Overview of CLARIN's key families of resources and tools for SSH research

Darja Fišer [1, 2], Jakob Lenardič [1]

[1] University of Ljubljana, [2] Jožef Stefan Institute
darja.fiser@ff.uni-lj.si, jakob.lenardic@ff.uni-lj.si

CLARIN ERIC
Common Language Resources and Technology Infrastructure

Background, motivation and goals

- Certain resources and tools are especially valuable because they can be used in many SSH disciplines
- A comprehensive transnational overview of the existing resources and tools is still missing
- An evaluation of their integration in the CLARIN infrastructure is needed
- Based on such surveys short- and long-term priorities can be made to improve and expand CLARIN infrastructure

Parliamentary corpora

- 21 corpora identified
- Available for all CLARIN languages except Italian (though smaller datasets exist)
- Roughly half are in CLARIN infrastructure

Czech Parliment Meetings, DK-CLARIN Almensprogligt korpus, Transcripts of Riigikogu, Eduskunta Corpus, Hellenic Parliament Meetings, Proceedings of Norwegian Parl Debates, Riksdag's Open Data, PTPARL, SlovParl, HNC, Hansard Corpus, Europarl Corpus

- Generally difficult to find parliamentary corpora through VLO (with keywords like "parliamentary")
- Incomplete metadata in some cases (e.g. unclear annotation for Estonian and Finnish corpora)
- Priority 1: improve metadata for existing corpora on national repositories so that they can be harvested by VLO (e.g. Riksdag's Open data)
- Priority 2: integrate the existing corpora missing in the infrastructure and make them findable through the VLO (e.g. Austrian parliamentary corpus)

Newspaper corpora

- In preparation
- 34 corpora identified thus far
- Collecting information untill 1 October
- Contributions and suggestions for further surveys warmly welcome!

Computer-mediated corpora

- 15 corpora identified
- Available for 13 languages – de (3), lt (3), nl (2), ee (1), fi (1), fr (1), pl (1), Welsh, 2 multilingual
- Most common data types: forums, blogs, tweets
- 8 part of CLARIN infrastructure

Estonian Mixed Corpus, Suomi24 corpus, LITIV, SoNaR New Media, NTAP, Dortmund Chat, Dereko

- 14 specialised datasets (8 are part of CLARIN), mostly from Twitter, developed for various tasks

5 for sentiment analysis, 1 for NER, 1 for entity linking, rest miscellaneous

- 13 tools identified, 10 in CLARIN (e.g. GATE Twitter collector, janes-ner), 3 not in CLARIN (Hunaccent, twython, dmi-tcat, Tweet NLP)

Parallel corpora

- 77 corpora identified
- Largest corpora in terms of represented languages: Parallel Bible Corpus (more than 100), OPUS Corpus (more than 50)
- Largest corpus in size: OPUS Helsinki Version, 2.7 billion tokens
- 51 are part of the CLARIN infrastructure
  - 11 through concordancers
  - 11 corpora for DL through LINDAT; 6 through CLARIN.SI; 1 CLARIN PL, 1 CLARINO; rest unavailable
- Various problems with metadata: unclear language direction, unclear alignment for roughly half of the corpora, unclear size, unclear licence (though most CC-BY)

A European Research Infrastructure

CLARIN ERIC was established in 2012; it is a landmark in the 2016 ESFRI roadmap.

www.clarin.eu

clarin@clarin.eu
@CLARINERIC
facebook.com/CLARINERIC
github.com/clarin-eric

Acknowledgements

We would like to thank everyone who has contributed to the overview by e-mailing us the data or filling out the spreadsheets. We would also like to thank the participants of the workshops for their invaluable comments.