

**CLARIN**



# Introducción a CLARIN-ERIC

infraestructura de investigación digital para apoyar  
la investigación basada en recursos lingüísticos:  
datos, herramientas y servicios

[www.clarin.eu](http://www.clarin.eu)  
[www.clariah.es](http://www.clariah.es)

**Mikel Iruskieta**  
HiTZ - UPV/EHU



INSTITUTO DA LINGUA GALEGA



**HiTZ**

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

# CLARIN en dos palabras

- Infraestructura común de recursos del lenguaje y tecnologías.  
*Common **L**anguage **R**esources and **T**echnology **I**nfrastructure*
- **ESFRI** roadmap 2006, **ESFRI** ERIC status 2012, Landmark 2016
- Acceso fácil y sostenible para la investigación de CCSS y Hum.
  - Datos digitales de lenguaje (texto, video o multimodal)
  - Herramientas para buscar, analizar y combinar datos allá dondequiera que estén
  - Acceso federado (con inicio de sesión único)
- Ecosistema para el intercambio de conocimiento
- Con servicios integrados en EOSC

# Principios FAIR

Findable  
Encontrable

Accesible

Interoperable

Reutilizable

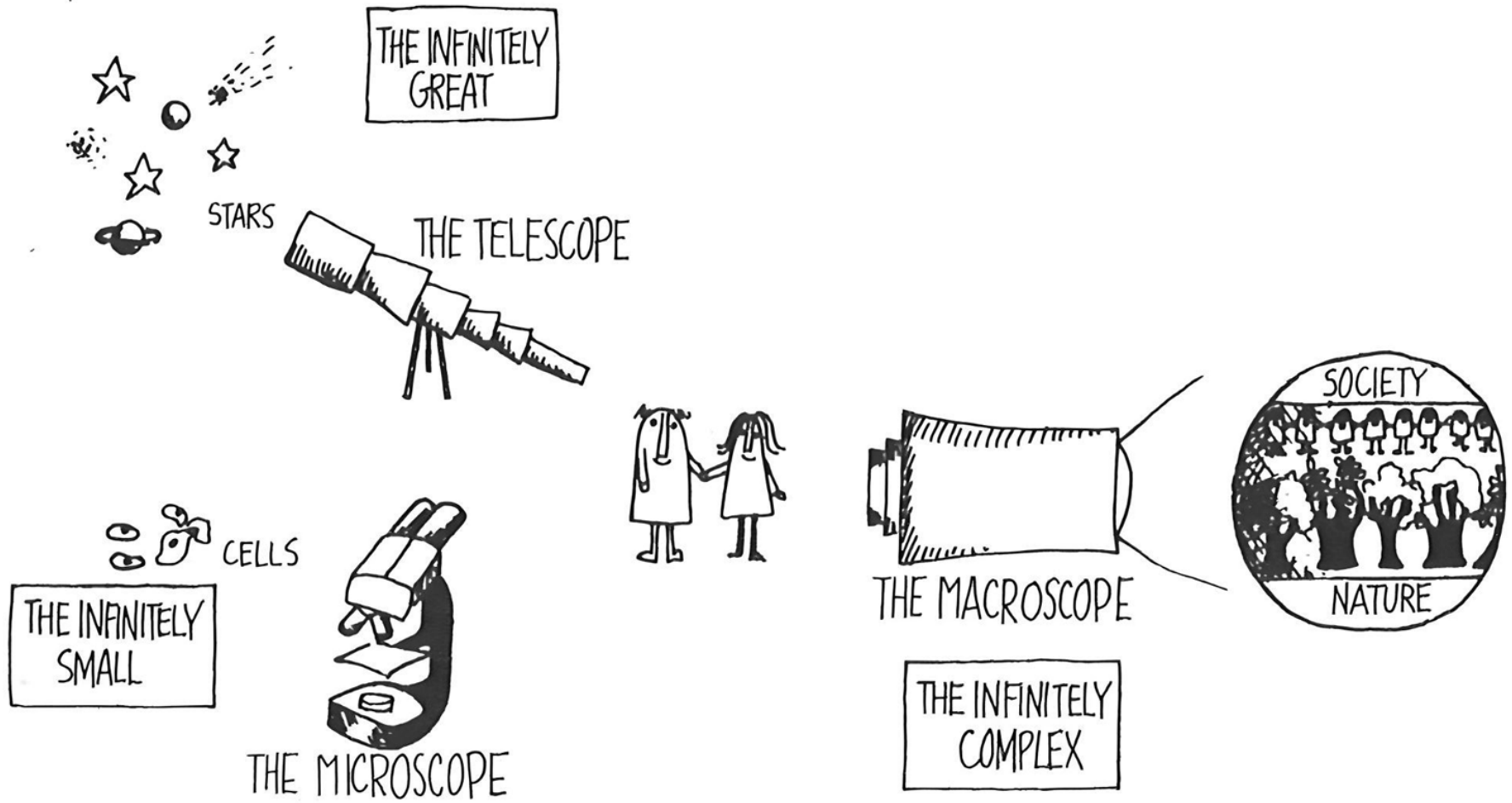
## Elementos clave

- Identificadores persistentes (PIDs)
- Plan de gestión de datos
- Metadatos
- Licencias
- Repositorios

# CLARIN y la ciencia abierta

- Promover el intercambio y la reutilización de datos lingüísticos a través de **registros de datos** sostenibles
- Mejorar e implementar la **interoperabilidad** de datos y servicios lingüísticos
  - Marco común de metadatos
  - Red distribuida de repositorios de datos lingüísticos certificados por FAIR
- Promover
  - métodos comparables
  - colaboración multidisciplinar
  - investigación transnacional
  - **ciencia de datos responsable**
- Apoyar la diversidad
  - lingüística de datos que cubren todos los **idiomas europeos** (y más)
  - herramientas multilingües
  - recursos lingüísticos multimodales e interdisciplinar
  - herramientas independientes (de disciplina e idioma)

# El potencial del macroscopio de CLARIN

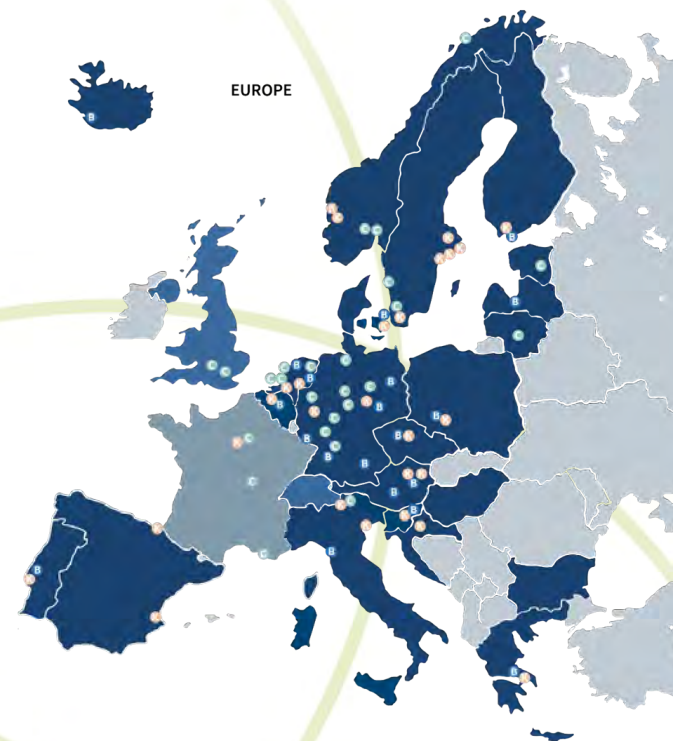


# Miembros y centros de CLARIN

- Tipo de consorcio: ERIC
  - 24 miembros, 2 observadores, 1 participante adjunto
- Red de 70 centros distribuidos
  - 21 centros de catos certificados CTS
  - Enfocado en principios FAIR e interoperabilidad
    - Acceso federado
    - Recolección central de metadatos para facilitar búsquedas
    - Servicios en cadena
  - 25 Centro de conocimiento (CLARIN K-centres)

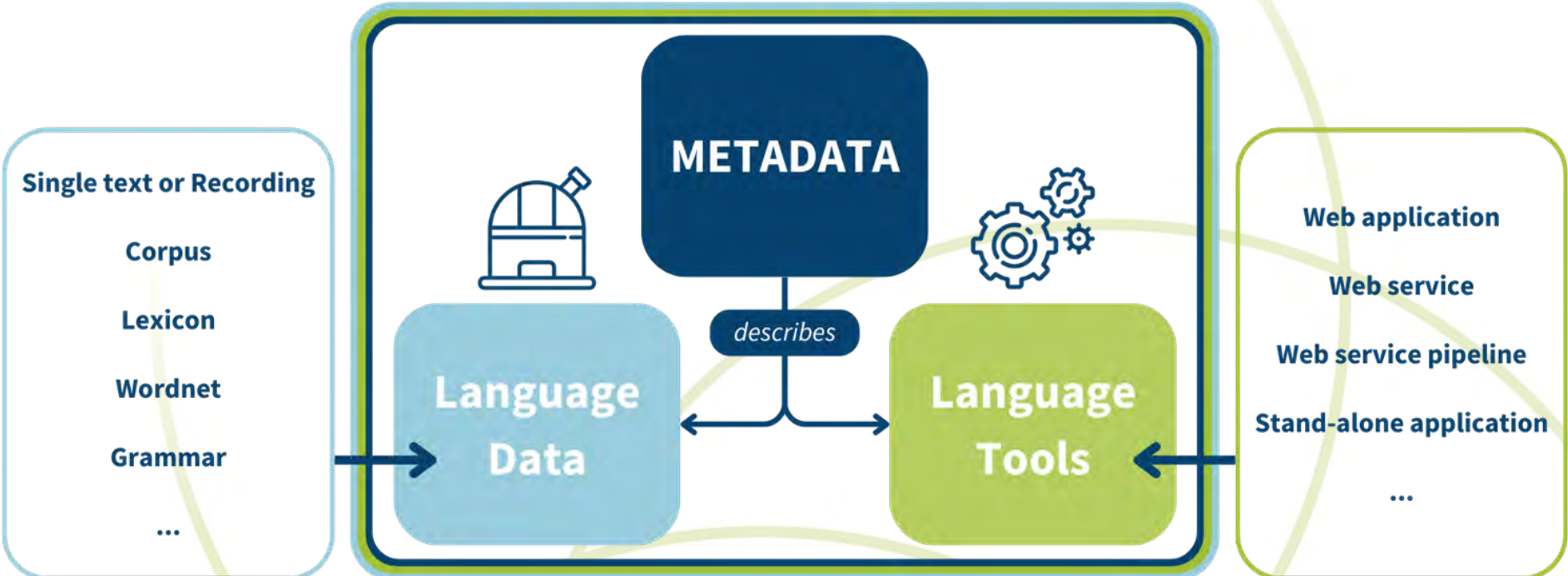


- ERIC members
- Observers
- Countries with participating centres
- Centre Providing Data
- Centre Providing Metadata
- Knowledge Centre



# Funcionamiento de CLARIN

## Repository at a CLARIN Centre





# Observatorio virtual del lenguaje (VLO)

<https://vlo.clarin.eu>

- Búsquedas complejas
- Enlaces a URLs de destino
- Opciones de descarga
- Información sobre licencias
- Características técnicas
- Descripción de las herramientas
- Información sobre cómo citar:

Showing 1 to 10 of 217 results within selection for: corpus Galician

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

Language

Type to filter or search for more

- Galician ✕
- English (54949)
- German (30577)
- French (12858)
- Unspecified (12501)
- Spanish; Castilian (8029)
- Dutch (5923)
- Chinese (5246)
- Japanese (3344)
- Indonesian (2007)

Corpus CLUVI  
(Part of LRT + Open Submissions Data & Tools)

Parallel corpus, 22 million words

Basque Catalan; Val... English French Galician ... (+3)

Landing page for this record

Corpus Técnico do Galego  
(Part of LRT + Open Submissions Data & Tools)

Domain-specific corpus (Law, Computing, Medicine, Economy, Sociology), 10 million words

Galician

Landing page for this record

## Corpus Técnico do Galego

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

BIBTEX CMDI

TALG Research Group (University of Vigo), 2014, *Corpus Técnico do Galego*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11372/LRT-615>.



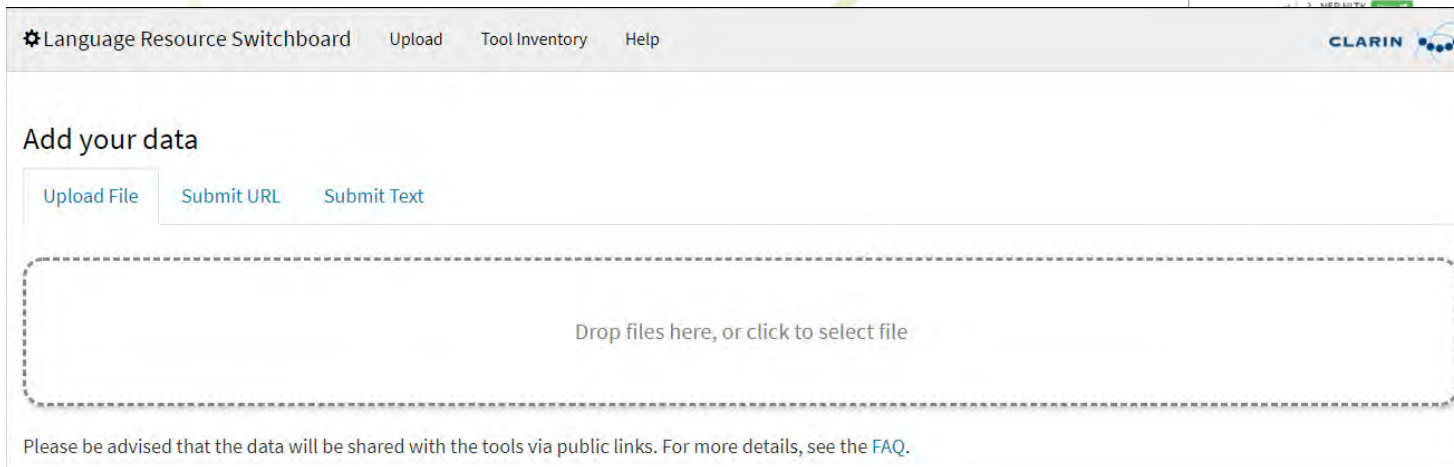
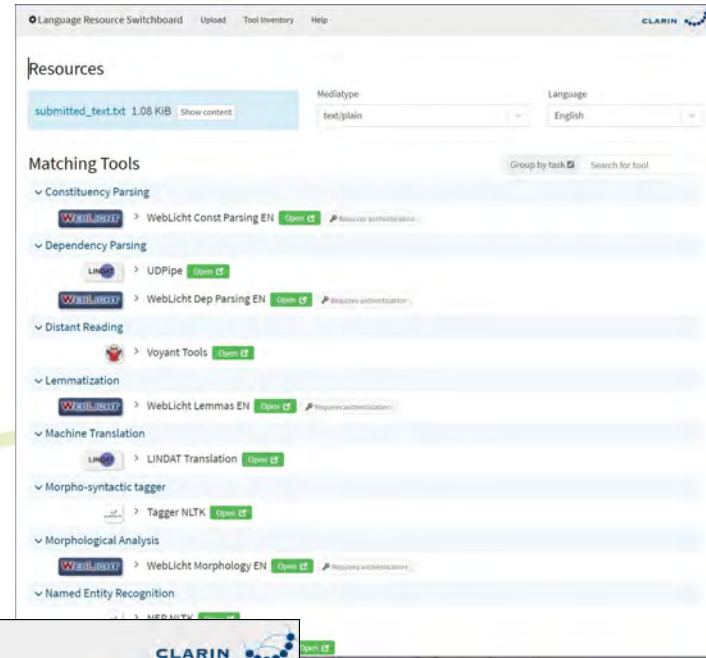




# Language Resource Switchboard

<https://switchboard.clarin.eu/>

- Prueba con un texto para buscar la herramienta más adecuada para ver que tareas de PNL se pueden realizar con ese archivo (formato/lengua)
- Se puede acceder directamente desde la VLO



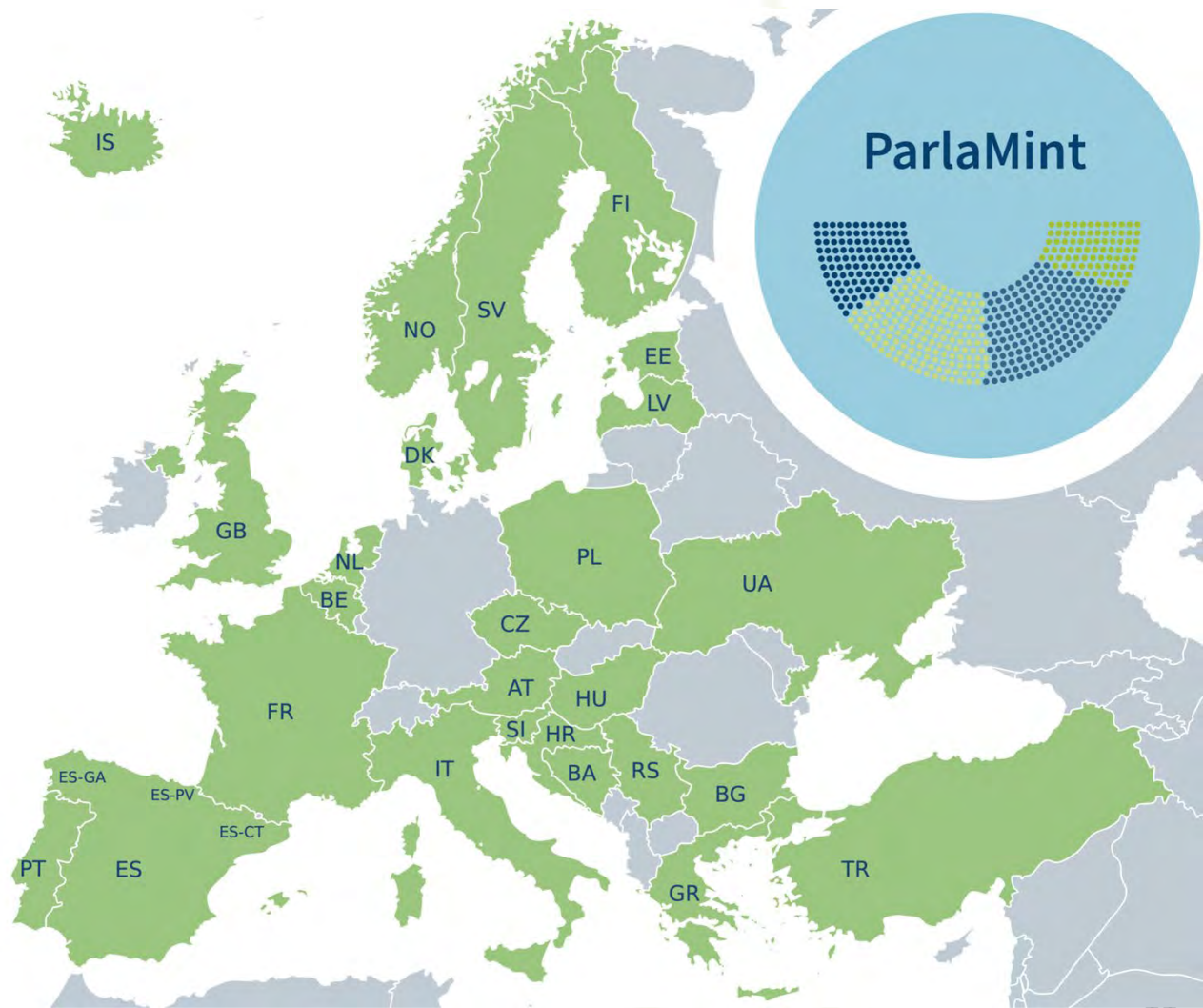
# Recursos CLARIN: *Resource Families*

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



# Proyecto emblematico de CLARIN

<https://www.clarin.eu/parlamint>



# ParlaMint

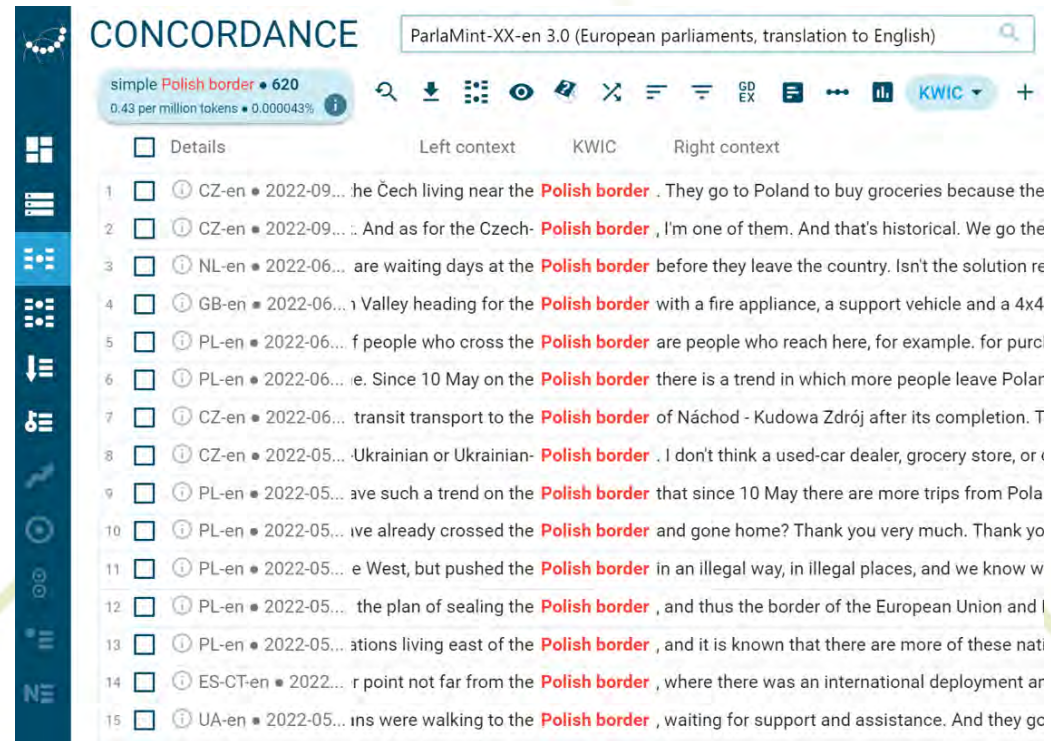
- Actas parlamentarias de 29 parlamentos europeos,
- Subcorpus:
  - Referencia: hasta el 30.01.2020
  - COVID: a partir del 31.01.2020
  - Guerra: a partir del 24.02.2022
- Corpus anotado lingüísticamente, nombres propios, metadatos
- Ejemplo de colaboración en CLARIN e interoperabilidad

# ParlaMint: interoperable y comparable

- Corpus **interoperable** ya que están:
  - Anotados con el mismo esquema TEI hecho *ad hoc*
- Corpus **comparable** ya que contienen:
  - El mismo período: 2015-2022
  - Los mismos metadatos (ponentes, genero, partidos, sesiones...)
  - El mismo paradigma de anotación lingüística
- Aumenta la interoperabilidad y la comparabilidad con la **traducción automática** al inglés

# ParlaMint-MT

- Producido con modelos OPUS-MT
- Posprocesamiento para corregir los nombres propios



The screenshot displays the ParlaMint Concordance tool interface. At the top, the title "CONCORDANCE" is visible next to a search bar containing "ParlaMint-XX-en 3.0 (European parliaments, translation to English)". Below the search bar, a status bar indicates the search term "simple Polish border" with 620 results and a frequency of 0.43 per million tokens (0.000043%). A toolbar with various icons for navigation and analysis is located below the status bar. The main area shows a list of 15 concordance entries, each with a checkbox, a language pair (e.g., CZ-en, NL-en, GB-en, PL-en, ES-CT-en, UA-en), a date, and a snippet of text where the search term "Polish border" is highlighted in red. The interface also includes options for "Details", "Left context", "KWIC", and "Right context".

# Para saber más



- CLARIAH-ES ([enlace](#))
- Noticias > Newsflash ([enlace](#))
- Tour de CLARIN ([enlace](#))
  - Spanish CLARIN K-centre ([enlace](#))
- Investigaciones de impacto > Impact Stories ([enlace](#))
  - Entrevista a un estudiante ([enlace](#))
- Centro de aprendizaje CLARIN > Learning Hub ([enlace](#))



# Referencias de interés

- Iruskieta, M., Estarrona, A., Farwell, A., & Rigau, G. (2022). INTELE: promoviendo la participación en las infraestructuras ERIC CLARIN y DARIAH. *Boletín de la ANABAD*, 72(2), 63-91.
- Bel, N. Gonzalez-Blanco, E. Iruskieta, M. (2016). [CLARIN Centro-K-español](#). *Procesamiento del Lenguaje Natural* 57: 151-154. ISSN: 1135-5948.
- Krauwer, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 1525-1531). European Language Resources Association (ELRA).
- Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- CLARIAH-ES: <http://www.clariah.es/>
- CLARIN: <https://www.clarin.eu/>