

# The CLARIN.SI research infrastructure

Darja Fišer

Faculty of Arts, University of Ljubljana  
Institute of Contemporary History  
Jožef Stefan Institute

Annual meeting of the Doctoral School of the University of Ljubljana  
2022-06-02

# 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 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 and 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 2013
- 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

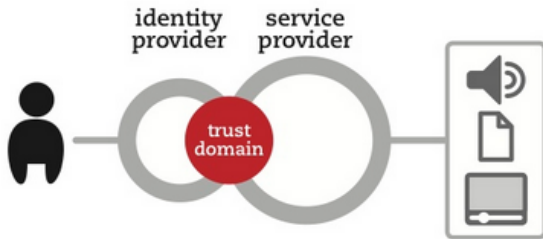


- Start of work in 2014
- Located at the Jožef Stefan Institute:
  - E8: Dept. for Knowledge Technologies
  - E3: Lab. for Artificial Intelligence
  - CMI: Networking Infrastructure Centre
- Organised as a consortium of 12 partners
  - 4 universities
  - 3 research institutes
  - 3 societies
  - 2 companies

# CLARIN.SI services

- Support for events:
  - Conferences "Language Technologies and Digital Humanities"
- 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 users can access most CLARIN services in the EU



## CLARIN.SI technical services

# Concordancers

- KonText + noSketch Engine
- currently provide over 50 corpora in 27 languages with over 14 billion words
- 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

# WebAnno

The screenshot displays the WebAnno web application interface. At the top, there is a navigation bar with the 'Curation' title and various tool icons for document management (Open, Re-create, Merge, Prev, Next, Export, Settings) and navigation (First, Prev, Go to, Next, Last). A status bar indicates 'Showing 7-8 of 41 sentences [document 23 of 50]'. The main workspace is divided into three sections: a left sidebar with a list of sentences (1-17), a central text area with annotations, and a right sidebar with 'Actions' (Layer, Forward annotation) and 'Annotation' (No annotation selected) options. The text area shows a sentence with two annotations: a 'TextAnnotation' for 'Lndim zajema vpliv učinkov ostenja' and a 'TextAnnotation' for 'Anketni vprašalnik'. The bottom of the image shows a Windows taskbar with the system clock at 19:13 on 07/11/2017.

- 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 over 300 LRTs

# 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:

BIBTEX

CMDI

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 : x +

https://www.clarin.si/repository/xmlui/handle/11356/1236

Repozitorij O repozitoriju Kontakt Prijava

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: f t s

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

ISCI

CLARIN.SI

Kaj lahko storite?

DEPOSIT CITE

Brskaj

Celoten repozitorij

Moj račun

Prijava

Statistike

Statistika Piwik BETA

- Citation info; Service integration; Basic metadata

# Anatomy of a resource landing page, 2

<b>Vrsta</b>	corpus	<b>Splošne informacije</b> <span>Prijava</span>	
<b>Velikost</b>	11351 texts, 1083233 utterances, 227896145 tokens		O vnosu v repozitorij
<b>Jezik(i)</b>	Slovenian		Citiranje
<b>Opis</b>	<p>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.</p> <p>This item comprises three datasets:</p> <ul style="list-style-type: none"><li>- the corpus in TEI (module Transcriptions of speech);</li><li>- the corpus in TEI with added automatic linguistic annotation: tokenisation, MSD tagging and lemmatisation;</li><li>- the linguistically 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.</li></ul> <p>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. <a href="http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html">http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html</a></p>		Življenjski cikel vnosa
<b>Izdajatelj</b>	Institute of Contemporary History		Pogosta vprašanja
		O repozitoriju	
		Pomoč uporabnikom	

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


# 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)



<b>Name</b>	siParl.zip	
<b>Size</b>	477.97 MB	
<b>Format</b>	application/zip	
<b>Description</b>	Corpus in TEI format	
<b>MD5</b>	08c83cecc1ac42e2cf5a28652aac996d	
<a>Download file</a> <a>Preview</a>		
<b>Name</b>	siParl-ana.zip	
<b>Size</b>	1.18 GB	
<b>Format</b>	application/zip	
<b>Description</b>	Corpus in TEI format, with linguistic annotations	
<b>MD5</b>	91929e140d2d7fa1426748700d21ebcb	
<a>Download file</a> <a>Preview</a>		

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



# Conclusions

# Conclusions

- You have just survived your very first intro into research infrastructures, data repositories and CLARIN.SI :)
- You will get a lot of more practical info from Kaja Dobrovoljc during the workshop
- A lot more useful info available on the CLARIN.SI website and the paper
- In case of any questions or problems, google first, then get in touch via [info@clarin.si](mailto:info@clarin.si) or [repo-help@clarin.si](mailto:repo-help@clarin.si)

# The CLARIN.SI research infrastructure

Darja Fišer

Faculty of Arts, University of Ljubljana  
Institute of Contemporary History  
Jožef Stefan Institute

Annual meeting of the Doctoral School of the University of Ljubljana  
2022-06-02