# Pretvorba korpusa ssj500k v **Univerzalno odvisnostno drevesnico za slovenščino**

**KAJA DOBROVOLJC**, ZAVOD ZA UPORABNO SLOVENISTIKO TROJINA

**TOMAŽ ERJAVEC**, ODSEK ZA TEHNOLOGIJE ZNANJA, INSTITUT "JOŽEF STEFAN"

**SIMON KREK**, LABORATORIJ ZA UMETNO INTELIGENCO, INSTITUT "JOŽEF STEFAN"

# Universal Dependencies

- pobuda za mednarodno **usklajeno skladenjsko razčlenjevanje besedil**
  - univerzalen nabor oblikoslovnih in skladenjskih oznak + smernice
  - možnost dodajanja oznak za jezikovno specifične pojave
- spodbujanje **napredka na področju procesiranja naravnih jezikov**
  - evalvacija obstoječih razčlenjevanih sistemov
  - razvoj večjezičnih razčlenjevalnih sistemov
  - medjezično učenje jezikovnih modelov
  - kontrastivne jezikoslovne analize
- rezultat drugih **predhodnih standardizacijskih projektov**
  - Stanford – skladenjske povezave (de Marneffe et al. 2014)
  - Google – besedne vrste (Petrov et al. 2012)
  - Interset – oblikoslovne lastnosti (Zeman 2008)

# Universal Dependencies v1.3

- 54 drevesnic
- 40 svetovnih jezikov
- 2 drevesnici za slovenščino
  - **drevesnica pisne slovenščine**
  - drevesnica govorjene slovenščine (Dobrovoljc in Nivre 2016)


# Skladenjsko razčlenjeni korpusi v slovenščini

1. **SDT** (Džeroski et al. 2005)
   - temelji na shemi **PDT**
   - 30.000 pojavnic (G. Orwell: 1984)
2. **jos100k** (Ledinek in Erjavec 2009)
   - temelji na shemi **JOS**
   - 100.000 pojavnic (vzorec FidaPLUS)
3. **ssj500k** (Krek et al. 2013)
   - temelji na shemi **JOS**
   - = jos100k + 400.000 novih pojavnic
   - 235.000 (skladenjsko razčlenjenih) pojavnic

**osnova za izdelavo drevesnice UD za slovenščino**

# Pretvorba ssj500k v UD za slovenščino

- v celoti avtomatizirana
- 3 zaporedni koraki
  - pretvorba formata (TEI → CONLLU)
  - pretvorba oblikoslovne ravni (JOS → UD)
  - pretvorba skladenjske ravni (JOS → UD)
- segmentacija, tokenizacija in lematizacija (za zdaj) nespremenjene

# **JOS** vs. **UD**: oblikoslovna raven

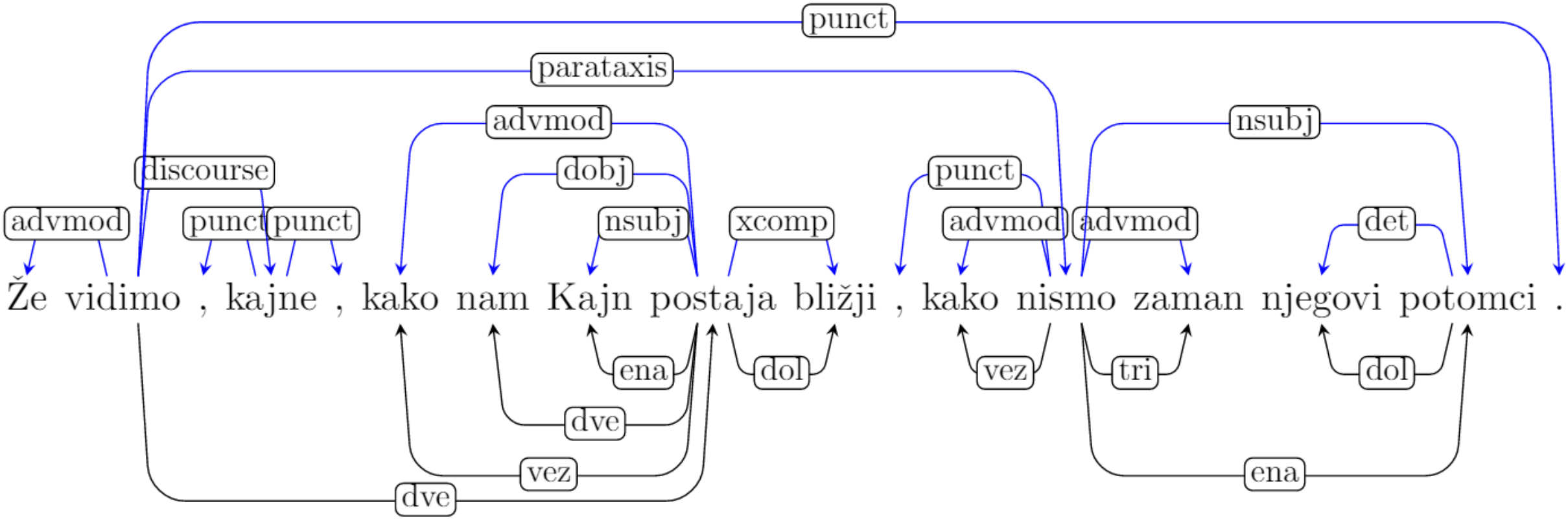
- sistema JOS in UD precej podobna
- **12** proti **17** univerzalnih besednih vrst
- posebnosti:
  - ločeno pomožni (**AUX**) in ostali (**VERB**) glagoli (**G**)
  - ločeno občna (**NOUN**) in lastna (**PROPN**) imena (**S**)
  - ločeno podredni (**SCONJ**) in priredni (**CONJ**) vezniki (**V**)
  - ločeno ločila (**PUNCT**) in simboli (**SYM**) (**U/<c>**)
  - okrajšava (**O**) kot del kategorije 'drugo' (**X**)
  - uvedba določilnikov (**DET**) (*moj, ta, enak; nekaj, toliko* ipd.)

# JOS vs. UD: oblikoslovna raven

- 22 (uporabljenih) oblikoskladenjskih lastnosti
  - 16/17 univerzalnih lastnosti (✘ `Voice`)
  - + 6 'jezikovnospecifičnih' lastnosti (`Gender[psor], Number[psor],Variant; Abbr; Foreign, NumForm`)
- nekaj novosti
  - drugačna kategorizacija glagolskih oblik (`Mood, Tense, VerbForm`), tudi pri ADJ in ADV
  - podrobnejše kategorije števnikov (`NumType=Card,Ord,Sets,Gen…`)
  - itd. (izčrpna dokumentacija že na spletu)

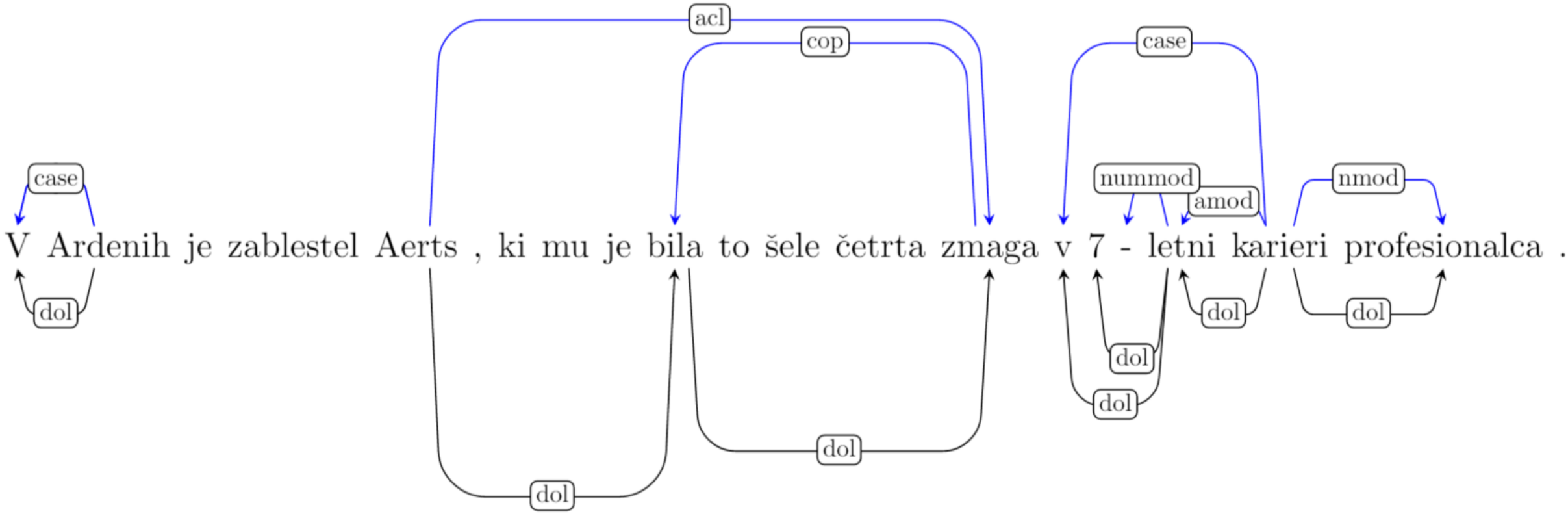
# JOS vs. UD: skladenjska raven

- sistema JOS in UD zelo različna, npr.:
  - ➢ daljši nabor tipov skladenjskih razmerij (**10** proti **40 oznak**)



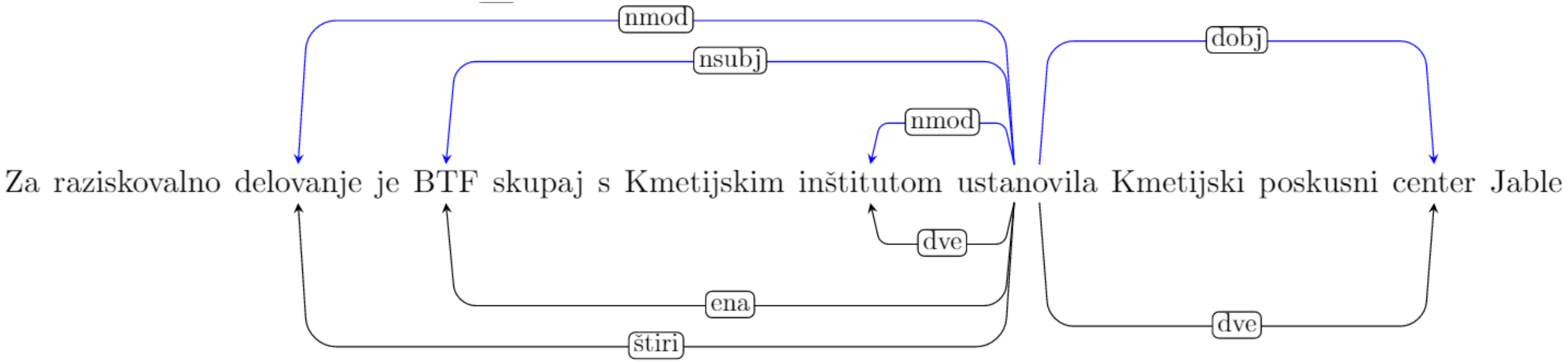
# JOS vs. UD: skladenjska raven

- podrobnejša skladenjska razmerja



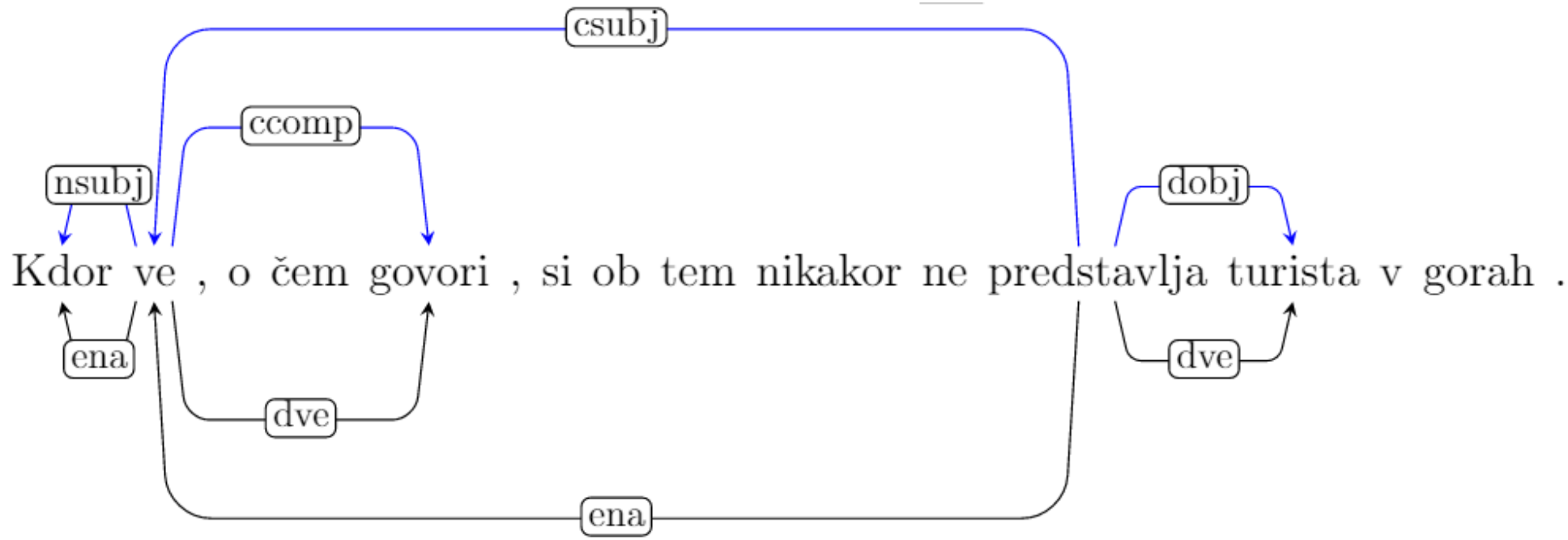
# JOS vs. UD: skladenjska raven

- razlikovanje med jedrnimi in ostalimi argumenti povedka



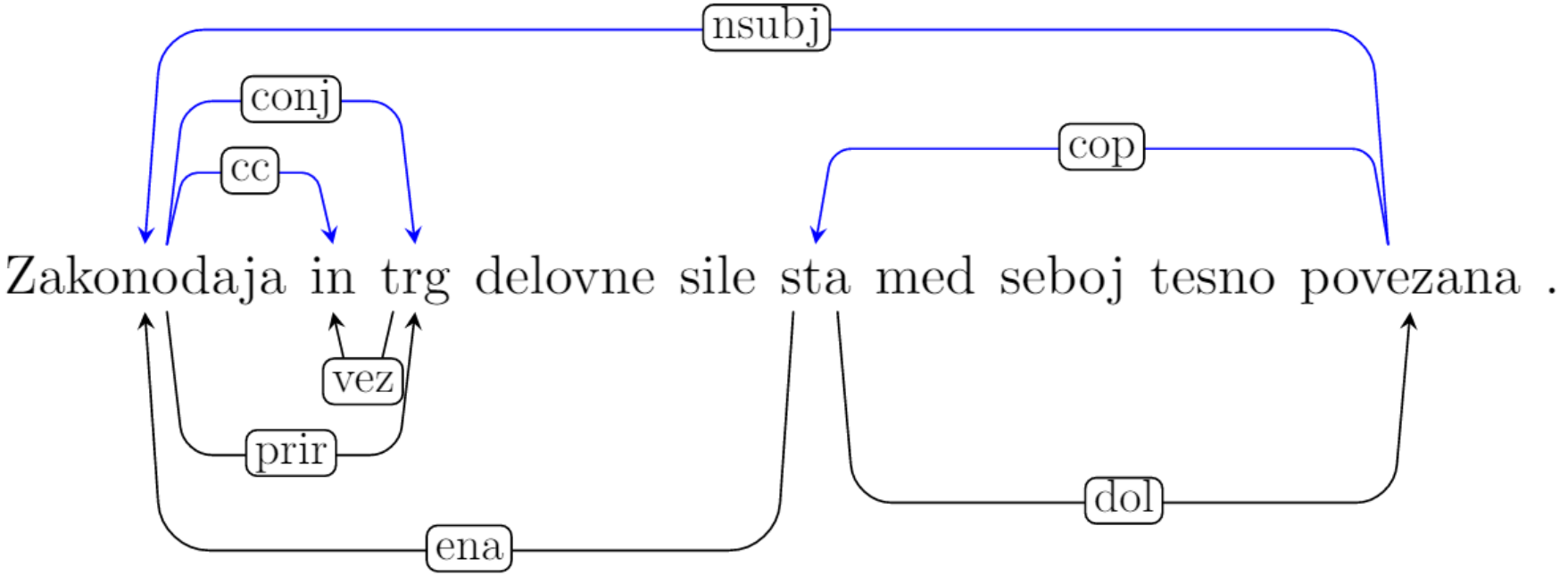
# JOS vs. UD: skladenjska raven

- razlikovanje med besednozveznimi in stavčnimi argumenti



# JOS vs. UD: skladenjska raven

- drugačna hierarhija (nekaterih) skladenjskih razmerij



# JOS vs. UD: skladenjska raven

- uporabljenih 31/40 univerzalnih povezav
  - ✗ neprepozna(v)ne ali neobstoječe
    - `nsubjpass, csubjpass, auxpass; list, dislocated, remnant, reparandum, vocative, goeswith`
  - 1 'jezikovnospecifična' povezava
    - `cc:preconj` (*tako* a *kot b*)
    - številne druge možnosti podoznak (cf. cs-ud)

# Pretvorba ssj500k v UD za slovenščino

- več sto pretvorbenih pravil na oblikoslovni in skladenjski ravni
  - kombinacije lastnosti pojavnice in konteksta
- izhodni format CONLL-U

vključene tudi oznake JOS

```
# sent_id ssj485.2607.9278
1   Joj          joj          INTJ    I          _                                                                                  5   discourse   _   SpaceAfter=No|Dep=0|Rel=Root
2   ,            ,            PUNCT   Z          _                                                                                  1   punct       _   Dep=0|Rel=Root
3   kako         kako         ADV     Rgp        Degree=Pos                                                                         5   advmod      _   Dep=4|Rel=Conj
4   sem          biti         VERB    Va-r1s-n   Mood=Ind|Negative=Pos|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin                 5   cop         _   Dep=0|Rel=Root
5   raztresen    raztresen    ADJ     Appmsnn    Case=Nom|Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing|VerbForm=Part             0   root        _   SpaceAfter=No|Dep=4|Rel=Atr
6   !            !            PUNCT   Z          _                                                                                  5   punct       _   Dep=0|Rel=Root
```

# Drevesnica UD za slovenščino (sl-ud.conllu)

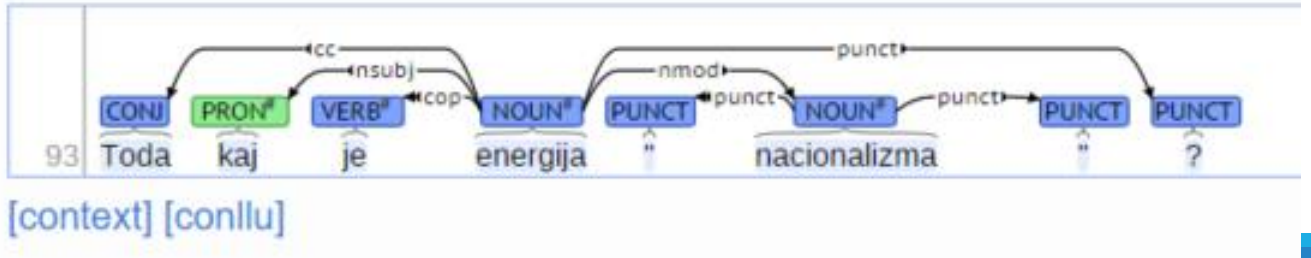
- natančnost ∝ omejena pokritost
  - manjša kot ssj500k
  - primerljiva z drugimi drevesnicami UD

| | **ssj500k** | **UD-sl** | **UD-hr** | **UD-eng** |
|---|---|---|---|---|
| pojavnic | 235.865 | 140.418 | 87.765 | 254.830 |
| stavkov | 11.411 | 7.996 | 3.957 | 16.622 |
| pojavnic/stavek | 20,7 | 17,6 | 22,2 | 15,3 |

# Drevesnica UD za slovenščino (sl-ud.conllu)

- del zbirke UD v1.2, UD v1.3 ...
  - prenos: http://hdl.handle.net/11234/1-1699


- vključena v razvoj številnih razčlenjeval
- vključena v različne spletne servise
  - brskanje po drevesnicah: SETS, PML Tree Query
  - označevanje neznanih besedil: UD Pipe




# UD za slovenščino: načrti

- izboljšava drevesnice
  - prehod na smernice v2 (CoNLL-ST 2017)
  - nova pravila za pretvorbo: manjkajoče povezave, preostanek ssj500k
  - ročni pregled
- izboljšava spletne dokumentacije
- vpliv spremembe sheme na referenčna označevalna orodja
  - razmislek o nadaljnjem vzdrževanju ločenih infrastruktur JOS-UD
- jezikovnotehnološke in (kontrastivne) jezikoslovne raziskave

Hvala. Vprašanja?

kaja.dobrovoljc@trojina.si

http://universaldependencies.org/

# Primer pravila: oblikoslovje

JOS: `msd=Gp.* & dep=del & head=G.*` → UD: `CPOS=AUX`

*JOS:* *Marko je*$_{\text{Gp}}$ *spal* vs. *Marko je*$_{\text{Gp}}$ *v Ameriki.*

*UD:* *Marko je*$_{\text{AUX}}$ *spal* vs. *Marko je*$_{\text{VERB}}$ *v Ameriki.*

# Primer pravila: skladnja

```
### cc: coordinating conjunction ##########################################
if cpostag(token) in ["CONJ", "PART", "X", "SCONJ"]:
    if jos_deprel(token) in ["Conj"]:
        if jos_head_jos_deprel(token) in ["Coord"]:
            token[6] = jos_head_jos_head(token)
            token[7] = "cc"

        elif cpostag(token) in ["CONJ"]:
            token[6] = jos_head(token) #verbs etc., but
            token[7] = "cc"
```