

CLARIAH-EUS

CLARIAH-EUS Humanitateetako eta Gizarte
Zientzietako euskarazko ikerketa-azpiegitura
marrazten HaMABI printzioei jarraituz

www.clarin.eu
www.dariah.eu

<http://ixa2.si.ehu.eus/intele/>

Ainara Estarrona
INTELE - HiTZ (UPV/EHU)
Mikel Iruskiet
INTELE - HiTZ (UPV/EHU)

Aurkezpenaren egitura



1. Sarrera
2. INTELE ikerketa-sarearen aurkezpena
3. Europako ikerketa-azpiegituren aurkezpena: CLARIN eta DARIAH
4. CLARIN-K zentroaren aurkezpena: erabiltzaileen esperientziak
5. Ikerketa errazten: kasu praktikoak
6. Eztabaida eta galderak

CLARIN



DARIAH-EU

SAIOAREN LABURPENA: CLARIAH-EUS ZERTARAKO ETA NOLA?

Helburua

- *European Open Science Cloud*: EOSC
- HaMABI printzipioen zientzia ([FAIR](#)): Hartu, Moldatu, Aztertu, Bistaratu eta Iraunarazi

Metodoa

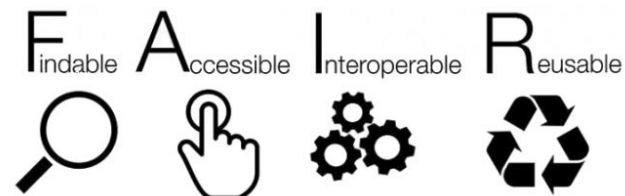
- Elkarreragingarritasuna azpiegituretan (CLARIN)
 - Europar garatuko ez den euskararako azpiegitura eraiki
 - ALL-LT-in-ONE-URL: baliabide guztiak, zerbitzu guztiak

Adibideak

- Elkarren artean komunikatzen diren azpiegitura europarretan ikertzeko erabilera kasuak eta tresnak (erabilerrazak)



**EUROPEAN OPEN
SCIENCE CLOUD**



Azpiegituren justifikazioa

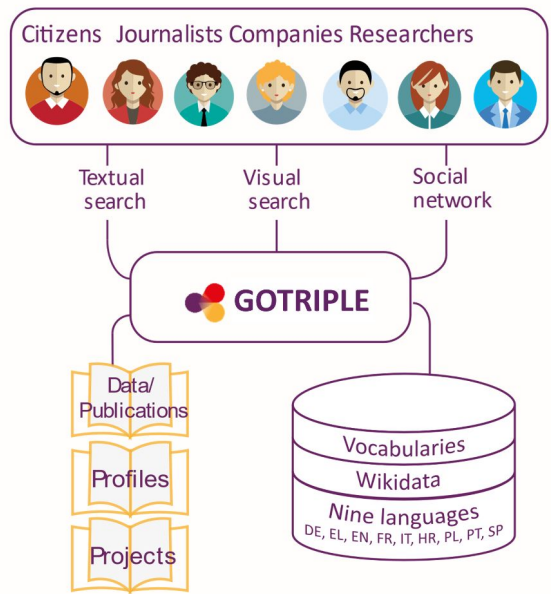
Zientzia: irekia, eraginkorra eta errepikagarria:

- Humanitateetako eta Gizarte Zientzietako ikerketaren zatiketa
- Humanitateetako eta Gizarte Zientzietako baliabideak biltegi desberdintean sakabanatuta
- Gizarte Zientzietan ikerketaren berrerabilpen gutxi
- Diziplinartekotasun gutxi
- Eragin sozial mugatua



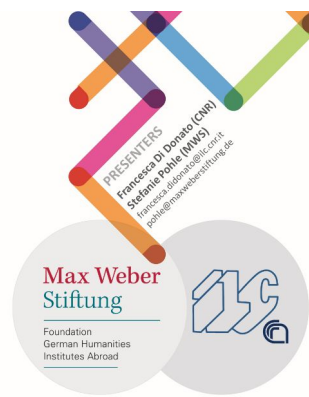
Transforming Research through Innovative Practices for Linked interdisciplinary Exploration

**The SSH discovery platform GOTRIPLE:
A future EOSC service**



WHY TRIPLE PROJECT?

- Strong fragmentation of SSH research
- SSH open scholarly resources (data, publications, other researchers' profiles and projects) currently scattered across local repositories
- Low use and reuse of SSH research
- Interdisciplinary collaboration possibilities are missed
- Societal impact is limited



www.gotriple.eu

TRIPLE will be a dedicated service of the OPERAS RI
 This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 853420.



INTELE: Red estratégica para la promoción de las infraestructuras de tecnologías del lenguaje en eHumanidades y Ciencias Sociales

- **INTELE**: CLARIN eta DARIAH europar azpiegituretan Espainiaren parte hartze ofiziala bultzatzeko ikerketa-sare estrategikoa
 - Humanitateetan eta Gizarte Zientzietan ikerketa sustatzea
 - Nazioarteko proiektuak eta programak sustatzea
 - Helburua: CLARIAH-ES
- **CLARIN**: Common Language Resources and Technology Infrastructure
 - ESFRI ERIC (2012) y ESFRI Landmark (2016)
- **DARIAH**: Digital Research Infrastructure for the Arts and Humanities

INTELE ikerketa-sarea



- Parte hartzaileak
 - 9 unibertsitate eta ikerketa zentro
 - EHU ([HiTZ](#)), UPF, UVigo, UNED, UCM, UAlcante ([BVC](#)), UJAEN, USC ([Instituto da Lingua Galega](#) y [CiTIUS](#)) eta CSIC ([Centro de Ciencias Humanas y Sociales](#), etc.)

- Babesa:
 - BNE, RAE, etc.
 - 35 unibertsitate baino gehiago eta 150 ikertalde eta ikerketa-sare baino gehiago

- Ekimenak:
 - 2020: Aurkezpen workshopa (online): 130 lagun baino gehiago
 - 2021: Nazioarteko web mintegiak (hilero)
 - 2021: Alor desberdinetako adituen workshopa
 - Hizkuntzalaritza, historia, zuzenbidea, liburutegiak, hezkuntza, etab.

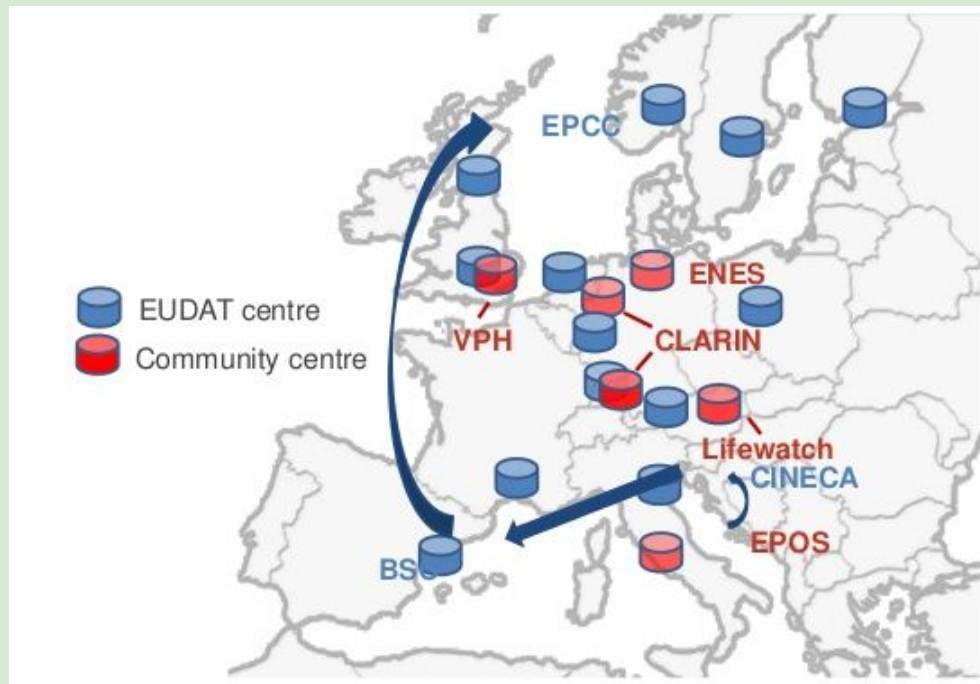


<https://ixa2.si.ehu.eus/intele/?q=home>

INTELE G1: Historia +

- **Teknika metodologiko kualitatiboetarako tresnak:**
 - Edukien analisia, testuena, diskurtsoarena, irudiena, kazetaritza testuen analisia, testu meatzaritza, sentimenduen analisia (sare sozialak), web scraping-a, ahozko iturriak, elkarrizketak, elkarrizketen transkripzioa, testuen sailkapena, dokumentuen etiketatzea, eztabaida-taldeak, etnografia, azterketa biografikoak
- **Teknika metodologiko kuantitatiboetarako tresnak:**
 - Inkestak, datu-baseak, estatistikak, testu corpusak, datuen meatzaritza (ikusentzunezko baliabideak eta baliabide digitalak), audientzien azterketa, big data, sare sozialak
- **Bistaratze eta lokalizazio tresnak:**
 - Informazio geografikoa: infografiak, mapa irudiztatuak, grafikak, mapa bizidunak, zuhaitz genealogikoak, arazo eta konponbideen zuhaitzak, sare sozialetatik eratorritako grafoak
- **Edukien biltegitzea:**
 - Iturriak: ikusentzunezkoak eta digitalak, datuen babesak eta ikerketaren kontserbazioa
- **Zerbitzuak:**
 - Formazioa, elkarreragingarritasuna, datu irekiak, itzultzaile automatikoak, konponbide informatikoen katalogoa
- **Datuak:**
 - Eduki irekiak, datuak deskargatzea modu malgu eta errazean

Europako ikerketa-ekosistema digitala



<https://www.slideshare.net/EUDAT/b2-safe-how-to-replicate-your-data>

Elkarreragingarritasuna!

SSHOC: Humanitate Digitaletako eta Gizarte Zientzietako e-azpiegituren konexioa



ESFRI - SSHOC



[CESSDA](#) ERIC - Consortium of European Social Science Data Archives

[CLARIN](#) ERIC - Common Language Resources and Technology Infrastructure

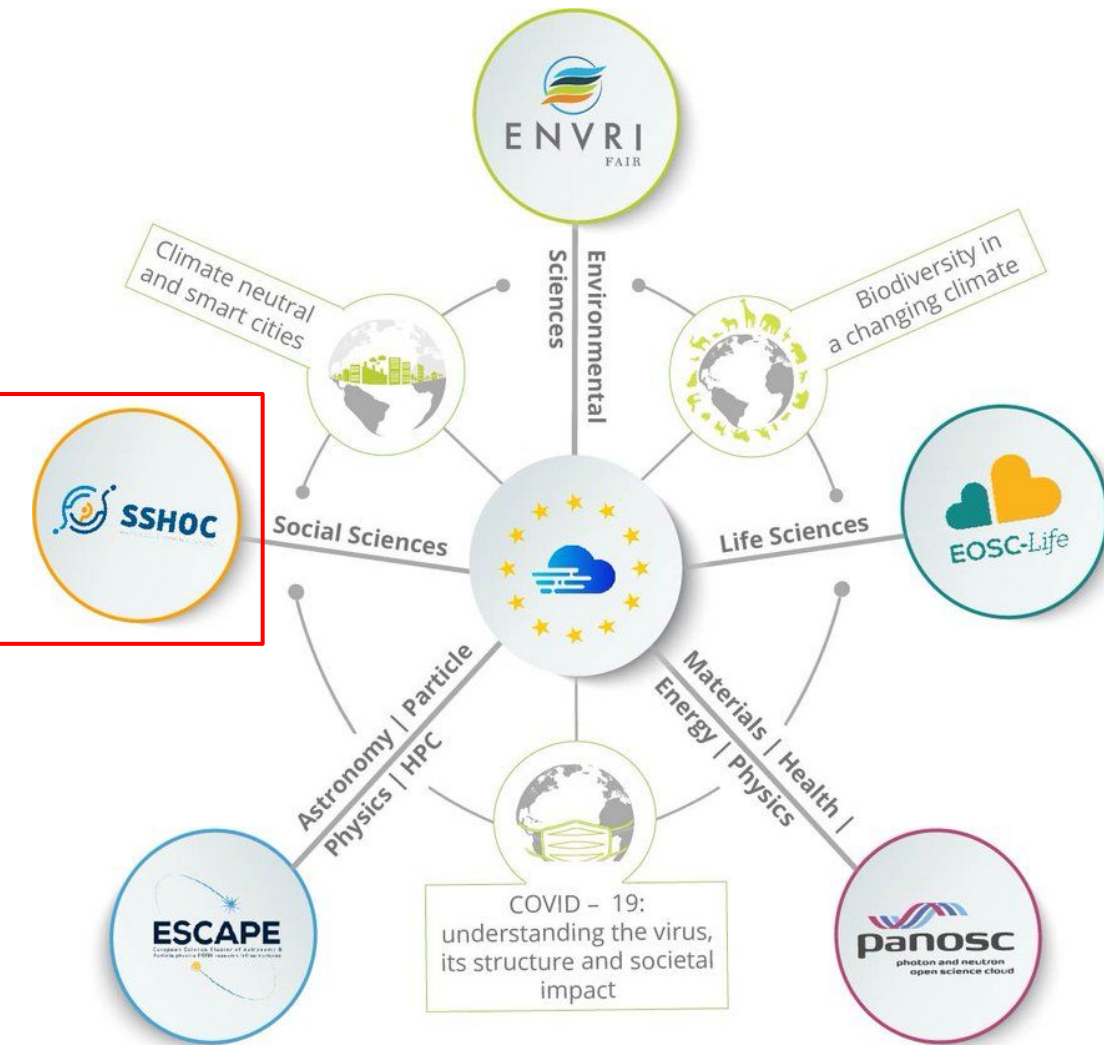
[DARIAH](#) ERIC - Digital Research Infrastructure for the Arts and Humanities

[ESS](#) ERIC - European Social Survey

[SHARE](#) ERIC - Survey of Health, Aging and Retirement in Europe

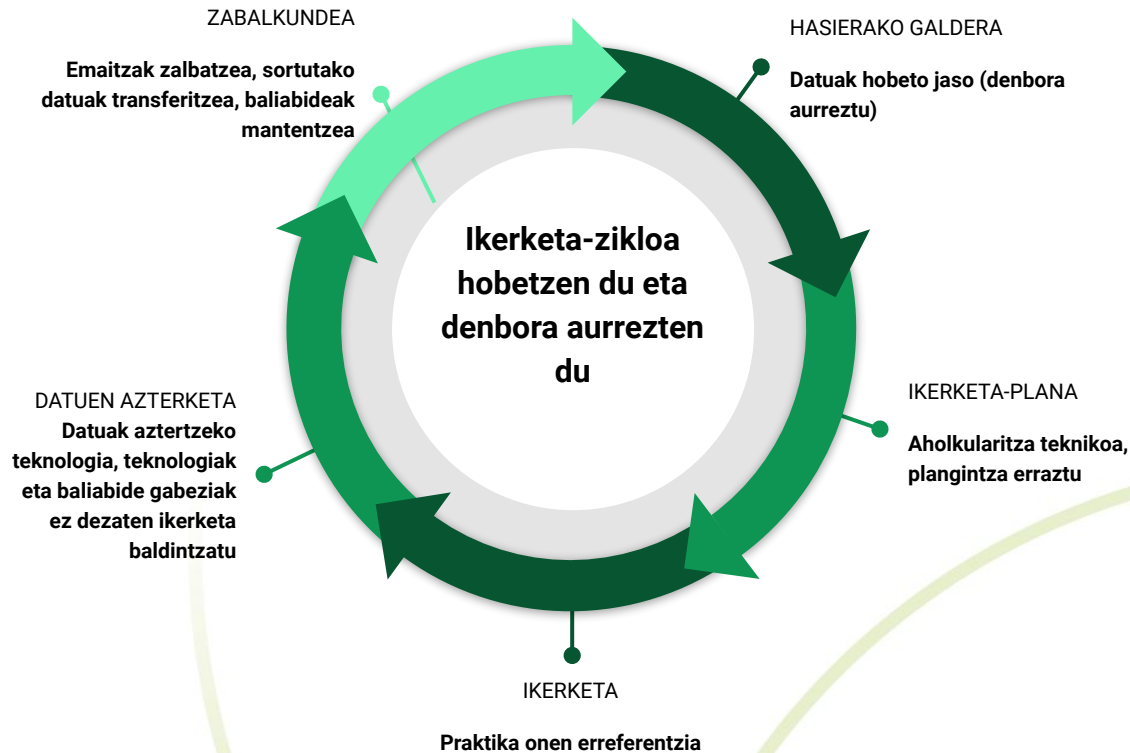
[E-RIHS](#) - European RI for Heritage Science - CSIC, CENIEH

EOSC azpiegituren mapa



[#SSHOC](#) Europako e-azpiegiturak **European Open Science Cloud** delakoan sendotzeko eta konektatzeko proiektua da.

Azpiegitura bat proiektu bat baino askoz gehiago da



- Sarbide konfederatua baliabide eta datu guztietara web-gune bakar batetik
- **Estandarrak**
- **Protokolo komunak**
- Paradigma aldaketarako laguntza
- Baliabide **estrategikoen** diseinua

- Corpusak: irekiak eta publikoak
- Irekiak akademiarentzat bakarrik
- Baimenduentzako bakarrik

PUB

AKA

AUT

Language Resource Switchboard

Zure hizkuntza datuentzat egokien den tresna aurkitzeko:

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

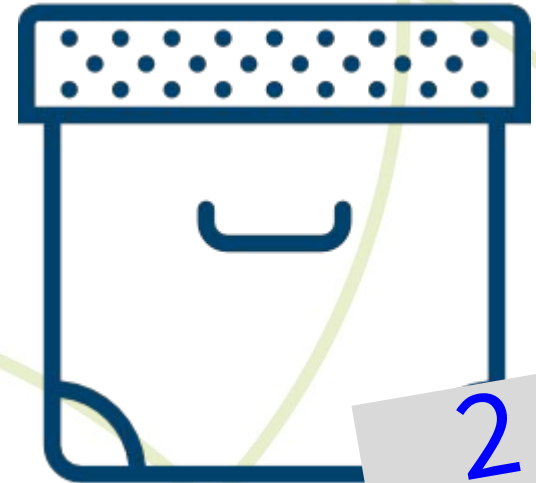


1

Depositing services

Corpusak eta baliabideak uzteko eta mantentzeko:

- www.clarin.eu/content/depositing-services



2

Language resources

Corpusak eta metadatuak:
kopuru handiak eta bilaketa
azkarrak:

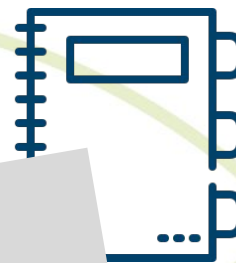
- vlo.clarin.eu/#tour
- contentsearch.clarin.eu
- <https://labur.eus/gZ1ld>



3



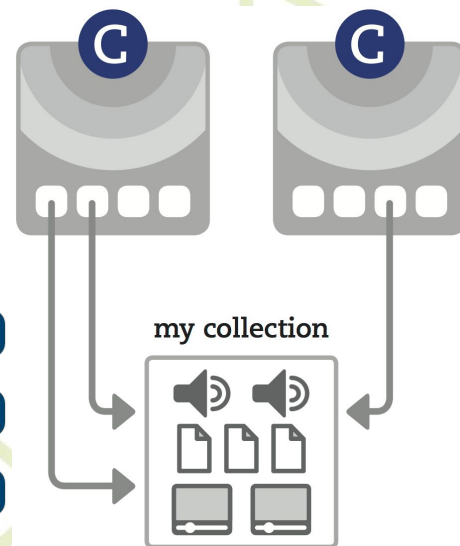
4



Virtual collections

Corpus birtualak sortu eta
horiek aipatu ahal izateko
(erreplikagarritasuna):

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

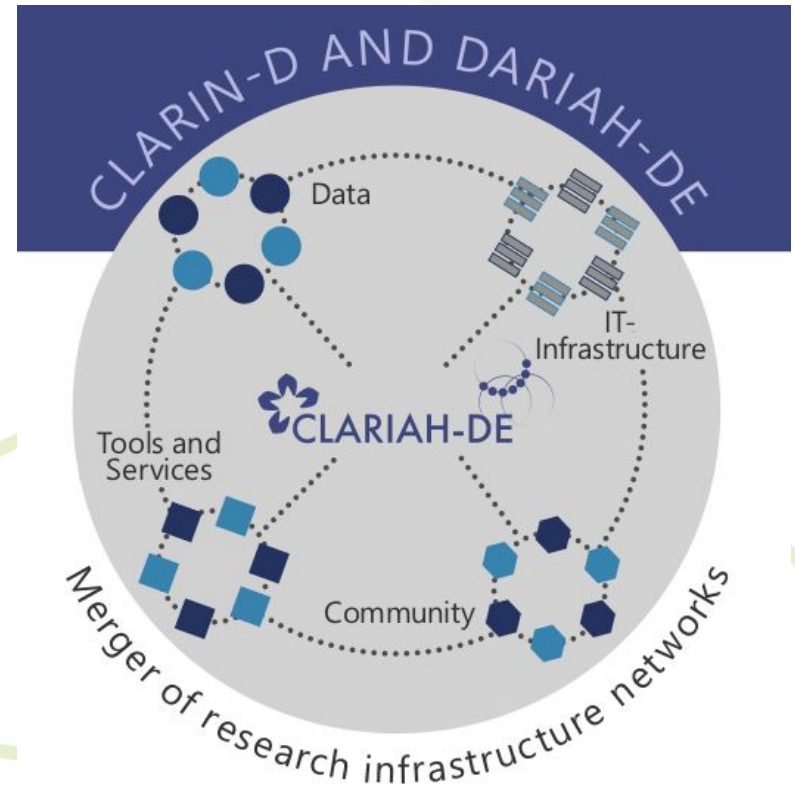


Azpiegitura eredua: CLARIAH (CLARIN + DARIAH)

THE NETWORK IN EUROPE



The CLARIN, DARIAH and CLARIAH research infrastructures are active on a national and the European level. CLARIN and DARIAH both have the status of a European Research Infrastructure Consortium (ERIC).



The content-related and technological foundations created by CLARIN-D and DARIAH-DE will be aligned, integrated, further developed and jointly maintained in CLARIAH-DE.

Hizkuntzen egoera CLARIN teknologian

<https://vlo.clarin.eu> milioi bat elementu baino gehiago daude (tresnak, corpusak...)

- Ingelesa (154.928)
- Alemana (143.271)
- Nederlandera (117.312)
- Daniera (109.962)
- Esloveniera (73.495)
- Poloniera (40.466)
- Frantsesa (24.432)
- Afrikaansa (7.870)
- ...

- Helburua eta konpromisoa
 - HaMABI
 - Garapen jasagarria

=

ALL LTs in ONE url

CLARIN-es hizkuntzetako baliabideak

- Gaztelania (14.444)
- Katalana (1.364)
- **Euskara** (498)
- Galegoa (216)

| DATA TYPE | |
|--------------------|-------|
| spoken | 99558 |
| speech | 4455 |
| writing | 1543 |
| gestures | 1307 |
| pointing-gestures | 454 |
| facial-expressions | 452 |
| emotional-state | 451 |

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE



4 QUALITY EDUCATION



8 DECENT WORK AND ECONOMIC GROWTH



16 PEACE, JUSTICE AND STRONG INSTITUTIONS



11 SUSTAINABLE CITIES AND COMMUNITIES



17 PARTNERSHIPS FOR THE GOALS



5 GENDER EQUALITY



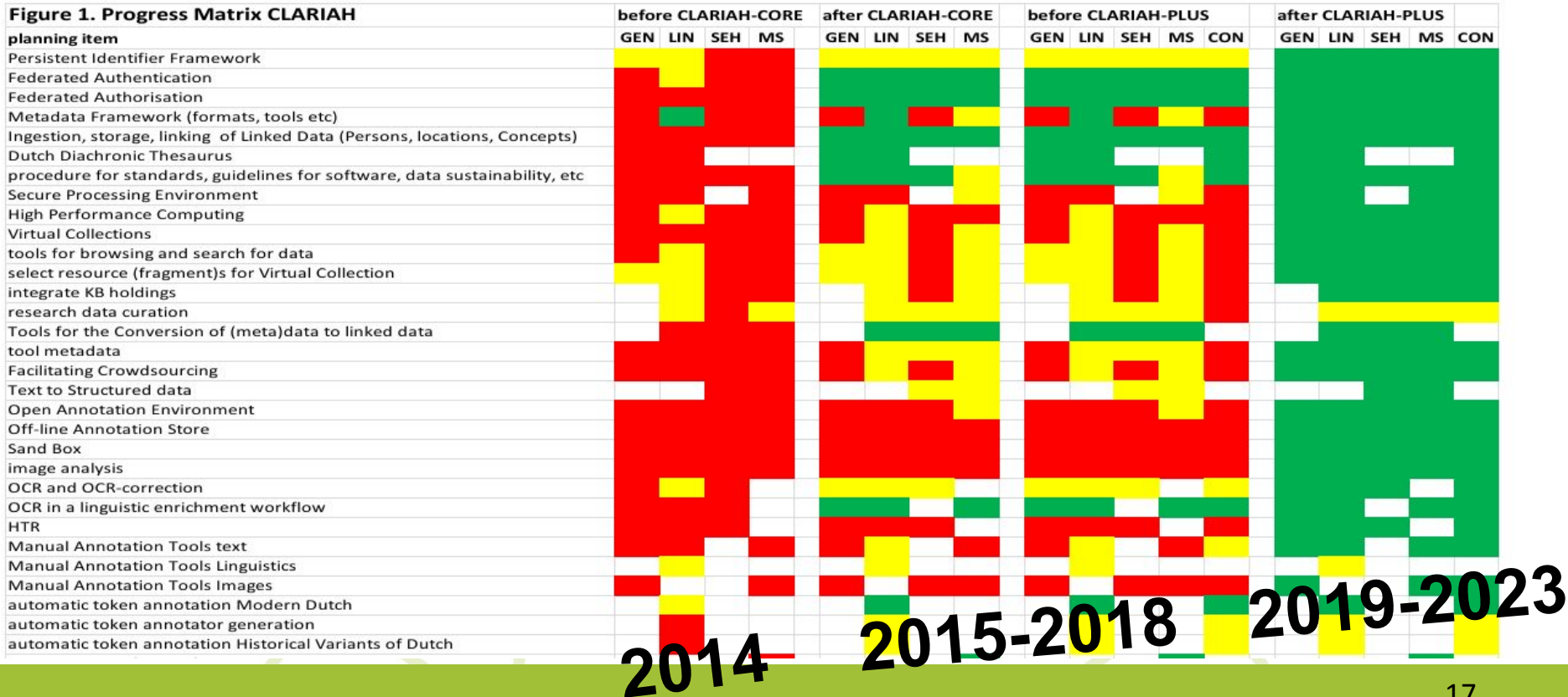
Nederlanderaren adibidea

National Roadmap for Large-Scale Research Infrastructure

1 General information

GENeric functionality, LINGuistics, Socio-Economic
 History and Media Studies, CONtent of texts: history, literary
 Iturria:
<https://www.clariah.nl/over/bestanden/downloads/send/10-folders/166-clariah-plus>

Figure 1. Progress Matrix CLARIAH



CLARIAH-EUS proposamena

- CLARIN eta DARIAH batzea
- Antolaketa eta finantziakoa
 - Unibertsitateak, ikerketa zentroak, erakundeak, etab.
 - Eusko Jaurlaritza, aldundiak, udalak, etab.
- Etorkizuneko CLARIAH-EUS azpiegituraren ekimenak
 - a) **Ikertzaileen koordinazioa eta zerbitzuak**
 - > 2 lagun (FTE)
 - b) **Zerbitzuak + Zabalkundea + Formazioa**
 - > Lagun 1 (FTE)
 - c) **Teknologia berriaren garapena**
 - > Lagun 1 (FTE), betiere garatu beharreko proiektu eta ekintzen arabera

CLARIAH-EUS: Bide-orria

- Proiektua aurkeztu
 - Unibertsitateetan, ikerketaldeen, ikerketa-zentroetan, euskal erakundeetan, instituzioetan, etab.
- CLARIAH-EUS partzuergoaren diseinua osatu
- **CLARIAH-EUS** CLARIN eta DARIAH azpiegituretan nodo proaktiboa izan

CLARIN-K zentroa eta bere nodoak

<http://clarin-es.org>

CLARIN
K CENTRE



[Centre](#) [Services](#) [Tools](#) [Materials](#) [Blog](#) [Private](#) [Contact](#)

Spanish CLARIN Centre-K

Spanish CLARIN Centre-K is a node of the European infrastructure CLARIN whose objective is to share knowledge and experience of the three founding constituent groups for research in humanities and social sciences.

We are the first Clarin Knowledge Centre and the representation of the CLARIN infrastructure in Spain.



1. IXA-EHU: <http://ixa2.si.ehu.eus/clarink/index.php?lang=es>
2. IULA-UPF: <http://www.clarin-cat-lab.org/index-en.html> & <http://services.iula.upf.edu/>
3. LINDH-UNED: <https://linhd.uned.es/clarin-centre-k/>
4. TALG-UVIGO: <http://sli.uvigo.gal/clarink/>

CLARIN: erabiltzaileen esperientziak

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

CLARIN through the eyes of the researchers

Tour de CLARIN 
Volume III



CLARIN-K zentroaren ad hoc zerbitzua

CLARINeko Tour delakoan argitaratua

- *Tesia Basque Center on Cognition, Brain and Language (BCBL)*
ikerketa-zentroan egiten du
- Gaia: “My PhD work focuses on the amount of exposure to each language within bilingual contexts, and how it shapes language acquisition at a cognitive and neural level”
 - **Tresnak:**
 - [ANALHITZA](#)
 - <https://switchboard.clarin.eu>

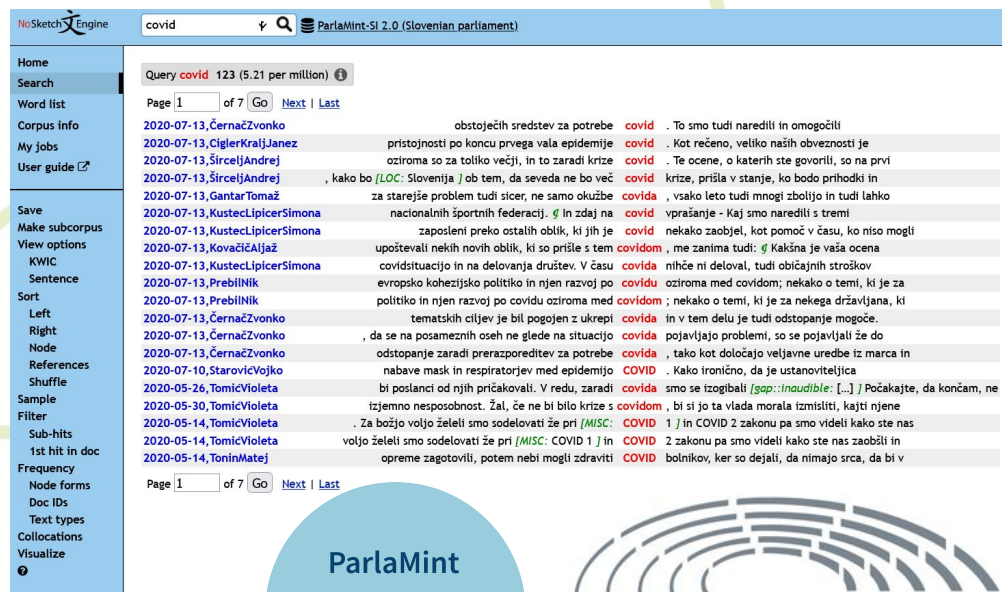


<https://www.clarin.eu/blog/tour-de-clarin-interview-jose-perez-navarro>

ParlaMint corpus eleanitza(testu idatzia)

1. Legebiltzarretako datuak eta metadatuak lortzea
2. ParlaMint eskemara bihurtzea
3. Anotazio linguistikoa (UDpipe eta NERC)
4. Bat datozen corpusen bidez corpus berri erabilgarriak sortzea (noSketch Engine / KonText) eta Parlameter

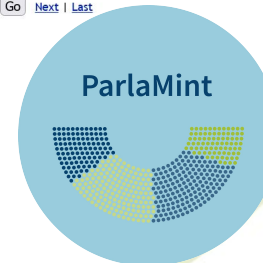
<https://www.clarin.si/noske/parlamint.cgi/>



The screenshot shows the ParlaMint search interface. The search query is 'covid', resulting in 123 hits (5.21 per million). The interface includes a sidebar with navigation options like Home, Search, Word list, Corpus info, My jobs, User guide, Save, Make subcorpus, View options, KWIC, Sentence, Sort, Left, Right, Node, References, Shuffle, Sample, Filter, Sub-hits, 1st hit in doc, Frequency, Node forms, Doc IDs, Text types, Collocations, and Visualize. The main content area displays a list of search results with dates, names, and snippets of text containing the word 'covid'.

| Date | Name | Snippet |
|------------|---------------------|--|
| 2020-07-13 | ČernačZvonko | obstoječih sredstev za potrebe covid . To smo tudi naredili in omogočili |
| 2020-07-13 | ČiglerKrajJanez | pristojnosti po koncu prvega vala epidemije covid . Kot rečeno, veliko naših obveznosti je |
| 2020-07-13 | ŠirčeljAndrej | oziroma so za toliko večji, in to zaradi krize covid . Te ocene, o katerih ste govorili, so na prvi |
| 2020-07-13 | ŠirčeljAndrej | , kako bo [LOC: Slovenija] ob tem, da seveda ne bo več covid krize, prišla v stanje, ko bodo prihodki in |
| 2020-07-13 | GantarTomaž | za starejše problem tudi sicer, ne samo okužbe covid , vsako leto tudi mnogi zbolijo in tudi lahko |
| 2020-07-13 | KustecLipicerSimona | nacionalnih športnih federacij. In zdaj covid vprašanje - Kaj smo naredili s tremi |
| 2020-07-13 | KustecLipicerSimona | zaposleni preko ostalih oblik, ki jih je covid nekoliko zaobjel, kot pomoč v času, ko niso mogli |
| 2020-07-13 | KovačičAljaž | upoštevali nekih novih oblik, ki so prišle s tem covidom , me zanima tudi: Kakšna je vaša ocena |
| 2020-07-13 | KustecLipicerSimona | covidstuaicjo in na delovanja društev. V času covidu nihče ni deloval, tudi običajnih stroškov |
| 2020-07-13 | PrebilNIK | evropsko kohezijsko politiko in njen razvoj po covidu oziroma med covidom; nekako o temi, ki je za |
| 2020-07-13 | PrebilNIK | politiko in njen razvoj po covidu oziroma med covidom ; nekako o temi, ki je za nekega državljana, ki |
| 2020-07-13 | ČernačZvonko | tematskih ciljev je bil pogojen z ukrepi covid in v tem delu je tudi odstopanje mogoče. |
| 2020-07-13 | ČernačZvonko | , da se na posameznih oseh ne glede na situacijo covid pojavljajo problemi, so se pojavljali že do |
| 2020-07-13 | ČernačZvonko | odstopanje zaradi prerazporeditev za potrebe covid , tako kot določajo veljavne uredbe iz marca in |
| 2020-07-10 | StaročVojko | nabave mask in respiratorjev med epidemijo covid . Kako ironično, da je ustanoviteljica |
| 2020-05-26 | TomičVioleta | bi poslanci od njih pričakovali. V redu, zaradi covid smo se izgubili [gop::inaudible: [...]] Počakajte, da končam, ne |
| 2020-05-30 | TomičVioleta | izjemno nesposobnost. Žal, če ne bi bilo krize s covidom , bi si jo ta vlada morala izmisliti, kajti njene |
| 2020-05-14 | TomičVioleta | Za božjo voljo želeli smo sodelovati že pri [MISC: COVID 1] in COVID 2 zakonu pa smo videli kako ste nas |
| 2020-05-14 | TomičVioleta | voljo želeli smo sodelovati že pri [MISC: COVID 1] in COVID 2 zakonu pa smo videli kako ste nas |
| 2020-05-14 | TorinMatej | opreme zagotovili, potem nebi mogli zdraviti covid bolnikov, ker so dejali, da nimajo srca, da bi v |

INTELEko web mintegian
ikerketagalderak:
<https://youtu.be/b0oNEIzBV9E>




Corpusa modu praktikoan erabiltzeko materiala

- 1 Introduction
- 2 Instructions for use
- 3 Corpora and concordancers
 - 3.1 Corpora
 - 3.2 Concordancers
- 4 Parliamentary records
 - 4.1 Parliamentary discourse
 - 4.2 Faithfulness of the records
 - 4.3 Know your research dataset
- 5 Language and gender
- 6 Corpus analysis
 - 6.1 The siParl 2.0 corpus
 - 6.2 TASK 1: Representation of women in the Slovenian Parliament
 - 6.2.1 Creating subcorpora
 - 6.2.2 Using frequency lists
 - 6.2.3 Comparative analysis
 - 6.3 TASK 2: Issues addressed by women
 - 6.3.1 Extracting keywords
 - 6.3.2 Analysing concordances
 - 6.3.3 Comparative analysis
 - 6.4 TASK 3: Topics related to women
 - 6.4.1 Working with frequencies
 - 6.4.2 Extracting collocations
 - 6.4.3 Comparative analysis

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



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

CLARIN-K IMPACT-CKC zentroaren erabilera

Interview | **Mikel Iruskieta**



Mikel Iruskieta is a computational linguist who is part of the Ixa Research Group and the Didactics of Language and Literature Department at the University of the Basque country. He has collaborated with the CLARIN IMPACT-CKC Knowledge Centre, which helped him and his colleagues digitize Basque texts.

Could you briefly describe your academic and research background?

My current research focuses on the didactics and analysis of Basque, mostly regarding discourse parsing and evaluation of discourse structure. For the last five years, I have mainly worked on adapting language technologies for teaching and learning purposes. With that goal, I have created and now co-lead a postgraduate programme in Basque (University Specialist in ICT and Digital Competences in Education, Continuing Education and Language Teaching), a research group working in Digital Humanities and Education. Our aim is to build a research community that will conduct research and teach in Basque by adopting a critical approach and using language technologies in a pedagogical context. In this postgraduate programme, my colleagues and I are developing a new framework of the socio-tech pedagogy for Basque that will cover the following topics:

<http://clarin-es.org/tour-de-clarin-vol-iii/>

- The Basics of Technology and Pedagogy;
- Formal Education and Technology;
- Continuing Education and Technology;
- Language Teaching and Technology Development;
- Society and Education, Opportunities and Risk of Technology;
- E-learning: Approaches and resources; and
- Digital Research: Methods and resources.

>

Does the fact that Basque is a language isolate have any bearing on the development of language tools tailored to it?

<

The history and current situation of the Basque language are both complex and interesting. Basque has a relatively small community of speakers (751,700 active and 1,185,500 passive speakers) which lives in contact with three powerful language communities, namely Spanish and French (as official languages in the Basque Country) and English (as a foreign language). It is also not supported enough by official language policies. As a result, Basque is still considered an under-resourced language. In this context, the work of the Ixa Group for NLP is highly valuable. They have developed basic resources for Basque (as well as for other languages) which are used by the research community, for example IXApipes (a modular set of NLP tools which provide easy access to NLP technology for several languages that can be used or exploit its modularity to pick and change different components) and ANALHITZA (a web service to analyse Basque, Spanish and English texts without needing any technical experience). Many more basic and advanced tools and resources for Basque can be found on the website of the HITZ: Basque Center for Language Technology.

>

How did you get involved with the IMPACT K-Centre and how did they help you with your research?

<

I learned about the IMPACT K-Centre when they joined CLARIN. Because I was working on several different digitization projects for Basque and for Spanish, I immediately got in touch with them and asked for their help. Isabel Martínez Sempere, the manager of IMPACT, helped me solve a digitization issue that I encountered when I was analysing the most frequently occurring words in *Pulgarcito*, which is a Cuban children's magazine written in Spanish from 1919 to 1920. This magazine consists of very diverse materials, such as drawings and handwritten texts, which are

IMPACT CLARIN K-centre eta BNE

Vida de Lazarillo de Tormes

 **IMPACT DATASET BROWSER**

This resource is property of

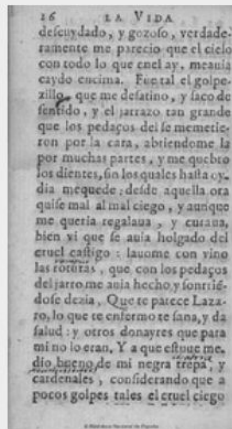


and distributed by the Impact Centre of Competence



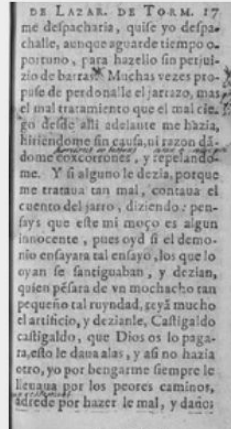
440435

[TIFF](#) [XML](#)



440436

[TIFF](#) [XML](#)



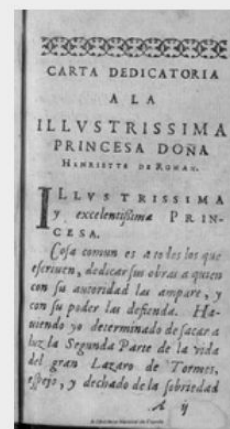
440437

[TIFF](#) [XML](#)



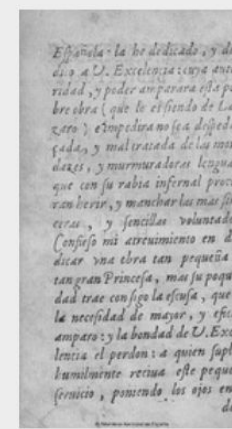
440438

[TIFF](#) [XML](#)



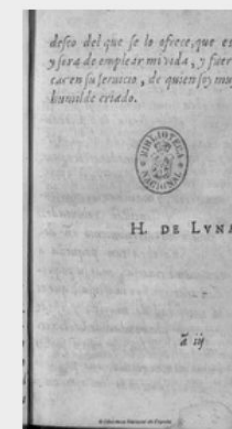
440439

[TIFF](#) [XML](#)



440440

[TIFF](#) [XML](#)

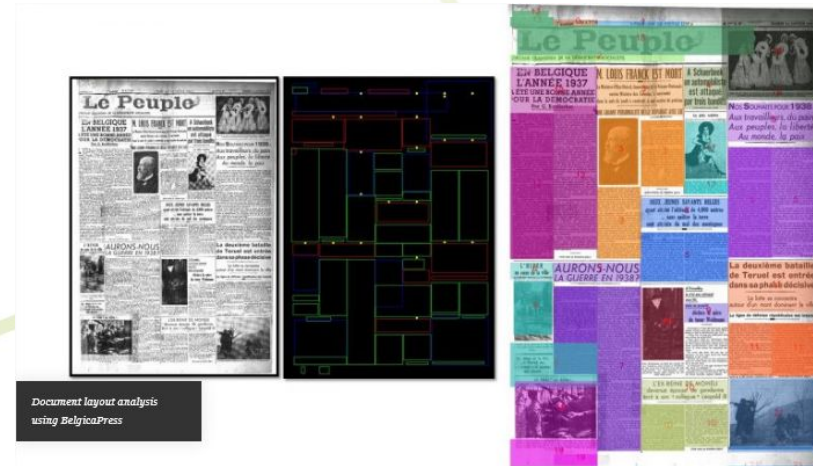


440441

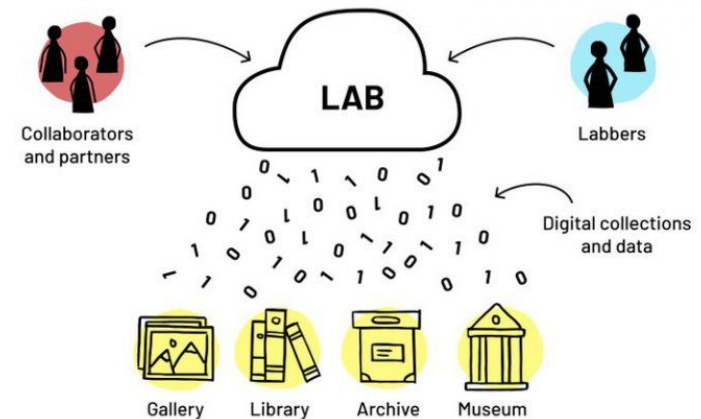
[TIFF](#) [XML](#)

DATA-KBR-BE: data as collection. DARIAH

- Datuak eskura jarri eta HDetan ikertzeko edizio digitalak sortu
 - Datu egokiak lortzeko lan-fluxua diseinatu
 - Open Data plataforma diseinatu
 - Bilduma digitalen inbentarioa
 - Datasetak argitaratu
 - Hackathon bat datasetak erabiliz



- 01
 - 01
 - alto
 - KB_JB840_1919-04-01_01-00001.xml
 - KB_JB840_1919-04-01_01-00002.xml
 - jpg
 - KB_JB840_1919-04-01_01-mets.xml
 - KB_JB840_1919-04-01_01.pdf
 - pdf
 - KB_JB840_1919-04-01_01-00001.pdf
 - KB_JB840_1919-04-01_01-00002.pdf
 - tif
 - KB_JB840_1919-04-01_01-00001.tif
 - KB_JB840_1919-04-01_01-00002.tif



Datuak bildumatzat. BVMC

CLARIN CENTRE K INTELE DARIAH-EU
Infraestructura de Tecnologías del Lenguaje

"Facilitando el acceso computacional a colecciones digitales"

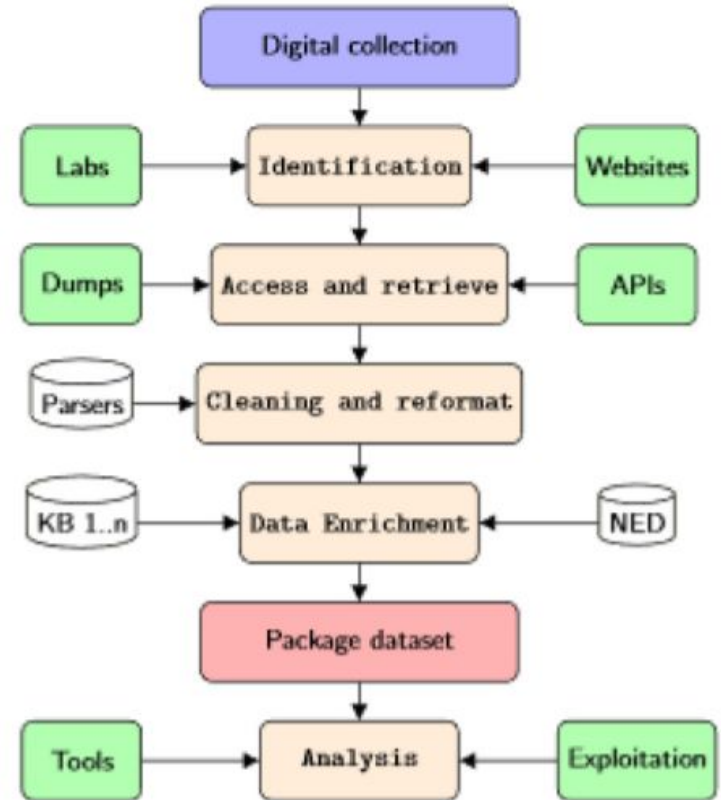
7 de junio de 2021
16:30h CET
(Online)

María Dolores Sáez
Gustavo Candela
María Pilar Escobar

Formulario de inscripción: http://ixa2.si.ehu.es/intele/form_enlace_webinar

INTELEko web mintegia: "Bilduma digitaletarako sarbide konputazionala erraztuz". (Biblioteca Virtual Miguel de Cervantes)

Saio praktikoa:
github.com/hibernator11/notebook-ph



SSHOC: Train-the-Trainer Bootcamp for Librarians:
<https://zenodo.org/record/3970799#.YMn2gqZ7mL1>



Erabilera-kasuak:
HaMABI printzipioekin lainoan
ikerketa errazten CLARIN eta DARIAH
azpiegituretan

Azpiegiturak erabiltzen

Digitalizaziorako tresnak

1. Eudat: datuak gorde eta partekatu >>> analizatu

Testu idatzia:

2. Eudat: FAIR data across borders and disciplines
3. Elkarreragingarria CLARINeko [Switchboard](#)arekin
4. Hizkuntza detekzioa eta tresna/metodoen aukera
5. Testua erraz aztertu eta bistaraatu/interaktuatu

Ahotsetik testura:

6. BAS zerbitzua CLARINen
7. Hamaika hizkuntza eta aldaera
8. Hamaika irteera formatu, ikertzen jarraitzeko: TXT, SCV, PRAAT, Video...

IMPACT beste CLARIN-K zentrua eta CLARINeko nodoen arteko lankidetzaz

IMPACT-CKC Helpdesk

CLARIN
K CENTRE



The [IMPACT centre of competence - CLARIN K-centre in digitisation](#) (IMPACT-CKC), as knowledge centre, offers its expertise and resources in digitisation and related fields to all institutions and particulars looking for advice. IMPACT resources include a [demonstrator platform](#) for testing online tools, a collection of [images and ground truth](#) associated, [historical lexica](#) for 10 languages among others.

For further information and advice on digitisation techniques, tools, language resources, image and [ground truth](#) collections, project proposals support, etc., please do not hesitate to contact us through our helpdesk by filling in the form below.

Your Name (required)

Your Email (required)

Basic information on data protection

- Responsible: IMPACT Centre of Competence in digitisation. Managed by Fundación Biblioteca Virtual Miguel de Cervantes (Postal address: Paseo de la Castellana, 103, 28046, Madrid, Spain. Email: info@digitisation.eu)

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



Title: Seis horas dentro de un taxi

Author: Andrés Carranque de Ríos

1

Transcription advancement 0%

2

Transcription advancement 0%

Transcribe the page automatically

This function creates an automatic transcription of the selected page

Recognition range



Overwrite the existing transcription page ▾

Recognition profile

- Latin ▾
- Polish
- English
- German
- French
- Latin
- Russian

Cancel

Tresna automatikoak eta eskuzko orrazketa

| No. | Page name | Number of rows | Verified | Correct | Incorrect | Last edition | Verification |
|-----|---|----------------|----------|---------|-----------|---|------------------------|
| 1 |  Seis_horas_dentro_de_un_taxi_Carranque_p.1.pdf | 87 | 0 | 0 | 0 | 2021-11-15 10:29 | Verify |
| 2 |  Seis_horas_dentro_de_un_taxi_Carranque_p.2.pdf | 87 | 0 | 0 | 0 | mikel.iruskieta@ehu.eus 2021-11-15 10:56 | Verify |

Fold

Transcribed object

Edit selection

Fold

Transcription



El encuentro no ha sido fuerte pero el conductor del otro auto...
documento por medio de un libro, y en caso de mirar, por temor a que...
donde se nos queda el micro, dispuesto a dudar, a realizar una...
que carabinero. La dama lo ve partir llena de satisfacción, y una...
Dijiste de Alta, número cuatro

Los inconvenientes que trae el encontrarse en un municipio...
Sufre los siete de la noche, cuando he sido elab la atención de que...
de un modo...
verdad, el taxi ha tenido que...
estar en un estado de...
de los bigotes debe ser un Landrö.

1 .

2 tomóvil se apea de su coche y empieza a gritar. Esto mismo

3 realiza | hombre joven, con aire de estudiante. Le pego con el codo

4 al conduc-

5 mi «jefe», y hasta el grupo de curiosos que comienza a rodearnos.

6 Por tor, y el auto que.la frenado. La dama. nos dice desde el

7 interior:

8 fin se arregla todo buenamente y regresamos al taxi. Entoaces, el

– "Vaya hacia la plaza de Santa Ana... No corra.

asombro so pinta en. nuestras caras. Resulta que cel señor vencia-

Pasada la plaza de Antón Martín noto que el chófer mo hace sc-

ble y la jovencita no aparecen por ningún lado. nas de que mire el

retrovisor. Observo por el espejito, y «o que la

–¡Atiza!– exclama el chófer–. Aquí había algún lío.. El viejo –

dama entrega un billete al adolescente, Después se efe t; i1 agia-

de los bigotes debe ser un Landrö. BOB - yep ni xi cERdiacct eras



SESS *

Monitorizaziorako tresnak

Title: 📄 Seis horas dentro de un taxi

Author: Andrés Carranque de Ríos

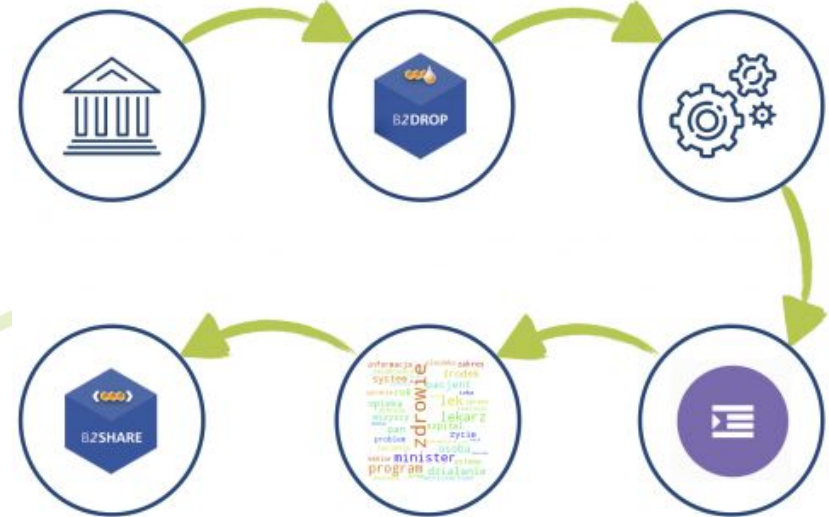
Advanced verification: Pages to verify at this property: 1 Verified rows: 87 / 174 Correct: 83 ✓ Incorrect: 4 ✗

| No. | Page name | Number of rows | Verified | Correct | Incorrect | Last edition | Verification |
|-----|---|----------------|----------|---------|-----------|--|------------------------|
| 1 |  Seis_horas_dentro_de_un_taxi_Carranque_p.1.pdf | 87 | 0 | 0 | 0 | 2021-11-15 10:29 | Verify |
| 2 |  Seis_horas_dentro_de_un_taxi_Carranque_p.2.pdf | 87 | 87 | 83 | 4 | mikel.iruskiet@ehu.eus 2021-11-15 10:59 | Verify |

Erabilera-kasua CLARINen: HaMABI printzipioak

EuDAT azpiegitura erabiliz

- Datuak:
 - “La edad de plata” por Jose Calvo Tello
 - Haurren corpusak
- Analisi sintaktikoa
- Analisia lainoan
- Argitalpen iraunkorrae



Interoperable

Originala ingelesez:

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



CLARINeko bilduma birtual multimodala



Virtual Collection Registry

Browse

Create

My Collections

Help



mikel.iruskieta_ehu.eus@clarin.eu

CLARIN



Euskarazko haurren corpusak

General

| | |
|------------------|-----------------------------|
| Name: | Euskarazko haurren corpusak |
| Type: | EXTENSIONAL |
| Creation date: | 2021-06-29 |
| Description: | Haurren euskarazko corpora |
| Purpose: | REFERENCE |
| Reproducibility: | INTENDED |
| Keywords: | • Haurrak |

Resources

Reference

[Frogs French Iduguine Corpus](#)

[Basque SotoValle Corpus - 040505](#)

[Basque Luque Corpus - 33cas3](#)

[Haur Hezkuntzako ipuin-bilduma](#)

[HDL 11304/f27f5e92-af01-4a37-a6d9-82cf14afa160](#)

Idatzizko testurako baliabideak lainoan



1. Testua jaitsi: Gitzip
2. Gorde corpora Eudaten
3. Aztertu [Switchboard](#)ekin
4. Aukeratu tresna bat aurrera egieko
5. ...

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

- Baliabideak
 - Euskararako: 3
 - Gaztelerarako: 4
 - Ingeleserako: 16
 - Alemanierarako: 13
 - Polonierarako: 26



Testu baten analisi sintaktikoa eta bistaratzea



↓ 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 á que la incierta luz de la ma-ñaña entre en hilos de claridad por las hendiduras de la puerta que da al campo , uno do Jos gatos del cortijo está en perspicaz acecho , con las dos manos estendijas hacia adelante , y la cabeza algo agachada , lo mismo que ai se hallara á la vista de algún fugitivo ratón .

| | | |
|-------------------------------------|------------------------|---|
| <input checked="" type="checkbox"/> | Hide empty attributes | x |
| deprel | nmod | |
| feats | Gender=Fem Number=Sing | |
| form | puerta | |
| head | 17 | |
| id | 20 | |
| lemma | puerta | |
| misc | TokenRange=109:115 | |
| upostag | NOUN | |
| xpostag | NOUN | |

Klik eginez, analisi-katea eraikiz



Main Page Chain 1 x + New Chain

Show tools with status: development production staging superseded withdrawn

Next Choices (Double-click on an icon to add it to the chain)

CLAR: TextCorpus2Lexicon
Language: Spanish
Document Type: Lexicon Forma
TCF Version: 0.4
entries.type: types

< >

i

Processing...

Input and Chain Selection

| Title [Plain Text] | SfS: To TCF Converter | SfS: BlingFire Tokenizer | SfS: OpenNLP Named Entity | SfS: Geolocation |
|--|---|--------------------------|---------------------------|--|
| Esperando á que la incierta luz de la ma-ñaña entre en bilos de claridad por las | TCF Version: 5 Language: Spanish Document Type: TCF Text | Sentences Tokens | Named Entities: OpenNLP | Geo - Capitals: Name Geo - Continents: Name Geo - Coordinates: Decimal Deg Geo - Countries: 2-Letter Countr |
| | | | | |

Distant Reading for European Literary

| Language | Last update | Texts | Words | AUTHORSHIP | | | | LENGTH | | | TIME SLOT | | |
|---------------------|-------------|-------|----------|------------|--------|---------|---------|--------|--------|------|-----------|---------|---------|
| | | | | Male | Female | 1-title | 3-title | Short | Medium | Long | 1840-59 | 1860-79 | 1880-99 |
| cze | 2021-04-09 | 100 | 5621667 | 88 | 12 | 62 | 6 | 43 | 49 | 8 | 12 | 21 | 39 |
| deu | 2021-04-11 | 100 | 12738842 | 67 | 33 | 35 | 9 | 20 | 37 | 43 | 25 | 25 | 25 |
| eng | 2021-04-09 | 100 | 12227703 | 49 | 51 | 70 | 10 | 27 | 27 | 46 | 21 | 22 | 31 |
| fra | 2021-04-09 | 100 | 8712219 | 66 | 34 | 58 | 10 | 32 | 38 | 30 | 25 | 25 | 25 |
| gsw | 2021-06-07 | 38 | 2392060 | 21 | 17 | 13 | 5 | 11 | 22 | 5 | 0 | 1 | 13 |
| hrv | 2021-03-22 | 21 | 1440018 | 21 | 0 | 4 | 0 | 6 | 12 | 3 | 6 | 12 | 2 |
| hun | 2021-04-09 | 100 | 6948590 | 79 | 21 | 71 | 9 | 47 | 31 | 22 | 22 | 21 | 27 |
| ita | 2019-11-21 | 34 | 3328244 | 32 | 2 | 19 | 3 | 13 | 10 | 11 | 5 | 12 | 10 |

Distant Reading for European Literary History

11 de junio de 2021
14:00h CET
(Online)

Rosario Arias

Borja Navarro

Christof Schöch

CLARIN CENTRE K INTELE DARIAH-EU
Infraestructura de Tecnologías del Lenguaje

INTELEko workshopa: "Distant Reading for European Literary History"

Saio praktikoa:
github.com/bncolorado/Processing-ELTeC-corp-us

TEI+XML fitxategiaren analisis 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](#) [Summary](#) [Documents](#) [Phrases](#) [Contexts](#) [Bubblelines](#) [Correlations](#)

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

El duende de la corte

o

Memorias de un fraile

Novela histórica original

de

Relative Frequencies

Document Segments (El duende de la corte...)

a más me mi no

Terms:

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

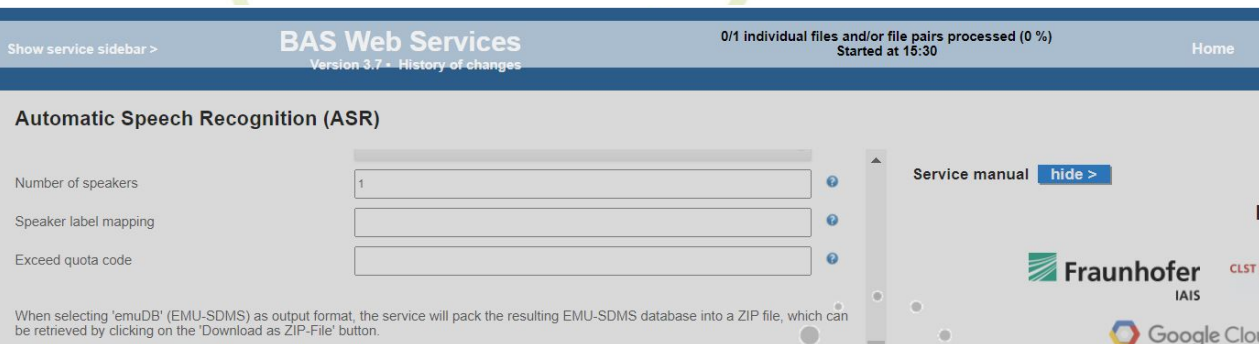
Voyant Tools · Stéfan Sinclair & Geoffrey Rockwell (© 2021) Privacy v. 2.4 (M55)

Ahotserako baliabideak lainoan

1. Deskargatu [EJko bideoa](#)
2. Transkribatu klik batekin [BAS](#)etik
3. Emaitzak aztertu



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



Euskera-gaztelania transkribapena



EU ES



Prefer

Albisteak eta ekitaldiak · Ekitaldiak eta gertaerak

2020 ira 14

EUSKO LEGEBILTZARRAREN 40. URTEURRE

Eusko Legebiltzarrak Euskal Herriko Unibertsitatearen (UPV-EHU) udako ik

Lekua MIRAMAR JAUREGIA

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

Ordua: 10:00 -18:00



urteurrena
aniversario
1980 - 2020

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

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



ITURRIA:

<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

OH Portal (CLARIN)

<https://oralhistory.eu/oh-portal>

0 1 0 0 0 0

Help Statistics Feedback

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



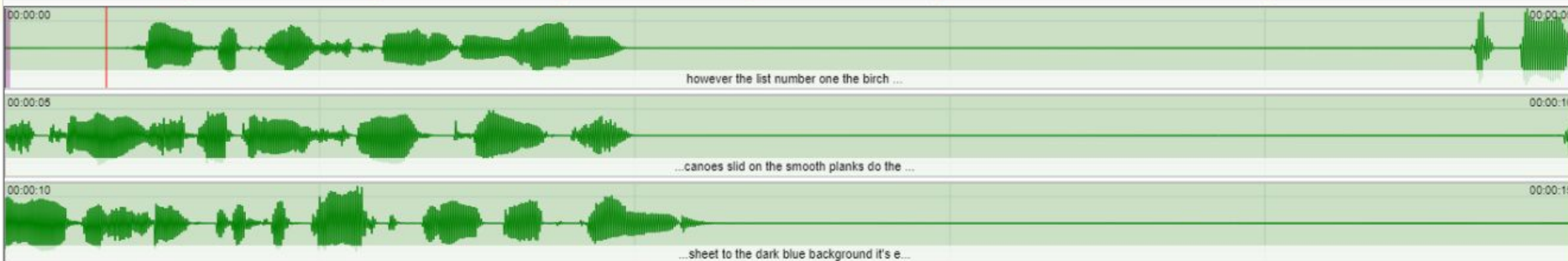
OCTRA v1.4.3 (url) Dictaphone Editor Linear Editor 2D-Editor

TRN Werkzeuge Exportieren DE

TASTENKOMBINATIONEN [ALT + 8]

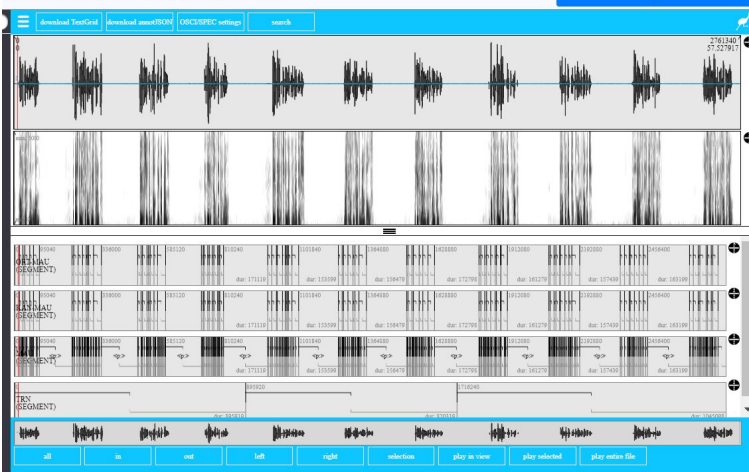
ÜBERSICHT [ALT + 0]

HILFE



0 1 0 0 0 0 Help Statistics Feedback

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



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

| | Poloniera | Alemaniera | Inglesa | Gaztelania | Euskara |
|-------------------------------------|-----------|------------|---------|------------|---------|
| 1. Constituency Parsing | | x | x | | |
| 2. Coreference Resolution | x | | | | |
| 3. Dependency Parsing | x | x | x | x | x |
| 4. Distant Reading | x | x | x | x | x |
| 5. Extraction of Polish terminology | x | | | | |
| 6. Inclusion detection | x | | | | |
| 7. Keyword Extractor | x | | | | |
| 8. Lemmatization | | x | x | | |
| 9. Machine Translation | | x | x | | |
| 10. Metadata Processing | | | | | |
| 11. Morpho-syntactic tagger | x | | x | | |
| 12. Morphological Analysis | x | x | x | | |
| 13. Named Entity Recognition | x | x | x | x | |
| 14. Named Entity Relation Detection | | | | | |
| 15. Part-Of-Speech Tagging | x | x | x | | |
| 16. Sentiment Analysis | x | | | | |
| 17. Shallow Parsing | x | | | | |
| 18. Spatial expression detection | x | | | | |
| 19. Speech Recognition | | | | | |

CLARIAH-EUS

MERSI **danke** 謝謝 **Spas**
 Баярлалаа **teşekkür ederim**
 спасибо **dank je** misaotra matondo **gracias** tapadh leat
 taafetai lava vinaka blagodarari **gracias** хвала
 nanni nandri kiitos dankie **ESKERRIK ASKO** tapadh leat asante manana
 dhanyavad **gracias** obriigada azokrate tenki
 bayarlalaa **gracie** **gracias** djiere dieuf lau mochchakkeram
bedankt **dziękuje** cnorakaloutioun **gracias** **agat**
 enkosi **obrigado** sobodi dekuji **sukriya** kop khun krap **taiku** **go raibh maith agat**
 didi madloba **sagolun** najis tuke **sukriya** **arigatō** **takk** **dakujem** trugarez
 kam sah hamnida **rahmat** **terima kasih** tanemirt rahmet **grazie** **arigatō** **takk** **dakujem** trugarez
 তোমাকে ধন্যবাদ **rahmat** **terima kasih** **merci** **merci** **merci** **merci**
THANK YOU **merci** **merci** **merci** **merci**
 감사합니다 xixie **merci** **merci** **merci** **merci**
شكريا
 Schukria

Erreferentziak

Iruskieta, M. (2019) [CLARIN Europako sarea: eHumanitateak eta zientzia sozialak lankidetzarako behar duten hizkuntza-azpiegitura sortzen.](#) Humanitate digitalak: aukerak, erakundeen rol berriak eta elkarlana. UEU. 2019ko ekainaren 20a. Bilbo. ULR: <https://www.youtube.com/watch?v=EEAwzxPL4GA&t=1346s>

Krauwert, 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., Krauwert, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Bel, N. Gonzalez-Blanco, E. Iruskieta, M. (2016). [CLARIN Centro-K-español.](#) Procesamiento del Lenguaje Natural 57: 151-154. ISSN: 1135-5948.

CLARIN: [https://www.clarin.eu/](https://www.clarin.eu)

DARIAH: [https://www.dariah.eu/](https://www.dariah.eu)

INTELE: <http://ixa2.si.ehu.es/intele/>