Damjan Popič
Univerza v Ljubljani
Filozofska fakulteta

Matejka Grgič
Univerza v Ljubljani
Filozofska fakulteta

DOI: 10.4312/jis.68.4.73-87
1.01

KORPUS LEKTORIRANIH SLOVENSKIH BESEDIL Z OBMOČJA ITALIJANSKO-SLOVENSKEGA JEZIKOVNEGA STIKANJA – STIKit: ZASNOVA, GRADNJA IN IZZIVI[1]

Prispevek predstavlja zasnovo, gradnjo in strukturo korpusa lektoriranih besedil s prostora jezikovnega stika med slovenščino in italijanščino, zlasti na območju poselitve avtohtone slovenske narodne skupnosti v Italiji. Slovenska skupnost v Italiji je razvila specifične jezikovne poteze, ki so bile predvsem v zadnjih dveh desetletjih predmet številnih raziskav, vendar pa so bile te omejene na manjše korpuse besedil oziroma na posamezne pojave jezikovnega stikanja. Na podlagi takih analiz se je oblikovala teza o jezikovnem separatizmu in secesionizmu slovenščine v Italiji, ki pa je zaradi nezadostne količine podatkov in neustrezne infrastrukture ni bilo mogoče ne dokazati ne ovreči.

Z namenom, da to hipotezo preverimo, smo se odločili zgraditi korpus lektoriranih besedil s področja jezikovnega stika med slovenščino in italijanščino in na ta način s kvantitativnimi metodami v sklopu projekta STIKit natančneje ugotoviti, a) v kolikšni meri se jezikovni separatizem pojavlja na območju poselitve slovenske skupnosti v Italiji in b) s katerimi jezikovnimi sredstvi se ta separatizem (najpogosteje) vzpostavlja. Splošni namen projekta pa je priprava in postavitev celostnega pregleda in jezikovne rabe na območju stika med slovenščino in italijanščino, ki bo v končni fazi prinesel globlje razumevanje procesov na tem območju ter – upamo – tudi jezikovnočrtovalske odzive in oblikovanje inovativnih jezikovnih orodij.

Ključne besede: STIKit, jezikovni stik, slovenščina v Italiji, sociolingvistika, jezikovne tehnologije

[1] Članek je nastal v okviru raziskovalnega programa št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave), ki ga je sofinancirala Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna. Tehnično in infrastukturno podporo za projekt STIKit in druge projekte, namenjene (jezikovnotehnološki) skrbi za slovenščino na območju jezikovnega stika med slovenščino in italijanščino, financira Centralni urad za slovenščino pri Avtonomni deželi Furlaniji - Julijski krajini, v sodelovanju s Slovenskim raziskovalnim inštitutom - SLORI.

A corpus of edited Slovenian texts from the area of Italian-Slovenian language contact – STIKit: design, construction, and challenges

This paper presents the design, construction and structure of a corpus of proofread texts from the area of linguistic contact between Slovene and Italian, especially in the area of settlement of the indigenous Slovene ethnic community in Italy. The Slovene community in Italy has developed specific linguistic traits that have been the subject of numerous studies, especially in the last two decades, but these have been limited to small corpora of texts or to individual phenomena of language contact. Such analyses have led to the development of a presumption of linguistic separatism and secessionism of Slovene in Italy, which, however, could not be proved or disproved due to insufficient data and inadequate infrastructure.

In order to test this hypothesis, we decided to build a corpus of proofread texts in the field of Slovenian-Italian linguistic contact and thus to determine more precisely a) to what extent linguistic separatism occurs in the area of Slovenian settlement in Italy, and b) by which linguistic means this separatism is (most often) established. The overall purpose of the STIKit project is, of course, to be able to develop and establish a comprehensive overview and understanding of language use in the area of Slovenian-Italian language contact, which will ultimately lead to a deeper understanding of the processes in this region and – hopefully – to better language-planning responses and the design of innovative language tools and resources.

Keywords: STIKit, language contact, Slovene in Italy, sociolinguistics, language technologies

1 Uvod

V zadnjih dveh desetletjih so bile opravljene številne raziskave (Jagodic idr. 2017; Pertot in Kosic 2014), ki poudarjajo, da je slovenščina v Italiji razvila samosvoje poteze, in sicer tudi na ravni tistih lokalnih različic, ki jih govorci in govorke dojemajo kot standardne (Grgič 2017). Izvor teh posebnosti so raziskave povečini pripisovale stiku z romanskimi jeziki in drugim dejavnikom, predvsem geografski obrobnosti, specifični politični zgodovini in prisotnosti državnih meja (Jagodic idr. 2017). Vendarle pa so nekatere novejše analize, npr. vzorčna raziskava rabe covidne terminologije (Grgič in Popič 2023), pokazale, da odstopanj, kot se izkazujejo v jezikovni rabi, ti dejavniki preprosto ne morejo zadovoljivo pojasniti, saj je šlo na primer pri epidemiji covida-19 za globalen pojav, pri terminologiji pa za izjemno dinamično in sočasno prevodno dejavnost iz (globalne) angleščine, pri čemer so se nove jezikovne rabe v medmrežju širile viralno.

Zaradi teh specifik je bila ravno covidna terminologija odlična priložnost za preverjanje trenutnih terminotvornih praks na območju jezikovnega stika. Raziskava je tako pokazala znatno odstopanje med slovensko covidno terminologijo v Sloveniji in slovensko covidno terminologijo v Italiji, obenem pa je pritrdila tudi prejšnjim raziskavam (Grgič 2018), ki so na sicer veliko manjših vzorcih ugotavljale, da se terminologija na tem območju prevaja iz italijanščine in se nato (le delno in povečini arbitrarno) prilagaja slovenskemu standardu.[2]

[2] Za prikaz (ne)razumljivosti terminologije navajamo nekaj terminov: *okrepljeno zeleno/covidno potrdilo* iz ital. (*super*) *green pass* (’pogoj/potrdilo PCT‘), *smart working/telematsko delo* (’delo na daljavo/od doma‘), *didaktika v prisotnosti* (’pouk v živo‘) iz ital. *didattica in presenza* itd.

Rezultati so bili vsaj delno presenetljivi, saj bi pri rabi terminologije, ki ne označuje lokalnospecifične predmetnosti, pričakovali malo oziroma bistveno manj razhajanja od jezikovnega standarda. Ta trend kaže, da je razhajanje med lokalnim in splošnim standardom večje od tega, kar smo doslej predvidevali, in da ni prisotno samo na ravni neformalnega sporazumevanja v lokalnih jezikovnih različicah, kjer je sicer variantnost pričakovana (Chambers 2003). Hkrati pa se je ob pregledu korpusnih podatkov, ki so kazali na skoraj izključno lokalno rabo, oblikovala hipoteza, da govorci in govorke to rabo dojemajo kot splošno/standardno – po vsej verjetnosti zaradi neustrezne izpostavljenosti nelokalnim sporazumevalnim praksam v slovenskem jeziku.

Da bi preverili to hipotezo, smo se odločili pripraviti sistematično ogrodje, ki bi omogočilo tako kvantitativno spremljanje razvoja slovenščine v Italiji kot tudi kvalitativni vpogled v sodobno pisno produkcijo obravnavanega območja. V ta namen smo zasnovali projekt in korpus *STIKit*, korpus lektoriranih (prevodnih in avtorskih) besedil z območja jezikovnega stika med slovenščino in italijanščino, s čimer projekt dopolnjuje že vzpostavljeno jezikovno orodje Loris (gl. Popič 2022),[3] s katerim prav tako sistematično spremljamo jezikovno rabo na tem območju, vendar se pri tem osredotočamo predvsem na jezikovne elemente, pri katerih pisci iščejo pomoč oziroma hipno jezikovno svetovanje.[4]

V nadaljevanju najprej izpostavimo temeljne značilnosti jezikovne rabe na področju jezikovnega stika med slovenščino in italijanščino, nato pa predstavimo metodološko in gradivno osnovo nastajajočega korpusa lektoriranih slovenskih besedil v Italiji.

2 Izhodišča

Posledice jezikovnih stikanj segajo onkraj neposrednih jezikovnih sprememb in vplivajo na družbeno-kulturno in kognitivno krajino vpletenih skupnosti. Učinki stika so večplastni in ne vplivajo zgolj na strukturo in rabo jezika, ampak tudi na sporazumevalne prakse in identitetne opcije govorcev in govork, pri čemer so jezikovne izbire pogosto ideološko pogojene (Meyerhoff 2008).

V nadaljevanju opišemo trenutno stanje slovenščine v Italiji oziroma proces t. i. jezikovnega separatizma, ki ga tu pojmujemo kot pospešeno in stalno oddaljevanje neke lokalne različice od splošnega standardnega jezika, pa tudi od nestandardnih in drugih lokalnih idiomov, ter jezikovnega secesionizma, ki ga tu dojemamo kot vedno pogosteje prisotno in izraženo stališče lokalne skupnosti govork in govorcev o edinstvenosti lokalnih jezikovnih pojavov, ki naj ne bi bili povezani s širšim (zalednim) jezikovnim kontinuumom.

---

[3] Povezava: https://www.jeziknaklik.it/loris/ (dostop 12. 12. 2023).

[4] Z zbiranjem besedil, ki jih uporabniki preverjajo v orodju Loris, namreč pridobivamo avtentična besedila, hkrati pa vidimo tudi, kateri jezikovni elementi jih zanimajo oziroma pri njihovi obdelavi potrebujejo pomoč.

2.1 Slovenščina v stiku med Slovenijo in Italijo

Slovenska skupnost v Italiji je razvila specifične jezikovne poteze, ki so bile predvsem v zadnjih dveh desetletjih predmet številnih raziskav (Grgič 2017; Jagodic idr. 2017; Mezgec 2016). Dosedanje raziskave pa so bile količinsko omejene na manjše korpuse besedil oziroma na preverjanje domnev o posameznih pojavih jezikovnega stikanja s kvalitativnimi raziskavami, npr. delno strukturiranimi intervjuji ali z opazovanjem z udeležbo (Grgič 2022). Zaradi neustrezne infrastrukture in omejene količine podatkov so analize sicer delno potrdile prisotnost tovrstnih pojavov, niso pa je mogle kvantitativno ovrednotiti (Popič 2022: 282).

Pretežno vzorčni podatki iz omejenih korpusov besedil so pokazali, da prihaja na območju slovenske poselitve v Italiji do postopnega razhajanja jezikovnega koda od osrednjega standarda slovenščine, tako da se lokalne rabe (tudi tiste, ki jih skupnost dojema kot standardne) »včasih že močno distancirajo od slovenskega jezikovnega kontinuuma« (Grgič 2017: 95). Pretekle raziskave (vse povzete v Jagodic idr. 2017) izpostavljajo predvsem dve značilnosti razhajanja zamejske slovenščine od osrednjega standarda oziroma vzpostavitve paralelne rabe, in sicer a) vpliv italijanskega jezika ter b) omejenost govornih položajev, v katerih se slovenščina v tem okolju uporablja, pri čemer je verjetno slednji vidik tudi (vsaj delni) povzročitelj prvega. V nadaljnji fazi so se prve raziskave, ki so temeljile na sicer le vzorčni analizi besedil (Pertot in Kosic 2014; Grgič 2017), že osredotočile predvsem na (samo)marginalizacijo zamejske skupnosti v odnosu do t. i. matice:

> /N/a naselitvenem območju slovenske manjšine v Italiji /so se/ že uveljavili procesi, ki imajo nekatere značilnosti pidžinizacije, folklorizacije, okamnitve, oslabitve in opuščanja (zamenjave) slovenskega jezika. Ti procesi niso posledica nezadostne pravno-formalne zaščite ali nižjega statusa oz. prestiža manjšinskega jezika, ampak predvsem pomanjkanja pasivne (input) in aktivne (output) izpostavljenosti različnim rabam slovenskega splošnosporazumevalnega jezika in drugih idiomov slovenskega jezikovnega kontinuuma (Grgič 2017: 97).

Take hipoteze lahko preverjamo po eni strani s kvalitativnimi metodami, npr. delno strukturiranimi intervjuji in analizo besedil, po drugi strani pa s kvantitativno analizo ustrezno zasnovanih reprezentativnih korpusov besedil.

2.2 Korpusnojezikoslovni pristop k proučevanju jezikovne podobe slovenščine v stiku z italijanščino

S korpusnojezikoslovnim pristopom lahko v obsežnejših zbirkah besedil preverimo, katere izbire govorcev in govork slovenščine v Italiji se razlikujejo od izbir govork in govorcev slovenščine v Sloveniji ter posledično od besedil, ki nastajajo v primerljivih okoliščinah in s primerljivim namenom. Tako lahko spremljamo stopnjo oddaljenosti dveh ali več kodov, variant oziroma idiomov znotraj danega jezikovnega kontinuuma (Grgič in Popič 2023). Oddaljenost je najbolj vidna v

besedilih, v katerih je pričakovana variantnost minimalna – to so strukturirana besedila z ustaljenimi sporazumevalnimi vzorci in visoko koncentracijo terminov za poimenovanje iste predmetnosti (eno-/istoreferenčni termini oziroma, drugače rečeno, termini z eno/isto referenco na obeh straneh meje).

S korpusnojezikoslovno raziskavo terminologije, vezane na epidemijo covida-19 (Grgič in Popič 2023), smo tako preverili, do kolikšne stopnje variantnosti prihaja v sodobnih zamejskih javnih besedilih: v korpus smo vključili dva letnika *Primorskega dnevnika* (2020 in 2021; okoli 20 milijonov besed). Korpusnojezikoslovna analiza je pokazala, da prihaja na področju terminologije in tudi v širši javni rabi do precejšnjega razhajanja med zamejskim in osrednjim jezikovnim standardom. Jezikovno (in specifično leksikalno) variantnost običajno pričakujemo pri manj formalnih sporazumevalnih kodih, pri referiranju na različno predmetnost/pojavnost in, seveda, pri jezikovnih različicah (idiomih), ki jih sami govorci in govorke prepoznavajo kot lokalne/regionalne (Fought 2006: 37). Korpusna analiza slovenske covidne terminologije na obeh straneh meje pa je pokazala, da je leksikalna variantnost v besedilih, ki nastajajo na območju poselitve slovenske skupnosti v Italiji, znatna tudi pri strokovni terminologiji (in bi torej pričakovali večjo leksikalno enotnost), za ubesedovanje predmetnosti/pojavnosti, ki se na obeh straneh meje ne razlikuje, ter za besedila, ki jih govorci in govorke dojemajo kot standardna.

To torej pomeni, da je leksikalno razhajanje med obema kodoma znatno, obenem pa to naznanja tudi precejšnjo verjetnost, da razlike najdemo tudi na drugih ravneh jezikovnega opisa, zato pa seveda potrebujemo širše ogrodje za analizo jezikovne rabe, k čemur stremi projekt STIKit. Izjemno pomemben vidik projekta pa je prav to, da poskušamo najti odgovor na to, v katero smer se razvija jezik govorcev in govork slovenščine v Italiji, ki se opredeljujejo za Slovence in Slovenke, pri čemer izhajamo iz izhodišča, da je vsak jezikovni kontinuum in vsak njegov (lokalni) segment artikulirana, kompleksna struktura sporazumevalnih sredstev, podkodov, ki jih govorci in govorke izbirajo in uporabljajo glede na njihovo pragmatično in simbolično funkcijo (Chambers 2003; Grgič 2022). Zaradi tega je torej nemogoče govoriti o splošnih »značilnostih slovenščine v Italiji«, obenem pa je težko določiti mejo med standardnimi in nestandardnimi različicami v jeziku (Davila 2016).

2.2.1 Korpus jezikovnih popravkov kot jezikovni vir o slovenščini v stiku z italijanščino

Slovenščina je razmeroma dobro opremljena s korpusi jezikovnih popravkov, še posebej glede na to, kako kadrovsko, časovno in finančno intenzivna je njihova gradnja.[5] Kot pravijo Arhar Holdt, Kosem in Stritar Kučuk (2022: 24), je

[5] Korpusi jezikovnih popravkov so korpus *Šolar* (Arhar Holdt idr. 2022), korpus besedil govorcev slovenščine kot tujega jezika *KOST* (Stritar Kučuk 2022) in korpus lektoriranih besedil *Lektor* (Popič 2014).

> /p/riprava korpusov z označenimi jezikovnimi popravki /…/ zapletena in počasna, saj poleg običajnih korakov priprave korpusnih besedil, kot so pravno urejeno pridobivanje besedil, strojno označevanje in formatiranje, zahteva tudi dodatne korake: od transkribiranja (pogosto ročno napisanih) besedil, anonimizacije do ročnega označevanja in vsebinskega kategoriziranja jezikovnih popravkov. Do nedavno so raziskovalci in raziskovalke, ki so gradili tovrstne korpuse, za naštete naloge iskali in prilagajali orodja, specializirana za kak drug namen, ter se spopadali z zapleteno metodologijo. Ta je pogosto vodila v napake pri ročnem delu in zahtevala redno tehnično podporo.

Večina izzivov, ki jih izpostavljajo avtorji, še vedno predstavlja znatno oviro pri gradnji tovrstnih (t. i. razvojnih) korpusov šolarjev in korpusov usvajanja tujega jezika (Leech 1998; Stritar Kučuk 2012), podobno pa velja tudi za rabo tovrstnih korpusov za preverjanje jezikovne rabe drugih skupin pišočih kot pri korpusu *STIKit* ali korektorskih praks jezikovnih delavcev kot pri korpusu *Lektor*.

Namen projekta STIKit je zato zastavljen zelo široko: zasnovati želimo obsežen jezikovni vir, ki bo služil kot osnova za nadaljnje raziskave slovenščine v Italiji, obenem pa tudi za jezikovne vire (kakršen je denimo jezikovni svetovalec Loris) in kot učna množica za velike jezikovne modele (npr. ChatGPT) ter seveda za jezikovne priročnike – predvsem tiste, namenjene učencem in učenkam šol s slovenskim jezikom in dvojezičnim italijansko-slovenskim poukom v Italiji ter specifičnim skupinam uporabnikov in uporabnic (novinarjev in novinark, prevajalcev in prevajalk itd.). Poleg tega nam bo tako obsežen korpus, ki bo združeval podatke o lokalni rabi in o lektorskih posegih vanjo, posredno omogočal primerjavo percepcije standarda v regionalnem in nadregionalnem okolju, kar bo omogočalo kvantifikacijo pojavov, kakršna sta jezikovni separatizem in secesionizem.

Pri projektu STIKit, ki združuje tako temeljnoraziskovalne kot aplikativne cilje, sodelujemo raziskovalci in raziskovalke Univerze v Ljubljani, kolegi in kolegice Slovenskega raziskovalnega inštituta (SLORI) v Trstu ter strokovno osebje Centralnega urada za slovenski jezik pri Avtonomni deželi Furlaniji - Julijski krajini.

Za gradnjo tako zahtevnega jezikovnega vira smo se odločili predvsem zaradi določenih prednosti, institucionalnih in tehničnih, ki jih v trenutnem konzorciju imamo, in sicer:

– proces zbiranja besedil je samodejen, saj se besedila zbirajo in lektorirajo na SLORI-ju ter Centralnem uradu za slovenski jezik, so avtentična, že digitalizirana, obenem pa so tudi popravki digitalni;
– zaradi dolgotrajnejše skrbi za slovenski jezik obeh ustanov v zamejstvu je fond nabranih besedil že obsežen, obenem pa tudi razmeroma strukturiran in organiziran;
– nedavni napredki na področju izdelave in priprave tovrstnih korpusov so močno olajšali in tehnično poenostavili samo označevanje korpusov;
– izdelava korpusa jezikovnih popravkov zamejskih besedil je logično nadaljevanje sistematizacije skrbi za jezik na območju jezikovnega stika med

slovenščino in italijanščino, ki ga tudi oba akterja, tako SLORI kot tudi Centralni urad, prepoznavata kot bistvenega za nadaljnje aktivnosti.

V nadaljevanju predstavimo strukturo in zasnovo projekta in korpusa *STIKit*.

3 Projekt in korpus *STIKit*

V tem poglavju predstavimo zasnovo in strukturo projekta STIKit, delotok priprave korpusa, podatkovno bazo in pa prikaz zvrstne distribucije besedil.

3.1 Besedilna osnova in metapodatki

Besedila, vključena v korpus *STIKit*, so avtorska in prevodna, vsa pa prihajajo z območja jezikovnega stika med slovenščino in italijanščino. Pri tem gre povečini za:
– lektorirana besedila s področja delovanja SLORI-ja (brezplačne lekture krajših besedil, ki nastajajo znotraj slovenske skupnosti v Italiji, npr. društev, ustanov, šol …);
– daljša lektorirana besedila naročnikov z območja jezikovnega stika (promocijsko gradivo, brošure, novinarska besedila, strokovna besedila ipd.);
– prevodi, ki jih v okviru svojega delovanja opravljajo Centralni urad za slovenski jezik ali druge prevajalske službe občin in javnih zavodov, vključenih v Mrežo za slovenski jezik v Furlaniji - Julijski krajini.[6]

V času priprave prispevka je bilo pri pripravi korpusa zbranih 846 različnih besedil s skupno 276.279 besedami, pri čemer smo žanrsko oznako do datuma priprave prispevka[7] med označevanjem pripisali 304 besedilom (skupno 189.059 besed). Trenutno zvrstno strukturo korpusa ponazarja tabela 1.

Zvrst	**Število besedil**
Leposlovno besedilo	5
Poljudnoznanstveno besedilo	74
Poročevalsko besedilo	232
Splošnosporazumevalno besedilo	69
Strokovno besedilo	2
Strokovno upravno-pravno besedilo	14
Skupna vsota	**396**

Tabela 1: Besedilnozvrstne oznake v korpusu STIKit

[6] Na spletišču Jezik na klik (https://www.jeziknaklik.it/obrazci-gradivo/#podatkovne-baze, dostop 12. 12. 2023) so na voljo pomnilniki prevodov z vključenimi zadnjimi različicami prevodnih besedil.

[7] Stanje dne 21. 11. 2023.

Kot ponazarja tabela 1, smo do trenutka priprave pričujočega prispevka besedilnozvrstne oznake pripisali 396 besedilom, pri tem pa je distribucija oznak močno nagnjena proti poročevalskim besedilom, kar je v veliki meri pričakovano in tudi odseva resnično stanje pisne produkcije na območju jezikovnega stika med slovenščino in italijanščino. Razmeroma dobro zastopana so še poljudnoznanstvena in splošnosporazumevalna besedila, medtem ko so druge zvrsti bistveno redkejše. Razlog za to je tudi v tem, da smo do tega trenutka obravnavali zgolj besedila, ki so bila lektorirana na SLORI-ju, pričakujemo pa, da se bo razmerje med zvrstmi nekoliko uravnotežilo, ko bodo obravnavana tudi besedila Centralnega urada za slovenski jezik, zlasti v prid strokovnih upravno-pravnih besedil.

Da bi pridobili kar najbolj podroben vpogled v zastopanost pojavov jezikovnega stika in jezikovno rabo nasploh v čim več žanrih besedil, besedilom pripisujemo tudi metapodatek o žanru. Trenutno žanrsko strukturo besedil ponazarja tabela 2.

Žanrska oznaka	**Število besedil**	**Skupaj besed**	**Besed na besedilo**
komentar/kolumna	58	45.235	780
poročilo	48	39.511	823
intervju	35	30.260	865
vabilo	25	6.348	254
novičnik	20	7.526	376
razstava (panoji)	20	8.663	433
obvestilo za javnost	17	5.290	311
dopis	16	5.173	323
obrazec	10	6.415	642
oglas	10	1.744	174
poljudnoznanstveni članek	7	7.451	1.064
vest	7	2.964	423
igra	6	4.652	775
tiskovno sporočilo	6	2.350	392
uvodnik za tiskani medij	5	2.357	471
turistična brošura	4	2.935	734
informativna zgibanka	3	3.253	1.084
pravilnik	3	229	76
življenjepis	2	543	272
anketni vprašalnik	1	301	301
didaktično gradivo	1	1.859	1.859
		185.059	

Tabela 2: Žanrske oznake v korpusu STIKit in število besed v posamezni kategoriji

Kot prikazuje tabela 2, smo besedilom v korpusu *STIKit* (zaenkrat) pripisali 21 različnih žanrskih oznak, predvsem v želji, da bi lahko na koncu ugotovili, ali je pojavnost posameznih jezikovnih elementov povezana s specifično zvrstjo ali zvrstmi. Še ena pomembna značilnost besedilne osnove korpusa *STIKit* je ta, da poskušamo zagotoviti tako žanrsko kot tudi besedilno pestrost, zato se ogibamo izrazito dolgim besedilom (npr. monografijam), ki bi lahko močno vplivala na distribucijo jezikovnih pojavov v korpusu. V ta namen namenoma izbiramo večje število krajših besedil, kar potrjuje tudi zadnji stolpec tabele 1, ki ponazarja, da je povprečna dolžina besedil v posameznih žanrih razmeroma enakovredna.

Besedila v korpusu pa opremljamo tudi s podatki o vrsti besedila, tj. z informacijo, ali gre pri posameznem besedilu za prevod ali avtorsko besedilo. To je z vidika proučevane problematike izjemno pomembno, saj želimo preveriti, ali se pri prevodnih besedilih uporabljajo drugačne strategije kot pri tvorjenju avtorskih besedil, čeprav so tudi slednja – kot smo pokazali v preteklih raziskavah – v resnici »prevedena«, saj nanje vplivajo italijanske predloge. Trenutno distribucijo avtorskih besedil in prevodov ponazarja tabela 3.

Vrsta (prevod/avtorsko)	**Število besedil**
avtorsko besedilo	339
neznano	2
prevod	100
Skupna vsota	**441**

Tabela 3: Razmerje med avtorskimi besedili in prevodi v korpusu STIKit

Kot prikazuje tabela 3, je razmerje med avtorskimi besedili in prevodi razmeroma močno nagnjeno proti avtorskim besedilom, kar bomo poskušali uravnotežiti z dodajanjem besedil Centralnega urada za slovenski jezik in Mreže za slovenski jezik Furlanije - Julijske krajine. Cilj je, da bi bila končna struktura korpusa čim bolj uravnotežena.

3.1.1 Ostali metapodatki

Poleg navedenih lastnosti besedil sistematično pripisujemo korpusu še nekatere druge metapodatke, s katerimi želimo omogočiti čim bolj učinkovito in sistematično križanje različnih vrst podatkov in iskanje potencialnih povezav. Ti podatki so naslednji:

– podatki o publikaciji,
– podatki o občini izvora,
– podatki o pokrajini izvora.

Podatki o publikaciji so pomembni zlasti z vidika jezikovne podobe različnih občil v zamejstvu, medtem ko želimo podatke o občini in pokrajini (Trst, Gorica in Videm) uporabiti za lokalno in regionalno umeščanje jezikovnih elementov ter ugotavljanje, ali je jezikovni stik na katerem od področij še posebej intenziven in s katerimi jezikovnimi sredstvi se vzpostavlja.

3.2 Označevanje korpusa *STIKit*

V tem odseku predstavimo nabor oznak in označevalni sistem za korpus *STIKit*. V želji, da bi pripravili nabor oznak, ki bi bil čim primerljiv z drugimi označevalnimi sistemi – in zaradi tega tudi pripravljen za računalniško procesiranje – smo kot izhodiščni sistem uporabili označevalni sistem korpusa *Šolar* (Arhar Holdt, Kosem in Stritar Kučuk 2022), pri tem pa smo ohranili zgolj vrhnje (splošnejše) kategorije, katerih oznake so predstavljene v tabeli 4.[8] Kategorije, vezane izrazito na šolska besedila, smo izpustili.[9]

PROBLEMI ZAPISA	**PROBLEMI SKLADNJE**	**PROBLEMI BESEDIŠČA**	**PROBLEMI OBLIKE**	**PROBLEMI ČRKOVANJA**
LOČILA	ODVEČ	MENJAVA	KATEGORIALNI POPRAVKI	KONZONANTI
MALA/VELIKA ZAČETNICA	IZPUST	GLAGOLI	PARADIGEMSKI POPRAVKI	VOKALI
SKUPAJ ALI NARAZEN	BESEDNI RED	SAMOSTALNIK	DODATNE OZNAKE	ČRKOVNI SKLOPI
ŠTEVILA	STRUKTURA	ZAIMEK		VARIANTNI PREDLOGI
KRAJŠAVE	DODAJANJE	PREDLOG		
		VEZNIK		
		DODATNE OZNAKE		
		PRIDEVNIK		
		PRISLOV		
		OSTALO		

Tabela 4: Nabor oznak za korpus STIKit

Tabela 4 prikazuje nabor oznak za korpus *STIKit*, ki je na vrhnji ravni identičen kot nabor oznak za korpus *Šolar*. Na nižjih ravneh, na katerih se oznake za korpus *Šolar* izrazito natančno delijo glede na najpogostejšo pisno prakso v šolskih okoljih, pa je označevalni sistem za korpus *STIKit* izrazito poenostavljen, saj

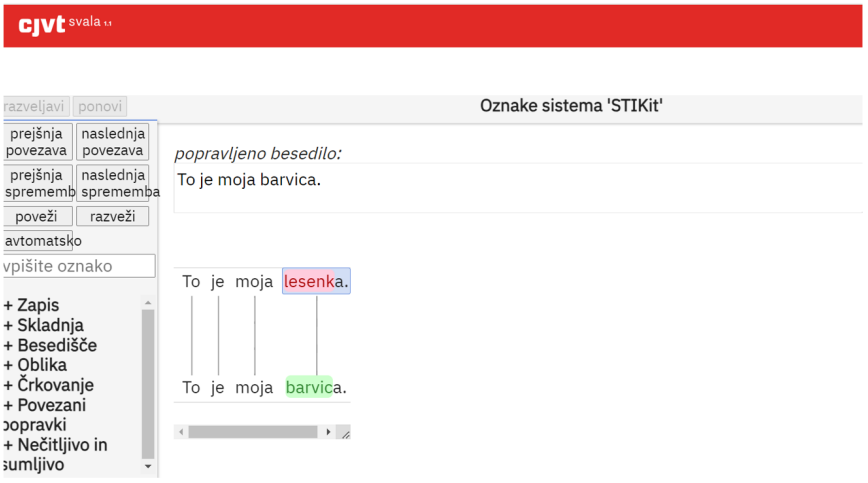
[8] Smernice za označevanje korpusa *Šolar* so na voljo na povezavi: http://hdl.handle.net/11356/1589 (dostop 12. 12. 2023).

[9] Oznake so podrobneje predstavljene v Arhar Holdt, Lavrič, Roblek in Goli (2022) in na povezavi: https://wiki.cjvt.si/books/11-jezikovni-popravki-solar/page/predstavitev-oznak (dostop 12. 12. 2023).

predvideva zgolj preproste binarizme tipa »stik : nestik«, kar poenostavljeno pomeni, da se označevalci – poleg tega, da jezikoslovno označujejo posamezne popravke – obenem še odločajo, ali je domnevni izvor za posamezen popravek jezikovni stik med slovenščino in italijanščino ali ne. Tovrstne odločitve so pogosto izjemno težke – pravzaprav lahko rečemo, da so težke v vseh primerih, razen pri najbolj očitnih primerih tipa *akantonacija*, *alikvota* ipd., ko leksem v standardni slovenščini ne obstaja in je beseda na območju stikanja tvorjena s podomačevanjem zapisa italijanskega leksema. To preprosto stikalo uporabljamo torej za označevanje primerov, pri katerih označevalec ali označevalka sumi, da gre za vpliv jezikovnega stika, kar nato nadalje preverjamo z dvojezičnim korpusom, ki je prav tako v nastajanju.

Glede na to, da se trenutno popravkom v korpusu *STIKit* pripisujejo vrhnje oznake, po koncu prve označevalske akcije načrtujemo kvantitativno obdelavo in kvalitativni pregled označenih besedil ter (najverjetneje) dopolnitev označevalske sheme s podrobnejšimi oznakami glede na rezultate vmesne analize. Pri tem bomo vrhnje kategorije obdržali usklajene s korpusom *Šolar*, tako da bo mogoča primerjava, tudi povsem avtomatizirana, nameravamo pa prilagoditi specifične oznake, če bo to potrebno.

Sam proces označevanja poteka v orodju Svala (Arhar Holdt idr. 2023), ki je bilo prilagojeno iz švedskega sistema Swell Editor (Volodina idr. 2019; Wirén idr. 2019), za različico 1.1 pa je bilo dopolnjeno še z možnostjo vnašanja komentarjev na besedilne elemente. To pomeni, da med označevanjem ne vnašamo zgolj popravkov, temveč lahko pripišemo tudi komentarje, ki jih je med popravljanjem vstavil lektor, učitelj ipd. Vmesnik orodja Svala ponazarja slika 1.



Slika 1: Vmesnik orodja Svala 1.1

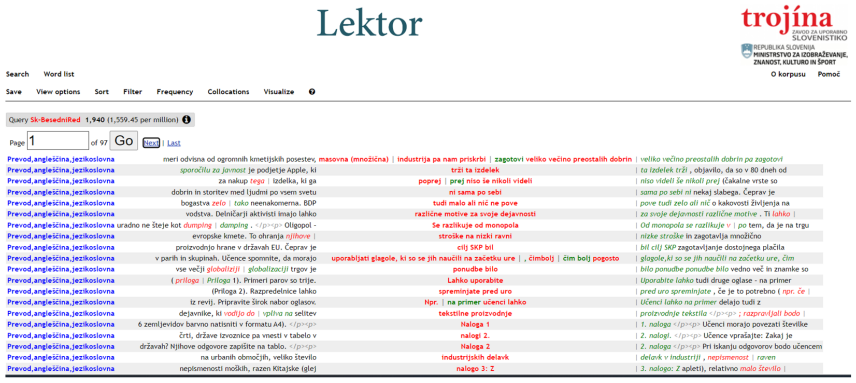
Kot ponazarja slika 1, je vmesnik orodja Svala za uporabnika bistveno prijaznejši kot prejšnje rešitve, ki so bile v največji meri vezane na urejanje izjemno dolge kode XML. Razlike med prvotnim in (večkrat) lektoriranim besedilom prikaže Svala grafično, tako da lahko brez večjih težav spremljamo spremembe. Na sliki 1 vidimo, da je bila beseda *lesenka* popravljena v *barvico*, kar bi označevalec označil z izbiro ustrezne oznake na levi strani (besedišče → samostalnik → stik). Ob tem bi bila kategorija zapisana v format JSON, ki ga po označevanju izvozimo in uporabimo za pripravo drugih formatov za konkordančnike in kvantitativno analizo, kot prikazuje spodnji izsek izhodnega formata.

```
{
   »source«: [
      {»id«: »s47«, »text«: »To »},
      {»id«: »s51«, »text«: »je »},
      {»id«: »s57«, »text«: »moja »},
      {»id«: »s67«, »text«: »lesenka. »}
   ],
   »target«: [
      {»id«: »t73«, »text«: »To »},
      {»id«: »t77«, »text«: »je »},
      {»id«: »t83«, »text«: »moja »},
      {»id«: »t91«, »text«: »barvica. »}
   ],
   »edges«: {
      »e-s47-t73«: {
         »id«: »e-s47-t73«,
         »ids«: [»s47«, »t73«],
         »labels«: [],
         »manual«: false
      },
      »e-s51-t77«: {
         »id«: »e-s51-t77«,
         »ids«: [»s51«, »t77«],
         »labels«: [],
         »manual«: false
      },
      »e-s57-t83«: {
         »id«: »e-s57-t83«,
         »ids«: [»s57«, »t83«],
         »labels«: [],
         »manual«: false
      },
      »e-s67-t91«: {
         »id«: »e-s67-t91«,
         »ids«: [»s67«, »t91«],
         »labels«: [»B/SAM/stik«],
         »manual«: false
      }
   }
}
```

3.3 Končni prikaz

Končni prikaz bo uporabnikom in uporabnicam omogočal hitro, učinkovito in zanesljivo spremljanje odstopanja jezikovne rabe od standarda. Za razliko od drugih sorodnih korpusov bo *STIKit* posebej omogočal spremljanje posebej označene lokalne rabe, ki jo skupnost dojema kot standardno, in kvantifikacijo odstopanj te rabe od splošnega slovenskega standarda.

Morda največja pomanjkljivost korpusov lektoriranih besedil v slovenščini je bila doslej prikaz podatkov in s tem možnost spremljanja sprememb. To je bržkone razumljivo, saj gre pri tovrstnih korpusih za izjemno kompleksen prikaz, pri katerem je treba sopostaviti tako popravljene jezikovne elemente kot tudi vstavljene, ob tem pa je pogosto v okviru enega jezikovnega popravka vključenih še več drugih. Slika 2 ponazarja rezultat iskanja v konkordančniku korpusa *Lektor* za popravke besednega reda.



Slika 2: Prikaz konkordančnika korpusa Lektor

Kot vidimo, je prikaz izjemno raztegnjen, saj so bili pri lekturah besednega reda popravljeni tudi izrazito dolgi segmenti, znotraj teh pa so bili popravljeni tudi drugi jezikovni elementi. Ker obstoječe rešitve torej ne odgovarjajo najbolj na zahteve sodobnih podatkovnih zbirk (poleg korpusa *STIKit* aktivno in redno potekajo tudi posodobitve korpusa *Šolar* in korpusa *KOST*), se za namene korpusa *KOST* pripravlja nova spletna rešitev za prikaz korpusov z jezikovnimi popravki. Pri vizualizaciji korpusa *STIKit* želimo uporabiti to rešitev, obenem pa jo obogatiti tudi s predmetnospecifičnimi metapodatki in informacijami ter vizualizacijami, vezanimi zlasti na geografski vidik rabe posameznega jezikovnega elementa.

4 Sklep

V prispevku so predstavljena sociolingvistična izhodišča, metodološki okvir in zgradba orodja *STIKit*, ki je namenjeno kvantitativnemu spremljanju pojavov italijansko-slovenskega jezikovnega stikanja na območju poselitve avtohtone

slovenske narodne skupnosti v Italiji. S tem namenom so v korpusu posebej označene lokalne rabe, ki jih skupnost dojema kot standardne. Z raziskovanjem teh rab lahko vidimo, v kolikšni meri so odstopanja od splošnega slovenskega standarda res posledica jezikovnega stika z italijanščino, v kolikšni meri pa jih je mogoče pripisati drugim okoliščinam, na primer pomanjkljivi izpostavljenosti nelokalnim rabam ali identitetnim izbiram govork in govorcev. Primerjava lokalno specifičnih in splošnih lektorskih popravkov v slovenščini pa bo omogočala spremljanje jezikovne kompetence tvorcev in tvork besedil.

Na podlagi tega bomo lažje načrtovali nova jezikovnoinformativna orodja, namenjena tej skupnosti: taka orodja naj ne bi bila normativna – njihov namen naj ne bi bilo predpisovanje ene same, kodificirane rabe, ampak podajanje informacij o kompleksnosti sporazumevalnih praks (npr. o variantnosti) v nekem okolju, ob upoštevanju celovite uporabniške izkušnje ciljnih skupin.

Hkrati pa bo lahko obdelava jezikovnih podatkov omogočala prilagajanje velikih jezikovnih modelov, temelječih na umetni inteligenci, potrebam manjših lokalnih skupnosti, jezikovnim manjšinam in drugim jezikovno ranljivim skupinam. To bo prav gotovo eden od izzivov naslednjega desetletja, saj so nova orodja zaenkrat oblikovana tako, da po eni strani omogočajo večjezičnost, po drugi pa vodijo v jezikovno uniformacijo.

Literatura

Arhar Holdt, Špela, Kosem, Iztok in Stritar Kučuk, Mojca, 2022: Metode in orodja za lažjo pripravo korpusov usvajanja jezika. Pirih Svetina, Nataša in Ferbežar, Ina (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik.* Obdobja 41. Ljubljana: Založba Univerze v Ljubljani. 23–30. DOI: https://doi.org/10.4312/Obdobja.41.23-30.

Arhar Holdt, Špela, Kosem, Iztok, Stritar Kučuk, Mojca, Krsnik, Luka, Jovan, Leon Noe, Bezgovšek, Luka in Popič, Damjan, 2023: *CJVT Svala (Kazalnik projekta Razvoj slovenščine v digitalnem okolju)*. V 1.1. https://orodja.cjvt.si/svala/. (Dostop 12. 12. 2023.)

Arhar Holdt, Špela, Lavrič, Polona, Roblek, Rebeka in Goli, Teja, 2022: *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar*. V 1.1 (12. avgust 2022). https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1589/Smernice-za-oznacevanje-korpusa-Solar_V1.1.pdf?sequence=3&isAllowed=y. (Dostop 12. 12. 2023.)

Chambers, Jack K., 2003: *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Oxford: Blackwell.

Davila, Bethany, 2016: The Inevitability of 'Standard' English: Discursive Constructions of Standard Language Ideologies. *Written Communication* 33/2. 127–148.

Fought, Carmen, 2006: *Language and ethnicity*. Cambridge: Cambridge University Press.

Grgič, Matejka in Popič, Damjan, 2023: Procesi jezikovnega separatizma pri čezmejnih jezikovnih manjšinah: prevzemanje, prilagajanje in prevajanje covidne terminologije med Slovenci in Slovenkami v Italiji. *Annales: anali za istrske in mediteranske študije. Series historia et sociologia* 33/1. 151–166.

Grgič, Matejka, 2017: Italijansko-slovenski jezikovni stik med ideologijo in pragmatiko. *Jezik in slovstvo* 62/1. 89–98.

Grgič, Matejka, 2018: Centralni urad za slovenski jezik in mreža storitev na obravnavanem geografskem območju: ocena trenutnega stanja, teoretski vidiki in organizacijski modeli. Janežič, Adriana in Jagodic, Devan (ur.): *Druga deželna konferenca o zaščiti slovenske jezikovne manjšine*. Trst: Regione autonoma Friuli Venezia Giulia, Consiglio regionale. 67–82.

Grgič, Matejka, 2022: Ne prvi, ne drugi, ne tuji, pa vendar naš, čeprav tuj: slovenski jezik v Italiji kot epistemološko vprašanje. Pirih Svetina, Nataša in Ferbežar, Ina (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik.* Obdobja 41. Ljubljana: Založba Univerze v Ljubljani. 109–116. DOI: https://doi.org/10.4312/Obdobja.41.109-116.

Jagodic, Devan, Kaučič Baša, Majda in Dapit, Roberto, 2017: Jezikovni položaj Slovencev v Italiji. Bogatec, Norina in Vidau, Zaira (ur.): *Skupnost v središču Evrope: Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja*. Trst: ZTT, SLORI. 66–88.

Leech, Geoffrey, 1998: Teaching and language corpora: A convergence. Wichmann, Anne idr. (ur.): *Teaching and language corpora*. London: Longmann. 1–23.

Meyerhoff, Miriam, 2008: Communities of Practice. Chambers, Jack K. idr. (ur.): *The Handbook of Language Variation and Change*. Oxford: Blackwell Publishing.

Mezgec, Maja, 2016: Linguistic landscape as a mirror: the case of the Slovene minority in Italy. *Razprave in gradivo: revija za narodnostna vprašanja* 77. 67–85.

Pertot, Susanna in Kosic, Marianna, 2014: *Jeziki in identitete v precepu: mišljenje, govor in predstave o identiteti pri treh generacijah maturantov šol s slovenskim učnim jezikom v Italiji*. Trst: Slovenski raziskovalni inštitut SLORI.

Popič, Damjan, 2014: Revising translation revision in Slovenia. Mikolič Južnič, Tamara idr. (ur.): *New horizons in translation research and education 2*. Joensuu: University of Eastern Finland. https://erepo.uef.fi/handle/123456789/14340.

Popič, Damjan, 2022: Digitalna podpora slovenskemu jeziku v stiku. Pirih Svetina, Nataša in Ferbežar, Ina (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik.* Obdobja 41. Ljubljana: Založba Univerze v Ljubljani. 281–290. DOI: https://doi.org/10.4312/Obdobja.41.281-290.

Stritar Kučuk, Mojca, 2012: *Korpusi usvajanja tujega jezika*. Ljubljana: Zveza društev Slavistično društvo Slovenije.

Stritar Kučuk, Mojca, 2022: KOST med korpusi usvajanja tujega jezika. Pirih Svetina, Nataša in Ferbežar, Ina (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik.* Obdobja 41. Ljubljana: Založba Univerze v Ljubljani. 323–334. DOI: https://doi.org/10.4312/Obdobja.41.323-334.

Volodina, Elena, Granstedt, Lena, Matsson, Arild idr., 2019: The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology* 6. 67–104.

Wirén, Mats, Matsson, Arild, Rosén, Dan in Volodina, Elena, 2019: SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. Skadina, Inguna in Eskevich, Maria (ur.): *Selected papers from the CLARIN Annual Conference 2018*, *Pisa, 8–10 October 2018*. Linköping: Linköping University Electronic Press. 227–239.