CLARIN K CENTRE

The CLASSLA knowledge centre for South Slavic languages

Nikola Ljubešić, Jožef Stefan Institute
Alpen-Adria-Universität
2020-05-20

CLASSLA overview

- Knowledge centre of the CLARIN ERIC infrastructure for language resources and technology
- Set-up by CLARIN.SI in March 2019, Bulgarian CLADA-BG joined in July 2019, hoping for Croatian CLARIN-HR to join as well
- Main goals:
  a. Coordination of development of language resources and technologies for the typologically related languages
  b. Joint training activities and helpdesk support for the underlying user base

Predecessor of CLASSLA - ReLDI

- ReLDI - Regional Linguistic Data Initiative
- Project funded by the Swiss Science Foundation (2015-2017)
- Partners - University of Zürich, University of Zagreb, University of Belgrade
- One of its goals - coordination of development of language resources for the HBS macro-language (Bosnian, Croatian, Montenegrin, Serbian)
- Some results of this coordination
  - hrLex and srLex inflectional lexicons (hrLex was developed in a previous FP7 project)
  - SETimes.SR manually annotated corpus of Serbian (pre-annotated with Croatian models)
  - ReLDI-NormTagNER Croatian and Serbian manually annotated Twitter datasets based on the Slovene JANES datasets (national project)
  - Whole linguistic processing pipeline developed initially for Croatian, adapted to Serbian and Slovene

CLASSLA components

- Helpdesk
- Frequently asked questions
- Data repository
- Concordancers
- Web services

CLASSLA components

- Helpdesk (helpdesk.classla@clarin.si)
  - 48-hours response time
  - Information on topics not covered in FAQs
  - Feedback on available resources and technologies (we just compiled a form for this)
  - Requests for new / improved resources or technologies
    - First two Montenegrin corpora (English-Montenegrin subtitles, web corpus)
    - Tokenizers for Slovene, Croatian, Serbian extended on Macedonian and Bulgarian
    - CLASSLA-StanfordNLP pipeline for Slovene, Croatian, Serbian extended to Bulgarian
    - Series of datasets deposited into the data repository (yearly funding available)
- Frequently asked questions (FAQs)
  - Slovene, Croatian, Serbian, extended to Bulgarian, Macedonian and Montenegrin to follow

CLASSLA components

- Data repository
  - CLARIN.SI repository
  - Already containing datasets of other South Slavic languages, expect for the diversity to improve through CLASSLA
  - Manually and automatically annotated corpora, lexicons, models
  - Manually annotated corpora (ssj500k, ReLDI-NormTagNER-hr)
    - Backbone of our language technologies - use machine learning (AI buzzword) to learn to automate a linguistic process (e.g., mapping surface forms to morphosyntactic tags)
    - High accuracy of the annotations, but still NEVER perfect (versioning)
      - The formalism of the linguistic phenomenon to be applied over data
      - The accuracy of the application of the formalism on data

CLASSLA components

- Data repository cont
  - Automatically annotated corpora (srWaC)
    - Accuracy of the annotations depends on the quality of the technologies
    - Morphosyntactic annotation during the last annotation of the srWaC corpus ~90%
    - Morphosyntactic annotation of Serbian nowadays ~95%
    - Question of strictness of technologies - should inflectional lexicons limit hypotheses
- Concordancers
  - NoSkE - open variant of the famous Sketch Engine, lacks user management, word sketches, compiling corpora, BootCat for compiling small web corpora etc.
  - During the ELEXIS project (2018-2022) SkE is free for academic purposes
  - Kontext - user management, different add-ons (spoken corpora, dependency trees etc.)

CLASSLA components

- Web services
  - Services for linguistic processing of text
  - Web interface of the current ReLDIanno services
  - Access via Python library (documentation)
  - Possibility of batch processing (zipped files) and different input formats (txt, doc(x), pdf)
  - Pre-processing for directly compiling corpora into concordancers an option for the future
  - Technologies are as good as
    i. The data they are trained on
    ii. The capacity of the technology to encode regularities from the training data
  - Current technologies in the web service from 2017, covering standard data only
  - Bleeding edge technologies on GitHub (https://github.com/clarinsi/)
  - CLASSLA-StanfordNLP neural pipeline holds currently the state-of-the-art for all South Slavic languages, working on adding it to the new version of the web services

The future of CLASSLA

- Short-term
  - New neural-based web services with models for more languages and different varieties
  - Even newer technologies to be developed inside “Razvoj slovenščine v digitalnem okolju” (2020-2022), natural language processing is currently a moving target!
  - In the same project automatic speech recognition for Slovene is to be developed, already working on ensuring data for training Croatian / HBS models (parliamentary recordings)
  - Increasing presence of Bulgarian, ensuring basic processing of Macedonian
  - Crawling systematically top-level domains to produce up-to-date comparable web corpora
- Mid-term
  - South-Eastern European?
  - South and West Slavic?
  - Slavic? Balto-Slavic?

Three main take-home messages

CLARIN K CENTRE

1. Synergistic approach to developing resources and technologies reduces regularly production cost to less than 50% (often 10%-20%) and improves both quality and comparability
2. Language technologies are as good as the data they are based on
3. A continuous communication between the infrastructure(s) and linguists is of paramount importance
    - Prioritising further developments
    - Working together on language resources - both the formalism and the application of the formalism (annotation) - direct impact on technologies!

CLARIN K CENTRE

The CLASSLA knowledge centre for South Slavic languages

Nikola Ljubešić, Jožef Stefan Institute
Alpen-Adria-Universität
2020-05-20