

What is CLARIN?

By Cristina Grisot (CLARINCH/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)

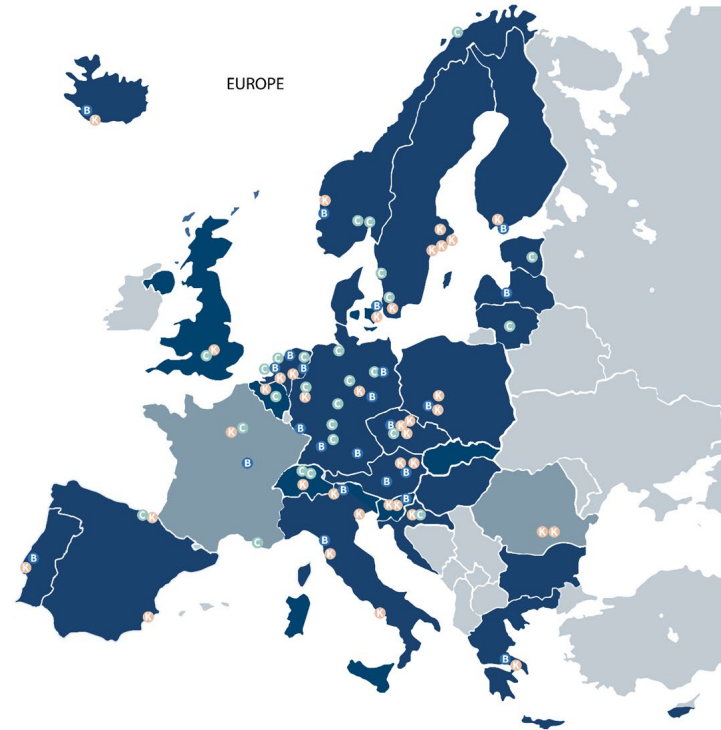
A distributed **European Infrastructure Consortium (ERIC)** consisting of:

- 27 members



CLARIN

- ERIC members
- Observers
- Countries with participating centres
- ⓑ Centre Providing Data
- ⓐ Centre Providing Metadata
- Ⓚ Knowledge Centre





CLARIN's Centres (August 2025)

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](#)

A distributed network of >70 centres

CLARIN
K-centre

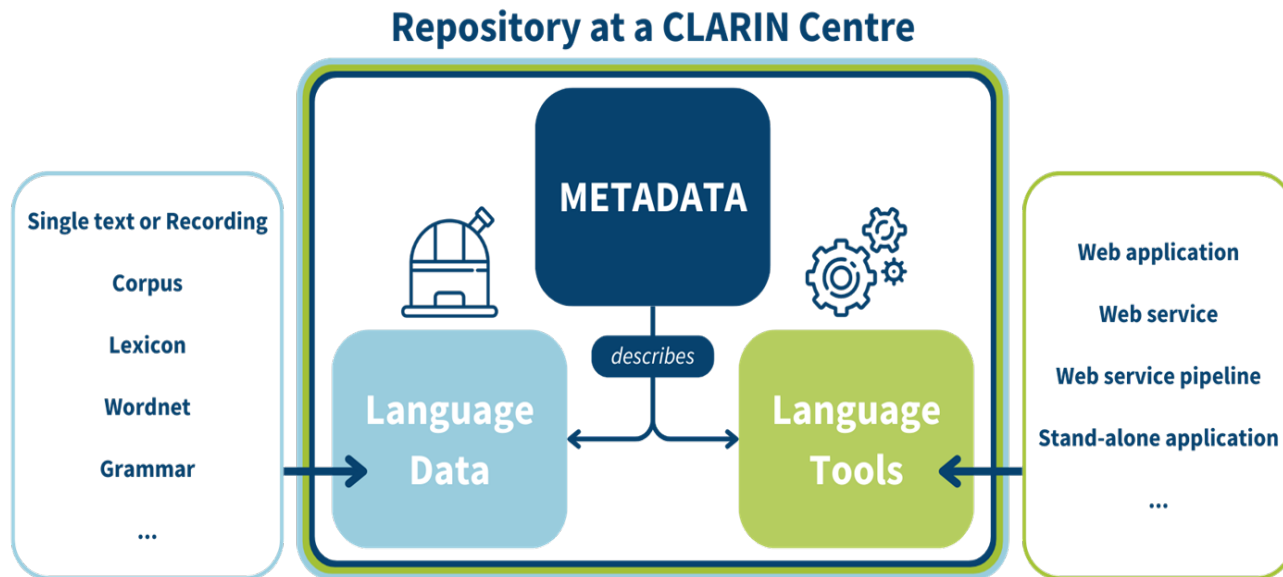


39 Knowledge centres

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



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**

The screenshot shows a search interface with the following elements:

- Search filters: Ukrainian, CLARIN.SI data & tools, corpus. Results per page: 10.
- Language filter: Ukrainian is selected and circled in red.
- Collection filter: A search box with the text "Type to filter or search for more".
- Search results:
 - Ukrainian-English parallel corpus MaCoCu-uk-en 1.0** (Part of CLARIN.SI data & tools) - This result is circled in red. It includes a description of the corpus and a landing page link.
 - Multilingual comparable corpora of parliamentary debates ParlaMint 4.0** (Part of CLARIN.SI data & tools) - Includes a description and a landing page link.

“ Please use the following text to cite this item or export to a predefined format:

BIBTEX CMDI

Bañón, Marta; et al., 2023, *Ukrainian-English parallel corpus MaCoCu-uk-en 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1858>.





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


The screenshot shows the 'Language Resource Switchboard' interface. At the top, there is a navigation bar with 'Language Resource Switchboard', 'Upload', 'Tool Inventory', and 'Help'. Below this, the 'Add your data' section contains three buttons: 'Upload File', 'Submit URL', and 'Submit Text'. A large dashed box below these buttons contains the text 'Drop files here, or click to select file'. At the bottom of the section, a small note reads: 'Please be advised that the data will be shared with the tools via public links. For more details, see the FAQ.'

Language Resource Switchboard Upload **Tool Inventory** Help


Tool Inventory

Constituency Parsing


 > WebLicht Const Parsing DE 


 > WebLicht Const Parsing EN 


Coreference Resolution

 > Concraft -> Bartek

Dependency Parsing

 > Concraft -> DependencyParser

 > MaltParser

 > UDPipe

 > WebLicht Dep Parsing DE 

 > WebLicht Dep Parsing EN 

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

CLARIN Single Sign-On (SSO)

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

Sign in via the CLARIN Service Provider Federation



Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account [here](#). For questions please contact spf@clarin.eu.

Home organisation list

Search for your home organisation... All countries

- clarin.eu website account
European Union
- AAI@EduHr Single Sign-On Service
Croatia
- Aalborg University
Denmark

BBAW CLARIN services
Version 2.0.3

- Tokenizer
- Tagger
- Syntax
- historische Personenerkennung
- GermaNet Browser
- Berliner Zeitung (1945-1993)
- Neues Deutschland (1946-1990)
- Neue Zeit (1945-1994)
- Natural Hazards and Earth System Sciences
- Jahrbuch des Schweizer-Alpenclubs (1864-2015)
- Imprint
- Privacy Policy



Easy Access to Protected Resources

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

Named Entity Recognition

Der Personennamenerkennung beruht auf dem SynCoPe-System (vgl. Syntax). Der Personennamenerkennung markiert in einem Dokument alle Zeichenfolgen, die Personennamen bezeichnen. Das Verfahren arbeitet sowohl regel- als auch listenbasiert. Da das System derzeit für historische Texte des 19. Jh. im Projekt "Deutsches Textarchiv" entwickelt wird, ist der Webservices nur für die Erkennung älterer Personennamen (bis Mitte des 20. Jahrhunderts) geeignet.

Input: Hans Clarin wurde 1929 als Kind von Johann Schmid und seiner Ehefrau Henny Klöcker in Wilhelmshaven geboren.

Analyse

Result: Hans Clarin wurde 1929 als Kind von Johann Schmid und seiner Ehefrau Henny Klöcker in Wilhelmshaven geboren.



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 (CLARINCH/UZH)

30 September 2025

CLARIN Annual Conference

Vienna, Austria

CLARIN

