CLARIN ERIC

Common Language Resources and Technology Infrastructure

Understanding culture and society with language resources and tools offered through the CLARIN Research Infrastructure

Open access to digital language resources that capture social and cultural diversity can help advance the **social sciences and humanities** at large

RESOURCE FAMILIES

The CLARIN Resource Families initiative provides a **user-friendly overview of the available language resources** in the CLARIN infrastructure for **researchers from digital humanities, social sciences and computer science.**

The overviews are **organized according to the types of data** in the re sources and include listings sorted by language.

The listings include the **most important metadata and brief descriptions,** such as resource size, text sources, time periods, annotations and licences as well as links to download pages and concordancers, whenever available.

Additional information:

- **Overview of other existing valuable language resources** which have not yet been integrated in the infrastructure.
- **Overview of other relevant materials** such as the thematic CLARIN workshops and tutorials and their accompanying videolectures, as well as a list of key publications on the resources surveyey.

NUMBERS in November 2019:

- 8 corpora families: Computer-mediated communication corpora, Historical corpora, L2 learner corpora, Literary corpora, Newspaper corpora, Parallel corpora, Manually annotated corpora, Parliamentary corpora, Spoken corpora
- 5 families of lexical resources: Lexica, Dictionaries, Conceptual Resources, Glossaries, Wordlists

Horizon 2020 projects

SSHOC

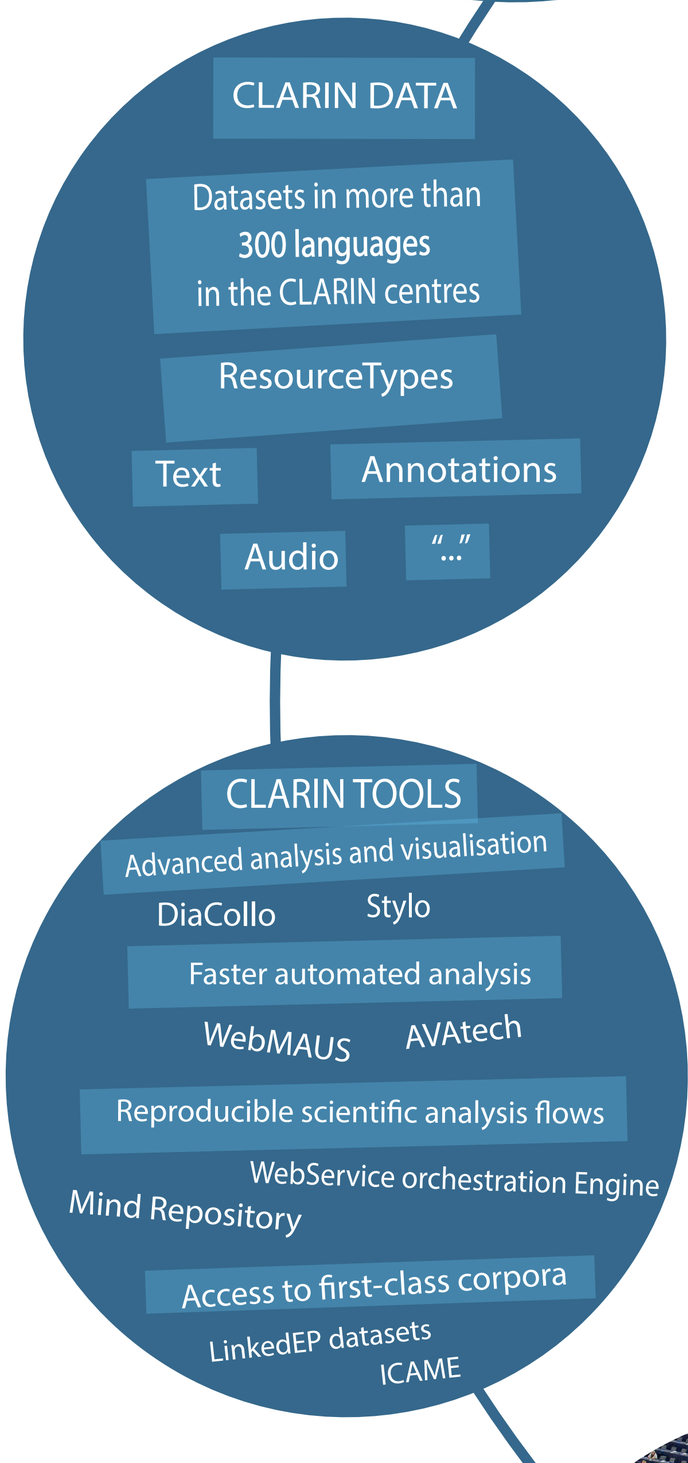
"Social Sciences and Humanities Open Cloud" (SSHOC) unites 20 partner organisations and a further 27 associates realising the vision articulated by the European Commission in 2016: To offer researchers in the social sciences and humanities seamless access to a full and unified panorama of flexible, scalable, relevant data and the services, tools and training required to make optimal use of that data.

CLARIN participation in SSHOC: CLARIN brings to SSHOC its expertise in Language Technology, and some of its core technical infrastructure components. Generalizing and adapting our services for **processing data from the social sciences and cultural heritage** where needed. The SSH will profit from this as CLARIN will from the assets and expertise of the other SSHOC stakeholder infrastructures. Together we will also assure the position of the SSH in the new to develop EOSC.

EOSC-hub

"EOSC-hub: services for the European Open Science Cloud" brings together multiple service providers to create the Hub: a single contact point for European researchers and innovators to discover, access, use and reuse a broad spectrum of resources for advanced data-driven research.

CLARIN participation in EOSC-hub: integration of CLARIN thematic services (Virtual Language Observatory, Virtual Collection Registry, Language Resource Switchboard) into the EOSC-hub. These are based on the Component Metadata framework. This will lead to closer connections between neighbouring research communities.



PARLIAMENTARY DATA IN CLARIN

22 parliamentary corpora surveyed

1 MULTILINGUAL: Europarl (21 languages)

17 MONOLINGUAL: **15** languages:

1 Czech	1 French	2 German	2 Norwegian	3 Slovenian
1 Danish	1 Estonian	2 Greek	1 Polish	1 Swedish
2 English	1 Finnish	1 Lithuanian	1 Portugues	1 Croatian

AVAILABILITY

- 3 through a concordancer only
- 10 for download
- 7 both

LICENCE

- 12 CC-BY
- 1 CLARIN ACA
- 1 CLARIN PUB

SIZE

- 7 small (<10 million tokens)
- 9 medium (10–100 million tokens)
- 5 large (>100 million tokens)

ANNOTATION

- 9 lemmatised
- 8 PoS-tagged

CLARIN in a nutshell

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for academics, researchers, students, journalists, and citizen-scientists in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a single sign-on environment
- that serves as an ecosystem for **knowledge sharing**
- ready for **integration in EOSC** (European Open Science Cloud)
- **CLARIN** adheres to the **FAIR** data principles

Virtual Language Observatory (VLO)

- The VLO covers many languages, both national and regional
- The advantage of the VLO is **faster identification of relevant resources**, allowing users to **re-use resources that already exist,** rather than having to produce their own.
- Additionally, the VLO allows users who **create or collect their own datasets** to make them better visible to others through publication of the metadata in the VLO.

Language Resource Switchboard (LRS)

- Tool that helps you to find a matching language processing web application for your data.

Long-term Preservation

- CLARIN infrastructure makes sure that language resources can be archived and made available to the community in a **safe and sustainable manner.**
- CLARIN helps researchers to store their resources, e.g. corpora, lexica, audio and video recordings, annotations, grammars, etc. **At least one CLARIN data centre in each country offers a depositing service.** These centres are willing to store the resources in their repository and assist with the technical and organisational details.

CLARIN ERIC

CLARIN ERIC, Drift 10, 3512 BS Utrecht, the Netherlands

clarin@clarin.eu

www.clarin.eu

Read further information about CLARIN Resource Families: