The role of CLARIN in supporting SSH scholars

Darja Fišer

21 November 2025

National EOSC Tripartite Event

Ljubljana, Slovenia

CLARIN

The CLARIN Dream

A humanities researcher in Maribor consults the CLARIN catalogue and:

- finds relevant text collections in data repositories in Slovenia, Austria and Hungary
- a named-entity recognition tool in the Czech Republic
- performs research on this material, with support from a centre of expertise in the Czech Republic
- uses them to build an enriched version of the text collections, benefitting from the CLARIN interoperability standards
- deposits the annotated text collections in a data repository in Slovenia for later use by others

… and can do all this from their own desk

CLARIN in a Nutshell

1. *‘Common Language Resources and Technology Infrastructure’*
2. **ESFRI** roadmap (2006), **ESFRI** ERIC status (2012), Landmark (2016)
3. Easy and sustainable access for scholars in Social Sciences and Humanities (SSH):
    4. **digital language data** (written, spoken, video or multimodal)
    5. **tools** to discover, analyse, combine data wherever they are located
    6. **single sign-on** environment (**you all can get an account**)
7. Ecosystem for **knowledge exchange**
8. Member of the **SSH Open Cluster**

https://www.clarin.eu/content/clarin-nutshell

CLARIN and Open Science

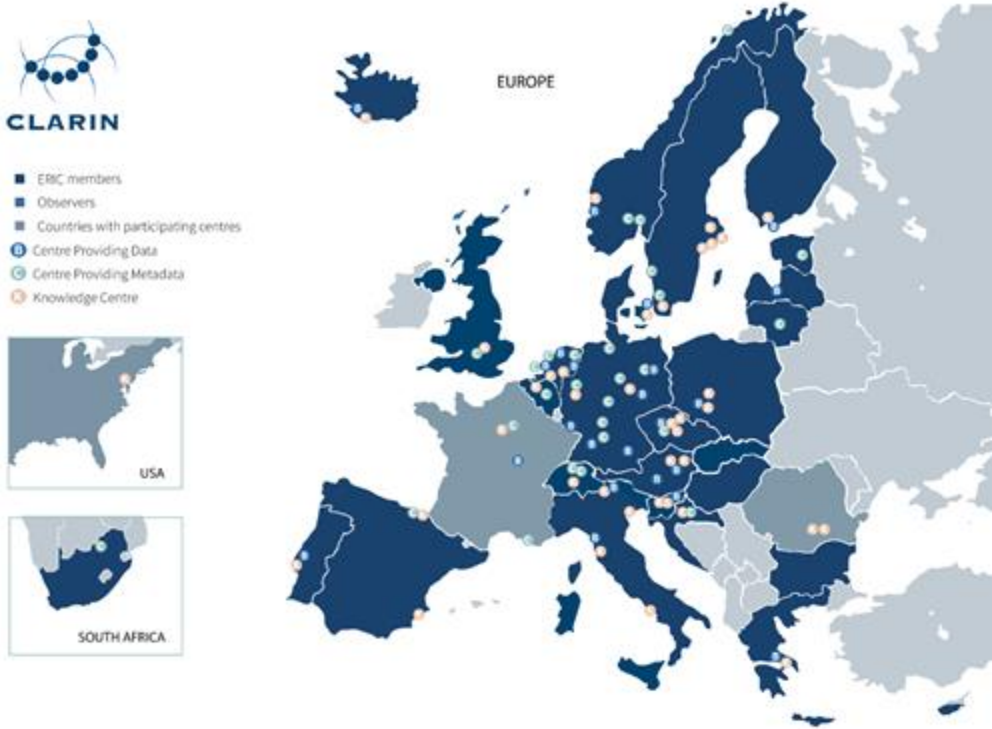
- Promotion of sharing and re-use of language data through **sustainable data registries**
- Adherence to **FAIR data principles**
- Enhancement and deployment of **interoperability** of **language data & services**
  - **common metadata framework**
  - distributed network of FAIR **certified data repositories** for language data

- Promotion of
  - studying language data from a **comparative** perspective
  - **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-** and **language-agnostic**

CLARIN’s Countries

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

- **27** members



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

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

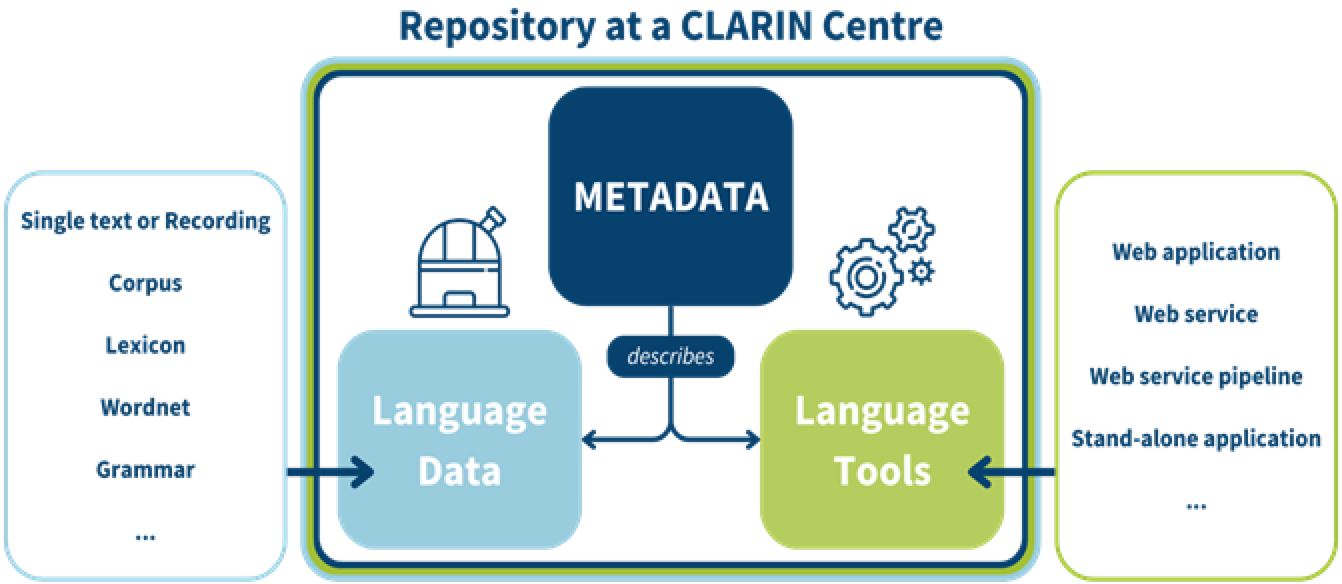
CLARIN K-centre

39 Knowledge centres

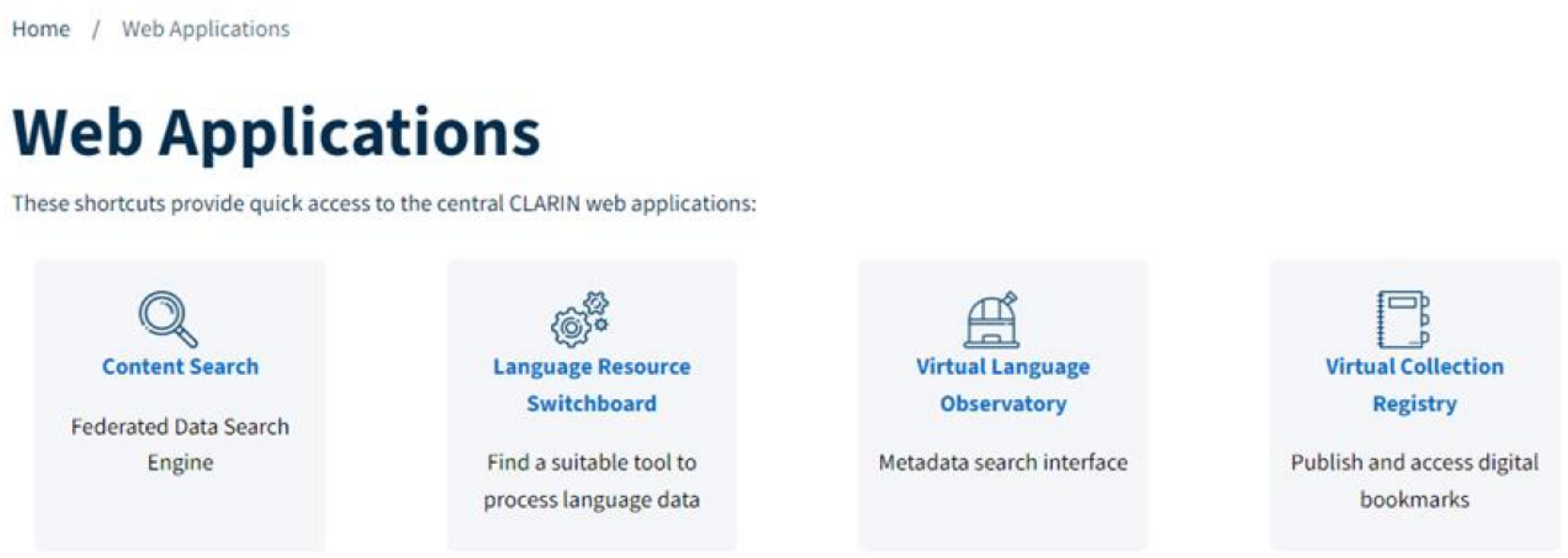
Operated by a single institute/group or as a distributed structure **covering a large number of research topics, languages and resource types.**

👉 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

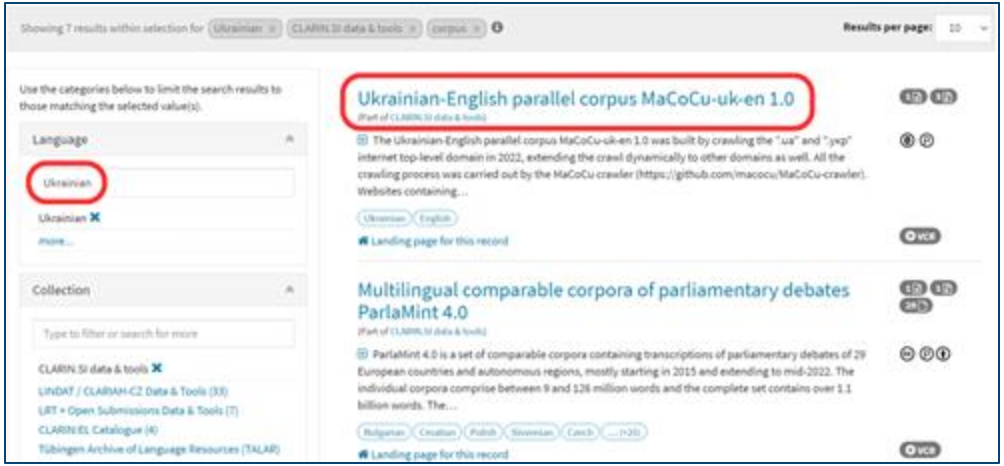
Metadata search interface

Virtual Collection Registry

Publish and access digital bookmarks

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**



Showing 7 results within selection for Ukrainian × CLARIN.SI data & tools × corpus ×

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Ukrainian

Ukrainian ✖

more...

Collection

Type to filter or search for more

CLARIN.SI data & tools ✖

LINDAT / CLARIAH-CZ Data & Tools (33)

LRT + Open Submissions Data & Tools (7)

CLARIN:EL Catalogue (4)

Tübingen Archive of Language Resources (TALAR)

Ukrainian-English parallel corpus MaCoCu-uk-en 1.0
(Part of CLARIN.SI data & tools)

The Ukrainian-English parallel corpus MaCoCu-uk-en 1.0 was built by crawling the ".ua" and ".укр" internet top-level domain in 2022, extending the crawl dynamically to other domains as well. All the crawling process was carried out by the MaCoCu crawler (https://github.com/macocu/MaCoCu-crawler). Websites containing...

Ukrainian English

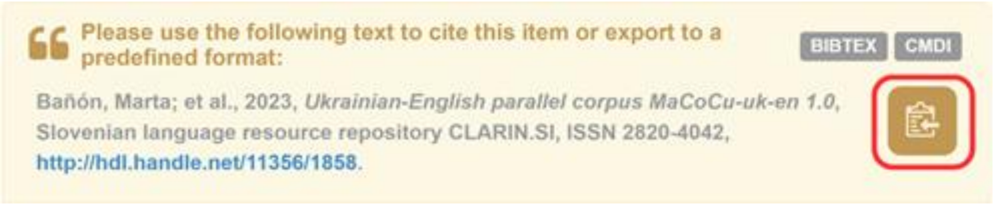
Landing page for this record

Multilingual comparable corpora of parliamentary debates ParlaMint 4.0
(Part of CLARIN.SI data & tools)

ParlaMint 4.0 is a set of comparable corpora containing transcriptions of parliamentary debates of 29 European countries and autonomous regions, mostly starting in 2015 and extending to mid-2022. The individual corpora comprise between 9 and 126 million words and the complete set contains over 1.1 billion words. The...

Bulgarian Croatian Polish Slovenian Czech ... (+20)

Landing page for this record

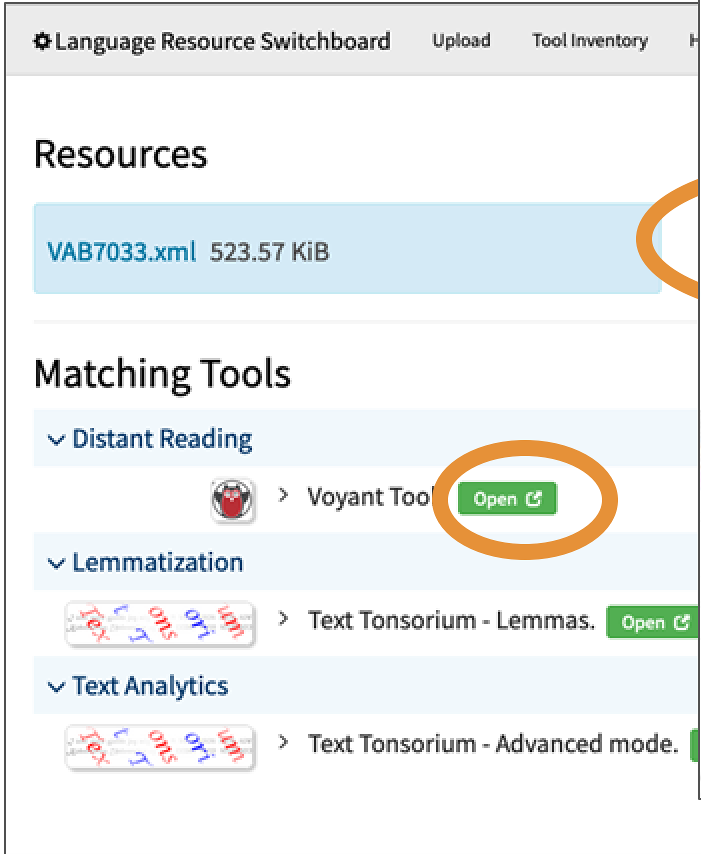


Please use the following text to cite this item or export to a predefined format:

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.

https://vlo.clarin.eu

Switchboard

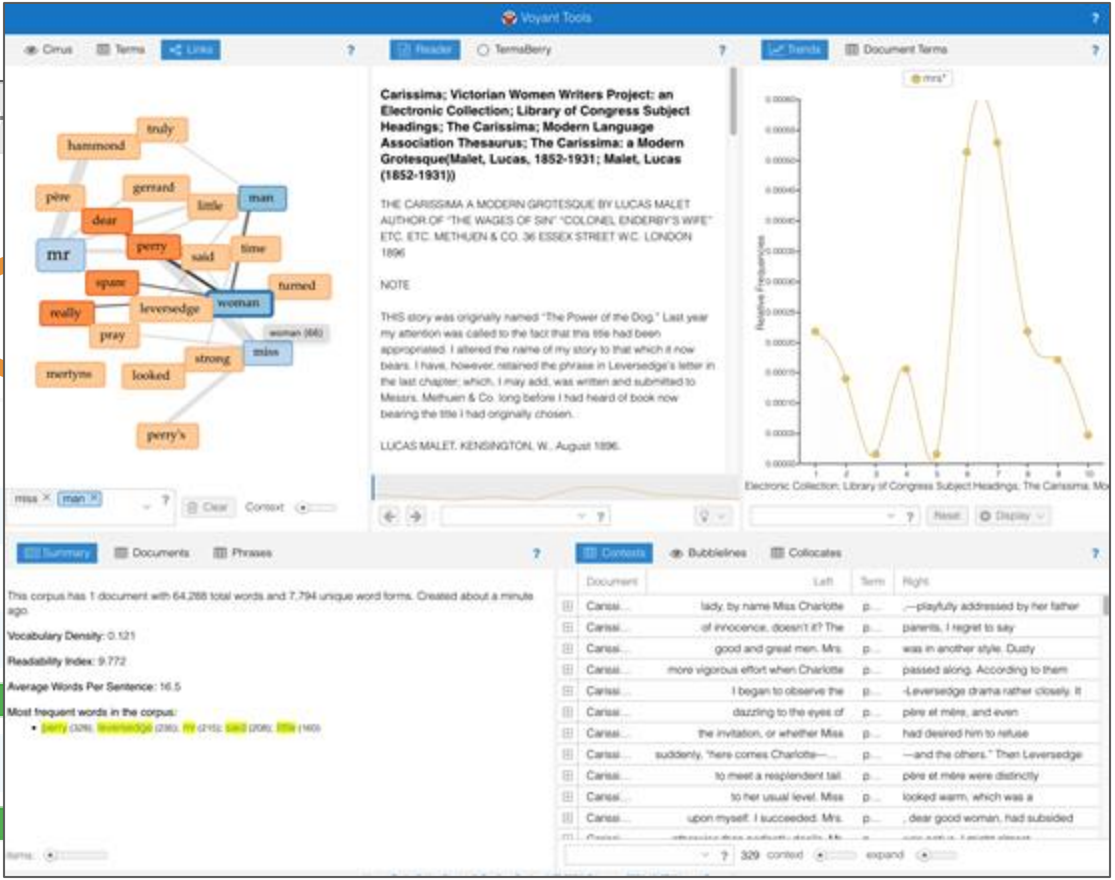


Resources

VAB7033.xml 523.57 KiB

Matching Tools

- Distant Reading
  - Voyant Tool[s] Open
- Lemmatization
  - Text Tonsorium - Lemmas. Open
- Text Analytics
  - Text Tonsorium - Advanced mode.



Carissima; Victorian Women Writers Project: an Electronic Collection; Library of Congress Subject Headings; The Carissima; Modern Language Association Thesaurus; The Carissima: a Modern Grotesque(Malet, Lucas, 1852-1931; Malet, Lucas (1852-1931))

THE CARISSIMA A MODERN GROTESQUE BY LUCAS MALET AUTHOR OF "THE WAGES OF SIN" "COLONEL ENDERBY'S WIFE" ETC. ETC. METHUEN & CO. 36 ESSEX STREET W.C. LONDON 1896

NOTE

THIS story was originally named "The Power of the Dog." Last year my attention was called to the fact that this title had been appropriated. I altered the name of my story to that which it now bears. I have, however, retained the phrase in Leversedge's letter in the last chapter; which, I may add, was written and submitted to Messrs. Methuen & Co. long before I had heard of book now bearing the title I had originally chosen.

LUCAS MALET. KENSINGTON, W., August 1896.

Electronic Collection; Library of Congress Subject Headings; The Carissima; Mo

This corpus has 1 document with 64,288 total words and 7,794 unique word forms. Created about a minute ago.

Vocabulary Density: 0.121

Readability Index: 9.772

Average Words Per Sentence: 16.5

Most frequent words in the corpus:

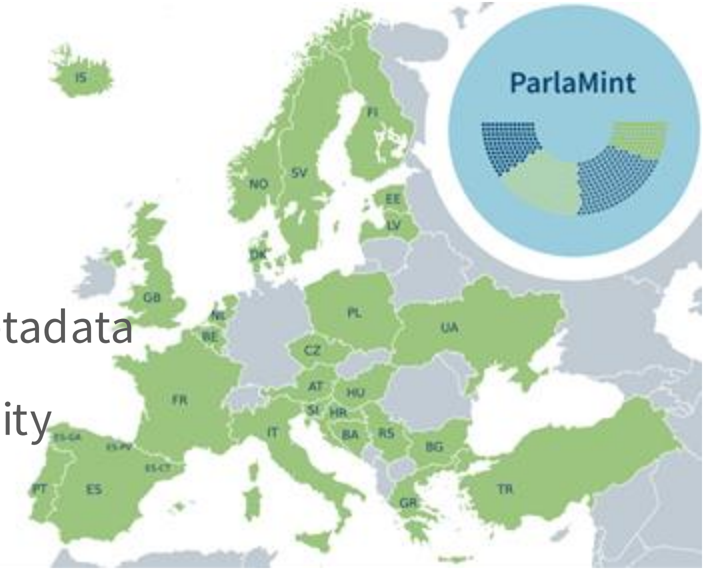
- perry (328); leversedge (236); mr (215); said (208); little (160)

Document	Left	Term	Right
Carissi...	lady, by name Miss Charlotte	p...	,—playfully addressed by her father
Carissi...	of innocence, doesn't it? The	p...	parents, I regret to say
Carissi...	good and great men. Mrs.	p...	was in another style. Dusty
Carissi...	more vigorous effort when Charlotte	p...	passed along. According to them
Carissi...	I began to observe the	p...	-Leversedge drama rather closely. It
Carissi...	dazzling to the eyes of	p...	père et mère, and even
Carissi...	the invitation, or whether Miss	p...	had desired him to refuse
Carissi...	suddenly, "here comes Charlotte—...	p...	—and the others." Then Leversedge
Carissi...	to meet a resplendent tall	p...	père et mère were distinctly
Carissi...	to her usual level. Miss	p...	looked warm, which was a
Carissi...	upon myself. I succeeded. Mrs.	p...	, dear good woman, had subsided

329 context expand

https://voyant-tools.org

Flagship Project - Parlamint



Parliamentary proceedings of 29 European parliaments

Corpora linguistically annotated, named entities, rich metadata

Great example of CLARIN collaboration and interoperability

The ParlaMint corpora are **interoperable**:

- annotated within a specifically developed TEI-based schema

The ParlaMint corpora are **comparable:**

- the same overlapping periods: 2015–2022
- the same metadata (related to speakers, parties, sessions, reactions, etc.)
- the same types of linguistic annotations

Interoperability and comparability are further increased also by **machine translations** into English

https://www.clarin.eu/parlamint

Thank you!
Questions & comments welcome!

www.clarin.eu
darja.fiser@clarin.eu

CLARIN