Spremljevalni korpus Trendi: metode, vsebina in kategorizacija besedil

Iztok Kosem,‡* Jaka Čibej,‡ Kaja Dobrovoljc,‡* Nikola Ljubešić‡

‡ Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
iztok.kosem@ijs.si, jaka.cibej@ijs.si, kaja.dobrovoljc@ijs.si, nikola.ljubesic@ijs.si
* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana

Povzetek

V prispevku opisujemo postopek gradnje korpusa Trendi – prvega spremljevalnega korpusa za slovenščino. Prva različica korpusa, imenovana Trendi 2022-05, vsebuje več kot 565 milijonov pojavnic iz več kot 1,4 milijona besedil. Namen korpusa je, da tako strokovni kot nestrokovni javnosti ponudi podatke o aktualni jezikovni rabi in omogoči spremljanje pojavljanja novih besed ter upadanja ali naraščanja rabe že obstoječih. Predstavimo metodologijo izdelave in vsebino korpusa ter prve korake pri načrtovani strojni klasifikaciji korpusnih besedil v kategorije (npr. gospodarstvo, okolje), s katerimi bo mogoče v korpusu spremljati jezikovno rabo tudi po tematskih področjih. Predstavimo tudi rezultate ankete, s katero smo preverili uporabniška pričakovanja o jezikovnem viru za spremljanje jezikovne rabe.

The *Trendi* Monitor Corpus of Slovene: Methods, Content, and Text Categorization

In the paper, we present the compilation of the Trendi corpus – the first monitor corpus of Slovene. The first version of the corpus, named Trendi 2022-05, contains over 565 million tokens coming from more than 1.4 million different texts. The purpose of the corpus is to provide both experts and non-experts with data on contemporary language use and enable the monitoring of the appearance of new words or the increase/decrease in the use of existing words. We present the methodology of corpus compilation, its content, and the first steps for the automatic classification of corpus texts into categories (such as economics and environment), which will enable the monitoring of language use by thematic areas. We also describe the results of a survey, the goal of which was to collect feedback on user expectations from a language monitoring resource

1. Uvod

Jezik se nenehno spreminja, pojavljajo se nove besede, obstoječe besede in besedne zveze dobivajo nove pomene, določene besede ali njihovi pomeni se prenehajo uporabljati ipd. V zadnjem času, tudi zaradi epidemije covida-19, ki je prinesla veliko novega izrazoslovja, je še posebej veliko pozornosti deležno področje neologije, tako leksikalne (nove besede) kot semantične (novi pomeni).

Za spremljanje sprememb v jeziku se tipično uporabljajo spremljevalni korpusi, ki vsebujejo najnovejša besedila v jeziku. Spremljevalni korpusi zapolnjujejo manko referenčnih korpusov, katerih izdelava zaradi raznovrstnosti besedil in njihovih formatov ter obsega traja dlje časa. V času tehnološkega napredka in ob dejstvu, da je zdaj zelo veliko besedil dostopnih na spletu, je izdelava spremljevalnih korpusov postala enostavnejša; kar je objavljeno danes, je lahko že jutri vključeno v korpus.

Za slovenščino kljub bogati opremljenosti na področju korpusov do zdaj nismo imeli spremljevalnega korpusa, čeprav se je med različnimi deležniki kazala jasna potreba po njem. Naslavljanja tega manka smo se lotili v okviru projekta *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED),[1] ki poteka od oktobra 2021 do novembra 2022 in ga sofinancira Ministrstvo za kulturo Republike Slovenije. Cilj projekta ni samo izdelati spremljevalni korpus, temveč tudi pripraviti infrastrukturo za njegovo redno posodabljanje.

V prispevku najprej ponujamo pregled nekaterih pomembnejših tujih spremljevalnih korpusov, nato pa predstavimo metodologijo in vsebino spremljevalnega korpusa Trendi. Sledi predstavitev klasifikacije tematskih kategorij, ki smo jo izdelali za pripravo modela za avtomatsko kategorizacijo besedil. V zadnjem delu predstavimo anketo med uporabniki o želenih statističnih izračunih iz korpusa. V zaključku predstavimo načrte za prihodnje delo.

2. Spremljevalni korpusi

V mednarodnem prostoru so spremljevalni korpusi prisotni že od 20. stoletja. Eden prvih je bil the Bank of English, ki je bil prvič objavljen leta 1991. Vsebuje več kot 650 milijonov besed[2] in je danes vključen v 4,5-milijardni korpus COBUILD založbe Collins. Korpus ni prosto dostopen, poleg zaposlenih na založbi Collins ga lahko uporabljajo tudi zaposleni in študentje na Univerzi v Birminghamu.

Za angleščino je danes pomemben predvsem korpus NOW (News on the Web; Davies, 2016-), ki vsebuje več kot 15 milijard besed iz spletnih časopisov in revij. Korpus zajema besedila od 2010 naprej. Kot je omenjeno na spletni strani,[3] korpus vsak mesec naraste za 180-200 milijonov besed.

Obsežna zbirka korpusov za spremljanje sprememb v jeziku, ki poleg angleščine pokriva še več kot 35 drugih jezikov, so korpusi Timestamped JSI. Korpusi vsebujejo novice, ki jih zbira JSI Newsfeed na Institutu "Jožef Stefan" (Trampuš in Novak, 2012). Korpusi za 18 jezikov so na voljo v orodju Sketch Engine (Kilgarriff et al., 2004),[4] v katerem imajo poleg ostalih funkcij orodja uporabniki na voljo tudi t. i. Trende (Herman, 2013), funkcijo, ki pomaga prepoznavati trende v rabi besed. Korpusi v Sketch Enginu

[1] https://sled.ijs.si/

[2] Žal nismo našli podatka, kdaj je bil korpus nazadnje posodobljen.

[3] https://www.english-corpora.org/now/

[4] https://www.sketchengine.eu/

vsebujejo besedila od 2014 do aprila 2021 (čas zadnje posodobitve) in so različnih velikosti; korpus angleščine na primer vsebuje približno 60 milijard besed.

Obstaja še precej drugih spremljevalnih korpusov, ki pa so pogosto na voljo zgolj za interno rabo. Primer takšnega korpusa je ONLINE, dinamični spremljevalni korpus češkega jezika, ki ga izdeluje Inštitut za češki nacionalni korpus.[5] Velik je približno 6,3 milijarde besed in vsebuje spletne novice, komentarje (pod spletnimi novicami), besedila s forumov in družabnih omrežij (Facebook, Twitter, Instagram). Korpus ONLINE je razdeljen na dva komplementarna korpusa: ONLINE_NOW in ONLINE_ARCHIVE. Prvi je posodobljen vsak dan in pokriva obdobje zadnjega meseca in preteklih šestih mesecev. ONLINE_ARCHIVE pokriva obdobje od februarja 2017 do prvega meseca, ki ga vsebuje ONLINE_NOW. Tako se vsebina zadnjega meseca po starosti v korpusu ONLINE_NOW na začetku vsakega meseca preseli v ONLINE_ARCHIVE.

Obstajajo tudi manjši in bolj specializirani spremljevalni korpusi, kakršen je npr. korpus Coronavirus (Davies, 2019-), ki zajema obdobje od januarja 2020 do danes in vsebuje več kot 1,4 milijarde besed. V njem so spletne novice v angleščini, vsak dan pa naraste za 3 do 4 milijone besed.

Do določene mere vlogo spremljevalnega korpusa opravljajo tudi diahroni korpusi, seveda pod pogojem, da vsebujejo čim novejša besedila. Kot primer lahko navedemo korpus sodobne ameriške angleščine (Corpus of Contemporary American English; Davies, 2008-), ki vsebuje besedila od leta 1990 do marca 2020 (zadnja posodobitev) in obsega več kot milijardo besed. Prednost korpusa je, da je žanrsko uravnotežen, saj vsebuje besedila iz osmih različnih žanrov (govorjeni jezik, leposlovje, revije, časopise, znanstvena besedila, televizijske in filmske podnapise, bloge in ostale spletne strani). Slovenski ekvivalent bi bil korpus Gigafida 2.0 (Krek et al., 2019),[6] ki obsega 1,13 milijarde besed, vendar pa je v primerjavi s korpusom sodobne ameriške angleščine manj ažuren (vsebuje samo besedila do leta 2018).

Za slovenščino do danes še ni obstajal pravi spremljevalni korpus. Obstajajo sicer viri, kot je Jezikovni sledilnik (Kosem et al., 2021),[7] ki že izkorišča najsodobnejše podatke o jezikovni rabi, v konkretnem primeru od JSI Newsfeeda, za izdelavo neke vrste začasnih korpusov, na katerih se potem izvajajo statistični izračuni. Taka ciljna raba je seveda tudi potrebna, vendar pa je namenjena nestrokovni javnosti; po drugi strani strokovna javnost, kot so leksikografi_ke, jezikoslovci_ke, drugi raziskovalci_ke potrebujejo dostop do izvirnih besedil, če želijo opravljati še druge analize.

3. Korpus Trendi

Izdelave prvega spremljevalnega korpusa za slovenščino, ki smo ga poimenovali Trendi, smo se lotili v okviru projekta SLED. Poleg izdelave in rednega posodabljanja korpusa Trendi ima projekt še dva cilja: pripravo na korpusnih podatkih temelječe statistike o različnih vidikih rabe besed in izdelavo orodja, ki bo besedila avtomatsko opremilo s podatkom o tematski kategoriji.

3.1. Metodologija in vsebina korpusa

Z metodološkega vidika smo pri snovanju korpusa Trendi morali sprejeti dve odločitvi: obdobje, ki ga bo korpus pokrival, in kako pogosto bo korpus posodobljen. Pri odločitvi o obdobju smo izhajali iz želje, da bi korpus Trendi vedno pokrival manko najnovejše različice referenčnega (pisnega) korpusa Gigafida, trenutno je to 2.0. V tem trenutku to pomeni, da bo Trendi vseboval besedila od 2019 naprej. To pomeni, da se ob objavi nove različice korpusa Gigafida (npr. korpus Gigafida 3.0 bo objavljen v sklopu projekta *Razvoj slovenščine v digitalnem okolju* - RSDO),[8] obdobje korpusa Trendi ustrezno prilagodi.

Tesna povezanost s korpusom Gigafida tudi pomeni, da bo korpus Trendi predstavljal standardno pisno slovenščino. Odločitev se nam zdi smiselna tudi zato, ker sta nestandardna oz. govorjena slovenščina pokrita s korpusi, kot sta JANES[9] in Gos,[10] in je torej njun razvoj predmet ločenih projektov. Navsezadnje pa ne gre pozabiti na nastajajoči korpus metaFida,[11] ki bo združil vse slovenske korpuse.

Pri pripravi seznama virov za vključitev v korpus Trendi smo izhajali iz seznama slovenskih spletnih virov, ki jih najdemo v servisu JSI Newsfeed. Izdelali smo seznam vseh virov od leta 2019 do konca 2021, pridobili smo tudi podatek o skupnem številu besedil na vir. Nato smo pri pripravi seznama za korpus Trendi podrobno analizirali vsakega od 243 virov. 90 virov smo izključili, ker je šlo za tuje ali slovenske spletne strani z vsebino v tujem jeziku. Nato smo s seznama odstranili še 34 virov, nekatere zato, ker niso vsebovali medijskih novic (blogi, spletne strani vladnih uradov in podjetij), druge zato, ker je njihova vsebina preveč specializirana (npr. repozitoriji akademskih publikacij so primernejši za korpuse, kot je Korpus akademske slovenščine). Ena od strani (preberi.si) je bila s seznama odstranjena zato, ker je agregator novic iz drugih virov. Končni seznam korpusa Trendi tako vsebuje 110 virov, med tistimi, ki so v obdobju 2019-2021 prispevali največ novic, so sta.si (260.080 besedil), rtvslo.si (97.924), siol.net (69.471), delo.si (65.415), 24ur.com (61.623), dnevnik.si (47.749) in vecer.com (45.548).

Seznam virov se bo redno posodabljal, saj lahko pričakujemo pojav novih spletnih strani, pa tudi ukinitev obstoječih. Kot primer lahko navedemo spletno stran necenzurirano.si, ki se je pojavila šele leta 2020 in je že 28. po številu novic (8.494). Dodajanje novih virov v korpus pomeni tudi večje število besed na mesečni ravni in posledično večji korpus Trendi. Trenutni okvirni izračuni kažejo, da se bo Trendi vsak mesec povečal za 10-15 milijonov pojavnic, pri čemer je bil povprečen mesečni obseg leta 2019 12,5 milijona pojavnic, leta 2021 pa že 21 milijonov pojavnic.

Zaradi narave korpusa Trendi bodo potrebne redne posodobitve, ki so zaenkrat predvidene na mesečni ravni, kot je praksa pri podobnih tujih korpusih. To se zdi trenutno realno, upoštevajoč časovno zahtevnost pridobivanja in

[5] https://korpus.cz/

[6] https://viri.cjvt.si/gigafida/

[7] https://viri.cjvt.si/sledilnik/slv/

[8] https://slovenscina.eu/

[9] https://www.clarin.si/kontext/query?corpname=janes

[10] http://www.korpus-gos.net/

[11] https://www.clarin.si/kontext/query?corpname=mfida01

označevanja besedil, pretvorb v potreben format in vključevanje korpusa v konkordančnike.

3.2. Priprava besedil

Za pripravo besedil smo pripravili cevovod, ki vključuje pridobivanje besedil, označevanje na različnih ravneh, združevanje po virih in obdobjih ter pretvorbo v različne formate. Pridobivanje besedil je zaenkrat vezano na servis JSI Newsfeed, ki uporablja protokol RSS novic, vendar pa smo sredi priprave lastnega postopka luščenja. Za to smo se odločili predvsem zato, ker smo odkrili, da so pri mnogih virih potrebne izboljšave pri pridobivanju besedil, npr. poleg besedila so izluščeni še drugi deli strani, besedilo ni pridobljeno v celoti ipd. Poleg tega strani včasih vsebujejo pomembne metapodatke o besedilu, ki trenutno niso del zajema. V novem postopku bomo ročno preverili rezultate pridobivanja besedil z vsakega vira in prilagodili algoritem za vsak vir, kjer se bo izkazala potreba po prilagoditvi.

Nekateri viri, kot so sta.si, delo.si itd. imajo vsebine zaklenjene oziroma so dostopne samo naročnikom. Pri pridobivanju prek protokola RSS so tako prosto dostopni samo povzetki ali prvih nekaj odstavkov, včasih celo samo naslov in podnaslov. Pri reševanju problema smo združili moči z ekipo, ki v okviru projekta RSDO oz. priprave korpusa Gigafida 3.0 sklepa pogodbe z besedilodajalci. Dogovor z besedilodajalci vključuje redno dostavljanje celotnih besedil. Posledično bo končna oblika cevovoda za korpus Trendi kombinacija priprave besedil, pridobljenih s spleta, in besedil, ki jih bodo v digitalni obliki poslali besedilodajalci.

Del postopka pridobivanja besedil je tudi deduplikacija, ki je trenutno omejena zgolj na raven vira besedila; del cevovoda je namreč preverjanje, da se besedilo z istim URL-jem ne ponovi. Zavedamo se, da zaradi pokrivanja istih dogodkov obstaja velika prekrivnost med viri. Še več, mnogi viri osnujejo številne novice na podlagi vsebin sta.si, kar pripelje do podvajanja besedila na ravni stavkov, odstavkov ali tudi celotne vsebine. Kljub temu za namene korpusa Trendi deduplikacija na ravni vsebine ni predvidena, saj želimo uporabnikom omogočiti analizo vsebin posameznih virov ter primerjalne analize med viri. Deduplikacija pa bo najbrž opravljena pri pripravi besedil za novo različico korpusa Gigafida, kot je bila praksa v preteklih različicah (Krek et al., 2019).

Sledi postopek strojnega označevanja besedil, za kar uporabljamo označevalni cevovod CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019),[12] ki se kot referenčno orodje za slovnično označevanje besedil v slovenščini aktivno razvija v okviru projekta RSDO. Orodje je nadgradnja odprtokodnega orodja Stanza (Qi et al., 2020), ki v primerjavi z izvorno programsko opremo podrobneje naslavlja specifike slovenščine, zlasti na ravni stavčne segmentacije, tokenizacije, oblikoskladenjskega označevanja in lematizacije po sistemu JOS (Erjavec et al., 2010). Poleg navedenih ravni orodje besedila tudi skladenjsko razčleni po sistemu Universal Dependencies (Dobrovoljc et al., 2017) in v njih označi imenske entitete (Zupan et al., 2017), kot so imena oseb, krajev, organizacij ipd.

Po končanem označevanju se v cevovodu opravi še pretvorba besedil iz privzetega formata označevalnega orodja (CONNL-U) v TEI XML, ki ga med drugim potrebujemo za statistične izračune s programom LIST (Krsnik et al., 2019). V ta proces sta vključena še dva povezana postopka združevanja besedil: združevanje besedil po viru na dan (vsakodneven postopek) in združevanje besedil istega vira za cel mesec (enkrat na mesec, na začetku novega meseca za nazaj). V zadnjem koraku, ki ga izvajamo enkrat mesečno in ga moramo pognati ločeno zaradi kombinacije XSLT in skripte Perl, je opravljena še pretvorba mesečnih datotek (razdeljenih po viru) v format VERT, ki ga uporabljata konkordančnika KonText (Machálek, 2020) in NoSketch Engine (Rychlý, 2007).

3.3. Prva različica korpusa Trendi

Prva različica korpusa Trendi, imenovana Trendi 2022-05, je bila objavljena junija 2022 in vsebuje 565.308.991 pojavnic oz. malo več kot 473 milijonov besed. V korpusu je 1.436.548 besedil od 48 izdajateljev, pri čemer imajo največje deleže Slovenska tiskovna agencija (337.484; 23,5 %), Delo d.o.o. (128.164; 8,9 %), Radiotelevizija Slovenija (124.861; 8,7 %), Media24 d.o.o. (100.587; 7 %), PRO PLUS d.o.o. (86.578; 6 %) in TSMedia d.o.o. (83.342; 5,8 %).

3.4. Dostopnost korpusa Trendi

Korpus Trendi je za brskanje prosto dostopen v treh konkordančnikih CLARIN.SI – konkordančniku KonText[13] in dveh različicah konkordančnika NoSketchEngine,[14] tako KonText kot NoSketch Engine imata več enakih funkcionalnosti (enostavno in napredno iskanje ipd.), vendar pa KonText ponuja možnost registracije in shranjevanje iskanj in priljubljenih korpusov, NoSketchEngine pa dodatne funkcionalnosti, kot je luščenje ključnih besed (angl. *keywords*) iz korpusov, za uporabo katerih ni potrebna registracija. Konkordančnik NoSketch Engine je na CLARIN.SI poleg starejše različice (Bonito) po novem na voljo tudi v novejši različici uporabniškega vmesnika (Crystal),[15] ki zagotavlja izboljšano uporabniško izkušnjo in dolgoročnejše vzdrževanje.

Odprto dostopna različica korpusa Trendi bo zaradi omejitev avtorskih pravic izdelana po isti metodi kot ccGigafida 1.0 (Logar et al., 2013), tj. vzorčeni bodo naključni odstavki posameznih besedil, in bo na voljo v repozitoriju CLARIN.SI.

Korpus bo v repozitoriju CLARIN.SI na voljo tako v formatu TEI kot v formatu CONNL-U, saj je slednji preferenčni format pri nalogah, ki vključujejo nadaljnje procesiranje podatkov, npr. strojno učenje, luščenje podatkov ipd.

3.5. Tematska kategorizacija besedil

Ena od aktivnosti projekta SLED je tudi izdelava orodja za avtomatsko kategorizacijo besedil glede na tematiko. Za izdelavo takšnega orodja oz. modela zanj potrebujemo dvoje: klasifikacijo kategorij in učno množico.

[12] https://pypi.org/project/classla/
[13] https://www.clarin.si/kontext/
[14] https://www.clarin.si/noske/
[15] https://www.clarin.si/ske/

Pri izdelavi nabora kategorij smo se opirali na podatke iz treh skupin virov:

- slovenskih novičarskih portalov, izbrali smo jih šest, tj. rtvslo.si, delo.si, sta.si, dnevnik.si, 24ur.com in vecer.com.
- nabora tematskih kod oz. kategorij Mednarodnega tiskovnega telekomunikacijskega sveta (IPTC).[16] S tem smo tudi želeli zagotoviti čim boljšo usklajenost naših kategorij z mednarodnim standardom.
- kategorij v sodobnih sinhronih in spremljevalnih korpusih, pri čemer sta bila relevantna predvsem češki korpus SYN_2015 (Křen et al., 2016) in estonski nacionalni korpus (Koppel in Kallas, v tisku).

Glavno vodilo pri pripravi klasifikacije je bilo pripraviti relativno majhen nabor kategorij, v katere lahko uvrstimo vse novice na različnih portalih. S tem bi zagotovili tudi boljše delovanje modela. Posledično smo pri analizi uporabljenih virov več pozornosti posvečali krovnim kategorijam, kar je bilo sploh potrebno pri naboru IPTC, ki ima približno 1.400 kategorij, razdeljenih v tri nivoje (s tem da krovni nivo sestavlja le 17 kategorij). Za ponazoritev smiselnosti uporabe zgolj krovnih kategorij lahko vzamemo kategorijo šport, ki ima na večini novičarskih portalov nadaljnje kategorije, od katerih se vedno pojavita samo *nogomet* in *košarka*, ostale pa le na nekaterih portalih, npr. dnevnik.si nima *zimskih športov*, ima pa ločeno podstran za novice o *Luki Dončiću*; rtvslo.si je edini, ki ima podstran za novice o *Formuli 1*, 24ur.si ima ločene podstrani za *Ligo prvakov* in *Ligo Evropa* (nogomet) ter *borilne športe*.

Naša končna klasifikacija vsebuje 12 kategorij:

- **umetnost in kultura**. Vključuje besedila o kulturi, umetnosti, filmih, knjigah, gledališču, pa tudi recenzije ipd.
- **črna kronika**. Naravne in ostale nesreče, človeški delikti, kriminal.
- **gospodarstvo**. Vključuje besedila s področja ekonomije, trgov, financ, zaposlitev ipd.
- **okolje**. Zajema okoljevarstvo, planet, energente, tudi kmetijske teme.
- **zdravje**. Fizično in mentalno zdravje ljudi, medicina, farmacija, zdravstvena infrastruktura.
- **prosti čas**. Hobiji, rekreacija, potovanja, turizem, ljubljenčki, dom in družina, bivanje.
- **politika in pravo**. Mednarodne in nacionalne novice s področna državne uprave, pravnih postopkov in družbenih razmerij, konfliktov, vojn.
- **znanost in tehnologija**. Znanstvena odkritja, zanimivosti, tehnološke inovacije, informacijska tehnologija, računalništvo.
- **družba**. Družbena vprašanja in razmerja, enakost, diskriminacija, religija, etika ipd.
- **šport**. Športni rezultati in zanimivosti z različnih športnih področij.
- **vreme**. Meteorološke napovedi, opisi vremenskih posebnosti, stanj, procesov.
- **zabava**. Estrada, moda, slog.
- **izobraževanje**. Procesi posredovanja in pridobivanja znanja ter veščin. Vse stopnje izobraževanja, od vrtca do univerzitetnega izobraževanja, pa tudi vseživljenjsko učenje.

Kot prikazuje primerjalna Tabela 1, obstaja precejšnja prekrivnost tako s kategorijami novičarskih portalov kot s kategorijami IPTC in tujih korpusov. V nekaterih primerih, npr. *gospodarstvo, prosti čas, politika* in *družba*, naša kategorija zajema več kategorij ostalih virov. Tako ima za prosti čas estonski korpus kar sedem ločenih kategorij. Edini primer, ko se eno od kategorij tujih virov lahko uvrsti v dve naši, sta *umetnost in kultura* ter *zabava*. Kategoriji smo namreč ločili po eni strani zato, ker ima veliko slovenskih novičarskih portalov ločene podstrani zanju, po drugi strani pa zaradi samega jezika - kulturno-umetniške vsebine so za razliko od zabavnih pogosto precej bolj strokovne.

Medtem ko v naše kategorije lahko umestimo vseh 17 kategorij IPTC, pa češki oz. estonski korpus določenih kategorij nimata, npr. estonski nima *črne kronike*, češki pa ne *okolja, zdravja, znanosti in tehnologije* ter *zabave*. Oba tudi nimata ločene kategorije za *vreme*, ki pa jo ima IPTC in smo jo dodali zato, ker jo ima večina slovenskih novičarskih portalov.

Če pogledamo še prekrivnost kategorij s stranmi oz. podstranmi šestih slovenskih novičarskih portalov, vidimo, da so problematične kategorije predvsem *politika*, *družba* in *izobraževanje*. Gre za sicer legitmne kategorije, ki pa na novičarskih portalih nimajo svojih podstrani, temveč so novice razpršene po drugih podstraneh, ki so večinoma opredeljene glede na geografski izvor novice, npr. Slovenija, Svet, Lokalno. Medtem ko so se avtorji češkega korpusa odločili slediti takšni delitvi tudi pri kategorijah (*current events, foreign news, domestic news, regional news*), smo se mi raje držali tematike. To za izdelavo učnih množic pomeni nekoliko več ročnega dela oz. iskanje drugih kazalcev, s katerimi lahko odkrijemo tematiko prispevka na posameznem portalu. Izjema je portal sta.si, ki že ima ustrezne kategorije, in sicer *Šolstvo* in *Družba*, za politiko pa *Državni zbor*, *Evropska unija*, *Mednarodna politika, Slovenska notranja politika* in *Slovenska zunanja politika*.

Učne množice smo izdelali z mapiranjem kategorij različnih virov novic na našo interno kategorizacijo. Tako lahko besedila iz določenih kategorij konkretnih virov uporabimo za učenje modela. Pri pripravi učnih množic bomo vzorčili tako količino podatkov iz posameznega vira kot količino podatkov v kategoriji in s tem zagotovili raznolikosti učnih množic, pa tudi robustnost končnega modela.

Za modeliranje bomo uporabili orodje fasttext (Joulin et al, 2016) z vložitvami CLARIN.SI (Ljubešić in Erjavec, 2018) in model SloBERTa (Ulčar in Robnik-Šikonja, 2021). Glede na razliko v rezultatih (pričakujemo, da se bo model SloBERTa odrezal boljše, a morda razlika v rezultatih ne bo tako opazna) in kompleksnosti klasifikatorja (fasttext je precej hitrejši in zahteva bistveno manj spominskih kapacitet), bomo izbrali klasifikator, ki ga bomo uporabili na novih besedilih.

[16] https://cv.iptc.org/newscodes/subjectcode

kategorija	slovenski portali (6)	češki korpus	estonski korpus	IPTC
umetnost in kultura	5	culture	culture & entertainment	arts, culture and entertainment
črna kronika	6	crime	/	disaster and accident
gospodarstvo	6	economy	economy, finance & business; agriculture; construction & real estate	economy, business and finance; labour
okolje	2	/	nature & environment	environmental issue
zdravje	3	/	health	health
prosti čas	4	leisure	beauty; cars; food & drinks; gambling & casinos; home, family & children; pets and animals; travel & tourism; video games	lifestyle and leisure
politika in pravo	1	politics	politics & government	politics; crime, law and justice; unrest, conflicts and war
znanost in tehnologija	5	/	science, technology & IT	science and technology
družba	1	social life	society; religion; sex; women	social issue; religion and belief; human interest
šport	6	sports	sports	sport
vreme	4	/	/	weather
zabava	4	/	culture & entertainment*	arts, culture and entertainment*
izobraževanje	1	/	education	education

Tabela 1: Primerjava tematskih kategorij projekta SLED z domačimi novičarskimi portali in tujimi viri.

3.6. Rezultati uporabniške ankete

Ker je Trendi prvi korpus svoje vrste v slovenskem okolju, smo ga želeli zasnovati karseda skladno z uporabniškimi pričakovanji. Ta smo v decembru 2021 preverili s pomočjo uporabniške ankete, s katero smo ugotovili, katerih podatkov o aktualni rabi jezika si raziskovalna skupnost želi in v kakšni obliki (npr. različni seznami, kot so kandidati za neologizme, besede in besedne zveze z najbolj izstopajočo rabo v določenem obdobju (dnevu, tednu, mesecu), izstopajoče besede in besedne zveze glede na vir ipd.).

Anketa[17] je bila izdelana na platformi 1KA, sestavljena pa je bila iz 9 vprašanj: med temi je bilo 5 vsebinskih, 4 pa so zbirala demografske podatke (spol, starost, področje delovanja). Diseminirana je bila po e-poštnih seznamih slovenskih jezikoslovnih raziskovalnih skupnosti (npr. SlovLit ter e-poštni seznam Slovenskega društva za jezikovne tehnologije) ter po družbenem omrežju Facebook (na uradni strani Centra za jezikovne vire in tehnologije Univerze v Ljubljani ter v neformalnih jezikoslovnih uporabniških skupinah, kot je *Prevajalci, na pomoč!*).

V celoti izpolnjenih vprašalnikov je bilo 100. Vzorec, ki ga je zajela anketa, zajema predvsem osebe ženskega spola (82 %), manjši delež pa je moških (18 %). Po starosti vzorec zajema predvsem generacije med 26. in 55. letom starosti (80 % vseh udeleženk_cev), največ med 26. in 35. letom (33 %) in med 46. in 55. letom (32 %). Večina

[17] Podrobnejše poročilo o izvedeni anketi je na voljo na spletni strani projekta: https://sled.ijs.si/wp-content/uploads/2022/02/SLED_anketa_porocilo_2022-2-03_final.pdf

udeleženk_cev je zaposlenih bodisi v javnem sektorju (61 %) bodisi je samozaposlena (20 %), le manjši delež ima še študentski status (3 %) ali pa so zaposleni v podjetjih (6 %), upokojeni (4 %) ali v iskanju zaposlitve (5 %). Po področju delovanja, pri katerem so udeleženci_ke lahko izbrali_e več možnosti, prednjačita lektoriranje (60 %) in prevajanje (46 %), visok delež pa imajo tudi ljubiteljsko raziskovanje jezika (38 %), strokovno in znanstveno pisanje (34 %), jezikoslovne raziskave (32 %) ter kreativno pisanje in blogerstvo (22 %). Skupno 40 % zajemajo tudi različne kategorije poučevanja jezika (slovenščina kot 1. jezik na osnovni ali srednji šoli, slovenščina kot 2. ali tuji jezik, jezikoslovni predmeti na višji/univerzitetni ravni). Vzorec nakazuje, da je anketa zajela različna področja jezikoslovno-raziskovalnega udejstvovanja.

V nadaljevanju predstavljamo podrobnejšo analizo odgovorov na vsebinska vprašanja.

3.6.1. Scenariji uporabe in uporabniško zanimanje

Anketiranci_ke so navedli_e, kateri podatki v orodju, ki bi spremljalo aktualno jezikovno rabo, bi jih najbolj zanimali, in pri vsakem od 6 predlaganih scenarijev uporabe (s konkretnimi primeri za lažjo predstavo) ocenili_e svojo stopnjo zanimanja (1 - sploh me ne zanima, 5 - zelo me zanima). Med scenariji so npr. *katere besede/besedne zveze so najznačilnejše za določeno obdobje v primerjavi z drugim obdobjem*? (npr. katere besede so se mnogo pogosteje uporabljale v februarju 2020 kot pa v februarju 2021); *v katerem obdobju je določena beseda/besedna zveza najpogostejša*? (npr. ali je bila beseda "tajkun" res najpogostejša v obdobju 2008-2009?); *ali raba besede/besedne zveze v zadnjem obdobju glede na trende narašča ali pada?* (npr. ali se "epidemija" uporablja vse pogosteje ali vse redkeje?).

Rezultati kažejo, da se anketirancem_kam vsi predlagani scenariji zdijo zanimivi: kategoriji "Zanima me" (4) in "Zelo me zanima." (5) namreč pri vsakem scenariju skupaj zajemata med 74 in 88 %. Po stopnji zanimanja najbolj izstopa scenarij, v katerem je mogoče primerjati trend rabe dveh ali več besed/besednih zvez (npr. *anticepilec* vs. *proticepilec*), enako pa anketiranke_ce zanima tudi, ali raba določene besede/besedne zveze v zadnjem obdobju glede na trende narašča ali pada.

Dobre tri četrtine vprašanih (76 %) je odgovorilo, da bi jim podatki o aktualni jezikovni rabi koristili pri delu, le 9 % tovrstni podatki ne bi koristili (15 % je neodločenih). Rezultati ankete torej potrjujejo, da jezikoslovno skupnost podatki o trendih jezikovne rabe zanimajo in da obstaja realna potreba po jezikovnem viru, ki tovrstne podatke prinaša sprotno in ažurno.

3.6.2. Načini prikaza podatkov

Na lestvici od 1 (sploh ni pomembno) do 5 (zelo pomembno) so anketiranci_ke ocenili_e tudi, kateri od predlaganih načinov prikaza podatkov (grafi s trendi jezikovne rabe. tabele s številskimi podatki, seznami besed oz. besednih zvez z naraščajočo/padajočo rabo, drugo) se jim zdijo pomembni. Če združimo deleže kategorij "pomembno" (4) in "zelo pomembno" (5), dobimo deleže 79 % za grafe, 64 % za tabele s številskimi podatki in 87 % za sezname besed s padajočo/naraščajočo rabo. Anketiranke_ce torej najbolj zanimajo preprosti seznami, najmanj pa napredne tabele s številskimi podatki.

3.6.3. Uporabniški predlogi

V odprtem vprašanju so imeli anketiranci_ke možnost izraziti predloge oz. dodatne scenarije, ki bi jih zanimali o aktualni jezikovni rabi. Dodatnih predlogov je bilo 15. Nanašajo se npr. na povezljivost orodja z drugimi jezikovnimi viri (npr. integracija v Slovenski oblikoslovni leksikon Sloleks in v korpus pisne standardne slovenščine Gigafida) in dostop do podatkov (npr. možnost dostopa do podatkov preko javnega API-ja), primerjavo sopomenskih različic besed oz. besednih zvez (npr. *oče* vs. *ata*), vključitev zgledov rabe in spremljanje jezikovne rabe daljših enot (npr. frazemov). Večina dodatnih predlogov sicer presega obseg projekta SLED, a predstavljajo pomembno povratno informacijo za razmislek o prihodnjem razvoju in integraciji spremljevalnega korpusa in iz njega izluščenih podatkov v ostale jezikovne vire.

4. Sklep in nadaljnje delo

V prispevku smo predstavili različne aktivnosti projekta SLED, s poudarkom na korpusu Trendi, nastajajočem spremljevalnem korpusu slovenskega jezika. Opisali smo metodologijo njegove izdelave, vsebino in oblike, v katerih je na voljo uporabnikom_cam. Predstavili smo tudi klasifikacijo tematskih kategorij, ki je bila oblikovana za namene izdelave modela za avtomatsko tematsko kategorizacijo besedil. Zadnji del je bil namenjen predstavitvi rezultatov ankete o uporabniških pričakovanjih o podatkih o aktualni rabi jezika, ki jih želi imeti zainteresirana skupnost.

V prihajajočih mesecih bomo nadaljevali z objavami mesečnih različic korpusa, pripravili prve statistične izračune in dokončali ter evalvirali algoritem za avtomatsko kategorizacijo besedil. Pomembno je, da smo veliko časa posvetili vzpostavitvi avtomatskih postopkov priprave besedil in izračunov, saj bo to pospešilo posodabljanje podatkov v konkordančnikih in na repozitorju CLARIN.SI.

Prav tako je ključna aktivnost izboljšava postopka pridobivanja besedil, ki bo poskrbela, da bodo odpravljene določene pomanjkljivosti trenutne metode. Ker bo vzpostavljena tesna povezanost med korpusom Trendi in referenčnim korpusom Gigafida, bo vsaka izboljšava postopkov koristila obema korpusoma.

S korpusom Trendi je slovenska jezikovna infrastruktura bogatejša za pomemben vir, ki bo relevanten tako za raziskovalno skupnost kot širšo javnost.

5. Zahvala

Projekt SLED (*Spremljevalni korpus in spremljajoči podatkovni viri*) financira Ministrstvo za kulturo Republike Slovenije kot del *Javnega razpisa za (so)financiranje projektov, namenjenih gradnji in posodabljanju infrastrukture za slovenski jezik v digitalnem okolju 2021–2022*. Raziskovalna programa št. P6-0411 (*Jezikovni viri in tehnologije za slovenski jezik*) in št. P6-0215 (*Slovenski jezik - bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

6. Literatura

Mark Davies. 2008–. The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/.

Mark Davies. 2016–. Corpus of News on the Web (NOW). Available online at https://www.english-corpora.org/now/.

Mark Davies. 2019–. The Coronavirus Corpus. Available online at https://www.english-corpora.org/corona/.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6thWorkshop on Balto-Slavic Natural Language Processing*, BSNLP@EACL 2017, str. 33–38.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: P*roceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).*

Ondrej Herman. 2013. *Automatic methods for detection of word usage in time.* Diplomska naloga. Masaryk University, Faculty of Informatics.

Armand Joulin, Edouard Grave, Piotr Bojanowski in Tomas Mikolov. 2016. *Bag of Tricks for Efficient Text Classification*. arXiv. https://arxiv.org/abs/1607.01759.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, Lorient, France*, str. 105-116. Lorient: Université de Bretagne Sud.

Kristina Koppel in Jelena Kallas. (v tisku). *Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu.* Eesti Rakenduslingvistika Ühingu aastaraamat.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt in Jaka Čibej. 2021. Language monitor: tracking the use of words in contemporary Slovene. V: I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek in C. Tiberius , ur., *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual*, str. 514–527. Brno: Lexical Computing CZ, s.r.o., https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_33_pp514-528.pdf.

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

Luka Krsnik et al. 2019. *Corpus extraction tool LIST 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1276.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička in Adrian Jan Zasina. 2016. SYN2015: Representative Corpus of Contemporary Written Czech. V: P*roceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 2522–2528, Portorož, Slovenia. European Language Resources Association (ELRA).

Nikola Ljubešić in Kaja Dobrovoljc. 2019. What does Neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, str. 29–34.

Nikola Ljubešić in Tomaž Erjavec. 2018. *Word embeddings CLARIN.SI-embed.sl 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1204.

Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, in Peter Holozan. 2013. *Written corpus ccGigafida 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1035.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. V: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, str. 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton in Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Pavel Rychlý. 2007. Manatee/Bonito-A Modular Corpus Manager. V: RASLAN, str. 65–70.

Mitja Trampuš in Blaž Novak. 2012. The Internals Of An Aggregated Web News Feed. V: *Proceedings of 15th Multiconference on Information Society 2012* (IS-2012). http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.

Matej Ulčar in Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. V: *Proceedings of the 24th International Multiconference – IS2021* (SiKDD). https://ailab.ijs.si/dunja/SiKDD2021/Papers/Ulcar+Robnik.pdf.

Katja Zupan, Nikola Ljubešić in Tomaž Erjavec. Smernice za označevanje imenskih entitet v slovenskem jeziku. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1238/SlovenianNER-slv-v1.0.pdf?sequence=7&isAllowed=y.