CLARIAH-EUS: a Cross-border CLARIAH Node for the Basque Language and Culture

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna, Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga and Jose Mari Arriola

CLARIAH-EUS, HiTZ Basque Center for Language Technology - Ixa NLP Group, University of the Basque Country (UPV/EHU), Manuel Lardizabal pasealekua, 1, 20018 Donostia-San Sebastian, Gipuzkoa, Basque Country

Abstract

CLARIAH-EUS is a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN and DARIAH, Europe's two principal digital research infrastructures for the humanities, arts, and social sciences. CLARIAH-EUS aims to sustain research in these fields of study that is related to Basque or Basque culture by supporting scholars with digital tools and resources. The node is unique because it seeks to service a language (Basque) and not a territory, making the infrastucture transnational in scope. In this article, we describe the motivations for creating CLARIAH-EUS, how it was constructed, the projects that are in currently in development, and future plans.

Keywords

Basque, Infrastructure, CLARIN ERIC, DARIAH ERIC, CLARIAH, Digital Humanities, Arts, Social Sciences

1. Introduction and motivation

The nature of research is in constant flux. This steady change is especially evident in fields where technology is required for research. Conversely, it is not always as pronounced in those where the presence of technology is deemed to be less essential; areas in which qualitative analysis often plays as significant a role as its quantitative counterpart. The humanities, arts, and social sciences, with some notable exceptions, have customarily fallen into the latter camp.

Over the past two decades, however, a small cadre within these disciplines has begun to take advantage of digital technology in ways that have given rise to new modes of research and, consequently, new lines of inquiry [1]. The advent of this "digital turn" are attested to by approaches and results that were once impossible. The use of digital tools and methods are, for example, reshaping how GLAMs (galleries, libraries, archives, and museums) engage with cultural heritage. The same is true of social scientists, who have adapted to the emergence of big data by devising novel techniques that take advantage of an ocean of digital information, casting new light on how data can represent reality [2]. Even the lines between disciplines have softened as digital humanities, in its broadest sense, has necessarily prompted unforeseen collaborative and transdisciplinary research, teaching, and publishing [3].

The results of this type of innovative work sometimes overshadow the fact that language technology is often at its core. More concretely, language technology specifically crafted to aid the humanities, arts, and social sciences. And the successful application of this particular specialization, it is worth remembering, depends to a great extent on the availability of tools, resources, and data for and in the languages that are being utilized for research. As may be appreciated, this is one reason why the development of language technologies is important for all languages, but absolutely crucial for languages that belong to relatively small populations [4].

Although Basque may be counted among the lesser-spoken languages, its situation in terms of language technology is comparatively favorable to languages of a similar size. This is largely thanks to significant progress in "fostering the necessary sociolinguistic conditions for the successful development and dissemination of LT," which has resulted in "state-of-the-art technology and robust, broad-coverage NLP for Basque" [5, 6]. Equally important in this regard are the several decades of collaborative work between research groups, foundations, language industry clusters, and regional institutions. Yet, despite these efforts, Basque remains in a precarious position with respect to research maturity and readiness.

The CLARIAH-EUS consortium was created to help

SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics, June 19-20, 2024, A Coruña, Spain.

jon.alkorta@ehu.eus (J. Alkorta); aritz.farwell@ehu.eus (A. Farwell); joseba.fernandezdelanda@ehu.eus (J. F. d. Landa); begona.altuna@ehu.eus (B. Altuna); ainara.estarrona@ehu.eus (A. Estarrona); mikel.iruskieta@ehu.eus (M. Iruskieta); xabier.arregi@ehu.eus (X. Arregi); xabier.goenaga@ehu.eus (X. Goenaga); josemaria.arriola@ehu.eus (J. M. Arriola)
0000-0003-0812-8618 (J. Alkorta); 0000-0003-0124-3007 (A. Farwell); 0000-0001-6067-3571 (J. F. d. Landa); 0000-0002-4027-2014 (B. Altuna); 0000-0002-1616-5665 (A. Estarrona); 0000-0002-6121-3902 (M. Iruskieta); 0000-0002-3359-1295 (X. Arregi); 0000-0003-2624-7143 (J. M. Arriola)

address these shortcomings. Its goals are twofold: 1) promote language technology for researchers involved in Basque-related humanities, arts, and social sciences and 2) foster and facilitate relationships between these researchers so that they may better exchange ideas and innovative approaches. For these reasons, CLARIAH-EUS is centered on language rather than territory or region, making it transnational in scope. Organizationally speaking, the network is a node within CLARIAH-ES, Spain's distributed infrastructure for CLARIN ERIC and DARIAH ERIC, Europe's two principal digital research infrastructures for the humanities, arts, and social sciences.

2. Objectives

As stated above, one of CLARIAH-EUS's objectives is to support language technology for researchers involved in Basque-related humanities, arts, and social sciences. Part of this activity involves producing digital resources for Basque and integrating them into the CLARIN ERIC and DARIAH ERIC infrastructures. Doing so will generate tools that are designed with Basque in mind and provide researchers with better access to them. A second, closely related area of activity, will focus on offering services to researchers who create or wish to utilize Basque language technology for the digital humanities.

Another ambition of CLARIAH-EUS is to establish a research community that is devoted to sustaining language technology for Basque-related humanities, arts, and social sciences. On the one hand, the purpose of this goal is to encourage collaboration between scholars that may lead to more impactful research or greater opportunities for participation in international projects. On the other, the aspiration stems from the belief that a strong research community is more likely to cultivate an environment that yields innovative approaches to Basque digital humanities and language technology for Basque.

3. Funding institutions

CLARIAH-EUS is focused on maintaining long-lasting and steady financial support so as to ensure short-, mid- and long-term research initiatives and guarantee the reusability of resources. Several public stakeholders share this perspective and have provided funding 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).

The UPV/EHU is represented by the Vice-Rectorate of Basque, Culture and Internationalization,[3] which participates as a financing entity, and by HiTZ, the Basque Center for Language Technology.[4] The latter, which also provides funding, is responsible for CLARIAH-EUS's administrative office and several of its members are part of the CLARIAH-EUS steering committee.

With the help of these institutions, CLARIAH-EUS has hired four staff members, who oversee the maintenance of the CLARIAH-EUS infrastructure and perform duties within the administrative offices of both CLARIAH-EUS and CLARIAH-ES (for which HiTZ is also responsible).[5]

4. Development of CLARIAH-EUS

The development of CLARIAH-EUS has occurred in two stages: a design phase (2021-2023) (see sections 4.1. and 4.2) and an implementation phase (2023-present), during which time CLARIAH-EUS has become an active node (see sections 4.3 and 4.4).

4.1. First workshop: needs and manifesto

On November 26, 2021, an initial workshop,[6] *Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen* (*Designing the CLARIAH-EUS infrastructure to develop language technology for Basque in the Humanities and Social Sciences*), was organized by HiTZ to create the CLARIAH-EUS infrastructure.

Its objective was to discuss opportunities and needs for different research areas. The workshop's activities included: 1) compiling a collection of use cases and posters of digital projects developed for anyone who wants to study Basque, 2) making a list of the strategic resources necessary for Basque and Basque research in different disciplines, and 3) obtaining the involvement of researchers to promote the CLARIAH-EUS research infrastructure.

Nine institutions and thirty-four researchers from twenty research groups participated and fourteen projects were presented. In addition, 134 organizations and individuals signed a manifesto[7] calling for the creation of a digital humanities infrastructure for Basque.

4.2. Weaving the network

From 2021 to 2023, the objective was to seek support from various organizations and research groups. This was obtained from ten entities: HiTZ (UPV/EHU), Udako Euskal Unibertsitatea (UEU), Iker research group, Elhuyar, Gogo

[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.es/
[6] https://www.clariah.eus/eu/1-workshop
[7] https://www.clariah.eus/eu/manifestua

Elebiduna research group (UPV/EHU), Elebilab research group (UPV/EHU), Aholab research group (UPV/EHU), Ixa research group (UPV/EHU), Soziolinguistika Klusterra, and the Unesco Chair in Human Rights and Public Powers (UPV/EHU). Nine of these organizations or research groups are from the southern Basque Country and one (Iker) is from the northern Basque Country.

During this time, the CLARIAH-EUS node was also defined in relation to CLARIAH-ES in Spain and the CLARIN ERIC and DARIAH ERIC infrastructures at the European level.

4.3. Second workshop: community and organization

CLARIAH-EUS's second workshop[8] was held in November 2023. In contrast to the previous workshop, its objective was to present the CLARIAH-EUS infrastructure and its aims, as well as to survey ongoing work in Basque digital humanities.

The event served as a kickoff ceremony for the founding members. CLARIAH-EUS's structural organization and road map were discussed, with a focus on the strategic lines that will be developed over next five years. In addition, two invited speakers gave talks and twenty-one posters were presented. A selection of these, along with descriptions of the research groups that took part in the workshop, will be described in a forthcoming publication.[9]

4.4. Action protocol

We have put in place an action protocol as part of CLARIAH-EUS's implementation. To use the service, a petitioner must first register a request at www.clariah.eus/contacto. Once a petition is activated, CLARIAH-EUS updates the status of the request as changes occur. The petitioner's opinion is solicited at the end of the collaboration or service through a survey: https://www.ixa.eus/events/clarink_survey. All node services are evaluated at the end of each year.

Furthermore, petitioners will also be asked to acknowledge the node in resulting publications or on websites by including the CLARIAH-EUS and HiTZ logos, along with the following statement: "SUPPORTED by CLARIN-EUS. HiTZ Center - University of the Basque Country UPV/EHU."

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

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

5. Projects and resources

As underscored above, one of CLARIAH-EUS's principal objectives is to support researchers by providing tools and resources[10] that can be employed in digital humanities and social sciences. Some of the tools and resources available through CLARIAH-EUS were produced before the infrastructure was constructed. These have been integrated into the infrastructure to multiply their outreach and usability. Others, however, have been or are being developed under the auspices of CLARIAH-EUS. The following examples fall under both categories.

5.1. Parlamint-ES-PV 4.0

ParlaMint 4.0 is a set of comparable corpora[11] containing transcriptions of parliamentary debates from twenty-nine European countries and autonomous regions, mostly dating from 2015 and to mid-2022. The individual corpora comprise between nine and 126 million words and the complete set contains over 1.1 billion words. CLARIAH-EUS has created the corpus [7] in Basque and Spanish utilizing data and metadata from the Basque Parliament.

5.2. Computational social science

Three datasets related to social media analysis are available for the purposes of experimentation and the development of tools for Basque: 1) the Heldugazte[12] [8] dataset, designed to identify the writing style of a specific text sequence; 2) the Heldugazte-Age[13] [9] dataset, meant to identify the age of Basque social media users by classifying them as either minors or adults; and 3) VaxxStance,[14] [10] which seeks to identify the stance expressed on social media regarding vaccines. Its objective is to determine whether a given tweet expresses an AGAINST, FAVOR, or NEUTRAL stance towards a previously defined topic.

5.3. BIM/SAHCOBA

Basque in the Making (BIM): A Historical Look at a European Language Isolate and *Syntactically Annotated Historical Corpus in Basque (SAHCOBA)* are two projects[15] for the construction of a morphosyntactically annotated Basque historical corpus [11]. BIM and SAHCOBA are interdisciplinary projects that include experts in linguistics and natural language processing. The BIM project aims to collect the most significant writings from the fifteenth century to the mid-eighteenth century (Archaic

[10] https://www.clariah.eus/eu/baliabideak_sailkapena

[11] https://www.clarin.si/repository/xmlui/handle/11356/1859

[12] https://github.com/joseba-fdl/heldugazte-corpus

[13] https://github.com/joseba-fdl/heldugazte-age-corpus

[14] https://vaxxstance.github.io/

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

and Old Basque), while the SAHCOBA project aims to extend this corpus from the mid-eighteenth century to the mid-twentieth century (Early and Late Modern Basque), when standard Basque appeared. The corpus comprises both part-of-speech and syntactic annotation, as well as a rich set of metadata structure. The database allows the annotated corpus to be searched by words, lemmas, grammatical categories, sequences of grammatical categories, and specific structural configurations.

6. The Future of CLARIAH-EUS

With regard to the institutional nature of CLARIAH-EUS, our objective is to restructure its current status as a CLARIN K-centre into a CLARIN B-centre in order to provide technical services as well as instructional guidance to researchers. With respect to future work, our current outlook is shaped by three criteria:

- Creating or adapting resources and services that researchers can access from the CLARIAH-EUS node.
- Creating or adapting resources and services that are strategically needed within the Basque community.
- Creating or adapting resources or services that can be articulated with CLARIN ERIC and DARIAH ERIC.

In the short term, our goal is to offer various resources and services in CLARIN and DARIAH by adapting existing resources. By way of example, we hope to integrate the Analhitza tool [12], the Euscrawl system [13], and ParlaMint. Additionally, we intend to offer several types of corpora, such as literature, historical texts, and social networks, as well as produce new resources, including a data repository and APIs. In the medium term, our main objective is to offer resources and tools for the field of education, while also working on other areas, such as integration with the Virtual Language Observatory (VLO) and the construction of language models. In the long term, we will attempt to fashion tools and resources for sociology, journalism, literature, and history. Ideally, this work will coincide with GLAM-related projects.

Acknowledgments

We wish to 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 generous support.

References

[1] M. Terras, Quantifying digital humanities, UCL Centre for Digital Humanities (2011).

[2] K. Crawford, K. Miltner, M. L. Gray, Critiquing Big Data: Politics, Ethics, Epistemology, International Journal of Communication 8 (2014) 1663—-1672. URL: https://ijoc.org/index.php/ijoc/article/view/2167/1164.

[3] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, J. Schnapp, Digital_Humanities, Mit Press, 2016.

[4] X. Arzoz, 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 (2015) 315–334. URL: 10.1007/978-3-319-10455-3_12.

[5] K. Sarasola, I. Aldabe, A. Díaz de Ilarraza, A. Estarrona, A. Farwell, I. Hernáez, E. Navas, Language Report Basque, in: European Language Equality: A Strategic Agenda for Digital Language Equality, Springer, 2023, pp. 95–98. doi:10.1007/978-3-031-28819-7_5.

[6] I. Gonzalez-Dios, B. Altuna, Natural Language Processing and Language Technologies for the Basque Language, Cuadernos Europeos de Deusto (2022) 203–230. doi:https://doi.org/10.18543/ced.2477.

[7] J. Alkorta, M. Iruskieta, Adding the Basque Parliament Corpus to ParlaMint Project, in: Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, 2022, pp. 107–110.

[8] J. Fernandez de Landa, R. Agerri, I. Alegria, Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case, Information 10 (2019). URL: https://www.mdpi.com/2078-2489/10/6/212. doi:10.3390/info10060212.

[9] J. Fernandez de Landa, R. Agerri, Social analysis of young Basque-speaking communities in twitter, Journal of Multilingual and Multicultural Development 0 (2021) 1–15. URL: https://doi.org/10.1080/01434632.2021.1962331. doi:10.1080/01434632.2021.1962331.

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

[11] A. Estarrona, I. Etxeberria, A. Soraluze, R. Etxepare, M. Padilla-Moyano, The first annotated corpus of historical Basque, Digital Scholarship in the Humanities 37 (2022) 391–404. URL: https:

//hal.science/hal-03505658.

[12] A. Otegi, O. Imaz, A. Díaz de Ilarraza, M. Iruskieta, L. Uria, ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research (2017).

[13] M. Artetxe, I. Aldabe, R. Agerri, O. Perez-de Viñaspre, A. Soroa, 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7383–7390. URL: https://aclanthology.org/2022.emnlp-main.499. doi:10.18653/v1/2022.emnlp-main.499.