The CLARIN.SI research infrastructure

Tomaž Erjavec

Dept. of Knowledge Technologies, Jožef Stefan Institute
Fran Ramovš Institute of the Slovenian Language, ZRC SAZU

DigiLing
2022-10-25

Introduction

Which studies need access to language data?

- Linguistics, e.g.:
  - Lexicography
  - Corpus linguistics
  - Language teaching
- Digital humanities, e.g.:
  - Literary studies ("distant reading")
  - Historical studies
  - Political studies
- Sociology, e.g.:
  - Survey data
  - Other textual data
- Computational linguistics
  - supervised machine learning
  - need manually annotated training (and testing) data

Language Resources

Corpora

- Uniformly encoded and documented collection of texts
- "Texts chosen according to explicit criteria"
- Annotated (metadata, linguistic annotations)
- Reference/specialised; mono/multilingual; written/speech

Lexicons

- Words/phrases; morphology, syntax, semantics, translations
- Machine readable dictionaries, computational lexicons, ontologies

Language models

Data for programs to enable them to annotate (analyse) texts in a certain language for some level of annotation (analysis)

Data re-use

Traditional approach

- Language resources made from scratch for each investigation
- The resource not available to other researchers

Downsides

- The compilation of a language resource can be very costly: waste of time and money to do it more than once
- Later researchers cannot check or improve the first results
- Monopoly of researchers and institutions that produced the resource
- The resources cannot be used for product development

Open access to the results of research projects

No barriers to access of research publications and data

Savings of time & money, avoiding duplication of work, encourages cooperation, transparency of the scientific process, innovation

FAIR principles

- Findable, Accessible, Interchangeable, Reusable
- EU projects for open data: EOSC
- FACT: fair, accurate, confidential, transparent

Problems to making language resources open

- Copyright on source texts
- Privacy protection (GDPR)
- Terms of use (of data providers)
- Much more work for the data compilers

CLARIN

Research infrastructures

What is a research infrastructure?

Equipment, resources and services used the the scientific community for undertaking state-of-the-art research.

In the field of Humanities:

- DARIAH ERIC: Digital Research Infrastructure for the Arts and Humanities
- **CLARIN ERIC: Common Language Resources and Technology Infrastructure**

CLARIN: Common Language Resources and Technology Infrastructure

- Vision: digital language resources and tools for all (European) language are available through a single sign-on for researchers in the humanities and social sciences
- Long-term preservation and access to language resources and technologies
- A contribution to maintaining and supporting the multi-lingual European cultural heritage
- A new paradigm of collaboration in the development of language resources and tools, enabling multiple use and adaptation to individual needs

Purpose

- Make existing tools and solutions available in a common infrastructure
- Support consulting an teaching on how to adapt tools and resources to specific research needs
- A contribution to standardisation of resources and tools

CLARIN ERIC

- Headquarters in the Netherlands
- 22 national consortia + 2 observer countries + 1 third party:
  - Slovenia member since 2015
- Board of Directors, National Coordinators Forum
- Working Groups (User involvement, Legal, Standards, ...)
- Most work is done in the scope of the national consortia
- Virtual Language Observatory:
  aggregates metadata from national CLARIN repositories

CLARIN ERIC offerings

- Annual conference:
  - CLARIN covers costs for 5 participants per country + authors
- CLARIN Mobility Grants
- Knowledge Centres:
  - K-centre for Corpus Linguistics
  - K-Centre for Diachronic Language resources
  - K-Centre for Speech Analysis
  - K-Centre for Terminology Resources and Translation Corpora
  - etc.
- Digital Humanities course registry
- Resource families
- VideoLectures
- etc.

CLARIN.SI

CLARIN.SI

- Start of work in 2014
- Located at the Jožef Stefan Institute:
  - E8: Dept. for Knowledge Technologies
  - E3: Lab. for Artificial Intelligence
  - CMI: Networking Infrastructure Centre
- Organised as a consortium of 12 partners
  - 4 universities (Ljubljana, Maribor, Nova Gorica, Koper)
  - 4 research institutes (ZRC SAZU, IJS, INZ, ZRS Koper)
  - 2 societies (SDJT, Trojina)
  - 2 companies (Amebis, Alpineon)

CLARIN.SI services

Three pillars:

1. **Repository**:
   Long term FAIR archiving of language resources (and tools)
2. Web services:
   Concordancers, GitLab & GitHub, WebAnno, etc.
3. Support & outreach:
   - Support of development of language resources and tools: annual project calls
   - CLASSLA K-centre for processing of South-Slavic languages
   - Conferences, e.g. "Language Technologies and Digital Humanities"
   - Presentations, tutorials, lectures

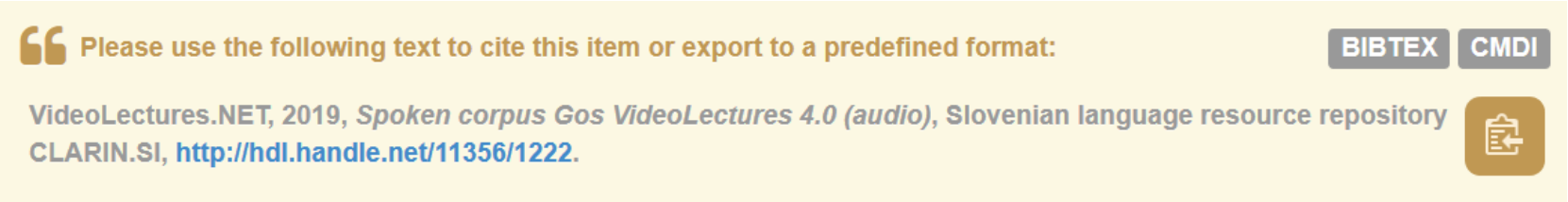
CLARIN.SI repository

Repository

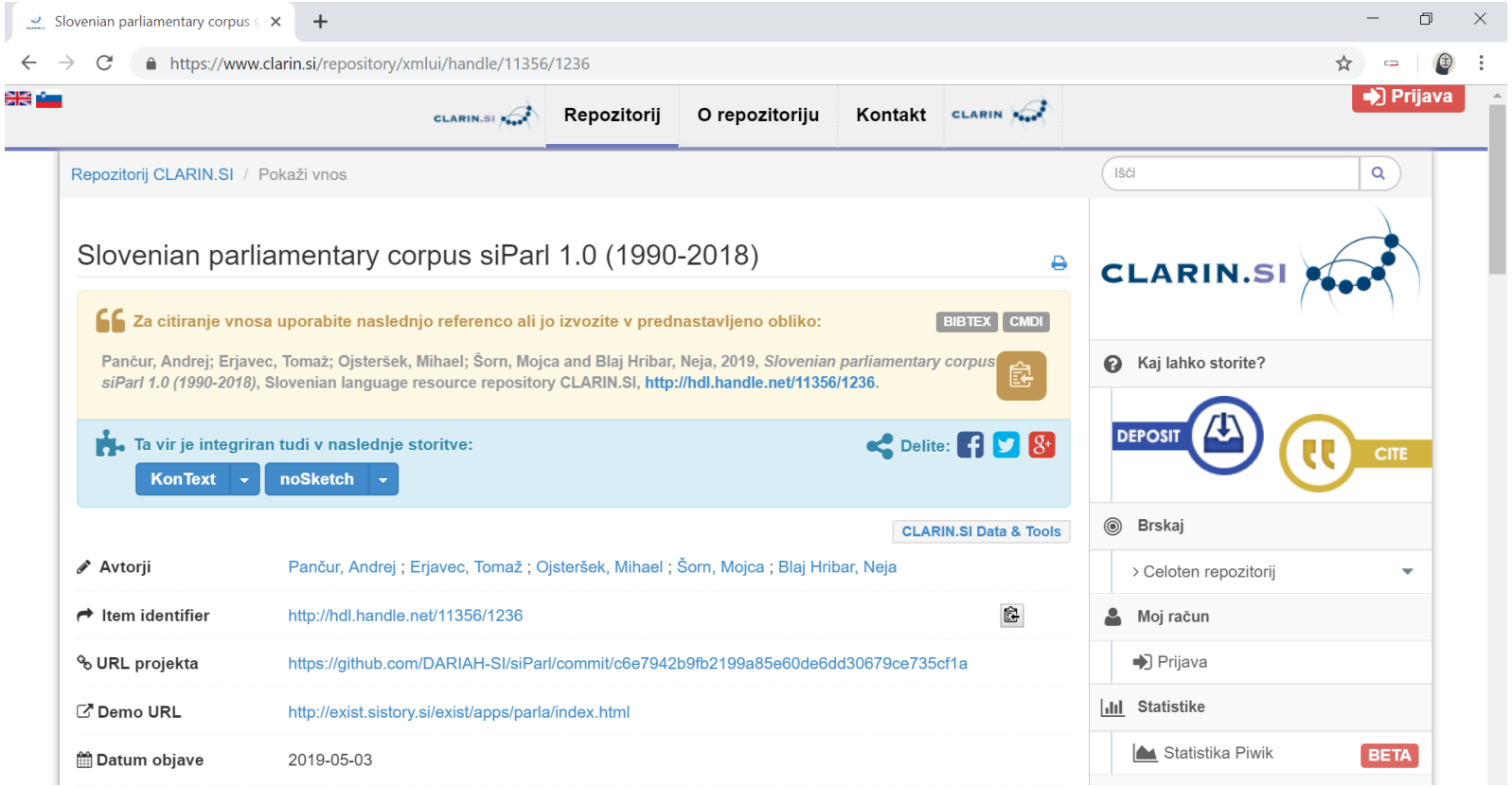
- Currently the most important CLARIN.SI service
- Long term and safe archiving of LRT (Core Trust Seal)
- Explicit rules of deposit and access (terms-of-use, licences)
- Ethical codex (code of conduct)
- Standardised meta-data
  - Component Metadata Infrastructure (CMDI)
  - Dublin Core (DC)
- Metadata harvesting
- Mostly standardised encoding of data (XML, TEI)
- Almost all resources available under CC licences
- Currently contains almost 450 entries

Permanent identifiers

- How to use URLs, so that they can be cited?
- DOI the most common way
- CLARIN uses the Handle system
- http://hdl.handle.net/11356/1222 → https://www.clarin.si/repository/xmlui/handle/11356/1222
- Important for correct citation of the resources



Anatomy of a resource landing page, 1



- Citation info; Service integration; Basic metadata

Anatomy of a resource landing page, 2

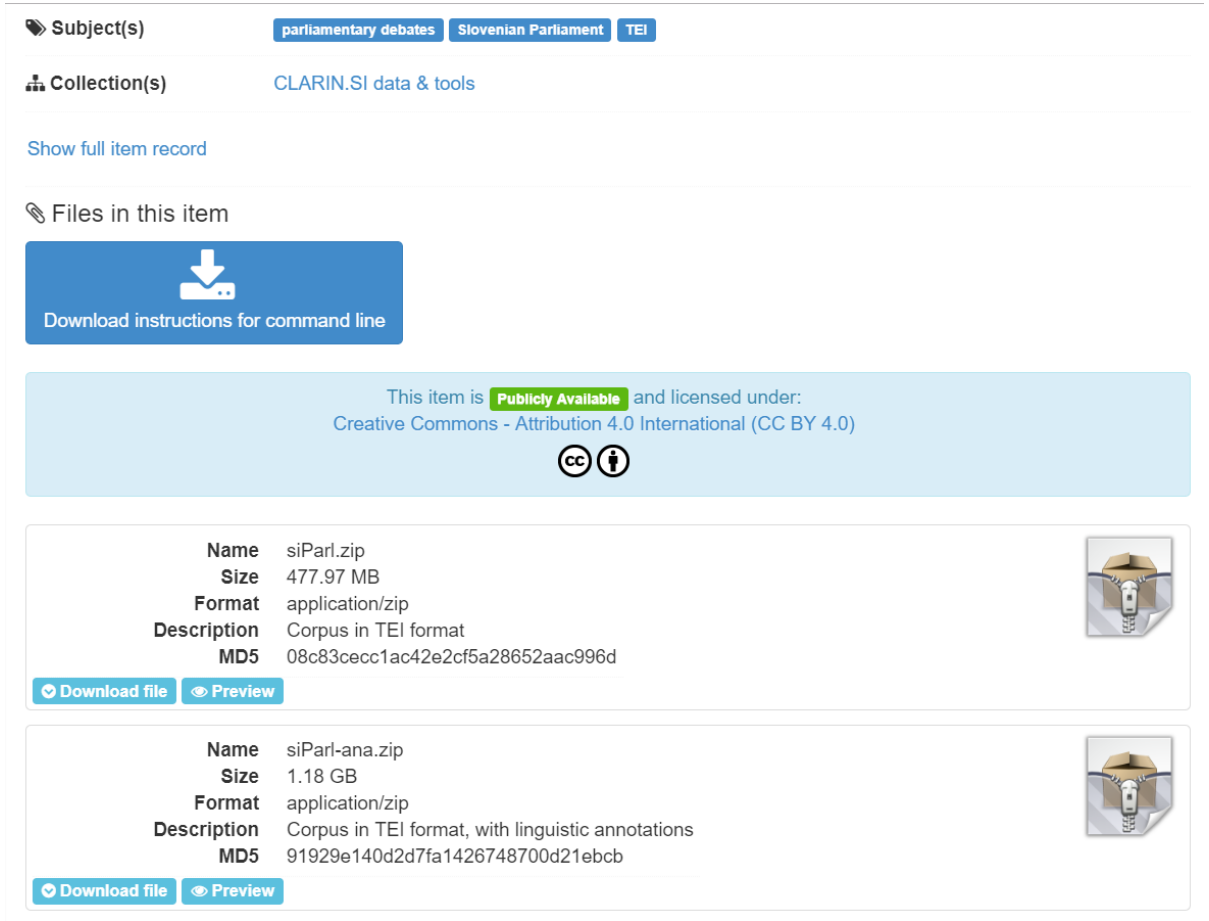
Vrsta	corpus
Velikost	11351 texts, 1083233 utterances, 227896145 tokens
Jezik(i)	Slovenian
Opis	The siParl corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative 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

Splošne informacije Prijava

- O vnosu v repozitorij
- Citiranje
- Življenjski ciklus vnosa
- Pogosta vprašanja
- O repozitoriju
- Pomoč uporabnikom

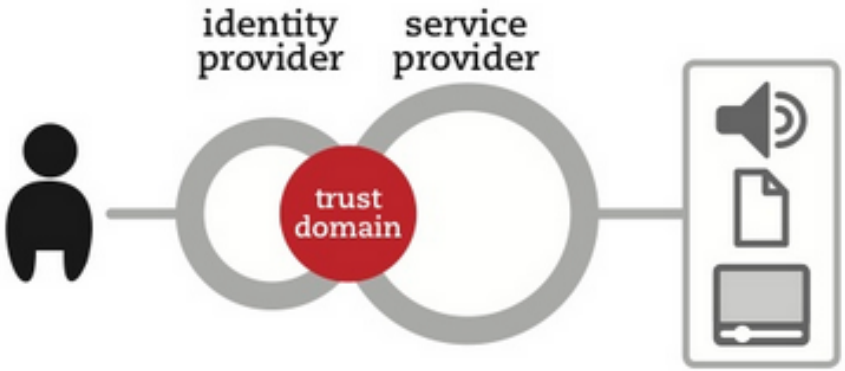
- Type, Size, Language, Description, Publisher of the data

Anatomy of a resource landing page, 3



- Keywords; Full Metadata;
- Licence, Downloading the data

Single sign-on



- Authentication and Authorisation Infrastructure (AAI)
- Single Sign-On (SSO): distinguish between service and identity provider
- The the user's identity is established by the Federation of Identity Providers (EduGain)
- Log-in necessary for CLARIN.SI repository download of protected resources and for up-load
- Slovene users can with their EduGain account access most CLARIN services in the EU

Using the repository: download resources

1. Find the resource: browse or search in the repository
2. Read the description and other metadata
3. Check the licence (e.g. CC NC, or CC ND)
4. Log-in necessary only for a few restricted resources with a more strict licence
5. Download the data
6. Use it
7. Properly acknowledge your use of the data!

Maybe the resource is in some other CLARIN repository?
Use CLARIN VLO: `https://vlo.clarin.eu/`

Using the repository: archiving resources

1. Carefully read the depositing guidelines for meta-data and data
2. Log-in and make your entry (i.e. enter required meta-data)
3. Decide on the licence
4. Upload the files (in the correct format!)
5. Finish the submission
6. A CLARIN.SI editor will review it and accept it or return it for corrections
7. Once ok, the editor will publish it

CLARIN.SI concordancers

Slovenian concordancers

- Best known is the concordancer for Gigafida and other concordancers at CJVT: Gos, Šolar, Lektor
- Translation Division of the General Secretariat of the Government of the Republic of Slovenia: Evroterm, a combinations of a terminological database and parallel corpus of EU law
- ISJFR ZRC: Nova Beseda
- 1 concordancer for 1 corpus (and so people often mix the two)

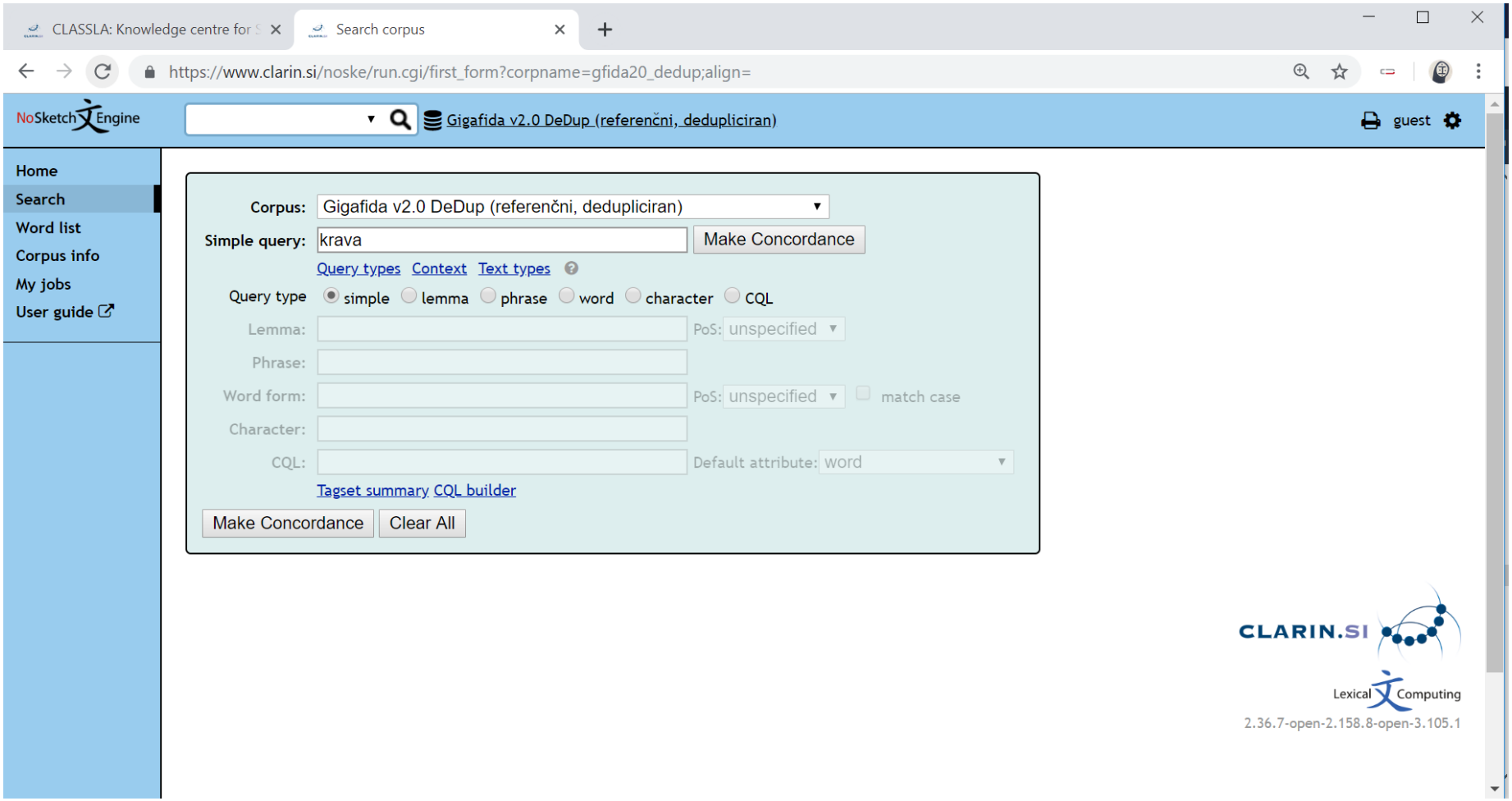
noSketch Engine and KonText

- Lexical Computing (Brno): company that offers the commercial concordancer Sketch Engine (SkE)
- They also offer an open source version of Sketch Engine, named noSketch Engine (noSkE)
- noSkE does not include some advances functionalities of SkE: Word sketches, Sketch differences, Thesaurus, BootCat
- Old noSkE (known as **Bonito**) is not longer maintained, but still offered by CLARIN.SI
- New noSkE (known as **Crystal**) has a very different user interface, also offered by CLARIN.SI
- **KonText** front-end develped by CNC, also offered by CLARIN.SI
- All three have the same back-end (Manatee) and use the same format for corpus files but have different front-ends (interfaces)

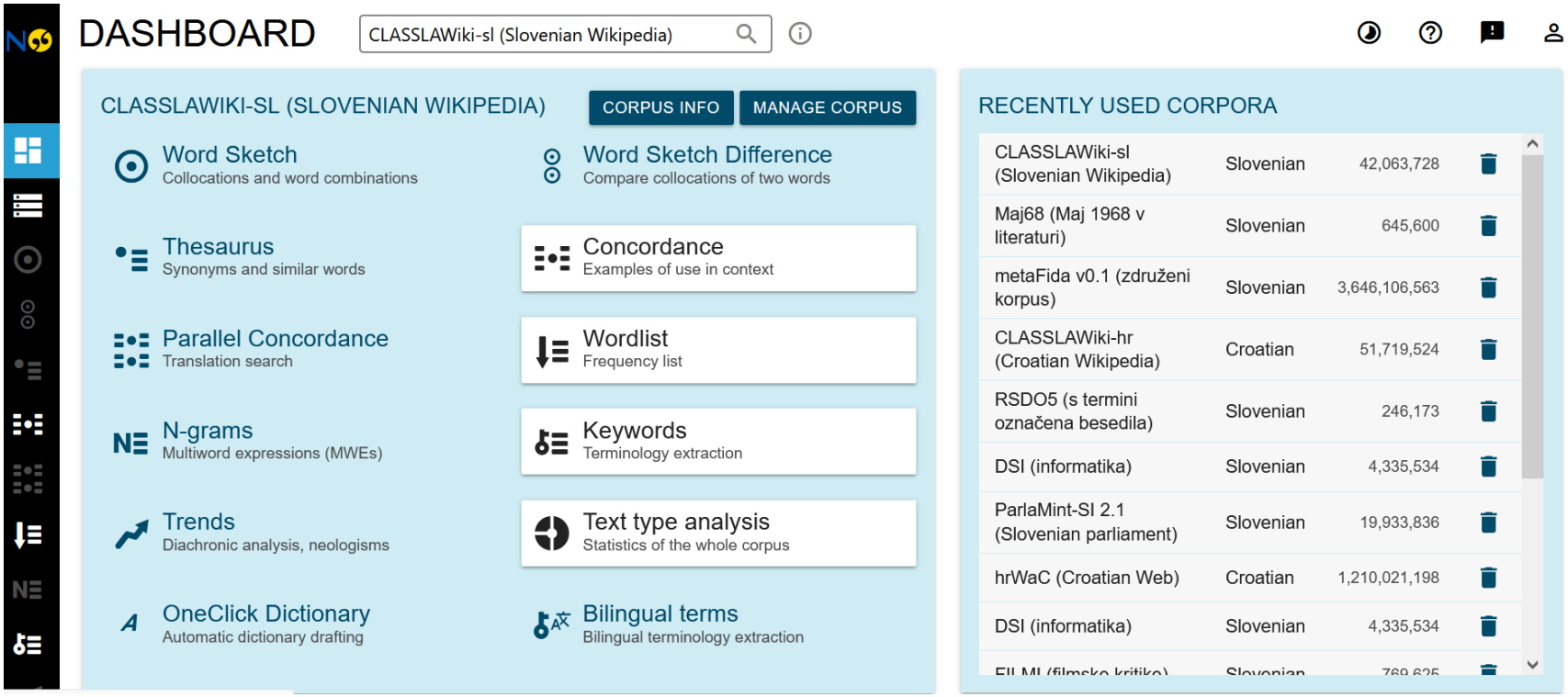
Concordancers @ CLARIN.SI

- The 3 concordancers:
  KonText + noSke Crystal + noSke Bonito
- All provide access to the same set of corpora
- Currently almost 100 corpora in 30 languages with over 20 billion words

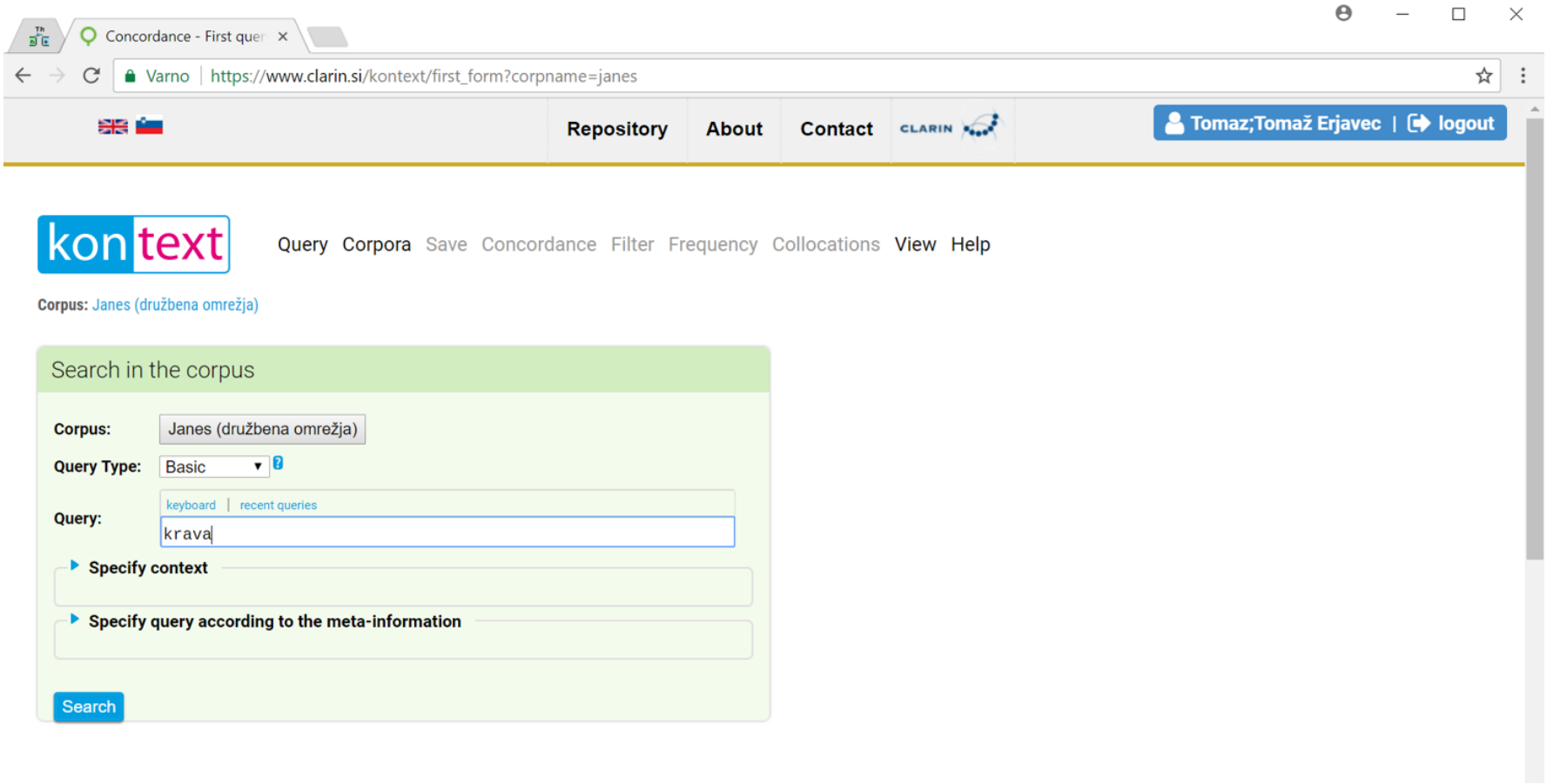
noSketch Engine Bonito



noSketch Engine Crystal



KonText


Concordance - First quer
Varno
https://www.clarin.si/kontext/first_form?corpname=janes
Repository
About
Contact
CLARIN
Tomaz;Tomaž Erjavec | logout
kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help
Corpus: Janes (družbena omrežja)
Search in the corpus
Corpus:
Janes (družbena omrežja)
Query Type:
Basic
Query:
keyboard | recent queries
krava
Specify context
Specify query according to the meta-information
Search

Comparison

	Bonito	Crystal	KonText
Documentation	No	Yes	No
Slovenian interface	Yes	(Yes)	Yes
Keywords	Yes	Yes	No
Maintained	No	Yes	Yes
Log-in	No	No (Yes)	Yes
SkE compatibility	No	Yes	No
JSON API	Yes	Yes	No
XML API	Yes	No	No
Links in other resources	Yes	No	No

Vertical files

- A noSkE corpus can have arbitrary structural and positional annotations
- The corpus registry file defines the structures, their attributes and positional attributes
- The vertical file contains the corpus, e.g.

```
<text corpus_id="imp" corpus="IMP (starejša besedila)"
      info="https://www.clarin.si/noske/...corpname=imp"
      id="ZRC_00001-1584" title="Biblija (vzorec)"
      author="Dalmatin, Jurij" year="1584">
<gap/>
<p id="ZRC_00001-1584.p2">
<s>
INu        in         in         Cc        Vp
on         on         on         Pp3msn    Zotmei
je         je         biti       Va-r3s-n  Gp-ste-n
ſvoje      svoje      svoj       Px-nsa    Zp-set
dvanajſt   dvanajst   dvanajst   Mlc-pa    Kbg-mt
Iogre      jogre      joger      Ncmpa     Sommt
k'         k          k          Sd        Dd
ſebi       sebi       se         Px---d    Zp---d
poklizal   poklical   poklicati  Vmep-sm   Ggdd-em
<g/>
,          ,          ,          Z         U
```

Corpus Query Language (CQL)

A rich corpus query language:

- Search a positional attribute, e.g. `[lemma="krava"]`
- Regular expressions, e.g. `[lemma="krav.*"]`
- Logical combination of conditions, e.g. `[lemma="krava" & word!="krava"]`
- Search for a sequence of tokens, e.g. `[lemma="zelo"] []{0,2} [lemma="krava"]`
- Constraints on structures, e.g. `[lemma="krava"] within <text year="2011"/>` or `<text year="2011"/> containing [lemma="krava"]`

All other types of queries (simple, lemma, phrase, character) and text-type selection can be translated to a CQL query

Some available corpora on concordancers

- Reference: GigaFida
- Scientific texts: KAS
- Speech: Gos, GosVL
- Historical Slovenian: IMP
- User-generated content: Janes
- Parliaments: siParl, 16-language ParlaMint
- Parallel: EU-DGT, Trans5, JaSlo, IsPac
- Other South-Slavic languages:
  Croatian, Serbian, Bosnian, Macedonian, Montenegrin
- Other languages: English, Japanese

MetaFida

- Researchers often want to search over several corpora but this is difficult and error-prone
- MetaFida: corpus of selected Slovenian corpora (RSDO project)
- Corpora chosen (=34 corpora, 3.5 billion words):
  - Many and varied corpora
  - Tokens annotated at least with their lemma and MSD
- Removal of duplicate paragraphs: deleted 12.5% paragraphs, 6.5% texts
- Sorted by year of text publication (those without year at the end)
- = MetaFida version 0.1
- Version 1.0 at end of RSDO project

Conclusions

Conclusions

- You have just survived your introduction into the CLARIN research infrastructures, esp. its data repository and concordancers :)
- A lot more information available on the CLARIN.SI website and the recent JT-DH 2022 paper
  Tomaž Erjavec, Kaja Dobrovoljc, Darja Fišer, Jan Jona Javoršek, Simon Krek, Taja Kuzman, Cyprian Laskowski, Nikola Ljubešić, Katja Meden: Raziskovalna infrastruktura CLARIN.SI
- In case of any remaining questions or problems of use, get in touch via info@clarin.si or repo-help@clarin.si

The CLARIN.SI research infrastructure

Tomaž Erjavec

Dept. of Knowledge Technologies, Jožef Stefan Institute
Fran Ramovš Institute of the Slovenian Language, ZRC SAZU

DigiLing
2022-10-25