

Na poti k skladienjskim analizam šolskega pisanja: skladienjski vzorci v korpusu Šolar 3.0

Tina Munda¹
Špela Arhar Holdt^{1,2}

1 Filozofska fakulteta UL
2 Fakulteta za računalništvo in informatiko UL

tina.munda@cjvt.si
spela.arharholdt@ff.uni-lj.si

Prikazujemo celostno luščenje skladienjskih vzorcev iz razvojnega korpusa šolskih pisnih izdelkov Šolar 3.0.

Metoda priprave korpusnih podatkov je osnovana na prosto dostopnem orodju za luščenje odvisnostnih drevesnic in analizo STARK.

Izluščeni podatki prikazujejo:

(a) besednozvezne skladienjske strukture, kjer je jedro vsaka izmed besednih vrst po označevalnem sistemu MULTEXT-East: samostalnik, glagol, pridevnik, prislov, zaimek, števnik, predlog, veznik, členek, medmet, okrajšava.

(b) medstavčne skladienjske strukture, in sicer: priredja, sorredja in podredja, skupaj s pripadajočimi vezniki.

Pridobljeni podatki so empirična osnova za skladienjske analize šolskega pisanja in dajejo dragocen vpogled v pisno zmožnost šolajoče se populacije, kar lahko pripomore k razvoju učinkovitejših šolskih učnih gradiv, pa tudi jezikovnotehnoloških orodij v podporo pisanju in opismenjevanju.

Omejitve metode: dobro poznavanje označevalnih shem UD in JOS-SYN; osnovna uporaba ukazne vrstice za zagon programa (ali IDE); poznavanje poizvedovalnega jezika dep_search, ki ga uporablja STARK; napake v strojnem označevanju; specifikke sheme UD.

Podatki bodo odprto na voljo na CLARIN.SI.

Afiliacija



Projekt



Financer



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 RS (ARIS). Zahvaljujemo se dr. Kaji Dobrovoljc za pomoč pri delu s programom STARK in recenzentom za koristne predloge.



STARK:
demo različica

STARK:
izvorna koda



Vhodna datoteka (format CoNLL-U):

```
# sent_id = solar1.1.1
1 1 1 NUM Kag NumForm=Digit|NumType=Card 0 modra
2 . . PUNCT U _ 0 modra
3 Šolska šolski ADJ Fpnzei Case=Nom|Degree=Pos|Gender=Fem|Number=Sing 4 dol
4 naloga naloga NOUN Sozei Case=Nom|Gender=Fem|Number=Sing 0 modra

# sent_id = solar1.2.1
1 Ne ne PART L _ Polarity=Neg 3 vez
2 da da SCOUNJ Vd _ 1 skup
3 sovražim sovražiti VERB Ggnspe Aspect=Imp|Mood=Ind|Number=Sing|Person=1|... 0 modra
4 , , PUNCT U _ 0 modra
5 da da SCOUNJ Vd _ 6 vez
6 ljubim ljubiti VERB Ggnspe Aspect=Imp|Mood=Ind|Number=Sing|Person=1|... 3 dve
7 sem biti AUX Gp-spe-n Mood=Ind|Number=Sing|Person=1|Polarity=Pos|... 6 del
8 na na ADP Dm Case=Loc 9 dol
9 svetu svet NOUN Somem Case=Loc|Gender=Masc|Number=Sing 6 štiri

# sent_id = solar1.3.1
1 " " PUNCT U _ 0 modra
2 Ne ne PART L _ Polarity=Neg 4 vez
3 da da SCOUNJ Vd _ 2 skup
4 sovražim sovražiti VERB Ggnspe Aspect=Imp|Mood=Ind|Number=Sing|Person=1|... 13 tri
5 , , PUNCT U _ 0 modra
6 da da SCOUNJ Vd _ 7 vez
7 ljubim ljubiti VERB Ggnspe Aspect=Imp|Mood=Ind|Number=Sing|Person=1|... 4 dve
8 sem biti AUX Gp-spe-n Mood=Ind|Number=Sing|Person=1|Polarity=Pos|... 6 del
9 na na ADP Dm Case=Loc 10 dol
10 svetu svet NOUN Somem Case=Loc|Gender=Masc|Number=Sing 7 štiri
11 " " PUNCT U _ 0 modra
12 , , PUNCT U _ 0 modra
13 izreče izreči VERB Ggdste Aspect=Perf|Mood=Ind|Number=Sing|Person=3|... 0 modra
14 Antigona Antigona PROPNI Slzei Case=Nom|Gender=Fem|Number=Sing 13 ena
15 v v ADP Dt Case=Acc 16 dol
16 bran bran NOUN Sometn Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing 13 štiri
17 kralju kralj NOUN Sometn Case=Dat|Gender=Masc|Number=Sing 13 dve
18 Kreonu Kreon PROPNI Slmed Case=Dat|Gender=Masc|Number=Sing 17 dol
19 za za ADP Dt Case=Acc 22 dol
20 svoje svoj DET Zp-set Case=Acc|Gender=Neut|Number=Sing|Poss=Yes|... 22 dol
21 nasprotno nasproten ADJ Fpnset Case=Acc|Degree=Pos|Gender=Neut|Number=Sing 22 dol
22 mišljenje mišljenje NOUN Soset Case=Acc|Gender=Neut|Number=Sing 13 štiri
23 pred pred ADP Do Case=Ins 24 dol
24 smrtjo smrt NOUN Sozeo Case=Ins|Gender=Fem|Number=Sing 22 dol
25 . . PUNCT U _ 0 modra
```

Konfiguracijska datoteka:

```
***** BASIC SETTINGS (see settings.md) ***
;---GENERAL SETTINGS---
input = sample/solar-orig_sl_short-xpos.conllu
output = sample/output_solar-orig_sl_samostalnik.tsv

;---TREE SPECIFICATIONS---
node_type = xpos
labeled = yes
label_subtypes = no
fixed = yes

;---TREE RESTRICTIONS---
size = 2-10
head = upos=S
;ignored_labels = punct
;allowed_labels = nsubj|obj|obl

;---SEARCH BY QUERY---
query = _ >advcl (_ >mark _)

;---ADDITIONAL STATISTICS---
association_measures = no
compare = sample/fr_gsd-ud-dev.conllu

;---VISUALISATION---
example = yes
grew_match = no
depsearch = no

;---OUTPUT THRESHOLD---
frequency_threshold = 5
max_lines = 100

***** ADVANCED SETTINGS (see advanced.md)
internal_saves = ./internal_saves
cpu_cores = 12
continuation_processing = no
greedy_counter = yes
complete = yes
processing_size = 1-7
```

STARK

luščenje prek shem
MULTEXT-East+JOS-SYN

Izhodna datoteka - besednozvezni skladienjski vzorci (samostalniško jedro):

Tree	Node A-xpos	Node B-xpos	Node C-xpos	Absolute f	Relative f	Order	Nr of nodes	Head node	Example
Dm < dol Sozem	Dm	Sozem		6486	34002	AB	2	Sozem	A[V] B[povedi] "Pokoril se je in si dopovedoval, da prijatelj[uje z Agico iz osebnih koristi]" lahko jasno vidimo, da je njen oče glavni razlog zakaj je z njo.
Dm < dol Somem	Dm	Somem		5633	29530	AB	2	Somem	Odpuliti smo in A[n] B[otoku] pustili glavne upornike.
Dt < dol Sozet	Dt	Sozet		5202	27270	AB	2	Sozet	Ko se jih je dotaknil, ga je zalotil oče Polikarp in mu A[za] B[kazen] odrezal mezinec.
									Toda tudi škof podležje njenim čarom in jo pošlje A[v] B[samostan] na poti v samostan se ustavi na pečini nad Renom in se vrže v reko, vitezi pa žalostno končajo v skali, od koder se ne morejo vrniti.
Dt < dol Sometn	Dt	Sometn		3422	17939	AB	2	Sometn	Danes se opazi A[občutna] B[razlika], saj je znanje veliko bolj pomankljivo.
Ppnzei < dol Sozei	Ppnzei	Sozei		2975	15596	AB	2	Sozei	Primarna socializacija poteka A[v] B[otroštvu], starši nas naučijo osnov vključevanja v družbo, nauči se temeljnih pravil, starši mu z kaznovanjem ali pohvalo/nagrado*
Dm < dol Sosem	Dm	Sosem		2648	13882	AB	2	Sosem	1. A[šolski] B[spis]
Ppnmeid < dol Somei	Ppnmeid	Somei		2300	12057	AB	2	Somei	Odnos med Simonom in Volodjo je bil zelo zavajajoč A[s] B[strani] Volodje kar se kaže tudi z tem stavkom "Simon: Delam na inštitutu za
Dr < dol Sozer	Dr	Sozer		2243	11758	AB	2	Sozer	
Dr < dol Somer	Dr	Somer		2086	10935	AB	2	Somer	Osredje sporočilo besedila je da lahko tudi ljudje, ki naj bi bili izobraženi za delo z ljudmi nimajo pravega odnosa A[do] B[sočloveka] in da se do njih obnašajo absurdno ter žaljivo. Ko mu je umrl dober prijatelj, A[Matija] B[Čop], je napisal lirsko-epsko pesnitev Krst pri Savici v 19. stoletju.
Slmei > dol Slmei	Slmei	Slmei		1982	10390	AB	2	Slmei	

luščenje prek sheme UD

Izhodna datoteka - medstavčni skladienjski vzorci (prislovni odvisnik):

Tree	Node A-upos	Node B-upos	Node C-upos	Absolute f	Relative f	Order	Nr of nodes	Head node	Example
VERB > advcl (SCONJ < mark VERB)	VERB	SCONJ	VERB	10052	52696	ABC	3	VERB	On je nekaj let mlajši od nje, A[spoznal] pa jo je v domu, B[ko] je C[obiskoval] obnemoglega brata.
(SCONJ < mark VERB) < advcl VERB	SCONJ	VERB	VERB	5547	29079	ABC	3	VERB	-če si v družbi so neka pravila, A[če] jih ne B[upoštevajš] te hitro C[izločijo] iz družbe
ADJ > advcl (SCONJ < mark VERB)	ADJ	SCONJ	VERB	1841	9651	ABC	3	ADJ	Bilo je bolj A[dolgočasno], B[čeprav] smo vsi komaj C[čakali], da prestopimo prag te šole.
VERB > advcl (ADV < mark VERB)	VERB	ADV	VERB	1197	6275	ABC	3	VERB	Rekel je: „Kaj boste A[storili] B[zdaj], ko vam C[hočejo] vzeti jezik?"
VERB > advcl (SCONJ < mark ADJ)	VERB	SCONJ	ADJ	1090	5714	ABC	3	VERB	Ti pari, ki so neizkušeni in ne pripravljeni na skupno in samostalno življenje brez pomoči staršev in drugih se hitro A[razpadejo], B[ker] niso C[pripravljeni] na izkušnje in probleme v svoji lasni družini, ki mora samostalno živeti in razmišljati s svojo glavo.
(SCONJ < mark ADJ) < advcl VERB	SCONJ	ADJ	VERB	584	3061	ABC	3	VERB	A[Ko] je bil Ivan Cankar B[majhen], je otroke zelo C[privačila] ljubljani.
VERB > advcl (SCONJ < mark NOUN)	VERB	SCONJ	NOUN	473	2480	ABC	3	VERB	Za mamo to A[predstavlja] dodaten stres, B[ker] je C[samohranika] in skrbelo jo je tudi nadaljno šolanje za Nino.
(SCONJ < mark VERB) < advcl ADJ	SCONJ	VERB	ADJ	466	2443	ABC	3	ADJ	A[Preden] sem B[postal] kralj Teb je bilo moje življenje C[umirjeno] in srečno, a vse to se je spremenilo, ko sta se nečaka Polinejk in Eteokelj po očetovi smrti skregala, kdo bo vladal Tebam, ker se nista mogla zneniti je prišlo do bojevanja v katerem sta oba umrla.
NOUN > advcl (SCONJ < mark VERB)	NOUN	SCONJ	VERB	408	2139	ABC	3	NOUN	Njegov smisel je A[lubezen] do Bogomile, pa B[čeprav] C[ve] da je nikoli več ne bo okusil.
ADV > advcl (SCONJ < mark VERB)	ADV	SCONJ	VERB	293	1536	ABC	3	ADV	A[Dopolodne], ko bomo v službi in popoldne, B[ko] se bomo C[ukvarjali] z drugimi rečmi.

Poster je bil predstavljen na konferenci Jezikovne tehnologije in digitalna humanistika 2024 (Ljubljana, Slovenija, 19. 9. 2024). Dodatne informacije, vključno z referencami, so na voljo v prispevku, ki je objavljen v konferenčnem zborniku.