What is CLARIN?

By Cristina Grisot (CLARIN ERIC/UZH)

30 September 2025

CLARIN Annual Conference

Vienna, Austria

CLARIN

CLARIN - Common Language Resources and Technology Infrastructure

1. **European Research Infrastructure Consortium (ERIC) since 2012**
2. Scholars in **Social Sciences and Humanities (SSH)** can use the infrastructure to access:
   a. **digital language data** (written, spoken, video or multimodal)
   b. **tools** to discover, analyse, combine data wherever they are located
   c. **through a single sign-on** environment (**you can get an account!**)
1. Ecosystem for **knowledge exchange and training**
2. Actively involved in shaping the **European Open Science Cloud (EOSC)** (clarin.eu/eosc)

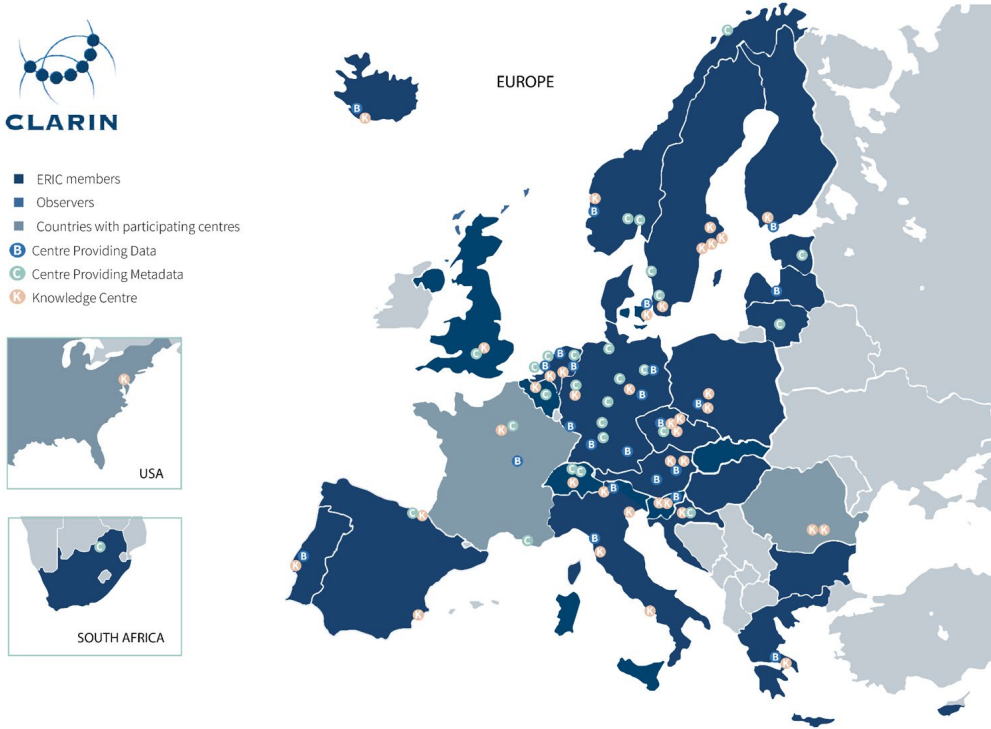
CLARIN for Open Science

- Promotion of sharing & reuse of language data through sustainable **data registries**
- Enhancement &deployment of **interoperability** of language data & services
  - common metadata framework
  - distributed network of **FAIR certified data repositories** for language data
- Promotion of
  - **comparative** perspectives
  - **multidisciplinary** collaboration
  - **transnational** research
  - **responsible** data science
- Support for **linguistic diversity**
  - data covering many languages
  - tools for many languages
  - language resources in all modalities
  - discipline- & language-agnostic

CLARIN's Network (August 2025)

A distributed **European Infrastructure Consortium (ERIC)** consisting of:

- **27** members



https://www.clarin.eu/content/national-consortia

CLARIN's Centres (August 2025)

A distributed network of >70 centres

CLARIN B-centre

22 CTS-certified data centres

Strong focus on **FAIRness** & **interoperability**:

- Federated login
- Central metadata harvesting for easy discovery
- Chained services

☞ **Search for B-centres**

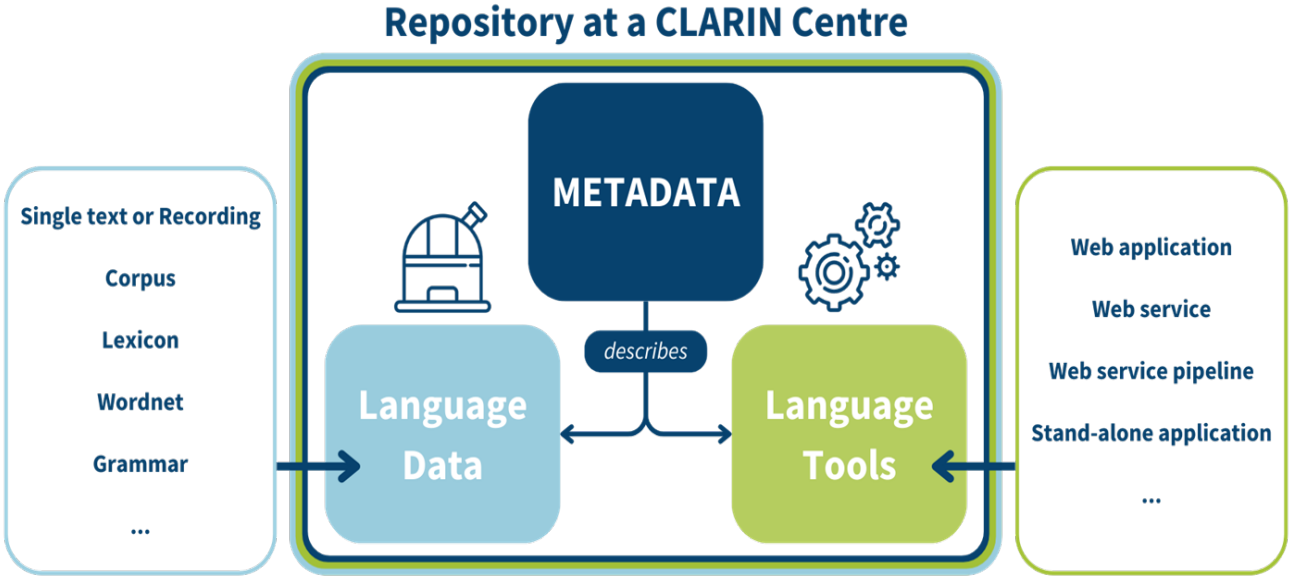
CLARIN K-centre

39 Knowledge centres

Operated by a single institute/group or as a distributed structure **covering a large number of research topics, languages and resource types.**

☞ **Search for K-centres**

How CLARIN Works



How CLARIN Works

Home / Web Applications

Web Applications

These shortcuts provide quick access to the central CLARIN web applications:

Content Search

Federated Data Search Engine

Language Resource Switchboard

Find a suitable tool to process language data

Virtual Language Observatory

Metadata search interface

Virtual Collection Registry

Publish and access digital bookmarks

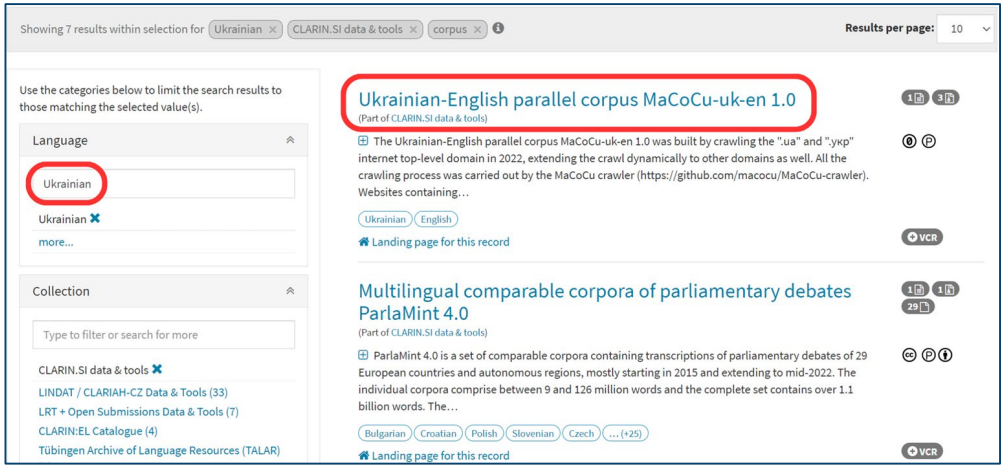
https://www.clarin.eu/content/web-applications

CLARIN and FAIR Principles

Findable	Virtual Language Observatory, Virtual Collection Registry, Federated Content Search
Accessible	Certified repositories, PIDs, Metadata, Federated Login
Interoperable	Language Resource Switchboard, Vocabularies, Metadata
Reusable	Licences, Standards & Formats, Active Metadata Curation

Virtual Language Observatory (VLO)

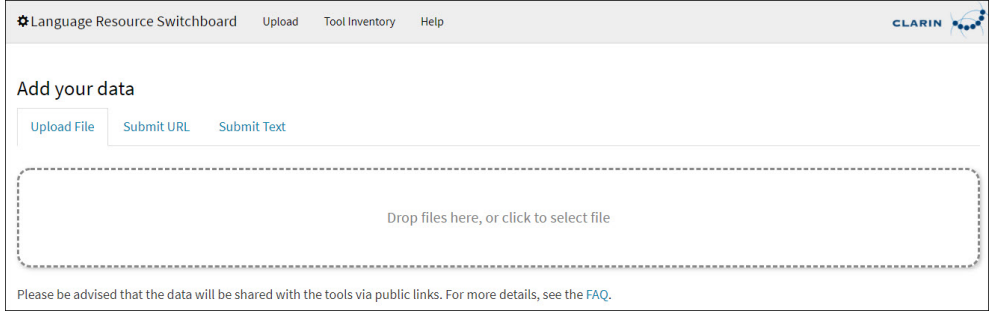
- Facet search to explore **>500,000 language resources, tools, services**
- Metadata to describe resources
- **Persistent Identifiers (HDL)** to access the landing pages of the resources and cite them
- Availability and/or **Licencing**
- Technical details
- Process with **Switchboard**
- Compile search results into a **Virtual Collection**

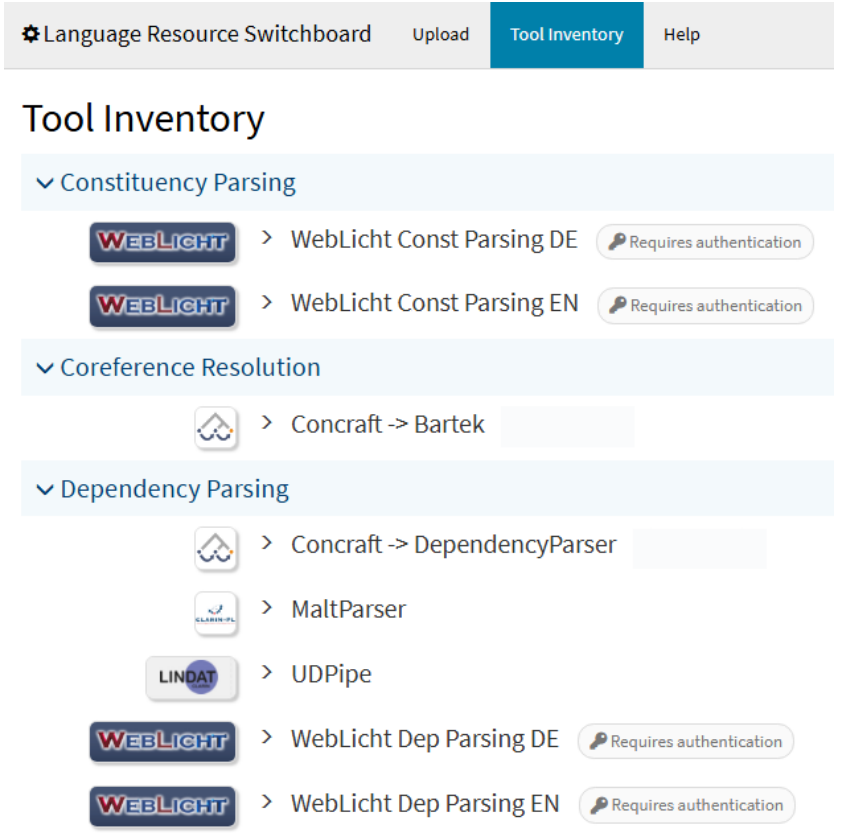


Language Resource Switchboard

Helps you find and use **NLP tools** that can process your data:

- Your own files
- Any data that is publicly accessible, e.g. a resource that you found in the VLO





https://switchboard.clarin.eu/

CLARIN Resource Families

Corpora

- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Corpora of Disordered Speech
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources

- Conceptual Resources
- Dictionaries
- Glossaries
- Language Models
- Lexica
- Wordlists

Tools

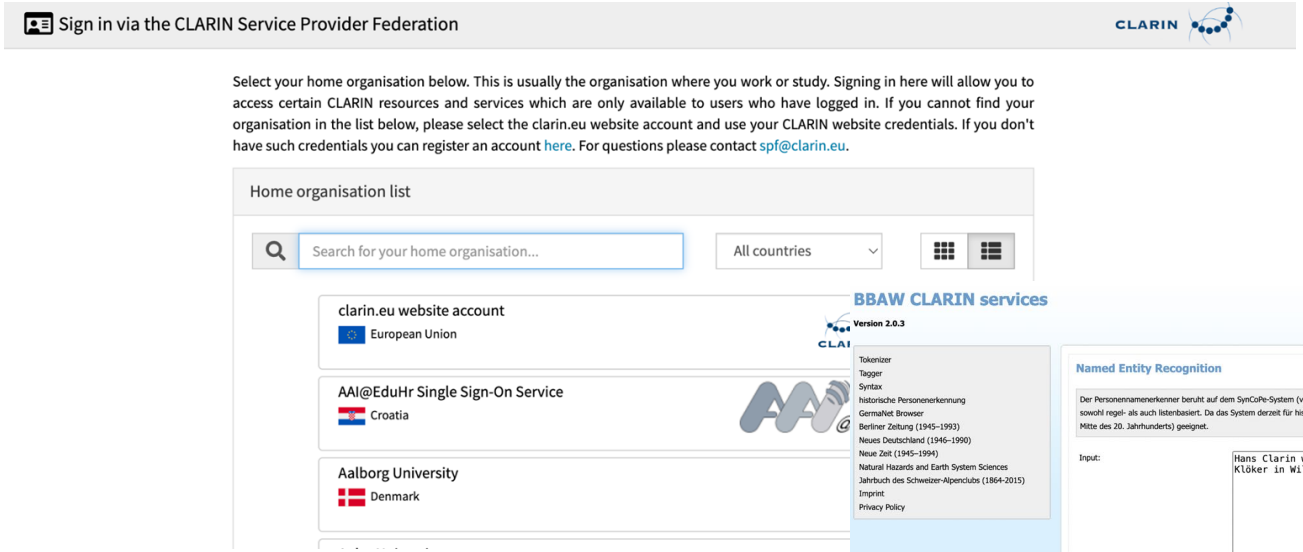
- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

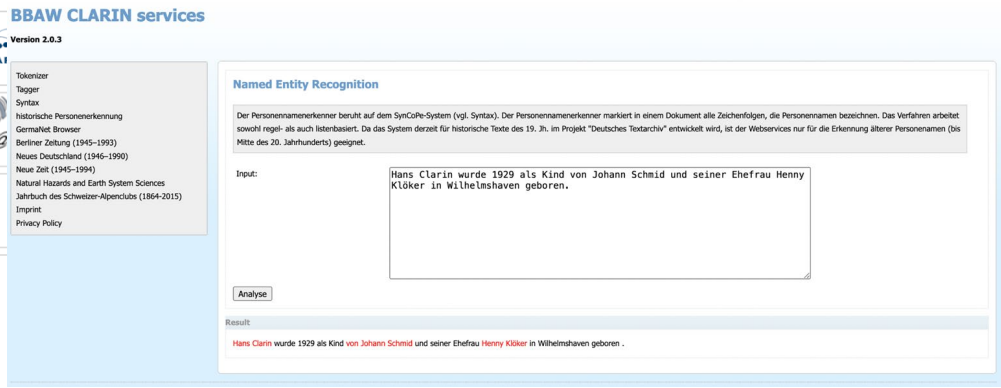
- Well-curated corpora and lexical resources organised per data type and language, which can be downloaded directly
- NLP tools
- Licence information
- Some corpora are available via concordancers, e.g. Korp, Corpuscle, or KonText, noSketch Engine.
- Links to publications and training materials

https://www.clarin.eu/resource-families

CLARIN Single Sign-On (SSO)

Restricted resources and tools in CLARIN can be accessed with your university credentials!





Easy Access to Protected Resources

Get easy access to protected resources, with your institutional username and password

https://www.clarin.eu/content/easy-access-protected-resources

Examples of Use

Possible scenarios for exploring the central services:

- ➔ I need to **find recordings** of spoken Catalan
- ➔ Can you help me **find the resources and tools** to research the **stylistic differences** in 20th century novels translated to Polish from English between British and American authors?
- ➔ Can you help me find evidence of a **specific linguistic expression in various corpora**?
- ➔ I have a **TEI file on my computer** and I would like to have dependency trees for all of the sentences that it contains.

What is CLARIN

By Cristina Grisot (CLARIN-CH/UZH)

30 September 2025
CLARIN Annual Conference
Vienna, Austria

CLARIN