LIST – orodje za kvantitativne slovnične raziskave korpusov

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256)

Jaka Čibej[123]

1 Institut Jožef Stefan 2 Filozofska fakulteta Univerze v Ljubljani 3 Fakulteta za računalništvo in informatiko Univerze v Ljubljani

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.

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256)

Kaj in zakaj

- Institut Jožef Stefan, Filozofska fakulteta UL, Fakulteta za računalništvo in informatiko UL
- Javna agencija za raziskovalno dejavnost Republike Slovenije (2017–2020)
- Nameni:
  - raziskati **jezikoslovne metodološke temelje** celostne **računalniške analize sodobne pisne in govorjene slovenščine**, kakršna je zajeta v **novih korpusih slovenskega jezika**
  - zagotoviti **empirično osnovo** za izdelavo novih **empirično zasnovanih slovničnih opisov** slovenskega jezika
  - izdelati obsežne prosto dostopne **korpusne baze podatkov**, ki bodo neposredno uporabne pri izdelavi bodočih jezikovnotehnoloških orodij in aplikacij za slovenski jezik
- Za uspešno raziskovanje znotraj novejših jezikoslovnih pristopov so potrebni zanesljivi empirični podatki o različnih jezikovnih pojavih, ki jih lahko zagotovi sodobno računalniško oz. korpusno jezikoslovje s strojno analizo obsežnih zbirk tako pisnega kot govorjenega jezika.
- Vse izluščene zbirke, programska oprema in drugi projektni rezultati bodo **prosto dostopni**.

Vsebina projekta

- 5 delovnih sklopov:
  - **DS1: Oblikoslovje in besedotvorje**
  - DS2: Kolokacije
  - DS3: Stalne besedne zveze
  - DS4: Vezljivost
  - DS5: Besedni nizi

Program LIST je nastal v okviru dejavnosti delovnega sklopa 1.

Delovni sklop 1: Oblikoslovje in besedotvorje

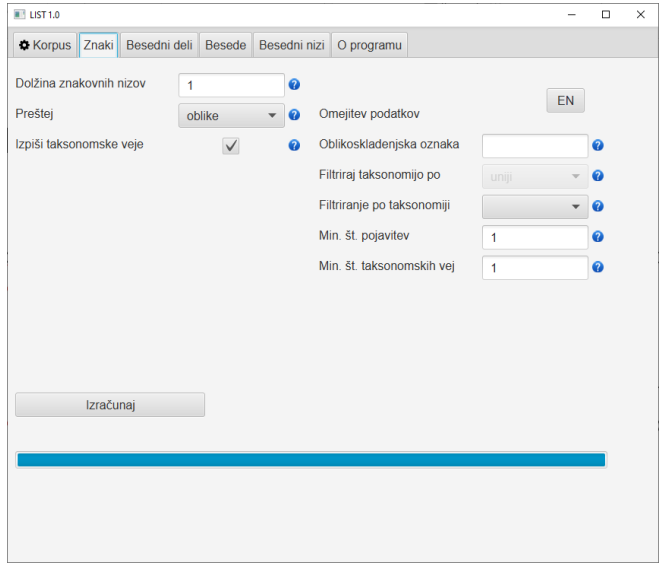
Vsebina Delovnega sklopa 1

- Cilji (med drugim):
  - razvoj metodologije za slovnični opis slovenščine na ravni oblikoslovja in besedotvorja
  - statistična analiza oblikoslovnih in besedotvornih procesov v sodobni standardni slovenščini
- Naloge:
  - metodologija **pridobivanja oblikoslovnih in besedotvornih korpusnih podatkov**
  - **zbirke** z oblikoslovnimi in besedotvornimi korpusnimi podatki
- Rezultati:
  - prostodostopna **programska oprema** za statistično obdelavo velikih korpusov
  - **baze podatkov** z izkorpusnimi oblikoslovnimi in besedotvornimi informacijami

Rezultati

- prostodostopna **programska oprema** za statistično obdelavo velikih korpusov
  - **LIST – korpusni luščilnik**

- **baze podatkov** z izkorpusnimi oblikoslovnimi in besedotvornimi informacijami
  - **frekvenčni seznami** iz korpusov Gigafida 2.0 in GOS 1.0 (odprto dostopni na CLARIN.SI)



Kaj je LIST?

O programu

- LIST je orodje za **luščenje frekvenčnih seznamov iz korpusov**
- računalniški program za kvantitativno analizo besedilnih korpusov
- omogoča izvažanje različnih frekvenčnih seznamov z upoštevanjem uporabniško določenih kriterijev za omejitev podatkov (npr. filtriranje po besedilnih zvrsteh, besednih vrstah, minimalnemu številu pojavitev).

Namen programa

- pridobivanje večjih količin frekvenčnih podatkov iz korpusov je zahtevno in/ali časovno potratno
- pomanjkanje uporabniške prijaznosti

- **Gigafida 2.0** – priprava seznamov podatkov iz najnovejše različice korpusa pisne standardne slovenščine, ki bodo na voljo vsem (raziskave in razvoj novih izdelkov)
- LIST **poenostavlja** postopek luščenja.
- Uporabniki lahko frekvenčne podatke luščijo sami, na lastnih osebnih računalnikih, brez tehnične podpore.

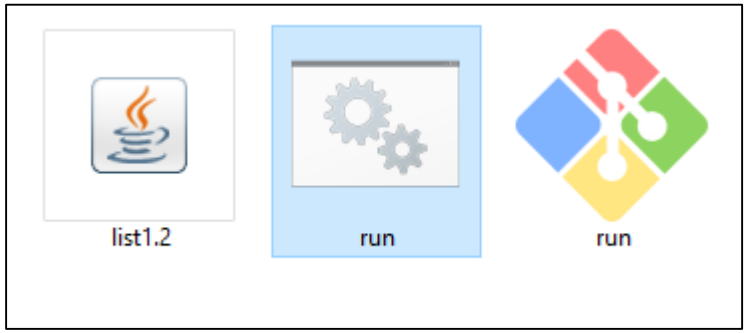
Zgodovina

- **2016**: pod imenom *CorpusStatistics* kot predmet diplomskega dela **Aleksandra Ključevška** z naslovom Statistična analiza slovenskih jezikovnih korpusov (FRI UL, mentor: dr. Marko Robnik Šikonja, somentor: dr. Simon Krek)
  - predstavljeno na konferenci JTDH 2018 (Ključevšek et al. 2018)
- Projekt *Nova slovnica*: **Luka Krsnik**, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Aleksander Ključevšek, Simon Krek, Marko Robnik Šikonja
  - preimenovanje v **LIST**
  - izpopolnjeno delovanje in vmesnik (uporabniška prijaznost)
  - podpora za najnovejši korpusni format (TEI P5 XML)
- Financiranje CLARIN.SI 2018: Projekt *Orodje za učinkovito analizo slovenskih korpusov*
  - dodan angleški vmesnik
  - dodana podpora za format VERT (SketchEngine)
  - podpora za tujejezične korpuse in nelatinične pisave

Kje je LIST na voljo?

- **Repozitorij CLARIN.SI:** https://www.clarin.si/repository/xmlui/
  - *Corpus extraction tool LIST*
- Priročnik za uporabo (v1.0): https://bit.ly/2qrdlZl

- Namestitev:
  - Naložimo datoteko .ZIP z repozitorija CLARIN.SI.
  - Razpakiramo datoteke v poljubno mapo.
  - Zaženemo run.bat (Windows) ali run.sh (Linux).

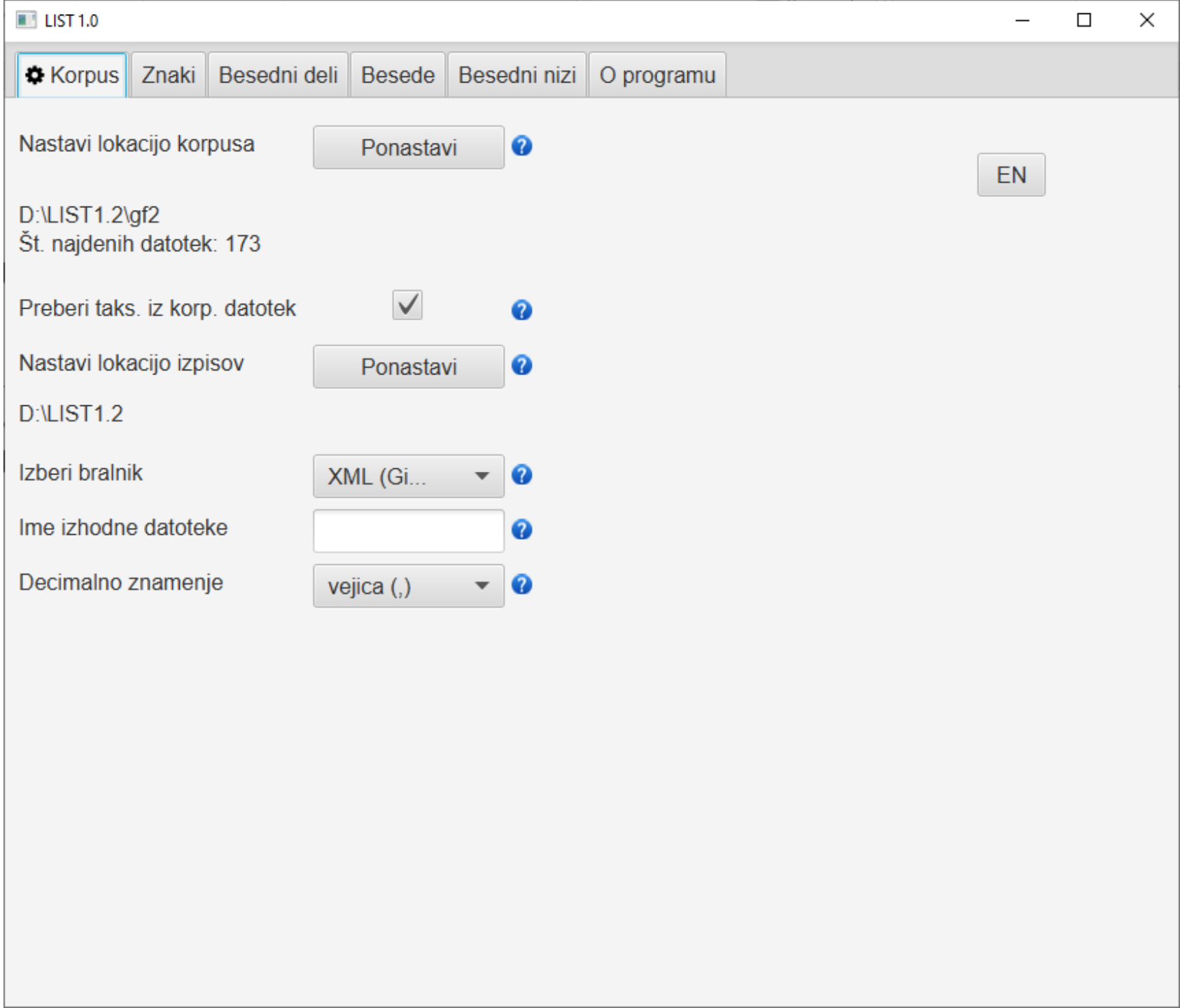


Kako deluje LIST?

cjvt

Zavihek Korpus

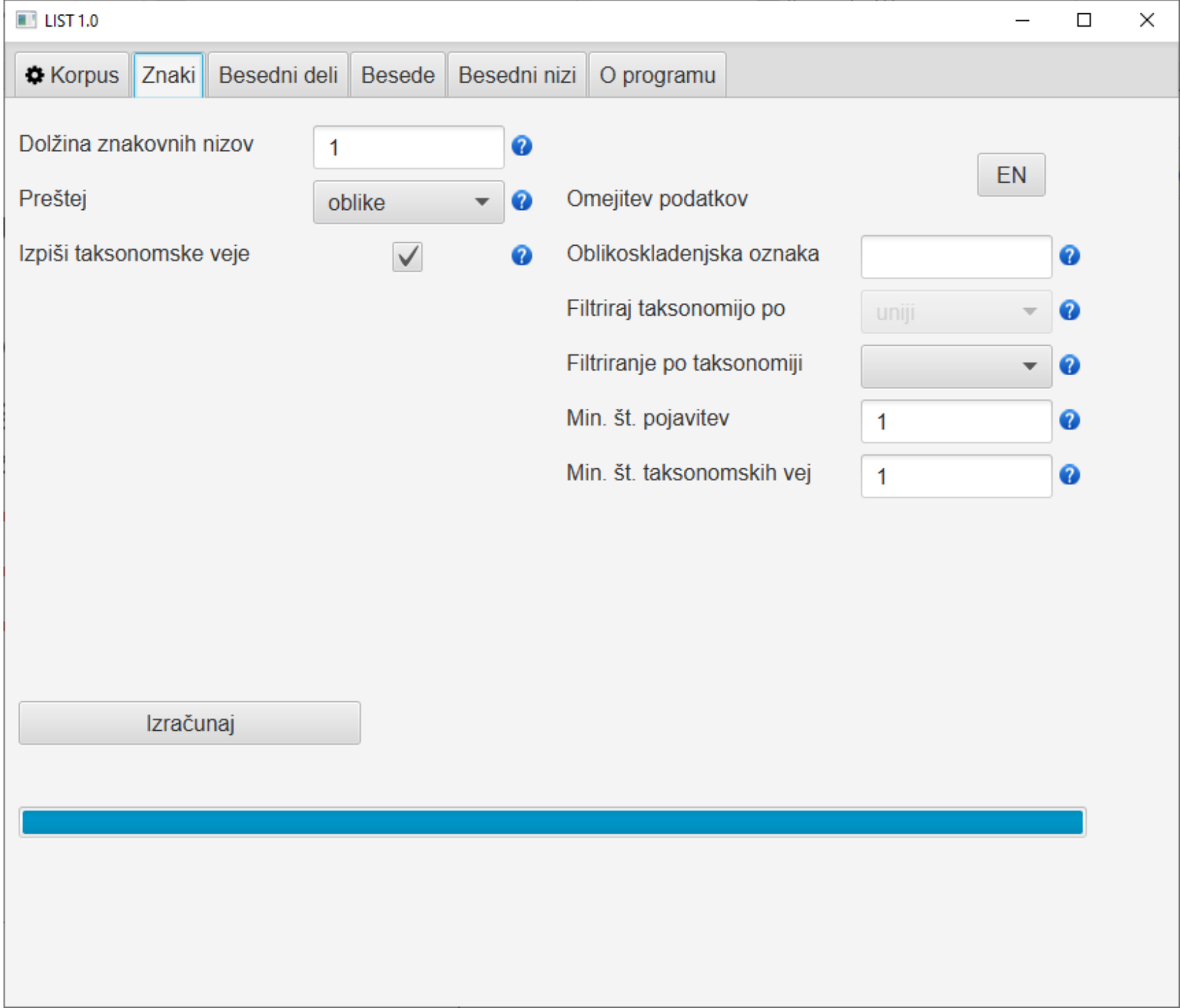
- Osnovne nastavitve
  - lokacija korpusa
  - lokacija izpisov
  - decimalno znamenje



cjvt

Zavihek Znaki

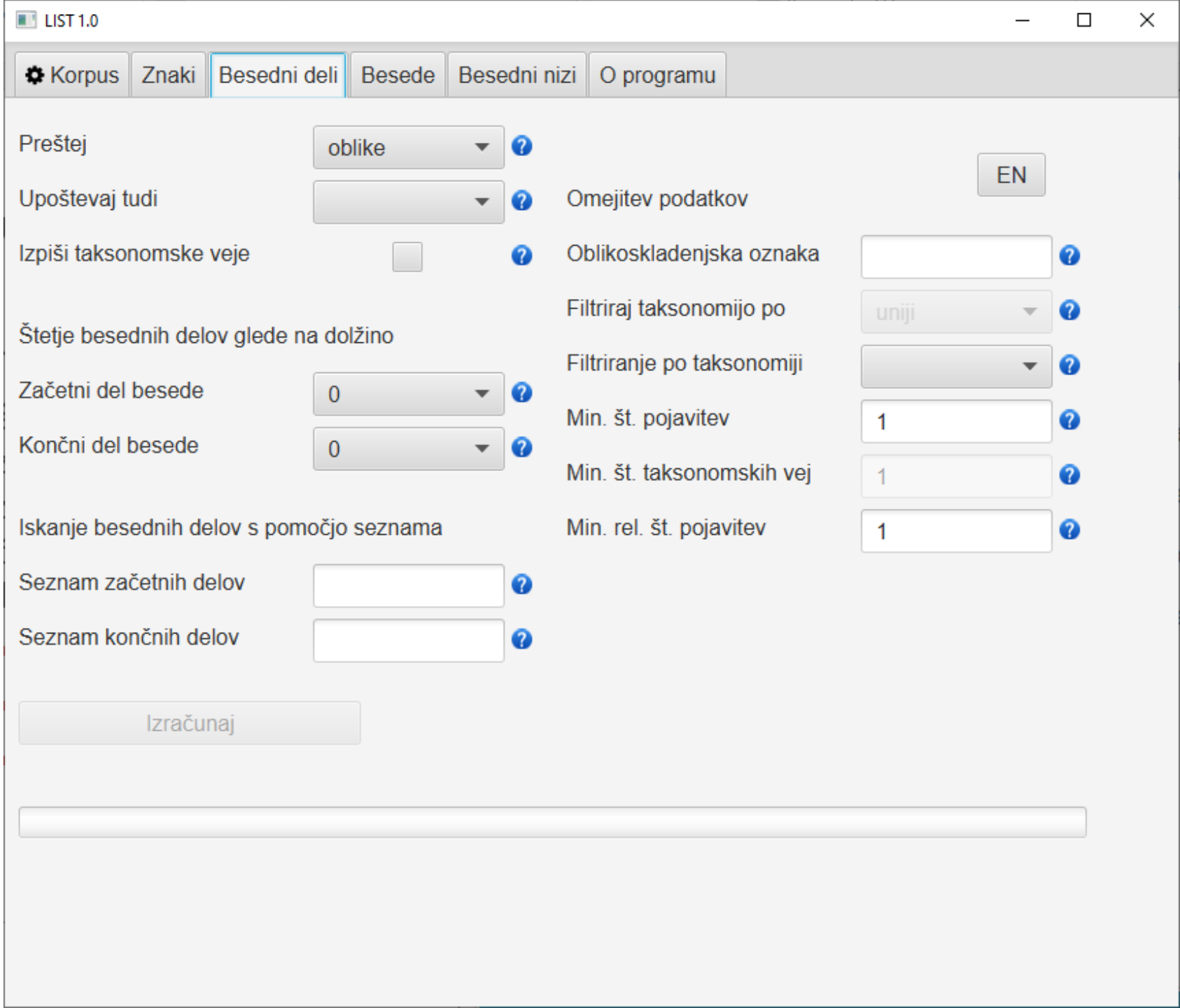
- Luščenje frekvenčnih seznamov znakovnih nizov
  - dolžina nizov
  - enote
  - taksonomija



cjvt

Zavihek Besedni deli

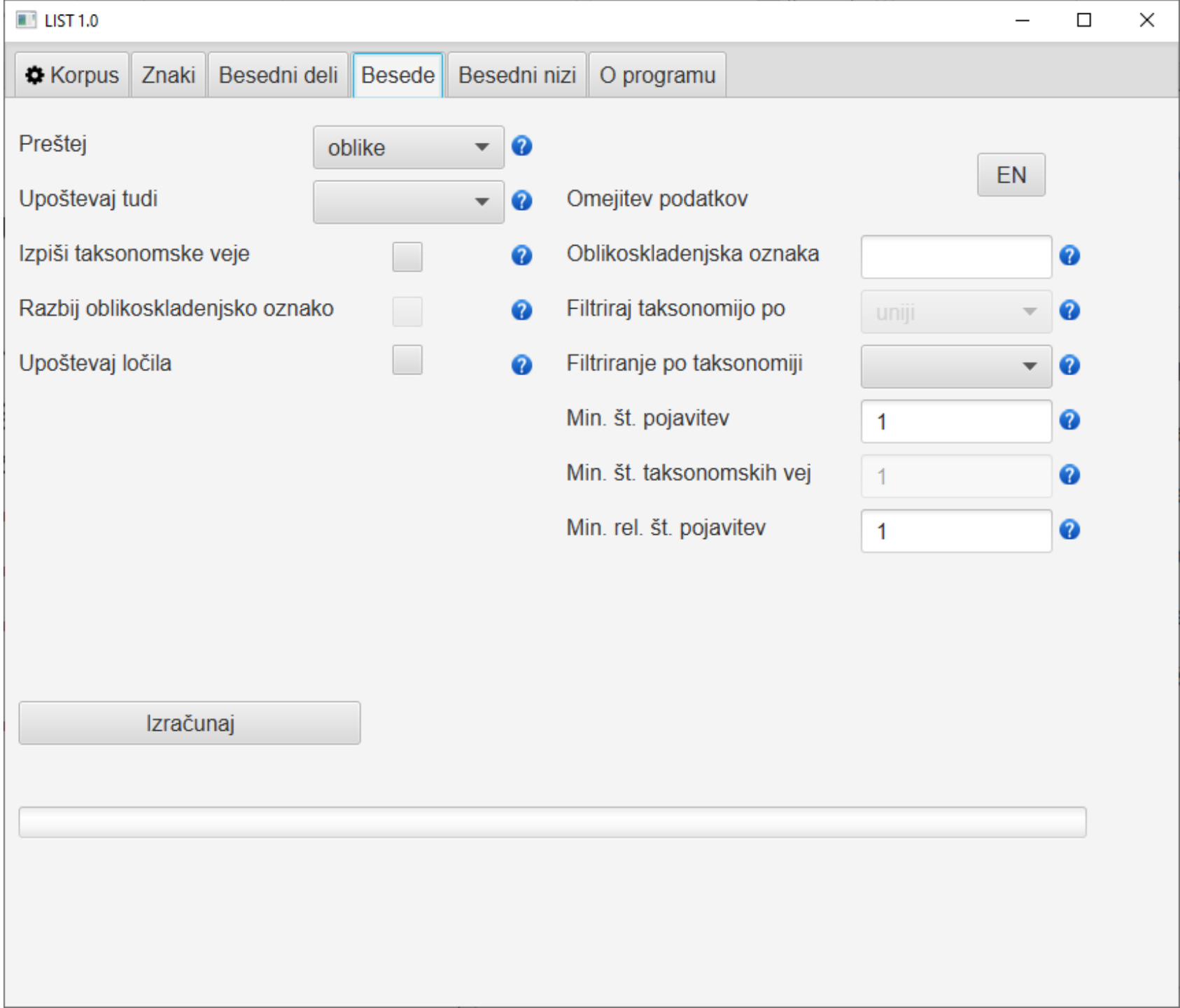
- Luščenje frekvenčnih seznamov enot, razcepljenih na začetne/končne besedne dele in preostanek enote
  - enote
  - taksonomija
  - štetje glede na dolžino
  - iskanje delov s seznamom



cjvt

Zavihek Besede

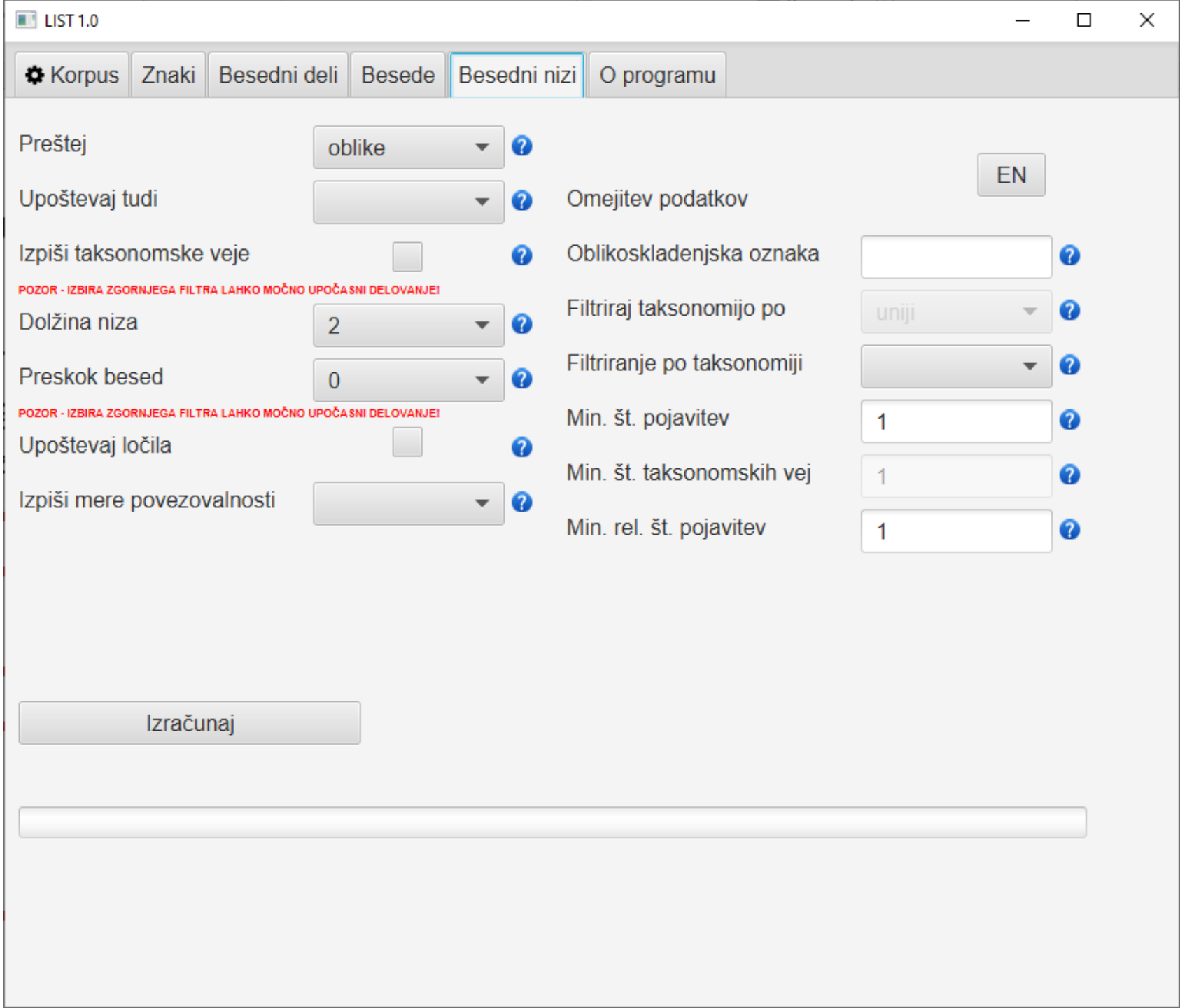
- Luščenje frekvenčnih seznamov enot
  - enote
  - dodatni izpisani podatki
  - taksonomija
  - razcep oblikoskladenjske oznake v ločene stolpce
  - upoštevanje ločil



cjvt

Zavihek Besedni nizi

- Luščenje frekvenčnih seznamov besednih nizov
  - enote
  - dodatni izpisani podatki
  - taksonomija
  - dolžina nizov (v besedah)
  - upoštevanje ločil
  - mere povezovalnosti



Frekvenčni seznami – Gigafida 2.0 in GOS 1.0

Gigafida 2.0

- **Seznami znakovnih n-gramov iz korpusa Gigafida 2.0: http://hdl.handle.net/11356/1272**
- **Seznami besednih delov iz korpusa Gigafida 2.0: http://hdl.handle.net/11356/1275**
- **Seznami besed iz korpusa Gigafida 2.0: http://hdl.handle.net/11356/1273**
- **Seznami besednih nizov iz korpusa Gigafida 2.0: http://hdl.handle.net/11356/1274**

GOS 1.0

- **Seznami znakovnih n-gramov iz korpusa GOS 1.0: http://hdl.handle.net/11356/1268**
- **Seznami besednih delov iz korpusa GOS 1.0: http://hdl.handle.net/11356/1270**
- **Seznami besed iz korpusa GOS 1.0: http://hdl.handle.net/11356/1269**
- **Seznami besednih nizov iz korpusa GOS 1.0: http://hdl.handle.net/11356/1271**