REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Korpus Janes-Norm 3.0

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.4

Avtorja: Jakob Lenardič[◊], Darja Fišer[◊]

[◊] Inštitut za novejšo zgodovino

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Jakob Lenardič in Darja Fišer: Projektni okvir

Poročilo oz. kazalnik *Korpus Janes-Norm 3.0* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirala Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Med besedilne korpuse, ki smo se jim na projektu posebej posvetili, sodi tudi specializiran korpus, ki ga opisujemo v tem kazalniku: učni korpus spletne komunikacije Janes-Norm 3.0 (Lenardič et al. 2022). Korpus omogoča empirično podprte jezikoslovne analize nestandardne spletne slovenščine ter izboljšanje jezikovnotehnoloških orodij za obdelavo besedil, napisanih v nestandardnem jeziku.

2 Jakob Lenardič in Darja Fišer: Opis projektnih aktivnosti

Aktivnost A1.4 se posveča dvema temeljnima besedilnima korpusoma pisne slovenščine: Gigafida besedilno reprezentira standardni jezik, Janes besedilno reprezentira uporabniško generirane spletne vsebine, mdr. v nestandardni slovenščini. Korpusa predstavljata osnovo za jezikovni opis, predpis, priročnike, jezikovne tehnologije in postopke vseh vrst. Poskrbeti je treba za dolgoročno osveževanje obeh korpusov, pri čemer je treba v načrtih upoštevati izkušnje deležnikov, ki korpusa uporabljajo za razvoj izdelkov. Temeljni premisleki so potrebni na ravni gradivne zastopanosti, opredeljevanja standardnosti besedil in s tem povezane vključitve slovenskega zamejstva. Vzporedno z načrti je treba

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

zagotoviti infrastrukturo, ki bo omogočala kontinuirano nadgrajevanje korpusov. Na eni strani to pomeni urejanje pravnih vprašanj glede pridobivanja besedil, vzpostavitev spletne platforme z informacijami za besedilodajalce, organizacijo mreže besedilodajalcev in opredelitev delotokov zbiranja. Na drugi strani je treba vzpostaviti cevovod za označevanje pridobljenih besedil, kar pokriva druga projektna aktivnost (A1.1 Razvoj označevalnega orodja CLASSLA-Stanza[1]). Ker ima označevanje nestandardnega jezika določene specifike, npr. vsebuje dodatni korak normalizacije nestandardnih jezikovnih prvin v njihove standardne ustreznice, je treba v projektu zagotoviti tudi učno množico za normalizacijo, ki bo osnova za izboljšavo tovrstnega avtomatskega označevanja.

Pričujoči kazalnik opisuje aktivnosti, povezane z razširitvijo korpusa uporabniških spletnih vsebin Janes. V kazalniku opisujemo vse 3 podaktivnosti, načrtovane v projektni prijavi:

- Načrtovanje uravnoteženosti korpusa.
- Ciljni dogovori z izbranimi gradivodajalci za zagotavljanje uravnoteženosti korpusa uporabniško generiranih vsebin.
- Anotacijska kampanja za normalizacijo nestandardnih jezikovnih prvin v uporabniško generiranih spletnih vsebinah.

Točka 2 je deloma postala nerelevantna, saj razširjeni korpus, zbran in označen v okviru RSDO, vsebuje zgolj eno vrsto besedil – slovenske tvite. To odločitev upravičujemo v razdelku 4.

REZULTAT AKTIVNOSTI

- Korpus Janes-Norm 3.0 (327.000 pojavnic) kot baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1733, objavljen v prostem dostopu pod licenco CC BY-NC-SA 4.0. Vsebuje ročne oznake na nivoju stavčne segmentacije, tokenizacije in normalizacije, ki so bile pripisane v označevalni kampanji, opisani v razdelku 5. Janes-Norm 3.0 vključuje poleg obstoječih virov Janes-Norm 1.2[2] (Erjavec et al. 2016) in Janes-Tag 2.1[3] (Erjavec et al. 2019) še

[1] https://github.com/clarinsi/classla
[2] http://hdl.handle.net/11356/1084
[3] http://hdl.handle.net/11356/1238

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Janes-RSDO, ki je bil razvit v sklopu tega delovnega sklopa. Kazalnik torej opisuje zasnovo slednje podmnožice v Janes-Norm 3.0.

3 Jakob Lenardič in Darja Fišer: Načrtovanje uravnoteženosti korpusa

Cilj aktivnosti je bilo načrtovanje razširitve obstoječega korpusa Janes-Norm 1.2 (Erjavec et al. 2016) z novim naborom besedil. Obstoječi korpus Janes-Norm 1.2, ki vsebuje ročne oznake na nivoju tokenizacije, stavčne segmentacije in normalizacije, obsega 7816 besedil oziroma 184.755 pojavnic, vzorčenih iz korpusa JANES (Fišer et al. 2020), ki žanrsko obsega več vrst besedil slovenske spletne komunikacije: tviti, forumske objave, komentarji na spletne novice, objave na blogih ter slovensko Wikipedijo (ibid: 232). V Janes-Norm 1.2 je polovica besedil jezikovno nestandardnih in polovica standardnih (Čibej et al. 2016a: 41).

Razširjeni del korpusa, tj. Janes-RSDO, ki je bil načrtovan v okviru tega delovnega sklopa, se žanrsko osredotoča zgolj na slovenske tvite (gl. tudi naslednji razdelek), ki jih je skupaj 11.960, kar ustreza 147.020 pojavnicam oziroma 125.000 besedam. Vsi tviti so iz leta 2020. Pri vzorčenju smo zopet upoštevali uravnoteženost standardnega in nestandardnega dela, pri čemer ima polovica vzorčenih tvitov izračunano vrednost nestandardnosti večjo kot 0,4 (tj. več kot 40 % jezikovno nestandardnih pojavnic v posameznem tvitu), polovica pa vrednost, ki se giblje med 0,2 ter 0,4 (tj. med 20 % in 39 % nestandardnih pojavnic v posameznem tvitu); gl. Ljubešić et al. (2015) za opis metode izračuna. Zgled tvita iz prvega – tj. bolj nestandardnega – nabora je npr. *@JureMakovec no sej to sm mislu :)*, v katerem so tri pojavnice od sedmih (tj. *sej*, *sm* in *mislu*) jezikovno nestandardne. Pojem nestandardnosti je v tem kontekstu opredeljen operativno: pojavnica je nestandardna, če jo je treba normalizirati (gl. razdelek 5). Zgled tvita iz drugega – tj. bolj standardnega – nabora je *@aleksandertusek no, sam une 3 prste bi jim počopala. #navijač*, v katerem sta samo 2 pojavnici (*sam* in *une*) od 12 nestandardni.

S tovrstnim vzorčenjem smo zagotovili, da se besedila v razširjenem delu korpusa ne osredotočajo zgolj na nestandardno komunikacijo, ampak da so reprezentativna za komunikacijo na slovenskem Tviterju na splošno.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4 Jakob Lenardič in Darja Fišer: Pridobivanje raznovrstnega uporabniško generiranega gradiva

Janes-RSDO vsebuje zgolj tvite. Razloga za to sta dva. Prvi je tehnične narave. Projektna ekipa je razvila namensko orodje TweetCat[4] (Ljubešić et al. 2014), s katerim so tehnični sodelavci zajeli tvite iz leta 2020. Orodje uporablja prostodostopen Twitter Search API[5], da najde uporabnike, ki tvitajo v ciljnem jeziku, tj. slovenščini. Pri pripravi besedil je bil izveden korak filtriranja uporabnikov. To poteka tako, da se s Pythonovim modulom *langid.py* v TweetCaT identificira jezik vsakemu zajetemu tvitu posameznega tviteraša in odstrani tiste uporabnike, pri katerih večinski jezik ni slovenščina. Kot že opisujejo Erjavec et al. (2018: 20), je to filtriranje potrebno, »da bi res zajeli čim več slovenskih in čim manj tujejezičnih tviterašev ob zavedanju, da je identifikacija jezika težak problem, toliko bolj za besedila na Twitterju, ki so zelo kratka, pogosto niso napisana v standardnem jeziku in lahko vsebujejo veliko tujejezičnih prvin«. Ključna prednost orodja TweetCaT je tudi v tem, da tvite zbira avtomatsko in kontinuirano, tako da za nov zajem tvitov (v nasprotju z drugimi platformami) ni potrebno vzpostaviti novega vzvoda.

Drugi razlog je pravne narave. Twitter, glede na pogoje uporabe za razvijalce[6], dovoli, da posamezen razvijalec oziroma raziskovalec v prostem dostopu objavi do 50.000 tvitov, medtem ko imajo druga družbena omrežja bistveno restriktivnejše pogoje uporabe – npr. na Facebooku, ki je sicer v Sloveniji najbolj razširjeno družbeno omrežje, »prevladuje zasebna komunikacija, za katero ponudnik izrecno prepoveduje zbiranje in distribucijo vsebin« (Erjavec et al. 2018: 19).

[4] https://github.com/clarinsi/tweetcat
[5] https://dev.twitter.com/rest/public/search
[6] https://developer.twitter.com/en/developer-terms/policy

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

5 Jakob Lenardič in Darja Fišer: Anotacijska kampanja za normalizacijo nestandardnih jezikovnih prvin

5.1 Delotok označevalne kampanje

Ročno označevanje je potekalo v okolju *Google Preglednice*. Podatki iz 11.960 tvitov oz. 147.000 pojavnic so bili pripravljeni v 64 datotekah, pri čemer je vsaka datoteka vsebovala okrog 2300 pojavnic, saj je bilo ocenjeno, da takšno število pojavnic posamezen označevalec povprečno pregleda v enem dnevu na dveh nivojih označevanja – tokenizaciji in normalizaciji. Podatke so pregledovali 4 študenti jezikoslovnih smeri (slovenistika in prevajalstvo), pri čemer sta dva študenta pregledala isto datoteko, kar je zagotavljalo večjo zanesljivost odločitve. Podatki so bili pripravljeni v tabelarični obliki, pri čemer je prvi stolpec vseboval avtomatsko razdeljene pojavnice, v preostala stolpca so pa označevalci vnašali popravke pri tokenizaciji in normalizaciji (Tabela 1).

Tabela 1: Ročni popravki tokenizacije in normalizacije; sosledje simbolov $0 označuje odvečno vrstico pri popravkih napačno razdruženih besed, znak | pa razdružitev dveh nestandardno zapisanih besed (npr.., *nauta* v *ne bosta*).

token_avtomatsko	**token_popravek**	**normalizacija**
zato		
tukó		tako
nauta		ne|bosta
s	s’m	sem
‘	$0	$0
m	$0	$0
pršva		prišla

Vsak označevalec je pregledal 5 datotek v enem tednu, tako da je označevanje trajalo 13 tednov med julijem in septembrom 2021. Končani kampanji je sledila kuracija. Posamezne odločitve označevalcev

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

pri kuraciji so bile zbrane v enotni datoteki. Naloga dveh kuratorjev (Jakob Lenardič in najbolj natančna označevalka) je bil pregled razhajanja pri tokenizaciji in normalizaciji ter pripis končne rešitve.

5.2 Označevalne smernice

Izhajali smo iz obstoječih označevalnih smernic za segmentacijo, tokenizacijo in normalizacijo nestandardne slovenščine (Čibej et al. 2016b: 2–7), ki so nastale v okviru projekta JANES (Fišer et al. 2020). Pred pričetkom kampanje so bile smernice označevalcem predstavljene na spletnem seminarju. Pripravljena je bila tudi nova različica smernic, ki je posodobljena na podlagi novo nastalih dilem pri tokenizaciji/segmentaciji (razdelek 4c) in normalizaciji (razdelek 4d) novega nabora.

5.3 Tokenizacija in stavčna segmentacija

Večina tokenizacijskih napak je bila enakih kot v kampanji, v kateri se je označevala prejšnja različica Janes-Norm 1.2; gl. Čibej et al. (2016a: razdelek 4.2) za povzetek napak. Tokenizator je tako velikokrat avtomatsko napačno ločil okrajšave (npr. niz *slov.* kot dve pojavnici “slo” + “.”) ter nize neenakih, v tvitu stičnih, emotikonov, ki smo jih združevali v eno pojavnico.

Nove tokenizacijske dileme so se pojavile pri izrazito nestandardnih tvitih, zlasti tistih, ki so vsebovali zapise fonoloških prvin v narečjih. Narečne besede, ki so vsebovale apostrof za označitev izpusta glasu, kot npr. *s’m* “sem”, *tolk’* “toliko”, *b*’ “bi”, so bile avtomatsko razdružene, tako da je ročni popravek vključeval združevanje (gl. Tabelo 1), medtem ko so bili narečni zapisi besed na koncu tvita pogosto avtomatsko združeni v eno pojevnico s piko (npr. gorenjski *vidu.*).

Nove tokenizacijske dileme so se pojavile tudi pri sledečih kategorijah:

- Tokenizacija inkluzivnega slovničnega spola, pri katerih je obrazilo ločeno z nekončnim ločilom, npr. *sam(a)*, *avtor/ica*, *s(m)o*, *pisatelj_ica*; tovrstni zgledi so bili zaradi skladnosti s krovnim korpus SUK ročno popravljeni tako, da so bili elementi razdruženi, npr. *sam + (+ a +)*.
- Raba vezaja oziroma pomišljaja pri ločevanju nebesednih elementov, npr. *o-ga-bno*, *med.over.net-ovski*, *@JureBrankovic-u*; tovrstni elementi so bili tudi popravljeni z razdruževanjem.
- Nestično zapisani emotikoni, npr. :); tovrstni elementi so bili združeni v eno pojavnico.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Pri ročnem označevanju stavčne (oz. natančneje povedne) segmentacije smo tudi izhajali iz prejšnjih smernic (Čibej et al. 2016b: razdelek 2), pri čemer novih dilem ni bilo. Posebno je bilo treba opozoriti na 2 sorodni vrsti težavnih primerov, ki zadevajo imena (@ime), emotikone/emojije ter heštege (#hešteg). Prvič, če se ti elementi pojavljajo sredi stavka, sami po sebi ne predstavljajo konec stavčnega segmenta. Na primer, slednja poved tako ustreza zgolj enemu segmentu: *neka baka :) uleti pa praša če loh gre kr naprej*. Če pa tovrstni elementi sledijo končnemu ločilu, pa predstavljajo svoj segment, tako da slednje sosledje povedi predstavlja 3 segmente; Segment 1: *Sonce, sneg in pot pod noge!* Segment 2: *:)* Segment 3: *Gremo v hribe!*

5.4 Normalizacija

Tako kot v prejšnji označevalski kampanji (gl. Čibej et al. 2016a: razdelek 4.3) smo pri normalizaciji izhajali iz načela minimalne intervencije, kar pomeni, da besedam nismo pripisovali standardnih sopomenk (npr. *pofarbat* je normalizirano v *pofarbati* in ne **pobarvati*), poleg tega pa smo ohranjali nestandardna skladenjska razmerja (*nism jo vidu* → *nisem jo videl*) in nismo popravljali slogovnih oziroma registrskih napak (npr. *rabiti* ni bil popravljen v *potrebovati* v relevantnih kontekstih).

Nove normalizacijske dileme, na podlagi katerih smo posodobili smernice[7], so se pojavili zlasti pri variantnosti zapisa. V obstoječih smernicah obstaja načelo, da se normalizacijska oblika pri nestandardnih besedah z več različicami zapisa (npr. *fovš, fouš, fauš*) izbere na podlagi oblike, ki je najpogostejša v krovnem korpusu JANES (Fišer et al. 2020), vendar je bilo to pravilo pri nekaterih primerih v konfliktu z neodvisnim pravilom, ki pravi, da "pregibane tujejezične besede, ki ohranjajo elemente izvornega zapisa [...], normaliziramo v najpogostejšo med različicami s tujimi prvinami zapisa v korpusa JANES (*sharati* → *sherati* in ne **šerati*)" (Čibej et al. 2016b: 7). Problem se je pojavil pri zgledih, kot so *Twiteraši*, *shandlati*, *twitnil*, saj ohranjajo tujejezični zapis zgolj v eni črki, sicer so pa normativno in oblikoskladenjsko povsem poslovenjeni; pri tovrstnih zgledih smo sledili splošnemu pravilu o najpogostejši obliki ne glede na tujejezično prvino, tako da se je npr. *twitnil* normaliziral v *tvitnil*, ki je pogostejša oblika kot *tweetnil*, s čimer so popravki ostali minimalni.

[7] https://wiki.cjvt.si/books/normalizacija/page/oznacevalne-smernice

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Najpogostejši zapis je bil tudi nekajkrat v konfliktu z referenčnimi viri. Na primer, *Slovar slovenskega knjižnega jezika* nudi iztočnico zgolj za slengovsko zaznamovani zapis *foter*, medtem ko je v rabi varianta *fotr* bistveno pogostejša; v izogib pravopisnemu preskriptivizmu smo pri normalizaciji upoštevali najpogostejšo obliko v dejanski rabi, tj. *fotr*.

Zaradi načela minimalne intervencije smo dodali novo pravilo, ki pravi, da naj se nestandardna obrazila, ki tvorijo novo besedno vrsto iz besedotvorne podstave, ne normalizirajo v standardna; npr. *včeraj je snegovalo*, četudi so takšna obrazila izjemno redka v rabi. Podobno smo dodali pravilo za normalizacijo nestandardnih vidskih predpon, tako da pri normalizaciji nismo odstranili morfemov, kot je *z-* v *zinštalirati*, čeprav so redki tako v standardni (korpus Gigafida 2.0) kot tudi nestandardni (korpus JANES) rabi.

Nazadnje smo tudi dopolnili seznam narečnih besed, pri katerih se besedišče lahko izjemoma spreminja. To so bili predvsem funkcijski ter zaimenski izrazi iz prekmurščine (ki jih v prejšnji različici korpusa ni bilo), kot je na primer *ge*, ki glede na kontekst lahko ustreza zaimkom *jaz*, *kje* ali *kjer*.

6 Prioritete za nadaljnji razvoj

Korpus Janes je zbirka javno objavljenih uporabniških spletnih vsebin, in sicer tvitov, forumskih sporočil, blogov, komentarjev na novice in pogovornih strani na slovenski Wikipediji. Podobno kot velja za razvoj Gigafide, je tudi za Janes potrebno zagotoviti:

- **Stalen delotok zbiranja relevantnih spletnih vsebin**

  Trenutno razvite delotoke in orodja za zbiranje, čiščenje, normalizacijo in jezikoslovno označevanje je potrebno vzdrževati in razvijati. Nove verzije korpusa Janes je potrebno pripraviti na vsaka tri leta.
- **Periodične nadgradnje podkorpusa Twitter**

  Potrebno je vzdrževati in dopolnjevati orodje TweetCaT, ki kontinuirano zbira tvite slovenskih uporabnikov. S tem orodjem je potrebno sprotno zajemanje objav in letno ali dvoletno dopolnjevati korpus Janes.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- **Vključitev še nepokritih družbenih omrežij**

  Zgradi je potrebno orodje za zajem drugih, doslej nepokritih družbenih platform, ki že imajo (npr. Facebook) ali šele pridobivajo zadostno število slovenskih uporabnikov (npr. Mastodon). Javna ali le raziskovalna dostopnost novih korpusov bo odvisna od značaja in licenčni pogojev posameznih platform.

7 Literatura

Čibej et al. 2016a = Jaka Čibej – Špela Arhar Holdt – Tomaž Erjavec – Darja Fišer. Razvoj učne množice za izboljšano označevanje spletnih besedil. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, 40–46.

Čibej et al. 2016b = Jaka Čibej – Špela Arhar Holdt – Tomaž Erjavec – Darja Fišer – Katja Zupan. Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, lematizacija in oblikoskladenjsko označevanje v1.0. Na spletu (2016). https://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-smernice-v1.0.pdf.

Erjavec et al. 2016 = Tomaž Erjavec – Darja Fišer – Jaka Čibej – Špela Arhar Holdt. CMC training corpus Janes-Norm 1.2. V: *Slovenian language resource repository CLARIN.SI* (2016). http://hdl.handle.net/11356/1084.

Erjavec et al. 2018 = Tomaž Erjavec – Nikola Ljubešić – Darja Fišer. Korpus slovenskih spletnih uporabniških vsebin Janes. V: *Viri, orodja in metode za analizo spletne slovenščine* (2018), 16–43.

Erjavec et al. 2019 = Tomaž Erjavec – Darja Fišer – Jaka Čibej – Špela Arhar Holdt – Nikola Ljubešić – Katja Zupan – Kaja Dobrovoljc. CMC training corpus Janes-Tag 2.1. V: *Slovenian language resource repository CLARIN.SI* (2019). http://hdl.handle.net/11356/1238.

Fišer et al. 2020 = Darja Fišer – Nikola Ljubešić – Tomaž Erjavec. The Janes project: language resources and tools for Slovene user generated content. V: *Language Resources & Evaluation* 54 (2020), 223–246. https://doi.org/10.1007/s10579-018-9425-z.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Lenardič et al. 2022 = Jakob Lenardič – Jaka Čibej – Špela Arhar Holdt – Tomaž Erjavec – Darja Fišer. CMC training corpus Janes-Norm 3.0. V: *Slovenian language resource repository CLARIN.SI* (2022). http://hdl.handle.net/11356/1733.

Ljubešić et al. 2014 = Nikola Ljubešić – Darja Fišer – Tomaž Erjavec. TweetCaT: a tool for building Twitter corpora of smaller languages. V: *Proceedings of LREC 2014*, 2279–2283.

Ljubešić et al. 2015 = Nikola Ljubešić – Darja Fišer – Tomaž Erjavec – Jaka Čibej – Dafne Marko – Senja Pollak – Iza Škrjanec. Predicting the level of text standardness in user-generated content. V: *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*, 371–378.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.