The research infrastructure CLARIN(.SI)

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

4. simpozij SCIMETH: Digitalni alati i resursi u jezikoslovlju
Filozofski fakultet u Rijeci
2019-05-15

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
- MRD, ..., ontology

Language models

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

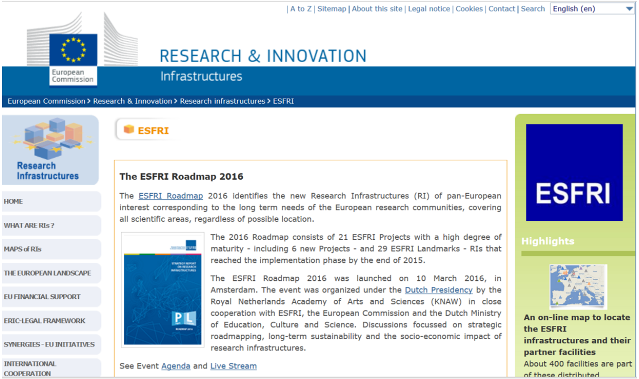
- Copyright on source texts
- Privacy protection (GDPR)
- Terms of use (of data providers)
- Much more work for the data compilers

CLARIN

Research infrastructures

What is a research infrastructure?

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


| A to Z | Sitemap | About this site | Legal notice | Cookies | Contact | Search
English (en)
European Commission
RESEARCH & INNOVATION
Infrastructures
European Commission > Research & Innovation > Research infrastructures > ESFRI
Research Infrastructures
HOME
WHAT ARE RIs ?
MAPS of RIs
THE EUROPEAN LANDSCAPE
EU FINANCIAL SUPPORT
ERIC-LEGAL FRAMEWORK
SYNERGIES - EU INITIATIVES
INTERNATIONAL COOPERATION
ESFRI
ESFRI
Highlights

The ESFRI Roadmap 2016

The ESFRI Roadmap 2016 identifies the new Research Infrastructures (RI) of pan-European interest corresponding to the long term needs of the European research communities, covering all scientific areas, regardless of possible location.

The 2016 Roadmap consists of 21 ESFRI Projects with a high degree of maturity - including 6 new Projects - and 29 ESFRI Landmarks - RIs that reached the implementation phase by the end of 2015.

The ESFRI Roadmap 2016 was launched on 10 March 2016, in Amsterdam. The event was organized under the Dutch Presidency by the Royal Netherlands Academy of Arts and Sciences (KNAW) in close cooperation with ESFRI, the European Commission and the Dutch Ministry of Education, Culture and Science. Discussions focussed on strategic roadmapping, long-term sustainability and the socio-economic impact of research infrastructures.

See Event Agenda and Live Stream

An on-line map to locate the ESFRI infrastructures and their partner facilities

About 400 facilities are part of these distributed

ESFRI Infrastructures

- European Strategy Forum on Research Infrastructures, founded in 2002
- The ESFRI Roadmap propsed 15 (in 2016: 21) RIs, some are established as ERICs (EU legal entity: European RI Consortium)
- 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
- 20 national consortia + 4 observer countries:
  - Slovenia member since 2013
  - Croatia member since 2018
- 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 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

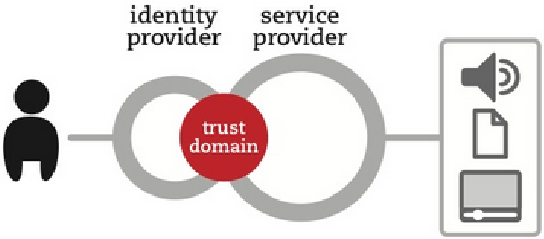
CLARIN.SI

- 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
  - 3 research institutes
  - 3 societies
  - 2 companies

CLARIN.SI services

- Support for events:
  - Conferences ”Language Technologies and Digital Humanities” (2016, 2018, ...)
  - JOTA @ VideoLectures
  - XVIII EURALEX International Congress, Lj., 17.-21.7.2018
- Support of development of language resources and tools:
  - making LRs ready to be included in the CLARIN.SI repo
  - first time in 2018: support of project to develop LRT
- Repository
  - Long term FAIR archiving of language resources (and tools)
- Two concordancers
- GitLab
- Manual annotation of corpora
- Automatic annotation of corpora
- Word 2 TEI conversion

Single sign-on



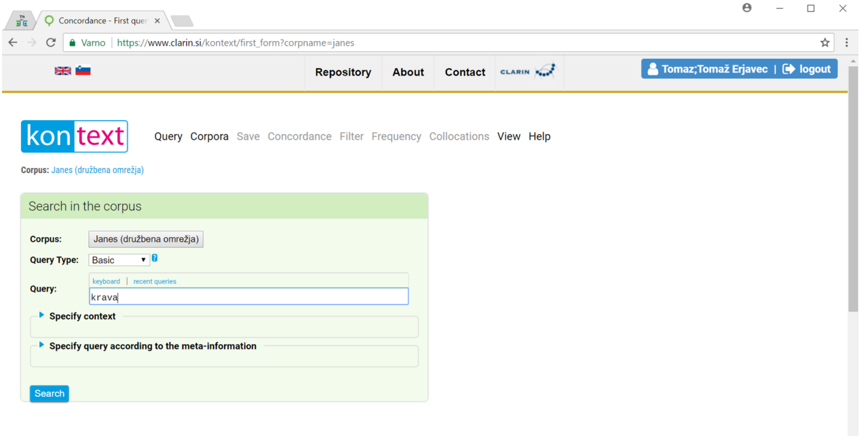
- Infrastructure for authentication and authorisation (AAI)
- Single Sign-On: Distinguish between the service provider and identity provider
- As opposed to classic web log-in here the identity of the user is known to the Federation of Identity Providers (EduGain)
- Easier access to resources and services for a global educational and research community
- Slovene and Croatian users can access most CLARIN services

CLARIN.SI technical services

Concordancers

- KonText + noSketch Engine
- Both use the same back-end: Manatee
- Can work with large corpora (> billion words)
- Corpora can be richly annotated:
  - structures: text, paragraph, sentence, term, name, etc.
  - metadata: text title, date of publication, type of sentence etc.
  - attributes of words: PoS tag, lemma, normalised form, etc.
- Rich query language: CQL (regular expressions, sequences, attributes, logical constructions)
- Various analyses and presentations
- RESTful, i.e. URLs can be quoted and fetched
- CLARIN.SI noSke & KonText currently provide over 50 corpora in 27 languages with over 14 billion words

KonText


Concordance - First quer
Varno https://www.clarin.si/kontext/first_form?corpname=janes
Repository
About
Contact
CLARIN
Tomaz;Tomaž Erjavec | logout
kon text
Query
Corpora
Save
Concordance
Filter
Frequency
Collocations
View
Help
keyboard | recent queries
▸ Specify context
▸ Specify query according to the meta-information
Search

Corpus: Janes (družbena omrežja)

Search in the corpus

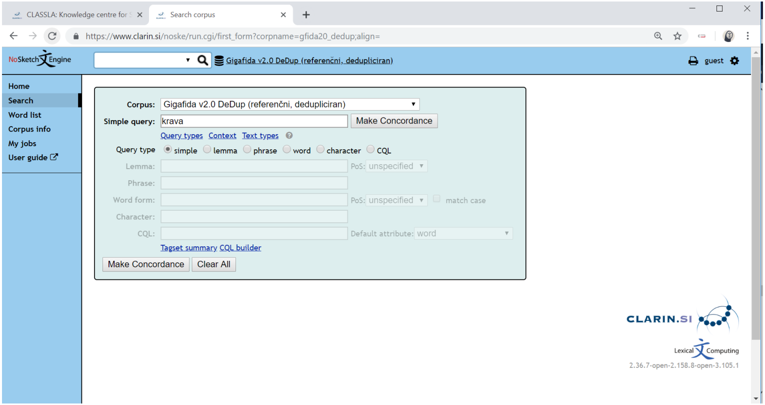
Corpus: Janes (družbena omrežja)

Query Type: Basic

Query: krava

- Developed by Czech CLARIN
- Allows log-in: saved queries, display settings, subcorpora
- Lacking some functionalities of noSketch Engine

noSketch Engine



CLASSLA: Knowledge centre for S ✕ | Search corpus ✕ | +

https://www.clarin.si/noske/run.cgi/first_form?corpname=gfida20_dedup;align=

NoSketch Engine | Gigafida v2.0 DeDup (referenčni, deduplicirani) | guest

Corpus: Gigafida v2.0 DeDup (referenčni, deduplicirani)

Simple query: krava [Make Concordance]

Query types Context Text types

Query type ● simple ○ lemma ○ phrase ○ word ○ character ○ CQL

Lemma: PoS: unspecified

Phrase:

Word form: PoS: unspecified □ match case

Character:

CQL: Default attribute: word

Tagset summary CQL builder

[Make Concordance] [Clear All]

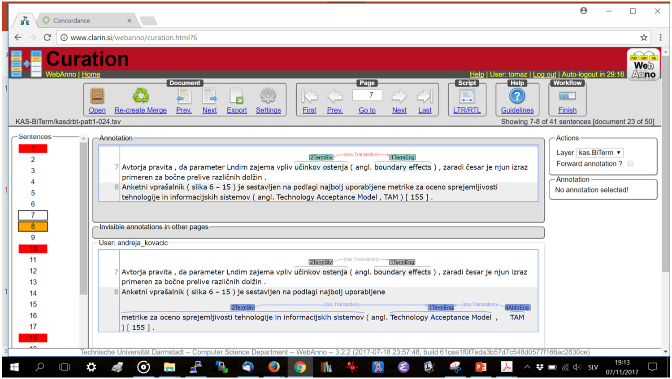
CLARIN.SI

Lexical Computing

2.36.7-open-2.158.8-open-3.105.1

- The open source version of the commercial Sketch Engine
- No log-in required or possible

WebAnno



www.clarin.si/webanno/curation.html?6

Curation

WebAnno | Home

Help | User: tomaz | Log out | Auto-logout in 29:16

KAS-BiTerm/kasdrbt-patt1-024.tsv

Showing 7-8 of 41 sentences [document 23 of 50]

Annotation

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

Invisible annotations in other pages

User: andreja_kovacic

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

Actions

Layer kas.BiTerm

Forward annotation ? □

Annotation

No annotation selected!

Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 3.2.2 (2017-07-18 23:57:48, build 61cea1f0f7eda3b57d7c548d0577f166ac2830ce)

- Tool for manual annotation of corpora
- Developed by German CLARIN
- Allows multiple annotators + curation phase
- At CLARIN.SI developed conversion TEI → TSV → TEI

Repository

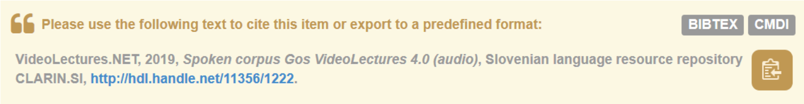
- Currently the most important CLARIN.SI service
- Long term and safe archiving of LRT (https, Nagios)
- Explicit rules of deposit and access (terms-of-use, licences)
- Ethical codex (Code of conduct)
- Standardised meta-data
  - Component Metadata Infrastructure (CMDI)
  - Dublin Core (DC)
- Metadata harvesting
- Mostly standardised encoding of data (XML, TEI)
- Almost all resources available under CC licences
- Currently contains about 100 LRTs

Repository platform

- Based on the DSpace platform for open digital repositories
- DSpace adapted for the needs of CLARIN repositories
- Developed by Czech CLARIN
- Development takes place on GitHub
- Also used by CLARIN Norway, Poland, Italy

Permanent identifiers

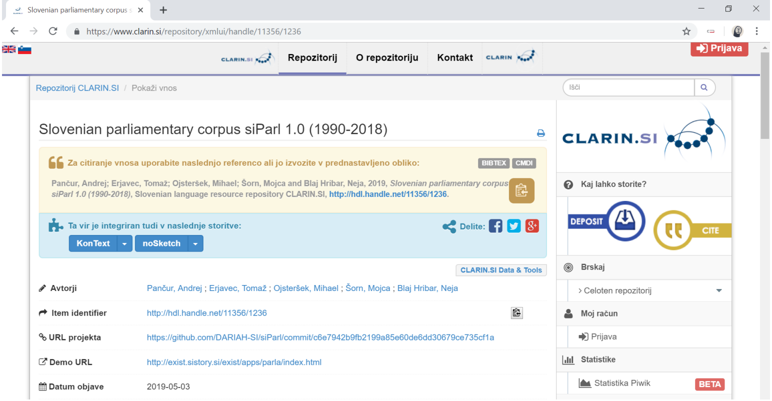
- How to use URLs, so that they can be cited?
- DOI the most common way
- CLARIN uses the Handle system
- http://hdl.handle.net/11356/1222 → https://www.clarin.si/repository/xmlui/handle/11356/1222
- Important for correct citation of the resources



Please use the following text to cite this item or export to a predefined format:

VideoLectures.NET, 2019, *Spoken corpus Gos VideoLectures 4.0 (audio)*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1222.

Anatomy of a resource landing page, 1



Slovenian parliamentary corpus s
https://www.clarin.si/repository/xmlui/handle/11356/1236
Repozitorij
O repozitoriju
Kontakt
Prijava
Išči
CLARIN.SI

Repozitorij CLARIN.SI / Pokaži vnos

Slovenian parliamentary corpus siParl 1.0 (1990-2018)

Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko: BIBTEX CMDI

Pančur, Andrej; Erjavec, Tomaž; Ojsteršek, Mihael; Šorn, Mojca and Blaj Hribar, Neja, 2019, *Slovenian parliamentary corpus siParl 1.0 (1990-2018)*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1236.

Ta vir je integriran tudi v naslednje storitve: KonText noSketch

Delite:

CLARIN.SI Data & Tools

Avtorji	Pančur, Andrej ; Erjavec, Tomaž ; Ojsteršek, Mihael ; Šorn, Mojca ; Blaj Hribar, Neja
Item identifier	http://hdl.handle.net/11356/1236
URL projekta	https://github.com/DARIAH-SI/siParl/commit/c6e7942b9fb2199a85e60de6dd30679ce735cf1a
Demo URL	http://exist.sistory.si/exist/apps/parla/index.html
Datum objave	2019-05-03

Kaj lahko storite?

DEPOSIT CITE

Brskaj

Celoten repozitorij

Moj račun

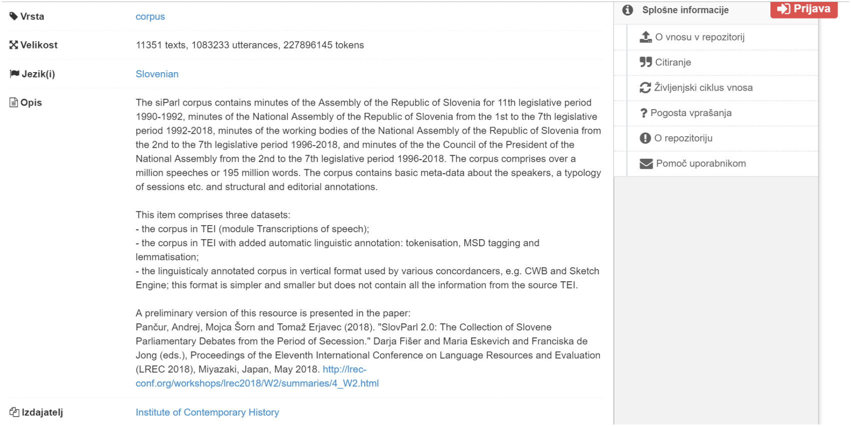
Prijava

Statistike

Statistika Piwik BETA

- Basic metadata; Citation; Service integration; Sharing; Project, Demo & Publication links
- Localisation, Toolbar, Login; Search, Basic info on repository, Browsing; Piwik

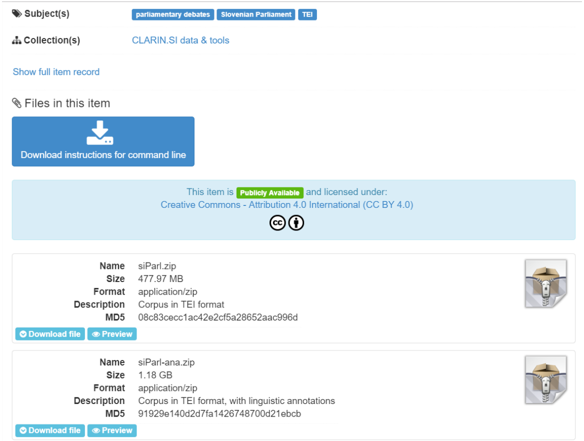
Anatomy of a resource landing page, 2



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

- Type, Size, Language, Description, Publisher of the data
- More information on repository

Anatomy of a resource landing page, 3



Subject(s) parliamentary debates | Slovenian Parliament | TEI

Collection(s) CLARIN.SI data & tools

Show full item record

Files in this item

Download instructions for command line

This item is **Publicly Available** and licensed under:
Creative Commons - Attribution 4.0 International (CC BY 4.0)

Name	siParl.zip
Size	477.97 MB
Format	application/zip
Description	Corpus in TEI format
MD5	08c83cecc1ac42e2cf5a28652aac996d

Download file | Preview

Name	siParl-ana.zip
Size	1.18 GB
Format	application/zip
Description	Corpus in TEI format, with linguistic annotations
MD5	91929e140d2d7fa1426748700d21ebcb

Download file | Preview

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

Conclusions

Conclusions

- The talk presented CLARIN and CLARIN.SI
- Not many words about CLARIN Croatia: while it is officially part of CLARIN, there are no Web page or services yet
- But the CLARIN.SI repository and concordancers already offer many Croatian language resources, which will be the subject of the next talk!

The research infrastructure CLARIN(.SI)

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

4. simpozij SCIMETH: Digitalni alati i resursi u jezikoslovlju
Filozofski fakultet u Rijeci
2019-05-15