CLARIAH-EUS: A Strategic Network Helping Basque Country Researchers to Participate in European Research Infrastructures

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna, Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga, Jose Mari Arriola
HiTZ Center - Ixa
University of the Basque Country, Spain
jon.alkorta,aritz.farwell,joseba.fernandezdelanda,begona.altuna, ainara.estarrona,mikel.iruskieta,xabier.arregi,xabier.goenaga, josemaria.arriola@ehu.eus

Inma Hernáez
HiTZ Center - Aholab
University of the Basque Country
Spain
inma.hernaez@ehu.eus

David Lindemann
Diachronic Linguistics, Typology, and the History of Basque Research Group
University of the Basque Country
Spain
david.lindemann@ehu.eus

Abstract

CLARIAH-EUS is a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN and DARIAH, Europe's leading digital research infrastructures for the humanities, arts, and social sciences. Focused on Basque or Basque culture-related research in these fields, CLARIAH-EUS offers scholars digital tools and resources. Distinct from other nodes, CLARIAH-EUS serves a language (Basque) rather than a specific territory, making the infrastructure transnational. This article outlines the rationale behind establishing CLARIAH-EUS, its development process, ongoing projects, and future plans.

1 Introduction

Academic research in fields driven by technology is often characterized by rapid change and the constant application of new methods. Disciplines traditionally less dependent on technology, however, have historically experienced more measured technological integration. Although it may be argued that this has generally been true for the humanities, arts, and social sciences, a "digital turn" over the past two decades is driving new modes of research and lines of inquiry in these areas. Digital tools, methods, and data are casting new light on complex social patterns and providing innovative techniques to interpret cultural heritage (Crawford et al., 2014; Terras, 2011). Moreover, the rise of digital humanities has blurred the boundaries between disciplines, spurring interdisciplinary collaborations in research, teaching, and publishing (Burdick et al., 2016).

Language technology, tools, and resources tailored specifically for the social sciences and humanities play a pivotal role in this pioneering work. Yet, much of this technological support is designed for use with English, creating an imbalance between techniques that can be applied when conducting English-language research and those that are available for research in other languages. This is especially true for languages spoken by smaller populations (Arzoz, 2015). Basque, one of these languages, has fortunately made significant strides in language technology due to deliberate efforts to foster the sociolinguistic conditions necessary for its successful development and dissemination. This includes sustained and proactive collaboration between research groups, foundations, industry clusters, and regional institutions (Gonzalez-Dios & Altuna, 2022; Sarasola et al., 2023). Nevertheless, Basque still faces challenges in terms of research maturity and availability of the wide-ranging resources needed to fully support social science and humanities projects.

The CLARIAH-EUS consortium was established to overcome these existing limitations. On the one hand, it seeks to encourage the use of language technology among researchers in Basque-related humanities, arts, and social sciences. On the other, it strives to strengthen and facilitate collaboration between these researchers, enabling them to share ideas and innovative approaches more effectively. To do so more effectively, CLARIAH-EUS took the strategic decision to orient itself towards language rather than geographical boundaries, making it transnational in scope. Furthermore, as a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN ERIC and DARIAH ERIC, CLARIAH-EUS is aligned with Europe's leading digital research platforms for the humanities, arts, and social sciences.

2 Objectives

As highlighted above, one of CLARIAH-EUS's objectives is to support language technology for Basque humanities, arts, and social sciences research. This effort translates into two key areas. The first is to build a repository that contains digital resources specifically for Basque. These resources will be integrated into the wider CLARIN and DARIAH infrastructures, ensuring that the Basque-focused tools that are developed become readily accessible to researchers. The second is to empower researchers by offering them dedicated services and training. We plan to provide users who are creating or utilizing Basque language technology for their projects with the resources to work autonomously in the digital domain.

By cultivating these two areas, CLARIAH-EUS intends to foment a vibrant research community that is dedicated to advancing Basque language technology for the humanities, arts, and social sciences. This focus on collaboration is designed to 1) open doors to greater participation in international projects by leveraging shared expertise to create more impactful outcomes and 2) nourish an environment that sparks groundbreaking approaches to Basque digital humanities and language technology.

3 Institutional Funding

CLARIAH-EUS prioritizes securing financial backing to ensure the viability of research initiatives across the short-, medium-, and long-term. This approach guarantees the ongoing usability and value of any resources that are created. The focus on sustainability has resonated with several public funding bodies, who have pledged support to the infrastructure. Currently, CLARIAH-EUS is supported by the Basque Government through its Department of Culture and Linguistic Policy,[1] the Provincial Council of Gipuzkoa,[2] and the University of the Basque Country (UPV/EHU). Backing by the UPV/EHU comes from the Vice-Rectorate of Basque, Culture and Internationalization,[3] and from HiTZ, the Basque Center for Language Technology.[4] HiTZ, in addition to providing financial support, also houses the infrastructure's administrative office. Furthermore, several of its members sit on the CLARIAH-EUS steering committee, contributing their guidance and expertise. Thanks to the support of these institutions, CLARIAH-EUS has assembled a team of four staff members, who play a crucial role in ensuring the smooth operation of both the CLARIAH-EUS infrastructure and the shared administrative office with CLARIAH-ES (also overseen by HiTZ).

4 Origins and Growth

To date, the evolution of CLARIAH-EUS has included a design phase (2021-2023) (see sections 4.1 and 4.2) and an implementation and consolidation phase (2023-present) (see sections 4.3 and 4.4).

4.1 First Workshop: Needs and Manifesto

CLARIAH-EUS's first workshop,[5] *Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen* (*Designing the CLARIAH-EUS infrastruc-*

[1] https://www.euskadi.eus/eusko-jaurlaritza/kultura-hizkuntza-politika-saila/

[2] https://www.gipuzkoa.eus/eu/

[3] https://www.ehu.eus/eu/web/nazioarteko-harremanak

[4] https://www.hitz.eus/eu

[5] https://www.clariah.eus/eu/1-workshop

ture to develop language technology for Basque in the humanities and social sciences) was organized by HiTZ on November 26, 2021 and laid the foundation for the future infrastructure.

The workshop aimed to foster discussion about opportunities and needs across various research areas related to Basque language and culture. It featured several activities: 1) the presentation of a collection of use cases and posters depicting digital projects relevant to Basque studies, which provided a platform for researchers to share their work; 2) collaborative breakout sessions focused on identifying the strategic resources most crucial for Basque research across different disciplines; and 3) engaging and building bridges between researchers that encouraged active participation in CLARIAH-EUS's future.

The event drew participants from nine institutions and thirty-four researchers representing twenty distinct research groups. Fourteen projects were presented and 134 organizations and individuals signed a manifesto.[6] This collective voice underscored the widespread demand for a dedicated digital humanities infrastructure for Basque research.

4.2 Assembling the Network

Between 2021 and 2023, CLARIAH-EUS's goal was to pursue the backing of several organizations and research groups. This was procured from seventeen entities: HiTZ (UPV/EHU), Basque Summer University (UEU),[7] Iker research group,[8] Elhuyar,[9] Gogo Elebiduna research group (UPV/EHU),[10] Elebilab research group (UPV/EHU),[11] Aholab research group (UPV/EHU),[12] Ixa research group (UPV/EHU),[13] Soziolinguistika Klusterra,[14] Diachronic Linguistics, Typology and the History of Basque (DLTB) research group (UPV/EHU),[15] Basque Research Group of Theoretical Linguistics (HiTT) (UPV/EHU),[16] Badalab,[17] Gizapedia,[18] Tralima-Itzulik research group (UPV/EHU),[19] General Directorate of Linguistic Equality (Prov. Council of Gipuzkoa),[20] the Public University of Navarre (UPNA),[21] and the UNESCO Chair in Human Rights and Public Powers (UPV/EHU).[22] Sixteen of these institutions and research groups are based in the southern Basque Country and one (Iker) is from the northern Basque Country (see Figure 1). During this time, CLARIAH-EUS's position within CLARIAH-ES in Spain, and CLARIN and DARIAH at the European level, was further solidified.

4.3 Second Workshop: Community and Organization

CLARIAH-EUS held its second workshop[23] in November 2023. We presented the CLARIAH-EUS infrastructure and existing Basque digital humanities projects. The workshop marked a significant milestone in the form of a kickoff ceremony for the founding members. The focus shifted from initial brainstorming to outlining the infrastructure's future. Key discussions centered on CLARIAH-EUS's organizational structure and a road map, which put an emphasis on strategic directions for the next five years. Two invited speakers shared their expertise and twenty-one research groups presented posters. A selection of these, along with details about the participating research groups, will be featured in a forthcoming publication, offering a valuable glimpse into the Basque digital humanities landscape.[24]

[6] https://www.clariah.eus/eu/manifestua

[7] https://www.ueu.eus/

[8] https://iker.cnrs.fr/?lang=en

[9] https://www.elhuyar.eus/en

[10] https://www.ehu.eus/HEB/

[11] https://www.ehu.eus/en/web/elebilab/

[12] https://aholab.ehu.eus/aholab/

[13] http://ixa.ehu.eus/

[14] https://soziolinguistika.eus/en/

[15] https://ekoizpen-zientifikoa.ehu.eus/grupos/24732/detalle?lang=en

[16] http://www.hittlinguistics.eus/en/

[17] https://badalab.eus/

[18] https://gizapedia.org/

[19] https://www.ehu.eus/en/web/tralimaitzulik/home

[20] https://www.gipuzkoa.eus/web/council

[21] https://www.unavarra.es/home

[22] http://katedraddhh.eus/en/

[23] https://www.donostiakultura.eus/eu/ikastaroak/clariah-eus-euskararako-ikerketa-azpiegitura-eraikitzen

[24] https://www.clariah.eus/eu/2-workshopa-azpiegitura-eraikitzen

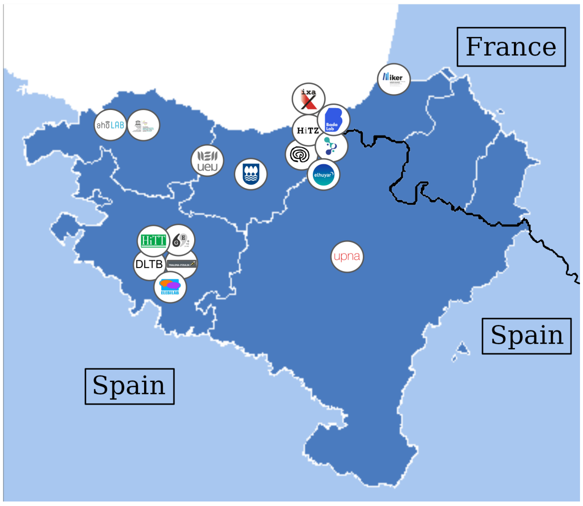


Figure 1: The CLARIAH-EUS Consortium

4.4 Third Workshop: A Vibrant Community, a Practical Infrastructure

For its third workshop,[25] which took place in November 2024, CLARIAH-EUS focused on strengthening the ties between those involved in local digital humanities projects and on highlighting ways the infrastructure can provide practical support. The workshop opened with a talk by invited speaker Mikko Tolonen, who shared Finland's experience in erecting FIN-CLARIAH and discussed how humanities-driven AI in the social sciences and humanities might be shaped. The event also included a roundtable discussion on technology and the social sciences, an introduction to CLARIAH-EUS's new B-centre, two presentations on CLARIN and DARIAH, a hands-on demonstration of Basque LLMs, and a poster session dedicated to digital humanities research currently being carried out in the Basque Country.

5 Projects, Tools, and Resources

As previously underscored, a core objective of CLARIAH-EUS is to empower researchers with the tools and resources[26] they need to excel in the digital humanities and social sciences. These resources fall into two categories: 1) resources that existed before CLARIAH-EUS was established, but which are now integrated into the network to maximize their reach and usability for the research community; and 2) newly developed resources created by CLARIAH-EUS members. The following section includes examples of both types, as well as a brief survey of several recent and ongoing projects that reflect current work being done on Basque-related topics.

5.1 Projects

5.1.1 CLARIAH-EUS-gArA

CLARIAH-EUS-gArA, funded by the Provincial Council of Gipuzkoa, is constructing a trustworthy conversational assistant for Basque news and research in the digital humanities. More specifically, the project explores how Latxa,[27] currently the largest and best performing LLM family for Basque, may be utilized in concert with retrieval-augmented generation (RAG) techniques to help guarantee the fidelity of responses to queries concerning news in Basque. The currently intrinsic tendency of LLMs to invent responses is an obstacle that CLARIAH-EUS-gArA also addresses. Accordingly, the project seeks to ensure that generated responses extracted from in-depth interactions with knowledge repositories include

[25] http://www.clariah.eus/eu/3.workshopa
[26] https://www.clariah.eus/eu/baliabideak_sailkapena
[27] https://huggingface.co/HiTZ/latxa-70b-v1.1

detailed annotations on the provenance of data so that researchers may fact-check, verify information, and assess reliability through accurate citations that provide direct access to original sources.

This ability to validate source material is not only a critical factor in restoring confidence in online content, but also a fundamental aspect of SSH research. To reach its objectives, CLARIAH-EUS-gArA is establishing methods to permanently crawl news as it is released and collect past news articles, offer synchronic and diachronic perspectives on events and opinions, and allow for multilingual comparison of Basque news with reporting in other languages. The technology developed for the project will be a step forward in this regard and for Basque DH in general because it will 1) create computational tools for Basque; 2) offer new means to browse news in Basque; 3) advance how Basque-language data are processed; and 4) support DH researchers in verifying sources and citations when searching for answers to complex research questions.

5.1.2 HarilkAI: Prototyping AI applications in the Social and Communication Sciences

HarilkAI's objective is to investigate the application of LLMs and AI tools to the field of teaching and education in the social and communication sciences.[28] To do so, a socio-technical infrastructure will be designed to experiment with AI tools and investigate how they may be better adapted to this area of study. HarilkAI addresses three main issues:

- the practice and epistemology of social and communication sciences. Participants will be asked to interact with AI tools, report on their effectiveness, and propose protocols, programs, and prototypes.
- digital sovereignty based on open-source philosophy and practices. This includes encouraging collaborative research and finding alternatives to proprietary software and platforms.
- the use of tools created from, and in, Basque within the context of LLMs.

Researchers and teachers from various research areas will be able to access HarilkAI's laboratory, which will adapt to their needs in an experimental and collaborative way. This includes developing software, implementing and configuring hardware, and constructing virtual spaces to test AI tools.

5.1.3 BIM/SAHCOBA

The projects BIM (*Basque in the Making: A Historical Look at a European Language Isolate*)[29] and SAHCOBA (*Syntactically Annotated Historical Corpus in Basque*) focused on developing a morphosyntactically annotated historical corpus of the Basque language (Estarrona et al., 2022). The interdisciplinary initiatives brought together experts in linguistics and natural language processing. BIM was dedicated to gathering key Basque texts from the fifteenth to mid-eighteenth centuries (Archaic and Old Basque), while the SAHCOBA project expanded this corpus to cover the period from the mid-eighteenth to the mid-twentieth centuries (Early and Late Modern Basque), coinciding with the emergence of standardized Basque. The corpus includes annotations for both parts of speech and syntax, along with extensive metadata. The database enables users to search the annotated corpus by word, lemma, grammatical category, sequences of grammatical categories, and specific syntactic structures. For example, the canonical order for negation in contemporary Basque is *ez da etorri* (has not come), that is ”negation particle (*ez*) + auxiliary verb (*da*) + verb in participle (*etorri*). But we know that in ancient texts the order *etorri ez da* appears. Thus, if we wanted to study the evolution of this syntactic structure over time, we could use the interface to search for the structure ”participle verb + negation particle (*ez*) + auxiliary verb” and we would retrieve examples of the type *etorri ez da*.

5.1.4 ZITERAUZI

Semantic web technologies (Linked Open Data) have enabled new opportunities for recording, exhibiting, and querying publication collections. The ZITERAUZI project (Astigarraga et al., 2025) aims to take advantage of this shift by creating a tool chain for citation extraction from scientific articles published

[28] http://www.clariah.eus/sites/default/files/posterrak/Poster_HarilkAI.pdf

[29] http://bim.ixa.eus/search

in Basque in order to track the use of the language within the academic sphere. Inguma, a Basque scientific production database, has served as a starting point. In collaboration with the Digital Humanities Center in Errenteria, the ZITERAUZI team enriched the scientific publication metadata on articles from the IkerGazte conference series currently found in Inguma by representing extracted citation relations. The project's objectives include representing these relationships in a directed graph. This proposed graph will serve as an infrastructure for different use cases, which together will advance the study of scientific production in Basque.

5.1.5 Historical Texts in Wiki-platforms as Linked Data

This project proposes a data model for storing Basque historical texts in a database following the Linked Data paradigm by utilizing Wikibase software as infrastructure (Lindemann & Alonso, 2025). On the one hand, the project models entities, describing corpus tokens and token spans on top of the preset Wikibase data model. On the other hand, it builds a Standard Basque dictionary, deploying the Ontolex-Lemon vocabulary on the same Wikibase instance. This model allows for linking tokens and token spans to the dictionary's entries on a lexical entry (lemma) level, lexical sense level, and inflected form level. At the same time, entities at these three levels may carry additional descriptions and other links. This extends a traditional way of morphosyntactic annotation with literal values towards linking dictionary elements as entities. In addition to this corpus-lexicon interface, other kinds of annotations are also modeled, such as philological and semantic annotations.

5.2 Tools and Resources

5.2.1 CORPErrore

CORPErrore is a resource that enriches the HABE-IXA corpus[30] by labeling errors in Basque over the corpus using the INCEpTION annotation tool (Klie et al., 2018). Through the CORPErrore website,[31] searches may be conducted by error or suberror, errors can be queried by tiers, and text searches may be performed.

5.2.2 ETEL

The ETEL system[32] is designed to analyze texts and obtain linguistic complexity measures in a simple way for researchers and research groups. ETEL allows research teams to collaborate on the same project with different user profiles. The system offers four main functionalities: 1) text analysis (linguistic complexity indicator selector and results visualizer), 2) text suggestion for selected text level, 3) corpus management tools (file manager and complexity-level manager), and 4) a user administrator. Various aspects of textual complexity may be analyzed using ETEL, including lexical phenomena (distinct words or lemmas), syllabic measurements, and PoS information, such as the number of verbal lemmas. ETEL's multifaceted analysis system compares analyzed text with the results of a pre-classified and analyzed corpus and graphically displays its position in relation to the corpus measurements already performed for each analyzed phenomenon.

5.2.3 IGARRITZ

IGARRITZ (Iruskieta et al., 2024) is an adapted web environment that employs AI techniques for Basque text prediction.[33] It is designed to facilitate the creation of texts in Basque for secondary school students with cerebral palsy. To achieve this goal, the project developed a web interface that is based on the HiTZ/roberta-eus-euscrawl-base-cased language model. This was retrained with an educational Basque corpus sourced from texts that appear in Gizapedia, Wikipedia, and the Basque newspaper *Berria*. Igarritz has been incorporated into the CLARIAH-EUS B-centre.

[30] https://b2share.eudat.eu/records/81433fddcd06405f8505c7606b29ff99

[31] https://corperrore.clariah.eus/

[32] https://etel.clariah.eus/etel/Analyzer

[33] https://igarritz.clariah.eus/

5.2.4 Contemporary Basque Student Handwritten Model

The Contemporary Basque Student Handwritten Model[34] is a Basque AI model in Transkribus designed to transcribe learners' handwriting in Basque. The dataset consists of school-based texts written by adolescent students aged 12–16. Original errors in the handwriting were preserved and transcribed verbatim. The model was trained on a corpus of 51,195 words in Basque, collected from various schools in the Basque Autonomous Community in 2023.

5.2.5 HiTZketan

Launched by HiTZ, HiTZketan (lit. *in conversation*) has developed a speech-to-speech translation system for Basque and Spanish. The system receives a speech signal in one of the two languages, translates it to the other, and then delivers this translation using speech with a personalized voice that imitates the original speaker. Accordingly, HiTZketan enables Basque-language support for state-of-the-art technologies in automatic speech recognition, machine translation, and personalized speech synthesis.[35]

5.2.6 Parlamint-ES-PV 4.0

ParlaMint 4.0 is a collection of comparable corpora[36] featuring transcripts of parliamentary discussions from twenty-nine European nations and autonomous regions, primarily spanning from 2015 to mid-2022. Each corpus contains between nine million and 126 million words, with the entire compilation exceeding 1.1 billion words. CLARIAH-EUS has developed the Basque and Spanish corpus (Alkorta & Iruskieta, 2022) using data and metadata sourced from the Basque Parliament.

5.2.7 Computational Social Science Resources

At least three datasets related to social media analysis are available for research and tool development in Basque: 1) the Heldugazte dataset,[37] which focuses on identifying the writing style of a given text sequence (Fernandez de Landa et al., 2019); 2) the Heldugazte-Age dataset,[38] designed to determine the age group of Basque social media users, classifying them as either minors or adults (Fernandez de Landa & Agerri, 2021); and 3) VaxxStance,[39] which analyzes social media posts to assess opinions on vaccines (Agerri et al., 2021). This dataset categorizes tweets based on their stance as AGAINST, IN FAVOR, or NEUTRAL regarding a predefined topic.

5.2.8 BERnaT: Modeling the Diversity of the Basque Language

BERnaT applies research done on linguistic diversity in the creation of language models to Basque.[40] Its objective is to develop Basque models that take into account the diversity of the language. By utilizing the Basque corpus EusCrawl as a foundation, along with the best-performing discriminative model (Artetxe et al., 2022), multiple Basque models have been created that incorporate varying levels of linguistic diversity. Four datasets were used to achieve this:

- EusCrawl, a clean and standardized Basque corpus.
- the Latxa corpus (Etxaniz et al., 2024), currently the largest available Basque corpus.
- a spontaneous corpus created specifically for this study, consisting of thousands of tweets from Basque users.
- a combined dataset integrating all the aforementioned corpora.

The newly trained models have been evaluated using the BasqueGLUE (Urbizu et al., 2022) benchmark. BERnaT aims to demonstrate that models trained with more diverse linguistic corpora can be beneficial for tasks involving more spontaneous language or greater linguistic variation.

[34] https://www.transkribus.org/model/contemporary-basque-student-handwritten
[35] http://www.clariah.eus/sites/default/files/posterrak/Hitzketan(2)Clariah.pdf
[36] https://www.clarin.si/repository/xmlui/handle/11356/1859
[37] https://github.com/joseba-fdl/heldugazte-corpus
[38] https://github.com/joseba-fdl/heldugazte-age-corpus
[39] https://vaxxstance.github.io/
[40] http://www.clariah.eus/sites/default/files/posterrak/Clariah-eus2024_posterra_bernat.pdf

5.2.9 C1 Automatic Evaluator for Basque

HiTZ has developed an automatic evaluator that determines whether Basque-language writings meet the C1 level or not. To develop the system, essays from candidates who took the HABE (Institute for Adult Literacy and Basque Relearning)[41] C1 exams were utilized, taking into account the grades assigned by examiners. Ten thousand automatically transcribed texts and around 600 manually transcribed compositions were obtained through an agreement between the IKERGAITU project[42] and HABE-HiTZ. The system is based on language models with a neural foundation. Experiments were conducted with different types of language models to identify the most suitable configuration: monolingual or multilingual, and encoder or decoder. Accuracy rate was utilized for automatic evaluation metrics. Results demonstrate that models based on the Latxa language model (Etxaniz et al., 2024) performed the best, with the top model achieving an overall accuracy rate of 79%. A demo has been created based on the most suitable system.[43]

5.2.10 Resources and Tools for the Development of High-Level Academic Texts

Researchers at HiTZ have developed various resources and tools to stabilize and promote the use of academic registers in Basque, without which the Basque linguistic community risks losing the specialized language that is essential within an academic context. These include: the Garaterm corpus,[44] the TZOS (Online System for the Service of Terminology)[45] terminological database, and the bilingual Academic Text Writing Support Tool (HARTA/TAILA).[46] These resources represent part of HiTZ's efforts to create a Basque work environment and foster collaboration through interoperable data sharing and a dynamic network of experts.

5.2.11 Children's Literature Corpus

The Basque Language Institute[47] and the research group Gogo Elebiduna (Bilingual Mind)[48] at the University of the Basque Country (UPV/EHU) have created a children's literature corpus to provide a resource for speech therapists, audiologists, and language teachers who are working with children with speech and language difficulties. In addition to the corpus, based on literature for children aged 0-8, the goal of the initiative is to develop a search system for utilizing the corpus, where professionals can select language materials according to various criteria.[49] The initial version of the corpus contains 428 books, categorized into two age groups (0-4 and 5-8), and a web interface that enables users to perform simple searches and phonemic queries. Future plans to expand the corpus include enriching it with a broader range of language materials, introducing visual media, and refining the search engine to allow for searches that encompass several phonological, lexical, semantic, morphosyntactic, and grammatical levels.

5.2.12 Basque PhD Abstracts and Abstracts Corpus Collection

The data collected in PhD theses completed at the University of the Basque Country, along with the tools created to analyze these data, are currently dispersed across university repositories, hindering their potential reuse. One of CLARIAH-EUS's objectives is to mitigate this issue by facilitating the searchability of thesis-produced data, enhancing their visibility and reusability for other researchers. To do so, CLARIAH-EUS aims to upload PhD abstract data into EuDat, subsequently making it accessible through CLARIN's VLO for wider utilization by researchers under the CC BY-NC 4.0 license. This exercise will include the development of novel methods and tools to generate and evaluate PhD abstracts.

[41] https://www.habe.euskadi.eus/hasiera/

[42] https://www.hitz.eus/iker-gaitu/

[43] https://huggingface.co/spaces/HiTZ/C1_sailkapen_demoa

[44] http://garaterm-corpusa.ixa.eus/

[45] https://tzos.ehu.es/?setuilang=eu

[46] https://harta.ixa.eus/

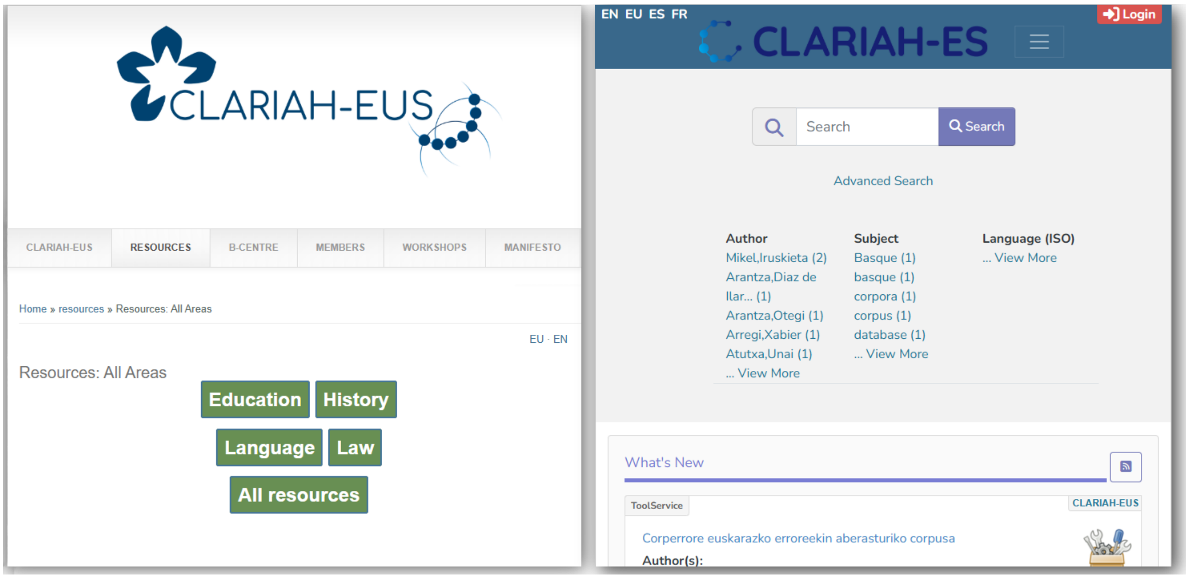
[47] https://www.ehu.eus/en/web/eins/

[48] https://www.ehu.eus/HEB/

[49] https://www.ehu.eus/ehg/08corpusa/

6 CLARIN B-Centre

The CLARIAH-EUS B-centre at HiTZ is part of the CLARIAH-ES infrastructure and, as such, will serve that larger community. Nonetheless, the need to establish a repository that can deliver digital language technology to those mainly engaged in Basque-language research and Basque Studies is a driving motivation behind the creation of the B-centre at HiTZ. Our hope is that the repository, currently in the final stages of construction, will help meet CLARIAH-EUS's desire to foster a collaborative space for those working with Basque or on Basque-related projects. With this in mind, the following brief survey of the CLARIAH-EUS B-centre highlights the initial steps taken in its ongoing construction.



Resources: All Areas

Author	Subject	Language (ISO)
Mikel,Iruskieta (2)	Basque (1)	... View More
Arantza,Diaz de Ilar... (1)	basque (1)	
Arantza,Otegi (1)	corpora (1)	
Arregi,Xabier (1)	corpus (1)	
Atutxa,Unai (1)	database (1)	
... View More	... View More	

Corperrore euskarazko erroreekin aberasturiko corpusa

Author(s):

Figure 2: CLARIAH-EUS B-Centre Webpages

6.1 Requirements

Several requirements must be fulfilled to be certified as a CLARIN B-centre and obtain compliance with CoreTrustSeal standards (CoreTrustSeal Standards and Certification Board, 2022). The latter provide a framework for building and maintaining trustworthy digital repositories that ensure the long-term accessibility and usability of digital data and metadata. HiTZ is currently working to put these guidelines and requirements in place. To begin with, we have defined our main objectives and mission in keeping with the goals of CLARIAH-EUS. Second, we have assigned staff to oversee the repository's construction and future management, with the expectation that we will hire and train new staff members in the near future. Qualified technicians from HiTZ are currently building the repository's infrastructure, housed at the University of the Basque Country's Department of Computer Science in San Sebastián, Spain. Third, and with respect to more technical questions, we are in the process of gathering information about data management, rights, legality, quality, suitability, and reliability in order to determine which measures and protocols are best suited to ensure CoreTrustSeal standards are met.

6.2 The B-Centre Server

The B-centre server's infrastructure must be configured to withstand accidental and temporary outages or crashes. This allows users and applications to continue operating without interruption and to access data and services. In our current setup, HiTZ has two servers available that we will configure in failover mode: one active and one passive, ready to take over if the first fails. However, to obtain a better High Availability (HA), we plan to integrate a third server for enhanced resilience and to strengthen our ability to withstand disruptions. Leveraging technology that links three servers, such as Ceph, can offer superior reliability compared to using only two servers.

A virtual environment manager that functions with infrastructures based on two servers must also be selected. Ideally, the VE should allow for straightforward management and, out of several options, we have chosen Proxmox. This choice ensures that all team members responsible for handling the group's infrastructure can easily manage the system. However, HiTZ has decided that one individual will be designated to oversee the process. After selecting the OS, we proceeded with the installation process and configured the cluster following these steps: 1) created the cluster and connected the servers; 2) configured Corosync to work with two nodes; 3) connected the servers with a straight cable at ten Gbps; 4) configured a zfs PoolStorage; 5) installed the guest machine VM (Rocky Linux); 6) implemented replication; and 7) enabled High Availability to migrate the guest VM on failure.

Each CLARIN B-centre is required to maintain data and software, necessitating the establishment of a repository. While various Digital Resource Managers are available, CLARIN does not mandate a specific choice. A common option among centers is the utilization of open source platforms like DSpace. We have chosen to utilize the CLARIN DSpace implementation[50] to meet CLARIN requirements and to facilitate connection to the Virtual Language Observatory (VLO). CLARIN DSpace 7[51] has been installed.

7 Future Steps for CLARIAH-EUS

In the near future, CLARIAH-EUS will complete the final phase in the implementation of its CLARIN B-centre, which will be added to its already operational CLARIN K-centre. This will allow us to offer technical services as well as valuable instructional guidance to researchers. As this process unfolds, three key criteria will guide CLARIAH-EUS's immediate development:

- **Building Resources**. CLARIAH-EUS will prioritize creating or adapting resources and services that are readily accessible to researchers through the CLARIAH-EUS node.
- **Strategic Focus**. The infrastructure will target resources and services that strategically address the needs of the Basque research community.
- **Collaboration**. CLARIAH-EUS will create or adapt resources and services that seamlessly integrated with CLARIN and DARIAH.

CLARIAH-EUS is focused on making an immediate impact by adapting existing resources and incorporating them into our B-centre. Specifically, we hope to include the Analhitza tool (Otegi et al., 2017), the Euscrawl system (Artetxe et al., 2022), and ParlaMint. Additionally, we plan to develop new resources and provide various corpora, including literature, historical texts, and social network data. Looking ahead, we intend to create tools and resources for sociology, journalism, literature, and history, ideally in alignment with GLAM-related initiatives.

Acknowledgments

We thank the Basque Government and its Department of Culture and Linguistic Policy, the Provincial Council of Gipuzkoa, the Vice-Rectorate of Basque, Culture and Internationalization at the University of the Basque Country (UPV/EHU), and the HiTZ center for their support, as well as the CLARIAH-ES Network (RED2022-134527-E), funded by the Spanish Ministry of Science, Innovation and Universities.

References

Agerri, R., Centeno, R., Espinosa, M., Fernandez de Landa, J., & Rodrigo, A. (2021). VaxxStance@IberLEF 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, *67*, 173–181. 10.26342/2021-67-15

Alkorta, J., & Iruskieta, M. (2022). Adding the Basque Parliament Corpus to ParlaMint Project. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 107–110.

[50] CLARIN DSpace implementation: https://github.com/ufal/clarin-dspace

[51] CLARIN DSpace installation https://github.com/ufal/clarin-dspace/wiki

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., & Soroa, A. (2022). Does Corpus Quality Really Matter for Low-Resource Languages? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7383–7390).

Arzoz, X. (2015). The Impact of Language Policy on Language Revitalization: The Case of the Basque Language. *Cultural and Linguistic Minorities in the Russian Federation and the European Union: Comparative Studies on Equality and Diversity*, 315–334.

Astigarraga, A., Lindemann, D., & Bidaguren, M. (2025). Ziterauzi: The tool chain for citation extraction from basque academic texts. *CLARIAH-EUS: Zientzia Sozialak eta Humanitate Digitalak gaur egun*.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2016). *Digital_humanities*. MIT Press.

CoreTrustSeal Standards and Certification Board. (2022, September). CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 Extended Guidance. https://doi.org/10.5281/zenodo.7051096

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*, *8*(0), 1663–1672.

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, *37*(2), 391–404.

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., & Soroa, A. (2024). Latxa: An open language model and evaluation suite for basque. https://arxiv.org/abs/2403.20266

Fernandez de Landa, J., & Agerri, R. (2021). Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, *0*(0), 1–15.

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, *10*(6).

Gonzalez-Dios, I., & Altuna, B. (2022). Natural Language Processing and Language Technologies for the Basque Language. *Cuadernos Europeos de Deusto*, (04), 203–230.

Iruskieta, M., de la Iglesia, I., Atutxa, U., & Ortiz, L. (2024). IGARRITZ: euskarazko testu iragarpenerako web ingurune egokitua. *Ekaia*, *Ale berezia*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Lindemann, D., & Alonso, M. (2025). Historical texts in wiki-platforms as linked data. *CLARIAH-EUS: Zientzia Sozialak eta Humanitate Digitalak gaur egun*.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., & Uria, L. (2017). ANALHITZA: A tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58), 77–84.

Sarasola, K., Aldabe, I., Díaz de Ilarraza, A., Estarrona, A., Farwell, A., Hernáez, I., & Navas, E. (2023). Language Report Basque. In *European Language Equality: A Strategic Agenda for Digital Language Equality* (pp. 95–98). Springer.

Terras, M. (2011). Quantifying digital humanities. *UCL Centre for Digital Humanities*.

Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., & Soroa, A. (2022, June). BasqueGLUE: A natural language understanding benchmark for Basque. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1603–1612). European Language Resources Association.