

# HUMANIDADES DIXITAIS:

recursos  
ferramentas  
e servizos

15-18 de xullo  
2024  
Aula C01  
Facultade de Filoloxía  
USC



CLARIN



# CLARIN-ERIC: Conocer y participar en la infraestructura Europea del Lenguaje

Mikel Iruskieta  
HiTZ - UPV/EHU

<https://orcid.org/0000-0002-6121-3902>

# 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

# Recomendaciones de la UNESCO para la ciencia abierta: valores y principios

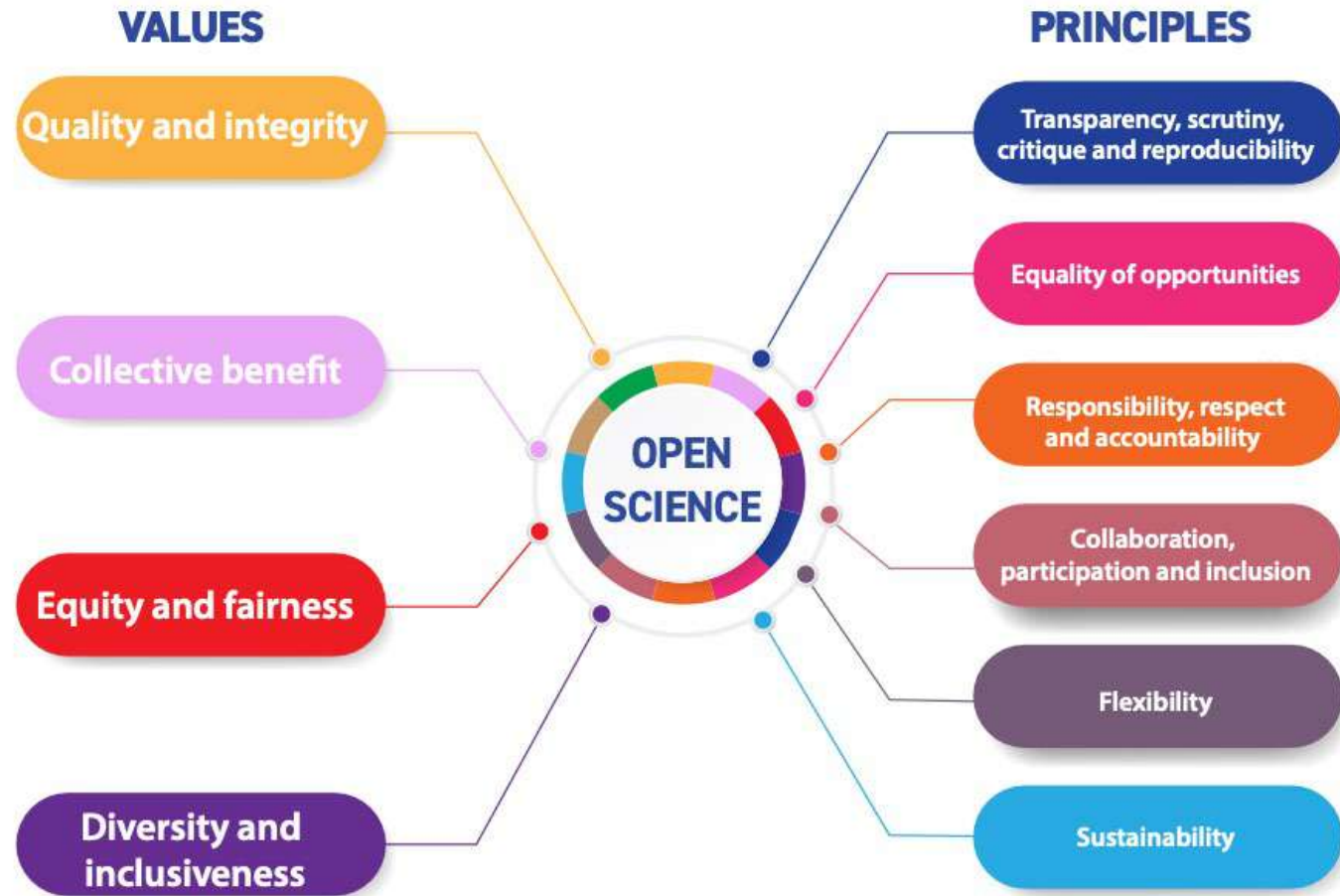


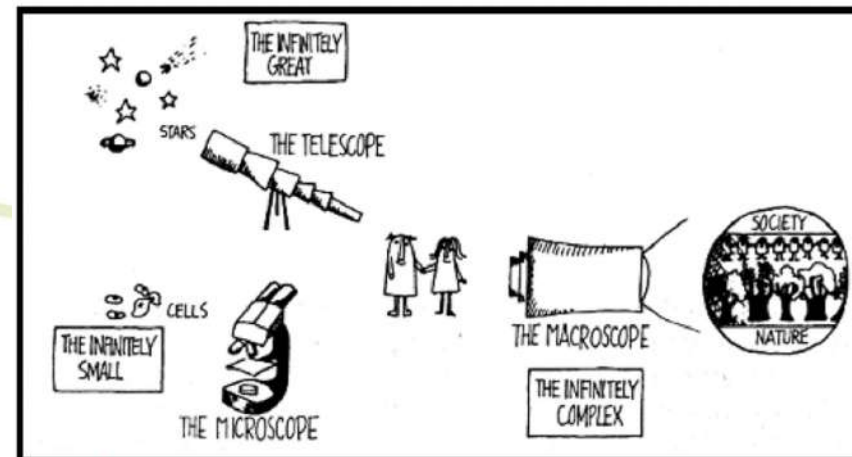
Image Source: [Wikimedia](#) Licence: [CC BY-SA 3.0](#)

# 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)

# Texto y voz como lupas para observar las CCSS

- La variación lingüística y el multilingüismo proporcionan potencialmente la base para la investigación comparada de fenómenos sociales y culturales que se reflejan en el uso del idioma
- Datos de texto y voz como datos sociales y culturales
- Colaboración entre infraestructuras es el **motor** para iniciativas multidisciplinares
- Algunos ejemplos:
  - Discurso parlamentario
  - Patrones de migración
  - Historia intelectual
  - Variación del idioma entre períodos y regiones
  - Dinámica en condiciones de salud mental



Fonte: (ROSNAY, 1979)

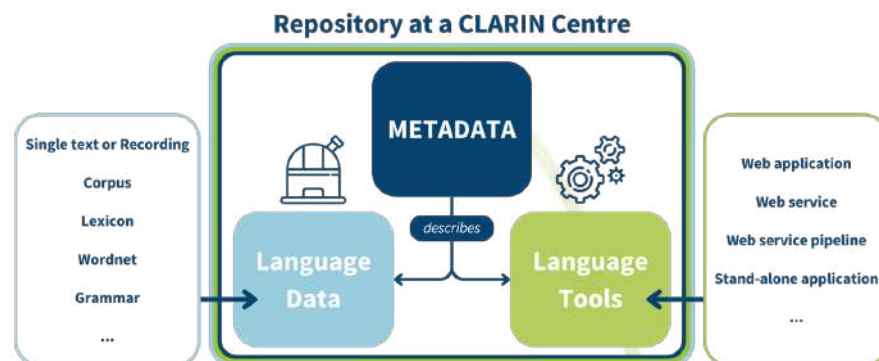
# La complejidad de las CCSS afecta la interoperabilidad

## Variedad de datos

- Lenguas y variaciones
- Modalidades (números, voz, texto, video...)
- Capas (contexto, datos, metadatos...)

## Agendas de investigación conjuntas

- Colaboración inter-estatal, -regional, -período
- Entornos multidisciplinares



## Ejemplos creados de manera flexible en respuesta a desafíos o crisis sociales



### Datos parlamentarios

- 29 parlamentos europeos
- Anotación: NER, sentimientos, etc.
- Temas
  - COVID: 2020-01-31 en adelante
  - WAR: 2022-02-24 en adelante

# El cluster abierto CCSS y la ciencia abierta

**SSHOC** es uno de los 5 clusters (H2020) creando el inventario de servicios de la EOSC



[www.science-clusters.eu](http://www.science-clusters.eu)

Intercambio y reutilización de datos y recursos a través de **servicios federados** y **centros certificados**

## Principios FAIR

Mejora y desarrollo de la **interoperabilidad** de datos y servicios

- Metadatos conjuntos
- **SSH Open Marketplace** como plataforma de descubrimiento

## Apoyando la **diversidad**

- Recursos de las comunidades europeas permiten la investigación transnacional
- Datos y herramientas en muchos idiomas
- Recursos en muchas modalidades y tipos de datos
- Métodos comparados y en colaboración multidisciplinar

## Ciencia abierta más allá de FAIR

- Ética de la ciencia de datos
- **Educando** las próximas generaciones

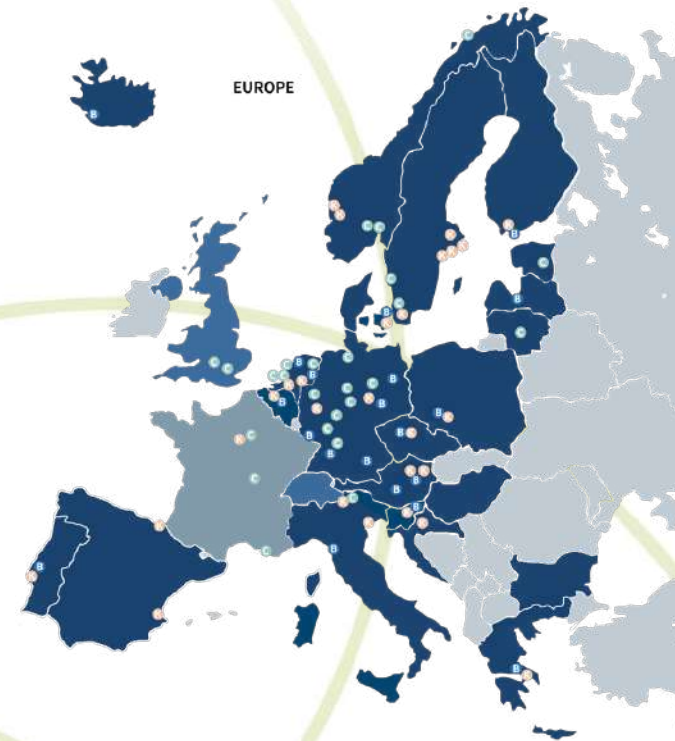


# Miembros y centros de CLARIN

- 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



# Estrategia de CLARIN 2024-2026

## Área 1: Uso académico

## Área 2: Uso no-académico

## Área 3: Calidad e interoperabilidad de los datos, herramientas y metadatos

## Área 4: Accesibilidad y usabilidad de los datos, herramientas y servicios

## Área 5: Colaboración e intercambio de ideas

- Obj1: Aumentar la repercusión de CLARIN
- Obj2: Ampliar el número de usuarios
- Obj 3: Desarrollar compromiso con comunidades no-académicas para impulsar innovaciones técnicas o sociales
- Obj 4: Mejorar la calidad de los metadatos y aumentar la capacidad de descubrimiento de recursos y herramientas.
- Obj 5: Mejorar la cobertura y calidad de los datos y herramientas
- Obj 6: Mejorar la interoperabilidad de metadatos, datos y herramientas
- Obj 8: Alcanzar un alto nivel de accesibilidad y usabilidad de los servicios técnicos centrales
- Obj 9: Consolidar la comunidad CLARIN para la colaboración y el intercambio de ideas
- Obj 10: Mejorar la colaboración entre “fronteras”: entre países, entre tipos de entidades y entre campos de especialización

# Datos con características propias en la investigación de las SSH

Según las **escuelas de pensamiento** de la Ciencia Abierta ([Fecher & Friesike, 2014](#)), las infraestructuras se organizan de diferente forma

- DARIAH > Escuela **pragmática**: investigación colaborativa
- CLARIN > Escuela **infraestructural**: arquitectura tecnológica, eficiencia de las herramientas y datos complejos
- Escuela de **medición**: impacto científico
- Escuela **democrática** y pública: acceso

Puede que dicha complejidad y desarrollo en el conocimiento nos lleve a organizarnos de una u otra forma.



# 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 (2002)

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

Parallel corpus, 22 million words

Besque 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, Sociol... 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>.





# Ejercicio VLO: Búsqueda de datos

<https://vlo.clarin.eu>

0



- Búsqueda:
- Euskara
- Texto
- HTML >> TXT

The screenshot shows the VLO search results page for the record 'Frogstory basque 10b 2011'. The page includes a breadcrumb trail 'VLO / Faceted search / Search results / Record: Frogstory basque 10b 2011' and indicates it is 'Record 11 of 68'. The main title is 'Frogstory basque 10b 2011'. Below the title are tabs for 'Record details', 'Links (4)', 'Availability', 'All metadata', and 'Technical Details'. A table lists the links:

Name
<a href="#">cocoon-79f85f0d-1426-32ad-9eb2-7167caf8310f</a>
<a href="#">DOI cocoon-79f85f0d-1426-32ad-9eb2-7167caf8310f</a>
<a href="#">1.17-510233</a>
<a href="#">10b_2011.xml</a>

At the bottom, there is an 'About' section with the version 'v4.11.4'.

On the right, the 'Switchboard' interface is visible, showing the selected record '10b\_2011.xml' (5.68 KiB) with a 'Show content' button. It includes dropdown menus for 'Mediatype' (text/rtf) and 'Language' (Basque). Under 'Matching Tools', there are sections for 'Distant Reading' (with 'Voyant Tools' and 'Open' button), 'Text Analytics' (with 'Text Tonsorium - Advanced mode' and 'Open' button), and 'WebLicht Advanced Mode' (with 'Open' button and 'Requires authentication' note).

# No todo es FAIR

47 bilingües y 39 multilingües

- El 38% no se buscan en VLO
- El 6% no se puede descargar o realizar búsquedas
- El 12% no se encuentra disponible
- El 13% no se sabe el tamaño
- El 31% no tiene información del nivel de alineación
- Al 7% le falta info de licencia
- 11 corpus no se encuentran en CLARIN

- Findable
- Accessible
- Interoperable
- Re-usable

(Fišer and Lenardič 2020)

178 Open ✓ 234 Closed Author Label Projects

- 🔍 Turkish sign language database CLARIAH-NL sign language resources  
#405 opened on Oct 20, 2022 by jakoble 2 tasks
- 🔍 "Exhibition Corpus" - Text, Sound, Sign CLARINO sign language resources  
#404 opened on Oct 20, 2022 by jakoble 1 task
- 🔍 DGS CORPUS CLARIN-D sign language resources  
#403 opened on Oct 20, 2022 by jakoble 1 task
- 🔍 Italian Sign Language Corpus CLARIN-D sign language resources  
#402 opened on Oct 20, 2022 by jakoble 1 task
- 🔍 MOCAP1 Huma-num  
#401 opened on Oct 20, 2022 by jakoble 1 task
- 🔍 Translations of the Bible and of the Church Manual into Finnish Sign Language #400  
clarin-eric/resource-families... on Oct 20, 2022  
<http://urn.fi/urn:nbn:fi:lb-2014073029> Missing size Unclear annotation  
FIN-CLARIN sign language resources
- 🔍 Translations of the Bible and of the Church Manual into Finnish Sign Language FIN-CLARIN sign language resources  
#400 opened on Oct 20, 2022 by jakoble 2 tasks
- 🔍 The Kipo Corpus FIN-CLARIN sign language resources  
#399 opened on Oct 20, 2022 by jakoble 1 task



# 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

The screenshot shows the 'Resources' page of the Language Resource Switchboard. At the top, there is a navigation bar with 'Language Resource Switchboard', 'Upload', 'Tool Inventory', and 'Help'. Below this, the 'Resources' section displays a file named 'submitted\_text.txt' (1.08 KIB) with a 'Show content' button. There are dropdown menus for 'Mediatype' (set to 'text/plain') and 'Language' (set to 'English'). The main area is titled 'Matching Tools' and lists various NLP tools categorized by task, such as Constituency Parsing, Dependency Parsing, Distant Reading, Lemmatization, Machine Translation, Morpho-syntactic tagger, Morphological Analysis, and Named Entity Recognition. Each tool entry includes a logo, the tool name, and a status indicator (e.g., 'Open' or 'Requires authentication').

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

# Recursos CLARIN: *Resource Families*

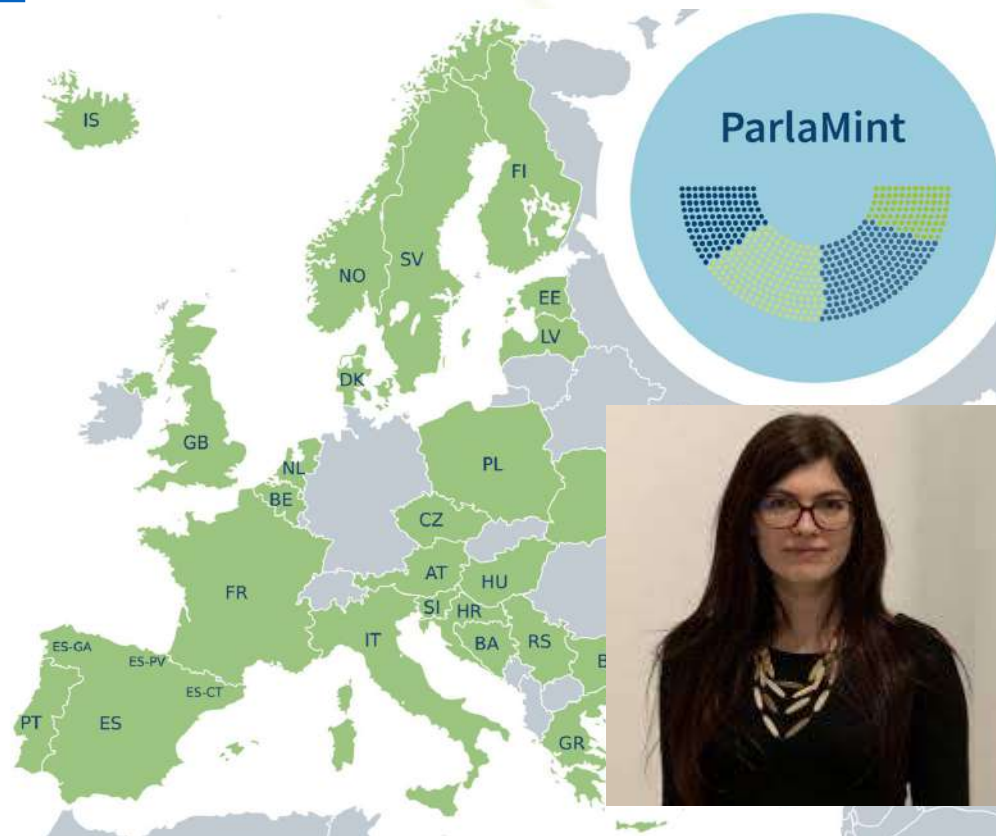
<https://www.clarin.eu/resource-families>





# Proyecto emblemático de CLARIN

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



**Adina Vladu** (Universidade de Santiago de Compostela)

*A explotación dun corpus para Humanidades e Ciencias Sociais: ParlaMint*   \*

15:00 - 17:30

# ParlaMint: interoperable y comparable

Proyecto interno y ejemplo de colaboración en CLARIN e interoperabilidad

- 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 **interoperable**, ya que están anotados con el mismo esquema TEI
- Corpus **comparable**, ya que contienen, similares periodos, metadatos, paradigmas de anotación lingüística
- Aumenta la interoperabilidad y la comparabilidad con la **traducción automática** al inglés

LENG	PALABRA CLAVE	Normaq	Relative density		Relative density	
		freq X M	F	M	Coalition	Oppo.
ES	<a href="#">día ? de la mujer</a>	1,77	168	72	161	79
GAL	Día Internacional da Muller	0,22	3	1	2	2

# Colab y ParlaMint:

<https://colab.research.google.com/drive/1XESKAMjMaa9hpLqfgaEirKKzoNSCFUVj?usp=sharing>



Permite programar y ejecutar Python en el navegador:

- Sin configuración
- Acceso a GPUs
- Compartir contenido

# Materiales para el uso práctico del corpus

<https://sidih.github.io/voices/toc.html>

## Voices of the Parliament A Corpus Approach to Parliamentary Discourse Research

»Prvič, sem političarka in  
ne politik, drugič pa ...«

Korpusni pristop  
k raziskovanju  
parlamentarnega  
diskurza

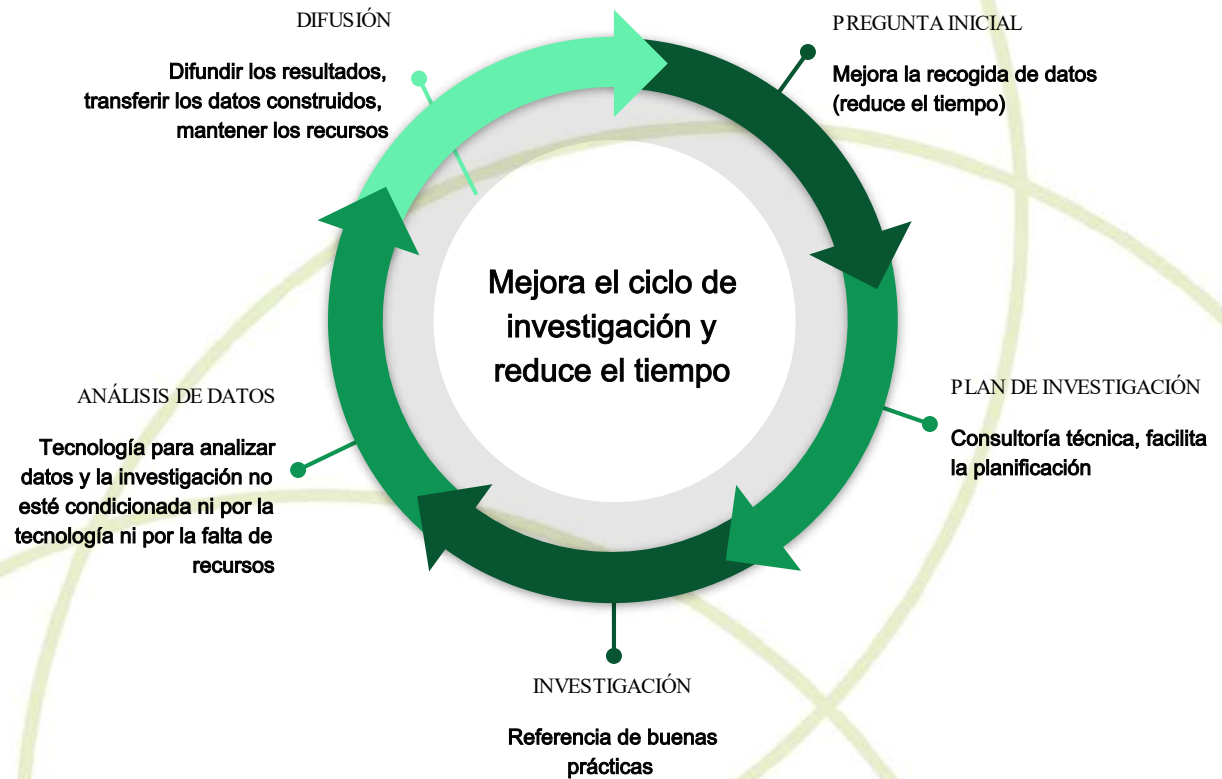




# Tareas

# Tareas para un curso de verano

- Identificar las necesidades de aprendizaje
  - Habilidades y competencias para investigadoras de SSH
- Resultados de aprendizaje



# CLARIN: experiencias de usuarios

<https://zenodo.org/record/4288980#.X9YDS7N7mDI>

## CLARIN through the eyes of the researchers

Tour de CLARIN 

Volume III



# Registrarse en CLARIN para acceder a toda la infraestructura

En algunos casos es necesario para acceder a datos, herramientas y servicios que están protegidos

## CLARIN account registration

Thank you for your interest in CLARIN. Please complete the form below.

After your registration is processed (normally within two working days), an automated email will be sent to your email address for verification. Click on the included link to verify your email address and activate your account **(if you do not receive this email please make sure to check your spam folder)**. After your account is active you can [download, explore and analyze password-protected language data](#) with the [CLARIN Identity Provider](#).

# Registrarse en EuDat

- Acceder en <https://b2access.eudat.eu>
- Guardar, compartir en <https://b2drop.eudat.eu>
- Publicar en <https://b2share.eudat.eu>





# Herramientas en el *market place* para la extracción de información de la web

- Busca que herramienta necesitas y para que



Activities [Capturing](#) [Gathering](#)

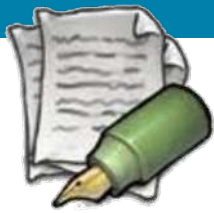
Keywords [data-extraction](#) [web-scraping](#)

import.io is a commercial tool for extracting and structuring web data.

[Read more](#)

# 1


# Manuales de PH



- OpenRefine y Wget

Programming Historian




ABOUT ▾ CONTRIBUTE ▾ LESSONS EVENTS SUPPORT US ▾ BLOG



## Fetching and Parsing Data from the Web with OpenRefine

Evan Peter Williamson

OpenRefine is a powerful tool for exploring, cleaning, and transforming data. In this lesson you will learn how to use Refine to fetch URLs and parse web content.

 Peer-reviewed  CC-BY 4.0  Support PH

Programming Historian

ABOUT ▾ CONTRIBUTE ▾ LESSONS EVENTS SUPPO



## Automated Downloading with Wget

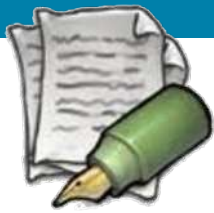
Ian Milligan

Wget is a useful program, run through your computer's command li  
online material.

 Peer-reviewed  CC-BY 4.0  Support PH

# 2

# Web-scraping (import.io)



## Choose your plan

### Trial Plan (14 days)

500 queries

- Point and click training
- Interactive extraction
- E-mail and ticket support
- Capture screenshots
- API access and Webhooks

Free

Selected

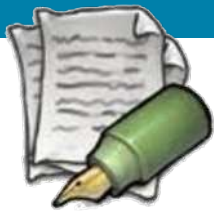
The screenshot shows the import.io dashboard with a dark sidebar on the left containing navigation icons for Home, Portal, Extractors, Reports, Account, and Help. The main content area is divided into several sections:

- Data Health:** 100.0% (Rows Extracted Today: 6, URLs Success/Failed: 2/0).
- Account Summary:** 4 queries used, Remaining 496, Trial Plan (14 days) ending June 05, 2024. Includes an Upgrade button.
- Extractors:** 2 total, 0 running.
- Extractor History Table:**

Extractor	Date/Time	Duration	URLs	Total Rows	Queries
argis.eu	11:06:52 May 23, 2024	00:00:27s	1 input	1 success/0 failed 2 rows extracted	2 queries used
berrie.eu	11:00:34 May 23, 2024	00:00:11s	1 input	1 success/0 failed 4 rows extracted	2 queries used

# 3

## Uso de las infraestructuras: texto escrito



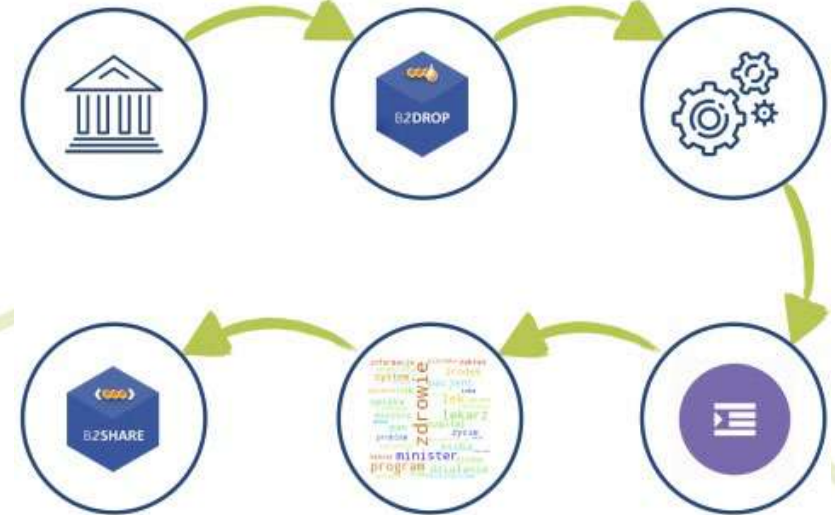
### Texto escrito:

1. Eudat: FAIR data across borders and disciplines
2. Interoperable con el [Switchboard](#) de CLARIN
3. Detección de lengua y propone herramientas/métodos
4. Analiza el texto y ofrece técnicas de visualización

# Investigar en la nube en CLARIN: principios FAIR

Ejemplo de compilación de novelas de varias fuentes en la infraestructura de EuDAT y del repositorio virtual de CLARIN

- Datos
- Análisis sintáctico
- Análisis en la nube
- Publicación persistente



Interoperable



Original en inglés:

<https://www.clarin.eu/showcase/eosc-portal-demonstration>

# Recursos en la nube para texto

1. Descargar texto: [Gitzip](#)
2. Corpus en [Eudat](#)
3. Pasar a [colección virtual](#)
4. Analizar con un clic en [Switchboard](#)
5. Elegir un recurso para el análisis de texto
6. ...



- Recursos (2020)
  - Para el castellano: 4
  - Para el inglés: 16
  - Para el alemán: 13
  - Para el polaco: 26

- Recursos (2023)
  - Para el castellano: 12
  - Para el inglés: 22
  - Para el alemán: 20
  - Para el polaco: 32



1. Constituency Parsing
2. Coreference Resolution
3. Dependency Parsing
4. Distant Reading
5. Extraction of Polish terminology
6. Inclusion detection
7. Keyword Extractor
8. Lemmatization
9. Machine Translation
10. Metadata Processing
11. Morpho-syntactic tagger
12. Morphological Analysis
13. Named Entity Recognition
14. Named Entity Relation Detection
15. Part-Of-Speech Tagging
16. Sentiment Analysis
17. Shallow Parsing
18. Spatial expression detection
19. Speech Recognition
20. Stylometry
21. TF, IDF, TF-IDF calculation
22. Text Analytics
23. Text Enhancement
24. Text Summarization
25. Tokenisation
26. Topic Modelling
27. Visualisation of Geographic Data
28. Word sense disambiguation

# Análisis de un texto con UDPipe



↓ Process Input ↓

Output Text    Show Table    Show Trees

Save Tree as SVG

Previous 1 2 3 4 5 6 7 8 9 10 11 12 ... Next

Esperando a que la incierta luz de la mañanita entre en hilos de claridad por las hendidas de la puerta que da al campo, uno de los gatos del cortijo está perspicaz acecho, con las dos manos estendidas hacia adelante, y la cabeza algo agachada, lo mismo que si se hallara a la vista de algún fugitivo ratón.

Hide empty attributes	x
deprel	nmod
fcats	Cender-FemNumber-Sing
form	puerta
head	17
id	20
lemma	puerta
misc	TokenRange=109..115
upostag	NOUN
xpostag	NOUN

# Ejemplo de análisis de TEI+XML del corpus ELTeC



## Resources

SPA2013\_OrtegaYFrias\_ElDuende.xml 1.12 MiB

Mediatype

application/tei+xml

## Matching Tools

▼ Distant Reading

> [Open](#) Voyant Tools

Voyant Tools

[Cirrus](#) [Terms](#) [Links](#)

[Reader](#) [TermsBerry](#)

[Trends](#) [Document Terms](#)



El duende de la corte : edición ELTeC(Ortega y Frías, Ramón (1825-1883))

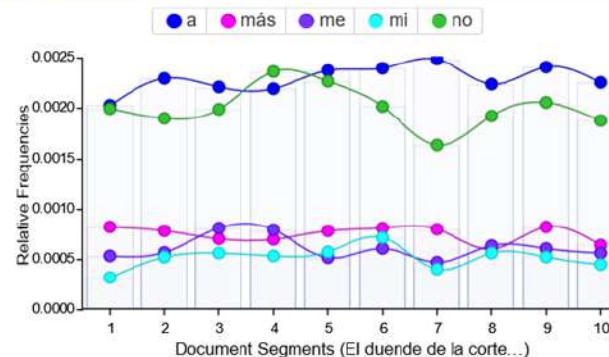
El duende de la corte

o

Memorias de un fraile

Novela histórica original

de



[Summary](#) [Documents](#) [Phrases](#)

[Contexts](#) [Bubblelines](#) [Correlations](#)

This corpus has 1 document with 178,052 total words and 14,073 unique word forms. Created now.

Vocabulary Density: 0.079

Average Words Per Sentence: 15.2

Most frequent words in the corpus: **a** (4049); **no** (3542); **más** (1316); **me** (1076); **mi** (909)

Document	Left	Term	Right
1) El du...	fácilmente sin necesidad de pedirlos	a	la imaginación del poeta, cuyas
1) El du...	vez se arranca una lágrima	a	los ojos, un suspiro al
1) El du...	una historia lo que voy	a	referir. No he tenido que
1) El du...	lectura de algunos párrafos convenció	a	mi amigo de que había
1) El du...	de cedérmelo. [1] Así vino	a	mis manos esta historia, y

4,049 context  expand



# Colección virtual ad hoc

## Colección virtual Edad de Plata

### General

Name: Colección virtual Edad de Plata  
Type: EXTENSIONAL  
Creation date: 2021-10-08  
Description: Presentación UCM  
Purpose: REFERENCE  
Reproducibility: INTENDED  
Keywords:

- UCM
- Edad
- de
- Plata

### Resources

Reference

ELTeC

CONSSA

Audio sobre novelas

T1-unamuno

T2-Fernan-Caballero

ELTeC (<https://github.com/COST-ELTeC/ELTeC-spa>)



CONSSA (<https://github.com/cligs/conssa>)



Audio sobre novelas (<http://contenidosdigitales.uned.es/fez/view/intecca:VideoCMAV-5a6f2b5bb111f57648b4cb3>)



Unamuno (<https://b2drop.eudat.eu/s/dLBDMkDBrs3HFT7>)



Caballero (<https://b2drop.eudat.eu/s/Hz5cP4aip5wobDP>)



Save Collection

Cancel



# Traducir documentos desde las colecciones virtuales

## LINDAT Translation

Translate

Docs

The translation service is available for *personal and non-commercial use* (see [terms of use](#) for more details).

### Source

English

### Target

French

advanced

### Input sentences

THE TWINS;  
A DOMESTIC NOVEL.  
BY  
MARTIN FARQUHAR TUPPER, A.M., F.R.S.  
AUTHOR OF  
PROVERBIAL PHILOSOPHY.  
HARTFORD:  
PUBLISHED BY SILAS ANDRUS & SON  
1851.

[page 13]

### Translation

LES TWINS;  
A DOMESTIC NOVEL.  
PAR  
MARTIN FARQUHAR TUPPER, A.M., F.R.S.  
AUTEUR  
PHILOSOPHIE PROVERBIALE.  
HARTFORD:  
PUBLIÉ par SILAS ANDRUS & SON  
1851.

[page 13]

Translate

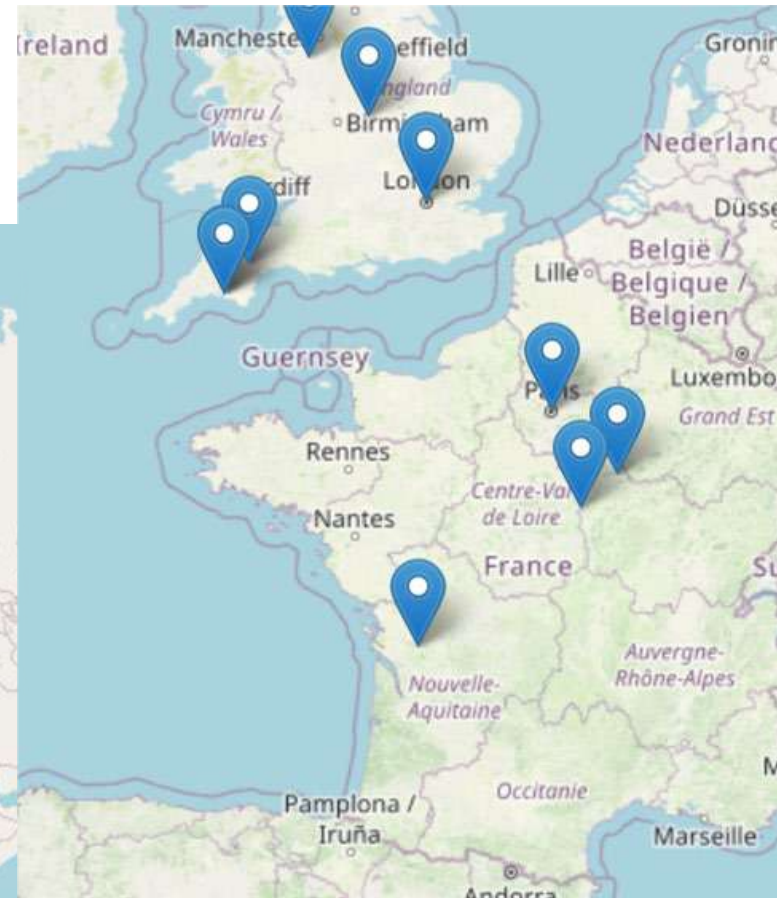
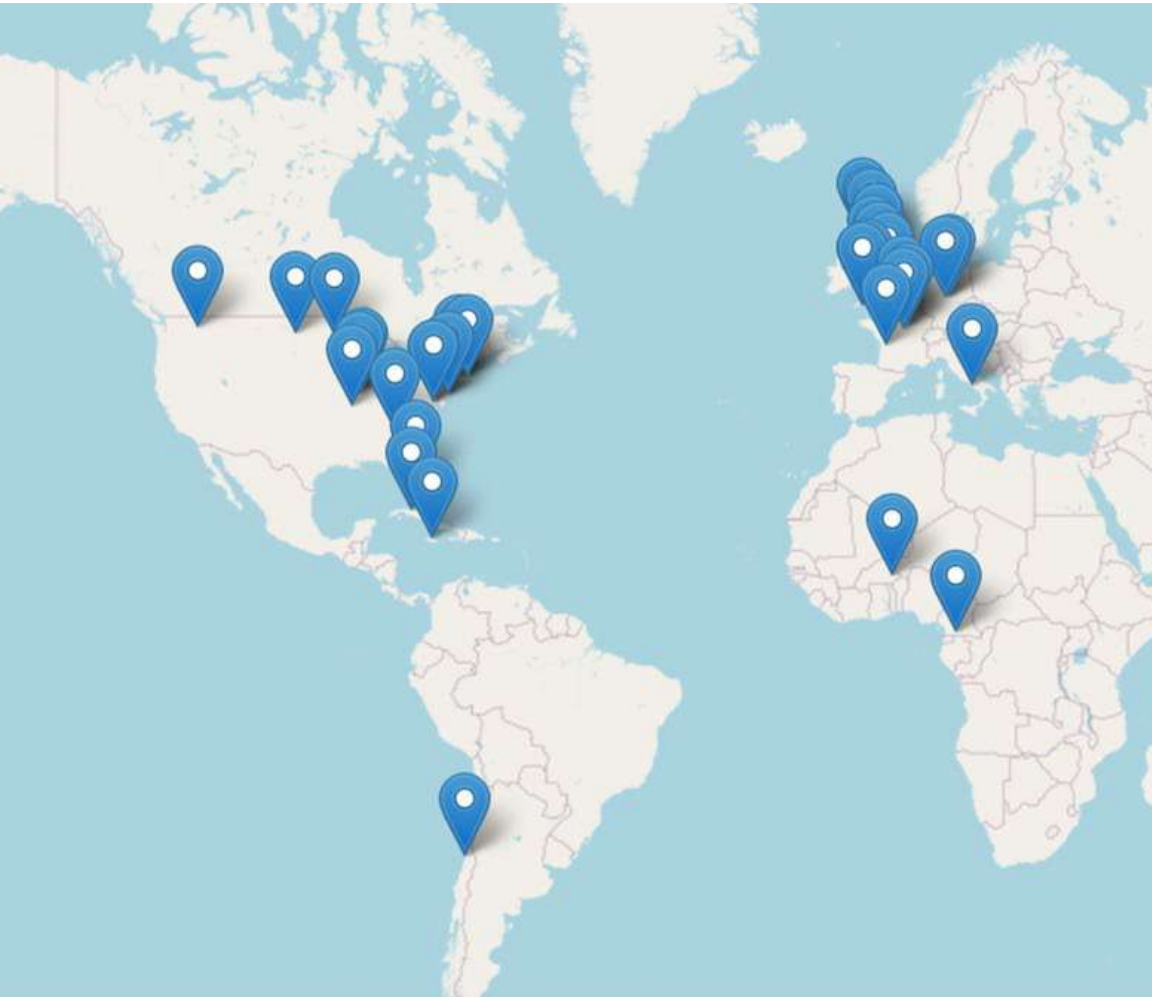
Choose file

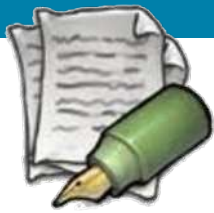
# NER: detección de nombres propios

PERS | LOC | ORG | DEMO | WORK | EVENT | OTHER

blooming in his cheeks; and, as if this was not miracle enough, he had brought his wife as been so superfluous as to purchase a new pair of double-barrelled pistols for Julian : the la as for Emily 's petit cadeau, it was a fifty guinea set of cameos, the choicest in their way t Oxford , to make inquiries after Charles : actually, good fortune had made him at once hu the arms of his wondering wife, as Paris might have flown to Helen , or Leander to his therefore he did not think to ask for Julian ; no doubt the boy was gone to bed. Indeed, he he could upon the feather-bed: he had need of poultices all over, and a quart of Friar's Ba hound had slunk to his kennel, and locked himself sullenly in, without even speaking to hi comforts lent the muddled man their aid. However, after the first rush of news to Mrs. Tra fall into his arms—for strangely did they love each other—suddenly asked, "But, where's —" "Oh! general, I'll tell you all about it to-morrow morning." "About what, madam? Gre what shall I say?" sobbed the silly mother. " Emily—Emily , poor dear Julian —" "What t singular calmness; probably for a dram of brandy. Saunders answered it so instantly, that I headed butler, anticipating all that he might say, she brushed past him, and hurriedly ran u

# NER y geoparsing





# 4

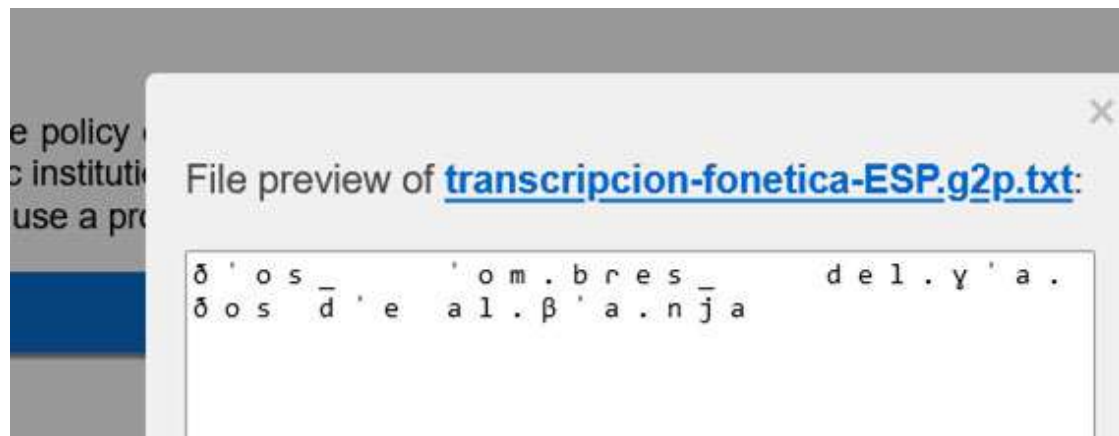
## Uso de las infraestructuras: voz

### De voz a texto:

1. Servicio WEBMaUS de [BAS](#) (CLARIN)
2. Decenas de lenguas y variaciones
3. Múltiples formatos de salida para seguir con la investigación: TXT, SCV, PRAAT, Video...

# Ejemplo de grafema > fonema

1. Subir un TXT
2. Elegir el formato de salida adecuado



9. /dos 'ombres del'gados de al'banja/

[do'sombrezðel'yaðoszðeal'βanja] o [do'sombrezøðel'yaðoşødeal'βanja]

Dos hombres delgados de Albania



Bavarian Archive for Speech Signals



# WebMAUS

## Ejemplo de ASR parlamentario

1. Descargar el [video](#)
2. Transcribir automáticamente en [BAS](#)
3. Elegir el formato de salida adecuado, para seguir investigando



1. Mary TTS
2. ASR
3. TextAlign
4. Pipeline without ASR
5. Pho2Syl
6. Chunker
7. AnnotConv
8. G2P
9. OCTRA - online text transcription system.
10. AudioEnhance
11. WebMAUS General
12. Chunk Preparation
13. Coala
14. WebMINNI
15. WebMAUS Basic
16. Anonymizer
17. TextEnhance
18. Formant Analysis
19. Subtitle
20. EMU Magic
21. Voice Activity Detection
22. EMU webApp - online labeling of speech data and more.
23. Pipeline with ASR
24. SpeakDiar

Show service sidebar >

**BAS Web Services**  
Version 3.7 • History of changes

0/1 individual files and/or file pairs processed (0 %)  
Started at 16:30

Home General Help + FAQs Publications Contact, About, Privacy

### Automatic Speech Recognition (ASR)

Number of speakers:

Speaker label mapping:

Exceed quota code:

Service manual [hide >](#)

Powered by:

Fraunhofer IAIS CLST | Centre for Language and Speech Technology Radboud University IBM

Google Cloud Platform

When selecting 'emuDB' (EMU-SDMS) as output format, the service will pack the resulting EMU-SDMS database into a ZIP file, which can be retrieved by clicking on the "Download as ZIP-File" button.

# ASR Bilingual Basque-Spanish



2020 ira 14

## EUSKO LEGEBILTZARRAREN 40. URTEURRE

*Eusko Legebiltzarrak Euskal Herriko Unibertsitatearen (UPV-EHU) udako il*

Lekua MIRAMAR JAUREGIA

Datak: leh, 26/10/2020 - art, 27/10/2020

Ordua: 10:00 -18:00



urteurrena  
40. ANIBERSARIO  
1980 - 2020

**EUSKO LEGEBILTZARRAREN 40. URTEURRENA:  
ATZERANZKO BEGIRADA**

**40 ANIVERSARIO DEL PARLAMENTO VASCO:  
UNA MIRADA RETROSPECTIVA**



SOURCE:

<https://www.legebiltzarra.eus/portal/eu/web/eusko-legebiltzarra/noticias-y-eventos/actos-y-eventos/-/buscador/content/40-aniversario-del-parlamento-vasco-una-mirada-retrospectiva>

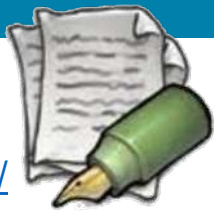
callGoogleASR: egun on guztioi eta ongi etorri abestia eusko legebiltzarrak euskal herriko unibertsitatearen udako ikastaroen baitan antolatu duen 2000 goiko ikastaro honetara eskerrak eman nahi dizkizuet jardunaldi hauetan parte hartu duzuen guztioi hizlari partehartzaile antolatzaileei ere gehiago covid-19 da gure bizitzak etengabe baldintzatzen dituen une honetan ikastaro hau horren lekuko eusko legebiltzarraren 40. urteurrena atzerako begirada da ikasturte honetarako aukeratutako gaia ezin ziteken besterik izan izan ere aurten 40 (...) izan gara eta legebiltzarrak horretan paper garrantzitsua izan du **en estos dos legislaturas el parlamento vasco se ha ido construyendo y consolidando dia a dia del mismo modo que este pueblo nuestro pueblo se ha ido reconstruyendo la trayectoria de la camara ha sido y es fiel reflejo de la evolucion social la presencia de la mujer (...)** eta konpromisoz aurre egiteko zuen ekarpenak helburu horretan lagunduko dugula sinetsita berriz ere eskerrak eman nahi dizkizuet guztioi



# 6

# Transcribe un audio en el OH Portal

<https://www.oralhistory.eu/>



OCTRA: plain.par , Language: eng-GB , Audio duration: 00:58



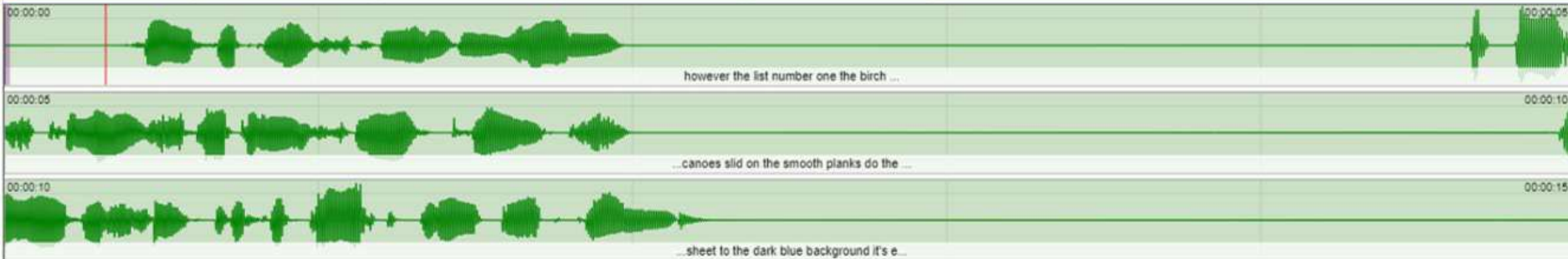
OCTRA v1.4.3 (url) — Dictaphone Editor Linear Editor 2D-Editor

TRN ⓘ ✂ Werkzeuge ⬇ Exportieren ⚙ ⓘ DE ▾

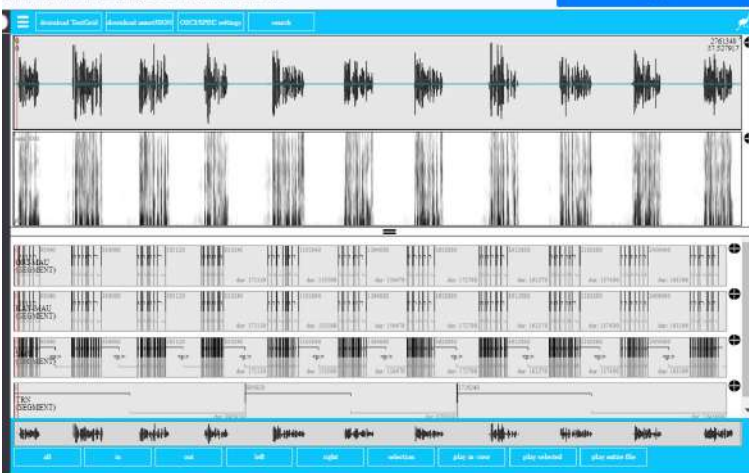
TASTENKOMBINATIONEN [ALT + 8]

ÜBERSICHT [ALT + 0]

HILFE



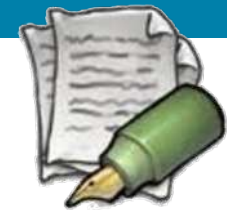
o: Harvard2.TextGrid , Language: eng-GB , Audio duration: 00:58



Webinar: <https://youtu.be/X6bFGJpMjVQ>

# 7

## LARKA: Búsqueda de herramientas para el aprendizaje de lenguas



Language Acquisition Reusing **Korp**

Write or paste a text into the field below.

- Hitex: complejidad oracional
- Coursebook editor: corpus de unidades didácticas
- Evaluación de textos
- Etiquetado lexicográfico: según CEFR
- SiWoCo: complejidad léxica
- Creador de pseudopalabras
- ...

What do you want to assess?

Learner essay

Text readability

Show all words of the following CEFR level(s)

A1

A2

B1

B2

C1

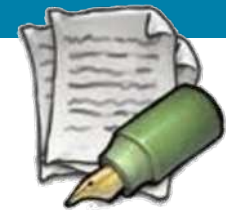
Additional options

Mark all potentially incorrect words

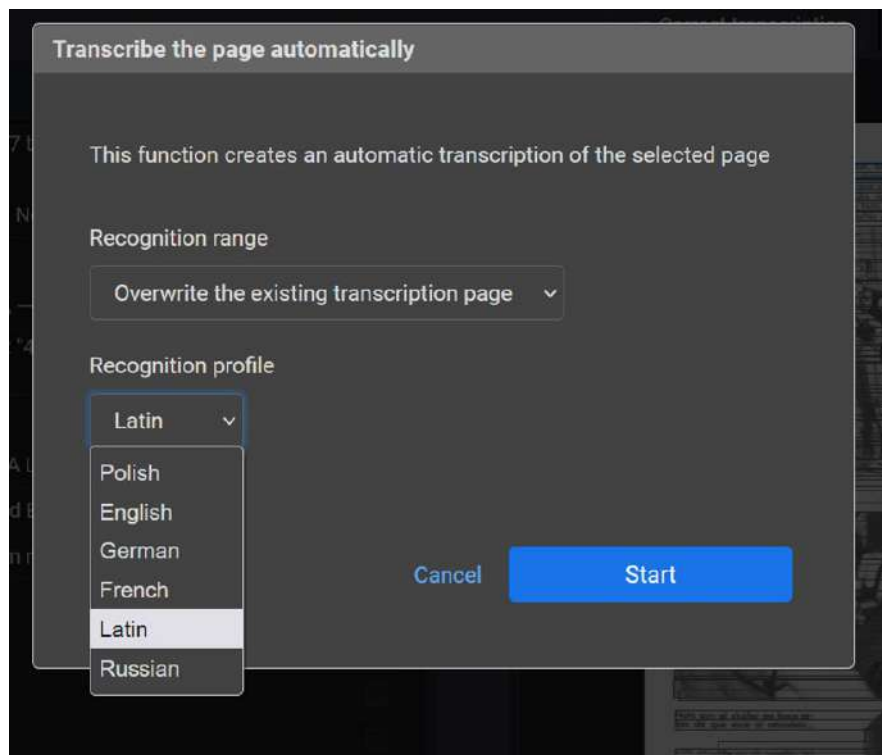
Use Spellchecker

# 8

## Digitalizar documentos con OCR/HTR Virtual laboratory

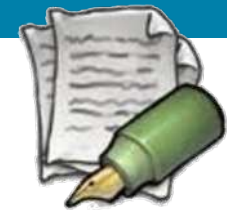


<https://wlt.pcss.pl/>



# 9

## Cierra el círculo de investigación



- Deposita tus datos tú mismo en
  - **CLARIN**
  - **EuDat (B2Share)**
  - Zenodo
  - European Language Grid
  - ELRC-SHARE
  - European Language Resources Association (ELRA)
  - META-SHARE
  - The Linguistic Data Consortium
  - Re3Data
  - CORE Trust Seal
- Hay 125 corpus no-catalogados
  - El 60% en LREC2020 (10% de artículos en proceedings)
  - 17% en la revista LRE. 2016-2020 (22 números)
  - 23% corpus encontrados por otros canales
- El 69% de los catalogados está en el CLARIN resource families
- El 52% se puede descargar de GitHub o de páginas personales
- El 13% se puede consultar online y en concordancias
- El 8% se puede solicitar al autor
- El 6% está en proceso de preparación
- El 20% que falta, no está claro

(Lenardič and Fišer 2022)

# Servicio de depósito: CLARIN y B2share

Decide el mejor lugar para depositar tus datos o herramientas:

- [www.clarin.eu/content/depositing-services](http://www.clarin.eu/content/depositing-services)



The following **certified CLARIN centres** offer depositing services:

Centre	Location	Depositing offer
ACDH-CH	Austria	<a href="#">Any linguistic and/or NLP data and tools</a>
LINDAT-CLARIAH/CZ	Czech Republic	<a href="#">Any linguistic and/or NLP data and tools</a> : corpora, machine translation systems, web services, etc
LINDAT-CLARIAH/CZ	Czech Republic	<a href="#">Language Resource Inventory</a> : An easy-to-use interface for submitting (meta)data and that it can be used immediately
CLARIN-DK-UCPH	Denmark	<a href="#">Danish language resources</a> : The focus is on writing texts with annotations, imdi-sessions containing other data.



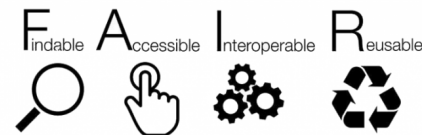
# 10 Preguntas a responder



- Podría la infraestructura...
  - ... evitar la fragmentación: **ALL-LT-in-ONE-URL**
- ganar interoperabilidad en (y entre) infraestructuras
  - para construir infraestructura **que no se desarrollará** en Europa
  - ¿ganar en **impacto social**?
  - ¿facilitar servicios para llegar mejor a la **comunidad**?
  - ¿desarrollar la infraestructura con **proyectos coordinados**?



**EUROPEAN OPEN  
SCIENCE CLOUD**



# 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

Altuna, B., Irukieta, M., Estarrona, A., Farwell, A., Arriola, JM., Alkorta J., Arregi., X (2022). CLARIAH-EUS: Building a Cross-border CLARIAH Node for the Basque Language. Digital Humanities Conference November 23-25. Budapest.

Irukieta, 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. Irukieta, 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/>



# HUMANIDADES DIXITAIS:

recursos  
ferramentas  
e servizos

15-18 de xullo  
2024  
Aula C01  
Facultade de Filloxía  
USC



CLARIN



# CLARIN-ERIC: Conocer y participar en la infraestructura Europea del Lenguaje

Mikel Iruskieta  
HiTZ - UPV/EHU

<https://orcid.org/0000-0002-6121-3902>