor, as well as their negative face needs, i.e., the need to act independently (Brown and Levinson, 1987). Epistemic modals show the following three pragmatic uses (Coates, 1987) related to Brown and Levinson's (1987) Politeness Theory. First, they are used as part of the negative politeness strategy to save the addressee's negative face, when for instance the speaker tries to facilitate open discussion by not assuming the addressee's stance on the conversational issue in advance. Second, epistemic modals can be used as an *addressee-oriented* positive politeness strategy, which involves the preservation of the positive image of the addressee and prevents them from feeling inferior to the speaker. Finally, they are used as part of a *speaker-oriented* positive politeness strategy, which involves the preservation of the positive image of the speaker by enabling the smooth withdrawal from a statement that can be perceived as a boast, threat, or similar.

Related to such politeness strategies, modals fulfil the conversational role of so-called hedging or boosting devices (Hyland, 2005). Epistemic modals function as hedges when the speaker uses them to reduce their commitment to the truth of the propositional content – i.e., to signal their hesitation or uncertainty with regard to what is being expressed, which is a type of face-saving strategy in and of itself (Gonzalvez Garcia, 2000; Hyland, 1998a). In terms of modal force, it is weak epistemic modals denoting possibility that typically correspond to hedges, though certain necessity modals can also acquire such a function in certain contexts, as we will show in the qualitative analysis.

Strong epistemic modals, which express the certainty or high commitment of the speaker to the truth of the utterance, typically function

as boosters and are used by the speaker to convince his or her audience, make his or her utterance argumentatively stronger, close the dialogue for further deliberation (Vukovic, 2014), stress the common knowledge and group membership (Hyland, 2005), and so forth. Such boosters can also be used manipulatively to boost a claim that is otherwise controversial or highly particular (Vukovic, 2014).

Deontic modality also fulfils interpersonal roles in communication. Because deontic modals express notions such as obligation and permission, they have to do with negotiating social power between an authority and the discourse participant to whom the permission is granted or obligation imposed upon (Winter and Gärdenfors, 1995). Deontic statements often involve a power imbalance between interlocutors (which is especially evident in cases when it is not in the interest of the agent to fulfil the obligation), so the use of deontic modals is often paired up with other pragmatic devices denoting politeness or face-saving. Politeness is thus "an overarching pragmalinguistic function that can be overtly or covertly marked in deontic and epistemic modal utterances" (Gonzalvez Garcia, 2000, p. 127).

3 Related work on modality in socially unacceptable discourse

The linguistic and pragmatic characteristics of modality in online socially unacceptable discourse have not yet been extensively explored in the literature. One exception is the work done by Ayuningtias, Purwati, and Retnaningdyah (2021), who analyse YouTube comments related to the 2019 Christchurch mosque shootings in New Zealand. They find that clauses with deontic modals outnumber those with epistemic modals, and that the main discursive strategy of commenters in socially unacceptable comments is to use deontic modals to incite violent action against members of the Muslim community.

Other corpus linguistic studies investigate modal markers from the perspective of stance. Chiluwa (2015), for example, analyses the stance expressed in the tweets of two radical militant groups, Boko Haram and Al Shabaab. Among other stance-related elements, she investigates the use of hedges (including weak epistemic modals) and

boosters (including strong epistemic modals). The results show that boosters are more frequent than hedges, although their overall frequency in the data was low. According to the author, the low frequency of hedges shows that radicalist discourse does not exhibit the tendency to mitigate commitment, which goes hand in hand with the slightly higher presence of boosters that are used as a rhetorical strategy to support (possibly unfounded) statements and to influence, radicalize and win over their readers by projecting assertiveness.

Another study on stance in this context is by Sindoni (2018), who looks at the verbal and multimodal construction of hate speech in British mainstream media. She analyses epistemic modal operators (among other related devices) in order to uncover the writer's stance and attitude towards the content conveyed in the news item. She finds that modality is strategically used to present the author's opinions as facts, while the opinions of others are reported as hypotheses and assumptions.

4 The *FRENK* corpus

4.1 Corpus make-up

For this study, we have used *FRENK*, a 270,000-token corpus of Slovenian Facebook comments of mostly socially unacceptable discourse (Ljubešić et al., 2019). The Facebook comments in the *FRENK* corpus concern two major topics – migrants, generally in the context of the 2015 European migrant crisis, and the LGBTQ community, mostly in the context of their civil rights – and are manually annotated for several different kinds of discourse.[4] The annotations distinguish whether the discourse is aimed towards a target's personal background, such as sexual orientation, race, religion, and ethnicity, or their belonging to a particular group, such as political party. They also distinguish the type of the discourse itself, which falls into four broad categories, one being acceptable discourse and the others different kinds of socially unacceptable discourse (Pahor de Maiti et al., 2019, p. 38):

- Acceptable discourse
- Socially unacceptable discourse

4 The annotations are performed on the comment level while also taking into account the features of the entire discussion thread.

- Offensive discourse, which corresponds to abusive, threatening or defamatory speech that is targeted towards someone on the basis of their background or group participation.
- Violent discourse, which contains threats or calls to physical violence and is often punishable by law (Fišer et al., 2017, p. 49).
- Inappropriate speech, which contains offensive language but is not directed at anyone in particular.

For our study, we have created two subsets of comments: the *acceptable subset* containing comments tagged as *acceptable*, and the *unacceptable subset* containing comments tagged as *offensive, violent* or *inappropriate*. This decision is based on the frequency distributions shown in Table 1.

Table 1: *The make-up of the FRENK corpus in terms of socially (un)acceptable discourse*

Subcorpus	Tokens	
Acceptable	92,922	34%
Offensive	143,948	53%
Inappropriate	1,471	1%
Violent	8,789	3%
Not relevant	24,572	9%
Σ	271,702	100%

The FRENK subcorpora are uneven in terms of size, with the violent and inappropriate sets contain significantly fewer comments than the acceptable and offensive sets. Because violent discourse is generally less frequent than offensive discourse in linguistic corpora,[5] it is difficult to annotate automatically (Evkoski et al., 2022), so one of the crucial features of *FRENK* is the fact that the annotations into discourse type were done manually, employing eight trained annotators per Facebook comment (Ljubešić et al., 2019, p. 9). Note that about 9% of the Facebook comments are marked as *not relevant*, which refers to comments with incorrect topic classification (*ibid., 5*).

5 This is also a result of the EU Code of conduct and terms of service of social media platforms, according to which content deemed illegal due to its hateful character needs to be taken down.

The latest, that is, version 1.1, of the *FRENK* corpus, which also includes texts in Croatian and English, is available for download from the CLARIN.SI repository (Ljubesić et al., 2021). However, the online version, which is accessible through CLARIN.SI's NoSketch Engine concordancer and which we have used for the purposes of this paper,[6] is not yet available to the public.

4.2 The modals analysed in the study

Table 2 shows that there are 13 modal expressions used in the study. We have selected the modals using two criteria.

The first criterion is the modal's tendency towards a single modal reading. As discussed in Section 2.1, modals are in principle ambiguous in terms of the modality type. However, corpus data show that certain modals have an overwhelming preference for a single reading. For instance, while the modal auxiliary *morati* can theoretically have both the epistemic and deontic interpretations (Roeder and Hansen 2006, pp. 162–163), as was shown in (3), the epistemic reading (3a) is actually extremely rare in attested usage, and in the case of the *FRENK* corpus completely non-existent.[7] Similarly, whenever the adverb *naj*$_{\text{IND}}$ is used in the indicative rather than conditional mood (glossed with the subscript IND in Tables 2 and 4 and through the rest of the paper), its meaning is always some shade of the deontic reading (command, wish, etc.). Thus, all the modals in Table 2 are either unambiguously deontic or unambiguously epistemic, so they function as a robust set for testing how deontic and epistemic modality manifests itself in different types of discourse without confounding examples with unintended interpretations.

6 https://www.clarin.si/noske

7 With the exception of *potrebno*, the frequency counts were performed on lemmas, as this is sufficient for distinguishing the part of speech as well. For instance, the lemma *mogoče* corresponds to the adverbial forms, whereas the lemma *mogoč* corresponds to the adjectival ones; however, the adjectival form when used predicatively is consistently ambiguous between the non-epistemic and epistemic interpretations, see Lenardič and Fišer (2021) for discussion and examples. In the case of *potrebno*, we queried the word form, as the lemma *potreben* also yields attributive uses (which are irrelevant because we are focusing on sentential modality), whereas the word form *potrebno* reliably yields the predicative uses.

Table 2: *The analysed modals; AF stands for absolute frequency*

Modal	Syntax	Modality	Force	AF
naj$_{IND}$	Adverb	Deontic	Likelihood	886
morati	Verb	Deontic	Necessity	489
treba	Adjective	Deontic	Necessity	306
smeti	Verb	Deontic	Possibility	150
verjetno	Adverb	Epistemic	Likelihood	123
mogoče	Adverb	Epistemic	Possibility	92
potrebno	Adjective	Deontic	Necessity	65
dovoliti	Verb	Deontic	Possibility	60
morda	Adverb	Epistemic	Possibility	46
najbrž	Adverb	Epistemic	Likelihood	29
ziher	Adverb	Epistemic	Necessity	25
dopustiti	Verb	Deontic	Possibility	19
zagotovo	Adverb	Epistemic	Necessity	16
Σ				2245

The catenative possibility verbs *dovoliti* and *dopustiti* also have the imperfective forms *dovoljevati* and *dopuščati*. As they are rarer than their perfective variants (i.e., 55 instances of *dovoliti* vs. five instances of *dovoljevati* and 11 instances of *dopustiti* vs. eight instances of *dopuščati*), they are counted in Table 2 under the morphologically less complex perfective forms. However, the aspectual distinction does not affect, at least in the *FRENK* data, the modality type, which stays deontic; the imperfective form in (5b) only seems to trigger or emphasize the continuous interpretation of the permission.

(5) a. Sloveniji ne bomo *dovolili* nastanka nasilnih band.
"We won't *allow* violent gangs to form in Slovenia."
b. Problem je v politiki, katera *dovoljuje* islamizacijo nase dezele
"The problem is in politics, which continues to *allow* the Islamization of our land."

The imperfective form of *dopustiti* is said to have an epistemic (or rather, doxastic) interpretation (Močnik, 2019), as shown in the constructed example in (6), which conveys that the speaker is not certain where the referent of the null subject of the embedded clauses is located.

(6) *Dopuščam*, da je notri, in *dopuščam*, da je zunaj.
"He *might* be inside and he *might* be outside."
(Močnik, 2019, p. 422)

However, in *FRENK* all instances of imperfective *dopuščati* convey the deontic interpretation of permission, as shown in the corpus example in (7). As with *dovoliti* and *dovoljevati*, the aspectual distinction does not affect the modal interpretation, which invariably stays deontic.

(7) [S]amo Slovenija *dopušča* izdajalcem i[n] koloborantom, da serjejo po državi
"Only Slovenia *allows* traitors and collaborators to shit all over the country."

The second criterion concerns the fact that some lexemes known to convey modal interpretations also occur frequently with a superficially similar propositional meaning that, however, is not modal. On such case is the adverb *itak*, as in example (8), also taken from *FRENK*.[8]

(8) Krscanstvo pa *itak* izvira iz istih krajev kot islam in juduizem *(sic)*.
"*Of course*, Christianity comes from the same place as Islam and Judaism."

This adverb differs from, for example, the certainty adverb *zagotovo* in that it does not convey the speaker's degree of certainty,[9] but

8 With these two criteria as filters, it does not appear that many modals have been excluded from our study – that is, in Slovenian, ambiguous modals do not outnumber the unambiguous ones, at least not considerably. For instance, most of the modals discussed by Roeder and Hansen (2007) in their overview of Slovenian modal expressions are included in our study (the two exceptions being *lahko*, which is consistently ambiguous between epistemic, ability i.e. "dynamic", and deontic readings, and *utegniti*, which is ambiguous between ability and epistemic readings). Note that the findings presented in this paper are not generalizable to the ambiguous modals like *lahko*; one would first have to manually annotate (a subset of) them to determine how frequent their different readings are, which is beyond the scope of the present paper.

9 *Zagotovo* has the synonym *gotovo*; we have excluded it from our overview because it is too frequently used in the non-modal sense, as in (1), which is mostly typical of non-standard Slovenian.
(1) Postrelit in *gotovo*.
"Shoot them all – that's the end of it."

rather simply intensifies whatever he or she knows to be actually the case (the historical-geographic source of Christianity). Because such non-modal readings are usually as frequent as the modal meaning in attested usage, we have omitted them from our study.

Lastly, note that in terms of syntactic category the modals in Table 2 do not constitute a homogeneous set. While most modals are syntactically adverbs (e.g., *morda, ziher*), some are verbs selecting for finite clausal complements, such as *dovoliti* in (2) and *dopuščati* in (7), verbs selecting for non-finite complements, such as *morati* in (3), and predicative adjectives (of the syntactic frame *It is necessary to*) selecting for non-finite complements, such as *treba* (see the examples in Section 6.1.1). However, such syntactic differences have no bearing on the modal interpretation – in all cases, the modals remain sentential operators that take semantic scope over the proposition denoted by the clause.

5 Quantitative analysis

5.1 The distribution of the modals between acceptable and unacceptable discourse

Tables 3 and 4 show how the Slovenian modals are distributed between the acceptable and unacceptable subsets for the unambiguously epistemic and deontic modals, respectively. The unacceptable subset brings together the three subtypes – offensive, inappropriate, and violent – introduced in Section 4.1. The acceptable and unacceptable sets contain 92,922 and 154,208 tokens, respectively.

In the epistemic set (Table 3), half of the modals – that is, the possibility modal *mogoče* and the necessity modals *ziher* and *zagotovo* – are more frequent in the corpus of unacceptable discourse, while the remaining 3 modals – that is, the possibility modal *morda* and the logically synonymous likelihood modals *najbrž* and *verjetno* – are more frequent in the subset of socially acceptable discourse. Overall, the six epistemic modals are 1.2 times more frequently used in acceptable discourse than they are in unacceptable discourse.

Table 3: *The distribution of epistemic modals in the FRENK corpus; AF stands for absolute frequency and RF for relative frequency, normalized to a million tokens*

	Acceptable		Unacceptable					
Modal	**AF**	**RF**	**AF**	**RF**	**A/U**	**LL**	**P**	**DIN**
verjetno	52	559.6	66	428.0	1.3	2.1	0.1508	13.3
morda	24	258.3	19	123.2	2.1	5.8	0.0156	35.4
mogoče	29	312.1	55	356.7	0.9	0.3	0.5581	−6.7
najbrž	12	129.1	13	84.3	1.5	1.1	0.2898	21.0
zagotovo	3	32.3	13	84.3	0.4	2.7	0.1011	−44.6
ziher	8	86.0	15	97.3	0.9	0.1	0.7791	−6.0
Σ	128	1,377.4	181	1,173.8	1.2	1.9	0.1676	7.9

The distribution is reversed in the set of unambiguously deontic modals (Table 4). As shown in the U/A column, all modals, save for the possibility verb *smeti* ("to allow"), are more characteristic of unacceptable rather than acceptable discourse, with the deontic necessity adjective *treba* and deontic likelihood adverb naj_{IND} ("should") showing the largest preference for the unacceptable set. Overall, the seven deontic modals are 1.3 times more frequently used in socially unacceptable discourse than they are in acceptable discourse.

Table 4: *The distribution of deontic modals in the FRENK corpus*

	Acceptable		Unacceptable					
Modal	**AF**	**RF**	**AF**	**RF**	**U/A**	**LL**	**p**	**DIN**
naj_{IND}	227	2,442.9	583	3,780.6	1.5	33.1	8.6×10^{-9}	−21.5
morati	151	1,625.0	292	1,893.6	1.2	2.4	0.1238	−7.6
	87	936.3	197	1,277.5	1.4	6.0	0.0139	−15.4
smeti	41	441.2	60	389.1	0.9	0.4	0.5364	6.3
potrebno	24	258.28	41	265.9	1.0	0.0	0.9101	−1.5
dovoliti	18	193.7	38	220.5	1.1	0.7	0.3939	−12.0
	4	43.1	15	97.3	2.3	2.4	0.12	−38.6
Σ	552	5,940.5	1226	7,924.5	1.3	33.4	6.5×10^{-9}	−14.5

Statistically, we have tested the differences in pairwise frequencies for all modals, as well as the overall differences between the unacceptable and acceptable sets in both the epistemic (Table 3) and deontic (Table 4) modals. We have used the log-likelihood statistic, which

"establish[es] whether the differences [between pairwise frequencies in two corpora with different sizes] are likely to be due to chance or are statistically significant" (Brezina, 2018, pp. 83–84). The formula for calculating the log likelihood statistic is given in (9), where the observed values $O_{1,2}$ correspond to the absolute frequencies of a modal in the unacceptable and acceptable sets. The loglikelihood value, labelled LL, is given in the antepenultimate column in each table.

(9) $2 \times (O_1 \times \ln(\frac{O_1}{E_1}) + O_2 \times \ln(\frac{O_2}{E_2}))$

In the epistemic set (Table 3), only one out of the six modals shows a statistically robust difference between the acceptable and unacceptable sets at $p < 0.05$ – that is, *morda*, whose LL value is 5.8 and $p = 0.0156$. The overall greater occurrence of epistemic modals in the acceptable set (AF = 128 tokens, RF = 1,377.4 tokens/million) than in the unacceptable set (AF = 181 tokens, RF = 1,173.7 tokens/million) is statistically unreliable; LL = 1.9, $p = 0.1676$.

In the deontic set (Table 4), only two out of the seven modals show a statistically robust difference – the likelihood modal naj_{IND} (LL = 33.1, $p = 8.6\times10^{-9}$) and the necessity modal *treba* (LL = 6.0, $p = 0.0139$). However, in contrast to the epistemic set, the overall greater occurrence of deontic modals in the unacceptable set (AF = 1226 tokens; RF = 7,924.5 tokens/million) than in the acceptable one (AF = 552 tokens; RF = 5,940.5 tokens/million) is statistically significant at the same cut-off point; log likelihood = 33.4, $p = 6.5\times10^{-9}$.

Using the online tool *Calc* (Cvrček, 2021), we have also calculated the Difference Index (DIN) – an effect-size metric – for all the pairwise differences in frequency. The DIN values, which are given in the final columns of Tables 3 and 4, are calculated using Fidler and Cvrček's (2015, p. 230) formula in (10), where RF_1 and RF_2 are the respective relative frequencies of the modals in the acceptable and unacceptable sets.

(10) $\mathbf{DIN} = 100 \times \frac{RF_1 - RF_2}{RF_1 + RF_2}$

The lines in (11) say how the DIN values are to be interpreted (*ibid.*).

(11) a. DIN = –100: the word is present only in corpus 2 (i.e., unacceptable set) but not in corpus 1 (i.e., acceptable set)
b. DIN = 0: the word occurs equally often in corpora 1 and 2 (i.e., in the acceptable and unacceptable sets)
c. DIN = 100: the word is present only in corpus 1 (i.e., acceptable set) but not in corpus 2 (i.e., unacceptable set)

The highest DIN values belong to the modals *morda* (35.4) and *najbrž* (21.0), indicating their greatest preference for the acceptable set, while the lowest belong to *zagotovo* (–44.6), *dopustiti* (–38.6) and naj_{IND} (–21.5), indicating the modals' preference for the unacceptable set.

Note that out of these five modals with high/low DIN values, only *morda* and naj_{IND} also show statistically robust differences according to the log-likelihood test, while *dopustiti*, *zagotovo*, and *najbrž* show statistically unreliable differences. It is likely that the insignificant differences are due to the relatively small size of the *FRENK* corpus, and the attendant fact that all modals save for naj_{IND} have quite low absolute frequencies (e.g., 19 frequencies of occurrence for *dopusititi* in Table 4). We do note, however, that even if we were to disregard naj_{IND} as a possible statistical confounder, the overall difference in the deontic set would remain significant, LL = 6.7, $p = 0.009$.

5.2 Comparison with Previous Work

The findings presented in the previous subsection are related to those in the literature (see Section 3) as follows. Just like in Ayuningtias, Purwati, and Retnaningdyah's (2021) work on socially unacceptable discourse in YouTube comments, our deontic modals outnumber epistemic modals in both the acceptable and unacceptable sets (e.g., 1,226 deontic modals vs. 181 epistemic modals in the unacceptable set). Second, both modals of epistemic necessity in Table 3 – that is, *zagotovo* and *ziher* ("certainly") – differ from most of the weaker modals, like *morda* ("possibly") and *najbrž* ("likely"), in that they are more frequent in unacceptable discourse; this is similar to the finding by Chiluwa (2015), who shows that strong epistemic modals are more frequent than weak ones in the case of tweets by radical militant

groups. However, and in contrast to Chiluwa (2015), our statistically robust finding is not the difference in modal force, but rather the difference in modality type, as discussed above.

6 Qualitative analysis

6.1 Deontic modals in violent discourse

In Section 5, it was shown that deontic modals are more typical of unacceptable discourse than they are of acceptable discourse, a finding that was shown to be statistically robust.

Table 5: *The distribution of deontic modals between the offensive and violent subsets of FRENK; the frequencies are relative and normalized to a million tokens*

Modal	Acceptable	Violent	Offensive
treba	936.3	4,437.4	1,083.7
potrebno	258.3	568.9	243.1
dovoliti	183.0	341.3	213.2
smeti	441.2	682.7	405.7
morati	1,625.0	1,479.1	1,910.4
naj$_{\text{IND}}$	2,442.9	6,371.6	3,647.2
Σ	5,893.7	13,881.0	7,503.3

To look at the pragmatics of deontic modals and their discursive role in relation to socially unacceptable discourse, let's first recall from Section 4.1 that the socially unacceptable discourse in the *FRENK* corpus is further subdivided into several subtypes. Here we focus on two – offensive discourse on the one hand and violent discourse on the other. It turns out that all of the surveyed deontic modals, with the exception of the auxiliary *morati*, are consistently more frequent in violent discourse than in offensive discourse.[10] This is shown in Table 5, where for instance *treba* is almost four times as frequent in the violent-speech subset (RF = 4,437.3 tokens per million) as it is in the offensive subset (RF = 1,083.7 tokens per million).

10 A caveat for comparison, of course, is that the violent subset is much smaller than the acceptable and offensive sets (see Table 1), which is why we report the general trend rather than focus on specific pairwise differences.

6.1.1 Deontic necessity

Let us begin with the observation that *treba* and *morati* are synonymous, possibly completely so, in terms of modal logic, as both entail necessities in terms of modal force and invariably have a deontic reading that has to do with a contextually determined obligation. However, despite the synonymy, *treba* is by far more frequent in violent speech than it is in offensive speech, while *morati* is the only deontic modal that is more prominent in offensive than in violent speech.

The difference in the distribution of the two synonymous modals can be tied to the fact that they vastly differ in their communicative function, which crucially is observable within the same subset. Put plainly, the chief difference is that *treba* occurs in considerably more hateful statements than *morati*, even though the statements all qualify as violent hate speech rather than offensive speech in that some kind of incitement towards violence is expressed in the modalized statement.

For instance, let's first consider some typical examples with *treba* from the violent subset:

(12) a. To golazen *treba* zaplinit, momentalno!!!!
"These vermin *must* be gassed at once!"
b. Pederčine je *treba* peljat nekam in postrelit.
"Faggots *must* be taken somewhere and shot."
c. Ni *treba* par tisoč Voltov, dovolj je 220, da ga strese in opozori, da bo čez par metrov stražar s puško.
"We don't *need* a couple of thousand volts; 220 is enough to electrocute them and warn them that, a couple of metres further on, an armed guard is waiting."

The chief linguistic characteristic of the *treba* examples boils down to lexical choice. The most prominent nominal collocate in the violent subset for the *treba* examples, calculated on the basis of the Mutual Information statistic, is *golazen* "vermin", which can be seen in example (12a), where migrants are referred to as such.[11] According to Assima-

11 As an anonymous reviewer notes, such collocational analysis should be taken tentatively because of how small the violent subset is in *FRENK*. For this reason, the presented collocations might not be a property of violent Slovenian discourse in general, and could be limited specifically to the *FRENK* data.

kopoulos, Baider, and Millar (2017, 41), such metaphoric expressions "are an intrinsic part of the Othering process, and central to identity construction".

In the case of animal metaphors such as MIGRANTS ARE VERMIN, migrants are conceptually construed and stereotyped as an invasive out-group that is maximally different from the in-group to which the speaker considers themselves to belong (*ibid.*). The other most prominent nominal collocate is *elektrika* ("electricity"); metaphors containing this lexeme or lexemes related to electricity (volts, to shock, etc.) often have implied reference, where the undergoers of the verbal event, i.e., migrants, are not directly mentioned, as shown in example (12c). Curiously, when the targets of violent speech are not migrants but members of the LGBT community, instead of metaphors like *golazen*, slurs such as *pedri* ("faggots") are used, as in example (12b).

Notice that in example (12c), *treba* occurs in a negated sentence. Here, negation takes semantic scope over necessity, which means the semantic composition of necessity and negation in relation to the proposition is "it is not *necessarily* the case that *P*" rather than "it is *necessarily* not the case that *P*". Pragmatically, negation in this example is interpreted in a similar manner to the so-called *metalinguistic negation* (Martins, 2020), as the commenter merely objects to the specific number of volts, but still condones the violent action i.e., the electrocution of migrants.

The examples with *morati*, on the other hand, are significantly less lexically charged, as shown in (13), and the statements framed in a more indirect way.

(13) a. Vse Evropske države bi *morale* bolj grobo udarit po migrantih.
"All European countries should *have to* strike back more strictly against migrants."
b. Kdo nas zaščitil a *moramo* mi tud nabavit pištolo.
"Who will protect us? Do we also *have to* buy a gun?"
c. Evropa bi *morala* stopiti skupaj hermeticno zapreti meje.
"Europe should *have to* come together and hermetically close the borders."

Even when the *morati* examples convey that it is necessary that some kind of action be taken against migrants, as in example (13a),

the verbs used are such that they no longer convey explicit violent acts, such as *postreliti* ("to shoot"), *zapliniti* ("to gas"), and *stresti* ("to electrocute") in the *treba* examples (12), but express non-violent acts, as in the case of the verbal phrase *zapreti meje* "close the borders" in (13c). Indeed, the calls to violent action with *morati* are significantly more tentative, as many of the cases of deontic *morati* are embedded under the conditional mood clitic *bi*, which leads to a composite meaning where the deontic necessity is interpreted as a suggestion rather a direct command, as in examples (13a) and (13c), which is also not the case with *treba*.

To sum up the discussion so far, we have observed that while *treba* and *morati* both convey deontic necessity (roughly an obligation that needs to be met), they occur in quite substantially different statements in terms of hateful rhetoric in the case of the same type of unacceptable discourse, i.e., violent speech. Further, *morati* is also the only deontic modal which is less typical of violent speech than it is of offensive speech.

We suggest that the difference is tied to the way the pragmatics of deontic modals interact with their core syntactic and semantic properties. As discussed in Section 2.2, pragmatically deontic modals fulfil the interpersonal function in communication. The interpersonal dimension has to do with the fact that the deontic necessity, i.e., obligation, is ascribed by the speaker to whoever corresponds to the agent of the verbal event in the modalized proposition; concretely, in the case of example (13a), the speaker says that it is European countries that have the obligation to strike back against migrants.

The chief difference between the *treba* (12) and the *morati* (13) examples, manifested in the discussed lexical differences, lies in this interpersonal pragmatic dimension, which is crucially influenced by the syntax of the expressions. *Treba* is an impersonal predicative adjective which, in contrast to *morati*, syntactically precludes the use of a nominative grammatical subject that would be interpreted as the agent in the modalized proposition (Rossi and Zinken, 2016). Consequently, all the statements in the *treba* set of examples are such that the agent has an undefined, arbitrary reference – for instance, it is unclear who is expected to "gas the vermin" in example (12a). What happens

pragmatically is that the subjectless syntax of the adjective *treba* allows the speaker to sidestep the ascription of obligation to a specific agent, thus largely obviating what is perhaps the core interpersonal aspect of deontic modality. This cannot be really avoided with *morati*, which is a personal verb that obligatorily selects for a grammatical subject in active clauses – in other words, because of its personal syntax, *morati* presents a bigger interpersonal burden on the speaker, as he or she needs to specifically name the person or institution that is required to fulfil the obligation.

Note that in the violent subset there is only one example where *morati* is used with the verb *dobiti* ("get"), which induces a passive-like interpretation (14). Here, the grammatical subject headed by *vsak* ("everyone") is interpreted as the target of the violent action rather that the agent. It is telling that this is also the only example with *morati* which is closer in the use of lexically charged items (i.e., being "shot in the head" rather than "the closing of borders" in the previous examples) to the *treba* examples, as this passive-like construction also precludes the use of an agentive noun phrase (unless it is introduced by the Slovenian equivalent of the *by*-phrase, but there are no such examples in the corpus).

(14) [V]sak, ki se približa našim ženskam in otrokom, *mora* dobiti metek v čelo.
"Everyone who gets close to our women and children *must* be shot in the head."

In short, the interpersonal structure influences the degree of hateful rhetoric, in the sense that speakers are more ready to use degrading metaphors, slurs and violent verbal expressions when they can avoid ascribing the obligation to someone specific. We follow Luukka and Markkanen (1997) by suggesting that impersonality has a similar hedging effect to epistemic modals, in the sense that the unexpressed agent in impersonals introduces a degree of semantic vagueness to the proposition, as does uncertainty brought about by the epistemic reading. Thus, with *treba* both deontic imposition and epistemic face-saving meet in one and the same lexeme.

6.1.2 Deontic likelihood

There is only one modal expressing deontic likelihood (i.e., a suggestion) in our set of modals; that is, *naj*$_{\text{IND}}$, which is by far the most frequent word under consideration (see Table 2). In this section, we discuss the possible reasons as to why this modal significantly favours unacceptable discourse (Table 4), and just like *treba* specifically favours the violent subtype of socially unacceptable discourse (Table 5).

Let's start by briefly presenting a lesser-known fact of the Slovenian modal system, which is that *naj*$_{\text{IND}}$ has an additional semantic component to it that is otherwise not present in *morati* or *treba*. This component, which is linguistically unexpressed (and thus hard to detect empirically, especially in corpus data), has to do with the fact that when a speaker/writer uses *naj*$_{\text{IND}}$, they not only convey a suggestion, but also the fact that it is the speakers themselves who are also the source of deontic authority (Stegovec, 2019). To see this, compare the constructed examples in (15), which are minimally different from one another save for the following two facts. On the one hand, in the last example (15c), *morati* is used instead of *naj*$_{\text{IND}}$. On the other hand, the subject of the adversative *ampak* clause is coreferential in (15a) and (15c) with the null subject of the initial clause but not in (15b).

(15) a. #Rekel je, da *naj* grejo stran, ampak noče, da grejo.
I ntended: "He said they *should* go away but he doesn't want them to."
b. Rekel je, da *naj* gredo stran, ampak nočem, da grejo.
"He said they *should* go away but I don't want them to."
c. Rekel je, da *morajo* iti stran, ampak noče, da grejo.
"He said that they *have to* go away but he doesn't want them to."
(Examples (15a) and (15b) from Stegovec, 2019, p. 60)

Because *naj*$_{\text{IND}}$ obligatorily involves speaker control, (15a) expresses a contradiction and is therefore semantically ill-formed (labelled with #). By contrast, (15b) is semantically coherent as the subjects of the two coordinated clauses are not coreferential. Lastly, (15b) also does not express a contradiction even under a coreferential

reading of the subjects of *reči* and *ne hoteti*, which shows that *morati* does not exhibit such subject-oriented control of the deontic suggestion.

For *FRENK*, we propose that the covert presence of speaker control explains an empirical gap in morphosyntactic agreement; that is, the fact that in all examples in the violent discourse *naj*$_{\text{IND}}$ exclusively patterns with a verb in the third person, as in (16). This is not the case for other inflected modals in violent discourse, which do pattern with first person agreement, as shown in (17). Relatedly, the *naj*$_{\text{IND}}$ clauses always contain an unexpressed subject that refers to a vague, arbitrary agent who is either a member of the in-group or part of a depersonalized collective, e.g. arbitrary migrants. In both cases, the speaker/writer, who is the deontic source, is excluded from this vague group that is expected to carry out what the speaker suggests.

(16) a. *Naj* postavljajo snajperiste na mejo
"They *should* put snipers on the border."
b. Kr pelte si jih domov pa vam *naj* posiljujejo žene.
"Take them home. They *should* rape your wives if you do."

(17) a. [A] *moramo* tud mi nabavit pištolo [...]
"Must we buy a gun?"
a. [...] da *dovolimo* tem pedercinam [...]
"... that we *allow* these faggots."

Pragmatically, the covertness of speaker control in *naj*$_{\text{IND}}$ acts as a hedge much like the impersonal syntax acts as a hedge for the obligatory meaning of *treba*, as discussed in the preceding section. This is also the reason why *naj*$_{\text{IND}}$ favours unacceptable discourse, and specifically the violent subtype. Even though *naj*$_{\text{IND}}$ is semantically weaker than either *treba* or *morati* in terms of modal force, this weakness is counterbalanced by the fact that speaker control is uniquely entailed. But because the entailment is not linguistically evident outside of constructed examples like (15a), it is also exempt from the surface discourse, in contrast to a first person verb like *dovolimo* "we allow" in (17b), which overtly spells out the deontic source. Consequently, in examples like (16) the speakers (or rather the authors of the Facebook comments, in this case) use *naj*$_{\text{IND}}$

to express a suggestion that is imposed exclusively on others, with the speakers themselves being the covert deontic source. This also seems to explain the exclusive patterning of *naj*$_{IND}$ with third-person agreement, which semantically excludes the speaker from the denotation of the null subject pronoun.

6.1.3 Deontic possibility

To wrap up our discussion of deontic modals, we finally turn to the three possibility verbs – *smeti*, *dovoliti*, and *dopuščati*.

There is one major difference in this set. The verb *smeti* always combines with negation in the violent subset, as in (18a). This is not the case with *dovoliti* and *dopuščati*, which appears both in negated and non-negated uses, as is exemplified with *dovoliti* in (17b) (preceding section) and (18b).

(18) a. Te horde *ne smemo* spustit v državo.
"We must *not allow* this horde into our country."
b. [P]olitiki *ne* bojo *dovolili*, da se jim zgodi kaj hudega.
"The politicians will *not allow* that anything horrible happens to them".

In (18a) and (18b), negation is interpreted above the modal, which pragmatically yields a reading in which permission is not given to "release this horde into the country" in (18a) and "allow anything to happen to the politicians" in (18b). Notice that negating *smeti* or *dovoliti* results in a stronger reading than in the examples in which *treba* is negated, as in (12c) in Section 6.1.1, where negation-over-necessity results in a lack-of-obligation interpretation.

However, the fact that *smeti* always patterns with negation in the *FRENK* violent data is likely not due to pragmatic considerations, but to lexico-syntactic ones, especially since the interpretation of modality in the examples with negated *dovoliti* and *smeti*, as in (18a) and (18b), is the same. If this is correct, negated *smeti* fills in a gap in the Slovenian lexical inventory, where the modal that is used to grant permission (and like *smeti* and unlike *dovoliti* does not select for a finite clause) – that is, *lahko* – is syntactically a positive-polarity item (Marušič and Žaucer,

2016),[12] which means that its distribution is limited to non-negated contexts, so the closest grammatically related word – that is, *smeti* – is used in negated contexts in its stead. That this is the case is suggested by the fact that the almost exclusive patterning with negation is not limited to the violent subset or even to the larger unacceptable superset, but holds for all occurrences of the modal within the corpus. Only five (3%) out of the 150 occurrences of *smeti* appear in non-negated clauses in the entire corpus, as in (19), whereas *dovoliti* and *dopustiti* are almost equally distributed between negated and non-negated clauses.

(19) [A]li se *smeta* dva istospolno usmerjena uradno poročiti
"Are two people of the same sex *allowed* to get married?"

6.2 Epistemic modals in offensive and acceptable discourse

Epistemic modals are slightly more frequent in acceptable comments, although the difference is not statistically robust, as shown in Section 5. In order to further explore the possible differences and similarities in the use of epistemic modals between different types of comments, we look at their distribution in three subcorpora, namely in acceptable, offensive and violent comments, and the distribution is shown in Table 6. We find that epistemic modals are very infrequent in the violent comments (even unattested for *morda* "possibly" and *najbrž* "likely") in contrast to deontic modals, which are more frequent almost across the board in the violent set (Table 5). On the other hand, the epistemic modals show a similar distribution between acceptable and offensive comments in contrast to violent ones. In fact, the main difference can be observed at the level of modal force. It can be observed that probability modals (*verjetno*, *najbrž*) are the only ones more frequently used in the acceptable comments, while certainty and possibility modals (*ziher*, *zagotovo* and *morda*, *mogoče*) appear more often in offensive comments. In order to investigate the possible reasons for such differences, we now look at the communicative functions that are realized by the investigated modals.

12 The modal *lahko* is exempt from our study because unlike *smeti* its interpretation varies between epistemic and deontic modality in the corpus data.

Table 6: *The distribution of epistemic modals between the acceptable, violent, and offensive subsets of FRENK*

Modal	Acceptable	Violent	Offensive
	258.3	0.0	169.3
mogoče	312.1	113.8	555.8
verjetno	559.6	341.3	451.6
najbrž	129.1	0.0	90.3
ziher	86.0	113.8	97.3
zagotovo	32.3	113.8	83.4
Σ	1,377.4	682.7	1,447.5

Note. The frequencies are relative and normalized to a million tokens.

6.2.1 Epistemic necessity

In offensive comments, the epistemic necessity modals *ziher* and *zagotovo* generally act as prototypical boosting devices, which means that in terms of their illocutionary force (Searle, 1975) they are used to emphasize the commenter's certainty in the assertion, as in example (20a). The meaning of certainty is also often emphasized stylistically, such as by the capitalization of the modal and the string of repeated exclamation marks in (20a).

(20) a. Begunca? Ekonomske migrante pa picke, ki se ne znajo borit za svoj kos zemlje *ZIHER* ne!!!!!!!
"Accepting a refugee? *CERTAINLY* not accepting economic migrants and cunts who don't know how to fight for their piece of land!!!!!!!"

b. Eni bi še radi denar od Slovenije, lahko dobite enosmerno karto za vašo vukojebino. Tam med ovcami se boste *zagotovo* počutili bolj domače
"Some would even like to get money from Slovenia. You can get a one-way ticket back to your shithole. There, you will *certainly* feel more at home among the sheep."

In lexically charged examples like (20a) and (20b), boosting also contributes to the positioning and legitimization of the commenter as an authoritative member of an in-group that is exclusionary of the

migrant out-group. In addition to the stylistically marked typography, such exclusiveness is emphasized by the contemptuous argumentation and explicitly through functional pronominal elements that are involved in the process of Othering by excluding the commenter and their associates, such as *vašo vukojebino* "your shithole" in (20b). Hyland (1998b), building on previous work by He (1993) and Myers (1989), claims that such authoritative use of boosters strategically establishes solidarity between the in-group members by making it seem as though the truth were arrived at by consensus, which makes it less easy to dispute by the members of the out-group.

In the acceptable comments, the two epistemic necessity modals often occur in ironic statements. This is shown by example (21) with the necessity adverb *ziher*, where the ironic interpretation is conveyed not only by the modal itself but also by the use of the intensifying adverb *itak* ("of course"), exaggeration by means of the collective reading of the plural pronoun *vsi* ("everyone"), the use of the verb in the first-person *dejmo* ("let's"), and the use of the emoticon at the end.

(21) Itak, dejmo vsi lagat, to je *ziher* prav :)
"Of course, let's all lie, that's *certainly* the right thing to do :)"

Irony, being a sophisticated rhetorical device in terms of discourse stylistics, enhances the persuasive effect of the proposition (Gibbs and Izett, 2005). It allows the speaker to adopt an attitude of emotional composure or detachment (Attardo, 2000), which helps the commenter to save face by superficially positioning them as an authority and masking the deficits of the argumentation, thus making the comment less disputable and the commenter less exposed to potential criticism. In addition, irony works as a hedge by attenuating the direct criticism conveyed in the comment. This allows the speaker not only to protect their face but also the face of the target, since ironic criticism is accepted better or in a friendlier way than direct critiques (Gibbs and Izett, 2005). Brown and Levinson (1987, p. 212) also note that "even fairly blatant indirectnesses [such as irony] may be defensible as innocent", as it enables the commenter to save face by distancing themselves from the proposition.

Although the boosting function of epistemic necessity modals predominates in the corpus data, there are few examples in which such modals are used as hedging devices even under a non-ironic reading. A case in point is example (22) from the set of offensive comments.

(22) [K]r k cerarju nej gredo *zihr* ma veliko stanovanje ... bedaki.
"They better go to the prime minister Cerar, he *surely* has a big flat ... assholes."

The modal in (22) hedges the propositional content by invoking the presumed shared knowledge of the in-group, which concerns the size of the prime minister's home. Here, the modal works as a face-saving device because it protects the speaker from the accusation of making an unfounded claim, as the modalized statement, despite entailing certainty, is still weaker than the unmodalized variant which would otherwise report that the speaker holds factual knowledge about the size of the prime minister's apartment.

6.2.2 Epistemic likelihood

Likelihood modals predominantly convey the epistemic meaning of low certainty without encoding additional communicative meanings. Such neutral usage is exemplified in (23), where the commenter uses *verjetno* to express that, according to their knowledge, it is likely (but not certain) that the addressee does not have the relevant evidence.

(23) O kredibilnosti posnetkov po katerih sodiš *verjetno* nimaš dokazov.
"You *likely* have no evidence to support the credibility of the recordings on which you base your judgments."

Occasionally, the likelihood modals fulfil additional communicative functions. For instance, the sentence in (24) uses a simile to communicate an offensive comparison to a thorny plant, which is attenuated with the likelihood modal. The negative representation of the target is therefore less categorical, which helps to preserve face both on the side of the commenter and the addressee.

(24) [N]a sliki imaš lep cvetoč travnik[,] ti si pa *verjetno* en blesav osat.
"The picture shows a nice blooming meadow, and you are *probably* one stupid thistle."

As in the case of epistemic necessity modals (see Section 6.2.1), there is frequent use of irony in the acceptable comments. For instance, example (25) has an ironic interpretation, as the commenter is facetiously suggesting that the addressee ask the reporters for concrete evidence on whatever it is that they are reporting on.

(25) Fajn, potem pa jim reci naj dokažejo, da se je to kar pišejo, res zgodilo v Ljubljani, ker *najbrž* imajo dokaze, če so napisali kar pač so.
"Great, then tell them to prove that this what they are writing about really happened in Ljubljana, because they *probably* have evidence if they wrote what they did."

Irony plays two roles here, both contributing to the face-saving dimension of the discourse. On the one hand, the use of irony positions the commenter as an apparent authoritative source of knowledge on the quality of the news media outlet which helps protect the commenter against potential criticism (see also example (21)). On the other hand, the irony, due to its non-literal meaning, allows for defensibility on part of the speaker (Brown and Levinson, 1987). Since the modal is part of a larger ironic context, its primarily communicative function is not to convey knowledge about the proposition; rather, it simply contributes to the overall attenuating effect of the complex, non-literal, communicative strategy used by the commenter.

6.2.3 Epistemic possibility

Epistemic possibility modals are used in offensive discourse to hedge impoliteness, as shown by example (26), which contains the epistemic adverb *morda*.

(26) Vi ste očitni preobremenjeni z nestrpnostjo in nehumanostjo. Jih *morda* vi pedenate?

"You are certainly preoccupied with intolerance and inhumanity. Is it *maybe* you who are looking after them?"

The second sentence in the example contains a question, but its communicative role is not to present an earnest query to the addressee, but rather functions as a rhetorical device. This is emphasized by the lexemes with strong negative connotations (e.g., *nehumanostjo* "inhumanity") and a certainty-denoting adverb (e.g., *očitno* "clearly") in the preceding sentence, which stands in stark contrast to the uncertainty that would be conveyed through questions and epistemic modal in a neutral, non-ironic context. What is happening here pragmatically is that the modal helps save the commenter's face by allowing them to appear less assertive, as the possibility semantics brought about by *morda* lessens the impoliteness of the question by way of (superficial) tentativeness.

A unique characteristic of possibility expressions is that they are also used to convey the so-called concessive interpretation, which is shown in (27) with *mogoče*.

(27) *Mogoče* sem malo staromoden, ampak tistemu, ki nekomu vzame življenje, ga je treba vzeti tudi njemu!
"*Maybe* I'm a bit old-fashioned, but whoever that takes somebody's life should also lose theirs!"

There is a debate in the semantic literature as to whether such concessive uses of epistemic expressions even constitute modality at all, or if they instead play some other discursive role. Palmer (2014, p. 31), for instance, claims that they do not, as "the speaker does not indicate doubt about the proposition, but rather accepts it as true, in order to contrast one state of affairs with another". By contrast, Baranzini and Mari (2019, p. 120) claim that "concessivity is to be understood at the discourse level and not as a meaning of the modal itself", and that the concessive reading arises from the underlying epistemic modal meaning.

In concessive examples with *mogoče* like (27), we believe that modality is still involved in the semantics. What the commenter concedes is not necessarily that a certain state of affairs holds in the actual world; instead, they use the concessive clause to indicate that they allow for

the possibility of their being old-fashioned.[13] From the perspective of the ongoing discourse, the concessive clause does not add anything to the at-issue meaning, but rather only adds to the interlocutors' conversational common ground (Green, 2017). In relation to face-saving, the concessive clause is used by the commenter to rhetorically agree with a possibility tied to the knowledge or belief state of the addressee, thereby establishing communicative rapport between them.

7 Conclusion

This paper has presented a corpus investigation of epistemic and deontic modal expressions in Slovenian Facebook comments in the *FRENK* corpus.

We have first proposed a set of Slovenian modals that show an overwhelming tendency towards a single modal reading. Because of such unambiguity, they constitute a robust set that allows for precise quantitative comparisons between different types of discourse without irrelevant confounding examples and for careful manual analysis of the corpus examples. Quantitatively, we have shown that deontic modals are a prominent feature of unacceptable discourse, and that they are especially prominent in discourse that concerns incitement to violent action, which is legally prosecutable.

In terms of discourse pragmatics, we have first shown that two deontic necessity modals, which are completely synonymous both in terms of force and modality type, nevertheless differ profoundly in the degree of hateful rhetoric in the same type of socially unacceptable discourse. We have shown that what makes a difference in such examples is the presence of impersonal syntax, which offers speakers the ability to linguistically obviate the ascription of the denoted obligation to a particular agent. We have suggested that this sort of face-saving strategy of ambiguity by way of impersonality correlates with the speaker's tendency to use dehumanizing language, such as slurs or degrading metaphors.

13 That the speaker does not concede that something is a fact is shown by the admissibility of the parenthetical clause in the constructed example in (1).
(1) *Mogoče* sem staromoden (kdo bi vedel), ampak …
"*Maybe* I'm old-fashioned (who knows), but …"

For deontic likelihood, we have claimed that the adverb/particle *naj* is well-suited for violent discourse because it implicitly conveys speaker control, and have proposed that this explains why *naj* exclusively occurs in comments in which the speaker expects others to carry out what is being suggested. For possibility modals, we have explored their interaction with negation, and showed how it leads to a stronger reading that involves the denial of permission.

Lastly, we have discussed the fact that epistemic necessity and likelihood modals do not only express truth commitment but also acquire additional communicative functions related to the face-saving strategy. In acceptable comments they often help signal irony, which pragmatically acts as a hedge. In offensive comments, they are used as boosters whereby they help position the speaker as an authoritative figure. Epistemic possibility modals also typically act as face-protecting hedges rather than just expressions of uncertainty, even when they occur in concessive clauses.

Acknowledgments

We would like to thank all three anonymous reviewers for their helpful comments and suggestions. The work described in this paper was funded by the Slovenian Research Agency research programme *P6-0436: Digital Humanities: resources, tools and methods* (2022–2027), the DARIAH-SI research infrastructure, and the national research projects *Z6-4616: Slovenian Verbal Valency: Syntax, Semantics, and Usage* (for author 1) and *N6-0099: LiLaH: Linguistic Landscape of Hate Speech* (for author 2).

References

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). *Online hate speech in the European Union: a discourse-analytic perspective*. Cham: Springer Nature.

Attardo, S. (2000). Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask*, *12*(1), 3–20.

Ayuningtias, D. I., Purwati, O., & Retnaningdyah, P. (2021). The Lexicogrammar of Hate Speech. In Y. Wirza et al. (Eds.), *Proceedings of Thirteenth Conference on Applied Linguistics (CONAPLIN 2020)* (pp. 114–120). Indonesia: Atlantis Press.

Baranzini, L., & Mari, A. (2019). From epistemic modality to concessivity: Alternatives and pragmatic reasoning per absurdum. *Journal of Pragmatics*, *142*, 116–138.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage (Vol. 4)*. Cambridge: Cambridge University Press.

Chiluwa, I. (2015). Radicalist discourse: a study of the stances of Nigeria's Boko Haram and Somalia's Al Shabaab on Twitter. *Journal of Multicultural Discourses*, *10*(2), 214–235.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. London and Canberra: Croom Helm.

Coates, J. (1987). Epistemic modality and spoken discourse. *Transactions of the Philological society*, 85(1): 110–131.

Cvrček, V. (2021). *Calc 1.03: Corpus Calculator*. https://www.korpus.cz/calc/. Last accessed: 20. 12. 2022.

Delgado, R. (2019). *Understanding words that wound*. New York: Routledge.

Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., & Kralj Novak, P. (2022). Retweet communities reveal the main sources of hate speech. *PloS one*, *17*(3), e0265602.

Fidler, M., & Cvrček, V. (2015). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic linguistics*, *23*(2), 197–239.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017): Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In Z. Waseem et al. (Eds.), *Proceedings of the first workshop on abusive language online* (pp. 46–51). Vancouver: Association for Computational Linguistics.

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341.

Gibbs Jr, R. W., & Izett, C. D. (2004). Irony as persuasive communication. In H. L. Colston and A. N. Katz (eds.): *Figurative Language Comprehe*nsion (pp. 143–164). New York: Routledge.

Gonzálvez García, F. (2000). Modulating grammar trough modality: a discourse approach. *ELIA*, *1*, 119–136.

Green, M. (2017). Conversation and common ground. *Philosophical Studies*, *174*(6), 1587–1604.

Halliday, M. A. (1970). Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of language*, *6*(3), 322–361.

He, A. W. (1993). Exploring modality in institutional interactions: Cases from academic counselling encounters. *Text-Interdisciplinary Journal for the Study of Discourse*, *13*(4), 503–528.

Hyland, K. (1998a). *Hedging in scientific research articles*. Amsterdam: John Benjamins Publishing Company.

Hyland, K. (1998b). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, *18*(3), 349–382.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse studies, 7*(2), 173–192.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford: Oxford University Press.

Krek, S. et al. (2019). *Corpus of Written Standard Slovene Gigafida 2.0*. http://hdl.handle.net/11356/1320. Slovenian language resource repository CLARIN.SI.

Lenardič, J., & Fišer, D. (2021). Hedging modal adverbs in Slovenian academic discourse. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *9*(1), 145–180.

Lenardič, J., & Pahor de Maiti, K. (2022). Slovenian Epistemic and Deontic Modals in Socially Unacceptable Discourse Online. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities* (pp. 108–116). Ljubljana: Institute of Contemporary History.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In K. Ekštein (Ed.), *International conference Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, v. 11697* (pp. 103–114). Cham: Springer.

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.1.* http://hdl.handle.net/11356/1462. Slovenian language resource repository CLARIN.SI.

Lorenzi-Bailly, N., & Guellouz, M. (2019). Homophobie et discours de haine dissimulée sur Twitter: celui qui voulait une poupée pour Noël. *Semen. Revue de sémio-linguistique des textes et discours*, *47*. doi: 10.4000/semen.12344

Luukka, M. R., & Markkanen, R. (1997). Impersonalization as a form of hedging. In R. Markkanen and H. Schröder (Eds.), *Hedging and Discourse, Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts* (pp. 168–187). Berlin: de Gruyter.

Martins, A. (2020). Metalinguistic negation. In V. Déprez & M. T. Espinal (Eds.), *The Oxford Handbook of Negation*. Oxford: Oxford University Press.

Marušič, F. L., & Žaucer, R. (2016). The modal cycle vs. negation in Slovenian. In F. L. Marušič & R. Žaucer: *Formal Studies in Slovenian Syntax* (pp. 167–192). Amsterdam: John Benjamins Publishing Company.

Močnik, M. (2019). Slovenian 'dopuščati' and the semantics of epistemic modals. In *Proceedings of 27th Formal Approaches to Slavic Linguistics meeting (FASL 27)*. Michigan: Michigan Slavic Publications.

Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied linguistics*, *10*(1), 1–35.

Pahor de Maiti, K., Fišer, D., & Ljubešić, N. (2019). How haters write: analysis of nonstandard language in online hate speech. In J. Longhi and C. Marinica: *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)*. Paris: CLARIN K-Center for CMC.

Palmer, F. R. (2001). *Mood and modality*. Cambridge: Cambridge University Press.

Palmer, F. R. (2014). *Modality and the English modals*. Oxon: Routledge.

Roeder, C. F., & Hansen, B. (2006). Modals in contemporary Slovene. *Wiener Slavistisches Jahrbuch*, *52*, 153–170.

Rossi, G., & Zinken, J. (2016). Grammar and social agency: The pragmatics of impersonal deontic statements. *Language, 92*(4), e296–e325.

Searle, J. R. (1975). A taxonomy of illocutionary acts. *Language, mind, and knowledge*, *7*, 344–369.

Siegel, A. A. (2020). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform* (pp. 56–88). Cambridge: Cambridge University Press.

Sindoni, M. G. (2018). Direct hate speech vs. indirect fear speech. A multimodal critical discourse analysis of the Sun's editorial "1 in 5 Brit Muslims' sympathy for jihadis". *Lingue e Linguaggi*, *28*, 267–292.

Stegovec, A. (2019). Perspectival control and obviation in directive clauses. *Natural Language Semantics*, 27(1), 47–94.

von Fintel, K. (2006). Modality and Language. In D. M. Borchert (Ed.), *Encyclopedia of Philosophy – Second Edition* (pp. 20–27). Detroit: MacMillan Reference USA.

Vukovic, M. (2014). Strong epistemic modality in parliamentary discourse. *Open Linguistics*, *1*(1): 37–52.

Winter, S., & Gärdenfors, P. (1995). Linguistic modality as expressions of social power. *Nordic Journal of Linguistics*, *18*(2), 137–165.

Slovnični in pragmatični vidiki naklonskosti v slovenščini v družbeno nesprejemljivih komentarjih na Facebooku

V članku predstavimo slovnično in pragmatično rabo epistemskih in deontskih naklonskih izrazov v korpusu družbeno sprejemljivih in nesprejemljivih komentarjev v slovenščini, ki so bili objavljeni na platformi Facebook. Za potrebe analize oblikujemo seznam naklonskih izrazov, ki pomensko pripadajo zgolj eni vrsti naklonskosti, kar nam omogoča učinkovite in točne korpusne poizvedbe in zanesljivo interpretacijo kvantitativnih rezultatov. V članku pokažemo, da so deontski naklonski izrazi, ne pa tudi epistemski, statistično značilno bolj pogosti v družbeno nesprejemljivih komentarjih, pri čemer še posebej izstopajo v komentarjih z nasilno vsebino. Kvantitativne izsledke nadgradimo s kvalitativno analizo diskurzivne vloge naklonskih izrazov. V kvalitativnem delu tako raziščemo, kako se pragmatične sporazumevalne strategije, med njimi pragmatično omejevanje in ojačevanje pomena propozicije ter blaženje potencialne grožnje posameznikovi integriteti, sklapljajo s temeljnimi skladenjskimi in pomenoslovnimi značilnostmi naklonskih izrazov, na primer njihovo naklonsko stopnjo in stavčno skladnjo.

Ključne besede: korpusno jezikoslovje, naklonskost, skladnja, pomenoslovje, pragmatika, sovražni govor

Referencing the Public by Populist and Non-populist Parties in the Slovene Parliament

Darja FIŠER
Institute of Contemporary History

Tjaša KONOVŠEK
Institute of Contemporary History

Andrej PANČUR
Institute of Contemporary History

The present moment raises many questions about the workings and resilience of parliamentary democracy in Western-type democracies, including the former socialist states of the East Central European region, where various forms of populism and illiberal democracy are taking shape. Among these, Slovenia is taken as a case study, since it is not only a former socialist state, but was also for a long time acknowledged as a post-socialist success story. Focusing on the central state institution in systems of parliamentary democracy, i.e. the parliament, and its members (MPs) this paper considers speech as performed during parliamentary sessions by MPs from populist and non-populist political parties between the years 1992 and 2018, the period of a fully democratic Slovene national parliament. It combines the methodological approaches of cultural history with corpus linguistics in order to map any possible differences in populist and non-populist discourse of MPs. Special attention is given to situations where MPs mentioned the public, thus testing the hypothesis that populist MPs engage more with the public as a part of their populist political style.

Keywords: political parties, populism, life-world, parliament, Slovenia

Fišer, D., Konovšek, T., Pančur, A.: Referencing the Public by Populist and Non-populist Parties in the Slovene Parliament. Slovenščina 2.0, 11(1): 69–90.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.69-90

1 Introduction

In the last two decades, the political scene in many democratic countries in Europe as well as around the globe has witnessed an increase in active populist political parties and a rise in their popularity among voters. Interpretations of this phenomenon vary from populism and illiberal democracy as serious threats to parliamentary democracy, to those who see it as a transitory phase of the otherwise firm rule of democracy.[1] Parallel to the spread of populism, many research fields have started to dedicate some of their attention to the phenomenon itself: its origins, developments, varieties, meanings and possible consequences. They attempted (and are still attempting) to map out populism in a variety of the spaces in which it appears, from media landscapes to popular and policy responses. Political science and sociological analyses pay special attention many of these, mainly through a lens of discursive practices of the most visible members of populist political parties.[2] By doing so, recent research has noted a clear difference between the discourses of members of populist and non-populist parties, especially when using social and other media.

However, much less is known about the relationship between populist and non-populist discourse in the speeches of members of parliament (MPs) in political systems of parliamentary democracy, in which parliaments are the central representative, legislative, and controlling state institutions. One of the most common interpretations of populism, especially combined with illiberal democracy in the area of East Central Europe, is the idea of the unfinished transition from state socialism to parliamentary democracy and market economy.[3] Inefficient, incomplete breaks with the past systems, the socialist mentality of the population and the corruption of political and economic elites stretching from the time of socialism to the present, are often

1 As one of the most resounding discussions on this topic, see: Ivan Krastev and Stephen Holmes, *The Light That Failed: A Reckoning*. London: Penguin Books, 2019.

2 Emanuela Fabijan in Marko Ribać, "Politični in medijski populizem v televizijskem političnem intervjuju," *Social Science Forum*, *Vol. 37*, Nr. 98 (2021), pp. 43–68.

3 Joachim von Puttkamer, Włodzimierz Borodziej and Stanislav Holubec (eds.), *From Revolution to Uncertainty: The Year 1990 in Central and Eastern Europe*. London: New York: Routledge, 2019. Rudi Rizman, *Uncertain Path. Democratic transition and consolidation in Slovenia*. College Station: Texas A&M University Press, 2006.

used as an explanations for current deviations from parliamentary democracy.

However, a great deal of research stresses the lack of empirical evidence to support such claims and has long been criticizing the general view underlining such an interpretation,[4] and this paper aims at contributing to this literature. The analysis conducted here is localized to the case of Slovenia as one of the former socialist states that has not (thus far) completely submitted to populism and illiberal democracy, yet its political and media space clearly exhibits some populist tendencies. To bring further nuances and critical understanding to the existence of modern-day populism, this paper is embedded around two key concepts.

This paper is an extended version of the conference paper of a conference contribution.[5] Here, we widened the methodological framework and strengthened the interdisciplinary nature of the analysis, embedding digital humanities deeper into historical interpretation. To achieve this, we first used the concept of the life-world to acknowledge the existence of a specific reality of MPs in which their speech, as analysed in this paper, is made. Second, we draw on the existing typology of populist and non-populist parties created by political scientists and sociologists to see how MPs from two different groups of political parties, i.e. populist and non-populist, construct their view of the public, thus taking into account the existing indications of populists' and populisms' unique connection to a public it perceives as its own.[6] The goal of this analysis is to detect any differences between populist and non-populist discourse, as observed through the lens of references to the general public.

2 Approach and methodology

To further investigate the connection between the speech of MPs, their image of the public, and their populist or non-populist origin, we

4 Valeria Bunce, "Should transitologists be grounded?", *Slavic Review, Vol. 54*, Nr. 1 (1995), pp. 126—127. Thomas Carothers, "The end of transition paradigm", *Journal of Democracy, Vol. 13*, Nr. 1 (2002), pp. 17—20.

5 Darja Fišer, Tjaša Konovšek in Andrej Pančur, "Referencing the Public by Populist and Non-Populist Parties in the Slovene Parliament," Darja Fišer and Tomaš Erjavec (eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities*. September 15 – 16 2022. Ljubljana, Slovenia. Ljubljana: Inštitut na novejšo zgodovino, 2022, pp. 243–247. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf (January 6 2022).

6 See further text below for references.

combine the methodological framework of cultural history of parliamentarianism with corpus linguistics.

From a historical perspective, we draw on recent developments in political history, focusing on the cultural side of the history of parliamentarism. This includes topics such as: the ideal image and workings of parliaments as an institution in the modern period as proposed by prominent scholars, thinkers and writers, the image of parliaments as architectural settings and as communicated by the media, parliaments as a space of specific communication, and – last but not least – parliament(arianism) as a concept in its own right.[7] In this paper we build on the concept of life-world (or *Lebenswelt*), which has to a small extent already been used in relation to parliamentarism – or, better, the people who shape it.[8]

The concept of life-world originated in philosophy, where it stressed that subjective experience of parliamentarism was identical with the reality of parliamentarism as such. In this respect, the life-world of MPs consists of their own experiences but also to a significant degree of how parliamentarism is seen by others.[9] The concept of life-world has been used in historiography to emphasize the circumstances in which parliamentarianism is experienced, focusing on MPs as historical actors.[10] In this case, this approach brings to the fore research questions about MPs' perceptions, education, and expectations; their political socialization, prior experiences, and everyday life; and the influence of collective opinions, public images, and the media on their work. In this paper, we focus on one of the aspects of MPs' life-world as used in historiography, namely the MPs' relationship with the public, through

7 Remieg Aerts, *The Ideal of Parliament in Europe Since 1800*. Cham: Palgrave Macmillan, 2019. Jure Gašparič, *Državni zbor 1992-2012: o slovenskem parlamentarizmu*. Ljubljana: Inštitut za novejšo zgodovino, 2012. Andreas Schulz and Andreas Wirsching (eds.), *Parlamentarische Kulturen in Europa. Das Parlament als Kommunikationsraum*. Düsseldorf: Droste Verlag, 2012. Pasi Ihalainen, Cornelia Ilie, and Kari Palonen (eds.), *Parliament and Parliamentarism. A Comparative History of a European Concept*. Berghahn, 2016.

8 Adéla Gjuričová and Tomáš Zahradníček, *Návrat parlamentu. Česi a Slováci ve Federálním shromáždění*. Praha: Argo, 2018.

9 Edmund Husserl, *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: eine Einleitung und die phänomenologische Philosophie*. Hamburg: Meiner, 1996 (1962). Jürgen Habermas, *The Theory of Communicative Action, Vol. 2, Lifeworld and system: a critique of functionalist reason.* Cambridge: Polity Press, 2007.

10 Adéla Gjuričová, Andreas Schulz, Luboš Velek and Andreas Wirsching (eds.). *Lebenswelten von Abgeordneten in Europa 1860–1990*. Düsseldorf: Droste Verlag, 2014.

the words they choose to refer to them. This, in turn, reveals a part of their self-understanding, including their intentionally or unintentionally expressed relation to populism.

Within the framework of life-world we further distinguish between Slovenian populist and non-populist parties on two axes. First, based on the profile of political parties, we draw on existing research for the criteria to determine which Slovenian political parties qualify as populist. Second, on the temporal axis, we acknowledge the profound political changes of 1990 (Slovene independence and the democratization and pluralization of political space), when the MPs of the Slovene assembly were freely elected for the first time; and the political changes of 2004 (when Slovenia joined the EU and NATO) as a year that witnessed the active beginnings of modern populism in the Slovene political space. We take into account the difference between modern populist parties, as they emerged in the last decade and a half, and their immediate precursors, which have existed since the early 1990s. Therefore, the analysis treats the Slovenian Democratic Party (SDS) and its predecessor, the Social Democratic Party of Slovenia (SDSS), along with the New Slovenia (NSi) and the Slovenian National Party (Slovenska nacionalna stranka, SNS), as populist parties, while all others were classified as non-populist.

We followed three basic and pre-established criteria in determining which political party may be considered populist in a present-day sense. First, a populist political party and its members must address the public as an exclusive group and reinforce their identity by emphasizing an external threat. Second, such politicians or political parties claim to be the only true representatives of the public, the best representatives of their interests, non-corrupt, and their guardians from the (real or imaginary) external threat. Third, populists actively undermine the central state institutions, the rule of law and the wider democratic mechanisms, such as the media.[11] In a historical sense, some uncer-

11 Danica Fink Hafner, *Populizem*. Ljubljana: Založba FDV, 2019. Ana Frank in Iztok Šori, "Normalizacija rasizma z jezikom demokracije: primer Slovenske demokratske stranke," *Časopis za kritiko znanosti, Vol. 43*, Nr. 260 (2015), pp. 89–103. Giovanna Campani in Mojca Pajnik, "Populism in historical perspectives". In Gabriella Lazaridis and Giovanna Campani (eds.), *Understanding the Populist Shift: Othering in a Europe in Crisis*. London: New York: Routledge, 2017, pp. 13–30.
Iztok Šori, "Za narodov blagor: skrajno desni populizem v diskurzu stranke Nova Slovenija," *Časopis za kritiko znanosti, Vol. 43*, Nr. 260 (2015), pp. 104–117. Jurij Hadalin, "Straight

tainty arose when deciding which predecessors of later or currently existing political parties to include in the analysis under populist. While there is no general criteria on how to measure the (dis)continuity of political parties through longer periods of time, we chose to take into consideration only those pairs of political parties and their predecessors that have a very strong continuity in their leadership and prominent tendencies towards populism or populist style of activity as described by the three criteria above. For example, the NSi party that has existed since 2000 on the one hand maintains strong continuity towards the Slovene Cristian Democrats (SKD), Slovene People's Party (SLS), and later towards the united Slovene People's Party and Slovene Christian Democrats (SLS+SKD); but on the other a much weaker one both in the sense of policymaking, political leadership, and public perception.

From a historiographical point of view, there is one more issue that needs to be addressed, namely the question of how to include breaks in historical development – usually clearly visible and often the centre of attention in qualitative analysis – into quantitative analysis in an interdisciplinary environment. In order to take advantage of a large dataset available for this analysis, some breaks in political history (such as the year 2004) are only indicated as points of change and not as full endpoints of a period. While searching for additional context for interpretation of the results of the present analysis in wider domestic as well as international political developments would certainly add value, we decided to maintain our focus on the Slovene parliament. Each parliamentary lifespan has its own specific periodization, stretching from one election to other, from one coalition formation to the next, and on the smaller scale from session to session. This remains the basic time frame of the present analysis which, at the same time, allowed us to make only limited assumptions about the possible outside influences on the Slovene parliamentary life before the analysis was carried out.

Talk. The Slovenian National Party's Programme Orientations and Activities," *Contributions to Contemporary History, Vol. 60*, Nr. 2 (2020), https://doi.org/10.51663/pnz.60.2.10. Jurij Hadalin, "What Would Henrik Tuma Say? From The Social Democratic Party of Slovenia to the Slovenian Democratic Party," *Contributions to Contemporary History* Vol. 61. Nr. 3 (2021), https://doi.org/10.51663/pnz.61.3.10. Marko Lovec (ed.), *Populism and attitudes towards the EU in Central Europe*. Ljubljana: Faculty of Social Sciences, 2019. Mojca Pajnik, "Media Populism on the Example of Right-Wing Political Parties' Communication in Slovenia," *Problems of Post-Communism, Vol. 66*, Nr. 1 (2019), pp. 21–32.

3 Analysing MPs' speech

Using corpus linguistics has, in this case, proven vital for managing an enormous set of data, i.e. the minutes of the parliamentary sessions of the Slovene assembly (formally named as the parliament in the 1991 Constitution) that were collected and made available for use through the CLARIN repository.[12] The analysis is based on the *Slovenian parliamentary corpus (1990-2018) siParl 2.0*, which contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative periods 1992-2018, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the 2nd to the 7th legislative periods 1996-2018, and minutes of the Council of the President of the National Assembly from the 2nd to the 7th legislative periods 1996-2018. The corpus comprises over 10,000 sessions, one million speeches or 200 million words.[13]

In our analysis we take into account the time span from 1992 when the first term of the Slovenian parliament started until 2018 when the seventh term ended. The time frame thus includes some important events that affected the development of Slovenian political parties, their governing style and, by extension, the actions of MPs, such as Slovenia's accession to the European Union in 2004,[14] the global financial crisis in 2007 and 2008, and the migrant crisis in 2015.[15] Using the typology advocated by sociologists and political scientists (see Section 2), we created subcorpora of populist and non-populist political parties for each parliamentary term, resulting in a total of 14 subcorpora. The subcorpora ranged between just under a million tokens in Term1 to 12 million tokens in Term7 for populist parties, and between 7 million tokens in Term1 to just under 15 million tokens in Term7 for non-populist parties.

12 *CLARIN Slovenia. Common Language Resources and Technology Infrastructure.* http://www.clarin.si/info/about/ (December 28, 2022).

13 Andrej Pančur, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn and Neja Blaj Hribar, *Slovenian parliamentary corpus (1990-2018) siParl 2.0*, Slovenian language resource repository CLARIN.SI (2020), ISSN 2820-4042, http://hdl.handle.net/11356/1300.

14 Gašparič, *Državni zbor*, pp. 108–151.

15 Benjamin Moffitt, "How to Perform Crisis: A Model for Understanding the Key Role of Crisis in Contemporary Populism," *Government and Opposition, Vol. 50*, Nr. 2 (2015), pp. 189–217.

The next step of the analysis presented a challenge, as there are no pre-existing wordlists of references to the general public that we could rely on. We therefore generated frequency lists of nouns for each subcorpus and manually selected those that refer to the public in the broadest sense (e.g. *person*, *citizen*, *inhabitant*) from the 1,000 most frequent nouns in each subcorpus. We only took into account the nouns that can only refer to people (groups or individuals), disregarding those that can also be used for institutions (e.g. *association*) or objects (e.g. *school*). We also checked their usage via a concordance search and discarded the expressions that could potentially be used for the general public but in this specific corpus predominantly refer to the MPs, the government or their staff (e.g. *proposer*).

As can be seen in Table 1, this yielded a total of 86 unique nouns with the total absolute frequency of 359,320 and relative frequency of 7,322.53 for the populist parties, and the total absolute frequency of 524,195 and relative frequency of 6,788.74 for their non-populist counterparts. Most (69) of the nouns are shared between both party groups (coloured in white, e.g. *victim, neighbour, human, Roma, patient*), in addition to 10 that are unique for the populist MPs (coloured in pink, e.g. *Croat, wife, Austrian*) and seven that are specific to non-populist MPs (coloured in blue, e.g. *stakeholder, recipient, tenant*).

Table 1: *List of specific and joint public-related words identified in the subcorpora of populist and non-populist speeches with their absolute and relative frequencies as well as the usage ratio*

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
P-ONLY	Hrvat *Croat*	1,341	**27.33**	0	0.00	/
	žena *woman*	397	8.09	0	0.00	/
	Avstrijec *Austrian*	318	6.48	0	0.00	/
	Diplomant *graduate*	300	6.11	0	0.00	/

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
P-ONLY	storilec *perpetrator*	232	4.73	0	0.00	/
	volivec *voter*	161	3.28	0	0.00	/
	delojemalec *employee*	36	0.73	0	0.00	/
	Neslovenec *Non-Slovenian*	31	0.63	0	0.00	/
	svojec *family member*	27	0.55	0	0.00	/
	delavka *worker (female)*	0	0.00	0	0.00	/
N-ONLY	deležnik *stakeholder*	0	0.00	1,784	**23.10**	/
	prejemnik *recepient*	0	0.00	1,191	15.42	/
	najemnik *tenant*	0	0.00	983	12.73	/
	dolžnik *debtor*	0	0.00	752	9.74	/
	vajenec *apprentice*	0	0.00	444	5.75	/
	kadilec *smoker*	0	0.00	290	3.76	/
	krajan *townsman*	0	0.00	172	2.23	/
JOINT	oče *father*	929	18.93	329	4.26	**4.44**
	obrtnik *craftsman*	1,187	24.19	540	6.99	**3.46**
	davkoplačevalec taxpayer	4,762	97.04	2,178	28.21	**3.44**
	migrant *migrant*	2,627	53.54	1,255	16.25	**3.29**
	vlagatelj *investor*	426	8.68	260	3.37	2.58
	podjetnik *enterpreneur*	3,880	79.07	2,671	34.59	2.29

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
JOINT	moški *man*	827	16.85	619	8.02	2.10
	ljudstvo *people*	3,089	62.95	2,376	30.77	2.05
	Italijan *Italian*	272	5.54	216	2.80	1.98
	Slovenka *Slovenian (female)*	1,432	29.18	1,143	14.80	1.97
	pacient *patient*	1,619	32.99	1,452	18.80	1.75
	zamejstvo *autochtonous Slovenian region*	1,067	21.74	966	12.51	1.74
	kmet *farmer*	6,839	139.37	6,739	87.28	1.60
	prijatelj *friend*	1,024	20.87	1,012	13.11	1.59
	naročnik *subscriber*	517	10.54	516	6.68	1.58
	Slovenec *Slovenian*	10,103	205.89	11,090	143.62	1.43
	dijak *student*	2,403	48.97	2,670	34.58	1.42
	kupec *buyer*	1,216	24.78	1,357	17.57	1.41
	državljan *citizen*	21,570	439.57	24,828	321.54	1.37
	priča *witness*	4,061	82.76	4,701	60.88	1.36
	državljanka *citizen (female)*	6,902	140.65	8,372	108.42	1.30
	narod *nation*	4,952	100.92	6,035	78.16	1.29
	žrtev *victim*	3,945	80.39	4,810	62.29	1.29
	sosed *neighbour*	738	15.04	928	12.02	1.25

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
JOINT	človek *human*	68,517	**1,396.30**	86,824	**1,124.44**	1.24
	Rom *Roma*	627	12.78	808	10.46	1.22
	bolnik *patient*	1,279	26.06	1,717	22.24	1.17
	prosilec *applicant*	343	6.99	468	6.06	1.15
	javnost *public*	16,248	331.12	22,367	289.67	1.14
	starš *parent*	5,732	116.81	7,893	102.22	1.14
	oseba *person*	16,836	343.10	23,762	307.74	1.11
	subjekt *subject*	3,406	69.41	4,866	63.02	1.10
	družina *family*	11,120	226.61	16,298	211.07	1.07
	otrok *child*	18,205	371.00	26,762	346.59	1.07
	gost *guest*	966	19.69	1,438	18.62	1.06
	begunec *refugee*	1,247	25.41	1,879	24.33	1.04
	mladina *youth*	1,384	28.20	2,101	27.21	1.04
	delničar *shareholder*	444	9.05	684	8.86	1.02
	tujec *foreigner*	3,169	64.58	4,908	63.56	1.02
	zavarovanec *insurance holder*	896	18.26	1,394	18.05	**1.01**
	volivec *voter*	3,478	70.88	5,544	71.80	**0.99**
	lastnik *owner*	8,031	163.66	12,814	165.95	**0.99**
	mati *mother*	320	6.52	512	6.63	0.98

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
JOINT	družba *society*	23,431	477.50	38,532	499.02	0.96
	študent *student*	4,973	101.34	8,202	106.22	0.95
	posameznik *individual*	7,367	150.13	12,307	159.39	0.94
	zavezanec *person liable*	2,437	49.66	4,096	53.05	0.94
	uporabnik *user*	3,441	70.12	5,866	75.97	0.92
	nosilec *holder*	2,211	45.06	3,812	49.37	0.91
	občan *resident*	1,558	31.75	2,688	34.81	0.91
	prebivalec *inhabitant*	5,318	108.37	9,404	121.79	0.89
	partner *partner*	4,580	93.34	8,312	107.65	0.87
	potrošnik *consumer*	1,657	33.77	3,060	39.63	0.85
	generacija *generation*	2,279	46.44	4,215	54.59	0.85
	delavec *worker*	10,768	219.44	20,055	259.73	0.84
	invalid *disabled person*	3,032	61.79	5,760	74.60	0.83
	prebivalstvo *population*	2,727	55.57	5,452	70.61	0.79
	manjšina *minority*	2,742	55.88	5,518	71.46	0.78
	učenec *pupil*	1,437	29.28	3,071	39.77	0.74
	ženska *female*	2,941	59.93	6,517	84.40	0.71
	upokojenec *retiree*	3,547	72.28	8,097	104.86	0.69
	skupnost *community*	16,208	330.30	38,163	494.24	0.67

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	**49,070,504**		**77,215,381**		
	#lemmas	**76**		**74**		
	LEMMA	**AF**	**RF**	**AF**	**RF**	**P:N ratio**
JOINT	pripadnik *member*	1,375	28.02	3,238	41.93	0.67
	upravičenec *beneficiary*	1,673	34.09	4,523	58.58	0.58
	upnik *creditor*	566	11.53	1,725	22.34	0.52
	podpisnik *signatory*	465	9.48	1,460	18.91	0.50
	udeleženec *participant*	500	10.19	1,685	21.82	**0.47**
	porabnik *consumer*	129	2.63	540	6.99	**0.38**
	populacija *population*	480	9.78	2,179	28.22	**0.35**
	Total	359,320	7,322.53	524,195	6,788.74	1.08

The list of populist-specific nouns contains words describing people according to their ethnic background (e.g. *Austrian, non-Slovenian*), family role (e.g. *relative, wife*) and employment status (e.g. *female worker, employee*). Non-populist-specific nouns contain expressions which describe the role or status of a person in an administrative or legal procedure (e.g. *stakeholder, recepient*), business transaction (e.g. *tenant, debtor*), origin (e.g. *local*), education (e.g. *apprentice*) or health status (e.g. *smoker*). Among the joint nouns, *father*, *craftsman*, *taxpayer* and *migrant* are used three times more frequently by populist MPs, whereas *beneficiary*, *participant*, *consumer* and *population* are used more than twice as frequently by non-populist MPs. *Insurance holder*, *voter* and *owner* are used nearly identically by both groups of MPs. This might reflect a difference between the populist and non-populist parties and their focus in their political base: while the first usually rally voters from rural areas, the latter are traditionally more successful in urban ones.

Table 2: *Absolute and relative frequencies of public-related words as used by populist and non-populist MPs per parliamentary term and with the statistical significance tests*

	T1	T2	T3	T4	T5	T6	T7	Total
Populist #tokens	950,851	4,917,224	7,291,606	8,607,268	8,598,006	6,622,380	12,083,169	49,070,504
Populist "public" AF	6,204	27,738	49,606	68,971	57,041	48,881	100,879	359,320
Populist "public" RF	6,525	5,641	6,803	8,013	6,634	7,381	8,349	7,323
Non-populist #tokens	7,323,569	11,387,486	8,838,299	14,394,700	11,452,223	8,869,712	14,949,392	77,215,381
Non-populist "public" AF	48,446	58,100	52,118	91,254	84,878	67,310	122,089	524,195
Non-populist "public" RF	6,615	5,102	5,897	6,339	7,411	7,589	8,167	6,789
P-value	0.3059	2.54E-43	6.61E-116	0	8.25E-94	2.81E-03	2.01E-07	1.41E-269
X^2 test	1.0482	190.4453	523.7064	2,181.3538	422.1633	21.9444	27.0286	1,230.5394
Statistical significance	NO	YES	YES	YES	YES	YES	YES	YES

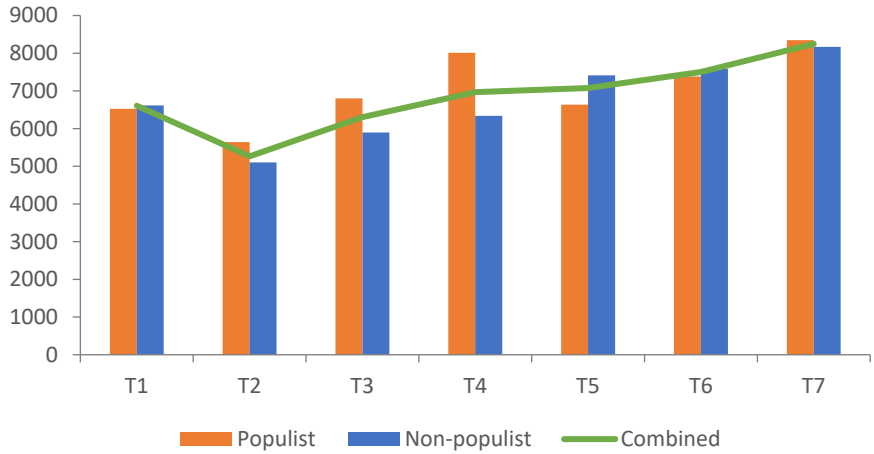


Figure 1: Relative frequency of nouns (y axis) referring to the public in speeches of MPs from populist and non-populist political parties in the Slovene parliament 1992–2018, by parliamentary term (x axis).

As can be seen from Table 2 and Figure 1, we observe a steady general upwards trend in the use of nouns, describing the public in both populist and non-populist parties over time. For all terms combined, populist MPs refer to the public statistically significantly more frequently than their non-populist counterparts (P-value 1.41E-269, X^2 test 1230.5394[16]), which confirms the main hypothesis. For all the MPs combined, the only, and quite substantial, drop in the frequency of references to the public can be observed from Term1 and Term2, which could be contributed to the early stages of the formation of the Slovenian political space. A lot of the discussions in parliament in these years were dedicated to shaping the new political system and (somewhat) changing the political culture, at the very least accepting newcomers to the parliamentary life and acquainting them with the work of an MP. However, this does not extend to a claim that the public was not important. On the contrary – already in the 1980s, before democratic changes took place, Slovenia was a stage for an active civil society that contributed towards legal and political democratization as well as proposed many possible solutions for the various political crises at the time, including drafts for a new constitution.[17]

16 *Calc: Corpus Calculator*. https://www.korpus.cz/calc/ (December 27 2022).
17 Božo Repe, *Slovenci v osemdesetih letih*. Ljubljana: Zveza zgodovinskih društev Slovenije, 2001.

Especially in Term1, the MPs had to face many questions of establishing the working of the new parliament itself. It took time before a new normality of the parliamentary work was established, before the MPs began to address the public more. While the early Slovene political transition exhibited a general consensus about the need to strengthen parliamentary democracy, the year after were and remain much less clear. In the years leading up to 2004, a new political aim became central, i.e. Slovenia's integration into the European Union, which to a certain extent offered a new common political goal for all the parliamentary parties. Nonetheless, the seemingly simple aim – to join the EU – contained a variety of different visions of how to actually achieve it, and political parties turned more and more to the public to try and gain (or maintain) their support throughout each voting cycle.[18]

After 2004, another major political shift took place. Slovenia was still an independent and sovereign state, but not as much as it has been in the previous years. Some of the state institutions had to adjust their functions and transmit a part of their decision-making process and jurisdictions on the supranational level of the EU.[19] With this earlier goal successfully achieved (Slovenia joined the EU on May 1 2004), the Slovene political space faced further polarization. This was reflected in the frequency and content of references of the public by the MPs, since they had to search for new contents of policy-making and ways of addressing their voters in an absence of a clear political goal.

As for individual terms, populist MPs refer to the public statistically significantly more often in Terms2-4 (1996–2008) and 7 (2014–2018) with Term4 as the biggest outlier, while the opposite is true of Terms5-6 with Term5 as the biggest outlier. In Term1, non-populist MPs use more public-denominating expressions but the difference is not statistically significant. Terms2-3 can be interpreted as the period of formation of populist parties (1992–2004), with Term4 (2004–2008) being the first parliamentary term working with a populist (SDS-led) government. The

18 Božo Repe, *Jutri je nov dan: Slovenci in razpad Jugoslavije*. Ljubljana: Modrijan, 2002, p. 7. Gašparič, *Državni zbor*, pp. 124–143.

19 Danica Fink-Hafner and Damjan Lah, *Proces evropeizacije in prilagajanja političnih ustanov na nacionalni ravni*. Ljubljana: Fakulteta za družbene vede, 2005, pp. 82-83. Danica Fink-Hafner and Damjan Lah, *Managing Europe from Home: The Europeanisation of the Slovene Core Executive*. Ljubljana: Fakulteta za družbene vede, 2003, p. 36.

switch towards more populist tendencies of SDS was at this time reflected in a new name (the party renamed itself from SDS – Socialdemokratska stranka/Social Democratic Party to SDS – Slovenska demokratska stranka/Slovene Democratic Party) and its succession to the European People's Party (EPP).[20] In turn, Term7 (2014–2018) could suggest the emergence of the second-wave growing power of populist parties in the face of the crisis of the non-populist parties – not only in Slovenia, but all across Europe.[21]

In Terms5-6 (2008–2014), when references to the general public prevailed in what sociologists and political scientists refer to as the non-populist discourse, the Slovenian political space witnessed an emergence of numerous new political parties, many of which entered parliament, which influenced the relation between populist and non-populist discourse.[22] Due to the safe-guards in parliamentary procedures which ensure equal opportunity of participation for opposition MPs regardless of their number, the speeches of MPs might also be influenced by the existence of populist and non-populist led governments and the strength of the populist and non-populist parties in the parliament at the time. While party strength is usually counted by the number of seats taken in the parliament, there are many more factors that influence it and make the correlation between the number of seats, coalition and opposition roles, and party strength.[23] Many of the parliamentary debates were influenced by the impact of the global economic crisis.[24]

20 Thomas Jansen, *At Europe's Service: the origins and evolution of the European People's Party*. Berlin: Heidelberg: New York: Springer, 2011.

21 James F. Downes and Edward Chan, "Explaining the electoral debacle of social democratic parties in Europe," *EUROPP. European Politics and Policy. London School of Economic*. https://blogs.lse.ac.uk/europpblog/2018/06/21/explaining-the-electoral-debacle-of-social-democratic-parties-in-europe/ (December 29, 2022). Stuart A. Brown, *The European Commission and Europe's Democratic Process. Why the EU's Executive Faces an Uncertain Future*. London: Palgrave Pivot, 2016, https://doi.org/10.1057/978-1-137-50560-6.

22 Organization for Security and Co-operation in Europe, *Observation of early parliamentary elections in Slovenia, 4 December 2011: OSCE/ODIHR Election Assessment Mission Final Report*. https://www.osce.org/odihr/elections/Slovenia/87786 (December 29 2022).

23 Giovanni Sartori, *Parties and party systems: a framework for analysis*. Colchester: ECPR, 2005. Alenka Krašovec, *Moč v političnih strankah: odnosi med parlamentarnimi in centralnimi deli političnih strank*. Ljubljana: Fakulteta za družbene vede, 2000.

24 Gašparič, *Državni zbor*, pp. 144-145.

4 Concluding remarks

While the results do confirm our initial hypothesis that populist parties refer to the public more, the difference between the two groups – populist and non-populist – appears to be smaller than the current findings of studies in sociology and political science suggest. Where research from these two fields mainly focuses on the speech of members of populist parties in (selected) television interviews, on social media, and other, less rigid environments, this contribution focused on taking into account all the speeches of MPs throughout the Slovenian parliament, which is a highly institutionalized and regulated environment that probably allows for less differentiation between MPs of different political orientation. Our results show that the same life-world of MPs, marked by their shared experience, social forms, norms, and a shared dialogue in plenary sessions, provides an environment with a strong unifying factor. Although there is little doubt that political parties themselves differ decisively from one another, the power of the institution, its rigidity and specificity, as well as MPs awareness of the target audience and reach of their speeches, proved to be decisive factors in MPs speech when speaking about the public.

According to political scientists and historians, the political space in Slovenia has been increasingly polarized since 1992. Again, our results show a somewhat more nuanced picture: while a growing difference between populist and non-populist discourse can be observed in Terms2-4, the gap narrows in Terms5-7. This challenges the dominant narrative of the Slovenian political space. The record high frequency of references to the public by populist MPs in Term4 coincides with the SDS winning the 2004 election for the first time after 1992, which happened immediately after the party went through its populist transformation in 2003. In Term5 the SDS witnessed a backlash with the non-populist coalition prevailing, while one of the populist parties, the NSi, did not even reach the parliamentary threshold.

The general public as well as the media frequently also refer to several of the more recent parties, such as Levica (The Left), as populist. While these parties do exhibit a certain populist appeal, their content, attitudes towards experts and state institutions, as well as their actions

in the parliament place them in the non-populist spectrum, with Levica gravitating more towards the democratic socialism[25] than to the same category of populism as defined by Mudde[26] which was the theoretical framework of this study. Another methodological issue is temporality: the modern populist shift is a phenomenon belonging to the 21st century; as such, the decade after 1992, included in our analysis, requires a separate interpretation and can only be understood as a preface to the later populist shift.[27]

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency research programmes No. P6-0436: Digital Humanities: resources, tools, and methods (2022–2027) and No. P6-0281: Political History, the CLARIN ERIC ParlaMint project (https://www.clarin.eu/parlamint) and the DARIAH-SI research infrastructure.

References

Aerts, R. (2019). *The ideal of parliament in Europe since 1800*. Cham: Palgrave Macmillan.

Brown, S. A. (2016). *The European Commission and Europe's Democratic Process. Why the EU's Executive Faces an Uncertain Future*. London: Palgrave Pivot. doi: 10.1057/978-1-137-50560-6

Bunce, V. (1995). Should transitologists be grounded? *Slavic Review*, *54*(1), 111–127.

Calc: Corpus Calculator. Retrieved from https://www.korpus.cz/calc/

Campani, G., & Pajnik, M. (2017). Populism in historical perspectives. In G. Lazaridis & G. Campani (Eds.), *Understanding the populist shift: othering in a Europe in crisis* (str. 13–30. London: New York: Routledge.

Carothers, T. (2002). The end of transition paradigm. *Journal of democracy, 13*(1), 5–20.

25 Alen Toplišek, "Between populism and socialism: Slovenia's Left party." In Giorgos Katsambekis and Alexandros Kioupkiolis (eds.), *The Populist Radical Left in Europe*. London: New York: Routledge, 2019, https://doi.org/10.4324/9781315180823-4.

26 Cas Mudde (ed.), *Racist Extremism in Central and Eastern Europe*. London: New York: Routledge, 2005. Cas Mudde, *Populist radical right parties in Europe*. Cambridge: Cambridge University Press, 2007.

27 Juan Francisco Fuentes, "Populism," *Contributions to the History of Concepts, Vol. 15*, Nr. 1 (2020), pp. 47–68.

Cas Mudde, *Populist radical right parties in Europe*. Cambridge: Cambridge University Press, 2007.

CLARIN Slovenia. Common Language Resources and Technology Infrastructure, http://www.clarin.si/info/about/

Fabijan, E., & Ribać, M. (2021). Politični in medijski populizem v televizijskem političnem intervjuju. *Social Science Forum*, *37*(98), 43–68.

Fink Hafner, D. (2019). *Populizem*. Ljubljana: Založba FDV.

Fink Hafner, D., & Lah, D. (2003). *Managing Europe from Home: The Europeanisation of the Slovene Core Executive*. Ljubljana: Založba FDV.

Fink Hafner, D., & Lah, D. (2005). *Proces evropeizacije in prilagajanja političnih ustanov na nacionalni ravni*. Ljubljana: Založba FDV.

Fišer, D., Konovšek, T., & Pančur, A. (2022). Referencing the Public by Populist and Non-Populist Parties in the Slovene Parliament. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities*, *September 15– 16 2022. Ljubljana, Slovenia* (pp. 243–247). Ljubljana: Inštitut na novejšo zgodovino. Retrieved from https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Frank, A., & Šori, I. (2015). Normalizacija rasizma z jezikom demokracije: primer Slovenske demokratske stranke, *Časopis za kritiko znanosti, 43*(260), 89–103.

Fuentes, J. F. (2020). Populism. *Contributions to the History of Concepts*, *15*(1), 47–68.

Gašparič, J. (2012). *Državni zbor 1992–2012: o slovenskem parlamentarizmu*. Ljubljana: Inštitut za novejšo zgodovino.

Gjuričová, A., & Zahradníček, T. (2018). *Návrat parlamentu. Česi a Slováci ve Federálním shromáždění*. Praha: Argo, 2018.

Gjuričová, A., Schulz, A., Velek, L., & Wirsching, A. (Eds.). (2014). *Lebenswelten von Abgeordneten in Europa 1860–1990*. Düsseldorf: Droste Verlag.

Habermas, J. (2007). *The Theory of Communicative Action*, *Vol. 2, Lifeworld and system: a critique of functionalist reason.* Cambridge: Polity Press.

Hadalin, J. (2020). Straight Talk. The Slovenian National Party's Programme Orientations and Activities. *Contributions to Contemporary History*, *60*(2). doi: 10.51663/pnz.60.2.10

Hadalin, J. (2021). What Would Henrik Tuma Say? From The Social Democratic Party of Slovenia to the Slovenian Democratic Party. *Contributions to Contemporary History 61*(3). doi: 10.51663/pnz.61.3.10

Husserl, E. (1996). *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: eine Einleitung und die phänomenologische Philosophie*. Hamburg: Meiner.

Ihalainen, P., Ilie, C., & Palonen, K. (Eds.). (2016). *Parliament and Parliamentarism. A Comparative History of a European Concept*. Berghahn.

Jansen, T. (2011). *At Europe's service: the origins and evolution of the European People's Party*. Berlin: Heidelberg: New York: Springer.

Krastev, I. & Holmes, S. (2019). *The light that failed: a reckoning*. London: Penguin Books.

Krašovec, A. (2000). *Moč v političnih strankah: odnosi med parlamentarnimi in centralnimi deli političnih strank*. Ljubljana: Fakulteta za družbene vede.

Lovec, M. (Ed.). (2019). *Populism and attitudes towards the EU in Central Europe*. Ljubljana: Faculty of Social Sciences.

Moffitt, B. (2015). How to Perform Crisis: A Model for Understanding the Key Role of Crisis in Contemporary Populism. *Government and Opposition, 50*(2), 189–217.

Mudde, E. (Ed.). (2005). *Racist Extremism in Central and Eastern Europe*. London: New York: Routledge.

Pajnik, M. (2019). Media Populism on the Example of Right-Wing Political Parties' Communication in Slovenia. *Problems of Post-Communism*, *66*(1), 21–32.

Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990–2018) siParl 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1300.

von Puttkamer, J., Borodziej, W., & Holubec, S. (Eds.). (2019). *From revolution to uncertainty: the year 1990 in central and Eastern Europe*. London: New York: Routledge.

Repe, B. (2002). *Jutri je nov dan: Slovenci in razpad Jugoslavije*. Ljubljana: Modrijan.

Repe, B. (2001). *Slovenci v osemdesetih letih*. Ljubljana: Zveza zgodovinskih društev Slovenije.Organization for Security and Co-operation in Europe, *Observation of early parliamentary elections in Slovenia, 4 December 2011: OSCE/ODIHR Election Assessment Mission Final Report*. Retrieved from https://www.osce.org/odihr/elections/Slovenia/87786

Rizman, R. (2006). *Uncertain Path. Democratic transition and consolidation in Slovenia*. College Station: Texas A&M University Press.

Sartori, G. (2005). *Parties and party systems: a framework for analysis*. Colchester: ECPR.

Schulz, A., & Wirsching, A. (Eds.). (2012). *Parlamentarische Kulturen in Europa. Das Parlament als Kommunikationsraum*. Düsseldorf: Droste Verlag.

Šori, I. (2015). Za narodov blagor: skrajno desni populizem v diskurzu stranke Nova Slovenija. *Časopis za kritiko znanosti,* 43(260), 104–117.

Toplišek, A. (2019). Between populism and socialism: Slovenia's Left party. In G. Katsambekis & A. Kioupkiolis (Eds.), *The Populist Radical Left in Europe*. London: New York: Routledge. doi: 10.4324/9781315180823-4

Sklicevanje populističnih in nepopulističnih strank na javnost v slovenskem parlamentu

Prispevek se posveča raziskovanju vprašanja različnih oblik in odstopanj od parlamentarne demokracije, pri čemer kot študijski primer jemlje Slovenijo med letoma 1992 in 2018. Osredotoča se na osrednjo institucijo parlamentarne demokracije, parlament, ter analizira delovanje poslancev oziroma njihovega govora v parlamentu vse od začetka prvega parlamentarnega mandata do izteka leta, ko so še na voljo podatki za analizo celotnega državnozborskega mandata. Skladno z domnevo, da je populizem svojevrsten politični slog, ki vzpostavlja posebno povezavo z množicami, torej javnostjo, si prispevek zastavlja vprašanje, ali so poslanci populističnih strank v primerjavi s poslanci nepopulističnih strank v svojem govorjenju kako drugače naslavljali splošno javnost. Rezultati empirične analize so pokazali, da so člani populističnih strank javnost naslavljaji pogosteje. Vendar je razlika med njimi in poslanci, ki so prihajali iz nepopulističnih strank, izrazito majhna, kar je moč pojasniti z močjo parlamenta kot institucije. To opozarja na pomembnost okolja, pričakovanj in navad (*Lebenswelt*), v katerem delujejo poslanci še tako različnih si strank.

Ključne besede: politične stranke, populizem, življenjski svet, parlament, Slovenija

Identifikacija metafore in metonimije v jezikovnih korpusih: poskus kategorizacije označenih metonimičnih prenosov v korpusu g-KOMET

Špela ANTLOGA
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Z jezikom nismo vedno zmožni neposredno ubesediti vsega, kar mislimo, zato za razlago pojavnosti uporabljamo različne jezikovno-kognitivne postopke, med drugim metafore in metonimije. Prepoznavanje vrednosti in razširjenosti metaforičnih in metonimičnih izrazov v jeziku je v zadnjih dvajsetih letih vodilo k povečanemu zanimanju za sistematično identifikacijo in luščenje tovrstnih figurativnih izrazov v korpusih posameznih jezikov. Izraze, pri katerih potekajo konceptualne preslikave, ki sodelujejo pri metaforičnih in metonimičnih procesih, je namreč težko izluščiti iz korpusa, ki niso posebej označeni za namene raziskovanja figurativnega jezika. V članku opredelim razumevanje konceptualne metafore in konceptualne metonimije, predstavim najpogostejše metode luščenja metaforičnih in metonimičnih izrazov iz jezikovnih korpusov ter na primeru korpusa g-KOMET, ki je ročno označen za metaforične izraze in metonimične prenose, ponazarjam poskus sistematizacije nekaterih najbolj prisotnih metonimičnih prenosov v slovenskem govorjenem jeziku.

Ključne besede: konceptualna metafora, konceptualna metonimija, metonimični vzorec, označevanje korpusa, govorjeni jezik

Antloga, Š.: Identifikacija metafore in metonimije v jezikovnih korpusih: poskus kategorizacije označenih metonimičnih prenosov v korpusu g-KOMET. Slovenščina 2.0, 11(1): 91–117

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.91-117

1 Uvod[1]

Jezik in mišljenje sta tesno povezana. Naše mišljenje je tako zapleteno, da z jezikom nismo vedno zmožni vsega »neposredno« izraziti, zato za razlago sveta uporabljamo različne jezikovno-kognitivne postopke, med drugim metafore in metonimije. Lakoff in Johnson (1980) poudarjata, da sta tako metafora kot tudi metonimija konceptualne narave in da gre za fenomen, ki igra osrednjo vlogo pri strukturiranju našega védenja o svetu. Korpusnih raziskav metafore in še posebej metonimije (ter tudi npr. ironije, sarkazma, pretiravanja, ki uporabljajo figurativni jezik) v slovenščini je malo. Čeprav so v zadnjem desetletju korpusne metode raziskovanja slovenščine postale uveljavljena empirična paradigma v jezikoslovju predvsem na področjih, povezanih z leksikologijo in slovnico ter jezikovno rabo, področje figurativnega jezika, ki je sicer na teoretski ravni dobilo zagon z razmahom teorije konceptualne metafore in metonimije (Lakoff in Johnson, 1980; Lakoff in Turner, 1989; Lakoff, 1993), pri tem trendu nekoliko zaostaja. Eden od možnih razlogov je pomanjkanje enotne in uspešne metode za sistematično identifikacijo metaforičnih in metonimičnih izrazov v že obstoječih korpusih, ki niso posebej označeni za konceptualne preslikave. S podobnimi težavami so se pred tem pri preučevanju metafore in metonimije soočali tudi v drugih jezikih, zato so se jezikoslovci za sistematično analizo konceptualnih struktur v jeziku zatekli k izgradnji korpusov z označenimi potencialnimi metaforičnimi in metonimičnimi izrazi. Gradnja in/ali označevanje takih korpusov sta časovno zamudna in zahtevata veliko prilagoditev označevalnih shem ciljnemu jeziku raziskovanja.

V prispevku bo predstavljeno razumevanje metafore in metonimije v okviru kognitivnega jezikoslovja, poleg tega bodo opisane različne bolj ali manj uveljavljene metode identifikacije metaforičnih in metonimičnih izrazov v obstoječih (splošnih) korpusih besedil z vsemi prednostmi in slabostmi. Kot eden od virov za sistematično analizo metaforičnih in metonimičnih izrazov v slovenskem jeziku bo predstavljen korpus

1 Članek je razširjena različica študentskega prispevka za konferenco Jezikovne tehnologije in digitalna humanistika 2022, ki je potekala 15. in 16. septembra 2022 v Ljubljani. Nadgrajen je z bolj poglobljenim in preglednim teoretičnim okvirom razumevanja in razlikovanja konceptualne metafore in metonimije ter figurativne in nefigurativne jezikovne rabe, z razširjenim opisom izdelave korpusa g-KOMET in obsežnejšo analizo metonimičnih prenosov v korpusu.

g-KOMET. Na primeru korpusa bo predstavljen poskus sistematizacije in klasifikacije najpogostejših označenih metonimičnih prenosov v slovenskem govorjenem jeziku.

2 Opredelitev metafore in metonimije

Konec sedemdesetih let prejšnjega stoletja se je tako zgodil t. i. kognitivni preobrat, ki je metaforo in metonimijo z jezikovne ravni prenesel na konceptualno, miselno raven. Ena ključnih sprememb takega pogleda na metaforo in metonimijo je odmik od slovnične (slovarske) razlage metafore in metonimije kot tipov pomenskih izpeljav h konceptualni metafori in metonimiji, ki nista samo del jezikovnega sporazumevanja, temveč sta način našega razmišljanja (pojave si razlagamo tako, da iščemo analogije in prenašamo podobnosti s poznanega na nepoznano). Jezikoslovcev torej ne zanimajo več zgolj posamezne besede z vidika pomenotvorne zmožnosti jezika, temveč tudi abstraktna podležeča razmerja med povezanimi koncepti, tj. načini mentalne organizacije konceptov, s pomočjo katerih človek osmišlja stvarnost, ki ga obdaja, in družbo, v kateri živi. Metafora in metonimija sta obravnavani kot konceptualni mehanizem, s pomočjo katerega se védenje o konkretnih pojavih in izkušnjah (ki so človeku lažje razumljiva in doumljiva) projicira na številne abstraktne domene, kot so recimo čas, čustva, organiziranost. Čas običajno konceptualiziramo kot prostor, čustva kot naravne sile, organizacije kot organizme ali stroje (Lakoff in Johnson, 1980; Lakoff, 1993).

Med različnimi teoretičnimi pristopi, ki so se posvečali preučevanju metafore, je ena najvidnejših teorija konceptualne metafore, ki so jo razvili George Lakoff in njegovi sodelavci (Lakoff in Johnson, 1980; Lakoff in Turner, 1989).[2] Po omenjenem teoretičnem modelu so **metafore** bistven element človekovega spoznavanja, ker je človekov konceptualni sistem po svoji naravi metaforičen, in sredstvo, ki nam omogoča,

2 V slovensko jezikoslovje je kognitivni pristop uvedla Kržišnik (1994) pri razlagi podstave frazeološke metafore s konceptualno metaforo. Obsežneje konceptualno metaforo obravnavata skupaj s Smolić (Kržišnik in Smolić, 1999), ko konceptualno metaforo prikažeta kot orodje za povezovanje in novo interpretacijo jezikovnega gradiva. Poleg omenjenih avtoric v še nekaterih sledečih objavah (Kržišnik in Smolić, 2000; Kržišnik, 2007) teorijo konceptualne metafore (in metonimije) v slovenščini razvijajo še Będkowksa-Kopczyk (2004, 2016), Bratož (2010) in Antloga (2020c).

da razumemo in doživljamo eno izkušenjsko področje ali domeno (*domain*) s pomočjo (v okviru) drugega. Prenos poteka s t. i. medpodročnimi preslikavami (*cross-domain mappings*) določenih elementov z izhodiščnega področja (*source domain*), ki je običajno konkretnejše, na ciljno področje (*target domain*), ki je bolj abstraktno. Konvencionalizirano to zapisujemo v obliki A JE B. V jeziku se konceptualne metafore (npr. ŽIVLJENJE JE TEKOČINA) realizirajo v obliki metaforičnih izrazov, npr. *biti poln življenja* (življenje kot tekočina, ki lahko napolni posodo), *uživati življenje z veliko žlico* (življenje je tako tekoče, da se prenaša z žlico), *življenje mi polzi skozi prste* (življenje je tako tekoče, da polzi po površini).

Tudi **metonimija** je bila v okviru tradicionalne retorike obravnavana predvsem kot retorična figura, torej je o njej razmišljala kot o jezikovnem pojavu, kot o predmetu figurativnega jezika. Aristotel je metonimijo pojmoval kot podtip metafore. Podobno definicijo metonimije zasledimo tudi v sodobnih slovarjih, npr. v Slovarju slovenskega knjižnega jezika.[3] Jakobson (1956) je poudaril inherentnost metonimije v jeziku in izpostavil pojem bližine kot temeljni princip metonimije. Kognitivni jezikoslovci, ki so se sicer bolj posvečali metafori kot metonimiji, se opirajo na ta in podobna stališča ter razširijo fenomen metonimije na pojmovno-pomenski mehanizem, ki omogoča strukturiranje jezika in mišljenja. Metonimije torej ne obravnavajo kot samo zamenjave enega leksema z drugim, ampak kot konceptualno orodje.

Če torej pri metafori prihaja do preslikave z enega konceptualnega področja na drugo na podlagi podobnosti (primerljivosti in skupnih lastnosti), do metonimičnih prenosov prihaja v okviru ene same domene na podlagi bližine (stičnosti, sorodnosti) – entitete, ki so si izkušenjsko blizu, lahko povezujemo oziroma zamenjujemo. Radden in Kovecses (1999, str. 21) ne govorita o *preslikavi*, temveč navajata, da je metonimija kognitivni proces, v katerem ena konceptualna entiteta *zagotavlja miselni dostop* do druge konceptualne entitete znotraj istega idealiziranega kognitivnega modela (IKM).[4] Dodajata, da so metonimične re-

3 »metonimija-e ž lit. besedna figura, za katero je značilno poimenovanje določenega pojma z izrazom za kak drug predmetno, količinsko povezan pojem«.

4 Po teoriji prototipov (Rosch v Lakoff 1987, str. 56) imajo kategorije v središču značilne predstavnike (prototipe), medtem ko se manj tipični predstavniki razvrščajo glede na to, v kolikšni meri odstopajo od lastnosti osrednjih primerkov.

lacije načini organizacije konceptualne strukture, in povzameta naslednje značilnosti metonimije (prav tam):

1. Metonimija je asociativno (s semiotičnega vidika indeksno) razmerje med pomenskimi enotami znotraj enega konceptualnega okvirja (v nasprotju z metaforo, ki je ikonično razmerje med dvema konceptualnima okvirjema).
2. Ker je metonimija konceptualna, je jezik samo ena od njenih uresničitev (najdemo jo v vizualnih umetnostih, gestah, glasbi itd.).
3. Izhodiščni pomen je konceptualno integriran v ciljnem pomenu kot rezultat metonimične operacije.
4. Izhodiščni in ciljni pomen sta konceptualno blizu.
5. Razmerje med izhodiščnim in ciljnim pomenom je naključno, tj. konceptualno nepogojeno.
6. Jeziki se lahko razlikujejo po tem, katere konceptualne povezave metonimično izkoriščajo.

Primeri metonimije se ne pojavljajo posamezno, ampak obstajajo sistematični metonimični koncepti, ki temeljijo na splošnih načelih. Eno takih splošnih načel, ki velja tudi za slovenščino, je, da lahko delo (knjižno, umetniško) uporabimo namesto avtorja tega dela (*Najraje berem Cankarja.*, *Od slovenskih impresionističnih del mi je najbolj všeč Grohar.*) Torej, avtor in njegova dela spadajo v IKM, ki ga lahko imenujemo IKM produkcije, v katerem so številne entitete, vključno s proizvajalcem (avtorjem), izdelkom (dela), krajem, kjer je izdelek narejen, itn. Vsi ti elementi tvorijo koherentno celoto v našem doživljanju sveta, saj se ponavljajo sočasno. Ker so v izkušnjah tesno povezane, se lahko nekatere entitete uporabijo za označevanje – to je za zagotavljanje miselnega dostopa do – drugih entitet znotraj istega IKM.

Znotraj domene je število konceptualnih preslikav, ki si jih delijo člani govorne skupnosti ali določene kulture, omejeno, npr. POVEČANA TELESNA TEMPERATURA ZA NAKLONJENOST (*Markovi možgani so si vtisnili v spomin njen dotik in ob misli nanjo mu je postalo vroče.*), POVEČAN SRČNI UTRIP ZA NAKLONJENOST (*Po tem, kako mi je razbijalo srce, ko sem ga zagledala, sem vedela, da ga še zdaleč nisem prebolela.*), RDEČICA ZA NAKLONJENOST (*V tistem trenutku zardim, ker vem, da me gleda ON.*), POTENJE ZA NAKLONJENOST (*Ko je stopil na oder,*

sem se od navdušenja začela potiti.), NEZMOŽNOST DIHANJA ZA NAKLONJENOST (»*Prvič v življenju mi je nekdo vzel dih, bilo je noro!*« *se čustvenega spoznanja spominja Mila.*), NEZMOŽNOST RAZMIŠLJANJA ZA NAKLONJENOST (*Ne pravijo zastonj, da smo zaradi ljubezni tudi malo nori, saj nam zamegli razum.*), NEZMOŽNOST PREHRANJEVANJA ZA NAKLONJENOST, NEZMOŽNOST POČITKA ZA NAKLONJENOST (*Tako sem vesela … in zaljubljena! Sploh ne morem jesti, sploh ne morem spati.*), ne pa na primer *NEZMOŽNOST UŽIVANJA PIJAČE IN NEZMOŽNOST PREBUJANJA ZA NAKLONJENOST (**Tako sem vesela … in zaljubljena! Sploh ne morem piti, zjutraj sploh morem vstati.*).

Tako metafora kot metonimija imata pomembno vlogo pri organizaciji pomena (enaka metonimična načela, ki povezujejo različne pomene besede, sodelujejo pri ustvarjanju novih pomenov v dejanski jezikovni rabi) ter pri produkciji in interpretaciji izreka (različna vloga metonimije v predikacijskih, propozicionalnih in ilokucijskih dejanjih, posebna vloga metonimije pri pragmatičnem sklepanju ipd.) (Littlemore, 2020).

Tabela 1: *Razlikovanje med metaforo in metonimijo. Povzeto po Feyaerts (2012)*

	metafora	**metonimija**
funkcija konceptualnega razmerja	sklepanje na podlagi podobnosti	referencialnost
narava konceptualnega razmerja	podobnost	(logična) povezava, bližina

Čeprav je iz tabele 1 razvidno razlikovanje med metaforo in metonimijo po funkciji in naravi konceptualnega razmerja, je včasih težko določiti mejo med njima. Npr. metafora JEZA JE VROČINA (*Kar zavrelo mu je, ko je videl, kaj je storila.*) je povezana s fiziološko reakcijo povišane telesne temperature pri doživljanju jeze in temelji na metonimičnem razmerju POSLEDICA ZA VZROK (povišana telesna temperatura za jezo). Goossens (1990) je za te primere skoval izraz metaftonimija (*metaphtonymy*). Barcelona (2001: 40) navaja, da obstajata dve prevladujoči vrsti metonimične motivacije za metaforo. (1) Metafora se v resnici ne razvije iz metonimije, ampak jo motivira in omejuje metonimični model ciljne domene. (2) Metafora nastane kot posplošitev metonimije. Dodaja (prav tam), da je metonimično motivirana večina

metafor za čustva (jeza, sreča, žalost, ljubezen, strah itd.) na podlagi fizioloških ali vedenjskih odzivov na čustva.

3 Dobesedni in nedobesedni (preneseni, figurativni) pomen pri metaforah in metonimijah

V kognitivni leksikalni semantiki je polisemija v okviru teorije IKM razumljena kot sistematično razmerje med različnimi pomeni (tako dobesednimi kot figurativnimi) oziroma kot rezultat konceptualne organizacije, kot je na primer kategorizacija. Tak pogled je privedel do različnih modelov za leksikalne mreže (Lakoff, 1987; Langacker, 1990), ki temeljijo na ideji, da vsak primerek besede vedno ohrani celotno paleto pomenov, ne glede na kontekst, v katerem se pojavi. Tako določena beseda pripada kompleksnemu pomenskemu omrežju, ki ga določajo različna področja in kognitivni procesi. Beseda predstavlja kategorijo različnih, a medsebojno povezanih pomenov, ki kažejo učinke prototipskosti. To dobro ponazarja primer analize prototipskosti angleške večpomenske besede *over* 'nad' (Lakoff, 1987, str. 416– 461). Njene pomene lahko sicer uvrstimo v eno kategorijo, vendar jih lahko ocenimo kot bolj (središčne) ali manj prototipske (periferne). Tudi pri slovenskem predlogu *nad* (Bratož 2010: 47–49) lahko opazimo izrazito večpomenskost (primeri v vrstnem redu od najbolj prototipnega (prostorskega) pomena (1), (2) k izražanju presežne mere (3), (4) do perifernih (prenesenih) pomenov za izražanje vzroka za določeno čustveno stanje (5) in družbenega hierarhičnega razmerja (6)): (1) *Slika je nad omaro.,* (2) *Nad naše kraje prihaja hladna fronta.* (3) *Velja za vse nad 18 let*. (4) *Slovenija je na vseh ravneh izobraževanja nad svetovnim povprečjem*. (5) *Zgražal se je nad njegovim nedostojnim obnašanjem*. (6) *Nad njim je le še generalni direktor.*

Če primerjamo metaforične in metonimične strukture, je pričakovati, da bomo metaforične izraze zaznali kot bolj *nedobesedne* v primerjavi z metonimičnimi. Nedobesednost je zelo podobna figurativnosti – metafora in metonimija sta namreč v številnih definicijah opredeljeni kot vrsti *figurativnega jezika*. Figurativne izraze (npr. *Zaradi božičnice je nekaterim zavrela kri*.) je mogoče prepoznati kot take le v sopostavitvi z njihovimi bolj dobesednimi izrazi (*Božičnica je nekatere zelo razjezila*.).

Množico jezikovnih izrazov si lahko namesto jasne dihotomije *dobesedni jezik* in *figurativni jezik* predstavljamo kot umeščene med dva pola – od skrajne dobesednosti do skrajne figurativnosti. *Figurativnost* že implicira dokaj visoko stopnjo nedobesednosti. Metonimije (*Živi le nekaj vrat naprej*.) so lahko v resnici blizu dobesednega pola in so bolj podobne strogo dobesednemu jeziku kot zelo figurativnemu jeziku (*Nogometaš J. Č. je ostal pred vrati kluba*.). Metonimični prenos med *vrati* kot delom hiše in celotno hišo je povsem *pravilen*; ne vzbuja nobenih podob. Ta jezikovni pojav je znan kot sistematična ali redna polisemija (Apresjan, 1974; v slovenskem prostoru predvsem Snoj, 2010; Vidovič Muha, 2000) ter sistemski in skoraj univerzalen mehanizem za označevanje konceptualno povezanih entitet na najbolj ekonomičen in naraven način. V tem smislu Gibbs in Colston (2012) navajata, da se zdi izraz *nedobesednost* primernejši kot *figurativnost*, saj zajame naravo zlasti številnih metonimičnih relacij.

4 Metode luščenja metaforičnih in metonimičnih izrazov v korpusih

V povezavi z metaforo in metonimijo sta zaradi pomanjkanja ustrezne metodologije problematična predvsem (sistematična) identifikacija in luščenje ustreznih podatkov iz splošnega jezikovnega korpusa. Konceptualne preslikave, ki sodelujejo pri metaforičnih in metonimičnih procesih, namreč niso neposredno povezane s posameznimi jezikovnimi oblikami in jih je težko izluščiti iz korpusov, ki niso posebej označeni za namene raziskovanja figurativnega jezika. S kombinacijo avtomatskega in ročnega luščenja podatkov iz splošnih korpusov so se v drugih jezikih izoblikovale različne metode identifikacije metaforičnih (in metonimičnih) izrazov (glej Stefanowitsch, 2006).

4.1 Ročno luščenje iz korpusa

Prva metoda z uporabo korpusa se je uveljavila zaradi potrebe po (bolj) sistematični analizi konceptualne metafore in metonimije. Iskanju po ključnih besedah že poznanih konceptualnih preslikav sledita branje besedila v korpusu ter sistematično izpisovanje metaforičnih in metonimičnih izrazov (Semino in Masci, 1996). Delo je zamudno in obsegovno

omejeno, predvsem pa neizkoriščeno z vidika količine podatkov v korpusu, a vsekakor bolj sistematično kot zanašanje na sporadične primere ali primere, ki ne izhajajo iz dejanske jezikovne rabe. Kljub temu je bila kognitivistom, ki so uporabljali ta pristop, očitana subjektivnost, neempiričnost in nekonsistentnost pri prepoznavanju (iskanju) in razlagi konceptualnih metafor in metonimij (npr. Tummers idr., 2005; Wasow in Arnold, 2005).

4.2 Iskanje po izhodiščni domeni

Metaforični izrazi so v izhodiščni domeni preslikave vedno povezani z neprenesenimi (nefigurativnimi) leksikalnimi enotami. Zato se je kot odziv na kritike oblikovala metoda iskanja izhodiščne domene po ključnih besedah oziroma identifikacija metafor na podlagi potencialnih izhodiščnih domen (pomensko polje, za katerega se predpostavlja oziroma je bilo že ugotovljeno, da sodeluje pri metaforičnih preslikavah, kot so na primer *srce*, *ogenj*, *boj*, *potovanje* ipd.). Iskanje lahko poteka preko posameznih besed v konceptualni strukturi ali preko skupine besed, ki so pomensko povezane (na primer *ogenj*, *plamen*, *vročina*, *pogoreti*, *zgoreti*, *plamteti*, *vzplamteti* ipd.). Z ročnim pregledovanjem rezultatov v korpusu je najprej določena potencialna metaforičnost izraza in nato ciljna domena metaforične preslikave (npr. LJUBEZEN, JEZA ipd.). Postopoma so nastajali seznami ključnih besed izhodiščnih domen za identifikacijo metafor v posameznih jezikih, kontekstih in diskurzih (Hanks, 2004; Koller, 2006). Za identifikacijo metonimije je izbrana pričakovano produktivna izhodiščna domena, za katero je že znano, da sodeluje v metonimičnih prenosih (npr. Hilpert, 2006, za izhodiščno domeno OKO).

Postopna uveljavitev identifikacije metaforičnih in metonimičnih izrazov v korpusih z iskanjem po ključnih besedah izhodiščne domene je vodila k zanimanju za raziskovanje figurativnega jezika v konkretnejših, bolj specifičnih domenah, npr. v političnem diskurzu, v ekonomiji, športu ipd.

4.3 Iskanje po ciljni domeni

Ob težnjah po identifikaciji metaforičnih izrazov v specifični domeni diskurza pristop, usmerjen v izhodiščno domeno, ni bil učinkovit, saj bi

zahteval predhodno poznavanje vira preslikave (izhodiščne domene), ki bi lahko bil potencialno najden v ciljni domeni. Zato se je uveljavila metoda iskanja ciljne domene s seznamom ključnih besed izhodiščnih domen. Za učinkovito identifikacijo metaforičnih in metonimičnih izrazov s ključnimi besedami ciljne domene je potrebna velika količina reprezentativnih in enotematskih besedil, ki so povezana z iskano ciljno domeno. To je relativno enostavno pri »konkretnih« ciljnih domenah, kot so POLITIKA, EKONOMIJA, ŠPORT, težje pa je iskanje metaforičnih in metonimičnih izrazov s ciljnimi domenami, kot so na primer ČUSTVOVANJE, UMSKA AKTIVNOST, ZAZNAVANJE. Dodaten problem, povezan s tovrstno identifikacijo metafor v korpusu, pa je, da so identificirane le izhodiščne domene, ki so povezane z izrazi, katerih pogostnost je v ciljni domeni tako visoka, da so se uvrstili na seznam ključnih besed. Analiza metaforičnih prenosov torej ni celovita in sistematična. Za identifikacijo metonimičnih prenosov tovrsten pristop ni učinkovit.

4.4 Iskanje po izhodiščni in ciljni domeni

Z združitvijo obeh predhodno navedenih metod se je uveljavila metoda iskanja stavkov oziroma delov besedil, ki vsebujejo ključne besede tako izhodiščne kot ciljne domene, predvsem v obliki avtomatskega luščenja metaforičnih izrazov (Philips, 2012). Kljub temu metoda še vedno zahteva poglobljen ročni pregled izluščenih podatkov zaradi možnih enakopisnic ali neprenesenega pomena obeh izrazov v stavku. Za tako iskanje je potreben zelo izčrpen seznam besed iz obeh domen, saj je sicer iskanje nepopolno. Poleg tega je metoda uporabnejša za raziskovanje že poznanih konceptualnih struktur, metafor in metonimij, manj pa za sistematično identifikacijo (novih oziroma vseh) konceptualnih struktur.

4.5 Iskanje po kazalnikih metaforičnosti

Nekaj poskusov identifikacije metaforičnih izrazov je potekalo tudi s t. i. kazalniki metaforičnosti. To so metajezikovni izrazi, ki napovedujejo oziroma signalizirajo metaforično rabo. Goatly (1997) kot metaforične signalizatorje navaja izraze, kot so *metaphorically*/*figuratively speaking* (metaforično/ figurativno rečeno, v prenesenem pomenu, preneseno

rečeno), *so to speak* (tako rekoč/če tako rečem), intenzifikatorje *literally* (dobesedno), *actually* (pravzaprav) ali celo ortografska znamenja, kot so narekovaji, poševni tisk ipd. S to metodo sicer izluščimo relativno malo metaforičnih izrazov, a osvetljuje jezikovne okoliščine, ko je metaforična raba v besedilu namerno (ali nenamerno) eksplicitno signalizirana (Skorczynska in Ahrens, 2015).

4.6 Iskanje po ustrezno označenem korpusu

Ena od zadnjih uveljavljenih ročnih metod je iskanje po korpusu, označenem s konceptualnimi preslikavami. Prvi korpus, označen s konceptualnimi preslikavami v obliki indirektne, direktne in implicitne metaforične besede[5] v štirih besedilnih tipih (časopisna besedila, strokovna besedila, literarna besedila in konverzacijska besedila) za angleški jezik,[6] je leta 2012 razvila skupina raziskovalcev, ki se je poimenovala Praglejazz. Ob tem je razvila postopek za ugotavljanje metaforičnih besed v besedilu, poimenovan MIPVU (Steen et al., 2010), da bi omogočila objektivnejšo, natančnejšo in bolj sistematično (jezikoslovno) analizo metaforičnih izrazov v različnih besedilih. Temeljno izhodišče za označevanje metaforičnih besed pri tem postopku je ugotavljanje razmerja med osnovnim in kontekstualnim pomenom besede. Pri tem je treba za vsako leksikalno enoto ugotoviti, ali se njen konkretni kontekstualni pomen razlikuje od njenega osnovnega pomena. Postopek je s prilagoditvami značilnostim posameznih jezikov sprožil zanimanje za identifikacijo metaforičnih izrazov in metafor v češčini (Pavlas in drugi, 2018), litovščini (Urbonaitė, 2016), madžarščini (Babarzy in Bencze, 2010), poljščini (Risinski in Mahula, 2015), srbščini (Bogetić, 2019) ter za izdelavo korpusov metafor v ruščini (Badryzlova in Lyashevskaya, 2017), hrvaščini (Despot in drugi, 2019) in kitajščini (Lu in Wang, 2017). Vsem navedenim je skupno, da so v korpus zajeta samo pisna besedila, da so označene samo metaforične besede, metafore ali domene preslikave (ne pa tudi metonimične besede, metonimični prenosi oziroma metonimije) in da je postopek identifikacije in označevanja metaforičnih besed prilagojen posebnostim ciljnega jezika. Eden od poskusov oblikovanja korpusa metafor v

5 Niso označene metafore, ampak besede, ki se potencialno lahko realizirajo kot metafore.
6 Gl. http://www.vismet.org/metcor/search/showPage.php?page=start.

slovenščini, ki bi omogočal jezikoslovno analizo metaforičnih izrazov in metafor v različnih besedilih, je korpus metafor KOMET 1.0 (Antloga, 2020a) in njegovo nadaljevanje z dodanimi transkripcijami govorjenega jezika, korpus g-KOMET (Antloga in Donaj, 2022).

5 Korpus g-KOMET

Korpus g-KOMET (korpus metaforičnih in metonimičnih izrazov v govorjenem jeziku)[7] s transkripcijami (po)govora v obsegu 52.529 besed je nadaljevanje pisnega korpusa metaforičnih izrazov in metafor KOMET 1.0. Besedilo za korpus je bilo izluščeno iz korpusa GOS 1.1.[8] Glede na želeno velikost korpusa (50.000 besed) smo iz vsake datoteke korpusa GOS izbrali 5 % besedila. Pri tem smo naključno izbrali začetno izjavo[9] govora in dodajali zaporedne izjave govora, dokler nismo dosegli želene velikosti. Če smo velikost dosegli sredi izjave, smo dodali tudi vse preostale besede v njej, zaradi česar je bila končna velikost korpusa 52.529 besed. Ohranili smo enako uravnoteženost besedila, kot je prisotna v korpusu GOS, tako po odstotku zastopanosti besedila v korpusu GOS 1.1 kot po zastopanosti žanrov. Korpus torej vključuje uravnotežen nabor transkripcij informativnega, izobraževalnega, razvedrilnega, zasebnega (telefonski pogovor, osebni stik) in nezasebnega (telefonski pogovor, osebni stik) diskurza. Če je bila beseda zapisana tako v pogovorni kot standardizirani obliki, smo prevzeli standardizirano obliko, ki je primernejša za označevanje. Pri luščenju besedila smo odstranili časovne oznake in oznake za menjavo govornih vlog, saj začetki in konci izluščenega dela besedila niso hkrati začetki in konci govornih vlog. Ohranili pa smo druge oznake, npr. smeh, hrup in prekinjene besede oz. napačne začetke. Za označevanje je bilo uporabljeno orodje Q-CAT (Brank, 2019).

5.1 Označevanje korpusa

Korpus je označevala ena oseba, ki je seznanjena s smernicami za označevanje metaforičnih besed in metonimičnih prenosov v besedilu

7 Projekt izdelave korpusa je bil financiran v okviru projekta CLARIN.si 2021. Korpus je dostopen na naslovu http://hdl.handle.net/11356/1293.

8 Dostopno na http://hdl.handle.net/11356/1438.

9 · [10] Izjavo in govorno vlogo razumemo, kot sta opredeljeni v specifikacijah za transkribiranje GOS (Zwitter Vitez idr., 2009).

v drugih jezikih in je postopek označevanja prilagodila slovenščini.[10] Označevanje je temeljilo na postopku za identifikacijo metaforičnih besed MIPVU (Steen et al., 2010),[11] ki je bil nadgrajen s postopkom za označevanje metonimičnih izrazov[12] in dodatno oznako pomenskega polja pri označenih metaforičnih izrazih. Označenost korpusa glede na vrsto označenih elementov je ponazorjena v tabeli 2.

Tabela 2: *Označenost korpusa g-KOMET*

Vrsta označenih elementov	**Število označenih besed (odstotek) ∑ = 52.529 besed**
metaforične besede	728 (1,38 %)
pomenska polja	65
metonimije	744 (1,42 %)

Kot prikazuje Slika 1, je postopek identifikacije (1–5) in označevanja (6–8) potekal po naslednjih korakih:

1. Branje celotnega besedila za splošno razumevanje pomena.
2. Določi posamezne leksikalne enote v besedilu.
3. Za vsako leksikalno enoto v besedilu določi njen pomen v kontekstu. Upoštevaj okolje, v katerem se leksikalna enota nahaja (kaj je pred in za njo).
4. Za vsako leksikalno enoto ugotovi, ali je njen kontekstualni pomen različen od njenega osnovnega pomena.[13] Osnovni pomeni so običajno konkretnejši (to pomeni, da si je tisto, kar evocirajo, lažje

10 Eden od ciljev ob prihodnji nadgradnji korpusa je, da bi vsaj del korpusa označilo več označevalcev. S tem bi lahko ugotovili stopnjo strinjanja med označevalci in posledično uspešnost označevalnega modela.

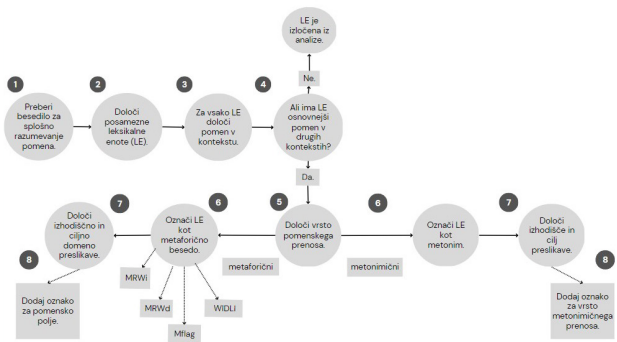
11 *Metaphor Identification Procedure Vrije Universiteit (MIPVU).*

12 Gre za predlog označevanja metonimičnih prenosov, in ne za izdelano označevalno shemo.

13 Osnovni pomen je bil določen na podlagi hierarhije slovarskih pomenov danega leksema v SSKJ. V uvodu v SSKJ (§ 75) je navedeno, da je na prvem mestu večpomenskih iztočnic osnovni pomen, to je tisti, ki je v sodobnem knjižnem jeziku najbolj nevtralen oziroma prevladujoč. Nadaljnji pomeni so razporejeni po stopnji odvisnosti od osnovnega pomena. Včasih na prvem mestu ni najmočnejši, ampak pomensko izhodiščni pomen iztočnice, zlasti če je to potrebno zaradi razvrstitve nadaljnjih pomenov (§ 76). Kadar ima beseda več pomenov, ki so enako močni, stoji na prvem mestu pomen, ki ima največ ali najmanj pomenskih prvin. Nadaljnji pomeni si sledijo, kakor število teh elementov ali upada ali narašča (§ 77). Manjši pomenski premiki so obdelani kot podpomeni ali pomenski odtenki v okviru nadrejenega pomena (§ 78). Pri slovničnih besedah odloča o razvrščanju razlag pogostnost, ob enaki pogostnosti pa ustaljeno slovnično zaporedje (§ 79). Pri označevanju metonimij so bila upoštevana tudi lastna imena, katerih osnovni pomen je bil določen na podlagi drugih dostopnih virov ali jezikovne presoje označevalca.

predstavljati, videti, slišati občutiti, vohati ali okusiti), povezani s telesnim delovanjem, bolj natančni in jasni. Če sta osnovni in kontekstualni pomen leksikalne enote enaka, to leksikalno enoto izloči iz nadaljnje analize.

5. Če ima leksikalna enota osnovnejši pomen v drugih kontekstih, določi vrsto pomenskega prenosa (metaforični ali metonimični).[14]
6. Če je pomenski prenos metaforičen, leksikalno enoto označi kot metaforično besedo in ji dodaj oznako za vrsto metaforičnega prenosa. Če je pomenski prenos metonimičen, označi leksikalno enoto kot metonim.
7. Označeni leksikalni enoti določi izhodišče in cilj preslikave oziroma prenosa pomena.
8. Če je označena leksikalna enota metaforična, dodaj oznako za pripadajoče pomensko polje. Če je označena leksikalna enota metonimična, dodaj oznako za vrsto metonimičnega prenosa.



Slika 1: *Postopek identifikacije in označevanja metaforičnih besed, pomenskih polj in metonimičnih prenosov v korpusu g-KOMET.*

5.1.1 Označevanje metaforičnih besed (oznake MRWi, MRWd, MFlag in WIDLI)

S predstavljenim postopkom so bili identificirani jezikovni izrazi, ki imajo potencial, da jih ljudje realiziramo kot metafore. V analizo so bili

14 Medtem ko je upoštevanje leksikalne enote kot enote označevanja smiselno z vidika sistematičnosti brez sprotnega odločanja o označevalni enoti, se je pri označevanju metonimije izkazalo, da je smiselno upoštevati širšo enoto pri sklicevanju na metonimični prenos, saj lahko več izrazov (že po definiciji metonimije) razumemo kot celoto.

vključeni glagoli, samostalniki, pridevniki in predlogi.[15] Identificirane metaforične besede so dobile naslednje oznake:

1) Indirektna metafora (*MRWi*) – pomen besede, označene kot *MRW*, ni osnovni pomen besede in je lahko pojasnjen kot preslikava z ene domene na drugo. Kontekstualni pomen je sicer že lahko konvencionaliziran in zato opredeljen v slovarju. Npr.: *Lepi, slavni in bogati* ***naravnost*** *obožujejo knežjo družino, ki vsakič poskrbi za* ***pikantne*** *govorice.*
2) Direktna metafora (*MRWd*) – beseda, označena kot MRW, je rabljena v svojem osnovnem pomenu, kontekstualni pomen pa je lahko pojasnjen kot preslikava z ene domene na drugo. Npr.: *Glasbenica Neisha je pred kratkim predstavila svoj novi videospot za pesem Prihaja maj, ki ga je med drugim snemala v ljubljanskem botaničnem vrtu, kjer se je med drugim zlila s tropskimi drevesi* ***kot*** *nekakšna* ***rajska ptica****.*
3) Mejni primer (*WIDLI*) – iz konteksta ni mogoče določiti, ali gre za metaforični ali osnovni pomen besede. Npr.: *»Do derbija se je že nakopičila utrujenost ob nizu zaporednih tekem.« Mladi hrvaški reprezentant* ***se*** *je tako kot večina najprej* ***dotaknil*** *bolečih mišic.* V tem primeru ni mogoče določiti, ali gre za osnovni pomen glagola *dotakniti se* ('približati se tako, da pride do dotika, stika') ali metaforični ('na kratko, nekoliko spregovoriti o čem').
4) Metaforični signalizator (MFlag) – metaforično jezikovno rabo lahko signalizirajo posamezne besede in slovnične strukture, npr. *kot*, *kakor*, *enačiti kaj z*, *metaforično rečeno* itd.

5.1.2 Uvrščanje v pomensko polje metaforičnega prenosa (oznaka frame*)*

Označeni metaforični izrazi so bili uvrščeni v pomenska polja, ki omogočajo, da te izraze umestimo v pomenske kategorije, ki motivirajo

15 V korpusu g-KOMET za razliko od korpusa KOMET 1.0 prislovi niso vključeni v analizo, ker je večina označenih prislovov v korpusu KOMET 1.0 dobila oznako 'WIDLI' (mejni primeri). Slovar namreč načeloma eksplicitno ne hierarhizira pomenov prislova, ampak skupaj s primeri navaja zgolj glagol ali pridevnik, iz katerega je tvorjen. Zgolj na podlagi tega osnovnega pomena posameznega prislova ni mogoče določiti, saj ne predpostavljamo, da se hierarhija pomenov ohranja pri prehodu v drugo besedno vrsto (enako velja za glagolnike, če nimajo določenih pomenov).

metaforični prenos (npr. naravni pojavi, čas, prostorska orientacija, družina, premikanje itd.). V označenem korpusu pa, obratno, pomensko polje omogoča, da znotraj pomenske kategorije poiščemo metaforične izraze, ki so lahko potencialno uresničitev neke konceptualne strukture (npr. uresničitve pomenske kategorije 'boj' (Antloga, 2020c): *osvoboditi se zamer* (ciljna domena: čustva), *ubraniti se očitkov* (ciljna domena: komunikacija), *izboriti si dodaten čas* (ciljna domena: čas), *boj za oblast* (ciljna domena: politika), *sklepni boj na volitvah* (ciljna domena: volitve), *boriti se z boleznijo* (ciljna domena: zdravje), *bitka s kilogrami* (ciljna domena: debelost), *pomočnik v boju proti sivim lasem* (ciljna domena: staranje) itd.). V korpusu g-KOMET je bilo označenim metaforičnim besedam in stalnim besednim zvezam določenih 65 pomenskih polj (premikanje, vidno zaznavanje, telo, čas, boj, oblika, trgovanje, tekočina, slušno zaznavanje, položaj, lov, medosebni odnosi itd.). Pomenska polja so bila deloma definirana vnaprej na podlagi najpogostejših izhodiščnih domen v drugih jezikih, deloma pa dopolnjena med označevanjem korpusa. Seznam ni končen in se lahko ob dodajanju novih besedil še dopolnjuje.

5.1.3 Označevanje metonimičnih prenosov

Metonimijo lahko kategoriziramo v različne vrste in podvrste, ki so lahko (skoraj) univerzalne in prisotne v več jezikih ali specifične v posameznem jeziku. V postopku identifikacije metonimično rabljenih izrazov nas je zanimalo, kako je organizirana konceptualna struktura oziroma kakšno je razmerje med pomenskimi enotami znotraj enega konceptualnega okvirja. Pri označevanju smo ugotavljali razmerje med obema entitetama preslikave znotraj ene domene v slovenščini (npr. INSITUTCIJA ZA OSEBO, KRAJ ZA OSEBO). Za opis vrste metonimičnega prenosa je bil vnaprej pripravljen delni seznam glede na izpričanost v drugih jezikih, ki smo ga sproti ob označevanju dopolnjevali. Ob pregledu označenih elementov smo za morebitne prilagoditve oznak upoštevali različico označevalne sheme, ki je nastala pri prvem označevanju. Ugotovljenim metonimičnim izrazom je bilo določenih 45 tipov metonimične preslikave (glej primere v tabeli 3).

Tabela 3: *Najpogostejši označeni metonimični prenosi v odstotkih glede na vse označene metonimične izraze v korpusu g-KOMET s primeri*

Vrsta metonimičnega prenosa	**Odstotek glede na vse označene metonimične izraze v korpusu g-KOMET**	**Primeri**
SPLOŠNO ZA SPECIFIČNO	16,8 %	*bolgari pa makedonci sploh nimajo sklonov* (njihovi jezik)
INSTITUCIJA ZA ČLANE	9,7 %	*s kakšnimi težavami se centri za socialno delo soočajo v praksi* (delavci)
DEL ZA CELOTO	7,1 %	*ti boš pa medtem mojega sinčka previjala* (skrbela za otroka)
REZULTAT DEJANJA ZA DEJANJE	6,4 %	*če si pa samo minus delam* (zapravljam)
AVTOR ZA DELO	6,3 %	*smo nadaljevali s Puškinom* (njegovo pesnitvijo)
LASTNOST ZA OSEBO	6 %	*ti si blond* (neumna)
CELOTA ZA DEL	3,6 %	*amerika ima ta optimizem in energijo* (prebivalci)
PREDMET ZA AKTIVNOST	3,6 %	*skos pred televizijo* (gledati program na televiziji)
LOKACIJA ZA OSEBO	3,5 %	*cela gerbičeva je bila tam* (študenti, ki živijo v študentskem domu na Gerbičevi)
LASTNOST ZA AKTIVNOST	2,1 %	*sem odprta za nove ideje*[16] (sprejemam nove ideje)
DEL TELESA ZA OSEBO	1,6 %	*šest evrov po glavi*
SREDSTVO DEJANJA ZA REZULTAT DEJANJA	1,3 %	*največkrat udari z besedami*[17] (s povedanim)
OSEBA ZA GIBANJE/ IDEOLOGIJO	1,3 %	*veliko navdušenje podpornikov obame* (njegove politike, ideologije)
DEJANJE ZA REZULTAT DEJANJA	1,2 %	*niti enkrat samkrat ni penzla v barvo namočil* (ni prebarval nečesa)
PODJETJE ZA DELAVCA	1,2 %	*merkator je kritiziral državno politiko* (zaposleni v Merkatorju)
LOKACIJA ZA DOGODEK	1,2 %	*greš na pokljuko* (na tekmo za svetovni pokal)

16 Gre tudi za metaforo IDEJE SO STVARI (in se lahko prenašajo), VIDNO JE ZAUPANJA VREDNO (kar je odprto, je vidno; kar vidimo, o tem se lahko prepričamo in temu zaupamo).

17 Gre tudi za metaforo BESEDNI SPOR JE SPOPAD.

6. Poskus sistematizacije označenih metonimičnih prenosov

Nadaljnja vsebinska delitev temelji na kategorizaciji glede na prevladujočo specifično pojmovno vsebino, ki omogoča dostop do drugega pomena preko metonimije.

STVAR ZA X

Metonimije STVAR ZA X so metonimije, katerih cilj (predvidena izhodiščna entiteta) je STVAR, do katere se dostopa s pomočjo referenčne vsebine, ki je z njo povezana v istem idealiziranem kognitivnem modelu. Metonimije STVAR ZA X lahko nadalje razdelimo v podvrste glede na konceptualno izhodišče metonimičnega prenosa:

- STVAR ZA STVAR: *da <u>kozica</u> vre dvajset do petindvajset minut* → POSODA (kozica) NAMESTO VSEBINE (vode v kozici)
- STVAR ZA ČLOVEKA: *pa so samo še <u>bobni</u> igrali* → INŠTRUMENT (bobni) NAMESTO GLASBENIKA, KI IGRA TA INŠTRUMENT
- STVAR ZA LASTNOST: *vidi <u>mercedesa</u> ko se pogleda v ogledalo* → AVTO (mercedes) NAMESTO VRLINE/POMANJKLJIVOSTI (samovšečnost, ošabnost)
- STVAR ZA DOGODEK: *na <u>rdeči preprogi</u>* (...) *znova zablestela* → STVAR NA DOGODKU (rdeča preproga) NAMESTO CELOTNEGA DOGODKA (podelitev nagrad)

LASTNOST ZA X

Pri metonimijah LASTNOST ZA X je za cilj (predvideni referent je LASTNOST) prenosa pomembna bližina posameznika ali skupine idealu, ki ga postavlja standardni referent, z neko (stereotipno) lastnostjo (ki nadomesti preostale lastnosti), največkrat vidno lastnostjo (ki lahko nadomesti čustvene lastnosti) ipd. Glede na konceptualno izhodišče metonimičnega prenosa v korpusu g-KOMET ločimo:

- LASTNOST ZA SKUPINO: *pa <u>črni</u> so tam* → ENA OD LASTNOSTI RASE (barva polti) NAMESTO PREDSTAVNIKOV RASE
- LASTNOST ZA OSEBO: *pa pa kekse sva jedli ko so ko sta nama jih prinesla z nizomzemske <u>stara</u> dva* → STAROST OSEBE NAMESTO OSEBE OSEBA ZA X

Metonimije OSEBA ZA X so pogoste metonimije, pri katerih prihaja do prenosa človekove dejavnosti, rezultatov dejavnosti, prostora dejavnosti ipd. na osebo, ki opravlja to dejavnost. Razdelimo jih lahko v naslednje podkategorije:

- OSEBA ZA AKTIVNOST: *sem zadnjič gledal nogometaše* → OSEBA, VKLJUČENA V AKTIVNOST (nogometaš), NAMESTO AKTIVNOSTI (nogometna tekma)
- OSEBA ZA TEORIJO: *vsi citirajo žižka* → PREDSTAVNIK TEORETIČNEGA PRISTOPA NAMESTO IZHODIŠČ, UGOTOVITEV TEGA PRISTOPA
- OSEBA ZA LOKACIJO: *pa pri zdravniku sto let čakala* (…) → OSEBA, KI OPRAVLJA DEJAVNOST (zdravnik), NAMESTO PROSTORA, KJER SE OPRAVLJA DEJAVNOST

LOKACIJA ZA X

Pri metonimijah LOKACIJA ZA X je lokacija uporabljena za priklic ene ali več entitet, ki so na tej lokaciji. Ker sta lokacija in to, kar se nahaja na lokaciji, v nekakšni prostorski relaciji, bi lahko vse tovrstne metonimije umestili k metonimijam DEL ZA CELOTO. Metonimije LOKACIJA ZA X lahko razdelimo v podvrste:

- LOKACIJA ZA DOGODEK: *to mi je ostalo od otočca* → KRAJ, KJER JE POTEKAL DOGODEK (Otočec), NAMESTO DOGODKA (festival Rock Otočec)
- LOKACIJA ZA INSTITUCIJO: se zmenijo na *šubičevi* → IME ULICE (Šubičeva ulica) NAMESTO INSTITUCIJE NA TEJ ULICI (Slovenski parlament)
- LOKACIJA ZA OSEBO: *v gostilni pa vse čisto tiho* → PROSTOR, KJER SE ZADRUŽUJE OSEBA, NAMESTO OSEBE V TEM PROSTORU

PROSTORSKA ORIENTACIJA ZA X

V to skupino uvrščam izraze, ki so celo ključni za govorjeno rabo (Verdonik, 2015), to so izrazi, s katerimi se orientiramo v prostoru in času ter imajo (tudi) deiktično funkcijo (prav tam), npr. tukaj/tule – tam, gor – dol, in so hkrati metonimično motivirani. Da so v govorjeni rabi pogostejši kot v pisni, lahko pojasnimo z večjo vpetostjo govorjene rabe v prostor in čas kot komunikacijski okoliščini (kot tudi nasploh z večjo vpetostjo govorjene rabe v kontekst).

- PROSTORSKA ORIENTACIJA ZA PROCES: *ja teran je kraški kraški teran je [...] a bova jutri spirala sinuse malo doli, kaj? Če bova spirala sinuse malo jutri ha* → SMER POTEKA PRENOSA TEKOČINE PO TELESU NAMESTO PROCESA PRENOSA TEKOČINE PO TELESU in ODSTRANJEVANJE STRUPENIH SNOVI NAMESTO PITJA ALKOHOLA
- PROSTORSKA ORIENTACIJA ZA REZULTAT DEJAVNOSTI: *pa saj imam to na računalniku gor* → REZULTAT DEJANJA NAMESTO PROCESA (*na računalniku gor* za proces prenosa programa, datoteke s pomnilniške naprave na trdi disk računalnika (*nalaganje*) in SMER PREMIKANJA NAMESTO PROCESA PREMIKANJA K ZGORNJI STRANI ČESA, TAKO DA NASTANE NEPOSREDEN STIK (*naložiti gor* za *naložiti na*)
- PROSTORSKA ORIENTACIJA ZA ČLOVEŠKO TELO: *pa še čisto do gor je zaprt* → SMER (*gor*) NAMESTO POZICIJE DELA TELESA (*vrat*) (izraz *biti do vratu zaprt* je sicer metaforičen)
- PROSTORSKA ORIENTACIJA ZA LOKACIJO: (pogovor o restavracijah) *Kje to? Aha. Zdaj se mi zdi da so ene štiri gor ne zdaj še ena se odpre in vse so dobre.* → SMER GIBANJA (gor, tj. z našega gledišča proti višjemu kraju) ZA CILJ (gor, tj. višja lokacija)

Označene vrste metonimije glede na metonimični prenos, kot so navedene v tabeli 3 in poglavju 6, kažejo na nekatere značilnosti govorjenega jezika:

- uporaba splošnih izrazov namesto specifičnih,
- posplošitev opisovanja dejanja na njegov rezultat,
- uporaba specifične leksike, ki je poznana govorni skupnosti, za splošnejši pojav (*iz krsta seveda ne v krstu obravnava pač vsebina tema krsta je pokristjanjevanje* (za Krst pri Savici), *pri pouku samo tega cankarja skozi* (za dela, besedila, odlomke, življenje Ivana Cankarja), *kar se zmenijo na šubičevi* (za dogovore, ki so sprejeti v Slovenskem parlamentu) ipd.),
- uporaba slogovno zaznamovane leksike, ki je motivirana metonimično (*blonda*, *moja deklina*),
- uporaba deiktičnih izrazov, ki so motivirani metonimično (*gor*, *dol*).

Metonimije lahko opazujemo tudi glede na vidik, ki določa izhodišče/sredstvo metonimičnega prenosa. Pogled izhaja iz predpostavke, da ima konceptualna metonimija izkustvene in spoznavne temelje, njene jezikovne uresničitve pa so samo ena od možnih oblik, skozi katere se izraža (v kontekstu že omenjenih idealiziranih kognitivnih modelov, ki predstavljajo abstrakcijo človekovih izkustev in delno zajemajo naše vedenje o svetu). Za kognitivne pristope je namreč primarno ravno vprašanje, zakaj izberemo prav določeno konceptualno entiteto za metonimični izraz, in ne neke druge.

Na tej podlagi lahko označene metonimične izraze opazujemo tudi z vidika povezave med pogostnostjo metonimičnega prenosa in človekovim izkustvom (npr. metonimični prenosi v korpusu g-KOMET *splošno za specifično* (125): *specifično za splošno* (3); *konkretno za abstraktno* (7) : *abstraktno za konkretno* (3); *definirano za nedefinirano* (2) : **nedefinirano za definirano* (0)). Zaradi lažjega razumevanja je bolj verjetno, da bodo metonimični prenosi potekali s splošnega na specifično, s konkretnega na abstraktno ipd. Povezanost z visoko pogostnostjo označenih tovrstnih metonimičnih prenosov v korpusu je ena od bistvenih (najpogostejših) funkcij metonimije, tj. referencialna funkcija, ki je nekakšna bližnjica za označevanje kompleksnega in abstraktnega pojava z enostavnejšim, konkretnejšim in razumljivejšim pojavom (izrazom).

Poleg tega lahko metonimične povezave opazujemo tudi z vidika povezave med pogostnostjo metonimičnega prenosa in kulturno preferenco:

- DEL TELESA, KI SODELUJE PRI PROCESU (*usta*), ZA PROCES (*govorjenje*) in PROCES (govorjenje) ZA DEJAVNOST (pripovedovanje): *Kako so se ohranjale ljudske pesmi? Eee govorili so si jih od ust do ust*),
- IZDELEK ZA LASTNOST: *pa ta ti je ko laško proti vodi* (je veliko boljši),
- KRAJ ZA SPECIFIČEN PROSTOR V TEM KRAJU in SPECIFIČNI PROSTOR V KRAJU ZA DEJVANOST: *Greš v kranjsko? Grem na cerkno letos.* (za smučanje na smučiščih v Kranjski gori in v Cerknem)).

Te pojmovne sheme združujejo posamezne elemente, povezane z našim kulturnospecifičnim védenjem o svetu, družbi, konvencijah

in običajih. V konkretni jezikovni situaciji pogosto kontekst in izkustvo določata, kateri segment enciklopedičnega védenja se bo profiliral kot pomemben in se jezikovno realiziral.

7 Sklep

Metonimija in metafora nista zgolj retorični figuri, ampak sta kognitivna pojava, ki imata pomembno vlogo tako pri organizaciji pomena kot tudi pri produkciji in interpretaciji izreka. Namen članka je nakazati poskus sistematizacije metonimičnih načel, ki povezujejo različne pomene besede, kot so bili označeni v korpusu govorjenih besedil. V naslednjih korakih bo potrebna analiza metonimičnih prenosov v dejanski jezikovni rabi tudi s pragmatičnega vidika. Čeprav so nekatere metaforične preslikave in metonimični prenosi univerzalni oziroma prisotni v več jezikih, je namreč unikatna njihova frekvenca pojavljanja v posameznih jezikih, njihova realizacija in vpetost kulturnospecifičnih elementov jezikovnega prostora. Za nadaljnjo analizo metaforičnih izrazov v slovenskem jeziku bo zanimiva tudi primerjava korpusa KOMET 1.0, v katerem so označene metaforične besede v zapisanem jeziku, in korpusa g-KOMET, ki vsebuje govorjena besedila v obliki transkripcij.

Literatura

Antloga, Š. (2020a). *Korpus metafor KOMET 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1293

Antloga, Š. (2020b). Korpus metafor KOMET 1.0. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference, 24.–25. september 2020* (str. 176–170). Ljubljana: Inštitut za novejšo zgodovino.

Antloga, Š. (2020c). Vloga metafor in metaforičnih izrazov v medijskem diskurzu: analiza konceptualizacije boja. V J. Vogel (ur.), *Slovenščina – diskurzi, zvrsti in jeziki med identiteto in funkcijo* (str. 27–34). Ljubljana: Znanstvena založba Filozofske fakultete.

Antloga, Š., & Donaj, G. (2022). *Korpus g-KOMET*. Slovenian language resource repository, CLARIN.SI. http://hdl.handle.net/11356/1490

Apresjan, J. (1974). Regular polysemy. *Linguistics: An inderdisciplinary journal of the language sciences, 12*(142), 5–32.

Babarczy, A., & Bencze, I. (2010). The automatic identification of conceptual metaphors in Hungarian texts: A corpus-based analysis. V *LREC 2010 Workshop on Methods for the Automatic Acquisition of Language Resources: Proceedings* (str. 31–36).

Badryzlova, Y., & Lyashevskaya, O. (2017). Metaphor Shifts in Constructions: the Russian Metaphor Corpus. V *Computational construction grammar and natural language understanding: Papers from the 2017 AAAI Spring Symposium*. The AAAI Press.

Barcelona, A. (2001). On the plausibility of claiming a metonymic motivation for conceptual metaphor. V A. Barcelona (ur.), *Metaphor and metonymy at the crossroads: cognitive approaches* (str. 31–58). Berlin in New York: Mouton de Gruyter.

Będkowska-Kopczyk, A. (2004). *Jezikovna podoba negativnih čustev v slovenščini*. Ljubljana: Študentska založba.

Będkowska-Kopczyk, A. (2016). Začutiti in občutiti: kognitivna analiza pomensko-skladenjskih lastnosti dveh predponskih tvorjenk iz glagola čutiti. V E. Kržišnik in M. Hladnik (ur.), *Toporišičeva obdobja* (str. 41–48). Ljubljana: Znanstvena založba Filozofske fakultete.

Bernjak, E., & Fabčič, M. (2018). Metonimija kot konceptualni in jezikovni pomen. *Anali PAZU HD 4*/(1–2), 11–23.

Bogetić, K. (2019). Linguistic metaphor identification in Serbian. V S. Nacey in T. Krennmayr (ur.), *MIPVU in Multiple Languages* (str. 203–226). Amsterdam: John Benjamins.

Brank, J. (2019). *Q-CAT Corpus Annotation Tool*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1262

Bratož, S. (2010). *Metafore našega časa*. Koper: Fakulteta za management.

Croft, W., & Cruise, A. D. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Despot, K., Tonković, M., Brdar, M., Perak, B, Ostroški Anić, A., Nahod, B., & Pandžić, I. (2019). MetaNet.HR: Croatian Metaphor Repository. V M. Bolognesi, M. Brdar in K. Š. Despot (ur.), *Metaphor and Metonymy in the Digital Age. Theory and Methods for Building Repositories of Figurative Language* (str. 123–146). Amsterdam: John Benjamins.

Feyaerts, K. (2012). Refining the Inheritance Hypothesis: Interaction between metaphoric and metonymic hierarchies. V A. Barcelona (ur.), *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective* (str. 59–78). Berlin: De Gruyter Mouton.

Gibbs, R. W. (1999). Researching Metaphor. V L. Cameron in G. Low (ur.), *Researching and applying metaphor* (str. 29–47). Cambridge: Cambridge University Press.

Gibbs, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning*. New York: Cambridge University Press.

Goatly, A. (1997). *The Language of Metaphors*. London in New York: Routledge.

Goossens, L. (1990). Metaphtonymy: the interaction of metaphor and metonymy in expressions for linguistic action, *Cognitive Linguisitcs, 1*(3), 323–342.

Gries, S., & Stefanowitsch, S. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics,* 9/1, 97–129.

Hanks, P. (2004). The syntamatics of metaphor and idiom. *International Journal of Lexicography 17*(3), 245–274.

Hilpert, M. (2006). Keeping and eye on the data: Metonymies and their patterns. V A. Stefanowitsch in S. Gries (ur.), *Corpus-Based Approaches to Metaphor and Metonymy* (str. 123–152). Berlin: De Gruyter Mouton.

Jakobson, R. (1956). The Metaphoric and Metonymic Poles. V R. Driven in R. Pörings (ur.), *Metaphor and Metonymy in Comparison and Contrast* (str. 41–47). Berlin/New York: Mouton de Gruyter.

Koller, V. (2006). Of critical importance: Using electronic text corpora to study metaphor in business media discourse. V A. Stefanowitsch in S. Gries (ur.), *Corpus-Based Approaches to Metaphor and Metonymy* (str. 237–266). Berlin: De Gruyter Mouton.

Kövecses, Z. (2002). *Metaphor: A practical Introduction*. Oxford/New York: Oxford University Press.

Kržišnik, E. (1994). *Slovenski glagolski frazemi*. Doktorska disertacija. Ljubljana.

Kržišnik, E., & Smolić, M. (1999). Metafore, v katerih živimo tukaj in zdaj. V E. Kržišnik (ur.), *35. seminar slovenskega jezika, literature in kulture* (str. 61–80). Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete.

Kržišnik, E., & Smolić, M. (2000): »Slike« časa v slovenskem jeziku. V I. Orel (ur.), *36. seminar slovenskega jezika, literature in kulture* (str. 7–19). Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete.

Kržišnik, E. (2007): »Metafore«, v katerih govorimo v slovenščini frazeološko – konceptualnometaforična analiza frazemov govorjenja. V E. Kržišnik in W. Eismann (ur.), *Frazeologija v jezikoslovju in drugih vedah* (str. 183–205). Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko.

Lakoff, G. (1987). Cognitive models and prototype theory. V U. Neisser (ur.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (str. 63–100). Cambridge University Press.

Lakoff, G. (1993). The contemporary theory of metaphor. V A. Ortony (ur.), *Metaphor and thought* (str. 202–251). Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Lakoff, G., & Turner, M. (1989). *More than Cool Reason: A Field Guide to Poetic Metaphor*. The University of Chicago Press.

Langacker, R. (1990). *Concept, Image, and Symbol: The Cognitive Basis of Grammar.* Berlin, New York: De Gruyter Mouton.

Littlemore, J. (2020). *Metaphors in the Mind: Sources of Variation in Embodied Metaphor*. Cambridge: Cambridge University Press

Lu, X., & Pin Yun Wang, B. (2017). Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation, 51*(3), 663–694.

Panther, K-U., & Radden, G. (1999). The potentiality for actuality metonymy in English and Hungarian V K. U. Panther in G. Radden (ur.), *Metonymy in Language and Thought* (str. 333–357). Amsterdam: John Benjamins.

Pavlas, D., Vrabeľ, O., & Kozmér, J. (2018). Applying MIPVU Metaphor Identification Procedure on Czech. V S. Kübler in H. Zinsmeister (ur.), *Proceedings of the Workshop on Annotation in Digital Humanities co-located with ESSLLI 2018* (str. 41–46). Sofia, Bulgaria.

Philips, G. (2012). Locating metaphor candidates in specialized corpora using raw frequency and keyword lists. V F. MacArthur, J. Oncins-Martínez, M. Sánchez-García in A. PiquerPíriz (ur.), *Metaphor in use: context, culture, and communication* (str. 85–106). Amsterdam: John Benjamins.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*(1), 1–39.

Radden, G., & Kövecses, Z. (1999). Toward a theory of metonymy. V K.-U. Panther in G. Radden (ur.), *Metonymy in language and thought* (str. 17–60). Amsterdam: John Benjamins.

Rosiński, M., & Marhula, J. (2015). MIPVU in Polish: On Translating the Method. *RaAM Seminar 2015*.

Ruiz de Mendoza, F., & Galera, A. (2014). *Cognitive modeling. A Linguistic Perspective*. Amsterdam: John Benjamins.

Semino, E., & Masci, M. (1996). Politics is football: metaphor in the discourse of Silvio Berlusconi in Italy. *Discourse and Society, 7*(2), 243–269.

Semino, E. (2017). Corpus linguistics and metaphor. V Dancygier, B. (ur.), *The Cambridge Handbook of Cognitive Linguistics* (str. 463–476). Cambridge: Cambridge University Press.

Skorczynska, H., & Ahrens, K. (2015). A corpus-based study of metaphor signaling variation in three genres. Text & Talk. *An Interdisciplinary Journal of Language Discourse Communication Studies, 35*(3), 359–381.

Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja. Dostopno prek www.fran.si

Snoj, J. (2010). *Metafora v leksikalnem sistemu*. Ljubljana: Inštitut za slovenski jezik Frana Ramovša.

Stallar, D. (1993). Two Kinds Of Metonymy. V *31st Annual Meeting of the Association for Computational Linguistics* (str. 87–94). Association for Computational Linguistics: Columbus, Ohio.

Steen, G. J., Dorst, A. G., Herrmann, B. J., Kall, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification. From MIP to MIPVU*. Amsterdam: John Benjamins.

Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. V A. Stefanowitsch in S. T. Gries (ur.), *Corpus-Based Approaches to Metaphor and Metonymy* (str. 1–17). Berlin: De Gruyter Mouton.

Tummers, J., Heylen, K. in Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory, 1*(2), 225–261.

Urbonaitė, J. (2016). Metaphor identification procedure MIPVU: an attempt to apply it to Lithuanian. *Taikomoji kalbotyra [Applied Linguistics], 7*, 1–25.

Verdonik, D. (2015). Govorjeni proti pisnemu ali katera leksika je »tipično govorjena«. V V. Gorjanc, P. Gantar, I. Kosem, S. Krek, M. Bratanić, W. Brownie in V. Cvrček (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 392–405). Ljubljana: Znanstvena založba Filozofske fakultete.

Vidovič Muha, A. (2000). *Slovensko leksikalno pomenoslovje*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete.

Warren, B. (2002). An alternative account of the interpretation of referential metonymy and metaphor. V R. Dirven, & R. Pörings (ur.), *Metaphor and Metonymy in Comparison and Contrast* (str. 113–133). Berlin: De Gruyter Mouton.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua, 115*, 1481–1496.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Stabej, M., & Krek, S. (2009). Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. V M. Stabej (ur.), *Infrastruktura slovenščine in slovenistike* (str. 437–442). Ljubljana: Znanstvena založba Filozofske fakultete.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2021). *Spoken corpus Gos 1.1*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1438.

Corpus Approaches to Metaphor and Metonymy Identification: The Case of Metonymy in g-KOMET

We are not always able to express everything we think directly, so we use various linguistic-cognitive operations, including metaphors and metonymies. Metaphorical and metonymical expressions are difficult to extract from a corpus that is not specifically annotated for the purposes of figurative language research. Among the attempts to create a corpus of metaphors in Slovene are the corpus of metaphors KOMET 1.0 and its continuation with transcriptions of the spoken language, corpus g-KOMET (corpus of metaphorical and metonymic expressions in the spoken language). The article presents the most common methods of extracting metaphorical and metonymic expressions from linguistic corpora, and illustrates an attempt to systematize some of the most common metonymic transfers in the Slovenian spoken language using the corpus g-KOMET.

Keywords: conceptual metaphor, conceptual metonymy, metonymic pattern, corpus annotation, spoken language

Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus

Andrejka ŽEJN
ZRC SAZU

Mojca ŠORLI
ZRC SAZU

This paper is a follow-up and elaboration of the paper published in the JTDH 2022 Conference Proceedings on manual semantic annotation of named entities based on a proposed set of annotations for a corpus of modernist literary texts. We first briefly describe the corpus and introduce the annotation scheme, then focus on the results of additional analyses, and conclude with further challenges and issues we identified with respect to established NER systems and practices of related projects. Overall, we identify several categories of proper names, foreign language elements, and bibliographic citations, but focus here on the challenges of annotating names of literary characters and place names, and provide examples of the results of preliminary analyses of these entities in the corpus.

Keywords: modernism, named entities, corpus analysis, Slovenian literature, *Tribuna*, *Problemi*, 1968

Žejn, A., Šorli, M.: Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus. Slovenščina 2.0, 11(1): 118–137.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.118-137

1 Introduction

In literary studies, named entities are most closely associated with research on literary characters and settings. A comprehensive picture of the way characters are named in literature and how place names are used in the text was obtained beyond the renderings of automatic recognition of "Named Entities" (hereafter NEs) by manually annotating these entities in literary texts, first by analyzing the annotation process, and then the data obtained from the annotated corpus itself. In this paper we report on an attempt to identify and annotate three groups of NEs in the "Corpus of 1968 Slovenian literature May68 2.0" (the May68 Corpus, for short)[1] (Juvan et al., 2022), expanding on the analyses first presented as a conference submission (see Šorli and Žejn, 2022). Section 1 provides a brief description of the corpus and the annotation procedure, followed by Section 2 that focuses on the preliminary results of an extended analysis of personal and place names. In Section 3, we discuss the potential for future annotation tasks and improvements to the annotation scheme, as well as the optimal application of the results.

In view of the significance for the Digital Humanities of controlling a large number of texts and their vertical reading, where patterns become visible that cannot be detected with the naked eye or traditional close reading, the corpus size is often seen as a key factor. At the same time, large volumes of text require automation of corpus processing for quantitative analysis, which includes different levels of (linguistic) annotation in the first phase and allows for additional levels of semantic annotation in later phases that enrich the text with metadata. In the presented approach, however, the annotation task is performed on a small, specialized corpus that is easier to control and allows for manual annotation. The identified and manually annotated NEs are distinguished based on semantic criteria, so we consider this an example of semantic annotation.

Together with the theoretical concept, the selection of annotation material, and the definition of guidelines for the annotation process (Pagel et al., 2020), the annotation scheme presented here constitutes a model for the extended annotation of NEs in modernist periodicals,

1 Corpus of 1968 Slovenian literature May68 2.0: http://hdl.handle.net/11356/1491

certain segments of which can be applied to other corpora of literary texts. We focus on both the identified inaccuracies and the advantages of manual annotation of selected groups of NEs in our specialized corpus (for more on the theoretical background and history of automated and manual annotation of NEs, including the different approaches, see Šorli and Žejn 2022: 188-189).

1.1 The May68 Corpus of Slovenian modernist literary texts – corpus description

The May68 Corpus is the result of a project on the literature of the avant-garde and modernism in the period of the worldwide student movement associated with May 1968, whose activities are also reflected in the transformation of literature. The corpus consists of Slovenian modernist literary texts from the late 1960s to the early 1970s and was created according to special criteria defined for the purposes of corpus and stylistic research of modernist texts. The student journals *Tribuna* and *Problemi*, from which the texts for the corpus were selected, played an important role in the theoretical and literary-artistic innovations of the Slovenian student movement. The May68 Corpus 1.0 contains 1,521 texts by 198 known authors published between 1964 and 1972 in the Slovenian periodicals *Tribuna*, *Problemi* and *Problemi.Literatura*. The version May68 Corpus 2.0, which has been further edited and corrected (metadata), contains 647 additional texts from *Tribuna* and *Problemi*. The texts contain complete bibliographic data, are classified by text and language type, degree of presence of non-standard Slovenian, foreign languages, modernism, and visual elements. Author details, i.e., gender and year of birth, are included with the texts. The presence of visual elements is also marked in the corpus (48 texts).[2]

1.2 Annotation procedure

Following the automatic pre-processing (the automatic linguistic annotation included lemmas, morphosyntactic descriptions from MULTEXT-East and morphological features and syntactic annotations from Universal Dependencies) of the May68 Corpus, further manual

2 A detailed description of the corpus was provided in Juvan et al. (2021: 60–64).

annotation was performed to capture more complex linguistic (semantic) phenomena and to provide a more sophisticated annotation model for proper nouns given the recurring representational problems. Manually annotated are (foreign) language variations and registers, but the focus of the present article is on the NEs denoting persons, including cited authors (sources), geographical locations, i.e., various real and fictitious place names, organizations, and miscellaneous entities.

The annotation was implemented using the WebAnno tool (Eckart de Castilho et al., 2016). WebAnno allows annotation of one sentence at a time, which is a disadvantage for longer instances of text marked by the use of foreign language(s). Each annotation round was curated by two curators. However, reiterative annotation was not foreseen, since the primary goal at this stage was not to improve automatic annotation, but to manually annotate the specialized corpus for optimal corpus analysis and stylistic studies.

The following sections and subsections introduce the types and categories of NEs, including the dilemmas encountered in the process of annotation and the rationale behind the decisions made. With a somewhat narrower notion of NER, for the purposes of this paper we are mainly talking about categories of"proper names (personal and place names)" rather than "named entities".

1.2.1 Named entity categories and resolution

At this first stage, a model for identifying and annotating the selected NEs was put in place, with a second stage of the project envisaged in which the texts will be annotated for the use of metaphor. We also discuss the practical treatment of proper names for the purposes of corpus linguistic and stylistic research, in the hope of improving the reliability of results and NLP models. As pointed out in Beck et al. (2020), representational problems in linguistic annotation arise from five different sources (ibid., 61): (i) Ambiguity is an inherent property of the data. (ii) Variation is also part of the data and can occur, for example, in different documents. (iii) Uncertainty is caused by lack of knowledge or information by the annotator. (iv) Errors may be found in the annotations. (v) Bias is a property of the annotation system as

a whole. We list a number of relevant annotated categories, their specific character, and representational problems associated with them. We focused on some open challenges in the annotation of NEs, and in particular problems related to the functional aspects of personal proper names and place names.

There is no universally accepted taxonomy for NEs, except for some coarse-grained categories (people, places, organizations). Since we are interested in a semantically oriented annotation and prefer more informative (fine-grained) categories, we opted for a three-level NE classification as shown in Table 1 (cf. Sevščíková et al., 2007). The first level in our annotation scheme corresponds to the three basic groups: 1. Proper names, 2. Foreign language and register variations, and 3. Cited authors. These groups are labelled as 1. NAME, 2. FOREIGN, 3. BIBLIO, respectively, with the first two further subdivided. The second and third levels provide a more detailed semantic classification. NE resolution is primarily linked to the category PER, which is labelled in terms of whether the character is text/plot-internal or -external. The NAME group includes the following types and subtypes:

- Person (PER), including the adjective derived from a person's name, is subdivided into fictional literary characters (PER- LIT), real characters referring to existing and historical or mythological persons or beings (PER-REAL), literary characters bearing a descriptive name (PER-DES), and members of national and social groups (PER-GROUP).
- Geographical location (GEO) comprises localities in Slovenia (GEO-SI), the former Yugoslavia (GEO- YU), Europe (GEO-EU), and in other countries, including fictitious place names (GEO-ZZ).
- Organizations and institutions (ORG).
- Miscellaneous (XXX).

Once the annotation process was completed, the labels were converted to TEI encoding in WebAnno.5 Following the conversion all proper names (personal names, place names, names of organizations) were labelled with <name>, then divided into types with @person, @ geo, @misc, @personGrp, and @org attributes, three subtypes for literary characters (@literary, @descriptive, @real), and for geographical names (@SI, @EU, @ZZ and @YU).

PERSON (PER) type is divided into PER-LIT, PER-REAL, PER-DES and PER-GRP. While the first three are categorized as subtypes of the same type, PER-GRP is defined as an independent type. As shown in Table 1, the most important NE resolution consists in the subdivision of the PER type (within the NAME group) into real, e.g., historical or real-life, persons appearing in the text, and fictional characters, each of which, however, is further specified according to semantic criteria. We have classified both historical and mythological names as non-fictional, that is, as PER-REAL, unlike, for example the Netscape project where variants of "legend", "mythological" and "fictional" are all subsumed under "fictitious" (cf. de Does et al. 2017, p. 364). The PER type includes names of people (and provisionally for pets), nicknames, pseudonyms, members of national and social groups.

Table 1: *The main categories of the May68 annotation scheme (WebAnno)*

Group	Type	Subtype	Description
NAME	**PERSON – PER**	**PER-REAL**	Real: Characters referring to real, i.e. existing and historical or mythological persons or beings, e.g. *Greta Garbo*, *Charlie Brown*, *hlapec Jernej*, *Maruška*
		PER-LIT	Literary: Fictional literary characters, e.g. *Ančika*, *Zobec*, *Janko*, *Polona*
		PER-DES	Descriptive: Literary characters that carry a descriptive name (e.g., *dolgolasec*, Eng. the long-haired guy)
		PER-GRP	Group: Members of national and social groups, e.g. *Kranjci*, *Slovenec, Američan*
	GEO	**GEO-SI**	Slovenia, e.g. *Ljubljana*, *slap Savica*, *Crngrob*
		GEO-YU	Former Yugoslavia (except for Slovenia), e.g. *Zagreb*, *Dajla*
		GEO-EU	Europe, e.g. *Frankfurt*, *Minsk*, *Vltava*
		GEO-ZZ	Other, e.g. *Peking*, *Kuba*, *Indija Koromandija*
	ORG	**–**	Names of organizations, institutions (e. g. *Klub nepismenih*, *Slovenska matica*, *Državna varnost*)
	XXX	**–**	Common proper nouns, including titles of books and other art works, artefacts, etc., e.g. *Rdeča kapica*, *Empire State Building*

Group	Type	Subtype	Description
FOREIGN	HBS	–	Serbo-Croatian
	EN	–	English
	DE	–	German
	FR	–	French
	IT	–	Italian
	LA	–	Latin
	XX	–	Other
	DIALECT	–	Dialect
	VERNACULAR	–	Vernacular
	SLANG	–	Slang
BIIBLIO	**–**	**–**	Quoted authors (Sources)

PER-REAL denotes both real, i.e. existing, persons and historical or mythological figures that are basically identifiable in encyclopaedic sources such as online lexicons of proper names, Wikipedia and the like. URL is an additional attribute of the NAME group and is given as a relevant source of information, such as a website, for a group of people appearing in the literary text. The assignment of an URL depends on the context or on extra-linguistic knowledge; we linked names to web resources only when a (personal) name was not assumed to be part of today's common cultural knowledge (e.g. Giorgio Albertazzi, Italian actor and director, or Dave Brubeck, American jazz pianist and composer), if a person can be assumed to be part of common (cultural) knowledge (Descartes, Nietzsche), we chose not to enrich the corpus with encyclopaedic data. All standard personal proper names are labelled as NAME and assigned to one of the closed subtypes.

The label PER-GRP with no subtype is assigned to members of a particular social group, most often nationality (*Slovenec*, *Nemec*), regional (*Kranjci*, *Štajerci*) or family (*Novakovi*) identity, but also smaller social groups defined on the basis of occupational or other criteria (*esesovec*, *vaščan*, *vojak*).

Of the categories introduced specifically for the purposes of the May68 Corpus, NAME / PER-DES proved, as expected, to be the most challenging subcategory. This group of names seems to be used to describe the personality and/or physical appearance of literary characters

(*govorancar*, *starec*, *brkati*), as well as their occupation (*načelnik*, *inšpektor*) or social status (*neznanec*).

Adjectives derived from personal proper nouns are annotated as the corresponding proper nouns, e.g., *Dimitrijev* (Dimitrij's), *Prešernov* (Prešeren's), *dolgolaščev* (pertaining to the long-haired one). Their derived character is revealed by morpho-syntactic tagging.

Given their statistical importance in the context of NER, the same annotation rules apply here as for characters in plays when they do not require special treatment with respect to their function. The labelling of personal names in plays depends on the status and/or function of the name. Names of individual characters that merely announce an individual character's speech, and thus his/her lines of dialogue, have not been annotated, while names in descriptions of their physical actions or behaviour are treated as ordinary proper names on the model of "sb does sth", etc. (*Pandolfo se oglеduje v zrcalu* / Pandolfo looks at himself in the mirror).

Compared to the categories of personal names, significantly fewer dilemmas occurred in the categorization and labelling of place names. Individual unresolved cases (e.g., fictitious places, names referring to localities or objects in space) were assigned to the category "Other".

Geographical names in the broadest sense spanning from names of streets (*84. ulica*), rivers (*Drava*), mountains (*Učka*, *Himalaja*), cities (*Piran*, *Rim*, *Dunaj*) to those of countries (*Slovenija*, *Japonska*) and continents (*Evropa*, *Južna Amerika*), but also abstract (e.g. space-related) or fictitious (text- or plot-internal) (planet *Tuku-Luka*) place names, were taken into account in the manual annotation. Adjectives derived from geographical names were also labelled following the scheme for personal proper names. Both place names and the derived adjectives were classified into four categories according to the wider geographical location: place names in Slovenia, in the former Yugoslavia (with the exception of Slovenia), in Europe and the rest.

Even before the advent of corpus linguistic research, which arises from methods that can be applied to larger literary corpora, including corpus stylistics, analyses of geographical names in literary studies took place in two fields: the first field is defined as the geography

of literature, or the study of the spatiality of literary works, which explores space at the level of textuality. Another field is so-called literary geography, which deals with the study of the place-bound nature of writing, publishing and reading and whose results are often presented in literary atlases (Perenič, 2012a, p. 259–260; Gregory et al., 2015, p. 6–8).[3] In recent decades, these two fields have further evolved within DH research, i.e., with distant reading approaches. The potential of entirely new modes and practices for literary scholarship has been suggested, with the aim of complementing existing work with the potential offered by large corpora of literary texts and the development of corpus linguistic and corpus stylistic methods, including the possibilities of data extraction from large machine-readable corpora (Gregory et al., 2015, p. 6–8).

The comparative study of the usage (patterns) and function of place names in literary works that includes both quantitative and qualitative aspects of geographical entities in literary works is called comparative literary onomastics (de Does et al. 2017, pp. 361–362). In the narratological approach of "distant reading", analyses of place names are part of broader research on the relationship between characters, plot, time and space, similar to the analyses within the framework of Text World Theory based on Bakhtin's concept of the chronotope (cf. Šorli and Žejn, 2021, p. 188 and the literature cited there) or the research within the "digital narratology of space" that established the study of space frames based on Lotman's concept of spatial semantics (cf. Viehhauser, 2020, p. 381). Modern quantitative research also includes other aspects in the analysis of space besides the classical narratological categories, such as the analysis of the connections between emotions and space (cf. Grisot and Hermann, 2022). As to the question of literary setting, it follows from the aforementioned types of research that the analysis of geographical entities is only one part of the necessary analysis. At the same time, the data that can be extracted from large corpora allow insights into literature that go beyond the limits of studying a limited corpus of selected "representative" texts.

3 For a survey of relevant research see Hladnik (2012) and, for more recent Slovenian studies, *Prostor slovenske književnosti* (cf. Perenič, 2012b).

2 Statistical analyses

2.1 Literary characters

From the previous analyses it appears that the May68 Corpus is clearly dominated by literary names, PER-LIT (68%), while the other two categories appear in relatively similar proportions: descriptive names, PER-DES (18%), and characters from the non-literary world, PER-REAL (14%). Moreover, a clear preponderance of male characters was found (on average about 80:20 in favour of male characters), and the analysis by the gender of the authors showed that this ratio was somewhat more balanced in female authors, with the exception of real-life characters, where the ratio is independent of gender, most likely due to the real and undisputed presence and roles of men and women in social and cultural history (for more details, see Šorli and Žejn, 2022, p. 193–194).

The annotated May68 Corpus contains texts from three basic literary genres and also enables searching by and comparing among these three basic categories. The following section presents some results of the analysis of the quantitative and proportional distribution of the three types of character names (literary, descriptive, and real names) in drama (see Figure 1), poetry (see Figure 2) and prose (see Figure 3).[4]

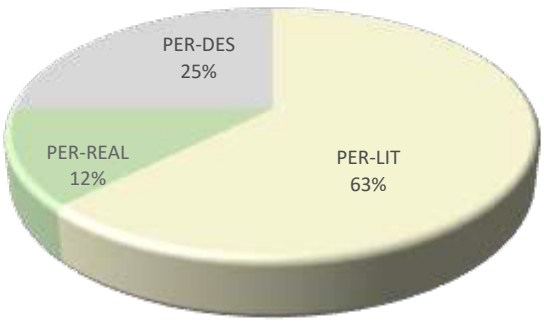


Figure 1: Distribution of character naming types in drama.

4 According to the number of occurrences, poetry accounts for 13.56%, prose for 66.09%, and drama for 18.56%. The remaining 1.77% represent hybrid genres, which are not considered in the analyses both because of their extremely low presence and because of their genre specificity.

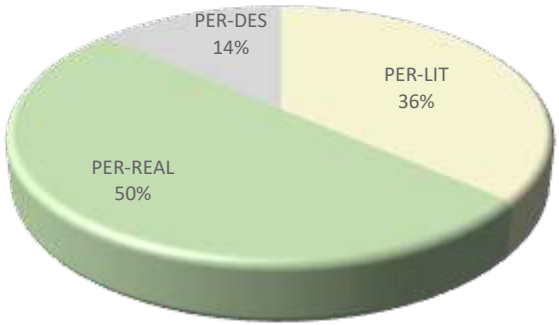


Figure 2: Distribution of character naming types in poetry.

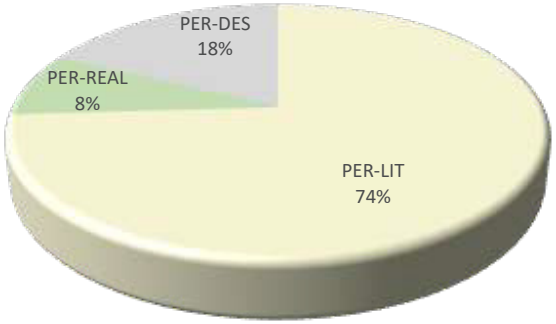


Figure 3: Distribution of character naming types in prose.

A comparison of the three charts shows significant differences in the distribution of the different types of character naming. Literary names (PER-LIT) are more prevalent in drama when compared to descriptive names (PER-DES) and real names (PER-REAL) (63% of all namings), and in prose (nearly 75% of all namings), while in poetry the proportion of literary names comes in third place, at 36%. The results show that in poetry there are fewer direct namings of literary characters and that, in general, these are not prominent. The proportion of descriptive names is largest in dramatic texts (25%), smaller in prose (18%), and even smaller in poetry (14%). It can be concluded that the higher proportion of descriptive names in drama is related to the fact that such texts are primarily intended to be performed on stage over two or three hours, and the descriptive names of characters are used as a means of "economic" characterization. In prose, because of the larger size of the text and the greater likelihood that descriptive nomenclature will take hold

in the text, it is more likely that a particular characteristic of a person, such as a physical trait, an occupation or social status, etc., will serve the function of a proper name.

The relationship between descriptive names in prose and poetry is also characterized by the fact that the first ten descriptive names by frequency (see Table 2) in drama are almost exclusively (with one exception) names that refer to an occupation and/or social status (e.g. chief, principal, mayor); in prose, almost half of the first ten designations indicate a particular physical characteristic of the character (e.g., one-armed, long-haired, old) – such designations are effectively replaced in dramatic texts by descriptions and instructions in the didaskalia or dramatic performance.

Table 2: *The first ten descriptive names in drama and prose by frequency*

DRAMA		**PROSE**	
Lemma	Frequency	Lemma	Frequency
načelnik	38	*senzal*	126
ravnatelj	27	*črni mož*	85
župan	21	*enoroki*	46
gospod šef	19	*inšpektor*	45
novi načelnik	17	*dolgolasec*	42
tovariš župan	16	*Zobčev*[5]	38
taščica	16	*Tomažev*	37
umetnik	14	*stotnik*	37
gospod namestnik	12	*kapitan*	35
bivši načelnik	12	*bela žena*	35
pisar	11	*stari*	31

The list of descriptive names in poetry shows that these descriptive names were annotated in only seven texts, and that a particular type of descriptive name predominates which is not generally a feature of drama or poetry.

In poetry, a large proportion of real-world persons is conspicuous, constituting the majority or even half of the names, while in drama and prose this category of proper names occupies the smallest proportion:

5 *Zobčev* and *Tomažev* are cases where women are named by their (husband's) second name.

12% in drama and only 8% in prose. These results could be an indication of a high degree of referentiality and intertextuality of the poetry in the corpus or of modernist poetry in general.

In the following, we present some analyses of the annotated place names, in accordance with the categories established for manual annotation. The results ensuing from the analysis by these four categories for the entire annotated corpus are shown in Figure 4. This shows that the largest proportion of place names is related to Slovenia (37%), followed by (the rest of) Europe (28%) and countries beyond Europe (25%), with an unexpectedly modest proportion of geographical locations classified in the territory of the former Yugoslavia (only 10%).[6]

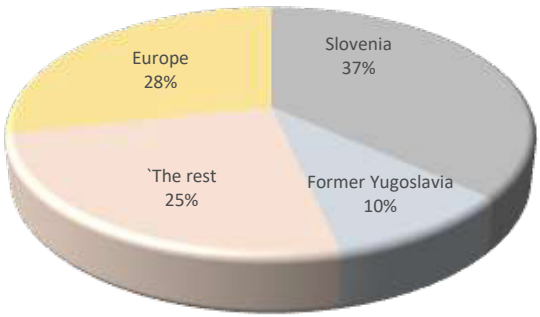


Figure 4: Place names according to the division into four major geographical units.

Similar to character names, we show the results below according to the percentage of geographical locations within each literary genre represented in the corpus.

6 Figure 1 shows data based on the number of occurrences, which show a more relevant picture, since recurring lemmas, in contrast with one or several mentions of a geographical location, are also an indicator of greater importance or presence in the text. The results by the frequency of lemmas show slightly different ratios: locations in Slovenia, Europe and the rest are almost equal each representing a little less than a third (Europe 30%, Slovenia and the rest 29% each); locations in the former Yugoslavia represent 12%. The analysis by the number of occurrences compared to the number of lemmas therefore shows a significantly greater role of geographical locations in Slovenia and the fact that locations outside Europe (the rest) are mostly just brief mentions and not so much actual places of action.

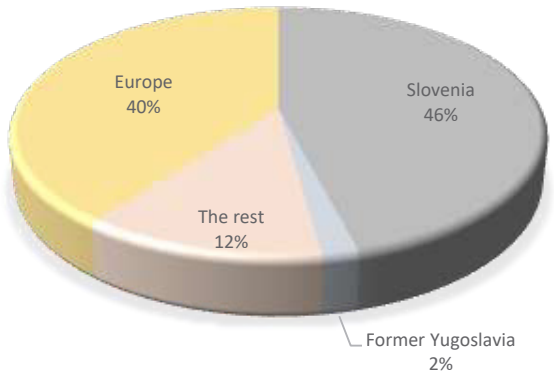


Figure 5: Proportional shares by classification in broader geographical units – Drama.

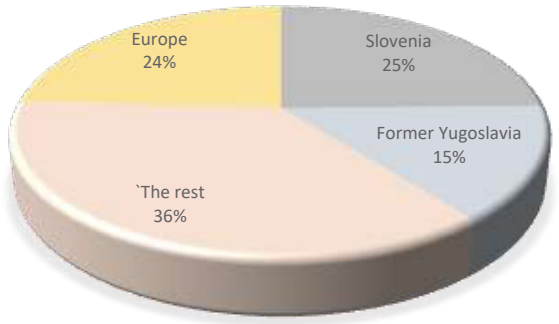


Figure 6: Proportional shares by classification in larger geographical units – Poetry.

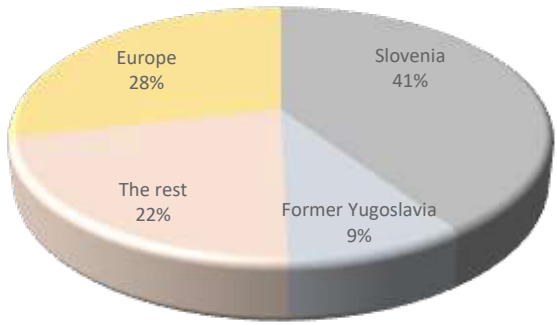


Figure 7: Proportional shares by classification in broader geographical units – Prose.

In the dramatic texts (see Figure 5), a disproportionate share of places in Europe (40%) and Slovenia (46%) is noticeable compared to poetry and prose, as well as a small proportion of places outside Europe (12%) and in the former Yugoslavia (a barely detectable 2%). The proportions within poetry (see Figure 6) are relatively even: places in Slovenia and Europe account for about a quarter, other places for slightly more (36%), and places in Former Yugoslavia for the least (15%). The distribution according to the general geographical classification in the prose (see Figure 7) corresponds most closely to the results for the entire corpus: the largest share is accounted for by geographical places in Slovenia (41%), followed by places in Europe (28%), other places (22%), and only 9% for places in the former Yugoslavia.

Since the number of geographical entities in the corpus is much smaller compared to personal names, the results in this segment are less representative and less likely to be generalizable to modernist literature or literature in general. Nonetheless, they suggest that there are some differences in the selection and listing of geographical locations across genres.

3 Potential benefits of corpus enlargement, additional annotation tasks, and further research

3.1 Conclusions and open issues

The main goal of our annotation task was to provide an adequate representation of a specific set of semantic data, i.e. named entities, and to fully exploit the potential of this type of corpus linguistic data in the context of future literary and linguistic analyses. To this end, we implemented a three-level annotation scheme. The preliminary results and additional analyses presented in this paper provide an argument for annotating the remaining part of the May68 Corpus, possibly with some adjustments to the scheme based on the experience of previous work. Accordingly, in the next phases of annotation we plan to improve the segments that show the lowest degree of consistency and annotator agreement, such as common nouns that serve the referential function of proper nouns and appear to act as a representational

continuum. We have yet to figure out how best to incorporate the various instances of descriptive names (PER-DES) into the annotation scheme, but they are certainly worth considering as a special (sub) category of the NAME group.

Compared to the categories of personal names, significantly fewer dilemmas occurred in the classification and labelling of place names. Individual unresolved cases (e.g., fictitious places, names referring to localities or objects in outer space) were assigned to the category "Other". Due to the high percentage of geographical entities labelled "Other", a further subdivision (in addition to Slovenia, Former Yugoslavia and Europe) of the wider geographical location must be proposed on the basis of the qualitative analysis of the results.

We conclude, on the basis of the high variation in referential expressions, that in potential future projects an additional step should be linking the different names of the same character (so-called "nesting"), the same applies to geographical entities, where several spelling variants occur (e.g., *Švica*, *švajc* = Switzerland).

Some NER projects report on the automatic linking of proper names to entries in Wikipedia (cf. de Does et al. 2017), which assists in the named entity resolution to distinguish between plot-internal and plot-external names. As shown in the introduction, we linked names to web resources only when a (personal) name was not assumed to be part of today's common cultural knowledge. This is another issue that needs to be resolved for future undertakings.

3.2 Future projects

The decision in favour of manual annotation of the May68 Corpus was based on the fact that this is a specialized corpus for which established automatic labelling of named entities would not predictably yield adequate and satisfactory results, as well as on experience from related research on named entities in various national literary corpora (cf. Stanković et al., 2019; Vala et al., 2015; Ketschik, 2020; Papay and Padó, 2020). As Won et al. (2018) have noted, using historical texts as an example, a single automatic tagging tool is not optimal for automatic tagging of place name and instead a clever combination of

multiple approaches is required. The fully annotated corpus will allow empirical testing of the differences between the results of manual and automatic annotation. Despite some adjustments, the three-level scheme for manual annotation is perhaps closest in granularity to the Janes-NER guidelines (CLARIN.SI) (cf. Zupan et al., 2017), whose categories are considered the standard for automatic annotation of Slovenian corpora. The results of this comparison could contribute to the optimization of tools for automatic labelling of named entities for corpora of literary texts.

Last but not least, the results of labelling, especially of character names and geographical entities, are crucial for the construction of an NE database. The database of the May68 Corpus could be the cornerstone for the compilation of a database of proper names in Slovenian literature of different literary genres, directions and periods, and the data on geographical entities could contribute to the research of spatiality of literary works on the level of textuality.

In a second stage of the project that is envisaged, the texts will also be annotated for the use of metaphor, which – financial means permitting – will result in a database of literary metaphor and metonymy. The goal of this additional annotation task will be to optimize the annotation procedure and apply the knowledge acquired about the use of metaphor in modernist literary texts for the purposes of future literary and linguistic analyses.

Acknowledgments

ARRS (Slovenian Research Agency) P6-0024 “Literarnozgodovinske, literarnoteoretične in metodološke raziskave [Research into literary history, literary theory and methodology].”

References

Beck, C. Booth, H., El-Assady, M., & Butt, M. (2020). Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In *The 14th Linguistic Annotation Workshop* (pp. 60–73). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.law-1.6.pdf

de Does, J., Depuydt, K., van Dalen-Oskam, K., & Marx, M. (2017). Namescape: Named Entity Recognition from a Literary Perspective. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 361–370). Ubiquity Press. Retrieved from http://www.jstor.org/stable/j.ctv3t5qjk.37

Eckart de Castilho, R., Mújdricza-Maydt, E., Muhie Yimam, S., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) (pp. 76–84). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/W16-4011.pdf

Gregory, I., Donaldson, C., Murrieta-Flores, P., & Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, *9*(1), 1–14. doi:10.3366/ijhac.2015.0135

Grisot, G., Herrmann, B. (2022). Emotions and space: an investigation of "urban" vs. "rural" emotional language in Swiss-German fiction around 1900. *Distant reading closing conference*. Accessed at https://www.distant-reading.net/events/conference-programme/

Hladnik, M. (2012). Prostor v slovenskih literarnovednih študijah: kritične izdaje klasikov. In U. Perenič (Ed.), *Prostor v literaturi in literatura v prostoru = Space in literature and literature in space* (pp. 271–282). Ljubljana: Slavistično društvo Slovenije. Retrieved from http://www.dlib.si/details/URN:NBN:SI:DOC-EFDJCFIF

Juvan, M., Šorli, M., & Žejn, A. (2021). Interpretiranje literature v zmanjšanem merilu: »Oddaljeno branje« korpusa »dolgega leta 1968«. *Jezik in slovstvo*, *66*(4), 55–76.

Juvan, M., Žejn, A., Šorli, M., Mandić, L., Tomažin, A., Jež, A., Balžalorsky Antić, v., & Erjavec, T. (2022). *Corpus of 1968 Slovenian literature Maj68 2.0, ZRC SAZU*, http://hdl.handle.net/11356/1430

Ketschik, N., Blessing, A., Murr, S., Overbeck, M., & Pichler, A. (2020). Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt* (pp. 203–236). Berlin, Boston: De Gruyter. Retrived from https://doi.org/10.1515/9783110693973-010

Pagel, J., Reiter, N., Rösiger, I., & Schulz, S. (2020). Annotation als flexibel einsetzbare Methode. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in*

der CRETA-Werkstatt (pp. 125–142). Berlin – Boston: De Gruyter. doi: 10.1515/9783110693973-010

Papay, S., & Padó, S. (2020). RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 835–841). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.104.pdf (1. 12. 2022)

Perenič, U. (2012a). Space in literature and literature in space. In U. Perenič (Ed.), *Space in literature and literature in space* (pp. 265–270). Ljubljana: Slavistično društvo Slovenije. Retrieved from: http://www.dlib.si/details/URN:NBN:SI:DOC-6P13WHOU

Perenič, U. (Ed.) (2012b). *Space in literature and literature in space*. Ljubljana: Slavistično društvo Slovenije. Retrieved from http://www.dlib.si/details/URN:NBN:SI:DOC-6P13WHOU

Stanković, R., Santos, D., Frontini, F., Erjavec, T., & Brando, C. (2019). Named Entity Recognition for Distant Reading in Several Languages. In G. Pálko (Ed.), *DH_Budapest_2019*. Budapest: ELTE. Retrieved from http://elte-dh.hu/dh_budapest_2019-abstract-booklet/

Ševščíková, M., Žabokrtský, Z., & Krůza, O. (2007). Named Entities in Czech: Annotating Data and Developing NE Tagger. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue: Proceedings of the 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007.* Berlin – Heidelberg: Springer-Verlag. Retrieved from https://ufal.mff.cuni.cz/~zabokrtsky/publications/papers/tsd07-namedent.pdf

Šorli, M., & Žejn, A. (2022). Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities 2022* (pp. 187–195) Ljubljana: Institute of Contemporary History. Retrieved from: https://www.sdjt.si/wp/dogodki/konference/jtdh-2022/zbornik/

Vala, H., Jurgens, D., Piper, A., & Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 769–774). Lisbon, Portugal: Association for Computational Linguistics.

Viehhauser, G. (2020). Zur Erkennung von Raum in narrativen Texten: Spatial frames und Raumsemantik als Modelle für eine digitale Narratologie des Raums. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte*

algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt (pp. 373–388). Berlin – Boston: De Gruyter. Retrieved from https://doi.org/10.1515/9783110693973-015

Won, M., Murrieta-Flores, P., & Martins B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities* 5. Retrieved from https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002

Zupan, K., Ljubešić, N., & Erjavec, T. (2017). *Annotation guidelines for Slovenian named entities: Janes-NER*. Technical report, Jožef Stefan Institute, September. Retrieved from https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf

Imenske entitete v modernističnih besedilih: ročno označevanje in analiza korpusa Maj68

V članku najprej predstavimo korpus Maj68, tj. korpus modernističnih literarnih besedil slovenskih avtorjev iz revij *Tribuna* in *Problemi* iz obdobja študentskega gibanja 1968. Korpus je bil avtomatsko oblikoskladenjsko označen, nato je sledila ročna semantična anotacija z namenom naprednejše analize korpusa. Cilj raziskave je bil, da v označeno gradivo zajamemo kompleksnejše semantične pojave in tem prilagodimo označevalni model, ki bi uspešno naslovil dileme označevanja literarnih besedil, in sicer dvoumnost, nejasnost in variantnost. Trinivojska označevalna shema ima tri osnovne kategorije, od katerih se prvi dve delita še nadalje: 1. lastna imena, 2. tuji jeziki in slovenske jezikovne varietete ter 3. bibliografske navedbe. Predstavljene so izbrane vsebinske analize imenskih entitet (imena likov in geografska imena) glede na tri temeljne literarne zvrsti. Rezultati analiz pokažejo določene razlike med zvrstmi, ki jih je mogoče interpretativno postaviti v širši literarni kontekst. V sklepih razmišljamo o možnostih izboljšave sheme, njene dodatne nadgradnje ter o potencialni nadgradnji rezultatov.

Ključne besede: modernizem, imenske entitete, korpusna stilistika, slovenska literatura, *Tribuna*, *Problemi*, 1968

Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov

David BORDON
Filozofska fakulteta, Univerza v Ljubljani

Namen prispevka je predstaviti raziskavo preverjanja razumljivosti nerevidiranih strojno prevedenih spletnih besedil. Primarni udeleženci v raziskavi so bili splošni bralci in ne izurjeni prevajalci ali popravljalci strojnih prevodov. Gre za prvo tovrstno raziskavo, ki je bila izvedena za slovenski jezik. Cilj raziskave je bil preveriti, v kolikšni meri so nerevidirani strojni prevodi razumljivi splošnemu bralstvu, pri čemer sem se posvetil tudi vplivu besedilnega in slikovnega konteksta. Preverjal sem prevode prevajalnikov Google Translate in eTranslation. Raziskava je bila izvedena z anketo, v kateri so udeleženci odgovarjali na vprašanja, ki so preverjala razumevanje spremljajočega besedilnega segmenta, v katerem je bila napaka. Rezultati nudijo vpogled v trenutno stopnjo razvoja strojnih prevajalnikov, ne z vidika storilnosti pri njihovem popravljanju, ampak z vidika, koliko jih razume ciljno bralstvo. Na koncu članka nudim novo evalvacijo izvornih segmentov, ki sem jih v začetku leta 2023 ponovno prevedel, tokrat še s prevajalnikom DeepL.

Ključne besede: nerevidirani strojni prevodi, evalvacija strojnih prevajalnikov, razumljivost pri končnih bralcih, Google Translate, eTranslation, DeepL

Bordon, D.: Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov. Slovenščina 2.0, 11(1): 138–159.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.138-159

1 Uvod

Pričujoči razširjeni članek nadgrajuje objavo v zborniku konference JTDH 2022 (Bordon, 2022) in mu dodaja poglavje št 6, v katerem evalviram prevodne rešitve izvorne raziskave z nadgrajenimi prevajalniki v letu 2023. V članku obravnavam raziskavo razumljivosti strojno prevedenih spletnih besedil pri bralcih, ki ne vedo, da prebirajo strojne prevode. Uporabil sem naključno izbrana angleška spletna besedila, slovenske prevode pa sem pridobil z nevronskima strojnima prevajalnikoma Google Translate in eTranslation. Prevodi niso bili revidirani, saj sem želel replicirati okoliščine, v katerih bi jih dejansko lahko našli – na spletu, kjer so zaradi (za nekatere) dovolj visoke kakovosti in cenovne nepremagljivosti (namreč so brezplačni) vedno pogostejša[1], kar velja tudi za prevajalske vtičnike, ki so vgrajeni v sodobne brskalnike in aplikacije. Uporabniki se nasploh vedno več poslužujejo strojnega prevajanja (Vieira et al., 2022).

Vprašanje razumljivosti v taki obliki je postalo aktualno samo v zadnjem času, saj so starejši, statistični modeli prevajalnikov slovnično nekonsistentni in jezikovno okorni, sodobni nevronski prevajalniki pa proizvajajo tekoča besedila, ki so težje ločljiva od človeških, hkrati pa je že profesionalnim pregledovalcem prevodov težje ugotoviti, kje so storili napako (Donaj in Sepesy Maučec, 2018).

Te napake nastanejo predvsem zaradi težav pri razdvoumljanju večpomenskih besed in prevajanju besed, ki jih ni v podatkovni zbirki, s katero smo prevajalnik učili (Thi-Vinh et al., 2019, str. 207; Koehn in Knowles, 2017, str. 28, 31–33; Sennrich et al., 2016, str. 3). Kljub morebitnim posamičnim napačno prevedenim besedam pa lahko ljudje pomen razberemo iz sobesedila. Pri preverjanju razumljivosti sem v vseh primerih vključil še kontekst, saj se v stvarnosti bralci nikoli ne srečujejo z izoliranimi besedami, ampak z zaključenimi besedili, ker pa se osredotočam na spletno okolje, sem besedilnemu kontekstu dodal še slikovnega, ki je pogost element na sodobnih spletnih straneh.

Namen članka je predstaviti grobo oceno razumljivosti prevodov NMT-sistemov (ang. *Neural machine translation*) v času, ko so taka

1 Nekaj primerov za »krompirjeve kline« (potato wedges) na spletu. sl.veg-recipes; sl.hiloved; sl.eathealthyeatgreek; sl.gastromium; sl.atomiyme – strojno prevedene spletne strani za mnogo različnih jezikov.

besedila na spletu vedno pogostejša, pri čemer me zanima predvsem, kako slikovno gradivo v besedilnem kontekstu vpliva na rezultate. Tovrstna raziskava za slovenščino še ni bila izvedena.

2 Sorodne raziskave

Raziskav na področju razumevanja nerevidiranih strojnih prevodov pri končnih bralcih je razmeroma malo, saj je z vidika omejenosti na stroko in gospodarstvu bolj zanimive analize storilnosti pri popravljanju prevodov veliko več raziskav osredotočenih zgolj na prevajalce.

Na Univerzi v Gentu je bila v sklopu projekta ArisToCAT izvedena raziskava o razumevanju izmišljenih besed in samostalniških besednih zvez (Macken et al., 2019). Primeri, ki so bili iz angleščine v nizozemščino prevedeni s strojnima prevajalnikoma Google Translate in DeepL, so bili predstavljeni samostojno ali v kontekstu povedi, pri tem pa udeleženci niso imeli dostopa do izvirnega besedila. V povprečju so udeleženci ugotovili pravilen pomen izmišljene besede v 23 % primerov, ko je beseda bila predstavljena brez konteksta. Če ji je bil dodan kontekst, je bilo 41 % odgovorov pravilnih; v scenariju, ko je bila izmišljena beseda predstavljena v povedi in so udeleženci morali izbrati pravilen pomen, je bilo pravilnih odgovorov 56 %.

V sklopu istega projekta je bila izvedena še analiza bralnega razumevanja človeškega prevoda na eni in nepopravljenega strojnega prevoda na drugi strani. Človeški prevodi so bili ocenjeni bolje z vidika jasnosti podajanja informacij, z vidika končnega razumevanja pa je bila razlika manjša (Macken in Ghyselen, 2018).

Castilho in Guerberof Arenas (2018) sta izvedli primerjalno analizo bralnega razumevanja za statistični in nevronski model strojnega prevajalnika v primerjavi s človeškim izvirnikom. Glede na omejen vzorec (6 udeležencev) in nedoslednost rezultatov je ugotovitev, da sistemi-NMT izkazujejo najboljše rezultate, občasno še boljše kot angleški izvirnik, nedokončna.

Martindale in Carpuat (2018) sta v raziskavi obravnavali odziv bralcev na tekočnost in natančnost nevronskih strojnih prevodov, ob tem pa sta preverjali stopnjo zaupanja informacijam v besedilu. Ugotovili sta, da bralce zelo zmotijo prevodi, ki niso tekoči.

Izsledke potrjuje tudi Popović (2020). V njenem eksperimentu so bralci v 30 % primerov zaradi zavajajoče tekočnosti sprejeli popolnoma napačno informacijo, še 25 % dodatnih primerov pa je bilo skoraj popolnoma (narobe) razumljivih.

Na tem mestu velja omeniti, da so se nedavno začele pojavljati bolj eksperimentalne metode prevajanja, katerih značilnost je upoštevanje multimedijskega konteksta, denimo zvočnega ali slikovnega. Lala in Specia (2018) sta razvila model multimedijskega leksikalnega prevajanja, katerega namen je prevajanje dvoumnih večpomenskih besed s pomočjo slikovnega konteksta. Sulubacak et al. (2020) so predstavili sorodne raziskave, uporabne podatkovne zbirke in metode raziskovanja na področju multimedijskega strojnega prevajanja, ki so vezane na prevajanje z zvokom, sliko in videom. Med novejšimi raziskavami Liu (2021) ponuja nevronski model vizualno-tekstovnega enkodiranja in dekodiranja.

Pričakujemo lahko, da se bo to področje v bodoče še hitreje razvijalo, predvsem zaradi tehnološkega napredka v drugih panogah (prepoznavanje slik, sinteza govora, avtomatsko podnaslavljanje ipd.).

3 Metoda

Raziskava je bila zasnovana okrog vprašalnika, ki je vseboval primere štirih vrst napak v slovenskih strojnih prevodih splošnih angleških spletnih besedil. Preverjal sem prevajalnika Google Translate in eTranslation, pri čemer je bil vsak zastopan z 12 vprašanji. Distribucijo vrst napak opredelim v podpoglavju 3.3. Poseben pomen sem posvetil slikovnemu gradivu v sobesedilu.

3.1 Izbor besedil

Besedila sem zbiral glede na verjetnost, da bi se bralci z njimi lahko dejansko srečali na spletu. Analiza prevajalskega trga je pokazala, da večje prevajalske agencije popolnoma obvladujejo sektorje, ki nudijo največ dobička in hkrati zahtevajo človeško revizijo (tehnika, zdravstvo, pravo, finance ipd.) (Evropska komisija, 2020). V manj dobičkonosnih sektorjih, kjer človeška revizija ni tako pomembna, obstaja večja verjetnost objave nerevidiranih strojnih prevodov.

Pregleda tržnega deleža spletnih iskalnikov, ki jih uporabljamo v Sloveniji, je pokazal, da 96 % vseh uporabnikov spleta uporablja iskalnik Google.[2] Na osnovi najbolj iskanih pojmov v brskalniku[3] sem izločil spletišča, ki nimajo prevodnega potenciala (družbena omrežja, spletni portali v slovenščini, slovenski mediji). S tem sem prišel do končnega izbora besedilnih področij: spletno nakupovanje, turizem, elektronika, multimedija in videoigre, luksuzne storitve, moda, osebno zdravje (telesna vadba in prehrana).

3.2 Prevodi besedil

Pri preizkušanju strojnih prevajalnikov se je izkazalo, da Googlov prevajalnik nudi drugačne prevodne rešitve glede na to, kako besedilo naložimo v obdelavo. Če besedilo prevajamo v pogovornem oknu vmesnika ali v brskalniku prevedemo spletno stran kot celoto, so rezultati boljši kot tisti, ki jih dobimo s funkcijo prevajanja dokumenta. Od štirih različnih specializiranih domen, ki jih nudi eTranslation, je najboljše rezultate nudil prevajalnik za splošna besedila (General Text). Uporabil sem najboljše možne prevode – omenjeno domeno v eTranslation, v Googlu pa sem prevajal v pogovornem oknu.

Tabela 1: *Razlike v prevodih glede na način obdelave; Google Translate*

Prevod iz vnosnega polja oz. samodejni prevod strani	**Prevod, pridobljen s funkcijo »prevedi dokument«**	**Izvirnik**
Naj bo topla - mikrovalovna pečica ohranja hrano, kot so zelenjava, juhe, jedi, graviža, omake in sladice, topla in okusna v pečici, dokler niso pripravljene za postrežbo.	Naj bo toplo funkcijo - Mikrovalovna ohranja živila, kot so zelenjava, juhe, nerazporejenega d'oeuvres, gravies, omake in sladice toplo in okusno v pečice, dokler oni propravljeni, da služijo.	Keep Warm Feature Maintains Food Temperature Keeps foods like vegetables, soups, hors d'oeuvres, gravies, sauces and desserts warm and delicious in the oven until they're ready to serve.

Tabela 2: *Prevod enakega segmenta; eTranslation*

Prevod modela »General Text« prevajalnika eTranslation
Ohraniti toplo funkcijo - Microwave ohranja hrano, kot so zelenjava, juhe, predjed d'oeuvres, omake, omake in sladice tople in okusne v pečici, dokler niso pripravljeni za postrežbo

2 https://gs.statcounter.com/search-engine-market-share/all/slovenia
3 https://ahrefs.com/keyword-generator

3.3 Kategorizacija napak

Prevode sem analiziral in določil štiri kategorije najpogostejših napak, ki niso vezane na jezikovni sistem oz. predpis – raziskava zanemarja slovnico in se osredotoča izključno na leksikalne napake.

- **Neprevedena beseda;** v prevodu se pojavlja beseda v enaki obliki kot v izvirniku. Dopustil sem možnost spremembe začetnih ali končnih morfemov, če je prevajalnik besedo samo preoblikoval.[4]
  - Primer 22 – samodejno namakanje – *loosens madeže*.
- **Napaka pri razdvoumljanju večpomenske besede;** denotativni pomen večpomenske besede ali besedne zveze ne ustreza pomenu v izvirniku.
  - Primer 11 – torba za pedal za bas boben – *primer vključen* - »case included«.
- **Hujša pomenska napaka;** napaka, ki otežuje razumevanje celotnega besedila.
  - Primer 18 – naprava za hranjenje hišnih ljubljenčkov. *Baterije vstavimo ali odstranimo*.
- **Izmišljena beseda;** prevajalnik si zaradi kodiranja na enote manjše od besede (subword encoding) pri ponovnem sestavljanju besede v fazi dekodiranja »izmisli« besedo, ki ni v rabi in je denimo ni moč najti v referenčnih korpusih ali v drugih jezikovnih virih – t. i. »nevronščina«[5].
  - Primer 15 – dvopojasni Wi-Fi – *dvopasovni Wi-Fi*.

Vsi primeri so vizualno predstavljeni v poglavju 3.6.

Končni nabor je obsegal 12 primerov na vprašalnik, skupno 24. Distribucija števila primerov glede na vrsto napake je bila sledeča:

- neprevedena beseda: 2 primera na prevajalnik;
- napaka pri razdvoumljanju: 4 primeri na prevajalnik;
- hujša pomenska napaka: 3 primeri na prevajalnik;
- izmišljena beseda: 3 primeri na prevajalnik.

4 Denimo, prevod za rob zaslona (ang. *bezel*, je prevajalnik prevedel kot »bezela«).
5 https://www.alternator.science/sl/daljse/z-nevronscino-v-prihodnost/

3.4 Kontekst

Izbranim besedilom sem glede na inherentne lastnosti spletne pojavitve dodal kontekst. Kontekst je lahko bil več vrst:

- izključno besedilni,
- besedilni in slikovni; slika ne vpliva na razumevanje,
- besedilni in slikovni; slika vpliva na razumevanje,
- izbor ene izmed več predlaganih slik glede na to, kaj piše v besedilu.

Slikovni kontekst sem vključil pri besedilih, ob katerih so se na spletu pojavljale fotografije, ki so pri nekaterih primerih bile zgolj vizualni dodatek, pri drugih pa je bilo pravilno razumevanje besedila vezano na prepoznavanje pravilnega vizualnega elementa.

V svoji raziskavi besed nisem nikoli predstavil v izolaciji, kot so to denimo storili v raziskavi Macken in drugi (2019), saj to niso realne okoliščine – napake v objavljenih strojnih prevodih bodo vedno del nekega besedila. Besedil nisem popravljal, anketirancem sem jih dal v branje vključujoč vse slovnične in pomenske napake, kot bi jih lahko sami prebrali na spletu.

3.5 Oblikovanje vprašalnika, format odgovorov na vprašanja in udeleženci

Anketo sem ustvaril na platformi Google Forms, ki nudi podporo za prikaz slik in dober vmesnik za pregled in izvoz rezultatov. Pomembno je poudariti, da anketirancem nisem razkril, da bodo brali strojno prevedena besedila. Omenil sem, da bodo »prebrali več kratkih besedil, ki so napisana v nekoliko okorni slovenščini«.

Vrste odgovorov so bile omejene s funkcionalnostjo platforme Google Forms in niso sledile nobeni logični metodi; določil sem jih subjektivno glede na vsebino primera in vrsto napake. Gre za najbolj nezanesljivo spremenljivko v metodi, saj bi s formulacijo vprašanja lahko sugeriral pravilen odgovor, zanimalo pa me je predvsem to, če prihaja do večjega odstopanja glede na tip odgovora, denimo, če bi bili odgovori odprtega tipa, kjer anketiranci vnesejo svoj odgovor v prazno vnosno polje, bistveno slabši kot tisti, kjer izbirajo med štirimi predlaganimi odgovori. S tem bi lahko preveril konsistenco pravilnosti oz. odstopanja glede na vrsto odgovora.

Vključil sem tri tipe podajanja odgovorov na vprašanja o razumljivosti besedil:

- odgovor odprtega tipa; anketiranci vpišejo odgovor v vnosno polje,
- odgovor zaprtega tipa (A, B, C ali D),
- izbor z razlago (A ali B, zakaj?).

Vprašalnik sem delil na družbenih omrežjih Facebook in Instagram in znance pozval, naj ga posredujejo naprej svojcem in svojim znancem, če je le mogoče starejšim. Demografskih podatkov nisem zbiral, izjema je zgolj podatek, če se oseba, udeležena v anketi, ukvarja s prevajanjem, kar je ena izmed pomanjkljivosti raziskave. Glede na razmeroma majhen vzorec sodelujočih in morebiten efekt odmevne komore bi bilo vsekakor raziskavo potrebno nadgraditi in ponoviti na bolj naključnem in predvsem večjem vzorcu, toda glede na čas zbiranja odzivov, ki je sovpadal s prvo omejitvijo gibanja vezano na epidemijo Covid-19, nisem imel druge izbire.

Na vprašalnik sem prejel 120 odgovorov.

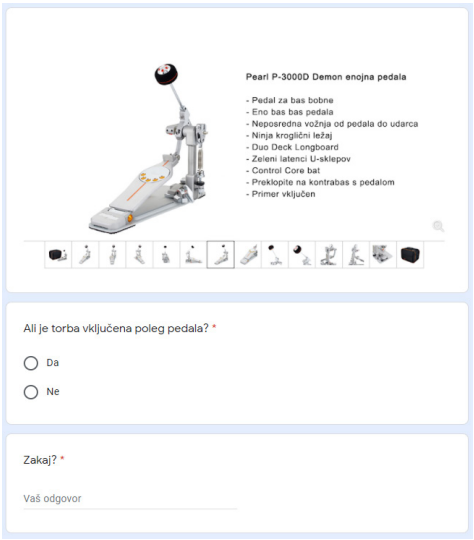
3.6 Primeri vprašanj in odgovorov v vprašalniku

V tem podpoglavju predstavljam par praktičnih primerov vprašanj iz ankete. Izbral sem tri različne primere; vsak ima različno vrsto odgovora, konteksta in napake. Celoten vprašalnik je dostopen na spletu[6]. Metodološki pristop je bil interpretativen – zavedam se mnogih morebitnih pomanjkljivosti, ki so vezane na to, kako so bila vprašanja izbrana, kako so formulirani potencialni odgovori in kako je razumevanje nekega besedila morda pogojeno s poznavanjem tehničnih vidikov področja, kateremu pripada. Bolje bi bilo vprašalnik standardizirati in uporabiti tako zasnovo ankete, ki obenem omogoča tudi boljšo statistično analizo.

Primer 11: torba za pedal za bas boben – *primer vključen* – »case included«.

- Vrsta napake: napaka pri razdvoumljanju večpomenske besede.
- Vrsta konteksta: besedilni in slikovni, kjer slika vpliva na razumevanje.
- Vrsta odgovora: izbor (A ali B) in razlaga.

6 https://forms.gle/NKnEHrcEgJa7Tydo9



Slika 1: Primer 11.

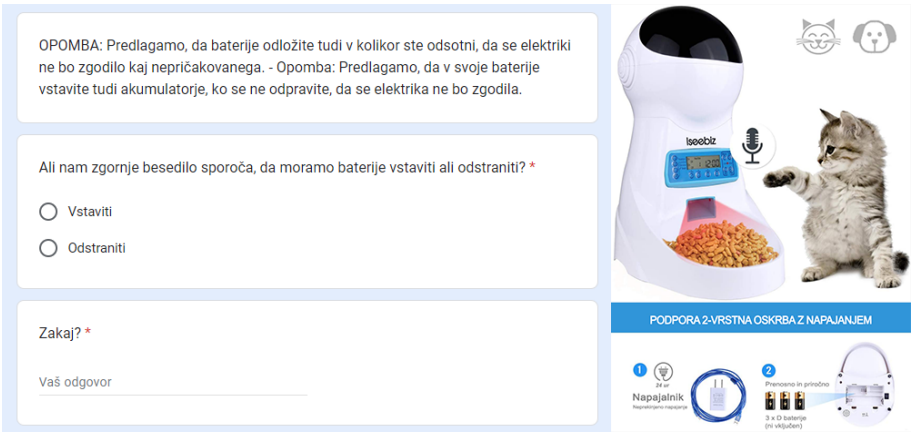
Primer 22: samodejno namakanje »loosens madeže«.

- Vrsta napake: neprevedena beseda.
- Vrsta konteksta: besedilni in slikovni, kjer slika ne vpliva na razumevanje.
- Vrsta odgovora: odprti tip.

Slika 2: Primer 22.

Primer 18: naprava za hranjenje hišnih ljubljenčkov. *Baterije vstavimo ali odstranimo*.

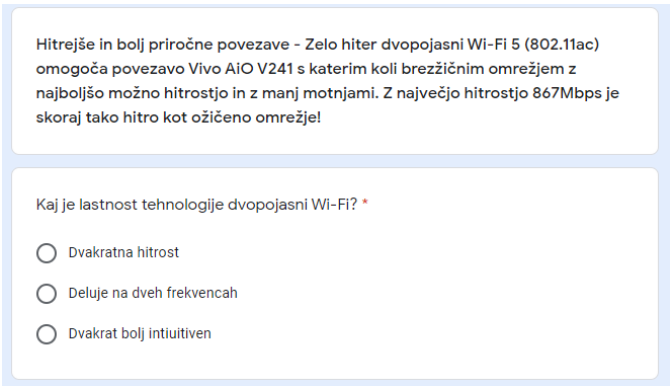
- Vrsta napake: hujša pomenska napaka.
- Vrsta konteksta: besedilni in slikovni, kjer slika ne vpliva na razumevanje.
- Vrsta odgovora: izbor (A ali B) in razlaga.



Slika 3: Primer 18.

Primer 15: dvopojasni Wi-Fi – *dvopasovni Wi-Fi*.

- Vrsta napake: izmišljena beseda.
- Vrsta konteksta: samo besedilni.
- Vrsta odgovora: zaprti tip.



Slika 4: Primer 15.

4 Rezultati

Rezultate predstavljam po naslednjih parametrih:
- splošno razumevanje,
- razumevanje glede na prevajalnik,
- razumevanje glede na tip napake,
- razumevanje glede na tip konteksta,
- razumevanje glede na tip odgovora.

4.1 Splošno razumevanje

Vprašalnik je obsegal 24 vprašanj, s 120 odzivi je bilo vseh možnih odgovorov 2.880. Vseh pravilnih odgovorov je bilo 1.697 oz. 58,96 %. Daljša razčlemba je na voljo v celotni raziskavi (Bordon, 2021).

4.2 Razumevanje glede na prevajalnik

Odgovori na vprašanja, vezana na prevajalnik Google Translate, so bili pravilni v 51,3 % primerov oz. 739 od 1.440 odgovorov. Prevajalnik eTranslation je pokazal boljše rezultate, delež pravilnih odgovorov je znašal 66,6 %.

4.3 Razumevanje glede na tip napake

V vprašalniku so bili vključeni štirje tipi različnih napak. V alinejah nizam tip napake in odstotek pravilnih odgovorov:
- izmišljena beseda: 48,5 %,
- neprevedena beseda: 64,8 %,
- napačno razdvoumljene večpomenske besede: 65,9 %,
- hujša pomenska napaka: 56,3 %.

4.4 Razumevanje glede na kontekst

V naslednjem segmentu predstavljam delež pravilnih odgovorov vezanih na kontekst:
- izključno besedilni: 60,4 %,
- besedilni in slikovni; slika ne vpliva na razumevanje: 44 %,
- besedilni in slikovni; slika vpliva na razumevanje: 69,8 %,

- izbor ene izmed več predlaganih slik glede na to, kaj piše v besedilu: 64,2 %.

4.5 Razumevanje glede na tip odgovora

V tem segmentu predstavljam rezultate glede na način izbora odgovora. Primarna funkcija te analize je preveriti konsistenco oz. morebitna odstopanja npr.; če so odgovori odprtega tipa, kjer anketiranci v prazno vnosno polje vnesejo poljuben odgovor, bistveno slabši kot tisti, kjer imajo na voljo denimo štiri predlagane odgovore, izberejo pa enega:

- odgovor odprtega tipa (vnosno polje): 36,3 %,
- odgovor zaprtega tipa (A, B, C ali D): 60,8 %,
- izbor z razlago (A ali B, zakaj?): 68,3 %.

Slabše rezultate pri odgovorih zaprtega tipa v primerjavi z ostalima dvema kategorijama je treba jemati z rezervo, saj so bili primeri s tako vrsto odgovora zgolj štirje. Samo določanje pravilnosti odgovora je pri takih primerih težje, osebno pa sem bil strog ocenjevalec, saj sem vse odgovore, ki niso bili popolnoma pravilni, označil za napačne – poleg tega, da je bil zbran pravi odgovor (A ali B) je razlaga v polju »zakaj?« morala odražati popolno razumevanje, da bi odgovor uvrstil med pravilne.

4.6 Skupina prevajalcev

Edini demografski podatek, ki sem ga zbiral, je, ali se oseba, ki odgovarja na vprašalnik, ukvarja s prevajanjem, več o tem v naslednjem poglavju. Pritrdilno je odgovorilo 24 udeležencev od 120. Pri teh osebah sem analiziral odgovore glede na vrsto napake in jih primerjal z neprevajalci. Nasploh so bili njihovi rezultati za 6 % boljši (63,7 %), po kategorijah pa:

- izmišljena beseda 53,5 % (+ 6,3 % boljše od neprevajalcev),
- neprevedena beseda 65,6 % (+ 1 %),
- razdvoumljanje večpomenske besede 70,8 % (+ 6,7 %),
- pomenska napaka 63,9 % (+ 9,6 %).

Ostalih demografskih podatkov nisem zbiral, kar je ena od slabosti raziskave. V primeru da bi podatki sovpadali z mojo predpostavko, da

niso relevantni, jih ne bi vključil, sedaj pa preprosto nimam podatkov, na katerih bi lahko utemeljil svojo odločitev.

5 Povzetek rezultatov

Pri pregledu rezultatov sem ugotovil, da je bil odstotek pravilnih odgovorov pri izboru strojno prevedenih segmentov, ki sem jih ocenjeval, 59 %. Od vseh 2.880 odgovorov je bilo 1.697 pravilnih.

Na tej točki je potrebno izpostaviti primer št. 6, ki je bil nasploh najslabše razumljen in je znižal povprečje rezultatov v vseh kategorijah, v katerih se je nahajal. Zelo verjetna razlaga zakaj je bil ta primer tako slabši od povprečja je dejstvo, da je bil ta primer glede modaliteta odgovora kategoriziran kot odgovor odprtega tipa (prazno vnosno polje) – če odgovor ni bil povsem točen, sem ga označil za napačnega – pravilna sta bila samo dva. Če bi v tem primeru anketirancem ponudil denimo zaprti tip odgovora, bi bil ta odstotek vsekakor višji, kar potrjuje, da bi iz vidika metodologije lahko bil pristop odlikovanja vprašanj in odgovorov boljši.

Tabela 3: *Primer št. 6; »Mednopni vložek«*

Slovenski prevod	Angleški izvirnik
En zmagovalec bo prejel grafično kartico GeForce RTX 2080 Ti Cyberpunk 2077 Edition.	One winner will receive the GeForce RTX 2080 Ti Cyberpunk 2077 Edition graphics card.
Vstop v predavanje je enostaven: 1. Prijavite se na forume ali ustvarite forumski račun . 2. Komentirajte to temo (BREZ CITIRANJA TE POSTAJE) in nam povejte, kaj želite narediti najbolj v Cyberpunku 2077. 3. Za potrditev vpisa vpišite svoje uporabniško ime v naš pripomoček za oddajo.	Entering the giveaway is easy: Sign in to the forums or create a forum account. Comment on this thread (WITHOUT QUOTING THIS POST) and tell us what you want to do most in Cyberpunk 2077. Sign your username in our giveaway widget to confirm your entry.
KAKO VSTOPITI: Če želite vstopiti, vnesite **mednopni vložek** in sledite navodilom za vstop v nagradne igrače.	HOW TO ENTER: To enter, submit **your entry during the Sweepstakes Period** and follow the directions to enter the Sweepstakes.

eTranslation je bil v povprečju za 15 % boljši od prevajalnika Google Translate, v katerem je bil omenjen primer. Nasploh pa je eTranslation kazal boljše rezultate. Najboljši rezultati glede na tip napake so bili vezani na razdvoumljanje besednega pomena (65,9 %), kar kaže, da

znamo ljudje nasploh dobro razbrati pomen iz sobesedila, na drugem mestu pa so bile neprevedene besede (64,8 %).

Rezultati so bili slabši, ko je prevajalnik napravil hujšo pomensko napako, ki je oteževala razumevanje celotnega segmenta (56,3 %), daleč najslabše rezultate pa je bilo moč opaziti v kategoriji izmišljena beseda (48,5 %), v kateri je sicer bil prej omenjeni primer št. 6.

Glede na tip konteksta so bili najboljši rezultati pri primerih, kjer je slika vplivala na razumevanje (69,8 %) in kjer so udeleženci morali izbrati sliko, na katero se je nanašalo besedilo (64,2 %). Rezultati so bili nekoliko slabši v izključno tekstovnem kontekstu (60,4 %), najslabši rezultati pa so bili v kategoriji, kjer je bila besedilu priložena slika, ki ne vpliva na razumevanje oz. potencialno zmede udeleženca (44 %) – v tej kategoriji je bil tudi primer št. 6. Izkazalo se je, da slikovni kontekst, ki lahko potencialno vpliva na razumevanje besedilnega segmenta, pri strojnih prevodih v realnih okoliščinah, torej na spletu, z vsem pomožnim gradivom, igra pomembno vlogo.

Udeleženci, ki se sicer ukvarjajo s prevajanjem, so na splošno odgovarjali boljše od povprečja. Njihov delež uspešnosti je bil največji v kategoriji hujša pomenska napaka (+9,6 %), kar bi lahko pojasnili s tem, da zaradi »poklicne deformacije« bolj učinkovito razumejo kontekst.

Pri tem velja omeniti, da je edini demografski podatek, ki sem ga v sklopu raziskave zbral to, če se oseba ukvarja s prevajanjem ali ne. Posledično je problematično sklepati, kakšen je bil denimo nivo znanja angleškega jezika, kakšna je bila starost udeleženih, nivo izobrazbe ipd. V tem primeru gre za veliko pomanjkljivost pri metodologiji, ki grobo omejuje sposobnost poročanja o morebitnih zaključkih. Vse predpostavke bi bile zelo subjektivne in brez empiričnih podatkov je njihovo navajanje brezpredmetno.

6 Stanje leta 2023

V začetku leta 2023 sem besedila po več kot dveh letih in pol (besedila sem namreč zbiral in strojno prevedel maja 2020) ponovno strojno prevedel. Zanimalo me je, če so se prevajalniki z nadgradnjami v tem času izboljšali in če so segmenti, kjer so proizvajali pomanjkljive prevode, sedaj bolje prevedeni.

Znova sem uporabil prevajalnika Google in eTranslation, dodal pa sem še prve rezultate iz prevajalnika DeepL, ki se je v zadnjih letih hitro umestil na sam vrh po kakovosti prevodnih rešitev in berljivosti.

Izpostaviti gre, da sem se pri ponovnem pregledu prevodnih rezultatov osredotočal zgolj na izbor primerov, ki sem jih vključil v sklopu raziskave. Rezultati v tem sklopu so zaradi tega morda nekoliko pristranski, saj če bi ocenjevali celotne segmente, bi lahko o razvoju kakovosti prevodov potegnili drugačne (slabše) zaključke, toda sem nad dotičnim izborom primerov, ki je morda nekoliko arbitraren, imel najboljši pregled in je omogočal relativno enostavno primerjalno analizo.

Primeri, ki sem jih uporabil v anketi, so pri obeh prevajalnikih, ki sem ju vključil v prvotno raziskavo, povzročali težave tako enemu kot drugemu sistemu. V anketi sem od 24 primerov vključil 12 primerov na prevajalnik, torej polovico, v tej evalvaciji pa primerjam takratno stanje z aktualnim za vseh 24 primerov. Za vsak izvorni primer sem označil, če je napaka prisotna ali ne – v določenih primerih je denimo napako storil Google in se s tem umestil v anketo, eTranslation pa je dal dobro rešitev. Za to evalvacijo sem označil vse primere za oba prevajalnika in jih po ponovnem prevodu analiziral. Rezultate sem umestil v kategorije:

- izvorna rešitev je dobra, sedanja je enako dobra,
- izvorna rešitev je dobra, sedanja je slabša,
- izvorna rešitev je slaba, sedanja je odlična,
- izvorna rešitev je slaba, sedanja je boljša,
- izvorna rešitev je slaba, sedanja je enako slaba.

6.1 eTranslation 2023

eTranslation je v prvotni raziskavi dal dobro rešitev pri štirih primerih, ostalih 20 je vsebovalo napako. Po ponovnem prevodu so rezultati sledeči:

- 2 izvorni rešitvi sta bili dobri in ostajata enako dobri,
- 2 izvorni sta bili dobri in sta zdaj slabi oz. vsebujeta napako,
- 6 napačnih rešitev je zdaj odličnih,
- 3 napačne rešitve so zdaj nekoliko izboljšane, vseeno pa ne povsem pravilni,
- ostalih 11 primerov ne kaže sprememb in ostaja napačnih.

6.2 Google Translate 2023

Prevajalnik Google je v izvorni raziskavi pravilno prevodno rešitev proizvedel pri dveh od vseh 24 primerov. Dobri dve leti kasneje kaže bistveno boljše rezultate:

- 2 izvorni rešitvi, ki sta bili dobri in ostajata enako dobri,
- 19 napačnih rešitev je zdaj povsem pravilnih,
- ena rešitev kaže izboljšanje, ni pa povsem pravilna,
- 2 napačni rešitvi ne kažeta sprememb in ostajata napačni.

6.3 DeepL 2023

Rezultate prevajalnika DeepL vključujem prvič, saj ni bil del izvorne raziskave – v tistem času še zdaleč ni užival takega ugleda in prominence kot sedaj, kar nakazuje na to, kako drastično se na tem področju dogajajo spremembe in izboljšave. Prevajalnik sem vključil predvsem zaradi tega, ker so primeri že izbrani in lahko zelo preprosto primerjam sodobne prevodne rešitve z ostalima dvema prevajalnikoma:

- pri 16 primerih je dal odlično rešitev;
- en primer je označen kot mejni, saj je izjemno dvoumen in ga je težko zares jasno umestiti v eno ali drugo kategorijo;
- ostalih 7 primerov je napačnih.

Pri tem gre takoj poudariti, da ima DeepL možnost spreminjanja leksemov v samem uporabniškem vmesniku. S klikom na besedo nam prevajalnik takoj ponudi morebitne alternative, kar je seveda orodje, ki primarno služi prevajalcem. V kontekstu, da bi bila besedila avtomatsko strojno prevedena in na spletu objavljena za končne uporabnike, bi lahko tovrstni tip orodja lahko še dodatno pomagal pri ugotavljanju pravega pomena oz. če bi se to orodje pojavilo v oblačku skupaj s segmentom v izvirniku, bi za osebo, ki ima že povprečno znanje angleškega jezika, verjetno že zadostovalo, da bi v veliki večini primerov prišla do pravega pomena.

Pri napačnih primerih in mejnem primeru sem kliknil na kritično besedo in mi je pri šestih od osmih ponudil pravilno rešitev med alternativami, kar mu daje potencial 22 pravilnih rešitev od 24.

6.4 Primerjava treh prevajalnikov v 2023

Na podlagi omejenega števila primerov in načina izbora opazovanja je težko z gotovostjo dejati, da se je kakovost strojnih prevodov radikalno izboljšala, je pa kljub vsemu na podlagi omenjenih primerov možno opaziti, da je v nekaj manj kot treh letih vsaj pri določenih primerih moč opaziti izboljšanje.

Če se osredotočimo na primere, ki sem jih vključil v raziskavo, je čas najbolje vplival na prevajalnik Google, ki v praksi kaže 20 pravilnih rešitev od 24. Prevajalnik DeepL ima trenutno pri danih primerih 8 napačnih rešitev, kot omenjeno pa kaže potencial, da bi lahko z dodatno funkcionalnostjo ponujanja ostalih prevodnih kandidatov dosegal skoraj popolno pravilnost, z izjemo enega mejnega rezultata. Bistveno slabše se je odrezal prevajalnik eTranslation – vsekakor kaže napredek, saj je v prvotni raziskavi dal pravilno rešitev samo v štirih primerih, to število se je januarja 2023 povzpelo na 8, še trije primeri pa kažejo blažjo izboljšavo – najbolj problematično je dejstvo, da je eTranslation pri dveh primerih celo nazadoval, vsi ostali pa so popolnoma enaki.

Z vidika razvoja najbolje kaže Googlu in prevajalniku DeepL, eTranslation pa je pri tem nekoliko bolj zadržan, se pa vseeno izboljšuje. Če bi želeli ugotoviti, kakšna je resnična stopnja razumevanja pri končnih uporabnikih, bi bilo treba eksperiment ponoviti in razširiti – ponavljam, da so rezultati, ki sem jih tu nanizal, arbitrarno presojeni iz vidika pravilnosti, podobno, kot sem to počel pri izvornih spletnih besedilih, ko sem zbiral »problematični« material za vprašalnik. Z današnjega vidika, vsi primeri, ki sem jih v tem segmentu označil za pravilne, ne bi sodili v anketo, saj se mi zdijo povsem neproblematični.[7]

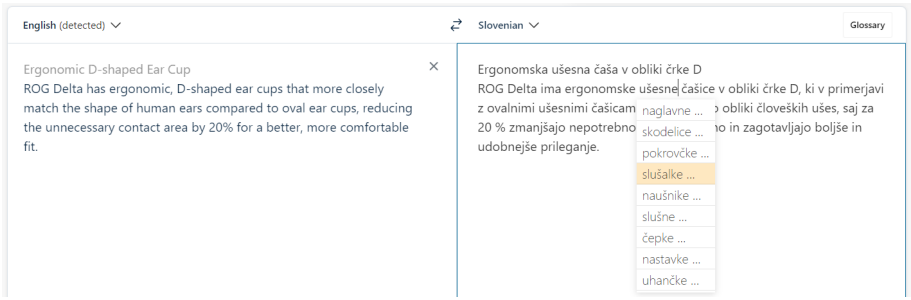
Tabela 4: *Primer nazadovanja eTranslation*

eTranslation 2020	Londonski stolp in Westminster sosednji lokalni pubi in tržnice ter časovno obdelani rituali, kot je menjava stražarjev, se odvijajo, ko vozači hitijo, da **ujamejo Tube**.
eTranslation 2023	Stolp v Londonu in Westminster sosednji lokalni pubi in trgi ter časovno obredi, kot je menjava stražarjev, se odvijajo, ko se vozači hitijo, da bi **ujeli cev**.
Angleški izvirnik	The Tower of London and Westminster neighbor local pubs and markets, and time-worn rituals like the changing of the guards take place as commuters rush to **catch the Tube**.

7 Preglednico z vsemi prevodnimi rezultati sem objavil na povezavi https://bit.ly/Preglednica_Nevronscina_2023.

Tabela 5: *Primeri izboljšav prevajalnika Google*

Google Translate 2020	• **Turška rižota**; • vnesite **mednopni vložek**; • **krompirjevimi klini**; • s kanadskim **šminkerjem** Simoneom Otisom; • **primer vključen**; • **dvopojasni** Wi-Fi 5; • na oprijemni površini je **izrezan diamant srednje globine**.
Google Translate 2023	• **Puranja rižota**; • **oddajte svojo prijavo**; • **rezinami sladkega krompirja**; • s kanadsko **vizažistko** Simone Otis; • **etui vključen**; • **dvopasovni** Wi-Fi 5; • prijemna površina ima srednje globoko **diamantno narebričenje**.
Angleški izvirnik	• **Turkey risotto**; • **submit your entry**; • sweet **potato wedges**; • Canadian **makeup artist** Simone Otis; • **case included**; • **dual-band** Wi-Fi 5; • the gripping surface features medium-depth **diamond knurling**.



Slika 5: Primer pravilno ponujene rešitve pri sprva napačnem prevodu – DeepL.

7 Sklep

V članku sem predstavil raziskavo o razumljivost nerevidiranih strojno prevedenih spletnih besedil pri končnih uporabnikih, ki niso bili posebej obveščeni, da prebirajo strojne prevode. Razumevanje besedilnih segmentov, ki so vključevali štiri različne tipe napak, ki nastanejo pri strojnem prevajanju NMT-sistemov, sem preverjal z anketo. Ta je vsebovala strojne prevode splošnih besedil, ki sem jih prevedel s prevajalnikoma Google Translate in eTranslation. Besedila so bila nerevidirana,

vsebovala so napake, ki so bile predstavljene v več različnih kontekstih, bodisi s slikovnim gradivom bodisi brez.

Rezultati so pokazali, da je splošna stopnja razumevanja 59 %, pri čemer se je izkazalo, da so prevodi eTranslationa nasploh razumljivejši od prevodov Googlovega prevajalnika. Število pravilnih odgovorov je bilo najvišje v kategoriji razdvoumljanja večpomenskih besed, kar nakazuje na to, da ljudje lažje razumemo pomen strojnih prevodov, če nam je dan kontekst. Pri tem je bilo najbolj učinkovito slikovno gradivo, s katerim so si lahko udeleženci v raziskavi pomagali razjasniti pomen določenega besedilnega segmenta.

Po analizi se je izkazalo, da je bil nekoliko problematičen način izbire odgovorov, saj sem anketirancem naključno vnaprej določil, na kakšen način bodo odgovarjali. Odgovori odprtega tipa so kazali slabše rezultate kot izbirni odgovori in odgovori zaprtega tipa, toda zaradi majhnega števila vprašanj je težko izpeljati kakšen razumen zaključek. Podobno velja za samo metodo odgovarjanja na anketo, ki je bila pogojena pandemičnemu času. Za bolj relevantne rezultate bi bilo potrebno izvajati test razumljivosti v živo, na razpravljalen način. Enako velja tudi za vzorec sodelujočih – večji in bolj raznolik vzorec bi dal jasnejše rezultate.

V bodoče bi bilo zanimivo raziskati, če se razumevanje nerevidiranih strojno prevedenih besedil izboljšuje skupaj z nadgradnjami strojnih prevajalnikov, hkrati pa bi se lahko osredotočil še na avtomatsko generirana besedila in jezik spletnih robotov.

Menim, da bo v prihodnje nekoliko manj raziskav storilnosti pri popravljanju strojnih prevodov in veliko več raziskav, ki bodo vezane na razumljivost strojno prevedenih ali avtomatsko generiranih besedil v praktičnih situacijah. Končni bralec se vedno bolj pogosto srečuje s takimi besedili, lahko pa pričakujemo, da bo zaradi še dodatnih izboljšav strojnih prevajalnikov, novih metod in razširjenosti prakse tovrstnih potencialnih stikov med stroji in bralci brez vmesnega posega človeškega popravljalca vedno več.

Zahvala

Raziskovalni program št. P6-0436 (Digitalna humanistika: viri, orodja in metode) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Bordon. D. (2022). Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov. V D. Fišer & T. Erjavec (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 286–291). Ljubljana: Inštitut za novejšo zgodovino. Pridobljeno s https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Bordon, D. (2021). »Razumevati nevronščino: Kako si ljudje razlagamo jezik strojnih prevajalnikov«. Magistrsko delo. Ljubljana: Univerza v Ljubljani. Pridobljeno s https://repozitorij.uni-lj.si/IzpisGradiva.php?id=125328

Castilho, S., & Guerberof Arenas, A. (2018). Reading Comprehension of Machine Translation Output: What Makes for a Better Read?. V J. A. Perez-Ortiz, F. Sanchez-Martinez, M. Espla-Gomis, M. Popovič, C. Rico, A. Martins, J. Van den Bogaert, M. L. Forcada (ur.), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (str. 79–88). Alacant, Španija. Pridobljeno s http://doras.dcu.ie/23071/

Donaj, G., & Sepesy Maučec, M. (2018). Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina. V D. Fišer & A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018* (str. 62–68). Ljubljana: Filozofska fakulteta, Inštitut za novejšo zgodovino. Pridobljeno s http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Donaj-et-al_Prehod-iz-statisticnega-strojnega-prevajanja-na-prevajanje-z-nevronskimi-omrezji-za-jezikovni-par-slovenscina-anglescina.pdf

Evropska komisija (2020). European Language Industry Survey 2020 Before & After Covid-19. Pridobljeno s https://ec.europa.eu/info/sites/default/files/2019_language_industry_survey_report.pdf

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. V *Proceedings of the First Workshop on Neural Machine Translation* (str. 28–39). Vancouver, Canada: Association for Computational Linguistics. Pridobljeno s https://arxiv.org/pdf/1706.03872.pdf

Lala, C., & Specia, L. (2018). Multimodal Lexical Translation. V *Proceedings of the 11th international conference on language resources and evaluation (LREC)* (str. 3810–3817). Miyazaki, Japonska: European Language Resources Association (ELRA). Pridobljeno s https://www.aclweb.org/anthology/L18-1602/

Lelner, Z. (2022). Machine Translation vs. Machine Translation Post-editing: Which One to Use and When?. Pridobljeno s https://blog.memoq.com/machine-translation-vs.-machine-translation-post-editing-which-one-to-use-and-when

Liu, J. (XX) Multimodal Machine Translation. Pridobljeno s https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9547270

Macken, L., & Ghysele, I. (2018). Measuring Comprehension and User Perception of Neural Machine Translated Texts: A Pilot Study. V *Translating and the Computer 40 (TC40), Proceedings* (str. 120–126). Geneva: Editions Tradulex. Pridobljeno s https://biblio.ugent.be/publication/8580951

Macken, L., Van Brussel, L., & Daems, J. (2019). NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. V *Computational Linguistics in the Netherlands Journal, 9,* 67–80. Pridobljeno s https://www.clinjournal.org/clinj/article/view/93

Martindale, M. J., & Carpuat, M. (2018). Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. Pridobljeno s https://arxiv.org/abs/1802.06041

Nunes Vieira, L., O'Sullivan, C., Zhang, X., & O'Hagan, M. (2022). Machine translation in society: insights from UK users. *Language Resources & Evaluation*. Pridobljeno s https://doi.org/10.1007/s10579-022-09589-1

Popović, M. (2020). Relations between comprehensibility and adequacy errors in machine translation output. V R. Fernández & T. Linzen, *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)* (str. 256–264). Pridobljeno s https://aclanthology.org/2020.conll-1.19.pdf

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. Pridobljeno s https://arxiv.org/abs/1508.07909

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. Pridobljeno s https://arxiv.org/abs/1911.12798

Thi-Vinh, N., Ha, T.-L., Nguyen, P.-T., & Nguyen, L.-M. (2019). Overcoming the Rare Word Problem for Low-Resource Language Pairs in Neural Machine Translation. V *Proceedings of the 6th Workshop on Asian Translation* (str. 207–214). Hong Kong, Kitajska: Association for Computational Linguistics. Pridobljeno s https://arxiv.org/abs/1910.03467

Voroniak, D. (2022). Post-Editing of Machine Translation: Best Practices. Pridobljeno s https://blog.crowdin.com/2022/03/30/mt-post-editing/

Zdarek, D. Machine Translation Post-editing Best Practices. Pridobljeno s https://www.memsource.com/blog/post-editing-machine-translation-best-practices/

Do you Speak Neuralese? How People Comprehend the Language of Modern MT Systems

The aim of this paper is to present a study on the comprehensibility of unedited machine-translated web texts. The primary participants in the study were general readers, not trained translators or post-editors, and it is the first study of its kind to be conducted for the Slovene language. The aim of the study was to examine the extent to which unedited machine translations are comprehensible to general readers, while giving focus to the influence of textual and pictorial context. The translations were obtained from Google Translate and eTranslation. The survey was conducted by means of a questionnaire, in which participants answered questions that tested their understanding of a text segment that included an error. The results provide an insight into the current state of development of machine translation engines, not from the point of view of PEMT, but from the point of view of how well machine translations are understood by the target readership. At the end of the article, I provide a new evaluation of MT output in the year 2023, including results for the DeepL MT engine.

Keywords: unedited machine translation, MT evaluation, Understandability by end readers, Google Translate, eTranslation, DeepL

Sklop 2
Jezikovni viri in tehnologije

Spremljevalni korpus Trendi in avtomatska kategorizacija

Iztok KOSEM
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan;
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Jaka ČIBEJ
Institut Jožef Stefan; Filozofska fakulteta, Univerza v Ljubljani

Kaja DOBROVOLJC
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

Taja KUZMAN
Institut Jožef Stefan

Nikola LJUBEŠIĆ
Institut Jožef Stefan; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani;
Inštitut za novejšo zgodovino

Prispevek predstavlja izdelavo korpusa Trendi, prvega spremljevalnega korpusa za slovenščino. Trenutna različica Trendi 2023-02 pokriva besedila od januarja 2019 do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic oz. več kot 586 milijonov besed. Namen korpusa je, da tako strokovni kot nestrokovni javnosti ponudi podatke o aktualni jezikovni rabi in omogoči spremljanje pojavljanja novih besed ter upadanja ali naraščanja rabe že obstoječih. Poleg same vsebine predstavimo tudi metodologijo in načela izdelave korpusa. Drugi del prispevka opisuje razvoj algoritma za avtomatsko kategorizacijo besedil z novičarskih portalov, ki je bil pripravljen za potrebe korpusa Trendi in tudi drugih korpusov s tovrstnimi besedili. Za namene algoritma je bil izdelan nabor 13 tematskih kategorij, ki so v veliki meri prekrivne z mednarodnimi standardi in

Kosem, I., Čibej, J., Dobrovoljc, K., Kuzman, T., Ljubešić, N.: Spremljevalni korpus Trendi in avtomatska kategorizacija. Slovenščina 2.0, 11(1): 161–188.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.161-188

kategorijami v primerljivih korpusih drugih jezikov. Na besedilih, označenih s kategorijami, smo naučili več različnih jezikovnih modelov in z najprimernejšim dosegli visoko zanesljivost določevanja tematike besedilom.

Ključne besede: spremljevalni korpus, avtomatska kategorizacija besedil, neologizmi, novičarski portali, slovenščina

1 Uvod

Jezik se nenehno spreminja, pojavljajo se nove besede, obstoječe besede in besedne zveze dobivajo nove pomene, določene besede ali njihovi pomeni se prenehajo uporabljati ipd. V zadnjem času, tudi zaradi epidemije covida-19, ki je prinesla veliko novega izrazoslovja, je še posebej veliko pozornosti deležno področje neologije, tako leksikalne (nove besede) kot semantične (novi pomeni).

Za spremljanje sprememb v jeziku se tipično uporabljajo spremljevalni korpusi, ki vsebujejo najnovejša besedila v jeziku. Spremljevalni korpusi zapolnjujejo manko referenčnih korpusov, katerih izdelava zaradi raznovrstnosti besedil in njihovih formatov ter obsega traja dlje časa. V času tehnološkega napredka in ob dejstvu, da je zdaj zelo veliko besedil dostopnih na spletu, je izdelava spremljevalnih korpusov postala enostavnejša; kar je objavljeno danes, je lahko že jutri vključeno v korpus (seveda pod pogojem, da besedilo ustreza kriterijem za vključitev in ga je tudi z vidika avtorskih pravic mogoče dodati v korpus).

Za slovenščino kljub bogati opremljenosti na področju korpusov do zdaj nismo imeli spremljevalnega korpusa, čeprav se je med različnimi deležniki kazala jasna potreba po njem. Naslavljanja tega manka smo se lotili v okviru projekta *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED),[1] ki je potekal od oktobra 2021 do novembra 2022 in ga je sofinanciralo Ministrstvo za kulturo Republike Slovenije.

Projekt SLED je naslovil še eno s korpusi povezano potrebo jezikoslovne skupnosti, in sicer algoritem za avtomatsko pripisovanje tematike korpusnim besedilom z novičarskih portalov. Tematsko modeliranje je dokaj razvita disciplina, ki je bila v slovenskem prostoru že uporabljena za primerjavo korpusov (Logar idr. 2015, Logar Berginc in

1 https://sled.ijs.si/

Ljubešić 2013), zbirke besedil z metapodatki o tematiki pa še niso bile obogatene s strojnimi metodami, čeprav je za to veliko potenciala. Metapodatek o tematiki, ki jo besedilo naslavlja, je koristen pri semantični analizi besed in besednih zvez, saj lahko takoj pokaže oz. opozori na morebitno omejenost rabe (npr. *dvojni dvojček* je omejen na področje športa). Pri spremljevalnih korpusih se besedila zbirajo dnevno (najpogosteje gre za besedila, ki jih objavljajo novičarski portali) in imajo le redko eksplicitno pripisano tematiko (podobno velja tudi za besedila v drugih korpusih). Ročno pripisovanje tematik bi bilo zaradi velike količine besedil izjemno zamudno in dolgoročno nevzdržno, zato je smiselno izdelati orodje za avtomatsko pripisovanje tematike.

V prispevku najprej ponujamo pregled nekaterih pomembnejših tujih spremljevalnih korpusov, nato pa predstavimo metodologijo in vsebino spremljevalnega korpusa Trendi. Sledi predstavitev izdelave orodja za avtomatsko kategorizacijo tematike besedil novičarskih portalov, pri čemer je bil pomemben del tudi izdelava nabora tematskih kategorij, ki je dovolj podroben, da razdeli besedila v smiselne in karseda celovite kategorije, in hkrati dovolj robusten, da za strojno kategorizacijo ni pretežaven. Na besedilih, označenih s kategorijami, smo naučili več različnih jezikovnih modelov in z najprimernejšim dosegli visoko zanesljivost določevanja tematike besedilom. V zaključku predstavimo načrte za prihodnje delo.

2 Spremljevalni korpusi

V mednarodnem prostoru so spremljevalni korpusi prisotni že od 20. stoletja, saj se je ideja sistematičnega na podatkih temelječega spremljanja sprememb v jeziku pojavila kmalu po nastanku prvih korpusov. Eden prvih spremljevalnih korpusov je bil angleški korpus Bank of English, ki je bil prvič objavljen leta 1991, z namenom rednega posodabljanja z novimi besedili iz pisnih in govorjenih besedil tako britanske kot ameriške angleščine. Korpus je bil kasneje vključen v 4,5-milijardni korpus COBUILD založbe Collins in po zadnjih podatkih obsega več kot 650 milijonov besed, a ima precej omejen dostop, saj ga lahko poleg zaposlenih na založbi Collins prosto uporabljajo zgolj zaposleni in študentje na Univerzi v Birminghamu.

Za angleščino je danes pomemben predvsem korpus NOW (News on the Web; Davies, 2016-), ki se dnevno posodablja z besedili različnih spletnih novičarskih strani (do 200 milijonov besed na mesec) in danes vsebuje več kot 16 milijard besed. Tako kot nekateri drugi v nadaljevanju omenjeni korpusi je korpus NOW prosto dostopen tako za brskanje preko spletnega portala[2] kot prenos podatkov na lokalni računalnik.[3] Med njimi je denimo tudi korpus Coronavirus (Davies 2019-), specializirani podkorpus korpusa NOW, ki se od januarja 2020 dnevno posodablja s 3–4 milijoni novimi besedami iz angleških spletnih novic na temo pandemije covid-19.

Obsežna zbirka korpusov za spremljanje sprememb v jeziku, ki poleg angleščine pokriva še več kot 35 drugih jezikov, so korpusi Timestamped JSI, ki vsebujejo spletne novice, zbrane preko storitve JSI Newsfeed na Institutu »Jožef Stefan« (Trampuš in Novak, 2012). Korpusi za 18 jezikov so na voljo v orodju Sketch Engine (Kilgarriff idr., 2004; Bušta idr., 2017), v katerem imajo poleg ostalih funkcij orodja uporabniki na voljo tudi t. i. Trende (Herman in Kovár, 2013), funkcijo, ki pomaga prepoznavati trende v rabi besed in je na voljo tudi za nekatere obsežnejše diahrone korpuse. Korpusi JSI Newsfeed v Sketch Enginu vsebujejo besedila od 2014 do aprila 2021 (čas zadnje posodobitve) in so različnih velikosti; korpus angleščine na primer vsebuje približno 60 milijard besed. V primerjavi z drugimi spremljevalnimi korpusi je posebnost te večjezične zbirke tudi to, da so poleg običajnih metapodatkov o datumu, viru in jeziku vsebovanih besedil besedila kategorizirana tudi glede na mesec, četrtletje, področje, lokacijo, ključne besede in druge podobne kategorije, relevantne za nadaljnjo analizo.

Obstaja še precej drugih spremljevalnih korpusov, ki so omejeni na določeno vrsto ali področje besedil (Mikko idr., 2018; Mauri idr., 2019; De Smedt, 2021) ali pa so na voljo zgolj za interno rabo. Primer takšnega korpusa je ONLINE, dinamični spremljevalni korpus češkega jezika, ki ga izdeluje Inštitut za češki nacionalni korpus. Velik je približno 6,3 milijarde besed in vsebuje spletne novice in komentarje pod njimi ter besedila s forumov in različnih družabnih omrežij. Korpus ONLINE je razdeljen na dva komplementarna korpusa: ONLINE_NOW

2 https://www.english-corpora.org/
3 https://www.corpusdata.org/

in ONLINE_ARCHIVE. Prvi je posodobljen vsak dan in pokriva obdobje preteklih šestih mesecev. ONLINE_ARCHIVE pokriva obdobje od februarja 2017 do prvega meseca, ki ga vsebuje ONLINE_NOW. Tako se vsebina zadnjega meseca po starosti v korpusu ONLINE_NOW na začetku vsakega meseca preseli v ONLINE_ARCHIVE.

Do določene mere lahko vlogo spremljevalnega korpusa opravljajo tudi diahroni korpusi, seveda pod pogojem, da vsebujejo čim novejša besedila. Kot primer lahko navedemo korpus sodobne ameriške angleščine (Corpus of Contemporary American English; Davies, 2008-), ki vsebuje besedila od leta 1990 do marca 2020 (zadnja posodobitev) in obsega več kot milijardo besed. Prednost korpusa je, da je žanrsko uravnotežen, saj vsebuje besedila iz osmih različnih žanrov (govorjeni jezik, leposlovje, revije, časopise, znanstvena besedila, televizijske in filmske podnapise, bloge in ostale spletne strani). Slovenski ekvivalent bi bil korpus Gigafida 2.0 (Krek idr., 2019), ki obsega 1,13 milijarde besed, vendar pa je v primerjavi s korpusom sodobne ameriške angleščine manj ažuren (vsebuje samo besedila do leta 2018). Kot hibridni pristop med zasnovo spremljevalnega korpusa na eni strani in statičnega na drugi lahko omenimo še češke korpuse serije SYN (Hnátková, 2014), ki se kot neprekrivni sinhroni korpusi izdajajo vsakih pet let (npr. SYN2000, SYN2005, SYN2010 itd.) in skupaj tvorijo referenčni korpus sodobne pisne češčine SYN, ki v zadnji verziji obsega več kot 6 milijard besed.

Za slovenščino pravi spremljevalni korpus do danes še ni obstajal. Obstajajo sicer spletne storitve, kot je Jezikovni sledilnik (Kosem idr., 2021), ki že izkorišča najsodobnejše podatke o jezikovni rabi, v konkretnem primeru podatke že omenjene storitve JSI Newsfeed, za izdelavo neke vrste začasnih korpusov, na katerih se potem izvajajo specifični statistični izračuni.[4] Taka ciljna raba je seveda smiselna, vendar pa je namenjena nestrokovni javnosti; po drugi strani strokovna javnost, kot so leksikografi_ke, jezikoslovci_ke in drugi raziskovalci_ke, potrebujejo tudi dostop do izvirnih besedil in kompleksnejših metapodatkov, če želijo opravljati še druge analize.

4 Podobno je zasnovana priljubljena storitev Google Ngram Viewer (https://books.google.com/ngrams/), ki omogoča analizo rabe besedišča skozi čas. Temelji na podatkovni zbirki Google Books NGram Corpus (Michel idr. 2011), ki vsebuje najpogostejše besede oz. besedne nize iz sicer nedostopne zbirke digitaliziranih besedil od začetka 17. stoletja dalje.

3 Korpus Trendi

Izdelave prvega spremljevalnega korpusa za slovenščino, ki smo ga poimenovali Trendi, smo se lotili v okviru projekta SLED. Poleg izdelave in rednega posodabljanja korpusa Trendi je imel projekt še dva cilja: pripravo na korpusnih podatkih temelječe statistike o različnih vidikih rabe besed in izdelavo orodja, ki besedila avtomatsko opremi s podatkom o tematski kategoriji (glej razdelek 4).

3.1 Metodologija in vsebina korpusa

Z metodološkega vidika smo pri snovanju korpusa Trendi morali sprejeti dve odločitvi: obdobje, ki ga bo korpus pokrival, in kako pogosto bo korpus posodobljen. Pri odločitvi o obdobju smo izhajali iz želje, da bi korpus Trendi vedno pokrival manko najnovejše različice referenčnega (pisnega) korpusa Gigafida, v času pisanja prispevka je bila zadnja različica 2.0. V tem trenutku to pomeni, da Trendi vsebuje besedila od januarja 2019 naprej. Ob objavi nove različice korpusa Gigafida se besedila iz korpusa Trendi dodajo v korpus Gigafida, obdobje, ki ga pokriva nova različica korpusa Trendi, pa se temu ustrezno prilagodi.

Tesna povezanost s korpusom Gigafida tudi pomeni, da bo korpus Trendi predstavljal standardno pisno slovenščino. Odločitev se zdi smiselna tudi zato, ker sta nestandardna oz. govorjena slovenščina pokrita s korpusi, kot sta JANES[5] in Gos,[6] in je torej njun razvoj predmet ločenih projektov.

Pri pripravi seznama virov za vključitev v korpus Trendi smo izhajali iz seznama slovenskih spletnih virov, ki jih najdemo v servisu JSI Newsfeed. Izdelali smo seznam vseh virov od leta 2019 do konca 2021, pridobili smo tudi podatek o skupnem številu besedil na vir. Nato smo pri pripravi seznama za korpus Trendi podrobno analizirali vsakega od 243 virov. 90 virov smo izključili, ker je šlo za tuje ali slovenske spletne strani z vsebino v tujem jeziku. Nato smo s seznama odstranili še več kot 40 virov, nekatere zato, ker niso vsebovali medijskih novic (blogi, spletne strani vladnih uradov in podjetij), druge zato, ker je njihova vsebina preveč specializirana (npr. repozitoriji akademskih publikacij

5 https://www.clarin.si/kontext/query?corpname=janes
6 http://www.korpus-gos.net/

so primernejši za korpuse, kot je Korpus akademske slovenščine – glej Žagar idr., 2022). Ena od strani (preberi.si) je bila s seznama odstranjena zato, ker je agregator novic iz drugih virov. Končni seznam korpusa Trendi tako vsebuje 107 virov, med tistimi, ki so v obdobju 2019–2021 prispevali največ novic, so sta.si (260.080 besedil), rtvslo.si (97.924), siol.net (69.471), delo.si (65.415), 24ur.com (61.623), dnevnik.si (47.749) in vecer.com (45.548).

Seznam virov se bo redno posodabljal, saj lahko pričakujemo pojav novih spletnih strani, pa tudi ukinitev obstoječih. Kot primer lahko navedemo spletno stran necenzurirano.si, ki se je pojavila šele leta 2020 in je že 28. po številu novic (8.494). Dodajanje novih virov v korpus pomeni tudi večje število besed na mesečni ravni in posledično večji korpus Trendi. Trenutni okvirni izračuni kažejo, da se bo Trendi vsak mesec povečal za 10-15 milijonov pojavnic, pri čemer je bil povprečen mesečni obseg leta 2019 12,5 milijona pojavnic, leta 2021 pa že 21 milijonov pojavnic.

Zaradi narave korpusa Trendi bodo potrebne redne posodobitve, ki so zaenkrat predvidene na mesečni ravni, kot je praksa pri podobnih tujih korpusih.[7] To se zdi trenutno realno, upoštevajoč časovno zahtevnost pridobivanja in označevanja besedil, pretvorb v potreben format in vključevanje korpusa v konkordančnike.

3.2 Priprava besedil

Za pripravo besedil smo pripravili cevovod, ki vključuje pridobivanje besedil, označevanje na različnih ravneh, združevanje po virih in obdobjih ter pretvorbo v različne formate. Pridobivanje besedil je zaenkrat vezano na servis JSI Newsfeed, ki uporablja protokol RSS novic. Nekateri viri, kot so sta.si, delo.si itd. imajo določene vsebine zaklenjene oziroma so dostopne samo naročnikom, zato so v teh primerih pri pridobivanju prek protokola RSS prosto dostopni samo povzetki ali prvih nekaj odstavkov, včasih celo samo naslov in podnaslov. Pri reševanju tega problema smo združili moči z ekipo, ki v okviru projekta *Razvoj slovenščine v digitalnem okolju* (RSDO) sklepa pogodbe z besedilodajalci.

7 Pri začetnih verzijah smo se še srečevali z določenimi hrošči pri pridobivanju in pretvarjanju besedil, zato je prihajalo do nekajmesečnih zamikov pri objavi korpusa.

Dogovor z besedilodajalci vključuje redno dostavljanje celotnih besedil. Posledično bo končna oblika cevovoda za korpus Trendi kombinacija priprave besedil, pridobljenih s spleta, in besedil, ki jih bodo v digitalni obliki poslali tisti besedilodajalci, ki preko protokola RSS ne omogočajo dostopa do celotnih besedil.

Del postopka pridobivanja besedil je tudi deduplikacija, ki je trenutno omejena zgolj na raven vira besedila; del cevovoda je namreč preverjanje, da se besedilo z istim URL-jem ne ponovi. Zavedamo se, da zaradi pokrivanja istih dogodkov obstaja velika prekrivnost med viri. Še več, mnogi viri osnujejo številne novice na podlagi vsebin sta.si, kar pripelje do podvajanja besedila na ravni stavkov, odstavkov ali tudi celotne vsebine. Kljub temu za namene korpusa Trendi deduplikacija na ravni vsebine ni predvidena, saj želimo uporabnikom omogočiti analizo vsebin posameznih virov ter primerjalne analize med viri. Deduplikacija pa bo najbrž opravljena pri pripravi besedil za novo različico korpusa Gigafida, kot je bila praksa v preteklih različicah (Krek idr., 2019).

Sledi postopek strojnega označevanja besedil, za kar uporabljamo označevalni cevovod CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019),[8] ki se kot referenčno orodje za slovnično označevanje besedil v slovenščini aktivno razvija v okviru projekta RSDO. Orodje je nadgradnja odprtokodnega orodja Stanza (Qi idr., 2020), ki v primerjavi z izvorno programsko opremo podrobneje naslavlja specifike slovenščine, zlasti na ravni stavčne segmentacije, tokenizacije, oblikoskladenjskega označevanja in lematizacije po sistemu JOS (Erjavec idr., 2010). Poleg navedenih ravni orodje besedila tudi skladenjsko razčleni po sistemu Universal Dependencies (Dobrovoljc idr. 2017) in v njih označi imenske entitete (Zupan idr., 2017), kot so imena oseb, krajev, organizacij ipd.

Po končanem označevanju se v cevovodu opravi še pretvorba besedil iz privzetega formata označevalnega orodja (CONNL-U) v TEI XML, ki ga med drugim potrebujemo za statistične izračune s programom LIST (Krsnik idr., 2019). V ta proces sta vključena še dva povezana postopka združevanja besedil: združevanje besedil po viru na dan (vsakodneven postopek) in združevanje besedil istega vira za cel mesec (enkrat na mesec, na začetku novega meseca za nazaj). V zadnjem koraku, ki ga

8 https://pypi.org/project/classla/

izvajamo enkrat mesečno in ga moramo pognati ločeno zaradi kombinacije XSLT in skripte Perl, je opravljena še pretvorba mesečnih datotek (razdeljenih po viru) v format VERT, ki ga uporabljata konkordančnika KonText (Machálek, 2020)[9] in NoSketch Engine (Rychlý, 2007)[10].

3.3 Zadnja različica in dostopnost korpusa Trendi

Prva različica korpusa Trendi, imenovana Trendi 2022-05, je bila objavljena junija 2022 in je vsebovala 565.308.991 pojavnic oz. malo več kot 473 milijonov besed. Trenutna zadnja različica Trendi 2023-02 pokriva besedila do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic oz. več kot 586 milijonov besed, kar je v skladu z našimi ocenami, da se bo korpus mesečno povečeval za 10–15 milijonov besed. V korpusu je 1.786.645 besedil 71 izdajateljev, pri čemer imajo največje deleže Slovenska tiskovna agencija (417.718 besedil; 23,4 %), Delo d.o.o. (164.300; 9,2 %), Radiotelevizija Slovenija (150.039; 8,4 %), Media24 d.o.o. (118.651; 6,6 %), PRO PLUS d.o.o. (108.736; 6,1 %) in TSMedia d.o.o. (101.846; 5,7 %).[11]

Korpus Trendi je za brskanje prosto dostopen v treh konkordančnikih CLARIN.SI – konkordančniku KonText in dveh različicah konkordančnika NoSketch Engine; tako KonText kot NoSketch Engine imata več enakih funkcionalnosti (enostavno in napredno iskanje ipd.), vendar pa KonText ponuja možnost registracije in shranjevanje iskanj in priljubljenih korpusov, NoSketch Engine pa dodatne funkcionalnosti, kot je luščenje ključnih besed (angl. *keywords*) iz korpusov, za uporabo katerih ni potrebna registracija. Konkordančnik NoSketch Engine je na CLARIN.SI poleg starejše različice (Bonito) po novem na voljo tudi v novejši različici uporabniškega vmesnika (Crystal),[12] ki zagotavlja izboljšano uporabniško izkušnjo in dolgoročnejše vzdrževanje.

Korpus Trendi kot podatkovna množica trenutno ni na voljo, saj avtorskopravna razmerja z izdajatelji besedil še niso urejena. Pogodbe z

9 https://www.clarin.si/kontext/query?corpname=trendi
10 https://www.clarin.si/noske/run.cgi/corp_info?corpname=trendi
11 Analiza pokaže, da so deleži izdajateljev iz meseca v mesec dokaj nespremenljivi. Najnovejši podatki so na voljo na https://www.clarin.si/ske/#text-type-analysis?corpname=trendi&wlminfreq=1&wlicase=1&include_nonwords=1&showresults=1&wlnums=frq&wlattr=text.publisher.
12 https://www.clarin.si/ske/#dashboard?corpname=trendi

besedilodajalci se sicer hkrati urejajo tudi za referenčni korpus Gigafida. Ko bo to urejeno, bo korpus Trendi na voljo pod odprto licenco z vzorčenimi odstavki (glej Logar idr. 2013[13] kot primer takšnega vira), v celoti pa za posamezne raziskovalce, ki bodo podpisali posebno pogodbo za uporabo korpusa v znanstvenoraziskovalne namene.

4 Avtomatska kategorizacija besedil

Sorodni spremljevalni korpusi, predstavljeni v 2. razdelku, poleg podatkov o letu nastanka besedil pogosto vsebujejo tudi podatek o področni kategorizaciji besedil, ki omogoča opazovanje jezikovnih trendov znotraj posameznih tematskih področij ali identifikacijo trendov, ki se pojavljajo zgolj na določenem področju. Korpusi Timestamped JSI tako vsebujejo informacijo o ključnih besedah in področjih glede na kategorizacijo ontologije DMOZ (Grobelnik idr., 2006), npr. družba, posel, šport, ki je besedilom avtomatsko pripisana z orodjem Enrycher (Štajner 2009).

Podobno tudi češki korpusi serije SYN in ONLINE vsebujejo podrobno večnivojsko kategorizacijo besedil (Cvrček idr., 2020) glede na skupino besedil (npr. leposlovje, stvarna besedila, publicistika), vrsto besedil (npr. znanstvena, strokovna, poljudna besedila), žanrsko skupino (npr. družboslovje) in žanrsko področje (npr. ekonomija, politika, pravo, psihologija), pri čemer sama metoda klasifikacije ni podrobneje dokumentirana.

Manj podrobna je kategorizacija besedil v korpusih NOW, COCA in Coronavirus, saj vsebujejo zgolj delitev na žanre (npr. govorjeni jezik, leposlovje, revije) in za določene med njimi (Sharoff 2018) tudi funkcijske tipe (npr. pravna besedila, navodila, recenzije, promocije), a to pomanjkljivost delno naslavlja njihov skupni konkordančnik, ki za dani konkordančni niz izpiše tudi topike oz. seznam ključnih besed glede na avtomatsko detekcijo pogosto ponavljajočih se besed v prikazanih besedilih (npr. simptom, alergija, pljuča).

4.1 Priprava tematskih kategorij

Ena od aktivnosti projekta SLED je bila tudi izdelava orodja za avtomatsko kategorizacijo besedil glede na tematiko. Za izdelavo takšnega orodja oz. modela zanj potrebujemo dvoje: nabor kategorij in učne množice.

13 https://www.clarin.si/repository/xmlui/handle/11356/1035

Pri izdelavi nabora kategorij smo se opirali na podatke iz treh skupin virov:

- slovenskih novičarskih portalov, izbrali smo jih šest, tj. rtvslo.si, delo.si, sta.si, dnevnik.si, 24ur.com in vecer.com;
- nabora tematskih kod oz. kategorij Mednarodnega tiskovnega telekomunikacijskega sveta (IPTC).[14] S tem smo tudi želeli zagotoviti čim boljšo usklajenost naših kategorij z mednarodnim standardom;
- kategorij v sodobnih sinhronih in spremljevalnih korpusih, pri čemer sta bila relevantna predvsem češki korpus SYN_2015 (Křen idr., 2016) in estonski nacionalni korpus (Koppel in Kallas, v tisku).

Glavno vodilo pri pripravi klasifikacije je bilo pripraviti relativno majhen nabor kategorij, v katere lahko uvrstimo vse novice na različnih portalih. S tem bi zagotovili tudi boljše delovanje modela. Posledično smo pri analizi uporabljenih virov več pozornosti posvečali krovnim kategorijam, kar je bilo sploh potrebno pri naboru IPTC, ki ima približno 1.400 kategorij, razdeljenih v tri nivoje (s tem da krovni nivo sestavlja le 17 kategorij). Za ponazoritev smiselnosti uporabe zgolj krovnih kategorij lahko vzamemo kategorijo *šport*, ki ima na večini novičarskih portalov nadaljnje kategorije, od katerih se povsod pojavita samo *nogomet* in *košarka*, ostale pa le na nekaterih portalih, npr. dnevnik.si nima *zimskih športov*, ima pa ločeno podstran za novice o *Luki Dončiću*; rtvslo.si je edini, ki ima podstran za novice o *Formuli 1*, 24ur.si ima ločene podstrani za *Ligo prvakov* in *Ligo Evropa* (nogomet) ter *borilne športe*.

Končna klasifikacija vsebuje 13 kategorij:

- **Umetnost in kultura**. Vključuje besedila o kulturi, umetnosti, filmih, knjigah, gledališču, pa tudi recenzije ipd.
- **Črna kronika**. Naravne in ostale nesreče, človeški delikti, kriminal.
- **Gospodarstvo**. Vključuje besedila s področja ekonomije, trgov, financ, zaposlitev ipd.
- **Okolje**. Zajema okoljevarstvo, planet, energente, tudi kmetijske teme.
- **Zdravje**. Fizično in mentalno zdravje ljudi, medicina, farmacija, zdravstvena infrastruktura.
- **Prosti čas**. Hobiji, rekreacija, potovanja, turizem, ljubljenčki, dom in družina, bivanje.

14 https://cv.iptc.org/newscodes/subjectcode

- **Politika in pravo**. Mednarodne in nacionalne novice s področna državne uprave, pravnih postopkov in družbenih razmerij, konfliktov, vojn.
- **Znanost in tehnologija**. Znanstvena odkritja, zanimivosti, tehnološke inovacije, informacijska tehnologija, računalništvo.
- **Družba**. Družbena vprašanja in razmerja, enakost, diskriminacija, religija, etika ipd.
- **Šport**. Športni rezultati in zanimivosti z različnih športnih področij.
- **Vreme**. Meteorološke napovedi, opisi vremenskih posebnosti, stanj, procesov.
- **Zabava**. Estrada, moda, slog.
- **Izobraževanje**. Procesi posredovanja in pridobivanja znanja ter veščin. Vse stopnje izobraževanja, od vrtca do univerzitetnega izobraževanja, pa tudi vseživljenjsko učenje.

Kot prikazuje primerjalna Tabela 1, obstaja precejšnja prekrivnost tako s kategorijami novičarskih portalov kot s kategorijami IPTC in tujih korpusov. V nekaterih primerih, npr. *gospodarstvo, prosti čas, politika in družba*, naša kategorija zajema več kategorij ostalih virov. Tako ima za prosti čas estonski korpus kar sedem ločenih kategorij. Edini primer, ko se eno od kategorij tujih virov lahko uvrsti v dve naši, sta *umetnost in kultura* ter *zabava*. Kategoriji smo namreč ločili po eni strani zato, ker ima veliko slovenskih novičarskih portalov ločene podstrani zanju, po drugi strani pa zaradi samega jezika – kulturno-umetniške vsebine so za razliko od zabavnih pogosto precej bolj strokovne.

Medtem ko v naše kategorije lahko umestimo vseh 17 kategorij IPTC, pa češki oz. estonski korpus določenih kategorij nimata, npr. estonski nima *črne kronike*, češki pa ne *okolja*, *zdravja*, *znanosti in tehnologije* ter *zabave*. Oba tudi nimata ločene kategorije za *vreme*, ki pa jo ima IPTC in smo jo dodali zato, ker jo ima večina slovenskih novičarskih portalov.

Tabela 1: *Primerjava tematskih kategorij projekta SLED z domačimi novičarskimi portali in tujimi viri*

kategorija	zastopanost na šestih slovenskih portalih	češki korpus	estonski korpus	IPTC
umetnost in kultura	5	culture	culture & entertainment	arts, culture and entertainment
črna kronika	6	crime	/	disaster and accident
gospodarstvo	6	economy	economy, finance & business; agriculture; construction & real estate	economy, business and finance; labour
okolje	2	/	nature & environment	environmental issue
zdravje	3	/	health	health
prosti čas	4	leisure	beauty; cars; food & drinks; gambling & casinos; home, family & children; pets and animals; travel & tourism; video games	lifestyle and leisure
politika in pravo	1	politics	politics & government	politics; crime, law and justice; unrest, conflicts and war
znanost in tehnologija	5	/	science, technology & IT	science and technology
družba	1	social life	society; religion; sex; women	social issue; religion and belief; human interest
šport	6	sports	sports	sport
vreme	4	/	/	weather
zabava	4	/	culture & entertainment*	arts, culture and entertainment*
izobraževanje	1	/	education	education

Če pogledamo še prekrivnost kategorij s stranmi oz. podstranmi šestih slovenskih novičarskih portalov, vidimo, da so problematične kategorije predvsem *politika*, *družba* in *izobraževanje*. Gre za sicer legitimne kategorije, ki pa na novičarskih portalih nimajo svojih podstrani,

temveč so novice razpršene po drugih podstraneh, ki so večinoma opredeljene glede na geografski izvor novice, npr. Slovenija, Svet, Lokalno. Medtem ko so se avtorji češkega korpusa odločili slediti takšni delitvi tudi pri kategorijah (*current events, foreign news, domestic news, regional news*), smo se mi raje držali tematike. To za izdelavo učnih množic pomeni nekoliko več ročnega dela oz. iskanje drugih kazalcev, s katerimi lahko odkrijemo tematiko prispevka na posameznem portalu. Izjema je portal sta.si, ki že ima ustrezne kategorije, in sicer *Šolstvo* in *Družba*, za politiko pa *Državni zbor, Evropska unija, Mednarodna politika, Slovenska notranja politika in Slovenska zunanja politika*.

4.2 Priprava učnih množic

Na podlagi pripisa kategorije s pomočjo URL-naslovov smo preverili, koliko besedil je na voljo v posamezni kategoriji. Izračunali smo njihovo minimalno, maksimalno in povprečno dolžino (v pojavnicah, tj. besedah, ločenih s presledki) ter dolžino besedil glede na kvartile. Kot kaže Tabela 2, je npr. v kategoriji šport najkrajše besedilo dolgo le 24 pojavnic, najdaljše pa 6028 pojavnic. 25 % vseh besedil v kategoriji šport je krajših od 150 pojavnic (q_1), 25 % pa je daljših od 374 pojavnic (q_3). Pri sestavi učne množice smo želeli vzorčiti predvsem besedila povprečne dolžine in se izogniti ekstremom, zato smo upoštevali le besedila z dolžino nad q_1 in pod q_3 (besedila srednje dolžine).

Tabela 2: *Dolžina in število besedil po kategorijah*

Kategorija	Min. št. pojavnic	Maks. št. pojavnic	q_1	q_2	q_3	Vsa besedila	Besedila srednje dolžine
Šport	24	6.028	150,0	226,0	374,0	70.617	35.099
Okolje	30	4.705	212,75	336,0	519,0	2.896	1.445
Gospodarstvo	26	7.309	98,0	222,0	401,0	11.489	5.725
Umetnost in kultura	26	7.264	236,0	362,0	535,0	17.207	8.596
Znanost in tehnologija	27	6.920	177,0	308,0	608,0	3.465	1.725
Zabava	50	5.944	117,0	156,0	222,0	16.210	7.975
Zdravje	7	31.221	129,0	189,0	503,0	25.134	12.510

Kategorija	Min. št. pojavnic	Maks. št. pojavnic	q_1	q_2	q_3	Vsa besedila	Besedila srednje dolžine
Črna kronika	50	2.301	117,0	154,0	236,0	7.342	3.656
Prosti čas	37	5.835	194,0	284,0	456,0	4.478	2.227
Vreme	56	1.934	128,0	146,0	1573,0	3.804	1.881
Izobraževanje	10	1.844	129,0	141,0	272,25	2.736	1.351
Politika in pravo	7	3.304	127,0	141,0	285,0	38.126	19.013
Družba	9	2.241	126,0	136,0	217,0	10.941	5.426

Najbolje zastopana kategorija je *Šport*, sledita pa ji *Politika* in *Zdravje*. Po drugi strani so slabše zastopane kategorije *Izobraževanje*, *Okolje* ter *Znanost in tehnologija*. Povprečne dolžine besedil (q_2) so med kategorijami podobne – daljša so predvsem besedila v kategorijah *Znanost in tehnologija*, *Okolje* in *Umetnost, kultura*. Na tej točki je treba omeniti, da lahko besedila po tem sistemu spadajo v natanko eno kategorijo – šlo je za pragmatično odločitev, saj tovrstni sistem odseva delitev pri večini virov, kjer je po eno besedilo uvrščeno le v eno kategorijo, obenem pa je implementacija enodimenzionalnih oznak pri klasifikaciji manj kompleksna, manj težav pa to povzroča tudi pri zapisu metapodatkov v končni format VERT. Poleg tega bi bilo treba opraviti tudi eksperiment ročnega označevanja z več kategorijami, da bi videli, ali je večplastna kategorizacija res potrebna in v kolikšni meri. V prvem koraku smo zato eksperimente omejili na enostavnejši scenarij z eno kategorijo.

Za razvoj modela za kategorizacijo tematike novičarskih besedil smo pripravili dve učni množici. Prva, manjša, je vsebovala približno 10.000 besedil (oz. po največ 800 besedil srednje dolžine za vsako od 13 kategorij). Druga, večja, pa približno 36.000 besedil – po 2.800 na kategorijo; pri podreprezentiranih kategorijah, v katerih ni bilo na razpolago vsaj 2.800 besedil srednje dolžine, smo v tem primeru deficit zapolnili tudi s krajšimi in daljšimi besedili. Večja množica je bila zato v primerjavi s prvo nekoliko slabše vzorčena, a smo želeli preveriti, ali lahko z večjo učno množico pri razvoju modela dosežemo višjo napovedno točnost.

Vzorčenje besedil (sestava množic je prikazana v Tabeli 3) je potekalo na naslednji način:

(1) Vzeli smo vsa besedila srednje dolžine znotraj posamezne kategorije.
(2) Vsaki kategoriji smo naključno vzorčili 3.000 besedil (750 besedil na leto za leta 2019, 2020, 2021 in 2022); če je bilo besedil manj kot 3.000, smo v vzorec zajeli vsa.
(3) Iz vzorca 3.000 besedil (iz točke (2)) smo naključno vzorčili 1.000 besedil (250 na leto).
(4) Iz teh 1.000 besedil smo naključno vzorčili 100 besedil za razvojno množico in še ločenih 100 besedil za testno množico.
(5) V manjšo učno množico smo dodali približno 800 besedil (iz vzorca 1.000 besedil iz točke (3) brez 200 besedil za razvojno in testno množico).
(6) V večjo učno množico smo dodali približno 2.800 besedil (iz točke (2)), a brez besedil, ki so bila v točki (4) dodana v razvojno in testno množico. Če ni bilo dovolj besedil srednje dolžine, smo naključno dodali nekaj krajših in daljših besedil, da smo dosegli približno 2.800 besedil (v primeru nekaterih slabše zastopanih kategorij te številke kljub temu ni bilo mogoče doseči).

Tabela 3: *Število besedil v učni, razvojni in testni množici*

Kategorija	Testna množica	Razvojna množica	Manjša učna množica	Večja učna množica
Črna kronika	100	100	800	2.800
Družba	100	100	800	2.800
Gospodarstvo	100	100	800	2.800
Izobraževanje	100	100	695	2.536
Umetnost in kultura	100	100	800	2.800
Okolje	100	100	711	2.696
Politika in pravo	100	100	800	2.800
Prosti čas	100	100	718	2.834
Šport	100	100	800	2.800
Vreme	100	100	726	2.800
Zabava	100	100	800	2.800
Zdravje	100	100	800	2.800
Znanost in tehnologija	100	100	753	2.814
Skupaj	**1.300**	**1.300**	**10.003**	**36.080**

Poskrbeli smo, da je vzorčenje besedil v največji možni meri upoštevalo enakomerno distribucijo tako med kategorijami (npr. črna kronika, politika) kot med viri (npr. rtvslo.si, delo.si). Ker smo besedila za učno množico črpali iz korpusa Trendi, ki v času pisanja tega članka zajema besedila med letoma 2019 in 2022, smo poskrbeli tudi, da so bila besedila karseda enakomerno vzorčena iz vseh štirih let – besedila iz let 2020 in 2021 so bila namreč zaradi pandemije novega koronavirusa v mnogo tematskih kategorijah zaznamovana s pandemsko vsebino; preveč besedil iz tega obdobja bi lahko v model vneslo neželene pristranskosti oz. znižalo njegovo robustnost.

4.3 Razvoj kategorizacijskih modelov in klasifikacija

Preizkusili smo več modelov za kategorizacijo, razvili pa smo jih tako z orodjem fastText (Joulin idr., 2016) kot z orodjem Simple Transformers (Rajapakse, 2019). Rezultate opišemo v nadaljevanju v ločenih razdelkih.

4.3.1 Modeli fastText

Za potrebe učenja modela z orodjem fastText smo učno, testno in razvojno množico predobdelali, tako da smo v besedilih vse znake za odstavke zamenjali s presledki, odstranili ločila in spremenili vse besede v male črke.

Glede na uporabljeno učno množico in glede na to, ali smo pri razvoju upoštevali vnaprej naučene vektorske vložitve za slovenščino na nivoju pojavnic po modelu fastText (Ljubešić in Erjavec, 2018), smo razvili štiri modele:

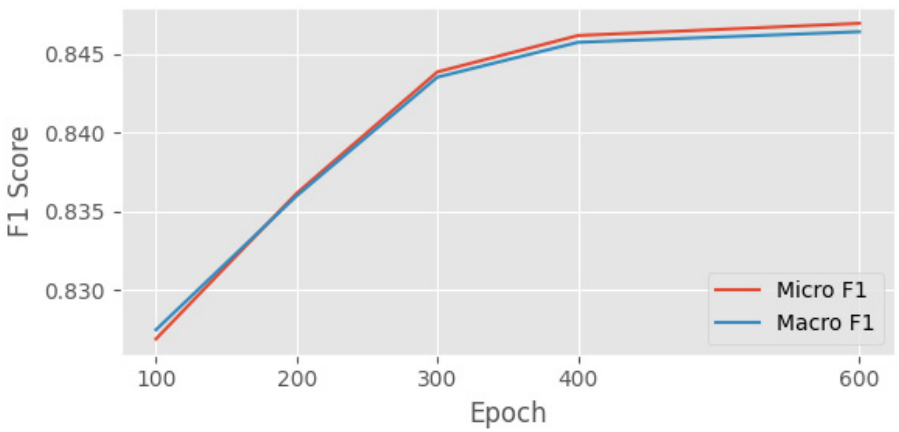
(1) z manjšo učno množico in brez vnaprej naučenih vektorskih vložitev;
(2) z večjo učno množico in brez vnaprej naučenih vektorskih vložitev;
(3) z manjšo učno množico in z uporabo vnaprej naučenih vektorskih vložitev;
(4) z večjo učno množico in z uporabo vnaprej naučenih vektorskih vložitev.

Tabela 4: *Delovanje modelov fastText*

Št. modela	Velikost učne množice	Vložitve	Mikro F1	Makro F1
1	manjša	ne	0,83	0,83
2	večja	ne	0,85	0,85
3	manjša	da	0,85	0,85
4	večja	da	0,85	0,85

Kot prikazuje Tabela 4, pri učenju na manjši učni množici uporaba vnaprej naučenih vektorskih vložitev nekoliko izboljša delovanje modela (2 odstotni točki pri merah makro F1 in mikro F1), pri učenju na večji učni množici pa ne pride do razlik. Učenje po drugi strani pri uporabi manjše učne množice traja le 15 minut, pri uporabi večje pa približno eno uro.

Pri učenju modelov fastText nismo avtomatsko optimizirali hiperparametrov, temveč smo pri večini upoštevali privzete nastavitve. Izjema je bilo število epoh, pri katerem smo model naučili večkrat na različnem številu epoh in preverili, kdaj sta meri F1 optimalni. Za optimiziranje hiperparametrov smo modele učili na učni množici, testirali pa na razvojni. Slika 1 npr. prikazuje meri F1 v odvisnosti od števila epoh za model 3 (manjša učna množica, vektorske vložitve). Meri F1 opazno naraščata do epohe 400, zatem se rast umiri.



Slika 1: Meri F1 v odvisnosti od števila epoh za model fastText z manjšo učno množico in vnaprej naučenimi vložitvami.

Za končni model smo izbrali model 4 (večja učna množica, vektorske vložitve), ki smo ga naučili na 1000 epohah. Model je pod imenom *fastText-Trendi-Topics 1.0* na voljo na repozitoriju CLARIN.SI (Kuzman idr., 2022). Tabela 5 prikazuje, kako se model odreže pri klasifikaciji besedil različnih kategorij na testni množici. Največjo točnost dosega pri kategorijah *Vreme*, *Šport* in *Črna kronika* (F1 > 0,90), najnižjo pa pri kategorijah *Znanost in tehnologija* ter *Prosti čas*. Razlika je pričakovana, saj kategorije z najvišjo točnostjo vsebujejo besedila, ki so si med seboj zelo podobna ne glede na vir ali čas objave (npr. opisi vremena, poročila o prometnih nesrečah, opisi športnih tekem); po drugi strani je kategorija *Prosti čas* precej bolj raznolika, enako pa velja tudi za *Znanost in tehnologijo*, ki pokriva različne discipline in opisuje vedno nova odkritja, zato so si besedila manj podobna.

Tabela 5: *Natančnost, priklic in F1 za različne kategorije besedil pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Kategorija	Natančnost	Priklic	F1
Vreme	0,98	0,96	0,97
Šport	0,94	0,97	0,96
Črna kronika	0,94	0,92	0,93
Umetnost in kultura	0,85	0,91	0,88
Zabava	0,91	0,86	0,88
Izobraževanje	0,90	0,83	0,86
Zdravje	0,86	0,83	0,85
Politika in pravo	0,83	0,83	0,84
Družba	0,87	0,79	0,83
Gospodarstvo	0,81	0,83	0,82
Okolje	0,77	0,88	0,82
Znanost in tehnologija	0,81	0,77	0,79
Prosti čas	0,65	0,70	0,67
Uravnoteženo povprečje	**0,86**	**0,86**	**0,85**

Matrika zamenjav je prikazana v Tabeli 6. Večina najpogostejših zamenjav je do neke mere smiselnih, npr. *Zabava - Umetnost* (6) in *Zabava - Prosti čas* (4) ter *Izobraževanje - Družba* (8) in *Družba - Politika* (7), *Družba - Okolje* (5), *Gospodarstvo - Okolje* (4), *Gospodarstvo - Politika* (7), *Okolje - Politika* (4), *Znanost - Okolje* (6), *Znanost*

- *Gospodarstvo* (4), *Prosti čas - Umetnost* (4). V kontekstu covidne krize so smiselne tudi zamenjave v kategoriji *Zdravje*, npr. *Zdravje - Politika* (4) in *Zdravje - Družba* (5). Manj smiselne so zamenjave iz kategorije *Prosti čas*, ki je bila tudi najslabše klasificirana: *Gospodarstvo - Prosti čas (5)*, *Okolje - Prosti čas* (4), *Znanost - Prosti čas* (7), *Prosti čas - Gospodarstvo* (5).

Tabela 6: *Matrika zamenjav pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Nap.→ Res.↓	**Vre.**	**Šp.**	**Črn.**	**Ume.**	**Zab.**	**Izo.**	**Zdr.**	**Pol.**	**Dru.**	**Gos.**	**Oko.**	**Zna.**	**Pro.**
Vre.	**97**	-	-	-	-	-	-	-	-	-	3	-	-
Šp.	-	**97**	-	-	1	-	-	-	-	-	1	-	1
Črn.	-	2	**91**	-	1	1	1	-	1	1	-	-	2
Ume.	-	1	-	**92**	2	-	-	-	1	-	-	2	2
Zab.	-	2	-	6	**85**	-	-	1	1	-	-	1	4
Izo.	-	-	-	1	2	**82**	1	1	8	1	1	-	3
Zdr.	-	-	1	1	-	-	**85**	4	5	-	-	2	2
Pol.	-	-	3	3	1	1	-	**91**	-	1	-	-	-
Dru.	-	1	-	1	1	3	2	7	**76**	-	5	-	-
Gos.	-	-	-	-	-	1	1	7	1	**80**	4	1	5
Oko.	1	-	-	1	-	-	-	4	-	1	**87**	2	4
Zna.	-	-	-	3	-	-	3	1	1	4	6	**75**	7
Pro.	1	5	2	4	2	1	4	-	1	5	4	5	**66**

4.3.2 Modeli tipa Bert

Jezikovni modeli tipa BERT temeljijo na nevronskih mrežah in se jih z učenjem na več milijardah besed pripravi na učinkovito reševanje različnih nalog obdelave naravnega jezika. Pripravljeni modeli, ki smo jih uporabili, so na voljo na repozitoriju Hugging Face, za nalogo kategorizacije tematike novičarskih besedil pa smo jih doučili z učenjem na učni množici, za kar smo uporabili orodje Simple Transformers.

Tudi pri učenju modelov tipa BERT smo uporabili privzete nastavitve hiperparametrov in preverili le optimalno število epoh pri učenju na učni množici in testiranju na razvojni. Za učenje smo uporabili slovenske vektorske vložitve SloBERTa 2.0 (Ulčar in Robnik-Šikonja,

2021), preizkusili pa smo tudi večjezične vložitve XLM-RoBERTa (Conneau idr., 2020), a ker so se odrezale slabše od enojezičnih, smo jih uporabili le v enem eksperimentu. Razvili smo tri modele (rezultati so prikazani v Tabeli 7):

(1) model z vložitvami XLM-RoBERTa in manjšo učno množico,
(2) model z vložitvami SloBERTa 2.0 in manjšo učno množico,
(3) model z vložitvami SloBERTa 2.0 in večjo učno množico.

Tabela 7: *Delovanje modelov tipa BERT*

Št. modela	Velikost učne množice	Vložitve	Mikro F1	Makro F1
1	manjša	XLM-RoBERTa	0,91	0,91
2	manjša	SloBERTa 2.0	0,93	0,93
3	večja	SloBERTa 2.0	0,94	0,94

Za model (1) se je kot najustreznejša nastavitev izkazalo 6 epoh, za model (2) 8 epoh in za model (3) 2 epohi. Učenje modelov (1) in (2) je trajalo približno 2 uri, modela (3) pa 3 ure.

Če primerjamo model SloBERTa, ki je bil učen samo na slovenskih podatkih (3,47 milijardah pojavnic), in večjezični model XLM-RoBERTa, ki je bil učen na 100 jezikih, od tega na 1,7 milijarde slovenskih pojavnic, rezultati iz tabele potrjujejo, da je tudi pri tej nalogi enojezični model primernejši od večjezičnega. To se sklada z ugotovitvami Ulčar idr. (2021), ki so primerjali modela na drugih pogostih nalogah, povezanih s procesiranjem naravnega jezika.

Čeprav sta si bila glede na rezultate modela (2) in (3) zelo podobna in je model (2) zahteval nekoliko manj časa za učenje, smo za končni model izbrali model (3), ki je bil naučen na večji učni množici in je morda kljub vsemu nekoliko bolj robusten. Na voljo je na repozitoriju CLARIN.SI pod imenom *SloBERTa-Trendi-Topics 1.0* (Čibej idr., 2022) ter na repozitoriju HuggingFace.[15] Rezultati klasifikacije po posameznih kategorijah so prikazani v Tabeli 8.

15 https://huggingface.co/cjvt/sloberta-trendi-topics

Tabela 8: *Natančnost, priklic in F1 za različne kategorije besedil pri modelu SloBERTa 2.0 z večjo učno množico in vektorskimi vložitvami*

Kategorija	Natančnost	Priklic	F1
Vreme	0,98	0,98	0,98
Šport	0,97	0,99	0,98
Črna kronika	0,99	0,98	0,98
Zabava	0,97	0,99	0,98
Umetnost, kultura	0,99	0,96	0,97
Politika	0,93	0,99	0,96
Družba	0,94	0,96	0,95
Zdravje	0,93	0,95	0,94
Znanost in tehnologija	0,89	0,94	0,92
Izobraževanje	0,93	0,88	0,90
Gospodarstvo, posel, finance	0,92	0,87	0,89
Okolje	0,89	0,89	0,89
Prosti čas	0,88	0,83	0,85
Uravnoteženo povprečje	**0,94**	**0,94**	**0,94**

Podobno kot modeli fastText se tudi izbrani model s SloBERTo najbolje odreže pri klasifikaciji besedil iz kategorij *Vreme*, *Šport* in *Črna kronika*. Najslabše se odreže pri kategorijah *Gospodarstvo, posel, finance* ter *Okolje* in *Prosti čas*, precej bolje pa klasificira besedila iz kategorije *Znanost in tehnologija* (F1 znaša 0,92, pri izbranem modelu s fastTextom pa je ta dosegla le 0,79). Tudi nasploh pri vseh kategorijah SloBERTa dosega občutno boljše rezultate – mera F1 je namreč v primerjavi s fastTextom pri posameznih kategorijah višja tudi za 9 odstotnih točk.

Matrika zamenjav je prikazana v Tabeli 9. Do zamenjav pri tem modelu za razliko od modela fastText (Tabela 6) prihaja le sporadično. Tudi tukaj je najpogostejša zamenjava smiselna: *Izobraževanje – Družba* (5), še vedno pa je najbolj problematična kategorija *Prosti čas*, npr. *Gospodarstvo – Prosti čas* (4), *Okolje – Prosti čas* (4).

Tabela 9: *Matrika zamenjav pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Nap.→ Res.↓	Vre.	Šp.	Črn.	Ume.	Zab.	Izo.	Zdr.	Pol.	Dru.	Gos.	Oko.	Zna.	Pro.
Vre.	**98**	-	-	-	-	-	-	-	-	-	2	-	-
Šp.	-	**99**	-	-	-	-	-	-	-	-	-	-	1
Črn.	-	-	**98**	-	-	-	-	-	-	-	1	1	-
Ume.	-	-	-	**94**	1	1	-	-	-	-	1	1	-
Zab.	-	-	1	-	**99**	-	-	-	-	-	-	-	-
Izo.	-	1	-	-	-	**88**	2	-	5	3	-	-	1
Zdr.	-	-	-	-	-	-	**95**	3	-	-	-	2	-
Pol.	-	-	-	-	-	-	-	**99**	-	1	-	-	-
Dru.	-	-	-	-	-	-	2	2	**95**	-	-	-	-
Gos.	-	-	-	-	-	3	1	1	1	**87**	2	1	4
Oko.	2	-	-	-	-	-	1	1	-	2	**87**	1	4
Zna.	-	-	-	-	-	1	1	-	-	-	3	**93**	1
Pro.	-	2	-	1	2	2	-	1	-	2	2	5	**82**

4.3.3 Klasifikacija korpusa Trendi

Za klasifikacijo besedil v korpusu Trendi (Kosem idr., 2022) smo uporabili model *SloBERTa-Trendi-Topics 1.0*, ki je sicer računsko požrešnejši in počasnejši, a v primerjavi z modelom *fastText-Trendi-Topics 1.0* tudi natančnejši. Na primer, za klasifikacijo verzije 2022-10 (1.635.614 besedil od 1. januarja 2019 do vključno 31. oktobra 2022) je bilo potrebnih približno 152 ur procesiranja z grafično procesno enoto (GPU) na spletni strani Kaggle.[16] Za klasifikacijo besedil enega meseca (npr. junij 2022; 41.262 zajetih besedil) smo potrebovali povprečno 3,5 ure (približno 3,3 sekunde na besedilo). Rezultati klasifikacije so prikazani v Tabeli 10.

Tabela 10: *Število klasificiranih besedil v korpusu Trendi 2022-10 po kategorijah*

Kategorija	Število besedil
Vreme	11.262
Šport	311.117
Črna kronika	116.208
Zabava	108.721
Umetnost in kultura	98.076
Politika in pravo	176.608

16 Kaggle: https://www.kaggle.com/ (Dostop: 16. december 2022)

Kategorija	Število besedil
Družba	25.681
Zdravje	131.001
Znanost in tehnologija	109.593
Izobraževanje	44.471
Gospodarstvo	186.385
Okolje	97.852
Prosti čas	218.634
Brez kategorije	5
Skupaj	1.635.614

Ob prvem pregledu korpusa se rezultati zdijo obetavni. Na podlagi že analiziranih leksikografskih podatkov smo opravili poizvedbe v korpusu Trendi in večinoma je kategorizacija potrdila jezikoslovne ugotovitve. Tako se na primer izraz *dvojni dvojček*, ki prihaja iz košarkarske terminologije, pojavlja skoraj izključno v besedilih, katerim je bila pripisana kategorija Šport (99,8 % zadetkov). Podobno *aforizem,* izraz s področja književnosti, pričakovano prevladuje v besedilih, ki jim je bila pripisana kategorija Kultura (67,3 %). Besedna zveza *državni zbor* kaže večjo razpršenost po kategorijah (pojavlja se v 12 od 13 kategorij), vseeno pa po pričakovanjih prevladuje v Politiki (49,4 %); a tudi v ostalih kategorijah najdemo primesi politike, npr. zadetki iz športnih besedil pogosto omenjajo politično potrjevanje zakonov in finančnih sredstev, relevantnih za določeno športno panogo.

5 Sklep

S spremljevalnim korpusom Trendi je jezikovna infrastruktura za slovenščino bogatejša za še en pomemben vir, ki omogoča spremljanje povsem sodobne rabe jezika. Najnovejša različica Trendi 2023-02 pokriva besedila od začetka 2019 do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic iz 107 virov. Izdelani cevovod omogoča dnevno zbiranje besedil, njihovo avtomatsko označevanje in pretvorbo, mesečno pa pretvorbo v format VERT, ki ga zahtevajo konkordančniki repozitorija CLARIN.SI, v katerih je korpus Trendi prosto dostopen.

Avtomatska kategorizacija besedil z novičarskih portalov glede na tematiko se je z uporabljenim naborom kategorij in z naučenimi orodji

izkazala za zelo uspešno – najboljši model pri večini kategorij dosega visoko točnost (nad 90 % oz. celo nad 98 %), do zamenjav med kategorijami pa prihaja redko, kar omogoča zanesljivo pripisovanje metapodatkov in gradnjo visoko zanesljivih tematskih podkorpusov, ki jih je mogoče uporabiti za nadaljnje analize distribucije jezikovne rabe besed. V okviru nadaljnjega dela na projektu bomo rezultate strojne klasifikacije tematik primerjali tudi z ročno pripisanimi oznakami, predvsem zato, da ugotovimo, ali tudi pri človeških označevalcih pogosto prihaja do določenih napak v kategorizaciji, zlasti med podobnimi ali delno prekrivnimi kategorijami (npr. *Zabava* in *Prosti čas* ter *Znanost in tehnologija* in *Okolje*). Upoštevati je treba tudi, da so bili modeli evalvirani na množici, ki je bila uravnotežena po kategorijah, zato je klasifikacija na realnih podatkih morda pogosteje lažno pozitivna za redkejše kategorije, kot je npr. *Okolje*. Ker dejanske porazdelitve besedil v kategorije v celotnem slovenskem medijskem prostoru ne poznamo, je treba uspešnost modela oceniti z ročno evalvacijo. Poleg tega je bil klasifikator razvit le za uporabo na člankih z novičarskih portalov, v prihodnje pa bi podobno metodo veljalo preizkusiti tudi za pripisovanje metapodatkov drugim besedilnim zvrstem, ki jih najdemo npr. v korpusu Gigafida 2.0 (npr. leposlovje, časopisni članki, revije).

Načrti za prihodnost vključujejo predvsem optimizacijo cevovoda, odpravo določenih hroščev (npr. nepričakovani simboli kot posledica napačnega kodiranja besedilnih datotek) in pa redno povečevanje korpusa. Del izboljšav korpusa je tudi zamenjava besedil, ki so bila pridobljena z virov s plačljivimi vsebinami, saj je pogosto brezplačno na voljo samo prvih nekaj odstavkov. Za tovrstna besedila je potrebno urediti pogodbo z besedilodajalci in besedila pridobiti izven obstoječega cevovoda. Korpus je lahko uporaben tudi za kompleksnejše naloge, kot so npr. analiza sloga, identifikacija avtorjev in ocenjevanje kompleksnosti besedila.

Zahvala

Projekt SLED (*Spremljevalni korpus in spremljajoči podatkovni viri*) je financiralo Ministrstvo za kulturo Republike Slovenije kot del *Javnega razpisa za (so)financiranje projektov, namenjenih gradnji in posodabljanju infrastrukture za slovenski jezik v digitalnem okolju 2021–2022*. Raziskovalna programa št. P6-0411 (*Jezikovni viri in tehnologije za*

slovenski jezik) in št. P6-0215 (*Slovenski jezik – bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Bušta, J., Herman, O., Jakubíček, M., Krek, S., & Novak, B. (2017). JSI Newsfeed corpus. *The 9th International Corpus Linguistics Conference*. University of Birmingham.

Caterina, M., Silvia, B., Eugenio, G., Massimo, C., & Francesco, S. (2019). KIParla corpus: a new resource for spoken Italian. *CEUR WORKSHOP PROCEEDINGS*. SunSITE Central Europe.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451).

Cvrček, V., Křen, M., Čermáková, A., Chlumská, L., Škrabal, M., in Kováříková, D. (2020). Overview of text classification in SYN2015. Pridobljeno s https://wiki.korpus.cz/doku.php/en:cnk:klasifikace_textu_syn2015

Čibej, J., Kuzman, T., Ljubešić, N., Kosem, I., Ponikvar, P., Dobrovoljc, K., & Krek, S. (2022). *Text classification model SloBERTa-Trendi-Topics 1.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1709.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA)*. Retrieved from https://www.english-corpora.org/coca/

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447–464.

Davies, M. (2016-). *Corpus of News on the Web (NOW)*. Pridobljeno s https://www.english-corpora.org/now/

Davies, M. (2019-). *The Coronavirus Corpus*. Pridobljeno s https://www.english-corpora.org/corona/

De Smedt, K. (2020). Contagious “Corona” Compounding by Journalists in a CLARIN Newspaper Monitor Corpus. *CLARIN Annual Conference*.

Grobelnik, M., Brank, J., Mladenić, D., Novak, B., & Fortuna, B. (2006). Using DMoz for constructing ontology from data stream. *28th International Conference on Information Technology Interfaces* (pp. 439–444).

Herman, O., & Kovár, V. (2013). Methods for Detection of Word Usage over Time. *RASLAN*.

Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: *Volume 2, Short Papers* (pp. 427–431).

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress* (pp. 105–116). Lorient, France.

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., Šinkec, M., & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1681

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. *Proceedings of the 12th Language Resources and Evaluation Conference*.

Kuzman, T., Čibej, J., Ljubešić, N., Kosem, I., Ponikvar, P., Dobrovoljc, K., & Krek, S. (2022). *Text classification model fastText-Trendi-Topics 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1710

Laitinen, M., Lundberg, J., Levin, M., & Martins, R. M. (2018). The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. *Digital Humanities in the Nordic Countries 3rd Conference*.

Ljubešić, N., & Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1204

Logar, N., Erjavec, T., Krek, S., Grčar, M. in Holozan, P. (2013). Written corpus ccGigafida 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1035

Logar Berginc, N., & Ljubešić, N. (2013). Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, *1*(1), 78–110.

Logar, N., Ljubešić, N., & Erjavec, T. (2015). Kres in Gigafida kot korpusna osnova za slovar: razlike in podobnosti. In M. Smolej (ur.), *Slovnica in slovar – aktualni jezikovni opis* (str. 479–486). Ljubljana: Znanstvena založba Filozofske fakultete.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., ..., & Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Rajapakse, T. C. (2019). *Simple Transformers*. Pridobljeno s https://github.com/ThilinaRajapakse/simpletransformers

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, *13*(1), 65–95.

Štajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenić D., & Grobelnik, M. (2010). A service oriented framework for natural language text enrichment. *Informatica, 34*(3), 307–313.

Trampuš, M., & Novak, B. (2012). Internals of an aggregated web news feed. *Proceedings of 15th Multiconference on Information Society*.

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., in Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. arXiv preprint arXiv:2107.10614. Pridobljeno s https://arxiv.org/pdf/2107.10614.pdf

Monitor Corpus Trendi and Automatic Text Categorization

The paper presents the compilation of the Trendi corpus, the first monitor corpus of Slovene. The current version (Trendi 2023-02) contains texts published between January 2019 and October 2023, with a total of over 700 million tokens (more than 586 million words). The purpose of the corpus is to provide linguists and non-linguists with data on current language use and to enable the monitoring of new words as well as the increase and decline in the use of existing words. In the paper, we present the contents of the corpus and the methods and criteria used in its compilation. The second part of the paper is focused on the development of a tool for categorizing text topics in news articles. The tool was developed specifically for the Trendi corpus but can be used for other corpora containing similar texts. A set of 13 thematic categories was developed for the tool. The set generally follows international standards and categories used in comparable corpora for other languages. Using texts annotated with these categories, we trained multiple language models and achieved a high classification accuracy when categorizing text topics.

Keywords: monitor corpus, automatic text categorization, neologisms, news sites, Slovene

DirKorp: A Croatian Corpus of Directive Speech Acts (v3.0)

Petra BAGO
Faculty of Humanities and Social Sciences, University of Zagreb

Virna KARLIĆ
Faculty of Humanities and Social Sciences, University of Zagreb

In this paper, we present recent developments on a new version (v3.0) of DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), the first Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks, a method of simulated communication that is implemented under pre-set conditions. This method is suitable for researching speech acts due to the ability to collect a great number of examples of such acts of equal propositional content and illocutionary purpose used in the same controlled situations. The presented situations are classified into two categories with regard to the relationship between the participants of the communication act: (1) situations involving interlocutors who are not in a familiar relationship; (2) situations involving interlocutors in a familiar relationship. Assignments of the two categories are organized into four pairs, asking respondents to share a speech act of similar propositional content. The respondents were 100 Croatian speakers, all undergraduate (63%) or graduate students (37%) of the Faculty of Humanities and Social Sciences (University of Zagreb). The corpus has been manually annotated on the speech act level, each speech act containing up to 14 features: (1) respondent ID, (2) familiarity/unfamiliarity, (3) utterance type, (4) directive performative verb in 1st person, (5) illocutionary force, (6) propositional content, (7) T/V form, (8) exhortative, (9) lexical marker of request, (10) lexical marker of apology, (11) lexical marker of gratitude, (12)

Bago, P., Karlić, V.: DirKorp: A Croatian Corpus of Directive Speech Acts (v3.0). Slovenščina 2.0, 11(1): 189–217.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.189-217

honorific title, (13) grammatical mood, and (14) modal verb in 2nd person. It contains 12,676 tokens and 1,692 types. The corpus is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the *Text Encoding Initiative Consortium* (TEI). DirKorp is available for download under the CC BY-SA 4.0 license from GitHub in TEI format. We describe applied pragmatic annotation as well as the structure of the corpus.

Keywords: corpus pragmatics, directive speech acts, DirKorp, Croatian language

1 Introduction

Corpus pragmatics is an interdisciplinary field of study that incorporates linguistic pragmatics and computer science, focusing on the development of natural language corpora in machine-readable form and their application for the purposes of studying pragmatics phenomena in written and spoken language. For a long time, linguists have regarded a corpus approach to language as incompatible with pragmatics (Romero-Trillo, 2008, p. 2). While the corpus approach to studying language implies processing authentic language material by implementing quantitative research methods, pragmatic research is still predominantly of a qualitative nature – based on the researcher's introspection, data obtained by elicitation methods, or an analysis of authentic linguistic material of small size. The application of corpus analysis in the research of pragmatics phenomena represents a major turnaround in the development of pragmatics, primarily because it allows a systematic analysis of language material of large size, and thus the detection of patterns of language use that "fly below the radar" through qualitative analyses (ibid.). In addition, it should be pointed out that the application of new technologies in linguistics, including pragmatics, did not only ensure, facilitate or accelerate numerous research processes but opened the door to a new, different way of thinking about language (Leech, 1992).

The application of corpus methods to large pragmatic corpora allows one to systematically carry out empirically based pragmatic research (Bunt, 2017, p. 327). While the implementation of corpus research can result in minor adjustments to existing theories on the one hand, it can lead to a rethinking of pragmatic concepts and theoretical frameworks on the other, such as the development of the theory of dialogue acts (ibid.).

According to Rühlemann and Aijmer (2015), one of the major methodological problems that corpus pragmatic researchers encounter is the disproportionate relationship between pragmatic functions and language forms by which these functions are expressed. One form can perform multiple pragmatic functions in discourse, while one function can be expressed by different forms, which makes the process of querying a corpus according to the pragmatic function criterion rather difficult. It is for this reason that corpus pragmatic researchers most often investigate conventional speech acts or functions performed by a limited number of language forms (Jucker et al., 2009, p. 4). The aim of this paper is to present the first Croatian corpus of directive speech acts, DirKorp, manually annotated for corpus pragmatic research.

The paper is structured as follows: Section 2 describes selected work related to corpus pragmatic research, Section 3 explores the definition, classification, and research methods of directive speech acts, while the subsequent three sections present the DirKorp corpus. Section 4 gives a description of the developed corpus, Section 5 describes 14 annotation features, and Section 6 presents the structure of the corpus encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2021). Finally, Section 7 contains the conclusions and some directions for future work.

This is a follow-up paper from the conference "Language Technologies and Digital Humanities" held in Ljubljana, Slovenia on 15th–16th September 2022, where we presented DirKorp v2.0. Here we present a new published version of the DirKorp (v3.0) with two additional annotation layers, as well as a new section clarifying the definition, classification, and research methods of directive speech acts (Section 3).

2 Related work

The number of large corpora with systematically implemented pragmatic annotation remains relatively small. Due to a disproportionate relationship between pragmatic functions and the language forms by which these functions are expressed, automatic corpus annotation does not produce satisfactory results. For this reason, only a few researchers have engaged in creating larger corpora of this sort. Generally, for the purposes

of corpus pragmatic research, specialized corpora of smaller size are produced for individual research purposes. In addition, pragmatic research is sometimes carried out on corpora without pragmatic annotation.

An example of a corpus that does not contain pragmatic annotation but was used for pragmatic research is the Birmingham Blog Corpus[1] (Kehoe and Gee, 2007; 2012). In fact, this is a subcorpus of a larger set of corpora being developed at the department *Research and Development Unit for English Studies* at the Birmingham City University. It consists of blog posts and reader comments, and includes some 500 million words in English that were collected between 2000 and 2010. Automatic POS annotation was performed using the Stanford Core NLP tools[2] and included lemma annotations and part-of-speech categories[3] based on the Universal Dependencies framework,[4] while the documents contain metadata of the publication date. Pragmatic research on speech acts has been conducted on this corpus. For example, Lutzky and Kehoe (2017a; 2017b) used it to analyse apologies as speech acts that contain formulaic expressions, which facilitate their querying in a corpus when using the available tools.

Similarly, we (Karlić and Bago, 2021) conducted research on the pragmatic functions and properties of imperatives using corpora without pragmatic annotation. We used hrWaC and srWaC (Ljubešić and Klubička, 2014), two large web corpora of the Croatian and Serbian languages with morphosyntactic annotation. For the purposes of the analysis, an additional pragmatic annotation of a representative sample of verbs in an imperative form was carried out manually. Other corpora of the Croatian spoken and written language with no pragmatic annotation have also been used as a resource for corpus pragmatic research. For example, Hržica, Košutar, and Posavec (2021) used the Croatian Corpus of the Spoken Language of Adults (HrAL) (Kuvač Kraljević and Hržica, 2016) and the Croatian National Corpus of the written language (HNK) (Tadić, 1996) for the search and analysis of connectors and discourse markers.

1 https://www.webcorp.org.uk/wcx/lse/corpora
2 https://stanfordnlp.github.io/CoreNLP/
3 See more about the POS tagset used for the Birmingham Blog Corpus: https://www.webcorp.org.uk/wcx/lse/guide.
4 https://universaldependencies.org/u/pos/index.html

According to Bunt (2017) the majority of corpora with pragmatic annotation contain labels on discourse relationships in written texts and on spoken dialogue acts. An example of such a larger corpus is the Penn Discourse Treebank or PDTB[5] (Prasad et al., 2018) which contains labels on discourse relations, i.e., discourse structure and its semantics. Discourse annotations were added to a subcorpus consisting of texts published in the newspaper *Wall Street Journal* with a total of around 1 million tokens, included in a bigger corpus *Penn Treebank* (PTB). Bunt (2017) states that there are corpora of other languages developed for the purposes of studying the co-occurrence of discourse labels, such as Chinese, Czech, Dutch, German, Hindi, and Turkish – emphasizing that these corpora are manually annotated and of modest size. Additionally, for each corpus a new schema was developed based on various theoretical starting points.

DialogBank[6] (Bunt et al., 2019) is one of a rare dialogue corpus annotated with an ISO 24617-2 standard. It contains already existing dialogue corpora annotated with various schemas. Four corpora are of English, namely HCRC Map Task (Anderson et al., 1991), Switchboard (Godfrey et al., 1992), TRAINS (Allen et al., 1995) and DBOX (Petukhova et al., 2014), and four of Dutch – DIAMOND (Geertzen et al., 2004), OVIS[7], Dutch Map Task (Caspers, 2000) and Schiphol (Prüst et al., 1984). Dialogue act annotation involves segmenting a dialogue into defined grammatical units and augmenting each unit with one or more communicative function labels.

Another example of a corpus with a pragmatic annotation is the *Engineering Lecture Corpus*[8] (Alsop and Nesi, 2013; 2014) which contains 76 transcripts based on hour-long video recordings of engineering lectures held in English at three universities. It is manually annotated for three pragmatic features: humour, storytelling, and summary.[9] Each feature can be augmented with one of the attributes containing additional information that describes the feature in more detail. Further, the corpus contains labels regarding significant breaks, laughter, writing or drawing on the board, etc.

5 https://doi.org/10.35111/qebf-gk47
6 https://dialogbank.lsv.uni-saarland.de/
7 http://www.let.rug.nl/vannoord/Ovis/
8 www.coventry.ac.uk/elc
9 https://www.coventry.ac.uk/research/research-directories/current-projects/2015/engineering-lecture-corpus-elc/annotations-and-mark-ups/

Finally, we present the SPICE-Ireland corpus (*Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland*) (Kallen and Kirk, 2012), a part of a larger set of corpora ICE-Ireland (*International Corpus of English: Ireland Component*) containing pragmatic, discourse, and prosodic features. The corpus contains various types of private and public, formal and informal dialogues and monologues of a length of about 2,000 words, with a size of some 625,000 words. It consists of spoken English. The pragmatic annotation of speech acts is based on Searle's classification (Searle, 1969; 1976): representatives, directives, commissives, expressives, and declaratives.

When it comes to corpus research of speech acts, researchers have two options: (1) to analyse examples from existing corpora of authentic linguistic material, or (2) to analyse examples of elicited linguistic material. In the second case, different types of data completion tests are usually applied, and based on the obtained results smaller custom-made corpora are created for the needs of individual research (and therefore not publicly available). This method is most often used in cross-linguistic, contrastive research, but it is also used in the study of individual languages (e.g., Barron, 2008; Trosborg, 1995). For an overview of pragmatic research of speech acts (including directives) on elicited linguistic material, see, for example, Wojtaszek (2008).

To the best of our knowledge, there exist no publicly available corpora of spoken or written Croatian with pragmatic annotation. So far, Croatian linguists have mostly dealt with speech acts from a theoretical perspective, referring primarily to the Austin's and Searle's theory (cf. Pupovac, 1990; Ivanetić, 1995; Miščević, 2018; Palašić, 2020). However, in recent years the number of research projects based on the qualitative and quantitative analysis of small-sized authentic linguistic materials (from literary texts and advertisements to email messages and political discourse in Croatian and other languages) has been increasing (cf. e.g., Pišković, 2007; Matić, 2011; Franović and Šnajder, 2012; Šegić, 2019).

3 Directive speech acts: definition, classification, and research methods

During verbal communication, speakers express their thoughts in the form of utterances, through which they convey information, express

their emotions and attitudes, or try to modify the addressee's behaviour (Capone, 2009, p. 1015).

Speech acts are utterances with specific properties and communicative functions (ibid.):

> A speech act (...) is not merely the expression of a thought. It is the vocalization of a certain representation of the world (external or internal) aimed at making official the display of an intention to change a state of things and at changing things by the public display of that intention.

Therefore, speech acts can be briefly defined as "actions performed via utterances" (Yule, 2002, p. 47). According to Searle (1975), there are five types of speech acts: (1) representatives – statements that can be evaluated as true or false; (2) directives, which speakers use to influence the addressee's wishes and actions; (3) commissives, through which speakers commit to perform some action in the future; (4) expressives, which speakers use to express their feelings or attitudes; and (5) declaratives or institutionalized declarations that formally change the state of affairs in extralinguistic reality (cf. Karlić and Bago, 2021).

Directive speech acts, or directives, are a type of speech act by which speakers express their "(...) desire/wish for the addressee to do something. (...) In using a directive, the speaker intends to elicit some future course of action on the part of the addressee, thus making the world match the words via the addressee" (Huang, 2009, p. 1004).

Directives differ with respect to their illocutionary force. The illocutionary force of a directive depends on how binding it is for the addressee. If the speaker insists on its realization, the illocutionary force of the directive is strong – and vice versa. According to this criterion, directive speech acts are classified into orders, commands, requests, pleads, incentives, advice, etc. (cf. Piper et al., 2005, p. 1021; Karlić and Bago, 2021, p. 37).

Directive speech acts can be direct or indirect. Direct directives contain an explicit directiveness marker – an imperative (*Close the window*) or a performative verb in the first person of the present tense (*I ask you to close the window*). Directiveness can be expressed implicitly, through assertions without a performative verb (*You should close the window*), interrogative utterances (*Can you close the window?*), or

elliptical utterances (*Um… the window…*). Just like illocutionary force, the propositional content of directive speech acts can also be expressed explicitly and implicitly (*It is cold here* [implicature: *Close the window*]) (cf. Huang, 2009, p. 1005; Karlić and Bago, 2021, p. 39).

According to Brown and Levinson (1987, p. 65–66), directives represent a typical example of face-threatening acts. For this reason, when using them, speakers often apply various politeness strategies that mitigate their illocutionary force (e.g., implicatures and lexical or grammatical modifiers of illocutionary force).

The foundations of speech act theory were laid by the philosophers John Austin and John Searle in works published in the 1960s and 1970s. Since then, numerous studies of speech acts have been conducted. In the beginning, they were non-empirical, based on the researcher's intuition. In recent years, however, the number of empirical studies of speech acts has grown significantly. Jucker (2009) distinguishes three types of data collection methods for the needs of empirical research of speech acts – field, laboratory, and armchair (Flöck and Geluykens, 2015, p. 10):

> While armchair approaches investigate participants' intuitions and attitudes about language use, field and laboratory approaches aim at studying actual language use. They differ, however, in the way language data are produced. While in laboratory approaches, language use is elicited by researchers (by employing role-plays or administering discourse completion tasks), field data are defined by the absence of such elicitation techniques. Field methods are therefore observational in nature, i.e., they require an authentic communicative intent by participants to produce language.

Each of the mentioned methods has its advantages and disadvantages. For the purposes of creating the DirKorp corpus, we applied the laboratory method of eliciting language production by role-playing. The main advantage of this method is that it gives "full variable control to the researcher (…), and can generate large amounts of data; however, participants use language without their own intrinsic communicative intent in fictional scenarios" (Flöck and Geluykens, 2015, p. 11). This method allowed us to collect a large amount of mutually comparable

directives with the same propositional content and produced in the same controlled circumstances.

In the following sections, we present a new version (v3.0) of DirKorp, the first Croatian corpus of directive speech acts.

4 Corpus description

DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*) (Karlić and Bago, 2021) is a Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks applying the method of simulated communication that is implemented under pre-set conditions. This method is suitable for researching speech acts due to the ability to collect a great number of examples of speech acts of equal propositional content and illocutionary purpose used in the same controlled situations. The questionnaire included eight closed-type role-playing tasks. These types of tasks imply recording the speaker's reactions (in this case in writing) to the stimulus without feedback. In each task, the participants are presented with one textually described hypothetical situation asking them to refer a directive speech act to their interlocutor. Their assignment was to imagine they were in the presented situation and to give a written statement they would use in the described situations. The presented situations are classified into two categories with regard to the relationship between the participants of the communication act: (1) situations involving interlocutors who are not in a familiar relationship (i.e., interlocutors who are not close and are not equal in terms of power relations, and communicate in more or less in/formal situations); (2) situations involving interlocutors in a familiar relationship (i.e., interlocutors in a close and equal relationship who communicate in more or less in/formal situations). Assignments of the two categories are organized into four pairs, asking respondents to share a speech act of similar propositional content: "I want you to return something that belongs to me" (for text of this role-playing task pair see Example 1 when interlocutors have (a) an unfamiliar relationship (label "NEFAM1") and (b) a familiar relationship (label "FAM1")); "I want you to answer my inquiry" (for text of this role-playing task pair see Example 2 when interlocutors have (a) an unfamiliar relationship (label "NEFAM2")

and (b) a familiar relationship (label "FAM2")); "I want you to change something that bothers me" (for text of this role-playing task pair see Example 3 when interlocutors have (a) an unfamiliar relationship (label "NEFAM3") and (b) a familiar relationship (label "FAM3")); "I want you to stop behaving inappropriately" (for text of this role-playing task pair see Example 4 when interlocutors have (a) an unfamiliar relationship (label "NEFAM4") and (b) a familiar relationship (label "FAM4"))[10].

Example 1

(a) Upravo si pojeo/la ručak u restoranu. Posluživao te stariji konobar koji se odnosio prema tebi ljubazno i profesionalno. Prilikom plaćanja računa konobar ti vraća 100 kuna manje nego što je trebao. Želiš da ti konobar vrati novac. Zamisli da se konobar nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You just ate lunch at a restaurant. You were served by an elderly waiter who treated you kindly and professionally. When paying the bill, the waiter refunds you 100 kunas less than he should have. You want the waiter to give you your money back. Imagine the waiter was in front of you and write what exactly you would say to him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

(b) Posudio/la si knjigu najboljem prijatelju (ili prijateljici). Rekao ti je da će ti je uskoro vratiti, no nije održao riječ. Sjedite zajedno u kafiću, situacija je opuštena, razgovarate o svakodnevnim stvarima. Želiš mu dati do znanja da ti treba čim prije vratiti knjigu. Zamisli da se tvoj prijatelj nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You lent a book to your best friend. (S)he told you (s)he'd give it back to you soon, but (s)he didn't keep her/his word. You are sitting together in a café, the situation is relaxed, you talk about everyday things. You want to let her/him know you need to get your book back as soon as possible. Imagine your friend was in front of you and write what exactly you would say to her/him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

10 Full texts of role-playing tasks are available in the corpus header as well.

Example 2

(a) Poslao/la si e-mail profesoru s upitom možeš li pohađati njegov izborni kolegij i hitno trebaš njegov odgovor i potvrdu u mailu. Međutim, profesor ne odgovara već tjedan dana, a rok za upis završava sutradan. Želiš ponovno zatražiti njegovu povratnu informaciju. Napiši kratak e-mail profesoru kakav bi mu uputio/la u navedenoj situaciji.
(Eng. *You sent an email to the professor asking if you can attend his elective course and urgently need his response and confirmation in the email. However, the professor has not responded for a week, and the admission deadline ends the next day. You want to ask for his feedback again. Write a short email to the professor as you would in this situation.*)

(b) Poslao/la si poruku (WhatsApp, Viber, Messenger) najboljem prijatelju (ili prijateljici) s pozivom na druženje sljedeće večeri. On je vidio poruku, ali nije odgovorio do sutradan. Želiš da ti odgovori čim prije kako bi mogao/la isplanirati ostatak dana. Napiši kratku poruku prijatelju kakvu bi mu uputio/la u navedenoj situaciji.
(Eng. *You sent a message (WhatsApp, Viber, Messenger) to your best friend with an invitation to hang out the next night. (S)he saw the message but did not respond until the next day. You want her/him to reply as soon as possible so you can plan the rest of the day. Write a short message to your friend as you would in this situation.*)

Example 3

(a) Voziš se u taksiju. Prozori su otvoreni i želiš da ih taksist zatvori jer ti je hladno. Zamisli da se nalaziš u navedenoj situaciji i napiši što bi točno rekao/la taksistu (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're riding in a cab. The windows are open, and you want the taxi driver to close them because you're cold. Imagine that you are in this situation and write down what exactly you would say to the taxi driver (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

(b) Voziš se u autu na suvozačkom mjestu. Vozač je tvoj najbolji prijatelj (ili prijateljica). Budući da vozi prebrzo i gleda u mobitel, ne osjećaš se ugodno i želiš da uspori. Zamisli da se nalaziš u danoj situaciji i napiši što bi mu točno rekao/la (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You are riding in the car in the passenger seat. The driver is your best friend. Because (s)he's driving too fast and looking at her/his cell phone, you don't feel comfortable and want her/him to slow down. Imagine that you are in a given situation and write what exactly you would say to her/him (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

Example 4

(a) Nalaziš se u dućanu i čekaš u redu pred blagajnom. Velika je gužva. Ispred tebe se u red ugura gospođa srednje dobi. Ljudi u redu iza tebe negoduju jednako kao i ti. Želiš da gospođa stane na kraj reda. Zamisli da se nalaziš u danoj situaciji i napiši što bi točno rekao/la gospođi (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're in the store waiting in line at the cash register. It's crowded. A middle-aged lady squeezes in front of you. The people in line behind you are just as resentful as you are. You want the lady to stand at the end of the line. Imagine that you are in a given situation and write down what exactly you would say to the lady (do not recount but formulate the statement as if you were directly addressing the interlocutor).*)

(b) Slušaš predavanje na fakultetu. Sjediš pored dvoje kolega s kojima si inače vrlo blizak/bliska. U jednom trenutku oni počinju glasno razgovarati i smijati se. Njihov razgovor ti smeta jer ne možeš pratiti predavanje, a i nastavnik pogledava u vašem smjeru. Želiš da prestanu. Zamisli da se nalaziš u danoj situaciji i napiši što bi im točno rekao/la (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).
(Eng. *You're listening to a lecture in college. You're sitting next to two colleagues with whom you are otherwise very close. At some point, they start talking loudly and laughing. Their conversation bothers you because you can't follow the lecture, and the lecturer looks in your direction. You want them to stop. Imagine that you are in a given situation and write down exactly what you would say to them (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

The respondents were 100 Croatian speakers, all undergraduate (63%) or graduate students (37%) of the Faculty of Humanities and

Social Sciences (University of Zagreb), aged between 18 to 33, with Croatian being the native language for the majority (96%). The questionnaire was administered in December 2020 and January 2021. Before completing the questionnaire, all the respondents were informed of the purposes of the study as well as what data would be collected. All the respondents voluntarily participated in the study and were made aware that they could withdraw from it at any time. By choosing to participate in the study, respondents gave informed consent for their data to be processed for the stated research purposes. The questionnaire was administered anonymously via an online survey, and the language material collected was used exclusively for research purposes.

The elicitation of language production by the role-playing method has its advantages and disadvantages. On the one hand, it enables the collection of a large number of speech acts with the same propositional content and illocutionary purpose. On the other hand, users of the corpus should keep in mind that the language material collected by this method does not reflect the features of actual language use, but instead shows what speakers think they would say and/or do in hypothetical situations.

DirKorp contains 12,676 tokens and 1,692 types.[11] Since it consists of 800 speech acts, it is a relatively small corpus compared to some of the corpora with pragmatic annotation presented in Section 3. However, as the first Croatian corpus with detailed pragmatic annotation, DirKorp can serve as a useful resource for researching the characteristics of speech acts on a formal and content level, the application of politeness strategies in communication in different situations, and the properties of other grammatical-pragmatic and lexical-pragmatic phenomena in the Croatian language that are annotated in the corpus. In addition, we believe that DirKorp can serve as a complement to research on speech acts that are conducted on authentic language materials and as a starting point for conducting contrastive research on the characteristics and use of speech acts in other languages. In addition, we hope that it will contribute to the development of larger corpora of

11 Respondents' answers contain utterances but also text about what they would do in the given situation. At this moment, the corpus contains no annotation of utterances of speech acts, and therefore we cannot analyse the average length of a response. Generally, we can only state that some speech acts contain only one utterance, while some contain more than one.

the Croatian language with pragmatic annotation, and that such work will encourage a wider application of the corpus-pragmatic research method.

In Karlić and Bago (2021), we have conducted corpus pragmatic analyses of the collected speech acts to investigate ways and means of expressing directives, and their pragmatic characteristics and functions. For example, we confirmed that indirect directives are more frequent than direct ones, especially among interlocutors who are not in a familiar relationship. Regarding a(n) (un)familiar relationship between interlocutors, we detected that explicit illocutionary force is more frequent in communication between interlocutors with a familiar relationship, while implicit illocutionary force is more frequent in communication between interlocutors with an unfamiliar relationship. Additionally, we have identified that imperative utterances are a more frequent type of direct directives than utterances with a directive performative verb in 1st person. For more such corpus pragmatic analyses see Karlić and Bago (2021).

5 Corpus annotation

The collected language material has been manually annotated on the speech act level by two independent annotators[12] with university graduate degrees in the field of philology. Annotators received oral and written instructions, including illustrative examples for all the features they had to annotate.

Basic categorization of speech acts (directive; direct and indirect; explicit and implicit) and their formal and pragmatic properties (i.e., performative verbs) was carried out according to the theory of speech acts by Austin (1962), Searle (1969; 1976), and their successors. The features and components of speech acts related to the phenomenon

12 When comparing the annotations of two annotators, in all categories, disagreements were found in at most 1.2% of examples and were mostly the result of accidental mistakes by one of the annotators. Once such mistakes were corrected, a consensus was reached among the annotators. The only category in which the disagreement was higher (2.5%) was the category "Illocutionary force". In most cases, these were examples with generalized conversational implicature (one annotator marked speech acts with this type of implicature as explicit, and another as implicit). Based on the instruction to label speech acts with all types of implicature as implicit, a consensus was reached among the annotators for this category as well.

of politeness (familiarity, use of T/V forms, and certain lexical modifiers of the illocutionary force of speech acts) are taken according to the politeness theory of Brown and Levinson (1987), while the grammatical characteristics (utterance type, grammatical mood, modal verbs) of speech acts are categorized according to the grammatical descriptions of contemporary Croatian and Serbian languages (Silić and Pranjković, 2007; Piper et al., 2005). For more on individual categories, see Karlić and Bago (2021). In the new version of DirKorp (v3.0), each speech act can contain up to 14 features. The first eight features were part of the corpus version v1.0, features nine to 12 were part of v2.0, while features 13 and 14 were newly added. Appendix A contains the frequency distribution of features two to 14. For a more detailed frequency distribution of all features see Karlić and Bago (2021).

(1) **Respondent ID** – This mandatory feature contains information on the identification of the respondent uttering the speech act.

(2) **Familiarity/unfamiliarity** – This mandatory feature contains information on the category of the proposed situation in which the speech act was uttered. Four situations are labelled 'unfamiliar' (involving interlocutors who are not in a familiar relationship), while the other four situations are labelled 'familiar' (involving interlocutors who are in a familiar relationship).

(3) **Utterance type** – This mandatory feature contains information on the utterance type regarding its structural organization. It contains six labels: (a) an imperative utterance, (b) an assertive utterance (a statement), (c) an utterance in the form of a question, (d) an utterance in the form of a predicate ellipsis[13], (e) a nonverbal signal, (f) a case of avoidance of executing a speech act (see Example 5).

13 Utterances in the form of a predicate ellipsis were singled out as a separate category due to: (1) the absence of a verb (and potentially other components of the sentence structure) and therefore the default indirectness and implicitness of the speech act, which makes them incomparable to other utterances in the corpus; (2) impossibility to determine the type of utterance for all examples due to their elliptical structure.

Example 5

(a) E vrati mi onu knjigu koju sam ti posudio.
(Eng. *Hey, give me back that book I lent you.*)
(b) Oprostite, ali mislim da ste mi krivo vratili novce.
(Eng. *Excuse me, but I think you gave me my money back wrong.*)
(c) Možete li molim vas zatvoriti prozore?
(Eng. *Could you please close the windows?*)
(d) E, moja knjiga??
(Eng. *Hey, my book??*)
(e) [Samo bih zavrtjela očima da vide moje neodobravanje, ali ne bih ništa rekla.][14]
(Eng. *[I'd just roll my eyes so that they see my disapproval, but I wouldn't say anything.]*)
(f) [Ne bih ništa rekao.]
(Eng. *[I wouldn't say anything.]*)

(4) **Directive performative verb in 1st person** – This optional feature contains information on the representation of a directive performative verb in 1st person as part of the speech act, only for assertive utterances and utterances in the form of a question. It contains two labels: (a) yes and (b) no (see Example 6).

Example 6

(a) Oprostite, molim da odete na kraj reda.
(Eng. *Excuse me, I am imploring you to go to the end of the line.*)
(b) Gospođo, morate na kraj reda stati.
(Eng. *Madam, you must move to the end of the line.*)

(5) **Illocutionary force** – The optional feature contains information on the explicitness or implicitness of the illocutionary force of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 7).

14 Descriptions of non-verbal situations can be found in Example 5 (e) and (f). All other examples contain actual utterances. DirKorp v3.0 does not contain annotations of utterances. Therefore, it is currently not possible to filter the speech acts with regard to actual utterances or descriptions of non-verbal situations.

Example 7

(a) Daj mi donesi više onu knjigu, treba mi!
(Eng. *Bring me that book already, I need it!*)
(b) Kaj je s onom knjigom koju sam ti posudio?
(Eng. *What happened to that book I lent you?*)

(6) **Propositional content** – This optional feature contains information on the explicitness or implicitness of the propositional content of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 8).

Example 8

(a) Gledaj na cestu, pusti mobitel.
(Eng. *Look at the road, leave the cell phone.*)
(b) Ti hoćeš da poginemo?
(Eng. *You want us to die?*)

(7) **T/V form** – This optional feature contains information on how the respondent addressed the interlocutor, using an informal (T-form) or a formal *you* (V-form). It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains three labels: (a) T-form, (b) V-form, and (c) impossible to determine (see Example 9).

Example 9

(a) Oprosti, dao si mi manje novca
(Eng. *Sorry*$_{T\text{-}form}$*, you*$_{T\text{-}form}$ *gave me less change.*)
(b) Oprostite, mislim da ste mi ipak još dužni 100 kuna.
(Eng. *Excuse*$_{V\text{-}form}$ *me, I think you*$_{V\text{-}form}$ *still owe me 100 kunas.*)
(c) Hmm... još 100 kuna, zar ne?
(Eng. *Hmm... another 100 kunas, right?*)

(8) **Exhortative** – This optional feature contains information on the representation of an exhortative as part of the speech act (a lexical

mean used to express encouragement, i.e., incentive particles). It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) yes and (b) no (see Example 10).

Example 10

(a) Daj mi više vrati knjigu, treba mi za knjižnicu.
(Eng. *Bring me back my book already, I need it for the library.*)

(b) Jel se sjećaš one knjige koju sam ti posudila? Potrebna mi je. Možeš li mi ju donijeti sutra na faks?
(Eng. *Do you remember that book I lent you? I need it. Could you bring it tomorrow to uni?*)

(9) **Request** – This optional feature contains information on whether the speech act includes a lexical marker of request (e.g., "please"). It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) yes and (b) no (see Example 11).

Example 11

(a) E da, jel bi mi mogao/la vratiti knjigu, molim te?
(Eng. *Oh yeah, could you bring the book back, please?*)
(b) Zaboravio si mi vratiti knjigu, jel se možeš idući put sjetiti?
(Eng. *You forgot to bring me back the book, can you remember next time?*)

(10) **Apology** – This optional feature contains information on whether the speech act includes a lexical marker of apology. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) yes and (b) no (see Example 12).

Example 12

(a) Oprostite, ovdje fali još 100 kuna
(Eng. *Excuse me, 100 kunas are missing here.*)

(b) Možete li molim vas pritvoriti prozore, hladno mi je?
(Eng. *Could you please close the windows, I'm cold?*)

(11) **Gratitude** – This optional feature contains information on whether the speech act includes a lexical marker of gratitude. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) yes and (b) no (see Example 13).

Example 13

(a) Molim te mi samo javi da znam zbog organizacije hoćeš li doći. Hvala ti!
(Eng. *Please just let me know whether you're coming so that I know because of the organization. Thank you!*)

(b) Heej, jel dolaziš večeras na druženje? Moram znati zbog organizacije. xoxo
(Eng. *Heeey, are you coming tonight to hang out? I need to know because of the organization. xoxo*)

(12) **Honorific title** – This optional feature contains information on whether the speech act includes an honorific title. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question, and in the form of an ellipsis). It contains two labels: (a) yes and (b) no (see Example 14).

Example 14

(a) Gospođo, kraj reda je dolje.
(Eng. *Madam, the end of the line is back there.*)

(b) Oprostite, tamo je kraj reda!

(Eng. *Excuse me, the end of the line is there!*)

(13) **Grammatical mood** – This optional feature contains information on grammatical mood used in a speech act. It is only applied to indirect speech acts (assertive utterances and utterances in the form of a question) since it is understood that direct imperative

speech acts contain verbs in the imperative mood. Accordingly, this feature contains two labels: (a) indicative mood and (b) conditional mood (see Example 15).

Example 15

(a) Oprostite, ali ovo nije kraj reda.
(Eng. *Excuse me, but this is not the end of the line*.)
(b) Oprostite jel bi mogli zatvorit prozore? Malo mi je hladno.
(Eng. *Excuse me, could you close the windows? I'm a little cold*.)

(14) **Modal verb in 2nd person** – This optional feature contains information on the representation of modal verb in 2nd person as part of a speech act. It is only applied to indirect speech acts (an assertive utterance and an utterance in the form of a question). It contains two labels: (a) yes and (b) no (see Example 16).

Example 16

(a) Oprostite, mislim da je došlo do pogreške, trebate mi vratiti još 100 kuna.
(Eng. *Sorry, I think there was a mistake, you have to return another 100 kunas*.)
(b) Malo je hladno ovdje, možemo možda zatvoriti prozor?
(Eng. *It's a little cold in here, can we possibly close the window?*)

6 Corpus format

DirKorp is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the Text Encoding Initiative Consortium (TEI) (TEI Consortium, 2021). The TEI document is comprised of a header and the body of the corpus. The content of the elements and attributes are in Croatian. The metadata of the corpus is given in the header including bibliographic information; the editorial practice; a structured taxonomy describing categories used for each of the 14 pragmatic features in the annotation process (see Figure 1 for an example), including the full text of the eight situations on the questionnaire; a list of questionnaire participants with information on their age, gender, undergraduate or graduate level of

study, enrolment in a philological/non-philological/combined study program and native language (see Figure 2 for an example); and a list of revisions of the DirKorp versions. The body of the corpus is composed of one division containing utterances with pragmatic features (see Figure 3 for an example).

DirKorp is available for download under the CC BY-SA 4.0 license from GitHub in TEI format (https://github.com/pbago/DirKorp).

```
<taxonomy xml:id="tiVi">
  <category xml:id="ti">
    <catDesc>Govorni čin sadržava obraćanje na ti (atribut se
    odnosi na tipove iskaza koji uključuju verbalna sredstva
    [imperativni, tvrdnja, upitni, eliptični]).</catDesc>
  </category>
  <category xml:id="vi">
    <catDesc>Govorni čin sadržava obraćanje na Vi (atribut se
    odnosi na tipove iskaza koji uključuju verbalna sredstva
    [imperativni, tvrdnja, upitni, eliptični]).</catDesc>
  </category>
  <category xml:id="persNeodredivo">
    <catDesc>Nije moguće odrediti sadržava li govorni čin
    obraćanje na ti ili Vi (atribut se odnosi na tipove iskaza
    koji uključuju verbalna sredstva [imperativni, tvrdnja,
    upitni, eliptični]).</catDesc>
  </category>
</taxonomy>
```

Figure 1: An example of a pragmatic feature description – how the respondent addressed the interlocutor (V-form, T-form, or impossible to determine, annotation feature 7 from Section 5).

```
<person xml:id="I001" sex="F">
  <p>ispitanik/ispitanica, 20 godina, spol Ž, preddiplomski
  studij Filozofskog fakulteta, nefilološko usmjerenje, materinji
  jezik hrvatski</p>
</person>
```

Figure 2: An example of participant information.

```
<u who="#I001" ana="#NEFAM1 #tvrdnja #dpg1N #isI #psI #vi
#adhorativN #molbaN #isprikaY #zahvalaN #honorifikN #gnI
#mg2N">Ispričavam se, pardon, fali još sto kuna. Oprostite.</u>
```

Figure 3: An example of a speech act containing all 14 pragmatic features.

7 Conclusion and future work

In this article we have presented DirKorp v3.0, the first Croatian corpus of directive speech acts, containing 800 elicited speech acts collected via an online questionnaire with role-playing tasks, specifically developed for pragmatic research studies. The respondents were 100 Croatian speakers, all students of the Faculty of Humanities and Social Science (University of Zagreb). The corpus has been manually annotated on the level of a speech act, with each speech act containing up to 14 features. It contains 12,676 tokens and 1,692 types. The corpus is available for download under the CC BY-SA 4.0 license from GitHub in TEI format.

Further work is planned on the corpus, which includes an evaluation of the developed schema for annotating directive speech acts (e.g., test-retest reliability on a sample of data to evaluate stability and consistency of the schema, domain experts reviewing the schema to determine if it adequately captures the relevant aspects of the data, reviewing the adequacy of encoding choices regarding attributes and its values), annotation at the levels smaller than a speech act, as well as augmentation with additional features such as information on various politeness strategies applied in a speech act.

Acknowledgments

This paper is generously co-financed by the institutional project of the Faculty of Humanities and Social Sciences (University of Zagreb) “South Slavic languages in use: pragmatic analyses” (principle researcher Virna Karlić). We also wish to thank our annotators for the time and effort.

References

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., …, & Traum, D. R. (1995). The TRAINS Project: A Case Study in Building a Conversational Planning Agent. *Journal of Experimental & Theoretical Artificial Intelligence, 7*(1),7–48.

Alsop, S., & Nesi, H. (2013). Annotating a Corpus of Spoken English: The Engineering Lecture Corpus (ELC). In *Proceedings of GSCP 2012: Speech and Corpora* (pp. 58–62). Firenze University Press, Florence.

Alsop, S., & Nesi, H. (2014). The Pragmatic Annotation of a Corpus of Academic Lectures. In *The International Conference on Language Resources and Evaluation 2014 Proceedings* (pp. 1560–1563). Reykjavik: European Language Resources Association.

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., ..., & Weinert, R. (1991). The HCRC Map Task Corpus, *Language and Speech*, *34*(4), 351–366.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford: Clarendon Press.

Barron, A. (2008). The structure of requests in Irish English and English English. In K. P. Schneider & A. Barron (Eds.), *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages* (pp. 35–68). John Benjamins Publishing Company.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Bunt, H. (2017). Computational Pragmatics. In *Oxford Handbook of Pragmatics* (pp. 326–345). Oxford University Press, New York.

Bunt, H., Petukhova, V., Malchanau, A., Fang, A. & Wijnhoven, K. (2019). The DialogBank: Dialogues with Interoperable Annotations. In *Language Resources and Evaluation*, *53*(2), 213–249.

Capone, A. (2009). Speech Acts, Classification and Definition. In *Concise Encyclopedia of Pragmatics* (pp. 1015–1017). Oxford: Elsevier.

Caspers, J. (2000). Melodic Characteristics of Backchannels in Dutch Map Task Dialogues. In *Proceedings, 6th International Conference on Spoken Language Processing* (pp. 611–614). Beijing: China Military Friendship Publish,. Retrieved from https://www.isca-speech.org/archive/icslp_2000/

Flöck, I., & Geluykens, R. (2015). Speech Acts in Corpus Pragmatics: A Quantitative Contrastive Study of Directives in Spontaneous and Elicited Discourse. In *Yearbook of Corpus Linguistics and Pragmatics* (pp. 7–37). Springer International Publishing.

Franović, T., & Šnajder, J. (2012). Speech Act Based Classification of Email Messages in Croatian Language. In *Proceedings of the Eighth Language Technologies Conference* (pp. 69–72). Ljubljana: Information Society.

Geertzen, J., Girard, Y., Morante, R., Van der Sluis, J., Van Dam, H., Suijkerbuijk, B., Van der Werf, R., & Bunt, H. (2004). The DIAMOND Project. In: *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004),* Barcelona.

Godfrey, J., Holliman, E. & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 517–520). San Francisco: IEEE Computer Society.

Hržica, G., Košutar, S., & Posavec, K. (2021). Konektori i druge diskursne oznake u pisanome i spontanome govorenom jeziku. *Fluminensia: časopis za filološka istraživanja*, *33*(1), 25–52.

Huang, Y. (2009). Speech Acts. In *Concise Encyclopedia of Pragmatics* (pp. 1000–1009). Oxford: Elsevier.

Ivanetić, N. (1995). *Govorni činovi*. Zagreb: FF-press, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Jucker, A. H. (2009). Speech Act Research between Armchair, Field and Laboratory: The Case of Compliments. *Journal of Pragmatics*, *41*, 1611–1635.

Jucker, A. H., Schreier, D., & Hundt, M. (Eds.). (2009). *Corpora: Pragmatics and Discourse*. Rodopi, Amsterdam.

Kallen, J. L., & Kirk, J. M. (2012). *SPICE-Ireland: A User's Guide*. Retrieved from https://pure.qub.ac.uk/en/publications/spice-ireland-a-users-guide

Karlić, V., & Bago, P. (2021). *(Računalna) pragmatika: temeljni pojmovi i korpusnopragmatičke analize*. Zagreb: FF Press. Retrieved from https://openbooks.ffzg.unizg.hr/index.php/Ffpress/catalog/book/125.

Kehoe, A., & Gee, M. (2007). New Corpora from the Web: Making Web Text More 'Text-Like'. In *Studies in Variation, Contacts and Change in English 2*. Retrieved from https://varieng.helsinki.fi/series/volumes/02/kehoe_gee/

Kehoe, A., & Gee, M. (2012). Reader Comments as an Aboutness Indicator in Online Texts: Introducing the Birmingham Blog Corpus. In: *Studies in Variation, Contacts and Change in English 12*. Retrieved from https://varieng.helsinki.fi/series/volumes/12/kehoe_gee/

Kuvač Kraljević, J., & Hržica, G. (2016). Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: časopis za filološka istraživanja*, *28*(2), 87–102.

Leech, G. N. (1992). Corpora and Theories of Linguistic Performance. In *Directions in Corpus Linguistics* (pp. 105–122). De Gruyter, Berlin.

Ljubešić, N., & Klubička, F. (2014). {bs, hr, sr}WaC-Web Corpora of Bosnian, Croatian and Serbian. In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Association for Computational Linguistics, Gothenburg. Retrieved from https://aclanthology.org/W14-0405.pdf

Lutzky, U., & Kehoe, A. (2017a). I Apologize for My Poor Blogging: Searching for Apologies in the Birmingham Blog Corpus. *Corpus Pragmatics*, *1*(1), 37–56.

Lutzky, U., & Kehoe, A. (2017b). Oops, I Didn't Mean to Be so Flippant. A Corpus Pragmatic Analysis of Apologies in Blog Data. *Journal of Pragmatics*, *116*, 27–36.

Matić, D. (2011). *Govorni činovi u političkome diskursu*. PhD thesis. Zagreb: Faculty of Humanities and Social Sciences.

Miščević, N. (2018). *Rođenje pragmatike*. Orion Art, Beograd.

Palašić, N. (2020). *Pragmalingvistika – lingvistički pravac ili petlja?* Zagreb: Hrvatska sveučilišna naklada.

Petukhova, V., Gropp, M., Klakow, D., Eigner, G., Topf, M., Srb, S., Motlicek, P., ... Potard, ..., & Schmidt, A. (2014). The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 252–258). European Language Resources Association, Reykjavik.

Piper, P. et al. (2005) = Предраг Пипер, Ивана Антонић, Бранислава Ружић, Срето Танасић, Људмила Поповић, Бранко Тошовић. 2005. *Синтакса савременог српског језика*. Проста реченица, Београд: Институт за српски језик САНУ, Београдска књига, Матица српска.

Pišković, T. (2007). Dramski diskurs između pragmalingvistike i feminističke lingvistike. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 33*(1), 325–341.

Prasad, R., Webber, B., & Lee, A. (2018). Discourse Annotation in the PDTB: The NextGeneration. In: *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 87–97). Santa Fe: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-4710.pdf

Prüst, H., Minnen, G. & Beun, R. (1984). Transcriptie dialooogesperiment juni/juli 1984, *IPORapport 481*. Eindhoven: Institute for Perception Research, Eindhoven University of Technology.

Pupovac, M. (1990). *Jezik i djelovanje*. Zagreb: Biblioteka časopisa Pitanja.

Romero-Trillo, J. (Ed.). (2008). *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. De Gruyter, Berlin.

Rühlemann, C., & Aijmer, K. (2015). Introduction. Corpus pragmatics: laying the foundations. In: *Corpus pragmatics* (pp. 1–28).

Searle, J. R. (1969). *Speech Acts*. Cambridge University Press, Cambridge.

Searle, J. R. (1975). A Taxonomy of Speech Acts. In: *Minnesota Studies in the Philosophy of Science* (Vol. 9, pp. 344–369). University of Minnesota Press.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, *5*, 1–23.

Silić, S. & Pranjković, I. (2007). *Gramatika hrvatskoga jezika za gimnazije i visoka učilista*. Zagreb: Školska knjiga.

Šegić, T. (2019). Tata kupi mi auto und Nivea Milk weil es nichts Besseres für die Hautpflege gibt. *Filologija*, *73*, 103–116.

Tadić, M. (1996). Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika, 41–42*, 603–611.

TEI Consortium (Ed.). (2021). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Trosborg, A. (1995). *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. Berlin; New York: Mouton de Gruyter.

Wojtaszek, A. (2016). Thirty years of Discourse Completion Test in Contrastive Pragmatics research, *Linguistica Silesiana, 37*, 161–173.

Yule, G. (2002). *Pragmatics*. Oxford, New York: Oxford University Press.

DirKorp: hrvaški korpus direktivnih govornih dejanj (v3.0)

V prispevku predstavljamo razvoj nove različice (v3.0) korpusa DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), prvega hrvaškega korpusa direktivnih govornih dejanj, ki je bil izdelan za namene raziskav pragmatike. Korpus vsebuje 800 govornih dejanj, ki so bila zbrana s spletnim vprašalnikom z nalogami igranja vlog – gre za metodo stimulirane komunikacije, ki poteka pod vnaprej določenimi pogoji. Metoda je primerna za raziskovanje govornih dejanj, saj lahko na ta način zberemo veliko število primerov z enako propozicijsko vsebino in ilokucijskim namenom, ki so uporabljeni v enaki kontrolirani situaciji. Predstavljene situacije razdelimo v dve kategoriji glede na odnos med udeleženci komunikacijskega dejanja: (1) situacije, ki vključujejo sogovorce, ki niso v sorodstvenem razmerju; (2) situacije z govorci v sorodstvenem razmerju. Naloge v obeh kategorijah so razdeljene v štiri pare, od sodelujočih pa zahtevajo, da pripišejo govorno dejanje s podobno propozicijsko vsebino. V vprašalniku je sodelovalo 100 govorcev hrvaščine; vsi so bili dodiplomski (63 %) ali podiplomski študenti (37 %) Fakultete za humanistiko in družbene vede (Univerza v Zagrebu). Korpus je bil ročno označen na ravni govornih dejanj, vsako dejanje pa vsebuje do 14 značilnosti: (1) ID sodelujočega, (2) sorodstveno/nesorodstveno razmerje, (3) tip izjave, (4) direktivni performativni glagol v prvi osebi, (5) ilokucijska sila, (6) propozicijska vsebina, (7) tikanje/vikanje, (8) prepričevalnost, (9) leksikalni označevalec za prošnjo, (10) leksikalni označevalec za opravičilo, (12) naziv spoštovanja, (13) slovnični naklon, (14) modalni glagol v drugi osebi. Korpus vsebuje 12.676 pojavnic in 1.692 različnic, enkodiran pa je v skladu s smernicami *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, ki jih razvija in vzdržuje konzorcij *Text*

Encoding Initiative Consortium (TEI). DirKorp je v formatu TEI na voljo za prenos pod licenco CC BY-SA 4.0 na platformi GitHub. V prispevku opišemo označevanje in strukturo korpusa.

Ključne besede: korpusna pragmatika, direktivna govorna dejanja, DirKorp, hrvaški jezik

Appendix A: Frequency distribution of annotated features 2-14

		Utterance type	∑	Directive performative verb in 1st person		Illocutionary force		Propositional content		T/V form		
				Yes	No	Explicit	Implicit	Explicit	Implicit	T-form	V-form	to determine
	NEFAM1	Imperative	2	N/A	N/A	2	0	0	2	0	2	0
		Assertive	88	3	85	3	85	11	77	2	86	0
		Question	10	0	10	0	10	0	10	0	9	1
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
A		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	FAM1	Imperative	22	N/A	N/A	22	0	21	1	22	0	0
		Assertive	15	1	14	1	14	7	8	15	0	0
		Question	60	0	60	0	60	33	27	60	0	0
		Ellipsis	3	N/A	N/A	0	3	2	1	2	0	1
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
B		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	NEFAM2	Imperative	0	N/A	N/A	0	0	0	0	0	0	0
		Assertive	87	39	48	39	48	56	31	0	87	0
		Question	13	0	13	0	13	12	1	0	87	0
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
C		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	FAM2	Imperative	40	N/A	N/A	40	0	38	2	40	0	0
		Assertive	8	2	6	2	6	4	4	8	0	0
		Question	46	0	45	0	46	5	41	45	0	1
		Ellipsis	3	N/A	N/A	0	3	1	2	1	0	2
		Nonverbal signal	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D		Avoidance	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	NEFAM3	Imperative	2	N/A	N/A	2	0	2	0	0	2	0
		Assertive	1	0	1	0	1	1	0	0	1	0
		Question	96	1	95	1	95	96	0	0	95	1
		Ellipsis	0	N/A	N/A	0	0	0	0	0	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
E		Avoidance	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	FAM3	Imperative	86	N/A	N/A	86	0	80	6	86	0	0
		Assertive	4	0	4	0	4	1	3	4	0	0
		Question	9	0	9	0	9	5	4	9	0	0
		Ellipsis	1	N/A	N/A	0	1	0	1	1	0	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
F		Avoidance	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	NEFAM4	Imperative	19	N/A	N/A	19	0	19	0	0	19	0
		Assertive	55	9	46	9	46	12	43	0	55	0
		Question	12	0	12	0	12	10	2	0	11	1
		Ellipsis	1	N/A	N/A	0	1	1	0	0	1	0
		Nonverbal signal	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
H		Avoidance	13	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	FAM4	Imperative	43	N/A	N/A	43	0	40	3	1	0	42
		Assertive	1	0	1	0	1	0	1	0	0	1
		Question	12	0	12	0	12	12	0	0	0	12
		Ellipsis	37	N/A	N/A	0	37	33	4	1	1	35
		Nonverbal signal	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
G		Avoidance	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Exhortative		Request		Apology		Gratitude		Honorific title		Grammatical mood		Modal verb in 2nd person	
Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Indicative	Conditional	Yes	No
1	1	2	0	1	1	0	2	0	2	N/A	N/A	N/A	N/A
0	88	3	85	88	0	0	88	4	84	83	5	11	77
0	10	2	8	8	2	0	10	1	9	8	2	9	1
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
17	5	7	15	0	22	0	22	0	22	N/A	N/A	N/A	N/A
1	14	1	14	0	15	0	15	0	15	8	7	2	13
2	58	7	53	3	57	0	60	0	60	55	5	28	32
0	3	0	3	0	3	0	3	0	3	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
0	87	44	43	10	77	40	47	66	21	54	33	1	86
0	13	6	7	2	11	6	7	10	3	12	1	11	2
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
14	26	13	27	0	40	2	38	0	40	N/A	N/A	N/A	N/A
1	7	2	6	0	8	1	8	0	8	6	2	1	7
0	46	2	44	0	46	0	46	0	46	45	1	13	33
0	3	0	3	1	2	1	2	0	3	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
0	2	0	2	1	1	1	1	0	1	N/A	N/A	N/A	N/A
0	1	0	1	0	1	1	0	0	1	1	0	0	1
0	96	42	54	63	33	4	92	3	93	71	25	84	12
0	0	0	0	0	0	0	0	0	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
59	27	16	70	0	86	0	86	0	86	N/A	N/A	N/A	N/A
1	3	0	4	0	4	0	4	0	4	2	2	0	4
0	9	2	7	0	9	0	9	0	9	8	1	7	2
1	0	1	0	0	1	0	1	0	1	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	16	11	8	4	15	0	19	14	5	N/A	N/A	N/A	N/A
0	55	10	45	27	28	0	55	35	20	52	3	4	51
0	12	3	9	7	5	0	12	5	7	11	1	9	3
0	1	1	0	0	1	0	1	1	0	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
27	16	11	32	1	42	0	43	0	43	N/A	N/A	N/A	N/A
0	1	0	1	0	1	0	1	0	1	1	0	0	1
1	11	6	6	1	11	1	11	0	12	12	0	12	0
22	15	2	35	0	37	0	37	0	37	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model

Kaja DOBROVOLJC
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

Luka TERČON
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Nikola LJUBEŠIĆ
Institut Jožef Stefan; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Universal Dependencies (UD) je mednarodno usklajena označevalna shema za medjezikovno primerljivo oblikoslovno in skladenjsko označevanje besedil po načelih odvisnostne slovnice, ki je bila ob več kot 130 drugih svetovnih jezikih uspešno uporabljena tudi za označevanje besedil v slovenščini. V prispevku predstavimo rezultate nedavnih aktivnosti v povezavi s shemo UD znotraj projekta Razvoj slovenščine v digitalnem okolju, v okviru katerega smo obstoječo infrastrukturo nadgradili s prenovo in podrobno dokumentacijo označevalnih smernic UD za slovenščino, razširitvijo drevesnice SSJ-UD za pisno slovenščino z novimi povedmi iz korpusov ssj500k in ELEXIS-WSD, izdelavo testne množice iz besedil korpusa SentiCoref za spletni portal SloBENCH ter polavtomatsko pretvorbo oblikoslovnih oznak referenčnih učnih korpusov SUK in Janes-Tag. Na razširjeni drevesnici SSJ-UD je bil naučen tudi novi napovedni model za skladenjsko razčlenjevanje v orodju CLASSLA-Stanza, ki ga v prispevku v

Dobrovoljc, K., Terčon, L., Ljubešić, N.: Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. Slovenščina 2.0, 11(1): 218–246.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.218-246

podporo nadaljnjim jezikoslovnim aplikacijam podrobneje ovrednotimo z vidika splošne natančnosti razčlenjevanja in najpogostejših tipov napak.

Ključne besede: slovnično označeni korpusi, odvisnostna slovnica, drevesnica, skladenjsko razčlenjevanje, obdelava naravnega jezika

1 Uvod

Jezikoslovno označeni korpusi, tj. digitalizirane zbirke besedil, ki poleg besed na površini vsebujejo tudi ročno pripisane podatke o njihovih slovničnih lastnostih na različnih ravneh jezikoslovnega opisa (Ide in Pustejovsky, 2017), predstavljajo enega izmed temeljnih jezikovnih virov za razvoj jezikovnotehnoloških orodij na eni strani in korpusnojezikoslovne raziskave na drugi. Slovnične lastnosti so besedilom tipično pripisane na podlagi vnaprej opredeljenih označevalnih shem oz. označevalnih sistemov, ki poleg nabora možnih oznak običajno vsebujejo tudi smernice za njihovo pripisovanje konkretnim slovničnim pojavom. Ker so v preteklosti označevalne sheme nastajale ločeno za posamezne jezike, slovnične teorije ali celo korpuse, je njihova posledična raznolikost onemogočala kakršnokoli neposredno primerjavo označenih podatkov ali na njih temelječih računalniških orodij.

Kot protiutež tovrstni razdrobljenosti je bila leta 2013 vzpostavljena označevalna shema Universal Dependencies,[1] ki si prizadeva za mednarodno oz. medjezično usklajeno slovnično označevanje besedil na oblikoslovni in skladenjski ravni, da bi pospešila razvoj večjezičnih jezikovnih tehnologij, medjezičnega strojnega učenja in kontrastivnih jezikoslovnih analiz. Znotraj sheme UD je bil tako vzpostavljen univerzalni nabor kategorij in smernic (17 besednih vrst, 24 oblikoskladenjskih lastnosti, 37 odvisnostnih skladenjskih relacij), ki odslej omogoča enotno označevanje podobnih slovničnih pojavov v različnih svetovnih jezikih, obenem pa dovoljuje tudi jezikovnospecifične izpeljave, če je to potrebno. Shema temelji na načelih odvisnostne slovnice, ki je v primerjavi s frazno skladnjo bolj primerna za jezike z bolj fleksibilnim besednim redom in za neposredno uporabo v različnih jezikovnotehnoloških aplikacijah (Jurafsky in Martin, 2021), njena teoretična izhodišča pa so podrobneje predstavljena v prispevku De Marneffe idr. (2021).

1 https://universaldependencies.org

Doslej je bilo z označevalno shemo UD ročno označenih že več kot 240 korpusov (t.i. odvisnostnih drevesnic, angl. *dependency treebanks*) v 130 svetovnih jezikih. Med njimi sta tudi univerzalni odvisnostni drevesnici pisne slovenščine SSJ (Dobrovoljc idr., 2017) in govorjene slovenščine SST (Dobrovoljc in Nivre, 2016), ki sta bili s tem neposredno vključeni v razvoj številnih najsodobnejših orodij za večjezično obdelavo naravnih jezikov (Zeman idr., 2018), kakor tudi raznolike primerjalnojezikoslovne raziskave (Futrell idr., 2015; Naranjo in Becker, 2018; Chen in Gerdes, 2018).

Glede na pomen razvoja slovenskih virov v tovrstnih mednarodnih standardizacijskih pobudah smo v okviru nacionalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO),[2] ki si prizadeva za zadovoljitev potreb po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik, obstoječe vire in povezano infrastrukturo za označevanje slovenskih besedil po sistemu Universal Dependencies bistveno nadgradili. Potek in rezultate te aktivnosti smo delno že predstavili v prispevku na konferenci Jezikovne tehnologije in digitalna humanistika 2022 (Dobrovoljc idr., 2022), v tem članku pa ga nadgradimo s posodobljeno in bolj poglobljeno analizo prvotnih rezultatov ter predstavitvijo novih povezanih podatkovnih množic.

V nadaljevanju članka tako po predstavitvi popisa najnovejših označevalnih smernic UD za slovenščino (2. razdelek) opišemo nastanek in vsebino štirih novih ročno označenih množic po sistemu Universal Dependencies (3. razdelek) – razširjene referenčne univerzalno skladenjsko razčlenjene drevesnice SSJ-UD, nove univerzalno skladenjsko razčlenjene drevesnice SloBENCH-UD ter novih univerzalno oblikoslovno označenih učnih korpusov SUK in Janes-Tag. Na novi različici korpusa SSJ-UD je bil naučen tudi novi napovedni model razčlenjevalnika CLASSLA-Stanza, ki ga v podporo nadaljnjim jezikoslovnim in jezikovnotehnološkim aplikacijam v 4. razdelku ovrednotimo z evalvacijo splošne natančnosti in kvalitativno analizo najpogostejših napak. Prispevek v 5. razdelku sklenemo s pregledom aktualne infrastrukturne podpore za jezikoslovne analize slovenskih UD korpusov in smernicami nadaljnjih raziskav.

2 https://slovenscina.eu/

2 Popis smernic UD za slovenščino

Splošne smernice UD, kakršne so dokumentirane na krovni spletni strani projekta,[3] so kot nadaljevanje predhodnih standardizacijskih iniciativ (Zeman, 2008; Petrov idr., 2012; de Marneffe idr., 2014) in večletnega kolaborativnega razvoja zasnovane tako, da skušajo na čim krajši način nasloviti oblikoslovne in skladenjske specifike čim širšega nabora jezikov. Tako v splošnih smernicah najdemo predvsem prototipične opredelitve posameznih oznak, opis najbolj tipičnih mejnih primerov in ponazoritve na primerih izbranih jezikov, naloga avtorjev drevesnic za posamezne jezike pa je, da te splošne smernice nato prenesejo na svoje konkretne jezikovne podatke. Pri tem infrastruktura UD omogoča, da se za vsak jezik ta načela popišejo kot jezikovnospecifične smernice na uradni spletni strani, vendar to ni obvezno, zato je dokumentacija označevalnih smernic UD za posamezne jezike prepuščena predvsem samoiniciativnosti avtorjev podatkov.

Za slovenščino so bile ob prvi objavi korpusa SSJ-UD (Dobrovoljc idr., 2017, gl. razdelek 3.1) tako dokumentirane zgolj smernice za pripisovanje besednih vrst in oblikoskladenjskih oznak, ki so odtlej ob prehodu s prve na drugo različico splošnih UD smernic (Nivre idr., 2020) že nekoliko zastarele. Po drugi strani smernice za pripisovanje univerzalnih skladenjskih relacij besedilom v slovenščini zaradi obsežnosti niso bile podrobneje dokumentirane oz. so bile razvidne zgolj implicitno iz pretvorbenih pravil na eni strani in označenosti objavljenega korpusa na drugi.

Prvi korak znotraj projekta RSDO je bil tako namenjen izčrpnemu popisu smernic UD za slovenščino na vseh treh ravneh označevanja (besedne vrste, oblikoskladenjske lastnosti in skladenjske relacije) v obliki priročnika (Dobrovoljc in Terčon, 2023), ki na slovenskih primerih razlaga in ponazarja uporabo posameznih oznak UD za označevanje besedil v slovenščini. Pri tem smo poleg opisa prvotnih smernic uvedli tudi nekaj manjših sprememb na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali neustrezna glede na splošne, jezikovno univerzalne smernice. Med njimi lahko izpostavimo predvsem spremembe v obravnavi primerjalnih struktur (jedro strukture je pridevnik

3 https://universaldependencies.org/guidelines.html

ali prislov, ki izraža primerjano lastnost), poudarjalnih členkov (razlikovanje med modifikatorji samostalnikov na eni in povedkov na drugi strani), besedilnih povezovalcev (razlikovanje glede na stavčno pozicijo) in zaimkov *se/si* (razlikovanje med zaimkom v vlogi predmeta na eni strani in prostim morfemom na drugi strani), ki so bili zaradi omejitev strojne pretvorbe iz sistema JOS (gl. razdelek 3.1) prvotno označeni drugače kot predvidevajo splošne smernice UD.

Priročnik s smernicami UD za slovenščino poleg opisov posamičnih slovničnih kategorij in načel njihovega pripisovanja slovenskim besedilom vsebuje še razdelek s podrobnejšo obravnavo težavnejših primerov (Dobrovoljc idr., 2023), ki se je dopolnjeval tudi skozi označevalne kampanje, opisane v 3. razdelku. V nekoliko strnjeni oz. tuji javnosti prilagojeni obliki so bile v angleščino prevedene slovenske smernice nato objavljene še na uradni spletni strani projekta UD,[4] kar omogoča neposredno primerjavo s smernicami za več deset drugih svetovnih jezikov, zbranimi na istem spletnem mestu.

V procesu dokumentacije slovenskih smernic so bila identificirana tudi nekatera odprta vprašanja, pri katerih bi dosledna implementacija splošnih smernic UD zahtevala znaten odmik od doslej uveljavljenih označevalnih praks v slovenskem prostoru, zlasti sistema JOS, in bi jih bilo zato smiselno nasloviti s širšo strokovno diskusijo. Nekaj več kot trideset tovrstnih vprašanj, ki segajo na vse ravni slovničnega opisa, od tokenizacije (npr. smiselnost razvezovanja naveznih zaimkov tipa *nanj* → *na njega*) do besednovrstne kategorizacije (npr. smiselnost premika členkov med prislove) in skladenjske analize (npr. smiselnost oz. način ločevanje trpniških struktur od povedkovodoločilnih), smo popisali v ločeni prilogi h krovnim smernicam, pri čemer so bila v sodelovanju z Univerzo v Novi Gorici za približno tretjino izbranih vprašanj že oblikovana nekatera izhodiščna priporočila za nadaljnje izboljšave.

4 Primer slovenskih smernic za pripisovanje oznake *nsubj* (samostalniški osebek) na krovnem portalu UD: https://universaldependencies.org/sl/dep/nsubj.html.

3 Novi ročno označeni korpusi UD za slovenščino

3.1 Nadgradnja univerzalno skladenjsko razčlenjenega korpusa SSJ-UD

Korpus SSJ-UD (Dobrovoljc idr., 2016; Dobrovoljc idr., 2017) je kot prvi univerzalno skladenjski korpus za slovenščino nastal na podlagi polavtomatske pretvorbe učnega korpusa ssj500k (Krek idr., 2020), prvotno označenega po shemi JOS (Erjavec idr., 2010). Za razliko od oblikoslovne ravni, ki jo je bilo mogoče s pravili za preslikavo iz enega v drug sistem pretvoriti v celoti,[5] je bilo zaradi robustnosti sistema za skladenjsko razčlenjevanje JOS v primerjavi s sistemom UD v celoti pretvorjenih zgolj 8.000 od izvorno 11.411 skladenjsko razčlenjenih povedi korpusa ssj500k, ki so bile nato kot korpus SSJ-UD prvič objavljene v zbirki UD v1.2.

Neobjavljene, polpretvorjene skladenjsko razčlenjene povedi korpusa ssj500k (razdelek 3.1.1) so tako predstavljale logično izhodišče za napovedano povečanje učnih podatkov znotraj projekta RSDO, ki sta mu sledili še razširitev s povedmi korpusa ELEXIS-WSD (3.1.2) in izboljšanje označenosti prvotnega korpusa (3.1.3). V vseh treh fazah je označevanje potekalo v označevalnem orodju Q-CAT (Brank, 2022), ki odslej podpira tudi uvoz korpusov v formatu CONLL-U, za primerjavo označenih datotek (kuriranje) pa smo uporabili lokalno inštalacijo orodja WebAnno (Eckart de Castilho idr., 2016), ki jo vzdržuje center CLARIN.SI (Erjavec idr., 2022).[6] Označevalni proces smo podrobneje že predstavili (Dobrovoljc in Ljubešić, 2022; Dobrovoljc idr., 2022), v nadaljevanju pa povzamemo zgolj najpomembnejše rezultate.

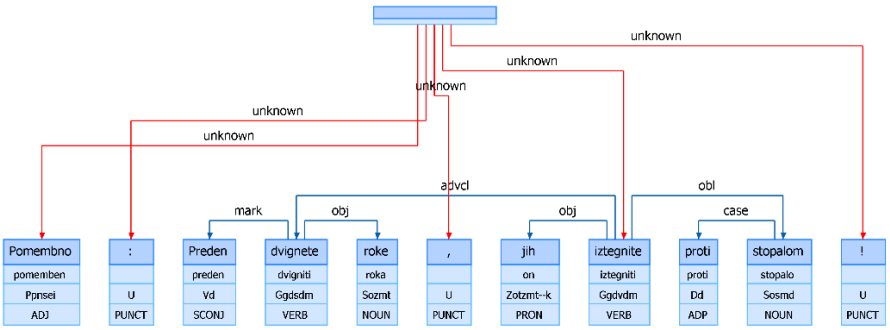
3.1.1 Razširitev s polpretvorjenimi povedmi iz korpusa ssj500k

V prvi fazi razširitve so označevalci ročno pregledali 3.411 polpretvorjenih povedi oz. 96.194 pojavnic korpusa ssj500k, med katerimi jih 22.377 (23,5 %) še ni imelo pripisane skladenjske relacije UD. Te so bile za potrebe lažje vizualizacije označene z relacijo *unknown* (Slika 1), označevalci (po dva na poved) pa so poleg ustvarjanja novih povezav preverjali tudi ustreznost že obstoječih (pretvorjenih) povezav.

5 Pravila in skripte za pretvorbo iz sistema JOS v UD so dokumentirana na povezavi https://github.com/clarinsi/jos2ud.

6 https://www.clarin.si/webanno/

Med pojavnicami, ki v izhodišču niso imele pripisane relacije UD, je bila skoraj polovica ločil (relacija *punct*), kar je bilo glede na pretvorbena pravila pričakovano, saj so bila ločila večinoma na relevantno jedro povezana šele po določitvi vseh drugih pojavnic v povedi, zlasti korena povedi (*root*), ki predstavlja tudi drugo najpogostejšo vrsto nepretvorjenih pojavnic (12 %). Tej sledita še relaciji *parataxis* (9 %) in *conj* (6 %), ki se uporabljata za povezovanje stavčnih soredij oz. priredij, torej struktur, kakršnih zgolj s pravili ni bilo mogoče pretvoriti z dovolj zanesljivo natančnostjo.



Slika 1: Primer prikaza polpretvorjene povedi iz ssj500k z manjkajočimi relacijami UD (unknown) v označevalnem orodju Q-CAT.

3.1.2 Razširitev s povedmi iz korpusa ELEXIS-WSD-SL

V drugi fazi širitve je bil skladenjsko razčlenjen še korpus ELEXIS-WSD--SL, tj. slovenski del paralelnega korpusa ELEXIS-WSD (Martelli idr., 2021; Martelli idr., 2022), razvitega za potrebe strojnega pomenskega razdvoumljanja, ki vsebuje v več evropskih jezikov prevedena besedila iz Wikipedie (Schwenk idr. 2021). Slovenski korpus ELEXIS-WSD vsebuje 2.024 povedi (31.237 pojavnic), ki so bile predhodno že ročno tokenizirane, lematizirane in oblikoskladenjsko označene po sistemu JOS, na podlagi česar smo korpus s pretvorbeno skripto samodejno pretvorili še v besedne vrste in oblikoskladenjske oznake UD, pojavitve glagola *biti*[7] pa razdvoumili ročno.

7 V primerjavi z označevalno shemo JOS, ki glagol *biti* ne glede na skladenjsko vlogo vedno označuje kot (pomožni) glagol, shema UD že na ravni določanja besednih vrst ločuje med pomožniki (AUX) in glavnimi glagoli (VERB). Podrobnejše smernice s primeri za

Tako označen korpus je bil izhodiščno skladenjsko razčlenjen z orodjem CLASSLA-Stanza (Ljubešić in Dobrovoljc 2019; Terčon in Ljubešić, 2023), pravilnost strojno pripisanih razčlemb pa so nato pregledali trije označevalci in končni kurator. Na ta način je bilo ročno popravljenih 1.534 (4,91 %) skladenjskih relacij, med katerimi so prevladovale strukture z oznakami *nmod* (samostalniški prilastki), *advmod* (prislovna določila), *obl* (odvisne samostalniške zveze), *conj* (priredja) in *punct* (ločila), kar se, kot bomo videli v nadaljevanju, sklada z najpogostejšimi tipi napak razčlenjevalnika nasploh (razdelek 4.2).

3.1.3 Izboljšanje označenosti v prvotnem korpusu SSJ-UD

Poleg dodajanja novih razčlenjenih povedi smo glede na rahlo spremembo smernic (2. razdelek), analizo ročnih popravkov pretvorjenih relacij (razdelek 3.1.1) in drugih identificiranih nedoslednosti izboljšali tudi označenost izhodiščne različice korpusa SSJ-UD.

Med približno 30 identificiranimi tipi napak oz. nedoslednosti so bile denimo pristavčne strukture, visok delež (neupravičenih) neprojektivnih povezav,[8] nedosledno ločevanje med sorednimi in priredno vezanimi stavki, med premimi in nepremimi predmeti, itd. Za vsako izmed kategorij smo s hevurističnimi poizvedbami ustvarili podkorpuse povedi s potencialno problematičnimi oznakami, ki so jih nato označevalci ročno pregledali in popravili v skladu s smernicami. Na ta način je bilo v izhodiščnem korpusu popravljenih 1.670 skladenjskih oznak, kar sicer predstavlja razmeroma majhen del celotnega korpusa (1,2 %).

3.1.4 Objava nove različice korpusa SSJ-UD

V zadnjem koraku smo izhodiščni korpus SSJ-UD z nekoliko izboljšano označenostjo (razdelek 3.1.3) združili z novimi povedmi iz korpusov ssj500k (3.1.1) in ELEXIS-WSD (3.1.2) ter tako dobili novo različico referenčne univerzalne odvisnostne drevesnice za pisno slovenščino

tovrstno razdvoumljanje glagola *biti* so na voljo tudi na GitHub repozitoriju CLARIN.SI: https://github.com/clarinsi/jos2ud/blob/master/Map/UD_biti_anno_navodila_v02.docx

8 Povezava med besedo A in besedo B je projektivna, če je beseda A posredno nadrejena tudi vsem drugim besedam med A in B – obstaja torej pot od A do vseh besed med A in B. Če si to predstavljamo grafično, se povezave v neprojektivnem drevesu med seboj križajo. To je v jezikih s prostim besednim redom, kot je slovenščina, sicer možen pojav (za primere gl. priročnik s smernicami), a vendarle redek.

SSJ-UD,[9] ki je bila s standardno delitvijo na učno, validacijsko in testno množico (več v Dobrovoljc in Ljubešić, 2022) prvič objavljena kot del uradnega izida UD v2.10 (Zeman idr., 2022).

Kot prikazuje tabela 1, nova različica v primerjavi s prvotno vsebuje 5.435 novih razčlenjenih povedi (+67,9 %) oz. skoraj enkrat večje število pojavnic (126.427, +89,9 %), s čimer se korpus SSJ-UD po številu pojavnic danes umešča v zgornjo osmino vseh UD drevesnic po svetu. Z razširitvijo je korpus SSJ-UD postal tudi bolj raznolik, saj se vsi trije podkorpusi (izvorne povedi iz ssj500k, nove povedi iz ssj500k, povedi iz ELEXIS-WSD) med seboj razlikujejo tako z vidika vrste vsebovanih besedil kot njihove skladenjske kompleksnosti, nenazadnje pa tudi z vidika izvora ročno pripisanih oznak (od pretvorb do popravljanja).

Tabela 1: *Zgradba nove različice korpusa SSJ-UD (od UD v 2.10 naprej)*

Podkorpus	Povedi	Pojavnice	Povp.
Prvotni SSJ-UD	8.000	140.670	17,58
Novo iz ssj500k	3.411	95.194	27,91
Novo iz ELEXIS-WSD	2.024	31.233	15,43
Skupaj novi SSJ-UD	**13.435**	**267.097**	**19,88**

Novi korpus SSJ-UD je bil obenem integriran tudi v novi referenčni učni korpus SUK (Arhar Holdt idr., 2022), v katerem univerzalno skladenjsko razčlenjene povedi po sistemu UD predstavljajo dobro četrtino celotnega korpusa. Ker korpus SUK vsebuje še številne druge ravni jezikoslovnih oznak, kot so skladenjske razčlembe po sistemu JOS, udeleženske vloge, večbesedne enote in imenske entitete, to odpira številne možnosti kompleksnejših čezravninskih analiz in raziskav.

3.2 Univerzalno skladenjsko razčlenjeni korpus SloBENCH-UD

Poleg nove, izboljšane in razširjene, drevesnice SSJ-UD smo izdelali še ločeno univerzalno odvisnostno drevesnico za spletni portal SloBENCH

9 Čeprav infrastruktura UD dopušča objavo poljubnega števila drevesnic, smo se namesto objave novih drevesnic UD za slovenščino namenoma odločili za priključitev novih povedi k že obstoječi drevesnici SSJ-UD, da bi zagotovili kar najbolj učinkovito izrabo teh podatkov v širši jezikovnotehnološki skupnosti, kjer se zaradi poenostavitve dela modeli pogosto razvijajo zgolj na izbrani, običajno največji, drevesnici nekega jezika.

(Žitnik in Dragar, 2021),[10] ki je bil razvit z namenom enotne primerjave uspešnosti orodij za različne jezikovnotehnološke naloge na slovenskih besedilih. Del vsake naloge je tudi vnaprej definirana testna množica besedil, na kateri razvijalci poženejo svoje orodje, evalvacijski sistem v ozadju pa nato strojne rezultate primerja z (javnosti skritimi) ročno pregledanimi rešitvami. V primerjavi s testno množico SSJ-UD, ki je skupaj z učno in validacijsko množico z oznakami vred javno objavljena kot del uradnega repozitorija UD, drevesnica SloBENCH-UD kot skrita testna množica zagotavlja bolj nepristransko evalvacijo, saj onemogoča namensko prilagajanje delovanja orodij specifičnim testnim podatkom.

Konkretno smo za izdelavo drevesnice SloBENCH-UD uporabili del korpusa SentiCoref (Žitnik, 2019), ki je bil za druge naloge na istem portalu že ročno pregledan na ravni segmentacije, tokenizacije, lematizacije in oblikoskladenjskih oznak JOS. Po ustaljenem postopku smo oznake JOS s pomočjo pretvorbenih pravil najprej preslikali v besedne vrste in oblikoslovne oznake UD ter nato množico strojno razčlenili z najnovejšim razčlenjevalnim modelom, naučenim na podatkih razširjenega korpusa SSJ-UD (razdelek 3.1). Strojno razčlenjene podatke je ročno pregledal in popravil ekspertni označevalec ter pri tem besednovrstno razdvoumil tudi glagol *biti* v vlogi pomožnega oz. glavnega glagola. Končna ročno pregledana množica SloBENCH-UD tako obsega 1.332 razčlenjenih povedi oz. 29.138 pojavnic in je tako po velikosti kot vsebini primerljiva s testno množico SSJ-UD. To navsezadnje potrjuje tudi zelo podobna stopnja natančnosti, ki jo na obeh testnih množicah dosega razčlenjevalni model CLASSLA-Stanza.

Množica je že bila integrirana v spletni portal SloBENCH, pri čemer je krovna naloga, tj. slovnično razčlenjevanje slovenskih besedil po shemi Universal Dependencies, zasnovana po vzoru tekmovanj CoNLL Shared Task 2017 in 2018 (Zeman idr., 2018), z evalvacijskimi metrikami vred.

3.3 Univerzalno oblikoslovno označena korpusa SUK in Janes-Tag

Kot smo omenili že v razdelku 3.1, so bila za pretvorbo korpusa ssj500k v prvo različico drevesnice SSJ-UD izdelana podrobna pravila za preslikavo

10 https://slobench.cjvt.si/

oblikoskladenjskih oznak JOS v besedne vrste in oblikoskladenjske lastnosti sistema UD. Zaradi precejšnje podobnosti med obema sistemoma je bilo s temi pravili mogoče natančno pretvoriti celotni korpus ssj500k, pozneje pa tudi druge sorodne vire z oznakami JOS, kot sta oblikoslovni leksikon Sloleks (Dobrovoljc idr., 2019) in ročno označeni učni korpus nestandardne slovenščine Janes-Tag (Erjavec idr., 2019). Čeprav tako označeni viri zaradi manka skladenjskih relacij ne morejo biti distribuirani kot del uradne zbirke drevesnic UD, kot to velja za slovenski drevesnici SSJ in SST, ti predstavljajo pomemben vir podatkov za učenje napovednih modelov na nižjih slovničnih ravneh (Dobrovoljc idr., 2019).

S tem namenom smo v univerzalne oblikoslovne oznake (besedne vrste in druge oblikoskladenjske lastnosti) pretvorili tudi novi referenčni učni korpus za standardno pisno slovenščino SUK 1.0 (Arhar Holdt idr., 2022) in razširjeni referenčni učni korpus za nestandardno pisno slovenščino Janes-Tag 3.0 (Lenardič idr., 2022), ki oba v primerjavi s prejšnjima različicama (ssj500k 2.3 in Janes-Tag 2.1) vsebujeta kar enkrat več ročno pregledanih lem in oblikoskladenjskih oznak JOS (približno 1 milijon oz. 190.000 pojavnic). Ker se pretvorbena pravila v času od nastanka prejšnjih različic korpusov niso spremenila, smo v okviru projekta RSDO pretvorbo opravili zgolj na novo dodanih besedilih in opravili ustaljeni ročni pregled povedi z glagolom *biti* za razdvoumljanje med pojavitvami pomožnega in glavnega glagola (po en označevalec na primer).

4 Novi razčlenjevalni model

V drugi fazi projekta smo na novi, bistveno večji različici ročno označenega korpusa SSJ-UD (razdelek 3.1) naučili tudi nov napovedni model skladenjskega razčlenjevanja po sistemu UD v označevalnem orodju CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić 2023),[11] ki se je kot temeljno programsko orodje za označevanje besedil v slovenščini prav tako razvijalo v okviru projekta RSDO. Gre za izpeljavo odprtokodnega orodja Stanza (Qi idr., 2020), ki v primerjavi z izvornim orodjem uvaja nekatere izboljšave na ravni tokenizacije, oblikoskladenjskega označevanja in lematizacije, skladenjski razčlenjevalnik pa se od izvornega (Dozat

11 https://github.com/clarinsi/classla

in Manning, 2016), ki temelji na nadgrajeni metodi dvosmernega dolgega kratkoročnega spomina (BiLSTM), razlikuje predvsem po uporabi besednih vložitev CLARIN.SI-embed.sl (Ljubešić in Erjavec, 2018), ki so bile naučene na slovenskih besedilih v obsegu 3,5 milijard besed.

Najnovejši razčlenjevalni model (Terčon in Ljubešić, 2023) je del najnovejše različice CLASSLA-Stanza 2.0. Modeli, ki so vključeni v to različico, so bili naučeni na učnem korpusu SUK (Arhar Holdt idr., 2022). Za učenje modela za skladenjsko razčlenjevanje je bil uporabljen le tisti del korpusa, ki ustreza korpusu SSJ-UD[12].

Primerjavo novega modela s predhodnim modelom, naučenim na prvotni različici SSJ-UD, sta podrobneje opisala že Dobrovoljc in Ljubešić (2022), ki ugotavljata, da je model, naučen na novi različici korpusa SSJ-UD,[13] zaradi povečanega obsega učnih podatkov in njihove diverzifikacije bistveno izboljšan v primerjavi z modelom, naučenim na prvotni različici.

Da bi osvetlili prednosti in pomanjkljivosti uporabe novega razčlenjevalnega modela v različnih jezikovnotehnoloških in jezikoslovnih aplikacijah ter obenem identificirali prioritete za njegove nadaljnje izboljšave, v nadaljevanju prispevka te ugotovitve nadgradimo s podrobnejšo evalvacijo splošne natančnosti modela (razdelek 4.1) na eni strani in analizo najpogostejših tipov napak (razdelek 4.2) na drugi.

Pri evalvaciji smo uporabili ročno označene podatke na nižjih ravneh označevanja (tokenizacija, stavčna segmentacija, oblikoskladenjsko označevanje, lematizacija), saj nas je v tej fazi razvoja razčlenjevalnika zanimala predvsem natančnost napovednega modela v izolaciji, brez vpliva napovednih karakteristik orodja na nižjih ravneh.

4.1 Splošna natančnost modela

Za kvantitativno evalvacijo splošne natančnosti modela smo uporabili standardni protokol, po katerem smo model, naučen na učni oz. validacijski množici uporabili za razčlenjevanje testne množice, napovedane oznake pa nato primerjali z ročno pripisanimi. Za poročanje o

12 Celoten proces učenja modelov za najnovejšo različico orodja CLASSLA-Stanza je natančneje opisan na GitHub repozitoriju: https://github.com/clarinsi/classla-training

13 V prispevku sta Dobrovoljc in Ljubešić (2022) sicer evalvirala predhodno neuradno različico modela, ki se od uradne razlikuje v količini in tipu učnih podatkov na nižjih ravneh, vendar je njuna splošna natančnost povsem primerljiva, saj sta bila naučena na identičnem skladenjsko razčlenjenm korpusu SSJ-UD.

natančnosti uporabljamo uveljavljeno metriko LAS (angl. *labeled attachment score*), ki prikazuje delež pojavnic s pravilno napovedano nadrejeno pojavnico in vrsto njunega skladenjskega razmerja, pri čemer ta delež povzemamo z oceno F1, ki prikazuje harmonično sredino med preciznostjo in priklicem.[14]

Rezultati, predstavljeni v tabeli 2, prikazujejo, da razčlenjevalni model dosega splošno natančnost 93,73 LAS F1, kar nekoliko poenostavljeno pomeni, da se model v povprečju na vsakih sto označenih pojavnic zmoti pri manj kot sedmih, tj. jim pripiše napačno nadrejeno pojavnico in/ali vrsto povezave med njima.

Kot prikazujejo rezultati za posamične tipe relacij v tabeli 2,[15] pa ta splošna ocena natančnosti ni reprezentativna za vse vrste skladenjskih struktur, saj je pri napovedovanju nekaterih relacij model bistveno natančnejši kot pri drugih.

Tabela 2: *Natančnost novega modela orodja CLASSLA-Stanza za skladenjsko razčlenjevanje po sistemu UD glede na metriko LAS*

Relacija	Izvorni opis	Slovenski prevod	Učna	Testna	LAS F1
acl	clausal modifier of noun	stavčni prilastki	3377	444	**83,48**
advcl	adverbial clause modifier	prislovni odvisniki	1927	239	**76,25**
advmod	adverbial modifier	prislovna določila (gl. op. 17)	16307	1935	**90,37**
amod	adjectival modifier	pridevniški prilastki	17.628	2165	**99,12**
appos	appositional modifier	pristavčna določila	1.505	163	**66,45**
aux	auxiliary verb	pomožni glagoli	9.773	1162	**99,01**
case	case marking preposition	predlogi	19.813	2415	**99,19**
cc	coordinating conjunction	priredni vezniki	7.294	923	**96,32**

14 Izračuni temeljijo na uradni evalvacijski skripti tekmovanja CoNLL Shared Task 2018 (Zeman idr., 2018), ki smo jo dodatno prilagodili tako, da poleg splošnega izračuna natančnosti vrača tudi rezultate za posamične skladenjske relacije, besedne vrste in druge relevantne oznake.

15 V Tabeli 2 ni relacij *goeswith* (napačno razdruženi deli besed), *reparandum* (samopopravljanja) in *compound* (zloženke), saj se v korpusu SSJ-UD ne pojavljajo. Pri relaciji *dislocated* podatka o natančnosti ni (oznaka *n/a*), saj se v testni množici ne pojavi. O natančnosti izpeljanih relacij oz. podoznak (npr. *flat:name*, *flat:foreign*) poročamo združeno z jedrno oznako (npr. *flat*).

Relacija	Izvorni opis	Slovenski prevod	Učna	Testna	LAS F1
ccomp	clausal complement	stavčna dopolnila (predmetni odvisniki)	1.544	187	**92,27**
conj	conjunct	priredno zloženi elementi	9.307	1108	**86,30**
cop	copula verb	vezni glagoli	4.244	542	**96,12**
csubj	clausal subject	osebkovi odvisniki	723	78	**85,53**
dep	unspecified dependency	nedoločena povezava	169	12	**20,00**
det	determiner	določilniki	4.724	616	**98,95**
discourse	discourse element	diskurzni členki	153	15	**75**
dislocated	dislocated element	dislocirani elementi	10	0	**n/a**
expl	expletive	ekspletivne besede	2.997	361	**96,31**
fixed	fixed multi-word expression	funkcijske zveze	944	89	**92,47**
flat	flat multi word-expression	eksocentrične zveze	152	5	**95,00**
iobj	indirect object	nepremi predmeti	873	87	**83,33**
list	list	seznami	583	16	**82,35**
mark	marker (subordinating conjunction)	podredni vezniki	6.415	804	**98,38**
nmod	nominal modifier	samostalniški prilastki	16.065	1887	**87,44**
nsubj	nominal subject	samostalniški osebki	10.585	1315	**96,05**
nummod	numeric modifier	številčna določila	3.543	311	**95,13**
obj	(direct) object	premi predmeti	9.733	1140	**96,37**
obl	oblique nominal (adjunct)	odvisne samostalniške zveze	14.049	1722	**92,10**
orphan	dependent of missing parent	elementi v eliptičnih strukturah	785	83	**71,90**
parataxis	parataxis	stavčna soredja	3.273	345	**73,26**
punct	punctuation symbol	ločila	32.116	3623	**94,09**
root	root element	koren povedi	10.903	1282	**96,80**
vocative	vocative	ogovori	63	1	**0**
xcomp	open clausal complement	odprta stavčna dopolnila	1.542	198	**92,50**
Vse relacije					**93,73**

Opomba. V 4. in 5. stolpcu je navedeno število pojavnic, označenih z dano relacijo, v učni oz. testni množici.

Med relacijami z najvišjo natančnostjo napovedovanja so po pričakovanju funkcijske besede, kot so predlogi (*case*; 99,19), pridevniški

prilastki (*amod*; 99,12), pomožni glagol *biti* (*aux*; 99,01), določilniški zaimki in prislovi (*det*; 98,95), podredni vezniki (*mark*; 98,38), ekspletivni zaimki (*expl*; 96,31) in priredni vezniki (*cc*; 96,32); skratka, pojavnice, ki se pojavljajo v zelo predvidljivih oblikah in skladenjskih položajih.

Poleg navedenih relacij model razmeroma dobro natančnost dosega tudi pri napovedovanju nekaterih jedrnih skladenjskih struktur, kot so samostalniški predmeti (*obj*; 96,37) in osebki (*nsubj*; 96,05), nadpovprečno uspešen pa je tudi pri identifikaciji korena povedi (*root*; 96,8), ki je običajno jedro povedka glavnega stavka, in veznega glagola *biti* (*cop*; 96,12), ki nastopa v strukturah s povedkovimi določili.

Med relacijami, pri napovedovanju katerih model dosega najslabše rezultate, pričakovano najdemo ogovore (*vocative*; 0,0), saj se v testni množici pojavi zgolj en primer, in nedoločene strukture (*dep*; 20,0), saj se ta oznaka kot skrajna možnost uporablja predvsem za povezovanje obrobnih, iregularnih pojavov, ki jim je nemogoče pripisati katerokoli drugo povezavo (npr. ostanki oštevilčenih strani pri digitalizaciji besedil).

Čeprav se je natančnost označevanja samostalniških pristavčnih določil (*appos*, 66,45), "osirotelih" stavčnih členov v povedih z glagolsko elipso (*orphan*; 71,90), stavčnih soredij (*parataxis*; 73,26), diskurznih členkov (*discourse*; 75,0), in naštevalnih seznamov (*list*; 82,35) z novo različico korpusa SSJ-UD bistveno izboljšala glede na prvotni model (Dobrovoljc, Ljubešić 2022), te relacije ostajajo med tistimi z najnižjo natančnostjo, kar je glede na njihovo ohlapnejšo slovnično povezanost s povedkom oz. nadrejenimi stavčnimi členi tudi pričakovano.

Med drugimi relacijami s podpovprečno natančnostjo označevanja lahko izpostavimo še podredne stavke različnih tipov, kot so prislovni (*advcl*; 76,25), prilastkovi (*acl*; 83,48), osebkovi (*csubj*; 85,53) in predmetni odvisniki (*ccomp*; 92,27). Poleg nepremih predmetov (*iobj*; 83,33), ki jih je težavno identificirati predvsem zaradi pomanjkljivosti trenutnih označevalnih smernic,[16] modelu precejšen izziv predstavlja-

16 V času nastanka smernic in označenih podatkov, ki jih opisuje ta prispevek, so splošne smernice UD zaradi kompleksnega prepletanja oblikoslovnih, skladenjskih in pomenskih razločevalnih lastnosti med premimi in nepremimi predmeti priporočale, da je v povedih z zgolj enim izraženim predmetom vedno označen kot premi predmet (*obj*), ne glede na sklon ali udeleženska vlogo. To robustno pravilo, ki je bilo identificirano tudi kot eno izmed odprtih vprašanj, omenjenih v 2. razdelku, je bilo pred kratkim opuščeno. Temu bodo sledile tudi prihodnje nadgradnje korpusa SSJ in lahko pričakujemo, da se bo posledično izboljšala tudi natančnost strojnega ločevanja med premimi in nepremimi predmeti.

jo tudi priredja, zlasti medstavčna (*conj*; 86,3), samostalniški prilastki (*nmod*; 87,44) ter prislovna določila povedkov, samostalnikov in pridevnikov (*advmod*; 90,37).

4.2 Najpogostejše napake modela

V drugem koraku evalvacije smo analizo zanesljivosti modela pri razčlenjevanju posameznih tipov relacij dopolnili še s podrobnejšo analizo najpogostejših tipov napak. Tabela 3 tako povzema distribucijo napak glede na to, pri katerem izmed obeh napovedanih podatkov (identifikator nadrejene pojavnice in vrsta skladenjske relacije med njima) se je model dejansko zmotil. Za vsak tip napake navajamo tudi najpogostejše podtipe glede na relacije, pri katerih se pojavlja, pri čemer štetje prikazujemo združeno za napake v obe smeri (npr. *obl-nmod* vključuje tako napovedovanje *obl* namesto *nmod* kot napovedovanje *nmod* namesto *obl*).

Identificirane pogoste tipe napak znotraj vsake kategorije na podlagi ročne analize napačno označenih primerov opišemo v nadaljevanju, pri čemer podrobneje predstavimo predvsem najpogostejše.

Tabela 3: *Distribucija napak razčlenjevalnega modela glede na tip napake*

Tip napake	Število napak
Napačno jedro	**845**
punct-punct	214
advmod-advmod	159
nmod-nmod	121
conj-conj	94
acl-acl	47
parataxis-parataxis	30
obl-obl	27
advcl-advcl	25
cc	23
cop-cop	20
Napačno jedro in oznaka	**493**
obl-nmod	140
parataxis-root	39
acl-advcl	19
root-nsubj	16
parataxis-appos	15

Tip napake	Število napak
Napačna oznaka	**257**
conj-parataxis	19
obl-nsubj	16
appos-conj	16
obj-iobj	14
obl-obj	11
Vse napake	**1595**

4.2.1 Napačna napoved nadrejenega elementa

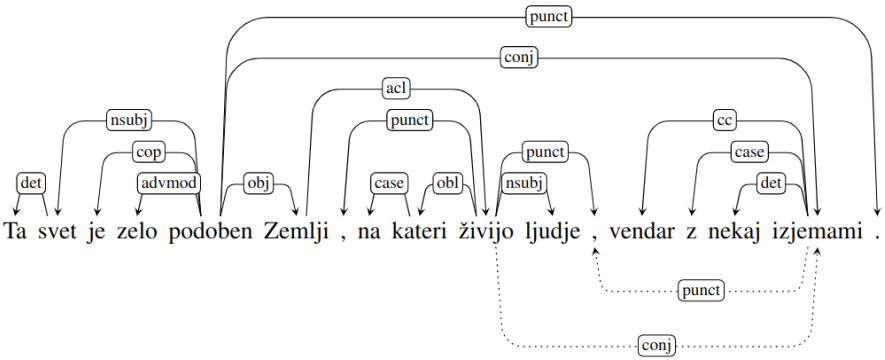
Kot prikazuje tabela 3, dobro polovico (52,8 %) predstavljajo napake, pri katerih je model pravilno napovedal skladenjsko vlogo pojavnice (pravilno relacijo oz. oznako), zmotil pa se je pri napovedi njenega nadrejenega elementa (jedra oz. izvora relacije).

Najpogostejša napaka pri določanju nadrejenega elementa je povezana z relacijo **punct**, ki označuje ločila. Po večini gre za primere, kjer so napačno določena tudi jedra drugih struktur v povedi, na katera se ločila praviloma povezujejo. Napačno povezana ločila so torej večinoma posledica napak razčlenjevanja njihovih nadrejenih struktur, kot prikazuje primer na Sliki 2, pri katerem razčlenjevalnik zadnji stavek zmotno interpretira kot priredje pred njim stoječega odvisnika, čemur ustreza tudi (napačno) označena vejica. Pri dolgih povedih, v katerih nastopa veliko podrednih elementov, obdanih z vejicami, se pojavljajo tudi napake, kjer razčlenjevalnik povzroči neprojektivnost z vezanjem vejic na napačno jedro.

Druga pogosta skupina je povezana s t.i. poudarjalnimi členki oz. prislovi, kot so besedice *tudi*, *še*, *le*, *že* idr., ki jim pripisujemo relacijo **advmod**,[17] njihova stava pa je v slovenščini razmeroma prosta – modificirajo lahko tako povedek kot posamezne stavčne člene, kar je pogosto mogoče razbrati šele iz konteksta ali prozodičnih poudarkov pri branju. Kot prikazuje primer na Sliki 3, razčlenjevalnik te besede namesto na poudarjeni samostalnik pogosto veže na povedek stavka. To ni

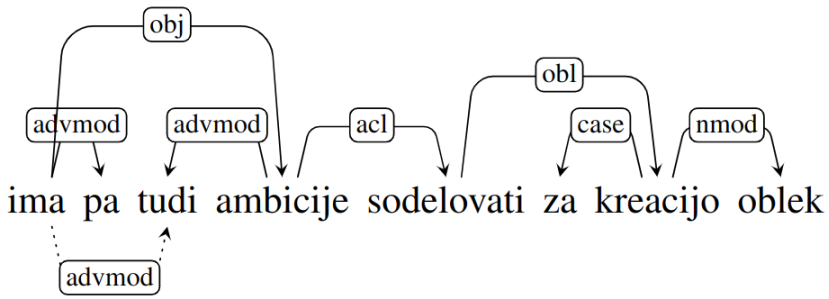
17 Relacija *advmod* se uporablja za označevanje prislovov v vlogi modifikatorjev, kar vključuje tako prislove v vlogi okoliščinskih dopolnil povedkov (kakršna Slovenska slovnica (Toporišič 2000) imenuje prislovna določila, npr. pridem *takoj*) kot prislove v vlogi modifikatorjev pridevniških, prislovnih ali samostalniških besednih zvez (prislovni prilastki, npr. *izjemno* prilagodljiv).

presenetljivo, glede na to, da gre za eno izmed kategorij, pri kateri so se označevalci najpogosteje razhajali, prav tako pa je bila nedosledno označena v prvotnem korpusu, v katerem so bile ob pretvorbi te pojavnice ne glede na vlogo vedno povezane na povedek.



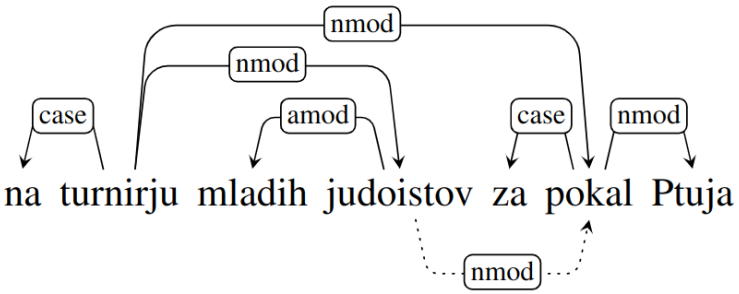
Slika 2: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) pripisanim jedrom relacije punct.

Pri relaciji **advmod** se napake pojavljajo, tudi ko nekemu prislovu sledi pridevnik. Razčlenjevalnik za nadrejeni element pogosto določi ta pridevnik, čeprav je pravi nadrejeni element povedek stavka. Zgodi pa se lahko tudi obratno: kot nadrejeni element je označen povedek, v resnici pa bi to moral biti sledeči pridevnik.



Slika 3: Primer napačne razčlembe poudarjalnih členkov (advmod zgoraj) kot prislovnih določil povedka (advmod spodaj).

Pri štirih relacijah s pogosto napačno pripisanim izvorom povezave, tj. **nmod**, **conj**, **acl** in **obl**, prihaja do podobne napake: razčlenjevalnik zanesljivo prepozna vrsto nadrejene strukture (npr. samostalniške zveze, pridevniške zveze ali povedki), vendar namesto prave strukture kot jedro izbere najbližjo ustrezno zvezo na levi, kar ni vedno prav, saj se včasih pravi izvor relacije v povedi pojavi že prej (Slika 4).



Slika 4: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) identificirano odnosnico predložne zveze v vlogi desnega prilastka (*nmod*).

Napake relacije **conj** se pogosto pojavijo, ko gre za zaporedje treh ali več priredno zloženih stavkov, kjer model ne ujame pravilnega nanašanja. Pojavijo se tako napake, kjer je pravilen izvorni stavek relacije pred ciljnim, kot tudi napake, kjer se pravilen izvoren stavek pojavi za ciljnim. Enaka napaka se pojavlja tudi pri stavčnih soredjih, ki so označena z relacijo **parataxis**.

Ko gre pri prislovnih določilih, označenih z **obl**, za izbiro med več kot enim možnim nadrejenim povedkom, označevalnik kot izvor relacije pogosto določi napačen povedek. Ti povedki se pojavijo tako pred ciljem relacije kot tudi za njim in sestavljajo najrazličnejše strukture.

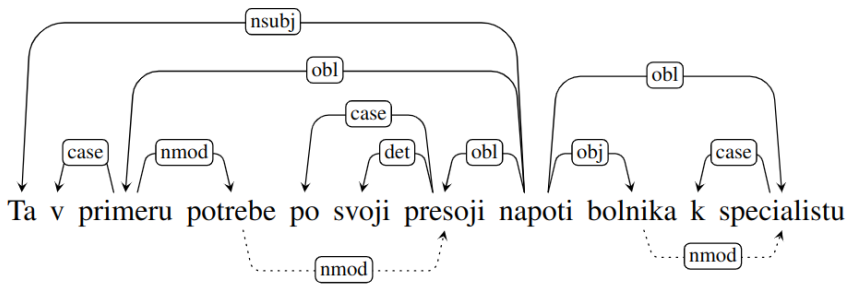
Pri relaciji **advcl** pogosto pride do napak v povezavi s stavčnimi primerjavami. Razčlenjevalnik za izvor primerjave ne določi primerjane lastnosti, pač pa povedek glavnega stavka, kar ni v skladu s trenutnimi smernicami. Do napak prihaja tudi pri strukturah z modalnimi in faznimi glagoli, kjer se odvisnik včasih pomotoma veže na nedoločnik in ne na modalni oz. fazni glagol, kot predpisujejo smernice. To nakazuje, da so med učnimi podatki pri teh strukturah morda nedoslednosti.

Včasih se zgodi, da razčlenjevalnik ne prepozna vseh enot priredja. Ker se priredni veznik vedno veže na drugi element priredja, ta napaka hkrati povzroči napako pri določanju jedra relacije **cc** (primer: *bilo je namenjeno predvsem in samo moškim* - tu sta v priredju besedi *predvsem* in *samo*, model pa je za drugi element napačno določil besedo *moškim*, ki je posledično dobila vlogo jedra relacije **cc**).

Napake relacije **cop** se pogosto pojavijo v obliki zamenjave osebka in povedkovega določila v povedih, v katerih je težko določiti, kateri člen izpolnjuje katero od teh dveh vlog. Te napake se pojavijo le ob veznem glagolu *biti*.

4.2.2 Napačna napoved nadrejenega elementa in relacije

Po pogostosti sledijo napake, pri katerih se je model zmotil tako pri napovedi nadrejene pojavnice kot njune skladenjske relacije (29,9 %). Med njimi najbolj izstopa zamenjevanje struktur z oznakama **obl**[18] in **nmod**, ki predstavlja tretji najpogostejši (pod)tip napak nasploh. Analiza primerov kaže, da gre večinoma za povedi, v katerih predložna zveza v vlogi prislovnega določila povedka (**obl**) stoji tik za neko samostalniško zvezo, model pa prislovno določilo napačno tolmači kot njen desni prilastek, za katere se uporablja relacija **nmod**, kot prikazuje primer na Sliki 5.



Slika 5: Primer napačne razčlembe predložnih prislovnih določil (*obl* zgoraj) kot desnih prilastkov (*nmod* spodaj).

18 Relacija *obl* se uporablja za odvisne samostalniške in predložne zveze, ki nastopajo v vlogi nejedrnih argumentov povedka. Poleg teh se s to relacijo označujejo tudi neglagolske strukture s primerjalnimi vezniki.

Manj pogoste v tej kategoriji so napake drugih kombinacij relacij. Pri **parataxis-root** gre za napake pri določanju glavnega stavka v nizu dveh ali več soredno zloženih stavkov, zlasti kadar gre za vrinjene stavke ali premi govor. Pri kombinaciji **acl-advcl** gre za napake ločevanja med prislovnodoločilnimi odvisniki in stavčnimi prilastki, pogosto v kombinaciji z veznikom *kot,* model pa včasih tudi ne prepozna določenih stavkov kot stavčnih primerjav in jih namesto tega veže na najbližjo samostalniško zvezo z relacijo **acl**. Kombinacija **root-nsubj** pogosto pomeni zamenjavo osebka in povedkovega določila v strukturah z veznim glagolom *biti,* včasih pa se te napake pojavijo tudi pri eliptičnih stavkih z nejasno strukturo. To so pogosto krajši stavki z več izpuščenimi členi.

4.2.3 Napačna napoved relacije

Med vsemi tremi kategorijami napak pa je najmanj takih, pri katerih je razčlenjevalnik pojavnico povezal s pravim nadrejenim elementom, a tej relaciji pripisal napačno oznako (17,3 %). V primerjavi s prvima dvema kategorijama so tukaj tipi glede na relacije razpršeni bolj enakomerno.

Do zamenjav oznak **conj** in **parataxis**[19] prihaja predvsem pri daljših povedih, pri katerih se med dva prirediono zložena stavka oz. med priredni veznik in drugi stavek v priredju vrivajo druge strukture (npr. odvisniki). Samostalniška prislovna določila (ki prejmejo relacijo **obl**)[20] so napačno označena kot osebki (**nsubj**) predvsem v zvezah z glagoli, kot so *imenovati*, *praviti*, idr., v katerih se pojavljajo v imenovalniku (npr. *pravimo jim mikroznaki*). Do zamenjave **obl-nsubj** velikokrat pride tudi pri strukturah z glagolom *biti* in samostalnikom v rodilniku (npr. v stavku *on je mnenja, da…* in *So bolj vesele narave…*). V takih primerih model pogosto določi samostalnik v rodilniku za osebek, kar lahko delno pojasnimo z dejstvom, da primerov tovrstnih struktur v učnih podatkih ni prav veliko.

Med drugimi tipi napačno pripisanih relacij je pogosta še dvoumnost med samostalniškimi zvezami v vlogi pristavčnih določil (**appos**)

19 Relacija *parataxis* se uporablja za označevanje stavčnih soredij različnih vrst. To so razmerja med besedo (običajno jedrom glavnega stavka) in drugimi elementi, ki z njo niso v priredju, podredju ali kateremkoli drugem jedrnem slovničnem razmerju.

20 Eno izmed odprtih vprašanj za prihodnje nadgradnje smernic je tudi jasnejša opredelitev kategorije *obl* in njene razmejitve glede na druge vrste samostalniških dopolnil glagola. Trenutno smernice namreč sledijo načelom opredelitve samostalniških ‘prislovnih določil’ znotraj sheme JOS-SSJ.

na eni in priredno povezanih elementov (**conj**) na drugi strani, zlasti kadar gre za naštevanje in zadnji element v brezvezniškem priredju stoji na koncu povedi (primer: *to je popoln poraz nekega koncepta, popoln poraz vseh nas*).

Pojavljajo se tudi napake ločevanja med prislovnimi določili in predmeti, predvsem pri samostalniških zvezah, ki izražajo časovni oz. prostorski okvir dogodka (**obl-obj**) in pa napačno določanje premega (**obj**) in nepremega predmeta (**iobj**).

5 Sklep

V prispevku smo predstavili nadgradnjo slovenskih slovnično razčlenjenih korpusov, ročno označenih po medjezikovno primerljivi shemi Universal Dependencies, v okviru katere smo po rahli prenovi in izčrpni dokumentaciji označevalnih smernic za slovenščino referenčno drevesnico pisne slovenščine SSJ-UD razširili z več kot 5.000 novimi povedmi, izdelali povsem novo testno množico za uporabo na evalvacijskem portalu SloBENCH in referenčna ročno oblikoskladenjsko označena korpusa SUK in Janes-Tag pretvorili v oblikoslovne oznake UD. Na novi različici drevesnice SSJ-UD smo naučili tudi nov napovedni model za skladenjsko razčlenjevanje slovenskih besedil, ki v splošnem dosega razmeroma visoko stopnjo natančnosti, pri čemer naša analiza kaže, da je pri členjenju nekaterih struktur mogoče pričakovati bistveno večjo zanesljivost rezultatov kot pri drugih.

Glede na mednarodno relevantnost sheme UD ti rezultati predstavljajo pomemben doprinos k nadaljnjemu razvoju jezikovnih tehnologij za slovenščino tako v slovenskem kot mednarodnem prostoru, saj je glede na odprti dostop in standardizirano distribucijo drevesnic UD mogoče pričakovati, da bodo novi podatki za slovenščino kmalu integrirani tudi v številna druga razčlenjevalna orodja oz. na njih temelječe aplikacije (npr. Honnibal in Montani, 2017; Nguyen idr., 2021). Poleg modelov za skladenjsko razčlenjevanje, kakršnega smo predstavili v tem prispevku, je skoraj enkrat večja količina učnih podatkov za slovenščino neprecenljiva tudi za nadaljnji razvoj modelov za lematizacijo in oblikoslovno označevanje po sistemu UD, ki v mednarodnem prostoru večinoma temeljijo zgolj na uradno izdanih drevesnicah UD, kot je SSJ-UD,

ne pa virih, ki so bili razviti oz. distribuirani v lokalnem kontekstu, kot sta denimo univerzalno oblikoslovno označena učna korpusa (ne)standardne slovenščine, SUK in Janes-Tag .

Čeprav je bila shema UD prvotno vzpostavljena predvsem za potrebe jezikovnotehnoloških raziskav, pa številne odmevne primerjalnojezikoslovne študije dokazujejo tudi njeno relevantnost na področju jezikoslovja, vključno s slovenistiko, kjer metodološki potencial skladenjsko razčlenjenih korpusov doslej še ni bil polno izkoriščen (Ledinek, 2018). Verjamemo, da izčrpno dokumentirane smernice, obsežni ročno označeni korpusi in sistematična evalvacija natančnosti na njih naučenih modelov predstavljajo pomemben doprinos k nadaljnjim jezikoslovnim raziskavam ročno in strojno razčlenjenih slovenskih korpusov

Pri tem je glede na kompleksno strukturo tovrstnih korpusov za doseganje tega cilja nujno vzpostaviti tudi ustrezno infrastrukturo za njihovo analizo. Poleg vključevanja korpusov SSJ-UD, SUK in Janes-Tag v konkordančnike CLARIN.SI (Erjavec, 2013), ki niso prilagojeni specifikam iskanja po odvisnostnih drevesih in njihovi vizualizaciji, je bilo v zadnjem času razvitih več namenskih orodij za analizo univerzalno skladenjsko razčlenjenih korpusov v slovenščini, kot sta orodje za statistično analizo skladenjsko razčlenjenih korpusov STARK (Krsnik idr., 2019) in spletni portal Drevesnik za napredno brskanje po slovenskih UD drevesnicah (Štravs in Dobrovoljc, 2022). Temu tipu oznak je bilo prilagojeno tudi označevalno orodje Q-CAT (Brank, 2022), za strojno slovnično označevanje novih besedil pa je bil pred kratkim vzpostavljen tudi spletni portal CJVT Označevalnik,[21] ki temelji na orodju CLASSLA-Stanza, a omogoča tehnično manj podkovanemu uporabniku prijaznejšo izbiro nastavitev in prikaz rezultatov.

Ne glede na v prispevku predstavljeno nadgradnjo jezikovne infrastrukture za medjezikovno primerljivo slovnično analizo slovenskih besedil pa je tako z vidika jezikovnotehnološke kot jezikoslovne uporabe te rezultate smiselno kontinuirano nadgrajevati tudi v prihodnje, kar vključuje tako izboljšavo izhodiščnih smernic na eni strani kot nadgradnjo in razvoj novih virov in tehnologij na drugi. Med drugim lahko pomembne rezultate pričakujemo tudi v okviru več tekočih nacionalnih projektov, ki se ukvarjajo s slovničnim označevanjem govorjenega jezika, ter v okviru

21 https://orodja.cjvt.si/oznacevalnik

evropske mreže COST UniDive,[22] ki se ukvarja z jezikovno raznolikostjo in univerzalnostjo v kontekstu jezikovnih tehnologij.

Zahvala

Predstavljeno delo so podprli projekt *Razvoj slovenščine v digitalnem okolju*, ki sta ga financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, ter raziskovalni program *Jezikovni viri in tehnologije za slovenski jezik* (št. P6-0411) in raziskovalni projekt *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (št. Z6-4617), ki ju financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Zahvala gre tudi označevalcem novih podatkov ter Tomažu Erjavcu, Luku Krsniku, Cyprianu Laskowskemu in Mihaelu Šinkcu za tehnično podporo.

Literatura

Arhar Holdt, S., & Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in Slovstvo*, *52*, 95–110.

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., Pori, E., Terčon, L., Munda, T., Žitnik, S., Robida, N., Blagus, N., Može, S., Ledinek, N., Holz, N., Zupan, K., Kuzman, T., Kavčič, T., Škrjanec, I., … Zajc, A. (2022). *Training corpus SUK 1.0*. http://hdl.handle.net/11356/1747

Brank, J. (2022). *Q-CAT Corpus Annotation Tool 1.4*. http://hdl.handle.net/11356/1684

Chen, X., & Gerdes, K. (2018). How Do Universal Dependencies Distinguish Language Groups? In J. Jiang & H. Liu (Eds.), *Quantitative Analysis of Dependency Structures* (pp. 277–294). De Gruyter Mouton. doi: 10.1515/9783110573565-014

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L., & Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0*. Pridobljeno s http://hdl.handle.net/11356/1745

de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 76–84). Pridobljeno s https://aclanthology.org/W16-4011

22 https://www.cost.eu/actions/CA21167/

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4585–4592). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. Pridobljeno s https://doi.org/10.1162/coli_a_00402

Dobrovoljc, K., Erjavec, T., & Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. *Zbornik Konference Jezikovne Tehnologije in Digitalna Humanistika, 29. September - 1. Oktober 2016, Filozofska Fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija* (str. 190–192). Pridobljeno s http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Dobrovoljc-et-al_Pretvorba-korpusa-ssj500k.pdf

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 33–38). doi: 10.18653/v1/W17-1406

Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2019). Improving UD processing via satellite resources for morphology. *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, (pp. 24–34). Pridobljeno s https://doi.org/10.18653/v1/W19-8004

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik-Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0*. Pridobljeno s http://hdl.handle.net/11356/1230

Dobrovoljc, K., & Ljubešić, N. (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, (pp. 15–22). Pridobljeno s https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1566–1573). Pridobljeno s https://aclanthology.org/L16-1248

Dobrovoljc, K., & Terčon, L. (2023). Universal Dependencies: Smernice za označevanje besedil v slovenščini. Pridobljeno s https://wiki.cjvt.si/attachments/23

Dobrovoljc, K., Marušič, F., Mišmaš, P. & Žaucer, R. (2023). Odprta vprašanja pri prenosu označevalne sheme Universal Dependencies na slovenska besedila: Priloga k smernicam. Pridobljeno s https://wiki.cjvt.si/attachments/25

Dozat, T., & Manning, C. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing. *5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings*. doi: 10.48550/arxiv.1611.01734

Erjavec, T. (2013). Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, *1*(1), 24–49. doi: 10.4312/slo2.0.2013.1.24-49

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022). Raziskovalna infrastruktura CLARIN.SI. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (pp. 47–54). Inštitut za novejšo zgodovino. Pridobljeno s https://nl.ijs.si/jtdh22/pdf/JTDH2022_Erjavec-et-al_Raziskovalna-infrastruktura-CLARIN.SI.pdf

Erjavec, T., Fišer, D., Čibej, J., Arhar Holdt, Š., Ljubešić, N., Zupan, K., & Dobrovoljc, K. (2019). *CMC training corpus Janes-Tag 2.1*. Pridobljeno s http://hdl.handle.net/11356/1238

Erjavec, T., Fišer, D., Krek, S., & Ledinek, N. (2010, May). The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10336–10341. doi: 10.1073/PNAS.1502134112/SUPPL_FILE/PNAS.1502134112.ST01.PDF

Guzmán Naranjo, M., & Becker, L. (2018). Quantitative Word Order Typology with UD. *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway* (pp. 91–104).

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Ide, N., & Pustejovsky, J. (2017). Handbook of linguistic annotation / Nancy Ide, James Pustejovsky, editors. In *Handbook of linguistic annotation*. Springer.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 3rd Edition Draft*. Prentice Hall, Pearson Education International.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., & Brank, J. (2020). The ssj500k training corpus for Slovene language processing. *Jezikovne Tehnologije in Digitalna Humanistika*, 24–33. Pridobljeno s http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2019). *Dependency tree extraction tool STARK 1.0*. Pridobljeno s http://hdl.handle.net/11356/1284

Ledinek, N. (2018). Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in Slovstvo*, *63*(2/3), 103–116. Pridobljeno s http://www.dlib.si/details/URN:NBN:SI:doc-N94NNL3K

Lenardič, J., Čibej, J., Arhar Holdt, Š., Erjavec, T., & Fišer, D. (2022). *CMC training corpus Janes-Norm 3.0*. Pridobljeno s http://hdl.handle.net/11356/1733

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). doi: 10.18653/v1/W19-3704

Ljubešić, N., & Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*. Pridobljeno s http://hdl.handle.net/11356/1204

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J.-L., Lipp, V., Váradi, T., Győrffy, A., László, S., … Munda, T. (2022). *Parallel sense-annotated corpus ELEXIS-WSD 1.0*. Pridobljeno s http://hdl.handle.net/11356/1674

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen, B. S., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovolijc, K., Ureña-Ruiz, R.-J., Sancho-Sánchez, J.-L., Lipp, V., Varadi, T., Györffy, A., … Munda, T. (2021). Designing the ELEXIS Parallel Sense--Annotated Dataset in 10 European Languages. *ELex 2021 Proceedings*. Pridobljeno s https://elex.link/elex2021/

Nguyen, M. van, Lai, V. D., Pouran Ben Veyseh, A., & Nguyen, T. H. (2021). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 80–90). doi: 10.18653/v1/2021.eacl-demos.10

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4034–4043). Pridobljeno s https://aclanthology.org/2020.lrec-1.497

Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv. doi: 10.48550/ARXIV.2003.07082

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1351–1361). doi: 10.18653/v1/2021.eacl-main.115

Štravs, M., & Dobrovoljc, K. (2022). *Service for querying dependency treebanks Drevesnik 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1715

Terčon, L., & Ljubešić, N. (2023). The CLASSLA-Stanza model for UD dependency parsing of standard Slovenian 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1769

Terčon, L. & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. arXiv. doi: 10.48550/arXiv.2308.04255

Toporišič, J. (2000). *Slovenska slovnica*. Založba Obzorja Maribor.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. doi: 10.18653/v1/K18-2001

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Algom, A., Andersen, E., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranes, G., … Ziane, R. (2022). *Universal Dependencies 2.10*, http://hdl.handle.net/11234/1-4758

Žitnik, S. (2019). *Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0*, http://hdl.handle.net/11356/1285

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework*, http://hdl.handle.net/11356/1469

Universal Dependencies for Slovenian: An Upgrade to the Guidelines, Annotated Data and Parsing Model

Universal Dependencies (UD) is an internationally coordinated annotation scheme for cross-linguistically comparable morphosyntactic annotation of corpora, which has been applied to more than 130 other languages worldwide, including Slovenian. In this paper, we present the results of recent activities related to Slovenian UD annotation within the Development of Slovene in a Digital Environment project. During the project, we upgraded the existing infrastructure with reviewed and detailed documentation of the Slovenian UD annotation guidelines and produced four new datasets, manually annotated in accordance with the scheme. Specifically, we expanded the SSJ-UD treebank for written Slovenian with new sentences from the ssj500k and ELEXIS-WSD corpora, and created a new hidden UD treebank based on the SentiCoref corpus to be used on the SloBENCH evaluation platform. In addition, the SUK and Janes-tag reference training corpora, originally annotated using the language-specific JOS annotation scheme, have been semi-automatically converted to UD part-of-speech categories and morphological features. The new version of the reference SSJ-UD treebank with more than 5,000 new sentences and double the original number of tokens was used to train a new dependency parsing model in the CLASSLA-Stanza annotation tool. This paper gives an in-depth evaluation of its performance with respect to the overall parsing performance, the relation-specific parsing performance and the most common types of errors produced.

Keywords: linguistic annotation, dependency grammar, treebanks, dependency parsing, natural language processing

Adapting an English Corpus and a Question Answering System for Slovene

Uroš ŠMAJDEK
Faculty of Computer and Information Science, University of Ljubljana

Matjaž ZUPANIČ
Faculty of Computer and Information Science, University of Ljubljana

Maj ZIRKELBACH
Faculty of Computer and Information Science, University of Ljubljana

Meta JAZBINŠEK
Faculty of Arts, University of Ljubljana

Developing effective question answering (QA) models for less-resourced languages like Slovene is challenging due to the lack of proper training data. Modern machine translation tools can address this issue, but this presents another challenge: the given answers must be found in their exact form within the given context since the model is trained to locate answers and not generate them. To address this challenge, we propose a method that embeds the answers within the context before translation and evaluate its effectiveness on the SQuAD 2.0 dataset translated using both eTranslation and Google Cloud translator. The results show that by employing our method we can reduce the rate at which answers were not found in the context from 56% to 7%. We then assess the translated datasets using various transformer-based QA models, examining the differences between the datasets and model configurations. To ensure that our models produce realistic results, we test them on a small

Šmajdek, U., Zupanič, M., Zirkelbach, M., Jazbinšek, M. Adapting an English Corpus and a Question Answering System for Slovene. Slovenščina 2.0, 11(1): 247–274.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.247-274

subset of the original data that was human-translated. The results indicate that the primary advantages of using machine-translated data lie in refining smaller multilingual and monolingual models. For instance, the multilingual CroSloEngual BERT model fine-tuned and tested on Slovene data achieved nearly equivalent performance to one fine-tuned and tested on English data, with 70.2% and 73.3% questions answered, respectively. While larger models, such as RemBERT, achieved comparable results, correctly answering questions in 77.9% of cases when fine-tuned and tested on Slovene compared to 81.1% on English, fine-tuning with English and testing with Slovene data also yielded similar performance.

Keywords: question answering, machine translation, multilingual models

1 Introduction

One of the goals of artificial intelligence is to build intelligent systems that can interact with humans and help them. One such task is reading the web and then answering complex questions about any topic with regard to the given context. These question answering systems could have a big impact on the way that we access information. Furthermore, open-domain question answering is a benchmark task in the development of artificial intelligence, since understanding text and being able to answer questions about it is something that we generally associate with intelligence.

Question answering (QA) is one of the disciplines in the broader field of natural language processing (NLP), which involves the automatic answering of questions posed in natural language. Thus, the goal of QA is the development of automated systems that can understand and respond to human questions in a way that is similar to how humans answer questions. The QA task is typically formulated as follows: given a question and a context, the system must identify the answer to the question within the given context.

Recently, pre-trained contextual embedding (PCE) models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have attracted plenty of attention due to their good performance in a wide range of NLP tasks, including QA. Compared to earlier

information retrieval and knowledge-based systems, modern QA systems are significantly less domain-dependent, as they do not require a specifically tailored database to function effectively. This has thus led to the development of multilingual question answering systems, where the same system can serve a multitude of languages.

However, multilingual QA tasks typically assume that answers exist in the same language as the question, and require a smaller corpus to fine-tune it for a given language and a broader domain (e.g. Wikipedia articles). Yet in practice, many languages face both information scarcity, where languages have few reference articles, and information asymmetry, where questions reference concepts from other cultures. Due to the sizes of modern corpora, performing human translations is generally not feasible, and therefore we often employ machine translations instead. However, machine translation has trouble interpreting the nuances of specific languages, such as culturally specific vocabulary (e.g. translating bird sanctuary as "ptičje zatočišče", where the correct translation is "ptičji rezervat"), the use of articles, proper nouns, abbreviations, and implicit relationships between words (Koehn and Knowles, 2017; Arnejšek and Unk, 2020). This is especially problematic in question answering, where the answer has to be found in its exact form within the context to be usable for training such a model.

The objective of our work is thus to reduce the impact of errors in the construction of a machine-translated (MT) dataset that can be used to both fine-tune and test a question answering (QA) model. Specifically, we focus on the translation of the popular SQuAD 2.0 (Rajpurkar et al., 2018) QA dataset. Moreover, we benchmark the accuracy of QA models fine-tuned using the proposed MT dataset by assessing them on a human-translated (HT) subset of the original data.

The main contributions of our work are:

- a pipeline for translation of an English QA dataset;
- performance comparison of the various monolingual and multilingual QA models fine-tuned on the original dataset and the English-to-Slovene MT datasets;
- comparison of the eTranslation and Google Cloud Translation services in terms of raw translation and QA performance using the data translated from English to Slovene; and

- evaluation of the QA performance of the resulting QA models on the HT subset.

This paper is a follow-up to our submission to JDTH 2022 (Zupanič et al., 2022). To improve upon the presented concept, we expanded our evaluation to include the corpus translated by the state-of-the-art Google Cloud Translation (Google CT) service to assess the impact of the quality of translations. In addition, to ensure the testing set is not influenced by the machine translation, we replaced the post-edited machine translation samples with a fully human-translated testing set. Lastly, we also experimented with additional model parameters during evaluation and improved the presentation of our method.

In Section 2 we present the related work. In Section 3 we present our dataset, the process of translation, and evaluate the quality of the translation. In Section 4 we give a brief overview of the models used in the evaluation. In Section 5 we present the evaluation and discuss the results in Section 6. In Section 7 we present the conclusions and give possible extensions and enhancements for future work.

2 Related work

Early question answering systems, such as LUNAR (Woods & WOODS, 1977), date back to the 1960s and the 1970s. They were characterized by a core database and a set of rules, both handwritten by experts of the chosen domain. Over time, with the development of large online text repositories and increasing computer performance, the focus shifted from such rule-based systems to using machine learning and statistical approaches, like Bayesian classifiers and Support Vector Machines.

An example of this kind of system that was able to perform question answering in the Slovene language was presented by Čeh and Ojsteršek (2009), where the authors used classification and answer retrieval in parallel. The system retrieved data from its own database, consisting of MS Excel files, local databases, and integrated web services. For question classification, they used Support Vector Machines. The problem was that the system was very limited in question answering, able to answer only specific predefined classes of questions.

Another major revolution in the field of question answering and natural language processing, in general, was the advent of deep learning approaches and self-attention. One of the most popular approaches of this kind is BERT (Devlin et al., 2019), a transformer model introduced in 2018. Since then it has inspired many other transformer-based models, such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and T5 (Raffel et al., 2020), XLM (Lample and Conneau, 2019) and XLNet (Yang et al., 2019).

Such models also have the advantage of being able to recognize multiple languages, giving rise to multilingual models and model variants, such as Multilingual BERT, XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021) and RemBERT (Chung et al., 2021). Nevertheless, the training requires large amounts of training data, which many languages lack, leading to varying performance between different languages. They have also been shown to perform worse than monolingual models (Martin et al., 2020; Virtanen et al., 2019). Ulčar and Robnik-Šikonja (2020) thus made an effort to strike a middle ground between the performance of monolingual models and the versatility of multilingual ones by reducing the number of languages in the multilingual model to three – two similar less-resourced languages from the same language family and English. This resulted in two trilingual models, FinEst BERT and CroSloEngual BERT.

In 2020, a Slovene monolingual RoBERTa-based model called SloBERTa (Ulčar et al., 2021) was introduced. It was trained on five different corpora, totalling 3.41 billion words. The latest version of the model is SloBERTa 2.0, augmenting the original model by more than doubling the number of training iterations. The authors evaluated its performance on named-entity recognition, part-of-speech tagging, dependency parsing, sentiment analysis, and word analogy, but not on question answering.

While the described advances of natural language processing models already offer us a partial solution for the lack of language-specific training corpora, namely the ability to train the model on a language where large corpora are present (e.g. English), the models still require language-specific fine-tuning, for which a sizable corpus is needed. In our work, we present a potential solution to this problem, by using

machine-translation methods to translate smaller corpora to Slovene, and then fine-tune and evaluate the results.

3 Dataset description and methodology

In this section, we describe the dataset used in our study and the methodology employed to create machine- and human-translated datasets for fine-tuning and testing the question answering models.

3.1 Stanford Question Answering Dataset (SQuAD 2.0)

SQuAD 2.0 (Rajpurkar et al., 2018) is a reading comprehension dataset. It is based on a set of articles on Wikipedia that cover a variety of topics, from historical, pharmaceutical, and religious texts to texts about the European Union. Every question in the dataset is associated with a segment of text, or span, from the corresponding reading passage. It consists of over 100,000 question-answer pairs extracted from over 500 articles. Unlike SQuAD 1.0, the dataset contains roughly twice as much data, and also includes unanswerable questions, which are designed to look similar to answerable ones, but lack an answer within the given text. Thus, for a system to perform well on SQuAD2.0, it must not only answer questions, when possible but also determine when no

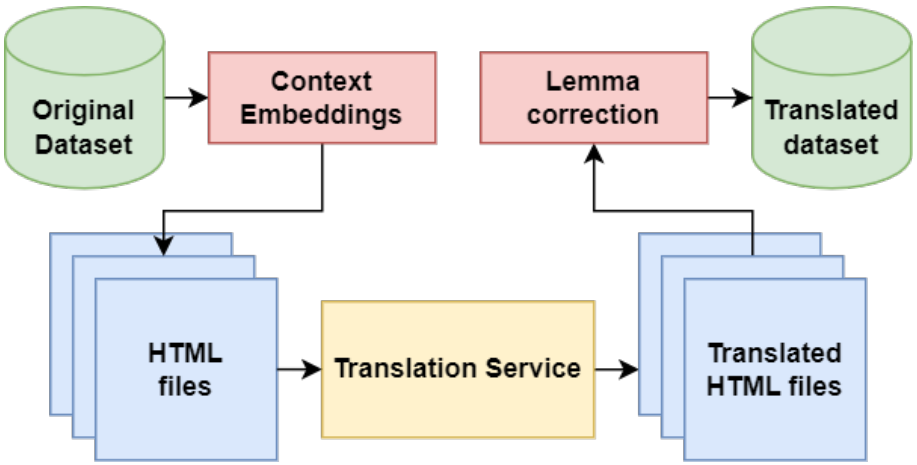


Figure 1: An overview of the machine translation pipeline.

answer is supported by the paragraph and abstain from answering. An example of a question-answer pair from the dataset is:

- **Question:** What is the name of the state referred to by historians during the Middle Ages as the Eastern Roman Empire?
- **Context:** During the Middle Ages, the Eastern Roman Empire survived, though modern historians refer to this state as the Byzantine Empire...
- **Answer:** the Byzantine Empire

3.2 Machine translation

In this subsection, we will describe the proposed machine-translation pipeline, a brief overview of which can be found in Figure 1. To translate the dataset into Slovenian we used two translation web services: eTranslation (European Commission, 2020) and Google CT. Due to the web services being primarily designed to translate webpages and short documents in DOCX or PDF format, our translation pipeline design was as follows:

1. Convert the corpus into HTML format. We wrapped context-question-answer coupling in separate HTML tags and placed them within a hierarchy resembling that in the original dataset. HTML tag attributes were used to preserve the unique question and answer identifiers for later evaluation. An example of the resulting structure can be seen in Appendix B.
2. Split the HTML file into smaller chunks. We found that 4 MB chunks work best, as larger chunks were often unable to be translated.
3. Send chunks to the translation service.
4. Use the original corpus file to compose the translated document in the original format.

The requirement for a context-question-answer coupling to be used to train a question answering model is that the answer has to be found in its exact form within the context. For example, if the answer entity is *Bizantinsko cesarstvo*, but the relevant part of the context was translated as *...v Bizantinskem cesarstvu...*, such a question-answer pair is unusable. This is due to the model's task being to find the index

of an answer to the question within the context. To improve upon basic translation, we employed two different methods.

The first was to correct the answers by breaking down both the answer and the context into lemmas and searching for the lemmatized form of the answer in the lemmatized form of the context. To accomplish this, the CLASSLA (CLARIN Knowledge Centre for South Slavic languages) language processing pipeline (Ljubešić and Dobrovoljc, 2019) was used. If a match was found, we replaced the bad answer with the matching original text in the context.

The second method was to embed the answers in the context before translation. This was done by replacing the answer entry of the untranslated document with a copy of the context entry which had the answer marked by a common HTML tag and a unique attribute to avoid mistaking it for a preexisting tag within the context. This allows the translation to also take into account the context surrounding the answer, greatly increasing the chance such an answer will be found in the original context. As the locations of answers within contexts are given by the dataset, finding the correct context entry is a trivial operation. For example, the untranslated answer entry '*the Byzantine Empire*' was replaced with the following: '*During the Middle Ages, the Eastern Roman Empire survived, though modern historians refer to this state as the Byzantine Empire...*'.

3.3 Human translation

When we added the Google CT service, we replaced the post-editing with the completely human translation of the excerpts so that the end comparison would be more objective and of better quality. The translation was done on a small number of automatically translated excerpts chosen randomly due to limited human resources.

The provided excerpts included original paragraphs or contexts, questions, and answers. Firstly, we translated the paragraphs and then the questions and answers since the answers had to match the text in the paragraph. As mentioned above in the description of the dataset, the topics of the original texts were very diverse and technical, covering different domains such as religion, history, politics, mathematics, and chemistry.

In total, there were 30 translated contexts with accompanying answerable and unanswerable questions, as well as impossible questions. The exact number of different segment types can be seen in Table 1.

Table 1: *Statistics for manually translated data subset*

Segment content	Number of instances
Context	30
Answerable question	142
Answer	435
Impossible question	143
Total number	**750**

4 Models

In this section, we present each of the five models that were used in the evaluation (Table 2). Since those transformer models are usable for a various number of natural language tasks we used Hugging Face's question answering pipeline to infer with question answering models.[1] The following models were selected as they are diverse in terms of properties and are publicly available, well documented, and have shown promising performance figures in the past.

Table 2: *Used models with the respective properties*

Model Name	Trained languages	No. of training tokens	No. of hidden layers
SloBERTa	1	3.47 billion	12
CroSloEngual BERT	3	5.9 billion	12
Multilingual BERT	104	No data	12
XLM-RoBERTa	100	6.3 trillion	24
RemBERT	110	1.8 trillion	32

4.1 SloBERTa

SloBERTa (Ulčar & Robnik-Šikonja, 2021) is a Slovene monolingual large pre-trained masked language model. It is closely related to the French CamemBERT model, which is similar to the base RoBERTa with

1 https://huggingface.co/docs/transformers/tasks/question_answering

12 hidden transformer layers but uses a different tokenization model. Since the model requires a large dataset for training, it was trained on five combined datasets with 3.47 billion tokens. It outperformed existing Slovene models.

4.2 CroSloEngual BERT

CroSloEngual BERT (Virtanen et al., 2019) is a trilingual model based on a BERT base with 12 hidden transformer layers and trained for Slovene, Croatian, and English, using 5.9 billion tokens from these languages. For those languages it performs better than multilingual BERT, which is expected, since studies show that monolingual models perform better than large multilingual ones (Virtanen et al., 2019).

4.3 Multilingual BERT

Multilingual BERT (M-BERT) (Devlin et al., 2019) is a version of BERT that has been trained on data from Wikipedia in 104 languages to satisfy the demand for multilingualism. It can perform tasks such as question answering, language classification, and many more for a wide range of languages.

The model was released in large form with 24 hidden transformer layers and in base form with 12 hidden transformer layers. We used the latter one in the current work.

4.4 XLM-RoBERTa

XLM-RoBERTa (XLM-R) (Conneau et al., 2020) is a pre-trained cross-lingual language model developed by Facebook AI. It is trained on 2.5 TB of CommonCrawl data, with a total of 6.3 trillion tokens in 100 languages, and based on RoBERTa (Robustly Optimized BERT Pretraining Approach) (Lample and Conneau, 2019). XLM-R, like M-BERT, uses a similar pretraining objective. However, XLM-R has a larger model size, shared vocabulary, and is trained using more training data from the web. XLM-R large, which we used in our work, has 24 hidden layers.

4.5 RemBERT

RemBERT (Chung et al., 2021) is a pre-trained multilingual model using a masked language modelling (MLM) objective. This model is pre-trained on 1.8 trillion tokens in 110 languages and is similar to mBERT. However, it differs in that its input and output embeddings are not tied. Instead, it uses small input embeddings and larger output embeddings, which makes the model more efficient because the output embeddings are discarded during fine-tuning. RemBERT, which has 32 hidden transformer layers, is the largest model we tested.

5 Evaluation and results

In this section, we present the evaluation results of our machine translation methods and the performance of the question answering models fine-tuned on the translated datasets.

5.1 Machine translation

To evaluate the quality of different translation methods, we measured how many answers can still be found within their respective context in their exact form. The results for the eTranslation service can be seen in Table 3. The resulting number of valid questions for both translation services, compared with the original dataset, are presented in Table 4.

Table 3: *Results for basic translation, lemma correction (LC), and context embedded (CE) translation of SQuAD 2.0 dataset by eTranslation*

Basic	LC	CE	LC+CE
44%	66%	93%	94%

Note. The percentages represent the number of answers that can be found within their respective context in their exact form.

Table 4: *Number of questions in the original SQuAD 2.0 dataset and our machine-translated datasets*

Dataset	Subset	AQ	AQ [%]	IQ	Total
Original	Train	86,821	66.6%	43,498	130,319
	Test	**5,928**	49.9%	5,945	11,873
eTranslation	Train	81,884	65.3%	43,498	125,382
	Test	**5,735**	49.1%	5,945	11,680
Google CT	Train	84,048	65.9%	43,498	127,546
	Test	**5,821**	49.5%	5,945	11,766

Note. AQ denotes the number of answerable questions and IQ the number of impossible questions.

5.2 Question answering

In order to evaluate the question answering performance of MT datasets obtained by eTranslation and Google CT, we first used both of them and the original English dataset, to fine-tune the following question answering models: M-BERT, XLM-R, RemBERT, SloBERTa 2.0, CroSloEngual BERT. This yielded 14 different fine-tuned model configurations, as showcased in the first two columns of Table 6. The fine-tuned models were then evaluated in two stages described later in the section. All tests were performed on a system with an Intel Xeon E5-2687W v4 @ 3.00GHz CPU and RTX 3090 24GB GPU. Before the evaluation, we removed all punctuation, leading and trailing white spaces, and articles from both ground truth and prediction. Both of them were also set in lowercase. The parameters used for fine-tuning are presented in Table 5.

The metrics used for the evaluation matched the official ones for the SQuAD 2.0 evaluation and were as follows:

- **Exact** - The fraction of predictions that matched at least of one the correct answers exactly.
- **F1** - The average overlap between prediction and ground truth, defined as an average of F1 scores for individual questions. The F1 score of an individual question is computed as a harmonic mean of the precision and recall, where precision was defined as T_M/T_P, and recall as T_M/T_{GP}, where T_M represents the maching tokens between

prediction and ground trouth, T_P number of tokens in prediction and T_{GP} number of tokens in the ground truth. A token is defined as a word, separated by white space.

In the first step of the evaluation, each of the models was tested using the HT subset of the original English testing set, and its untranslated counterpart. Additionally, we also repeated each of the tests on MT testing subsets, in order to assess their viability to be used as testing sets. The F1 scores can be seen in Table 6, whereas a full result overview is presented in Appendix C, Table 10.

In the second step of the evaluation, we repeated the tests with full original and MT testing sets to account for the potential discrepancies between the difficulty of the original dataset and the subset in the first set of experiments. The results of this step can be seen in Table 7 and Appendix C, Table 11. By comparing this set of results with the ones obtained in the first step we can see that the full set gives slightly better results, implying that the questions chosen in the subset were more difficult on average.

Table 5: *Parameters used to fine-tune the evaluated models*

Model Name	B	MS	LR	E
XLM-R Large	4	265	1e-5	3
M-BERT	8	256	1e-5	3
CroSloEngual BERT	8	320	1e-5	3
RemBERT	4	256	1e-5	3
SloBERTa 2.0	8	320	1e-5	3

Note. B denotes the number of batches used during fine-tuning, MS the maximum sequence length, LR the learning rate, and E the number of epochs.

6 Discussion

6.1 Quantitative analysis

By comparing the results of matching entries in Tables 6 and 7, we can observe that the results are consistently better when using the entire dataset instead of the randomly chosen subset for testing. The

differences are relatively minor though and both tables still show the same trends, which we would interpret as a positive indicator of the results in Table 6 being a good representation of the behaviour of the entire dataset.

6.1.1 Manual versus machine translation

By comparing the results of tests using HT data of a model fine-tuned on MT data, with the same model fine-tuned on original data (Table 6) we can see that the impact of fine-tuning with MT datasets varies depending on the size and the inherent performance of the model. The largest performance gain from fine-tuning a multilingual model on MT data as opposed to original data can be observed with M-BERT (F1 score of 74.0% as opposed to 58.2%) and CroSloEngual BERT (F1 score of 73.6% as opposed to 65.1%). However, for the latter, this is only true when using the set translated by Google CT. The impact is less noticeable for the larger RemBERT (F1 score of 81.5% as opposed to 79.5%) model, while worse for the XLM-R Large (F1 score of 81.4% as opposed to 81.6%) model. We reason this to be the result of the inherent ability of those two models to perform well when trained and tested on two different languages. This is clearly visible in our case as these two models, unlike smaller ones, retained much of their ability to answer English questions even when fine-tuned on MT data. It is harder to evaluate the impact of using the MT fine-tuning set on SloBERTa 2.0 since we cannot benchmark it against the same model fine-tuned on the original data, but comparing its results to other similarly sized models – M-BERT and CroSloEngual BERT – we can see that it outperforms both of them, which we would consider as a positive indicator for the viability of using MT data to fine-tune the model. Taking all of this into account we would conclude that using MT data to fine-tune question answering models is a superior option to using original English data, especially if one is constricted to using small models. Additionally, it has the benefit of enabling the use of Slovene monolingual models, which tend to outperform multilingual models of the same size.

6.1.2 Comparison of translation services

Comparing the results of the models fine-tuned with data translated by eTranslation and Google CT and tested on HT data in Table 6, we can observe that in most cases Google CT outperforms eTranslation. The exception to this is M-BERT, where eTranslation slightly outperformed its counterpart, but the difference is not significant. The magnitude of the differences varies from almost insignificant with M-BERT and XLM--R Large, to noticeable with CroSloEngual BERT and RemBERT. We suspect this is due to the inherent differences in the model structure and the data they were pre-trained on.

By observing the results of different models and their variations when tested on the data translated by MT as opposed to the results obtained by testing on HT data (Table 6), we can see that the results vary significantly. We suspect that this is due to the various grammatical errors present in the MT data, as shown in Section 6.2, which are not present in the data that was used to pre-train the models, and thus the models have a harder time recognizing the structure of the context. This is further reinforced by the fact that models consistently yield better results when tested using data translated by Google CT as opposed to eTranslation (Tables 6 and 7) since the former contains fewer grammatical and semantic mistakes (Section 6.2). Additionally, we can also observe a bias where models fine-tuned with one translation service perform better when tested on the same translation service as compared to the model fine-tuned with the other translation service tested on that testing set. All this points us toward concluding that while the MT datasets are a viable solution for fine-tuning they are less suitable for testing, especially if the resulting translation is of lower quality, such as when using the eTranslation service.

Table 6: *Comparison of the F1 scores of various models and their fine-tuning configurations on the human-translated subset of SQuAD 2.0 (N=285), and the subsets containing the same question from the original English dataset and the two machine-translated datasets*

Model Name	Fine-Tuning Dataset	Original [%]	eTranslation [%]	Google CT [%]	Human Transl. [%]
M-BERT	Original	78.4	48.2	55.9	58.2
	eTranslation	62.6	64.5	73.4	74.0
	Google CT	65.1	64.8	71.4	73.6
CroSloEngual BERT	Original	75.5	60.8	63.4	65.1
	eTranslation	63.1	58.8	64.5	63.6
	Google CT	58.8	66.5	66.6	73.6
SloBERTa 2.0	eTranslation	65.0	72.2	76.1	74.9
	Google CT	65.2	68.0	72.9	78.3
XLM-R Large	Original	85.5	69.1	75.8	**81.6**
	eTranslation	82.6	**73.1**	76.9	81.1
	Google CT	82.3	70.9	**77.4**	81.4
RemBERT	Original	**87.2**	71.4	74.3	79.5
	eTranslation	84.1	72.9	76.6	78.6
	Google CT	84.8	71.6	76.0	**81.5**

Note. Specific parameters used in fine-tuning are presented in Table 5.

Table 7: *Comparison of F1 scores of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and the two Slovene machine-translated SQuAD 2.0 evaluation datasets (N=11.680)*

Model Name	Fine-Tuning Dataset	Original [%]	eTranslation [%]	Google CT [%]
M-BERT	Original	78.9	59.2	61.9
	eTranslation	68.2	68.3	70.7
	Google CT	68.9	67.9	71.3
CroSloEngual BERT	Original	76.3	63.5	66.8
	eTranslation	68.2	65.5	68.3
	Google CT	65.7	66.5	70.0
SloBERTa 2.0	eTranslation	64.7	73.7	76.8
	Google CT	66.9	72.8	77.0
XLM-R Large	Original	86.3	74.8	78.5
	eTranslation	83.0	**75.6**	78.3
	Google CT	84.4	**75.5**	**80.1**
RemBERT	Original	**87.5**	71.4	74.3
	eTranslation	83.9	72.9	76.6
	Google CT	84.5	71.6	76.0

Note. The English dataset only contains the questions pre-set in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 5.

6.2 Qualitative analysis of translations

A comparison of the differences and the types of mistakes in the two machine translations and the human translation was made. The representativeness of these differences cannot be determined, but by looking at more examples some general mistakes of machine translations can be noted. It should also be considered that some of the mistakes of machine translation are more severe than others, and that in some segments there is a much greater number of them than in others, so the mistakes could not be counted exactly.

Firstly, the segments with contexts are very long and this normally led to more grammar, syntactic and stylistic mistakes in machine translations. The eTranslation MT yielded the worst results, as can be seen in Appendix A, example 1, as there was a wrong gender agreement and a big semantical mistake ('caving in' translated as 'jamarstvo'), which did not occur with Google CT. This was expected to a certain degree, as Google CT uses state-of-the-art translation methods, while eTranslation does not. Additionally, eTranslation is designed to perform best when working with texts on EU-related matters (European Commission, 2020) while our dataset is comprised of technical texts which cover a wide variety of topics.

There was also a great dissimilarity between the translations of answerable and impossible questions, because machine translation provided incoherent results. The changes are more notable because they affect the overall understanding of potential readers. These segments are shorter, but in both MT examples the word 'plants' was translated literally, so we can see in the example in the Appendix A that the HT translation is still the best one. Nevertheless, there was a larger number of instances where Google CT performed better than eTranslation at the grammatical and syntactical levels.

The segments with answers were the most similar ones, most probably because they are shorter. The contextual mistakes in the answers were for the most part already corrected in the contexts. More severe mistakes include semantic mistakes (e.g. plants translated as 'rastline', not 'naprave') and completely wrong answers (e.g. empty segment instead of 'Fermilab' or 'in' instead of '1,388'). Some frequent mistakes also occurred in translations of the names of movements, books, projects, or other names (e.g. 'Bricks for Varšava' was left untranslated by

eTranslation MT, Google CT did translate it to 'Opeke za Varšavo' and was changed in HT to 'Zidaki za Varšavo'). There were some punctuation errors, but the most interesting are grammatical mistakes of both MT services, especially when the wrong grammatical case, gender, or number is used. The answers had to be in the exact same form, so many answers do not sound coherent, which is of course not the case for English, where the conjugation does not change the words as much (e.g. with eTranslation 'Which part of China had people ranked higher in the class system?' — 'Northern' — 'V katerem delu Kitajske so bili ljudje višje v razrednem sistemu?' — 'Severni'). On the other part, some corrected segments were identical even though the source was different due to the use of articles in the English language (e.g. 'North Sea' and 'the North Sea' were both translated as 'Severno morje'). This occurred with both MT, but in some cases Google CT performed better, producing more exact matches. It was also better at capturing the same amount or length of answers as in the original. The answer of eTranslation for 'harvests of their Chinese tenants' was: 'čemer je dohodek od žetve kitajskih najemnikov', whereas Google CT captured only 'žetve njihovih kitajskih najemnikov'.

It should also be noted that the database SQuAD 2.0 is not entirely reliable. From the batch of randomly sampled 142 test question and answer groups, there were 14 occurrences where at least one of the given answers was not correct (e.g. 'Advanced Steam movement' instead of 'pollution' as an answer to 'Along with fuel sources, what concern has contributed to the development of the Advanced Steam movement?').

6.3 Qualitative analysis of predictions

By observing the individual cases of incorrect predictions, the main source of errors seems to stem from the grammatical and stylistic errors of the machine translation and occasionally its inability to convey the right meaning. The correct predictions are most likely the ones where the answer to the question is short and the words are not conjugated, i.e. numbers and names, even though there are some exceptions.

In the examples provided, we can see that there are two types of errors that we looked at. The first is when there is a wrong answer, but a right prediction (in Table 8), and the second is the correct answer and

the wrong prediction (Table 9). Most of the time, the wrong answers and predictions occur with the eTranslation service, and improvement of Google CT and HT is visible from a few representative examples, but sometimes, when the questions are more complicated, even the Google CT and HT do not provide a prediction at all, while sometimes only HT provides the correct prediction.

Table 8: *Examples of correct predictions with wrong answers*

#	Dataset	Question	Answer	Prediction
1	ENG	How many of Warsaw's inhabitants spoke Polish in 1933?	833,500	833,500
	ET	Koliko prebivalcev Varšave je leta 1933 govorilo poljsko?	prebivalcev	833.500
	GCT	Koliko prebivalcev Varšave je leta 1933 govorilo poljsko?	833.500	833.500
	HT	Koliko prebivalcev Varšave je leta 1933 govorilo poljski jezik?	833.500	833.500
2	ENG	Who recorded "Walking in Fresno?"	Bob Gallion	Bob Gallion je
	ET	Kdo je posnel „Walking in Fresno?"	Bob	Bob Gallion
	GCT	Kdo je posnel "Walking in Fresno"? Kdo je posnel »Walking in Fresno«?	Bob Gallion	Bob Gallion
	HT	Kdo je posnel "Walking in Fresno"? Kdo je posnel »Walking in Fresno«?	Bob Gallion	Bob Gallion

Note. ENG denotes the English dataset, ET one translated by the eTranslation service, GCT one translated by the Google Cloud translation service, and HT one translated by a human.

Table 9: *Examples of correct answers with wrong predictions. ENG denotes the English dataset, ET one translated by the eTranslation service, GCT one translated by the Google Cloud translation service, and HT one translated by a human*

#	Dataset	Question	Answer	Prediction
1	ENG	How many total acres is Woodward Park?	300 acres	300 acres
	ET	Koliko hektarjev je Woodward Park?	300 hektarjev	235 hektarjev
	GCT	Koliko skupno hektarjev obsega Woodward Park?	300 hektarjev	300
	HT	Koliko akrov skupaj obsega park Woodward?	300 akrov	300
2	ENG	How many miles, once completed, will the Lewis S. Eaton trail cover?	22 miles	22
	ET	Koliko kilometrov, ko bo konč ano, bo pokrivalo Lewis S. Eaton?	22 milj	(35 km
	GCT	Koliko milj bo pot Lewisa S. Eatona pokrivala, ko bo konč ana?	22 milj	22
	HT	Koliko milj bo, ko bo dokonč ana, dolga pot Lewisa S. Eatona?	22 milj	22

7 Conclusion

In this work, we presented a method for the machine translation of question answering datasets. To evaluate the method, we applied it to the SQuAD 2.0 dataset and used the results to train and test the following question answering models: Multilingual BERT, CroSloEngual, SloBERTa 2.0, XLM-RoBERTa, and RemBERT. In order to discern the impact of the quality of the translated data we performed the translation with two different translation services: eTranslation and state-of--the-art Google Cloud Translation. To evaluate the models using as close to real data as possible, we took a small subset of the original testing set and manually translated it to Slovene, which formed the basis for the performance comparisons.

The results show that using machine-translated data for evaluation led to notably worse results as compared to the human-translated data. Moreover, we noticed that while multilingual models fine-tuned using machine-translated data performed better than ones fine-tuned on English data when given a task of answering the machine-translated question, the situation was in most cases reversed when given a task of answering human-translated questions. This leads us to conclude that machine translation, at least as available via the eTranslation service, is not particularly suitable for training multilingual models. Of all the models, SloBERTa 2.0 produced the best results on both machine- and human-translated data, while the RemBERT gave comparable results even when only fine-tuned on the English dataset.

The results show that the application of machine-translated data produced by our method leads to better results on smaller multilingual question answering models, as compared to fine-tuning them using the original, English, data. On the other hand, the results for larger models were mostly unaffected by the language of the dataset used to train them. The most notable benefit is the ability to fine-tune monolingual models, which would otherwise be unusable. Our experiments also show that this machine-translated data is not suitable for the purpose of testing the models. The impact of the quality of the translation service is minor and varies depending on the model.

The testing procedure could be improved by using a dataset that was already manually translated to Slovene, which would allow us to

benchmark our results against it as well. The experiment could also be expanded by including more datasets, such as Natural Questions (Kwiatkowski et al., 2019), and other models, such as Microsoft's mDeBERTaV3. Additionally, further effort could be dedicated to ascertaining the optimal parameters for fine-tuning the question answering models.

References

Arnejšek, M., & Unk, A. (2020). Multidimensional assessment of the eTranslation output for English–Slovene. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 383–392). Lisboa: European Association for Machine Translation. Retrieved from https://aclanthology.org/2020.eamt-1.41

Chung, H. W., Févry, T., Tsai, H., Johnson, M., & Ruder, S. (2021). Rethinking Embedding Coupling in Pre-trained Language Models. *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., . . ., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747

Čeh, I., & Ojsteršek, M. (2009). Developing a question answering system for the Slovene language. *WSEAS Transaction on Information science and applications*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (Vol. 1, pp. 4171–4186). Minneapolis: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

European Commission. (2020). CEF Digital eTranslation. *CEF Digital eTranslation*.

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Vancouver: Association for Computational Linguistics. doi: 10.18653/v1/W17-3204

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., . . ., & Petrov, S. (2019). Natural Questions: a Benchmark for Question

Answering Research. *Transactions of the Association of Computational Linguistics*.

Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=H1eA7AEtvS

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . ., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Florence: Association for Computational Linguistics. doi: 10.18653/v1/W19-3704

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., . . ., & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203–7219). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.645

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . ., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*, 1–67. Retrieved from http://jmlr.org/papers/v21/20-074.html

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. doi: 10.48550/ARXIV.1806.03822

Ulčar, M., & Robnik-Šikonja, M. (2020). Finest BERT and CroSloEngual BERT. *International Conference on Text, Speech, and Dialogue* (pp. 104–111).

Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. *In Proceedings of Data Mining and Data Warehousing, SiKDD*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., . . ., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Woods, W. A., & WOODS, W. A. (1977). Lunar rocks in natural English: Explorations in natural language question answering.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., . . ., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483–498). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems, 32*.

Zupanič, M., Zirkelbach, M., Šmajdek, U., & Jazbinšek, M. (2022). Preparing a corpus and a question answering system for Slovene. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (pp. 353–359). Ljubljana, Inštitut za novejšo zgodovino. Retrieved from https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Appendix A: Translation examples

Below are some examples of two machine translations and a human translation, where some specific differences, which occur more times, can be seen.

1. **Example of context segment (excerpt)**
    - Original
        - The Northern Chinese were ranked higher and Southern Chinese were ranked lower because southern China withstood and fought to the last before caving in.
    - eTranslation
        - Severna Kitajci so bili uvrščeni višje in južna Kitajci so bili uvrščeni nižje, ker je južna Kitajska zdržala in se borila do zadnjega pred jamarstvom.
    - Google Translate
        - Severni Kitajci so bili uvrščeni višje, južni Kitajci pa nižje, ker je južna Kitajska zdržala in se borila do zadnjega, preden je popustila.
    - Human translation
        - Severni Kitajci so bili uvrščeni višje in južni Kitajci so bili uvrščeni nižje, ker se je južna Kitajska pred predajo upirala in se borila do zadnjega.
2. **Examples of answerable question segment**
    - Original
    - Who was Al-Banna's assassination a retaliation for the prior assassination of?
    - What plants create most electric power?
    - eTranslation
    - Kdo je bil Al-Bannin umor maščevanja zaradi predhodnega umora?
    - Katere rastline ustvarjajo največ električne energije?
    - Google CT
    - Komu je bil atentat Al-Banne povračilo za prejšnji atentat?
    - Katere rastline proizvajajo največ električne energije?
    - Human translation
    - Al-Bannov umor je bil maščevanje za čigav predhodni umor?
    - Kateri obrati ustvarjajo največ električne energije?

3. **Example of impossible question segment**
    - Original
        - What Book of the Bible is knowledge of the law traced back to?
    - eTranslation
        - Do katere knjige Svetega pisma je znano pravo?
    - Google CT
        - Od katere svetopisemske knjige sega znanje o zakonu?
    - Human translation
        - V kateri svetopisemski knjigi že zasledimo poznavanje prava?

Appendix B: HTML Structure

```
<data class=0>
   <paragraph class=0>
      <context>The Normans (Norman: Nourmands; French: Normands; Latin:
      Normanni) were the people who in the 10th and 11th centuries gave
      their name to Normandy, a region in France. They were descended
      from Norse ("Norman" comes from "Norseman") raiders and pirates
      from Denmark, Iceland and Norway who, under their leader Rollo,
      agreed to swear fealty to King Charles III of West Francia. Through
      generations of assimilation and mixing with the native Frankish
      and Roman-Gaulish populations, their descendants would gradually
      merge with the Carolingian-based cultures of West Francia. The
      distinct cultural and ethnic identity of the Normans emerged ini-
      tially in the first half of the 10th century, and it continued to
      evolve over the succeeding centuries.</context>
   <qas class=0>
   <question>In what country is Normandy located?</question>
   <answer class=0>
   <text>France</text>
   </answer>
   </qas>
   <qas class=1>
   <question>When were the Normans in Normandy?</question>
   <answer class=0>
   <text>10th and 11th centuries</text>
      <in_context>The Normans (Norman: Nourmands; French: Normands; La-
      tin: Normanni) were the people who in the <b>10th and 11th cen-
      turies</b> gave their name to Normandy, a region in France. They
      were descended from Norse ("Norman" comes from "Norseman") raiders
      and pirates from Denmark, Iceland and Norway who, under their le-
      ader Rollo, agreed to swear fealty to King Charles III of West
      Francia. Through generations of assimilation and mixing with the
      native Frankish and Roman-Gaulish populations, their descendants
      would gradually merge with the Carolingian-based cultures of West
      Francia. The distinct cultural and ethnic identity of the Normans
      emerged initially in the first half of the 10th century, and it
      continued to evolve over the succeeding centuries.</in_context>
```

```
   </answer>
   <answer class=1>
   <text>in the 10th and 11th centuries</text>
      <in_context>The Normans (Norman: Nourmands; French: Normands; La-
      tin: Normanni) were the people who <b>in the 10th and 11th cen-
      turies</b> gave their name to Normandy, a region in France. They
      were descended from Norse ("Norman" comes from "Norseman") raiders
      and pirates from Denmark, Iceland and Norway who, under their le-
      ader Rollo, agreed to swear fealty to King Charles III of West
      Francia. Through generations of assimilation and mixing with the
      native Frankish and Roman-Gaulish populations, their descendants
      would gradually merge with the Carolingian-based cultures of West
      Francia. The distinct cultural and ethnic identity of the Normans
      emerged initially in the first half of the 10th century, and it
      continued to evolve over the succeeding centuries.</in_context>
   </answer>
   </qas>
   ...\\
   </paragraph>
   ...\\
</data>
```

Appendix C: Detailed results

Table 10: *Full comparison of the results of various models and their fine-tuning configurations on the human-translated subset of SQuAD 2.0 (N=285), and the subsets containing the same question from the original English dataset and the two machine-translated datasets*

Model Name	Fine-Tuning Dataset	Original [%]		eTranslation [%]		Google CT [%]		Human Transl. [%]	
		Exact	F1	Exact	F1	Exact	F1	Exact	F1
M-BERT	Original	76.1	78.4	42.8	48.2	53.0	55.9	55.1	58.2
	eTranslation	58.2	62.6	58.6	64.5	68.7	73.4	70.9	74.0
	Google CT	59.3	65.1	58.9	64.8	66.7	71.4	69.4	73.6
CroSloEngual BERT	Original	73.3	75.5	53.0	60.8	57.5	63.4	61.1	65.1
	eTranslation	59.6	63.1	51.6	58.8	58.2	64.5	60.0	63.6
	Google CT	55.8	58.8	60.7	66.5	61.8	66.6	70.2	73.6
SloBERTa 2.0	eTranslation	59.3	65.0	64.9	72.2	69.5	76.1	70.9	74.9
	Google CT	61.8	65.2	61.4	68.0	66.3	72.9	73.0	78.3
XLM-R Large	Original	83.5	85.5	61.4	69.1	70.9	75.8	**78.6**	**81.6**
	eTranslation	79.3	82.6	**66.0**	**73.1**	71.2	76.9	76.1	81.1
	Google CT	79.3	82.3	63.5	70.9	**71.2**	**77.4**	76.8	81.4
RemBERT	Original	**84.9**	**87.2**	64.2	71.4	69.1	74.3	74.0	79.5
	eTranslation	80.0	84.1	64.9	72.9	70.2	76.6	71.9	78.6
	Google CT	80.4	84.8	63.2	71.6	70.2	76.0	**77.9**	**81.5**

Note. Specific parameters used in fine-tuning are presented in Table 5.

Table 11: *Full comparison of the results of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and the two Slovene machine-translated SQuAD 2.0 evaluation datasets (N=11.680)*

Model Name	Fine-Tuning Dataset	Original [%]		eTranslation [%]		Google CT [%]	
		Exact	F1	Exact	F1	Exact	F1
M-BERT	Original	75.8	78.9	52.7	59.2	57.1	61.9
	eTranslation	63.7	68.2	61.7	68.3	65.7	70.7
	Google CT	64.4	68.9	61.4	67.9	66.9	71.3
CroSloEngual BERT	Original	72.9	76.3	56.2	63.5	61.5	66.8
	eTranslation	63.6	68.2	58.3	65.5	62.3	68.3
	Google CT	61.4	65.7	59.8	66.5	65.3	70.0
SloBERTa 2.0	eTranslation	60.6	64.7	66.6	73.7	71.4	76.8
	Google CT	63.9	66.9	65.6	72.8	72.3	77.0
XLM-R Large	Original	83.4	86.3	67.1	74.8	73.4	78.5
	eTranslation	79.0	83.0	**68.0**	**75.6**	72.3	78.3
	Google CT	80.9	84.4	**68.0**	**75.5**	**75.3**	**80.1**
RemBERT	Original	**84.5**	**87.5**	67.1	71.4	69.1	74.3
	eTranslation	79.1	83.9	66.8	72.9	70.2	76.6
	Google CT	80.1	84.5	67.0	71.6	70.2	76.0

Note. The English dataset only contains the questions pre-set in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 5.

Prilagoditev angleškega korpusa in sistema za odgovarjanje na vprašanja za slovenski jezik

Pomanjkanje ustreznih podatkov za učenje je ena od ključnih težav pri razvoju slovenskih modelov za odgovarjanje na vprašanja (QA). Sodobna orodja za strojno prevajanje lahko to težavo rešijo, vendar pa se pri njihovi uporabi soočimo z novih izzivom: odgovori se morajo natančno ujemati z deli danega konteksta, kjer ta odgovor je, saj model odgovorov ne generira, temveč le išče. Kot rešitev predlagamo metodo, kjer odgovore prevajamo skupaj s kontekstom, kar poveča verjetnost, da bo odgovor preveden v enaki obliki. Učinkovitost te metode ocenjujemo na naboru podatkov SQuAD 2.0, prevedenem z uporabo storitev eTranslation in Google Cloud, kjer se z njeno uporabo delež neujemanj odgovora in konteksta zmanjša s 56 % na 7 %. Prevedene podatke nato ocenimo z uporabo različnih QA modelov, ki temeljijo na arhitekturi transformer, in preučimo razlike med podatkovnimi nizi in

konfiguracijami modelov. Da zagotovimo čim bolj realistične rezultate, modele testiramo na človeških prevodih majhnega deleža izvirne zbirke podatkov. Rezultati kažejo, da se glavne prednosti uporabe strojno prevedenih podatkov pokažejo pri natančnem prilagajanju (angl. fine-tuning) manjših večjezičnih modelov in enojezičnih modelov. Večjezični CroSloEngual BERT model je na primer dosegel 70,2 % točnih ujemanj pri testiranju na slovenskih podatkih v primerjavi s 73,3 % točnih ujemanj pri testiranju na angleških podatkih. Medtem ko so bili rezultati pri večjih modelih podobni, pri čemer je RemBERT dosegel 77,9 % točnih ujemanj na slovenskih podatkih v primerjavi z 81,1 % na angleških podatkih, so se ti obnesli podobno tudi pri natančnem prilagajanju na angleških podatkih, kar pomeni, da jih strojno prevedeni podatki niso bistveno izboljšali.

Ključne besede: sistemi za odgovarjanje na vprašanja, strojno prevajanje, večjezični modeli

Praktični vidiki uporabe podbesednih enot v strojnem prevajanju slovenščina-angleščina

Gregor DONAJ
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Mirjam SEPESY MAUČEC
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Večina sodobnih sistemov za strojno prevajanje temelji na arhitekturi nevronskih mrež. To velja za spletne ponudnike strojnega prevajanja, za raziskovalne sisteme in za orodja, ki so lahko v pomoč poklicnim prevajalcem v njihovi praksi. Čeprav lahko sisteme nevronskih mrež uporabljamo na običajnih centralnih procesnih enotah osebnih računalnikov in strežnikov, je za delovanje s smiselno hitrostjo potrebna uporaba grafičnih procesnih enot. Pri tem smo omejeni z velikostjo slovarja, kar zmanjšuje kakovost prevodov. Velikost slovarja besednih enot je še posebej pereč problem visoko pregibnih jezikov. Rešujemo ga z uporabo podbesednih enot, s katerimi dosežemo večjo pokritost jezika. V članku predstavljamo različne metode razcepljanja besed na podbesedne enote z različno velikimi slovarji in primerjamo njihovo uporabo v strojnem prevajalniku za jezikovni par slovenščina-angleščina. V primerjavo vključujemo še prevajalnik brez razcepljanja besed. Predstavljamo rezultate uspešnosti prevajanja z metriko BLEU, hitrosti učenja modelov in hitrosti prevajanja ter velikosti modelov. Dodajamo pregled praktičnih vidikov uporabe podbesednih enot v strojnem prevajalniku, ki ga uporabljamo skupaj z orodji za računalniško podprto prevajanje.

Ključne besede: strojno prevajanje, velikost slovarja, podbesedne enote, grafične procesne enote

Donaj, G., Sepesy Maučec, M. Praktični vidiki uporabe podbesednih enot v strojnem prevajanju slovenščina-angleščina. Slovenščina 2.0, 11(1): 275–301.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.275-301

1 Uvod

Strojno prevajanje je postopek avtomatskega prevajanja besedil iz enega naravnega jezika v drugega. Dolgo časa je bilo v praksi zavrnjeno zaradi slabe kakovosti prevodov ali pa je celo ponujalo vir humorja zaradi nesmiselnih in smešnih prevodov. V zadnjih letih se je kakovost strojnega prevajanja bistveno izboljšala, kar je privedlo do njegove uporabe v prevajalski praksi.

1.1 Pristopi strojnega prevajanja

Od svojih začetkov do danes so se pristopi strojnega prevajanja naravnega jezika večkrat spremenili (Sepesy Maučec in Donaj, 2019). Začetki segajo v 50. leta prejšnjega stoletja, ko so prvi sistemi za prevajanje temeljili na pravilih. Kasneje so se pojavili sistemi, ki so izhajali iz primerov prevodov. Algoritmi so iskali podobne pare stavkov na obeh straneh jezikovnega para. Sledilo je strojno prevajanje na podlagi statističnih modelov (Brown idr., 1993) in nekje od leta 2015 naprej strojno prevajanje, ki temelji na uporabi nevronskih mrež in trenutno daje najboljše rezultate (Bahdanau idr., 2014; Vaswani idr., 2017). Uporaba nevronskih mrež zahteva dovolj veliko računsko zmogljivost sistemov in velike množice učnega gradiva, ki morajo biti na voljo, da lahko naučimo modele za prevajanje. Prav razpoložljivost obojega v zadnjih letih je bila povod za nedavni razcvet nevronskega strojnega prevajanja (Stahlberg, 2020).

Sodobni pristopi strojnega prevajanja temeljijo na treh osnovnih arhitekturah nevronskih mrež: nevronske mreže s povratno zanko (Recurrent Neural Network – RNN), konvolucijske nevronske mreže (Convolutional Neural Network – CNN) in arhitektura transformerjev s samo-pozornostjo (self-attention), ki so danes najbolj pogoste.

1.2 Tehnični izzivi

Nevronsko strojno prevajanje prinaša tehnične izzive. V praksi je nujna uporaba grafičnih procesnih enot (Graphical Processing Unit – GPU), ki imajo omejeno velikost delovnega pomnilnika, zaradi česar v določenih primerih ne moremo uporabljati poljubno velikih nevronskih mrež.

Velikost nevronske mreže oz. modela za prevajanje v strojnem prevajalniku je odvisna od izbrane arhitekture, nastavitev hiperparametrov nevronske mreže in velikosti slovarja. Omejena velikost delovnega pomnilnika nam tako omejuje velikost slovarja, kar pomeni slabo pokritost besedišča jezika in posledično dodatne napake pri prevajanju. Uporaba sistemskega delovnega pomnilnika ni praktična, saj je komunikacija z njim nekajkrat počasnejša in v nekaterih orodjih ni podprta. Tudi jeziki, med katerimi prevajamo, predstavljajo različne izzive zaradi svojih specifičnih lastnosti. Raziskave se večinoma osredotočajo na lastnosti, ki izvirajo iz morfologije jezika. Takšni lastnosti sta na primer pregibanje besed (češčina, slovenščina) in sestavljanje besed (nemščina, kitajščina). Obe imata za posledico velik slovar besed (Tamchyna idr., 2017).

1.3 Orodja za računalniško podprto prevajanje

Orodja za računalniško podprto prevajanje (Computer-Aided Translation – CAT) omogočajo uporabo pomnilnikov prevodov – seznamov že nastalih prevodov besed ali besednih zvez, ki jih prevajalec uporablja v nadaljevanju dela, ne omogočajo pa prevajanja neznanih fraz, ki niso na seznamu. Sodobna orodja podpirajo tudi strojno prevajanje. OPUS-CAT MT Engine[1] je primer uporabniku prijaznega sistema za strojno prevajanje, ki ga lahko uporabljamo na računalniku s sistemom Windows. Za ta sistem je na voljo vtičnik za različna orodja CAT, kot so Trados Studio, memoQ in OmegaT. S tem lahko strojno prevajanje približamo poklicnim prevajalcem. Naloga prevajalca je v tem primeru popravljanje napak, ki se pojavijo v strojnih prevodih. Če so strojni prevodi v osnovi dovolj dobri, prevajalec porabi manj časa za popravljanje, kot bi ga potreboval za ročno prevajanje (Etchegoyhen idr., 2014). Vtičnik prevajalcu omogoča tudi učenje modelov za strojno prevajanje oz. njihovo prilagajanje na izbrano domeno. To pa hkrati pomeni, da mora prevajalec poznati osnovne tehnične izzive učenja modelov. Nekatere med njimi naslavljamo v tem članku.

Takšen način uporabe strojnega prevajanja je smiseln tudi v primeru zaupnih besedil, saj le-ta ne zapustijo računalnika prevajalca. To pa ne drži v primeru uporabe spletnih storitev za prevajanje.

1 https://helsinki-nlp.github.io/OPUS-CAT/

Še vedno je za tipičnega prevajalca uporaba lastnega strojnega prevajalnika morda bolj zahtevna kot uporaba preprostih pomnilnikov prevodov. Del izziva je izbira pravega modela prevajanja. Želimo, da daje model dovolj kakovostne prevode, hkrati pa moramo zagotoviti, da ga uporabnik sploh lahko zažene na svoji strojni opremi.

1.4 Struktura članka

V članku bomo predstavili različne podatkovno vodene metode za razcepljanje besed, s katerimi zmanjšamo velikost slovarja in v celoti pokrijemo jezik. Izbrali smo metode, ki so dobro znane in uveljavljene, vendar temeljijo na različnih optimizacijskih kriterijih. Te metode bomo uporabili na primeru strojnega prevajanja med slovenščino in angleščino. Predstavili bomo rezultate uspešnosti prevajanja, hitrosti učenja in prevajanja ter velikosti izdelanih modelov in njihove porabe pomnilnika GPU. Vse metode bomo primerjali z besednim modelom brez razcepljanja.

V diskusiji bomo dodali praktične vidike uporabe takšnih pristopov v strojnem prevajanju ob uporabi lastne strojne opreme.

Ta članek je razširjena verzija prispevka na konferenci Jezikovne tehnologije in digitalna humanistika 2022 in predstavlja podrobnejšo predstavitev uporabljenih metod, dodatne rezultate in razširjeno diskusijo.

2 Slovarske enote v strojnem prevajanju

Najbolj intuitivna izbira slovarske enote prevajalnika je beseda, ki je najpogosteje izbrana osnovna enota v drugih postopkih jezikovnih tehnologij. Prinaša pa številne izzive. Za dovolj dobro pokritost besedišča jezika so potrebni veliki slovarji, kar je še posebej izrazit problem pri visoko pregibnih jezikih, kot je slovenščina. Posledica premajhnih slovarjev je visok delež neznanih besed oz. besed izven slovarja, ki močno zmanjša kakovost prevodov.

Za obvladovanje omenjenih težav so bile predlagane različne alternativne slovarske enote. Kot najmanjša slovarska enota je bila uporabljena črka oz. znak, ki se je izkazal kot zelo robustna enota, manj občutljiva na šum in razlike v domeni učnega in testnega korpusa (Heigold, Varanasi, Neumann, & van Genabith, 2018; Gupta, Besacier,

Dymetman, & Gallé, 2019). Potrebne so določene prilagoditve v arhitekturi nevronske mreže, saj je dolžina segmenta nekajkrat daljša od segmenta, ki kot slovarske enote uporablja besede. Posledica je slabše modeliranje odvisnosti na velikih razdaljah v besedilih.

Preizkušene so bile tudi slovarske enote, ki so po velikosti med črko in besedo – podbesedne enote. Podbesedne enote, dobljene s podatkovno vodenim razcepljanjem, ki kot enoto ohranja pogosta zaporedja črk, so se v splošnem izkazale kot najbolj učinkovite, saj v večji meri ohranjajo sintaktične in semantične lastnosti (Sennrich idr., 2016; Banerjee in Bhattacharyya, 2018). Ker je beseda lahko razcepljena na več različnih načinov, je bila predlagana metoda regulacije razcepljanja (Kudo, 2018). Kot slovarsko enoto bi lahko uporabili tudi jezikoslovno enoto morfem, vendar bi za pravilno delitev besed v algoritmu razcepljanja na morfeme potrebovali slovnično znanje. Podobno lahko rečemo za delitev besed na zloge.

Smiselnost razcepljanja besed na manjše slovarske enote kažejo tudi druga sorodna dela, ki se lotevajo nevronskega prevajanja za morfološko bogate jezike (Tukeyev idr., 2020; Marco idr., 2022).

Osnova za učenje modelov prevajanja so vzporedni korpusi besedil, ki vsebujejo poravnane segmente besedila. Poravnanost segmentov pomeni, da imata oba segmenta enak pomen in sta prevod drug drugega. Običajno so segmenti posamezne povedi, vendar so pogosti primeri, kjer je ena poved v nekem jeziku poravnana z dvema ali celo več povedmi v drugem jeziku. Korpus ni uporaben le za učenje prevajalnika, ampak tudi za učenje modelov za razcepljanje besed na podbesedne enote.

V nadaljevanju opisujemo različne postopke razcepljanja, ki smo jih uporabili v članku. Pri tem bo poudarek na postopku BPE, ki se najpogosteje uporablja. Vsi opisani postopki so podatkovno vodeni. To pomeni, da so modeli za razcepljanje besed naučeni izključno na korpusu, sami postopki pa so neodvisni od jezika.

2.1 Postopek Byte-Pair Encoding

Postopek BPE (Byte-Pair Encoding) je bil prvotno razvit kot postopek za stiskanje podatkov, ki deluje z iterativno zamenjavo najpogostejših

parov simbolov z novimi posameznimi simboli. Sennrich in drugi (Sennrich idr., 2016) so ta algoritem priredili za namen razcepljanje besed.

V postopku se najprej inicializira slovar, ki vsebuje vse črke in druge znake (številke, ločila, matematični simboli itd.), ki se pojavijo v korpusu, ter posebni simbol za konec besede. Vsebina korpusa se tako obravnava kot zaporedje simbolov, ki so v prvem koraku le črke in drugi znaki. Nato sledi iterativni postopek, v katerem algoritem poišče najpogostejši par zaporednih simbolov in se le-ta povsod nadomesti z novim simbolom. Te iterativne korake imenujemo združevanja, pri čemer pogostost parov simbolov izračunavamo na celotnem učnem korpusu. Pri postopku preskočimo združevanja, ki bi vključevala simbol za konec besede, kar v končnem korpusu prepreči združevanje besed, namesto njihovega razcepljanja. Tako nastane končni slovar podbesednih enot (združeni simboli) ter model, ki vsebuje seznam združevanj.

Parameter postopka je število združevanj in neposredno vpliva na velikost slovarja. Natančna velikost slovarja je enaka vsoti števila združevanj in števila simbolov v začetnem koraku. Dovolj veliko število združevanj pomeni, da se nekatere besede v celoti združijo nazaj v en simbol, enak izvorni besedi.

Najpogostejše besede v korpusu se pri uporabi modela ohranijo kot celote, redkejše pa se razcepijo na enote iz nastalega slovarja. Ker so v tem slovarju tudi posamezne črke, je skoraj zagotovljeno, da bo vsaka beseda ustrezno razcepljena in bo delež enot izven slovarja enak nič. Izjeme so zelo redke in se lahko pojavijo, ko v testnem besedilu zasledimo znak, ki se v učnem korpusu ne pojavi.

1:	d-r-ž-a-v-e	č-l-a-n-i-c-e	b-o-d-o	p-r-e-g-l-e-d-a-l-e	s-e-z-n-a-m-e	i-n	o-d	i-z-d-a-j-a-t-e-l-j-i-c-e
2:	d-r-ž-a-v-e	č-l-a-n-i-c-e	b-o-d-o	pr-e-g-l-e-d-a-l-e	s-e-z-n-a-m-e	i-n	o-d	i-z-d-a-j-a-t-e-l-j-i-c-e
5:	d-r-ž-a-v-e	č-l-an-i-c-e	b-o-d-o	pr-e-g-l-e-d-a-l-e	s-e-z-n-a-m-e	i-n	o-d	i-z-d-a-j-a-t-e-l-j-i-c-e
10:	d-r-ž-av-e	č-l-an-i-c-e	b-o-d-o	pr-e-g-l-ed-a-l-e	s-e-z-n-a-m-e	i-n	o-d	i-z-d-a-j-a-t-e-l-j-i-c-e
20:	d-r-ž-av-e	č-l-an-i-c-e	b-od-o	pr-e-g-l-ed-a-l-e	s-e-z-na-m-e	in	o-d	i-z-d-a-j-a-te-lj-i-c-e
50:	dr-ž-av-e	č-l-an-i-c-e	b-od-o	pre-g-l-ed-al-e	se-z-na-m-e	in	o-d	iz-d-a-j-a-te-lj-i-c-e
100:	držav-e	č-l-an-i-c-e	b-od-o	pre-g-l-ed-al-e	se-z-na-m-e	in	o-d	iz-da-ja-te-lj-i-c-e
200:	držav-e	član-ic-e	b-od-o	pre-g-led-al-e	se-z-na-m-e	in	o-d	iz-da-ja-te-lj-ic-e
500:	države	članice	bodo	pre-gled-al-e	se-z-na-me	in	od	iz-da-ja-te-lj-ic-e
1000:	države	članice	bodo	pregled-ale	se-z-na-me	in	od	iz-da-ja-te-lj-ic-e
2000:	države	članice	bodo	pregled-ale	se-zna-me	in	od	iz-da-ja-telj-ic-e
5000:	države	članice	bodo	pregled-ale	sezna-me	in	od	izda-ja-telj-ice
10000:	države	članice	bodo	pregled-ale	sezna-me	in	od	izda-ja-telj-ice

Slika 1: Ponazoritev delovanja postopka BPE na izbranem segmentu besedila z različnim številom združevanj od 1 do 10.000.

Na Sliki 1 so prikazani rezultati razcepljanja izbranega segmenta besedila s postopkom BPE, če uporabimo različna števila združevanj pri

učenju modela. Meje med podbesednimi enotami so na sliki prikazane z vezaji. Brez združevanj se vse besede najprej razdelijo na posamezne črke. V prvi vrstici na sliki je rezultat po enem koraku združevanja. Ker je najpogostejši par zaporednih simbolov »p« in »o«, ki se v tem segmentu besedila ne pojavi, so vse besede še vedno razdeljene na črke. Drugi najpogostejši par simbolov v učnem korpusu je »p« in »r«, kar lahko opazimo v drugi vrstici na sliki, kjer sta ta dva simbola združena v besedi »pregledale«. Z večanjem števila združevanj se začnejo združevati še nadaljnji simboli. Na primer, med vrsticama, ki predstavljata 500 in 1000 združevanj, lahko v besedi »pregledale« vidimo, da sta se združila para simbolov »pre« in »gled« ter »al« in »e«.

Avtorji v (Sennrich idr., 2016) so predstavili implementacijo algoritma in predlagali možnost skupnega učenja razcepljanja (Joint BPE), kjer uporabimo besedilo v obeh jezikih vzporednega korpusa kot učno gradivo za model razcepljanja. Tako tvorimo en model, ki vsebuje vsa združevanja, in po en slovar za vsak jezik v paru. Tako kot prej pri tem postopku nastavimo število združevanj in s tem skupno število združenih simbolov. Ni nujno, da se vsi združeni simboli pojavijo v obeh jezikih. Posledično sta slovarja za oba jezika manjša od števila združevanj.

2.2 Orodje Morfessor

Program Morfessor (Creutz in Lagus, 2002) je bil razvit v želji po razcepljanju besed v kompleksnih jezikih na podbesedne enote, ki približno ustrezajo morfemom – najmanjšim enotam besede, ki nosijo pomen. Želja je bila, imeti podatkovno voden postopek, ki deluje za več jezikov brez dodatnega slovničnega znanja in zgradi slovar jezikovnih enot, ki je manjši in bolj splošen kot slovar besed.

Predpostavka delovanja algoritma je, da so besede sestavljene iz zaporedja več segmentov, kot je to tipično v aglutinativnih jezikih. Razvita sta bila dva algoritma. Prvi temelji na principu najkrajše dolžine opisa, drugi pa na principu največje verjetnosti. Uporabili smo prvega.

Cilj algoritma je najti slovar podbesednih enot, ki daje optimalno vrednost funkcije cene (cost function), ki vsebuje dva dela: ceno izvornega besedila T in ceno slovarja V. Ceno opišemo z

$$C = C(T) + C(V) = -\sum_{m_i \in T} \log p(m_i) + \sum_{m_j \in V} k \cdot l(m_j),$$

kjer je *m* podbesedna enota, *T* je besedilo, *V* je slovar, $l(m_j)$ je dolžina podbesedne enote m_j (število črk) in *k* je število bitov, ki so potrebni za predstavitev ene črke in se lahko postavi na 5. Verjetnost posamezne podbesedne enote v besedilu $p(m_i)$ izračunamo z oceno največje verjetnosti kot razmerje med absolutno frekvenco te enote in številom vseh enot v besedilu.

V svojem delu smo uporabljali novejšo implementacijo programa – Morfessor 2.0 (Virpioja idr., 2013). Iskalni algoritem v tej implementaciji poišče nabor podbesednih enot, ki optimizirajo funkcijo cene, pri tem lahko ročno izbiramo uteži za obe komponenti funkcije cene ali pa izberemo želeno velikost slovarja. Mi smo v vseh primerih uporabe izbrali velikost slovarja.

2.3 Unigramski modeli

Zadnja metoda, ki smo jo pogledali, je razcepljanje besed na podlagi uporabe unigramskega modela (Kudo, 2018). V unigramskem modelu je verjetnost besedila oz. zaporedja podbesednih enot $\boldsymbol{x}=(x_1, x_2 \ldots x_M)$ modelirana kot produkt verjetnosti posameznih enot tega zaporedja:

$$P(\boldsymbol{x}) = \prod_{i=1}^{M} p(x_i),$$

kjer je *M* dolžina besedila in $p(x_i)$ verjetnost *i*-te enote v besedilu. Pri tem so vse podbesedne enote v slovarju in vsota verjetnost vseh enot mora biti enaka 1.

Najverjetnejše razcepljanje x^* besed vhodnega besedila *X* je tisto, za katerega velja

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in S(X)}{arg\,max}\, P(\boldsymbol{x}),$$

kjer je *S*(*X*) množica vseh možnih razcepljanj besed v besedilu *X*.

Verjetnosti posameznih unigramov podbesednih enot lahko določimo z algoritmom EM (Expectation Maximization), optimalno razcepljanje besed pa najdemo z Viterbijevim algoritmom (Kudo, 2018).

Primer implementacije opisanega postopka najdemo v orodju Sentence Piece (Kudo in Richardson, 2018), v katerem so implementirani še drugi postopki, vključno z BPE. Tudi v tem orodju lahko izhajamo iz želene velikosti končnega slovarja.

2.4 Izbira in primerjava postopkov

Pri izvedbi eksperimentov smo izbrali 4 kombinacije razcepljanja besed in uporabljenega orodja:

- Joint BPE – postopek BPE s skupnim učenjem združevanja na vzporednem korpusu in implementacijo Rica Sennricha, imenovano Subword NMT.
- MRF – postopek na principu najkrajše dolžine opisa, kjer se uteži v funkciji cene prilagodijo glede na želeno velikost slovarja in implementacijo Morfessor 2.0.
- SP-BPE – postopek BPE z ločenim učenjem združevanja na vsaki strani vzporednega korpusa in implementacijo v orodju SentencePiece.
- SP-UG – postopek na podlagi unigramskih jezikovnih modelov in implementacijo v orodju SentencePiece.

Dodatno smo vse eksperimente ponovili še z modelom, v katerem ne uporabljamo razcepljanja besed.

3 Eksperimentalni sistem

3.1 Korpusi

Eksperimenti so bili izvedeni na prosto dostopnem vzporednem korpusu ParaCrawl, različica 7.1 (Bañón idr., 2020). Korpus je bil zgrajen iz besedil s spletnih strani, ki imajo dovolj dvojezične vsebine. Strani in besedila so bili najdeni s pomočjo obstoječih statistik organizacije CommonCrawl. Korpus je bil samodejno poravnan. Za jezikovni par slovenščina-angleščina vsebuje približno 3,7 milijona poravnanih

segmentov, kar predstavlja 60,9 milijona besed na slovenski in 65,5 milijona besed na angleški strani. Do danes je bil korpus ParaCrawl že razširjen in za jezikovni par slovenščina-angleščina vsebuje približno 7,5 milijona segmentov.

Korpus smo razdelili na 3 dele: učni korpus, razvojni korpus in testni korpus. Razvojni korpus je namenjen sprotni validaciji tekom učenja strojnega prevajalnika, testni korpus pa končnemu testiranju in vrednotenju rezultatov. Za vsakega izmed teh dveh korpusov smo izbrali 2.000 naključnih segmentov besedila iz izvornega korpusa. Preostanek je bil uporabljen kot učni korpus.

3.2 Predprocesiranje

Nad vsemi deli korpusov smo izvedli standardno predprocesiranje za strojno prevajanje: čiščenje, normalizacijo ločil, tokenizacijo in *truecasing*. Učni korpus smo uporabili tudi za učenje modela za *truecasing*.

V koraku čiščenja smo iz datotek izločili odvečne presledke, prazne vrstice in vrstice, ki presegajo določeno dolžino (število besed) ali pa presegajo razmerja med številom besed v tem segmentu v enem jeziku in številom besed v istem segmentu v drugem jeziku. Takšni segmenti bi lahko ovirali algoritem učenja. Velikost predprocesiranega korpusa je zato rahlo manjša od velikost izvornega korpusa.

V koraku normalizacije ločil smo uredili stičnost ločil ter zamenjali nekatera ločil, ki jih ne najdemo v naboru ASCII znakov, z ustreznimi ločili iz nabora. S tem se zmanjšata raznolikost ločil in pomanjkanje podatkov za nekatere redke pojavnice v jeziku, kar lahko predstavlja težavo pri učenju velikega števila parametrov v modelih.

V koraku tokenizacije smo besedilo razdeli na pojavnice tako, da smo vrinili presledke med besede in stična ločila, nismo jih pa vrinili med besede in pike, ki skupaj predstavljajo krajšave, kot so: npr., dr., št. itd., saj krajšave predstavljajo eno besedo in jih želimo po tokenizaciji obdržati kot eno pojavnico.

V koraku *truecasing smo* odpravili velike začetnice na začetku povedi, razen če so te besede lastna imena ali druge besede, ki jih vedno pišemo z veliko začetnico (npr. svojilni pridevniki ali v angleščini zaimek »I«). Ponovno je namen zmanjšanje pomanjkanja podatkov.

```
or: Dr. Ivan Gašparovič, predsednik Slovaške republike (julij 2005) — skulptura Venetski konj.
nm: Dr. Ivan Gašparovič, predsednik Slovaške republike (julij 2005) - skulptura Venetski konj.
tk: Dr. Ivan Gašparovič , predsednik Slovaške republike ( julij 2005 ) - skulptura Venetski konj .
tc: dr. Ivan Gašparovič , predsednik Slovaške republike ( julij 2005 ) - skulptura Venetski konj .
```

Slika 2: Segment besedila iz slovenske učne množice v različnih korakih predprocesiranja: or – originalni zapis iz korpusa, nm – zapis po normalizaciji ločil, tk – zapis po tokenizaciji, tc – zapis po truecasing.

Na Sliki 2 je prikazan primer segmenta iz korpusa v posameznih korakih predprocesiranja, kjer vidimo spremembo pomišljaja v vezaj, spremembo stičnosti ločil in odpravo velike začetnice na začetku povedi. Končne velikosti vseh predprocesiranih korpusov so predstavljene v Tabeli 1.

Tabela 1: *Število segmentov besedila v učnem, razvojnem in testnem korpusu*

Korpus	Število segmentov
Učni	3.714.473
Razvojni	1.987
Testni	1.990
Skupaj	3.718.450

Večina korakov predprocesiranja je namenjena zmanjšanju pomanjkanja podatkov. Besede, ki imajo vedno isti pomen, se brez predprocesiranja lahko pojavijo v različnih oblikah glede na to, ali so na začetku povedi (velika začetnica) ali ne in ali ob njih stoji stično ločilo. Prevajalnik ločuje med temi oblikami zapisa. Težave se lahko pojavijo pri redkejših besedah, saj se morda v učni množici ne pojavijo v vseh oblikah dovolj pogosto.

Po prevajanju dobimo rezultat v *truecase* obliki, zato je potrebno postprocesiranje prevedenega besedila, kjer obnovimo velike začetnice na začetkih povedi in opravimo detokenizacijo tako, da poskrbimo za ustrezno stičnost ločil.

3.3 Razcepljanja besed

Pri razcepljanju besed smo uporabili orodja, ki so opisana v prejšnjem poglavju. Učni del korpusa smo uporabili za učenje modela razcepljanja, nato smo naučene modele uporabili za razcepljanje vseh delov korpusa. Tako smo dobili različice razcepljenih korpusov.

Ker smo izhajali iz želje po različnih velikostih končnih slovarjev, smo spreminjali ustrezne parametre pri uporabi orodij za učenje razcepljanja. Pri tem orodja uporabljajo te parametre na različne načine, kar pomeni, da velikosti končnih slovarjev niso natančno ustrezale nastavljenim vrednostim parametrov. Želene vrednosti, ki smo jih nastavili, so: 10.000, 15.000, 20.000, 25.000, 30.000, 40.000, 50.000, 60.000, 80.000, 100.000, 120.000 in 150.000. V Tabeli 2 so prikazane natančne velikosti slovarjev, ki jih dobimo na slovenski in angleški učni množici pri teh nastavitvah.

Tabela 2: *Velikost izdelanih slovenskih (sl) in angleških (en) slovarjev*

Nastavitev velikosti	Joint BPE (sl)	MRF (sl)	SP-BPE (sl)	SP-UG (sl)	Joint BPE (en)	MRF (en)	SP-BPE (en)	SP-UG (en)
10k	11.384	18.670	17.064	17.814	11.556	18.405	16.909	17.358
15k	16.273	28.251	25.525	26.716	15.739	27.822	25.455	26.177
20k	21.101	37.934	33.561	35.534	19.631	37.664	33.595	34.779
25k	25.883	46.879	41.297	44.175	23.299	46.298	41.395	43.051
30k	30.625	55.438	48.822	52.717	26.760	55.994	48.946	51.204
40k	39.890	73.766	63.478	69.530	33.593	73.960	63.132	66.839
50k	49.063	93.726	77.520	86.111	40.115	90.248	76.515	82.082
60k	58.155	109.989	91.015	102.242	46.404	105.558	89.312	96.924
80k	76.152	143.018	117.134	133.892	58.788	133.679	113.496	125.572
100k	93.938	174.026	142.294	164.877	71.043	159.419	136.198	153.190
120k	111.646	205.658	166.442	195.155	82.972	182.895	157.987	180.256
150k	138.006	238.620	201.334	239.515	101.013	210.425	188.859	218.140

Na Sliki 3 je prikazan primer segmenta, kjer smo besede razcepili z uporabo vseh štirih postopkov. Uporabili smo ciljno velikost slovarja 20.000, saj so pri tej velikosti razcepljanja besed bolj pogosta in lažje prikažemo več razlik v enem segmentu. Na sliki so mesta delitve besed nakazana z vezaji.

```
Joint BPE:     države  članice  bodo pregle-dale  sezna-me in od izdaja-te-ljice ...
Morfessor:     držav-e članic-e bodo pregled-a-le seznam-e in od izdajatelj-ice  ...
SP - BPE:      države  članice  bodo pregleda-le  sezna-me in od izdaja-telji-ce ...
SP - Unigram:  države  članice  bodo pregleda-le  seznam-e in od izdajatelj-ice  ...
```

Slika 3: Primer segmenta besedila iz testne množice z razcepljenimi besedami.

V Tabeli 3 imamo navedena števila razcepljanj na testni množici. V primeru, da je bila beseda razcepljena na dve podbesedni enoti, to štejemo kot eno razcepljanje. Če je bila razdeljena na več podbesednih enot, ustrezno štejemo več razcepljanj. Iz podatkov vidimo padanje števila razcepljanj z večanjem nastavitve velikosti slovarja. Po pričakovanjih vidimo več razcepljanj na slovenskem besedilu, kot na angleškem.

Slovenska testna množica vsebuje 36.994 pojavnic, angleška pa 39.874. Glede na podatke v tabeli lahko ugotovimo delež razcepljanj glede na število pojavnic. Pri velikostih slovarjev okoli 100.000 na slovenskem besedilu oz. 60.000 na angleškem pade ta delež pod 10 %. Izjema je pri razcepljanjih z orodjem Morfessor, kjer je ta delež še vedno večji.

Tabela 3: *Število razcepljanj besed na testni množici*

Nastavitev velikosti	Joint BPE (sl)	MRF (sl)	SP-BPE (sl)	SP-UG (sl)	Joint BPE (en)	MRF (en)	SP-BPE (en)	SP-UG (en)
10k	21.317	26.579	17.802	17.715	14.786	17.045	12.546	11.514
15k	16.893	23.086	13.816	13.688	11.198	13.608	9.215	8.518
20k	14.035	20.970	11.461	11.510	9.019	11.504	7.503	7.005
25k	12.045	19.450	9.893	10.028	7.634	10.088	6.430	6.115
30k	10.588	18.362	8.781	8.980	6.715	9.347	5.680	5.507
40k	8.559	16.266	7.227	7.533	5.388	7.914	4.789	4.732
50k	7.186	15.079	6.187	6.646	4.474	7.162	4.217	4.271
60k	6.185	13.852	5.413	5.965	3.830	6.145	3.834	3.983
80k	4.862	12.673	4.455	5.161	3.034	5.469	3.358	3.612
100k	3.963	10.874	3.803	4.655	2.482	4.699	3.082	3.436
120k	3.374	9.998	3.367	4.370	2.090	4.422	2.900	3.345
150k	2.713	9.459	2.910	4.084	1.697	4.003	2.742	3.249

V modelih brez razcepljanja besed smo uporabili natančne velikosti slovarjev: 60.000, 80.000, 100.000, 125.000, 150.000, 200.000, 250.000 in 300.000.

V naslednjem koraku smo zgradili slovarje za vse različice razcepljenih učnih korpusov in za nerazcepljen besedni učni korpus. Medtem ko v razcepljenih korpusih slovarji pokrijejo celotni korpus, se pri besednem korpusu pojavijo besede izven slovarja. V Tabeli 4 smo prikazali deleže besed izven slovarja (Out of Vocabulary – OOV) na testnem delu korpusa za oba jezika. Po pričakovanjih vidimo, da so deleži večji na slovenski strani in da padajo z večanjem slovarja.

Tabela 4: *Delež besed izven slovarja pri besednih slovarjih na slovenskem (sl) in angleškem (en) testnem korpusu*

Slovar	OOV (sl) [%]	OOV (en) [%]
60k	6,66	2,57
80k	5,38	2,07
100k	4,44	1,77
125k	3,74	1,50
150k	3,22	1,30
200k	2,53	1,08
250k	2,11	0,95
300k	1,82	0,85

3.4 Prevajalnik

Model prevajalnika je v vseh primerih nevronski strojni prevajalnik. Izbrali smo model *Amun* v orodju Marian NMT, ki sledi pristopu, opisanem v (Bahdanau, Cho, & Bengio, 2014), in temelji na arhitekturi RNN s celicami GRU (gated recurrent unit) z dimenzijo skritega stanja 1024 in dimenzijo vgrajenih vektorjev 512 (privzete nastavitve orodja). Naše dosedanje izkušnje na tej množici so kazale, da z uporabo arhitekture transformer in samo-pozornosti ne dosežemo bistvenih izboljšav. V nadaljevanju članka opisujemo tudi eksperimente in rezultate, ki pojasnjujejo te izkušnje. Dolžine segmentov besedila smo pri učenju omejili na 80 pojavnic (besed in ločil oz. podbesednih enot in ločil), kar pomeni, da upoštevamo 99,7 % vseh segmentov v učni množici brez razcepljanja. Pri modelih, kjer uporabljamo razcepljanje, tako upoštevamo med 96,3 % in 99,5 % vseh segmentov. Omejitve dolžine segmentov nismo povečevali, saj glede na omenjeno pokritost predvidevamo, da ne bi prišlo do znatnih sprememb rezultatov.

Učenje smo izvajali 10 epoh s preverjanjem rezultata na razvojni množici na vsakih 100 posodobitev parametrov modela. Najboljši model glede na razvojno množico smo nato uporabili pri vrednotenju rezultatov na testni množici.

Pri prevajanju smo uporabljali mini serije (mini-batch) velikosti 64, medtem ko je pri učenju uporabljena fleksibilna velikost, ki je prilagojena velikosti delovnega pomnilnika enote GPU, na kateri izvajamo učenje. Izbrali smo ciljno porabo pomnilnika 20.000 MiB. Pri tem so bile

povprečne velikosti mini serij za različne modele odvisne predvsem od velikost slovarja. Najmanjše mini serije so bile v povprečju dolge 124 segmentov pri besednem slovarju s 300.000 besedami in obema smerema prevajanja, najdaljše pa 1132 pri modelu s skupnim razcepljanjem BPE pri prevajanju iz angleščine v slovenščino in slovarjem z 20.000 podbesednimi enotami.

3.5 Ocenjevanje uspešnosti prevajanja

Uspešnost prevajanja oz. kakovost prevodov smo ocenjevali z metriko BLEU, ki omogoča avtomatsko vrednotenje in se v praksi tudi največ uporablja (Papineni idr., 2002). BLEU neodvisno oceni prevod vsakega segmenta, ocena BLEU celotnega testnega korpusa je uteženo geometrijsko povprečje ocen posameznih segmentov. Ocena segmenta je enaka modificirani natančnosti (precision) ujemanja n-gramov (zaporedij pojavnic dolžine n) med strojnim in referenčnim prevodom. Izračun natančnosti je modificiran tako, da se pri štetju ponovitev n-gramov upošteva največ toliko ponovitev, kot jih je prisotnih v referenčnem prevodu. Privzeto se uporabljajo n-grami do reda 4.

Ocena segmenta favorizira kratke prevode, zato je metriki dodana kazen prekratkih prevodov (brevity penalty).

3.6 Orodja

Za predprocesiranje (čiščenje, normalizacijo, tokenizacijo in *truecasing*) ter postprocesiranje (*detruecasing* in detokenizacijo) smo uporabljali skripte iz programskega paketa MOSES (Koehn idr., 2007). Za učenje prevajalnikov in prevajanje smo uporabljali orodje Marian NMT (Junczys-Dowmunt idr., 2018), ki smo ga poganjali na grafičnih procesnih enotah Nvidia Tesla V100. Za vrednotenje rezultatov z metriko BLEU smo uporabljali orodje SacreBLEU (Post, 2018), ki kot del vrednotenja izvaja ponovno tokenizacijo in vrednoti tokenizirana besedila. Orodja in algoritmi za razcepljanje besed na podbesedne enote so opisani v poglavju 2.

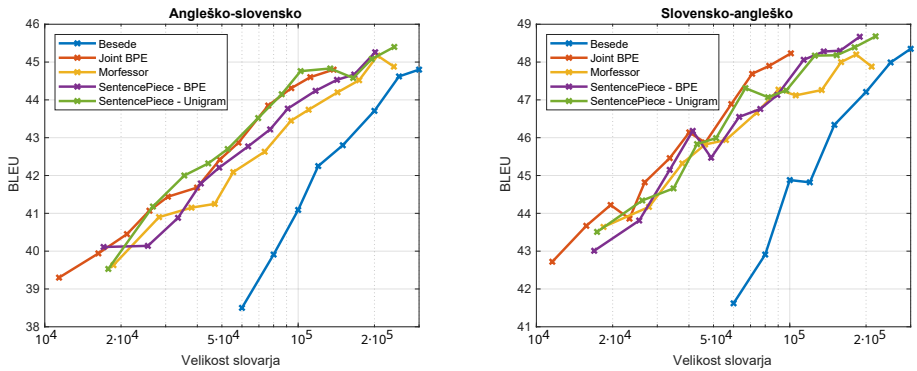
4 Rezultati

4.1 Uspešnost prevajanja

Ker je bil osnovni namen uporabe podbesednih enot zmanjšanje velikosti slovarja in s tem izvedljivost uporabe nevronskih strojnih prevajalnikov, najprej prikazujemo primer rezultatov na tipičnih velikostih slovarjev. Za besedni slovar smo izbrali velikost 60.000 besed. Velikosti med 60.000 in 65.000 so pogosto uporabljene v procesiranju naravnega jezika. V Tabeli 5 primerjamo rezultate prevajanja med besednim modelom in modelom Joint BPE z enako velikostjo slovarja. Vidimo izboljšanje uspešnosti prevajanja z uporabo podbesednih enot, kot jo tipično zasledimo v obstoječi literaturi, npr. v (Sennrich idr., 2016). Na tej točki smo še dodali rezultate vrednotenja, ki jih dobimo z metriko ChrF (β = 3) (Popović, 2015). Čeprav se ta metrika uveljavlja za vrednotenje prevajanja pri morfološko kompleksnih jezikih, smo preostale rezultate predstavili le z metriko BLEU, ki je še vedno uveljavljena in zadostuje za medsebojno primerjavo naših modelov.

Tabela 5: *Primer rezultatov uporabe besednega modela in modela z uporabo Joint BPE pri slovarju velikost 60.000*

Model	BLEU (en-sl)	BLEU (sl-en)	ChrF (en-sl)	ChrF (sl-en)
Besedni	38,50	41,62	58,43	60,68
Joint BPE	42,87	45,87	63,13	65,76

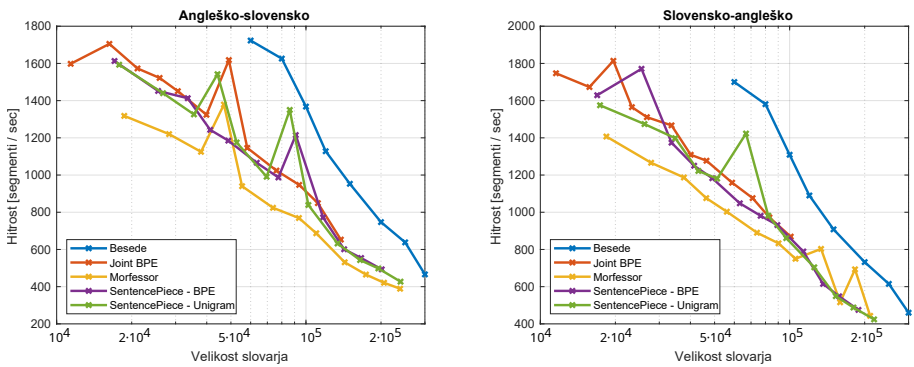


Slika 4: Rezultati uspešnosti prevajanja za vse modele.

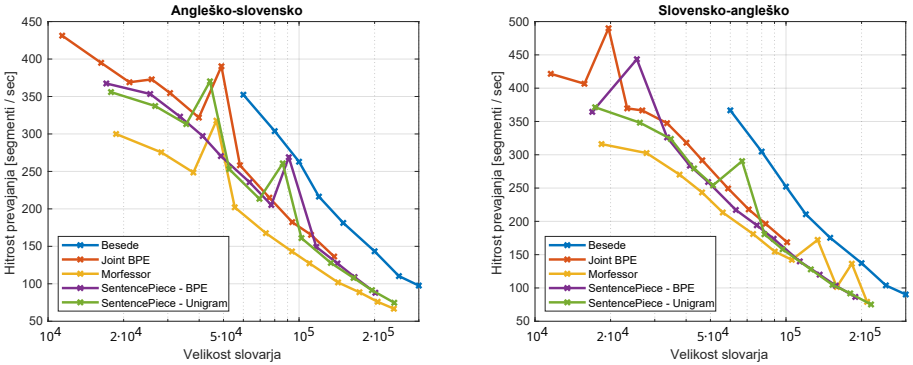
Na Sliki 4 so prikazani rezultati uspešnosti prevajanja v odvisnosti od velikosti slovarja za vse sisteme. Na slikah so velikosti slovarjev ponazorjene v logaritemskem merilu. Pri tem smo vedno upoštevali dejansko velikost slovarja, kot je razvidna iz Tabele 2.

Na splošno lahko opazimo, da uspešnost prevajanja narašča z večanjem slovarjev tako pri sistemih brez razcepljanja kot pri sistemih z razcepljanjem. Nekateri sistemi odstopajo od tega trenda, npr. prevajalnik iz slovenščine v angleščino s slovarjem 120.000 besed, ki daje slabši rezultat kot sistem slovarjem 100.000 besed. Pomemben rezultat, ki ga opazimo, je ta, da uspešnost prevajanja pri uporabi besednih modelov narašča hitreje in se pri največjih slovarjih precej približa uspešnosti prevajalnikov z razcepljanjem.

Ko primerjamo sisteme, ki uporabljajo različne načine razcepljanja besed, vidimo manjše razlike. Kljub temu lahko opazimo, da pri prevajanju iz angleščine v slovenščino večinoma daje najboljše rezultate orodje Sentence Piece z razcepljanjem na osnovi unigramov (Sentence Piece – Unigram), v nasprotni smeri pa orodje Subword NMT s skupnim učenjem (Joint BPE).



Slika 5: Hitrost učenja prevajalnika za vse modele.



Slika 6: Hitrost prevajanja pri uporabi različnih modelov.

4.2 Hitrosti učenja in prevajanja

Slika 5 prikazuje hitrost učenja modela prevajalnika, Slika 6 pa hitrost prevajanja pri njegovi uporabi. Vsi rezultati so dobljeni pri uporabi grafične procesne enote. V vseh primerih uporabljamo kot merilo za hitrost število obdelanih segmentov besedila na sekundo, saj se zaradi različnih razcepljanj število pojavnic razlikuje med sistemi. Število besed na sekundo pri učenju dobimo, če upoštevamo, da je povprečno število besednih pojavnic (besed in ločil) na segment v slovenskem besedilu 20,2, v angleškem besedilu pa 18,7.

Opazimo lahko, da se vse hitrosti zmanjšujejo z večanjem slovarja. Hitrost pri besednih modelih je večja kot hitrost pri ostalih modelih, vendar se tudi tukaj razlika pri večjih slovarjih zmanjšuje. Vidimo, da so najpočasnejši modeli tisti, ki za razcepljanje korpusa uporabljajo orodje Morfessor, vendar so počasnejši le približno 10 do 20 % glede na druge modele z razcepljanjem.

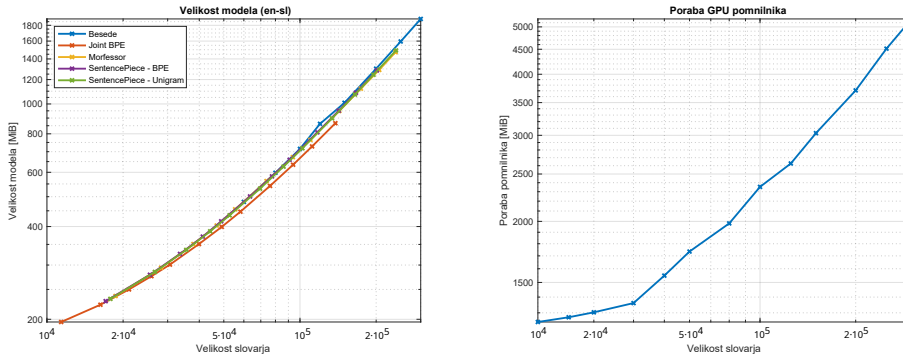
V rezultatih opazimo več točk, ki močno odstopajo od trendov. Predvidevamo, da so odstopanja nastala zaradi naključnih začetnih nastavitev nekaterih parametrov pri učenju, morebitnih odstopanj na uporabljeni strojni opremi in specifičnih lastnosti programske opreme za učenje modelov, med prilagajanjem velikosti mini serije pri različnih velikostih slovarja ali drugih vplivov.

Prikazane hitrosti prevajanja ne upoštevajo predprocesiranja in postprocesiranja besedila, ki zahtevata bistveno manj časa. Osnovno

predprocesiranje in postprocesiranje delujeta na sodobnem osebnem računalniku s hitrostjo približno 10.000 segmentov na sekundo.

Hitrost prevajanja se bistveno upočasni, če uporabljamo prevajanje na centralnih procesnih enotah računalnika CPU. Hitrost prevajanje iz angleščine v slovenščino pri besednem slovarju z 80.000 besedami je nekaj čez 300 segmentov na sekundo na enoti GPU. Če uporabimo sodobno enoto CPU lahko ta hitrost pade na približno 1 segment na sekundo. Čeprav je takšna hitrost v določenih primerih lahko še sprejemljiva za prevajanje, enote CPU niso uporabne za učenje modelov. Podobno lahko vidimo bistveno zmanjšanje hitrosti tudi pri drugih modelih.

Pri teh analizah je potrebno pripomniti, da so hitrost odvisne od optimizacije programske opreme, strojne opreme in nekaterih nastavitev orodja pri uporabi. V splošnem lahko v prihodnosti pričakujemo povečanje hitrosti prevajanja.



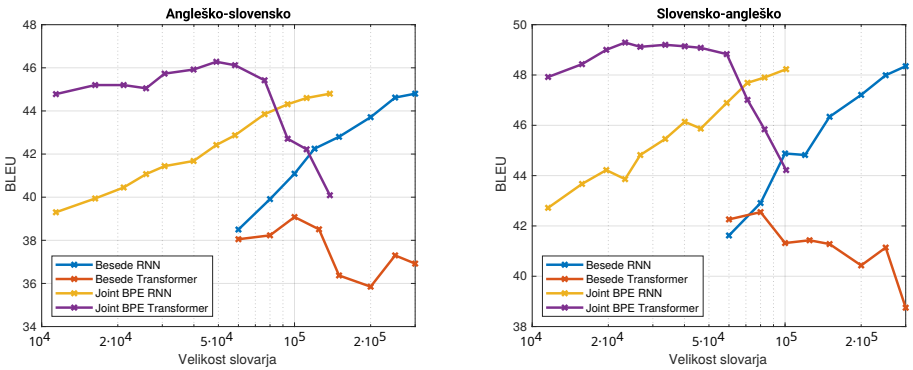
Slika 7: Velikost izdelanega modela za vse modele ter poraba pomnilnika na grafični procesni enoti med prevajanjem.

4.3 Velikosti modelov

Na Sliki 7 so na levi prikazane velikosti datotek za vse modele prevajanja in njihova poraba pomnilnika enote GPU za besedne modele. Velikosti datotek naraščajo skoraj linearno z velikostjo slovarja (na grafu sta obe osi v logaritemskem merilu). Vsakemu modelu sta pridruženi dve datoteki z obema slovarjema, ki sta bistveno manjši. Prikazane velikosti so za modele prevajanja iz angleščine v slovenščino. Modeli v nasprotni smeri imajo primerljive velikosti.

Desno na Sliki 7 je prikazana še poraba pomnilnika pri uporabi modelov pri prevajanju. Vidimo, da ima orodje osnovno porabo pomnilnika, kar se kaže v manjših spremembah porabe pri malih slovarjih. Pri večjih slovarjih poraba pomnilnika linearno narašča. Prikazana je poraba pomnilnika za besedne modele, ki je enaka v obeh smereh prevajanja. Porabe pomnilnika pri drugih modelih so primerljive glede na velikost slovarja. Pri preverjanju porabe pomnilnika je bila uporabljena mini serija velikosti 64. Pri učenju je bila poraba pomnilnika okvirno trikrat večja.

Vsi do sedaj predstavljeni eksperimenti in rezultati so ohranjali arhitekturo samega prevajalnika. V zadnjem delu eksperimentov smo preizkusili še arhitekturo transformer in spremembe vrednosti hiperparametrov modelov.



Slika 8: Primerjava rezultatov med izbranimi modeli z arhitekturo RNN in modeli z arhitekturo transformer.

4.4 Arhitekture in hiperparametri

Na Sliki 8 je prikazana primerjava rezultatov dveh arhitektur modelov pri uporabi besednih modelov in modelov z razcepljanjem BPE s skupnim slovarjem. Arhitekturi sta RNN, za katero so rezultati prikazani že na Sliki 4, in arhitektura transformer. Vidimo, da so rezultati pri uporabi besednih enot in transformer modelov bistveno slabši. Pri uporabi deljenja besed po postopku BPE pa vidimo, da obstaja za velikost slovarja interval, znotraj katerega so modeli boljši od ostalih modelov. Vidimo tudi, da pri zelo velikih slovarjih modeli z arhitekturo RNN dohitevajo modele transformer.

Tabela 6: *Uspešnost prevajanja in velikosti modelov pri uporabi večjih vrednosti hiperparametrov*

Dimenzija vgrajenih vektorjev	Dimenzija skritega stanja	BLEU (en-sl)	BLEU (sl-en)	Velikost modela [MiB]	Poraba pomnilnika GPU pri prevajanju [MiB]	Poraba pomnilnika GPU pri učenju [MiB]
512	1024	44,80	48,35	1888	5.191	18.041
1024	1024	45,40	48,36	3673	7.427	25.171
1024	2048	**46,40**	49,34	4033	8.171	26.617
2048	1024	46,33	**49,59**	7248	11.131	32.229

Rezultati v Tabeli 6 se nanašajo na uporabo različnih nastavitev hiperparametrov modelov. Vsi rezultati v tabeli so dobljeni z modeli z arhitekturo RNN ter besednimi enotami s slovarjem velikosti 300.000. Iz tabele vidimo večanje uspešnost prevajanja pri povečevanju vrednosti hiperparametrov. Krepko označena podatka v tabeli predstavljata najboljše pridobljene rezultate za vsako stran prevajanja v tej raziskavi. Vidimo še, da so ti rezultati boljši kot najboljši rezultati pri uporabi modelov z arhitekturo transformer iz Slike 8, kjer sta najboljša rezultata 46,12 in 49,29.

Sorodne raziskave kažejo velik doprinos uporabe arhitekture transformer (Vaswani idr., 2017), vendar so te raziskave uporabljale deljenje besed in manjše velikosti slovarjev. Naši rezultati prikazujejo, da sta to pomembna faktorja pri uspešnosti teh modelov. V predhodnih eksperimentih smo uporabljali večje slovarje in smo zato dobili vtis, da transformer modeli dajejo slabše rezultate.

Z večanjem hiperparametrov se poveča velikost modela, kot je prikazano v tabeli, in sorazmerno tudi poraba pomnilnika GPU. Še večje vrednosti hiperparametrov pripeljejo do tega, da učenja modelov več ni možno izvajati na izbrani strojni opremi, čeprav bi bila uporaba še vedno možna.

Zmanjšali sta se tudi hitrost učenja in prevajanja. Ti sta pri modelu z največjimi vrednostmi hiperparametrov približno dvakrat manjši kot pa pri uporabi privzetih vrednosti (Slika 6).

5 Diskusija

Za praktično uporabo modelov prevajanja je pomembna lastnosti poraba pomnilnika GPU. Če model potrebuje večji pomnilnik, kot ga je na

voljo, ga ne moremo uporabljati na enoti GPU. Delovanje na centralni procesni enoti s sistemskim delovnim pomnilnikom (RAM) je bistveno počasnejše in predstavlja izzive za tehnično manj izkušene prevajalce.

Rezultati na Sliki 7 kažejo potrebno velikost pomnilnika GPU za uporabo modelov s privzetimi nastavitvami hiperparametrov. Tudi naši največji modeli imajo porabo pomnilnika med prevajanjem pod 6 GiB. Tako velike pomnilnike dandanes najdemo v grafičnih karticah nižjega do srednjega cenovnega razreda. S tega vidika za večino prevajalcev ne bo ovire v uporabi velikih slovarjev. To se spremeni z uporabo večjih vrednosti hiperparametrov nevronske mreže. V eksperimentih na Sliki 7 smo uporabljali RNN z dimenzijo skritega stanja 1024 in dimenzijo vgrajenih vektorjev 512. Z večanjem teh dveh parametrov lahko dosežemo izboljšane rezultate, vendar se sorazmerno poveča tudi velikost modela in s tem poraba grafičnega pomnilnika, kot to kaže Tabela 6. Naši najboljši modeli potrebujejo za delovanje pomnilnik GPU velikosti, ki je tipična za enote GPU višjega cenovnega razreda. V prihodnosti lahko pričakujemo, da bodo enote GPU za uporabo takih modelov zlahka dosegljive tudi na prenosnih računalnikih.

Potrebe po pomnilniku so bistveno večje pri učenju, kjer potrebujemo enote GPU višjega cenovnega razreda. Za naše eksperimente s povečanimi vrednostmi hiperparametrov smo potrebovali profesionalno opremo, ki tipičnemu posameznemu uporabniku ni dosegljiva za nakup.

Več parametrov pomeni, da za oceno njihovih vrednosti potrebujemo večje učne množice, sicer lahko učenje vodi v prekomerno prilegaanje. Kljub temu smo s povečanimi vrednostmi hiperparametrov dobili najboljše rezultate.

Sama hitrost uporabe modela pri prevajanju je po naših izkušnjah odvisna predvsem od pasovne širine za komunikacijo med grafičnim procesorjem in grafičnim pomnilnikom. Običajne potrošniške grafične kartice imajo lahko do petkrat manjšo pasovno širino od enote GPU, ki smo jo uporabljali v tem članku. Hitrosti so še nekoliko manjše pri enotah GPU za prenosne računalnike. Pri najpočasnejših modelih to še vedno pomeni več kot 10 segmentov besedila na sekundo. Potrebni čas za strojno prevajanje tako predstavlja neznatni delež skupnega časa, ki ga prevajalec potrebuje za uporabo prevajalnika in pregled prevodov.

Hitrost je zelo pomembna pri učenju modelov prevajanja. V eksperimentih smo izvajali učenje na korpusu s 3,7 milijona segmentov. Hitrosti učenja so se precej razlikovale. Celotni postopek učenja je trajal približno 6 ur pri najhitrejših in nekaj več kot 1 dan pri najpočasnejših modelih, ki so dajali najboljše rezultate. Na potrošniških grafičnih karticah se ponovno podaljša čas za faktor približno 10.

Strojno prevajanje je možno izvajati tudi brez standardnega predprocesiranja in postprocesiranja, vendar se pri tem lahko uspešnost prevajanja zmanjša. Razlika postaja manjša z večanjem učnih korpusov. S tem zmanjšamo potrebe po dodatnih korakih in orodjih pri uporabi strojnega prevajanja na lastni opremi.

6 Sklep

V članku smo obravnavali problem pokritosti besedišča z različnimi slovarji osnovnih slovarskih enot. To je še posebej pereč problem jezikov z velikim številom pregibnih besednih vrst. Prikazali in primerjali smo nekatere najpogostejše podatkovno vodene metode za razcepljanje besed in njihovo uporabo na primeru nevronskega strojnega prevajalnika. Analiza je pokazala, da vsi štirje načini razcepljanja dajejo primerljive rezultate z le majhnimi razlikami. Pri prevajanju iz angleščine v slovenščino je najboljši Sentence Piece – Unigram, medtem ko pri prevajanju v obratni smeri Subword NMT s skupnim učenjem. Naši rezultati kažejo, da z razcepljanjem besed še vedno dosegamo boljše rezultate kot s prevajalniki brez razcepljanja, tudi v primerih, ko jim večamo slovarje. Trend kaže, da lahko z besednimi prevajalniki z nadaljnjim večanjem slovarja dohitimo modele z razcepljanjem besed. Ob trenutnem trendu razvoja in večanja pomnilniških zmogljivosti enot GPU bo takšne modele v prihodnje možno naučiti in uporabljati na potrošniških enotah GPU.

Prikazani rezultati lahko služijo raziskovalcem in uporabnikom kot orientacija pri izbiri velikosti slovarja, arhitekture modela in vrednosti hiperparametrov za strojne prevajalnike, če želijo upoštevati uspešnost prevajanja, hitrost prevajanja in velikost modela. Slednja je pomembna zaradi omejitev strojne opreme.

Natančne izbire parametrov in velikosti slovarja bodo odvisne od specifičnega primera uporabe, kjer upoštevamo: domeno prevajanja,

morebitno potrebo po adaptaciji na domeno, razpoložljiva strojna oprema prevajalca ali prevajalske agencije, razpoložljivi korpusi. Kot splošno vodilo lahko predlagamo modele s čim večjimi slovarji in vrednostmi hiperparametrov, ki jih razpoložljiva oprema še dopušča.

Za boljše razumevanje uporabnosti razcepljanja besed v strojnem prevajanju bi bilo treba izvesti še nadaljnje raziskave. V tem članku smo le v omejenem obsegu preizkušali modele transformer in spremembe vrednosti hiperparametrov. Izvedli smo le postopke podatkovno vodenega razcepljanja, ki jih v nespremenjeni obliki lahko uporabimo tudi pri drugih pregibnih jezikih. V nadaljevanju lahko proučujemo še metode razcepljanja na podlagi slovničnega znanja ali kombiniranje komplementarnih metod. Pomemben prispevek h kakovosti prevajanja pri besednih modelih imata lahko večanje učne množice in večanje hiperparametrov modela. Slednje pomeni povečanje velikosti modela in njegovo počasnejše delovanje. Nadaljnje raziskave bodo tudi vključevale podrobnejšo analizo napak, ki se pojavljajo pri različnih metodah razcepljanja.

Zahvala

Raziskovalni program št. P2-0069, v okvirju katerega je nastala ta raziskava, je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Avtorji se zahvaljujejo konzorciju HPC RIVR (www.hpc-rivr.si) za sofinanciranje raziskave z uporabo zmogljivosti sistema HPC MAISTER na Univerzi v Mariboru (www.um.si).

Zahvaljujejo se tudi avtorjem vzporednega korpusa ParaCrawl za njegovo prosto dostopnost.

Literatura

Bahdanau D., Cho K., & Bengio Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*.

Banerjee, T., & Bhattacharyya, P. (2018). Meaningless yet meaningful: Morphology grounded subword-level nmt. In *Proceedings of the second workshop on subword/character level models* (pp. 55–60). Retrieved from https://aclanthology.org/W18-1207.pdf

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., ..., & Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4555–4567). doi: 10.18653/v1/2020.acl-main.417

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263–311.

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the workshop on morphological and phonological learning of ACL-02* (pp. 21–30). doi: 10.3115/1118647.1118650

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Loenhout, G. V., Pozo, A. D., ..., & Volk, M. (2014). Machine translation for subtitling: A large-scale evaluation. In N. C. C. Chair et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation* (LREC'14). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/463_Paper.pdf

Gupta, R., Besacier, L., Dymetman, M., & Gallé, M. (2019). Character-based NMT with transformer. *arXiv preprint arXiv:1911.04997*.

Heigold, G., Varanasi, S., Neumann, G., & van Genabith, J. (2018). How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th conference of the association for machine translation in the Americas* (Vol 1, pp. 68–80). Retrieved from https://aclanthology.org/W18-1807.pdf

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., ..., & Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL2018, system demonstrations* (pp. 116–121).

Koehn, P., Hoang, H.T., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., ..., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180).

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 66–75). doi: 10.18653/v1/P18-1007

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). doi: 10.18653/v1/D18-2012

Marco, M. W. D., Huck, M., & Fraser, A. (2022). Modeling Target-Side Morphology in Neural Machine Translation: A Comparison of Strategies. In *Proceedings of the Conference on Machine Translation (WMT)* (Vol 1, pp. 56–67).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318). Retrieved from https://aclanthology.org/P02-1040.pdf

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 392–395). doi: 10.18653/v1/ W15-3049

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Retrieved from https://aclanthology.org/W18-6319.pdf

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (Vol. 1*,* pp. 1715–1725). doi: 10.18653/v1/P16-1162

Sepesy Maučec, M., & Donaj, G. (2019). Machine Translation and the Evaluation of Its Quality. In A. Sadollah & T. S. Sinha (Eds.), *Recent Trends in Computational Intelligence*. IntechOpen.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69, 343–418.

Tamchyna, A., Marco, M. W. D., & Fraser, A. (2017). Modeling target-side inflection in neural machine translation. In *Proceedings of the Conference on Machine Translation (WMT)* (Vol. 1, pp. 32–42).

Tukeyev, U., Karibayeva, A., & Zhumanov, Z. H. (2020). Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, *7*(1), 1856500. doi: 10.1080/23311916.2020.1856500

Vaswani A., Shazeer,N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems. (pp. 5998–6008).

Virpioja, S., Smit, P., Grönroos, S.-A., & Kurimo, M. (2013). *Morfessor 2.0: Python implementation and extensions for morfessor baseline*. Aalto University.

Practical Aspects of Using Subword Units in Slovene-English Machine Translation

Most modern machine translation systems are based on neural networks. This includes web-based applications, research tools and translation tools for translators. Although neural net systems can be used on common central processing units in computers and servers, reasonable speech can only be achieved by using graphics processing units (GPUs). Given today's technology for GPUs, neural machine translation systems can only use a limited vocabulary, with negative effects on translation quality. The use of subword units can alleviate the problems of vocabulary size and language coverage. However, with further technological developments the limited vocabulary and the use of subword units are losing significance. This paper presents different word splitting methods with different final vocabulary sizes. We apply these methods to the machine translation task for the Slovene-English language pair and compare them in terms of translation quality, training and translation speed, and model size. We also include a comparison with word-based translation models, and present some practical aspects on the use of subword units when machine translation is used with computer-aided translation tools.

Keywords: machine translation, vocabulary size, subword units, graphical processing units