Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

Delavnica Odprta znanost
Pravna fakulteta
23. januar 2020

CLARIN.SI

`http://www.clarin.si/`

- Začetek dela v 2014
- Institut "Jožef Stefan":
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 11 partnerjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
  - 2 podjetji: Amebis, Alpineon
  - 1 društvo + 1 zavod: SDJT, Trojina

CLARIN.SI storitve

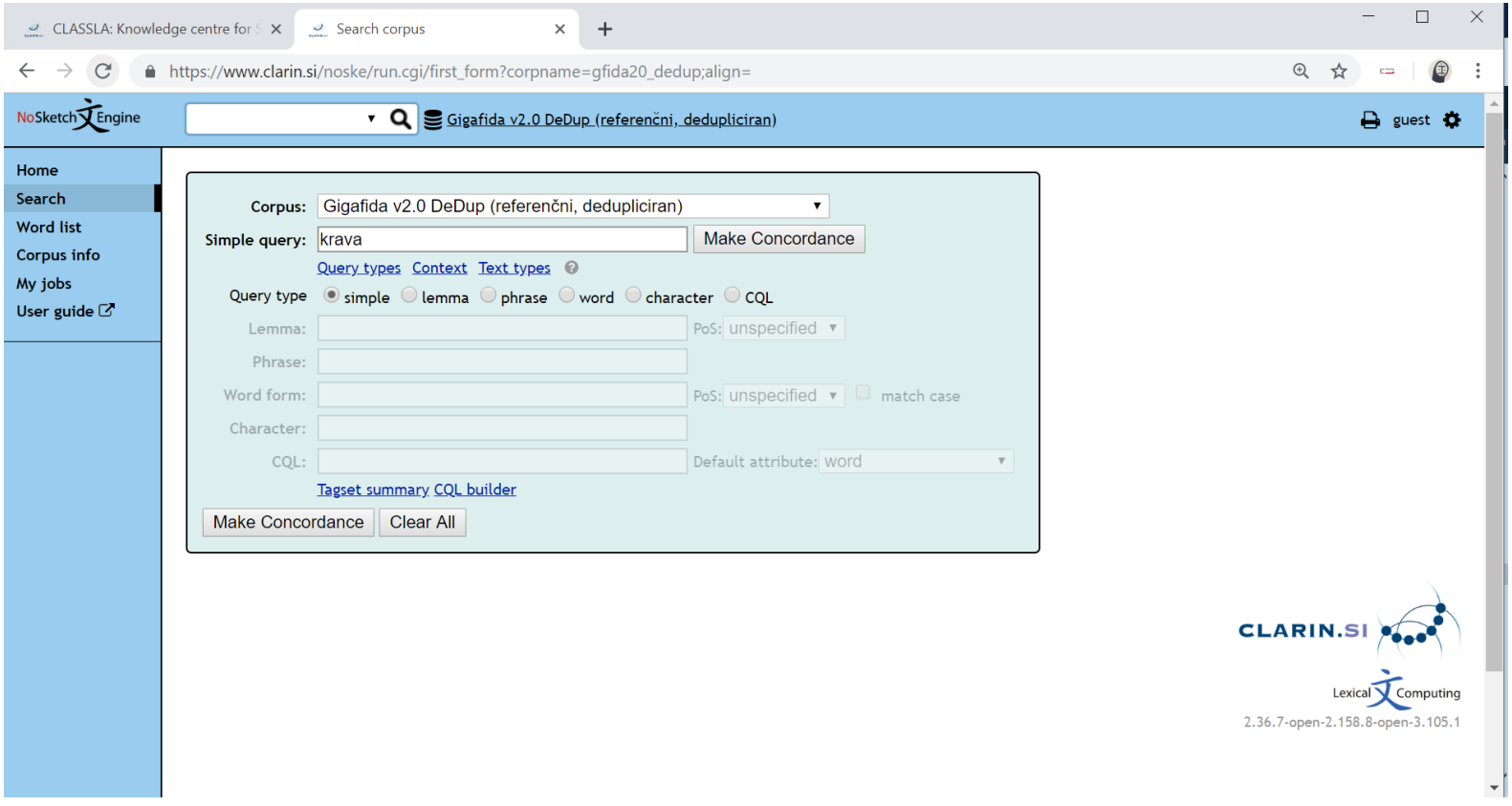
Trije stebri:

- Certificiran repozitorij jezikovnih virov in orodij
  - dolgotrajno hranjenje
  - avtentikacija in avtorizacija
  - stalni identifikatorji
  - eksplicitni pogoji uporabe in licence
  - principi FAIR
- Spletne storitve
  - dva konkordančnika (spletne analize korpusov)
  - orodja za označevanje besedil
  - itd.
- Podpora:
  - financiranje priprave virov za vključitev v repozitorij
  - večji projekti: 30.000 EUR letno: 7 projektov v 2018, 6 v 2019
  - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
  - CLASSLA K-centre: CLARIN.SI center znanja za računalniško obdelavo južnoslovanskih jezikov

Konkordančnika

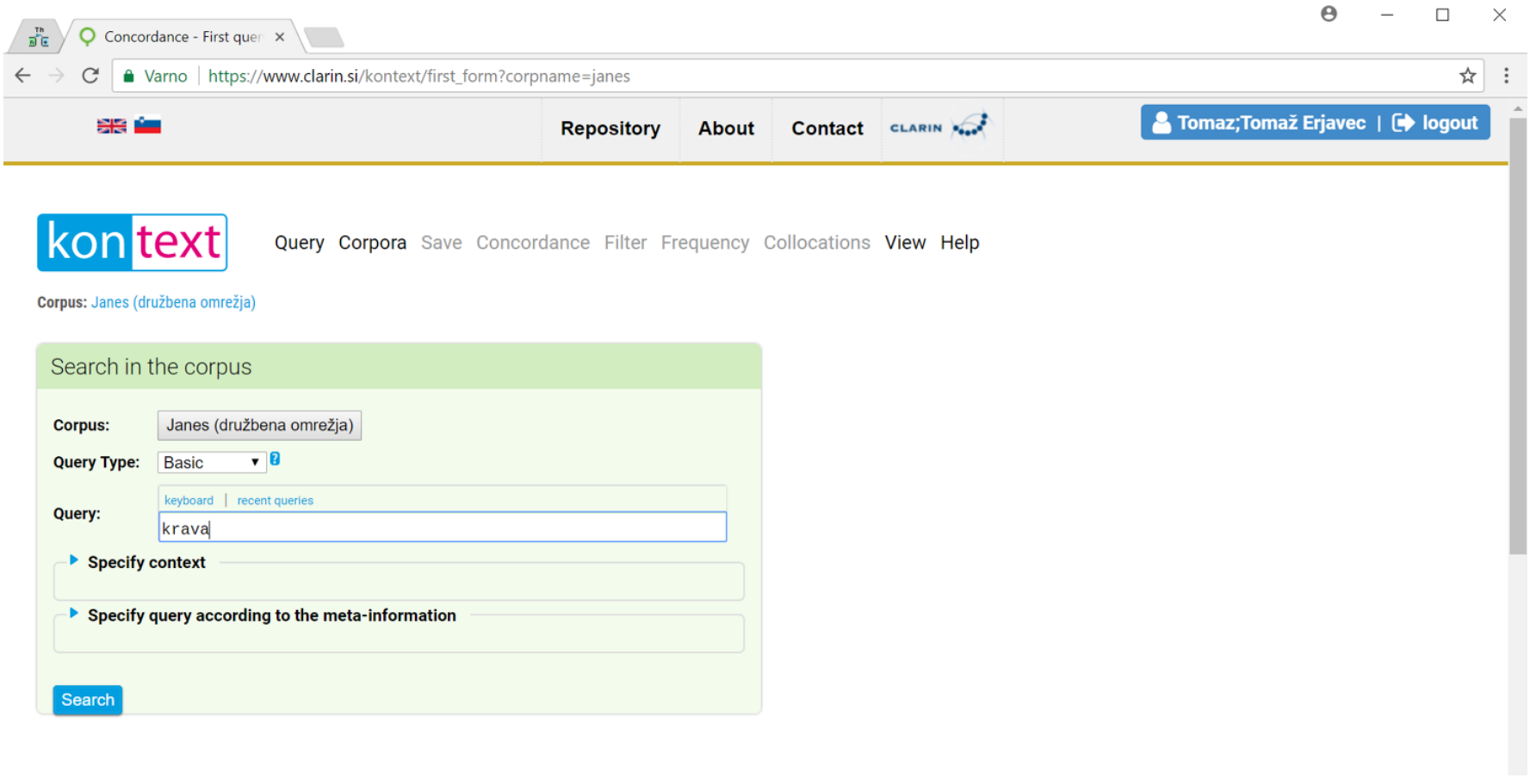
- noSketch Engine + KonText
- Obdelava velikik korpusov (več milijard besed)
- Korpusi so lahko bogato označeni:
  - strukture (besedilo, odstavek, termin, . . .)
  - metapodatki (datum, vrsta besedila, spol avtorja, standardnost,. . .)
  - atributi pojavnic (oblikoskladenjska oznaka, lema, normalizirana oblika,. . .)
- Bogat poizvedovalni jezik: CQL
- Raznovrstni izpisi in analize
- RESTful, tj. poizvedbe prek URLjev
- Dostopnih 70 korpusov v 27 jezikih in s prek 15 milijard besed

noSketch Engine



- Odprtokodna različica Sketch Engine
- Prijava ni niti potrebna niti mogoča

KonText



- Razvil ga je češki CLARIN
- Prijava: shranjene poizvedbe, nastavitve zaslona, podkorpusi
- Nima nekaterih funkcionalnosti noSketch Engine

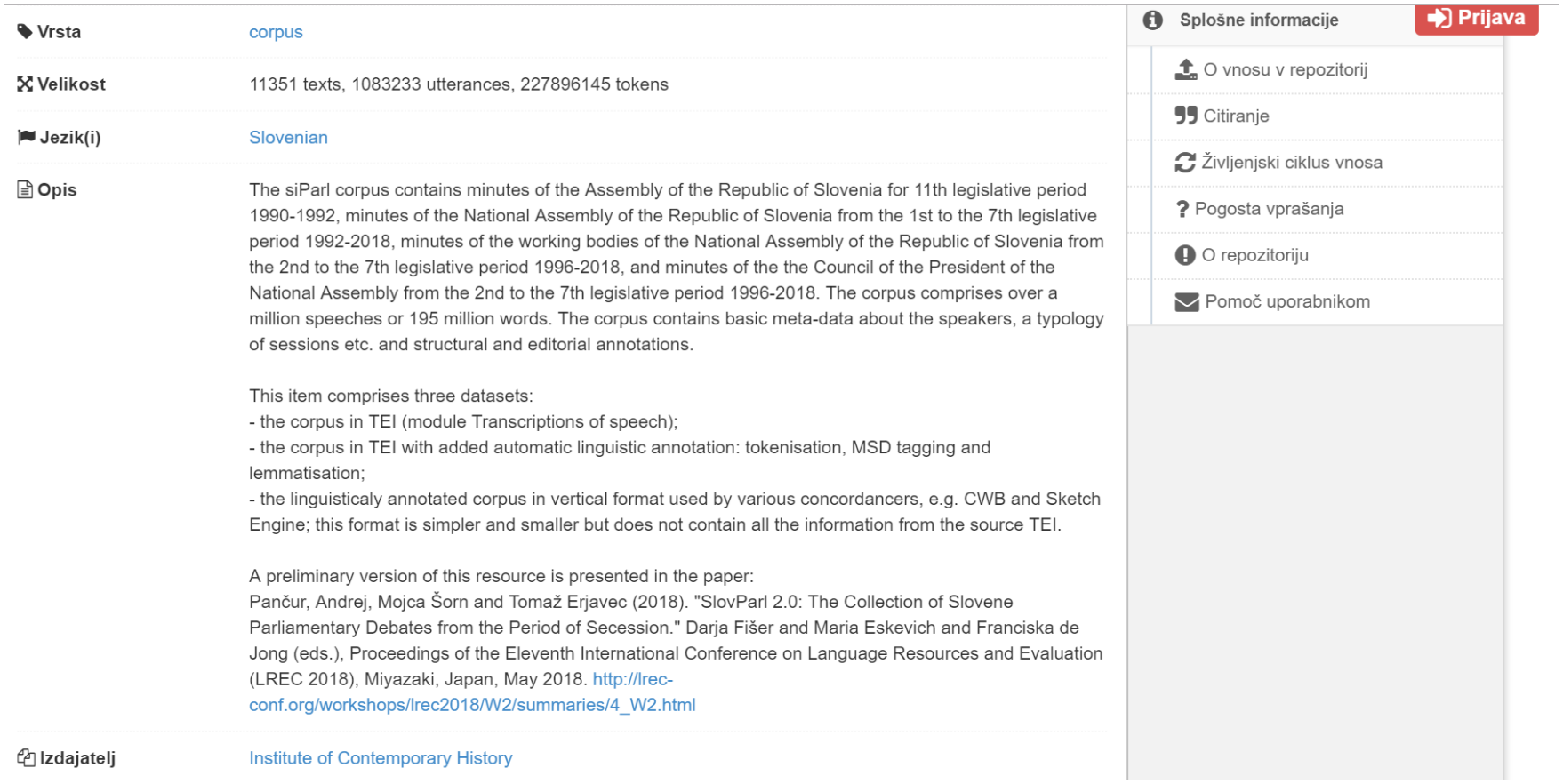
Repozitorij

- Zaenkrat najbolj pomembna storitev CLARINa
- Stalna in varna hramba jezikovnih virov (https, Nagios)
- Eksplicitni pogoji uporabe (ToS, licenca), etični kodeks (CoCo)
- Standarden zapis metapodatkov: Component Metadata Infrastructure (CMDI) Dublin Core (DC)
- Žetev metapodatkov (metadata harvesting)
- Večinoma standarden zapis podatkov (XML, TEI)
- Trenutno vsebuje 159 virov, od tega 121 slovenskih

Anatomija vnosa, 1

- osnovni metapodatki; citiranje; integracija s storitvami; povezave na projekt, demo & objave
- lokalizacija, orodna vrstica, prijava; iskanje, osnovne informacije o repozitoriju, brskanje; Piwik

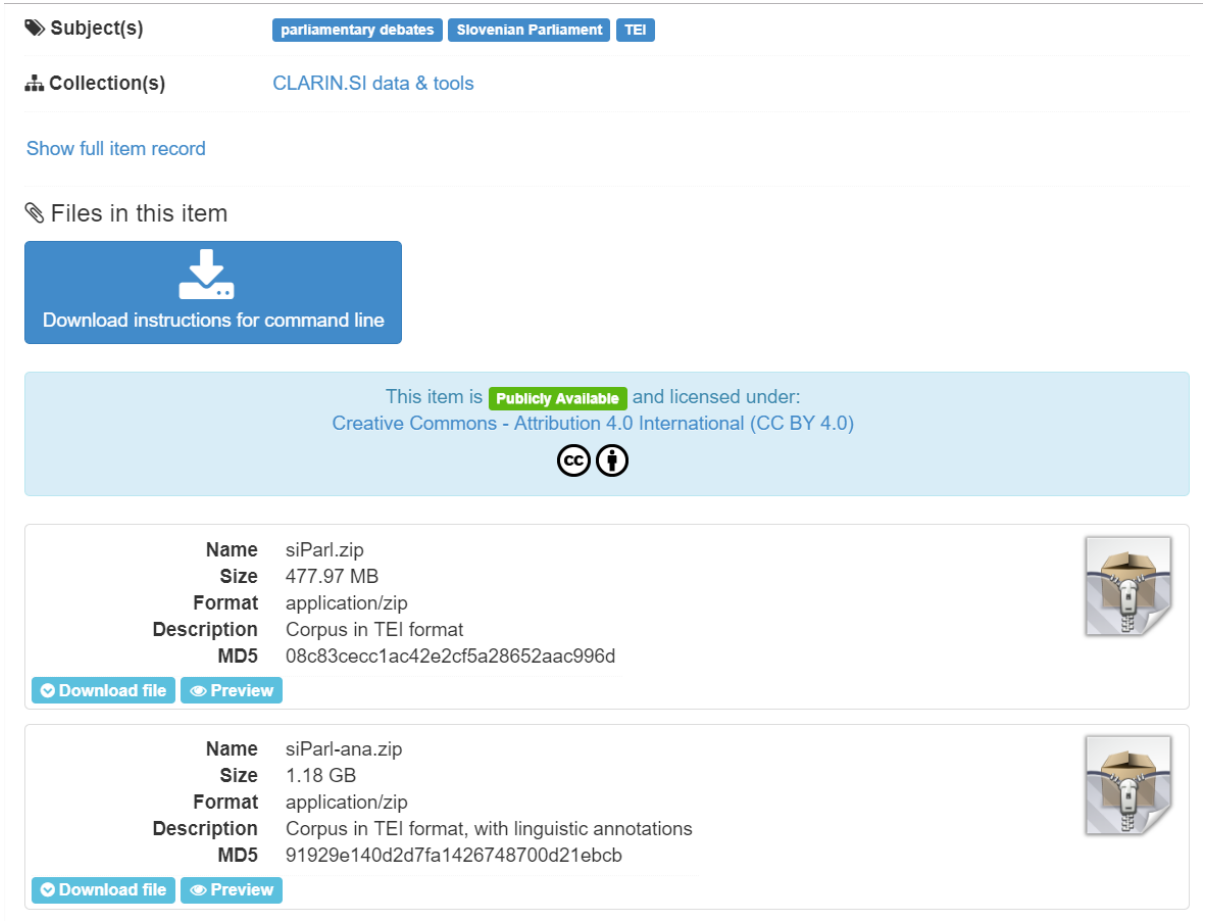
Anatomija vnosa, 2



Vrsta	corpus
Velikost	11351 texts, 1083233 utterances, 227896145 tokens
Jezik(i)	Slovenian
Opis	The siParl corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative period 1992-2018, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the 2nd to the 7th legislative period 1996-2018, and minutes of the the Council of the President of the National Assembly from the 2nd to the 7th legislative period 1996-2018. The corpus comprises over a million speeches or 195 million words. The corpus contains basic meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations. This item comprises three datasets: - the corpus in TEI (module Transcriptions of speech); - the corpus in TEI with added automatic linguistic annotation: tokenisation, MSD tagging and lemmatisation; - the linguisticaly annotated corpus in vertical format used by various concordancers, e.g. CWB and Sketch Engine; this format is simpler and smaller but does not contain all the information from the source TEI. A preliminary version of this resource is presented in the paper: Pančur, Andrej, Mojca Šorn and Tomaž Erjavec (2018). "SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession." Darja Fišer and Maria Eskevich and Franciska de Jong (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html
Izdajatelj	Institute of Contemporary History

- vrsta, velikost, jezik, opis, založnik
- več informacih o repozitoriju

Anatomija vnosa, 3



- ključne besede, vsi metapodatki
- licenca, prevzem podatkov

Nekaj vnosov

- **ccGigafida**: Referečni korpus standardne slovenščine
  vzorčeni odstavki korpusa Gigafida (100 mil. besed)
- **Janes**: Korpusi družbenih omrežij
  tviti, blogi, komentarji (250 mil. besed)
- **IMP**: Jezikovni viri starejše slovenščine
  600 knjig, 1700 - 1918, (17 mil. besed)
- **siParl**: Slovenski parlament
  zapisniki slovenskega parlamenta 1990-2018 (195 mil. besed).
- **Gos VideoLectures**: Govorni korpus
  55 predavanj, 22 ur govora (180,000 besed)

Zaključki

- Namen CLARIN(.SI) je spodbujati raziskave, ki potrebujejo dostop do jezikovnih podatkov
  - digitalna humanistika in družboslovje
  - jezikovne tehnologije
  - vse ostale vede, kjer je jezik pomemben
- Odprt dostop do virov, orodij in storitev
- Avtentikacija AAI, citiranje (handle)
- Prek 70 korpusov na konkordančnikih
- Prek 150 virov v repozitoriju
- Slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

Delavnica Odprta znanost
Pravna fakulteta
23. januar 2020