Saška Štumberger (ur.): ***Predpis in norma v jeziku***.
Ljubljana: Založba Univerze v Ljubljani, 2024
zbirka Obdobja, 43
ISBN 978-961-297-439-8, 978-961-297-442-8 (PDF)
ISSN: 1408-211X, e-ISSN: 2784-7152
© Univerza v Ljubljani, Filozofska fakulteta, 2024. Vse pravice pridržane.

PRIMERJAVA SISTEMOV ZA OZNAČEVANJE JEZIKOVNIH POPRAVKOV V ŠTIRIH SLOVENSKIH BESEDILNIH KORPUSIH

Špela Arhar Holdt
Filozofska fakulteta in Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana
spela.arharholdt@ff.uni-lj.si

Damjan Popič
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
damjan.popic@ff.uni-lj.si

Mojca Stritar Kučuk
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
mojca.stritarkucuk@ff.uni-lj.si

DOI:10.4312/Obdobja.43.11-20

Za slovenščino so na voljo oz. v procesu gradnje štirje besedilni korpusi, ki vsebujejo oznake jezikovnih popravkov: Šolar, KOST, Lektor in STIKit. Popravki v teh korpusih so označeni po različnih označevalnih sistemih, prilagojenih specifikam korpusnega gradiva. Prispevek analizira označevalne sisteme, identificira podobnosti in razlike v označevalnih kategorijah in opredeli možnosti za medsebojne preslikave oznak ter primerjalne analize korpusnega gradiva.

jezikovni popravki, KOST, Šolar, Lektor, STIKit

For Slovenian, four text corpora that contain linguistic error annotations are available or under construction: Šolar, KOST, Lektor, and STIKit. The errors and corrections in these corpora are labeled with different annotation systems, each adapted to the specific characteristics of the corpus material. This article analyses the systems, identifying similarities and differences in the annotation categories, and it explores possibilities for label-mapping and comparative analyses of the corpus material.

error annotation, KOST, Šolar, Lektor, STIKit

1 Uvod

Besedilni korpusi, ki vsebujejo oznake jezikovnih napak oz. popravkov, predstavljajo empirično podstat za različne teoretične in aplikativne jezikoslovne raziskave, mdr. s področja normativistike in jezikovne didaktike. V primerjavi z drugimi evropskimi jeziki ima slovenščina zgledno stanje glede tovrstnih jezikovnih virov; na voljo ali v procesu gradnje so štirje besedilni korpusi: 1) korpus besedil, ki so nastala v slovenskih osnovnih in srednjih šolah, Šolar, 2) korpus besedil govorcev,

ki se učijo slovenščino kot drugi ali tuji jezik, KOST, 3) korpus besedil z lektorskimi popravki Lektor in 4) korpus besedil, ki nastajajo v zamejskem stiku slovenščine in italijanščine, StikIT. Čeprav so našteti jezikovni viri vsebinsko in oblikovno podobni, nastajajo z različnimi nameni in v različnih (projektnih) okoliščinah. V zadnjih letih smo dobili priložnost za usklajevanje metodologije njihove priprave, poskrbeli smo za enotno strojno označevanje in korpusni format ter nova orodja, kot sta označevalnik CJVT Svala[1] in specializirani konkordančnik.[2]

Jezikovni popravki v naštetih korpusih so označeni po različnih označevalnih sistemih, razlike in podobnosti med sistemi pa še niso bile sistematično raziskane in popisane. Zato je cilj tega prispevka analizirati označevalne sisteme ter identificirati njihove podobnosti in razlike. S tem želimo vzpostaviti pogoje za primerjalne analize tako jezikovnih težav kot tudi jezikovne korekcije v različnih vrstah korpusnih besedil, obenem pa nam tovrstna primerjava omogoča tudi načrtovano in kontinuirano raziskovanje normativne podobe slovenskega jezika med različnimi skupinami uporabnikov – ter seveda jezikovnonačrtovalske odzive na dognanja.

V nadaljevanju najprej predstavimo vse štiri korpuse in osnovne značilnosti njihovih označevalnih sistemov, nato pa jih primerjamo po jezikovnih ravninah. Prispevek zaključimo s sklepom in priporočili za nadaljnje delo.

2 Slovenski korpusi z jezikovnimi popravki

2.1 Šolar

Razvojni korpus Šolar, ki obstaja in se nadgrajuje že dobro desetletje (Rozman idr. 2012; Kosem idr. 2016; Arhar Holdt, Kosem 2023), je trenutno dostopen v različici 3.0 (Arhar Holdt idr. 2022a). Vsebuje 5485 besedil (večinoma esejev, pa tudi praktičnosporazumevalnih besedil), ki so jih napisali učenci slovenskih osnovnih in srednjih šol. V korpusu je tudi 36.570 avtentičnih učiteljskih jezikovnih popravkov, s pomočjo katerih je mogoče opazovati podajanje povratne informacije v kontekstu razvoja pisnih zmožnosti.

Sistem označevanja jezikovnih težav za ta korpus je bil pripravljen na osnovi samih popravkov, s kategorizacijo po principu od spodaj navzgor. Označevalni sistem je hierarhičen in tronivojski. Povsem vrhnji nivo temelji na klasifikaciji, ki je bila preizkušena za poskusni korpus slovenščine kot tujega jezika PiKUST (Stritar 2012). Skupina raziskovalcev je učiteljske popravke nato združevala v manjše skupine glede na vrsto jezikovnega problema (Kosem idr. 2020), kategorije pa so bile na koncu celovito pregledane in hierarhično urejene v 180 različnih tipov (Goli 2018), ki jih opisujejo označevalne smernice (Arhar Holdt idr. 2022b).

2.2 KOST

Pisni korpus slovenščine kot drugega in tujega jezika KOST (prim. Stritar Kučuk 2023a) se razvija od leta 2019. Vključuje spise in eseje udeležencev lektoratov,

1 https://orodja.cjvt.si/svala/
2 https://kost.cjvt.si/, https://solar.cjvt.si/

tečajev in izpitov iz slovenščine na Filozofski fakulteti Univerze v Ljubljani. Zaradi naraščajoče rabe strojnih prevajalnikov in drugih spletnih pripomočkov pri pisanju domačih nalog se vedno bolj omejuje na besedila, nastala v bolj nadzorovanih izpitnih situacijah (npr. pri testih). Tvorci govorijo več kot 30 prvih jezikov, večina pa prihaja z južnoslovanskega govornega območja. KOST 2.0 vključuje več kot 8300 besedil oz. več kot 1,5 milijona pojavnic. V njem so bile med gradnjo korpusa na več kot 2000 besedilih ročno popravljene in označene jezikovne napake, ki jih pri pisanju v slovenščini delajo tvorci.

Taksonomija napak, uporabljena pri tem označevanju, je precej robustna in kombinira klasifikacijo po jezikoslovnih kategorijah s formalnimi oznakami po površinski strukturi (Granger 2003: 467). Temelji na že prej omenjeni klasifikaciji, ki je bila preizkušena za poskusni korpus slovenščine kot tujega jezika PiKUST (Stritar 2012), prilagojena za prvo verzijo korpusa Šolar in prilagojena tudi zahtevam označevalnega orodja Svala (Arhar Holdt idr. 2022c). Pri označevanju napak v korpusu KOST velja načelo minimalnega popravka (prim. Granger idr. 2022; Reznicek idr. 2012): izhodiščno besedilo je spremenjeno samo na mestih, na katerih je nerazumljivo oz. slovnično nesprejemljivo. Vsaka napaka dobi oznako glede na taksonomijo napak, dokumentirano v priročniku za označevanje (Stritar Kučuk 2023b).

2.3 Lektor

Korpus Lektor[3] (prim. Popič 2014) je korpus lektorskih popravkov avtorskih in prevodnih besedil, ki zajema milijon besed in okoli 30.258 kategoriziranih popravkov (od teh je 15.414 različnih). Nastal je komplementarno korpusu Šolar, saj želi jezikovne zadrege učencev sopostaviti jezikovnim zadregam odraslih pišočih, predvsem tistih, ki se s pisanjem oz. prevajanjem tudi poklicno ukvarjajo. Osnovni namen korpusa Lektor je torej preveriti, v kolikšni meri se jezikovne zadrege pri pisanju in prevajanju razlikujejo od besedil šolske populacije (in, posledično, ali se torej jezikovne težave »vlečejo« vse od šolskih let naprej), obenem pa tudi, ali so lektorski posegi normativno utemeljeni ali slogovno motivirani. Prevodoslovni vidik korpusa Lektor pa leži predvsem v analizi prevodne revizije kot globljega procesa ali zgolj izvajanja lekture na prevodno besedilo brez upoštevanja izvirnika.

Zaradi teh dveh vidikov je nabor oznak popravkov v Lektorju specifičen, saj poleg jezikovnih ravnin natančno opredeljuje tudi slogovne popravke in jih poskuša čim bolj razdelati. Obenem je dodana kategorija pragmatičnih popravkov, v kateri najdemo pomenske, prevodne, faktografske in podobne popravke. V korpusu Lektor je torej pet vrhnjih kategorij (slogovni, oblikoslovni, pravopisni, skladenjski in pragmatični popravki), skupno pa shema zajema 50 podkategorij. Oznake so bile razvite postopoma, s testnim označevanjem besedil in vzpostavljanjem označevalske sheme, tako da se je označevanje korpusa začelo s skupno deveto različico nabora kategorij, med samim označevanjem pa so bile uvedene še številne spremembe.

3 https://www.korpus-lektor.net/

2.4 STIKit

STIKit je projekt in nastajajoči korpus lektoriranih besedil s prostora jezikovnega stika med slovenščino in italijanščino (prim. Popič, Grgič 2023), zlasti na območju poselitve avtohtone slovenske narodne skupnosti v Italiji. Namen projekta in korpusa je z objektiviziranimi analizami jezikovne rabe na tem področju natančneje ugotoviti, 1) v kolikšni meri se slovenščina v Italiji odmika od slovenščine v Sloveniji in 2) s katerimi jezikovnimi sredstvi se divergentni razvoj slovenščine v Italiji (najpogosteje) vzpostavlja. Splošni namen projekta pa je priprava in vzpostavitev celostnega pregleda jezikovne rabe na območju stika med slovenščino in italijanščino, ki bo prinesel globlje razumevanje procesov na tem območju ter omogočil premišljene jezikovnonačrtovalske odzive in oblikovanje inovativnih jezikovnih orodij za podporo zamejski skupnosti.

Korpus STIKit je od vseh obravnavanih korpusov najnovejši in zaradi tega v veliki meri sloni na tehnologijah in metodologijah predhodnih projektov, zlasti Šolarja (orodje Svala in spletni konkordančnik). Zaradi inherentne želje po naboru oznak, ki bi bil čim primerljivejši z drugimi označevalnimi sistemi – in zaradi tega tudi pripravljen za računalniško procesiranje in primerjavo podatkov med različnimi podatkovnimi zbirkami –, je bil kot izhodiščni sistem uporabljen označevalni sistem korpusa Šolar (Arhar Holdt idr. 2022b). Pri tem so ohranjene zgolj vrhnje (splošnejše) kategorije, medtem ko je na nižji ravni trenutno omogočena le izbira med binarnima možnostma »stik : nestik«, tj. ali je bil določen jezikovni pojav, ki je bil med lekturo popravljen, nestandarden zaradi (domnevnega) vpliva jezikovnega stika med italijanščino in slovenščino ali ne. Trenutno poteka analiza prvih označenih besedil (ok. 50.000 besed), na podlagi katere bo shema ustrezno dopolnjena in popolnjena tudi na nižjih ravneh.

3 Podobnosti in razlike označevalnih sistemov

3.1 Zapis, črkovanje, pravopis

V korpusu KOST kategorija *zapis* vsebuje popravke ločil, črkovanja, zapisa besed skupaj ali narazen oz. z vezajem, velike oz. male začetnice in krajšav. Primerljive skupine najdemo v korpusu Šolar, ki težave črkovanja sicer obravnava kot samostojno skupino, pod *zapis* pa umešča popravke ločil, zapisa male oz. velike začetnice, pisanja skupaj oz. narazen in krajšav; dodatna skupina so popravki zapisa števil (menjave zapisa s številko/besedo ipd.). Naštete skupine se v Šolarju še dodatno členijo. V Lektorju se vrhnja kategorija imenuje *pravopis*. Popravki ločil se ločijo na stavo in zamenjavo, kategorijo zapis z malo/veliko začetnico nadomešča natančnejša členitev, ki je v veliki meri prekrivna s podkategorijami v sistemu korpusa Šolar. Ujemata se tudi kategoriji *skupaj/narazen* in *krajšave*, namesto kategorije *črkovanje* pa korpus Lektor loči popravke zapisa, (domnevne) tipkarske napake in spremembe izrazne oblike (npr. *6* → *šest* ali obratno). Primerjavo kategorij prikazuje Tabela 1.

KOST	**Šolar**		**Lektor**	**STIKit**
Zapis	**Zapis**	**Črkovanje**	**Pravopis**	**Zapis**
Ločilo	*Ločila*: 11 podkategorij		*Ločilo – stava* *Ločilo – zamenjava*	*Ločila*
Črkovanje		*Konzonanti*: 9 podkategorij *Vokali*: 6 podkategorij *Sklop*: 6 podkategorij *Variantni predlogi*: 2 podkategoriji *Zapis [u̯, w, ɯ]*: 4 podkategorije	*Zapis* *Tipkarska napaka* *Sprememba izrazne oblike*	
Skupaj/ narazen	*Skupaj/ narazen*: 8 podkategorij		*Skupaj/narazen* *Pretvorba besedne zveze v tvorjenko* *Pretvorba tvorjenke v besedno zvezo* *Sprememba zveze ali tvorjenke: priredno – podredno*	*Skupaj/ narazen*
Mala/velika začetnica	*Mala/velika začetnica*: 11 podkategorij, mdr.: občna, lastna, stvarna, zemljepisna, pridevnik na -ski		*Začetnica pri zapisu občnega imena* *Začetnica pri zapisu imen bitij* *Začetnica pri zapisu stvarnega lastnega imena* *Začetnica pri zapisu zemljepisnega lastnega imena* *Začetnica pri zapisu pridevnika*	*Mala/velika začetnica*
Krajšave	*Krajšave*		*Krajšave*	*Krajšave*
	Števila			*Števila*
			Drugo	

Tabela 1: Primerjava označevanja jezikovnih problemov na ravni zapisa, črkovanja oz. pravopisa.

3.2 Oblika

Popravki oblike so v korpusu KOST omejeni na končnice pregibnih besednih vrst, kar odraža tudi označevalni sistem, prikazan v Tabeli 2. V kategorijo *ostalo* sodi pregibanje glavnih številnikov in kratic. Korpus Šolar precej drugače oblikovnih

popravkov ne omeji na končnice, deli pa jih glede na kategorialne (pri katerih učenec uporabi v jeziku obstoječo obliko, vendar z neustreznimi kategorialnimi lastnostmi, npr. v napačnem sklonu, času) ter paradigmatske (pri katerih učenec uporabi obliko, ki v standardnih pregibnih paradigmah ne obstaja, npr. zaradi pregibanja po neustreznih analogijah). V sistemu so na voljo tudi neobvezne oznake za pogoste napake in popravke oblikovnih variant. Analiza je pokazala, da kategorije med korpusi niso enostavno povezljive, saj v Šolarju presegajo meje posamezne besedne vrste, prav tako pa se del Šolarjevih paradigmatskih popravkov v korpusu KOST pojavi pri oznakah besedišča (prim. razdelek 3.3). Vendar bi bilo mogoče vsaj del podatkov povezati tako, da se kategorialne popravke v korpusu Šolar razvrsti glede na besedno vrsto s pomočjo oblikoskladenjskih oznak, ki so pripisane v korpusnih besedilih. Medtem ko korpus STIKit v celoti sledi vrhnji strukturi korpusa Šolar, pa Lektor interpretira in združuje del poimenovanj (poimenovanja bitij, zemljepisna in stvarna lastna imena). Ostala podaja glede na besedno vrsto.

<table>
<tr><th>KOST</th><th>Šolar</th><th>Lektor</th><th>STIKit</th></tr>
<tr><th>Oblika</th><th>Oblika</th><th>Oblika</th><th>Oblika</th></tr>
<tr><td>Samostalnik</td><td rowspan="2">Kategorialni popravki:
20 podkategorij, mdr.: čas, določnost, način, naklon, spol, oseba
Paradigmatski popravki:
6 podkategorij, mdr.: glagolska osnova, glagolska končnica, neobstojni vokal
Dodatne oznake:
za popravke besed mati, hči, otrok in oblikovne variante</td><td>Pregibanje domačih imen bitij
Pregibanje tujih imen bitij
Pregibanje domačih zemljepisnih lastnih imen
Pregibanje tujih zemljepisnih lastnih imen
Pregibanje stvarnih lastnih imen/občnih besed</td><td rowspan="2">Kategorialni popravki
Paradigmatski popravki
(brez podkategorij)</td></tr>
<tr><td>Pridevnik</td><td>Pregibanje pridevnikov</td></tr>
<tr><td>Glagol</td><td></td><td>Pregibanje glagolov</td><td></td></tr>
<tr><td>Zaimek</td><td></td><td>Pregibanje zaimkov</td><td></td></tr>
<tr><td>Prislov</td><td></td><td>Pregibanje nepregibnih/ funkcijskih besed</td><td></td></tr>
<tr><td>Ostalo</td><td></td><td>Pregibanje/zapis številkov</td><td></td></tr>
</table>

Tabela 2: Primerjava označevanja jezikovnih problemov na ravni oblike.

3.3 Besedišče, slog, pragmatika

Popravki besedišča v korpusu KOST se nanašajo na neustrezen koren oz. osnovo besede in so razdeljeni v podkategorije na osnovi besedne vrste, kot kaže Tabela 3. Popravki zajemajo rabo besed z napačnim pomenom, besed, ki so zaznamovane glede na kontekst, ali neustrezno tvorjenih oz. v slovenščini neobstoječih besed. Tudi sistem korpusa Šolar ločuje po besednih vrstah, pri čemer so posebej obravnavani popravki prek besedne vrste (npr. menjave polnopomenske besede v zaimek ali obratno). Šolar za razliko od KOST-a ponuja ločeno, dodatno oznako za zaznamovano besedišče, neustrezno tvorjene oz. neobstoječe oblike pa se obravnavajo pri popravkih oblike

(prim. razdelek 3.2). Precej drugače je zasnovan sistem Lektorja, ki loči popravke sloga in pragmatike s primeri, ki se v Šolarju in KOST-u pojavljajo deloma na ravni besedišča in deloma skladnje. Analiza je pokazala, da bi se kljub velikim razlikam v sistemih na ravni besedišča vsaj določene skupine (npr. *tujka*, *prevzemanje*, *pomen*) dalo povezati s pomočjo informacij v oblikoslovnih oznakah, druge (npr. *izbris*, *dodajanje*) pa povezati na ravni skladnje, čeprav so v korpusu Lektor vsa dodajanja in izbrisi obravnavani kot izrazito slogovni popravki, če jih ni mogoče jezikovnosistemsko ali normativno razložiti.

<table>
<tr><th>KOST</th><th>Šolar</th><th colspan="2">Lektor</th><th>STIKit</th></tr>
<tr><th>Besedišče</th><th>Besedišče</th><th>Slog</th><th>Pragmatika</th><th>Besedišče</th></tr>
<tr><td>Samostalnik</td><td>Samostalnik:
3 podkategorije</td><td rowspan="10">Dvojnica/variantni zapis
Tujka
Kolokacija
Izbris
Dodajanje
Prevzemanje
Vezljivost/vezava
Besednovrstna pretvorba
Koreferenca
Drugo</td><td rowspan="10">Prevajalska napaka
Pomen
Faktografija
Komentar</td><td>Samostalnik</td></tr>
<tr><td>Glagol</td><td>Glagol:
4 podkategorije</td><td>Glagol</td></tr>
<tr><td>Zaimek</td><td>Zaimek:
5 podkategorij</td><td>Zaimek</td></tr>
<tr><td>Pridevnik</td><td>Pridevnik</td><td>Pridevnik</td></tr>
<tr><td>Prislov</td><td>Prislov</td><td>Prislov</td></tr>
<tr><td>Predlog</td><td>Predlog:
4 podkategorije</td><td>Predlog</td></tr>
<tr><td>Veznik</td><td>Veznik:
4 podkategorije</td><td>Veznik</td></tr>
<tr><td></td><td>Menjave prek besedne vrste:
8 podkategorij</td><td>Menjave prek besedne vrste</td></tr>
<tr><td></td><td>Dodatne oznake:
za zaznamovano besedišče</td><td>Dodatne oznake</td></tr>
<tr><td>Ostalo</td><td>Ostalo</td><td>Ostalo</td></tr>
</table>

Tabela 3: Primerjava označevanja jezikovnih problemov na ravni besedišča, sloga oz. pragmatike.

3.4 Skladnja

Tabela 4 primerja označevalne sisteme na ravni popravkov skladnje. Korpusa KOST in Šolar prinašata skladne kategorije, ki so v Šolarju še dodatno členjene, npr. glede na to, katere vrste jezikovni element je bil pri popravljanju dodan, izpuščen ali besednoredno premeščen. Napake strukture presegajo raven ene same besede in zajemajo širok spekter problemov, v katerih tvorec neustrezno tvori oz. uporabi besedno zvezo, stavek ali poved. Korpus Šolar prinaša tudi dodatne, neobvezne oznake za popravke vsebinskih napak in pomensko praznih struktur, ki so deloma, ne pa v celoti primerljive s popravki, ki jih Lektor umešča na raven pragmatike; izjema

so seveda oznake popravkov, vezanih na medjezikovni in medkulturni prenos (npr. prevajalske napake).

<table>
<tr><th>KOST</th><th>Šolar</th><th>Lektor</th><th>STIKit</th></tr>
<tr><th>Skladnja</th><th>Skladnja</th><th>Skladnja</th><th>Skladnja</th></tr>
<tr><td>Besedni red</td><td>Besedni red:
8 podkategorij</td><td>Besedni red</td><td>Besedni red</td></tr>
<tr><td>Struktura</td><td>Struktura:
7 podkategorij</td><td rowspan="3">Razvezava stavkov
Združitev stavkov
Zamenjava veznika
Pretvorba skladenjskega razmerja
Pretvorba neosebne/brezosebne oblike v tvorno obliko
Pretvorba v neosebno/brezosebno obliko
Vezava
Stavčno ujemanje/ujemanje naslonskih oblik
Predlog
Drugo</td><td>Struktura</td></tr>
<tr><td>Izpuščeni jezikovni elementi</td><td>Izpust:
14 podkategorij</td><td>Izpuščeni jezikovni elementi</td></tr>
<tr><td>Odvečni jezikovni elementi</td><td>Odveč:
21 podkategorij</td><td>Odvečni jezikovni elementi</td></tr>
<tr><td></td><td>Dodatne oznake:
za odstranjene pleonastične in pomensko prazne strukture, za popravke vsebinskih napak</td><td></td><td></td></tr>
</table>

Tabela 4: Primerjava označevanja jezikovnih problemov na ravni skladnje.

3.5 Povezani popravki

V vseh označevalnih sistemih se pojavlja oznaka za povezani popravek, tj. spremembo v besedilu, ki nastane kot posledica izhodiščnega jezikovnega popravka. Če se npr. po popravku napačno izbranega predloga spremeni tudi sklon samostalniške besedne zveze, se slednje označi kot povezani popravek. Razlika je v tem, da se v korpusu Šolar ta oznaka pripisuje samostojno, v ostalih pa se pojavlja le kot dodatna oznaka, kar je treba upoštevati pri primerjanju frekvenčnih podatkov iz različnih korpusov.

4 Sklep in nadaljnje delo

V prispevku smo primerjali označevalne sisteme za jezikovne popravke v štirih slovenskih korpusih. Rezultati analiz kažejo delno povezljivost med vrhnjimi kategorijami, nekoliko več med korpusoma Šolar in KOST oz. Šolar in STIKit, medtem ko se korpus Lektor s Šolarjem povezuje tudi prek nekaterih podkategorij. Če bi želeli primerjati podatke med korpusi, bi bilo potrebno dodatno delo, na eni strani z vključevanjem informacij iz oblikoskladenjskih oznak, na drugi s prerazvrščanjem označevalnih podkategorij. Ker preslikave niso povsem enoznačne, se zdijo

metodološko smiselne predvsem kvalitativne analize, pri katerih je mogoče natančneje preučiti označevalne smernice in označeno korpusno gradivo. Kvantitativne primerjave med korpusi so sicer možne (zlasti pri pogostnosti vrhnjih kategorij), vendar zahtevajo precejšnjo pozornost, tako zaradi razlik med kategorijami kot tudi (ne)doslednosti jezikovnih popravkov in različnih praks pri pripisovanju dvojnih oznak. Predstavljeni prispevek lahko služi kot izhodišče za ustrezno načrtovanje primerjalnih raziskav.

Čeprav je ključno skrbeti za standardizacijo in prenos metodologije korpusne gradnje, se celostno enotenje označevalnih sistemov ne kaže kot dobra rešitev. Možne bi bile manjše uskladitve, vendar so razlike, ki jih je pokazala analiza v prispevku, večinoma utemeljene z namenom korpusne gradnje in specifikami korpusnega gradiva. Tako je za nadaljnje delo ključna zlasti skrb za skladnost korpusnega formata, ki omogoča, da za vse korpuse uporabljamo enaka orodja (npr. za strojno označevanje, luščenje podatkov, iskanje in vizualizacijo v konkordančnikih). Nekoliko drugačen scenarij uporabe korpusnih podatkov pa je za namene ocenjevanja strojnega popravljanja jezikovnih napak, ki zahteva usklajene in primerljivo označene podatkovne množice. Iz korpusa Šolar je že bila pripravljena množica Šolar-Eval (Gantar idr. 2023), ki lahko služi kot model za nadaljnje delo.

Literatura

ARHAR HOLDT, Špela, ROZMAN, Tadeja, STRITAR KUČUK, Mojca, KREK, Simon, KRAPŠ VODOPIVEC, Irena, STABEJ, Marko, PORI, Eva, GOLI, Teja, LAVRIČ, Polona, LASKOWSKI, Cyprian, KOCJANČIČ, Polonca, KLEMENC, Bojan, KRSNIK, Luka, KOSEM, Iztok, 2022a: *Developmental Corpus Šolar 3.0*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1589.

ARHAR HOLDT, Špela, LAVRIČ, Polona, ROBLEK, Rebeka, GOLI, Teja, 2022b: *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar*. Rezultat projekta Razvoj slovenščine v digitalnem okolju. Različica 1.1. https://wiki.cjvt.si/books/11-jezikovni-popravki-solar/page/oznacevalne-smernice.

ARHAR HOLDT, Špela, KOSEM, Iztok, STRITAR KUČUK, Mojca, 2022c: Metode in orodja za lažjo pripravo korpusov usvajanja jezika. Nataša Pirih Svetina, Ina Ferbežar (ur): *Na stičišču svetov: Slovenščina kot drugi in tuji jezik. Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani. 23–30.

ARHAR HOLDT, Špela, KOSEM, Iztok, 2023: *Šolar, the Developmental Corpus of Slovene*. Preprint. https://doi.org/10.21203/rs.3.rs-3274669/v1.

GANTAR, Polona, BON, Mija, GAPSA, Magdalena, ARHAR HOLDT, Špela, 2023: Šolar-Eval: evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in slovstvo* LXVIII/4. 89–108.

GOLI, Teja, 2018. *Na napakah se učimo: kategorizacija učiteljskih popravkov v korpusu Šolar 2.0*. Magistrsko delo. Ljubljana: Filozofska fakulteta.

GRANGER, Sylviane, 2013. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* XX/3. 465–480. https://doi.org/10.1558/cj.v20i3.465-480.

GRANGER, Sylviane, SWALLOW, Helen, THEWISSEN, Jennifer, 2022: *The Louvain Error Tagging Manual Version 2.0*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain. https://cdn.uclouvain.be/groups/cms-editors-cecl/Granger%20et%20al._Error%20tagging%20manual_v2.0_2022.pdf.

KOSEM, Iztok, STRITAR KUČUK, Mojca, MOŽE, Sara, ZWITTER VITEZ, Ana, ARHAR HOLDT, Špela, ROZMAN, Tadeja, 2020: *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Znanstvena založba Filozofske fakultete. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/229/329/5311-1.

POPIČ, Damjan, 2014: Revising Translation Revision in Slovenia. Tamara Mikolič Južnič, Kaisa Koskinen, Nike Kocijančič Pokorn (ur.): *New Horizons in Translation Research and Education 2*. Joensuu: University of Eastern Finland. 72–89. http://epublications.uef.fi/pub/urn_isbn_978-952-61-1657-0/urn_isbn_978-952-61-1657-0.pdf.

POPIČ, Damjan, GRGIČ, Matejka, 2023: Korpus lektoriranih slovenskih besedil z območja italijansko-slovenskega jezikovnega stikanja – STIKit: Zasnova, gradnja in izzivi. *Jezik in slovstvo* LXVIII/4. 73–87. https://doi.org/10.4312/jis.68.4.73-87.

REZNICEK, Marc, LÜDELING, Anke, KRUMMES, Cedric, 2012: *Das FALKO-Handbuch: Korpus Aufbau und Annotationen, Version 2.01*. Berlin: Humboldt-Universität zu Berlin.

ROZMAN, Tadeja, STRITAR, Mojca, KOSEM, Iztok, 2012: Šolar – korpus šolskih pisnih izdelkov. Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar, Iztok Kosem (ur.): *Empirični pogled na pouk slovenskega jezika.* Ljubljana: Trojina, zavod za uporabno slovenistiko. https://doi.org/10.4312/9789610603511.

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

STRITAR KUČUK, Mojca, 2023a: Prvi korpus slovenščine kot tujega jezika KOST 1.0. Špela Arhar Holdt, Simon Krek (ur.): *Razvoj slovenščine v digitalnem okolju.* Ljubljana: Založba Univerze v Ljubljani. 93–117.

STRITAR KUČUK, Mojca, 2023b: *Priročnik za označevanje napak*. https://www.cjvt.si/korpus-kost/wp-content/uploads/sites/24/2023/10/Prirocnik-za-oznacevanje-napak-v-KOST-u-2023-10-17.pdf (dostop 26. 10. 2023).

Projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159) in program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Projekt izdelave korpusa STIKit finančno podpira Centralni urad za slovenski jezik pri Avtonomni pokrajini.