Good research needs good infrastructure: The CLARIN.SI and CLASSLA options for supporting research on South Slavic languages

Part 1: The CLARIN(.SI) research infrastructure

Tomaž Erjavec

National coordinator of CLARIN.SI

Dept. of Knowledge Technologies, Jožef Stefan Institute

Ljubljana

Alpen-Adria-Universität, 20th May, 2020

Overview of the lecture

1. Introduction
2. The CLARIN EU research infrastructure
3. The CLARIN.SI research infrastructure

I. Introduction

- Language technologies
  - main paradigm: supervised machine learning
  - need training (manually annotated) language resources
  - need test data
- Empirically supported linguistic investigations:
  - based on real (and, if possible, annotated) language data
- Annotated language resources needed for each language
- Where can we get such resources for Slovene and other South-Slavic languages?

Language resources

1. Corpora:
   - uniformly encoded and document collection of texts
   - explicit criteria for text selection
   - annotated (morphosyntax, lemmatisation, syntax, named entities, …)
   - reference/specialised; mono/multilingual; text/speech
2. Lexicons:
   - the vocabulary of a language
   - words / phrases
   - morphosyntax, syntax, semantics, translations, external and internal links
3. Models:
   - data that enables a program to annotate text in a certain language for a certain level of annotation
   - e.g. Stanford-NLP model for parsing of Slovene; Moses model translating Slovene to English

Resource reuse

- Traditional approach:
  - develop language resources for each project separately
  - resources unavailable to other researchers
- Disadvantages:
  - the development of a language resource can be very costly: waste of time and money if it is done several times
  - later researchers cannot replicate or improve the initial results
  - supports the monopoly of institutions that produced the resources
  - the resources cannot be used to help in the development of products

Open access to the results of research projects

- No barriers to publications and data:
  - saves of time and money;
  - avoids repetition of work;
  - encourages cooperation;
  - makes the research process more transparent
  - stimulates innovation
- A very strong trend in EU (H2020) projects
- Problems in enabling open access to language resources:
  - copyright on texts
  - privacy protection (GDPR), including the right to be forgotten,
  - terms-of-use by owners of social media platforms (e.g. Twitter)

Research infrastructures

Research Infrastructures are facilities that provide resources and services for research communities to conduct research and foster innovation.

Research infrastructures

- Beginning, 2002: ESFRI (European Strategy Forum on Research Infrastructures),
- Roadmap: proposed 15 (2016: 21) RIs, some already established as ERICs (EU legal entity: European RI Consortium)
- Humanities and Social Sciences:
  - CLARIN ERIC: Common Language Resources and Technology Infrastructure
  - DARIAH ERIC: Digital Research Infrastructure for the Arts and Humanities
  - CESSDA: Consortium of European Social Science Data Archives

II. CLARIN ERIC

Common Language Resources and Technology Infrastructure

Common Language Resources and Technology Infrastructure

- Vision: digital language resources and technologies for all (European) languages are available for researchers in the humanities and social sciences
- Repository for long-term, extensive archiving and enabling access to language resources and technologies
- Contribution to preserving and supporting the European multilingual cultural heritage
- A collaborative paradigm in the compilation of language resources and the development of language tools, enabling re-use, experiment replicability and reproducibility

CLARIN

- Enable access to existing solutions in a unified infrastructure
- Consulting & teaching how to adapt tools and resources to specific research needs
- Legal, technical aspects of distribution
- Contribution to standardisation of resources and tools

CLARIN ERIC

- 21 member states + 4 observers
- Based in the Netherlands:
  director, support staff, strong DH / CL community
- Committees: BoD, NCF, SCTC, …
- Aggregators: Virtual Language Observatory
- Most work is done by the national consortia
- Annual conference:
  - authors of accepted paper go for free
  - session for PhD students
  - book of abstracts (post-conference papers), posters, bazaar, invited talks etc.

III. CLARIN.SI

CLARIN.SI

- CLARIN Slovenia, start of work in 2014
- Organised as a consortium of (currently) 11 partners:
  - 4 universities: Ljubljana, Maribor, Nova Gorica, Primorska
  - 4 research institutes: ZRC SAZU, IJS, INZ, Trojina
  - 2 companies: Amebis, Alpineon
  - 1 society: Slovenian society for language technologies, SDJT
- Headquarters at IJS:
  - E8: Dept. for Knowledge Technologies
  - E3: Laboratory for Artificial Intelligence
  - CMI: Networking Infrastructure Centre

CLARIN.SI

- Repository
  - long term archiving of language resources (and tools)
  - also, for software and manually annotated datasets: CLARINSI GitHub virtual organisation & http://gitlab.clarin.si
- Web services:
  - 2 concordancers (corpus analysis)
  - automatic annotation
  - WebAnno platform for manual annotation (e.g. training sets)
- Support for events:
  - Conference „Language Technologies and digital humanities“ (1998, …, 2016, 2018, 2020)
  - JOTA lectures “Jezikovnotehnološki abonma”: VideoLectures
  - XVIII EURALEX International Congress, Ljubljana, 2018
  - 22nd Intl. Conf. on Text Speech and Dialogue, Ljubljana, 2019
- Support for development and archiving language resources and tools
  - support for resource update for archiving in the repository (cca 500 EUR)
  - larger projects for development: 2018: 8, 2019: 7 projects (cca 6,000 EUR)

IV. Conclusions

- The purpose of CLARIN(.SI) is to support research that need access to language data
  - Digital humanities and social sciences
  - Language Technologies (~ Computational Linguistics)
  - All other fields where language is important
- Open access to resources, tools and services
- CLARIN(.SI) financial support:
  - Organising various types of events
  - Work on specific topics incl. outreach
  - Development or modification of resources
  - Attendance at CLARIN conferences