Slovnične analize ročno označenega korpusa ssj500k z orodjem Q-CAT

Kaja Dobrovoljc

1 Center za jezikovne vire in tehnologije, Filozofska fakulteta, Univerza v Ljubljani

2 Laboratorij za umetno inteligenco, Inštitut Jožef Stefan

Dogodek “Kvantitativne in kvalitativne korpusne analize z novimi orodji iz nacionalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256)”, Ljubljana, 19. 11. 2019

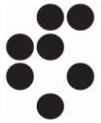
Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

O (kvalitativnem) korpusnem jezikoslovju

- korpusno jezikoslovje
  - preučevanje jezika na podlagi realnih jezikovnih podatkov
  - ni samo kvantitativno

- analiza **manjšega** števila **konkretnih** primerov rabe
  - obvladljivejše delo
  - **poglobljenejša analiza**
    - upoštevanje konteksta
    - identifikacija izjem in mejnih pojavov

- idealna kombinacija obeh pristopov



Načrt predavanja

1. **Predstavitev korpusa ssj500k**
   - 1.1. Nastanek in vsebina
   - 1.2. Ravni označenosti
   - 1.3. Dostopnost korpusa

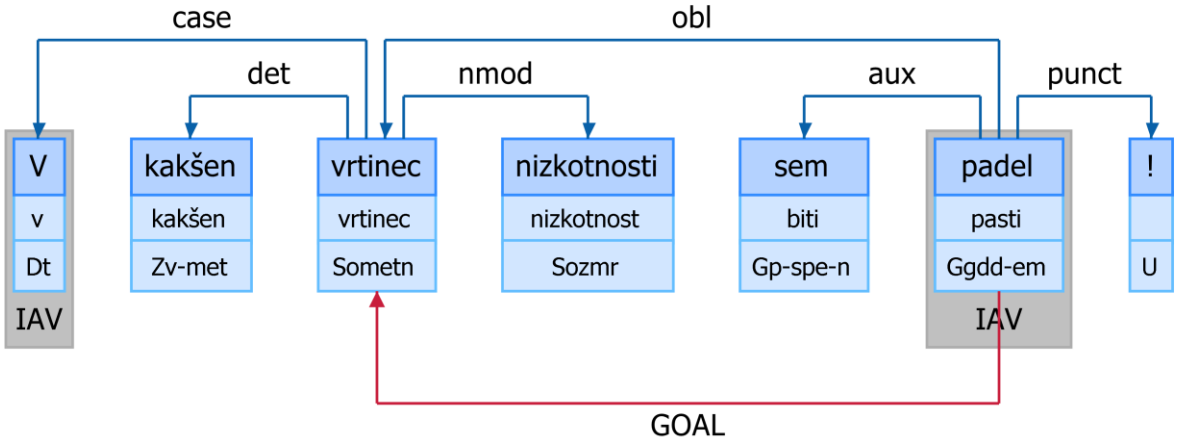
2. **Predstavitev orodja Q-CAT**
   - 2.1. Namestitev
   - 2.2. Pregledovanje
   - 2.3. Iskanje
   - 2.4. Označevanje

3. **Zaključek**

Učni korpus **ssj500k**

Kaj je korpus ssj500k

- korpus slovenskih besedil z **(ročno) pripisanimi slovničnimi oznakami** na različnih ravneh



- prosto dostopen
- približno 500.000 besed

Za kaj je uporaben

- temeljna podatkovna zbirka za **razvoj jezikovnih tehnologij**
  - osnova za nadzorovano strojno učenje
  - orodja za slovnično analizo (npr. samodejno oblikoskladenjsko označevanje)
  - druge aplikacije (npr. samodejno prevajanje, povzemanje …)

- pomembna tudi za **razvoj jezikoslovnega opisa** slovenščine
  - preverjanje obstoječih teorij in tipologij na avtentičnem gradivu
  - sistematična detekcija mejnih pojavov
  - razvoj novih tipologij
  - jezikoslovne analize

Zgodovina nastajanja

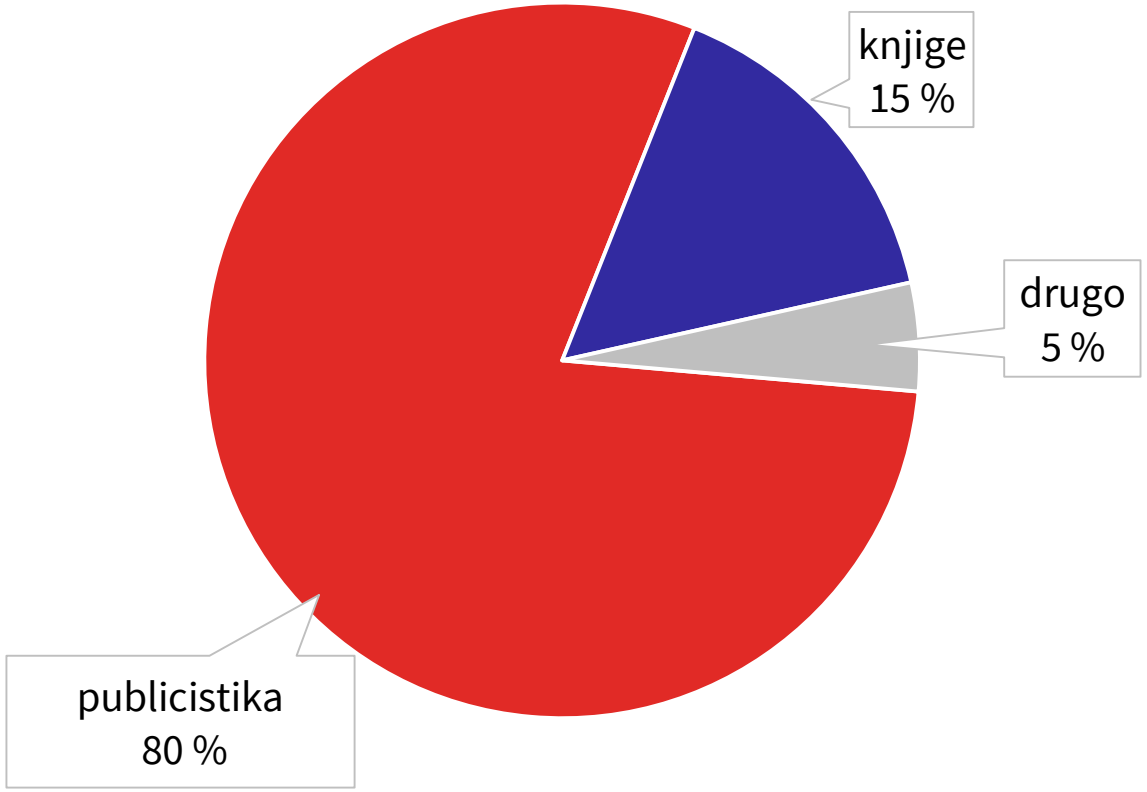
- več kot dve desetletji nenehnega razvoja
- 1998-: **MULTEXT-East** – razvoj prvih jezikovnih virov za oblikoskladenjsko označevanje slovenščine (specifikacije, prvi korpus – roman *1984*)
- 2007-2009: **Jezikoslovno označevanje slovenščine (JOS)** – nadgradnja smernic za lematizacijo, oblikoskladenjsko in (novo) skladenjsko označevanje, prenos na novi korpus jos100k (vzorec korpusa FidaPLUS)
- 2008-2013: **Sporazumevanje v slovenskem jeziku** (SSJ) – označenih še dodatnih 400.000 besed (ssj500k), dodane tudi imenske entitete
- 2013-2018: niz manjših projektov
  - **Označevanje udeleženskih vlog** v slovenščini in hrvaščini
  - **PARSEME**: glagolske večbesedne enote
  - **UD**: mednarodno usklajeno (obliko)skladenjsko označevanje
  - **JANES**: imenske entitete
- itd.

→ trenutna različica: **2.2** (Krek et al. 2019)

Besedilna zgradba

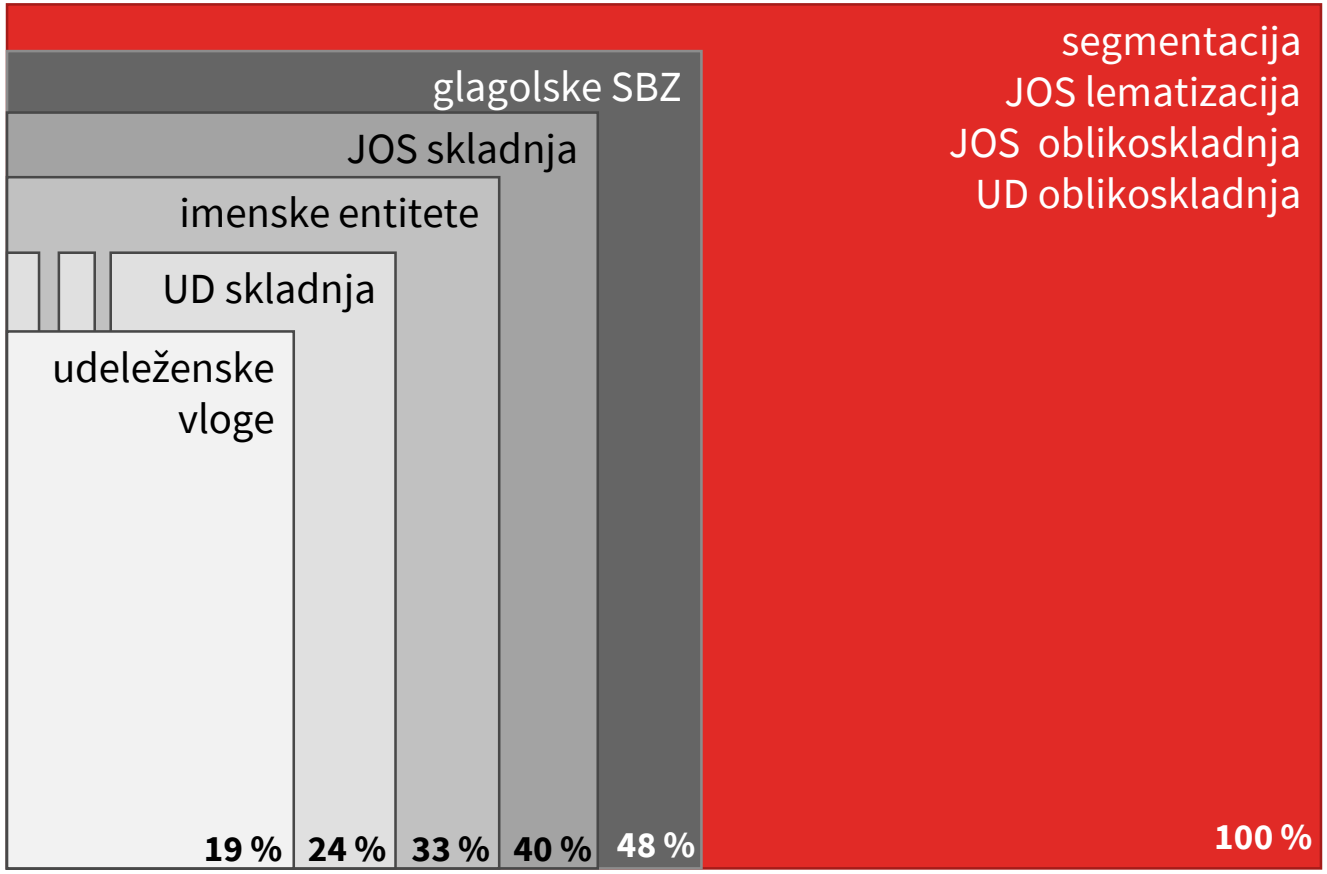
- vzorec referenčnega korpusa sodobne pisne slovenščine **FidaPLUS** (Arhar in Gorjanc 2007)

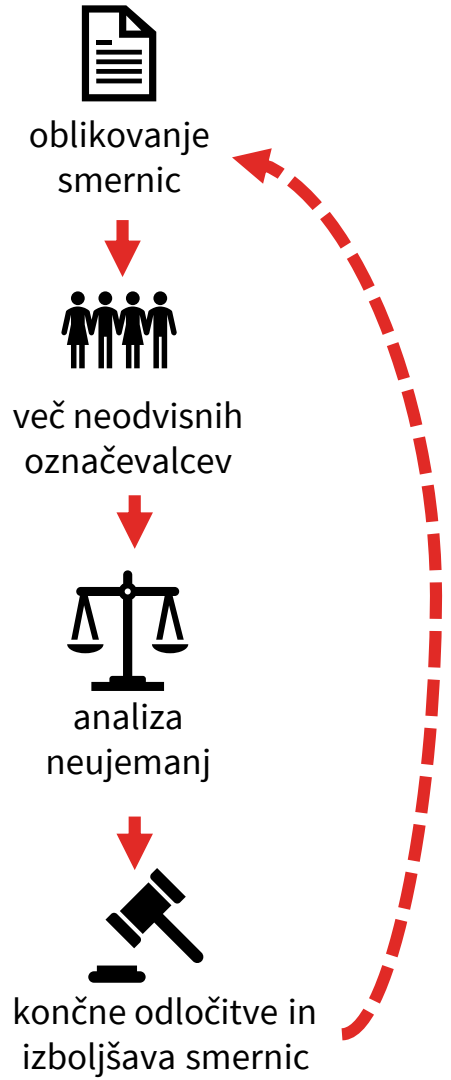
Enota	Pojavitev
Besedil	**1.655**
Odstavkov	8.137
Povedi	**27.829**
Pojavnic	586.248
Besed	**500.295**



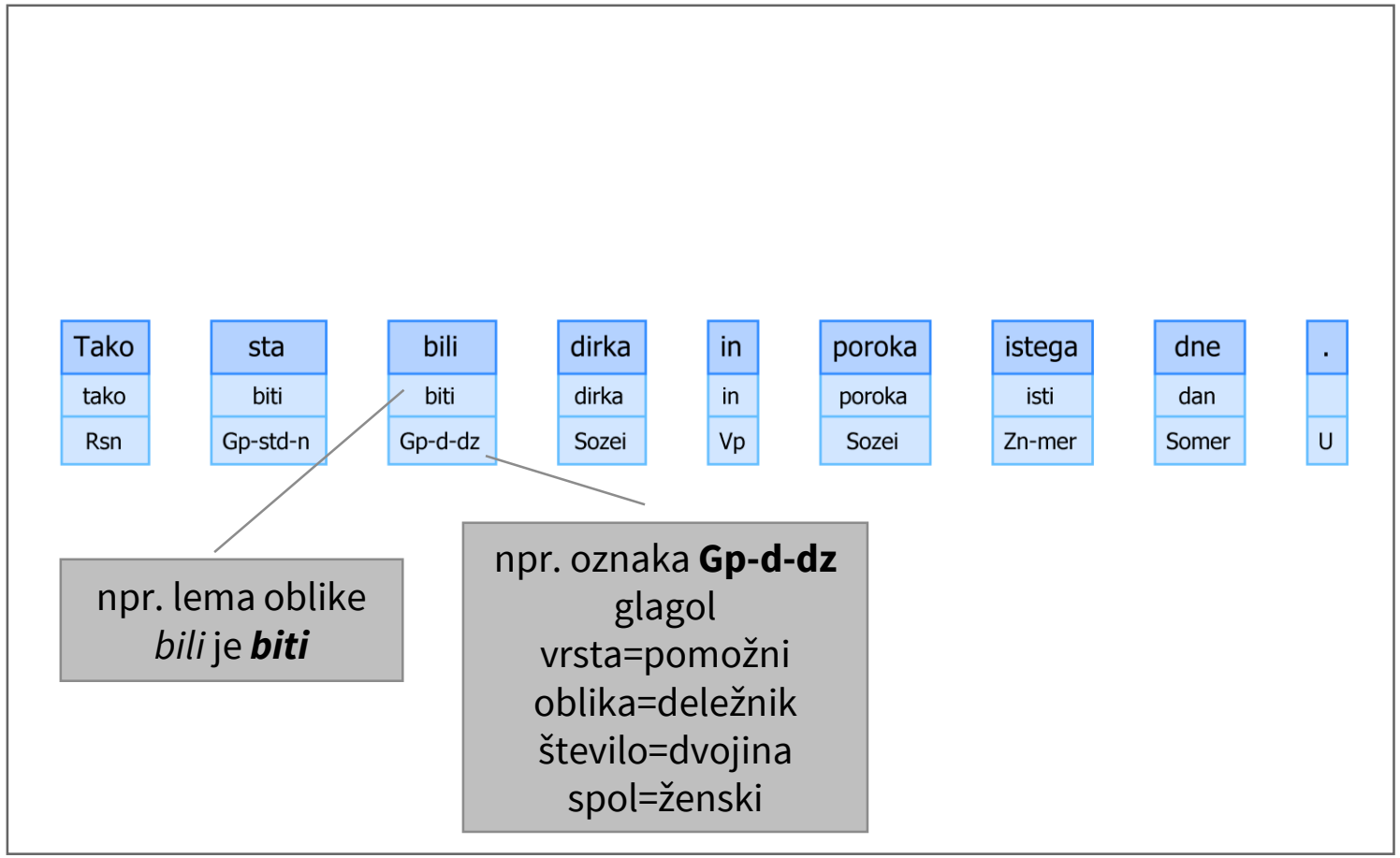
Označenost korpusa ssj500k

- več ravni označenosti različnega obsega



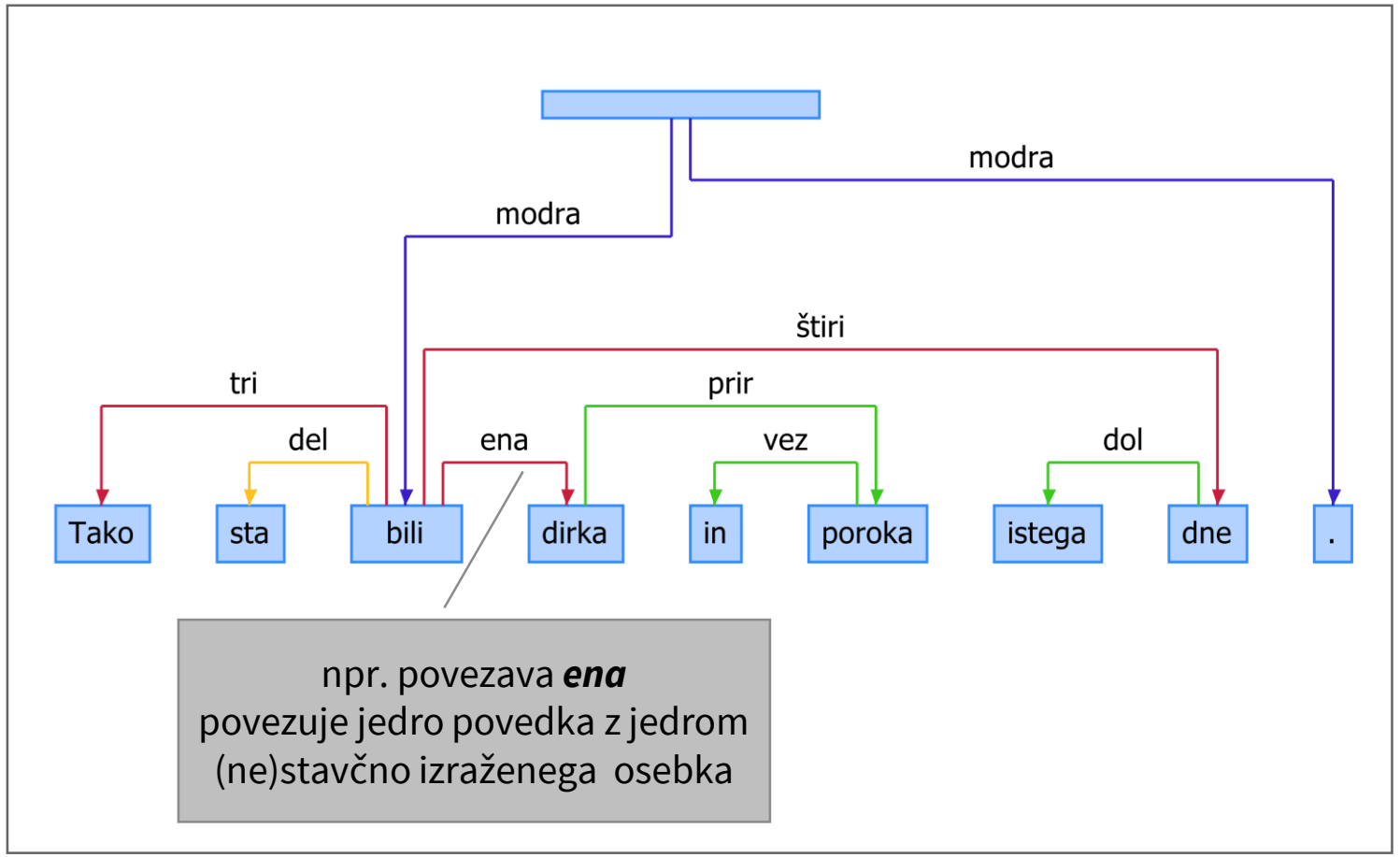


Oblikoskladnja JOS

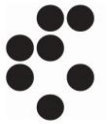


- pripisovanje **osnovnih oblik** besed (lem) in njihovih **oblikoslovnih lastnosti** (oblikoskladenjskih oznak)
- označevalni sistem **JOS** (Erjavec in Krek 2008; Holozan et al. 2008) – opredelitev nabora besednih vrst, oblikoskladenjskih lastnosti in vrednosti za slovenščino, pravila lematizacije
- http://nl.ijs.si/jos/
- označen celoten korpus

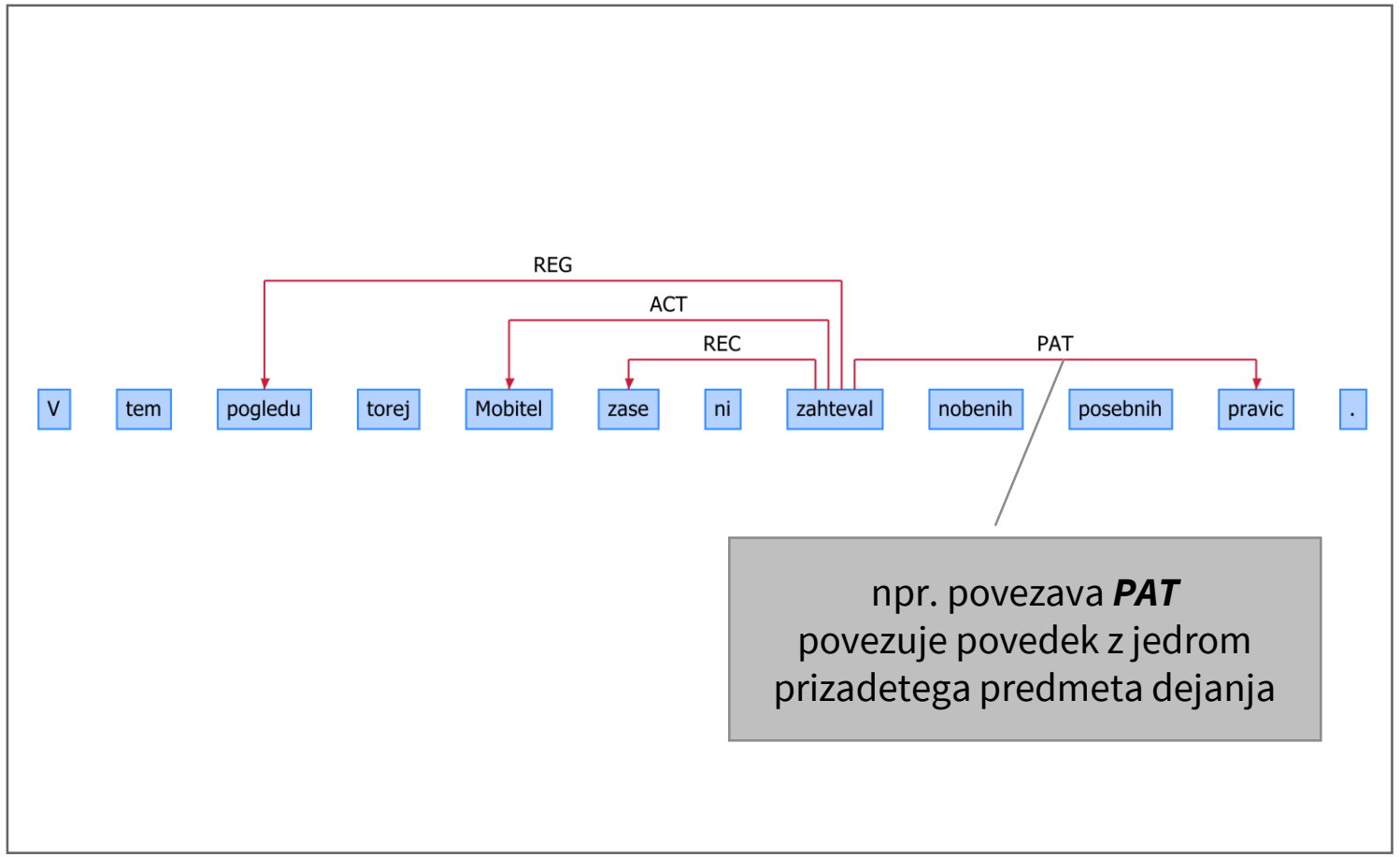
Skladnja JOS



- razčlenjevanje površinske **skladenjske zgradbe povedi** (teorija odvisnostne skladnje)
- označevalni sistem **JOS** (Erjavec et al. 2010; Holozan et al. 2008) – opredelitev **deset jedrnih tipov** skladenjskih povezav
- osredotočenost na **besednozvezno skladnjo** in **glagolsko vezljivost**
- označenih pribl. 40 % korpusa (11.400 povedi)

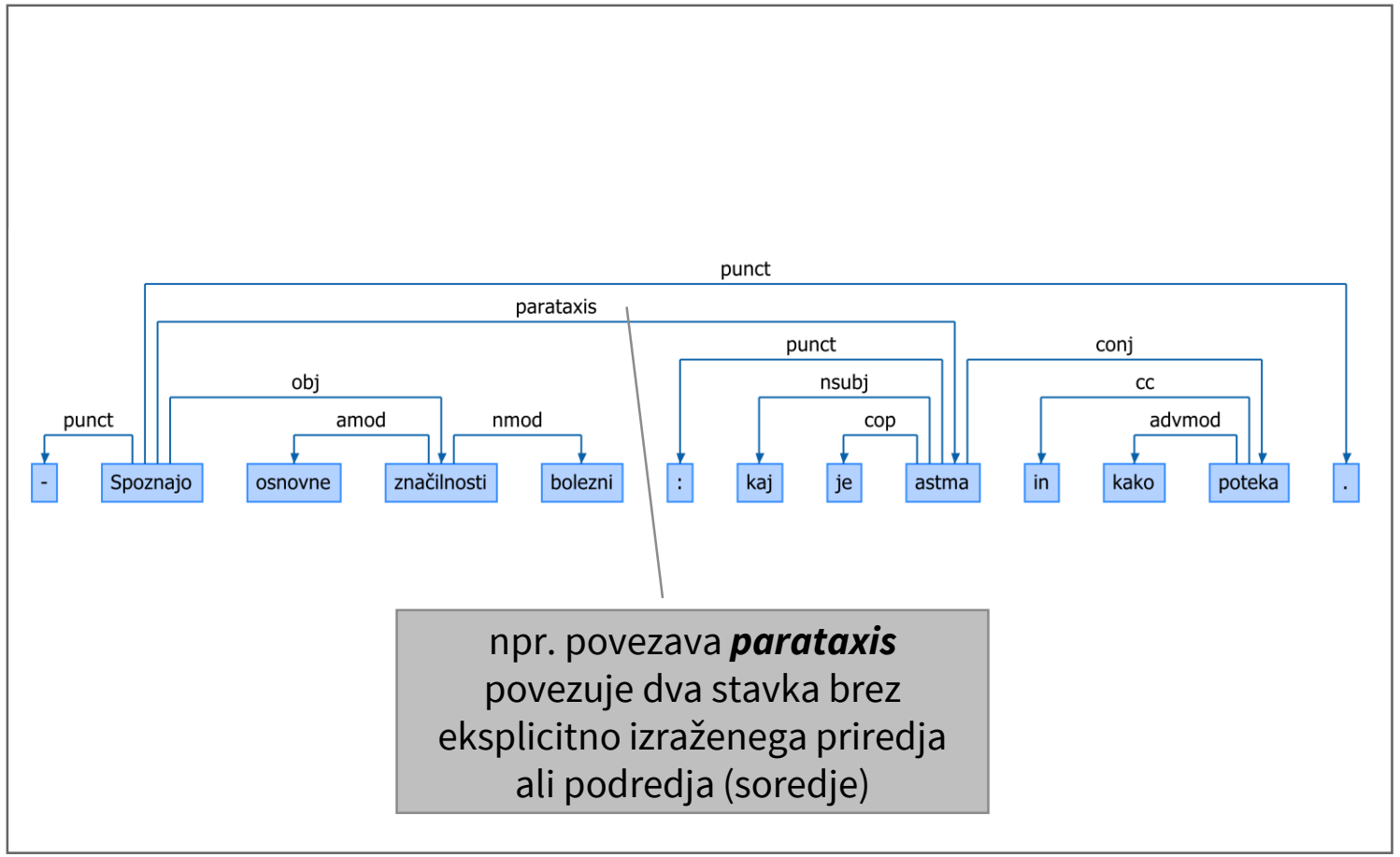


Udeleženske vloge



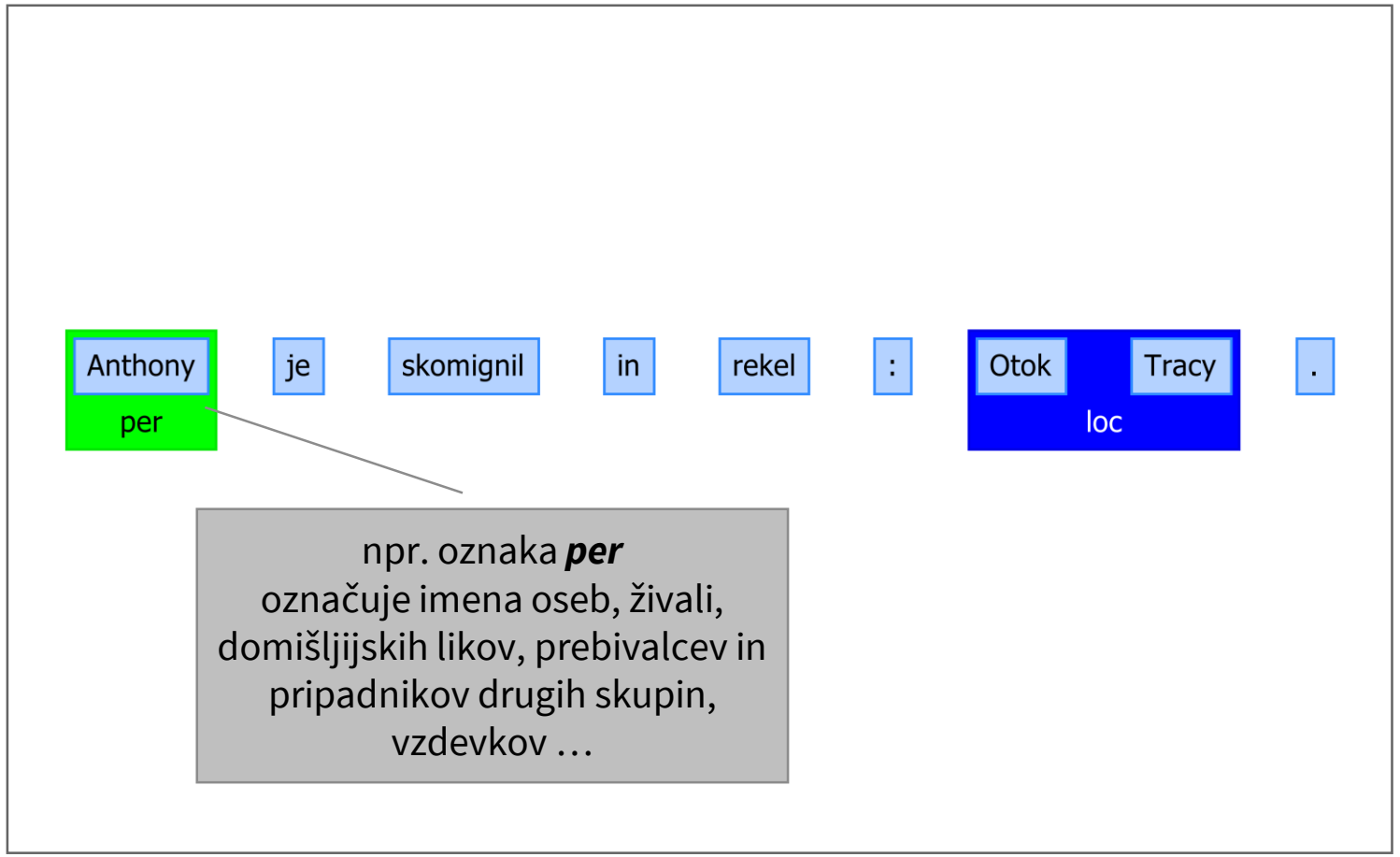
- razčlenjevanje **pomenske zgradbe povedi** – povedje in njegova določila
- slovensko-hrvaška **označevalna shema SRL** (Gantar et al. 2018) na podlagi češke in drugih sorodnih tipologij
- **25 udeleženskih vlog**: 5 delovalnikov, 17 prislovnih dopolnil, 3 povezave za stalne zveze
- označenih pribl. 19 % korpusa (5.501 povedi)

Universal Dependencies



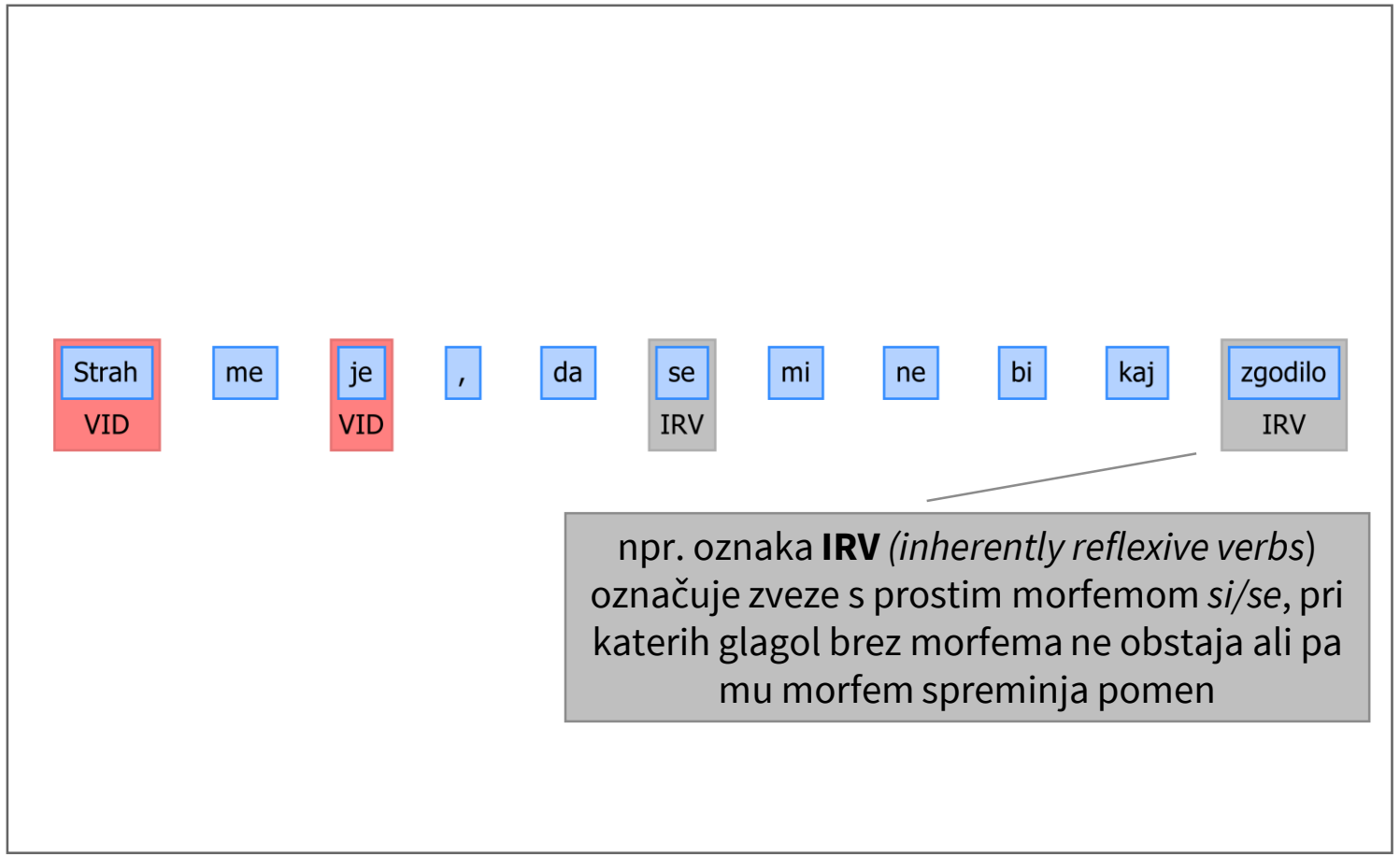
- **mednarodno poenoten** sistem oblikoslovnega in skladenjskega označevanja korpusov
- 37 tipov skladenjskih povezav, ki pokrivajo **celotno strukturo povedi**
- **strojna pretvorba** na podlagi pravil preslikave med JOS in UD (Dobrovoljc et al. 2017)
- označenih 100 % (oblikoslovje) oz. 24 % (8.000 razčlenjenih povedi) korpusa

Imenske entitete



- označevanje **lastnoimenskih** besed in besednih zvez
- smernice, razvite znotraj projekta **JANES** (Fišer et al. 2018): http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf
- **5 tipov entitet** (osebe in izpeljani svojilni pridevniki, kraji, organizacije, drugo)
- označenih pribl. 33 % korpusa (9.488 povedi, 7.016 entitet)

Glagolske večbesedne enote



- označevanje **stalnih besednih zvez** z glagolsko sestavino
- mednarodno usklajen sistem označevanja, razvit znotraj akcije COST **PARSEME** (Candito et al. 2016; Gantar et al. 2018)
- **4 tipi zvez**: idiomi**,** zveze s pomensko oslabljenimi glagoli, predložnimi zvezami in povratnimi zaimki
- označenih pribl. 48 % korpusa (13.511 povedi, 3.364 enot)

Dostopnost korpusa ssj500k 2.2

- **podatkovna baza** na repozitoriju CLARIN.SI
  - http://hdl.handle.net/11356/1210 
  - različni formati (XML TEI, .vert, CONLL-U)
- konkordančnika **noSketchEngine** in **Kontext**
  - https://www.clarin.si/noske/run.cgi/corp_info?corpname=ssj500k22
  - https://www.clarin.si/kontext/first_form?corpname=ssj500k22
  - brskanje po besednih oblikah in njihovih oznakah
  - ogled konkordanc, izdelava frekvenčnih seznamov
- označevalno orodje **Q-CAT**
  - http://hdl.handle.net/11356/1262
  - naprednejše brskanje po kombinacijah oznak
  - spreminjanje ali dodajanje oznak

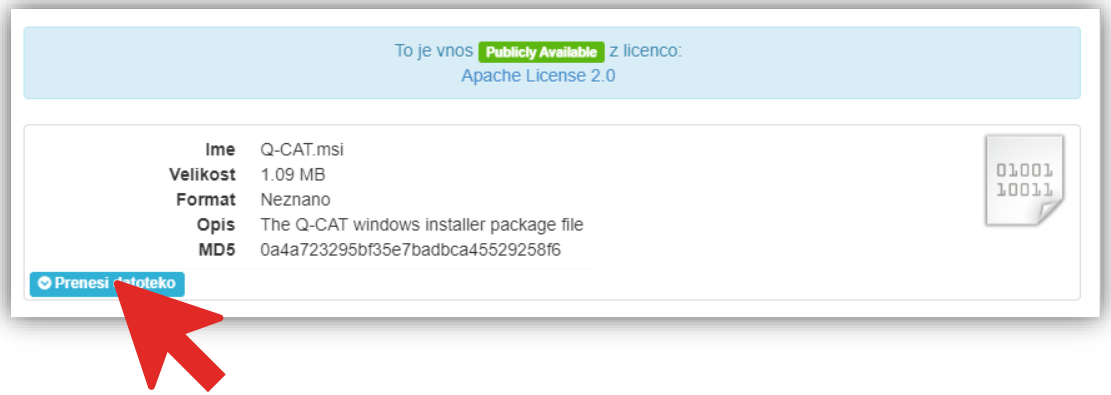
Orodje **Q-CAT**

Kaj je orodje Q-CAT

- Q-CAT: (*Querying-Supported) Corpus Annotation Tool*
- računalniški **program za izdelavo in analizo** ročno označenih besedilnih korpusov
- avtor: **Janez Brank** (IJS)
- prvotno razvit v okviru projekta Sporazumevanje v slovenskem jeziku (2008-2013)
  - pod imenom **SentenceMarkup**
  - skladenjsko razčlenjevanje in pripisovanje imenskih entitet
- nadgrajen v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode (2017–2020)
  - uvoz poljubnega korpusa s poljubnimi ravnmi označevanja
  - dinamično spreminjanje nastavitev
  - omogočena kompleksnejša iskanja
  - izboljšana uporabniška izkušnja

Dostopnost in namestitev

- prosto dostopen na repozitoriju CLARIN.SI
  - http://hdl.handle.net/11356/1262
- deluje na operacijskem sistemu **Windows**
- preprost za namestitev
  - prenos in zagon datoteke Q-CAT.msi
- dobro dokumentiran



Priročnik za uporabo:
https://bit.ly/32ypbhy

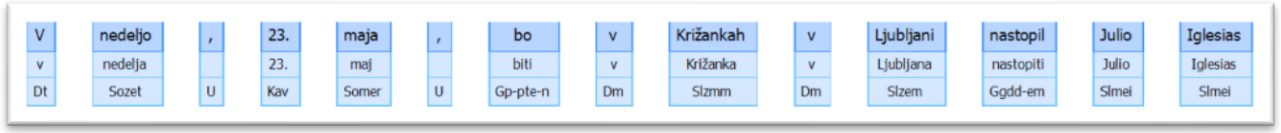
Prikaz uporabe na primeru korpusa ssj500k

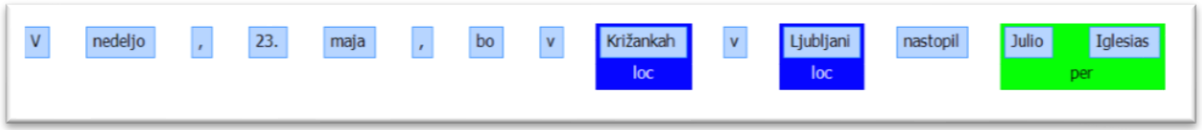
- okvirni scenarij
  1. **Uvoz** korpusa in **brskanje**
  2. **Iskanje** po označenem korpusu
     - iskanje po posameznih ravneh
     - kombiniranje iskanj po več ravneh
     - pregledovanje in shranjevanje zadetkov
  3. **Označevanje**
     - popravljanje ali dodajanje že opredeljenih oznak
     - oblikovanje nove ravni označevanja

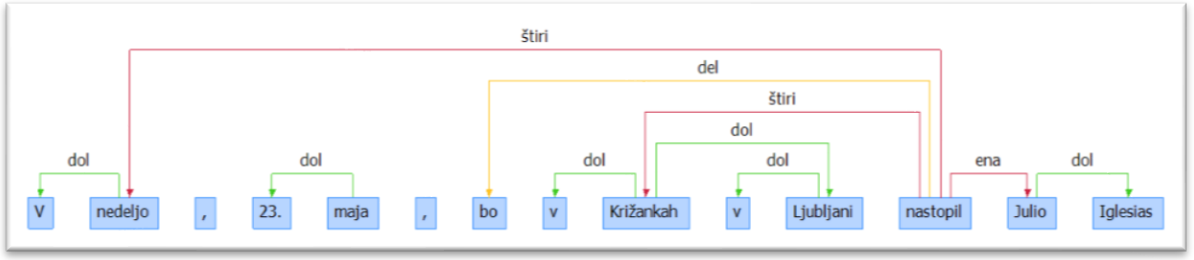
Priročnik za uporabo:
https://bit.ly/32ypbhy

Vrste oznak

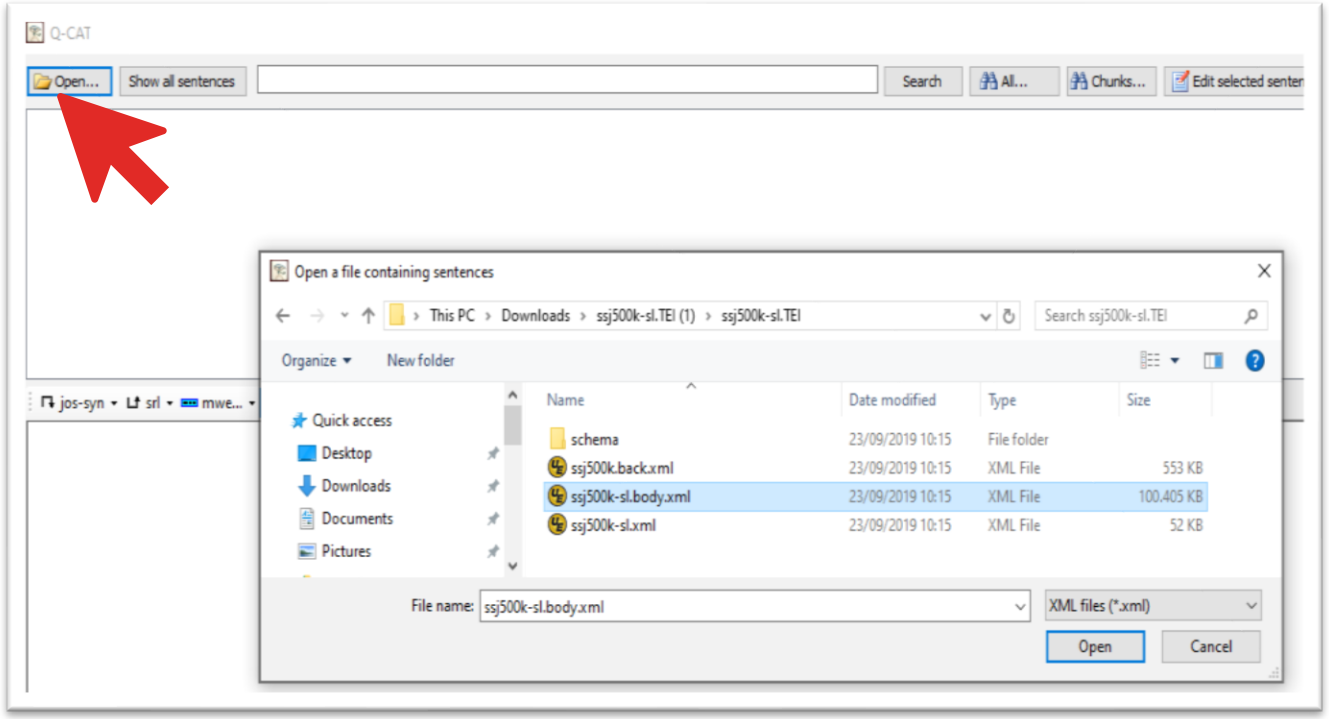
- **oznake oblik**
  - npr. oblika, lema, msd
  - modra polja
- **oznake nizov** (angl. *chunks*)
  - npr. stalne besedne zveze
  - barvne ploščice
- **oznake povezav** (angl. *links*)
  - npr. skladenjske povezave
  - barvne puščice



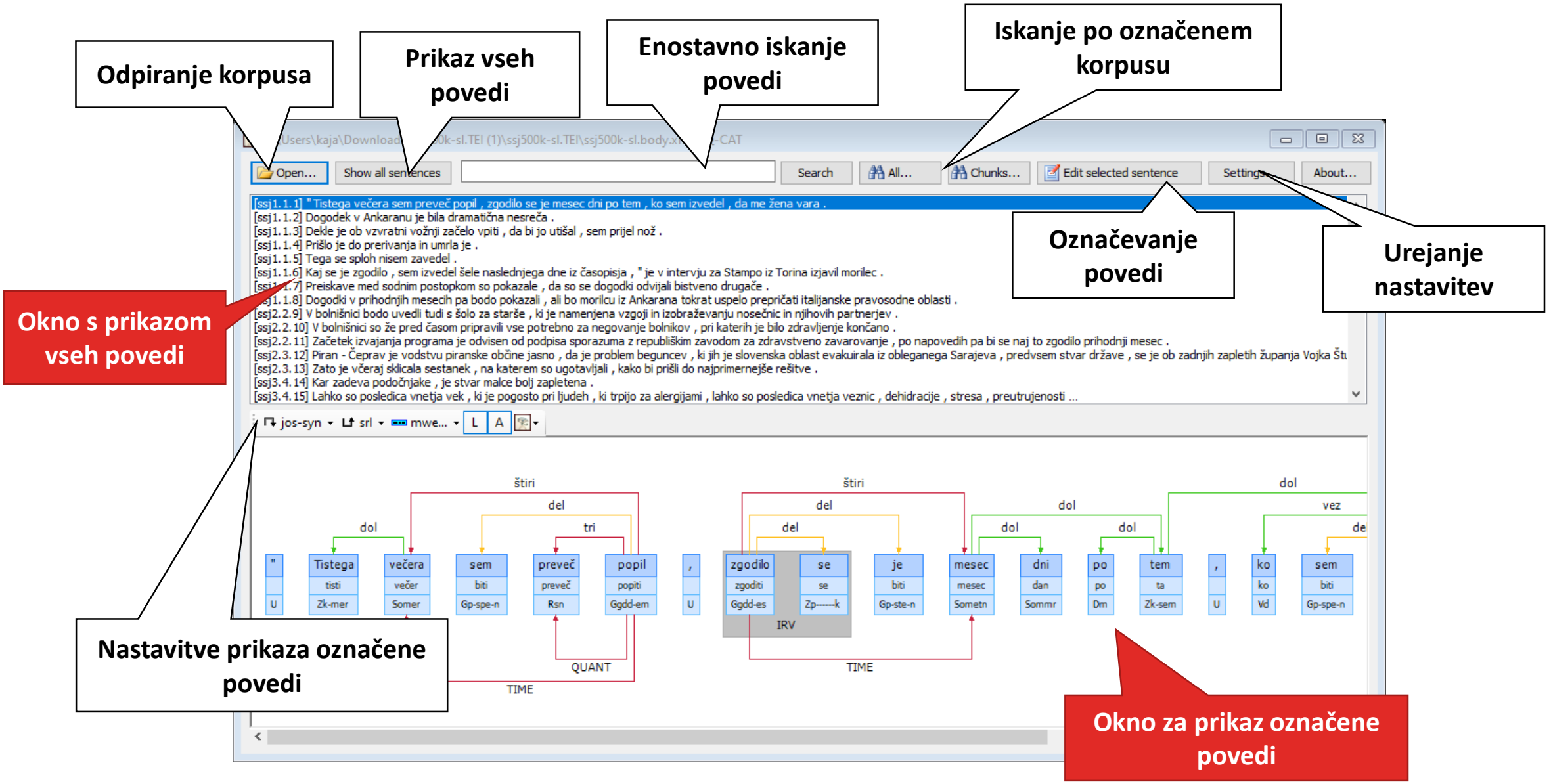




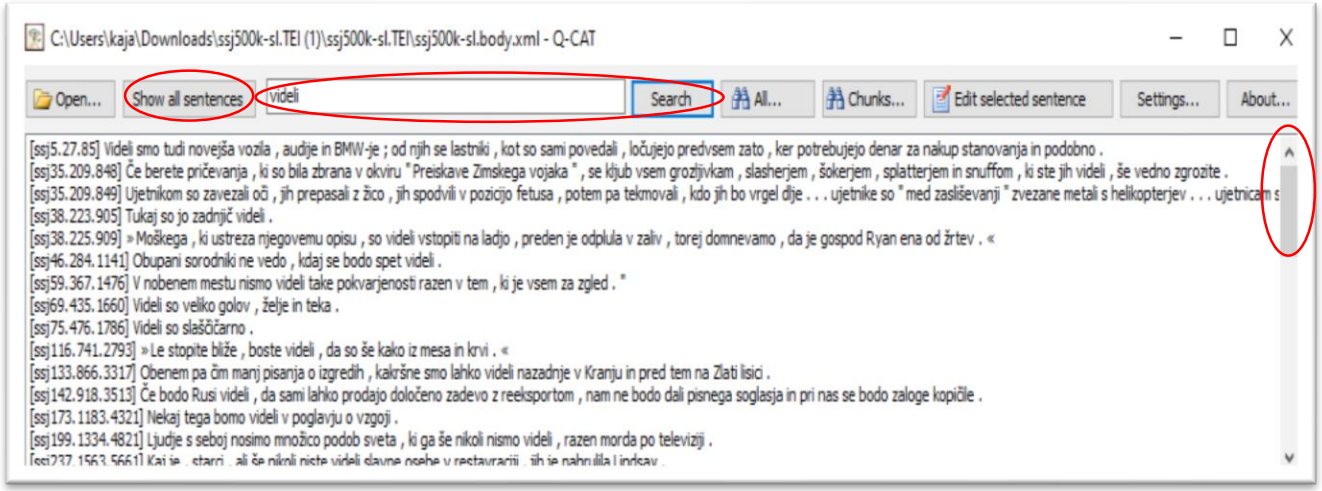
Uvoz korpusa



- zagon datoteke **QCat.exe**
- izbira korpusa v format XML (obvezno **segmentiran na besede in povedi**)
- brez dodatnih nastavitev lahko uvozimo
  - korpus **brez slovničnih oznak**
  - korpus **z eno ali več ravnmi označenosti ssj500k**
- v primeru uvoza korpusa z nepoznanimi oznakami, potrebne predhodne spremembe nastavitev

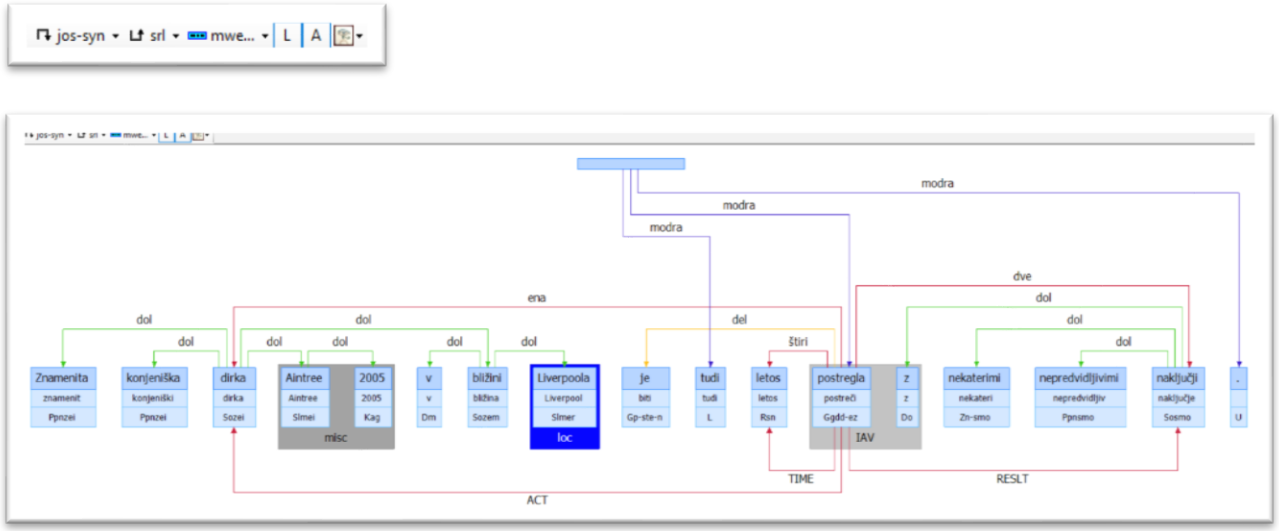

Odpiranje korpusa
Prikaz vseh povedi
Enostavno iskanje povedi
Iskanje po označenem korpusu
Okno s prikazom vseh povedi
Označevanje povedi
Urejanje nastavitev
Nastavitve prikaza označene povedi
Okno za prikaz označene povedi
Open...
Show all sentences
Search
All...
Chunks...
Edit selected sentence
Settings
About...
[ssj1.1.1] " Tistega večera sem preveč popil , zgodilo se je mesec dni po tem , ko sem izvedel , da me žena vara .
[ssj1.1.2] Dogodek v Ankaranu je bila dramatična nesreča .
[ssj1.1.3] Dekle je ob vzvratni vožnji začelo vpiti , da bi jo utišal , sem prijel nož .
[ssj1.1.4] Prišlo je do prerivanja in umrla je .
[ssj1.1.5] Tega se sploh nisem zavedel .
[ssj1.1.6] Kaj se je zgodilo , sem izvedel šele naslednjega dne iz časopisja , " je v intervju za Stampo iz Torina izjavil morilec .
[ssj1.1.8] Dogodki v prihodnjih mesecih pa bodo pokazali , ali bo morilcu iz Ankarana tokrat uspelo prepričati italijanske pravosodne oblasti .
[ssj2.2.9] V bolnišnici bodo uvedli tudi s šolo za starše , ki je namenjena vzgoji in izobraževanju nosečnic in njihovih partnerjev .
[ssj2.2.10] V bolnišnici so že pred časom pripravili vse potrebno za negovanje bolnikov , pri katerih je bilo zdravljenje končano .
[ssj2.2.11] Začetek izvajanja programa je odvisen od podpisa sporazuma z republiškim zavodom za zdravstveno zavarovanje , po napovedih pa bi se naj to zgodilo prihodnji mesec .
[ssj2.3.12] Piran - Čeprav je vodstvu piranske občine jasno , da je problem beguncev , ki jih je slovenska oblast evakuirala iz obleganega Sarajeva , predvsem stvar države , se je ob zadnjih zapletih županja Vojka Štu
[ssj2.3.13] Zato je včeraj sklicala sestanek , na katerem so ugotavljali , kako bi prišli do najprimernejše rešitve .
[ssj3.4.14] Kar zadeva podočnjake , je stvar malce bolj zapletena .
[ssj3.4.15] Lahko so posledica vnetja vek , ki je pogosto pri ljudeh , ki trpijo za alergijami , lahko so posledica vnetja veznic , dehidracije , stresa , preutrujenosti ...
jos-syn
srl
mwe...
L
A
štiri
del
dol
tri
QUANT
TIME
IRV
vez
Tistega
večera
sem
preveč
popil
zgodilo
se
je
mesec
dni
po
tem
ko
sem

Brskanje po povedih



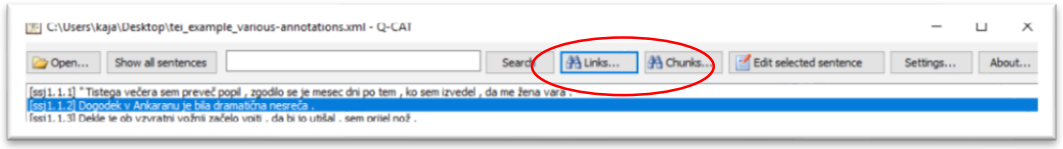
- po seznamu povedu se premikamo s puščicami ali drsnikom na desni
- iskalno okno za enostavno iskanje povedi išče po **besednih oblikah** (npr. *videli*)
- v primeru vnosa več besednih oblik (npr. *so videli*) program poišče povedi, ki jih vsebujejo, ne glede na dejanski vrstni red

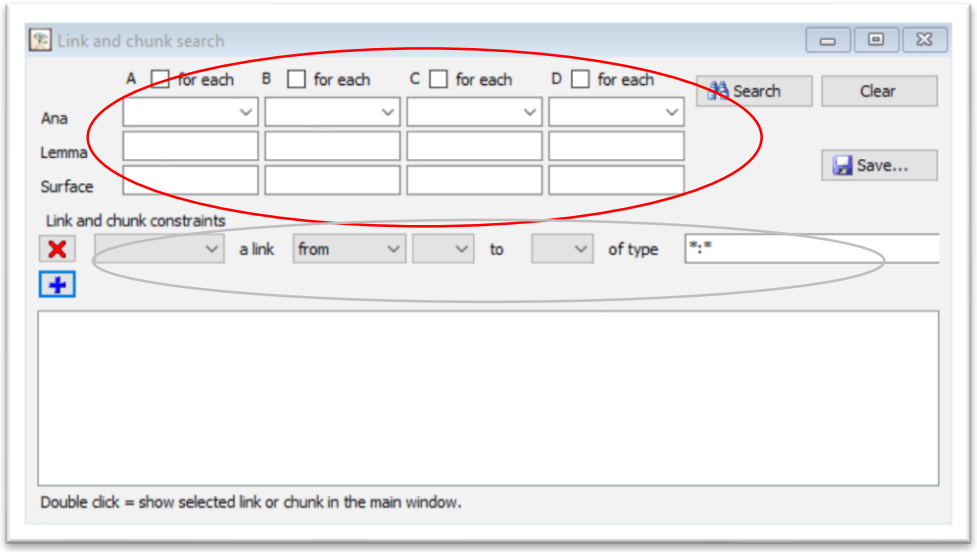
Prikaz označene povedi



- ob izbiri povedi se v spodnji polovici vmesnika prikaže poved z morebitnimi oznakami
- oznake (npr. lema, msd, povezave, stalne zveze …) lahko **poljubno vklapljamo in izklapljamo**

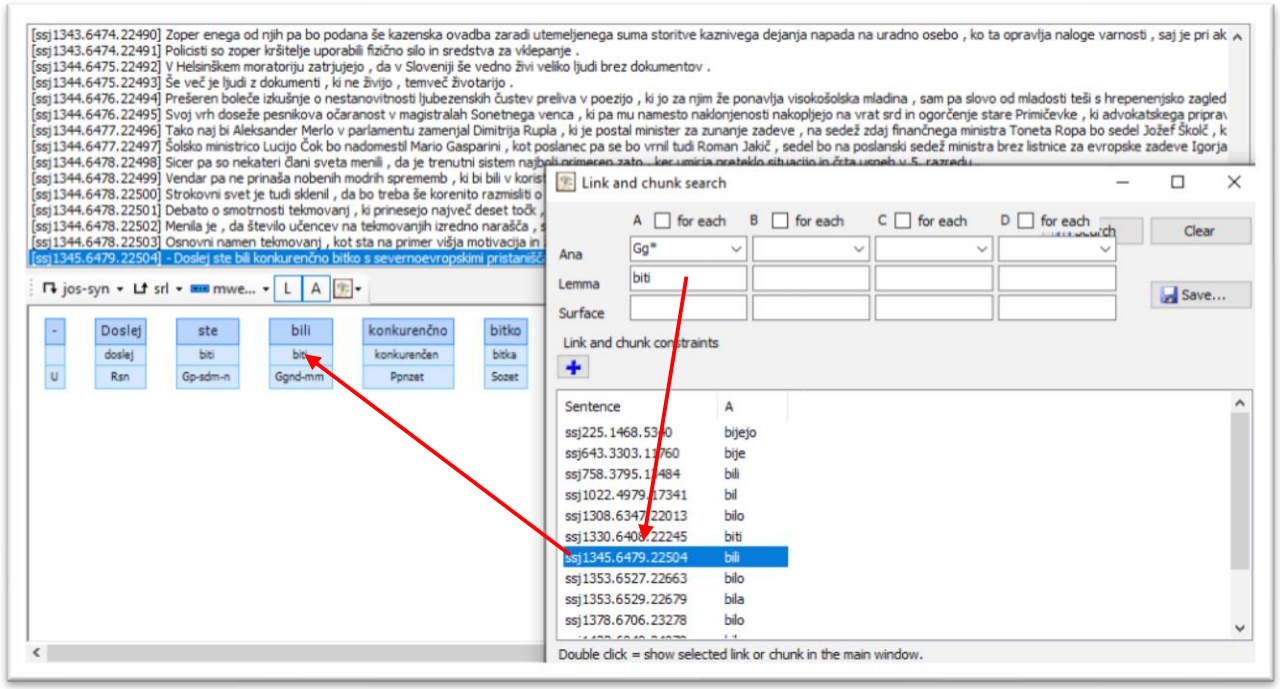
Naprednejša iskanja po korpusu





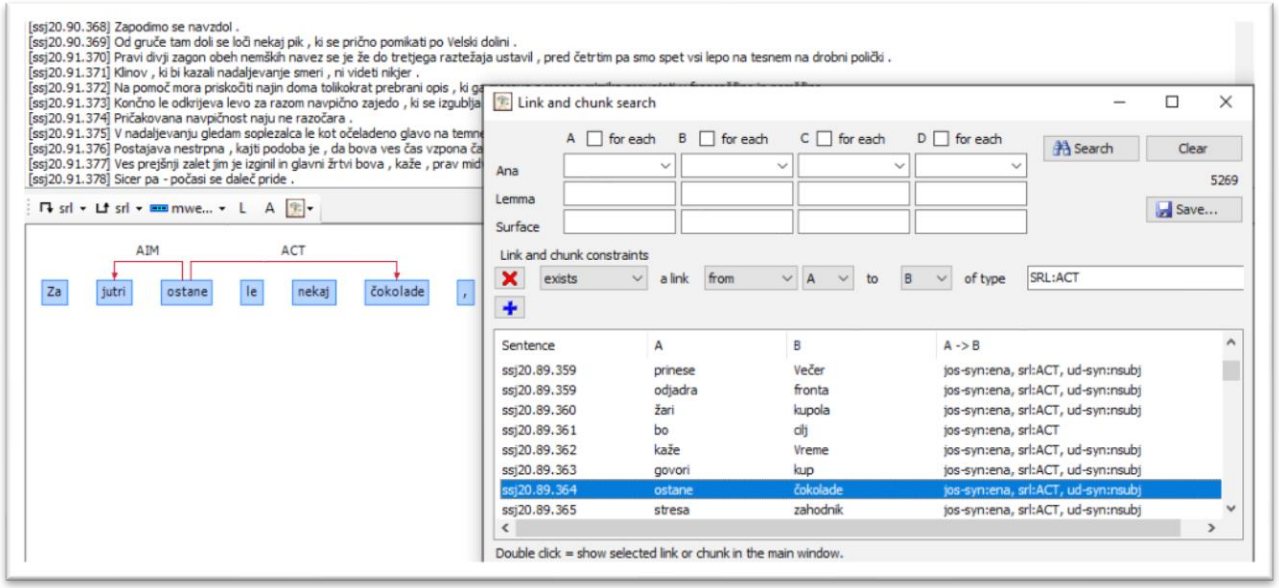
- orodje omogoča dve vrsti naprednejših iskanj po korpusu: **splošno iskanje** (*All*) in **iskanje zgolj po nizih** (*Chunks*)
- vmesnik za iskanje je sestavljen iz območja za opredelitev **lastnosti besed** (zgoraj) in območja za opredelitev **oznak nizov/povezav** (spodaj)

Rezultati iskanja



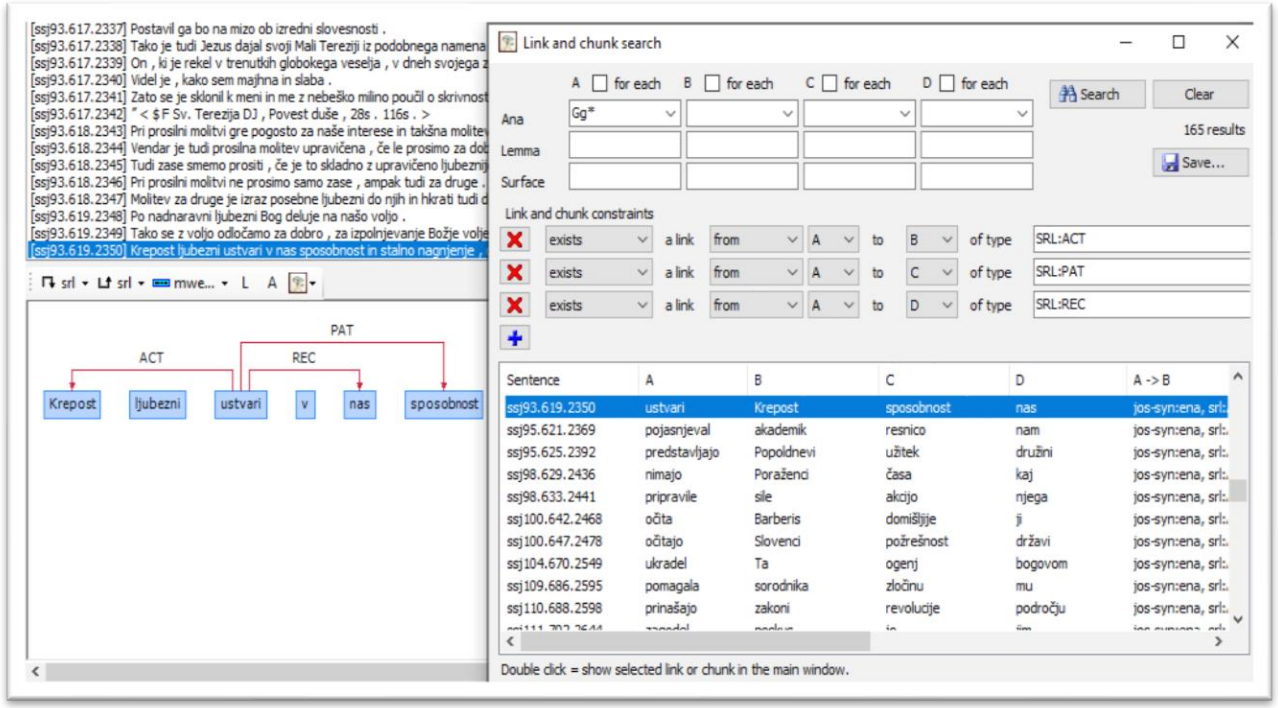
- iskanje vrne **število zadetkov** (desno zgoraj) in seznam vseh **ustreznih povedi** in opredeljenih pojavnic
- z dvojnim klikom na zadetek se v glavnem oknu v ozadju prikaže ustrezna poved

(Splošno) iskanje po eni vrsti povezave



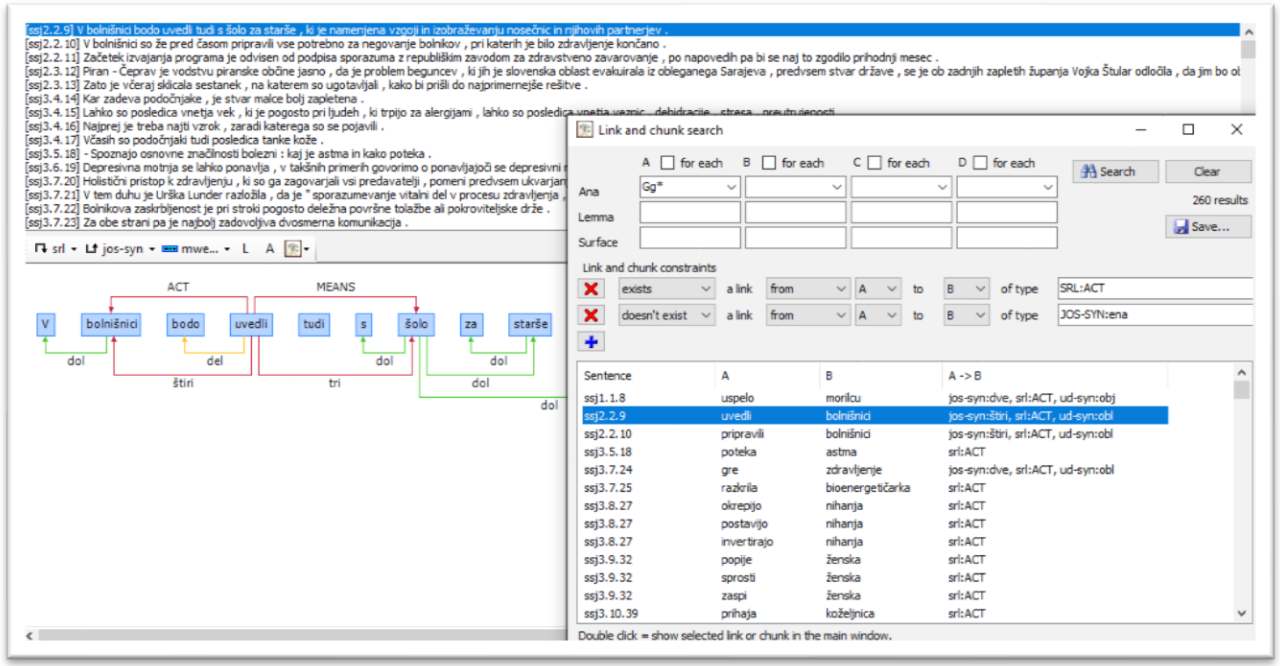
- S klikom na **modri znak plus** se odpre območje za omejevanje oznak povezav ali nizov, v katerem lahko opredelimo:
  - obstoj (*exists*) ali neobstoj (*doesn't exist)* določene povezave med dvema pojavnicama
  - konkretno (*from)* ali poljubno (*between*) smer povezave med dvema pojavnicami
  - označevalno raven in oznako povezave (polje *:*), npr. SRL:ACT
- iskanje lahko kombiniramo tudi z opredelitvijo dodatnih lastnosti pojavnic (zgoraj), npr. Gg*

(Splošno) iskanje po več povezavah hkrati



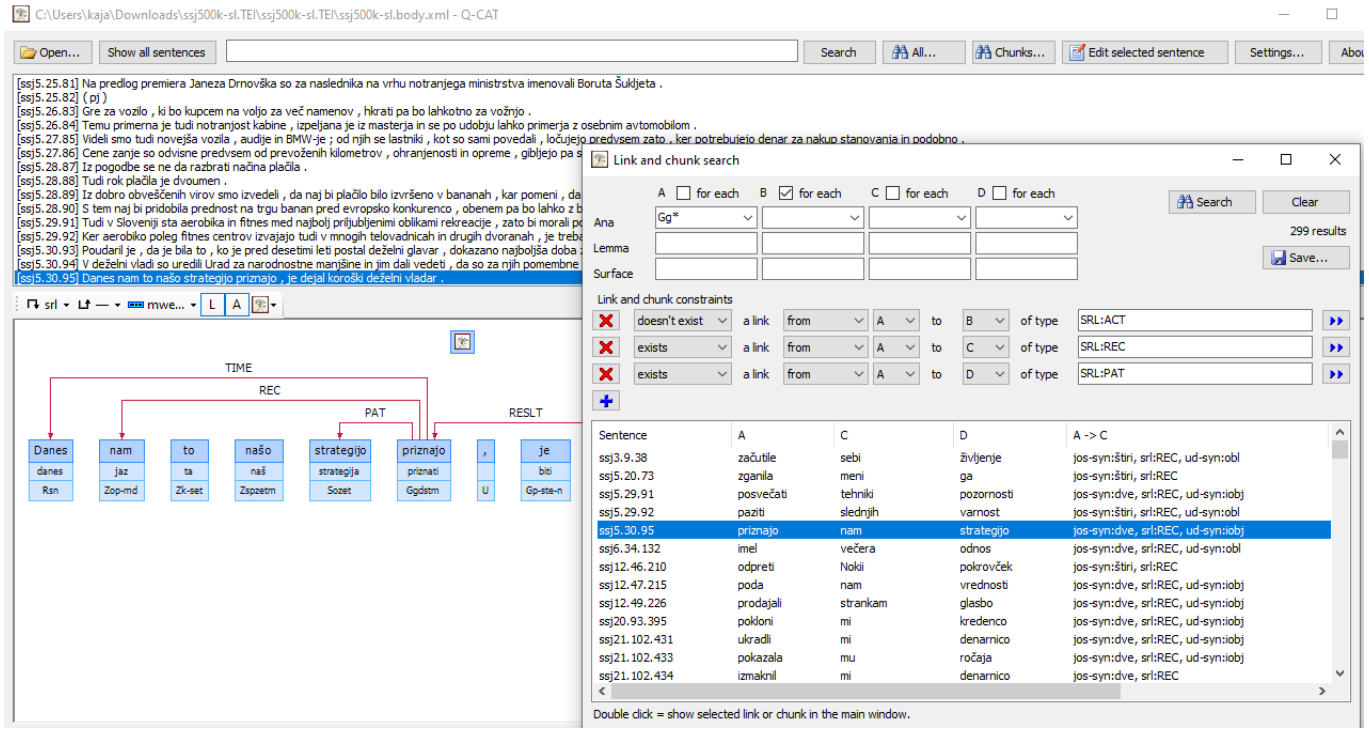
- S klikom na **modri znak plus** odpiramo nova polja za vnos dodatnih iskalnih pogojev
- npr. SRL:ACT + SRL:PAT + SRL:REC, če nas zanimajo povedi s povedki z vzorcem *kdo komu kaj*

(Splošno) iskanje po različnih ravneh hkrati



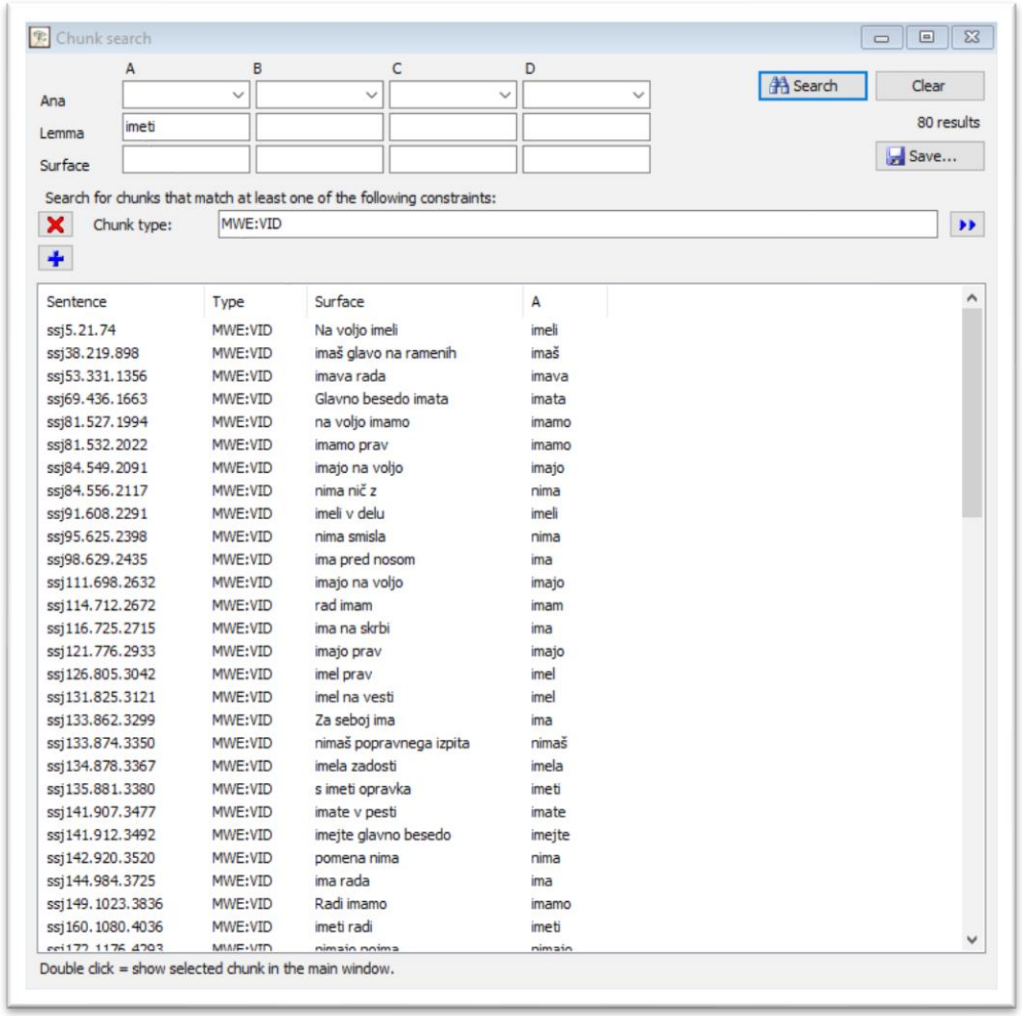
- hkratno iskanje lahko izvedemo tudi s povezavami na **več različnih ravneh označevanja**
- npr. iskanje vršilcev (SRL:ACT), ki niso osebki (JOS-SYN:ena)
- npr. iskanje vršilcev (SRL:ACT) v stalnih besednih zvezah (MWE:*)

Kompleksnejše iskanje z negacijo



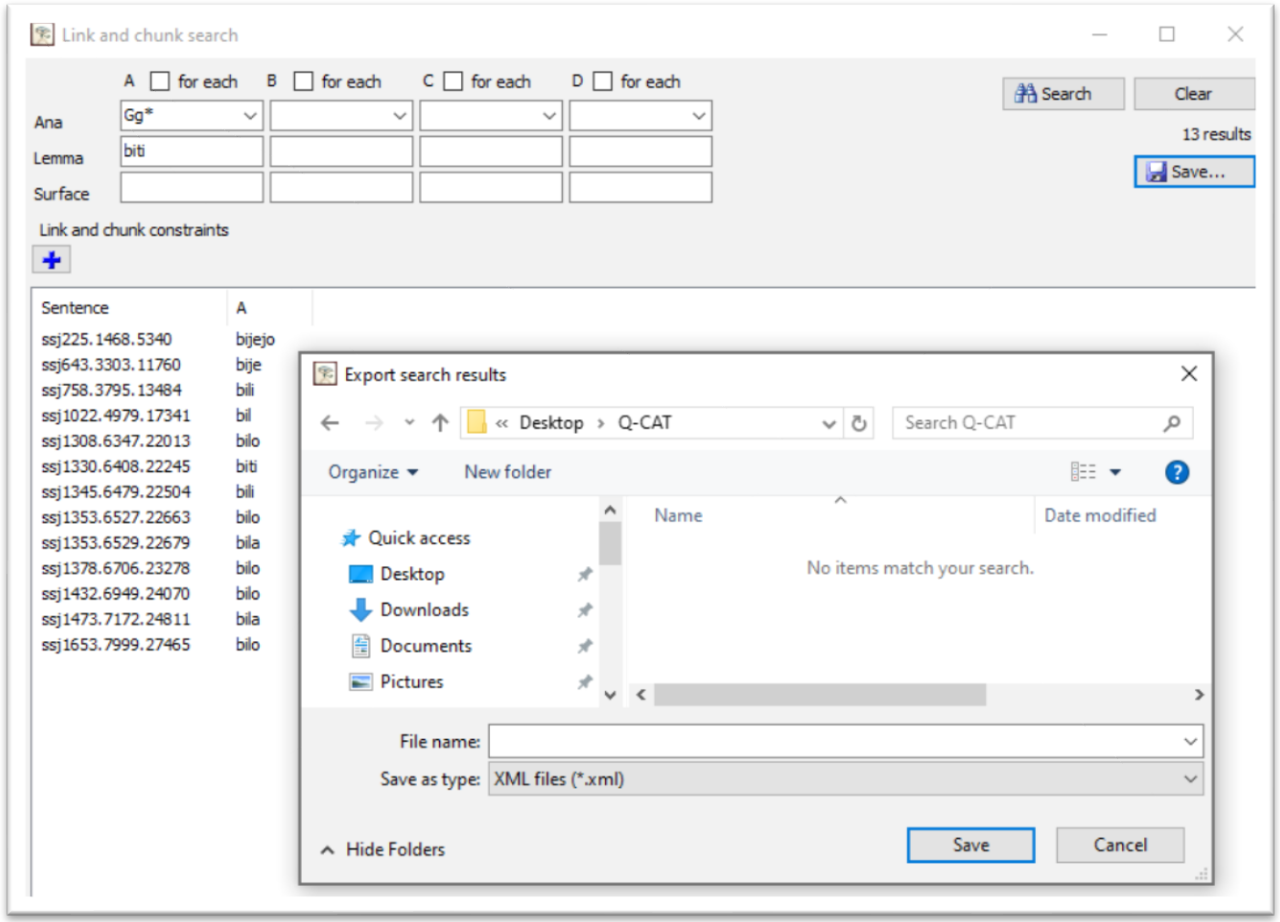
- kadar funkcijo negacije uporabimo za povezave med več različnimi pari pojavnic (npr. A in B ter A in C), moramo opredeliti tudi, ali negacija velja za vse možne kombinacije pojavnic ali ne.
- Temu je namenjeno polje ***for each*** pri vsaki pojavnici, ki ga izberemo, če mora izbrani pogoj veljati za vsako pojavnico v povedi.

Iskanje po nizih



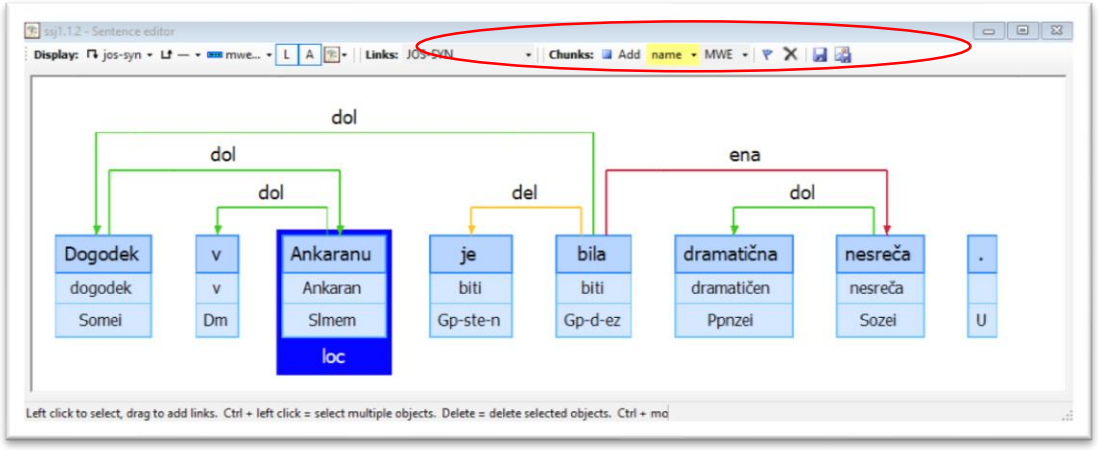
- deluje podobno kot splošno iskanje, razen:
  - vsa **iskanja se izvajajo izključno po pojavnicah z oznakami nizov** (npr. imenske entitete, MWE)
  - pri izpisu rezultatov se v stolpcu *Surface* **izpiše celotna zveza**
- npr. iskanje leme *imeti* vrne vse stalne zveze s to lemo
- npr. iskanje leme *imeti* s povezavo MWE:VID vrne vse frazeme s to lemo

Shranjevanje rezultatov iskanj



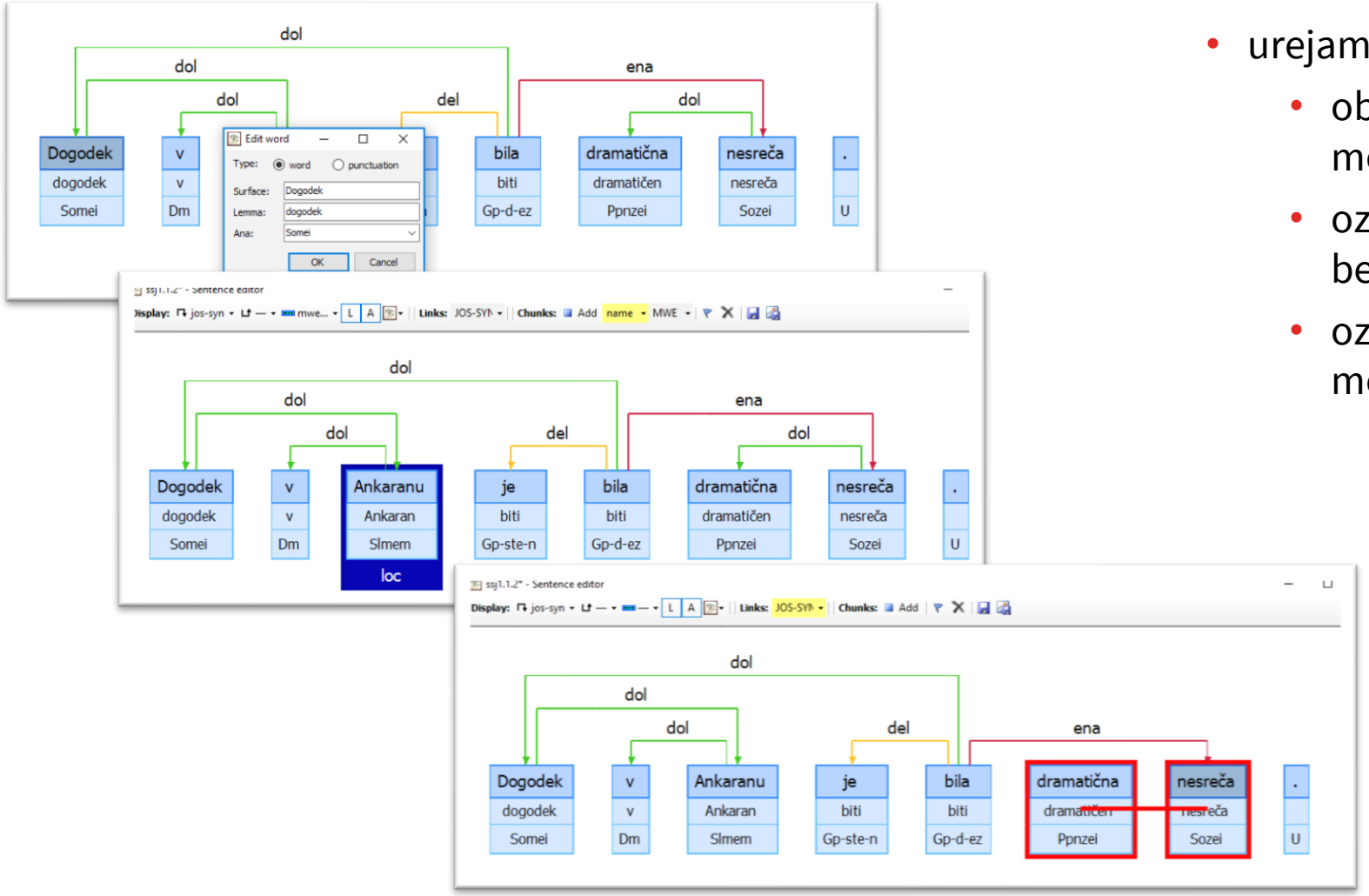
- rezultate iskanj lahko shranimo s klikom na gumb *Save*
- Z izbiro shranjevanja v obliki **XML** shranimo vse povedi v enaki obliki, kot je poved zapisana v izhodiščnem korpusu, z vsemi oznakami vred. S tem načinom shranjevanja torej **ustvarimo podkorpus** izhodiščnega korpusa.
- Z izbiro shranjevanja v obliki **.txt** shranimo vse rezultate v obliki podobnega **seznama**, ločenega s tabulatorji, kot se prikaže pri rezultatih iskanja v orodju. Uporabno za nadaljni pregled v Excelu.

Označevanje korpusa



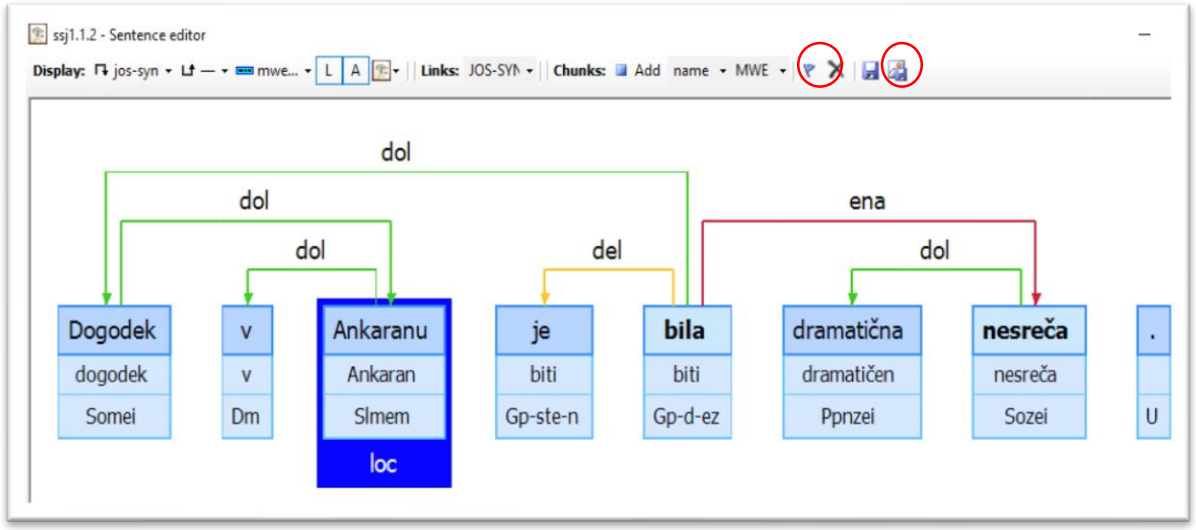
- s klikom na *Edit selected sentence* v zgornji vrstici vmesnika se odpre okno za urejanje povedi
- v njem **vklopimo vsaj prikaz ravni, ki jo želimo urejati**, lahko pa tudi druge

Označevanje korpusa



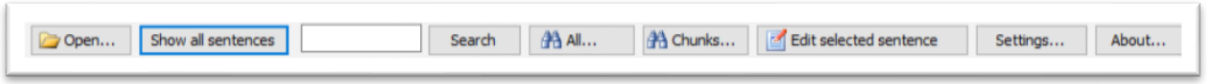
- urejamo lahko:
  - oblike, leme, msd-je (dvoklik na modra polja pod besedilom)
  - oznake nizov (klik na ploščice pod besedami ali besednimi zvezami)
  - oznake povezav (klik na puščice med besedama)

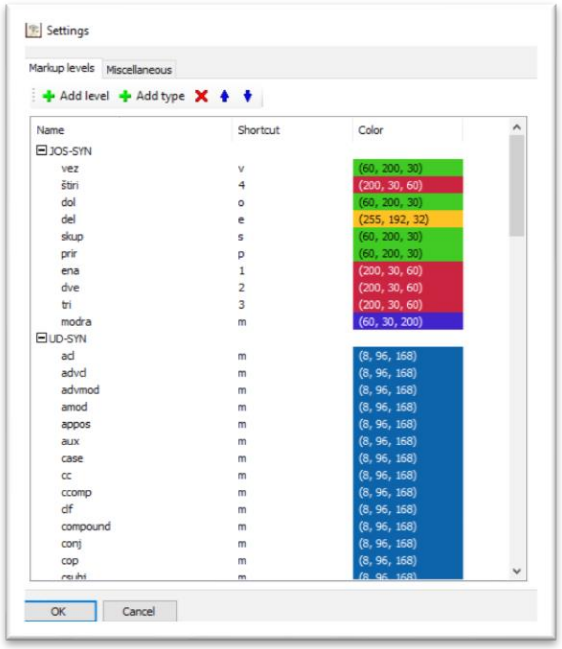
Shranjevanje in izvoz slike



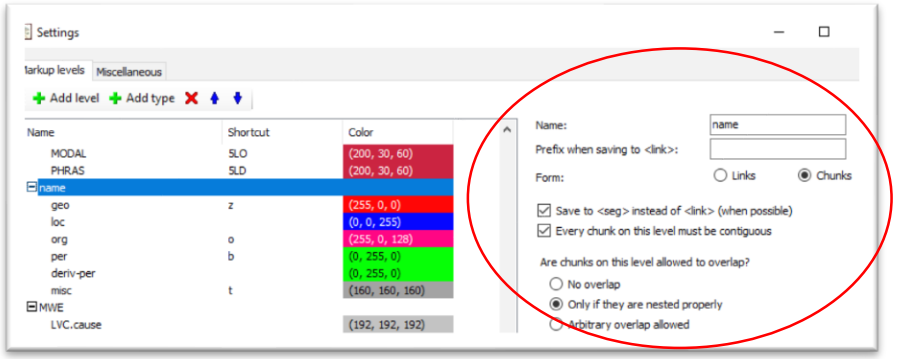
- spremembe shranimo s klikom na ikono za shranjevanje oz. nas na to opozori program ob izhodu iz okna za urejanje. Če ne določimo drugače, se spremembe **shranijo v izvorno datoteko**.
- označeno poved lahko tudi **shranimo kot sliko** visoke ločljivosti

Nastavitve označevanje





- v vmesniku za nastavitve označevanja (gumb *Settings*) lahko **spremenimo nastavitve obstoječih ravni** označevanja, npr.
  - dodajanje/brisanje/omejevanje oznak
  - urejanje bližnjic
  - urejanje barv
- ali **ustvarimo nove ravni** označevanja



Format

- XML TEI (pomoč na info@clarin.si)
- korpus mora biti vsaj segmentiran na povedi/besede, lahko pa vsebuje tudi:
  - oznake oblik (npr. lema, msd)
  - ena ali več označevalnih ravni iz ssj500k
  - poljubne druge oznake nizov/povezav – spremembe nastavitev pred začetkom dela
- primeri v mapi Programi/IJS/Q-CAT/Samples

Zaključek

- predstavili **korpus ssj500k**, največjo prostodostopno zbirko slovenskih besedil z ročno pripisanimi slovničnimi oznakami
- poleg uporabnosti za razvoj jezikovnih tehnologij ima korpus velik potencial za **kvalitativne korpusne analize** slovničnih pojavov v slovenščini
- ta analiza bistveno olajšana z **orodjem Q-CAT**, ki je prenosljivo tudi na druge korpuse

- poleg obstoječih ravni označenosti ssj500k **v pripravi** še:
  - stalne besedne zveze po smernicah LBS (Gantar 2015; Gantar et al. 2019)
  - oznake koreferenčnih povezav (Žitnik in Bajec 2018)
  - odstranjevanje nestandardnih besedil (projekt NSSS)
  - …

- ssj500k kot preizkusni kamen za druge jezikoslovne teorije in tipologije v našem prostoru?

Bibliografija

- Krek, S. et al. (2019). **The ssj500k Training Corpus for Slovenian Language Processing**. *V pripravi*.
- Dobrovoljc, K. (2019). **Q-CAT: Orodje za ročno označevanje in analizo besedilnih korpusov. Priročnik za uporabo.** https://bit.ly/32ypbhy

- Arhar, S. and Gorjanc, V. (2007). Korpus FidaPLUS: Nova generacija slovenskega referencnega korpusa (The FidaPLUS Corpus: A New Generation of the Slovene Reference Corpus). *Jezik in slovstvo*, 52(2):95–110.
- Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, pages 33–38.
- Erjavec, T. and Krek, S. (2008). The JOS morphosyntactically tagged corpus of Slovene. In *LREC 2008*.
- Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Erjavec, T. (2004). MULTEXT-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).
- Fišer, D., Ljubešić, N., and Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, Sep.
- Gantar, P., Arhar Holdt, Š., Čibej, J., Kuzman, T., and Kavčič, T. (2018a). Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. In *Proceedings of the conference on Language Technologies Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.
- Gantar, P., Despot, K. S., Krek, S., and Ljubešić, N. (2018b). Towards semantic role labeling in slovene and croatian. In *Proceedings of the conference on Language Technologies Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.
- Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S., and Velušček, A. (2008). Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku". Technical report.
- Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, , Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2019). *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/ 1210.

Hvala.

info@cjvt.si

Center za jezikovne vire in tehnologije

Večna pot 113
1000 Ljubljana
Slovenija

www.cjvt.si
00386 14798299
info@cjvt.si