The research infrastructure CLARIN(.SI)

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

4. simpozij SCIMETH: Digitalni alati i resursi u jezikoslovju
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 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 by the scientific community for undertaking state-of-the-art research.
ESFRI Infrastructures

- European Strategy Forum on Research Infrastructures, founded in 2002
- The ESFRI Roadmap proposed 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
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
- 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.
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

- Developed by Czech CLARIN
- Allows log-in: saved queries, display settings, subcorpora
- Lacking some functionalities of noSketch Engine
The open source version of the commercial Sketch Engine

No log-in required or possible
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

- 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 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.


- **Izdajatelj**: Institute of Contemporary History

- **Type, Size, Language, Description, Publisher of the data**
- **More information on repository**
Keywords; Full Metadata;
Licence, Downloading the data
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