Jezikoslovne dileme pri strojnem procesiranju slovenščine

Simon Krek

Univerza v Ljubljani, Center za jezikovne vire in tehnologije

Institut "Jožef Stefan", Laboratorij za umetno inteligenco

Strojno procesiranje slovenščine

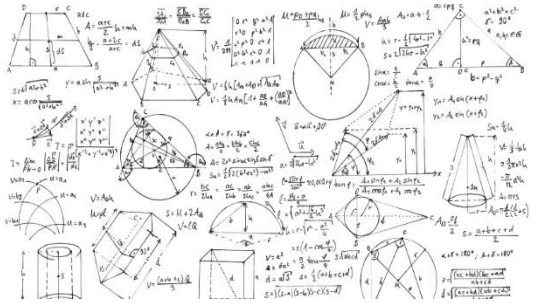
- pet minut za teorijo
- segmentacija in tokenizacija
- lematizacija
- oblikoskladenjsko označevanje
- skladenjsko razčlenjevanje
- označevanje semantičnih vlog

Repozitorij CLARIN.SI

- **Učni korpus ssj500k 2.2**
  - https://www.clarin.si/repository/xmlui/handle/11356/1210

- **Program za vizualizacijo Q-CAT**
  - https://www.clarin.si/repository/xmlui/handle/11356/1282

Ključne besede

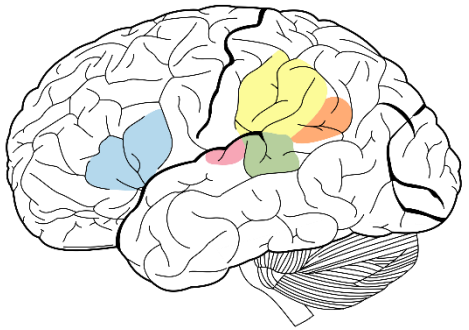


slovnica

jezikovni podatek

slovar

korpus

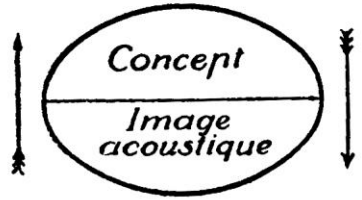


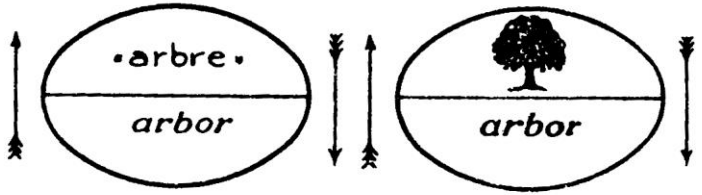
20. stoletje

- "/R/azlika med besedoslovnim in skladenjskim dejstvom lahko izgine /.../"
    - Ferdinand de Saussure [1916]
- "/I/skanje semantično zasnovane definicije 'gramatičnosti' bo jalovo /.../"
    - Noam Chomsky [1957]
- "Tradicionalno ločevanje med besediščem in slovnico je napačno."
    - Ray Jackendoff [2005]

Narava jezika – semiotika

- Da bi raziskovanje /.../ dobilo teoretsko osnovo, je ključna odločitev, kako resno vzeti vsako jezikovno dejanje kot pripadajoče in izhajajoče iz nekega sistema jezikovnih znakov, ki je namenjeno (pomenski) interpretaciji in tudi ima interpreta.

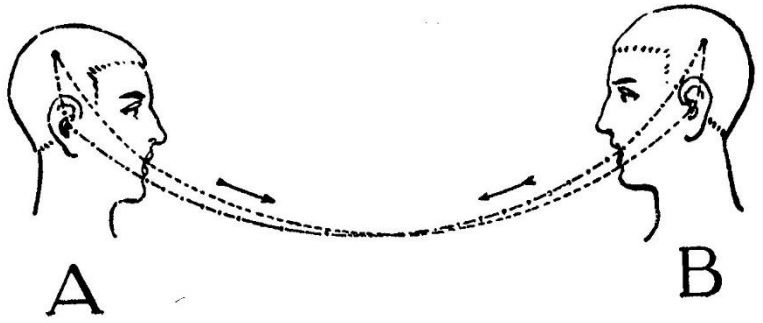




Ferdinand de Saussure: Cours de Linguistique Générale

Narava jezika – *langagague* (jezik. realizacije)

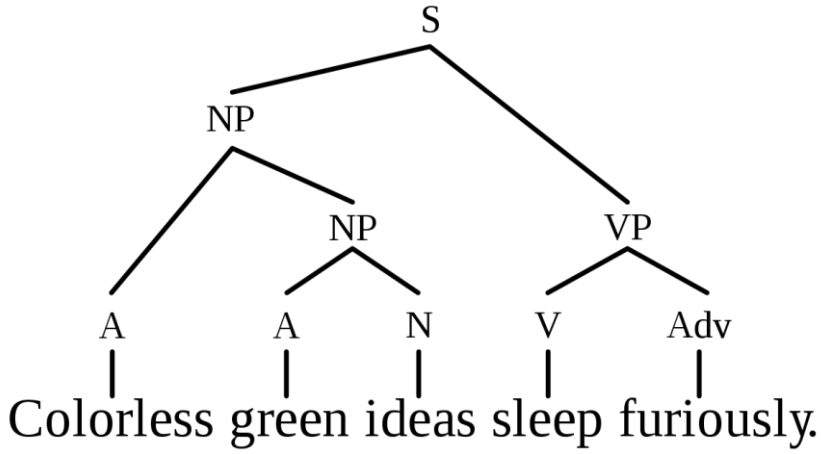
- Če lahko katerokoli instanco zavržemo kot neinterpretabilno s sistemskega stališča in s tem opredeljeno kot ne-jezik, ne-skladnja itd., smo se s tem a priori odpovedali možnosti teoretske refleksije jezika oz. govorice v saussurjevskem pomenu (tj. *langagague*).



Ferdinand de Saussure: Cours de Linguistique Générale

Narava jezika – možgani, iskalci pomena

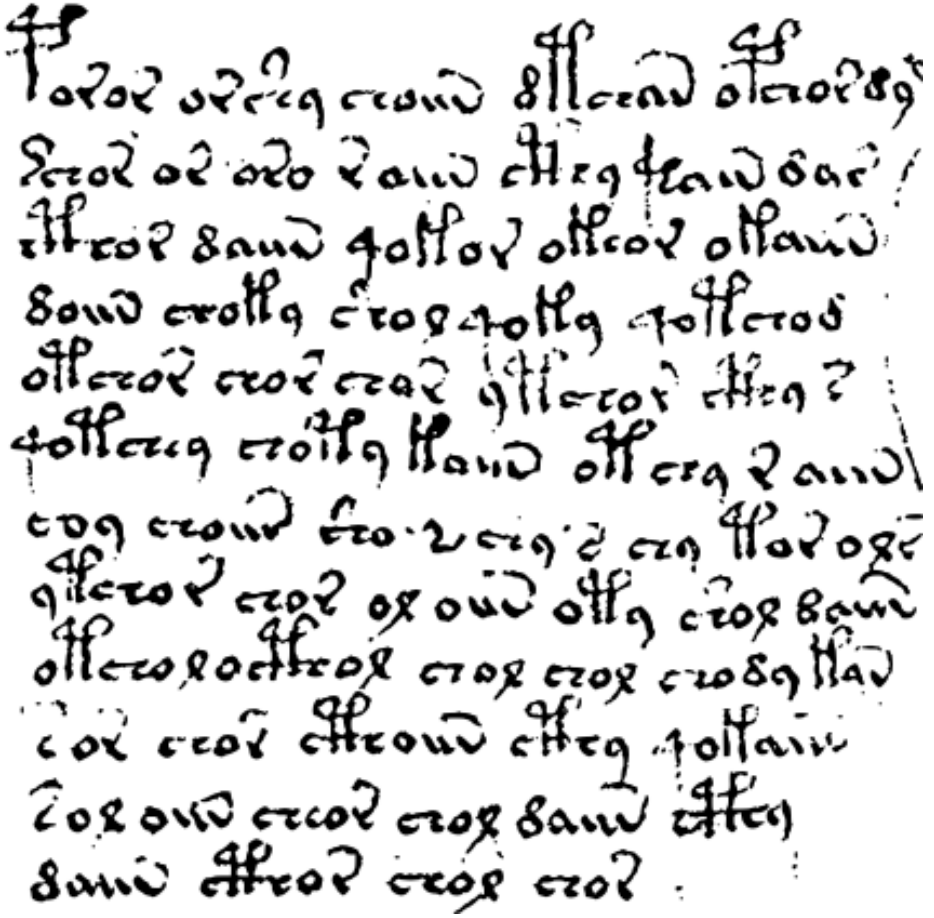
- Če pa pristanemo na holistični pristop, jezika ne moremo več dojemati kot inventar formalnih struktur, temveč kot vir ustvarjanja pomena, ki v končni fazi izhaja iz sistemskih vzorcev, ki jih opredeljuje izbirnost.
- Ti sistemski vzorci niso opredeljeni le kot izključujoča možnost ali nemožnost, temveč so interpretabilni na lestvici od posamezne instanciacije do takorekoč neskončnih ponovitev.



Noam Chomsky: Syntactic Structures

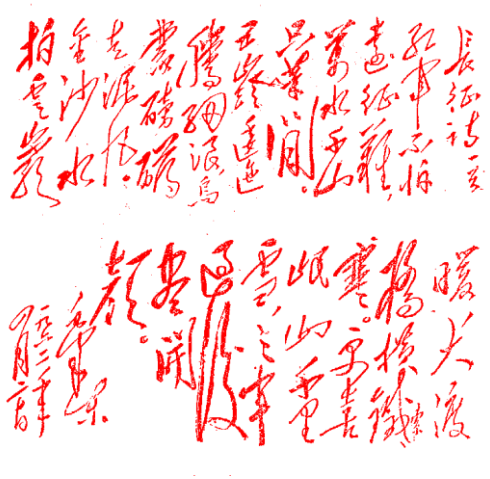
Narava jezika – interpretativni potencial

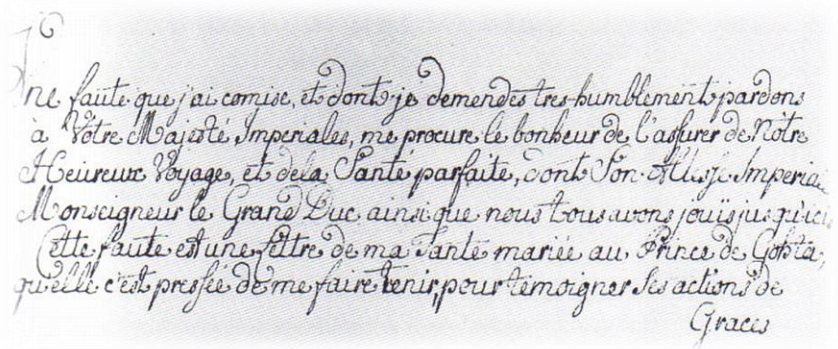
- Sistem izbir pa je družbeno pogojen le kot potencial, ki ga posamezni govorec lahko izkoristi za svoj – v osnovi komunikacijski – namen.
- Na ta način ni ovir za preverjanje tako rekoč katerekoli hipoteze jezikoslovne narave na empiričnem materialu – korpusu.



Vojničev rokopis (1404–1438)

Korpus – empirično izhodišče





Primorske novice – 17. januar 1997

- "Tistega večera sem preveč popil, zgodilo se je mesec dni po tem, ko sem izvedel, da me žena vara. Dogodek v Ankaranu je bila dramatična nesreča. Dekle je ob vzvratni vožnji začelo vpiti. Da bi jo utišal, sem prijel nož. Prišlo je do prerivanja in umrla je. Tega se sploh nisem zavedel. Kaj se je zgodilo, sem izvedel šele naslednjega dne iz časopisja," je v intervju za Stampo iz Torina izjavil morilec. Preiskave med sodnim postopkom so pokazale, da so se dogodki odvijali bistveno drugače. Dogodki v prihodnjih mesecih pa bodo pokazali, ali bo morilcu iz Ankarana tokrat uspelo prepričati italijanske pravosodne oblasti.

Tokenizacija in segmentacija

- celoten alfanumerični niz od prvega do zadnjega znaka v besedilu razdelimo na smiselne enote za potrebe računalniške obdelave in analize besedil
- preprost in učinkovit odgovor: (oboje)stičnost pri kombinacijah alfanumeričnih znakov in ločil
- slediti čim bolj intuitivnim in razumljivim pravilom, da se izognemo presenečenjem in nerazumevanju pri kasnejših obdelavah tokeniziranih besedil

Primorske novice – 17. januar 1997

- "Tistega večera sem preveč popil, zgodilo se je mesec dni po tem, ko sem izvedel, da me žena vara. Dogodek v Ankaranu je bila dramatična nesreča. Dekle je ob vzvratni vožnji začelo vpiti. Da bi jo utišal, sem prijel nož. Prišlo je do prerivanja in umrla je. Tega se sploh nisem zavedel. Kaj se je zgodilo, sem izvedel šele naslednjega dne iz časopisja," je v intervju za Stampo iz Torina izjavil morilec. Preiskave med sodnim postopkom so pokazale, da so se dogodki odvijali bistveno drugače. Dogodki v prihodnjih mesecih pa bodo pokazali, ali bo morilcu iz Ankarana tokrat uspelo prepričati italijanske pravosodne oblasti.

Primorske novice – 17. januar 1997

- "Tistega|večera|sem|preveč popil,|zgodilo|se|je|mesec|dni|po|tem,|ko|sem|izvedel,|da|me|žena|vara.|//Dogodek|v|Ankaranu|je|bila|dramatična|nesreča.|//Dekle|je|ob|vzvratni|vožnji|začelo|vpiti.|//Da bi|jo|utišal,|sem|prijel|nož.|// Prišlo|je|do|prerivanja|in|umrla je.|// Tega|se|sploh|nisem|zavedel.|//Kaj|se|je|zgodilo,|sem|izvedel|šele|naslednjega|dne|iz|časopisja,|je|v|intervju|za|Stampo|iz|Torina|izjavil|morilec.|//Preiskave|med|sodnim|postopkom|so|pokazale,|da|so|se|dogodki|odvijali|bistveno|drugače.|//Dogodki|v|prihodnjih|mesecih|pa|bodo pokazali,|ali|bo|morilcu|iz|Ankarana|tokrat|uspelo|prepričati|italijanske|pravosodne|oblasti. //

Revija Ekipa

- 9. etapa (Savigliano-Sestriere, 190 km): 1. Rujano (Ven/Selle Italia) 5;49:30, 2. Simoni (Ita, Lampre) + 0:26, 3. Di Luca (Ita, Liquigas) + 1:37, 4. Garate (Špa, Saunier) + 1:53, 5. Van Huffel (Bel, Davitamon-Lotto) + 1:55, 6. Gončar (Ukr, Domina Vacanze), 7. Savoldelli (Ita, Discovery Channel), 8. Valjavec (Slo, Phonak, vsi isti čas), 9. Cano (Kol, Davitamon-Lotto) + 2:38, 10. Sella (Ita, Panaria) + 5:07, 78. Gorazd Štangelj (Slo/Lampre) + 29:11, 145. Uroš Murn (Slo/Phonak) + 39:27: 20. etapa (Albese con Cassano - Milano, 119 km): 1. Petacchi (Ita, Fassa Bortolo) 3;24:08, 2. Zabel (Nem, T-Mobile), 3. Forster (Nem, Gerolsteiner), 4. Lorenzetto (Ita, Domina Vacanze), 5. Velo (Ita, Fassa Bortolo), 6. Grillo (Ita, Ceramica Panaria), 7. Lopez (Špa, Illes Balears), 8. Renshaw (Avs, FDJ), 9. Mori (Ita, Saunier Duval)

Projekt Jezikoslovno označevanje slovenščine (JOS) – učni korpus (100.000 pojavnic)

Znak	Opis	Število	Obojestično med poj.	V pojavnici	Obojestično v poj.
<c>,</c>	vejica	8933	5	108	108
<c>.</c>	pika	5354	0	653	91
<c>(</c>	oklepaj	644	0	0	0
<c>)</c>	zaklepaj	638	4	0	0
<c>-</c>	vezaj	549	196	88	77
<c>:</c>	dvopičje	419	0	88	88
<c>"</c>	narekovaj	410	0	0	0
<c>»</c>	narekovaj	318	0	0	0
<c>«</c>	zarekovaj	313	1	0	0
<c>?</c>	vprašaj	200	0	0	0
<c>;</c>	podpičje	119	0	0	0
<c>!</c>	klicaj	118	0	0	0
…	…	…	…	…	…

Vejica, dvopičje

- vejica/dvopičje postane del pojavnice, če je obojestična in če je niz znakov od presledka do presledka v celoti sestavljen iz števk

Dnevnik,časopis	dosegla prejšnji mesec na mitingu v Belgiji. Z	**1:55,19**	ima v lasti daleč najboljši letošnji rezultat
Delo Revije,revija	5 Panizzi (Fra, Peugeot 206 WRC) +	**2:41,6**	Touran je nastal na platformi, ki
drugo,strokovno	SH SYNOPSIS .B xmplay Nov 12	**21:82:82**	localhost cardmgr[l52]: socket 8: ATA/IDE
Delo,časopis	64: 52 (33:26), Turčija : Latvija 63:59 (	**34:24**	). 23 204 Ljubljana -
Delo,časopis	turnirju NLB za leto 1998 je zmagal Leon Mazi	**8,5**	pred Šifrerjem, Podlesnikom in Mohrom po

Opuščaj

- opuščaj je vedno del pojavnice, če je rabljen obojestično, levo- ali desnostično

Krtina,strokovno	ruski biolog Alexey Kondrashov, v knjigi	**Mendel's**	demon: gene justice and the complexity
Dnevnik,časopis	petični zdraviliški gosti segali po vrhunskem	**slat'nčanu**	in kakšno penino afrodito kupili tudi kot
Dnevnik,časopis	po najbolj znanem prodajalcu hitre hrane	**McDonald's**	, sta že devetič družno pripravila liga
Dnevnik,časopis	Exit, Nothing is Real but the Girl, The	**Dream's**	Lost on Me, Under the Gun ter Maria, ki
Dnevnik,časopis	Grande Motte nad Tignesom v bližini Val	**d'Isera**	v Franciji. Ker v sezoni 1999-2000 za alpske
drugo,revija	vedno imamo stare fane, ki so z nami vse od	**Gast'r'bajtr'sov**	, je pa še vedno na vsakem koncertu kup
24ur.com,internet	jasamtaj pa kaj si ti	**f'u'k'nj'e'n**	slovenija je znana po svetu

Vezaj realno

DZS,strokovno	RAZVEJENOST OGLJIKOVODIKOVE VERIGE V MOLEKULAH	**2-METILPROPAN-1-OLA**	ZARADI STERIČNIH OVIR ZMANJŠUJE MOŽNOST
DZS,strokovno	vezjo CC (alken). Število točk10-	**98-76-54-32**	-0Predlagana ocena54321 Preglednica
DZS,strokovno	vezjo CC (alken). Število točk10-98-	**76-54-32-0Predlagana**	ocena54321 Preglednica rezultatov
DZS,strokovno	lahko nevtraliziramo z reakcijo s suhim	**2-metilpropan-2-olom**	(glej 16. pogl.). ??????????????
drugo,leposlovno	Ljubljana : UMco, 2008. (Zbirka Suspenz) ISBN	**978-961-6445-82**	-5 239115776 Andreju v spomin.
drugo,leposlovno	: UMco, 2008. (Zbirka Suspenz) ISBN 978-	**961-6445-82-5**	239115776 Andreju v spomin.
Adria Media,revija	Cosmo, 10/08, p. 143, by Victoria Lucia,	**Blow-his-mind-tip**	(Describe a sexy dream) Opiši mu

Vezaj

- osnovna težnja podobna kot pri obravnavi presledkov – ločenost na več pojavnic, kjer je to mogoče
- posamezna pojavnica naj bo čim bolj samozadosten (leksikalni) element – v korpusu (ali oblikoslovnem leksikonu) se pojavlja kot samostojna enota
- nesamozadostni oz. leksikalno nesamostojni element naj nosi informacije, ki so pomembne bodisi za
    - kasnejše pridobivanje jezikoslovnih informacij iz korpusa ali
    - za izvajanje enega od procesov jezikoslovnega označevanja korpusov
- pogostejši tipi rabe vezaja so pri uveljavljanju teh načel pomembnejši od manj pogostih

Rezultat

- a. zaznamovanje nesamostojnih pomenskih delov besede
  - predpon (pre-, pri-, pra-), končnic (-a, -e, -i), priponskih obrazil (-ost, -oba, -ota, -oča), medpon (-o-, -e-), vpon (-k-, npr. ti-k-ati) ipd.;
- b. vezaj med sestavinami zloženk, nastalih iz podredne zveze
  - 25-letnica, 4-urna (seja), 48-kilometrska (proga), 12-kratna (premoč), 100-odstoten; C-vitaminski, B-diplomski (izpit), C-dur, C-vitamin, TV-drama, c-mol; tako tudi 14-krat
- c. za naveznim členkom le-
- d. med kratičnim imenom in končnico
- e. med črkovno ali številčno podstavo ali osnovo in njunim končajem

drugo,časopis	z ambientalnimi odisejadami, zbranimi na	**LP-ju**	»Refused«, ki ga lahko od prejšnjega ponedeljka
drugo,revija	mogoč dostop do seznama tistih, ki so s	**TunA-jem**	nekje v bližnji okolici. Sporočila prebiramo
drugo,revija	enostavno. Prijaviš se na spletnih straneh	**CORDISS-a**	, izpolniš nekaj obrazcev in že postaneš
drugo,revija	klasičen in znan primer spin-offa navaja	**IBM-ovo**	podjetje za proizvodnjo osebnih računalnikov
drugo,revija	* * Z odpravo prepovedi izvoza orožja iz	**EU-ja**	na Kitajsko se odpira fronta trgovinske
drugo,časopis	lestvice. Vlado Nachbar, oče slovenskega	**NBA-jevca**	Boštjana Nachbarja, ki je funkcijo direktorja
drugo,časopis	– nova politika je le nova krinka starih	**LDS-ovcev**	, ki hočejo imeti pod svojim okriljem samo
drugo,časopis	bi prav prišel reduktor. A ker lastniki	**SUV-ov**	svoje jeklene konjičke večkrat razkazujejo
Mladinska knjiga,leposlovno	Sašev rokopis z ozkimi, zašiljenimi nemškimi	**E-ji**	in R-ji, z odločnimi potezami, ki določajo
Mladinska knjiga,leposlovno	rokopis z ozkimi, zašiljenimi nemškimi E-ji in	**R-ji**	, z odločnimi potezami, ki določajo moža

Pika

- a. pika je del pojavnice:
  - če je obojestična in niz vsebuje le števke
  - če je obojestična in gre za spletni naslov ali naslov elektronske pošte
  - če je levostična in gre za okrajšavo, vrstilni številnik ali zaporednostni prislov
  - če se stavek konča z okrajšavo, vrstilnim številnikom ali zaporednostnim prislovom
- b. večdelne okrajšave se delijo na posamezne pojavnice ne glede na (desno)stičnost.

drugo,časopis	zgodovinarska dr. Jera Vodušek Starič in v dosjeju	**ing.**	Mačkovšek št. 7 na strani 121 zasledila
drugo,časopis	Vodušek Starič in v dosjeju ing. Mačkovšek	**št.**	7 na strani 121 zasledila, kdo je glavni
drugo,časopis	novoizvoljenemu predsedniku, svetovljanu	**dr.**	Danilu Türku in predsedniku vlade Janezu
drugo,časopis	prihodnosti. Revija Mladina je 20. 10. 2003 v	**št.**	42 na strani 6 sklenila prispevek o »kraljevi
drugo,časopis	6 sklenila prispevek o »kraljevi vojski	**oz.**	slovenskih četnikih v Grčaricah: »… na
drugo,časopis	nedvomno dobro spozna. Foto: Atka,	**d.**	o. o, Gosposka ulica 13, Celje Lastnik
drugo,časopis	nedvomno dobro spozna. Foto: Atka, d.	**o.**	o, Gosposka ulica 13, Celje Lastnik
drugo,časopis	izvira iz časa obnove te nepremičnine v	**90.**	letih 20. stoletja, in jo tudi obravnava
drugo,časopis	objavljen niz fotografij izkopa in pogreba	**t.**	i. bizoviških žrtev iz maja 1942. 17. maja
drugo,časopis	objavljen niz fotografij izkopa in pogreba t.	**i.**	bizoviških žrtev iz maja 1942. 17. maja

Tokenizacijska pravila

- 1. korak: prepoznavanje in označevanje ločil in simbolov
- 2. korak: označevanje pojavnic na podlagi alfanumeričnega niza med presledki
- 3. korak: združevanje spletnih naslovov in naslovov elektronske pošte
- 4. korak: združevanje (akronimov, številk itd.) s stičnimi vezaji
- 5. korak: združevanje vejic, pik in dvopičij v številskih nizih in združevanje okrajšav
- 6. korak: segmentacija

```
<text><body>
    <div xml:id="F0028708" n="1 8 103 122">
        <p xml:id="F0028708.9" n="8 103 122">
            <s xml:id="F0028708.9.1" n="19 24">
      <c xml:id="F0028708.9.1.1">                    “          </c>
      <w xml:id="F0028708.9.1.2">                    Tistega    </w><S/>
      <w xml:id="F0028708.9.1.3">                    večera     </w><S/>
      <w xml:id="F0028708.9.1.4">                    sem        </w><S/>
      <w xml:id="F0028708.9.1.5">                    preveč     </w><S/>
      <w xml:id="F0028708.9.1.6">                    popil      </w>
      <c xml:id="F0028708.9.1.7">                    ,          </c><S/>
      <w xml:id="F0028708.9.1.8">                    zgodilo    </w><S/>
      <w xml:id="F0028708.9.1.9">                    se         </w><S/>
      <w xml:id="F0028708.9.1.10">                   je         </w><S/>
      <w xml:id="F0028708.9.1.11">                   mesec      </w><S/>
      <w xml:id="F0028708.9.1.12">                   dni        </w><S/>
      <w xml:id="F0028708.9.1.13">                   po         </w><S/>
      <w xml:id="F0028708.9.1.14">                   tem        </w>
      <c xml:id="F0028708.9.1.15">                   ,          </c><S/>
      <w xml:id="F0028708.9.1.16">                   ko         </w><S/>
      <w xml:id="F0028708.9.1.17">                   sem        </w><S/>
      <w xml:id="F0028708.9.1.18">                   izvedel    </w>
      <c xml:id="F0028708.9.1.19">                   ,          </c><S/>
      <w xml:id="F0028708.9.1.20">                   da         </w><S/>
      <w xml:id="F0028708.9.1.21">                   me         </w><S/>
      <w xml:id="F0028708.9.1.22">                   žena       </w><S/>
      <w xml:id="F0028708.9.1.23">                   vara       </w>
      <c xml:id="F0028708.9.1.24">                   .          </c><S/>
```

Lematizacija

- proces pripisovanja osnovne oblike korpusnim pojavnicam pri tistih besednih vrstah, ki so pregibne in tvorijo oblikoslovno paradigmo
- ločevanje med oblikoslovnimi in besedotvornimi paradigmami
    - npr. stopnjevanje prislovov kot besedotvorna izpeljava iz pridevniških oblik primernika in presežnika
- zapis (velike/male črke) kot inherentna lastnost leme

Strojna lematizacija

- Strojni lematizatorji:
  1. lematizator podjetja Amebis, d.o.o., ki je integralni del slovničnega analizatorja
  2. lematizator spletnega servisa ToTaLe, razvit v okviru projekta "Jezikoslovno označevanje slovenščine" (Tomaž Erjavec, Sašo Džeroski)
  3. lematizator LemmaGen, razvit na IJS (Matjaž Juršič)
  4. lematizator (izvor: LemmaGen) kot del označevalnika Obeliks, razvit v okviru projekta „Sporazumevanje v slovenskem jeziku“ (Miha Grčar)
  5. lematizator Reldi kot del označevalnika Reldi, razvit v okviru projekta Reldi (Nikola Ljubešić)
  6. lematizator **StanfordNLP (CLARIN.SI)**
- Strojno berljiv leksikon besednih oblik:
  - Multext-East – 15.000 osnovnih oblik
  - Sloleks – 100.000 osnovnih oblik

Analiza (JOS)

- **razdvoumljanje**: napačne odločitve, kjer se je strojni mehanizem odločal med več možnimi lemami pri pojavnicah, ki imajo pogoste skupne leme
  - delo-del, grafik-grafika, kraj-kraja, predlog-predloga, premier-premiera, ..
- **ugibanje**: napačne odločitve, kjer ne gre za izbiro med več znanimi lemami, temveč za proces ugibanja nove leme na podlagi razgradnje pojavnice na komponente
  - ja (jama), dekleto (dekle), fitno (fitnes), izdelk (izdelek), kotičk (kotiček), magisterje (magisterij), me (menih), ospred (ospredje), bonaec (bonaca), desetvaljnika (desetvaljnik), ...
- **tujejezično:** citatno rabljene besede iz tujih jezikov, pri katerih v slovenščini bodisi uporabimo sklanjatvene paradigme ali pa pojavnice ostanejo v citatni obliki
  - aga (age), dahlio (dahlia), enterpris (enterprise), alpina (alpine), chica (chic), ...
- **vprašanje**: avtentično jezikoslovno vprašanje, kaj pri takih pojavnicah dejansko pričakujemo kot osnovno obliko

Ali je lematizacija lahko neodvisen postopek?

- Osnovna oblika pri samostalnikih
  - lastna imena
    - velika/mala začetnica
      - *trenutno najstarejši kaveljc je 80-letni Boris **Mlinar [Mlinar?]***
      - *Nastale so klasike Rozmarijin otrok, Izganjalec duhov in slasher **Noč** čarovnic. [**noč?**]*
    - spol
      - ***Saša** ni doma. **[Saša/Sašo]***
    - paradigma v srednjem ali ženskem spolu
      - *Razstava bo letos gostovala še v Mehiki, ZDA, **Japonski [Japonska]***
      - *Po avanturi na **Japonskem** ste ponovno nosili majico Maribora. **[Japonsko]***
  - posamostaljeni pridevniki
    - brezposelni [brezposeln], zaposleni [zaposlen], dežurni [dežuren]

Samostalnik

- prevladujoče množinske oblike, parni organi itd.
  - oči [oko?], očesa [oko?], ustnice [ustnica?], testenina-testenine, toplica-toplice, oblica-oblice, smet-smeti, spomin-spomini
- variantne osnovne oblike in velik del paradigme podoben
  - cigaret-cigareta, copat-copata

ŽENSKI SPOL	ednina	dvojina	množina
imenovalnik	cigareta	cigareti	cigarete
rodilnik	cigarete	cigaret	cigaret
dajalnik	cigareti	cigaretama	cigaretam
tožilnik	cigareto	cigareti	cigarete
mestnik	cigareti	cigaretah	cigaretah
orodnik	cigareto	cigaretama	cigaretami

MOŠKI SPOL	ednina	dvojina	množina
imenovalnik	cigaret	cigareta	cigareti
rodilnik	cigareta	cigaretov	cigaretov
dajalnik	cigaretu	cigaretoma	cigaretom
tožilnik	cigaret	cigareta	cigarete
mestnik	cigaretu	cigaretih	cigaretih
orodnik	cigaretom	cigaretoma	cigareti

Glagol

- Pričakujemo nedoločnik (25 oblik):
  - sedanjik, gl. oseba, gl. število
  - opisni deležnik na -l, sestavljeni glagolski časi, gl. spol, gl. število
  - velelnik, gl. oseba (prva, druga), gl. število (brez 1. os. ed.)
  - namenilnik
  - nedoločnik
- Ostale oblike, ki jih še povezujemo z glagolom:
  - deležja (-e, -aje, -ši) --> oblika = lema
  - deležnik na -č/-ši --> pridevniška lema
  - deležnik stanja na -l/-n/-t) --> pridevniška lema
  - glagolnik --> samostalniška lema

Pridevnik

- Ali želimo dati prednost tradicionalni slovarski moški obliki pridevnika in združevati veliko število oblik pod eno samo lemo?
  - 167 oblik za eno pridevniško lemo
    - spol (18 x 3 = 54)
    - tri stopnje = skupaj 162 oblik
    - osnovnik, im.|tož. m. sp. -> določna/nedoločna oblika (+2)
    - vse stopnje v tož. ed. m. sp. -> obliko za živost (+3)
- Alternativa: pripis osnovne oblike glede na
  - spol
  - stopnjo
- Elativ?
  - predober, preslab

Drugo

- zaimek
  - pregiben, zahteven, zaprta lista
- števnik
  - pregiben, nezahteven, (pol)odprta besedna vrsta
- prislov
  - stopnjevanje prislovov
  - besedotvorni proces iz ustrezne stopnje pridevnika?
    - oster/ostro, ostrejši/ostreje, najostrejši/najostreje

- <text><body> <p xml:id="F0028708.9" n="8 103 122"><s xml:id="F0028708.9.1" n="19 24">

<c xml:id="F0028708.9.1.1">	“	</c>
<w xml:id="F0028708.9.1.2" lemma="tisti">	Tistega	</w><S/>
<w xml:id="F0028708.9.1.3" lemma="večer">	večera	</w><S/>
<w xml:id="F0028708.9.1.4" lemma="biti">	sem	</w><S/>
<w xml:id="F0028708.9.1.5" lemma="preveč">	preveč	</w><S/>
<w xml:id="F0028708.9.1.6" lemma="popiti">	popil	</w>
<c xml:id="F0028708.9.1.7">	,	</c><S/>
<w xml:id="F0028708.9.1.8" lemma="zgoditi">	zgodilo	</w><S/>
<w xml:id="F0028708.9.1.9" lemma="se">	se	</w><S/>
<w xml:id="F0028708.9.1.10" lemma="biti">	je	</w><S/>
<w xml:id="F0028708.9.1.11" lemma="mesec">	mesec	</w><S/>
<w xml:id="F0028708.9.1.12" lemma="dan">	dni	</w><S/>
<w xml:id="F0028708.9.1.13" lemma="po">	po	</w><S/>
<w xml:id="F0028708.9.1.14" lemma="ta">	tem	</w>
<c xml:id="F0028708.9.1.15">	,	</c><S/>
<w xml:id="F0028708.9.1.16" lemma="ko">	ko	</w><S/>
<w xml:id="F0028708.9.1.17" lemma="biti">	sem	</w><S/>
<w xml:id="F0028708.9.1.18" lemma="izvedeti">	izvedel	</w>
<c xml:id="F0028708.9.1.19">	,	</c><S/>
<w xml:id="F0028708.9.1.20" lemma="da">	da	</w><S/>
<w xml:id="F0028708.9.1.21" lemma="jaz">	me	</w><S/>
<w xml:id="F0028708.9.1.22" lemma="žena">	žena	</w><S/>
<w xml:id="F0028708.9.1.23" lemma="varati">	vara	</w>
<c xml:id="F0028708.9.1.24">	.	</c><S/>

Oblikoskladenjsko označevanje

- Pripisovanje oblikoskladenjskih oznak (*POS-tagging*, *part-of-speech tagging*, *word-class tagging*), je ena od najstarejših in najpogostejših oblik avtomatiziranega dodajanja interpretativnih informacij jezikoslovne narave besedilom, pri čemer posamezni pojavnici v besedilu pripišemo, v kateri osnovni **besednovrstni razred** spada v specifičnem jezku ter v nekaterih primerih tudi **lastnosti**, ki jih izkazuje znotraj razreda.

Klasifikacija / kategorizacija

- klasifikacija in kategorizacija sta različna koncepta
- klasifikacija je pripisovanje objektov vnaprej definiranim razredom
- kategorizacija je začetna identifikacija teh razredov in se torej mora zgoditi pred klasifikacijo

Naloga

- **vse besede** nekega jezika razdeliti na razrede po izbranih kriterijih

- Kaj je beseda?
- Kaj je razred?
- Kaj je kriterij?

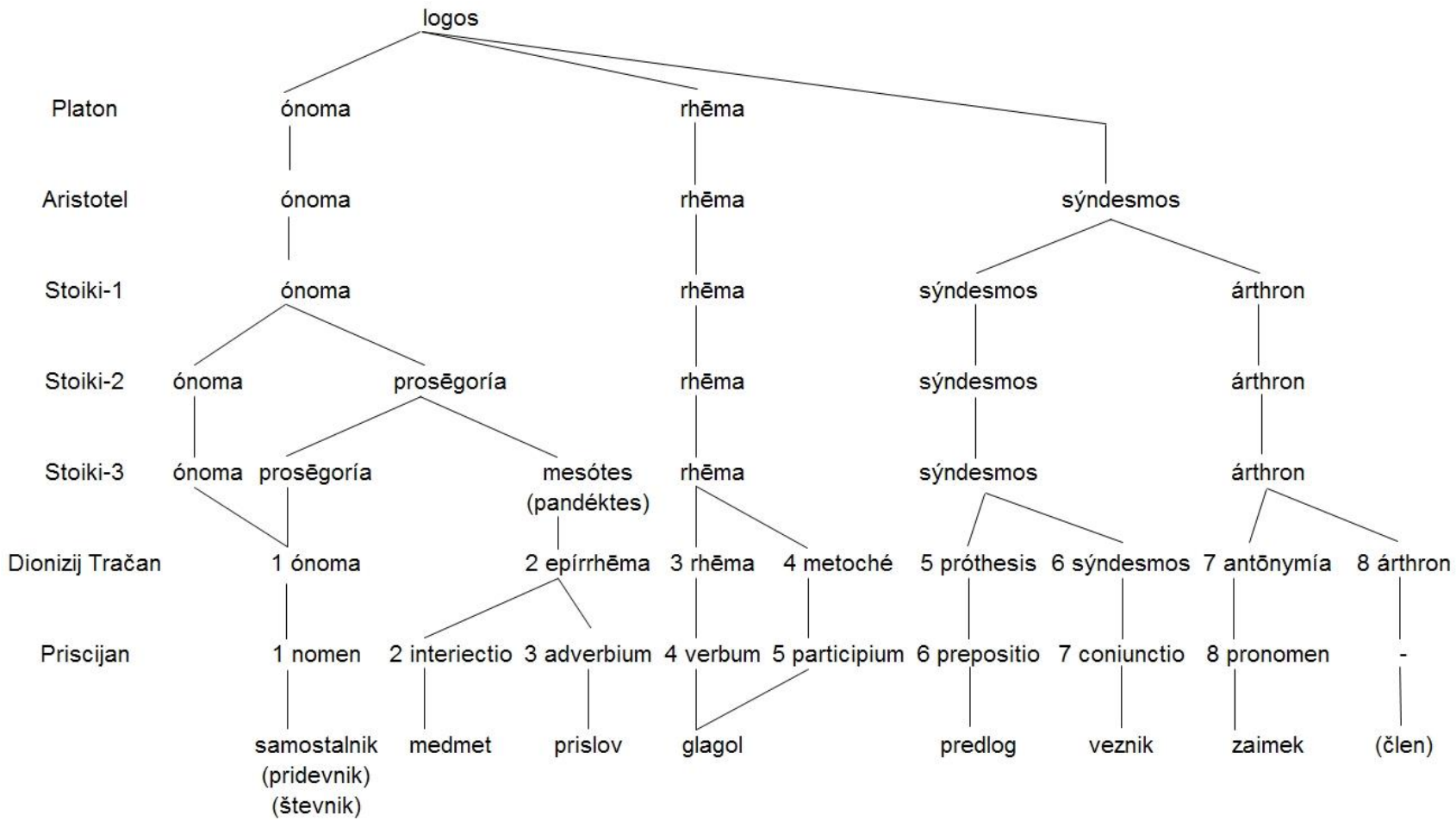
Sapir (1921), Bloomfield (1933), Crystal (1967)

- Na žalost, ali na srečo, noben jezik ni tiransko konsistenten. Vse slovnice puščajo.
    - Edward Sapir: Language: An Introduction to the Study of Speech. 1921.
- /.../ nemogoče je vzpostaviti povsem konsistentno shemo besednih vrst, ker se razredi besed prekrivajo in križajo.
    - Leonard Bloomfield: Language. 1933.
- /.../ zaključiti moramo, da so lahko besedne vrste ozke ali široke, kakor narekuje posamezna situacija, in da nobena klasifikacija ni absolutno boljša od druge /.../ različni jezikoslovci bodo za različne namene izdelali bolj ali manj detajlne klasifikacije.
    - David Crystal: English. *Lingua 17*. 1967.

Manning in Schütze: 1999

- Besedna vrsta je pravzaprav kompleksen pojem, kajti motiviran je z različnih izhodišč, npr. s semantičnega (imenovanega tudi pojmovno), distribucijskega skladenjskega ali oblikoslovnega. Takšna pojmovanja besednih vrst so pogosto v konfliktu.
    - Manning in Schütze: Foundations of statistical natural language processing. 1999.

Malce zgodovine



Slovenščina

- (slovnična) kategorizacija:
  - Adam Bohorič (1584)
  - Marko Pohlin (1768/1783)
  - Jernej Kopitar (1808/9)
  - Anton Janežič (1854)
  - Anton Breznik (1916)
  - Anton Bajec, Rudolf Kolarič, Mirko Rupel (1956/1964)
  - Jože Toporišič (1976/2000)
- (slovarska) klasifikacija:
  - Pleteršnik: Slovensko-nemški slovar (1894/5)
  - Slovenski pravopis (1899 | 1920 | 1935 | 1950 | 1962 | 2001)
  - Slovar slovenskega knjižnega jezika (1970/1991)
  - Slovenski pravopis (2001)
  - Slovar slovenskega knjižnega jezika (2014)

Besednovrstne kategorije

	Kopitar (1808/9) Janežič (1854) Breznik (1912) Bajec, Kolarič, Rupel (1956)	Toporišič (2000)
1	samostalnik	samostalniška beseda
2	pridevnik	pridevniška beseda
3	zaimek	povedkovnik
4	številnik	členek
5	glagol	glagol
6	prislov	prislov
7	predlog	predlog
8	veznik	veznik
9	medmet	medmet

Jože Toporišič

- Besedne vrste so v tej knjigi obravnavane kot pojmi za množice besed z enakimi skladenjskimi vlogami in drugimi lastnostmi (npr. tvorjenost, slovnične kategorije, konverznost ipd.). Po tej teoriji je v slovenskem knjižnem jeziku 9 besednih vrst;
    - Jože Toporišič, Slovenska slovnica. 2000.

- /.../ težav in nedoslednosti ter nepopolnosti pa je rešena teorija besednih vrst, ki se opira na skladenjska merila. Taka teorija besedne vrste določa izključno in enotno le po skladenjskih načelih /.../
    - Jože Toporišič, Oblikoslovne razprave. 2003.

Kriteriji


distribucijski
- skladenjski
predlog
veznik
prislov
zaimek
samostalnik
pridevnik
glagol
števnik
pomenski
oblikovni

Kriteriji


distribucijski
- skladenjski
povedkovnik
členek
pomenski
oblikovni

Povedkovnik

- /.../ povedkovniki oz. povedkovi dopolnilniki niso nič drugega kot pomenske determinante povedkov, zato ostajajo na stavčnočlenski ravni (priložnostna pomenskoskladenjska raba ne more biti besednovrstno odločilna).
    - Andreja Žele, Slovarska obravnava povedkovnika. 2004.
- SSKJ I (1970/1991)
    - 0: »v povedni rabi« (496) ali »v povedno-prislovni rabi« (54): čudno (je), dolgočasno (je), dvomljivo (je), čudež (je), greh (je) itd.
- Slovenski pravopis (2001)
    - 467: (biti) pes, (biti) hrupno, (biti) lepa, (biti) kvit, (biti) proti itd.
- SSKJ II (2014)
    - 10: bòt, kós, kvít, predôlgčas, preškóda, pretemà, prerés, premràz, tréba, všéč

Členek

- SSKJ 1 (1970/1991)
  - 7
  - naj, si, ga, bodi, koli, le, bi
- Slovenski pravopis (2001)
  - 180 (206)
- SSKJ 2 (2014)
  - 193

Členek in SSKJ 2

izt	zgled	SSKJ 1	SP 2001	SSKJ 2
aja	aja, zdaj sem se spomnil	medmet	medmet	členek
alias	Janez Kotar, alias Maček	prislov	prislov	členek
bore	o njem vemo bore malo	prislov	prislov	členek
čuda	čuda gosta trava	prislov	prislov	členek
ergo	to je sovražnik, ergo ga je treba uničiti	prislov	veznik	členek
izvestno	knjiga bo izvestno kmalu izšla	prislov	-	členek
malce	je malce zmedena	prislov	prislov	členek
mar	mar bi bil tam ostal …	prislov	prislov	členek
namreč	midva, namreč žena in jaz …	prislov	veznik	členek
praviloma	praviloma voziti po desni	prislov	prislov	členek
samkrat	enkrat samkrat sta se sprla	prislov	prislov	členek
večidel	sadje je večidel še trdo	prislov	prislov	členek

Procesiranje naravnega jezika

- usmerjenost v praktične aplikacije
    - strojno prevajanje, informacijsko poizvedovanje, luščenje informacij, avtomatsko povzemanje, avtomatsko odgovarjanje na vprašanja, prepoznava govora, sinteza govora itd.
- omejitve računalniškega procesiranja
    - „česar ne zna razvrščati človek…“
- medjezična primerljivost
    - MULTEXT(-EAST) (1994-1996 / 1995-1997)
    - Universal Dependencies (2014-)

Tabele oznak za slovenščino

- SLON-13 (1991-), Jure Zupan (KI)
- MULTEXT-EAST (1993-1995, 1995-1997), *Multilingual Text Tools and Corpora for Central and Eastern European Languages*
  - IJS (+Amebis, FF UL)
- NOVA BESEDA (pr. 1997-1999)
  - Inštitut za slovenski jezik „Frana Ramovša“ ZRC
- LC-STAR (2002-2004), *Lexica and Corpora for Speech-to-Speech Translation Components*
  - Univerza v Mariboru (FERI)
- JOS (2007-2009), Jezikoslovno označevanje slovenskega jezika
  - IJS (+FF UL)
- Universal Dependecies (2014-)

št	kategorija	JOS	Multext-East	LC-STAR	ISJFR	SLON-13
1	samostalnik	+	+	+	+	+
	lastno ime	Ø	Ø	Ø	+	Ø
2	glagol	+	+	+	+	+
	pomožni glagol	Ø	Ø	+	Ø	+
3	pridevnik	+	+	+	+	+
4	zaimek	+	+	+	+	+
5	števnik	+	+	+	+	+
6	prislov	+	+	+	+	+
7	veznik	+	+	+	+	+
8	predlog	+	+	+	+	+
9	medmet	+	+	+	+	+
10	členek	+	+	+	+	+
11	okrajšava	+	+	+	+	+
	neuvrščeno	+	+	Ø	Ø	+

Samostalnik (JOS)

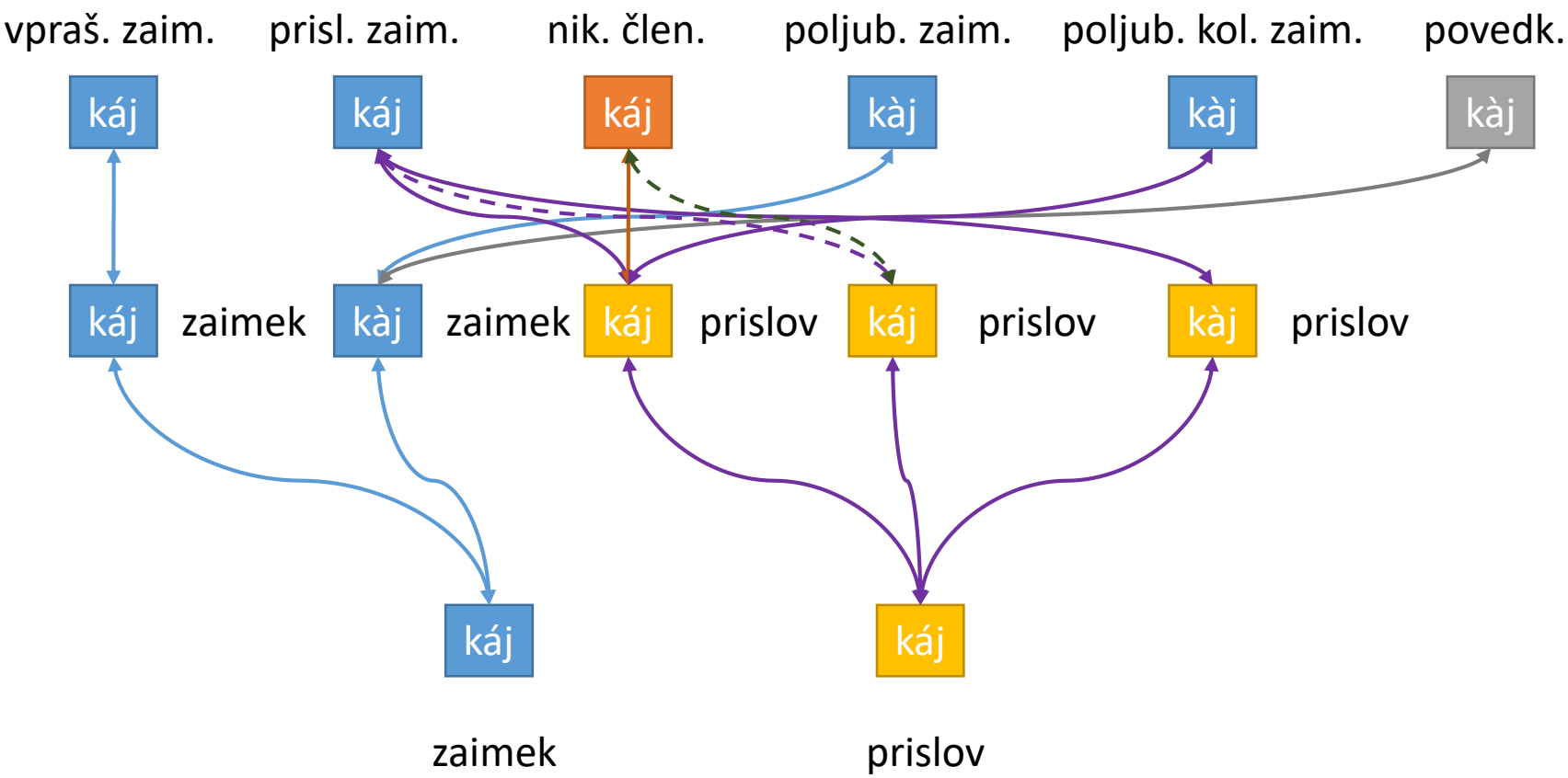
P	atribut	vrednost	koda
0	besedna_vrsta	samostalnik	S
1	vrsta	občno_ime, lastno_ime	o, l
2	spol	moški, ženski, srednji	m, z, s
3	število	ednina, dvojina, množina	e, d, m
4	sklon	imenovalnik, rodilnik, dajalnik, tožilnik, mestnik, orodnik	i, r, d, t, m, o
5	živost	ne, da	n, d

Pozicijske oznake (JOS)

	Besedna vrsta	vrsta	spol	število	sklon	živost
razlaga	samostalnik	občno_ime	moški	ednina	tožilnik	ne
koda	**S**	**o**	**m**	**e**	**t**	**n**
zgled	Oblekla si je bila kavbojke in pleten **pulover/Sometn**					

	Besedna vrsta	vrsta	spol	število	sklon	živost
razlaga	samostalnik	lastno_ime	moški	ednina	tožilnik	da
koda	**S**	**l**	**m**	**e**	**t**	**d**
zgled	**Davida/Slmetd** sva spoznala preko prijateljevega prijatelja					

SP – SSKJ – JOS


vpraš. zaim.
prisl. zaim.
nik. člen.
poljub. zaim.
poljub. kol. zaim.
povedk.
káj
kàj
zaimek
prislov
zaimek
prislov

Zaimek – ssj500k

Ft.Z.N.N	račun cigaret prihranite , in si za ta denar	kaj	/Zv-set	lepega kupite . ¶ Učili so nas , da je človeško
Ft.Z.N.N	račun delovanja starega NK Olimpija tudi	kaj	/Zv-set	zasluži . Iz omenjenih dokumentov je namreč
Ft.Z.N.N	ruskem pregovoru : Kdor se naokrog hvali ,	kaj	/Zv-set	bo naredil , mu načrti prej ali slej propadejo
Ft.Z.N.N	času . Informacija je tu , zdaj pa čakamo ,	kaj	/Zv-set	bodo pokazali specializirani muzeji . Štirideset
Ft.Z.N.N	dokler me ne vznemiri kakšen šef , češ ,	kaj	/Zv-sei	da je to sedaj s tem mojim tekstom . Saj
Ft.Z.N.N	imenom rastlin , da bomo spomladi vedeli ,	kaj	/Zv-set	smo kam posadili , saj je pomlad še daleč
Ft.Z.U.R	Mat pa foter bga nej raj mal povprašala ,	kva	/Zv-set	misl s svojim življenjem nardit . Sam ne
Ft.Z.N.N	prijavila , takrat pa se bo tudi pokazalo ,	kaj	/Zv-set	si lahko obetava od sezone . Na domačem
Ft.Z.N.S	tistimi okronano najlepšimi , pogumne kot le	kaj	/Zv-sei	in se ne branijo Playboyevega objektiva
Ft.Z.N.N	sem bil rad v družbi ali tam , kjer se je	kaj	/Zv-sei	dogajalo . Tako je tudi zdaj , ko sem upokojen
Ft.Z.N.N	je pač iz *[geo:* ZDA *]* in jo prav malo briga ,	kaj	/Zv-sei	se dogaja v naših logih . Zato pa držimo
Ft.Z.N.N	ima za seboj že 2000 let zgodovine , je	kaj	/Zv-set	takega vsekakor mogoče pričakovati . Ali
Ft.Z.N.N	uredniki in novinarji Slovenskega Primorja . ¶	Kaj	/Zv-sei	bo z blejskim gradom in kaj z otokom ,
-	tudi pri različnih oblikah evtanazije . ¶	Kaj	/Zv-set	menite o Krstu pri Savici in dilemi Črtomir
Ft.Z.N.N	kar je počel , v prid državi in da se nima	česa	/Zv-ser	sramovati . Hkrati je znova odločno zavrnil
Ft.Z.N.S.H	morejo pustiti hladne . Ali lahko obstaja	kaj	/Zv-sei	takega , kot je formalna racionalnost ?
Ft.Z.N.N	naslovom Mason & amp ; Dixon . Vsi imajo danes	kaj	/Zv-set	povedati o Thomasu Pinchonu ; še več ,
Ft.Z.U.R	razumem , zakaj pride tolikokrat do prepirov ,	kaj	/Zv-set	bi gledali : vsakdo bi rad gledal kaj svojega
Ft.Z.N.N	te konkurence preselil bliže , v Trst .	Kaj	/Zv-set	to pomeni za koprsko pristanišče ? ¶ Glave
Ft.Z.U.R	človeka ! " (Thomas Carlyle) ¶ K vprašanju "	Kaj	/Zv-sei	je cilj ? " nujno sodi tudi določanje časovnega

Prislov – ssj500k

Ft.Z.N.N	da je lahko tudi širša javnost izvedela	kaj	/Rsn	več o njih , vendar pa končne odločitve
Ft.Z.N.N	Priznam , da od mene svet ne bo več imel	kaj	/Rsn	prida : ko bom končal univerzo , jih bom
Ft.Z.N.N	kar sicer , kot si boste mislili , ne pove	kaj	/Rsn	dosti . A ko je pred kratkim padla slovenska
Ft.Z.N.N	izpovedovati svojih bravuroznih domislic ,	kaj	/Rsn	šele kritične distance ? Tudi pisma bralcev
Ft.Z.N.S.N	otrok ne traja dolgo . Prva topla sapica	kaj	/Rsn	hitro spremeni veličastnega sneženega moža
Ft.Z.N.S.N	mero navdušenja . Surovi kavčuk je namreč	kaj	/Rsn	muhasta snov . Pri nizkih temperaturah
Ft.Z.N.N	, da ji to povem . ɠ No , se zdaj počutite	kaj	/Rsn	drugače ? Se vam pred očmi bliska 25 rumenih
Ft.Z.N.N	tedaj reporterju kaj zgodilo , bi bili dami	kaj	/Rsn	na tesnem zaradi govoric , tako pa je bilo
Ft.Z.N.N	predsednikovih pogovorov . Res so njegovi sodelavci	kaj	/Rsn	hitro ugotovili , da manjkajo prepisi nekaterih
Ft.Z.N.N	veseljak . Aktualna politika ga ne zanima	kaj	/Rsn	prida , bolj ga skrbijo planetarne zadeve
Ft.Z.U.R	Prav , da nisi mogu zadovolit tud sebe ne ,	kva	/Rsn	šele kakšno pičko z unim štempselnom ,
Ft.Z.N.N	po modi naravnan tip človeka . Sicer pa ,	kaj	/Rsn	ni rekla , da so tudi špargljі v modi ?
Ft.Z.N.N	skrivnostnih želvjih poteh lahko izvedeli	kaj	/Rsn	več , so znanstveniki označili tudi obe
Ft.Z.N.N	reševalno operacijo . ɠ Si se prestrašil ,	kaj	/Rsn	, Fred , je zaslišal z druge strani žice
Ft.Z.N.N	komunalnega prispevka za gradnjo garažne hiše ,	kaj	/Rsn	šele z njegovo višino . Kljub temu pa je
Ft.Z.N.S.H	le malo k vzdržljivosti . Džoging pa ni	kaj	/Rsn	prida za moč , a izvrstno povečuje vzdržljivost
Ft.Z.N.N	precej kruto vprašanje , saj mi ostane komaj	kaj	/Rsn	prostega časa . Toda ko ga imam , ga posvetim
Ft.Z.N.N	naredili še kakšno o tem , bi jim postalo	kaj	/Rsn	hitro jasno , da je večina klobas *[other:*
Ft.Z.N.N	. ɠ Z letošnjimi prazniki si sicer človek	kaj	/Rsn	malo lahko pomaga , saj so se kot naročeno
Ft.Z.U.R	pomembni zadevi . ɠ Globoko , žalostno in nič	kaj	/Rsn	olajšano sem vdihnil , kot da bi v stanovanju

- <text><body> <p xml:id="F0028708.9" n="8 103 122"><s xml:id="F0028708.9.1" n="19 24">

<c xml:id="F0028708.9.1.1"> “ </c>
<w xml:id="F0028708.9.1.2" lemma="tisti" msd="Zk-mer"> Tistega </w><S/>
<w xml:id="F0028708.9.1.3" lemma="večer" msd="Somer"> večera </w><S/>
<w xml:id="F0028708.9.1.4" lemma="biti" msd="Gp-spe-n"> sem </w><S/>
<w xml:id="F0028708.9.1.5" lemma="preveč" msd="Rsn"> preveč </w><S/>
<w xml:id="F0028708.9.1.6" lemma="popiti" msd="Ggdd-em"> popil </w>
<c xml:id="F0028708.9.1.7"> , </c><S/>
<w xml:id="F0028708.9.1.8" lemma="zgoditi" msd="Ggdd-es"> zgodilo </w><S/>
<w xml:id="F0028708.9.1.9" lemma="se" msd="Zp------k"> se </w><S/>
<w xml:id="F0028708.9.1.10" lemma="biti" msd="Gp-ste-n"> je </w><S/>
<w xml:id="F0028708.9.1.11" lemma="mesec" msd="Sometn"> mesec </w><S/>
<w xml:id="F0028708.9.1.12" lemma="dan" msd="Sommr"> dni </w><S/>
<w xml:id="F0028708.9.1.13" lemma="po" msd="Dm"> po </w><S/>
<w xml:id="F0028708.9.1.14" lemma="ta" msd="Zk-sem"> tem </w>
<c xml:id="F0028708.9.1.15"> , </c><S/>
<w xml:id="F0028708.9.1.16" lemma="ko" msd="Vd"> ko </w><S/>
<w xml:id="F0028708.9.1.17" lemma="biti" msd="Gp-spe-n"> sem </w><S/>
<w xml:id="F0028708.9.1.18" lemma="izvedeti" msd="Ggdd-em"> izvedel </w>
<c xml:id="F0028708.9.1.19"> , </c><S/>
<w xml:id="F0028708.9.1.20" lemma="da" msd="Vd"> da </w><S/>
<w xml:id="F0028708.9.1.21" lemma="jaz" msd="Zop-et--k"> me </w><S/>
<w xml:id="F0028708.9.1.22" lemma="žena" msd="Sozei"> žena </w><S/>
<w xml:id="F0028708.9.1.23" lemma="varati" msd="Ggnste"> vara </w>
<c xml:id="F0028708.9.1.24"> . </c><S/>

Tistega

<w lemma="tisti" msd="Zk-mer">Tistega</w><S/>

- zaimek
  - vrsta=kazalni
  - spol=moški
  - število=ednina
  - sklon=rodilnik

večera

<w lemma="večer" msd="Somer">večera</w><S/>

- samostalnik
  - vrsta=občno_ime
  - spol=moški
  - število=ednina
  - sklon=rodilnik

sem

- <w lemma="biti" msd="Gp-spe-n">sem</w><S/>

- glagol
  - vrsta=pomožni
  - oblika=sedanjik
  - oseba=prva
  - število=ednina
  - nikalnost=nezanikani

preveč

- <w lemma="preveč" msd="Rsn">preveč</w><S/>

- prislov
  - vrsta=splošni
  - stopnja=nedoločeno

popil

- <w lemma="popiti" msd="Ggdd-em">popil</w>

- glagol
  - vrsta=glavni
  - vid=dovršni
  - oblika=deležnik
  - število=ednina
  - spol=moški

Universal Dependencies

- Razvoj medjezično čim bolj konsistentnega sistema skladenjskega razčlenjevanja za čim več jezikov
- Temelji na:
  - *Google universal PoS tags* (Petrov, Das, McDonald 2012)
    - https://code.google.com/p/universal-pos-tags/
  - *(universal) Stanford dependencies*
    - http://nlp.stanford.edu/software/stanford-dependencies.shtml
  - *Interset interlingua for morphosyntactic tagsets* (Zeman, 2008)
- Nahaja se na:
  - http://universaldependencies.github.io/docs/

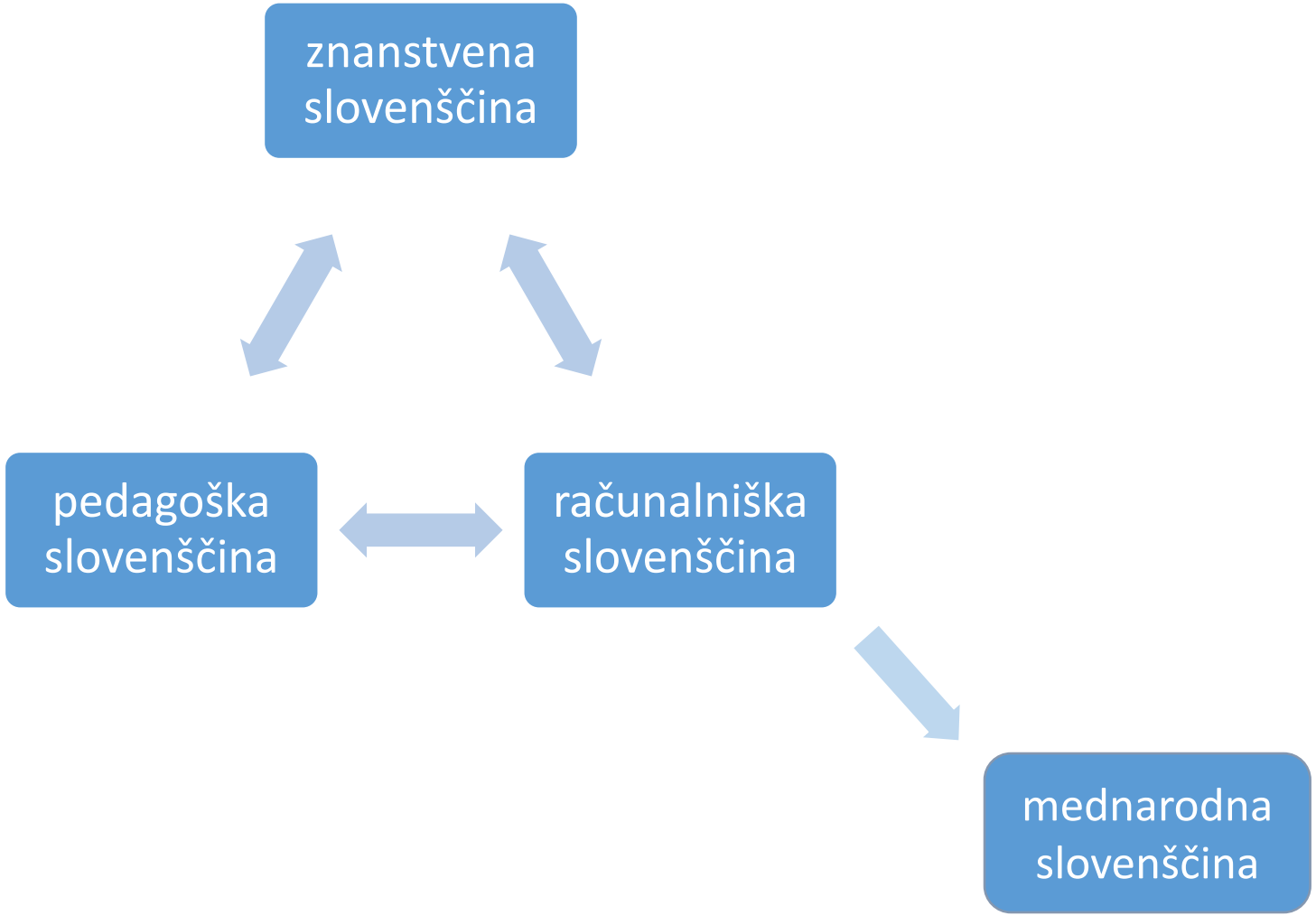
Universal POS tags

Open class words	Closed class words	Other
ADJ – pridevnik	ADP – predlog	PUNCT – ločilo
ADV – prislov	AUX – pomožni glagol	SYM – simbol
INTJ – medmet	CONJ – veznik	X – neuvrščeno
NOUN – samostalnik	DET – Ø	
PROPN – lastno ime	NUM – števnik	
VERB - glagol	PART – „členek“	
	PRON – zaimek	
	SCONJ – podredni veznik	

UD – določilnik in členek

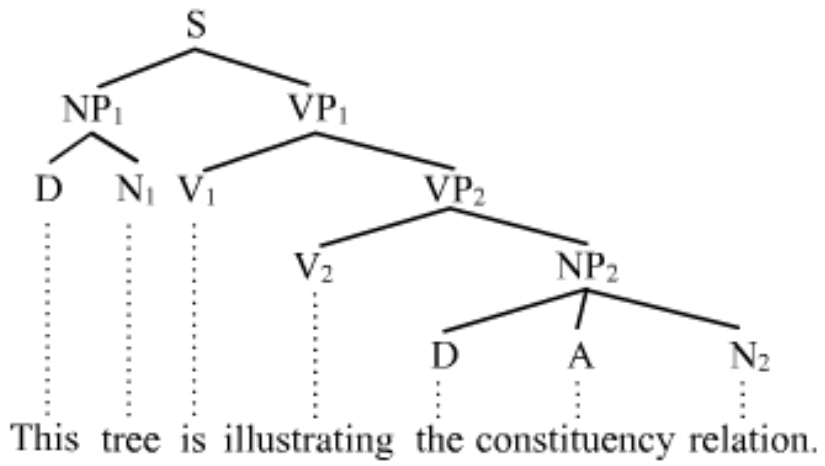
- „Note that the notion of **determiners** is unknown in grammars of some languages (e.g. Czech); words equivalent to English determiners may be traditionally classified as pronouns and/or numerals in these languages. In order to annotate the same thing the same way across languages, the words satisfying our definition of determiners should be tagged DET in these languages as well.“
- **„Particles** are function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech.“

Oblikoslovje: problem treh slovenščin

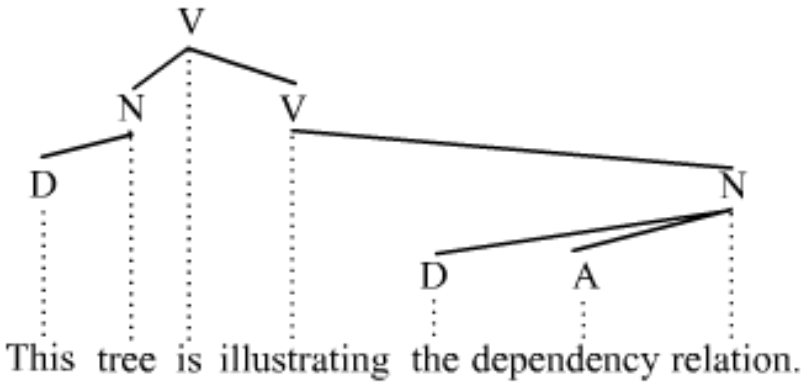


Skladenjsko razčlenjevanje

- avtomatsko pripisovanje skladenjskih razmerij med pojavnicami v povedi
  - korpusna raziskava temeljnih skladenjskih pojavov v jeziku
  - podpora kompleksnejšim jezikovnim tehnologijam



Constituency relation (PSG)



Dependency relation

Jezikoslovni model

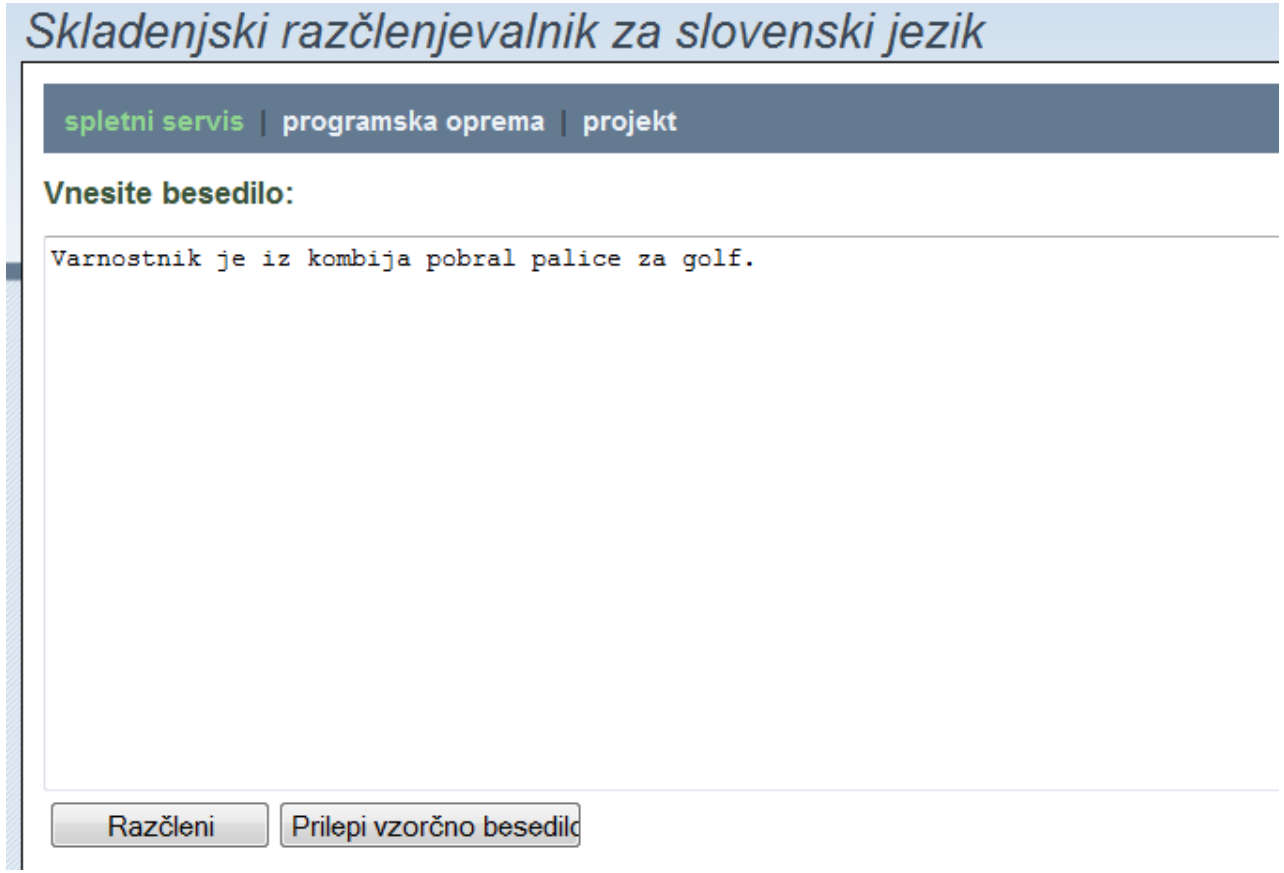
- sistem odvisnostne drevesnice
- **Jezikoslovno označevanje slovenščine** (http://nl.ijs.si/jos)
- **Sporazumevanje v slovenskem jeziku** (http://www.slovenscina.eu)
- upoštevanje tradicionalnih skladenjskih opisov slovenščine in specifične narave strojne obdelave besedil
- 10 tipov skladenjskih razmerij

Tipi skladenjskih povezav

Skupina	Tip	Kaj povezuje
1. nivo	*dol*	jedro in določilo besednih zvez
	del	deli zloženega povedka
	prir	jedra v prirednih zvezah znotraj stavka
	vez	besede ali ločila v vezniški vlogi.
	skup	nepolnopomenske besede, ki imajo zelo močno tendenco po sopojavljanju
2. nivo	*ena*	osebek stavka
	dve	predmet stavka
	tri	prislovno določilo lastnosti
	štiri	ostala prislovna določila
3. nivo	*modra*	hierarhično najvišje pojavnice, skladenjsko manj predvidljive in oddaljene strukture, vrinki, ločila …

Spletni servis

http://razclenjevalnik.slovenscina.eu/



Razčlenjeno besedilo:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <text>
    <body>
      <p xml:id="0">
        <s xml:id="0.0">
          <w lemma="varnostnik" msd="Somei" xml:id="0.0.1">Varnostnik</w>
          <S />
          <w lemma="biti" msd="Gp-ste-n" xml:id="0.0.2">je</w>
          <S />
          <w lemma="iz" msd="Dr" xml:id="0.0.3">iz</w>
          <S />
          <w lemma="kombi" msd="Somer" xml:id="0.0.4">kombija</w>
          <S />
          <w lemma="pobrati" msd="Ggdd-em" xml:id="0.0.5">pobral</w>
          <S />
          <w lemma="palica" msd="Sozmt" xml:id="0.0.6">palice</w>
          <S />
          <w lemma="za" msd="Dt" xml:id="0.0.7">za</w>
          <S />
          <w lemma="golf" msd="Sometn" xml:id="0.0.8">golf</w>
          <c xml:id="0.0.9">.</c>
          <links>
            <link afun="ena" dep="0.0.1" from="0.0.5" />
            <link afun="del" dep="0.0.2" from="0.0.5" />
            <link afun="dol" dep="0.0.3" from="0.0.4" />
            <link afun="štiri" dep="0.0.4" from="0.0.5" />
            <link afun="modra" dep="0.0.5" from="0.0.0" />
            <link afun="dve" dep="0.0.6" from="0.0.5" />
            <link afun="dol" dep="0.0.7" from="0.0.8" />
            <link afun="dol" dep="0.0.8" from="0.0.6" />
            <link afun="modra" dep="0.0.9" from="0.0.0" />
          </links>
```

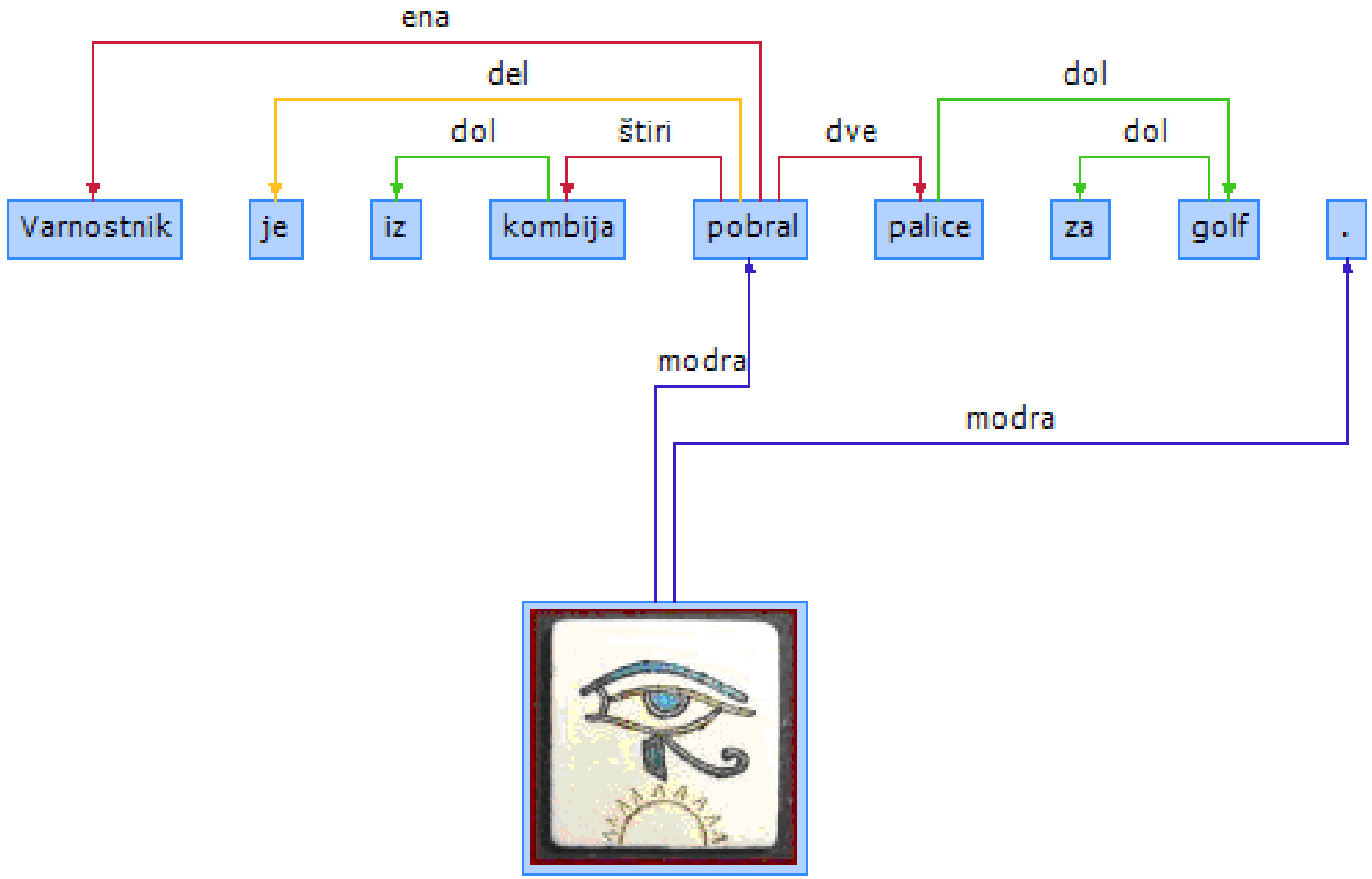
lematizacija in oblikoskladenjsko označevanje

skladenjske povezave

korenski element (ničta pojavnica)

Vizualizacija:

Varnostnik je iz kombija pobral palice za golf.



Programska oprema

- **skladenjski razčlenjevalnik MSTparser:** http://razclenjevalnik.slovenscina.eu/Programska_oprema.aspx; licenca Apache License 2.0
  - natančnost (F1) – 87.52

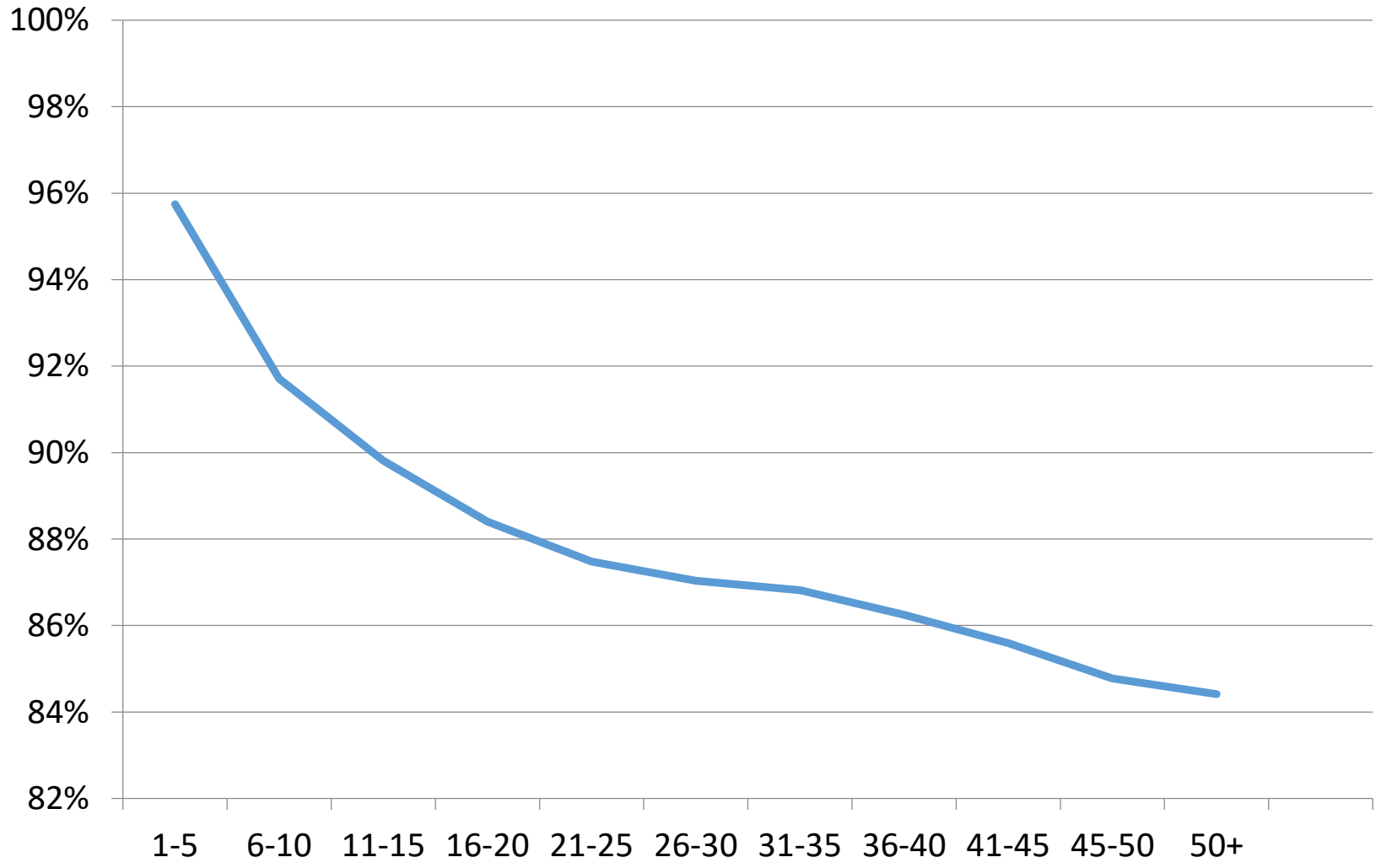
- **skladenjski razčlenjevalnik StanfordNLP:** https://github.com/clarinsi/classla-stanfordnlp; licenca Apache License 2.0
  - natančnost (F1) – 92.68

Razčlenjevalnik in učni korpus

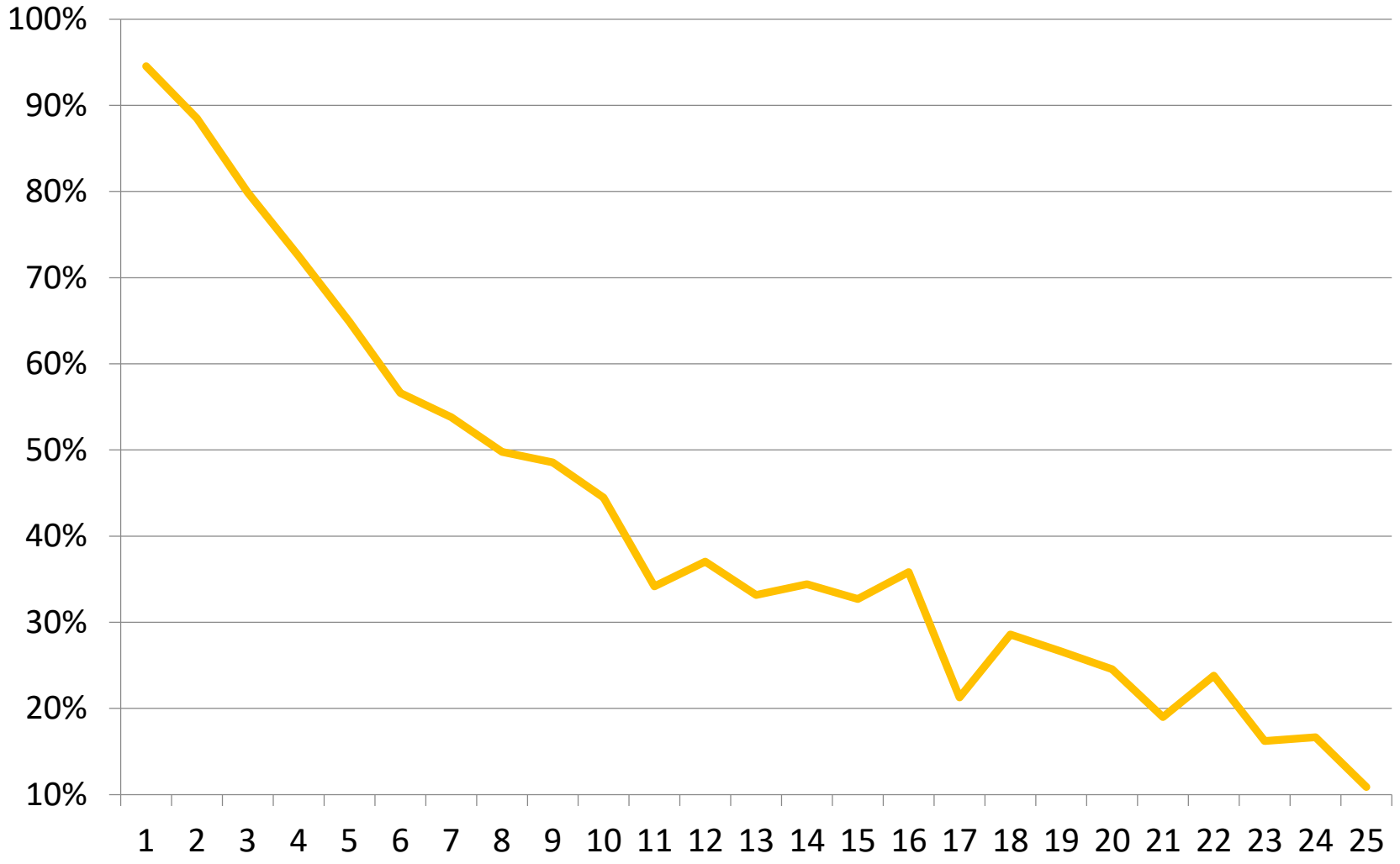
- MSTParser (Minimum Spanning Tree Parser, McDonald, Lerman in Pereira, 2006)
- statistični razčlenjevalnik
- učni korpus ssj500k; skladenjsko označeni del:

oznaka	opis	število
<w>	beseda	200.320
<c>	ločilo/simbol	35.545
<w> + <c>	pojavnica	235.865
<links>	element s skladenjskimi povezavami	11.411
<link>	skladenjska povezava	235.865

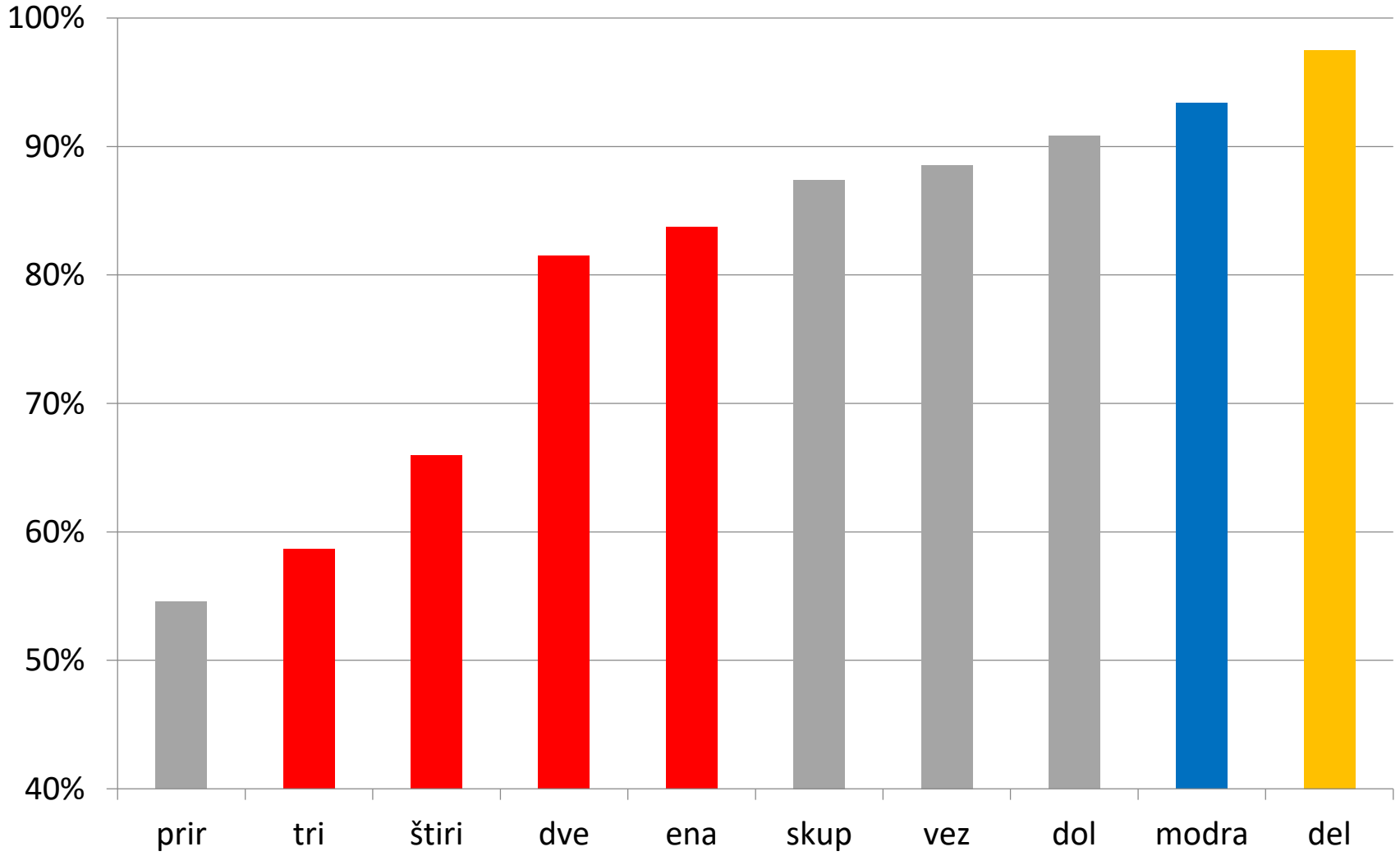
Dolžina povedi

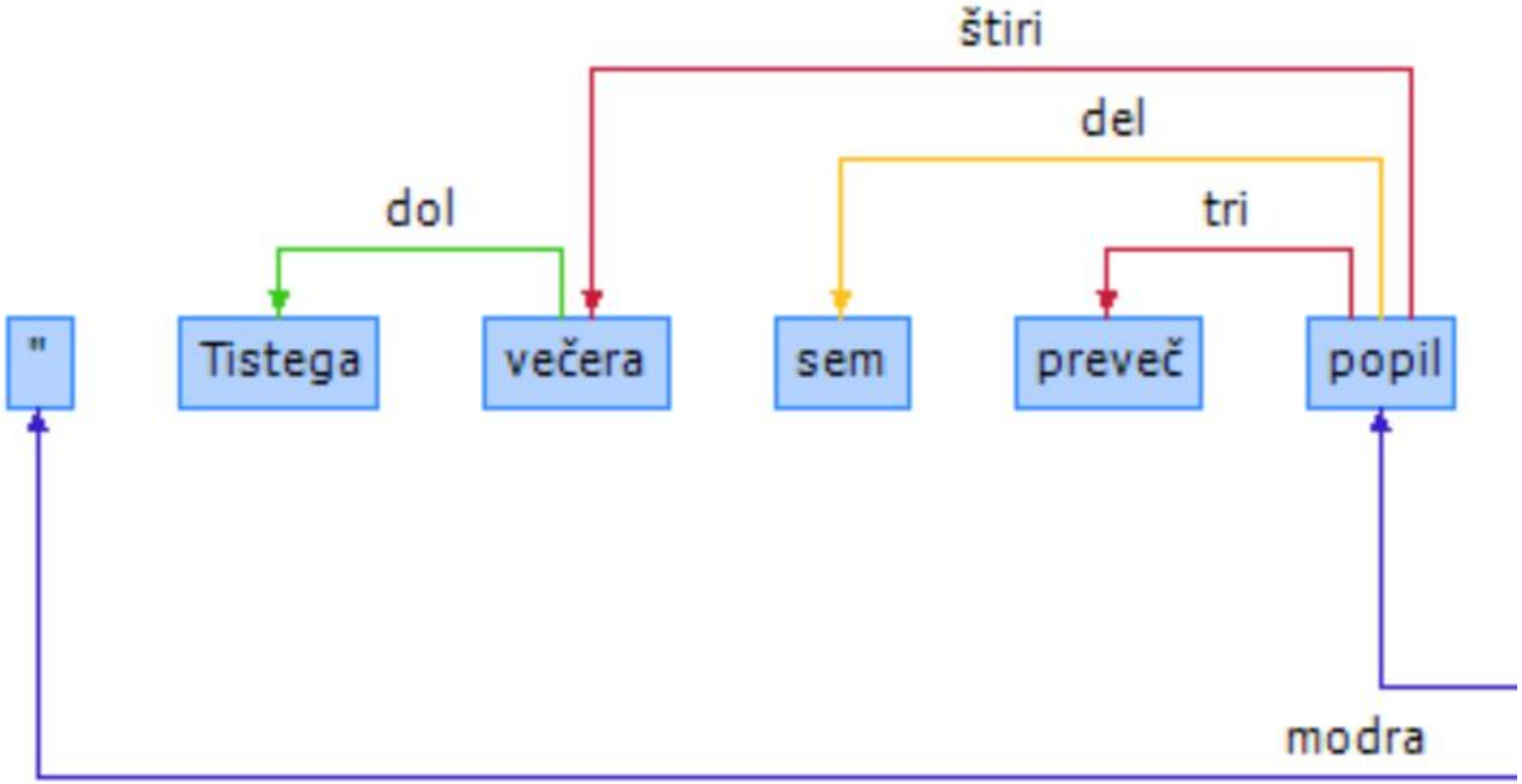


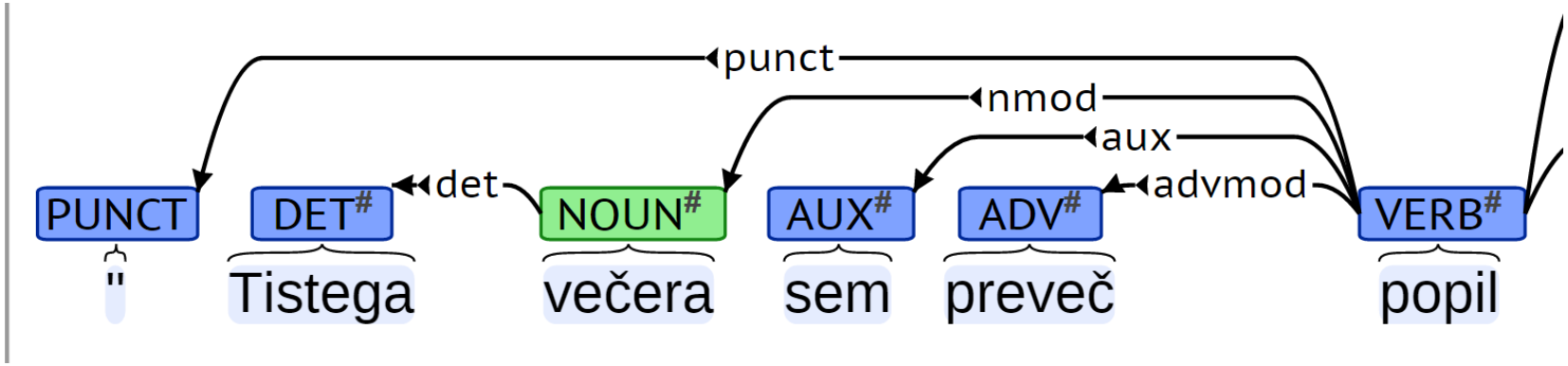
Dolžina povezave

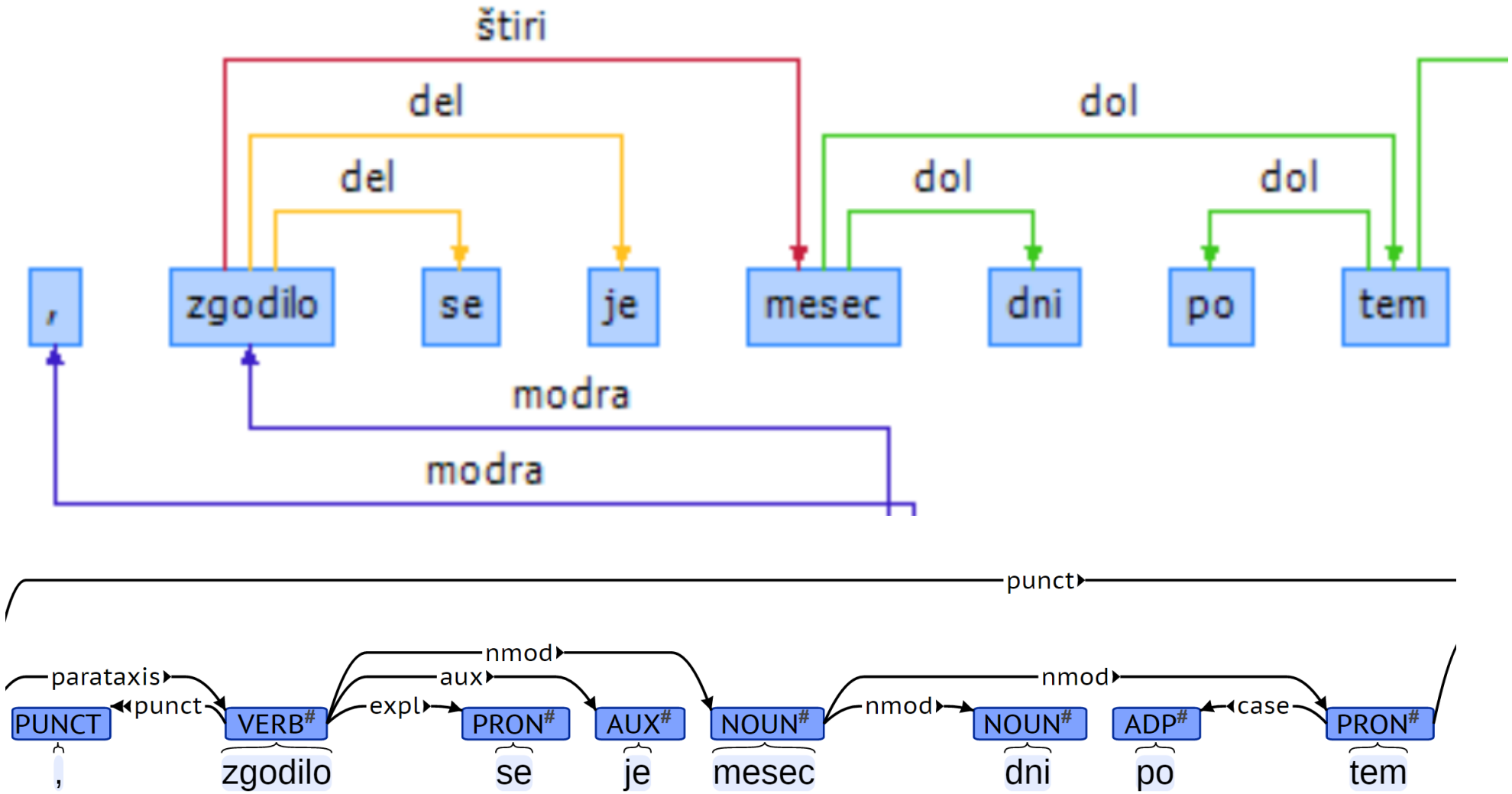


Tip povezave

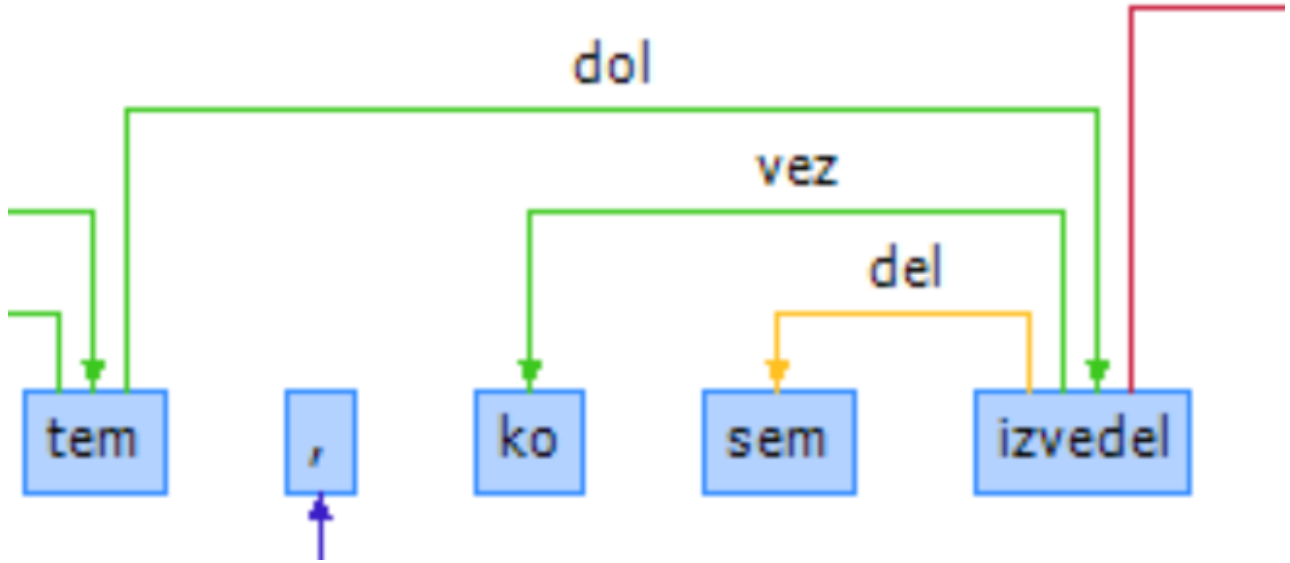


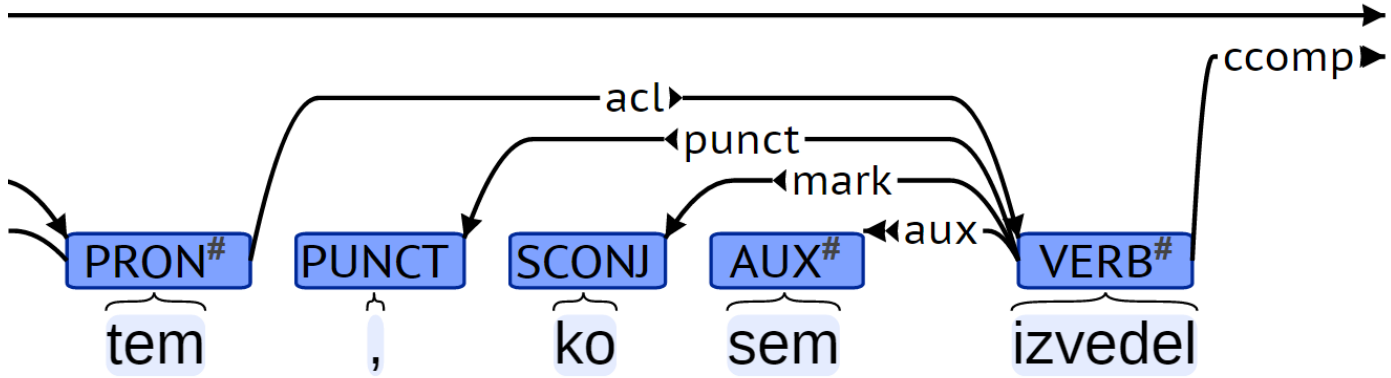

štiri
del
dol
tri
"
Tistega
večera
sem
preveč
popil
modra

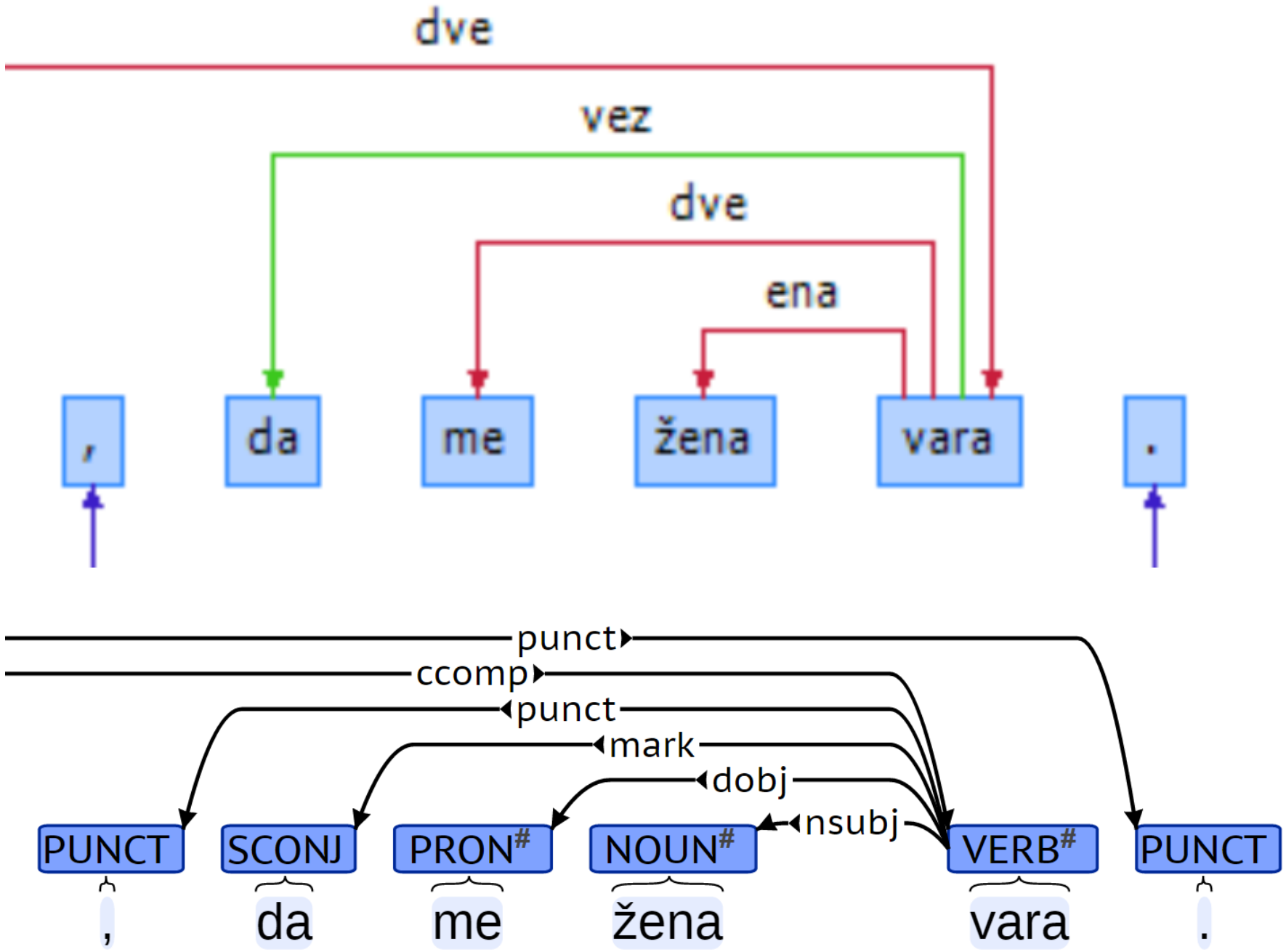

punct
nmod
aux
det
advmod
PUNCT
DET#
NOUN#
AUX#
ADV#
VERB#
"
Tistega
večera
sem
preveč
popil


štiri
del
del
dol
dol
dol
,
zgodilo
se
je
mesec
dni
po
tem
modra
modra

punct
parataxis
punct
nmod
aux
expl
nmod
nmod
case
PUNCT
VERB#
PRON#
AUX#
NOUN#
NOUN#
ADP#
PRON#
,
zgodilo
se
je
mesec
dni
po
tem


dol
vez
del
tem
,
ko
sem
izvedel


ccomp
acl
punct
mark
aux
PRON#
PUNCT
SCONJ
AUX#
VERB#
tem
,
ko
sem
izvedel


dve
vez
dve
ena
,
da
me
žena
vara
.
punct
ccomp
punct
mark
dobj
nsubj
PUNCT
SCONJ
PRON#
NOUN#
VERB#
PUNCT
,
da
me
žena
vara
.

Označevanje semantičnih vlog

- Prepoznavanje semantičnih udeležencev (udeleženskih vlog), ki so povezane s stavčnim predikatom oz. glagolom, in pripisovanje teh vlog posameznim stavčnim elementom

Kdo	***naredi***	**kaj**	**komu**	**kdaj**	**kje**	**kako**	**zakaj**
ACT		PAT	REC	TIME; DUR; FREQ	LOC	MANN	CAUSE
delovalniki				okoliščine			

Udeleženske vloge v učnem korpusu za slovenščino

DELOVALNIKI
aktanti, argumenti

ACT
PAT
REC
ORIG
RESLT

OKOLIŠČINE
cirkumstanti, adjuncts

ČAS & PROSTOR

TIME
DUR
FREQ

EVENT

LOC
SOURCE
GOAL

QUANT

VZROK & NAČIN

AIM
CAUSE
CONTR
COND
REG

ACMP
RESTR
MANN
MEANS

GLAGOLSKE ZVEZE

MWPRED
MODAL
PHRAS

Nabor udeleženskih vlog - delovalniki

Oznaka	Opis		Primer
ACT	aktant, vršilec	delujoči udeleženci, povzročitelji ali nosilci dejanja	Oče ACT dela.
PAT	prizadeto	prizadeti predmet dejanja	Kuhajo kosilo PAT
REC	prejemnik	prejemnik, posredni udeleženec dejanja; nedelovalniški udeleženec, ki mu je dejanje v škodo ali v prid	Prijatelju REC sem poslal darilo.
ORIG	izvor	izhodišče, izvor/vir/povod dejanja	Plašč je dobil od očeta ORIG
RESLT	učinek	učinek, rezultat, cilj dejanja	Imenovali so ga za predsednika RESLT

Nabor udeleženskih vlog – okoliščine: čas in prostor

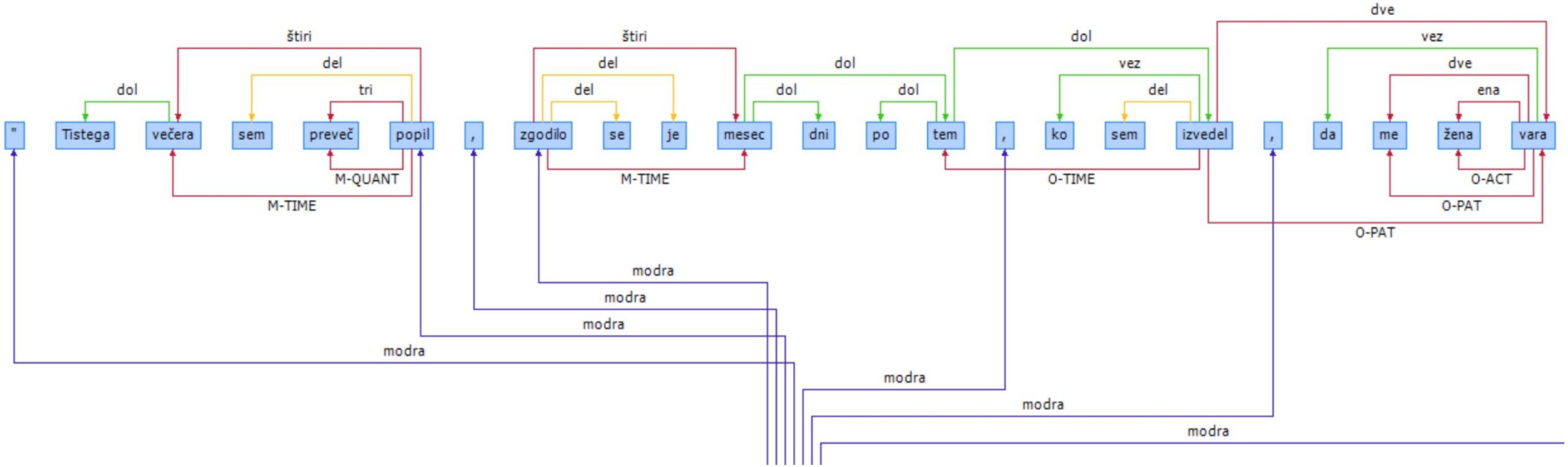
Oznaka		Opis		Primer
ČAS	TIME	čas	konkretni trenutek ali interval dejanja; kdaj	Med počitnicami $_{\text{TIME}}$ ni niti enkrat posijalo sonce.
	DUR	trajanje	trajanje stanja, dejanja koliko časa	Prišel je za en mesec $_{\text{DUR}}$
	FREQ	pogostnost	frekvenca dejanja; kako pogosto, kolikokrat	Vsak dan $_{\text{FREQ}}$ se mučimo s tem.
PROSTOR	LOC	kraj	konkretna lokacija, kraj, mesto dejanja; kje	V bližini vasi $_{\text{LOC}}$ stoji kozolec.
	SOURCE	začetna lokacija	začetna točka v prostoru; od kod	S stropa $_{\text{SOURCE}}$ odpada omet.
	GOAL	končna lokacija	končna točka v prostoru; kam	Prišel je domov $_{\text{GOAL}}$

Oznaka		Opis		Primer
VZROČNOST	AIM	namen	namen dejanja; čemu, s kakšnim namenom	Telovadi, da bi shujšala $_{AIM}$
	CAUSE	vzrok	vzrok dejanja; zakaj	Umrl je zaradi srčne kapi $_{CAUSE}$
	CONTR	protivnost	nepričakovana posledičnost dejanja; kljub čemu	Medtem ko se plače nižajo, $_{CONTR}$ cene rastejo
	COND	pogojnost	pogoj za obstoj dejanja ali dogodka	Če neha deževati $_{COND}$ gremo na kavo.
	REG	ozir	glede na, primerjava	Preiskava je potekala v skladu z zakonom$_{no\text{-}REG}$
NAČIN	ACMP	spremstvo	predmet, oseba ali dogodek, ki spremlja dejanje ali druge udeležence	Mama je s sinom$_{-ACMP}$ odšla v cirkus.
	RESTR	omejitev	izjema, omejitev	Vsi so bili tam razen tebe $_{RESTR}$
	MANN	način	načinovna lastnost dejanja,	Dela prepočasi $_{MANN}$
	MEANS	sredstvo	sredstvo ali orodje za izvedbo dejanja	Piše s peresom $_{MEANS}$
KOLIČINA	QUANT	količina	količina, razlika	Cena goriva se je podražila za 3 cente $_{QUANT}$

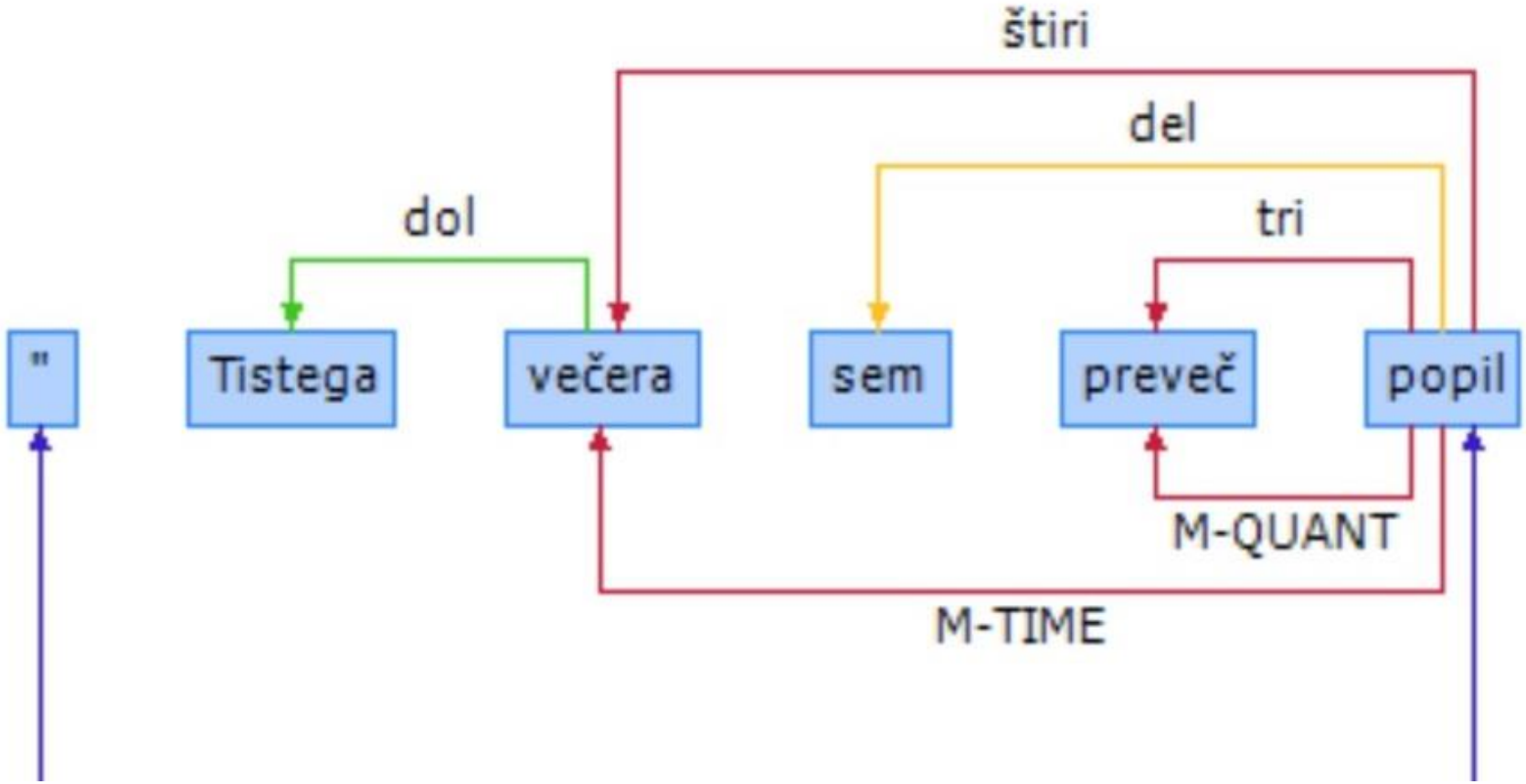
Nabor udeleženskih vlog – glagolske zveze

Oznaka		Opis	Primer
GLAGOLSKE ZVEZE	MWPRED	zveze z nedoločniki	Dati vedeti$_{MWPRED}$
	MODAL	zveze biti + modalnega prislova/pridevnika	je treba, je mogoče
	PHRAS	pomensko neprozorne zveze	Iti na živce $_{PHRAS}$

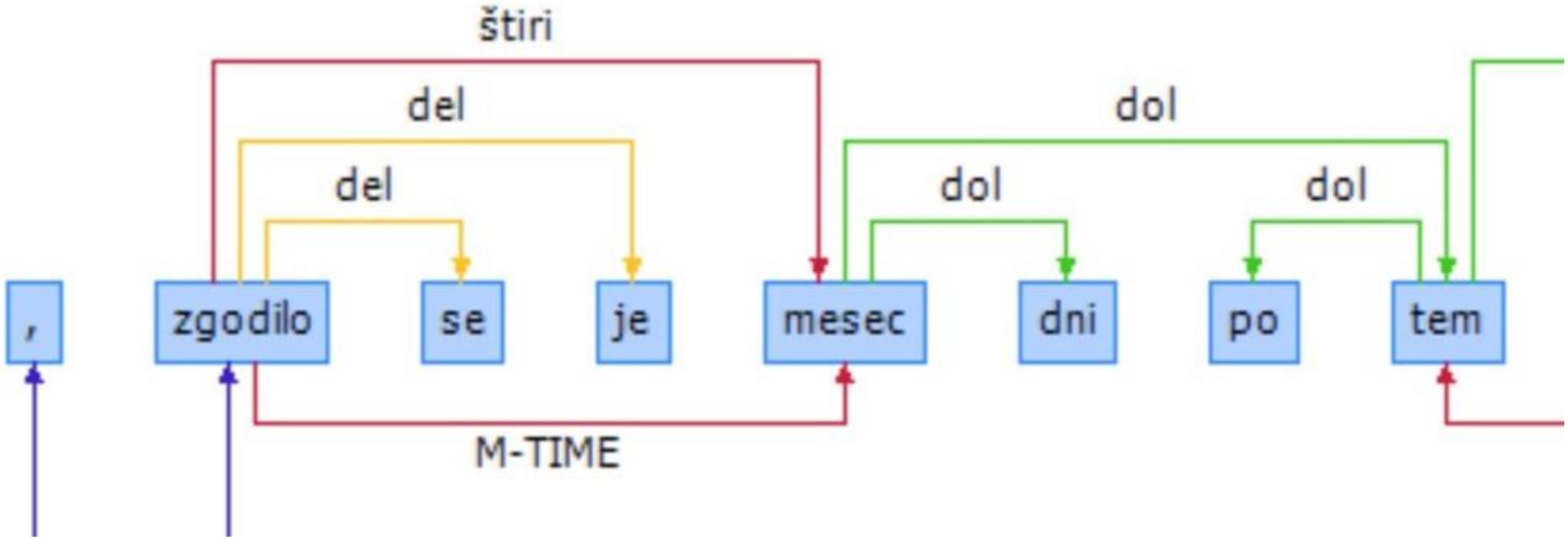
Primer označenega stavka



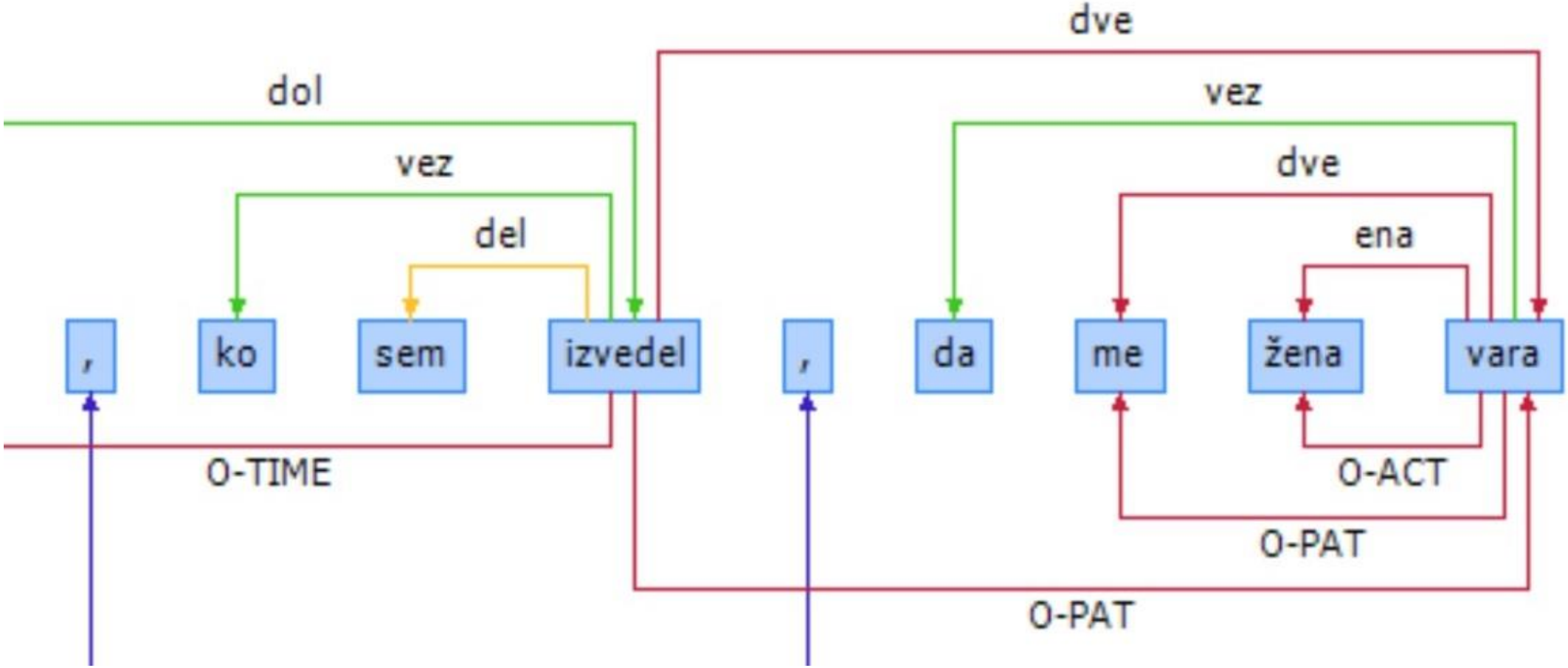
TIME / QUANT


štiri
del
dol
tri
"
Tistega
večera
sem
preveč
popil
M-QUANT
M-TIME

TIME



ACT / PAT / TIME


dve
dol
vez
vez
dve
del
ena
,
ko
sem
izvedel
,
da
me
žena
vara
O-TIME
O-ACT
O-PAT
O-PAT

Označevanje semantičnih vlog

- Razvoj označevalnika semantičnih vlog
  - Bilateral-srl (CLARIN.SI)
  - https://github.com/clarinsi/bilateral-srl
- Učni korpus ssj500k
  - Ročno označen 5.501 stavek
- Q-CAT
  - Dodan nivo za ročno označevanje semantičnih vlog

Drugi nivoji označevanja

- Vsebovane v ssj500k
  - Imenske entitete
  - Večbesedne enote (glagolske) – projekt PARSEME

- Druge
  - Semanični tipi (smreka -> drevo -> rastlina)
  - Pomensko razdvoumljanje (word sense disambiguation)

Pozor

IŠČEM KANDIDATA ALI KANDIDATKO ZA DOKTORSKI ŠTUDIJ, KI ŽELI RAZISKOVATI SEMANTIKO SLOVENSKEGA JEZIKA, V POVEZAVI Z JEZIKOVNIMI TEHNOLOGIJAMI. PIŠITE MI: KONTAKT.

Povezave

- Jezikoslovno označevanje slovenščine (JOS)
  - http://nl.ijs.si/jos/
- Sporazumevanje v slovenskem jeziku
  - http://www.slovenscina.eu/
- CLARIN.SI (repozitorij, GitHub)
  - http://www.clarin.si/
- Universal Dependencies
  - http://universaldependencies.org/

- e-pošta
  - simon.krek@ijs.si
- Center za jezikovne vire in tehnologije Univerze v Ljubljani
  - http://www.cjvt.si/

- StanfordNLP (tokenizacija, segmentacija, označevanje, lematizacija, razčlenjevanje): https://github.com/clarinsi/classla-stanfordnlp