Korpusni viri in orodja za skladenjske raziskave slovenskega jezika

Kaja Dobrovoljc Zor

Univerza v Ljubljani, FF UL / FRI UL
Institut Jožef Stefan

Simpozij *Na poti do nove slovenske slovnice*, 9. oktober 2026

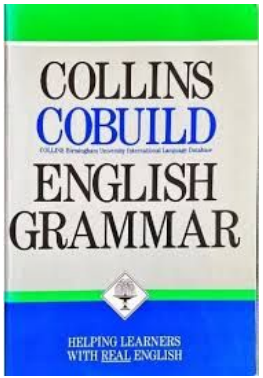
Današnja predstavitev

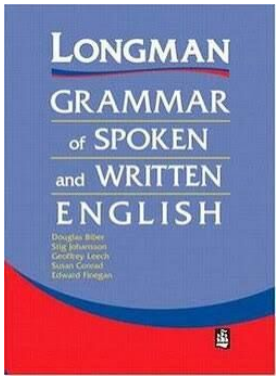
1. **Vloga korpusov** v slovničnih raziskavah jezika
2. Izbrani **skladenjsko razčlenjeni korpusi** slovenskega jezika (drevesnice)
3. Razpoložljiva **orodja za analizo** skladenjsko razčlenjenih korpusov

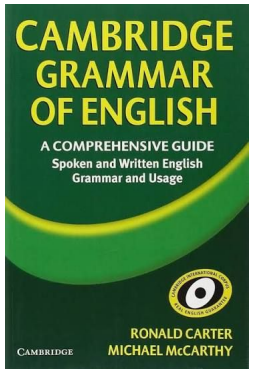


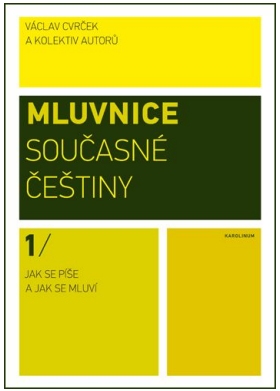
Od jezikovnega sistema k jezikovni rabi

- funkcijski pogled na jezik
  - opis dejanske rabe v realnih sporazumevalnih okoliščinah
- korpus kot empirična osnova
  - kaj se pojavlja, kako pogosto, kje
- porast korpusnih slovnic v številnih jezikih
  - tudi prvi temelji za slovenščino (projekt ARIS *NSSS* 2017–2020)









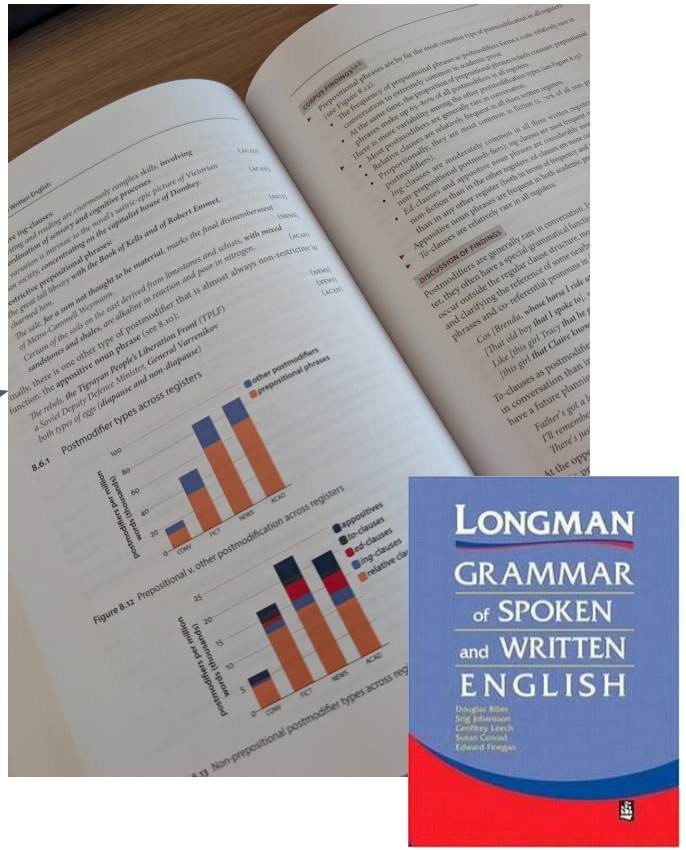


Kaj korpus doda slovničnemu opisu?

- preverjanje hipotez in intuicij
- ločevanje tipičnega od obrobnega
- opis variabilnosti v rabi (zvrsti)
- odkrivanje novih pojavov
- avtentični primeri

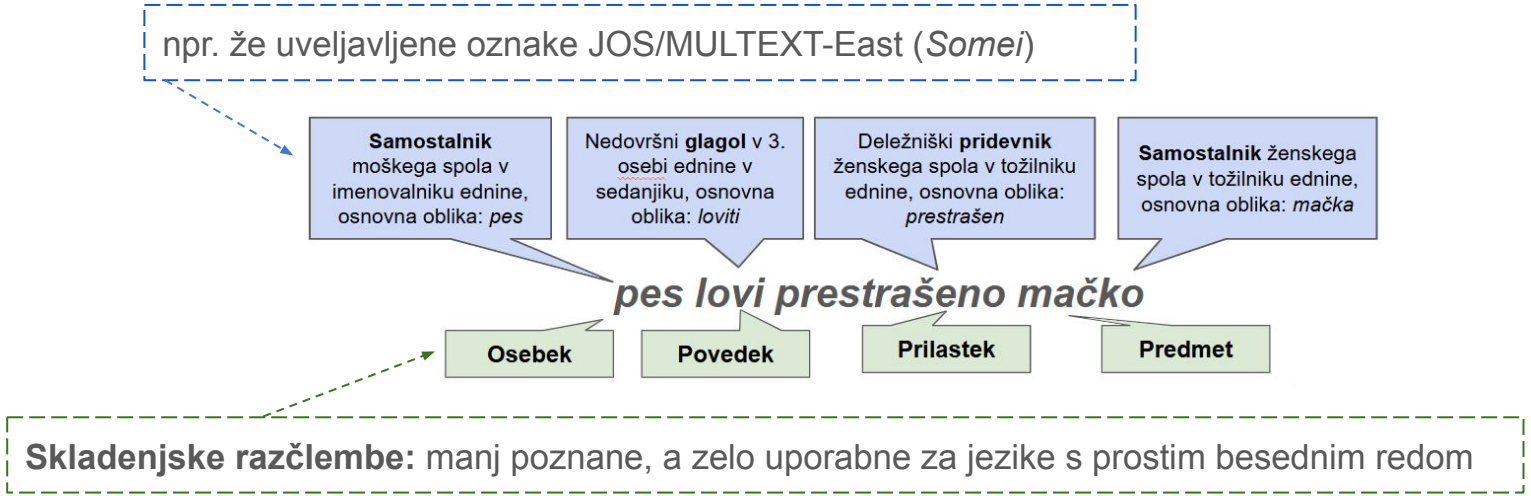
Primer iz **Biber et al. (1999)**
Longman Grammar of Spoken and Written English

slovnični opis sistematično dopolnjujejo s korpusno analizo (primeri, pogostost, razlike med registri)



Slovnično označevanje korpusov

- dodajanje slovničnih interpretacij (oznak) besedam na površini
- uporabno za: priklic pojavov, kompleksne analize, strojno učenje
- tipično: **oblikoslovna** in **skladenjska** razčlemba povedi



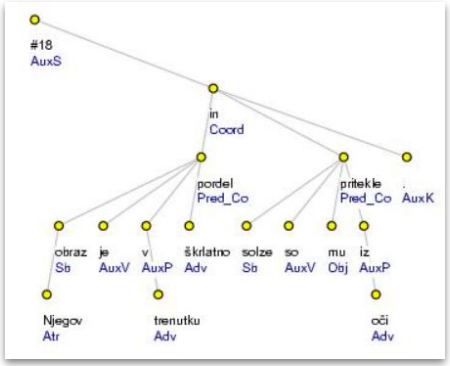
Skladenjsko razčlenjeni korpusi za slovenščino

- več pristopov, vsi temeljijo na **odvisnostni slovnici** (Tesnière 1959)
  - označevanje skladenjskih razmerij med besedami (skladenjskih odvisnosti)

Drevesnica SDT

(Džeroski et al. **2006**)

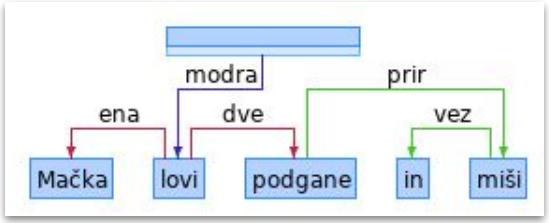
- po vzoru *Prague Dependency Treebank*
- 3.000 povedi iz romana *1984*



Drevesnica JOS / ssj500k

(Ledinek et al. **2010**)

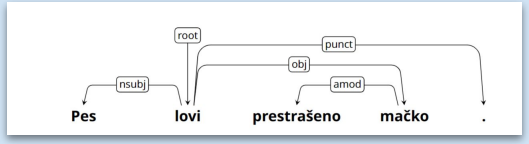
- poenostavljena lokalna shema
- 11.000 povedi iz korpusa FidaPLUS



Drevesnice UD

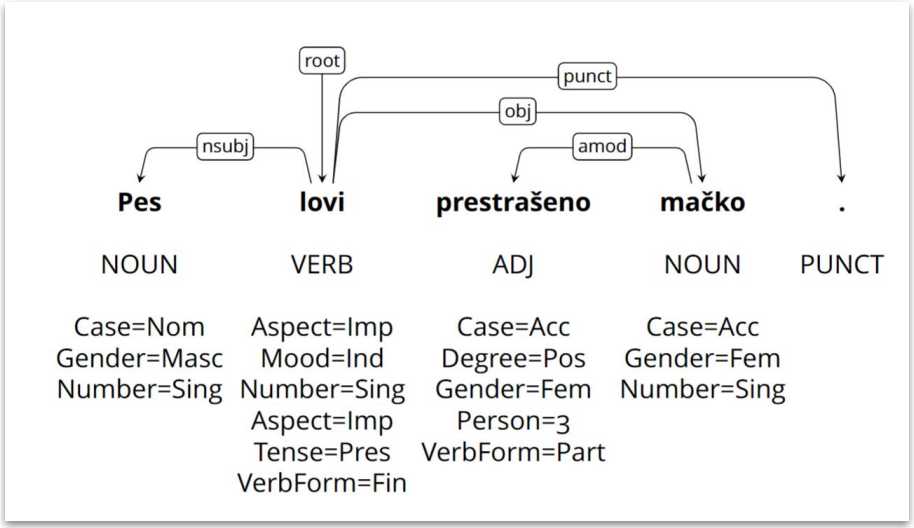
(Dobrovoljc et al. **2015+**)

- **Universal Dependencies**: mednarodno usklajen pristop
- 13.500 povedi pisne slovenščine (JOS+Elexis)
- 6.000 izjav govorjene slovenščine (GOS)



Shema Universal Dependencies (UD)

- shema za **medjezikovno primerljivo slovnično označevanje** korpusov
  - besedne vrste (17 oznak)
  - oblikoslovne lastnosti (24 lastnosti)
  - skladenjske odvisnosti (37 relacij)
  - (+ lokalne oznake JOS/MTE!)
- uporabljena že za **več kot 190 jezikov**, tudi slovenščino
  - smernice, dokumentirane v priročniku
  - več ročno in strojno označenih korpusov



Pes lovi prestrašeno mačko

Ročno označeni korpusi za slovenščino



Drevesnica **SSJ** (SUK): **pisna**

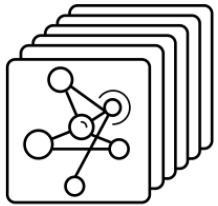
- **13.500 razčlenjenih povedi**
  - 267.000 besed
  - publicistična, literarna, stvarna, spletna, wiki besedila
  - polavtomatska pretvorba iz drevesnice JOS + novi podatki iz RSDO
- tudi kot del učnega korpusa SUK
  - + udeleženske vloge, stalne besedne zveze, imenske entitete ...



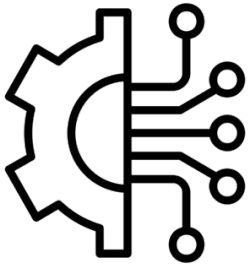
Drevesnica **SST** (ROG): **govorjena**

- **6.000 razčlenjenih izjav**
  - 76.000 besed, 10 ur posnetkov
  - spontani govor iz korpusa GOS 2
  - uravnotežen glede na vrsto diskurza, kanal in značilnosti govorcev
  - samopopravljanja, diskurzni označevalci in drugi tipično govorjeni pojavi
- tudi kot del učnega korpusa ROG
  - + prozodične enote, netekočnosti, dialoška dejanja …

Strojno označeni korpusi za slovenščino



UČNI KORPUSI
(SSJ-SUK, SST-ROG)



RAČUNALNIŠKI MODELI
(CLASSLA-Stanza, SPOT-Trankit, SpaCy-SL … in drugi **razčlenjevalniki**)

Danes že precej natančni:

- Leme: **99** %
- Besedne vrste: **99** %
- Oblikoslovne lastnosti: **98** %
- Skladenjske relacije: **95** %



STROJNO RAZČLENJENI KORPUSI (novi, večji)

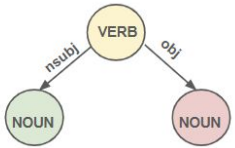
- **Gigafida** (1B): referenčni pisni
- **GOS** (2M): referenčni govorni
- **Trendi** (spremljevalni): novice
- **Šolar** (1M): šolski pisni izdelki
- **…**

Prenos in uporaba

- večina razčlenjenih korpusov je v formatu CONLL-U **prosto dostopnih** na repozitorijih CLARIN.SI, LINDAT/CLARIAH-CZ, GitHub
- **gradivni vir** v številnih jezikoslovnih in jezikovnotehnoloških raziskavah
  - preprost primer: variabilnost zaporedja osebka, predmeta in povedka (MA N. Hüll)

Korpusi:
SSJ, SST, JANES, Šolar, Šolar-GPT

Iskalni pogoj (drevo):



SVO: *mama kupuje jabolka*

SOV: *mama jabolka kupuje*

OVS: *jabolka kupuje mama'*

OSV: *jabolka mama kupuje*

VSO: *kupuje mama jabolka*

VOS: *kupuje jabolka mama*

	pisna	govorjena	spletna	šolska	generirana
SVO	**72,5**	**64,3**	**75,2**	**74,0**	**89,3**
SOV	6,1	10,7	5,9	6,5	3,8
OVS	14,2	10,7	9,3	8,7	4,5
OSV	1,7	2,4	2,4	2,2	0,8
VSO	3,8	9,5	4,5	7,9	1,4
VOS	1,7	2,4	2,9	0,7	0,2
entropija	0,94	1,16	0,93	0,93	0,48

Spletna orodja za analizo razčlenjenih korpusov

BRSKANJE PO KORPUSIH

“Kdaj stoji osebek za povedkom?”

- ❑ Drevesnik - manjši korpusi, vizualizacije
- ❑ CQP-Tree - večji korpusi na clarin.si/SkE

STATISTIČNA ANALIZA KORPUSOV

“Kateri tipi samostalniških zvez so najpogostejši?”

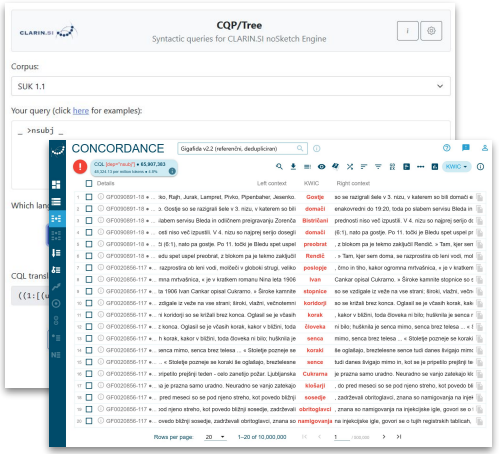
- ❑ STARK za induktivno luščenje struktur

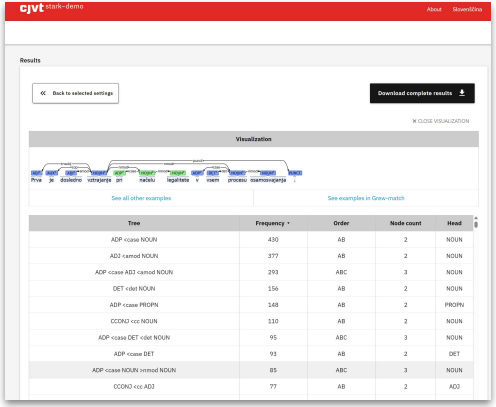
STROJNO OZNAČEVANJE BESEDIL

“Besedilu pripiši leme, besedne vrste in ...”

- ❑ Označevalnik (CLASSLA-Stanza)



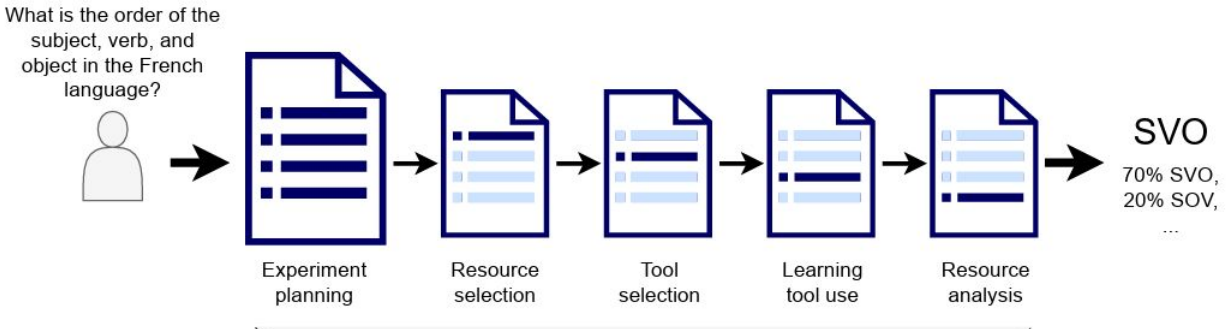






Nova generacija orodij: **veliki jezikovni modeli**

Tradicionalni pristop: jezikoslovec analizo in sintezo korpusnih podatkov izvede sam



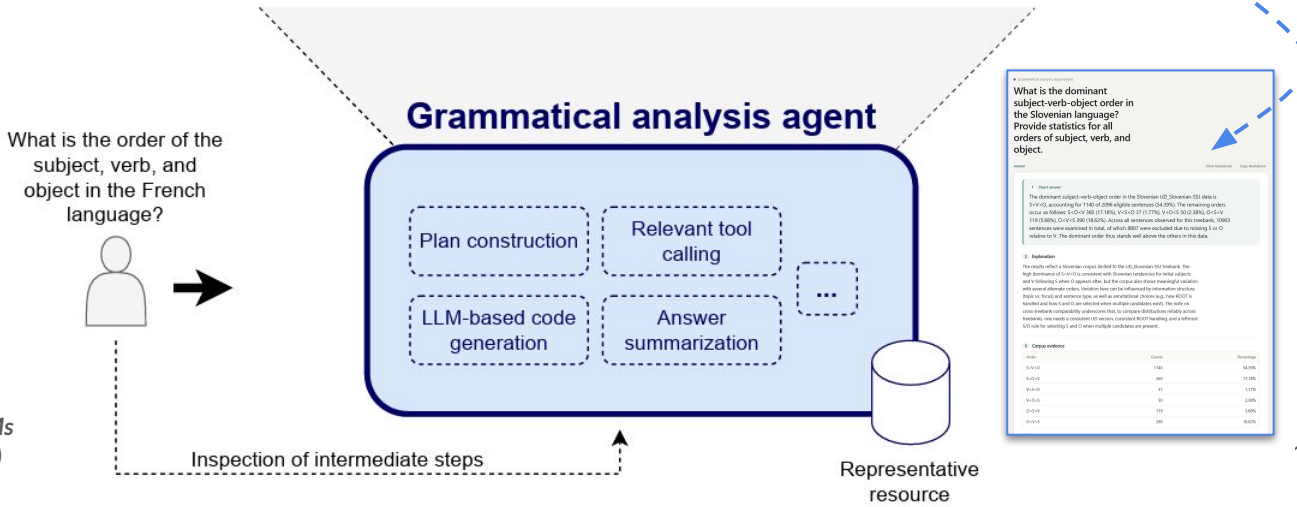
Naš predlog: korpusno analizo in sintezo opravi specializirani jezikovni model (**slovnični ‘agent’**)

Klemen et al. ***Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis*** (SLIDE 2026)

Nova generacija orodij: **veliki jezikovni modeli**

Uporabniku prijazna sinteza (v razvoju):
- **kratki odgovor** na vprašanje
- **razlaga** pojava z avtentičnimi **primeri**
- **povzetek gradiva** (tabele, grafi)
- **metoda** (opis postopka, koda)
- **dodatno branje** (sklici)

Naš predlog:
korpusno analizo in sintezo opravi specializirani jezikovni model (**slovnični 'agent'**)



Klemen et al. ***Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis*** (SLIDE 2026)

Ključni poudarki

1. Danes imamo za slovenščino na voljo številne skladenjsko označene **korpuse** ter **orodja** za njihovo kvalitativno in kvantitativno analizo.
   → Vsi so prosto dostopni in na voljo za nadaljnje izboljšave in prilagoditve specifičnim ciljem.
2. Z njihovo pomočjo lahko slovnični opis slovenščine **preverjamo** in **dopolnjujemo** z analizo dejanske jezikovne rabe.
3. Številne nove metodološke možnosti prinaša tudi **umetna inteligenca**, npr.
   - lažja in učinkovitejša analiza korpusnih podatkov
   - integracija obstoječih virov znanja in jezikovnih podatkov
     - dosedanje slovnice, članki, spletni viri ↔ korpusi (predpogoj: odprti dostop!)
   - interaktivna in uporabniku prilagojena slovnična ‘poglavja’

To infrastrukturo so skozi leta gradili številni sodelavci

Tjaša Arčon, Špela Arhar Holdt, Luka Bezgovšek, Kristina Bizjak, Živa Blaževič, Klara Canzutti, Lea Cibrič, Jaka Čibej, Kaja Dobrovoljc, Niklas Deworetzki, Tadeja Dušej, Tomaž Erjavec, Ivana Fekeža, Polona Gantar, Nanika Holz, Nives Hüll, Leon Noe Jovan, Urška Kamenšek, Matej Klemen, Iztok Kosem, Simon Krek, Luka Krsnik, Andreja Košir, Robert Kuret, Iztok Lebar Bajec, Nina Ledinek, Jakob Lenardič, Andrej Lovšin, Nikola Ljubešić, Boštjan Marhold, Arianna Masciolini, Nina Mikulin, Barbara Modrijan, Tina Munda, Sara Može, Joakim Nivre, Tanja Novak, Lea Peršič, Ina Poteko, Tanja Radovič, Rebeka Roblek, Marko Robnik-Šikonja, Peter Rupnik, Matija Škofljanec, Mihael Šinkec, Simona Šinkovec, Miha Štravs, Luka Terčon, Jerneja Umer, Gašper Uršič, Darinka Verdonik, Urška Vranjek, Daniel Zeman, Karolina Zgaga, Petra Žalodec … avtorji korpusov Gigafida in GOS ter drugi.

CLARIN.SI

AI4DH CENTRE OF EXCELLENCE IN AI FOR DIGITAL HUMANITIES

cjvt

Hvala za pozornost!

kaja.dobrovoljc@ijs.si
kaja.dobrovoljc@ff.uni-lj.si

Rezultati, predstavljeni v tej predstavitvi, so bili delno financirani s podporo Agencije RS za raziskovalno in inovacijsko dejavnost v okviru raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411) ter projektov *Veliki jezikovni modeli za digitalno humanistiko* (GC-002) in *Določanje sklonskega ustroja v slovenščini in drugih jezikih* (J6-70213). Aktivnosti je delno podprla tudi Evropska unija v okviru Centra odličnosti za umetno inteligenco v digitalni humanistiki (HORIZON-WIDERA-2023-TALENTS-01-01).

ARIS Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Funded by the European Union

Literatura in druge uporabne povezave

- Izbrani članki
  - De Marneffe et al. 2020: Universal Dependencies
  - N. Ledinek, 2014: Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine
  - N. Ledinek, 2015: Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi
  - K. Dobrovoljc et al. 2022: Universal Dependencies za slovenščino
  - K. Dobrovoljc 2024: Uporaba drevesnice SST v raziskavah govorjene slovenščine
- Spletni seminar o shemi UD (za začetnike):
  - https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:webinar-1
- Smernice za označevanje po shemi UD:
  - splošne (v angleščini): https://universaldependencies.org/
  - slovenske (s primeri): Universal Dependencies: Smernice za označevanje besedil v slovenščini.
- Aktualni raziskovalni projekti
  - SPOT (ARIS 2022-2025): spot.ff.uni-lj.si
  - LLM4DH (ARIS 2024-2027): https://www.cjvt.si/llm4dh/projektni-sklopi/projektni-sklop-2/#aktivnost-2.3