

# CLARIN-IT

the Italian Common Language Resources and Technology Infrastructure



## CLARIN ERIC and CLARIN-IT

Monica Monachini - *National Coordinator CLARIN-IT*

Francesca Frontini - *CLARIN Board of Directors*

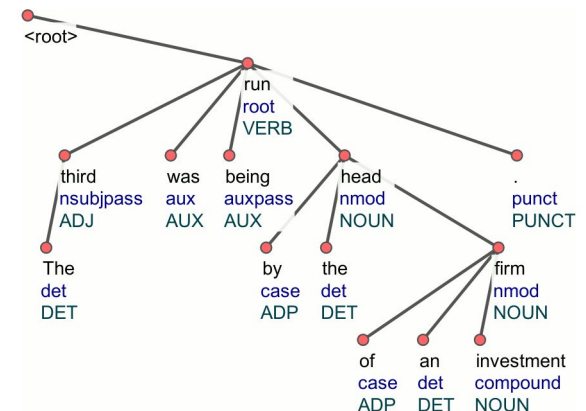
Istituto di Linguistica Computazionale - ILC CNR

08/06/2021

*Scienza aperta e gestione dei dati per le scienze umane e del patrimonio culturale: corso di formazione a distanza*



- CLARIN ERIC - di cosa si tratta?
- Cosa offre CLARIN-IT?
- Alcuni esercizi per mostrarvi le varie possibilità e i servizi che offre CLARIN
  - link in chat al documento
  - <https://docs.google.com/document/d/1XZoacpC1wd0YISKIJSLIFKTYNxo6AmYG>



# CLARIN in due parole



- detiene lo status di **ESFRI ERIC** dal 2012 e **Landmark** dal 2016
- attraverso modalità intuitiva e sostenibile, fornisce agli studiosi delle **scienze umane** e **scienze sociali** accesso a:
  - dati linguistici digitalizzati (in formato scritto, parlato, video o multimodale)
  - **strumenti avanzati** per scoprire, esplorare, annotare, analizzare o confrontare/combinare tali dati a prescindere di dove si trovino
  - attraverso un sistema di controllo di accesso **single sign-on**
- costituisce un ecosistema ideale per la condivisione di conoscenze
- è parte integrante dell' **European Open Science Cloud**
  - See [clarin.eu/eosc](http://clarin.eu/eosc)

# CLARIN: dati e comunità



- Newspaper archives
- Literary texts
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- ...

For the CLARIN Resource Families initiative, see:  
<https://www.clarin.eu/resource-families>

- Digital humanities
- Linguistics and Philology
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public
- ...

# CLARIN oggi



- **21 membri:** (AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI)
- **3 observers:** FR, UK, ZA
- **> 68 centri (25 centri certificati CTS)**



*accesso federato*



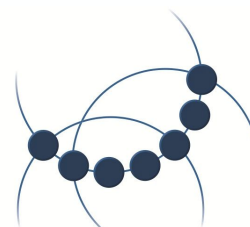
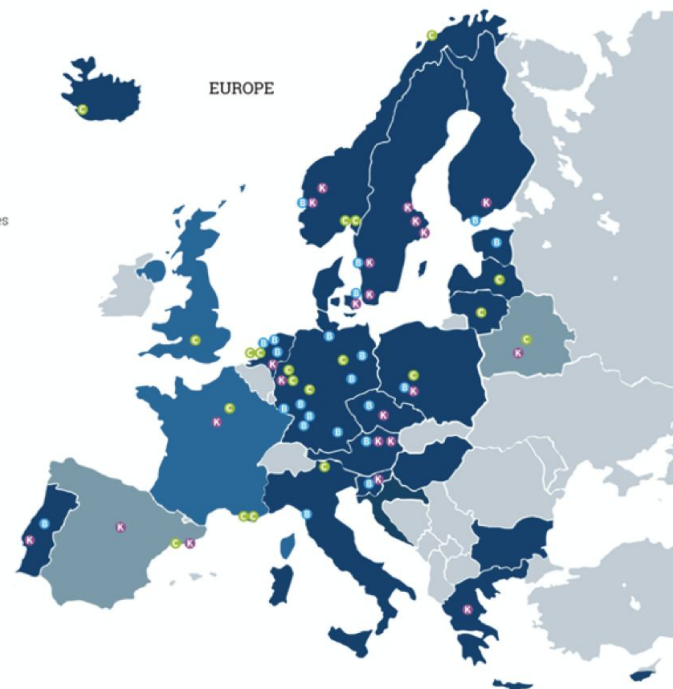
*raccolta centralizzata di metadati*



*servizi concatenati*



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



**CLARIN**

Common Language Resources and  
Technology Infrastructure



# CLARIN-UK

Infrastructure for Digital Language Resources and Tools

[Home](#)[Centres](#)[Resources](#)[People](#)[News](#)[Events](#)[Courses](#)[Blog](#)[About](#)

[Read more about CLARIN-UK](#)

## Data & Tools

[More](#)

CorCenCC

Corpws  
Cenedlaethol  
Cymraeg  
Cyfoes – the  
National Corpus  
of  
Contemporary  
Welsh

## Latest News

[More](#)

New  
members of  
the  
CLARIN\_UK  
consortium

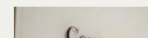
18 November  
2020  
Welcome  
aboard!

## Events

[More](#)

Lancaster  
Symposium  
on Innovation  
in Corpus  
Linguistics  
2021

Wednesday 23  
June, Online



Corpus

<https://www.clarin.ac.uk/>





# CLARIN-UK

Infrastructure for Digital Language Resources and Tools

[Home](#)[Centres ▾](#)[Resources ▾](#)[People](#)[News](#)[Events](#)[Courses](#)[Blog](#)[About](#)[#LancsBox](#)[BNC](#)[CKLD](#)[CLAWS](#)[CLiC](#)[CorCenCC](#)[CQPweb](#)[ELAR](#)[GATE](#)[GATE Cloud](#)[Hansard at  
Huddersfield](#)[Read more about CLARIN-UK](#)

## Data & Tools

[More](#)

CorCenCC

Corpws  
Cenedlaethol  
Cymraeg  
Cyfoes – the  
National Corpus  
of  
Contemporary  
Welsh

[More](#)

## Events

[More](#)

Lancaster  
Symposium  
on Innovation  
in Corpus  
Linguistics  
2021

Wednesday 23  
June, Online



Corpus  
Linguistics

<https://www.clarin.ac.uk/>



# CLARIN Resource families



## Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

## Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

## Tools

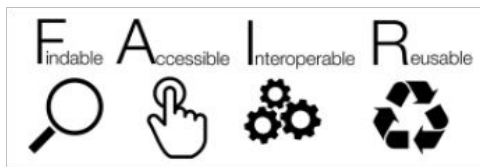
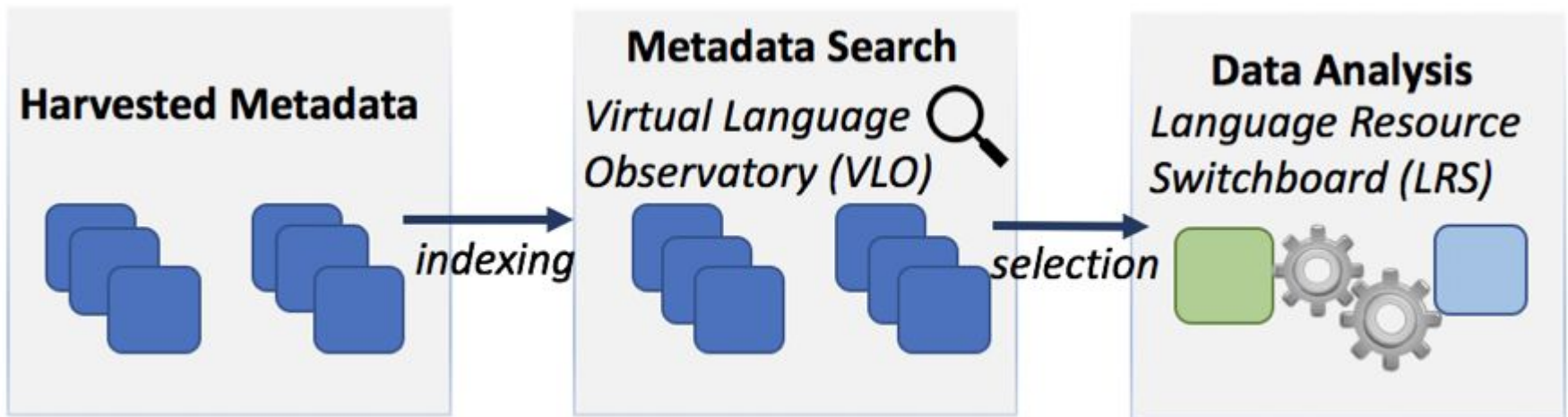
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

## Spoken corpora in the CLARIN infrastructure

### Corpora with transcriptions and audio recordings

Corpus	Language	Description	Availability
<a href="#">Arabic Speech Corpus</a> <b>Licence:</b> CC BY 4.0	Arabic	The corpus is available for download from a dedicated webpage.  For a relevant publication, see <a href="#">Halabi (2016)</a> .	<a href="#">Download</a>
<a href="#">DIALEKT v1: dialectal corpus with multi-tier transcription</a> <b>Size:</b> 100,000 words  <b>Annotation:</b> orthographically and phonetically (dialect features) transcribed, MSD-tagged, lemmatised  <b>Licence:</b> Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains traditional dialectological material, mostly unprepared monologue-type speech.  The corpus is available download (upon request) and through the concordancer KonText.  For a related publication, see <a href="#">Komrsková et al. (2018)</a> .	<a href="#">Concordancer</a> <a href="#">Download</a>

# L'infrastruttura tecnica



[clarin.eu/fair](http://clarin.eu/fair)



[vlo.clarin.eu](http://vlo.clarin.eu)



[switchboard.clarin.eu](http://switchboard.clarin.eu)