

NA POTI K SKLADENJSKIM ANALIZAM ŠOLSKEGA PISANJA: SKLADENJSKI VZORCI V KORPUSU ŠOLAR 3.0

Tina MUNDA,¹ Špela ARHAR HOLDT^{1, 2}

¹Filozofska fakulteta, Univerza v Ljubljani

²Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

1 UVOD

Razvojni korpus Šolar vsebuje besedila, ki so jih učenke in učenci slovenskih osnovnih in srednjih šol samostojno napisali pri pouku, s tem pa empirične podatke o značilnostih šolske pisne produkcije. Korpus omogoča pripravo učinkovitejših, k uporabniku usmerjenih jezikovnih učnih gradiv in pristopov ter razvoj jezikovnotehnoloških orodij, ki podpirajo pisanje in opismenjevanje. Korpus, ki je po obsegu in sestavi primerljiv jezikovnim virom za druge evropske jezike (npr. Abel in sod., 2014; Barbagli in sod., 2016; Glaznieks in sod., 2022; Granger, 1998; Pala in sod., 2003), se razvija že dobro desetletje (Arhar Holdt in Kosem, 2024; Kosem in sod., 2016; Rozman in sod., 2012) in je trenutno dostopen v različici 3.0 (Arhar Holdt in sod., 2022). V okviru projekta *Empirična podlaga za digitalno podprto razvoj pisne jezikovne zmožnosti* (ARIS, J7-3159)¹ razvijamo metodologijo pridobivanja korpusnih podatkov za različne raziskovalne in razvojne namene, pri čemer nas zanimajo tudi podatki za skladenjske analize šolskega pisanja.

Če je še pred kratkim za slovenščino vladalo pomanjkanje skladenjsko označenih podatkov, pa tudi prosto dostopnih orodij za njihovo analizo in vizualizacijo (Ledinek, 2018), se v zadnjih letih stanje bistveno izboljšuje. Skladenjsko označevanje, ki je del označevalnika CLASSLA-Stanza 2.1 (Arhar Holdt in so., 2023; Terčon in Ljubešić, 2023), dosega 90-odstotno natančnost skladenjskega označevanja po sistemu Universal Dependencies in skoraj 94-odstotno po sistemu JOS-SYN. (Obliko)skladenjske oznake so na voljo v različnih besedilnih korpusih za slovenščino in so bile uporabljene za napredno pridobivanje podatkov, npr. na ravni kolokacij (Krek in sod., 2022), in za jezikoslovne analize, npr. analize skladenjskih značilnosti govorjenega jezika (Dobrovoljc, v objavi), računalniško

¹<https://www.cjvt.si/prop/>

posredovane slovenščine (Arhar Holdt, 2018) in udeleženskih vlog (Gantar, 2023). Na voljo so tudi zmogljiva orodja za luščenje oblikoskladenjskih in skladenjskih podatkov, kot so denimo Q-CAT (Brank, 2023), LIST (Krsnik in sod., 2019), Drevesnik (Štravs in Dobrovoljc, 2024) in novonastali STARK (Krsnik in sod., 2023),² ki ga uporabljamo in opišemo tudi v tem prispevku.

V nadaljevanju predstavljamo frekvenčne sezname skladenjskih struktur iz korpusa Šolar 3.0 in orišemo postopek njihovega luščenja, ki predstavlja enega prvih poskusov pridobivanja skladenjskih podatkov z orodjem STARK. Cilj je bil pridobiti besednozvezne skladenjske vzorce, kjer je jedro vsaka izmed besednih vrst izbranega označevalnega sistema, in medstavčne vzorce: priredja, soredja in podredja, skupaj s pripadajočimi vezniki (kar ponuja možnost podrobnejšega razvrščanja gradiva in na veznikih temelječe raziskave).

2 PRIPRAVA PODATKOV

2.1 O orodju STARK

STARK³ je prosto dostopno orodje, zgrajeno na osnovi Pythona in upravljano prek ukazne vrstice, ki iz vhodne datoteke – skladenjsko razčlenjenega korpusa (drevesnice) – proizvede izhodno datoteko – frekvenčni seznam skladenjskih struktur (dreves) (Dobrovoljc in sod., 2023). Vsebino izhodne datoteke narekujejo prilagodljivi luščilni kriteriji, ki omogočajo tako splošna kot podrobno usmerjena iskanja. Določiti je možno dolžino skladenjskega vzorca, npr. pri nastavitvi *size = 2-5* orodje vrne vse možne skladenjske strukture z 2–5 besedami (vozlišči). Iskanje lahko poljubno nadgrajujemo: določiti je možno lastnosti jedra strukture, in sicer je to lahko lema in/ali oblikoskladenjska oznaka (podprti so operatorji ‘|’ (ali), ‘&’ (in), ‘!’ (ne)); prikaz vozlišč v izluščenih strukturah, ki so lahko oblike, leme, oblikoskladenjske oznake, morfološke lastnosti ali skladenjske relacije – odvisno od namena (če za prikaz določimo obliko, bo drevo v izhodu zapisano npr. *Poiščeš >dve rešitev*, če pa določimo oznako MULTEXT-East, pa *Ggdsde >dve Sozet*); iskati je možno drevesa, pri čemer dovolimo le nekatere skladenjske povezave, lahko pa poljubne relacije v izluščenih vzorcih

²Na voljo na: <https://github.com/clarinsi/STARK>

³Več o delovanju orodja in nastavitvah luščenja je na voljo na <https://github.com/clarinsi/STARK> in v Dobrovoljc in sod. (2023).

tudi prepovemo. Možno je tudi iskanje s poizvedbo (ang. *query*), ki temelji na poizvedovalnem jeziku `dep_search`, ki ga uporablja tudi Drevesnik.⁴

STARK je bil sprva zasnovan za skladiščno razčlenjevanje korpusov, označenih z jezikovno univerzalno shemo Universal Dependencies⁵ (v nadaljevanju: UD), v različici 3.0 pa je omogočeno luščenje tudi prek drugih shem skladišnega označevanja. Če je prejšnja verzija orodja podpirala luščenje podatkov po sistemu z natanko enim korenskim elementom (ang. *root*) v povedi (kar je inherentna značilnost skladišne sheme UD), nova različica podpira luščenje prek shem z več koreni v povedi. Med te sodi tudi slovenščini prilagojen sistem JOS-SYN⁶ (Erjavec in sod., 2010), s katerim je mdr. označen korpus Šolar.

2.2 Besednozvezne skladišne strukture

Besednozvezne skladišne strukture iz korpusa Šolar 3.0 smo luščili prek oblikoskladišnih oznak (v nadaljevanju: MSD-oznake) MULTEXT-East (Erjavec, 2012). Vozlišča so v pridobljenih frekvenčnih seznamih ponazorjena z MSD-oznakami, skladišne relacije pa so po sistemu JOS-SYN (Erjavec in sod., 2010), ki je komplementarna dopolnitev omenjenega sistema oblikoskladišnega označevanja. Uporabili smo STARK 3.0, saj ta v nasprotju s prejšnjima različicama omogoča luščenje prek označevalne sheme z več kot enim korenskim elementom (gl. razdelek 2.1).

Pred luščenjem smo korpus predprocesirali. Orodje namreč oblikoskladišne oznake, ki naj jih uporabi za luščenje, pričakuje v stolpcu, kjer so v izvorni datoteki Šolarja UD-oznake. Ker želimo luščiti po sistemu MULTEXT-East, smo vrednosti v tem stolpcu prepisali z oznakami MULTEXT-East iz sosednjega stolpca, vendar so bile tja shranjene le prve črke oznake, ki kažejo na besedno vrsto. Tako smo se pri določanju jedra besedne zveze v nastavitvah luščenja izognili zvrščanju vseh obstoječih MSD-oznak za vsako besedno vrsto (samostalniki jim ima 104, npr. *Somei*, *Somer* itd.). Stolpec s celotnimi MSD-oznakami je ostal nespreme-

⁴Več o sestavljanju iskalnih nizov je na voljo na: <https://orodja.cjvt.si/drevesnik/help/en/> in <https://github.com/clarinsi/STARK/blob/master/settings.md>

⁵<https://universaldependencies.org/>; več o označevalnem sistemu je na voljo na: <https://wiki.cjvt.si/books/07-universal-dependencies>

⁶Več o označevalnem sistemu je na voljo na: <https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn>

njen, kar pomeni, da bodo v izvozu kot vozlišča skladišne strukture prikazane celotne MSD-oznake.

V Tabeli 1 je prikazan primer besednozvezne skladišne strukture v izhodni datoteki. V nadaljevanju predstavljamo tipe podatkov, ki jih najdemo v izvozu.

- *Tree* prikazuje eno izmed izluščenih skladišnih relacij (odvisnostnih dreves), ki ustreza iskalnemu nizu. Predstavljena je v obliki vozlišča A <relacija vozlišča B,⁷ kar nam pove: (a) vozlišča A je podrejeno vozlišču B (puščica nakazuje smer odvisnosti) in (b) med vozliščema A in B je skladišna razmerje [relacija].
- *Node A-xpos* in *Node B-xpos* izpišeta vsako vozlišče izluščene relacije posebej.
- *Absolute frequency* je absolutna pogostost izluščene relacije v korpusu.
- *Relative frequency* je relativna pogostost izluščene relacije v korpusu.
- *Order* prikazuje vrstni red vozlišč v izluščeni strukturi.
- *Number of nodes* prikazuje število vseh vozlišč v izluščeni strukturi.
- *Head node* izpiše vozlišče, ki predstavlja koren v izluščeni strukturi.
- *Example* izpiše naključno korpusno poved, v kateri je bila prepoznana skladišna struktura iz prvega stolpca; v povedi so vozlišča strukture tudi označena, tako da se vidi, katere pojavnice zavzemajo katero mesto v skladišni strukturi.
- *MI*, *MI3*, *Dice*, *logDice*, *t-score* in *simple-LL* so mere povezanosti, ki nam povejo moč statistične povezanosti med vozlišči drevesa znotraj korpusa, ki ga analiziramo. Te mere so tipično v rabi za merjenje jakosti kolokacij, lahko pa podajo uporabne informacije tudi za analizo skladišnih struktur.

⁷Vozlišča so lahko prikazana kot besedna oblika, lema ali oblikoskladišna oznaka.

Tabela 1: Ponazoritev najpogostejšega skladenjskega vzorca s samostalniškim jedrom v korpusu Šolar 3.0, skupaj s podatki, ki jih v izhodni datoteki vrne orodje. Tipi podatka, ki so v izvozu v stolpcih, so za namene jasnejše ponazoritve tukaj predstavljeni v vrsticah, v stolpcu pa so vrednosti tipa podatkov za eno strukturo. Najpogostejši skladenjski vzorec s samostalniškim jedrom je kombinacija predloga v mestniku (Dm) in občnoimenskega samostalnika ž. sp. edn. v mestniku (Sozem).

<i>Tree</i>	Dm <dol Sozem
<i>Node A-xpos</i>	Dm
<i>Node B-xpos</i>	Sozem
<i>Absolute frequency</i>	6486
<i>Relative frequency</i>	3400.2
<i>Order</i>	AB
<i>Number of nodes</i>	2
<i>Head node</i>	Sozem
<i>Example</i>	Samo A[v] B[smrti] je videla rešitev.
<i>MI</i>	3.90
<i>MI3</i>	29.22
<i>Dice</i>	0.19
<i>logDice</i>	11.59
<i>t-score</i>	75.14
<i>simple-LL</i>	3121.30

Končni rezultat je 11 izvozov z besednozveznimi skladenjskimi vzorci iz korpusa Šolar, po en izvoz za vsako od besednih vrst po sistemu MULTEXT-East: samostalnik, glagol, pridevnik, prislov, zaimек, števnik, predlog, veznik, členek, medmet, okrajšava (kategoriji neuvrščeno in ločilo nista vrnila rezultatov). Vsakemu izvozu pripada še izhodna datoteka s prepoznanimi drevesi v vseh povedih korpusa.⁸

2.3 Medstavčne skladenjske strukture

Medstavčne strukture smo luščili z verzijo orodja 2.0 na osnovi skladenjskih oznak UD, saj so oznake sistema JOS-SYN, ki so na voljo v korpusu Šolar 3.0, zasnovane za souporabo z oblikoskladenjskimi oznakami MULTEXT-East, STARK pa ne podpira možnosti, da bi jih bilo mogoče učinkovito uporabiti pri izdelavi

⁸Konfiguracijska datoteka bo ob koncu projekta objavljena poleg izvozov.

luščilnih parametrov. Korpus smo pred luščenjem stavčnih struktur zato morali strojno označiti še s skladenjskimi relacijami po sistemu UD, s CLASSLO-Stanza 2.1 (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić, 2023).

Za namene bodočih jezikoslovnih analiz na ravni medstavčne skladnje smo v korpusu identificirali razmerja, ki kažejo na sestavo povedi: priredje, soredje in podredje. Pri podredjih smo posebej luščili osebkov odvisnik, predmetni odvisnik in prislovne odvisnike. Slednji po sistemu skladnje UD združuje vse tipe prislovnih odvisnikov, zato v iskalni pogoj kot vozlišča vključujemo tudi veznike, s pomočjo katerih je možno natančneje urejati in analizirati izvožene podatke.

Definiranje poizvedbe za luščenje osebkovega (csubj), predmetnega (ccomp) in prislovnih odvisnikov (advcl) skupaj z vezniki je bilo relativno preprosto, saj v osnovi povezujejo samo povedke in ni treba dodatno definirati izvora in cilja povezave. Upoštevati pa je treba tudi večbesedne veznike. Pri luščenju smo tako uporabili tri poizvedbe, kot kaže spodnji primer za prislovne odvisnike:

a. `_ >advcl (_ >mark _)` vrne prislovni odvisnik z veznikom (cilj relacije mark), čemur ustreza skladenjski vzorec v Tabeli 2.

b. `_ >advcl (_ >mark (_ >fixed _))` dodatna relacija fixed povezuje dele večbesednih enot, kamor po UD spadajo tudi večbesedni vezniki; kot vozlišči sta določeni dve vezniški besedi (npr. *Jazonu je ostala zvesta in mu A[pomagala] do prestola, B[tako] C[da] je D[ubila] kralja.*).

c. `_ >advcl ((_ >fixed _ >fixed _) <mark _)` dve relaciji fixed ustrezata trobesednim veznikom (npr. *Antigona je A[pokopala], B[kljub] C[temu] D[da] je E[vedela] kakšne so posledice sodbe.*).

Ker se našteje poizvedbe delno prekrivajo, se lahko iste povedi pojavijo v več izvozi. Poizvedba a. namreč določa kot vozlišče veznik (cilj povezave mark), ne določa pa naslednjega vozlišča (ki je lahko tudi druga beseda dvo- ali trobesednega veznika). Tako se lahko v rezultatu a. pojavi tudi stavek z dvobesednim veznikom, pri čemer prva beseda ima vozlišče, druga pa ne. Primer v Tabeli 2 je rezultat poizvedbe a., vendar isto poved najdemo tudi v izvozu b., kjer sta vozlišči obe besedi veznika. Zato smo uvedli fazo poprocesiranja, kjer smo iz

predstavljenih treh izvozov izbrisali duplikate in obdržali le unikatno poved z največjim številom vezniških vozlišč.

Tabela 2: Ponazoritev skladenjskega vzorca s prislovnim odvisnikom v korpusu Šolar 3.0, skupaj s podatki, ki jih v izhodni datoteki vrne orodje. Vzorec (*Tree*), predstavljen z UD-oznakami in relacijami kot vozlišči, prinaša povezavo glagola glavnega stavka z glagolom podrednega stavka (advcl je povezava za prislovni odvisnik), slednji pa je povezan s podrednim veznikom (cilj relacije mark). Podatki, ki so v izvozu v stolpcih, so za jasnejšo ponazoritev tukaj predstavljeni v vrsticah, v stolpcu pa so vrednosti tipa podatka za eno strukturo; za razlago tipov podatka gl. razdelek 2.2.

<i>Tree</i>	(ADV <mark VERB) <advcl VERB
<i>Node A-upos</i>	ADV
<i>Node B-upos</i>	VERB
<i>Node C-upos</i>	VERB
<i>Absolute frequency</i>	212
<i>Relative frequency</i>	111.1
<i>Order</i>	ABC
<i>Number of nodes</i>	3
<i>Head node</i>	VERB
<i>Example</i>	A[Medtem], ko sta se B[prepirala], je vonj po piščancu C[zavohal] tudi jež.
<i>MI</i>	-2.69
<i>MI3</i>	12.77
<i>Dice</i>	0,00
<i>logDice</i>	4.23
<i>t-score</i>	-79.36
<i>simple-LL</i>	1967.72

Definiranje poizvedbe za priredje (conj) in soredje (parataxis) je bilo zahtevnejše, saj ti relaciji po shemi UD označujeta tudi druge, ne nujno stavčne strukture. Tako je bilo treba za pridobitev priredij in soredij v iskalnem pogoju določiti tudi izvor in cilj povezave, ki morata biti glagolska. Pokriti je bilo treba tako stavke s polnopomenskim glagolom (upos=VERB) kot tiste s povedkovim določilom (ta vzorec smo definirali s povezavo cop, ki ima na izvoru relacije vezni glagol *biti*, na cilju pa besedo v vlogi povedkovega določila). Da smo pokrili vse možne kombinacije, smo uporabili štiri poizvedbe, kot kaže spodnji primer za soredja:

- a.u pos=VERB >parataxis upos=VERB za soredja, kjer je polnopomenski glagol (VERB) na izvoru in cilju povezave.
- b.u pos=VERB >parataxis (_ <cop _) za soredja, kjer je na izvoru povezave polnopomenski glagol, na cilju pa povedkovo določilo.
- c. _ >cop _ >parataxis upos=VERB za soredja, kjer je na izvoru povezave povedkovo določilo, na cilju pa polnopomenski glagol.
- d. _ >cop _ >parataxis (_ <cop _) za soredja, kjer je povedkovo določilo na izvoru in cilju povezave.

Pri luščenju priredij smo dodali še vozlišča za veznike, ki so spet lahko eno- ali večdelni. Za ponazoritev navajamo poizvedbo za priredje, kjer je na izvoru povezave polnopomenski glagol, na cilju povedkovo določilo, povezana pa sta z dvobesednim prirednim veznikom: upos=VERB >conj ((_ <fixed _) <cc _ <cop _). Tudi pri izvozi priredij smo izvedli zgoraj opisani postopek deduplikacije. To vpliva na statistično sliko izvoženih podatkov, kar je treba upoštevati pri interpretaciji rezultatov.

Končni rezultat je 5 izvozov z medstavčnimi skladenjskimi vzorci iz korpusa Šolar. Več izvozov za istovrstno medstavčno relacijo, ki so nastali zaradi podrobne obravnave veznikov ali vključitve povedkovega določila kot jedra stavka, je bilo po fazi poprocesiranja združenih v enega. Vsakemu izvozu pripada še izhodna datoteka s prepoznanimi drevesi v vseh povedih korpusa.⁹

2.4 Dostopnost podatkov

Izvozi skladenjskih struktur iz korpusa Šolar 3.0, ki so v obliki frekvenčnih seznamov, bodo po koncu projekta na voljo na repozitoriju CLARIN.SI; dodali bomo tudi predlog za njihovo analizo oz. vizualizacijo s programom MS Excel.

2.5 Omejitve

Uporaba predstavljene metodologije luščenja skladenjskih korpusnih podatkov predpostavlja dobro poznavanje označevalnih shem UD in JOS-SYN in osnovno uporabo ukazne vrstice za zagon programa oz. uporabo uporabniku prijaznejšega integriranega razvojnega okolja (ang. *integrated development environment*;

⁹Konfiguracijska datoteka bo ob koncu projekta objavljena poleg izvozov.

IDE), kot je npr. PyCharm. Nadalje je za luščenje skladijskih dreves treba usvojiti poizvedovalni jezik `dep_search`, ki ga uporablja STARK. Osnovni napotki so na voljo na strani GitHub.¹⁰ Ker poizvedovalni jezik temelji na poizvedovalnem jeziku Drevesnika, spletnega vmesnika za iskanje po slovenskih skladijskih drevesnicah, je dobro poznati tudi tega.¹¹

Pri orodju STARK je za naš scenarij rabe umanjala možnost uporabe operatorja za negacijo (!) pred puščico (pri sestavi poizvedbe (ang. *query*)). Na voljo je negacija relacije, ki v primeru `_ >!fixed _` prepoveduje pojav druge besede večbesedne enote, ni pa možnosti negacije katere koli podrejene relacije. Npr. z iskalnim nizom:

`_ >advcl (_ >mark (_ !>fixed _))` bi izločili vse večbesedne veznike oz. izluščili le enobesedneveznike. Ta možnost bi v našem primeru zmanjšala število poizvedb in s tem čas ter procesno breme, hkrati pa ne bi bilo potrebno poprocesiranje izvozov zvezniki (gl. razdelek 2.3).

Upoštevati je treba, da luščimo na osnovi strojno označenih korpusnih podatkov, ki lahko vsebujejo napake v pripisanih skladijskih povezavah. Upoštevati je treba tudi specifične sheme UD. Primerjalni odvisniki s 'kot', npr. *Kakor vem je A[več] pogovora o problemu, B[kot] C[pa] da bi bili takoj D[nasilni]*, so označeni kot prislovni odvisniki (`advcl`), pri čemer pa je izvor povezave lastnost, na katero se primerjava nanaša, in ne jedro povedke glavnega stavka, kot velja pri drugih primerih. Za navedeno poved se tako na mestu, kjer bi pričakovali glagol, pojavi kot vozlišče 'več' (označen kot izvor povezave `advcl`).

3 NADALJNJI KORAKI

Naslednji korak je pregled skladijskih podatkov iz korpusa Šolar 3.0 in njihova ureditev glede na stopnjo šolanja. Po predstavljeni metodologiji bo opravljeno tudi luščenje iz korpusa učbenikov (Kosem in sod., 2022), kar bo omogočilo primerjalne analize produkcijskega in pričakovanega recepcijskega dometa na različnih stopnjah šolanja. Izvoženi, urejeni in dokumentirani podatki, ki jih je v bodoče mogoče bogatiti še z drugimi informacijami, npr. merami skladijske kompleksnosti, bodo skupnosti odprto na voljo in bodo kot taki lahko služili kot

¹⁰Na voljo na: <https://github.com/clarinsi/STARK/blob/master/README.md> in <https://github.com/clarinsi/STARK/blob/master/settings.md>

¹¹<https://orodja.cjvt.si/drevesnik/>

empirična podstat za različne namene, od korpusnih skladenjskih analiz, preverjanj hipotez o značilnostih šolskega pisanja do priprave didaktičnih smernic, učnih gradiv in drugih aplikativnih ciljev.

4 ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* (J7-3159) in raziskovalnega programa Jezikovni viri in tehnologije za slovenski jezik (P6-0411), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS). Zahvaljujeva se dr. Kaji Dobrovoljc za pomoč pri delu s programom STARK in recenzentom za koristne predloge.

LITERATURA

- Abel, A., Glaznieks, A., Nicolas, L. in Stemle, E. (2014). Koko: An l1 learner corpus for german. V *Proceedings of the 9th international conference on language resources and evaluation* (str. 2414–2421). Reykjavik, Iceland.
- Arhar Holdt, Š. (2018). Korpusni pristop k skladnji računalniško posredovane slovenščine. V D. Fišer (Ur.), *Viri, orodja in metode za analizo spletne slovenščine* (Zv. 1, str. 228–252). Ljubljana: Znanstvena založba Filozofske fakultete. <https://doi.org/10.4312/9789610600701>
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon, L., & Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v SUK 1.0. V Š. Arhar Holdt in S. Krek (Ur.), *Razvoj slovenščine v digitalnem okolju* (str. 119–156). Založba Univerze. <https://doi.org/10.4312/9789612972561>
- Arhar Holdt, Š. in Kosem, I. (2024). Šolar, the developmental corpus of slovene. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-024-09758-4>
- Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L., & Kosem, I. (2022). *Developmental corpus šolar 3.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1589>
- Barbagli, A., Lucisano, P., Dell’Orletta, F., Montemagni, S. in Venturi, G. (2016). Cita: An l1 italian learners corpus to study the development of writing competence. V *Proceedings of the 10th international conference on language resources and evaluation* (str. 88–95). Portorož, Slovenia.
- Brank, J. (2023). *Q-CAT corpus annotation tool 1.5*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1844>

- Dobrovoljc, K. (v objavi). Uporaba drevesnice sst v raziskavah govornjene slovenščine: prednosti in omejitve. *Jezik in slovstvo*.
- Dobrovoljc, K., Krsnik, L. in Šikonja, M. R. (2023). Stark: a tool for dependency tree extraction and analysis. V *Unidive 1st general meeting: universality, diversity and idiosyncrasy in language technology* (str. 1-3). Paris-Saclay University. https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:2023-saclay:abstracts:62_dobrovoljc_et_al_stark_a_tool_for_dependency_tree.pdf
- Erjavec, T. (2012). Multext-east: morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, 46(1), 131–142. <https://doi.org/10.1007/s10579-011-9174-8>
- Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010). The jos linguistically tagged corpus of slovene. V *Proceedings of the seventh conference on international language resources and evaluation (Irec'10)*. Valeta, Malta. <https://aclanthology.org/L10-1087/>
- Gantar, P. (2023). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidik. V M. Smolej in M. Schlamberger Brezar (Ur.), *Prispevki k preučevanju slovenske skladnje* (Zv. 1, str. 77–97). Ljubljana: Založba Univerze. <https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/421/764/8503>
- Glaznieks, A., Frey, J. C., Stopfner, M., Zanasi, L. in Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of italian, german and english. *International Journal of Learner Corpus Research*, 8(1), 97–120.
- Granger, S. (1998). The computer learner corpus: A versatile new source of data for sla research. V S. Granger (Ur.), *Learner english on computer* (str. 3–18). Addison Wesley Longman.
- Kosem, I., Pori, E., Žagar, A. in Arhar Holdt, Š. (2022). *Corpus of slovenian textbooks ccUčbeniki 1.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1693>
- Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P. in Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. V T. Erjavec in D. Fišer (Ur.), *Zbornik konference jezikovne tehnologije in digitalna humanistika* (str. 95–100). Ljubljana: Znanstvena založba Filozofske fakultete. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf
- Krek, S., Gantar, P. in Kosem, I. (2022). Extraction of collocations from the gigafida 2.1 corpus of slovene. V A. Klosa (Ur.), *Euralex 2022, proceedings of the xx euralex international congress* (str. 240–252). Mannheim, Germany: IDS-Verlag. <https://euralex.org/publications/extraction-of-collocations-from-the-gigafida-2-1-corpus-of-slovene/>

- Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S. in Robnik-Šikonja, M. (2019). *Corpus extraction tool LIST 1.2*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1276>
- Krsnik, L., Dobrovoljc, K. in Robnik-Šikonja, M. (2023). *Dependency tree extraction tool STARK 2.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1899>
- Ledinek, N. (2018). Skladijska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in slovstvo*, 63(2/3), 103–116. <https://doi.org/10.4312/jis.63.2-3.103-116>
- Ljubešić, N. in Dobrovoljc, K. (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). Florence, Italy: Association for Computational Linguistics. <https://www.aclweb.org/anthology/W19-3704>
- Pala, K., Rychlý, P. in Smrž, P. (2003). Text corpus with errors. V V. Matoušek in P. Mautner (Ur.), *Text, speech and dialogue (tsd 2003) lecture notes in computer science* (Zv. 2807, str. 90–97). Springer.
- Rozman, T., Stritar, M. in Kosem, I. (2012). Šolar – korpus šolskih pisnih izdelkov. V T. Rozman in sod. (Ur.), *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina, digitalna različica Založba Univerze v Ljubljani. <https://doi.org/10.4312/9789610603511>
- Štravs, M. in Dobrovoljc, K. (2024). *Service for querying dependency treebanks drevesnik 1.1*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1923>
- Terčon, L. in Ljubešić, N. (2023). *Classla-stanza: The next step for linguistic processing of south slavic languages*. arXiv. <https://doi.org/10.48550/arXiv.2308.04255>

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna.

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International.

<https://creativecommons.org/licenses/by-sa/4.0/>

