CLARIN

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

Mikel Iruskieta
HiTZ - UPV/EHU
https://orcid.org/0000-0002-6121-3902

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INSTITUTO DA LINGUA GALEGA

Universidad del País Vasco Euskal Herriko Unibertsitatea

HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

CLARIN en dos palabras

- Infraestructura común de recursos del lenguaje y tecnologías.
  ***C**ommon **La**nguage **R**esources and Technology **In**frastructure*
- **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

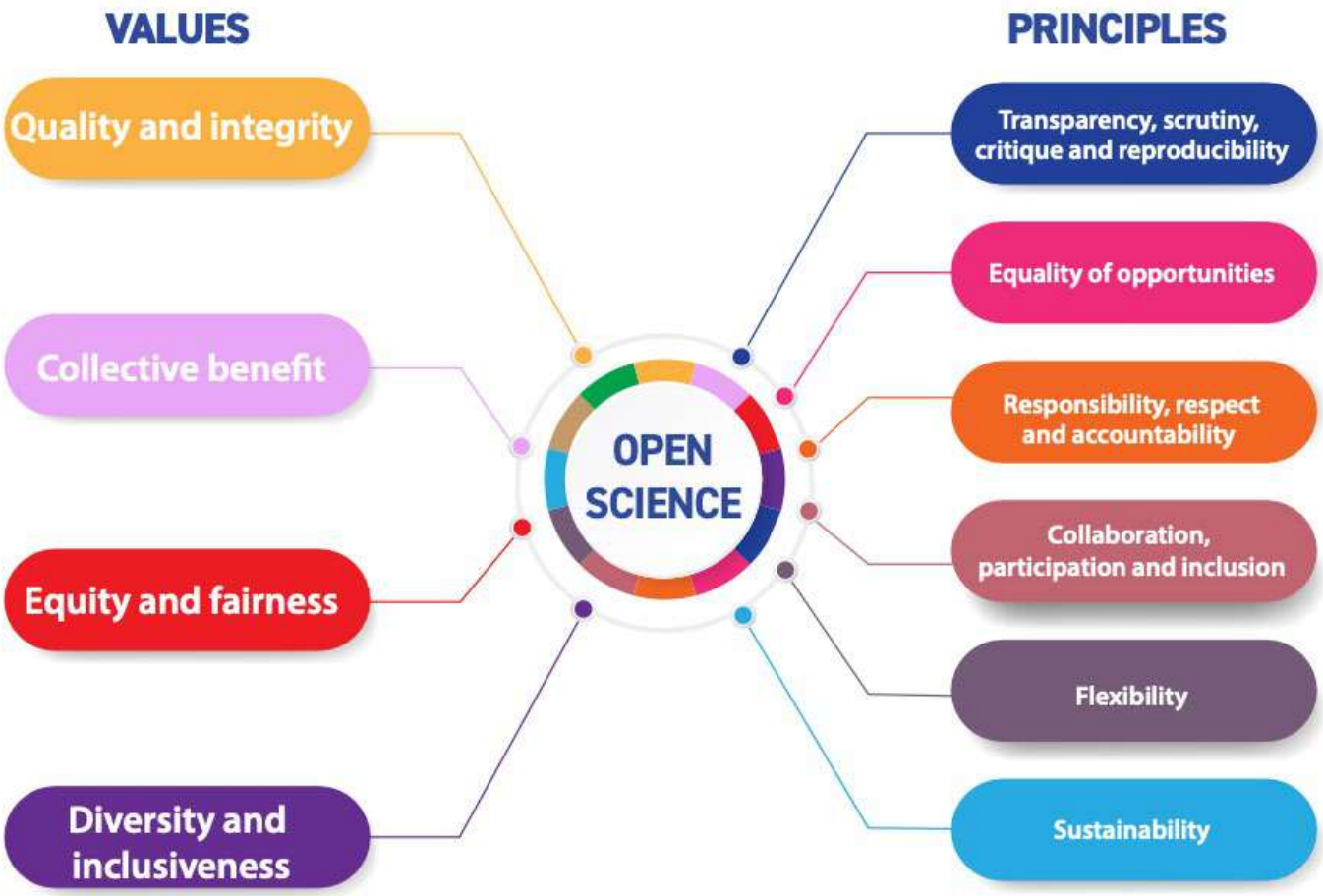
Interoperable

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



CLARIN y la ciencia abierta

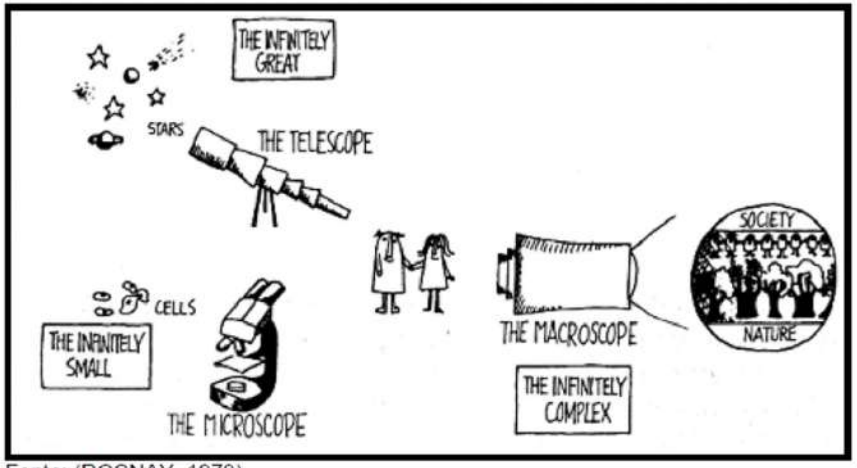
- Promover el intercambio y la reutilización de datos lingüísticos a través de **registros de datos** sostenibles
- Mejorar e implementar la **interoperabilidad** de datos y servicios lingüísticos
  - Marco común de metadatos
  - Red distribuida de repositorios de datos lingüísticos certificados por 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 multidisciplinarias

- 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)

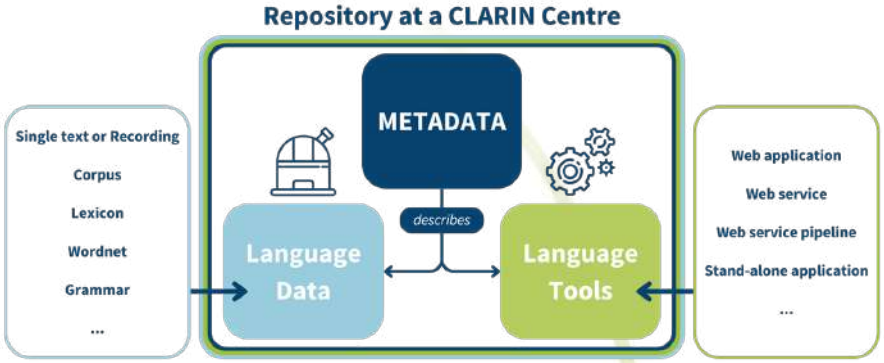
La complejidad de las CCSS afecta la interoperabilidad

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

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**

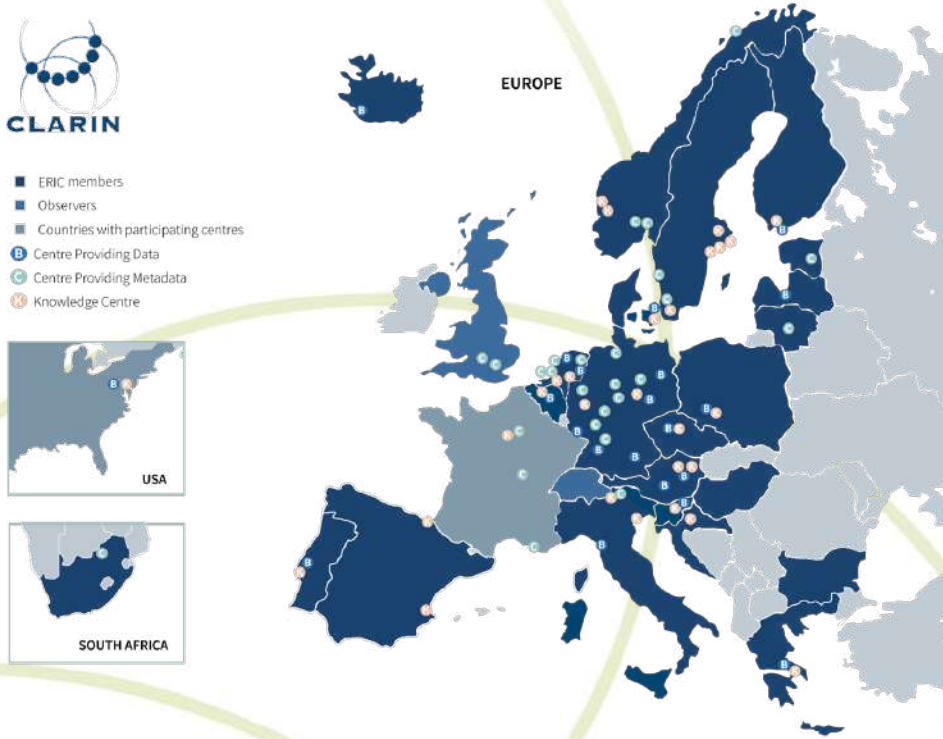
- Recursos de las comunidades europeas permiten la investigación transnacional
- Datos y herramientas en muchos idiomas
- Recursos en muchas modalidades y tipos de datos
- Métodos comparados y en colaboración multidisciplinar

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)



Estrategia de CLARIN 2024-2026

Área 1: Uso académico

- Obj1: Aumentar la repercusión de CLARIN
- Obj2: Ampliar el número de usuarios

Área 2: Uso no-académico

- Obj 3: Desarrollar compromiso con comunidades no-académicas para impulsar innovaciones técnicas o sociales

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

- 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

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

- Obj 8: Alcanzar un alto nivel de accesibilidad y usabilidad de los servicios técnicos centrales

Área 5: Colaboración e intercambio de ideas

- 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

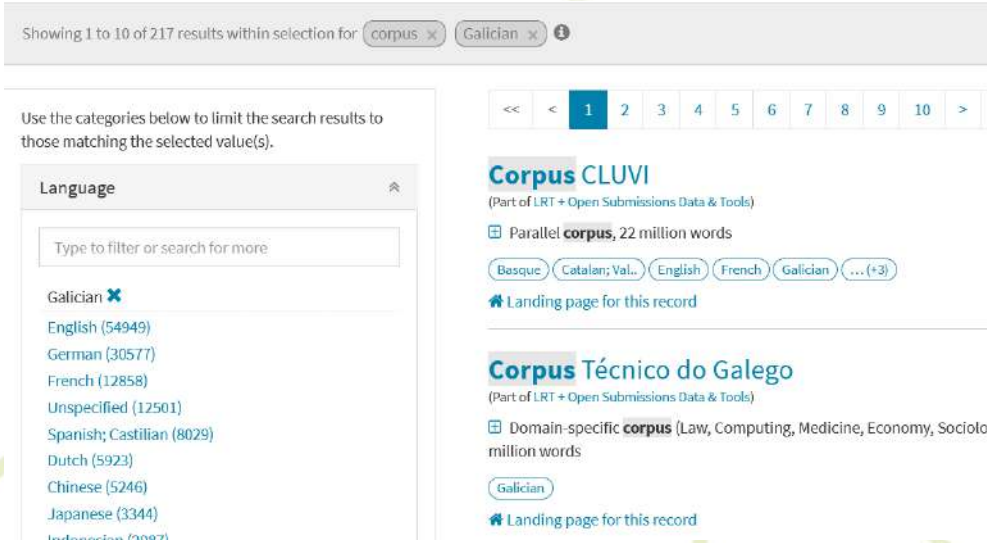
Según las **escuelas de pensamiento** de la Ciencia Abierta (Fecher & Friesike, 2014), las infraestructuras se organizan de diferente forma

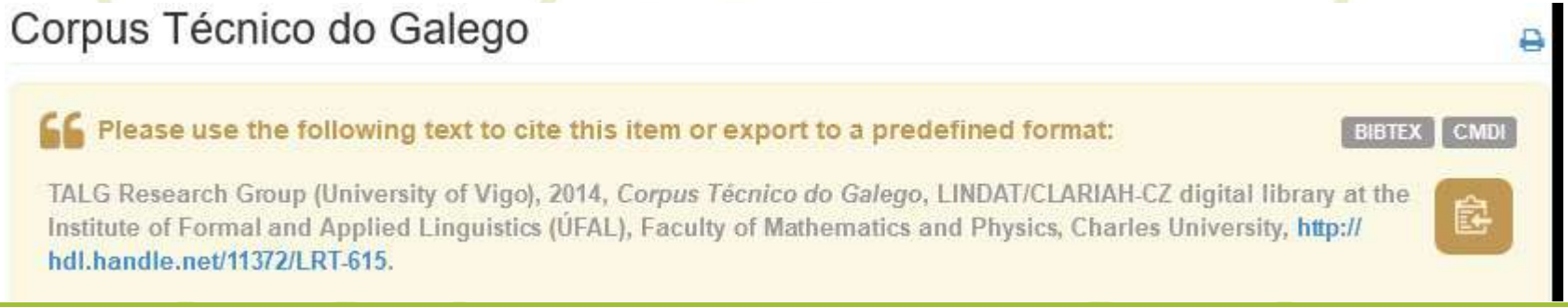
- DARIAH > Escuela **pragmática:** investigación colaborativa
- CLARIN > Escuela **infraestructural:** arquitectura tecnológica, eficiencia de las herramientas y datos complejos
- Escuela de **medición:** impacto científico
- Escuela **democrática** y pública: acceso

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:



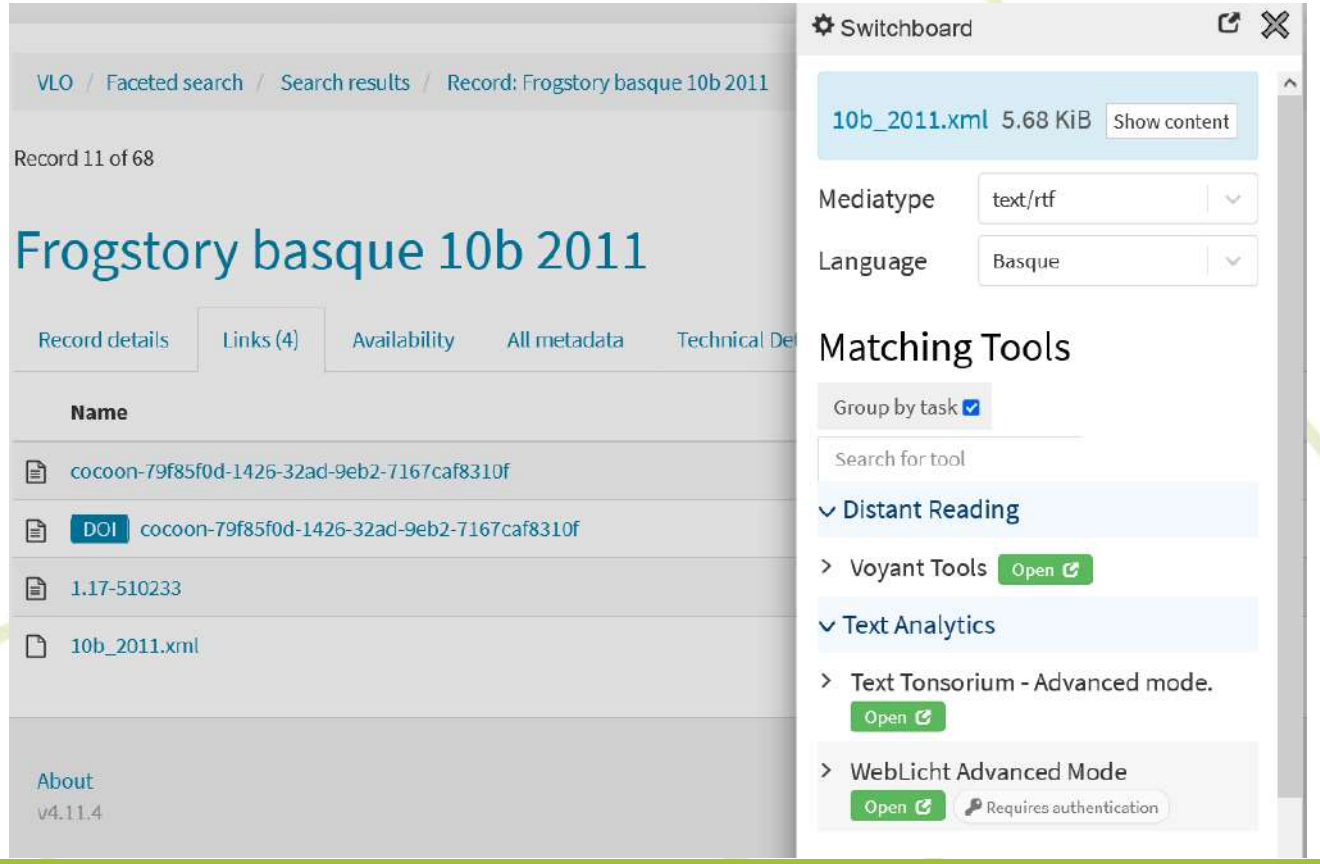


Ejercicio VLO: Búsqueda de datos

https://vlo.clarin.eu

0

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



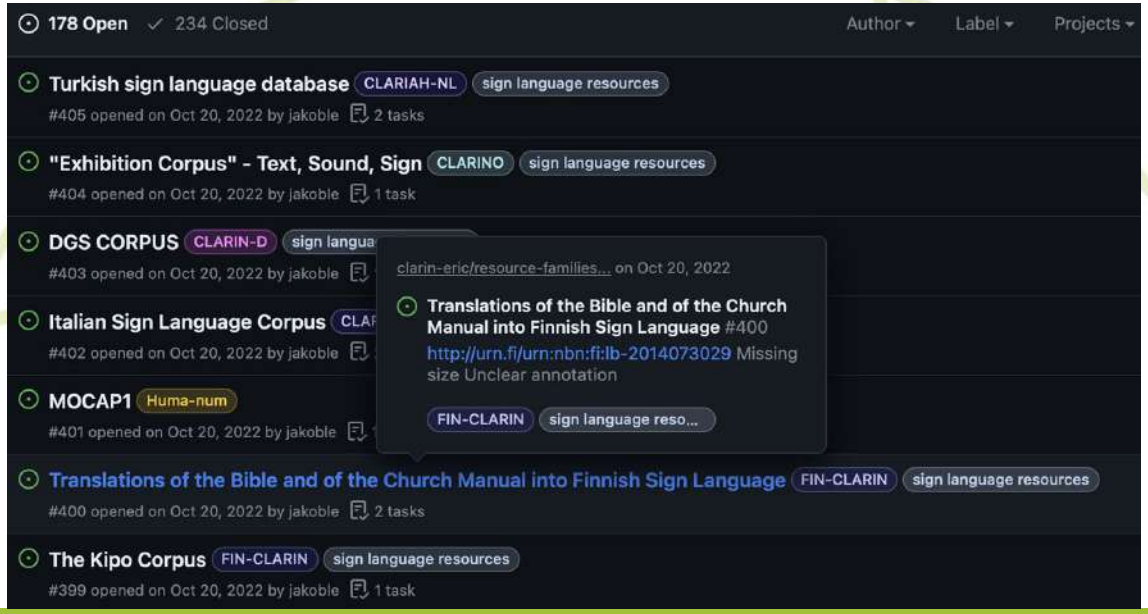
No todo es FAIR

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

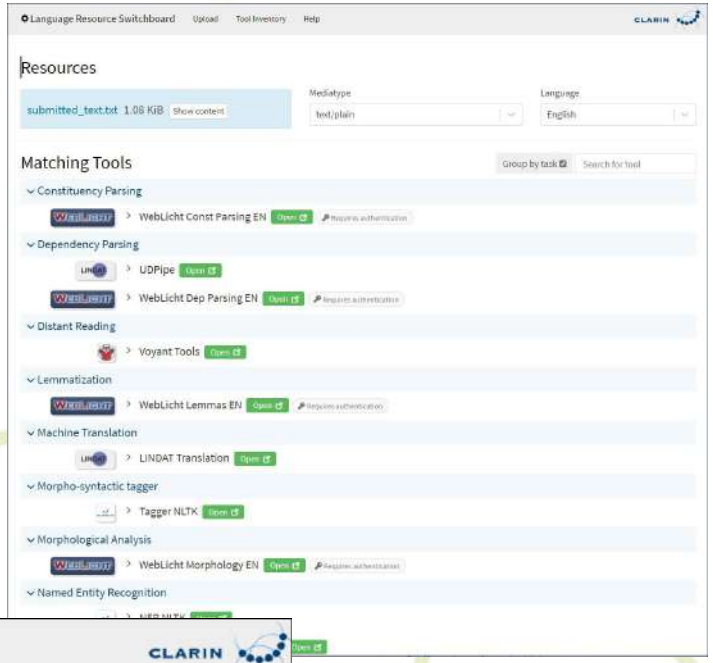
- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**e-usable

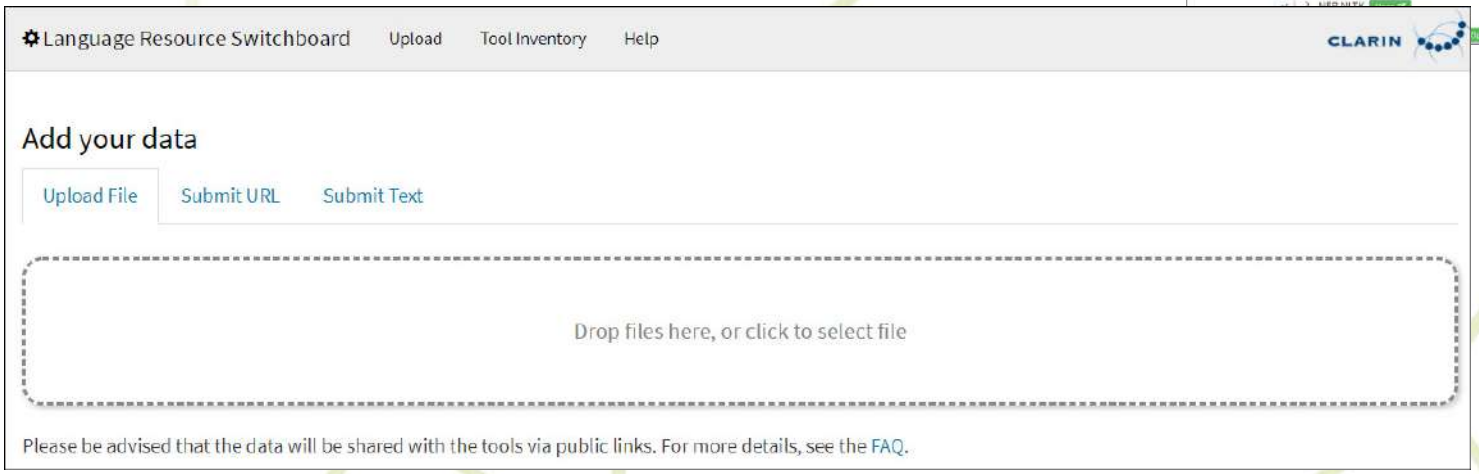
(Fišer and Lenardič 2020)



Language Resource Switchboard
https://switchboard.clarin.eu/

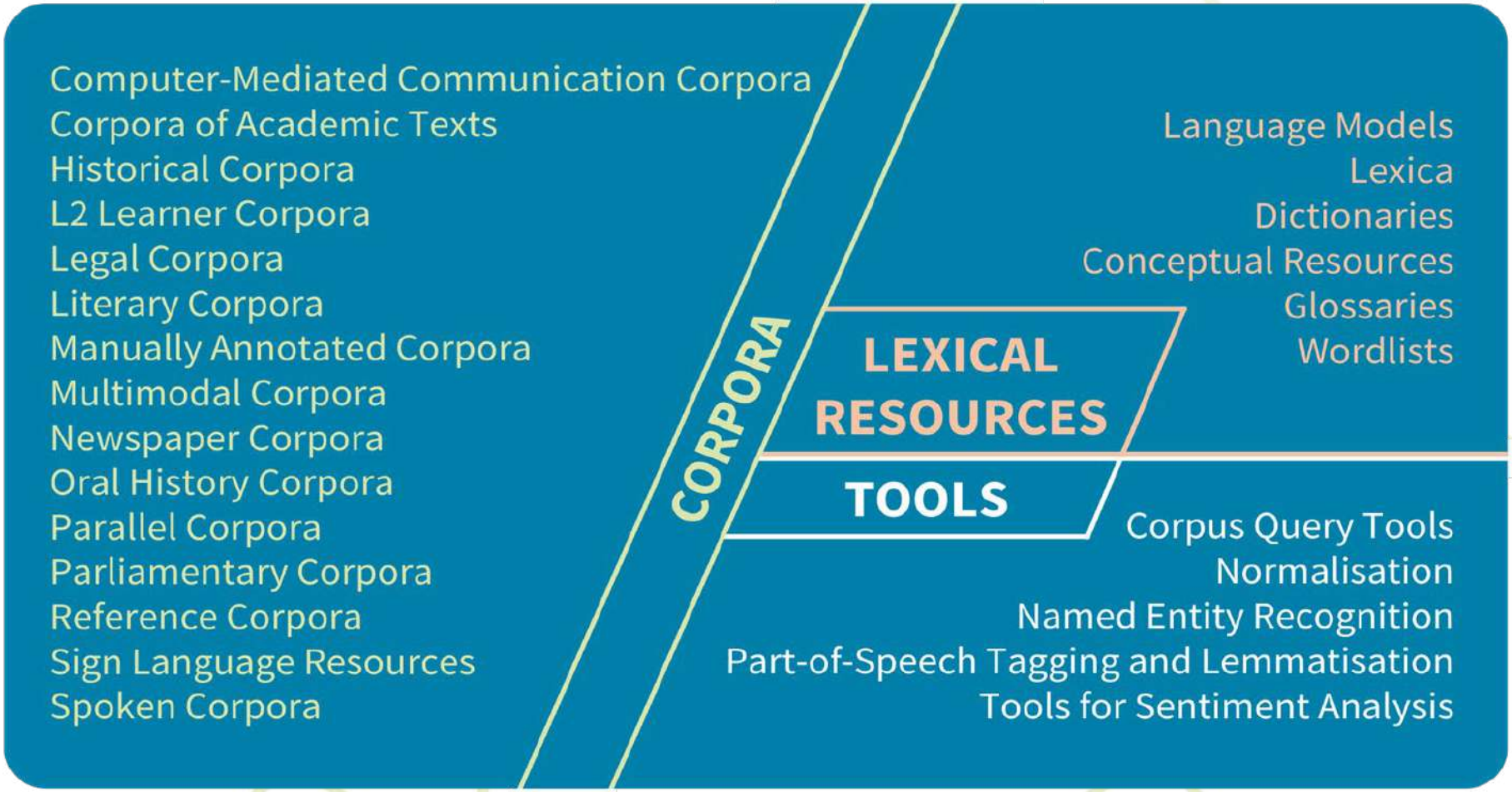
- Prueba con un texto para buscar la herramienta más adecuada para ver que tareas de PNL se pueden realizar con ese archivo (formato/lengua)
- Se puede acceder directamente desde la VLO





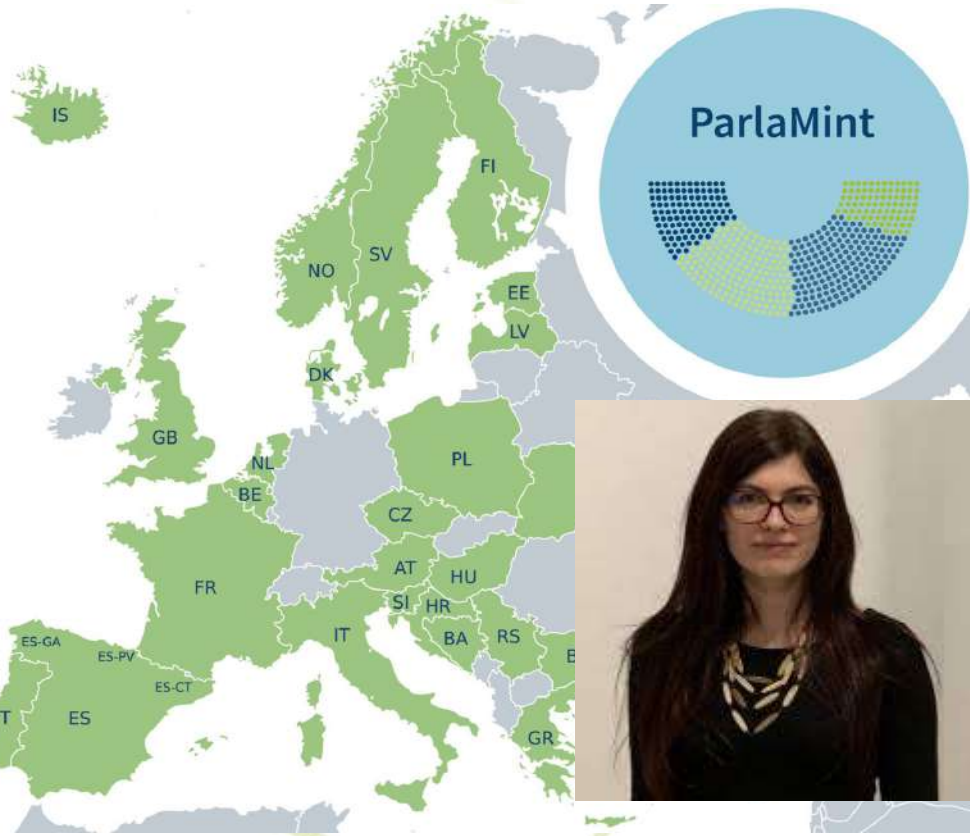
Recursos CLARIN: *Resource Families*

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



Proyecto emblemático de CLARIN

https://www.clarin.eu/parlamint



Adina Vladu (Universidade de Santiago de Compostela)

15:00 - 17:30 *A explotación dun corpus para Humanidades e Ciencias Sociais: ParlaMint**

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

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

- 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

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

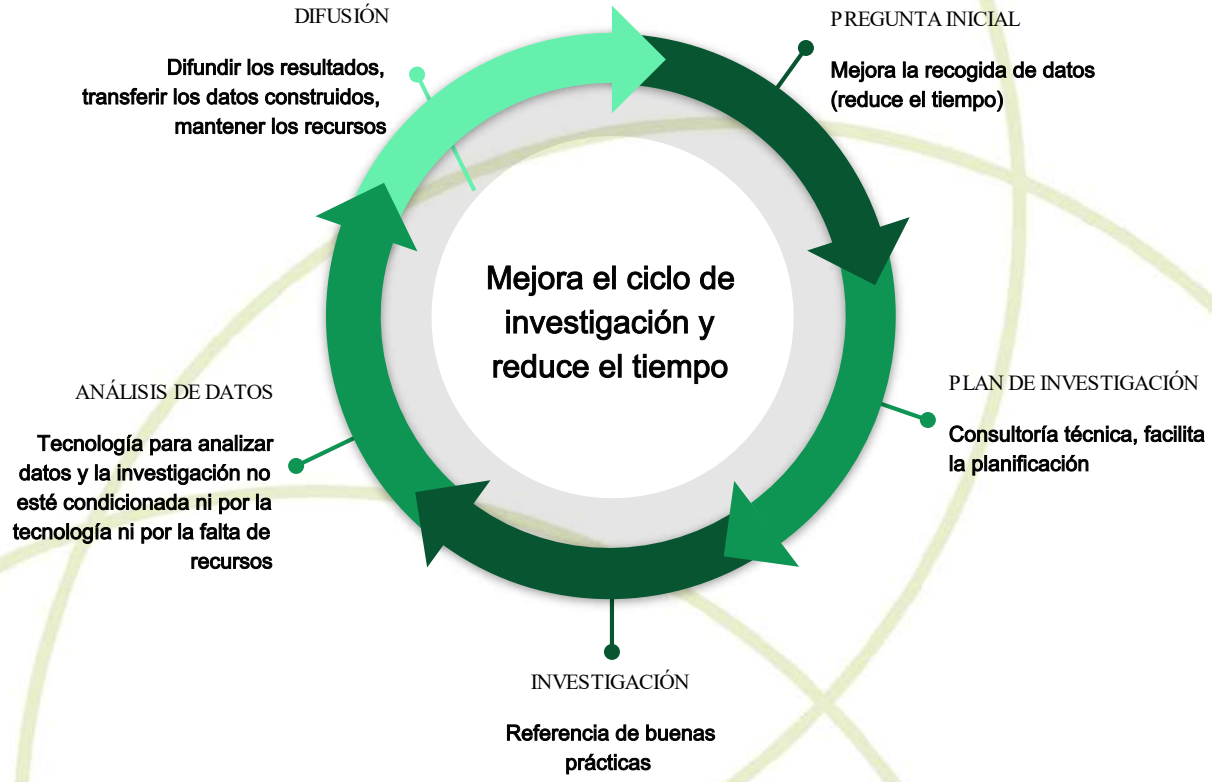
https://sidih.github.io/voices/toc.html

Tareas

Pixabay: cstibi

Tareas para un curso de verano

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

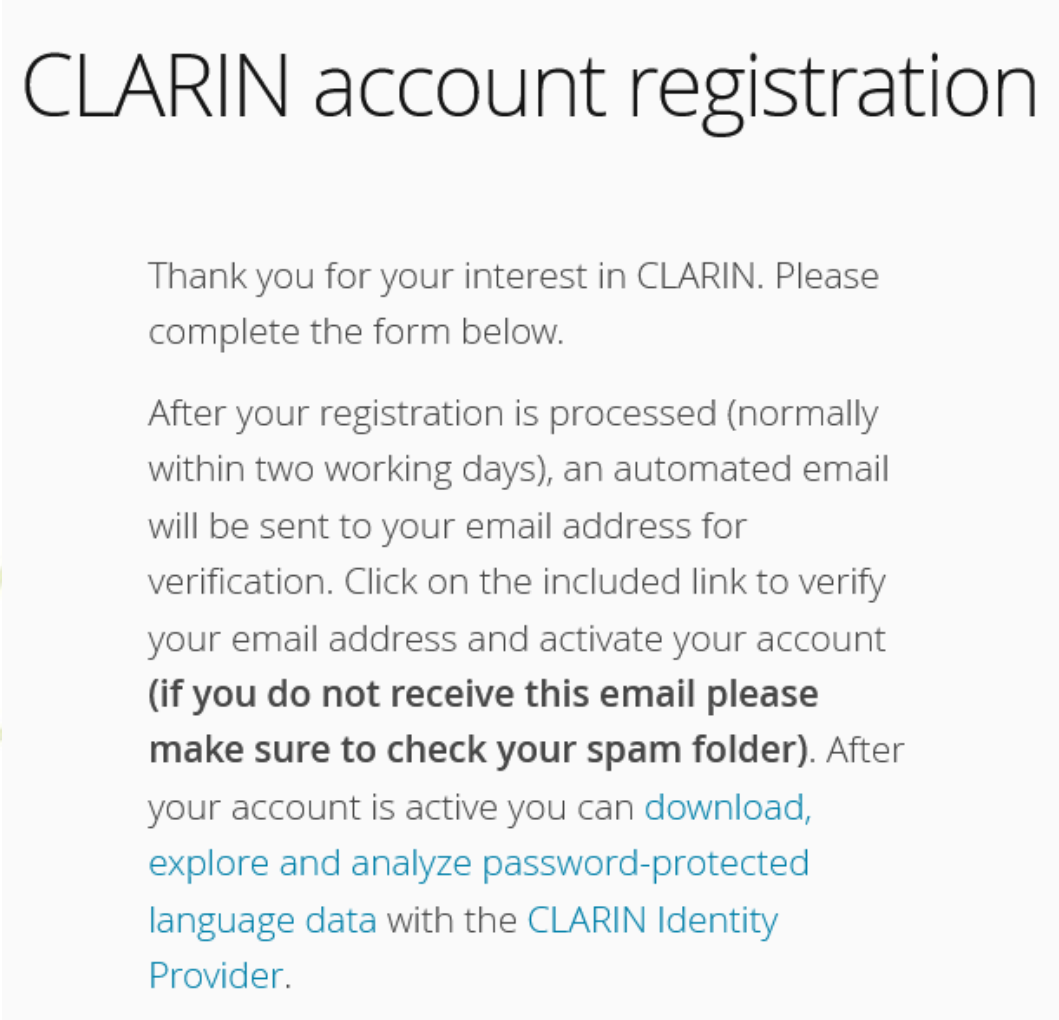


CLARIN: experiencias de usuarios

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

Registrarse en CLARIN para acceder a toda la infraestructura

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



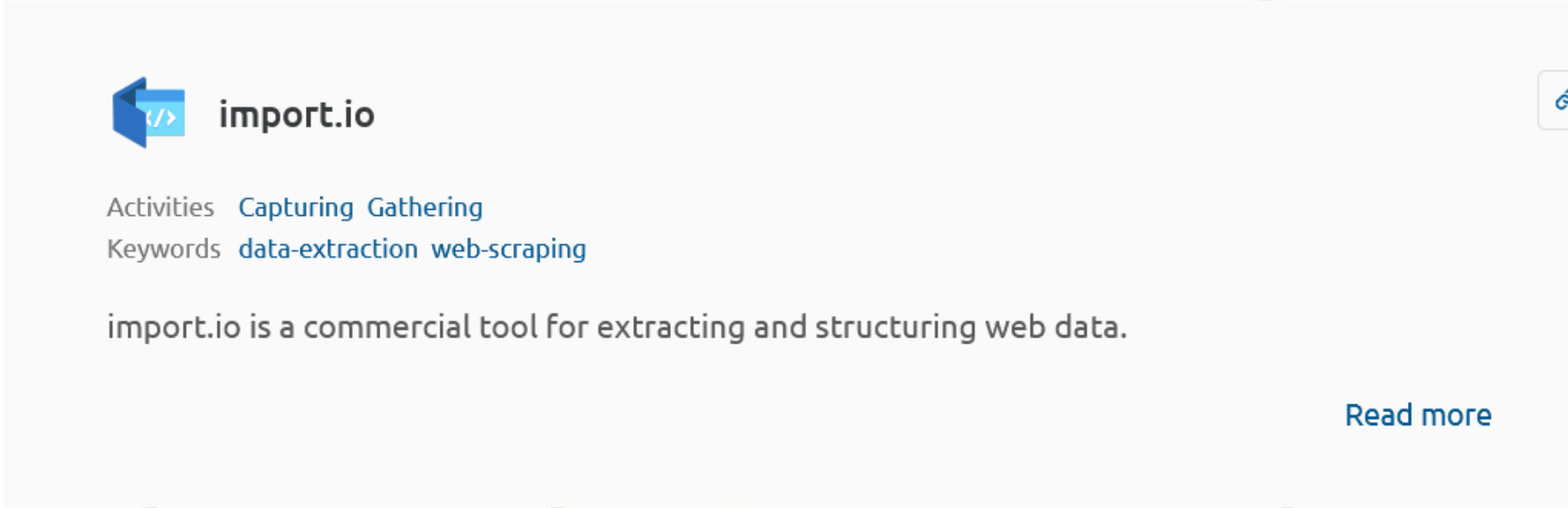
https://idm.clarin.eu/unitygw/pub#!registration-CLARIN%20Identity%20Registration

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

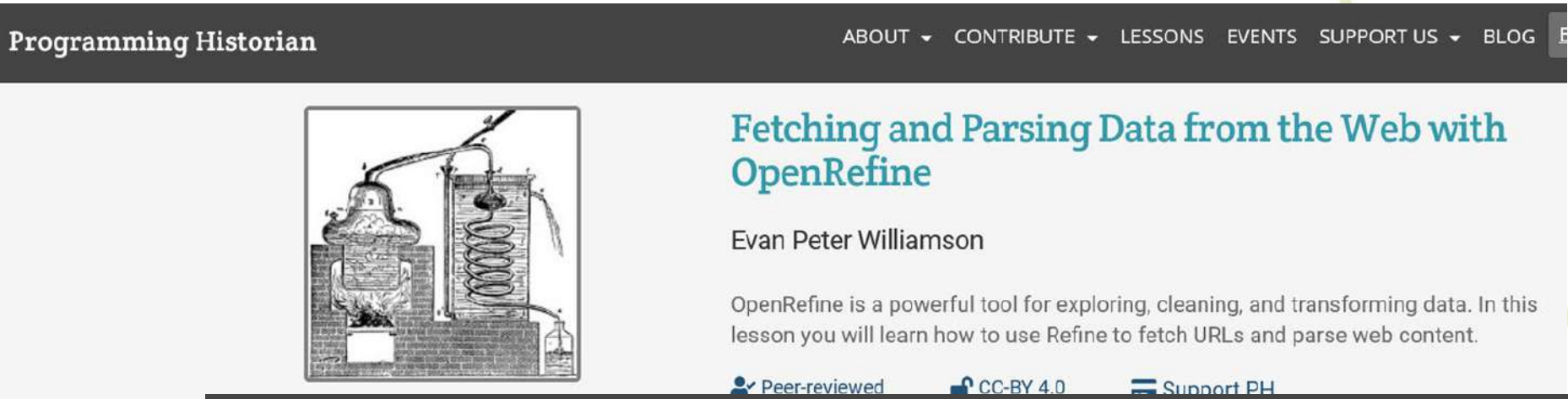


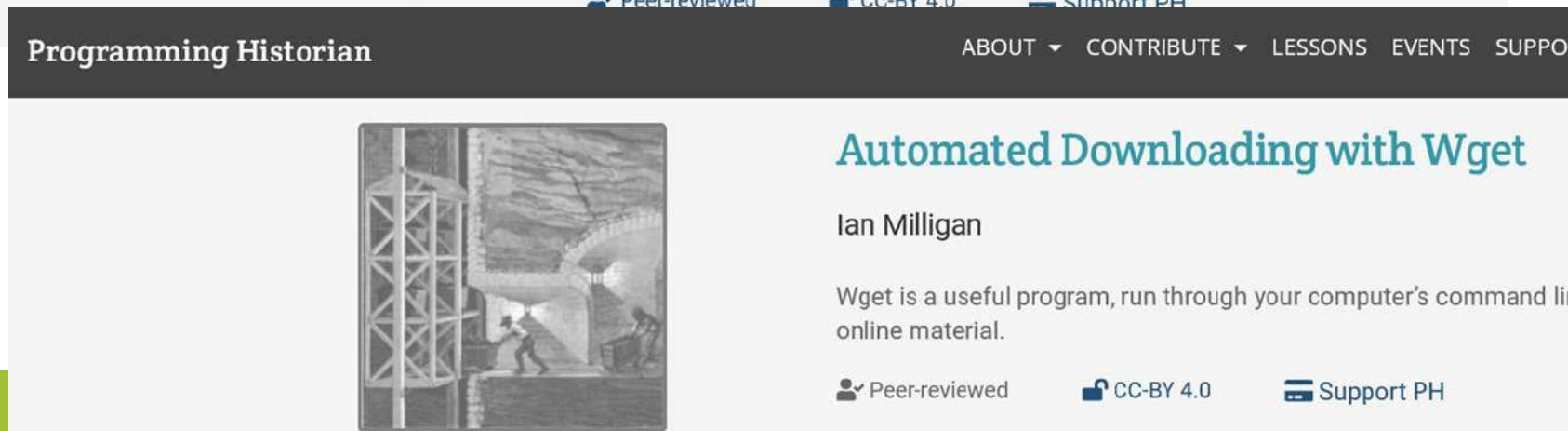
https://marketplace.sshopencloud.eu/search?q=web-scraping&categories=tool-or-service

1

Manuales de PH

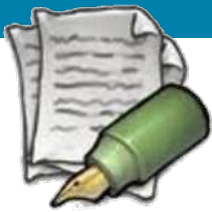
- OpenRefine y Wget

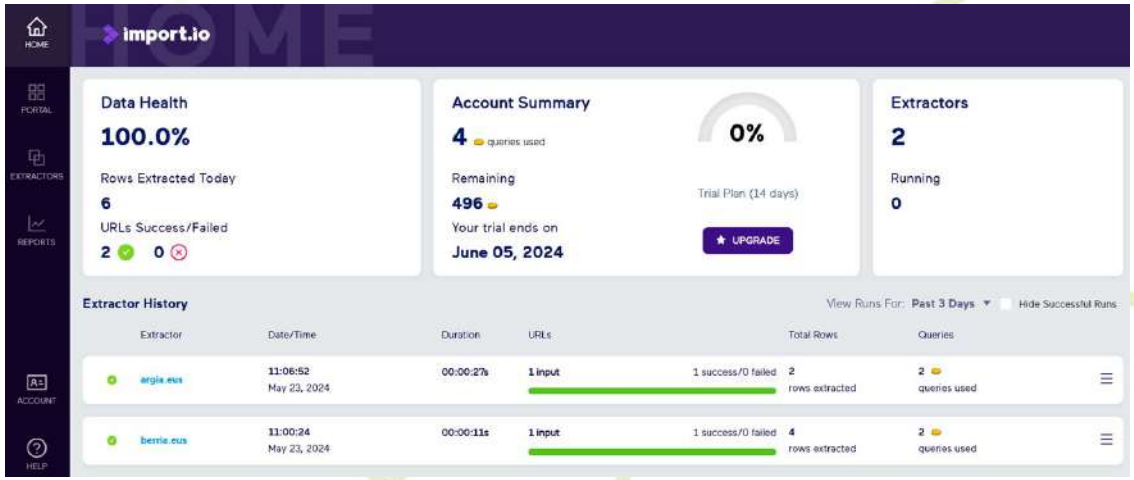




2

Web-scraping
(import.io)





https://app.import.io/dash/portal

Choose your plan

Trial Plan (14 days)

500 queries

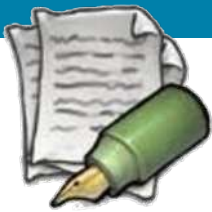
- Point and click training
- Interactive extraction
- E-mail and ticket support
- Capture screenshots
- API access and Webhooks

Free

Selected

3

Uso de las infraestructuras: texto escrito



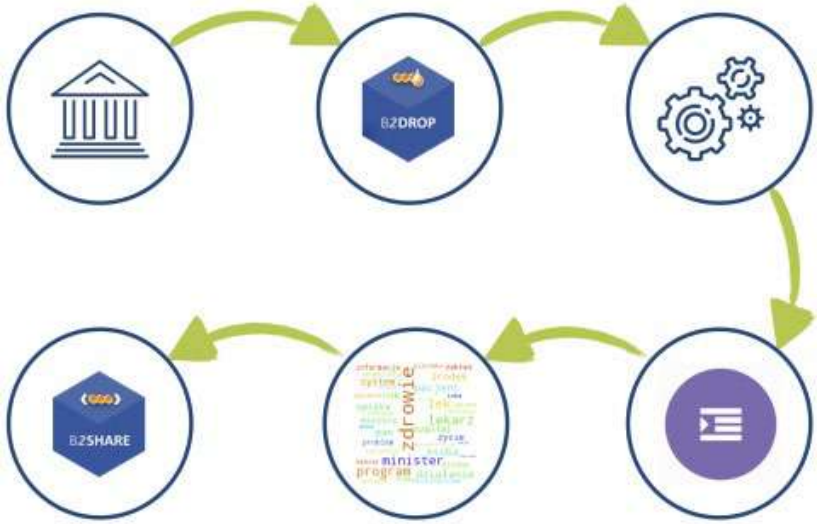
Texto escrito:

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

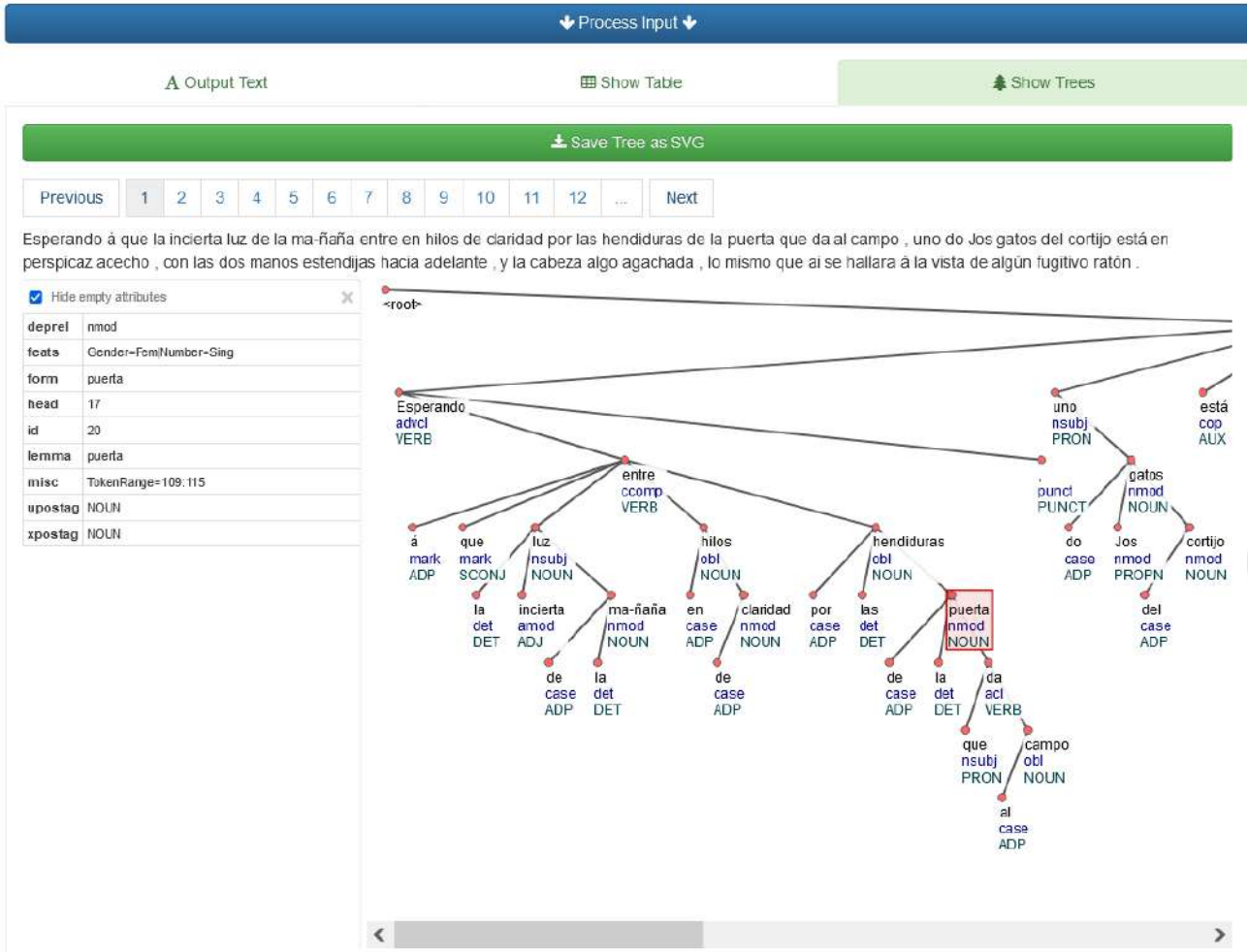
1. Descargar texto: Gitzip
2. Corpus en Eudat
3. Pasar a colección virtual
4. Analizar con un clic en Switchboard
5. Elegir un recurso para el análisis de texto
6. …

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

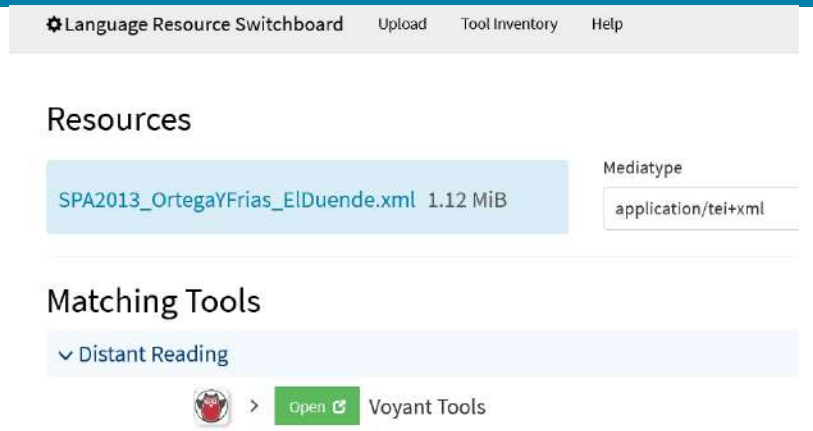
- Recursos (2023)
  - Para el castellano: 12
  - Para el inglés: 22
  - Para el alemán: 20
  - Para el polaco: 32

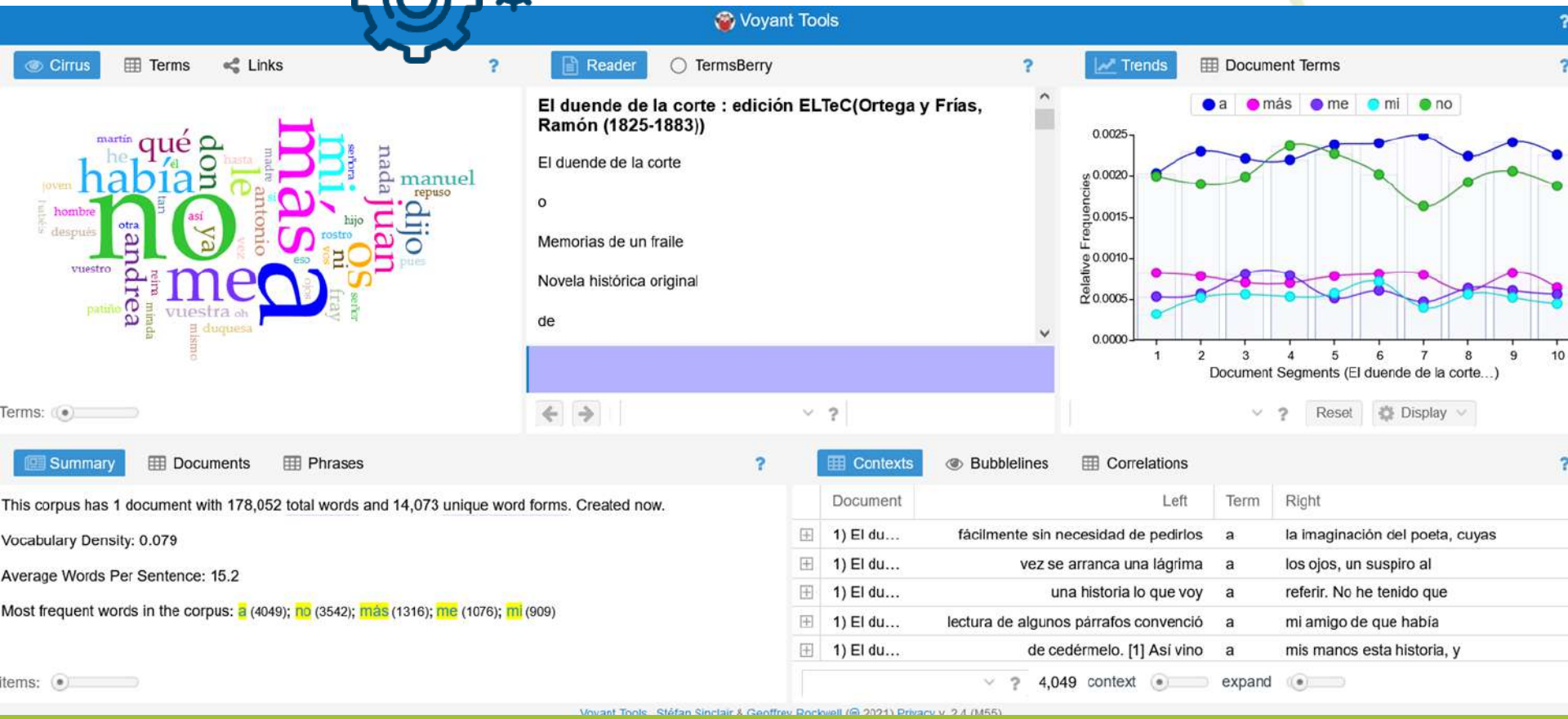
1. Constituency Parsing
2. Coreference Resolution
3. Dependency Parsing
4. Distant Reading
5. Extraction of Polish terminology
6. Inclusion detection
7. Keyword Extractor
8. Lemmatization
9. Machine Translation
10. Metadata Processing
11. Morpho-syntactic tagger
12. Morphological Analysis
13. Named Entity 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



Ejemplo de análisis de TEI+XML del corpus ELTeC





Colección virtual ad hoc

Virtual Collection Registry | Browse | Create | My Collections | Help | mikel.iruskieta_ehu.eus@clarin.eu | CLARIN

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)

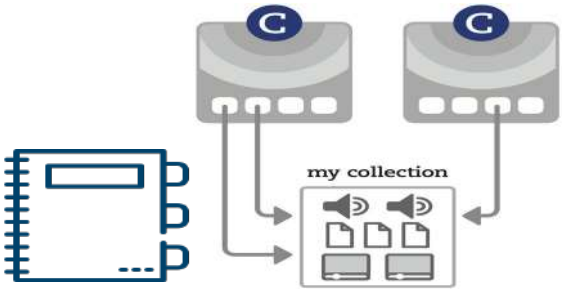
CONSSA (https://github.com/cligs/conssa)

Audio sobre novelas (http://contenidosdigitales.uned.es/fez/view/intecca:VideoCMAV-5a6f2b5bb1111f57648b4cb3)

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

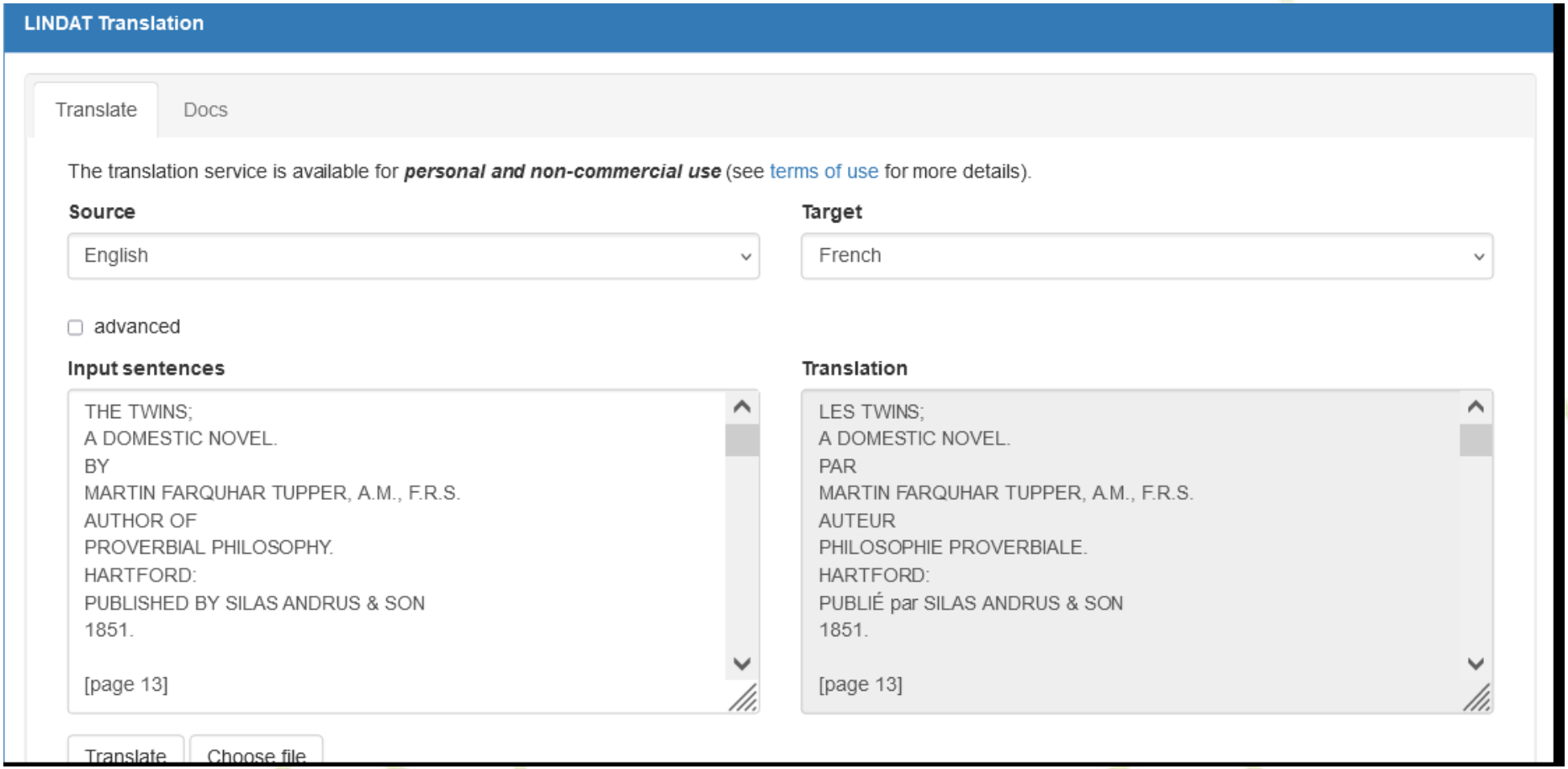
Caballero (https://b2drop.eudat.eu/s/Hz5cP4aip5wobDP)

Save Collection | Cancel



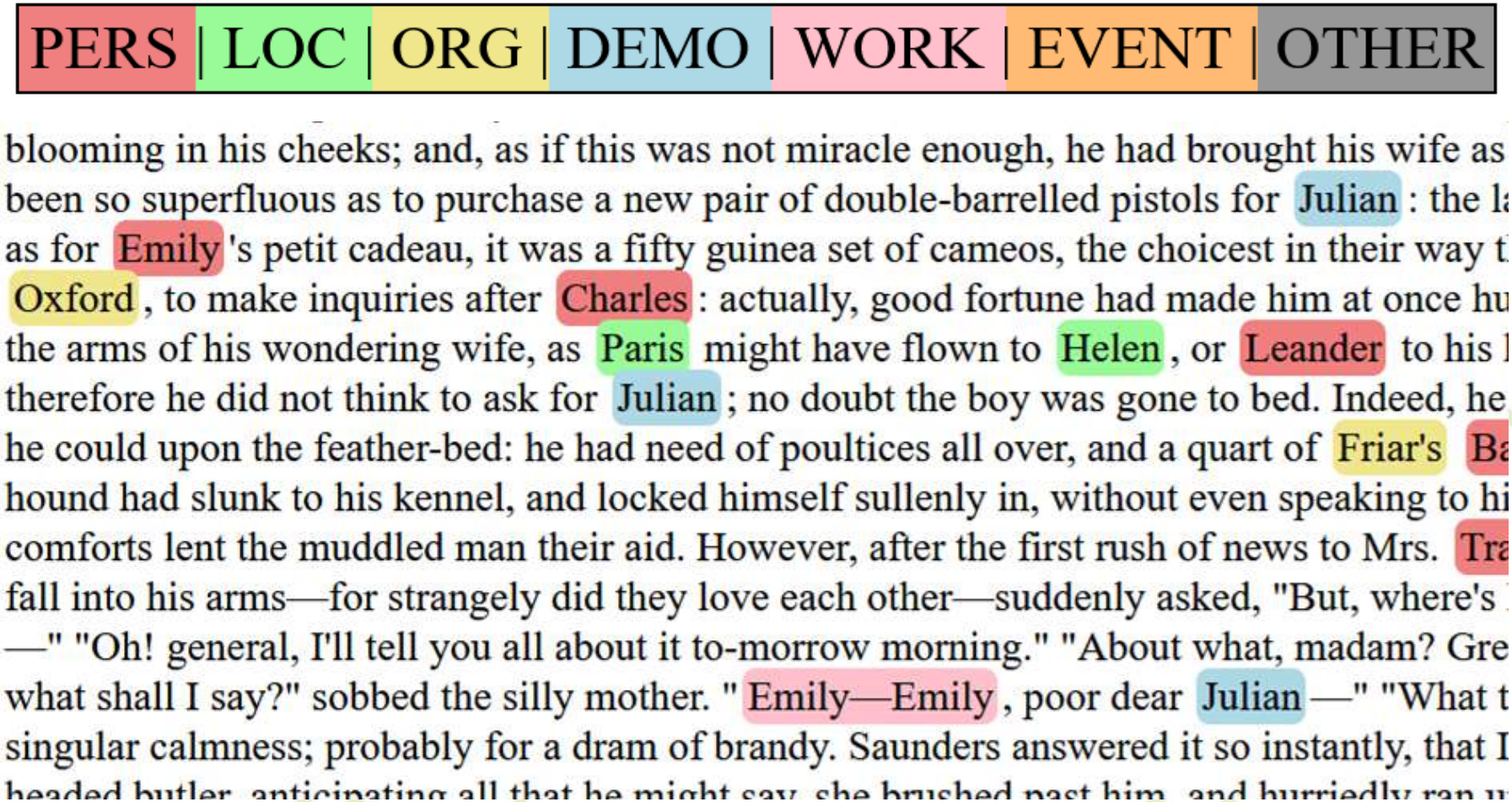
https://www.clarin.eu/content/virtual-collections

Traducir documentos desde las colecciones virtuales

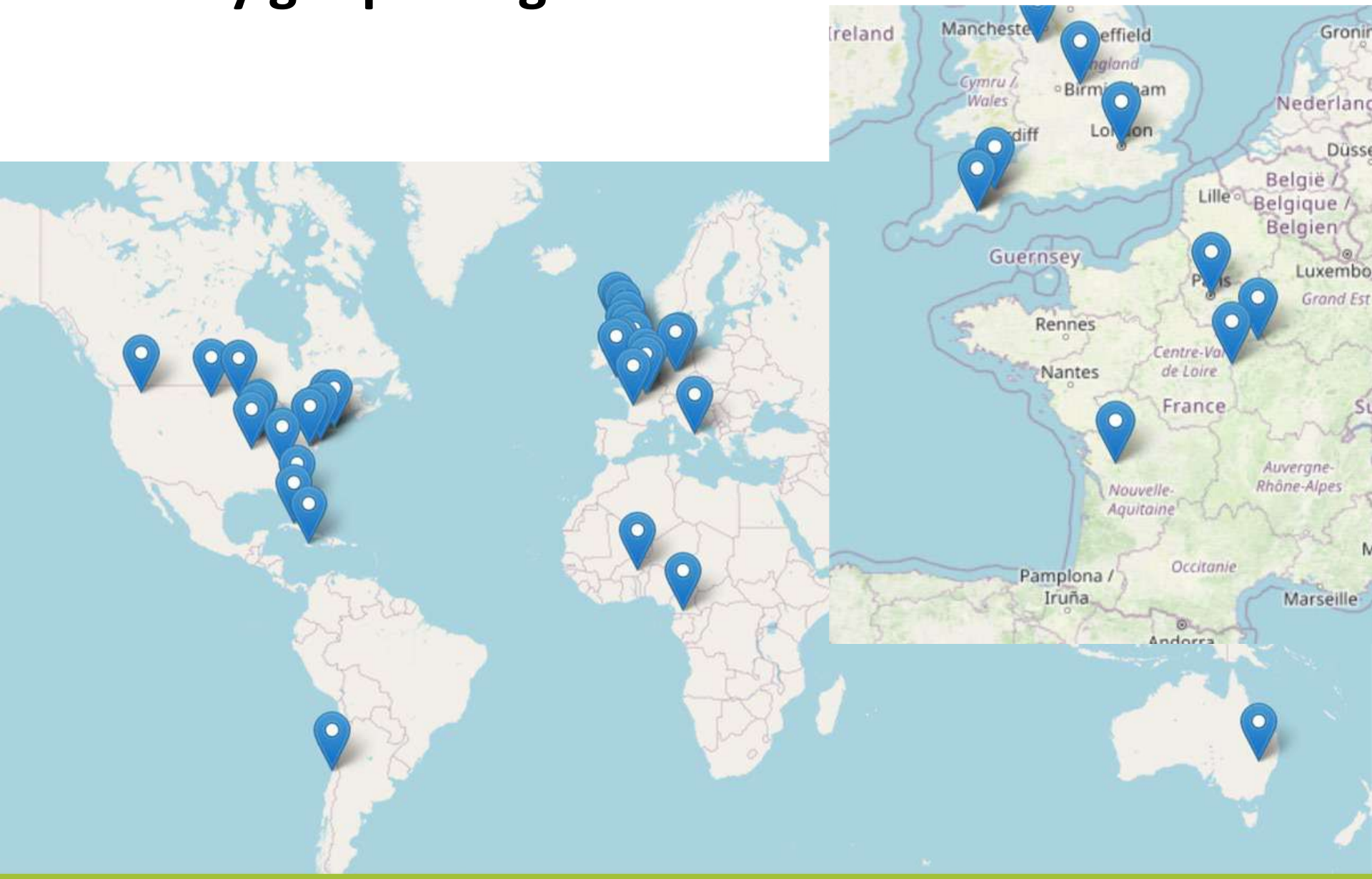


https://lindat.mff.cuni.cz/services/translation

NER: detección de nombres propios

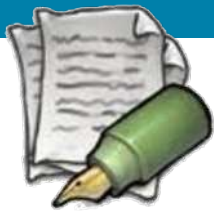


NER y geoparsing



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 ˈombres delˈgados de alˈbanja/
[doˈsombrezðelˈɣaðoszðealˈβanja] o [doˈsombreʂðelˈɣaðoʂðealˈβanja]
Dos hombres delgados de Albania

Bavarian Archive for Speech Signals

CLARIN-D

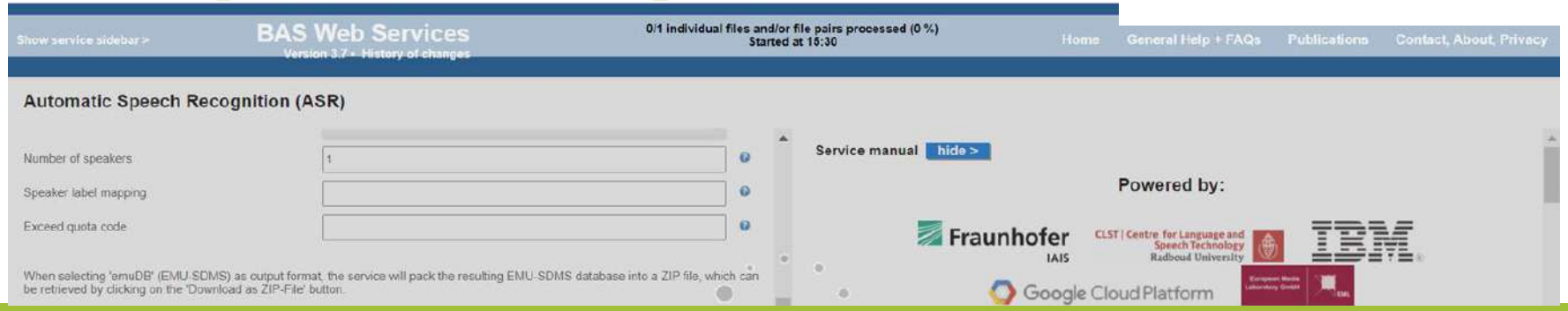
IPS INSTITUTE OF PHONETICS AND SPEECH PROCESSING

https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Grapheme2Phoneme

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



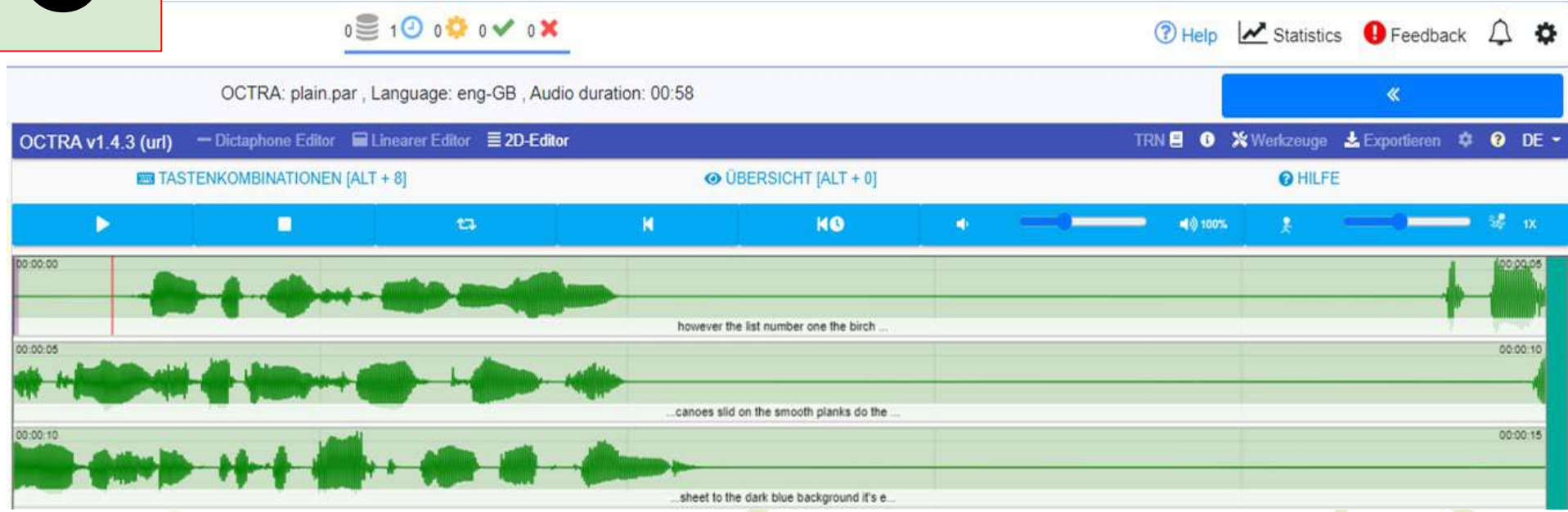
ASR Bilingual Basque-Spanish

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

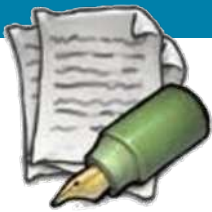
https://www.oralhistory.eu/



Webinar: https://youtu.be/X6bFGJpMjVQ

7

LARKA: Búsqueda de herramientas para el aprendizaje de lenguas



Language Acquisition Reusing Korp

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

Write or paste a text into the field below.

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

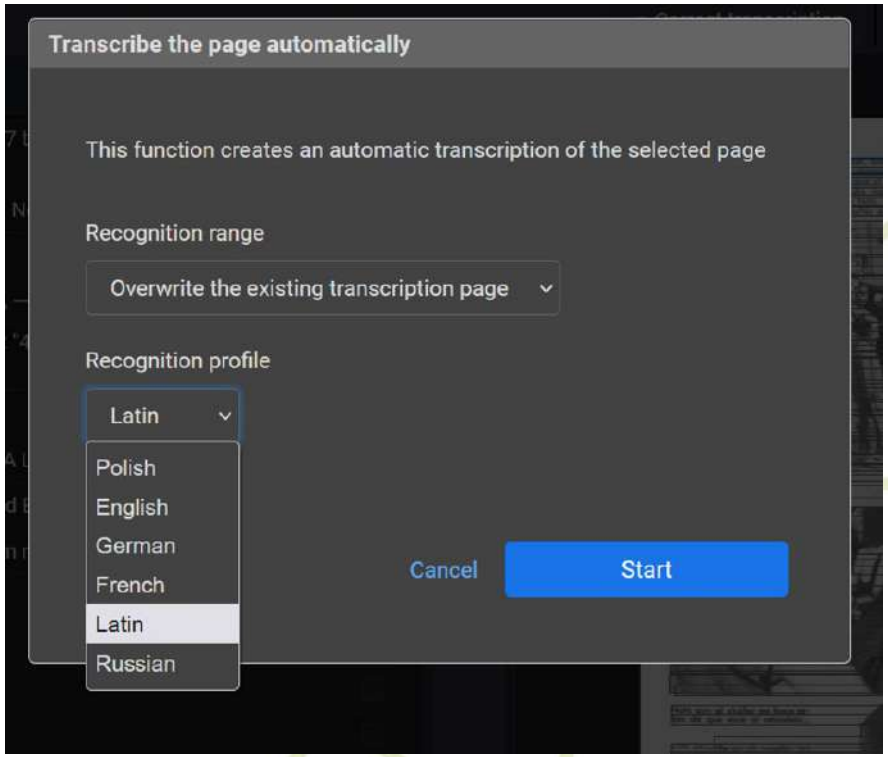
- Mark all potentially incorrect words
- Use Spellchecker

https://spraakbanken.gu.se/larkalabb/texteval

8

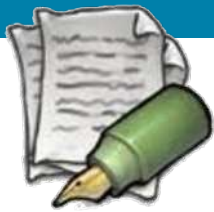
Digitalizar documentos con OCR/HTR
Virtual laboratory

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

The following certified CLARIN centres offer depositing services:

Centre	Location	Depositing offer
ACDH-CH	Austria	Any linguistic and/or NLP data and tools
LINDAT-CLARIAH/CZ	Czech Republic	Any linguistic and/or NLP data and tools: corpo machine translation systems, web services, etc
LINDAT-CLARIAH/CZ	Czech Republic	Language Resource Inventory: An easy-to-use i through submissions and submit metadata. It ((meta)data and that it can be used immediatel
CLARIN-DK-UCPH	Denmark	Danish language resources: The focus is on wri texts with annotations, imdi-sessions containi other data.

B2SHARE

EUDAT

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

Findable Accessible Interoperable Reusable

CLARIN

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., Iruskieta, 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.

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/

CLARIN

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

Mikel Iruskieta
HiTZ - UPV/EHU
https://orcid.org/0000-0002-6121-3902

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INSTITUTO DA LINGUA GALEGA

Universidad del País Vasco Euskal Herriko Unibertsitatea

HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology