slovenščina 2.0, letnik 12, številka 1

Slo2.0

2024_1

UNIVERZA V LJUBLJANI
FF Filozofska fakulteta

Slovenščina 2.0
Letnik/Volume 12, Številka/Issue 1, 2024

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. (ARIS)/This journal is published with the support of the Slovenian Research and Innovation Agency (ARIS).

Kazalo vsebine

RAZPRAVE

LOGIČNO SKLEPANJE V NARAVNEM JEZIKU ZA SLOVENŠČINO 1
Tim KMECL, Marko ROBNIK-ŠIKONJA

KORPUSNE OZNAKE ZA OPIS KONTEKSTA GOVORNIH DOGODKOV V SLOVENSKIH GOVORNIH KORPUSIH 54
Andreja BIZJAK

POMEMBNOST REALISTIČNE EVALVACIJE: PRIMER POPRAVKOV SKLONA IN ŠTEVILA V SLOVENŠČINI Z VELIKIM JEZIKOVNIM MODELOM 106
Timotej PETRIČ, Špela ARHAR HOLDT, Marko ROBNIK-ŠIKONJA

POROČILA

LREC-COLING 2024: ZDRUŽENA MEDNARODNA KONFERENCA RAČUNALNIŠKEGA JEZIKOSLOVJA TER JEZIKOVNIH VIROV IN EVALVACIJE 95
Mojca BRGLEZ, Matej KLEMEN, Luka TERČON, Špela VINTAR

MUCH MORE THAN AN INTRODUCTION TO THE LANGUAGE-SPECIFIC LEXICAL SEMANTICS 131
Igor MARKO GLIGORIĆ

Logično sklepanje v naravnem jeziku za slovenščino

Tim KMECL
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Na področju strojnega razumevanja naravnega jezika so v zadnjih letih najuspešnejši veliki jezikovni modeli. Pomemben problem s tega področja je logično sklepanje v naravnem jeziku, za reševanje katerega morajo modeli vsebovati dokaj široko splošno znanje, strojno generiranje razlag sklepov pa nam omogoča dodaten vpogled v njihovo delovanje.
Preizkusili smo različne pristope za logično sklepanje v naravnem jeziku za slovenščino. Uporabili smo dva slovenska velika jezikovna modela, SloBERTa in SloT5, in mnogo večji angleški jezikovni model GPT-3.5-turbo. Za učenje modelov smo uporabili slovensko podatkovno množico SI-NLI, strojno pa smo prevedli še 50.000 primerov iz angleške množice ESNLI.
Model SloBERTa, prilagojen na SI-NLI, doseže na testni množici SI-NLI klasifikacijsko točnost 73,2 %. Z vnaprejšnjim učenjem na prevodih ESNLI smo točnost izboljšali na 75,3 %. Ugotovili smo, da modeli delajo drugačne vrste napak kot ljudje in da slabo posplošujejo med različnimi domenami primerov. SloT5 smo na množici ESNLI prilagodili za generiranje razlag pri logičnem sklepanju. Ustreznih je manj kot tretjina razlag, pri čemer se model dobro nauči pogostih stavčnih oblik v razlagah, večinoma pa so pomensko nesmiselne. Predvidevamo, da so slovenski veliki jezikovni modeli z nekaj sto milijoni parametrov zmožni iskanja in uporabe jezikovnih vzorcev, njihovo poznavanje jezika pa ni povezano s poznavanjem resničnosti.

Kmecl, T. et al.: Logično sklepanje v naravnem jeziku za slovenščino. Slovenščina 2.0, 12(1): 1–53.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2024.1.1-53

Za uvrščanje primerov in generiranje razlag smo uporabili tudi večji model GPT-3.5-turbo. Pri učenju brez dodatnih primerov doseže na testni množici SI-NLI točnost 56,5 %, pri pravilno uvrščenih primerih pa je ustreznih 81 % razlag. V primerjavi z manjšimi slovenskimi modeli kaže ta model dokaj dobro razumevanje resničnosti, pri čemer pa ga omejuje slabše poznavanje slovenščine.

Ključne besede: logično sklepanje v naravnem jeziku, veliki jezikovni modeli, arhitektura transformer, SloBERTa, SloT5, GPT-3.5-turbo, ChatGPT, razlage, slovenščina, prilagajanje modelov

1 Uvod

Procesiranje naravnega jezika je vse od začetkov računalništva v sredini prejšnjega stoletja pomembno raziskovalno in aplikativno področje. Zajema številne probleme, ki so povezani z naravnim jezikom. Problemi so lahko klasifikacijske ali generativne narave. Pod klasifikacijske naloge spada na primer označevanje besednih vrst, razpoznava čustev v besedilu in zaznavanje neželene pošte. Pri generativnih problemih je izhod besedilo, kot na primer pri povzemanju vhodnega besedila in odgovarjanju na vprašanja. V zadnjih letih so na tem področju najuspešnejši veliki jezikovni modeli (angl. *large language models*, kratica *LLM*).

Veliki jezikovni modeli so posebna vrsta globokih nevronskih mrež z od nekaj deset milijonov do več sto milijard parametri, naučenih za modeliranje jezika, konkretno za napovedovanje naslednje ali manjkajoče besede v nizu. Učijo se na ogromnih korpusih besedil, najpogosteje s svetovnega spleta. Po splošnem učenju, ki jim zagotovi poznavanje jezika, jih lahko prilagodimo za različne naloge. V središče pozornosti širše javnosti so veliki jezikovni modeli vstopili novembra 2022 s predstavitvijo spletnega klepetalnega robota ChatGPT (OpenAI, 2022), ki temelji na velikem jezikovnem modelu GPT-3.5. Ker zna ChatGPT odgovarjati na vprašanja, slediti navodilom in ima znanje s praktično vseh področij človeškega delovanja, daje marsikomu vtis inteligence, podobne človeški.

Ljudje pri svojem razmišljanju in reševanju (jezikovnih) problemov med drugim uporabljamo zdravorazumsko sklepanje in poznavanje

sveta, pridobljeno skozi izkušnje in neposredno interakcijo s svetom. Po drugi strani pa se veliki jezikovni modeli učijo le na korpusih besedil in neposrednega dostopa do resničnega sveta nimajo. To poraja vprašanje, ali se modeli naučijo zgolj različnih jezikovnih vzorcev in hevristik, ki zadoščajo za reševanje različnih nalog, ali pa učenje le iz besedila omogoča globlje razumevanje resničnosti. Problem, ki nam lahko da vpogled v to, je logično sklepanje v naravnem jeziku (angl. *natural language inference*, kratica *NLI*).

Pri tipični formulaciji problema logičnega sklepanja v naravnem jeziku sta vhod dve povedi. Prvo imenujemo premisa, drugo hipoteza. Cilj je ugotoviti, v kakšnem logičnem razmerju sta podani povedi. Če je hipoteza logična posledica premise, oziroma če lahko ob predpostavki, da je premisa resnična, utemeljeno sklepamo na resničnost hipoteze, imenujemo to razmerje *implikacija* (angl. *entailment*). Če so informacije v hipotezi v nasprotju s tistimi v premisi, oziroma lahko iz resničnosti premise utemeljeno sklepamo na neresničost hipoteze, to imenujemo *kontradikcija* (angl. *contradiction*). Če pa iz premise ne moremo sklepati niti na resničnost niti na neresničnost hipoteze, torej če informacije v premisi hipoteze niti ne potrjujejo niti ne zavračajo, je tak primer *nevtralen* (angl. *neutral*). Tako formuliran problem je v osnovi klasifikacijski, lahko pa ga razširimo tako, da zahtevamo poleg klasifikacije (uvrščanja) še razlago zanjo. Če zmore jezikovni model, ki problem rešuje, odgovor utemeljiti, omogoča to dodaten vpogled v njegov pristop k reševanju problema.

Na primer, če je dana premisa *Mož in žena sedita v dnevni sobi* in hipoteza *V dnevni sobi sedita zakonca*, gre za implikacijo, ker sta mož in žena zakonca. Če bi bila hipoteza namesto tega *Dnevna soba je prazna*, bi bil to primer kontradikcije, ker sta glede na premiso v sobi dva človeka in zato ne more biti prazna. Če pa bi bila hipoteza *Ura je šest popoldne*, bi šlo za nevtralen primer, ker v premisi podatek o času ni podan, ljudje pa v dnevni sobi lahko sedijo kadarkoli.

V angleščini obstajajo številne podatkovne množice s primeri logičnega sklepanja v naravnem jeziku (Bowman idr., 2015; Camburu idr., 2018; McCoy idr., 2019). Prav tako so se z uporabo velikih jezikovnih modelov za reševanje tega problema ukvarjali mnogi raziskovalci (H. Liu idr., 2023; McCoy idr., 2019; Poth idr., 2021; Wang idr., 2021;

Zhong idr., 2023), tudi s strojnim generiranjem razlag zanje (Camburu idr., 2018; Kumar in Talukdar, 2020).

Logično sklepanje v naravnem jeziku za slovenščino je precej manj raziskano področje. Obstaja le ena izvirno slovenska podatkovna množica primerov logičnega sklepanja, imenovana SI-NLI (Klemen idr., 2022). Na spletni platformi SloBench (CJVT UL, 2023) sta objavljena dva rezultata vrednotenja na testni množici SI-NLI, oba pristopa uporabljata model SloBERTa (Ulčar in Robnik Šikonja, 2021). S strojnim generiranjem razlag pri logičnem sklepanju se v slovenščini ni ukvarjal še nihče.

Naš cilj je preizkusiti več pristopov za reševanje tega problema v slovenščini, pri tem pa testirati več velikih jezikovnih modelov, tako na slovenski podatkovni množici kot na strojnem prevodu angleške. Zanima nas, kako uspešni so različni modeli pri reševanju klasifikacijskega problema, kako dobro zmorejo generirati razlage in ali so sposobni posploševanja med različnimi domenami primerov istega problema. Na osnovi rezultatov skušamo poleg vrednotenja pristopov ugotoviti tudi, ali veliki jezikovni modeli, uporabni za slovenščino, premorejo dejansko razumevanje sveta ali pa se naučijo le jezikovnih vzorcev. Poskuse razdelimo v štiri sklope.

V prvem sklopu uporabimo slovensko množico SI-NLI za učenje klasifikacijskega modela SloBERTa. Preizkusiti želimo, kako zmogljiv je ta model, ali je število učnih primerov zadostno, kako na učenje vplivajo primeri, ki jih narobe uvrstijo ljudje, in ali so napake, ki jih naredi model, podobne človeškim.

V drugem sklopu poskusov se ukvarjamo z uporabo strojnega prevajanja iz angleščine, konkretno prevoda množice ESNLI (Camburu idr., 2018), in prenosom znanja. Model SloBERTa učimo na prevodih te množice in primerjamo rezultate s tistimi iz prejšnjega sklopa. Zanima nas, kako dobro jezikovni modeli posplošujejo med različnimi učnimi množicami, ali je strojno prevajanje lahko primeren nadomestek slovenskim podatkovnim množicam in ali lahko strojne prevode uporabimo za izboljšanje napovedovanja na SI-NLI.

Cilj tretjega sklopa je prilagajanje generativnega slovenskega modela SloT5 (Ulčar in Robnik-Šikonja, 2023) za generiranje razlag, ki jih kvalitativno ocenimo in s tem skušamo razložiti, kako jezikovni modeli rešujejo problem logičnega sklepanja.

Zadnji sklop poskusov temelji na uporabi angleškega modela GPT-3.5-turbo, ki poganja tudi ChatGPT. Uporabimo ga tako za klasifikacijo kot za generiranje razlag. Ugotoviti želimo, ali je ta model uporaben za slovenščino in ali mu nekaj redov velikosti več parametrov in učnih podatkov omogoča boljše razumevanje in logično sklepanje, tudi če ni bil naučen specifično za to nalogo.

Članek je razdeljen na sedem razdelkov. V razdelku 2 najprej predstavimo delovanje arhitekture transformer, ki je osnova za velike jezikovne modele. Zatem predstavimo tri jezikovne modele, ki smo jih uporabili za reševanje problema – slovenska modela SloBERTa in SloT5 ter angleški GPT-3.5-turbo. V razdelku 3 predstavimo in analiziramo slovensko podatkovno množico SI-NLI in angleško množico ESNLI ter opišemo njeno strojno prevajanje v slovenščino. V razdelku 4 opišemo postopek učenja modelov in izbiro parametrov zanje po štirih, prej opisanih sklopih poskusov. Predstavimo še evalvacijske metrike in način vrednotenja rezultatov. Rezultati vrednotenja so po sklopih podani v razdelku 5, ki vsebuje kvantitativne in kvalitativne ocene pristopov ter interpretacijo rezultatov. V razdelku 6 združimo najpomembnejše rezultate in jih postavimo v širši kontekst. Članek zaključimo v 7. razdelku s povzetkom narejenega in predlogi za nadaljnje delo ter izboljšave.

2 Veliki jezikovni modeli in predhodno delo

V tem razdelku predstavimo velike jezikovne modele, ki smo jih uporabili za logično sklepanje v naravnem jeziku. Začnemo s predstavitvijo modelov SloBERTa in SloT5, slovenskih različic modelov BERT in T5. Nato predstavimo največji model, ki smo ga uporabili, angleški GPT-3.5-turbo. Na koncu predstavimo še predhodno delo s področja uporabe velikih jezikovnih modelov za logično sklepanje v angleščini.

2.1 Modela BERT in SloBERTa

BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin idr., 2019) je jezikovni model, ki temelji na uporabi kodirnika arhitekture nevronskih mrež transformer (Vaswani idr., 2017). Pri osnovni različici je kodirnik sestavljen iz 12 plasti in uporablja vektorje dimenzije 768. Skupno ima model 110 milijonov parametrov.

Za njegovo učenje je uporabljeno t. i. samonadzorovano učenje. Naučen je za dve nalogi, napovedovanje zaporednosti dveh stavkov in napovedovanje maskirane besede. Pri prvi je vhod v model sestavljen iz dveh stavkov, model pa mora ugotoviti, ali sta stavka zaporedna ali ne. Pri napovedovanju maskirane besede se 15 % členov na vhodu zamenja s posebnim členom [mask], model pa napoveduje člen, ki je bil tam pred zamenjavo. Z uporabo le druge naloge in večje količine podatkov kot pri originalnem modelu BERT je bila naučena izboljšana različica, imenovana RoBERTa (Y. Liu idr., 2019).

Učenje tako naučenih velikih jezikovnih modelov se kasneje nadaljuje na drugih nalogah, pri čemer je potrebna znatno manjša količina učnih primerov kot sicer, saj model že vsebuje neko razumevanje jezika, ki ga je treba zgolj prilagoditi konkretni nalogi. To imenujemo prilagoditev modela (angl. *fine-tuning*), za model, ki ga imamo za osnovo, pa rečemo, da je vnaprej naučen (angl. *pre-trained*).

Na enaki metodi učenja in enaki arhitekturi kot RoBERTa temelji slovenski model SloBERTa (Ulčar in Robnik-Šikonja, 2021). Ta model je bil naučen na korpusih slovenskih besedil, in sicer Gigafida 2.0 (Krek idr., 2020) (besedila iz knjig, revij, časopisov in interneta), Janes (Fišer idr., 2016) (besedila z družabnih omrežij), KAS (Erjavec idr., 2021) (akademska besedila), siParl (Pančur in Erjavec, 2020) (parlamentarni transkripti) in slWaC (Ljubešić in Erjavec, 2011) (slovenske spletne strani). Učna množica skupno vsebuje približno 3,4 milijarde besed oziroma 4,7 milijard členov. Učenje modela je trajalo 98 epoh (angl. *epochs*). Ta model smo uporabili kot vnaprej naučeni model in ga prilagodili za reševanje klasifikacijskega problema logičnega sklepanja v naravnem jeziku.

2.2 Modela T5 in SloT5

T5 (*Text-to-Text Transfer Transformer*) (Raffel idr., 2020) je družina modelov več velikosti, ki po zgradbi sledijo arhitekturi transformer in vsebujejo tako kodirnik kot dekodirnik. Ideja pristopa T5 je, da se vse naloge, tudi klasifikacijske, obravnava kot transformacijo enega besedila v drugo, npr. polnega besedila v povzetek ali besedila iz enega

jezika v drugega. Model so vnaprej učili na množici besedil s spleta, pri čemer je vhod modela besedilo, v katerem so nekatere besede zamenjane s posebno oznako, izhod pa mora biti besedilo, ki vsebuje manjkajoče besede v pravem zaporedju. Tako vnaprej naučene modele se nato prilagaja za različne naloge, kot so povzemanje besedil, prevajanje, razpoznava čustev in odgovarjanje na vprašanja.

Slovenska različica modela T5 se imenuje SloT5 (Ulčar in Robnik-Šikonja, 2023). Pravzaprav gre za dva modela, ki sta po zgradbi enaka dvema modeloma iz družine originalnih modelov T5. Manjši T5-sl-small ima 8 plasti v kodirniku in 8 v dekodirniku, skupaj 60 milijonov parametrov, večji model T5-sl-large pa ima 24 plasti v kodirniku in 24 v dekodirniku, skupaj 750 milijonov parametrov. Za učenje so bili uporabljeni isti korpusi kot za učenje modela SloBERTa. Manjši model so učili 5 epoh, večjega pa eno, pri čemer je učenje manjšega na 4 grafičnih karticah A100 s 40 GB pomnilnika trajalo 12 dni, večjega pa tri tedne.

Oba slovenska modela T5 smo uporabili za generiranje razlag za primere logičnega sklepanja.

2.3 Model GPT-3.5-turbo

Model GPT-3 (*Generative Pre-trained Transformer 3*) (Brown idr., 2020) temelji na uporabi dekodirnika arhitekture transformer, pri čemer je vhod v model že na začetku vstavljen v dekodirnik. Model je naučen na korpusu besedil s spleta CommonCrawl, ki vsebuje približno 400 milijard členov, z nalogo napovedovanja naslednje besede v besedilu. Ta model je bistveno večji od modelov, predstavljenih v prejšnjih razdelkih, saj vsebuje kar 175 milijard parametrov. Model je v lasti podjetja OpenAI in ni prosto dostopen, uporablja pa se ga lahko prek programskega vmesnika API, ki ga podjetje ponuja.

GPT-3 se lahko brez dodatnega učenja uporablja za številne naloge z dvema tehnikama. Pri prvi, imenovani učenje brez dodatnih primerov (angl. *zero-shot learning*), se modelu kot vhod posreduje navodilo oziroma opis naloge v naravnem jeziku in morebitni kontekst. To je lahko na primer besedilo nekega članka in navodilo, naj model članek povzame. Zaradi velikosti modela in velike količine

učnih podatkov, kar mu omogoča dobro posploševanje, zna model mnogim navodilom pravilno slediti in dobimo pravilen izhod. Za razliko od prilagoditve, ki se uporablja pri modelih arhitekture BERT in T5, pri učenju brez dodatnih primerov ne potrebujemo učnih primerov za želeno nalogo. Druga tehnika uporabe GPT modelov brez dodatnega učenja je uporaba nekaj dodatnih primerov (angl. *few-shot learning*), ki je podobna prejšnji, le da tu poleg navodil in konteksta na vhod dodamo še nekaj že rešenih primerov, ki lahko modelu olajšajo razumevanje naloge.

InstructGPT (Ouyang idr., 2022) je družina modelov, katerih največji je po velikosti enak GPT-3. Osnova je vnaprej naučen model GPT-3, s prilagoditvijo pa so izboljšali sposobnost modela za odgovarjanje na vprašanja in s tem tudi rezultate, ki jih lahko dosežemo z učenjem brez ali z nekaj dodatnimi primeri. Izboljšanje je bilo doseženo z uporabo spodbujevanega učenja s človeško povratno informacijo (angl. *reinforcement learning from human feedback*, RLHF). RLHF poteka v več korakih. Na začetku pripravimo podatkovno množico različnih navodil, za katera bi želeli, da jih model zna upoštevati pri uporabi, na primer ukaz, da napiše povzetek nekega besedila. Nato človeški demonstratorji spišejo odgovore za vsako od navodil, ki se uporabijo za nadzorovano učenje vnaprej naučenega modela (prilagoditev). V naslednjem koraku za vsako od navodil s tem modelom generiramo odgovor, človeški ocenjevalci odgovore ocenijo, ta informacija pa se nato uporabi za dodatno učenje z algoritmom spodbujevanega učenja, ki ne zahteva zvezne funkcije izgube.

GPT-3.5-turbo je eden najzmogljivejših modelov, ki jih ponuja OpenAI (OpenAI, 2023b). Je variacija zmogljivega modela InstructGPT, točni podatki o zgradbi in učenju pa niso objavljeni. To je model, ki poganja znani spletni vmesnik ChatGPT (OpenAI, 2022).

Čeprav model GPT-3 za razliko od predstavljenih v prejšnjih razdelkih nima slovenske različice, je med učnimi podatki tudi nekaj slovenščine. Modela ne prilagajamo za specifično nalogo, saj do njega nimamo neposrednega dostopa. Model GPT-3 nam služi za preizkus, ali več redov velikosti večje število parametrov in učnih podatkov pri vnaprejšnjem učenju lahko odtehta te pomanjkljivosti.

2.4 Uporaba jezikovnih modelov za logično sklepanje

Področje logičnega sklepanja v naravnem jeziku je v angleščini dobro raziskano. Zadnja leta najboljše rezultate dosegajo veliki jezikovni modeli, predstavljeni v prejšnjih razdelkih.

Poth idr. (2021) so vnaprej naučena modela BERT in RoBERTa prilagajali za reševanje različnih nalog, med drugim so testirali tudi več podatkovnih množic za logično sklepanje v naravnem jeziku. Z modelom RoBERTa, s katerim so dosegli boljše rezultate, so dosegli klasifikacijsko točnost 41,5 % na množici ANLI, 87,5 % na množici MNLI in 91,1 % na SNLI. Točnost na SNLI je bila do takrat najvišja dosežena.

Njihov rezultat so z lastnim velikim jezikovnim modelom nadgradili Wang idr. (2021). Uporabljeni model je po zgradbi podoben modelu RoBERTa, uporabili pa so drugačen način učenja. Z namenom povečanja učne množice so primere iz podatkovnih množic za druga področja razumevanja naravnega jezika, nepovezana z logičnim sklepanjem, reformulirali kot probleme logičnega sklepanja, te pa so nato uporabili za vnaprejšnje učenje modela. S to prilagoditvijo so na množici SNLI dosegli najboljši objavljen rezultat, 93,1 %.

Zhong idr. (2023) so primerjali zmogljivost modelov RoBERTa in GPT-3.5-turbo z učenjem brez ali z nekaj dodatnimi primeri. Z učenjem brez dodatnih primerov so presegli rezultate modela RoBERTa, njihova točnost je znašala 89,3 %, uporaba enega ali pet dodatnih primerov pa rezultata ni izboljšala. Liu idr. (2023) so primerjali modele RoBERTa, GPT-3.5-turbo in GPT-4 na podatkovnih množicah LogiQA in ReClor. Tudi oni so ugotovili, da točnost GPT-3.5-turbo za nekaj odstotkov preseže model RoBERTa, še večjo pa doseže GPT-4.

Z generiranjem razlag za logične sklepe so se ukvarjali Camburu idr. (2018). Pokazali so, da so pri tej nalogi veliki jezikovni modeli arhitekture transformer boljši od prejšnjih pristopov. Z učenjem lastnega modela na množici ESNLI so na tej množici dosegli klasifikacijsko točnost 81,7 %, pri čemer je pri pravilno klasificiranih primerih ustreznih 64 % razlag. Njihov pristop sta izboljšala Kumar in Talukdar (2020) z uporabo vnaprej naučenega modela GPT-2, predhodnika GPT-3.5-turbo. Preizkusila sta različne pristope, med drugim tudi z učenjem treh ločenih modelov za generiranje razlag za primere posameznih

razredov in dodatnega modela, ki na koncu izbere najboljšo. S tem sta izboljšala tako klasifikacijsko točnost kot tudi delež ustreznih razlag.

Problem NLI za slovenščino je manj raziskan. Na spletni platformi SloBench (CJVT UL, 2023) sta objavljena dva rezultata vrednotenja na testni množici SI-NLI, oba pristopa uporabljata prilagoditev modela SloBERTa. S strojnim generiranjem razlag pri logičnem sklepanju se v slovenščini ni ukvarjal še nihče.

3 Uporabljene podatkovne množice

V tem razdelku predstavimo podatkovni množici s področja logičnega sklepanja v naravnem jeziku, ki smo ju uporabili za učenje jezikovnih modelov. Najprej, v razdelku 3.1, opišemo in analiziramo sestavo slovenske množice SI-NLI, v razdelku 3.2 pa predstavimo še angleško množico ESNLI, ki vsebuje tudi razlage. Opišemo tudi postopek prevajanja te množice v slovenščino.

3.1 Slovenska množica SI-NLI

SI-NLI (Klemen idr., 2022) je podatkovna množica s skupno 5937 pari povedi v slovenščini. Vsak par vsebuje premiso in hipotezo ter je označen z oznako *entailment* (implikacija), *neutral* (nevtralno) ali *contradiction* (kontradikcija), ki označuje razmerje med povedmi. Pri konstrukciji množice so sodelovali človeški označevalci (angl. *annotators*). Množica je bila ustvarjena na osnovi povedi, ki se pojavijo v korpusu ccKres (Logar idr., 2013), ki vsebuje različne tipe besedil, kot so članki v časopisih in revijah, literarna in neliterarna besedila ter besedila z interneta. Označevalci so spreminjali hipoteze tako, da so ustrezale vsaki od možnih treh kategorij. Po en primer za vsako oznako je prikazan v Tabeli 1.

Delitev SI-NLI na učno, validacijsko in testno množico je predstavljena v Tabeli 2. Avtorji zagotavljajo, da so težji in lažji primeri enakomerno razporejeni med vsemi tremi množicami. Vse tri množice vsebujejo premiso in hipotezo za vsakega od primerov. Učna in testna množica vsebujeta poleg tega še oznako oz. klasifikacijo primera in po tri stolpce s klasifikacijami posameznih označevalcev, njihovimi identifikatorji ter morebitnimi komentarji. Testna množica tega ne vsebuje.

Tabela 1: *Trije primeri iz podatkovne množice SI-NLI, po eden za vsako oznako*

Premisa	Med večletnimi širokolistnimi pleveli prevladujejo slak, osat, gabez in ščavje, najpogostejše večletne trave v koruzi pa so pirnica in divji sirek.
Hipoteza	Slak, gabez in osat spadajo med širokolistni plevel, enako tudi ščavje, na drugi strani pa med večletne trave v koruzi prištevamo pirnico in divji sirek.
Oznaka	*implikacija*
Premisa	"Res je," je zavzdihnila in se zravnala na sedežu.
Hipoteza	Z globokim poraženim izdihom je morala priznati, da dejstev ne gre zanikati.
Oznaka	*nevtralno*
Premisa	Večina delničarjev ga je potrdila za predsednika, Janez Pestotnik pa je postal novi predsednik nadzornega sveta Banke Karantanija.
Hipoteza	Ker se delničarji s svojimi glasovi niso uspeli uskladiti, je Banka Karantanija še vedno brez predsednika nadzornega sveta.
Oznaka	*kontradikcija*

Tabela 2: *Število primerov v učni, validacisjki in testni množici SI-NLI, ESNLI in ESNLIsi*

Podatkovna množica	*Učna*	*Validacijska*	*Testna*
SI-NLI	4392	547	998
ESNLI	550.000	10.000	10.000
ESNLIsi	49.922	3000	3000

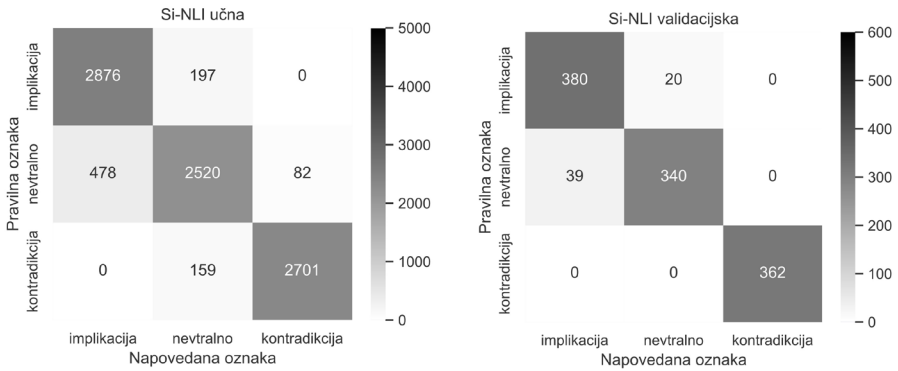
Analiza

Analizirali smo sestavo učne in validacijske množice ter morebitne razlike med oznakami primerov in klasifikacijami posameznih označevalcev, torej primere, v katerih so označevalci glede na dodeljeno oznako storili napako.

V učni množici s 4392 primeri je 34,6 % vseh implikacij, 32,5 % nevtralnih in 33,0 % kontradikcij. Množica je torej skoraj uravnotežena. Podobno je pri validacijski množici, kjer 35,3 % primerov predstavlja implikacije, 31,6 % je nevtralnih in 33,1 % opisuje kontradikcije.

Vsakega od primerov v učni množici sta označila dva ali trije označevalci. Od oznake primera se vedno razlikuje največ ena oznaka označevalcev. 89,8 % oznak označevalcev se ujema z oznako primera (torej jih lahko privzamemo za pravilne oziroma ta odstotek predstavlja točnost človeških označevalcev), 79,1 % primerov pa je takih, da se nobena od oznak označevalcev ne razlikuje od oznake primera

(soglasni primeri). Na levi strani Slike 1 je predstavljena matrika zamenjav (angl. *confusion matrix*). Vidimo lahko, da ni napak, kjer bi označevalec zamenjal razreda implikacija in kontradikcija. Največ napak, skoraj tri četrtine, je posledica zamenjave med razredi implikacija in nevtralno, najpogostejša napaka pa je označitev nevtralnega primera kot primera implikacije.



Slika 1: Matriki zamenjav za oznake označevalcev na učni in validacijski množici SI-NLI.

V validacijski množici je soglasnih primerov 89,2 %, pravilnih oznak označevalcev pa 94,8 %. Struktura napak, prikazana na desni strani Slike 1, je podobna.

Edine napake so posledice zamenjav med razredoma implikacija in nevtralno, ki so tudi v učni množici najpogostejše, največ napak pa je tudi tu predstavljala označitev nevtralnega primera za implikacijo. V nadaljevanju skušamo ugotoviti, ali veliki jezikovni modeli delajo podobne tipe napak kot človeški označevalci in kako vpliva odstranitev primerov, kjer označevalci niso soglasni.

Metrika, ki je za oceno strinjanja med različnimi ocenjevalci bolj robustna od odstotka soglasnih primerov, je Cohenova kappa, ki ima nabor vrednosti med –1 in 1, kjer 0 pomeni količino ujemanj, ki jih lahko pripišemo naključju, 1 pa popolno ujemanje (McHugh, 2012). Podrobno analizo strinjanja med označevalci so opravili Klemen idr. (2024), ki so za vse pare izračunali Cohenovo kappo. Njihovo povprečje znaša 0,74, kar kaže na visoko konsistentnost med označevalci.

Ker oznake primerov testne množice niso javno objavljene, smo za naše potrebe validacijsko množico uporabili kot testno, učno množico pa smo razdelili na učno in validacijsko. V učni množici SI--NLI se večina premis ponovi trikrat, nobena premisa pa se ne pojavi hkrati v učni in validacijski množici, kar smo zagotovili tudi pri naši delitvi. Naključno smo izbrali 200 premis iz učne množice ter pripadajoče pare (590 primerov) shranili kot validacijsko množico, preostalih 3802 pa kot učno. V nadaljevanju se učna, validacijska in testna množica SI-NLI nanašajo na našo delitev, razen kjer je posebej navedeno drugače. Dodatno smo shranili tudi podmnožico naše učne množice, pri kateri so bili označevalci soglasni in vsebuje le 3015 primerov.

3.2 Množica z razlagami ESNLI

ESNLI (Camburu idr., 2018) je angleška podatkovna množica s 570 tisoč primeri (njena delitev je prikazana v Tabeli 2), od katerih vsak vsebuje premiso, hipotezo, eno od treh možnih oznak in razlago. Osnova te množice je angleška množica SNLI (Bowman idr., 2015), v kateri so premise opisi slik, človeški označevalci pa so jim dopisali po eno hipotezo za vsako od treh kategorij. ESNLI dodatno vsebuje še razlage, zakaj dani par premise in hipoteze pripada dodeljeni kategoriji. Razlage so pisali ljudje, želeli pa so, da so samozadostne, torej da za njihovo razumevanje ni potrebno predhodno prebrati premise in hipoteze. Primer take razlage je *Kdorkoli lahko plete, ne le ženske*, primer neustrezne pa *Ne moremo sklepati, da so to ženske* (Camburu idr., 2018). Pisci razlag so morali tudi označiti, katere besede v premisi in hipotezi so ključne za izbor dane oznake. Primeri v učni množici vsebujejo vsak po eno razlago, tisti v validacijski in testni pa po tri. Po en primer za vsako oznako (samo z eno razlago) iz množice je podan v Tabeli 3.

Tabela 3: *Trije primeri iz podatkovne množice ESNLI*

Premisa	Mlad fant poljublja starca na čelo.
Hipoteza	Tam je fant, ki izkazuje naklonjenost staremu moškemu.
Razlaga	Poljubljanje je način izkazovanja naklonjenosti.
Oznaka	*implikacija*
Premisa	Črno-beli pes skače čez rdeče-belo palico.
Hipoteza	Pes spi.
Razlaga	Psi ne skačejo, ko spijo.
Oznaka	*kontradikcija*
Premisa	Mlada ženska, oblečena v belo jopico in kratke hlače, sedi na robu ploščadi, ki je dvignjena nad vodno površino.
Hipoteza	Ženska sedi na pomolu in opazuje sončni zahod.
Razlaga	Ni nujno, da ženska, ki sedi na ploščadi, opazuje sončni zahod.
Oznaka	*nevtralno*

Opomba. Primeri so bili prevedeni v slovenščino, po eden za vsako oznako.

3.2.1 Prevajanje

Ker se osredotočamo na logično sklepanje v slovenščini, smo del množice ESNLI strojno prevedli. Najprej smo se morali odločiti, kateri strojni prevajalnik uporabiti. Na voljo sta bili dve storitvi v oblaku, Google Prevajalnik (Google Prevajalnik, 2023) in prevajalnik DeepL (DeepL Translate API, 2023), ki ponujata programski vmesnik API. Dodatno smo imeli na voljo še prostodostopni strojni prevajalnik NeMo iz projekta RSDO (Lebar Bajec idr., 2022), naučen za prevajanje iz angleščine v slovenščino.

Za odločitev, katerega od prevajalnikov uporabiti, smo najprej prevedli manjše število primerov z vsakim od treh prevajalnikov, prevode uporabili za učenje klasifikacijskih modelov in na osnovi uspešnosti posameznega modela izbrali najboljšega. Učenje modelov na treh različnih prevodih je opisano v razdelku 4.2.1. Na podlagi njihovih rezultatov, predstavljenih v istem razdelku, smo za prevajanje večje množice uporabili Google Prevajalnik.

Naključno je bilo izbranih 50 tisoč primerov iz učne množice, tri tisoč iz validacijske in tri tisoč iz testne. Njihovi prevodi so predstavljali učno, validacijsko in testno množico za učenje modelov. V nekaterih izbranih

primerih učne množice so manjkale razlage ali hipoteze, poleg tega pa so bile nekatere zelo kratke razlage zelo slabe zaradi nesmiselnosti ali nerazumljivosti (npr. *his new it*, *man and guy*, *Runs in runs* ...), podobno pa so bile nesmiselne tudi nekatere zelo kratke hipoteze, ki so bile ali napačne oziroma zmotno skrajšane ali pa je šlo za primer le ene besede brez konteksta (npr. *Two wom*, *Fetch*, *f*, *A baby is* ...). Prevajanje takšnih primerov bi bilo nesmiselno, ročno preverjanje vsakega od njih pa preveč zamudno. Učno množico smo zato filtrirali tako, da smo odstranili vse primere, kjer je bila razlaga krajša od 14 znakov ali hipoteza krajša od 10 znakov. Naša ocena je namreč, da je večina tako odstranjenih primerov nesmiselna, če pa katero od mej zvišamo, bi odstranili dosti primerov, ki vsebujejo kratke, a ustrezne razlage oziroma hipoteze. Po filtriranju vsebuje učna množica 49.922 primerov.

Vse tri množice smo nato prevedli z Google Prevajalnikom, pri čemer smo za primere iz testne in validacijske množice prevedli le prvo od treh razlag. Množice, ki vsebujejo identifikator primera, izvirno angleško hipotezo, premiso in razlago ter njihove prevode v slovenščino, smo v formatu TSV objavili na spletu.[1] V nadaljevanju bo ta podatkovna množica imenovana ESNLIsi.

Napake v prevodih Po prevajanju smo naključno izbran vzorec prevodov še pregledali, da bi ocenili njihovo kvaliteto. Čeprav je lahko naša ocena nekoliko subjektivna, menimo, da je večina primerov prevedenih ustrezno. Pri nekaterih je prevedena formulacija rahlo nerodna, vendar kljub temu slovnično in pomensko pravilna. V nekaj odstotkih prevodov se pojavljajo napake. Primere napačnih prevodov prikazuje Tabela 4.

Prva vrsta napake je napačen prevod ene od besed, kar lahko vidimo v razlagi prvega primera. Tam je prevajalnik angleško besedo *batter* prevedel kot *udarec* namesto *udarjalec*. Razlaga je zato nerazumljiva, če pa do podobne napake pride v premisi ali hipotezi, je lahko nerazumljiv cel primer.

V drugem primeru je prikazano nekonsistentno prevajanje iste besede. Besedna zveza *gave up* je enkrat prevedena z glagolom *opustil*, drugič pa *obupal*. Razlaga je zato neustrezna, saj bi morala biti uporabljena ali enaka formulacija (kot to velja v angleškem izvirniku) ali pa bi morala

1 https://github.com/timkmecl/nli-slovene?tab=readme-ov-file#podatkovne-mno%C5%BEice

razlaga vsebovati še dodatno obrazložitev, da kdor obupa, nekaj opusti.

V nekaterih primerih prevajanje celo spremeni kategorijo, v katero bi bilo primer pravilno uvrstiti, kar je prikazano v tretjem primeru tabele. Tam je angleška besede "*man*" v premisi prevedena kot "*človek*" namesto "*moški*". Izvirnik je označen kot kontradikcija, saj je kraljica ženska in ne moški (razlaga je ustrezno prevedena). Ker pa je kraljica človek, v slovenskem prevodu pa je namesto angleške besede "*his*", uporabljene v izvirni premisi, spolno nevtralna oblika pridevnika "*svoj*", bi bila ustrezna klasifikacija slovenskega prevoda za nevtralni primer zaradi dejstva, da je kraljica tudi človek.

Tabela 4: *Primeri napak pri strojnem prevajanju podatkovne množice ESNLI*

Oznaka	*implikacija*
Premisa	A man wearing a red shirt with the number 54 hits a baseball, while a catcher prepares to catch the ball.
Hipoteza	A red shirted batter hits the pitch.
Razlaga	A man in a red shirt who hits a baseball is known as a **batter**.
Premisa	Moški, oblečen v rdečo majico s številko 54, udari žogico za baseball, medtem ko se lovilec pripravlja, da ujame žogo.
Hipoteza	Udarjalec v rdeči majici pride na igrišče.
Razlaga	Moški v rdeči majici, ki udari bejzbolsko žogico, je znan kot **udarec**.
Oznaka	*kontradikcija*
Premisa	An elderly male is blowing air into an object.
Hipoteza	The elderly man **gave up** blowing air into the object
Razlaga	If he **gave up** how can he be blowing air.
Premisa	Starejši moški piha zrak v predmet.
Hipoteza	Starejši moški je **opustil** vpihovanje zraka v predmet
Razlaga	Če je **obupal**, kako lahko piha zrak.
Oznaka	*kontradikcija*
Premisa	**A man** sits on his throne behind the drums.
Hipoteza	The Queen of England sits behind some drums.
Razlaga	A Queen is a woman not **a man**.
Premisa	**Človek** sedi na svojem prestolu za bobni.
Hipoteza	Angleška kraljica sedi za bobni.
Razlaga	Kraljica je ženska in ne **moški**.

Opomba. Za vsakega od treh primerov je najprej podana oznaka primera, potem premisa, hipoteza in razlaga v izvirniku, nato pa še njihovi slovenski prevodi. Napake in nekonsistentnosti so označene odebeljeno.

4 Učenje in vrednotenje modelov

V tem razdelku v štirih sklopih predstavimo načine, kako smo učili in uporabili jezikovne modele. Pristopi prvih treh podrazdelkov temeljijo na prilagoditvi (angl. *fine-tuning*) vnaprej naučenih slovenskih velikih jezikovnih modelov. V podrazdelkih 4.1 in 4.2 je opisano učenje jezikovnega modela SloBERTa, najprej več pristopov učenja na množici SI-NLI, nato pa še na prevedeni množici ESNLIsi. V podrazdelku 4.3 je opisan poskus generiranja razlag z generativnimi modeli družine SloT5, ki smo jih učili na razlagah množice ESNLIsi. V podrazdelku 4.4 je opisana uporaba velikega angleškega modela GPT-3.5-turbo. Na koncu, v razdelku 4.5, pa je predstavljen še način vrednotenja pristopov.

Cilj vrednotenja je bil preveriti, kako uspešni so različni pristopi pri klasifikaciji primerov iz testne množice SI-NLI, ki služi kot merilo za vse pristope. V razdelkih 4.3 in 4.4 je predstavljena ocena uspešnosti generiranja razlag, rezultati so predstavljeni v razdelku 5.

Vsa programska koda je bila napisana v programskem jeziku Python v okolju Jupyter Notebook. Za učenje modelov smo uporabili storitev Kaggle Notebooks, ki omogoča zaganjanje datotek Jupyter Notebook v oblaku. Vsi modeli so bili učeni na virtualnem stroju na grafični kartici Nvidia Tesla P100 s 16 GB pomnilnika. Za delo s podatki je bila uporabljena knjižnica Pandas, za učenje modelov pa HuggingFace Transformers (Wolf idr., 2020). Koda je javno dostopna na spletu.[2]

4.1 Učenje klasifikatorja SloBERTa na SI-NLI

V tem sklopu smo najprej prilagodili model SloBERTa[3] na učni množici SI-NLI. Ta model služi kot izhodišče za primerjavo ostalih pristopov. Učenje modela smo trikrat ponovili ob različnih delitvah na učno in validacijsko množico, da bi ocenili občutljivost na izbiro učnih podatkov. Z učenjem modela na podmnožici učne množice, ki vsebuje le primere, pri katerih nobeden od označevalcev ni storil napake, skušamo odgovoriti na vprašanje, kako ti primeri vplivajo na uspešnost učenja

2 https://github.com/timkmecl/nli-slovene
3 https://huggingface.co/EMBEDDIA/sloberta

in ali gre pri napakah le za človeško površnost oziroma ali so ti primeri inherentno težji od ostalih ali dvoumni. Za primerjavo je naučen še en model na naključno izbrani podmnožici iste velikosti.

4.1.1 Izbira parametrov in učenje modelov

Parametre učenja smo poskušali nastaviti tako, da bi z učenjem dosegli čim večjo klasifikacijsko točnost na validacijski množici, pri čemer pa smo želeli, da se model neha izboljševati prej kot v 20 epohah (angl. *epoch*), saj bi sicer učenje trajalo predolgo. Za učenje vseh modelov tega razdelka smo uporabili enake parametre, zato jih navajamo le enkrat. Število epoh učenja smo tako nastavili na 20, stopnja učenja (angl. *learning rate*), ki se je izkazala za najbolj uspešno, je 10^{-5}, delež ogrevanja (angl. *warmup ratio*) pa 0,05. Uporabljena velikost paketa (angl. *batch size*) za učenje je 32. Nastavili smo jo tako, da je čim večja, saj vzporedna obravnava večjega števila primerov pohitri učenje, hkrati pa je dovolj majhna, da zahteve po pomnilniku grafične kartice ne presežejo 16 GB. Uporabljen je privzeti algoritem učenja AdamW s privzetimi parametri $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$. Po vsaki epohi učenja model generira napovedi za validacijsko množico, kot končni model pa je shranjen model tiste epohe, po kateri je bila klasifikacijska točnost na validacijski množici največja. Po enakem postopku smo model SloBERTa učili tudi na podmnožici učne množice s soglasnimi označevalci.

Zgolj neposredna primerjava rezultatov tega modela z osnovnim ne bi bila ustrezna, saj lahko na kvaliteto modela vpliva tudi količina učnih podatkov. Zato smo iz učne množice naključno izbrali toliko primerov, kot jih je v podmnožici s soglasnimi ocenjevalci (3015). Še na teh primerih smo učili model SloBERTa, rezultati pa dajejo oceno, ali bi s povečanjem števila primerov v podatkovni množici SI-NLI lahko bistveno izboljšali rezultate na njej učenih modelov ali pa je že pri trenutnem številu primerov omejujoč dejavnik zmogljivost modela in ne velikost učne množice.

4.2 Učenje s prenosom iz ESNLIsi

V tem sklopu se ukvarjamo z učenjem s prenosom, za kar uporabljamo podatkovno množico ESNLI in lasten strojni prevod njenega dela v slovenščino, ESNLIsi. Zanima nas predvsem, koliko je pri logičnem sklepanju v slovenščini uporabno prevajanje angleških podatkovnih množic kot potencialna rešitev za majhno količino ustreznih učnih primerov, ki so na voljo v slovenščini. Vira stavkov pri SI-NLI in ESNLI sta povsem različna (članki, dokumenti, knjige itd. pri SI-NLI in opisi slik pri ESNLI (Bowman idr., 2015; Klemen idr., 2022; Logar idr., 2013)). To dejstvo smo uporabili za ugotavljanje, kako dobro modeli, naučeni za logično sklepanje na specifičnem tipu stavkov, posplošujejo na drugačen tip. Metoda predobdelave podatkov, izbire parametrov, učenja modelov in izbire končnega modela je enaka kot v prejšnjem razdelku, zato so v tem razdelku navedene le morebitne razlike v metodi oziroma izbranih parametrih.

4.2.1 Izbira prevajalnika

Izvedba zastavljenih ciljev je zahtevala slovenski prevod angleške množice ESNLI, za kar je bila najprej potrebna izbira enega od treh strojnih prevajalnikov (DeepL, Google in NeMo).

Naključno smo izbrali 200 primerov vsakega od treh razredov iz testne množice in po 200 iz validacijske ter tako dobili testno in validacijsko množico s po 600 primeri. Naključno smo izbrali tudi po 500 primerov iz vsakega od treh razredov iz učne množice in dobili učno množico s 1500 primeri. Nato smo vse tri množice prevedli z vsakim od prevajalnikov. Za prevajanje z Google Prevajalnikom in DeepL smo uporabili njihove vmesnike API za Python, za prevajalnik NeMo pa smo najprej prenesli objavljeni model[4] in ga zagnali na lastnem računalniku z uporabo modula nemo_toolkit za Python.

Na vsakem od prevodov smo prilagodili model SloBERTa pri stopnji učenja 10^{-5} in deležu ogrevanja 0,02. Točnosti modelov na pripadajočih testnih množicah so predstavljene v Tabeli 5. Glede na majhno količino učnih primerov je razlika med njimi relativno majhna, dodatno pa bi težava z uporabo teh točnosti za izbiro najboljšega prevoda lahko

4 https://www.clarin.si/repository/xmlui/handle/11356/1736

bila uporaba prevodov istega prevajalnika tudi za vrednotenje. Tako bi lahko prevajalnik, katerega prevodi so slabši, a bolj konsistentni, morda dal boljši rezultat kot prevajalnik, ki je boljši, a morda prevaja na manj konsistenten način.

Tabela 5: *Točnost napovedi v % za tri strojne prevajalnike*

	Deepl	*Google*	*NeMo*
Prevedena množica ESNLIsi	**74,3**	73,0	71,2
Slovenska množica SI-NLI	50,2	**52,7**	48,0

Temu se lahko izognemo z vrednotenjem modelov na primerih, ki so izvorno v slovenščini. V ta namen smo generirali napovedi za validacijsko množico SI-NLI, kot to prikazuje Tabela 5. Na osnovi teh podatkov smo za prevajanje izbrali storitev Google Prevajalnik. Dodatna prednost tega prevajalnika je, da je bilo, za razliko od storitve DeepL, prevajanje dane količine podatkov brezplačno, v primerjavi s prevajalnikom NeMo, kjer prevajanje poteka na lastnem računalniku, pa precej hitrejše. Postopek izbire podatkov in prevajanje, s katerim smo dobili množico ESNLIsi, je opisan v razdelku 3.2.1.

4.2.2 Učenje modelov

Na slovenskih prevodih učne množice ESNLIsi smo 10 epoh pri stopnji učenja 10^{-5} in deležu ogrevanja 0,02 učili model SloBERTa. Kriterij za selekcijo končnega modela je bila klasifikacijska točnost na validacijski množici ESNLIsi.

Podobno kot v prejšnjem sklopu smo tudi tu želeli ugotoviti, koliko bi učenje modela na večjem številu primerov izboljšalo rezultat, s čimer bi določili tudi smiselnost nadaljnjega prevajanja primerov iz ESNLI. Model SloBERTa smo zato učili na podmnožici 40 tisoč naključno izbranih primerov (80 %) iz učne množice ESNLIsi z istimi parametri kot prvič in z nespremenjeno validacijsko množico.

Poleg tega smo želeli za primerjavo s SI-NLI določiti še zmogljivost modela, ki bi bil naučen na podobnem številu primerov. V ta namen smo učili še en model SloBERTa na naključno izbranih 4000 primerih učne množice ESNLIsi (velikost učne množice SI-NLI je 3802 primerov) pri stopnji učenja 10^{-5} in deležu ogrevanja 0,02, za validacijsko

množico pa smo iz validacijske množice ESNLIsi naključno izbrali 600 primerov (velikost validacijske množice SI-NLI je 590).

Z vsemi tremi modeli smo generirali napovedi za testno množico ESNLIsi. Za vrednotenje učenja s prenosom smo z njimi generirali še napovedi za testno množico SI-NLI. Z osnovnim modelom, naučenim na SI-NLI, pa smo generirali napovedi za testno množico ESNLIsi.

4.2.3 Prilagajanje na SI-NLI

Ugotoviti smo želeli, če oziroma koliko lahko z uporabo ESNLIsi, ki ima več kot 10-krat več primerov kot SI-NLI, izboljšamo model glede na osnovnega, naučenega le na SI-NLI. To smo poskusili storiti s prilagajanjem modela, ki smo ga najprej učili na učni množici ESNLIsi (gl. 4.2.2). Učenje modela se je nato nadaljevalo na učni množici SI-NLI, za validacijsko množico pa je bila uporabljena validacijska množica SI-NLI, saj je bil cilj dobiti model, ki bi čim bolje napovedoval na testni množici SI-NLI. Model smo učili 6 epoh, tokrat pri manjši stopnji učenja $4 \cdot 10^{-6}$. S tako dobljenim modelom smo generirali napovedi za testno množico SI-NLI.

4.3 Generiranje razlag s SloT5

V prejšnjih dveh sklopih je bil cilj čim bolj uspešna klasifikacija primerov. Problem logičnega sklepanja pa lahko razširimo tako, da ne zahtevamo le končne oznake primera, temveč želimo imeti še razlago, ki nam pojasni, zakaj je ta oznaka primerna. Naloga je torej generativnega in ne klasifikacijskega tipa, zato v tem delu uporabljamo vnaprej naučena slovenska generativna modela družine SloT5 (manjši model t5-sl-small[5] in večji t5-sl-large[6]). Ker podatkovna množica SI-NLI ne vsebuje razlag primerov, je učenje potekalo le z uporabo množice ESNLIsi. Končni namen je uporaba tako naučenih modelov za generiranje razlag za primere iz testne množice SI-NLI.

Postopek predobdelave podatkov, učenja modelov in izbire parametrov zanje je podoben kot pri učenju modelov SloBERTa, opisanih v prejšnjih razdelkih, le da so tu uporabljeni ekvivalentni razredi

5 https://huggingface.co/cjvt/t5-sl-small
6 https://huggingface.co/cjvt/t5-sl-large

knjižnice HuggingFace, namenjeni učenju generativnih modelov. Razlika je v tem, da generativni model pri učenju kot ciljni izhod zahteva besedilo, podano kot zaporedje členov. Ciljno besedilo je v našem primeru razlaga primera brez dodatne predobdelave.

Naučili smo tri modele. Za učenje dveh smo začeli z modelom t5-sl-small. Učenje prvega je potekalo na naključno izbrani podmnožici učne množice ESNLIsi s 4000 primeri iz prejšnjega razdelka. Stopnja učenja je bila $8 \cdot 10^{-5}$, padanje uteži (angl. *weight decay*) 0,01, velikost paketa 32, ogrevanje pa ni bilo uporabljeno.

Model se je učil 10 epoh. Drugi model smo učili na celotni učni množici ESNLIsi pri stopnji učenja $4 \cdot 10^{-5}$ in ostalih parametrih, enakih kot v prejšnjem primeru.

Na podmnožici s 4.000 primeri smo učili še model t5-sl-large s stopnjo učenja 10^{-4}, padanjem uteži 0,01 in velikostjo paketa 4, kar je bila največja možna velikost za pomnilnik uporabljene grafične kartice. Model se je učil tri epohe.

Z vsemi tremi modeli smo generirali razlage za primere testne množice ESNLIsi. Za vrednotenje razlag smo uporabili prvih 50 primerov te množice (19 primerov implikacije, 19 kontradikcije in 12 nevtralnih primerov). Zaradi slabih rezultatov že na tej množici modelov nismo vrednotili še na primerih iz množice SI-NLI. Pri tem bi namreč pričakovali še slabše rezultate, saj bi šlo za učenje s prenosom med dvema različnima podatkovnima množicama — to se je pokazalo že pri vrednotenju klasifikacijskih modelov prejšnjega sklopa.

Zaradi rezultatov opisanih treh modelov, predstavljenih v razdelku 5.3, smo se odločili, da večjega SloT5 modela ne bomo učili na celotni množici ESNLIsi. To učenje bi zahtevalo velik časovni vložek, saj bi trajalo več ur, hkrati pa boljšega rezultata od že predstavljenih ni pričakovati.

4.4 Uporaba GPT-3.5-turbo

Vsi do sedaj opisani pristopi so temeljili na prilagoditvi vnaprej naučenih jezikovnih modelov SloBERTa in SloT5 na določeni podatkovni množici. Uporabljeni modeli so bili vnaprej naučeni na slovenščini, vsebujejo pa nekaj sto milijonov parametrov. V tem razdelku je predstavljen alternativni pristop. Za napovedovanje oznak v testni množici

SI-NLI uporabimo učenje brez dodatnih primerov (angl. *zero-shot learning*) in z nekaj dodatnimi primeri (angl. *few-shot learning*) z modelom GPT-3.5-turbo, ki ni vnaprej naučen specifično na slovenščini, temveč primarno na angleških besedilih, vsebuje pa nekaj redov velikosti več parametrov od že uporabljenih slovenskih modelov. Enak pristop je bil uporabljen tudi za generiranje razlag.

Pri tem pravzaprav ne gre za učenje nevronskih mrež v običajnem pomenu besede. Pri učenju brez dodatnih primerov modelu zgolj postavimo vprašanje oziroma mu kot vhod damo navodilo, izhod pa je že generirano besedilo, odgovor oziroma napoved. Učna množica, na kateri bi se model učil, tako ni potrebna. Pri učenju z nekaj dodatnimi primeri vhod modela poleg navodila vsebuje še nekaj primerov problema s podanim odgovorom. Tu učno množico potrebujemo le kot vir primerov, zato je lahko zelo majhna (zgolj nekaj primerov). Skrajna različica učenja z nekaj dodatnimi primeri je učenje v kontekstu (angl. *in-context learning*), kjer model kot del vhoda dobi tudi več sto rešenih primerov. Pri obeh tipih učenja je pomembna izbira navodila, saj razumljivost navodila vpliva na kvaliteto odgovorov.

Testiranja smo se lotili v več korakih. V prvem koraku smo uporabili spletni vmesnik OpenAI Playground, v katerem smo na nekaj primerih iz učne množice SI-NLI ročno preizkušali različna navodila. Cilj tega je bil predvsem izločitev navodil, ki bi bila modelu očitno nerazumljiva. Želeli smo tudi doseči, da bi bil odgovor modela v želenem formatu, npr. ena sama beseda pri klasifikaciji in jasno ločena razlaga in končna napoved oznake pri generiranju razlag.

V drugem koraku smo izbrali naključno množico 100 primerov iz učne množice SI-NLI (naključni izbor je vseboval 41 primerov implikacije, 35 kontradikcije in 24 nevtralnih primerov). Napisali smo program, ki za vsak primer generira vhodni tekst za model na podlagi določenega navodila, nato z uporabo programskega vmesnika OpenAI API za Python (razred openai.ChatCompletion) avtomatsko pošilja vhodna besedila modelu, od njega prejme odgovore in shrani napovedane oznake (oziroma generirane razlage). Parameter temperatura (angl. *temperature*) smo nastavili na 0, kar zagotavlja deterministične odgovore, saj pri generiranju teksta model za naslednjo besedo vedno izbere tisto, ki je označena kot najbolj verjetna.

Zadnji korak je bila uporaba enakega pristopa za napovedovanje oznak za celotno testno množico SI-NLI, kjer smo uporabili le navodila, katerih rezultati prejšnjega koraka so bili najboljši. S tem, da so bili primeri v prejšnjem koraku vzeti iz učne množice in ne iz testne, smo preprečili, da bi prek izbire navodil prišlo do prekomernega prileganja podatkom v testni množici.

Žal je bil v času naše uporabe programski vmesnik OpenAI zelo nestabilen, verjetno zaradi preobremenjenosti strežnikov. Strežnik namreč programu pogosto ni vrnil odgovora, zato je bilo treba izvajanje prekiniti. Poleg tega je vmesnik namesto odgovora večkrat javljal napako, da je strežnik preobremenjen. Po eni takšni napaki je bilo za nadaljevanje uporabe modela potrebno čakanje, pogosto nekajminutno, sicer je ob vsaki naslednji poslani zahtevi strežnik vrnil isto napako. Zaradi tega ni bilo možno implementirati avtomatskega ponovnega zaganjanja v primeru napake. Pogostost prekinitev je bila večja v primeru daljših vhodnih besedil in daljših generiranih izhodov, kjer je do napake prišlo na vsakih nekaj primerov. Celovito testiranje pristopa v teh okoliščinah bi bilo bistveno preveč zamudno, zato smo omejili število različnih pristopov, na celotni učni množici pa smo kasneje vrednotili le navodilo, ki se je pri vrednotenju na 100 primerih izkazalo za najboljše.

4.4.1 Učenje brez dodatnih primerov

Najprej smo poskusili učenje brez dodatnih primerov. Na množici 100 primerov smo testirali štiri različna navodila:

– *navodilo-1-en*, ki so ga uporabili Zhong idr. (2023):

```
Given the sentence „{premise}“, determine if the following statement is entailed or contradicted or neutral: „{hypothesis}“
The answer (entailed or contradicted or neutral) is:
```

– *navodilo-1-si*, slovenski prevod prejšnjega:

```
Glede na stavek „{premise}“ ugotovite, ali je naslednja izjava posledica, kontradikcija ali nevtralna: „{hypothesis}“
Odgovor (posledica ali kontradikcija ali nevtralna) je:
```

– *navodilo-2-en*, prirejeno po H. Liu idr. (2023):

```
Instructions: You will be presented with a premise and a
    hypothesis about that premise in Slovene. You need to
    decide whether the hypothesis is entailed by the premise
    by choosing one of the following answers: 'entailment':
    The hypothesis follows logically from the information
    contained in the premise. 'contradiction': The hypothe-
    sis is logically false from the information contained in
    the premise. 'neutral': It is not possible to determine
    whether the hypothesis is true or false without further
    information. Read the passage of information thorough-
    hly and select the correct answer from the three answer
    labels. Read the premise thoroughly to ensure you know
    what the premise entails.
Premise: {premise}
Hypothesis: {hypothesis}
Answer (just one word, either entailment/neutral/
    contradiction):
```

– *navodilo-2-si*, slovenski prevod prejšnjega:

```
Navodila: Predstavljena vam bo premisa in hipoteza o tej
    premisi. Odločiti se morate, ali je hipoteza posledica
    premise, tako da izberete enega od naslednjih odgovo-
    rov: 'posledica': Hipoteza logično sledi iz informacij,
    ki jih vsebuje premisa. 'kontradikcija': Hipoteza je
    logično napačna glede na informacije, ki jih vsebuje
    premisa. 'nevtralno': Brez dodatnih informacij ni mo-
    goče ugotoviti, ali je hipoteza resnična ali napačna.
    Natančno preberite odlomek informacij in izberite pra-
    vilen odgovor med tremi oznakami odgovorov. Temeljito
    preberite predpostavko, da boste vedeli, kaj sledi iz
    predpostavke.
Premisa: {premise}
Hipoteza: {hypothesis}
Odgovor (samo ena beseda, bodisi posledica/nevtralno/
    kontradikcija):
```

V vseh prikazanih navodilih je za vhod v model niz {premise} zamenjan s premiso, {hypothesis} pa s hipotezo primera (tako pri angleških kot pri slovenskih navodilih sta premisa in hipoteza v slovenščini). S testiranjem v OpenAI Playground smo predhodno ugotovili, da je izhod modela kdaj res le ena beseda kot zahtevano (npr. *kontradikcija*), kdaj pa odgovori v celem stavku (npr. *Izjava je kontradikcija*.). Končna klasifikacija je zato določena na podlagi pojavitve podniza v izhodu oziroma odgovoru. Če odgovor vsebuje podniz *posledica* ali *entail*, štejemo, da je napovedana oznaka implikacija, če vsebuje *kontradikcija* ali *contradict* je kontradikcija, če *nevtral* ali *neutral* pa je nevtralna. Za procesiranje odgovorov smo napisali funkcijo v jeziku Python. Ta način klasifikacije odgovorov je bil zadosten, saj je v vseh primerih odgovor vseboval enega od teh podnizov.

Kot najboljše se je izkazalo *navodilo-1-en*, posamezni rezultati so predstavljeni v naslednjem razdelku. Z uporabo tega navodila so bile na enak način kot prej napovedane oznake za celotno testno množico SI-NLI.

Za razliko od vnaprej naučenih modelov iz prejšnjih razdelkov, ki so bili učeni na slovenščini, je GPT-3.5-turbo v največji meri učen na angleških besedilih. Zato smo želeli preveriti, če in koliko slabši je rezultat zaradi morebitnega slabšega razumevanja v učnih podatkih bistveno manj zastopane slovenščine v primerjavi z angleščino. Napovedi za izbrano množico 100 primerov smo tako generirali še z uporabo njihovih angleških prevodov (in navodila *navodilo-1-en*).

4.4.2 Učenje z nekaj dodatnimi primeri

Najboljše navodilo (*navodilo-1-en*) smo uporabili za osnovo testiranja učenja z nekaj dodatnimi primeri. Sledili smo načinu, ki so ga uporabili Zhong idr. (2023). Iz učne množice smo naključno izbrali po en primer vsake od treh oznak ter njihove premise in hipoteze vstavili v navodilo. Na koncu navodil smo dodali odgovor entailed, če je bil primer v razredu implikacija, če kontradikcija *contradicted* in *neutral* za nevtralne primere. Isti nabor treh dopolnjenih navodil smo dodali na začetek vhoda za vse primere pri napovedovanju.

Na ta način smo generirali napovedi za prej opisano množico 100 primerov. Poskus smo skupaj ponovili trikrat, vsakič z drugimi tremi naključnimi izbranimi primeri. Rezultati so predstavljeni v razdelku 5.4. Ker ta pristop v primerjavi z učenjem brez dodatnih primerov skoraj ni izboljšal rezultatov, zaradi tehničnih težav strežnikov OpenAI, opisanih na začetku razdelka, tega pristopa nismo dodatno vrednotili na testni množici SI-NLI. Zaradi počasnosti strežnika tudi nismo poskušali uporabiti več kot treh podanih primerov naenkrat.

4.4.3 Generiranje razlag

Podobno metodo smo uporabili še za hkratno generiranje razlag in klasifikacijo. S pomočjo testiranja znotraj vmesnika OpenAI Playground smo *navodilo-1-en* modificirali tako, da je model kot izhod najprej podal razlago v slovenščini, v naslednji vrstici pa klasifikacijo. Navodilo, ki smo ga na koncu uporabili, je bilo oblike:

```
Given the sentence „{premisa}“, determine if the following
    statement is entailed or contradicted or neutral: „{hi-
    poteza}“. First give a short, one sentence reasoning or
    explanation for the decision in slovene, and then the
    final answer (one english word - „entailed“ or „contra-
    dicted“ or „neutral“) in a new line after that.

Razlaga v slovenščini:
```

Niza {premisa} in {hipoteza} sta kot prej vsakič zamenjani s premiso in hipotezo danega primera. Izhod modela nato razdelimo na dva dela glede na znak za novo vrstico. Prvi del vzamemo za razlago, drugi del pa obravnavamo enako kot odgovor pri prej opisanem pristopu in mu dodelimo oznako glede na vsebovan podniz. Pri ročnem testiranju smo ugotovili, da kljub zahtevi v navodilu, naj bo končni odgovor beseda v angleščini, kdaj odgovori v slovenščini, pri čemer namesto besede *entailed* uporabi besedo *potrjeno*. Zato tudi primere, ki vsebujejo podniz *potrjen*, pri napovedovanju uvrstimo kot implikacijo.

Z opisano metodo smo generirali napovedi in razlage za množico 100 primerov. Na koncu smo izbrali podmnožico 50 primerov, ki jo bomo uporabili za kvalitativno vrednotenje generiranih razlag. Želeli smo, da pogostost posameznih razredov v njej vsaj približno odraža pogostost razredov v celotni podatkovni množici, zato smo iz množice stotih naključno izbrali 18 primerov implikacije, 16 kontradikcije in 16 nevtralnih primerov.

4.5 Način vrednotenja

V tem razdelku opišemo, kako smo vrednotili različne pristope vrednotili. Najprej predstavimo načine za vrednotenje klasifikatorjev. Definiramo nekaj metrik in utemeljimo primernost njihove uporabe za naš problem. Opišemo tudi način vrednotenja generiranih razlag. Rezultati opisanega vrednotenja so predstavljeni v naslednjem razdelku.

4.5.1 Evalvacijske metrike za klasifikacijo

Za kvantitativno vrednotenje klasifikatorjev poznamo več različnih metrik. Temeljna je klasifikacijska točnost (angl. *classification accuracy*) ali samo točnost, ki nam pove delež primerov, ki jih je klasifikator uvrstil v pravilen razred.

Zgolj ta podatek nam pogosto ne da zadostne informacije o delovanju klasifikatorja, predvsem v primeru neuravnoteženih podatkovnih množic. Poznamo metrike, s katerimi lahko bolje ovrednotimo tudi tovrstne primere. Natančnost (angl. *precision*) pove, kolikšen delež primerov, uvrščenih v izbrani ciljni razred, temu razredu dejansko pripada. Priklic (angl. *recall*) pove, kolikšen delež primerov ciljnega razreda je vsebovan med pozitivnimi napovedmi tega razreda. Ocena F_1 (angl F_1-*score*) je harmonična sredina natančnosti in priklica in služi kot nekakšen povzetek sposobnosti klasifikatorja za napovedovanje izbranega razreda, kar je lahko dobra alternativa klasifikacijski točnosti (Müller in Guido, 2016).

Če imamo pri klasifikacijskem problemu le en ciljni razred, lahko za vrednotenje napovedi uporabimo le metrike za ta razred. Pri problemu klasifikacije v več razredov pa jih najprej izračunamo za vsak razred posebej, nato pa izračunamo njihovo povprečje. Uporaba mikro

povprečja je priporočena, kadar je enako pomemben vsak posamezen primer, če pa je enako pomemben vsak posamezen razred, se uporabi makro povprečje (Müller in Guido, 2016).

4.5.2 Vrednotenje klasifikatorjev

Obe uporabljeni testni množici sta skoraj uravnoteženi, zato smo pri vrednotenju klasifikatorjev najbolj upoštevali klasifikacijsko točnost in o njej poročamo pri vseh pristopih. Dodatno pri vsakem pristopu navajamo še povprečje ocene F_1, povprečje natančnosti in priklicev.

Pri problemu logičnega sklepanja v naravnem jeziku, s katerim se ukvarjamo, nobeden od treh razredov ni privilegiran glede na ostale. Ker so razredi za nas enakovredni, je smiselna uporaba makro povprečja metrik. V nadaljevanju besedila zato ocena F_1 danega pristopa pomeni makro povprečje ocen F_1 za vse tri razrede, enako velja tudi za natančnost in priklic. Glavni merili za primerjavo pristopov sta tako klasifikacijska točnost in ocena F_1, in sicer na testni množici SI-NLI, ki je izvirno slovenska, človeško ustvarjena množica.

Ker nas zanima tudi, kakšne tipe napak delajo modeli in ne le njihova sposobnost napovedovanja razredov, pri nekaterih rezultatih dodamo še matriko zamenjav (angl. *confusion matrix*), ki vsebuje po en stolpec in eno vrstico za vsakega od možnih razredov. Element matrike v stolpcu A in vrstici B pove število primerov razreda B, za katere je klasifikator napovedal razred A. Vsota diagonale te matrike je število pravilno uvrščenih primerov.

V poskusih iz drugega sklopa, kjer se ukvarjamo z učenjem s prenosom, navajamo še klasifikacijske točnosti modelov na testni množici ESNLIsi.

V poskusih uvrščanja z GPT-3.5-turbo zaradi opisanih težav za vse pristope izračunamo točnost, oceno F_1 in matrike zamenjav za izbrano množico 100 primerov, ki smo jo uporabili za vrednotenje. Ker v tej množici primeri niso tako enakomerno razporejeni med razredi kot v celotnih testnih množicah, je tu relativno bolj pomembna ocena F_1. Prej navedene metrike za celo testno množico SI-NLI izračunamo samo za najuspešnejši pristop na manjši množici, kot je opisano v razdelku 4.4.

4.5.3 Vrednotenje generiranja razlag

Za generirana besedila obstaja več metrik, ki jih lahko izračunamo na podlagi primerjave generiranega besedila s ciljnimi odgovori, ki jih imamo v testni množici. Pristopi temeljijo na podobnosti dveh besedil. Ker lahko za en primer obstajata dve ali več pravilnih razlag, ki se med seboj razlikujejo, poleg tega pa lahko le ena beseda povsem spremeni pravilnost razlage, takšen povsem kvantitativen pristop tu ne bi bil ustrezen. Namesto tega smo 50 razlag za vsak pristop ročno pregledali in jih ocenili kot ustrezne ali ne.

Da je razlaga ocenjena kot ustrezna, mora biti pravilna, torej mora najprej dati argument za pravilno klasifikacijo primera. Zgolj pravilnost pa ne zadošča. Razlage, kot je na primer *Druga trditev pove isto kot prva* (kot utemeljitev, zakaj je nek primer implikacija) ali *Navedbe v drugi trditvi nasprotujejo prvi* (za kontradikcijo), so lahko sicer pravilne, a ne podajajo nobene dodatne informacije. Zahtevali smo, da je iz razlage razvidno neko razumevanje, torej da razlaga izpostavi, kaj v premisi in hipotezi privede do tega, da se primer uvršča v nek razred.

Pri tem nismo zahtevali, da je razlaga popolna. Na primer, če je razlogov za kontradikcijo več, zadošča navedba enega. Prav tako za pozitivno oceno nismo zahtevali slovnične pravilnosti, želeli smo le razumljivost, na oceno pa ni vplival niti slog.

Modele SloT5, učene na ESNLIsi, iz tretjega sklopa smo vrednotili na prvih 50 primerih testne množice ESNLIsi, ki vsebujejo 19 primerov implikacije, 19 kontradikcije in 12 nevtralnih primerov. Razlage, generirane z GPT-3.5-turbo, smo vrednotili na množici 50 primerov iz SI-NLI, in sicer 18 naključno izbranih primerov implikacije, 16 kontradikcije in 16 nevtralnih primerov.

5 Rezultati

V tem razdelku po sklopih predstavimo rezultate, pridobljene z metodami, opisanimi v prejšnjem razdelku. Pristope ocenimo tako kvantitativno kot tudi kvalitativno. Rezultate interpretiramo in na njihovi podlagi skušamo odgovoriti na izhodiščna vprašanja.

5.1 Učenje klasifikatorja SloBERTa na SI-NLI

Na testni množici SI-NLI smo vrednotili vse tri v tem sklopu naučene modele. Model SloBERTa, prilagojen na celotni učni množici SI-NLI, imenujemo model *SI-NLI-celotna*. Model *SI-NLI-soglasni* je model SloBERTa, prilagojen le na primerih učne množice, pri katerih so bili vsi označevalci med seboj soglasni in so hkrati izbrali pravilno oznako primera. Model *SI-NLI-manjša* pa je model SloBERTa, ki smo ga prilagodili na naključno izbrani podmnožici učne množice iste velikosti (približno 80 % cele množice). Rezultati vrednotenja so prikazani v Tabeli 6. Model *SI-NLI-celotna*, prilagojen na celotni učni množici SI--NLI, doseže najvišje ocene pri vseh metrikah.

Pri poskusu še trikratne naključne delitve izvirne učne množice SI-NLI na učno in validacijsko smo ugotovili, da znaša povprečje klasifikacijskih točnosti štirih različnih delitev 73,2 %, standardni odklon pa 0,8 %. Točnost je torej nekoliko odvisna od izbire učnih primerov, kar moramo upoštevati pri primerjavi modelov.

Tabela 6: *Metrike v %, izračunane na napovedih za testno množico SI-NLI*

Model	Točnost	F_1	Natančnost	Priklic
SI-NLI-celotna	**73,2**	**73,2**	**73,3**	**73,2**
SI-NLI-manjša	72,2	72,2	72,4	72,4
SI-NLI-soglasni	72,9	73,0	73,3	72,8

Opomba. Za tri modele SloBERTa, učene na celotni učni množici SI-NLI in dveh podmnožicah.

Model *SI-NLI-manjša* ima za 1 % manjšo klasifikacijsko točnost od modela *SINLI-celotna*, podobno tudi ostale metrike. To je vpliv zmanjšanja učne množice na 80 % prvotne. Količina učnih primerov, vsebovana v podatkovni množici SI-NLI (nekaj tisoč), je trenutno torej tako majhna, da bi z razširjanjem množice z dodatnimi primeri rezultate modelov, učenih na njej, še lahko izboljševali.

Oglejmo si vpliv primerov, na katerih se označevalci tudi motijo, na učenje modela. Vidimo lahko, da izločitev takšnih primerov iz učne množice zmanjša točnost za manj kot odstotek. To zmanjšanje je le malo manjše kot takrat, ko so izločeni primeri naključno izbrani, razlika je manjša od standardnega odklona. Sklepamo lahko, da gre

predvsem za posledico manjše učne množice. Delež takih primerov v učni množici torej ne vpliva na uspešnost učenja.

V Tabeli 7 podajamo še točnosti vseh treh modelov na podmnožici učne množice s soglasnimi označevalci, podmnožici ostalih (torej tistih primerih, kjer je kateri od označevalcev naredil napako), in razlike teh dveh točnosti. Izkaže se, da so razlike med modeli le pri podmnožici soglasnih odločitev, medtem ko dajejo enake rezultate na ostalih primerih.

Tabela 7: *Primerjava točnosti napovedi v % za soglasne in nesoglasne primere testne množice SI-NLI*

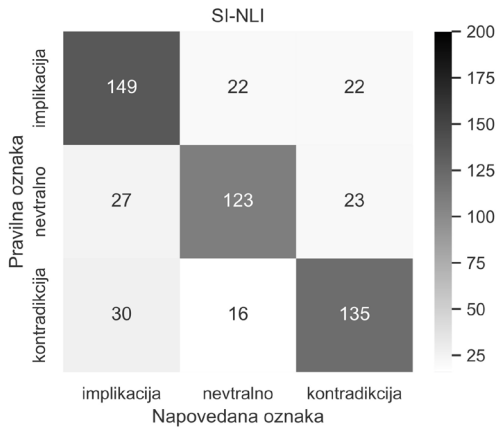
Model	Točnost na soglasnih	Točnost na ostalih	Razlika
SI-NLI-celotna	**76,0**	61,0	**15,0**
SI-NLI-manjša	73,6	61,0	12,5
SI-NLI-soglasni	74,4	61,0	13,4

Pri vseh treh modelih je točnost na soglasnih primerih skoraj 15 % večja kot na ostalih. Vidimo, da primere, pri katerih so imeli težave ljudje, slabše klasificirajo tudi naučeni modeli. To, ali so bili primeri takšne vrste vsebovani v učni množici ali ne, na točnost pri napovedovanju na njih ne vpliva. Najverjetneje gre vsaj delno za dvoumne primere (takšne, ki jih tudi ljudje razumemo na različne načine), kar hkrati razloži tako napake nekaterih človeških označevalcev kot slabše rezultate modelov.

Drugačne oznake nekaterih označevalcev tako vsaj do neke mere niso napake, temveč različne interpretacije istega primera. Posledično bi to lahko upoštevali pri učenju modelov. Možen pristop bi lahko namesto le enega ciljnega razreda pri učenju takim primerom podal verjetnostno porazdelitev, kjer bi bila verjetnost vsakega od razredov delež človeških oznak tega razreda. S tem se nismo ukvarjali in je predlog za nadaljnje delo.

Iz matrik zamenjav modela *SI-NLI-celotna* na Sliki 2 (napake drugih modelov so podobne) vidimo, da so vrste napak, ki jih modeli naredijo, relativno enakomerno razporejene med šestimi možnostmi (ne glede na točno izbiro tipa ali količine učnih primerov). Nobena vrsta napake posebej ne izstopa, prav tako modeli nimajo večjih težav s

primeri enega razreda kot z drugimi. Po tem se očitno razlikujejo od človeških označevalcev (Slika 1), ki razredov implikacija in kontradikcija sploh ne zamenjujejo, največkrat pa zamenjajo primere implikacije in nevtralne primere. Jasno je, da se strojni način reševanja problema razlikuje od človeškega. To opažanje skušamo dodatno razložiti prek generiranja razlag v razdelku 5.3.



Slika 2: Matrika zamenjav za napovedi modela, učenega na celotni učni množici SI-NLI.

5.2 Učenje s prenosom iz ESNLIsi

V tem poskusu smo evalvirali tri modele SloBERTa, prilagojene na množici ESNLIsi: model *ESNLIsi-celotna*, učen na celotni učni množici ESNLIsi; model *ESNLIsi-40k*, učen na 80 % primerov iste množice; model *ESNLIsi-4k*, učen na 4 tisoč primerih te množice, kar je podobno številu primerov v množici SI-NLI; in model *ESNLIsi-SI-NLI*, najprej učen na celi učni množici ESNLIsi, zatem pa prilagojen še na učni množici *SI-NLI*.

5.2.1 Klasifikator za ESNLIsi

V Tabeli 8 so prikazane klasifikacijske točnosti na testni množici ESNLIsi za prve tri modele in model *SI-NLI-celotna* iz prejšnjega razdelka. Največjo točnost ima model *ESNLIsi-celotna*, le malo manjšo *ESNLIsi-40k*. Daleč najmanjša je točnost modela *SI-NLI-celotna*, kjer gre v tem primeru za prenos znanja iz ene množice na drugo.

Tabela 8: *Točnost napovedi v % za testno množico ESNLIsi treh modelov, učenih na ESNLIsi, in modela, učenega na SI-NLI*

Model	*ESNLIsi-celotna*	*ESNLIsi-40k*	*ESNLIsi-4k*	*SI-NLI-celotna*	*Poth idr.*	*Wang idr.*
Točnost	85,7	85,4	80,0	49,3	91,1	**93,1**

Opomba. Dodane so še točnosti, ki so jih na množici ESNLI dosegli Wang idr. (2021) z lastno arhitekturo in Poth idr. (2021) z modelom RoBERTa.

Naša največja dosežena točnost je manjša od največje točnosti na množici SNLI, ki so jo z lastno arhitekturo jezikovnega modela dosegli Wang idr. (2021). Manjša je tudi od točnosti, ki so jo s prilagoditvijo modela RoBERTa dosegli Poth idr. (2021). Njihov model je po zgradbi in velikosti enak modelu SloBERTa, ki smo ga uporabili mi. Razlika je delno posledica tega, da smo uporabili 10-krat manjšo učno množico (prevedli smo le del množice ESNLI), delno posledica napak pri prevajanju (opisanih v razdelku 3.2.1), delno pa zaradi vnaprejšnjega učenja modela na drugih korpusih v različnih jezikih. Razlika znaša približno 5 %, iz česar lahko sklepamo, da je to približek za delež primerov v podatkovni množici, pri katerih je prevod napačen do te mere, da ga ni več mogoče uvrstiti v ustrezni razred.

Točnost modela *SI-NLI-celotna* na množici SI-NLI v prejšnjem sklopu je več kot 5 % manjša od točnosti modela *ESNLIsi-4k* na ESNLIsi. Oba modela sta bila učena na podobni količini primerov, zato razlika ni posledica razlike v velikosti učnih množic. Prav tako ni posledica napak, ki bi jih povzročilo strojno prevajanje množice ESNLI iz angleščine, saj bi to kvečjemu zmanjšalo točnost na ESNLIsi. Množica SI-NLI vsebuje povedi, pridobljene iz različnih vrst besedil, medtem ko povedi v ESNLI temeljijo na opisih slik. SI-NLI je bolj raznolika, hkrati pa so primeri daljši in pogosto bolj abstraktni. Iz primerjave primerov vidimo, da so primeri logičnega sklepanja v množici ESNLIsi dejansko lažji kot v SI-NLI. Napovedovanje na bolj raznolikih in abstraktnih primerih je za model SloBERTa težje.

Povečanje števila primerov s 4 na 40 tisoč poveča klasifikacijsko točnost za nekaj več kot 5 %, dodatno povečanje na 50 tisoč pa za manj kot pol odstotka. Količina 50 tisoč primerov je za obravnavan par modela in podatkovne množice zadostna. S prevajanjem dodatnih primerov iz angleščine se sposobnost napovedovanja z uporabo

prilagoditve modela SloBERTe ne bi bistveno povečala, zato menimo, da nadaljnje prevajanje iste podatkovne množice ni smiselno.

5.2.2 Učenje s prenosom za napovedovanje na SI-NLI

V Tabeli 9 podajamo klasifikacijske točnosti in tri ostale metrike, izračunane na testni množici SI-NLI za štiri modele tega sklopa. Pri prvih treh modelih gre v tem primeru za prenos znanja. Točnost in ocena F_1 se večata z večanjem števila učnih primerov. Vse metrike so daleč največje za model *ESNLIsi-SI-NLI*, ki je edini še prilagojen na množici SI-NLI.

Najpogostejša napaka modelov je napačna napoved nevtralnega razreda namesto drugih dveh razredov. Model, naučen na večji učni množici, dela takšnih napak manj. Struktura napak modela *ESNLIsi-SI-NLI* je podobna strukturi napak modelov prejšnjega sklopa, učenih le na SI-NLI.

Vidimo, da ima tu večanje števila učnih primerov večji vpliv kot pri vrednotenju, ko učna in testna množica pripadata isti podatkovni množici. Iz razlike med točnostjo in oceno F_1 modelov *ESNLIsi-celotna* in *ESNLIsi-40k* lahko sklepamo, da bi tu s prevajanjem dodatnih primerov množice ESNLI lahko rezultate še nekoliko izboljšali.

Tabela 9: *Metrike v %, izračunane na napovedih za testno množico SI-NLI*

Model	Točnost	F_1	Natančnost	Priklic
ESNLIsi-celotna	65,4	65,2	67,1	65,2
ESNLIsi-40k	64,0	63,8	67,6	64,1
ESNLIsi-4k	55,9	55,5	61,6	56,6
ESNLIsi-SI-NLI	**75,3**	**75,3**	**75,3**	**75,4**

Opomba. Za tri modele, učene na ESNLIsi, in model, ki je bil na koncu prilagojen še na SI-NLI.

Kljub temu, da se je model *ESNLIsi-celotna* učil na več kot 10-krat večji množici kot *SI-NLI-celotna* iz prejšnjega sklopa, sta zaradi učenja s prenosom tako točnost kot ocena F_1 na množici SI-NLI manjši za skoraj 10 %.

Še slabša sta rezultata modelov *SI-NLI-celotna* in *ESNLIsi-4k*, ki sta bila učena na nekaj tisoč primerih, na tuji množici. Čeprav je množica SI-NLI bolj raznolika in težja za napovedovanje, je prenos znanja modela, učenega na njej, celo nekoliko slabši. To je morda posledica

prevajalskih napak v testni množici ESNLIsi, ki so lahko vzrok za napačne klasifikacije in posledično slabše rezultate modela *SI-NLI-celotna*, vrednotenega na njej.

Iz teh ugotovitev lahko zaključimo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja v naravnem jeziku relativno slab. Izboljšuje se z večanjem števila učnih primerov. Tudi z za red velikosti večjim številom učnih primerov pa ne dosegamo rezultatov modelov, učenih na istem tipu primerov, kot jih vsebuje testna množica. Model SloBERTa torej relativno slabo posplošuje z ene podatkovne množice na drugo, oziroma s problema logičnega sklepanja na povedih enega izvora na povedi drugega izvora.

Z uporabo množice ESNLIsi smo kljub temu uspeli izboljšati rezultat modela SloBERTa. Če smo to množico uporabili za vnaprejšnje učenje, nato pa model prilagodili še na SI-NLI, sta točnost in ocena F_1 za približno odstotek večja kot brez njene uporabe. Tudi ta model pa ne dosega najboljšega rezultata, objavljenega na SloBench (točnost 77,2 %) (CJVT UL, 2023), ki je sicer učen na nekoliko večji množici, ki vsebuje tudi našo testno množico. Za izboljšanja napovednih modelov za množico SI-NLI oziroma za logično sklepanje v slovenščini na splošno z uporabo učenja s prenosom bi bilo tako smiselno nadaljnje prevajanje množic, ki so v primerjavi z ESNLI po izvoru povedi bolj raznolike.

Kljub vsemu je rezultat učenja na 50 tisoč prevodih klasifikator, ki na novi, povsem drugače sestavljeni množici, dve tretjini primerov pravilno uvrsti. Ta pristop je enostavno posplošiti na druge jezike z malo viri, v katerih podatkovne množice primerov logičnega sklepanja ne obstajajo.

5.3 Generiranje razlag s SloT5

Vrednotili smo razlage, generirane za prvih 50 primerov testne množice ESNLIsi, generirane s tremi modeli: modelom *t5-large-4k*, dobljenim s prilagoditvijo večjega modela SloT5 (t5-sl-large) na podmnožici učne množice ESNLIsi s 4 tisoč primeri; modelom *t5-small-4k*, dobljenim s prilagajanjem manjšega modela SloT5 (t5-sl-small) na isti množici; in modelom *t5-small-50k*, dobljenim s prilagajanjem manjšega modela SloT5 na vseh 50 tisoč primerih učne množice ESNLIsi.

Tabela 10: *Rezultati vrednotenja razlag treh modelov SloT5 na množici 50 primerov*

Model	*t5-large-4k*	*t5-small-4k*	*t5-small-50k*
Ustrezne razlage izmed 50	8 (16%)	8 (16%)	**14** (28%)

V Tabeli 10 so navedena števila ustreznih razlag. Kriterij za ustreznost razložimo in utemeljimo v razdelku 4.5.3. Vidimo, da je število ustreznih razlag zelo majhno. Nekoliko večje je le pri učenju na 50 tisoč primerih, a je delež ustreznih razlag še vedno manjši od tretjine. Večji model ni bolj uspešen kot manjši. Vzrok za to je verjetno podučenost večjega modela. Ulčar in Robnik-Šikonja (2023) namreč ugotavljata, da je manjši od modelov SloT5 pri večini testiranih nalog boljši zaradi podučenosti modela t5-sl-large. Zaradi te ugotovitve in naših lastnih rezultatov večjega modela nismo učili na celotni učni množici ESNLIsi, saj bi to zahtevalo prevelik časovni vložek glede na pričakovane rezultate.

Trije primeri iz testne množice ESNLIsi s premiso, hipotezo, podano razlago in razlagami, generiranimi s tremi modeli, so prikazani v Tabeli 11.

Tabela 11: *Trije primeri iz testne množice ESNLIsi (prvi kontradikcija, drugi nevtralen in tretji implikacija) in zanje generirane razlage treh modelov SloT5*

Premisa	Dve osebi hodita po ulicah.	
Hipoteza	Dve osebi sedita notri.	
Razlaga	Ljudje ali hodijo ali sedijo. Lokacija je zunaj na ulicah ali znotraj.	
t5-large-4k	Ljudje so bodisi na ulicah ali v zaprtih prostorih.	✓
t5-small-4k	Ljudje ne morejo hoditi in hoditi hkrati.	×
t5-small-50k	Ljudje ne morejo hoditi in sedeti hkrati.	✓
Premisa	V ospredju stoji semafor z dvema simboloma kolesa, zeleno puščico, ki zavija levo, in rdečim signalom za naprej, medtem ko moški v belem klobuku in modri srajci kolesari po rdeči poti.	
Hipoteza	Ta moški je športnik.	
Razlaga	Ne morete sklepati, da je športnik.	
t5-large-4k	Ni vsak moški športnik.	✓
t5-small-4k	Človek ne more biti športnik.	×
t5-small-50k	Človek ne more biti hkrati športnik in športnik.	×
Premisa	Kuhar srednjih let je v belih vratih mobilne kuhinje.	
Hipoteza	V bližini kuhinje je kuhar.	

Razlaga	Kuhar na vratih kuhinje je blizu kuhinje.	
t5-large-4k	Kuhar je v belih vratih mobilne kuhinje.	×
t5-small-4k	Če je kuhar, potem je v bližini kuhinje.	×
t5-small-50k	Kuhar srednjih let je kuhar.	×

Opomba. Ustrezne razlage so označene s ✓, napačne pa z ×.

5.3.1 Kvalitativna ocena razlag

Ob pregledu generiranih razlag nismo opazili očitnih značilnosti, ki bi razlage modelov medsebojno razlikovale. Kvalitativno so si zelo podobne, zato podajamo splošna opažanja.

Razlage so slovnično pravilne in modeli SloT5 se dobro naučijo forme razlag. V učni množici se pogosto pojavljajo razlage določenih oblik, npr. *X ne more Y in Z hkrati.* ali *X ne more biti Y* za kontradikcijo, *Ni vsak X Y* za nevtralne ali *Če je X, potem je Y* za implikacijo. Modeli si te oblike ali predloge zapomnijo in jih uporabljajo, a jih izbirajo na videz naključno. Pogosto je glede na oznako primera napačna že sama izbira oblike. To lahko vidimo pri drugi in tretji razlagi za drugi primer (Tabela 11).

Modeli se naučijo tudi tega, da morajo v razlagah uporabiti besede, ki se pojavijo v premisi in hipotezi. To pogosto privede do povsem nesmiselnih povedi, kot je druga razlaga za prvi primer in tretja razlaga za drugi primer v Tabeli 11. V nekaterih primerih model tako ustvari resnično poved, a ni ustrezna kot razlaga, takšna je tretja razlaga za tretji primer (Tabela 11).

Sklepamo, da modeli SloT5 nimajo zadostnega dejanskega poznavanja sveta, da bi ga lahko uporabili za generiranje razlag pri logičnem sklepanju. Naučijo se zgolj forme, niso pa zmožni pisati pomensko ustreznih razlag. To kaže na to, da so zmožni iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni povezano s poznavanjem resničnosti.

Poizkus generiranja razlag z modeli SloT5 ocenjujemo kot neuspešen.

McCoy idr. (McCoy idr., 2019) domnevajo, da jezikovni modeli za reševanja problema logičnega sklepanja uporabljajo različne jezikovne hevristike, ki temeljijo na vsebovanosti delov hipoteze v premisi. Uporaba takšnih hevristik lahko razloži nezmožnost generiranja ustreznih razlag naših modelov, saj gre pri njihovi uporabi le za jezikovno analizo in ne za dejansko razumevanje.

Uporaba podobnih hevristik namesto dejanskega razumevanja sveta je zato verjetno razlaga za slab prenos znanja in slabo posploševanje ter drugačno strukturo napak klasifikatorjev glede na človeške, ki smo jih omenjali v prejšnjih dveh sklopih. Njihov pristop k reševanju zastavljenega problema po tej domnevi temelji na procesiranju jezika namesto na poznavanju zakonitosti resničnega sveta in zdravorazumskega sklepanja, kot to počnemo ljudje. Naučene hevristike na eni podatkovni množici morda ne delujejo na drugi, če se domena povedi med njima preveč razlikuje.

5.4 Uporaba GPT-3.5-turbo

V tem razdelku podajamo rezultate vrednotenja različnih načinov uporabe modela GPT-3.5-turbo na množici 100 primerov podatkovne množice SI-NLI. Na celotni testni množici SI-NLI je ovrednoten najboljši od pristopov.

5.4.1 Učenje brez dodatnih primerov

Kot lahko vidimo iz prvih štirih vrstic Tabele 12, je za uspešnost pri učenju brez dodatnih primerov pomembna izbira navodila, saj to lahko spremeni točnost napovedi za skoraj 10 %. Navodilo nekoliko vpliva tudi na vrsto napak. Ne glede na izbiro navodila model GPT največ primerov napačno uvrsti kot implikacijo. Najmanjkrat model pravilno klasificira nevtralni razred.

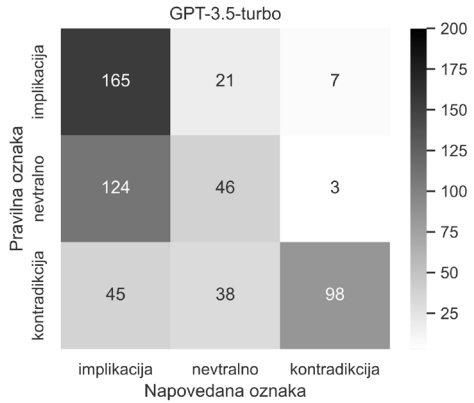
Tabela 12: *Rezultati klasifikacije v % pri uporabi modela GPT-3.5-turbo z učenjem brez dodatnih primerov za različna navodila*

Navodilo	Točnost	F_1	Natančnost	Priklic
navodilo-1-en	**59**	**54,6**	57,3	**54,6**
navodilo-1-si	51	48,7	53,4	47,4
navodilo-2-en	51	47,3	**58,6**	50,5
navodilo-2-si	54	48,4	54,6	48,4
navodilo-razlaga	49	47,2	49,4	47,7
navodilo-1-en angleški	*66*	*55,3*	*59,1*	*58,6*

Opomba. Metoda navodilo-razlaga najprej generira razlago primera, nato pa primer še klasificira (glej razdelek 5.4.3). Pri pristopu v zadnji vrstici tabele so bile napovedi generirane na angleških prevodih primerov, zato pristop ni neposredno primerljiv s prejšnjimi.

Tabela 13: *Rezultati v % pri uporabi modela GPT-3.5-turbo z učenjem brez dodatnih primerov in uporabo navodila-1-en za napovedovanje oznak v celotni testni množici SI-NLI*

Točnost	F_1	Natančnost	Priklic
56,5	54,5	61,3	55,4



Slika 3: Matrike zamenjav za napovedi GPT-3.5-turbo z učenjem brez dodatnih primerov in uporabo navodila-1-en za celotno testno množico SI-NLI.

Ker dosega od štirih testiranih navodil *navodilo-1-en* na množici 100 primerov najvišjo točnost in oceno F_1, smo to vrednotili na celotni testni množici SI-NLI. Metrike so podane v Tabeli 13, matrika zamenjav pa na Sliki 3. Najpogostejša napaka je tako kot pri človeških označevalcih (Slika 1) označitev nevtralnega primera za implikacijo. Model najslabše uvršča nevtralne primere.

Rezultati so očitno slabši kot rezultati modela SloBERTa, prilagojenega na SI-NLI. Vseeno so rezultati minimalno boljši kot pri uporabi učenja s prenosom z nekaj tisoč učnimi primeri (model *ESNLIsi-4k* v Tabeli 9), vendar slabši kot pri učenju s prenosom z nekaj deset tisoč učnimi primeri (model *ESNLIsi-40k*). GPT-3.5-turbo je torej sposoben reševanja problemov s področja logičnega sklepanja v naravnem jeziku, čeprav ni bil učen ali prilagojen za to nalogo. Učenje zelo velikih modelov na veliki količini podatkov s spleta da modelu zadostno razumevanje pomena jezika in razumevanje sveta, da je zmožen relativno uspešno reševati nalogo sklepanja v naravnem jeziku.

Glede na to, da je pri navodilih *navodilo-2-** kot del navodila podana tudi razlaga oznak, napake pri uvrščanju niso posledica nerazumevanja pomena treh NLI oznak. Če primerjamo metrike v Tabeli 12, vidimo, da uporaba angleških prevodov izboljša točnost za 7 %, izboljša pa tudi preostale metrike. Sklepamo lahko, da ima GPT-3.5-turbo nekaj težav z razumevanjem slovenščine. Očitno slovenščino pozna, saj tudi pri uporabi slovenskih izvirnikov doseže primerljive rezultate kot učenje s prenosom slovenskega modela SloBERTa, je pa zaradi majhne zastopanosti slovenščine njeno poznavanje slabše.

5.4.2 Učenje z nekaj dodatnimi primeri

V Tabeli 14 so prikazane metrike za tri različne naključne izbire dodatnih primerov, za primerjavo pa še za isto navodilo brez dodatnih primerov. Vidimo, da ta pristop rahlo poveča točnost, v dveh primerih od treh pa zmanjša oceno F_1. Različne izbire dodatnih primerov metrike spremenijo za nekaj odstotkov. V primerjavi z učenjem brez dodatnih primerov je še večja pristranskost k označevanju primerov kot implikacije, manj primerov pa označi kot nevtralne.

Iz primerjave rezultatov ocenjujemo, da učenje s tremi dodatnimi primeri glede na učenje brez dodatnih primerov bistveno ne izboljša napovedovanja. Brown idr. (2020) so pri testiranju modela GPT-3 na različnih nalogah ugotovili, da učenje z enim dodatnim primerom v povprečju izboljša dosežke modela, še bolj pa ga izboljša učenje s 50 dodatnimi primeri. V prihodnje bi bilo zato smiselno preveriti, ali bi z uporabo večjega števila dodatnih primerov lahko izboljšali rezultate.

Tabela 14: *Rezultati v % pri uporabi modela GPT-3.5-turbo z učenjem z nekaj dodatnimi primeri*

Pristop	Točnost	Ocena F1	Natančnost	Priklic
nekaj-primerov 1	60	53,0	59,3	53,9
nekaj-primerov 2	**63**	**55,2**	**62,7**	**56,9**
nekaj-primerov 3	61	53,3	60,7	54,5
brez in navodilo-1-en	59	54,6	57,3	54,6

Opomba. V prvih treh vrsticah so rezultati za tri različne naključne izbire dodatnih primerov, v zadnji pa so za primerjavo rezultati z istim navodilom brez dodatnih primerov.

5.4.3 Generiranje razlag

Za preverjanje uspešnosti generiranja razlag je model GPT-3.5-turbo z uporabo učenja brez dodatnih primerov z navodilom *navodilo-razlaga*, prirejenem po *navodilo-1-en*, za 100 primerov najprej generiral razlage, nato pa je primere še klasificiral. Ker je razlaga generirana najprej, je zaradi mehanizma pozornosti v arhitekturi transformer vplivala tudi na klasifikacijo.

Klasifikacijske metrike za ta pristop so podane v predzadnji vrstici Tabele 12. Predhodno generiranje razlage napovedovanja ne izboljša. Vidimo, da je točnost manjša kot pri navodilih brez razlag, predvsem zaradi pogostega napačnega uvrščanja primerov drugih dveh razredov kot kontradikcije. Pogostost posameznih vrst napak je povsem spremenjena, kar še dodatno kaže na to, da je uspešnost reševanja problema z učenjem brez dodatnih primerov zelo občutljiva na točno formulacijo navodila in zahtevan pristop reševanja. S preizkušanjem več različnih navodil za ta pristop bi lahko rezultate verjetno izboljšali.

Na podmnožici 50 primerov smo razlage ročno ovrednotili. Za pravilno klasificirane primere je bilo ustreznih 81 % razlag (86 % za pravilno klasificirane primere implikacije, 85 % za primere kontradikcije in 71 % za nevtralne primere; opozarjamo, da je uporabljen vzorec premajhen za zanesljivo primerjavo po razredih).

Sposobnost logičnega sklepanja tako za razliko od manjših modelov, uporabljenih v prejšnjih treh sklopih, pri tem modelu ni pogojena le z uporabo relativno preprostih jezikovnih hevristik, pač pa dejansko daje rezultate, ki kažejo na razumevanje.

Kvalitativno vrednotenje Trije primeri z ustreznimi razlagami so prikazani v Tabeli 15. Razlage so jasne in jezikovno pravilne, nekatere pa so slogovno nerodne (npr. tretji primer v Tabeli 15). Več lahko izvemo iz pregleda napačnih razlag, saj nam to razloži vzroke napak.

Trije primeri s tipičnimi neustreznimi razlagami so prikazani v Tabeli 16. Prvi primer ima razlago, ki je sicer tehnično skoraj pravilna, a neustrezna. Hipoteza je res le parafrazirana premisa in je zato implicirana, vendar bi razlaga morala to bolj jasno ponazoriti (npr. poudariti, da *veliki čezmerni odmerki* pomenijo *prekomerno zaužitje*). Drugi

primer vsebuje napačno razlago, kljub temu pa je klasifikacija pravilna. Model očitno ne ve, da je stoletnica specifična vrsta obletnice, in verjame, da je to vzrok za kontradikcijo. V tretjem primeru sta napačni tako razlaga kot klasifikacija. Model je v hipotezi pri besedni zvezi *samo policisti* obravnaval le drugo besedo in zanemaril prvo, ki pa je ključna za pravilno klasifikacijo.

Napake, podobne tisti v prvem primeru, bi verjetno lahko odpravili z bolj natančnimi navodili ali podajanjem nekaj primerov ustreznih razlag poleg navodil. Te napake so namreč posledica tega, da model nima informacije o tem, kakšna mora biti ustrezna razlaga.

Napaka v drugem primeru kaže na pomanjkljivo razumevanje odnosov med slovenskimi besedami. Pogosta vrsta napake je, da model besedo iz hipoteze in besedo iz premise, ki sta sopomenki ali je ena nadpomenka in ena podpomenka, obravnava kot različni in povedi zato kot kontradiktorni ter na tej osnovi uvrsti primer kot kontradikcijo. To razloži tudi prepogosto uvrščanje primerov v razred kontradikcija v matriki zamenjav.

Iz vrednotenja generiranja razlag lahko zaključimo, da model GPT-3.5-turbo kaže dokaj dobro razumevanje resničnosti, pri čemer pa ga omejuje slabše poznavanje slovenščine. Večino klasifikacij je zmožen ustrezno utemeljiti, po čemer se razlikuje od manjših slovenskih modelov. Pristop bi verjetno lahko še izboljšali s preizkušanjem različnih, morda bolj natančnih navodil, in z uporabo učenja z dodatnimi primeri.

Glede na neuspeh slovenskih modelov SloT5 pri isti nalogi menimo, da je za generiranje razlag pri logičnem sklepanju v slovenščini trenutno najboljši pristop uporaba zelo velikih jezikovnih modelov z več milijardami parametrov, četudi niso bili naučeni specifično za uporabo na slovenščini.

Tabela 15: *Primeri iz množice SI-NLI s pravilnimi razlagami in klasifikacijami, generiranimi z modelom GPT-3.5-turbo*

Oznaka	*nevtralno*
Premisa	Vstala je in stopila k oknu, s katerega se je videlo na dvorišče.
Hipoteza	Pogledala je skozi okno, da bi videla, kaj je na dvorišču.
Odgovor modela	Iz podanega stavka ni mogoče sklepati, ali je pogledala skozi okno ali ne, zato je izjava nevtralna. – *nevtralno*
Oznaka	*kontradikcija*
Premisa	Pisni izpit ocenijo zunanji ocenjevalci po merilih, ki jih pripravi DPK SM za matematiko.
Hipoteza	Vsak učenec je svoje pisno preverjanje znanja ocenil sam s pomočjo sošolcev, pri ocenjevanju so postavili kar svoje lasnte kriterije.
Odgovor modela	Izjava je v nasprotju s podanim stavkom, saj pravi, da so učenci sami ocenili svoje pisno preverjanje znanja, medtem ko je v podanem stavku navedeno, da pisni izpit ocenijo zunanji ocenjevalci. – *kontradikcija*
Oznaka	*implikacija*
Premisa	Slovenija je še vedno najmanj zadolžena članica držav evrobmočja in se je tej ravni zadolženosti zaradi hitre gospodarske rasti doslej približevala postopoma.
Hipoteza	V Evropski uniji Slovenija še vedno velja za eno izmed držav z najmanj dolgovi, k čemur je pripomogel hiter, a postopen, vzpon gospodarstva.
Odgovor modela	Izjava je v skladu s podanim stavkom, saj oboje trdi, da je Slovenija ena izmed držav z najmanj dolgovi v Evropski uniji, pri čemer je k temu prispevala hitra gospodarska rast. – *implikacija*

Tabela 16: *Primeri iz množice SI-NLI z napačnimi ali neustreznimi razlagami, generiranimi z modelom GPT-3.5-turbo*

Oznaka	*implikacija*
Premisa	Veliki čezmerni odmerki lahko povzročijo močno zaspanost, utrujenost in občutek vročine.
Hipoteza	Prekomerno zaužitje lahko sproži intenzivno utrujenost, zaspanost, celo občutek povišane temperature.
Odgovor modela	Izjava je enaka kot izvirna, zato je izjava potrjena. – *implikacija*
Oznaka	*kontradikcija*
Premisa	To častitljivo obletnico so počastili s številnimi prireditvami in z izdajo obsežnega zbornika.
Hipoteza	Praznovali so stoletnico in ob tem pripravili dva koncerta ter izdali kratko brošuro.
Odgovor modela	Izjava je v nasprotju s podanim stavkom, saj omenja stoletnico, medtem ko je v podanem stavku omenjena obletnica. – *kontradikcija*
Oznaka	*kontradikcija*
Premisa	Kraj nesreče sta si ogledali tudi preiskovalna sodnica in državna tožilka.
Hipoteza	Okoliščine nesreče so si prišli ogledat samo policisti.
Odgovor modela	Iz podanega stavka ni mogoče sklepati ali so si okoliščine nesreče ogledali samo policisti ali ne, saj se o tem ne govori. – *nevtralno*

Opomba. Prvi primer ima razlago, ki je sicer tehnično skoraj pravilna, a neustrezna. Drugi primer vsebuje napačno razlago, kljub temu pa je klasifikacija pravilna. V tretjem primeru sta napačni tako razlaga kot klasifikacija.

6 Diskusija

V tem razdelku združimo najpomembnejše rezultate iz različnih sklopov poskusov, ki so bolj podrobno predstavljeni v prejšnjem razdelku. Dodatno jih interpretiramo in jih postavimo v širši kontekst obstoječega dela.

6.1 Uvrščanje

Za uvrščanje primerov smo preizkusili več pristopov z uporabo modelov SloBERTa in GPT-3.5-turbo. Metrike so predstavljene v Tabeli 17.

Tabela 17: *Metrike v %, izračunane na napovedih za testno množico SI-NLI, za štiri različne pristope*

Model	Točnost	Ocena F1	Natančnost	Priklic
SloBERTa (*SI-NLI-celotna*)	73,2	73,2	73,3	73,2
SloBERTa (*ESNLIsi*)	65,4	65,2	67,1	65,2
SloBERTa (*ESNLIsi SI-NLI*)	**75,3**	**75,3**	**75,3**	**75,4**
GPT-3.5-turbo	56,5	54,5	61,3	55,4

Opomba. V oklepajih so navedene učne množice, pri zadnjem pristopu dodatnega učenja ni bilo.

Model SloBERTa, učen na množici SI-NLI, je dosegel klasifikacijsko točnost 73,2 %, bi pa lahko dosegel boljše rezultate, če bi bila učna množica večja. Izločitev učnih primerov, na katerih se človeški označevalci motijo, zmanjša točnost napovedi, podobno oz. enako kot izločitev enakega števila naključno izbranih primerov. Na primerih, kjer se človeški označevalci motijo, se pogosteje kot pri ostalih motijo tudi jezikovni modeli. Domnevamo, da gre vsaj delno za dvoumne primere, torej takšne, ki jih tudi ljudje razumejo na različne načine; drugačne oznake nekaterih označevalcev torej niso napake, pač pa različne interpretacije istega primera. Posledično bi to informacijo lahko upoštevali pri učenju modelov in morda tako izboljšali rezultate.

Model SloBERTa, učen na prevodih ESNLI, je, kljub znatno večji učni množici, na testni množici SI-NLI dosegel skoraj 10 % manjšo točnost. Ugotavljamo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja relativno slab, izboljšuje pa se z večanjem števila učnih primerov. Kljub temu pa model, učen na strojnih

prevodih, pravilno uvršča skoraj dve tretjini primerov na drugi množici. Ta pristop je enostavno posplošiti na druge jezike z malo viri, v katerih podatkovne množice primerov logičnega sklepanja morda ne obstajajo, zato je pristop, kljub slabšim rezultatom, uporaben.

Najboljši rezultat dosežemo z vnaprejšnjim učenjem modela SloBERTa na množici ESNLI in z nadaljnjo prilagoditvijo na množici SI--NLI; na testni množici SI-NLI je bila dosežena točnost 75,3 %. Angleške prevode podatkovnih množic torej lahko uporabimo za izboljšanje rezultatov na SI-NLI, vendar pa naš rezultat ne dosega najboljšega rezultata, objavljenega na SloBench (točnost 77,2 %) (CJVT UL, 2023).

Z uporabo GPT-3.5-turbo smo dosegli slabši rezultat, tako pri učenju brez dodatnih primerov kot tudi z nekaj dodatnimi primeri, čeprav je model mnogo večji od modela SloBERTa. Občutljiv je na izbiro ukaznega navodila in na izbor primerov pri učenju z nekaj dodatnimi primeri. Z bolj širokim testiranjem različnih navodil bi rezultat najverjetneje lahko izboljšali.

6.2 Generiranje razlag

Razlage smo poskusili generirati z dvema različno velikima modeloma, SloT5 in GPT-3.5-turbo. Poskus generiranja razlag z modelom SloT5 je bil neuspešen. Ustrezne razlage so modeli generirali za manj kot tretjino primerov (28 %). Ugotovili smo, da tudi večji model SloT5 ni boljši od manjšega. Modeli se dobro naučijo zgolj forme, niso pa zmožni ustvariti pomensko smiselnih razlag.

Sklepamo, da sta ta dva slovenska velika jezikovna modela z do nekaj sto milijonov parametrov zmožna iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni v celoti povezano s poznavanjem resničnosti. To je v skladu z ugotovitvami McCoy idr. (2019), da jezikovni modeli za reševanja problema logičnega sklepanja uporabljajo različne jezikovne hevristike.

Uporaba podobnih hevristik namesto dejanskega razumevanja sveta je zato verjetno razlaga za slab prenos znanja, slabo posploševanje, drugačne tipe napak glede na človeške in nezmožnost generiranja pravilnih razlag. Pristop velikih jezikovnih modelov k reševanju zastavljenega problema po tej domnevi temelji na procesiranju jezika

namesto na poznavanju zakonitosti resničnega sveta in zdravorazumskega sklepanja, kot to počnemo ljudje.

Boljše rezultate smo dosegli z modelom GPT-3.5-turbo in učenjem brez dodatnih primerov. Pri pravilno uvrščenih primerih, ki so predstavljali približno polovico, je bilo ustreznih 81 % razlag. Glede na ta rezultat in slab rezultat SloT5 menimo, da je za nadaljnje raziskovanje generiranja razlag pri logičnem sklepanju za slovenščino najbolj smiselna uporaba velikih jezikovnih modelov z več milijardami parametrov, tudi če ti niso bili učeni specifično za slovenščino. Koristil bi tudi veliki model, naučen na zadostnem številu slovenskih besedil.

GPT-3.5-turbo je torej sposoben reševanja problemov s področja logičnega sklepanja v naravnem jeziku, čeprav ni bil učen ali prilagojen za to nalogo. Kaže dokaj dobro razumevanje resničnosti, pri čemer pa ga za našo nalogo omejuje slabše poznavanje slovenščine. Večino pravilnih klasifikacij je zmožen ustrezno utemeljiti, v čemer se razlikuje od manjših slovenskih modelov. Učenje zelo velikih modelov na veliki količini podatkov s spleta da modelom zadostno razumevanje pomena jezika in razumevanje sveta za uspešno reševanje problema logičnega sklepanja in utemeljitev sklepov.

7 Zaključek

Raziskovali smo različne pristope k logičnemu sklepanju v naravnem jeziku za slovenščino. Preizkusili smo več velikih jezikovnih modelov za uvrščanje primerov in generiranje razlag in v slovenščino strojno prevedli ter objavili 50.000 primerov iz angleške podatkovne množice ESNLI.

Ugotovili smo, da je prenos znanja med različnimi množicami primerov logičnega sklepanja relativno slab, izboljšuje pa se z večanjem števila učnih primerov. Najboljši rezultat smo dosegli z vnaprejšnjim učenjem modela SloBERTa na množici ESNLI in nadaljnjo prilagoditvijo na množici SI-NLI. Angleške prevode podatkovnih množic torej lahko uporabimo za izboljšanje rezultatov na slovenski množici SI-NLI.

Poskus generiranja razlag z modelom SloT5 je bil neuspešen, bolj uspešen pa je bil z mnogo večjim GPT-3.5-turbo. Sklepamo, da so slovenski veliki jezikovni modeli z nekaj sto milijoni parametrov zmožni iskanja in uporabe jezikovnih vzorcev, poznavanje jezika pa ni v celoti

povezano s poznavanjem resničnosti. Za nadaljnje raziskovanje generiranja razlag pri logičnem sklepanju za slovenščino predlagamo uporabo velikih jezikovnih modelov z več milijardami parametrov, tudi če niso bili učeni specifično za slovenščino. Koristil bi tudi veliki model, naučen na zadostni množici slovenskih besedil.

Testirane pristope bi lahko še izboljšali. Če bi namesto ali poleg prevodov množice ESNLI za vnaprejšnje učenje modela SloBERTa uporabili prevode množice, ki je bolj raznolika od ESNLI ali pa so primeri v njej po izvoru bolj podobni tistim v SI-NLI, bi verjetno lahko na testni množici SI-NLI dosegli boljše rezultate. Prav tako bi rezultate morda lahko izboljšali z uporabo informacije o dvoumnih primerih oziroma o nestrinjanju označevalcev. Oboje zahteva nadaljnje testiranje.

Največ potenciala za izboljšanje je pri generiranju razlag in uporabi GPT-3.5-turbo. Z bolj širokim testiranjem različnih navodil, tako za klasifikacijo kot za generiranje razlag, bi najverjetneje lahko izboljšali rezultate pri obeh nalogah. Prav tako bi bilo smiselno preizkusiti učenje z več deset ali več sto dodatnimi primeri, ki ponavadi izboljša dosežke modela (Brown idr., 2020). To zlasti velja za generiranje razlag, saj smo tam preizkusili le eno navodilo in učenje brez dodatnih primerov.

V času našega testiranja je bil dostop do modela GPT-4 (OpenAI, 2023a), ki je izboljšana različica modela GPT-3.5-turbo, omejen. Ta model dosega boljše rezultate pri večini nalog. V prihodnje bi bilo smiselno preveriti, če in koliko boljše napovedi in razlage bi lahko dobili z njim.

Tako GPT-4 kot GPT-3.5-turbo sta v lasti podjetja OpenAI in sta dostopna le prek vmesnika tega podjetja, točni podatki o njuni zgradbi pa niso znani. Obstajajo javno dostopni odprti modeli, ki so po zmogljivosti primerljivi ali skoraj primerljivi z njima, kakršen je na primer LLaMa-2 (Touvron idr., 2023). Podobne preizkuse bi lahko izvedli tudi na katerem od teh odprtih modelov.

Zahvala

Delo je podprla Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) iz državnega proračuna preko raziskovalnega programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in projekta PROP – Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (št. J7-3159).

Literatura

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, ..., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems*, *31*.

CJVT UL. (2023). *SloBench – Natural language inference (SI-NLI) leaderboard*. Pridobljeno s https://slobench.cjvt.si/leaderboard/view/9

DeepL Translate API. (2023). Pridobljeno s https://www.deepl.com/pro-api

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (str. 4171–4186).

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Lang. Resour. Eval.*, *55*(2), 551–583.

Fišer, D., Erjavec, T., & Ljubešić, N. (2016). JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *4*(2), 67–99.

Google Prevajalnik. (2023). Pridobljeno s https://translate.google.com/?hl=sl

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2022). *Slovene Natural Language Inference Dataset SI-NLI*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1707

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2024). SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia (str. 14859–14870). ELRA and ICCL. Pridobljeno s https://aclanthology.org/2024.lrec-main.1294.pdf

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 Twelfth Language*

Resources and Evaluation Conference, Marseille, France (str. 3340–3345). European Language Resources Association. Pridobljeno s https://aclanthology.org/2020.lrec-1.409

Kumar, S., & Talukdar, P. (2020). NILE: Natural Language Inference with Faithful Natural Language Explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (str. 8730–8742). Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771

Lebar Bajec, I., Repar, A., Demšar, J., Bajec, Ž., Rizvič, M., Kumperščak, B., & Bajec, M.(2022). *Neural Machine Translation model for Slovene-English language pair RSDO-DS4-NMT 1.2.6*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1736

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. Pridobljeno s file:///C:/Users/student1/Downloads/Evaluating_the_Logical_Reasoning_Ability_of_ChatGP.pdf

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, ..., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. Pridobljeno s https://arxiv.org/pdf/1907.11692

Ljubešić, N., & Erjavec, T. (2011). hrWaC and slWac: compiling web corpora for Croatian and Slovene. *Proceedings of the 14th International Conference on Text, Speech and Dialogue* (str. 395–402). doi: 10.1007/978-3-642-23538-2_50

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). *Written corpus ccKres 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1034

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy (str. 3428–3448). Association for Computational Linguistics. doi: 10.18653/v1/P19-1334

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Müller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.

OpenAI. (2022). *Introducing ChatGPT*. Pridobljeno s https://openai.com/blog/chatgpt

OpenAI. (2023a). GPT-4 Technical Report. Pridobljeno s https://arxiv.org/pdf/2303.08774

OpenAI. (2023b). *Models – OpenAI API*. Pridobljeno s https://platform.openai.com/docs/models/gpt-3-5

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., ..., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. Pridobljeno s https://arxiv.org/abs/2203.02155

Pančur, A., & Erjavec, T. (2020). The siParl corpus of Slovene parliamentary proceedings. *Proceedings of the Second ParlaCLARIN Workshop* (str. 28–34).

Poth, C., Pfeiffer, J., R"uckl'e, A., & Gurevych, I. (2021). What to Pre-Train on? Efficient Intermediate Task Selection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (str. 10585–10605).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., ..., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. Pridobljeno s https://arxiv.org/abs/2307.09288

Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021* (str. 17–20).

Ulčar, M., & Robnik-Šikonja, M. (2023). Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*, 1–12. doi: 10.3389/frai.2023.932519

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as few-shot learner. Pridobljeno s https://arxiv.org/pdf/2104.14690

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., ..., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (str. 38–45).

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. Pridobljeno s https://arxiv.org/pdf/2302.10198

Natural language inference for Slovene

In recent years, large language models have been the most successful approach to natural language processing. An important problem in this field is natural language inference, which requires models to contain relatively broad general knowledge. Moreover, the requirement for models to explain their reasoning can offer additional insights into their functioning. We tested several approaches for natural language inference in Slovene. We used two Slovene large language models, SloBERTa and SloT5, as well as a much larger English model GPT-3.5-turbo. Training data consisted of the Slovene dataset SI-NLI and an additional 50,000 machine-translated samples from the English dataset ESNLI. The SloBERTa model was fine-tuned on both datasets. Fine-tuning it on the SI-NLI dataset achieved a classification accuracy of 73.2% on the SI-NLI test set. Pretraining it on the ESNLI dataset improved its accuracy to 75.3%. We observe that models make different types of errors compared to humans and that they generalize poorly across different datasets.

The SloT5 model was also fine-tuned on ESNLI to generate explanations for natural language inference samples. Less than a third of explanations were appropriate, with the model learning common sentence patterns from the domain and producing semantically meaningless explanations. We assume that the tested Slovene large language models with up to several hundred million parameters are capable of identifying and using language patterns, but their language understanding is not necessarily sufficient to understand reality. When the considerably larger GPT-3.5-turbo was used both for classification and explanation generation, it achieved an accuracy of 56.5% on the SI--NLI test set using zero-shot learning, but with 81% of the explanations being appropriate for the correctly classified samples. In comparison with smaller Slovene models, this model shows a reasonable understanding of reality but is limited by its lower Slovene proficiency.

Keywords: natural language inference, large language models, transformer architecture, SloBERTa, SloT5, GPT-3.5-turbo, ChatGPT, explanations, Slovene, fine-tuning

Korpusne oznake za opis konteksta govornih dogodkov v slovenskih govornih korpusih

Andreja BIZJAK
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Zaradi časovno in finančno zahtevne priprave govornega korpusa je ob zasnovi potreben temeljit razmislek o njegovi sestavi in kategorizaciji beleženih metapodatkov. Raznoliki govorni dogodki, vključeni v nacionalni referenčni korpus, naj bi v čim večji meri odražali raznolikost sodobnega govorjenega jezika. Zanimalo nas bo, na kakšen način kategorizirati oznake za opis konteksta govornih dogodkov, da bi to reprezentativnost dosegli, ne da bi se popolnoma odrekli medsebojni primerljivosti podatkov. Premišljena zasnova nam omogoča, da je ob kasnejših korpusnih nadgradnjah potrebnih čim manj časovno zamudnih prilagoditev oznak. Izvedli bomo primerjalno analizo domačih in tujih govornih korpusov, s katero bomo kritično ovrednotili štiri temeljne kategorije oznak za opis konteksta govorne situacije. Pregledali bomo zasnovo tujih referenčnih govornih korpusov FOLK, BNC2014, ORAL2013, Nizozemskega govornega korpusa in C-ORAL-ROM ter jih primerjali z aktualnim referenčnim korpusom govorjene slovenščine Gos 2.1. Problematizirali bomo izbrane oznake in izpostavili težavnejša mesta, ki bi zahtevala dodatne premisleke in potencialno prekategorizacijo v prihodnje.

Ključne besede: govorni korpusi, zasnova korpusa, govorni dogodki, kategorizacija oznak

Bizjak, A.: Korpusne oznake za opis konteksta govornih dogodkov v slovenskih govornih korpusih. Slovenščina 2.0, 12(1): 54–94.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2024.1.54-94

1 Uvod

Govorni jezikovni viri so pomembni ne le za jezikoslovno raziskovanje sodobnega govorjenega jezika, temveč tudi za področje razvoja govornih tehnologij in usposabljanja modelov strojnega učenja. V vse bolj digitalizirani družbi so nepogrešljivi pri razvoju razpoznave in sinteze govora, analize sentimenta, prevajalnikov ter velikih jezikovnih modelov. Govorni jezikovni viri, med njimi tudi govorni korpusi, poleg multimodalnih posnetkov in zapisa govora vsebujejo metapodatke o posnetkih, govorcih in govornih dogodkih (npr. lokacija in čas snemanja, spol in starost govorca, tip govora in govornega dogodka). Izbor metapodatkov, ki jih beležimo, med drugim omogoča preverjanje uravnoteženosti in reprezentativnosti govornega vira glede na govorce in govorne situacije. Vsebinske odločitve, katere kategorije metapodatkov pri zasnovi korpusa zajeti in kako podrobno jih opisati, so odvisne od vrste gradiva in namena govornega vira. Tako se posledično ob združevanju različnih govornih jezikovnih virov pojavljajo težave, izhajajoče iz vsebinske nezdružljivosti zabeleženih metapodatkov (Verdonik idr., 2022).

Prav govorni korpusi so tisti, ki med drugim omogočajo primerjalne analize različnih jezikovnih rab glede na regionalno ali dialektološko obarvanost, čas in prostor, demografijo govorcev, družbeno-kulturni kontekst itd. Obsežno množico korpusnih podatkov najpogosteje analiziramo s pomočjo konkordančnikov, napredne avtomatske analitične tehnike, kot je rudarjenje besedil, pa nam omogočajo, da iz velikih količin podatkov izluščimo vzorce in informacije, ki niso takoj razvidni. Posamezne tehnike rudarjenja besedil, kot je npr. tematsko modeliranje v sklopu parlamentarnih razprav (Pretnar Žagar idr., 2022), lahko uporabimo za identifikacijo tematik (Chizhik in Sergeyev, 2021), modeliranje argumentacije (Petukhova idr., 2015), za analizo čustvene zaznamovanosti govora (Rheault idr., 2016) ali stališč govorcev (Abercrombie in Batista-Navarro, 2020). Govorni korpusi raziskovalcem omogočajo razumevanje procesiranja naravnega jezika, izvedbo pragmatičnih ali sociolingvističnih analiz gradiva (Gorjanc in Fišer, 2013), kjer je treba zabeležiti čim več podatkov o kontekstu, ter pravorečne, žanrske in leksikološke analize – korpusne raziskave postajajo vse pomembnejše na različnih raziskovalnih področjih (Vintar, 2010). Po Čermáku lahko

jezikovne podatke delimo na zunanje, neobdelane (pred vključitvijo v korpus) in na notranje, obdelane podatke (zdaj že del korpusa), za katere je značilno poenoteno označevanje (Gorjanc in Krek, 2005). Če so oznake za opis govornih situacij nedosledne, premalo natančne in nekoherentne, imajo raziskovalci in drugi uporabniki težave pri uporabi gradiva ali pa jih celo zavedemo k neustreznemu iskanju po korpusu oz. k neustreznim raziskovalnim rezultatom. Temeljit premislek o sestavi govornega korpusa in kategorizaciji govornih dogodkov, ki naj v čim večji meri odražajo najrazličnejše kontekste, v katerih govorci komunicirajo, pa je ključnega pomena tudi zato, ker je priprava govornega korpusa časovno in finančno izjemno zahtevna.

Kategorije, navezujoče se na situacijski kontekst in udeležence, so pri zasnovi govornih korpusov najzahtevnejši tip kategorij. Hkrati namreč odražajo raznolike značilnosti konteksta v širšem pomenu kot tudi značilnosti govorcev, vključenih v točno določeni govorni dogodek. Še zlasti so problematične podkategorije, navezujoče se na tematiko, predmet pogovora itd., saj ni na voljo nobene splošno priznane taksonomije, ki bi jo lahko upoštevali, zaradi česar se zdijo tako rekoč neskončne (Cermák, 2009). Z dodatnim izzivom se soočimo že pri samem poimenovanju kategorije, ki opisuje govorni dogodek, saj so v rabi različni izrazi (tip ali opis govornega dogodka, domena, žanr itd.). Pojem govorni dogodek, uporabljen v korpusu Gos 1.1, izhaja iz sociolingvistike (Verdonik, 2013), natančneje iz dela o etnografiji komunikacije D. Hymesa (Hymes, 1974) in njegove opredelitve govornega dogodka (*speech event*). V korpusu Artur (Verdonik idr., 2023) se kategorija vrsta govornega dogodka nanaša na najvišjo raven delitve govora (npr. na javni, nejavni, brani govor), kar v Gos 2.1 ustreza kategoriji tip govora (npr. javni razvedrilni, nejavni zasebni). Kategoriji, ki je v nekaterih tujih korpusih poimenovana kot domena (*domain*), v Arturju ustreza kategorija opis govornega dogodka (npr. okrogla miza, seja državnega zbora, prosti dialog med dvema sogovornikoma), v korpusi Gos 2.1 pa tip govornega dogodka (npr. intervju, predavanje, pogovor med prijatelji/znanci). V korpusu FOLK je za isti namen uporabljeno poimenovanje *Gattung*, za razlikovanje med zasebnim, javnim in institucionalnim govorom pa kategorija *Interaktionsdomäne*. V korpusu BNC2014 je za opis različnih tipov uporabljeno poimenovanje *conversation type*,

za podrobnejši ročni vsebinski opis govornega dogodka pa *activity description*. V korpusu ORAL2013 je bil izbran izraz *typ komunikační situace*, v Nizozemskem govornem korpusu *component*, v C-ORAL-ROM pa *genre*. V tem prispevku bomo uporabljali poimenovanja iz najaktualnejše različice referenčnega govornega korpusa pri nas, tj. korpusa Gos 2.1. Za taksonomijo na najvišji ravni delitve govora bomo uporabljali kategorijo tip govora (npr. informativno-izobraževalni govor), za poimenovanja različnih govornih dogodkov pa tip govornega dogodka (npr. osnovnošolska učna ura). Za prosti in podrobnejši opis govornega dogodka bomo uporabili kategorijo opis dogodka (npr. splošno predavanje za prvi letnik prevajalstva) ter kategorijo kanal za označevanje načina zajemanja ali reprodukcije zvočnih podatkov.

Zaradi pomanjkanja poglobljenega premisleka o ustreznosti kategorizacije metapodatkov v slovenskih govornih korpusih je cilj tega prispevka kritično oceniti nabor oznak, ki v korpusih opisujejo kontekst govorne situacije. V procesu vključevanja posnetkov iz korpusov Gos 1.1, Gos VideoLectures in Artur v nadgrajeno različico korpusa Gos 2.1 smo zaznali nekaj nedoslednosti in neusklajenosti, predvsem pa manko kakršne koli razprave na to temo. Z namenom kritičnega pretresa izbranih oznak bomo v 2. poglavju najprej pregledali zasnovo petih tujih referenčnih govornih korpusov in korpusa Gos 2.1. V 3. poglavju bomo predstavili uporabljeno metodologijo, v 4. poglavju pa skušali identificirati kritična mesta izbranih oznak v korpusu Gos 2.1 ter analizirati njihovo konsistentnost v primerjavi z gradivom, kasneje dodanim iz korpusov Gos 1.1, Gos VideoLectures in korpusa Artur. Osredotočili se bomo na primerjalno analizo štirih temeljnih kategorij za opis konteksta govornih dogodkov, to so tip govora (*taxonomy*), tip govornega dogodka (*domain*), opis govornega dogodka (*title*) in kanal (*channel*). Primerjali bomo razlike med odločitvami snovalcev domačih in tujih govornih virov ter analizirali, ali so bolj naklonjeni prostim vnosom ali sistematičnemu beleženju podatkov z uporabo vnaprej opredeljenih odgovorov. V 5. poglavju bomo predstavljene rezultate kritično ovrednotili ter podali potencialne razrešitve identificiranih šibkih mest, ki bi jih bilo pri zasnovi oz. nadgradnji govornih korpusov za slovenščino smiselno upoštevati v prihodnje. Glavne ugotovitve bomo povzeli v 6. poglavju.

2 Pregled področja

Na področju kategorizacije oznak za opis konteksta govornih dogodkov je bila izvedena raziskava (Kopřivová idr., 2019), ki ponuja celovit pregled najrelevantnejših kriterijev, ki jih je pri zasnovi govornega korpusa smiselno upoštevati. Ob upoštevanju različnih teoretičnih pristopov in ob pregledu obstoječih govornih korpusov je bil pripravljen pragmatično utemeljen nabor govornih dogodkov, ki si ne prizadeva biti univerzalen ali dokončen, temveč služi kot vodilo in konceptualni okvir za spodbujanje upoštevanja raznolikosti govornih dogodkov. V ospredju so kriteriji, o katerih lahko sklepamo neposredno iz situacijskega konteksta govornih dogodkov, ne da bi bilo treba za podatke zaprositi udeležence. Dvoumni izrazi, kot je npr. „spontano“, s pomenom neformalno ali nepripravljeno, niso bili vključeni. Kriterije za kategorizacijo govornih dogodkov so avtorji raziskave (Kopřivová idr., 2019) razvrstili glede na kategorijo, povezano z lokacijo snemanja (*setting-related criteria*), in kategorijo, povezano z udeleženci. V prvo sodijo kriteriji, kot so stopnja uradnosti, ki je odvisna od družbene vloge govorca in tega, ali govor poteka v imenu institucije/na slovesnosti (npr. ravnateljev govor na začetku šolskega leta, rojstnodnevna zdravica). Naslednji kriteriji so stopnja javnosti, ki je odvisna od velikosti občinstva in dostopnosti za ožjo ali širšo javnost (npr. politična razprava na TV vs. usposabljanje na delovnem mestu vs. pogovor z odvetnikom); posredovanje komunikacije (govorci so lahko fizično prisotni na istem mestu ali pa govorijo na daljavo preko telefona ali komunikacijskih platform, kot je Skype) ter sinhronost, ki je odvisna od tega, ali pogovor poteka sinhrono v istem času za oba govorca ali pa je čas produkcije govora različen od časa njegove percepcije (npr. posnetki na TV in spletu, ki ne potekajo v živo).

Med kriterije, povezane z udeleženci, so uvrščeni število (aktivnih) govorcev (monolog vs. dialog); stopnja pripravljenosti, ki je odvisna od tega, ali govorec vnaprej pozna tematiko in namen pogovora ali pa je nepripravljen ter odgovore tvori sproti; število naslovnikov (eden ali več); stopnja naslovnikove aktivnosti (dialog vs. občinstvo pri snemanju oddaje v živo vs. gledalci pred televizijskim zaslonom); odnos med udeleženci, ki je odvisen od stopnje skupnega znanja in izkušenj (npr.

razlikovanje med prijatelji, strokovnimi sodelavci in neznanci); stopnja medsebojnega poznavanja ali zaupnosti (družinski člani in prijatelji vs. udeleženci razgovora za zaposlitev); simetrija družbenih vlog, ki je odvisna npr. od starosti in izobrazbe (pogovor med prijatelji iste starosti vs. pogovor med nadrejenim in podrejenim) in socialno-demografske značilnosti (npr. spol, starost, izobrazba, kraj bivanja, poklic itd.).

Da bi s pomočjo primerjalne analize kritično ocenili nabor oznak, ki v slovenskih govornih korpusih opisujejo kontekst govorne situacije, bomo v nadaljevanju najprej pregledali zasnovo petih tujih referenčnih govornih korpusov. Ker zaradi časovnih omejitev ne moremo analizirati vseh, smo izbrali prepoznavnejše in široko uporabljene referenčne govorne korpuse, ki niso domensko omejeni in torej pokrivajo širok spekter govorjenega jezika, hkrati pa so relevantni po svojem obsegu in vplivu, ki so ga po izdaji imeli na oblikovanje drugih korpusov. Eden od kriterijev za izbor je bila tudi dostopnost relevantne znanstvene literature o zasnovi govornega korpusa in njegovih oznakah. Za referenčne korpuse je namreč metodologija gradnje, ki predvideva celo mrežo kriterijev, natančneje izdelana (Gorjanc, 2005). Ker je »kakovost korpusa določena z avtentičnostjo besedil« (prav tam), smo v analizo vključili čim bolj avtentičen spontani govor. Izbrali smo evropske govorne korpuse, ki so nam jezikovnokulturno in znanstvenoraziskovalno blizu, kar nam omogoča bolj neposredno primerjavo in potencialno implementacijo primerov dobrih praks. V analizo smo želeli zajeti govorjeni jezik različnih jezikovnih skupin, da bi v čim večji meri izključili možnost vpliva le-te na označevanje konteksta govornih dogodkov. Analizirali bomo korpuse FOLK, BNC2014, ORAL2013, Nizozemski govorni korpus in C-ORAL-ROM. Slednji je še zlasti ustrezen za kontrastivne študije, saj gre za primerljivi korpus (Gorjanc in Fišer, 2013), ki vključuje primerljiva besedila v štirih romanskih jezikih.

2.1 Korpus FOLK

Korpus FOLK (Forschungs- und Lehrkorpus gesprochenes Deutsch)[1] je korpus pogovorov v nemškem jeziku, obsegajoč širok nabor interakcij v zasebnem, institucionalnem in javnem okolju (Schmidt, 2014).

1 https://agd.ids-mannheim.de/folk.shtml

Opazujemo in beležimo lahko raznolike značilnosti govornih dogodkov, pri čemer za opredelitev določenega govornega dogodka vse seveda niso enako relevantne. Deppermann in Hartung (2012) podajata pregled značilnosti, ki so bile v začetnih fazah snovanja korpusa FOLK del širšega premisleka, posamezne značilnosti pa nato vključene tudi v samo izgradnjo korpusa. Med analizirane značilnosti v fazi načrtovanja sodijo tip govornega dogodka (*Gattung*) in družbeni sektorji, v katere so ti razvrščeni (npr. izobraževanje, gospodarstvo, medicina, prosti čas, religija, politika, umetnost in mediji). Na govorne dogodke vplivajo tudi družbene vloge govorcev v smislu pravic in obveznosti, ki so konstitutivne za njihov vstop v zasebne ali institucionalne interakcije in ki se ne tvorijo šele skozi pogovor (npr. mati, otrok v pogovorih za družinsko mizo; pogovor v krogu ožjih prijateljev, sodnik/obtoženec v sodnih postopkih). Nadaljnja analizirana značilnost je, da so govorni dogodki za udeležence različno dostopni. Razlikujemo zaprte govorne dogodke (intimne/sorodstvene vezi ali pogojevanje dostopa s specifičnimi institucionalnimi vlogami: npr. šepet v postelji, posvet zaprtega tipa), (delno) javne dogodke (npr. sestanek fakultete) in javne govorne dogodke (govor v politični kampanji). Pomembna je tudi zaupnost med udeleženci (neznanci ob prvem stiku, znanci, zaupni odnosi med prijatelji in družino ali mešani odnosi).

Kriterij institucionalnosti govorne dogodke deli glede na to, ali so vloge udeležencev v pogovoru vezane na institucije (npr. svetovalec in svetovanec v svetovalnem razgovoru). Stopnja priprave loči spontane govorne dogodke brez priprave (pogovor na zabavi), vnaprej pripravljene (razgovor za službo), vnaprej oblikovane govorne dogodke (molitev) in glasno branje (novice). Govorni dogodek je lahko vezan na tematiko, kar je odvisno od vrste govornega dogodka in ne konkretnega primera (npr. tematsko fiksirani govorni dogodki, kot sta televizijska politična razprava in sodna obravnava; dogodki, vezani na določeno tematsko področje, kot sta pogovor med zdravnikom in pacientom ali pogovorna oddaja; in nefiksirani govorni dogodki, kot sta pogovor za družinsko mizo in pogovor v gostilni). Opredeljen je lahko tudi namen pogovora (npr. postavitev diagnoze, psihološko svetovanje, pritoževanje znotraj družinske razprave). Pri določitvi opazovanih metapodatkov je treba biti pozoren na število udeležencev – diadni

pogovor (npr. psihoanaliza), triadni (mediacijska pogajanja) ali večosebni pogovor (npr. skupinska terapija), glede na menjavo govorca pa razlikujemo monolog in dialog. Smiselno je upoštevati prisotnost oz. odsotnost občinstva in njihovo vlogo (npr. pogovori z udeleženci, ki primarno niso verbalno aktivni – primer pogovorne oddaje, kjer je občinstvo razpršeno: razlikovanje med gosti pogovorne oddaje, gosti v studiu in televizijskimi gledalci).

Ena od opazovanih značilnosti je lahko tudi medij oz. tehnični vidik prenosa (npr. pogovor iz oči v oči, telefonski pogovor, posredovano preko množičnih medijev). Pri pogovornih oddajah lahko opazujemo notranji komunikacijski krog, kjer bi kot kanal opredelili osebni stik, in zunanji komunikacijski krog, kjer gre za prenos preko množičnih medijev. Naslednji značilnosti sta kraj, ki podrobneje opredeljuje vrsto govornega dogodka (npr. pridiga v cerkvi, družinski pogovor v zasebnem domu, sodna obravnava na sodišču), in čas (npr. pogovor pri zajtrku, ponedeljkov jutranji krožek v osnovni šoli, novoletni govor predsednika). Dodatna značilnost, vezana na čas, je časovna omejenost, ki razlikuje omejeno trajanje govornega dogodka od neomejenega oz. od trajanja, ki ni vnaprej določeno (institucionalni in javni pogovori so npr. skoraj vedno časovno omejeni). Kot zadnjo značilnost obravnavata praktično referenco/sklicevanje (*empraktischer Bezug*), ki opredeljuje govorne dogodke, pri katerih govorjenje ni v središču, ampak ima le dopolnilno oz. koordinacijsko vlogo (Deppermann in Hartung, 2012).

Pri sprejemanju odločitev o tem, katere metapodatke bomo v govornem korpusu dejansko beležili, je treba upoštevati splošno sprejeto soglasje na področju korpusnega jezikoslovja, in sicer da reprezentativnost v statističnem smislu ne more biti cilj zasnove korpusa (Lemnitzer in Zinsmeister v Deppermann in Hartung, 2012). Prvi izziv pri pripravi nacionalnega govornega korpusa, kot navajata Deppermann in Hartung (2012), se pojavi pri sami opredelitvi osnovne populacije, ki naj bi jo vzorčili. Kvantitativno sestavo jezikovne resničnosti je glede na številne parametre izjemno zahtevno smiselno oceniti, npr. koliko voznikov med vožnjo preklinja in kako pogosto, koliko jih poje in kako pogosto. Poleg tega se pri vsakodnevnem sporazumevanju nenehno pojavljajo novi govorni dogodki, kot je npr. hiphop dvoboj, obstoječi pa se preoblikujejo ali izginjajo. Nejasno je tudi, po katerih kriterijih naj

bi posamezne govorne dogodke utežili, npr. dialog vs. skupinska debata; pogovor v živo vs. komuniciranje preko množičnih medijev, kjer ima peščica posameznikov aktivno, množica pa pasivno vlogo. Eden od nadaljnjih izzivov, identificiranih ob pripravi korpusa FOLK (Deppermann in Hartung, 2012), je, da ne obstaja celovit seznam vseh govornih dogodkov, ki se pojavljajo v komuniciranju, saj je znanje o tem razpršeno, velikokrat intuitivno in nikjer v celoti popisano. Zato je iluzorno pričakovati, da bi lahko v korpus vključili vse mogoče kombinacije tako govornih dogodkov kot vseh kategorij empiričnih metapodatkov o govorcih, kot so spol, starost itd. Spodaj navajamo pregled ključnih parametrov, ki so bili vključeni v zasnovo korpusa FOLK (Kaiser, 2018).

Tabela 1: *Stratifikacijski parametri v korpusu FOLK*

	Parameter	Vrednosti
Vodilna stratifikacija	Interakcijska domena	zasebno institucionalno javno drugo
	Življenjsko področje (*Lebensbereich*)	**zasebno**: zasebno (ni določeno) **institucionalno**: izobraževanje, organi/uprava, medpoklicno sporazumevanje, klubske/društvene dejavnosti (*Vereinsleben*), religija/cerkev, kultura (umetnost/zabava/šport), storitvene dejavnosti, medicina/zdravje **javno**: politika, zabava, znanost, gospodarstvo **drugo**: drugo (ni določeno)
	Aktivnosti	**zasebno**: usmerjeno v aktivnosti[2]: odprt vnos (npr. obnavljanje, načrtovanje dopusta); neusmerjeno v aktivnosti **institucionalno**: usmerjeno v aktivnosti: odprt vnos (npr. sestanek, učna ura vožnje) **javno**: usmerjeno v aktivnosti: odprt vnos (npr. mediacija, panelna razprava) **drugo**: usmerjeno v aktivnosti: *maptask*, intervju, eksperimentalna igra

2 Ta oznaka pomeni, da potek aktivnosti določajo npr. neke vnaprej znane tematike ali specifični tipi nalog.

	Parameter	Vrednosti
Dodatno	Medij	iz oči v oči telefon posredovano preko množičnih medijev
	Število udeležencev	natančna specifikacija delitev na interakcijo z dvema, tremi ali več osebami
	Občinstvo	da ne odprto polje
	Zaupnost	poznan neznan zaupen drugo
	Družbene vloge in odnosi	odprt vnos
	Praktična referenca	da ne neznano/mešano
	Jezik(i) interakcije	odprt vnos

Najpogostejše oznake v korpusu FOLK, ki so bile preko prostega vnosa zabeležene za opis tipa govornega dogodka na področju zasebne interakcije, so pogovor ob kavi, pogovor med partnerjema, pogovor v družini, pogovor med prijatelji, pogovor med študenti, pogovor na počitnicah/izletu, pogovor med gospodinjenjem, odrasli igrajo družabne igre, igranje iger z otroki in branje otrokom. Na področju interakcij v šoli/na univerzi gre za učne ure na zasebni srednji šoli, učne ure na poklicni šoli, ustne izpite na univerzi ter povratne informacije med učitelji. Najpogostejši govorni dogodki na področju interakcije na delovnem mestu so sestanek, menjava izmene v bolnišnici, usposabljanje v dobrodelni organizaciji in pogovor na policijski postaji, na področju javne interakcije pa mediacijski pogovori. Drugi tipi interakcije so še navigacijske naloge (*maptask*) in biografski intervjuji (Schmidt, 2014).

2.2 Britanski nacionalni korpus

Na pobudo britanske vlade konec 80. in v začetku 90. let prejšnjega stoletja je v sodelovanju akademskih institucij z gospodarstvom nastal

obsežen Britanski nacionalni korpus (BNC)[3], katerega 10-odstotni delež sestavlja govorjeni jezik, tj. okoli 10 milijonov besed (Zemljarič Miklavčič, 2008), izmed katerih jih približno polovica sodi v spontani govor. BNC je enojezikovni, vzorčni, mešani, sinhroni korpus, ki pa navkljub zavidanja vredni uravnoteženosti predstavlja zgolj zamrznjeno sliko angleščine, kakršna je bila v rabi nekako od sredine 70. do sredine 90. let prejšnjega stoletja (Gorjanc, 2005).

Eden najvplivnejših referenčnih virov za gradnjo govornih korpusov danes je Britanski nacionalni korpus 2014 (BNC2014)[4], katerega govorni podkorpus vključuje več kot 11 milijonov besed. Sestavljen je iz demografsko in kontekstualno uravnoteženega dela (Zemljarič Miklavčič idr., 2015). Ob izgradnji govornega korpusa so snovalci sprejeli odločitev, da bodo zbirali podatke, ki se pojavljajo zgolj v neformalnih kontekstih, saj so ugotovili, da obstaja večja potreba po uporabi in raziskovanju tovrstnih podatkov (Love idr., 2017). Govorni del BNC2014 vsebuje prepise posnetih pogovorov, ki so jih med letoma 2012 in 2016 zbrali predstavniki britanske javnosti. Pogovori so bili posneti v neformalnem okolju (običajno doma) in so potekali med prijatelji in družinskimi člani. Inovativni vidik korpusa je, da so govorci svoje pogovore posneli z vgrajeno napravo za snemanje zvoka v svojih pametnih telefonih. Korpus obsega 1 251 pogovorov, v katerih je sodelovalo skupaj 672 govorcev.[5]

Love in sodelavci (2018) so ugotovili, da obstajajo metapodatki, na osnovi katerih lahko besedila razvrstimo v posamezne kategorije. Takšni metapodatki so število govorcev, snemalno obdobje, leto snemanja in podatki o transkriptorjih. Po drugi strani pa nekaterih metapodatkov, zabeleženih na ravni besedila, ni mogoče uporabiti za kategorizacijo besedil, saj so lahko (in pogosto so) za vsako besedilo različni. Med njimi so opis aktivnosti (možnost prostega vnosa v BNC2014 – snemalci so bili pozvani, naj sami opišejo, kaj se je med snemanjem dogajalo, npr. „par se sprehaja po podeželju in se pogovarja“, njihovi opisi pa so dokumentirani kot dobesedni zapisi); razmerje med govorci (snemalci so na vnaprej

3 http://www.natcorp.ox.ac.uk/
4 http://corpora.lancs.ac.uk/bnc2014/
5 http://corpora.lancs.ac.uk/bnc2014/

določenem seznamu potencialnih razmerij med govorci izbrali eno od možnosti, npr. ožja družina, partnerji, najbližji prijatelji; prijatelji, širši družinski krog; kolegi; znanci); ID-ji govorcev; dolžina posnetka in lokacija posnetka. Nadaljnji metapodatki so še izbrani opisi, značilni za tip pogovora (seznam izbirnih polj, na katerem so sodelujoči lahko izbrali več govornih dejanj, ki so na posnetku, npr. razpravljanje in pojasnjevanje, poizvedovanje, pritoževanje, zahtevanje, vabljenje, svetovanje, razglašanje, pripovedovanje anekdot, dogovarjanje, opravičevanje, kupovanje/prodaja, pripovedovanje šal) in teme (možnost prostega vnosa, kamor so snemalci lahko zapisali vse teme, ki so bile zajete v pogovoru, npr. računalniško programiranje, hrana, vino, savne, odpiranje daril), katerih zapisi so bili dobesedno dokumentirani (Love idr., 2018). Metapodatki o snemanju, ki so jih v celoti izpolnili avtorji posnetkov po svoji lastni presoji, kot na primer posamezne teme pogovora, so, kot rečeno, dobesedno objavljeni v dokumentaciji o korpusu, brez poskusov shematizacije ali standardizacije (Love idr., 2017).

Ker je govor v demografskem delu korpusa pretežno vsakdanji/neformalen, je bil v kontekstualnem delu korpusa namerno dodan nabor jezikoslovno motiviranih besedilnih vrst (Burnard, 2000). Na najvišji ravni tipologije je bila izvedena delitev na štiri enako velike kontekstualno utemeljene kategorije, od katerih je vsaka razdeljena na podkategoriji monolog (v obsegu 40 %) in dialog (v obsegu 60 %):

- **izobraževanje/informiranje**: predavanja, razprave, učne predstavitve, novičarski komentarji, interakcije v razredu
- **poslovanje/gospodarstvo**: pogovori in intervjuji v podjetjih, pogovori s člani sindikata, prodajne predstavitve, poslovni sestanki, posveti (na področju zdravstva, prava, gospodarstva, stroke)
- **javno/institucionalno**: politični govori, pridige, javne razprave, seje sveta, verska srečanja, parlamentarni in sodni postopki
- **prosti čas**: govori, predvajani športni komentarji, pogovori znotraj klubov, predvajane pogovorne oddaje in telefonski pogovori, klubska srečanja

Znotraj vsake podkategorije je bil opredeljen razpon besedilnih vrst, ki ni bil fiksno določen, temveč zasnovan dovolj prožno, da je omogočal vključitev dodatnih besedilnih vrst. Splošen cilj je bil doseči uravnotežen izbor besedilnih vrst ob upoštevanju kategorij, kot so regija, spol in tema. Druge lastnosti, kot je namen, so bile dodane na osnovi naknadnih presoj (Burnard, 2000).

2.3 Korpus ORAL2013

Kot smo v uvodu že navedli, so pri oblikovanju govornih korpusov kategorije, navezujoče se na situacijski kontekst in udeležence, najzapletenejša vrsta kategorij. V Tabeli 2 je prikazanih 12 besedilnih značilnosti oz. atributov, predstavljenih v obliki binarnih opozicij in poskusno razdeljenih v štiri večje skupine. Pri konkretnem posnetku ali besedilu so navedene besedilne značilnosti lahko bodisi prisotne (plus +) bodisi odsotne (minus –), kar pa ne pomeni, da včasih ne obstaja tudi tretja možnost in da je razlikovanje med njimi vedno povsem jasno in enostavno določljivo. Če je vseh 12 kategorij prisotnih (+), naj bi šlo za prototipsko govorjeni jezik, v obratnem primeru za pisna besedila, vendar pogosto prihaja do prekrivanj (Cermák, 2009).

Tabela 2: *Razvrstitev 12 besedilnih značilnosti*

PLUS (+)	**MINUS (–)**
Izvor besedila:	
govorjeno (tj. izvorno)	brano
dialog (tj. izvorno, tipično)	monolog
Medosebni, družbeni odnosi med govorci in fizični kontekst:	
poznavanje/bližina (prijatelji, družina)	govorci si niso blizu
enakost govorcev (družbena, poklicna)	neenakost
zasebno (nejavno)	javno
neformalno	formalno
interaktivno	enosmerno (predavanje, govor, ukazi)
prisotnost (fizična bližina)	oddaljenost (npr. telefon)
eden-na-enega	eden-na-več (občinstvo)
Pristop k temi/situaciji:	
spontano (improvizirano)	pripravljeno (bolj ali manj pripravljeno)

PLUS (+)	MINUS (–)
sproščeno (neformalno)	ustaljeno (*regular*)/uradno (obrednost, protokol)
Zavedanje o snemanju:	
govorec se snemanja ne zaveda	govorec se snemanja zaveda

Korpus ORAL2013[6] je zasnovan kot reprezentativni korpus spontane govorjene češčine, ki se uporablja v neformalnih komunikacijskih situacijah. Vsebuje 291 ur posnetkov in več kot 130 000 besed.

Dejavnike, ki vplivajo na naravo govorjenega jezika, lahko za namen sestave korpusa razdelimo v dve glavni skupini. V prvo skupino uvrščamo predvsem situacijske dejavnike, ki so pomembni za zagotavljanje neformalnosti in so zato merilo za pridobitev posnetkov in njihovo vključitev v korpus. Pri tem imajo ključno vlogo neformalnost komunikacijske situacije ter njena nejavna in neuradna narava; drugi dejavniki vključujejo zlasti zasebno okolje, dialoškost izjav, fizično prisotnost govorcev in njihovo tesno medsebojno povezanost, nepripravljenost in spontanost. Vsi posnetki, vključeni v korpus ORAL2013, izpolnjujejo navedena merila, kar zagotavlja kar največjo prototipičnost zajetih govorjenih posnetkov. Najpogostejši govorni dogodki v korpusu so obisk, pogovor doma, v restavraciji, med skupno dejavnostjo itd. Podatkov iz telefonskih intervjujev, komunikacije prek Skypa ali drugih podobnih situacij namenoma niso zbirali. Glavni cilj je bil ohraniti čim večjo avtentičnost govorcev in njihovih govorov, kar je seveda pogojeno z naravnim okoljem. Zato govorci o snemanju običajno niso bili obveščeni vnaprej, temveč šele po končanem snemanju (Lucie idr., 2015).

2.4 Nizozemski govorni korpus

Korpus govorjene nizozemščine (Spoken Dutch Corpus)[7] ni bil zgrajen za poseben namen ali v interesu (ene same) točno določene skupine uporabnikov. Obsega okoli 800 ur posnetkov in 9 milijonov besed. Glavni parameter, na katerem je korpus zasnovan, je družbeno-situacijsko okolje, v katerem se jezik uporablja. Na osnovi tega lahko

6 https://wiki.korpus.cz/doku.php/en:cnk:oral2013
7 https://lands.let.ru.nl/cgn/doc_English/topics/project/pro_info.htm

razlikujemo več kategorij, ki jih lahko opredelimo glede na situacijske značilnosti, kot so sporazumevalni cilj, medij, število govorcev ter odnos med govorcem in poslušalcem (Oostdijk, 2000).

Tabela 3: *Taksonomija Nizozemskega govornega korpusa*

dialog/ multilog	zasebno		spontano	neposreden stik	pogovor (iz oči v oči)
					intervju
				na daljavo	telefonski pogovor
					poslovna transakcija/posel
	javno	predvajano	bolj ali manj pripravljeno		intervju in diskusija
		nepredvajano (*non-broadcast*)	spontano		diskusija, debata, sestanek
					predavanje
monolog	zasebno		bolj ali manj pripravljeno		opis slike
	javno		predvajano	spontano	spontani komentar
				pripravljeno	novice, programi o aktualnem dogajanju
					informativni bilten
					komentar
			nepredvajano	pripravljeno	predavanje, govor
					brano besedilo

Pri določanju obsega posameznih kategorij korpusa so snovalci korpusa upoštevali več vidikov. Zaradi velike potrebe po podatkih o spontanem govorjenem jeziku je bilo zbiranje naklonjeno nepripravljenemu, spontanemu govoru. Ker interakcija velja za tipično značilnost govorjene komunikacije, so bili obširneje zastopani dialogi in multilogi. Da bi zajeli razlike med posameznimi jezikovnimi zvrstmi v smislu številčnosti njihovih variacij, so bolj heterogene zvrsti v korpusu na splošno zastopane z večjim številom vzorcev kot bolj homogene. O dolžini posnetkov, ki se razlikuje od kategorije do kategorije, zaradi česar je nemogoče napovedati optimalno dolžino posnetka, so se pogosto odločali na osnovi intuitivne presoje. Ob tem ne gre zanemariti dejstva, da je posamezne vrste podatkov lažje zbrati kot druge. Da bi zadostili potrebam določenih skupin uporabnikov, je bilo treba za posamezne kategorije zbrati minimalno količino podatkov, kar še posebej velja za kategorije, ki se uporabljajo za razvoj tehnoloških aplikacij, tj. telefonski pogovori in brana besedila (Oostdijk, 2000).

2.5 C-ORAL-ROM

Korpus C-ORAL-ROM[8] uporablja dve različni strategiji vzorčenja, eno za predstavitev govorjenega jezika v formalnem kontekstu in drugo za neformalni del. Žanr (oz. področje rabe) je z zaprtim spustnim seznamom strogo opredeljen samo pri formalnem govoru, kar pa ne velja za neformalni govor, kjer je omogočen prosti vnos (Cresti idr., 2004).

Tabela 4: *Parametri za opis govorjenega jezika v shematski zasnovi korpusa C-ORAL-ROM*

Kategorija *(type)*	Pod-kategorija	Pod-pod-kategorija
neformalno	zasebno	dialog in multilog
neformalno	zasebno	monolog
neformalno	javno	dialog in multilog
neformalno	javno	monolog
formalno	naravni kontekst (osebni stik)	politični govor, politična razprava; pridiga; poučevanje; strokovna razprava; konferenca; gospodarstvo; pravo
formalno	mediji	pogovorna oddaja; znanstveni tisk; reportaža; intervju; šport; novice; vremenska napoved
	telefon	zasebni pogovor; interakcija človek–stroj

V nasprotju s formalnimi situacijami je za množico situacij, v katerih se uporablja neformalni jezik, značilno, da je odprta in da noben kontekst zanjo ni bolj tipičen od drugega. Zato je bila pri zasnovi korpusa C-ORAL-ROM sprejeta odločitev, da žanri in področja rabe niso vnaprej izrecno opredeljeni. Na ta način je bila na teoretični ravni zagotovljena možnost, da je v korpus vključen kateri koli pomemben žanr (oz. tip govornega dogodka) oz. da noben žanr ni bil obsojen na nično verjetnost pojavitve (Cresti in Moneglia, 2005).

Vzorčenje štirih romanskih jezikovnih virov za korpus C-ORAL-ROM je temeljilo na nizu spremenljivih parametrov, ki tvorijo semiološko in sociološko strukturo korpusa spontanega govora. Ti so dialoška struktura (monologi, dialogi, multilogi); družbeni kontekst rabe, saj so posnetki pogovorov znotraj družinskega in zasebnega življenja ločeni od tistih, ki potekajo v javnosti; kanal (interakcije iz oči v oči, posnetki medijske produkcije in telefonski posnetki), področje rabe,

8 http://www.elda.org/en/proj/coralrom.html

pri čemer gre za področje družbenega okolja, aktivnosti in poklicev, kot so pravo, gospodarstvo, raziskovanje, poučevanje, cerkev itd. Opazovani parametri so še register, saj so posnetki s standardnim jezikom ločeni od posnetkov, na katerih prevladuje neformalna, nestandardna raba jezika, in metapodatki o govorcu, preko katerih se beležijo glavne sociolingvistične značilnosti govorcev, kot so starost, spol, izobrazba, poklic in geografski izvor. Formalni register opredeljujejo dejavniki, kot so javna raba govora, profesionalna raba govora v skladu z družbenimi vlogami, ki jih ima govorec v skupnosti, in namera o izvedbi govorjenega besedila, ki predvideva obravnavo določene tematike, argumentacijo, zaključke itd. Za formalni govor velja, da je lahko tudi spontan, če ta ni (delno) vnaprej pripravljen (Cresti in Moneglia, 2005).

2.6 GOS 2.1

Prvi referenčni govorni korpus za slovenščino, tj. korpus Gos, je izšel leta 2011 (Verdonik idr., 2013). Različica Gos 1.1[9] obsega približno 120 ur transkribiranega govora, posnetega v različnih situacijah: radijske in televizijske oddaje, učne ure in predavanja, zasebni pogovori med prijatelji ali v družini, delovni sestanki, posvetovanja, prodajni pogovori itd. Gos 1.1 vsebuje več kot milijon besed. Med letoma 2017 in 2021 je bil kot dodatek h korpusu Gos izdan korpus Gos VideoLectures (Verdonik, 2018). Govor v korpusu Gos VideoLectures 4.2[10] je javni akademski govor, ki obsega 22 ur v obliki 55 javnih predavanj, dostopnih prek spletnega portala Videolectures.net. Artur 1.0[11] je govorna podatkovna zbirka, zasnovana za potrebe avtomatske razpoznave govora za slovenski jezik. Podatkovna baza vsebuje 1067 ur govora, od katerih je transkribiranih 884 ur. Bazo sestavlja 573 ur branega govora, 62 ur javnega govora (npr. medijski posnetki, spletne konference, delavnice, spletna predavanja), 74 ur nejavnega oz. zasebnega govora (npr. monologi, dialogi kot sproščeni pogovori ali prosti govor o različnih temah) ter 201 ura parlamentarnega govora.

Podatki iz korpusa Artur so bili uporabljeni za nadgradnjo slovenskega referenčnega govornega korpusa Gos (Verdonik idr., 2013). V

9 https://www.clarin.si/repository/xmlui/handle/11356/1438
10 https://www.clarin.si/repository/xmlui/handle/11356/1222
11 https://www.clarin.si/repository/xmlui/handle/11356/1776

zadnjo različico Gos 2.1[12] je vključenih 185 ur iz korpusa Artur 1.0, in sicer vsi transkribirani terenski posnetki javnega in nejavnega govora. Da bi bili podatki čim bolj uravnoteženi, je v sklopu parlamentarnega govora v Gos 2.1 vključenih največ 4000 besed na posameznega govorca. Ti izbrani posnetki so bili skupaj z vsem gradivom Gos Videolectures 4.2 in Gos 1.1 združeni v korpus Gos 2.1. Referenčni korpus Gos 2.1 tako obsega približno 300 ur govora in 2,4 milijona besed. Na voljo je v repozitoriju CLARIN.SI in v novem korpusnem konkordančniku[13]. Brani govor iz korpusa Artur, ki ni bil vključen v referenčni govorni korpus Gos 2.1, pomeni dragoceno podatkovno bazo za fonetične in fonemske raziskave v prihodnje (Verdonik, 2021).

3 Metodologija

S ciljem kritično ovrednotiti nabor oznak za opis konteksta govorne situacije v govornih korpusih Gos 1.1, Gos VideoLectures, Artur in Gos 2.1 bomo skušali odgovoriti na dve temeljni raziskovalni vprašanji. Najprej nas bo zanimalo, kako je nabor oznak primerljiv z izbranimi tujimi referenčnimi govornimi viri, nato pa bomo v izbranih kategorijah oznak skušali identificirati kritična mesta, ki bi zahtevala premislek o njihovi potencialni prekategorizaciji v prihodnje. V primerjalni analizi petih referenčnih tujih govornih korpusov, FOLK, BNC2014, ORAL2013, Nizozemski govorni korpus ter C-ORAL-ROM, in govornih korpusov za govorjeno slovenščino se bomo posvetili štirim kategorijam oznak, ki se v največji meri navezujejo na kontekst govorne situacije. To so tip govora, tip govornega dogodka, njegov podrobnejši opis in kanal. Zanimali nas bodo shematska zasnova primerjanih korpusov, vsebinsko-formalne razlike med posameznimi oznakami ter strategija pristopa k zajemu metapodatkov. Ugotavljali bomo, ali so katere od oznak nezadovoljivo zastopane, nekonsistentne, redundantne, dvoumne ali nejasno poimenovane. Nazadnje bomo nabor oznak ovrednotili še v smislu njihove odprtosti za zajem kar najobsežnejše množice različnih govornih dogodkov v primerjavi z rabo zaprtih spustnih seznamov z omejenim številom vnaprej določenih oznak, ki beleženim metapodatkom

12 https://www.clarin.si/repository/xmlui/handle/11356/1863
13 https://viri.cjvt.si/gos/

omogočajo večjo sistematičnost in medsebojno primerljivost.

Pri analizi zajetih metapodatkov o kontekstu govornih dogodkov v korpusih govorjene slovenščine se bomo osredotočili tudi na postopek preslikave oznak, ki so bile opravljene ob pripravi aktualne različice Gos 2.1. Da bi bili podatki medsebojno čim bolj primerljivi in kompatibilni, so bile posamezne oznake iz korpusov Gos 1.1, Gos Videolectures in Artur ob prenosu v korpus Gos 2.1 prekategorizirane, preimenovane ali dodane. V ta namen bomo analizirali več kot 115 avdio posnetkov, pri katerih smo v predhodno opravljeni primerjalni analizi identificirali težavnejša mesta. Na osnovi vsebine avdio posnetkov bomo oblikovali smernice za morebitne prekategorizacije ali dopolnitve oznak, ki jih bomo predlagali v diskusiji. Pri analizi bomo poleg avdio posnetkov, njihovih zapisov in pregleda tuje literature upoštevali tudi razpoložljivo dokumentacijo o zasnovi in metapodatkih obravnavanih govornih korpusov.

4 Rezultati

Najprej si bomo ogledali, kako je nabor oznak v slovenskih govornih korpusih primerljiv z izbranimi tujimi referenčnimi govornimi viri, v nadaljevanju pa bomo skušali identificirati kritična mesta izbranih oznak in analizirati, iz katerih vidikov nabor oznak ustrezno opisuje kontekst govorne situacije, kateri vidiki pa zahtevajo morebitne dopolnitve.

4.1 Primerjava nabora oznak za opis govorne situacije v domačih in tujih govornih korpusih

Ker so se snovalci korpusov BNC2014 in ORAL2013 odločili, da bodo v govorni korpus vključili zgolj posnetke, ki se pojavljajo v neformalnih kontekstih, neposredna primerjava s tipi govora v korpusu Gos 2.1 ni mogoča. Zanimivejše je opažanje, da so nekateri posnetki, ki so v Gos 2.1 označeni kot nejavni nezasebni govor, v korpusu FOLK uvrščeni v kategorijo institucionalno, v korpusu BNC2014 pa v kategorijo javno/institucionalno. V primerjavi korpusa FOLK s korpusom Gos (Kaiser, 2018) je izpostavljeno, da je komuniciranje, ki je v korpusu FOLK opredeljeno kot institucionalno, v Gos ex negativo opredeljeno kot

nejavno nezasebno.[14] Interakcije na področju izobraževanja, ki v korpusu FOLK sodijo na področje institucionalnega, so v korpusu Gos »iz nejasnih razlogov označene kot javne«.[15] Avtor primerjave se kritično opredeli tudi do kategorije javni govor, znotraj katere pogreša posnetke telefonskih pogovorov. Telefonski pogovori na delovnem mestu so v korpusu FOLK kategorizirani kot institucionalni, v korpusu Gos pa kot nejavni nezasebni (Kaiser, 2018). V nadaljevanju bomo izpostavili še mejno kategorizacijo parlamentarnega govora v Gos 2.1 in jo primerjali z izbranimi rešitvami v tujih korpusih.

V korpusu C-ORAL-ROM je taksonomija v večji meri primerljiva s korpusom Gos 2.1. Poleg zasebnih posnetkov, ki se odvijajo v domačem okolju, so v omenjenem korpusu znotraj neformalnega govora vključeni še posnetki, ki se odvijajo v javnem življenju. V korpusu Gos 2.1 pa so pri nejavnem govoru poleg zasebnih posnetkov vključeni še posnetki nezasebnega govora, navezujoči se na posnetke interakcij na delovnem mestu in v izobraževalnih institucijah. Glede na analizo tujih govornih korpusov bi bilo smiselno razmisliti, da bi morda že na ravni taksonomije v referenčne korpuse za govorjeno slovenščino v prihodnje vključili razlikovanje med družbenimi sektorji oz. področji, v katerih se govorni dogodek odvija (npr. gospodarstvo, politika, zabava, mediji, izobraževanje). Ideja v slovenskem prostoru ni nova, saj se v kontekstu splošnih načel gradnje korpusov, tako pisnih kot govornih, znotraj lastnosti oz. kategorije **besedilni kontekst** kot potencialne oznake omeni izobraževanje, dom, delo in prosti čas (Gorjanc in Logar, 2007).

14 Kot primer za tovrsten tip govornega dogodka navajamo formalni delovni sestanek.
15 Sem sodita na primer osnovnošolska in srednješolska učna ura.

Tabela 5: *Pregled taksonomije oz. tipov govora v izbranih tujih korpusih*

FOLK	zasebno	institucionalno		javno	drugo
	ni opredeljeno	izobraževanje, uprava, medpoklicno sporazumevanje, klubsko/društveno življenje, religija, kultura (zabava, umetnost, šport), storitvene dejavnosti, medicina		politika, zabava, znanost, gospodarstvo	ni opredeljeno
BNC2014	demografski del	kontekstualni del			
	pretežno vsakdanji/neformalen govor	izobraževanje, informiranje	poslovanje, gospodarstvo	javno, institucionalno	prosti čas
		predavanja, učne predstavitve, novičarski komentarji, interakcije v razredu, možnost dodatnega vnosa	pogovori in intervjuji v podjetjih, prodajne predstavitve, poslovni sestanki, posveti, možnost dodatnega vnosa	politični govori, pridige, javne razprave, seje sveta, verska srečanja, sodni postopki, možnost dodatnega vnosa	govori, predvajani športni komentarji, pogovorne oddaje, telefonski pogovori, klubska srečanja, možnost dodatnega vnosa
ORAL2013	neformalno (nejavno)				
Nizozemski govorni korpus	zasebno			javno	
				predvajano (preko množičnih občil) *broadcast*	
				nepredvajano (preko množičnih občil) *non-broadcast*	
C-ORAL-ROM	formalno	neformalno			
		zasebno	javno		
			predavanje univerzitetnega profesorja, diskusija v pogovorni oddaji itd.		

Tipi govornih dogodkov, vključenih v korpus FOLK, so precej podrobno opredeljeni, saj so bili pridobljeni preko prostega vnosa, npr. pogovor med opravljanjem gospodinjskih opravil; odrasli igrajo družabne igre; učna ura na zasebni srednji šoli; ustni izpit na univerzi; menjava izmene v bolnišnici; mediacijski pogovor; biografski intervju itd. (Schmidt, 2014). Slednje potrjujejo tudi izkušnje pri pripravi korpusa BNC2014. Pri pridobivanju metapodatkov so različne značilnosti, nanašajoče se na tip govornega dogodka, kategorizirali v čim manjši meri, saj so ugotovili, da so te lahko (in pogosto so) za vsako besedilo drugačne (Love idr., 2018). Če namreč želimo čim bolj celovito prikazati spontani govor, je bistveno, da z merili za sestavo korpusa omogočamo kar najobsežnejšo raznolikost govornih kontekstov in da si prizadevamo v kar najmanjši meri vplivati na opredelitev kategorije govornega dogodka (Cresti in Moneglia, 2005). Na osnovi pregleda zasnove tujih korpusov se pri poimenovanju tipov govornega dogodka zdi torej smiselno, da zapisovalce metapodatkov v čim manjši meri usmerjamo. Omogočiti jim je treba, da v odprtem, prostem poimenovanju zajamejo čim več vsebinskih podatkov o govornem dogodku, tako da uporabnik korpusa že na osnovi tipa govornega dogodka pridobi informacijo o dogajanju na posnetku.

Glede na zgornjo ugotovitev so tipi govornega dogodka v Gos 2.1 ustrezno kategorizirani, saj odslikavajo raznolik spekter govorjenega jezika. V poimenovanjih posameznih tipov govornega dogodka lahko razberemo celo podatek o družbenem kontekstu (npr. fakultetno predavanje), številu udeležencev (npr. prosti dialog med dvema sogovornikoma) in stopnji pripravljenosti (npr. prosti monološki govor). Sistematizacija tipov govornih dogodkov je uporabna predvsem zaradi medsebojne primerljivosti metapodatkov in možnosti hitrih analiz. Vendar pa je pri tem priporočljivo, da ob vnaprej opredeljenih spustnih seznamih dodamo vsaj še oznako *drugo*. Druga možnost je, da pri beleženju tipa govornega dogodka dopuščamo prost vnos, po zaključenem procesu pridobivanja posnetkov pa te smiselno poimenujemo po posameznih tipih ali pa sistematiziramo zgolj proste opise, pridobljene pod oznako *drugo*.

Nujno se zdi ohraniti dodatno kategorijo, opis govornega dogodka, kjer označevalci metapodatkov na kratko prosto opišejo kontekst

govorne situacije. Takšen način je bil implementiran že v prvi različici korpusa Gos 1.1, podobno rešitev pa najdemo tudi v nekaterih tujih korpusih (npr. BNC2014). S prostim opisom prihodnjim raziskovalcem in drugim uporabnikom zagotovimo čim bolj natančne in celovite vsebinske informacije za raziskovalno delo ali razvoj novih rešitev.

Tabela 6: *Pregled kategorije tip govornega dogodka (domain) v izbranih tujih korpusih*

FOLK	neusmerjeno v aktivnosti	usmerjeno v aktivnosti: **prosti vnos:** načrtovanje dopusta, učna ura vožnje, panelna razprava itd.
BNC2014	demografski del:[16] prosti opis dogodka oz. kratek naslov	
	prosti vnos: par se sprehaja po podeželju in se pogovarja; pogovor o odnosih na kavi s prijatelji; pogovor s sostanovalci med pripravo obeda itd.	
ORAL2013	najpogostejši govorni dogodki: obisk, pogovor doma, v restavraciji, med skupno dejavnostjo itd.	
Nizozemski govorni korpus	opredeljenih je **14 kategorij**: pogovor (iz oči v oči); intervju; telefonski pogovor; poslovna transakcija oz. izmenjava/posel; intervju in diskusija; diskusija, debata, sestanek; predavanje; opis slike; spontani komentar; novinarsko poročilo; programi, ki poročajo o aktualnem dogajanju; informativni bilten; komentar; predavanje, govor; brano besedilo	
C-ORAL-ROM	neformalni govor	formalni govor
	prosti vnos	politični govor; politična razprava; pridiga; poučevanje; strokovna razprava; konferenca; gospodarstvo; pravo; pogovorna oddaja; znanstveni tisk; reportaža; intervju; šport; novice; vremenska napoved

Pri kategoriji kanal je treba upoštevati razlikovanje med govorci in končnimi naslovniki oz. uporabniki. Kanal lahko opredelimo glede na to, kdo so najaktivnejši oz. primarni govorci ali pa glede na to, komu je posnetek namenjen. Kanal radio in televizija na primer določajo končni naslovniki, ne pa osebni stik govorcev v studiu. Na to razlikovanje opozarjata tudi Deppermann in Hartung (2012), ki predlagata razlikovanje med notranjim in zunanjim komunikacijskim krogom. Pri pogovorni oddaji lahko kanal opredelimo glede na notranji komunikacijski krog kot osebni stik, medtem ko ga glede na zunanji komunikacijski krog opredelimo kot množični mediji. Govorce ločujemo na več nivojih: vabljeni gostje pogovorne oddaje (primarno verbalno aktivni

16 Razpon besedilnih vrst znotraj štirih krovnih kontekstualno utemeljenih kategorij ni bil fiksno določen, s čimer je omogočal vključitev dodatnih besedilnih vrst (Burnard, 2000).

govorci), občinstvo v studiu in gledalci pred domačimi TV-zasloni. Tovrstno stopenjsko razvrščanje občinstva je v korpusu FOLK omogočil prost vnos kot dodatna možnost poleg beleženja prisotnosti ali odsotnosti občinstva. Za označevanje samega kanala so bile vnaprej ponujene možnosti: iz oči v oči/osebni stik, telefon in posredovano preko množičnih medijev (+ mešano). Zabeleženi kanal v korpusih BNC2014 in ORAL2013 je osebni stik, v Nizozemskem govornem korpusu se mu pridružuje še oznaka na daljavo (npr. telefonski pogovor). Formalni del govornega korpusa C-ORAL-ROM pri kategoriji kanal razlikuje naravno okolje oz. osebni stik, telefon in množični mediji. Kanali, označeni v korpusu Gos 1.1, so televizija, radio, osebni stik in telefon, v korpusu Gos 2.1 pa mu je zlasti zaradi posnetkov, dodanih iz korpusa Artur, ki so bili posneti v času pandemije covida-19, dodan kanal internet. Pred tem internet v nasprotju s telefonom, osebnim stikom, avdiem in videem še ni bil upoštevan kot poseben prenosnik, saj naj bi bil dojet zgolj kot kanal za prenos zvoka in/ali slike (Zemljarič Miklavčič, 2008).

4.2 Analiza različnih vidikov nabora oznak za opis govorne situacije v slovenskih govornih korpusih

Taksonomija korpusa Gos 2.1, ki je utemeljena na **tipih govora** iz korpusa Gos 1.1, se deli na javni govor in nejavni govor. V javni govor sodita informativno-izobraževalni in razvedrilni govor, v nejavnega pa nezasebni in zasebni govor. Posnetki, ki so bili v korpus Gos 2.1 vključeni iz korpusa Gos VideoLectures, so bili označeni kot informativno-izobraževalni govor. Tako so bili označeni tudi posnetki javnega govora, vključeni v Gos 2.1 iz korpusa Artur, medtem ko so bili posnetki nejavnega govora iz Arturja označeni kot nejavni zasebni govor. Kot omenjeno, je nekoliko diskutabilen parlamentarni govor, ki je bil ob vključitvi iz korpusa Artur v Gos 2.1 označen kot nejavni nezasebni govor. Razmisliti bi bilo smiselno o njegovi prekategorizaciji v javni govor, podrobneje v informativno-izobraževalni govor. Argumentacijo za tovrstno določitev lahko najdemo v korpusu BNC2014, kjer so parlamentarni in sodni postopki ter seje sveta uvrščeni med javna/institucionalna besedila, in v korpusu FOLK, kjer področje „politike“ spada k javni interakcijski domeni.

Ob nadgradnji korpusa Gos v različico 2.1 je bil pripravljen interni predlog za prestrukturiranje oznak znotraj kategorije, imenovane **tip govornega dogodka**. Oznake posnetkov javnega, zlasti medijskega govora, so bile združene v skupine s podobnimi lastnostmi. Predlagane oznake, ki natančneje od izvirnih pojasnjujejo kontekst govornega dogodka, so televizijski novinarski prispevek namesto novinarski prispevek (npr. *POP TV, Preverjeno*), televizijski športni prenos namesto zgolj športni prenos, moderirani radijski program namesto moderirani program (*npr. Val 202, Aktualno*) in jezikovni tečaj namesto tečaj. Precej splošna in za končnega uporabnika nejasna oznaka moderirani pogovor naj bi bila prekategorizirana v televizijski intervju/soočenje (npr. *RTV Slovenija, Odmevi*), televizijsko razvedrilno oddajo (npr. *POP TV, As ti tud not padu*) in radijski intervju (npr. *Ognjišče, Naš gost*). Posamezni posnetki, v Gos 1.1 opredeljeni kot moderirani program, naj bi v skladu s predlaganim prestrukturiranjem sodili k radijskim intervjujem, namesto moderirane oddaje in resničnostnega šova pa naj bi bilo v Gos 2.1 uvedeno poenoteno preimenovanje televizijska razvedrilna oddaja (*npr. POP TV, Vzemi ali pusti*). Implementacija predlaganih sprememb bi vsekakor pripomogla k večji informativnosti in nedvoumnosti metapodatkov in k izboljšanju uporabniške izkušnje. Preimenovanje bi bilo smotrno tudi zato, ker posamezni tipi govornega dogodka s svojo novo oznako hkrati implicirajo tudi že kanal (npr. televizijski novinarski prispevek, radijski intervju).

Posnetkom iz korpusa Gos VideoLectures, ki tipov govornih dogodkov niso imeli beleženih, je bila ob njihovi vključitvi v Gos 2.1 dodana enotna oznaka javno predavanje.[17] Posnetki, dodani iz korpusa Artur, so ohranili svoj izvorni tip govornega dogodka (npr. seja državnega zbora, novinarska konferenca), vendar je treba opozoriti, da v Gos 2.1 niso bili vključeni posnetki, ki še niso transkribirani (npr. seminarji in predavanja, katerih vir sta Arnes in Univerza v Mariboru). Gos 2.0, dostopen v konkordančniku na CJVT,[18] vključuje spodaj navedene tipe

17 V konkordančniku za korpus Gos 2.0, dostopnem na CJVT, jim je bila dodana oznaka predavanje.

18 https://viri.cjvt.si/gos/

govornih dogodkov[19] (nejasno je, zakaj se kot tip govornega dogodka dvakrat pojavi oznaka moderirani pogovor):

Seja državnega zbora
Pogovor med prijatelji/znanci
Predavanje
Spletni dogodek
Intervju
Prosti monološki govor
Moderirani pogovor
Pogovor v družini
Prosti dialog med dvema sogovornikoma
Moderirani program
Razlaganje in opisovanje
Moderirani pogovor
Okrogla miza
Osnovnošolska učna ura
Srednješolska učna ura
Storitev
Fakultetno predavanje
Novinarska konferenca
Neformalni delovni sestanek
Športni prenos
Novinarski prispevek
Prodaja/trgovina
Formalni delovni sestanek
Moderirana oddaja
Konzultacija
Svetovanje
Javno predavanje
Informacije
Resničnostni šov
Formalni razgovor
Tečaj
Tajništvo
Nagovor na dogodku

Skozi analizo smo identificirali nekaj diskutabilnih oznak za opis tipa govornega dogodka v korpusu Gos 2.1, za katere bomo v diskusiji predlagali nekaj potencialnih rešitev oz. predlogov za prekategorizacijo metapodatkov v prihodnjih nadgradnjah slovenskih govornih korpusov. Pri javnem govoru v nadaljevanju predlagamo preimenovanje oznak predavanje, intervju, spletni dogodek in (osnovnošolska, srednješolska) učna ura. Tip govornega dogodka predavanje je nejasno označen kar s tremi različnimi oznakami, in sicer se poleg oznak javno predavanje in predavanje v konkordančniku Gos 2.0 pojavlja še oznaka fakultetno predavanje.

Pri nejavnem nezasebnem govoru ugotavljamo, da bi posamezne tipe govornih dogodkov lahko združili, spet druge pa podrobneje vsebinsko opredelili. Mednje sodijo neformalni delovni sestanek, storitev, prodaja/trgovina, informacije, svetovanje in tajništvo. Cilj prekategorizacije tipov govornih dogodkov bi vsekakor moral biti poenotenje, ko gre za govorne situacije v skoraj identičnih kontekstih, saj na ta način dosežemo večjo primerljivost in strukturiranost metapodatkov. V dodatnem opisu govornih dogodkov pa naj zapisovalci prosto in podrobneje zabeležijo podatke o tem, kdo je sodeloval v govornem dogodku,

19 Posamezne oznake v korpusu Gos 2.0 v konkordančniku na CJVT in v korpusu Gos 2.1 v repozitoriju Clarin niso poenotene, kot je npr. oznaka akademski, družboslovje vs. fakultetno predavanje; gimnazija, družboslovje vs. srednješolska učna ura; javno predavanje vs. predavanje.

kdaj in kje se je ta odvijal in kakšne so bile glavne teme ali namen pogovora.

Pri nejavnem zasebnem govoru je analiza pokazala nekaj nekonsistentnosti pri tipih govornega dogodka, označenih kot prosti dialog med dvema sogovornikoma, razlaganje in opisovanje ter prosti monološki govor. Glavni vzrok za nekonsistentno poimenovanje posnetkov z oznako razlaganje in opisovanje je ne povsem usklajen proces pridobivanja posnetkov na več fakultetah.[20] Med snemanjem so se govorcem na zaslonu sproti izpisovala vprašanja, ki so določala potek njihovega govora. Nekaj primerov tovrstnih vprašanj:

- Kateri film vam je bil všeč in zakaj?
- Na kratko opišite zgodbo tega filma.
- Kateri izletniški kraj vam je všeč in zakaj?
- Opišite pot do tega kraja.
- Opišite pripravo jedi, ki vam je všeč.
- Narekujte primer vabila na rojstnodnevno zabavo.

V primerjavi z zgoraj opisanim načinom so imeli govorci pri snemanju posnetkov z oznako prosti monološki govor več svobode pri izbiri tematike. Vendar pa tudi tukaj monologi niso povsem prosti, saj so jih s (pod)vprašanji usmerjali snemalci ali pa so se vprašanja izpisovala na zaslonu. Za boljše razumevanje konteksta snemalne situacije navajamo del navodil:

a) prosto narekovanje pametnim napravam,[21] brez predloge. Govorec si je moral predstavljati, kot da narekuje pametnemu telefonu ali drugi pametni napravi, ta pa njegovo besedilo zapisuje (npr. prosto narekovanje SMS-sporočil, govorno iskanje po spletu). Isti govorec je lahko v enem posnetku uporabil več različnih tem:

    - Kako pripraviš svojo najljubšo jed?
    - Po kateri poti greš od doma do npr. šole, fakultete, službe ali kakšne turistične destinacije?

20 Koordiniranje snemanja nejavnega govora za govorno bazo Artur je potekalo na FE UL in FERI UM.

21 Koordiniranje snemanja je potekalo na FERI UM.

- Pametni napravi narekuj vabilo na dogodek s podatki kaj, kdaj, kje.
- Pametnemu asistentu narekuj katero koli drugo primerno temo (vsebovati ne sme sovražnega ali žaljivega govora).

b) prosti monološki govor, delno voden s sprotnimi kratkimi podvprašanji s strani moderatorja.[22] Monološki govor je sproti usmerjal moderator ali pa so bile okvirne tematike govorcem predlagane vnaprej. Na posnetkih se pojavljajo poljubne kombinacije teh tematik, včasih pa tudi povsem prosti monolog o poljubnih temah, kot je npr. opis postopka izgradnje objekta in priprave dokumentacije zanj, opis nakupovanja ali nastanka glasbenega banda itd.

Pri kategoriji **Opis govornega dogodka** gre za prosti opis govornega dogodka po presoji snemalca oz. zapisovalca metapodatkov. Prosti opisi iz korpusa Gos 1.1 so bili v nespremenjeni obliki preslikani v aktualno različico korpusa Gos 2.1, enako velja za naslove predavanj iz korpusa Gos VideoLectures. Pri izdelavi korpusa Artur zbiranje metapodatkov za to kategorijo ni bilo predvideno, zato bi jo bilo treba posnetkom dodati ročno v eni od prihodnjih nadgradenj. S tem bi podatke poenotili ter raziskovalcem in drugim uporabnikom referenčnega korpusa Gos 2.1 omogočili pridobivanje natančnejših informacij o kontekstu govornega dogodka. Nekateri primeri takšnih opisov iz Gos 2.1 so obnavljanje zavarovalne police za hišo; glasbena oddaja o domači glasbi z gosti in nagradno igro ter predavanje ob dnevu fakultete o najnovejših odkritjih v vesolju.

Ker želimo, da referenčni korpus čim bolj celovito odslikava raznolikost spontanega govora, moramo po vzoru tujih korpusov dopuščati kar najbolj prosto opredelitev oznak za opis konteksta govornega dogodka. V primeru rabe formalnega jezika namreč lahko predvidimo vrsto tipičnih govornih dogodkov, v katerih je glede na določen družbeno-zgodovinski kontekst formalna raba jezika prednostna, kar pa ne velja za področje neformalnega govora. Formalni govor je zato mogoče učinkovito prepoznati tako, da opredelimo njegove najbolj tipične kontekste rabe. V nasprotju s tem pa je množica situacij, v katerih se uporablja neformalni jezik, odprta in tipov govornega dogodka ne moremo opredeliti s

22 Koordiniranje snemanja je potekalo na FE UL.

pomočjo seznama tipičnih kontekstov rabe. Noben kontekst ni namreč bolj tipičen od drugega. Z natančno opredelitvijo žanra (oz. tipa govornega dogodka) je nedvomno mogoče doseči višjo stopnjo primerljivosti podatkov, vendar pa lahko s tem izgubimo oz. izločimo pomembna področja rabe neformalnega govora, kjer so posamezni žanri še vedno v veliki meri neraziskani (Cresti in Moneglia, 2005).

Oznake znotraj kategorije **kanal** so pri preslikavi iz korpusov Gos 1.1 in Gos VideoLectures v Gos 2.1 ostale nespremenjene. Kanala televizija in radio sta v korpusu Gos 2.1 določena glede na končne uporabnike, medtem ko sta kanala osebni stik in telefon opredeljena glede na govorce. Kategorija kanal je bila dodana pri posnetkih iz korpusa Artur, saj metapodatki tam niso bili beleženi. Pri govornih dogodkih seja državnega zbora, okrogla miza, nagovor na dogodku, novinarska konferenca, prosti dialog med dvema sogovornikoma in prosti monološki govor je bil dodan kanal osebni stik. Pri govornem dogodku intervju je bil kanal pri preslikavi oznak opredeljen kot radio, pri intervjujih, domensko poimenovanih kot spletni dogodek, pa je kanal označen kot internet. Pri slednjih gre pretežno za intervjuje, posnete z namenom objave na spletu, podobni primeri bi lahko bili tudi podkasti. Pri govornih situacijah, kot sta okrogla miza in novinarska konferenca, je priporočljivo razmisliti o podrobnejši podkategorizaciji udeležencev, in sicer na primarne govorce, to so vabljeni govorniki, za katere je tipičen kanal osebni stik, od fizično prisotnega (bolj ali manj aktivnega) občinstva ter uporabnikov, ki dogodek spremljajo ali preko televizije ali v digitalnem formatu na internetu.

5 Diskusija

Na osnovi primerjalne analize s tujimi govornimi korpusi in pregleda obstoječega nabora oznak v slovenskih govornih korpusih podajamo nekaj predlogov za potencialno prekategorizacijo metapodatkov za opis konteksta govorne situacije v prihodnje. Znotraj kategorije **tip govornega dogodka** bi zaradi večje medsebojne primerljivosti podatkov oznaki javno predavanje in predavanje poenotili v oznako javno predavanje. Slednja namreč izraža, da je predavanje namenjeno širšemu občinstvu. Oznaka fakultetno predavanje je po našem mnenju

ustrezna, saj natančneje opredeljuje tip predavanja. Glede na to, ali je predavanje predvajano preko interneta ali poteka v živo, se primerno določi kanal, zaradi vse večje digitalizacije pa bi bilo vendarle smiselno razmisliti tudi o vpeljavi nove oznake spletno predavanje. Skoraj vsi intervjuji v korpusu Gos 2.1 so radijski intervjuji, zato bi jih tako tudi preimenovali. Za intervjuje, predvajane preko interneta, bi bilo smiselno preimenovanje v spletne intervjuje oz. spletne pogovore. Slednja oznaka je primerna tudi za posnetke, na katerih več udeležencev debatira preko spletne platforme (npr. *posnetki STA kluba*).

Kot navedeno, je oznaka spletni dogodek ena od najbolj neustreznih oznak za poimenovanje tipa govornega dogodka v korpusu. Ob nadgradnji korpusa Gos 2.1 je bila prenesena iz korpusa Artur in podaja informacijo o kanalu, ne pa o govornem dogodku samem. Razlog za to odločitev je treba razumeti v duhu časa epidemije covida-19, ko je večina posnetih dogodkov za bazo Artur potekala preko interneta in se je oznaka spletni dogodek zdela ustrezna. V času po epidemiji ugotavljamo, da bi bilo treba posnetke ponovno analizirati in za vsak govorni dogodek posebej na novo določiti oznako. Za posnetke, katerih vir je STA, predlagamo preimenovanje v spletni posvet ali spletno soočenje (*npr. predstavitev kandidatov za predsednika Atletske zveze Slovenije*) in (moderirani) spletni pogovor (*npr. o rakavih bolnikih in pripravah na pandemijo covida-19*). Za posnetke, ki sta jih v korpus Artur prispevala Univerza v Mariboru in ZRC SAZU, predlagamo oznaki spletna konferenca in spletna (panelna) razprava, nekatere dodatne možnosti so še video navodilo (*tutorial*), spletna delavnica in spletni seminar (*webinar*).

Premisliti je treba tudi, ali tip govornega dogodka, kot sta osnovnošolska in srednješolska učna ura, sodi v javni govor ali bi bilo morda bolj ustrezno, da ga vključimo v nejavni nezasebni govor. V korpusu FOLK spada oznaka izobraževanje na področje, ki se imenuje institucionalno, v korpusu BNC2014 pa na področje, imenovano izobraževanje/informiranje. Ustreznejše poimenovanje za neformalni delovni sestanek, ki označuje posnetke, na katerih se zaposleni v trgovini (*npr. lovski ali ribiški*) sproščeno pogovarjajo, bi morda bilo pogovor med sodelavci. Po vzoru tujih korpusov, ki s poimenovanjem tipov govornih dogodkov precej natančno opredeljujejo vsebino govornega dogodka, in z namenom doseganja čim večje obvestilnosti bi predlagali

prekategorizacijo oznake storitev v prodajna predstavitev, pogovor med uslužbencem in stranko ter v inštruiranje. Sem spadajo posnetki iz korpusa Gos 1.1, kot so pogovor arhitekta s stranko o izgradnji bazena, fizioterapevtke s pacientko in pogovor v frizerskem salonu. Glede na vse večjo digitalizacijo delovnih procesov bi med metapodatke lahko umestili dodatno oznako, in sicer spletni sestanek. V primeru, da bi na višji ravni taksonomije dodali predlagani podatek o družbenem sektorju, kot je npr. gospodarstvo, storitvene dejavnosti ali trgovina, bi tip govornega dogodka prodaja/trgovina preimenovali v (telefonski) pogovor med prodajalcem in stranko. Gre za govorne dogodke, kot so pogovor med prodajalcem prevlek in stranko, nakup rož v cvetličarni, (telefonski) pogovor med stranko in prodajalcem v zlatarni ali pogovor med stranko in uslužbenko ponudnika telekomunikacijskih storitev. Posamezni posnetki, v Gos 1.1 označeni kot informacije, so vsebinsko skoraj identični dogodkom, označenim kot prodaja/trgovina, zato bi jih bilo smiselno poenotiti v (telefonski) pogovor med prodajalcem/informatorjem in stranko, podrobnejše podatke o govornih dogodkih, kot so bančna uslužbenka posreduje informacije o ponudbi banke ali podajanje informacij s strani klicnega centra podjetja, ki trži telekomunikacijske storitve, pa bi navedli v kategoriji opis govornega dogodka. Oba posnetka z oznako svetovanje bi vsebinsko natančneje opredelili kot (telefonski) pogovor med svetovalcem in stranko (npr. posnetek o svetovanju glede energetsko varčne gradnje), oba posnetka z oznako tajništvo pa kot telefonski pogovor med sodelavci (npr. telefonski pogovor med tajnico in predavateljem).

Če bi želeli tipe govornega dogodka v različnih jezikovnih virih, kot sta Artur in Gos 2.1, čim bolj poenotiti, bi prosti dialog med dvema sogovornikoma preimenovali v pogovor v družini ali pogovor med prijatelji/znanci. Če se za preimenovanje ne odločimo, je navedba med dvema sogovornikoma v vsakem primeru redundantna, saj dialog že nakazuje število udeležencev. Za oznako razlaganje in opisovanje predlagamo, da se poimenovanje poenoti v vodeni monološki govor (oz. vodeni monolog). Posamezni posnetki nejavnega govora v korpusu Artur, ki so sicer vsebinsko primerljivi, so bili namreč zaradi različnih načinov snemanja na več fakultetah nekonsistentno poimenovani. V primerjavi s posnetki z oznako razlaganje in opisovanje so imeli govorci

pri snemanju posnetkov z oznako prosti monološki govor več svobode pri izbiri tematike, vendar, kot smo že videli, tudi tu ne gre za povsem proste monologe. Na osnovi tega bi bilo smiselno tip govornega dogodka preimenovati v delno vodeni monološki govor (oz. delno vodeni monolog).

Tabela 7: *Predlog preimenovanja oznak za kategorijo tip govornega dogodka*

Obstoječe poimenovanje	**Alternativna poimenovanja**
Javni govor	
• javno predavanje • predavanje	javno predavanje
nova oznaka	spletno predavanje
intervju	• radijski intervju • spletni intervju • spletni pogovor
spletni dogodek	• spletni posvet • spletno soočenje • (moderirani) spletni pogovor • spletna konferenca • spletna (panelna) razprava • video navodilo (tutorial) • spletni seminar (webinar)
Nejavni nezasebni govor	
neformalni delovni sestanek	pogovor med sodelavci
storitev	• prodajna predstavitev • pogovor med uslužbencem in stranko • inštruiranje
nova oznaka	spletni sestanek
prodaja/trgovina	(telefonski) pogovor med prodajalcem in stranko
informacije	• (telefonski) pogovor med prodajalcem in stranko • (telefonski) pogovor med informatorjem in stranko
svetovanje	(telefonski) pogovor med svetovalcem in stranko
tajništvo	telefonski pogovor med sodelavci
Nejavni zasebni govor	
prosti dialog med dvema sogovornikoma	• pogovor v družini • pogovor med prijatelji/znanci • prosti dialog
razlaganje in opisovanje	• vodeni monološki govor • vodeni monolog
prosti monološki govor	• delno vodeni monološki govor • delno vodeni monolog

Ugotovili smo, da je pri kategoriji **opis govornega dogodka** priporočljivo dopustiti, da zapisovalci v kratkem zapisu s svojimi besedami podrobneje opišejo kontekst govornega dogodka. V korpusu Artur opisi govornega dogodka niso bili beleženi, zato bi jih bilo v prihodnje smiselno ročno dodati. Za govorni tip prosti monološki govor bi bilo v opisu smiselno izpostaviti pogosto ponavljajoče se teme:

- Opis priprave (najljubše) jedi
- Narekovanje ali podajanje ukazov pametni napravi
- Govorno iskanje po spletu
- Opis poti
- Opis poteka (delovnega) dne
- Opis dela na vrtu
- Opis vsebine najljubšega filma
- Opis najljubšega kraja in njegovih znamenitosti itd.

Na osnovi analize zasnov tujih govornih korpusov podajamo najprej v tabelarični in nato še v opisni obliki nekaj usmeritev za vključitev priporočenih sklopov podatkov, pri čemer ni nujno, da vsak opis govornega dogodka zajame vse sklope. Zapisovalce metapodatkov lahko usmerimo in jim priporočimo zapis podatkov, ki so za opis govorne situacije še posebej relevantni, hkrati pa jim dopuščamo kritično možnost vnosa glede na konkreten govorni dogodek in glede na že zabeležene metapodatke.

Tabela 8: *Predlog atributov za prosti opis govornega dogodka*

Atribut	Opis	Primeri
Zaupnost med govorci	Označuje stopnjo medsebojnega poznavanja med govorci.	družinski člani, prijatelji, neznanci
(Družbene) vloge udeležencev	Označuje vnaprej določene pravice in obveznosti, konstitutivne za različen tip govorčevih interakcij.	mati in otrok, sodnik in obtoženec, moderator in intervjuvanec
Namen oz. cilj	Označuje namen govornega dogodka.	izmenjava mnenj, prepričevanje, postavitev diagnoze, pritoževanje, pripovedovanje šal
Tematika	Označuje glavno in/ali dodatne teme govornega dogodka.	računalniško programiranje, kulinarika

Atribut	Opis	Primeri
Stopnja pripravljenosti	Označuje stopnjo vnaprejšnje pripravljenosti vsebine in/ali strukture govornega dogodka.	(delno) spontani, (delno) vodeni, vnaprej pripravljeni
Interaktivnost oz. vloga občinstva	Označuje morebitno prisotnost občinstva in možnost njegovega vključevanja v govor.	vabljeni gostje – primarni govorci, občinstvo v studiu, gledalci v domačem okolju
Število aktivnih govorcev	Označuje število govorcev, ki sodelujejo v govornem dogodku.	monolog, dialog, multilog
Kraj, čas	Označuje kraj in čas govornega dogodka.	v šoli, v tihem studiu, ponedeljkov jutranji krožek

Ena od pomembnih oznak je seznanjenost govorcev oz. medsebojna zaupnost, ki opredeljuje, kako dobro se govorci med seboj poznajo. V korpusu FOLK je opredeljenih več oznak z različnimi stopnjami zaupnosti, kot so npr. poznan, neznan in zaupen. V korpusu BNC2014 je bil implementiran zaprt spustni seznam z vnaprej določenimi oznakami, kot so ožja družina, partnerji, najbližji prijatelji vs. prijatelji in širši družinski krog, medtem ko so bili v korpus ORAL2013 vključeni zgolj govorci, ki so medsebojno tesno povezani, se dobro poznajo ali so intimni. Na govorni dogodek vplivajo tudi (družbene) vloge udeležencev. Te se, kot v korpusu FOLK, navezujejo na pravice in obveznosti, konstitutivne za govorčev vstop v zasebne ali institucionalne interakcije, zanje pa je značilno, da se ne tvorijo šele med samim govornim dogodkom, temveč so znane vnaprej. Takšne vloge so npr. mati in otrok v domačem okolju, sodnik in obtoženec v sodnem postopku, moderator in intervjuvanec, učitelj in učenec, uradnik in stranka itd.

Korpusa BNC2014 in Nizozemski govorni korpus v opisu navajata tudi namen oz. cilj govornega dogodka. Sem lahko uvrščamo izmenjavo mnenj, pogajanje, prepričevanje, pojasnjevanje, postavitev diagnoze, psihološko svetovanje, pritoževanje, poizvedovanje, opravičevanje, pripovedovanje šal itd. Po vzoru korpusa BNC20214 sodi v opis govorne situacije tudi oznaka tematike pogovora (omemba glavne in dodatnih tem), kot sta npr. računalniško programiranje in kulinarika. Govorne dogodke lahko razlikujemo glede na to, ali so vezani na specifično tematiko/področje ali pa so tematsko neopredeljeni. Govorni dogodek je odvisen tudi od stopnje pripravljenosti, saj je dogodek

(delno) spontan, (delno) voden ali pa moderiranje poteka v živo in sproti. Njegova struktura je lahko določena, kar pomeni, da se govorci nanj pripravijo vnaprej ali pa se jim vprašanja sproti izpisujejo oz. se jih sproti glasovno usmerja. Korpus ORAL2013 vsebuje izključno spontani, nepripravljeni govor, Nizozemski govorni korpus pa ločuje oznake spontani, pripravljeni in bolj ali manj pripravljeni.

Naslednja oznaka, ki podrobneje določa govorni dogodek, je interaktivnost oz. vloga občinstva. Navesti je priporočljivo morebitno prisotnost občinstva in v nadaljevanju opredeliti, ali ima občinstvo možnost takojšnjega in interaktivnega podajanja povratnih informacij ali ne. Če doslej metapodatka o številu in menjavi aktivnih govorcev (tj. tistih, ki kaj povedo) še nismo zajeli, ga je treba navesti vsaj na tem mestu (npr. monolog, dialog, multilog). Ključne informacije o kontekstu govornega dogodka pridobimo tudi iz oznak, kot sta kraj in čas govornega dogodka. Korpusa BNC2014 in C-ORAL-ROM snemalcem omogočata prosti opis kraja (npr. v tihem studiu ali pridiga v cerkvi) in časa (npr. ponedeljkov jutranji krožek v osnovni šoli), korpus BNC2014 pa je za leto snemanja uporabil spustni seznam. Beležimo lahko celo časovno omejenost, kjer razlikujemo govorne dogodke, ki so časovno omejeni (npr. govorilne ure), od tistih, katerih trajanje je neomejeno.

Kar se kategorije **kanal** tiče, bi bilo smiselno razlikovati med tem, ali je dogodek prvenstveno namenjen izvedbi v živo in je na spletu objavljen zgolj njegov posnetek (v tem primeru bi kanal označili glede na govorce oz. govorni dogodek) ali pa je prvenstveno namenjen za objavo na spletu in se v živo odvija zgolj zato, da je posnetek bolj zanimiv ali avtentičen, na primer v prisotnosti občinstva (v tem primeru bi kanal označili glede na končne uporabnike, torej kot internet). Takšen primer so seje državnega zbora, kjer bi kanal glede na končne uporabnike lahko bil opredeljen kot televizija ali internet, saj so seje na voljo tudi preko spletnih strani Parlameter[23]. Podobna dilema se pojavi tudi pri predavanjih, objavljenih na spletni platformi VideoLectures. V Gos 2.1 je kanal v obeh primerih označen kot osebni stik, čeprav bi ga glede na končne uporabnike lahko opredelili tudi kot internet. Posamezna netranskribirana predavanja, ki sicer niso bila dodana v korpus Gos 2.1, najdemo pa jih v korpusu Artur, so bila zaradi epidemije

23 https://parlameter.si/

covida-19 od samega začetka snemalnega procesa načrtovana kot spletna predavanja. Smiselno je torej, da je oznaka za kanal pri tovrstnih posnetkih internet in ne osebni stik. Takšni govorni dogodki so na primer nekatera predavanja, objavljena na Arnesovem spletišču in spletna predavanja za zaposlene na Univerzi v Mariboru.

6 Zaključek

Da bi kritično ovrednotili nabor oznak za opis konteksta govorne situacije, smo v raziskavi primerjali nabor oznak v slovenskih govornih korpusih z izbranimi rešitvami v tujih referenčnih govornih virih. Z analizo oznak v izbranih kategorijah smo identificirali posamezna kritična mesta in zanje podali nekaj usmeritev in predlogov za potencialno prekategorizacijo. Dotaknili smo se tudi internih priporočil za prestrukturiranje oznak znotraj kategorije tip govornega dogodka, ki so bila oblikovana ob nadgradnji korpusa Gos v različico 2.1. Ker zgolj delna implementacija omenjenih priporočil pomeni nekonsistentnost označevanja, bi bilo temu v prihodnje priporočljivo nameniti dodatno pozornost.

V zvezi s poimenovanjem tipov govornih dogodkov predlagamo čim manj usmerjanja pri beleženju metapodatkov, kar omogoča celovitejši odraz raznolikosti sodobnega govorjenega jezika. Zapisovalcem omogočimo prosti vnos ali pa na spustnih seznamih, ki povečujejo medsebojno primerljivost podatkov, dodamo vsaj še oznako *drugo*. Gre za kategorije, kakršni sta tudi besedilna vrsta in govorni položaj, ki so v kontekstu korpusnega jezikoslovja (pisni in govorni korpusi) razumljene kot »jezikovno in kulturno najbolj specifične« (Gorjanc in Logar, 2007). Z vnaprejšnjo kategorizacijo oznak tvegamo, da posamezne govorne dogodke izpustimo ali pa jih nezadostno in nekoherentno opišemo. Pri posnetkih, označenih kot spletni dogodek, predlagamo ponovno analizo in določitev nove oznake, saj trenutna podaja informacijo o kanalu in ne o govornem dogodku. Razmislek o potencialni prekategorizaciji je potreben tudi pri posnetkih, ki so v Gos 2.1 označeni kot nejavni nezasebni govor, v tujih korpusih pa uvrščeni v kategorijo institucionalno (FOLK) ali javno/institucionalno (BNC2014). Podobno velja za javne govorne dogodke, kot so npr. osnovnošolske in srednješolske učne ure, za katere je najbrž primernejša oznaka nejavni

nezasebni tip govora. Druga možnost je, da na ravni taksonomije dodamo diferenciacijo po (družbenih) področjih, kot so gospodarstvo, politika, zabava, mediji in izobraževanje. Na področje politike bi uvrstili parlamentarni govor, pri katerem po vzoru tujih govornih korpusov podajamo pobudo za prekategorizacijo v javni govor. Nekaj predlogov za alternativna preimenovanja oznak kategorije tip govornega dogodka podajamo v diskusiji v Tabeli 7, strnjen pregled priporočenih smernic za opis govornega dogodka pa v Tabeli 8. Ročne opise govornega dogodka bi bilo v prihodnje treba dodati vsem posnetkom, ki so bili ob nadgradnji v korpus Gos 2.1 dodani iz korpusa Artur. Kanal lahko diferenciramo glede na (fizično) prisotnost ali odsotnost občinstva ali glede na stopnjo aktivnosti govorcev. Pri posnetkih sej državnega zbora in predavanj s spletišča VideoLectures bi tako kanal lahko opredelili kot televizija/internet in ne kot osebni stik.

Zaradi hitrega razvoja velikih jezikovnih modelov, ki temeljijo na obsežnih zbirkah besedil, bo vse večjo vlogo pridobivala avtomatska identifikacija žanrov (*automatic genre identification*), ki bo avtomatsko pridobljena besedila, značilna za velike spletne korpuse, obogatila z oznakami o tipu govornega dogodka (npr. promocijsko, pravno besedilo). Obetavne rezultate nudijo že veliki jezikovni modeli sami, ki za avtomatsko identifikacijo žanrov niso učeni (*zero-shot*). To velja npr. za modela GPT-3.5 in GPT-4, ki sta se izkazala tudi na področju zunajdomenskih in večjezičnih virov ter pri identifikaciji žanrov znotraj manjših jezikov, kot je slovenščina. Vendar se je treba zavedati njihovih omejitev, kot sta zamuden ročni pregled dobljenih rezultatov in skrita arhitektura, ki onemogoča poznavanje kriterijev za razvrščanje. Za učinkovito rabo teh modelov potrebujemo strokovno znanje o tvorjenju učinkovitih ukazov (*prompt engineering*), pri nekaterih, kot so Falcon, pa tudi visoko zmogljivo strojno opremo. Vse to otežuje dostopnost raziskav posameznikom in organizacijam z omejenimi viri (Kuzman et al., 2023). Natančnejše in hitreje generirane rezultate zagotavlja prosto dostopen, na XLM-RoBERTa temelječ in preko ročno označenih podatkov prilagojen (*fine-tuned*) žanrski klasifikator (*genre classifier*),[24] ki omogoča avtomatsko identifikacijo tipov govornih dogodkov v številnih jezikih. V sklopu raziskave je bil podan tudi zanimiv

24 https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

predlog za enotno žanrsko shemo za združevanje žanrsko raznolikih podatkovnih baz s pomočjo naslednjih oznak: informacije/razlaga (npr. raziskovalni članek, biografija), navodila (npr. recept, tehnična pomoč), pravni tekst (npr. licenca za programsko opremo, pogodba), novice (npr. športno poročilo, policijsko poročilo), mnenje/argumentacija (npr. blog, politična propaganda), promocija (npr. oglas, vabilo na dogodek), proza/lirika (npr. pesem, šala), forum (npr. forum, vprašanja in odgovori – Q&A) in drugo (Kuzman et al., 2023). Ne glede na to, ali tipe govornih dogodkov označujemo avtomatsko ali ročno, pa si je treba prizadevati, da zabeleženi metapodatki čim celoviteje odražajo sodobno govorjeno slovenščino v vseh njenih najsubtilnejših različicah in pestrosti kontekstualno pogojenih pojavitev.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science, 3*(1), 245–270.

Burnard, L. (2000). The British national corpus users reference guide. In: Oxford University Computing Services Oxford.

Cermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics, 14*, 113–123. Pridobljeno s https://www.jbe-platform.com/content/journals/10.1075/ijcl.14.1.07cer

Chizhik, A. V., & Sergeyev, D. A. (2021). Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach. International Conference on Digital Transformation and Global Society.

Cresti, E., & Moneglia, M. (2005). C-ORAL-ROM: *Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins Publishing Company. Pridobljeno s https://books.google.si/books?id=ybc5AAAAQBAJ

Cresti, E., Nascimento, F. B. d., Moreno-Sandoval, A., Véronis, J., Martin, P., & Choukri, K. (2004). The C-ORAL-ROM CORPUS. A Multilingual Resource

of Spontaneous Speech for Romance Languages. International Conference on Language Resources and Evaluation.

Deppermann, A., & Hartung, M. (2012). Was gehört in ein nationales Gesprächskorpus? : Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim).

Gorjanc, V. (2005). *Uvod v korpusno jezikoslovje*. Izolit.

Gorjanc, V., & Fišer, D. (2013). *Korpusna analiza* (2. izd. ed.). Znanstvena založba Filozofske fakultete.

Gorjanc, V., & Krek, S. (Ur.). (2005). *Študije o korpusnem jezikoslovju: zbornik* (1. izd. ed., Vol. 130). Krtina.

Gorjanc, V., & Logar, N. (2007). Od splošnih do specializiranih korpusov-načela gradnje glede na njihov namen. *Razvoj slovenskega strokovnega jezika*, 637–650. Pridobljeno s https://doi.org/https://repozitorij.uni-lj.si/Dokument.php?id=182750&lang=slv

Hymes, D. H. (1974). Foundations in Sociolinguistics: An Ethnographic Approach.

Kaiser, J. (2018). Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung*, *19*, 515–552. Pridobljeno s https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8668/file/Kaiser_Zur_Stratifikation_des_FOLK-Korpus_2018.pdf

Kopřivová, M., Komrsková, Z., Poukarová, P., & Lukeš, D. (2019). Relevant Criteria for Selection of Spoken Data: Theory Meets Practice. *Journal of Linguistics/Jazykovedný casopis*, *70*(2), 324–335. doi: 10.2478/jazcas-2019-0062

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149-1175. Pridobljeno s https://www.mdpi.com/2504-4990/5/3/59

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*, 319–344. Pridobljeno s https://www.jbe-platform.com/content/journals/10.1075/ijcl.22.3.02lov?crawler=true

Love, R., Hawtin, A., & Hardie, A. (2018). *The British National Corpus 2014: User Manual and Reference Guide (version 1.1)*. ESRC Centre for Corpus Approaches to Social Science.

Lucie, B., Michal, K., & Martina, W. (2015). Korpus spontánní mluvené češtiny ORAL2013.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and First Evaluation. International Conference on Language Resources and Evaluation,

Petukhova, V., Malchanau, A., & Bunt, H. (2015). Modelling argumentation in parliamentary debates. Proceedings of the 15th Workshop on Computational Models of Natural Argument, Principles and Practice of Multi-Agent Systmes Conference (PRIMA 2015), Bertinoro, Italy,

Pretnar Žagar, A., Pahor de Maiti, K., & Fišer, D. (2022). What's on the agenda?: topic modelling parliamentary debates before and during the COVID-19 pandemic = Kaj je na dnevnem redu?. Pridobljeno s https://sidih.github.io/agenda/index-sl.html

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, *11*(12), e0168843.

Schmidt, T. C. (2014). The Research and Teaching Corpus of Spoken German — FOLK. International Conference on Language Resources and Evaluation,

Verdonik, D. (2013). Koncept konteksta v jezikoslovnih in diskurznih teorijah. *Slavistična revija*, *61*(4), 631–650. Pridobljeno s https://srl.si/sql_pdf/SRL_2013_4_08.pdf

Verdonik, D. (2018). Korpus in baza Gos Videolectures.

Verdonik, D. (2021). *Govorni viri za pravorečje*. 1. slovenski pravorečni posvet. Pridobljeno s https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., ..., & Bordon, D. (2023). *ASR database ARTUR 1.0 (audio)*. Faculty of Electrical Engineering and Computer Science, University. Pridobljeno s https://www.clarin.si/repository/xmlui/handle/11356/1776

Verdonik, D., Bizjak, A., Žgank, A., & Dobrišek, S. (2022). Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. Pridobljeno s https://nl.ijs.si/jtdh22/pdf/JTDH2022_Verdonik-et-al_Metapodatki-o-posnetkih-in-govorcih-v-govornih-virih-primer-baze-Artur.pdf

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*, 1031–1048. Pridobljeno s https://link.springer.com/content/pdf/10.1007/s10579-013-9216-5.pdf

Vintar, Š. (Ur.). (2010). *Slovenske korpusne raziskave* (1. natis ed.). Znanstvena založba Filozofske fakultete.

Zemljarič Miklavčič, J. (2008). *Govorni korpusi* (1. natis ed.). Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Zemljarič Miklavčič, J., Stabej, M., Krek, S., & Zwitter Vitez, A. (2015). *Kaj in zakaj v referenčni govorni korpus slovenščine*. Pridobljeno s http://www.korpus-gos.net/Content/Static/Kaj_in_zakaj_v_referencni_govorni_korpus_slovenscine.pdf

Corpus annotations for describing the context of speech events in Slovene speech corpora

The time-consuming and costly preparation of a speech corpus requires careful consideration of its composition and the categorization of the recorded metadata at the time of its design. The variety of speech events included in the national reference corpus should reflect the diversity of contemporary spoken language as much as possible. We will be interested in how to categorize the annotations used to describe the context of the speech events in order to achieve this representativeness without completely giving up the intercomparability of the data. A thoughtful design allows us to minimize the time-consuming tag adjustments required in subsequent corpus upgrades. We will carry out a comparative analysis of domestic and foreign speech corpora to critically evaluate the four basic categories of annotations used to describe the context of a speech situation. We will review the design of the foreign reference speech corpora FOLK, BNC2014, ORAL2013, the Spoken Dutch Corpus and C-ORAL-ROM and compare them with the current reference corpus of spoken Slovene Gos 2.1. We will problematize the selected annotations and highlight the more problematic areas that would require further consideration and potential future re-categorization.

Keywords: speech corpora, corpus design, speech events, categorization of annotations

LREC-COLING 2024: združena mednarodna konferenca računalniškega jezikoslovja ter jezikovnih virov in evalvacije

Mojca BRGLEZ,[1] Matej KLEMEN,[2] Luka TERČON,[1, 2] Špela VINTAR[1]

1 Filozofska fakulteta, Univerza v Ljubljani

2 Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

1. Uvod

Konec maja, natančneje 22.–25. 5., je v piemontskem Torinu potekala megakonferenca LREC-COLING 2024. Letošnja konferenca je bila edinstvena, saj je združila sicer samostojni mednarodni konferenci LREC in COLING, ki sta poprej na vsaki dve leti potekali v istem letu, odslej pa se bosta vsako leto izmenjevali. Prva je usmerjena v jezikovne vire in evalvacijo in je bila že 14. zapovrstjo. Konferenco organizira združenje ELRA[1], prej Evropsko združenje za jezikovne vire (European Language Resources Association), od letos preimenovano v regijsko neomejeno Združenje za jezikovne vire ELRA (ELRA Language Resources Association). Druga konferenca, COLING, se osredotoča na računalniško jezikoslovje in je pod okriljem Mednarodnega odbora za računalniško jezikoslovje ICCL[2] (International Committee on Computational Linguistics) potekala že tridesetič. Dogodek je bil umeščen v kongresni

1 https://www.elra.info/

2 https://ufal.mff.cuni.cz/iccl

Brglez, M. et al.: LREC-COLING 2024: združena mednarodna konferenca računalniškega jezikoslovja ter jezikovnih virov in evalvacije. Slovenščina 2.0, 12(1): 95–105.

1.19 Recenzija, prikaz knjige, kritika / Review, book review, critique

DOI: https://doi.org/10.4312/slo2.0.2024.1.95-105

center kompleksa Lingotto, ki je nekoč služil kot tovarna avtomobilov FIAT. Konferenca je potekala hibridno, del udeležencev je dogajanje spremljal in prispevke predstavljal prek spleta. Sicer, morda zaradi že tako velikega števila prispevkov in dogajanja, ni bilo čutiti velikega mešanja med vsebinami/udeleženci v živo in tistimi na daljavo. Glavni del programa je zajemal vrsto predstavitev v obliki posterja ali govorne predstavitve. Na začetku in koncu konference sta se odvili uvodna in zaključna slovesnost, ki sta bili namenjeni predvsem formalnim vidikom. Poleg glavnega, vsebinskega dela konference sta bila organizirana tudi dva dogodka, namenjena mreženju in spoznavanju italijanske glasbe in hrane.

Na uvodni slovesnosti je programski odbor predstavil analizo sprejetih člankov. Na glavni del konference je bilo prijavljenih skupno kar 3.471 prispevkov, od tega je bilo sprejetih 44 % ali 1.554 prispevkov. Letošnji delež sprejetih prispevkov se je tako gibal med bolj restriktivno politiko COLING in bolj permisivno LREC. Države z največjim zastopanjem na konferenci so bile Kitajska, ZDA, Nemčija, Francija in Japonska. Slovenija je po drugi strani slavila po največjem deležu sprejetih prispevkov, saj je bilo pri prijavi uspešnih 12 od 15 prispevkov oziroma kar 80 %. Glede obiskanosti ocenjujemo, da se je konference v živo udeležilo približno 2000 ljudi, skupaj z udeleženci na daljavo pa bi številka znala znašati vsaj dvakrat toliko. Zbornik konference z rekordno dolžino 18.801 strani je na voljo na https://aclanthology.org/2024.lrec-main.

Poleg poročila in splošnih obvestil je predsednik odbora ELRE Simon Krek na uvodni slovesnosti podelil tudi nagrado Antonio Zampolli za izjemni prispevek k napredku jezikovnih tehnologij in evalvacije jezikovnih tehnologij na področju človeških jezikovnih tehnologij (»Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation within Human Language Technologies«). Prejemnik letošnje nagrade je Nizar Habash, redni profesor na Univerzi New York v Abu Dhabiju, ki se prvenstveno ukvarja z obdelavo arabskega jezika in arabskih narečij. Kot prejemnik nagrade je svoje delo predstavil v okviru posebnega predavanja na konferenci.

Zaključna slovesnost je bila namenjena predvsem zahvali vsem, ki so pomagali pri organizaciji megakonference: lokalnim organizator-

jem, vsebinskim koordinatorjem, prostovoljcem, recenzentom, avtorjem in obiskovalcem. Temu so sledila še vabila na naslednji, ločeni izvedbi konferenc LREC in COLING ter vabila na ostale prihajajoče jezikovnotehnološke konference.

2. Obkonferenčna srečanja

Pred začetkom glavne konference in dan po njej so potekali številni seminarji in kolokviji, skupaj se je zvrstilo kar 36 dogodkov. V nadaljevanju na kratko opišemo zgolj nekaj izbranih, ki smo se jih udeležili avtorji prispevka.

Med seminarji smo se najprej udeležili *Navigating the Modern Evaluation Landscape: Considerations in Benchmarks and Frameworks for Large Language Models,* na katerem so Leshem Choshen, Ariel Gera, Yotam Perlitz, Michal Shmueli-Scheuer in Gabriel Stanovsky temeljito predstavili metode za evalvacijo jezikovnih modelov in izzive, ki jih predstavljajo novejši generativni modeli. Z bolj dinamičnimi oblikami generiranih vsebin in novimi možnostmi uporabe namreč starejše metrike za vrednotenje, kot so natančnost, priklic, BLEU in podobne, po eni strani niso več ustrezne za vrednotenje vse bolj kreativnih vsebin, po drugi pa ne naslavljajo drugih, nejezikovnih zmogljivosti, ki jih nudijo jezikovni modeli, npr. jezikovni model kot pogovorni partner ali kot vršilec dejanj v digitalnem svetu. Še posebej veliko zanimanja je med poslušalci vzbudila debata o ugotovitvah številnih novih raziskav, ki razkrivajo, kako občutljivi so sodobni generativni jezikovni modeli na obliko poziva (angl. *prompt*), ki služi kot navodilo za ustvarjanje želenega rezultata. Že najmanjše variacije v obliki lahko namreč privedejo do velikih razlik v rezultatih, ki jih isti modeli dosežejo na različnih evalvacijskih lestvicah.

Na seminarju z naslovom *Towards a Human-Computer Collaborative Scientific Paper Lifecycle* so avtorji Qingyun Wang, Carl Edwards, Heng Ji in Tom Hope predstavili dobre in slabe plati uporabe umetne inteligence v okviru znanstvenega raziskovanja tako pri samem raziskovalnem procesu kot tudi pri pisanju prispevkov. Opisali so že preizkušene metode vključevanja sodobnih jezikovnih modelov za vsako od petih stopenj izdelave znanstvenega prispevka: pregled literature, iz-

delava hipotez, načrtovanje eksperimenta, pisanje in končni pregled. Izpostavljeni so bili številni izzivi, s katerimi se raziskovalci srečujejo ob uporabi umetne inteligence pri izvedbi raziskav. Predvsem sodobni jezikovni modeli pogosto generirajo popolnoma neresnične podatke (t. i. halucinacije), izkazujejo zelo malo domensko-specifičnega znanja in le redko proizvedejo podatke visoke kakovosti.

Med kolokviji smo najprej obiskali CogALex (*Cognitive Aspects of the Lexicon*), ki združuje raziskovalce kognitivnih vidikov jezika. Na kolokviju so bile predstavljene najnovejše raziskave o tem, kako se besede in pomeni hranijo, povezujejo in obdelujejo v našem miselnem leksikonu ter kako ta zapletena, široka in dinamična omrežja pomenov in besed predstavljati v leksikografskih virih. H kolokviju smo prispevali tudi slovenski raziskovalci, in sicer s člankom *How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene* (Žagar idr., 2024a).

Na prvi izdaji kolokvija CL4Health (*Patient-Oriented Language Processing*) smo lahko poslušali raziskave, ki se osredotočajo na obdelavo jezika za potrebe pacientov. Naraščajoče zanimanje za to temo namreč izhaja iz uporabe avtomatiziranih pomočnikov in pogostejšega iskanja zdravstvenih informacij na spletu, ki so danes veliko bolj dostopne. K področju sta prispevala tudi slovenska raziskovalca, in sicer z raziskavo *Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction* (Knez in Žitnik, 2024).

V okviru kolokvija NLPerspectives *The 3rd Workshop on Perspectivist Approaches to Natural Language Processing* je bilo govora o podatkovnem perspektivizmu. Temeljna ideja perspektivizma je ta, da nesoglasja pri označevanju podatkov niso vedno posledica napačnega odločanja označevalca ali slabih navodil za označevanje, pač pa so lahko naravno pričakovana in prisotna zaradi razlogov, kot so osebna prepričanja ali inherentna subjektivnost/dvoumnost problema, ki ga je mogoče interpretirati iz različnih perspektiv. Večina člankov je predstavljala nove vire, ki omogočajo učenje modelov na neagregiranih (torej ne-povprečenih, ne-združenih) oznakah, nekaj manj člankov pa je predstavljalo modele, naučene na podlagi tovrstnih oznak. Ideja perspektivizma je sicer prisotna tudi v mnogih člankih, predstavljenih na glavnem delu konference.

V Torinu je potekal tudi kolokvij ParlaCLARIN, med organizatorji katerega sta bila Darja Fišer, Institut za novejšo zgodovino/CLARIN ERIC, ter David Bordon, Filozofska fakulteta Univerze v Ljubljani. Na dogodku se je zvrstilo kar nekaj prispevkov, pri katerih so sodelovali slovenski raziskovalci: *Parliamentary Discourse Research in Political Science: Literature Review* (Skubic in Fišer, 2024), *Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines* (Çöltekin idr., 2024), *ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments* (de Jong idr., 2024), *Historical Parliamentary Corpora Viewer* (Kavčič idr., 2024).

Na kolokviju First Workshop on Reference, Framing, and Perspective so se predstavili še Ivačič idr. (2024) s člankom *Comparing News Framing of Migration Crises using Zero-Shot Classification*, na srečanju SIGUL za manj podprte jezike pa Ljubešić idr. (2024) s člankom *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining*.

3. Glavna konferenca

Na konferenci se je velika večina sprejetih prispevkov predstavila v obliki plakata (posterja), manjši del prispevkov pa z govornimi nastopi. Članki so bili umeščeni v 26 tematskih sklopov, od katerih so bili najbolj zastopani *Corpora and Annotation*, *Applications Involving LRs and Evaluation* ter *Evaluation and Validation Methodologies*, ki sodijo bolj v domeno konference LREC. Izbire je bilo resnično preveč, saj je običajno hkrati potekalo vsaj pet sej predstavitev in prav toliko razstav plakatov. Prvi dan je bilo mnogo teh sej na daljavo. Nekoliko neugodna je bila umestitev razstav plakatov v ločeni zgradbi, saj je to pomenilo manjšo zamudo s hojo do razstavne hale, kdaj pa nam je močan dež celo popolnoma onemogočil premik tja.

Ker je bilo poslušanja in branja vrednih raziskav ogromno, naj tukaj omenimo zgolj tiste, pri katerih so sodelovale slovenske ustanove in raziskovalci[3]. Med prispevke, ki so razgrnili nove jezikovne vire, lahko uvrstimo: *SUK 1.0: A New Training Corpus for Linguistic Annota-*

3 Število člankov s slovenskimi soavtorji, ki smo jih našli, se ne ujema z uradno statistiko, a uradnega kriterija štetja organizatorji niso navedli.

tion of Modern Standard Slovene (Arhar Holdt idr., 2024a), *Towards an Ideal Tool for Learner Error Annotation* (Arhar Holdt idr., 2024b), *SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation* (Klemen idr., 2024), *CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation* (Ljubešić in Kuzman, 2024), *The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings* (Mochtak idr., 2024), *A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection* (Dobranić idr., 2024), *MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations* (Gromann idr., 2024) in *Gos 2: A New Reference Corpus of Spoken Slovenian* (Verdonik idr., 2024). Preostala polovica prispevkov je predstavila nove metode in raziskave na področju jezika in jezikovnih tehnologij, to so bili: *A Computational Analysis of the Dehumanisation of Migrants from Syria and Ukraine in Slovene News Media* (Caporusso idr., 2024), *When Cohesion Lies in the Embedding Space: Embedding-Based Reference-Free Metrics for Topic Segmentation* (Ghinassi idr., 2024), *Denoising Labeled Data for Comment Moderation Using Active Learning* (Pelicon idr., 2024), *LLMSegm: Surface-level Morphological Segmentation Using Large Language Model* (Pranjić idr., 2024), *Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across Languages* (van Noord idr., 2024) in *SENTA: Sentence Simplification System for Slovene* (Žagar idr., 2024b).

ELRA in COLING sta najboljšim prispevkom letos podelila dve glavni nagradi. Nagrado za najboljši prispevek so prejeli Taiga Someya, Ryo Yoshida in Yohei Oseki za *Targeted Syntactic Evaluation on the Chomsky Hierarchy*, nagrado za najboljši študentski prispevek pa Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, in Rachel Bawden za *When Your Cousin Has the Right Connections: Unsupervised Bilingual Lexicon Induction for Related Data-Imbalanced Languages*. V prvem članku avtorji predstavijo izsledke o tem, kako dobro veliki jezikovni modeli prepoznajo različne lastnosti jezikov v hierarhiji Chomskega, v drugem pa avtorji izpostavijo omejitve obstoječih nenadzorovanih metod za ustvarjanje dvojezičnih slovarjev v jezikih z zelo malo jezikovnimi viri in predstavijo metodo, ki deluje bolje.

4. Vabljeni govorci

Vsak dan glavne konference je bilo moč poslušati vabljeno predavanje. Med tujima vabljenima govorcema sta bila Roger Levy z univerze MIT in Li Juanzi z univerze Tsinghua v Pekingu. Prvi je v sredo izvedel predavanje z naslovom *Large Language Models and Human Cognition*, na katerem je predstavil raziskave, ki naj bi nam prav s pomočjo jezikovnih modelov odstirale pogled na kompleksne procese človeške kognicije pri učenju in procesiranju jezika s tako imenovanim povratnim inženirstvom (ang. *reverse engineering*) pa tudi z razvijanjem modelov, ki simulirajo odraslo razumevanje otroškega govora.

Drugi dan konference je pred udeleženci nastopil Nizar Habash, prejemnik nagrade Antonio Zampolli, ki je poleg svojih znanstvenih udejstvovanj, predvsem na področju obdelave arabskega jezika, predstavil tudi svoje zasebne hobije, med drugim izmišljanje novih jezikov in umetniški projekt Palisra. V tem raziskuje možnost palestinsko-izraelske enotnosti preko ustvarjalnega združevanja obeh jezikov, pisav, kultur. Zelo zanimivo je, da na univerzi v Abu Dhabiju poučuje predmet, v katerem študenti snujejo povsem nove jezike, pri čemer se pokaže pomembnost spoznavanja prvin in struktur tujih jezikov ter kritična obravnava kdaj samoumevnih prvin in struktur lastnega maternega jezika.

Popoldne je predaval lokalni vabljeni govorec, Michele Loporcaro, ki deluje na univerzi v Zürichu. V predavanju *The Language Landscape of Italy as a Linguistic Data Mine* je predstavil delo z romanskimi narečji italijanskimi narečji, ki so večinoma nenapisani jeziki, a hranijo edinstvene jezikovne prvine in strukture, ki niso značilne niti za romansko skupino niti za indoevropsko vejo jezikov nasploh in tako raziskovalcem ponujajo neprecenljiv uvid v tipologijo jezikov.

Tretji dan konference je pred zbranimi nastopila Li Juanzi. V predavanju z naslovom *Knowledge in the LLM Era: Actuality, Challenge and Potentiality* je predstavila delo tamkajšnje raziskovalne skupnosti pri razumevanju zmogljivosti velikih jezikovnih modelov in govorila o izzivih, ki jih morajo modeli še premagati za globlje razumevanje in uporabo znanja.

5. Sklep

Ni presenetljivo, da so bili pri večini prispevkov megakonference LREC-COLING v središču pozornosti jezikovni modeli, ki vse bolj nadomeščajo tradicionalnejše pristope k obdelavi naravnega jezika, vse bolj pa vstopajo tudi na področja evalvacije in učenja jezikov. Skladno s tem se je več konferenčnih sekcij posredno ali neposredno posvečalo vprašanjem pristranskosti, etike in moralnih načel, ki naj bi jih jezikovni modeli vsebovali, ob vsem tem pa se še vedno, morda celo vse bolj, kažejo razlike med angleščino kot dominantnim jezikom in dolgim repom manjših jezikov, ki iz angleščine »podedujejo« ne le znanje, ampak tudi celo vrsto kulturnih in družbenih norm.

Spodbudno pa je, da je bilo na konferenci opaziti zelo raznolik nabor prispevkov iz najrazličnejših jezikov. Raziskovalci nekaterih manj zastopanih jezikov, kot je na primer kazaščina, so na konferenci namreč predstavili prve, pionirske korake na področju razvoja jezikovnih virov in orodij. Kljub v svetovnem merilu izjemno majhnemu številu govorcev nam to kaže, da slovenščina na področju jezikovnotehnoloških raziskav nikakor ni v slabem položaju.

Naslednjo izvedbo konference COLING bodo med 19. in 24. januarjem 2025 gostili Združeni arabski emirati, do naslednje konference LREC v letu 2026 na še nerazkriti lokaciji pa bo treba še nekoliko počakati. Do takrat vas vabimo k branju prispevkov s slovenskih ustanov, navedenih v seznamu literature.

Literatura

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon, L., & Žitnik, S. (2024a). SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 15428–15435). Pridobljeno s https://aclanthology.org/2024.lrec-main.1340/

Arhar Holdt, Š., Erjavec, T., Kosem, I., & Volodina, E. (2024b). Towards an Ideal Tool for Learner Error Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 16392–16398). Pridobljeno s https://aclanthology.org/2024.lrec-main.1424

Caporusso, J., Brglez, M., Hoogland, D., Koloski, B., Purver, M., & Pollak, S. (2024). A Computational Analysis of the Dehumanisation of Migrants from Syria and Ukraine in Slovene News Media. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 199–210). Pridobljeno s https://aclanthology.org/2024.lrec-main.18/

Çöltekin, Ç., Kopp, M., Meden, K., Morkevicius, V., Ljubešić, N., & Erjavec, T. (2024) Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 94–100). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.14/

de Jong, A., Kuzman, T., Larooij, M., & Maarten, M. (2024). ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC--COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 110–115). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.16

Dobranić, F., Evkoski, B., & Ljubešić, N. (2024). A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 695–703). Pridobljeno s https://aclanthology.org/2024.lrec-main.61/

Ghinassi, I., Wang, I., Newell, C., & Purver, M. (2024). When Cohesion Lies in the Embedding Space: Embedding-Based Reference-Free Metrics for Topic Segmentation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 17525–17536). Pridobljeno s https://aclanthology.org/2024.lrec-main.1524/

Gromann, D., Gonçalo Oliveira, H., Pitarch, L., Apostol, E.-S., Bernad, J., Bytyçi, E., ..., & Zdravkova, K. (2024). MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 11783–11793). Pridobljeno s https://aclanthology.org/2024.lrec-main.1029/

Ivačič, N., Purver, M., Lind, F., Pollak, S., Boomgaarden, H., & Bajt, V. (2024). Comparing News Framing of Migration Crises using Zero-Shot Classification. *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024* (str. 18–27). Pridobljeno s https://aclanthology.org/2024.rfp-1.3

Kavčič, A., Stojanoski, M., & Marolt, M. (2024). Historical Parliamentary Corpora Viewer. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC--COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 127–132). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.19

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2024). SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 14859–14870). Pridobljeno s https://aclanthology.org/2024.lrec-main.1294/

Knez, T., & Žitnik, S. (2024). Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction. *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024* (str. 131–136). Pridobljeno s https://aclanthology.org/2024.cl4health-1.16/

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 3271–3282). Pridobljeno s https://aclanthology.org/2024.lrec-main.291/

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (str. 189–203). Pridobljeno s https://aclanthology.org/2024.sigul-1.23

Mochtak, M., Rupnik, P., & Ljubešić, N. (2024). The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 16024–16036). Pridobljeno s https://aclanthology.org/2024.lrec-main.1393/

Pelicon, A., Karan, M., Shekhar, R., Purver, M., & Pollak, S. (2024). Denoising Labeled Data for Comment Moderation Using Active Learning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 4626–4633). Pridobljeno s https://aclanthology.org/2024.lrec-main.413/

Pranjić, M., Robnik-Šikonja, M., & Pollak, S. (2024). LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 10665–10674). Pridobljeno s https://aclanthology.org/2024.lrec-main.933/

Skubic, J., & Fišer, D. (2024). Parliamentary Discourse Research in Political Science: Literature Review. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 1–11). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.1/

van Noord, R., Kuzman, T., Rupnik, P., Nikola Ljubešić, N., Esplà-Gomis, M., Ramírez-Sánchez, G., & Toral, A. (2024). Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across Languages. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 5221–5234). Pridobljeno s https://aclanthology.org/2024.lrec-main.465/

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 7825–7830). Pridobljeno s https://aclanthology.org/2024.lrec-main.691/

Žagar, A., Brglez, M., & Vintar, Š. (2024a). How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024* (str. 42–48). Pridobljeno s https://aclanthology.org/2024.cogalex-1.5/

Žagar, A., Klemen, M., Robnik-Šikonja, M., & Kosem, I. (2024b). SENTA: Sentence Simplification System for Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 14687–14692). Pridobljeno s https://aclanthology.org/2024.lrec-main.1279/

Pomembnost realistične evalvacije: primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom

Timotej PETRIČ
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

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

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Med napake pri pisanju v standardni slovenščini sodi raba neustreznega slovničnega sklona ali števila. S pomočjo velikega jezikovnega modela SloBERTa smo razvili novo metodologijo za strojno prepoznavo tovrstnih težav, ki smo jo preizkusili na neustrezni rabi tožilnika namesto rodilnika in množine namesto dvojine. Za vrednotenje in spreminjanje besednih oblik v vhodnih povedih smo uporabili standardna orodja za obdelavo naravnega jezika, kot sta oblikoskladenjski označevalnik CLASSLA-Stanza in leksikon besednih oblik Sloleks. Predlagani popravki temeljijo na statistiki besednih oblik pri uporabi napovedovanja maskirane besede z velikim jezikovnim modelom. Zaradi pomanjkanja zadostne količine učnih podatkov smo napovedne modele učili na umetno generiranih napakah. Uspešnost strojnega popravljanja smo najprej ovrednotili na umetnih množicah in korpusu Lektor, kasneje pa še na novoustvarjeni evalvacijski množici Šolar-Eval. Evalvacija na prvih dveh množicah

Petrič, T; Arhar Holdt, Š.; Robnik-Šikonja, M.: Pomembnost realistične evalvacije: primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom. Slovenščina 2.0, 12(1): 106–130.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2024.1.106-130

je pokazala visoko uspešnost razvite metodologije (zaznanih več kot 90 % napačno nastavljenih besed), Šolar-Eval pa je razkril mnogo slabšo uspešnost na realističnih podatkih (zaznanih le 29,5 % težav tipa rodilnik-tožilnik in 11,4 % težav tipa dvojina-množina). V celoti rezultati kažejo na nevarnost pretiranega prilagajanja podatkovnim množicam in pomembnost evalvacije na ciljno grajenih avtentičnih podatkih, ki pa so za slovenščino še vedno pomanjkljivi.

Ključne besede: strojno slovnično pregledovanje, slovnični sklon, slovnično število, veliki jezikovni modeli, evalvacija

1 Uvod

Slovnični pregledovalniki – programi, ki preverjajo slovnično ustreznost pisnih besedil, opozarjajo na potencialne jezikovne težave in predlagajo popravke – so ena od temeljnih jezikovnih tehnologij in predstavljajo pomembno digitalno infrastrukturo za sodobne jezike, tudi slovenščino (Krek, 2023). Tipična težava, na katero slovnični pregledovalniki opozarjajo, je raba besednih oblik, ki glede na kontekst ne ustrezajo po sklonu, številu ali kaki drugi slovnični lastnosti. Za preizkus novega pristopa k strojnemu slovničnemu pregledovanju smo izbrali neustrezno rabo tožilnika namesto rodilnika ter množine namesto dvojine. Tovrstne menjave oblik so v procesu razvoja jezikovnih kompetenc v standardni slovenščini relativno pogoste; frekvenčni seznam jezikovnih popravkov v razvojnem korpusu Šolar 3.0 (Arhar Holdt idr., 2022b) pokaže, da so popravki menjave rodilnika in tožilnika na 2. mestu vseh oblikoslovnih kategoričnih popravkov (predstavljajo 341 od 2.916 popravkov), popravki dvojine in množine pa na 5. mestu (246 od 2.916 popravkov), pri čemer podatki kažejo tudi trdoživost obravnavanih jezikovnih problemov, ki se pojavljata tako v osnovni kot vse do konca srednje šole.[1]

V zadnjih letih je povečana zmogljivost vzporednega računanja z grafičnimi procesorji povzročila nov val uspehov na področju umetne inteligence, tudi pri obdelavi naravnega jezika. Trenutno najpomembnejša arhitektura nevronskih mrež, ki prevladujejo pri obdelavi jezika,

1 Na ostalih prvih mestih so heterogenejše skupine jezikovnih težav: na 1. mestu različne menjave sklonov, ki v označevalnem sistemu niso dobile lastne označevalne kategorije, na 3. mestu raznoliki primeri menjav ednine in množine, na 4. mestu pa raznoliki popravki glagolskega časa.

je *transformer* (Vaswani idr., 2017). Modeli, kot je BERT (Devlin idr., 2019), se z napovedovanjem manjkajočih besed v povedih na veliki količini gradiva naučijo jezikovnih značilnosti besedil. V članku smo tak slovenski model, imenovan SloBERTa (Ulčar in Robnik-Šikonja, 2021b), uporabili za generiranje predlogov popravkov neustrezno rabljenega slovničnega sklona in števila.

Ideja predlaganega pristopa poskusi izkoristiti zmožnost modela SloBERTa, da napoveduje maskirane (skrite) besede v povedi. Pristop najprej poišče potencialno problematične besede, ki jih želimo strojno preveriti, v našem primeru besede v tožilniku ali množini. Te besede obravnavamo kot skrite. S pomočjo modela SloBERTa napovemo možne besede na tem mestu povedi in s pomočjo označevalnika CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019) pridobimo njihove oblikoskladenjske lastnosti. Iz statistike napovedanih oblikoskladenjskih lastnosti nato napovemo najverjetnejši potencialni popravek oblike izvorne besede, npr. če je večina napovedanih besed v dvojini, potencialno problematična beseda pa je v množini, predlagamo njen popravek v dvojino. Besedno obliko z želenimi oblikoskladenjskimi lastnostmi, ki jo program predlaga kot ustrezno, pridobimo iz oblikoslovnega leksikona Sloleks (Dobrovoljc idr., 2019).

Pristop je metodološko nov, preliminarna evalvacija na umetnih podatkih je pokazala, da je uspešen in potencialno uporaben za različne vrste oblikoslovnih napak. Najprej smo ga ovrednotili na povedih iz korpusa Lektor (Popič, 2014), ko se je pojavila nova evalvacijska množica Šolar-Eval (Gantar idr., 2023), pa še na njej. Evalvacijo na korpusu Lektor smo izvedli tako, da smo v povedih nastavili različno število besed v (obravnavanem) napačnem sklonu ali številu. Izračunane metrike natančnosti, priklica in ocene F_1 so pokazale dobro pravilnost in potencialno praktično uporabnost predlaganih popravkov. Evalvacijo na korpusu Šolar-Eval smo izvedli kvalitativno, z identifikacijo in analizo zaznanih in nezaznanih napak. Ti rezultati so, v nasprotju s prejšnjimi, pokazali dokaj nizko uspešnost preizkušenega pristopa. Kljub temu menimo, da raziskava prinaša koristen uvid v pomemben jezikovnotehnološki problem, ki vključuje tudi pomen kakovostne evalvacije in ciljno grajenih evalvacijskih množic.

Članek je sestavljen iz sedmih razdelkov. V razdelku 2 predstavimo

sorodna dela, v razdelku 3 uporabljene jezikovne vire in tehnologije ter v razdelku 4 predlagano metodologijo napovedovanja slovničnih popravkov. V razdelku 5 opišemo postopek evalvacije, v razdelku 6 pa njene rezultate. Članek zaključimo z razdelkom 7, kjer povzamemo opravljeno delo in začrtamo smer za nadaljnje izboljšave.

2 Sorodna dela

Trenutno za slovenski jezik ne obstaja brezplačen slovnični pregledovalnik. Najbolj dodelano orodje, na voljo v uporabniško prijaznem vmesniku, ki napake in popravke tudi vizualizira, je komercialno razvita Amebis Besana.[2] Program, ki temelji na ročno sestavljenih jezikovnih pravilih in podatkovni zbirki Ases (Romih in Holozan, 2002), v brezplačni testni različici omogoča strojno preverbo krajših besedil (do 500 znakov). Vmesnik Besane je podoben komercialnim izdelkom za tuje jezike, kot sta Grammarly in ProWritingAid, ki zaznavajo in popravljajo težave pri rabi ločil, črkovanju in slovnici, prav tako pa ponujajo predloge za izboljšanje sloga pisanja, besedne raznolikosti ter jasnosti in učinkovitosti sporočil.

S strojnimi popravki jezikovnih napak z uporabo globokih nevronskih mrež se je ukvarjalo že več avtorjev. Božič je v okviru diplomskega dela razvil model za avtomatsko popravljanje vejic v slovenskem jeziku (Božič, 2020), ki je pravilno napovedal 92,5 % primerov. Njegov še nekoliko izboljšan program Vejice je prosto dostopen na spletni strani Centra za jezikovne vire in tehnologije Univerze v Ljubljani.[3] Rizvič se je ukvarjal z avtomatskim postavljanjem ločil v tekstu, pridobljenim iz prepoznavalnika govora (Rizvič, 2020). Najboljše rezultate je dosegel z uporabo vektorskih vložitev ELMo in modela BERT. Dosežena ocena F_1 je bila 91,0 %, 91,6 % in 72,0 % za napovedovanje mesta vejice, pike in vprašaja. S predlogi popravkov končnih ločil se je ukvarjal Velikonja (Velikonja, 2021), ki je s pomočjo modela SloBERTa napovedoval tip in mesto postavitve končnih ločil. Naučeni model je dosegel oceno F_1 96,4 % za postavljanje pike in 85,1 % za postavljanje vprašaja. Napovedovanje klicaja ni bilo uspešno.

2 Spletna stran programa: https://besana.amebis.si/.
3 Dostopno na https://orodja.cjvt.si/vejice/.

S pomočjo modelov SloBERTa in SloT5 (Ulčar in Robnik-Šikonja, 2023) ter uporabo orodja CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019) in leksikona Sloleks (Dobrovoljc idr., 2019) je Mokotar razvil metodologijo za zaznavanje, prepoznavanje in popravljanje različnih vrst jezikovnih napak (Mokotar, 2023), pri čemer je bila uporabljena nekoliko poenostavljena tipologija napak iz korpusa šolskih besedil Šolar (Arhar Holdt idr., 2022a). Modela zaznavanja in prepoznavanja sta dosegla oceni F_1 88 % in 14 %, model za popravljanje pa oceno GLEU 50 %.

Z napovedovanjem in popravljanjem jezikovnih napak v drugih jezikih se je ukvarjalo več avtorjev. Rozovskaya idr. (2014) so razvili sistem za prepoznavanje napačne oblike glagola za angleščino s klasičnimi pristopi strojnega učenja. V zadnjem času se za detekcijo in korekcijo napak uporabljajo izključno nevronski pristopi, predvsem temelječi na velikih jezikovnih modelih. Zhang idr. (2020) so prilagodili arhitekturo modela BERT za napovedovanje jezikovnih napak na primeru kitajščine. Pred kratkim so pregledni članek o strojnem pregledovanju jezikovnih napak pripravili Bryant idr. (2023).

V zadnjem času so se pojavili tudi pristopi s še večjimi velikimi jezikovnimi modeli, kot sta GPT-3 (Brown idr., 2020) in LLaMA (Touvron idr., 2023). V javnosti znani izdelki, kot je ChatGPT, ki v času priprave prispevka uporablja GPT-3.5, so naučeni predvsem na angleškem jeziku. Za slovenščino še ne obstajajo dovolj obsežni jezikovni viri, s katerimi bi naučili primerljivo zmogljiv jezikovni model, zato je uporaba za popravljanje slovenskih besedil trenutno omejena. Evalvacije (Fang idr., 2023; Wu idr., 2023) so pokazale, da je ChatGPT nagnjen k pretiranemu popravljanju oz. spreminjanju izvornega besedila (ang. *over-correcting*), kar je lahko moteče, kadar želimo minimalno in transparentno jezikovno intervencijo (npr. za potrebe jezikovne didaktike).

V tem članku preizkušeni pristop se od zgoraj omenjenih del razlikuje po novi metodologiji, ki uporablja statistiko predlaganih besed maskirnega jezikovnega modela.

3 Uporabljeni viri in tehnologije

Predlagana metodologija za strojno slovnično pregledovanje temelji na jezikovnih virih in tehnologijah, ki jih opisujemo v tem razdelku. V

razdelku 3.1 opišemo jezikovni model BERT, v razdelku 3.2 pa njegovo slovensko različico SloBERTa, ki jo uporabljamo za napovedovanje potencialnih zamenjav maskiranih besed. V razdelku 3.3 predstavimo zbirko orodij CLASSLA-Stanza, ki jo uporabljamo za segmentacijo na povedi in besede ter za oblikoskladenjsko označevanje. Na koncu, v razdelku 3.4, predstavimo še oblikoslovni leksikon Sloleks.

3.1 Model BERT

BERT (Devlin idr., 2019) je vnaprej naučeni nevronski maskirni jezikovni model, ki temelji na arhitekturi *transformer* (Vaswani idr., 2017). Kot odprtokodno ogrodje je na voljo za različne naloge strojne obdelave naravnega jezika. Naučen je s pomočjo velikih besedilnih korpusov in zaradi svoje arhitekture in načina maskiranega učenja vsebuje predstavitev besed v kontekstu. To znanje lahko uporabimo za reševanje mnogih nalog, med drugim strojno označevanje sentimenta, odgovarjanje na vprašanja in tudi napovedovanje manjkajočih besed vhodnega besedila, kar smo uporabili v našem delu.

3.2 Model SloBERTa

Veliki maskirni jezikovni model SloBERTa (Ulčar in Robnik-Šikonja, 2021a; 2021b) uporablja arhitekturo robustne inačice modela BERT, imenovane RoBERTa (*A Robustly Optimized BERT Pretraining Approach*) (Liu idr., 2019). Arhitektura RoBERTa pri učenju namesto statičnega maskiranja besed uporablja dinamično maskiranje, kar pomeni, da se maskiranje besed ne zgodi samo enkrat – v fazi predpriprave vhodnih besedil – ampak večkrat, med posameznimi epohami učenja; RoBERTa tudi opusti nalogo napovedovanja, ali sta dve vhodni povedi sosednji v besedilu, ki je prisotna v modelu BERT. SloBERTa je enojezikovni model, naučen na 3,47 milijarde pojavnic (besed in ločil) iz vhodnih besedil. Slovar pojavnic, ki jih model uporablja za pretvorbo besedila v sezname vektorskih vložitev, ima 32.000 vnosov. Celotno učenje na besedilih izbranih slovenskih korpusov, kot so Gigafida 2.0 (Krek idr., 2020), siParl 2.0 (Pančur idr., 2020) in KAS (Erjavec idr., 2019), je obsegalo 98 epoh. Implementacija modela SloBERTa je med drugim na voljo v programski knjižnici HuggingFace, ki omogoča

odprtodostopen prenos modela ter preprosto uporabo v programskem jeziku Python.[4]

3.3 Zbirka orodij CLASSLA-Stanza

CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019) je zbirka orodij za procesiranje in jezikoslovno označevanje besedil. Med drugim omogoča segmentacijo, lematizacijo, oblikoskladenjsko in skladenjsko označevanje ter označevanje imenskih entitet v (standardnih in nestandardnih) slovenskih, hrvaških, srbskih, bolgarskih in deloma tudi makedonskih besedilih. Temelji na knjižnici Stanza (Qi idr., 2020). Označevalnik CLASSLA-Stanza v našem delu uporabljamo za segmentacijo besedila na povedi in besede ter oblikoskladenjsko označevanje besednih oblik po sistemu Multext-East v6 (Erjavec, 2017).[5] Sistem vsebuje nabor oznak (na kratko *oznake MSD*), ki določijo besedno vrsto, nato pa niz pri tej besedni vrsti izkazanih slovničnih lastnosti, kot so denimo spol, sklon in število.

3.4 Leksikon Sloleks

Sloleks je odprtodostopni leksikon besednih oblik za slovenščino, ki poleg osnovne oblike besede vsebuje nabor pregibnih oblik, podatke o pogostosti leme in pregibnih oblik iz referenčnega pisnega korpusa, zbir standardnih in nestandardnih oblikoslovnih variant ter povezave na besedotvorno sorodne besede (Dobrovoljc idr., 2015). Verzija 2.0,[6] ki je dostopna na repozitoriju CLARIN.SI (Dobrovoljc idr., 2019), vsebuje 100.805 leksikonskih enot. Oblikoslovni leksikon v našem delu uporabljamo za pridobivanje besednih oblik v želenem sklonu ali številu.

Uporabljena podatkovna množica ima oblikoskladenjske informacije po sistemu MULTEXT-East (Erjavec, 2017). V stolpcih po vrsti vsebuje besedne oblike, leme, oznake MSD in frekvence pojavitve. V zadnjem stolpcu so kategorije, ki jih vsebujejo oznake MSD, izpisane z besedo. Za lažjo predstavo na Sliki 1 prikažemo izsek dela leksikona iz tekstovne datoteke.

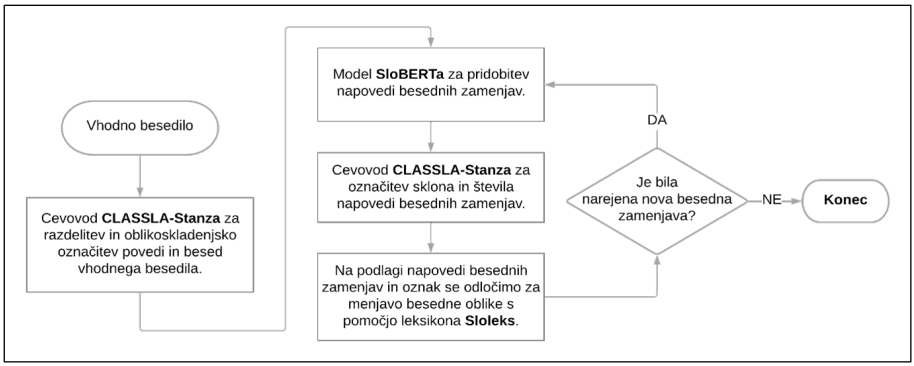
4 Dostopno na https://huggingface.co/EMBEDDIA/sloberta.
5 Verzija 6 je dostopna na http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.
6 V času med razvojem metodologije in objavo prispevka je bila na CLARIN.SI objavljena tudi že nadgrajena različica Sloleks 3.0.

BESEDA	LEMA	MSD	FREKVENCA	PODROBNEJŠI OPIS MSD
abscesa	absces	Ncmsg	0.000011	Noun Type=common Gender=masculine Number=singular Case=genitive NOUN Case=Gen Gender=Masc Number=Sing
absces	absces	Ncmsn	0.000063	...
abscesi	absces	Ncmpn	0.000031	...

Slika 1: Primer informacij v leksikonu Sloleks (vir: Dobrovoljc idr., 2019). Izmed podatkov uporabljamo predvsem stolpce BESEDA, LEMA in MSD. Glede na lemo in želeno spremembo sklona in/ali števila s pomočjo oznake MSD poiščemo ustrezno obliko obravnavane besede.

4 METODOLOGIJA SLOVNIČNEGA POPRAVLJANJA

V tem poglavju predstavimo cevovod za napovedovanje slovničnih popravkov in opišemo posamezne gradnike. Shema cevovoda je prikazana na Sliki 2.



Slika 2: Shema cevovoda za napovedi jezikovnih popravkov. Model SloBERTa napove 10 najverjetnejših besed namesto zamaskirane. To so lahko različne besede (ne le dejansko zamaskirana), vendar za vse, na podlagi konteksta, s pristopom CLASSLA-Stanza napovemo sklon in število. Na podlagi teh napovedi določimo najverjetnejši sklon in število, potem pa za ta sklon in število v Sloleksu poiščemo pravo obliko za zamaskirano besedo.

4.1 Segmentacija in izbor kandidatov za popravke

Vhodno besedilo najprej obdelamo z orodji CLASSLA-Stanza za segmentacijo na povedi in besede. Za vse besede vsake povedi določimo še njihove oblikoskladenjske oznake. Z modelom SloBERTa (ali

SloBERTaMCD, opisanim v razdelku 4.2) za vsako besedo vsake povedi, ki ji je mogoče določiti sklon ali število, preverimo, če je možno pripisati tudi drugačen sklon ali število glede na vrednosti parametrov *caseXY* in *numberXY*, npr. če je parameter *case42* nastavljen na vrednost 1, zamenjave besed iz tožilnika v rodilnik ne iščemo, za vrednosti med 0 in 1 pa jih poskusimo najti; podobno velja npr. za parameter *number32*, kjer vrednost 1 pomeni, da zamenjav iz množine v dvojino ne iščemo, za vrednosti manjše od ena pa sprožimo iskanje potencialnih popravkov. Parameter *caseXY* določa verjetnostni prag prepričanosti modela v trenutni sklon besede.

Za besedo v tožilniku in mejno vrednost parametra nastavljeno na 0,5, torej *case42=0,5* bi denimo iskali njeno zamenjavo v rodilniku in jo uveljavili, če bi prepričanost napovednega modela za to spremembo presegla prag 0,5. Če bi za vse mejne vrednosti izvornega sklona rodilnik veljalo *case4{1,2,3,5,6}=1*, zamenjav za besede v tožilniku ne bi iskali. Na ta način iščemo le zamenjave besed v sklonu ali številu, za katere smo naučili napovedne modele. Trenutno iščemo popravke le za *case42* in *number32* – torej popravke sklona iz tožilnika v rodilnik in števila iz množine v dvojino.

Mejne vrednosti smo naučili v procesu optimizacije hiperparametrov. Omejili smo se le na iskanje optimalnih mejnih vrednosti za *case42, number32* in še za nastavitev kombinacije obeh mejnih vrednosti hkrati: *case42* in *number32*.

4.2 Zanesljivejše napovedi s SloBERTaMCD

Namesto modela SloBERTa lahko za napovedi zamenjav maskiranih besed iz vhodnih besedil uporabimo postopek SloBERTaMCD. Postopek uporablja model SloBERTa (razdelek 3.2), ki ga uporabimo na način Monte Carlo Dropout (MCD) (Miok idr., 2022). Pristop uporablja izpust nevronov (angl. *dropout*) v fazi napovedovanja kot tehniko *regularizacije* (Gal in Ghahramani, 2016). To storimo tako, da vsako poved z maskirano besedo 50-krat obdelamo z modelom SloBERTa in na koncu vrnemo vektor povprečij vseh napovedi besednih zamenjav. Pri uporabi MCD ima izhodni vektor napovedi zamenjav bolje kalibrirane verjetnosti napovedi uteži, kar lahko pomaga pri razvrstitvi in izboru najbolj primernih zamenjav.

4.3 Možne zamenjave

Če za besedo iščemo zamenjave, jo zamaskiramo in celotno poved v obliki niza dodamo v nov seznam povedi (ki imajo vse po eno maskirano besedo). Ta seznam povedi obdelamo z modelom SloBERTa (ali SloBERTaMCD). Kot izhod za vsako vhodno poved s po eno maskirano besedo prejmemo po 10 napovedi najbolj primernih besednih zamenjav (to so lahko poljubne besede). Vsaka napoved ima pripisano verjetnost in besedno zamenjavo.

Za vsako napoved besedne zamenjave pridobimo podatke o sklonu in številu s pomočjo orodja CLASSLA-Stanza. Tokrat je povedi 10-krat več kot pri začetnem označevanju, saj imamo za vsako poved s po eno maskirano besedo 10 možnih besednih zamenjav. CLASSLA-Stanza nam oblikoskladenjsko označi le primere, ki so na mestu izvorno maskirane besede. Za vsako besedo, ki ji je možno pripisati sklon ali število, imamo zdaj po 10 predlogov besednih zamenjav in njihove oblikoskladenjske lastnosti. Kot končni predlog zamenjave izberemo sklon in število, ki imata največjo vsoto verjetnosti. Če imata predlagana sklon in število dovolj veliko vsoto verjetnosti v primerjavi z mejnimi vrednostmi, bomo za izhodiščno besedno obliko predlagali zamenjavo.

4.4 Ponavljanje in ustavitev postopka

Zgoraj opisani postopek ponavljamo, dokler sistem še predlaga nove besedne zamenjave. V vsakem dodatnem obhodu obdelujemo le besede, za katere še nimamo besednih zamenjav. Ponavljanje je potrebno, ker v enem obhodu ponavadi ne naslovimo vseh potencialnih težav, npr. popravimo samostalnik, pridevnika pred njim pa ne. Primer zaznanih napak (v rdečem), predlaganih popravkov (v zelenem) in izpisa po koncu procesiranja prikazujemo spodaj. Kot vidimo v zadnjih dveh povedih, se lahko zgodi tudi, da sistem ne zazna vseh napak v besedilu.

Hotel je preizkusiti dve testne|testni vožnje|vožnji z avtom. Opazil je, da mu manjkata dve iztočnice|iztočnici. Včeraj nisem videl Petro|Petre. Že dolgo nisem jedel tako dobro|dobre solato|solate. Ne morem jo videti! Naredil je dve čudne|čudni napake.

Hotel je preizkusiti dve **testni vožnji** z avtom. Opazil je, da mu manjkata **dve iztočnici.** Včeraj nisem videl **Petre.** Že dolgo nisem jedel tako **dobre solate.** Ne morem **jo** videti! Naredil je dve **čudni napake.**

5 Postopek evalvacije

V razdelku 5.1 opišemo korpus Lektor, ki ga uporabljamo za evalvacijo umetno ustvarjenih podatkov in učenje mejnih vrednosti parametrov. Šolar-Eval, ki ga uporabljamo za kvalitativno evalvacijo na avtentičnem gradivu, opišemo v razdelku 5.2. Postopek nastavljanja napačnih besed za evalvacijo opišemo v razdelku 5.3. Razdelek 5.4 predstavi uporabljene evalvacijske metrike.

5.1 Korpus Lektor

Lektor (Popič, 2014) je korpus lektoriranih avtorskih besedil in prevodov. V njem so zbrana novejša neliterarna (strokovna in poljudnoznanstvena) besedila različnih avtorjev in prevajalcev, ki so jih lektorirali različni lektorji. Vsako korpusno besedilo vsebuje metapodatke o avtorju, publikaciji, lektorju, pripisane pa ima tudi leme, oblikoskladenjske oznake in vsebinske kategorije lektorskih popravkov. Kot podatkovno bazo v formatu XML smo ga za raziskovalne namene dobili od avtorja korpusa Damjana Popiča.

Korpus smo pretvorili v tekstovno datoteko, kjer je v vsaki vrstici po ena lektorirana poved – skupno 28.744 povedi. Povedi so po vrsticah razvrščene naključno. Datoteko smo razdelili na dva dela: prvi del ima 80 % oz. 22.995 vseh povedi in je bil uporabljen za učenje mejnih vrednosti parametrov, drugi ima preostalih 20 % oz. 5.749 povedi in je bil uporabljen pri evalvaciji programske rešitve.

5.2 Korpus Šolar-Eval

Med delom se je izkazalo, da obstoječi korpusi z jezikovnimi popravki ne vsebujejo dovolj podrobno ali dosledno označenih podatkov, da bi bili neposredno uporabni pri razvoju slovenskih črkovalnikov in slovničnih pregledovalnikov (Gantar idr., 2023: 91). Kot rešitev je bila pripravljena evalvacijska množica Šolar-Eval (Arhar Holdt idr., 2023). Ta

vsebuje 109 esejev, ki so jih napisali slovenski osnovnošolci in srednješolci, ter vključuje 9.808 jezikovnih popravkov, ki so jih podrobno in usklajeno označili jezikoslovci. Vsebuje tudi metapodatke o korpusnih besedilih in raznolike jezikoslovne oznake.

V evalvacijo smo vključili povedi, v katerih je najti avtentične popravke tožilnika v rodilnik (46 povedi), množine v dvojino (57 povedi) in oboje (1 poved). Številne od teh povedi vsebujejo tudi težave in popravke drugih jezikovnih značilnosti.

5.3 Nastavljanje napačnih besednih oblik za evalvacijo

Za model smo naučili tri kombinacije mejnih vrednosti parametrov: mejno vrednost *case42, number32* in sočasno nastavitev kombinacije obeh mejnih vrednosti *case42* in *number32* (opisano v razdelku 4.1). Vsako od kombinacij mejnih vrednosti parametrov smo evalvirali posebej. Za evalvacijo smo izbrali 3.030 povedi izmed 5.749 iz testne množice korpusa Lektor (opisana v razdelku 5.1). V množici je bilo skupno 61.180 besed, med njimi 7.629 besed v tožilniku, 9.060 v rodilniku, 12.732 v množini in 690 v dvojini.

Za vsako besedo, ki je bila v iskanem sklonu ali številu, smo poiskali vse možne oblike te besede v neustreznem sklonu oz. številu in jih shranili v seznam. Na primer, ko smo evalvirali model z nastavljeno mejno vrednostjo *case42*, smo za vse besede v rodilniku poiskali alternativne besedne oblike v tožilniku. Podobno smo storili tudi za evalviranje modela z nastavljeno mejno vrednostjo *number32*. Ko smo evalvirali model s hkratno nastavitvijo obeh mejnih vrednosti, smo za vsako besedo v seznam shranili vse možne alternativne besedne oblike (tožilnik ali množina). Shranjevali smo samo alternativne besedne oblike, ki so se razlikovale od izvornih (ne pa tudi enakopisnih oblik).

Napačne besedne oblike smo nastavili, ko je bilo možno za izbrano poved pridobiti vsaj toliko besed z alternativnimi besednimi oblikami, kot smo želeli. Če je bilo za določeno besedo na voljo več alternativnih oblik, smo naključno izbrali le eno. Nastaviti napačno besedno obliko torej pomeni, da smo za ustrezno obliko v rodilniku nastavili še neustrezno obliko v tožilniku. Kasneje je program za vse

primere, ki so bili v tožilniku že izvorno ali pa so imeli nastavljene tožilniške oblike, iskal popravke iz tožilnika v rodilnik (oz. iz množine v dvojino).

5.4 Ocenjevanje modela

Najprej smo programsko rešitev testirali s pomočjo podatkov, ustvarjenih iz korpusa Lektor. Za evalvacijo napovednih modelov smo v povedih nastavili različno število besed v želenem napačnem sklonu ali številu. Izbrali smo samo povedi, v katerih se je neustrezna besedna oblika formalno razlikovala od ustrezne (ne pa primerov z enakopisnimi oblikami). Povedi smo obdelali s cevovodom, opisanim v razdelku 4, in dobili strojno predlagane slovnične popravke.

Rezultate smo razdelili v štiri skupine. Besede, ki so imele nastavljeno napačno obliko in dobile ustrezen popravek, smo označili za dejansko pozitivne primere (angl. *true positive oz. TP*). Besede, ki so imele nastavljeno napačno obliko in nepravilen ali nenastavljen popravek, smo označili za napačno negativne primere (angl. *false negative oz. FN*). Besede, ki niso imele nastavljene napačne oblike, vendar so imele popravek, smo označili za napačno pozitivne primere (angl. *false positive oz. FP*). Besede, ki niso imele nastavljene napačne oblike in tudi ne popravka, pa smo označili za dejansko negativne (angl. *true negative oz. TN*).

Iz teh štirih skupin smo izračunali metrike natančnost (angl. *precision*), priklic (angl. *recall*) in oceno F_1 (Jurafsky in Martin, 2024).

V naslednjem primeru povedi je 14 besed. Zeleno so obarvane izvorne besede, rdeče so nastavljene napačne besedne oblike in modre so predlogi popravkov, ki so izhod iz programa. Besedi *znamenite* in *trikotne* sta šteti kot dejansko pozitivni. Beseda *Amerikami* je šteta za napačno negativni primer, saj program ni predlagal nobenega popravka. Ostale besede so štete za dejansko negativne.

Trgovina s sužnji je postala del znamenite|znamenito|znamenite trikotne|trikotno|trikotne trgovine med Amerikama|Amerikami, Evropo in Afriko.

V naslednjem koraku smo postopek evalvirali še na avtentičnem gradivu korpusa Šolar-Eval. Korpusne povedi smo obdelali s

cevovodom, opisanim v razdelku 4, in dobili strojno predlagane slovnične popravke. Obravnavane 104 povedi smo strojno slovnično pregledali še s spletno različico slovničnega pregledovalnika Amebis Besana 4.32.8. Oboje rezultate smo primerjali z jezikoslovnimi popravki, podanimi v korpusu, in ročno analizirali podobnosti in razlike.

Med analizo smo popravke razdelili v dve skupini: (a) primeri, v katerih je jezikovna težava opazna in rešljiva na ravni same povedi, npr. *Potem pa sva se ob skodelici mamine kave pogovarjale o imenih, barvi las in podobno*, (b) primeri, kjer je za popravek treba poznati širši besedilni kontekst ali pa sta v standardnem jeziku sprejemljivi tako popravljena kot nepopravljena različica, npr. *Zgodbe sem prebral ker sem jih moral*, pri čemer je iz predhodnega besedila razvidno, da je učenec prebral samo dve zgodbi. Ker gre za različno zahtevne primere, jih v Rezultatih prikažemo ločeno.

6 Rezultati

V tem razdelku predstavimo rezultate evalvacije. Najprej v razdelku 6.1 predstavimo rezultate na korpusu Lektor z nastavljenimi napakami tipa tožilnik-rodilnik in množina-dvojina. Sledijo rezultati evalvacije na avtentičnem gradivu v razdelku 6.2, nato pa še evalvacija hitrosti obdelave v razdelku 6.3.

6.1 Rezultati evalvacije na umetnem gradivu

V Tabelah 1, 2 in 3 so rezultati testiranja na množici s 3.030 povedmi, kjer smo primerjali tudi uporabo napovedovanja z uporabo metode MCD (oz. SloBERTaMCD iz razdelka 4.2) v primerjavi z običajnim modelom SloBERTa (v tabelah označeno z *MCD* in *neMCD*).

Ker vse povedi iz testne množice nimajo vedno iskanega števila besed v konfiguraciji pravilnega sklona oz. števila (npr. nimamo povedi z vsaj tremi besedami v pravilnem rodilniku, ki jim lahko nastavimo nepravilno obliko v tožilniku), se zgodi, da se za te povedi ne nastavijo nobene napačne besede. Te povedi vseeno vključimo v evalvacijo, ker model za besede v iskanem izvornem sklonu oz. številu še vedno poizkusi najti popravek – torej jih lahko beležimo kot potencialno napačno pozitivne primere. Prav tako teh povedi nismo

želeli odstraniti ali spreminjati zato, ker bi s tem vnašali dodatno pristranskost.

Metrike v Tabeli 1 prikazujejo rezultate popravljanj slovničnega števila. V primerjavi s popravljanjem sklonov (Tabela 2) prikazujejo občutno slabše rezultate. To je verjetno posledica dejstva, da je bilo besed v dvojini, ki bi jih lahko uporabili za nastavljanje napačnih besed in evalvacijo, v testni množici bistveno manj. Prav tako je raba dvojine v slovenščini manj pogosta kot raba tožilnika, zato so imeli tudi besedilni korpusi, ki so bili uporabljeni v procesu vnaprejšnjega učenja modela SloBERTa, na voljo manj tovrstnega gradiva.

Tabela 1: *Natančnost napovedi samo za popravljanje slovničnega števila* ***množina-dvojina*** *z različnim številom nastavljenih napačnih besed v povedi v množini*

	Št. napačnih besed v povedi					
	1		**2**		**3**	
	MCD	**neMCD**	**MCD**	**neMCD**	**MCD**	**neMCD**
Napačne besede	171	171	212	212	180	180
Natančnost	89,3	**90,4**	91,8	**92,2**	90,7	**90,9**
Priklic	**78,4**	77,2	**79,7**	78,3	75,6	**77,8**
F-ocena	**83,5**	83,3	**85,4**	84,7	82,4	**83,8**

**Opomba.* Model je imel nastavljeno mejno vrednostjo number32=0,056 (oz. number32=0,032 za neMCD).

Tabela 2: *Natančnost napovedi samo za popravljanje sklona* ***tožilnik-rodilnik*** *z različnim številom nastavljenih napačnih besed v povedi v tožilniku*

	Št. napačnih besed v povedi					
	1		**2**		**3**	
	MCD	**neMCD**	**MCD**	**neMCD**	**MCD**	**neMCD**
Napačne besede	1.994	1.994	2.858	2.858	2.853	2.853
Natančnost	98,2	**98,4**	98,5	**98,9**	98,4	**98,7**
Priklic	**96,2**	95,4	93,3	**93,5**	**92,6**	92,5
F-ocena	**97,2**	96,9	95,8	**96,1**	95,4	**95,5**

**Opomba.* Model je imel z nastavljeno mejno vrednostjo case42=0,0097 (oz. case42=0,007 za neMCD).

Tabela 3: *Natančnost napovedi za popravljanje sklona **tožilnik-rodilnik** in števila **množina-dvojina***

	Št. napačnih besed v povedi					
	1		2		3	
	MCD	**neMCD**	**MCD**	**neMCD**	**MCD**	**neMCD**
Napačne besede	2.051	2.051	2.986	2.986	3.075	3.075
Natančnost	97,3	**97,7**	98,0	**98,5**	98,0	**98,3**
Priklic	**95,4**	95,0	91,4	**92,6**	91,4	**92,4**
F-ocena	**96,3**	**96,3**	94,6	**95,5**	94,6	**95,3**

**Opomba.* Model je imel nastavljene mejne vrednosti case42=0,006 in number32=0,03 (oz. case42=0,0013 in number32=0,036 za neMCD).

Rezultati v Tabeli 3 kažejo, da se je model dobro naučil napovedovanja besednih zamenjav. V testni množici 3.030 povedi je približno 13-krat več besed v rodilniku kot besed v dvojini. Naučene mejne vrednosti *case42=0,006* in *number32=0,03* (oz. *case42=0,0013* in *number32=0,036* za neMCD) kažejo na to, da se je model v procesu učenja mejnih vrednosti naučil, da besedam v množini postavi višji prag kot besedam v tožilniku. Opozorimo naj, da mejne vrednosti pragov ne predstavljajo kalibriranih verjetnosti, zato jih ne presojamo kot gotovosti modelov za dane odločitve.

Primerjava med napovedovanjem z in brez MCD pokaže dokaj majhne in nekonsistentne razlike. Glede na precej večjo računsko zahtevnost metode MCD zato metode napovedovanja z MCD za naš problem ne moremo priporočiti.

6.2 Rezultati evalvacije na avtentičnem gradivu

Rezultate kvalitativne evalvacije na gradivu korpusa Šolar-Eval prikazujeta Tabeli 4 in 5. Kot je bilo omenjeno v razdelku 5.4, so rezultati ločeni glede na to, ali gre za primere, ki so rešljivi na ravni same povedi, ali primere, kjer je za interpretacijo potrebno več sobesedila.

Tabela 4: *Število avtentičnih jezikovnih popravkov rabe množine namesto dvojine v 58 obravnavanih povedih in število ustreznih strojnih rešitev, ki jih dobimo bodisi samo z novim pristopom, samo z Besano, z obema programoma ali z nobenim od njiju*

	Št. problemov	Novi pristop	Besana	Oba	Nerešeno
Dvoumni primeri	63	**1**	0	0	62
Nedvoumni primeri	25	**7**	3	2	13
Skupaj	88	**8**	3	2	75

Tabela 5: *Število avtentičnih jezikovnih popravkov rabe tožilnika namesto rodilnika v 47 obravnavanih povedih in število ustreznih strojnih rešitev, ki jih dobimo bodisi samo z novim pristopom, samo z Besano, z obema programoma ali z nobenim od njiju*

	Št. problemov	Novi pristop	Besana	Oba	Nerešeno
Dvoumni primeri	7	**1**	0	0	6
Nedvoumni primeri	54	8	**24**	9	13
Skupaj	61	9	**24**	9	19

Tako novi pristop kot Besana pri dvoumnih primerih dosegata izredno nizko uspešnost, vendar tudi nedvoumni primeri z obema postopkoma pogosto ostajajo nezaznani: to velja za 13 od 25 popravkov množine v dvojino ter za 13 od 54 popravkov tožilnika v rodilnik. Pri tem je treba izpostaviti, da na uspešnost popravljanja lahko vplivajo tudi preostali odstopi od standarda, ki se pojavljajo v obravnavanih povedih. Za ponazoritev navajamo nekaj primerov neuspešno identificiranih problemov:

- Ko smo šli prek železnice jaz in prijatel **smo ušli** vlaku, ampak ona pa je bila skoraj pod vlakom.
- Vendar se Hamlet izogiba **vsak kontakt** z Ofelijo.
- Ampak usaj nimaš slabe vesti o tem, kar si naredil in **kar** nisi.

Kot je razvidno iz Tabel 4 in 5, je novi pristop pri problemu dvojine malenkost boljši od Besane, medtem ko je pri težavah rodilnika Besana znatno uspešnejša. Jezikoslovna analiza ni pokazala vzorcev, ki bi nakazovali razloge za (ne)uspešnost enega ali drugega pristopa pri posameznih primerih. Za ponazoritev navajamo tri primere, ki jih je rešil novi pristop, ne pa Besana:

- Ko smo prišli do dveh aprtmajev, ki smo **jih** rezervirali.
- Oba cutita ljubezen drug do drugega ampak jih ta verski spopad ločuje drug od drugega.
- Izhaja iz bolj revne družine, zato si **marsikaj** nemore privoščiti.

Nato pa še tri primere, ki jih je rešila Besana, ne pa novi pristop:

- Če se zaposlim nemorem pričakovati **zlati zaslužek**.
- Ob problemu ne dajo samo **kruto kazen** ampak se o tem pogovorijo in jim svetujejo.
- Jaz in Sternfeldovka sva sklenili, da bo ona šla v grm in ko bo čas bo prišla ven, da bova Tulpenhajna **razkrinkale**.

6.3 Hitrost obdelave besedil

Za nekatere vrste uporabe prepoznavalnikov in popravljalnikov jezikovnih napak je pomembna tudi hitrost izvajanja, npr. pri popravljalnikih, ki delujejo v okviru sprotnega prepoznavanja govora, zato v omejenem obsegu analiziramo tudi ta aspekt razvitih modelov. Hitrost obdelave vhodnih besedil je odvisna predvsem od časa, ki ga program potrebuje, da besedilo (večkrat) obdela z modelom SloBERTa oz. SloBERTaMCD in orodji CLASSLA-Stanza. Meritve smo opravili na računalniku z dvema procesorjema *Intel(R) Xeon(R) Silver 4214*, torej z 48 nitmi. Na računalniku so bile 4 GPE *NVIDIA GeForce RTX 2080 Ti*. Oba modela smo naložili na eno GPE, istočasno pa se na kartici niso izvajali drugi programi.

Ob uporabi SloBERTaMCD smo v povprečju ob obdelavi besedila velikosti približno 61.000 besed dosegli hitrost obdelave 13,06 besed na sekundo. Ob uporabi navadnega modela SloBERTa se je hitrost podvojila na 26,02 besed na sekundo. Hitrost je bila višja, ker SloBERTaMCD uporablja večkratno povprečenje izhodnih vrednosti (opisano v razdelku 4.2) in je napovedovanje besednih zamenjav zato počasnejše.

7 ZAKLJUČEK

Razvili smo metodologijo za napovedovanje slovničnih popravkov, ki temelji na maskirnem jezikovnem modelu. Model vrne verjetnostno distribucijo popravkov slovničnega števila in sklonov, na podlagi katere ocenimo vrsto popravka in njegovo verjetnost. Za dejansko predlagane popravke uporabljamo parameter, katerega vrednost predstavlja prag verjetnosti ustreznosti popravka. Za razvoj metodologije smo uporabili orodje CLASSLA-Stanza (razdelek 3.3), napovedi različnih

besednih oblik z modelom SloBERTa (razdelek 4.2) in slovenski oblikoslovni leksikon Sloleks (razdelek 3.4). Za izboljšanje napovedi verjetnostne distribucije napak smo neuspešno preizkusili metodo MCD, ki je sicer okoli dvakrat počasnejša od napovedi običajnega modela SloBERTa. Izvorna koda programa je odprtodostopna.[7] Metodologijo smo preizkusili na popravkih tipa množina-dvojina in tožilnik-rodilnik in na različnih evalvacijskih množicah dosegli zelo različne rezultate.

Program ob hkratni uporabi naučenih mejnih vrednosti za popravljanje sklona tožilnik-rodilnik in števila množina-dvojina na umetnih podatkih doseže F_1-oceno med 95 % in 96 %. Pravilno je popravil od 92 % do 95 % napačno nastavljenih besed – odvisno od števila nastavljenih napačnih besed v evalvacijski množici. Na drugi strani na avtentičnih podatkih, ki so lahko vpeti v sobesedilo in vsebujejo tudi druge odstope od jezikovnega standarda, program uspešno naslovi le 29,5 % težav z rodilnikom (oz. 31,5 % primerov, ki so rešljivi brez širšega konteksta) in 11,4 % težav z dvojino (oz. 36 % primerov, ki so rešljivi brez širšega konteksta). Če je prva evalvacija kazala na presežno uspešnost predlaganega pristopa, pa je evalvacija na avtentičnem jeziku ugotovitve ustrezno relativizirala.

Rezultati našega dela kažejo, da je evalvacije na umetno pripravljenih podatkih, ki so na področju obdelave naravnega jezika sicer pogoste, nujno dopolnjevati z analizami na avtentičnem jeziku. Kvalitetna evalvacija bi morala upoštevati širši kontekst uporabe modelov, saj osredotočenost na ozko področje določene evalvacijske množice lahko vodi v pretirano prilagajanje specifični jezikovni situaciji na račun širše potencialne uporabe. Raba samo umetnih množic in množic, ki se le delno prekrivajo z nameravanim področjem uporabe, lahko privede do preveč optimističnih ocen uspešnosti delovanja. Na drugi strani je izziv raznolikost izražanja v naravnem jeziku, saj »pravilni odgovori« v jezikovnih podatkih vsebujejo le eno od izraznih možnosti, ne pa drugih prav tako pravilnih možnosti. Ta značilnost lahko modele strojnega učenja vodi do preozke usmerjenosti in zanemari variantne ali kreativnejše možnosti izražanja. Pri evalvaciji lahko vodi do preveč pesimističnih ocen uspešnosti delovanja ali do favoriziranja modelov, ki so preozko usmerjeni.

7 Na voljo na: https://github.com/timopetric/EssayHelper.

Izziv pri evalvaciji nalog s področja obdelave in razumevanja naravnega jezika sta tudi subjektivnost in (ne)konstistentnost človeških odločitev, ki se v evalvacijskih podatkih obravnavajo kot pravilne. Problem je še večji, če metodologija priprave ni transparentno popisana ali množice niso javno na voljo, da bi bile evalvacije ponovljive in razložljive. Ker je priprava kvalitetnih evalvacijskih virov zahtevna in zamudna, jih za številne jezike, tudi slovenščino, še vedno primanjkuje. Šolar-Eval, ki je bil pripravljen posebej za evalvacijo slovenskih črkovalnikov in slovničnih pregledovalnikov, ustreza vsem želenim kriterijem, vendar vsebuje izključno besedila šolajoče se populacije. Podobne množice bi bilo po enotni metodologiji treba v prihodnje razviti tudi za druge vzorce jezikovne rabe.

Ker so se v zadnjem času pojavili mnogo večji generativni jezikovni modeli, kot sta GPT-4 in Gemini, nekateri tudi prostodostopni, npr. LLama-3, se nadaljnji razvoj predstavljene ideje ne zdi smiseln, razen morebiti za samostojno delovanje slovničnih pregledovalnikov na računalnikih z malo računskimi viri. V nadaljnjem delu se bomo osredotočili predvsem na večje odprtodostopne modele in na razvoj učnih množic zanje. Ti modeli lahko, ob ustreznem učenju, naslovijo tudi v članku izpostavljeno in v jeziku pogosto težavo preozkega konteksta, ko je za ustrezno identifikacijo jezikovne težave treba upoštevati širše značilnosti sobesedila, ne le posamezne povedi.

Zahvala

Dostop do celotnega korpusa Lektor nam je omogočil vodja projekta izgradnje korpusa, doc. dr. Damjan Popič, s Filozofske fakultete Univerze v Ljubljani. Raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) in projekta Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159) in Veliki jezikovni modeli za digitalno humanistiko (GC-0002) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., ..., & Kosem, I. (2022a). *Developmental corpus Šolar 3.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1589

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., ..., & Kosem, I. (2022). *Frequency list of language problems from Šolar 3.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1716

Arhar Holdt, Š., Gantar, P., Bon, M., Gapsa, M., Lavrič, P., & Klemen, M. (2023). *Dataset for evaluation of Slovene spell- and grammar-checking tools Šolar-Eval 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1902

Božič, M. (2020). *Globoke nevronske mreže za postavljanje vejic v slovenskem jeziku* (Diplomska naloga). Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Pridobljeno s https://repozitorij.uni-lj.si/IzpisGradiva.php?id=119034&lang=slv

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, ..., Askell, A., idr. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, *49*(3), 643–701. doi: 10.1162/coli_a_00478

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (str. 4171–4186). doi: 10.18653/v1/N19-1423

Dobrovoljc, K., Krek, S., & Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, P. Gantar, I. Kosem & S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Pridobljeno s https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/15/47/489

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*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1230

Erjavec, T. (2017). MULTEXT-East. V *Handbook of Linguistic Annotation* (str. 441–462). Springer.

Erjavec, T., Fišer, D., Ljubešić, N., Ferme, M., Borovič, M., Boškovič, B., Ojsteršek, M., & Hrovat, G. (2019). *Corpus of Academic Slovene KAS 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1244

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *ArXiv*. doi: 10.48550/arXiv.2304.01746

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International conference on machine learning*, 1050–1059.

Gantar, P., Bon, M., Gapsa, M., & Holdt, Š. A. (2023). Šolar-Eval: Evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in slovstvo*, *68*(4), 89–108. doi: 10.4312/jis.68.4.89-108

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing (3rd ed. draft)*. Pridobljeno s https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Krek, S., Holdt, Š. A., 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. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth language resources and evaluation conference, LREC 2020, Marseille, France* (str. 3340–3345). The European Language Resources Association (ELRA).

Krek, S. (2023). Language Report Slovenian. In *European Language Equality: A Strategic Agenda for Digital Language Equality* (str. 211–214). Cham: Springer International Publishing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*. doi: 10.48550/arXiv.1907.11692

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*, 29–34. doi: 10.18653/v1/W19-3704

Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To BAN or not to BAN: Bayesian attention networks for reliable hate speech detection. *Cognitive Computation*, *14*(1), 353–371.

Mokotar, R. (2023). *Obvladovanje slovničnih napak v šolskih pisnih izdelkih z metodami za obdelavo naravnega jezika* (Diplomska naloga). Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Pridobljeno s https://repozitorij.uni-lj.si/IzpisGradiva.php?id=144932&lang=slv

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

Popič, D. (2014). Revising translation revision in Slovenia. *New Horizons in Translation Research and Education 2,* 72–89. University of Eastern Finland Joensuu.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rizvič, M. (2020). *Avtomatsko postavljanje ločil v surovem tekstu* (Magistrsko delo). Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Pridobljeno s https://repozitorij.uni-lj.si/IzpisGradiva.php?id=117687&lang=slv

Romih, M., & Holozan, P. (2002). Infrastruktura za razvoj jezikovnih tehnologij-korpus FIDA in sistem ASES. V T. Erjavec, J. Žganec Gros (ur.), *Jezikovne tehnologije, 14.–15. oktober, Ljubljana, Slovenija* (str. 166). Pridobljeno s http://nl.ijs.si/isjt02/zbornik/sdjt02-D02amebis.pdf

Rozovskaya, A., Roth, D., & Srikumar, V. (2014). Correcting grammatical verb errors. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (str. 358–367).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., …, Azhar, F., idr. (2023). Llama: Open and efficient foundation language models. *ArXiv*. doi: 10.48550/arXiv.2302.13971

Ulčar, M., & Robnik-Šikonja, M. (2021a). SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of Slovenian KDD Conference, SiKDD 2021, part of Information Society*.

Ulčar, M., & Robnik-Šikonja, M. (2021b). *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1397

Ulčar, M., & Robnik-Šikonja, M. (2023). Sequence to sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*. doi: 10.3389/frai.2023.932519

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Velikonja, N. (2021). *Segmentacija in postavljanje končnih ločil v slovenskih stavkih z modeli tipa BERT* (Diplomska naloga). Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Pridobljeno s https://repozitorij.uni-lj.si/IzpisGradiva.php?id=130323&lang=slv

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *ArXiv.* doi: 10.48550/arXiv.2303.13648

Zhang, S., Huang, H., Liu, J., & Li, H. (2020). Spelling Error Correction with Soft-Masked BERT. V D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (ur.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *July 2020* (str. 882–890). Association for Computational Linguistics. Pridobljeno s https://aclanthology.org/2020.acl-main.pdf

The importance of realistic evaluation: an example of correcting Slovene grammatical case and number with large language models

Frequent grammar errors in standard Slovene include using an incorrect grammatical conjugation or number. Using the large language model SloBERTa, we have developed a new methodology for the machine detection of such problems and tested it on incorrect use of the accusative instead of the genitive case and the plural instead of the dual. We applied standard natural language processing tools for Slovenian to evaluate and modify word forms in the input sentences, such as morphosyntactic tagger CLASSLA-Stanza and Slovenian word form lexicon Sloleks. The proposed corrections are based on word form statistics when using masked word prediction with a large language model. Due to the lack of sufficient training data, we trained the prediction models on synthetically generated errors. We first evaluated the performance of machine correction on synthetic data and the Lektor corpus, and later on a newly developed evaluation dataset Šolar-Eval. The evaluation on the first two datasets showed the excellent performance of the developed methodology (more than 90% of detected synthetically introduced errors), while with Šolar-Eval it had a far worse performance (only 29.5% of the problems with the genitive-accusative grammatical case were detected, and just 11.4% of those with the dual-plural grammatical number). Overall, the results show the danger of overfitting to datasets and the importance of evaluating on purposefully designed authentic datasets, which are still rare for Slovene.

Keywords: grammatical error correction, grammatical case, grammatical number, large language models, evaluation

Much more than an introduction to the language-specific lexical semantics

Igor Marko GLIGORIĆ
Faculty of Humanities and Social Sciences, University of Zagreb

An exceptional contribution to Croatian linguistics, Slavistics, and undoubtedly linguistics in general, was published by Matrix Croatica (*Matica Hrvatska*) in 2024. The book, *Introduction to Lexical Semantics of the Croatian Language* (*Uvod u leksičku semantiku hrvatskoga jezika*) by Tatjana Pišković, currently an associate professor at the Department of Croatian Language at the Faculty of Humanities and Social Sciences, University of Zagreb, represents a landmark achievement. The manuscript was reviewed by three distinguished linguists – Maja Bratanić, Branimir Belaj, and Ida Raffaelli – all esteemed figures in their fields. This book cannot be simply categorized as a scientific monograph or a university textbook on lexical semantics, as its scope – both in depth and broadness of the approach – surpasses such classifications, moving toward an encyclopaedic treatment of the central topic.

In the *Preface*, the author highlights the complexity of studying meaning, carefully delineating fundamental concepts and the domain explored in the detailed examination and analysis that follows. The various levels at which different branches of linguistic semantics are named overlap with one another (cf. Pišković, 2024, p. 6). This acknowledgment underscores the author's awareness that semantics

Gligorić, I. M.: Much more than an introduction to the language-specific lexical semantics. Slovenščina 2.0, 12(1): 131–137.

1.19 Recenzija, prikaz knjige, kritika / Review, book review, critique

DOI: https://doi.org/10.4312/slo2.0.2024.1.131-137

resists rigid boundaries far more than grammar does. Consequently, the phenomena analysed and interpreted cannot be classified with equal clarity and precision. Pišković identifies this very fact as one reason why certain linguistic approaches have entirely distanced themselves from meaning. However, the author recognizes this area as a space that can and should inspire curiosity and creativity, as she describes it as something that "liberates and delights" (ibid.).

Pišković views lexical semantics as the study of word meanings, which can, to some extent, be contrasted with morphology, the study of word forms. The book is structured into two main parts: the first provides a *Brief History of Lexical Semantics* (*Kratka povijest leksičke semantike*), while the second analyses the *Basic Topics in Lexical Semantics* (*Osnovne teme leksičke semantike*). It is worth noting that the *Brief History of Lexical Semantics* could perhaps have been expanded – though this is not a critique – yet it still includes all significant elements relevant to the diachronic development of this linguistic discipline. On the other hand, in the section on *Basic Topics in Lexical Semantics*, the author explores fundamental concepts and lexical-semantic relations. However, her approach, supported by examples that substantiate her conclusions and illustrate lexical-semantic phenomena, opens up much more complex and broader considerations of lexical semantics than the term "basic" might imply.

As indicated by the title, the first part of the book offers a brief history of lexical semantics. Drawing on relevant scholars, Pišković divides the history of lexical semantics into five major periods. She explains that the focus is on illustrating the beginnings and foundations of each approach discussed, rather than recent or specialized studies within these frameworks. This method of presenting theoretical foundations is both understandable and justified, particularly given the book's primary purpose.

Despite the deliberate exclusion of detailed analyses of each approach to word meaning mentioned, the first part of the *Introduction* exceeds readers' expectations. Pišković provides a chronological and causal description of pre-structuralist diachronic semantics, structuralist semantics, generative semantics, neo-structuralist semantics,

and cognitive semantics, offering readers a clear diachronic overview of the study of word meaning. She highlights, discusses, and illustrates the approaches of M. Bréal and H. Paul, classifies semantic changes, and identifies some of the lasting contributions of pre-structuralist diachronic semantics, including those in a Croatian context. She also extensively explores structuralist critiques of pre-structuralism, the theory of lexical fields, as well as European and American componential analysis. Separate subsections are devoted to relational semantics and the general contributions of structuralist semantics, with a specific focus on those in Croatian. The clarity and methodology employed in the first two subsections reveal the author's consistent approach throughout the text.

In the subsection on generative semantics, Pišković correctly identifies the dominance of syntax over semantics during this period. She examines Katzian semantics, generative and interpretative semantics, and the contributions inherited from this era. Regarding neo-structuralist semantics, she systematically analyses two perspectives: the neo-structuralist componential approach to lexical meaning and the relational approach. The former includes subsections on natural semantic metalanguage, conceptual semantics, two-level semantics, and generative lexicon theory. The further chapters consider WordNet, meaning-text theory, and corpus-based distributional analysis. As with the other sections, Pišković concludes this part by identifying neo-structuralist semantic influences in Croatian.

When discussing cognitive semantics, Pišković observes the blurring of boundaries between meaning and usage, that is, the boundaries traditionally embodied in the distinction between semantics and pragmatics. She analyses core concepts of cognitive linguistics, including prototype theory, conceptual metaphor and metonymy, and conceptual integration. She also examines various structures of encyclopaedic knowledge within cognitive semantics, such as the idealized cognitive model and frame semantics. Following her established methodology, Pišković also explores cognitive semantics in Croatian.

The final subsection considers the development of lexical semantics as cyclical. The author notes that the first three subsections

follow one another chronologically, while the last two emerge simultaneously with the third. She also emphasizes that all these approaches still coexisted at the time of writing this book, which she explicitly states (ibid.: 153). The *Preface* makes it clear that the first part of the book serves to justify the theoretical and methodological choices made in the second part (ibid.: 7). Furthermore, it announces that all the mentioned approaches will be integrated into an understanding of lexical-semantic concepts, a promise that is consistently fulfilled.

The second part of this encyclopaedic publication is motivated by what Pišković describes as propaedeutic practice. It begins with the definition of key terms such as word, lexeme, and lexicon, as well as lexical meaning and traditionally recognized lexical-semantic relations and phenomena.

In the first chapter of the second part, the concepts of words, lexemes, grammar, and lexicons are examined from various perspectives. The author questions the division into open and closed word classes, distinguishes lexicon, lexical inventory, and dictionary, and discusses the status of lexemes in the dictionary and the concept of listemes.

The second chapter is dedicated to lexical meaning. The author attempts to define it in terms of meaning alone, then in terms of lexical and pragmatic meaning. Various triadic models – models of lexical triangles – are illustrated and commented upon. Attention is then focused on different types of lexical meaning and the metalexical aspect, where Pišković addresses the issue of defining words, lexemes, and lexical meaning.

As the author herself notes in the *Preface* (ibid.: 7), the central chapter is devoted to polysemy. The chapter begins with defining this, offering not only definitions but also thoughtful interpretations of the reasons and mechanisms behind lexical polysemy. Highlighting the cognitive basis of polysemy, she particularly addresses the mechanisms of polysemy: lexical metaphor, lexical metonymy, and lexical synecdoche. It is especially worth emphasizing the examples provided by the author, as well as her original and clear distinction between metonymy and synecdoche within the cognitive-linguistic framework.

The next subsection deals with homonymy. It is important to note that, methodologically, homonymy is consistently contrasted with polysemy, with the often-blurred boundaries and the relationship between polysemy and homonymy in lexicography and psycholinguistics being highlighted. The discussion – faithfully following the principles of scientific elaboration – is preceded by a problematization of the origin of homonyms.

The next subsection is dedicated to what many consider one of the prototypical lexical-semantic – i.e. conceptual – relations: antonymy. It is discussed as a perfect lexical-semantic relation (ibid: 391–395). As readers might expect from the earlier elaboration of semantic concepts, Pišković offers a dual classification of antonyms: morphological and semantic. Moreover, she does not disappoint more demanding readers by examining the relationship between polysemy and polyantonymy.

On the other hand, synonymy is addressed. Understandably, it is defined as an imperfect lexical-semantic relation. Again, the author adheres to a consistent methodological framework. Within this, she identifies synonymy as a graded category or phenomenon, distinguishing various levels of synonymity. More or less predictably – depending on the reader's profile – explanations are provided for the relationships between polysemy and polysynonymy, synonymy and the dictionary, and synonymy and paronymy.

The hierarchy of lexical units is examined in two chapters: one on hyponymy and the other on meronymy. In discussing hyponymy, Pišković analyses lexical hierarchy and lexical inclusion, taxonomy, and other types of hyponymic relations. Here, taxonomy is understood as a representative example of hyponymy. The author also explores the phenomena of quasi-hypoymy, polysemy, and polyhyponymy, as well as the relationship between hyponymy and the dictionary.

The final chapter is dedicated to meronymy. Partonomy and other types of meronymic relations are central to the discussion of this lexical-semantic relation. Similarly to the discussion on hyponymy, quasi-meronymy is specifically analysed and illustrated, with particular attention being paid to its similarity to quasi-hyponymy. The chapter concludes with definitions and examples of automeronymy,

supermeronymy, and polymeronymy, as well as an analysis of the relationship between meronymy and the dictionary.

At the end of the book, there is an extensive bibliography. The book also includes an index of names and terms, followed by a note about the author.

In the *Preface*, the author notes that the book is partly an introductory work, designed as a university textbook – a manual for teachers and students, a book intended to benefit each of these two target groups in a way that meets their specific needs and expectations. However, it is clear that Pišković's latest work goes beyond the framework of a university textbook or scientific monograph, and is much more than that. The interesting and contemporary examples, with which the author makes her highly professional and linguistically articulated text more accessible to readers, are a notable strength of this publication. Additionally, the book consistently considers dictionary definitions of lexemes and their concrete usage, balancing these two aspects of the linguistic nature of language units.

A review of a book like *Introduction to the Lexical Semantics of the Croatian Language* (*Uvod u leksičku semantiku hrvatskoga jezika*) could be written in at least two ways. Here, we could have highlighted the most interesting and important elements according to subjective judgment. However, we chose to present the entirety and broadness encompassed by Pišković's 615 pages of original content. While we acknowledge that this approach might seem less precise – and it surely is at some extent – we consider cataloguing the topics covered by the author in her book to be far more important.

This reflects the totality of the author's knowledge of lexical semantics, her outstanding understanding of concepts and her ability to navigate the field she writes about, and thus interested readers will not be disappointed by the book. The target audience is primarily Crostists, Slavists, linguists and students of these (philological) studies. However, given the nature of meaning as not only a linguistic category, Pišković's work will also enlighten readers who view the linguistic formations of conceptual material differently: sociologically, anthropologically, philosophically, psychologically, and so on, or simply in a practical way, as users of language. This

monographic-textbook-encyclopaedic work will thus find its audience in the most diverse reader groups, who will find much to discover, learn and consider between its covers.

Reference

Pišković, T. (2024). *Uvod u leksičku semantiku hrvatskoga jezika*. Zagreb: Matica hrvatska.