Euskorpora: A Strategic Framework for Digital Sovereignty and Linguistic Inclusion of Basque in the Era of AI

Victoria Arranz, Sara Arregi, Leire Barañano, Aitor García-Pablos

Euskorpora
Paseo Mikeletegi 65, 20009 Donostia-San Sebastian, Spain
{victoria.arranz, sarregi, lbaranano, agarcia}@euskorpora.eus

Abstract

Euskorpora is a pioneering initiative designed to establish a comprehensive digital infrastructure for the development of speech and language technologies in Basque. Built upon European, Spanish, and Basque strategies, it addresses the scarcity of linguistic data, foundational models, and technological resources for this non-Indo-European, low-resourced language. The project integrates large-scale data collection from public institutions and private organisations, creating extensive multimodal corpora that cover the linguistic, dialectal, and domain diversity of Basque. These resources support the training of open language models for speech, translation, and language understanding, as well as the establishment of an interoperable infrastructure aligned with European initiatives such as the European Language Data Space (LDS). By combining linguistic research, artificial intelligence, and data governance, Euskorpora ensures the digital sovereignty and inclusion of the Basque language within the global AI ecosystem. Beyond its regional focus, it stands as a transferable model for advancing linguistic diversity, technological innovation, and equitable digital transformation in multilingual Europe.

Keywords: Basque language, under-resourced, digital sovereignty, infrastructure, LLMs, speech and language resources

1. Introduction: The Status of Basque and Its Language Technology Landscape

The Basque language is a thriving unique non-Indo-European language spoken by approximately 800,000 individuals across Northern Spain and Southern France[1]. Despite its long history and rich cultural heritage, Basque has faced significant challenges in terms of standardisation, official recognition, and digital presence. The European Parliament resolution of 11 September 2018 on language equality in the digital age[2] emphasized the need for a "large-scale, long-term coordinated funding programme for research, development, and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands." It also highlighted Europe's ambition "to secure its leadership in language-centric AI."

In response to this resolution, the European Language Equality (ELE)[3] undertook detailed studies of European languages, including Basque (Sarasola et al., 2022). These studies evaluated the level of support that the different languages receive in terms of Language Resources (LRs) and Language Technologies (LTs). The findings provide a comprehensive picture of the current state, usage, and technological infrastructure of Basque, while also identifying gaps and challenges for further development. The general status of technological support for European languages is depicted in Figure 1, with that for Basque language highlighted with an orange arrow.

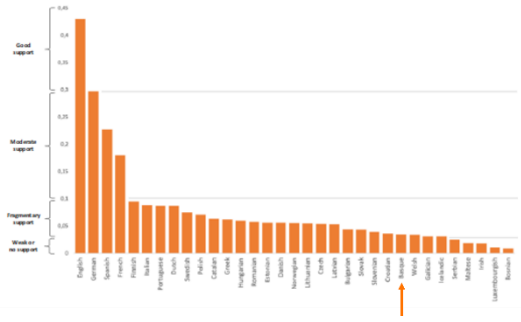

Figure 1: General Status of Technological Support for European Languages (2022)[4]

Basque exhibits a high level of digital engagement. However, it remains underrepresented in professional, business, and entertainment domains, highlighting the need for further development of digital resources and content. Furthermore, as indicated by (Rehm and Way, 2023), languages that cannot keep up with the rapid evolution of digital technologies on equal terms with English and other major languages are at serious risk of facing digital extinction.

In addition to the language preservation and presence-oriented context, there is the focus on the digital market for language technologies. These technologies are now present everywhere,

[1] https://es.wikipedia.org/wiki/Euskera
[2] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html
[3] https://european-language-equality.eu/

[4] Source: European Language Equality: A Strategic Agenda for Digital Language Equality

from voice assistants to machine translation, text and voice interfaces, digital learning environments, and industrial applications. According to the Language Technology Market Study[5] developed by *Nimdzi Insights*[6] in the framework of the DIGITAL Europe Programme[7], we are experiencing an unprecedented acceleration in the field of language technologies. The sector of language-focused artificial intelligence (AI) is expected to exceed 50 billion euros in the coming years, with a particular emphasis on large language models (LLMs), speech technologies, and machine translation. For this reason, the digital corpora that make it possible to train language and speech models will be a decisive factor for competitiveness and linguistic inclusion in the AI era.

2. European Context: Language Resources Infrastructures

Language resource infrastructures and catalogues play a fundamental role in enabling the development, sharing, and sustainability of linguistic data and tools across languages. At the European level, major initiatives such as **ELRA**[8], **CLARIN**[9], **META-SHARE**[10], **ELRI**[11], **ELRC**[12], and the **European Language Grid (ELG)**[13] provide a structured ecosystem for the collection, standardisation, and distribution of language resources. Complementing these, specialised repositories such as **OPUS (Open Parallel Corpus)**[14], constitute one of the largest open platforms of parallel corpora for machine translation research. National infrastructures also play a key part in fostering linguistic research and innovation, such as **Språkbanken (Swedish Language Bank)**[15]**,** which serves as Sweden's central repository for language data and tools and is fully integrated within the CLARIN network; **Kielipankki (Language Bank of Finland)**[16] which is part of FIN-CLARIN, and the **Taalbank Nederlands / Instituut voor de Nederlandse Taal (IVNT)**[17]. On a global scale, the **Open Language Archives Community (OLAC)**[18] acts as a federated metadata aggregator, interconnecting repositories worldwide and ensuring interoperability, discoverability, and long-term preservation of linguistic assets. Together, these infrastructures constitute the backbone of the multilingual digital ecosystem, setting the standards for resource sharing and technological cooperation across linguistic and national boundaries. Nowadays, there is also the Common European Data Space (CEDS) ecosystem, here represented by the **European Language Data Space (LDS)**[19]. Within this well-established landscape of language resource infrastructures, Euskorpora introduces a distinctive, regionally grounded yet internationally aligned initiative. Its added value lies in addressing the specific needs of a less-resourced, non-Indo-European language while adhering to the technical, ethical, and interoperability standards defined by European and global infrastructures. Euskorpora contributes by expanding the linguistic diversity represented in digital repositories, as well as by developing open, high-quality Basque corpora (***Euskorpus***), AI models, and an interoperable data infrastructure that can serve as a blueprint for other minority and low-resourced languages. By integrating Basque into the broader European digital scenario, Euskorpora strengthens the multilingual AI ecosystem and ensures that linguistic equity and technological innovation progress hand in hand.

3. Spanish Strategies Supporting Language Technologies

National (Spanish) and regional strategies on language technologies demonstrate a consolidated commitment to inclusive digital transformation and innovation in linguistic diversity, thus supporting less-resourced languages. At the national level, the *Language Technology Plan (Plan TL)*[20] and the *PERTE New Language Economy*[21] promote the development of language processing tools, open linguistic resources, and foundational models that enhance the digital presence of all languages. Complementing these initiatives, projects like *ILENIA*[22] and *ALIA*[23] focus on applied research and technological transfer, fostering infrastructure and methodologies for advanced language model training. Within this framework, the *Basque Government's Digital Transformation Strategy*

[5] This study was presented at the Language Technology Landscape Conference 2025.
[6] https://www.nimdzi.com/
[7] https://digital-strategy.ec.europa.eu/en/activities/digital-programme
[8] https://www.elra.info/
[9] https://www.clarin.eu/
[10] http://www.meta-share.org/
[11] https://www.vicomtech.org/en/rdi-tangible/projects/project/european-language-resource-infrastructure
[12] https://language-data-space.ec.europa.eu/related-initiatives/elrc_en
[13] https://live.european-language-grid.eu/
[14] https://opus.nlpl.eu/
[15] https://xn--sprkbanken-35a.se/
[16] https://www.kielipankki.fi/language-bank/
[17] https://ivdnt.org/
[18] http://www.language-archives.org/
[19] https://language-data-space.ec.europa.eu/index_en
[20] https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx
[21] https://portal.mineco.gob.es/es-es/ministerio/PlanRecuperacion/pertes/Paginas/PERTE_de_la_lengua.aspx
[22] https://proyectoilenia.es/
[23] https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf

2025 (ETDE2025)[24] identifies language technologies as a strategic driver for innovation, inclusion, and linguistic sovereignty, reinforcing the digital presence of Basque and aligning regional policy with European objectives on linguistic diversity and AI-driven language infrastructures. The convergence of these initiatives positions the Basque Autonomous Community as a reference model for the sustainable digital development of minority languages.

4. Euskorpora: An Innovative and Pioneering Infrastructure for a Low-Resourced Language

Built upon the strategic frameworks established by the Spanish and Basque language technology initiatives, Euskorpora emerges as a comprehensive and forward-looking not-for-profit association dedicated to support the development of language and speech technologies for Basque. Conceived as an infrastructural and scientific initiative, Euskorpora aims to endow the Basque language with the necessary digital resources, computational models, and technological infrastructure to ensure its full participation in the emerging multilingual AI ecosystem.

Euskorpora is a large-scale initiative designed to create, structure, and sustain the linguistic data and foundational models required for the advancement of artificial intelligence technologies in Basque. It functions both as a **corpus-building effort**, dedicated to the systematic collection, annotation, and curation of speech and text data, and as a **technological framework**, supporting the training, evaluation, and deployment of open-source AI models for speech, translation, and language processing.

The mission of Euskorpora is to guarantee the digital sovereignty and equality of the Basque language by providing the scientific, technical, and infrastructural foundations necessary for its integration into the global digital ecosystem. By bridging the gap between language policy, linguistic research, and technological innovation, Euskorpora seeks to strengthen the presence of Basque in all domains of digital interaction, promoting both linguistic diversity and technological competitiveness.

Euskorpora represents a pioneering effort in the field of minority language technologies, addressing critical gaps in digital resources, computational models, and infrastructure. Its innovative nature lies in the creation of comprehensive, high-quality Basque language corpora, the development of open-source foundational models for speech and language processing, and the establishment of an integrated technological platform capable of supporting research, industry applications, and cross-sectoral adoption. By combining scientific rigor, technological sophistication, and strategic vision, Euskorpora sets a benchmark for similar initiatives in other language communities, demonstrating a scalable model for fostering linguistic inclusion and technological innovation simultaneously.

4.1 Objectives

Euskorpora's general objectives can be summarised as follows:

- To establish a sustainable ecosystem of linguistic resources and AI technologies that support the use of Basque in digital environments.
- To ensure equitable access to advanced speech and language processing tools for Basque, comparable in quality and scope to those available for high-resource languages.
- To promote collaboration between academia, industry, and public institutions in the creation, governance, and exploitation of linguistic data and models.
- To contribute to the European objectives of multilingual digital inclusion and linguistic data infrastructure interoperability.

An initial workplan has been established until 2030 with the following specific objectives:

- Compilation of large-scale **multimodal corpora** in Basque, covering diverse dialects, registers, and domains of use.
- Development and open release of **foundational models** for speech recognition, synthesis, translation, and language understanding in Basque.
- Establishment of a **dedicated infrastructure** for the storage, versioning, and validation of linguistic datasets as well as the evaluation of AI models.
- Integration of Basque language technologies into **key strategic sectors**, such as education, health, industry, public administration, and media, fostering real-world adoption.
- Formulation of a **methodological framework** transferable to other minority and low-resourced languages, consolidating Euskorpora as a reference model for inclusive AI development.

[24] https://www.euskadi.eus/digitalanaiz-estrategia-para-la-transformacion-digital-de-euskadi/web01-a2lehiar/es/

4.2 Approach

Euskorpora adopts an integrated and interdisciplinary approach that bridges linguistic research, artificial intelligence, and digital infrastructure development to strengthen the technological ecosystem of the Basque language. Grounded in collaboration among academic, institutional, and industrial actors, the initiative encompasses the entire cycle of language technology creation, from corpus compilation and data annotation to the training, evaluation, and deployment of foundational models. Its methodology is guided by three core principles: **comprehensiveness**, addressing all stages of resource and model development; **openness**, ensuring that data and models are accessible to foster collaboration and transparency; and **alignment**, harmonising efforts with regional, national, and European frameworks for digital innovation and multilingual AI. By integrating these dimensions, Euskorpora seeks to create a sustainable, interoperable, and inclusive environment for the advancement of Basque within the global landscape of language technologies.

A key step on the setting up of Euskorpora's ecosystem is the integration of members in both its decision-making and its data-provision and technology-deployment bodies.

As part of this member ecosystem, Euskorpora has the privilege to count on the presence of the Royal Academy of the Basque Language (*Euskaltzaindia*)[25], which is the official body responsible for the Basque language.

5. Euskorpora's Status and Ongoing Work

5.1 Data Production

The data production work addressed within Euskorpora comprises several steps that are described in the upcoming subsections, where the 2025 achievements are also presented.

5.1.1 Data Negotiation and Clearing

As described earlier, stakeholders joining the Euskorpora association define the data they can contribute towards the constitution of the target corpus *Euskorpus* in collaboration with Euskorpora's team. This corpus is being built with all the different types of data (audio, video, text, for different domains and language varieties) that are collected from the different data providers and later processed for LT development.

The first step involves a strategical negotiation stage during which Euskorpora:

1. introduces its mission and objectives;
2. describes the full workplan to potential stakeholders individually;
3. provides an overview on LTs and the use of AI in their businesses with the aim of guiding the data providers to understand the benefits of embracing LT in their services;
4. reassures data providers guaranteeing full compliance of the legal framework defined for their data.

Stakeholders who are interested in joining the association:

- Establish a licence defining their specific terms of collaboration and the conditions for the donation of their data.
- The licence established between each data provider and the association defines the **type of data** that will be donated as well as the **conditions of use**. A data provider may require specific conditions of use, such as not using their data in certain domains or for a certain application. One such example is the case of audio data provided by the Basque Radio and Television (EiTB)[26], institution specifically stating that their data should not be used for voice cloning.
- These discussions count on the presence of Euskorpora's legal experts to guarantee the clear and compliant establishment of the licences.
- The data provider's legal team is also invited to the discussions so as to both establish a collaboration rooted in mutual trust and define the conditions of use in a reassuring collaborative and transparent manner.
- Licences are then tailored to meet the specific requirements of each collaboration and to secure the above-mentioned relationship of trust and transparency between the involved parties.
- Private organisations and industrial stakeholders pay a fee to become members of the association while public and academic institutions join for free.
- The stakeholders' data contribution to the association is free of charge. The objective is to support Euskorpora's mission and objectives and benefit from its outputs and results in the long term, as well as be a part of the technology ecosystem that is being set-up.
- The provision of language data is agreed upon on a continuous basis, which allows for a long-term enrichment of the corpus being built under a controlled versioning protocol.

5.1.2 Data Collection

Data collection is planned to be done through an online platform, called "*Bildaltzaile*" ("Sender" in Basque), which has concluded its beta version. Each Euskorpora's member that has relevant content to contribute will receive a unique set of credentials to provide their data. They will log into the platform and upload the content they wish to

[25] https://www.euskaltzaindia.eus/

[26] https://www.eitb.eus/es/

share. The data will be received on Euskorpora's end as a raw data collection to be further processed, both manually and automatically, to be organised, classified and curated.

In the meantime, data contributions are done through Wasabi[27], a cloud space which has allowed starting work on data collection and annotation while developing the platform.

During 2025, both speech and text raw data have been collected from some of Euskorpora's members and collaborators. Below follows a selection of it:

Regarding **speech**:

- Almost 2,000 hours of both speech and video data were collected from the Basque public broadcast service EiTB.
- 4,000 hours of radio speech have also been contributed by EiTB.
- Further speech and video contents have also been provided by other members of the association and are currently being revised and categorised.

Regarding **text**:

- EiTB has also contributed large numbers of their web content (articles, news).
- Members like Baleuko, Mondragon Korporazioa, Euskaltel, CAF, Innobasque, Petronor, Gureak and the University of the Basque Country (EHU) have started providing a wide variety of written contents, such as articles, manuals, books, reports, regulations, scripts, etc.
- Over 2,000 scientific articles from EHU's Digital Archive for Teaching and Research (ADDI)[28] have been collected holding an open unrestricted-use CC licence.

In 2025, Euskorpora's work on data annotation focused on audio transcription of a selection of EiTB's television programmes (ranging from news, sports, interviews, and cultural programmes, among others): 1,396 hours of annotated speech were accomplished out of the ~2,000 hours collected (cf. next section).

Work on the transcription of the radio audio has been initiated in 2026 and, as of today, several teams are currently working on it: an initial set of 490 hours is targeted by the end of April 2026.

5.1.3 Data Annotation

To create certain types of datasets, data must be annotated. The required annotation varies depending on the nature of the data and its intended purpose. For example, audio data containing voice must be segmented into speaker turns, transcribed and otherwise be characterised. Or, for example, monolingual texts might be translated to/from Basque to obtain translation datasets that can be used in either MT system training or evaluation. Further, untranslated monolingual Basque texts may also be processed to constitute the digital corpus of Basque *Euskorpus*.

The annotation tasks need to be carefully designed, which include the devise of their guidelines, the formation of the human labellers, and the selection of the appropriate labelling tools and formats.

The **transcription** workflow followed by Euskorpora on EiTB's speech data in 2025 can be outlined as follows:

- All data were pre-segmented and pre-transcribed with Trebe's transcription tool[29].
- Transcription guidelines were defined which are language specific and cover the transcription of several Basque dialects (known as "**euskalkis**"[30]). Basque has multiple geographical varieties which differ in pronunciation, vocabulary and grammar. Although there is a standard variety named **batua** (unified), daily life preserves Basque language diversity, which is an additional challenge for language technology systems. Our work addresses the following varieties: **gipuzkera, bizkaiera, nafarrera, nafar-lapurtera** and **zuberera**.
- Guidelines for the use of the ***Transcriber*** transcription tool (Barras et al., 2001) were developed. *Transcriber* is the tool chosen for transcription, but a more evolved and user-friendly *Transcriber*-inspired annotation tool is currently being developed to 1) ease usability (web-based) and 2) address *Transcriber*'s limitations in terms of functionalities.
- Guidelines for the validation of the transcription work were also defined.
- Euskorpora published public tenders to set-up external teams of Basque language experts to revise all pre-transcriptions manually.
- Revision and transcription work through external teams involved training language experts who were, generally, linguistic experts from the translation and interpreting domains. This also allowed to convert language experts to a new professional field.
- Six companies were selected and subcontracted under this public

[27] https://wasabi.com/es
[28] https://www.ehu.eus/es/web/biblioteka/addi-artxibo-digitala
[29] https://www.trebe.org/es/
[30] https://eu.wikipedia.org/wiki/Euskalki

procedure, and they set up teams comprising 85 language experts to work on the task.
- Euskorpora supervised and validated the external language experts' work with its in-house team.
- As earlier mentioned, 1,396 hours of video speech were achieved in 2025 using this procedure.

Further audio transcription is currently being done (on the radio audio content described in the previous section) as part of Euskorpora's objectives for 2026. These annotated speech data will be part of the LLM development strategy which is currently being deployed.

Regarding the preparation of language data for **text** technologies, over 1,600,000 words of Spanish text from ~700 documents from the ADDI repository have been selected for translation into Basque with the aim of creating both training and evaluation datasets for Spanish2Basque Machine Translation. These benchmarks will be safeguarded by Euskorpora who will lead the evaluation tasks, thus, avoiding evaluation data contamination.

The following four domains have been chosen for this task:

- Cultural Studies and Anthropology
- Education and Pedagogy
- History and Archaeology
- Public Health and Epidemiology

These data were extracted from academic documents such as articles and theses which were originally published in PDF format and had to undergo a conversion and cleaning pre-processing. Content was gathered in a balanced manner from the documents used, trying to obtain a varied and rich representation from the documents covering those domains. Translation work is starting in March 2026.

As with the work of translators in transcription tasks, translators require specific guidance in the production of this type of translations which will be used for technological development. This is also part of Euskorpora's workflow, together with the definition of guidelines, publication of public tenders and final quality control (final validation once the translation team has finished their work).

5.1.4 Data Validation

Annotated data is reviewed and validated by humans to assure it complies with the labelling guidelines and other quality standards. A validation protocol has been defined for this purpose for each of the annotation tasks, i.e., audio transcription and text translation.

Data validation involves manual revision and automatic checks (the latter, for instance, in formal/format aspects). The objective is to progressively provide further automated assistance to human reviewers, to better scale the efforts, but never fully replacing the final human validation.

Given the newly-trained nature of the transcription revision teams, the implementation of this quality control phase has been twofold:

- Transcription-revision work during the first steps of the project has been fully revised and corrected by Euskorpora, providing direct feedback to the transcription-revision teams for improvement and as part of their training.
- Once teams are fully operational, they undergo the validation of their work based on representative sampling. Any mistake still found is brought back to the language expert team(s) having done the work.

In the assessment of text translation, validation guidelines have also been drafted to allow for both a qualitative and a quantitative validation. These guidelines focus on methodology (e.g., quality assurance protocols) as well as the content to apply the methodology.

5.2 LLM Development

With the generated corpora and datasets, Euskorpora will train and develop open base models for different tasks, potentially in collaboration with other relevant agents of the Basque ecosystem. The objective is to turn the gathered content into something actionable, and to explore the possibilities of the technology at each step leveraging the available data. Approaches will be explored considering the context of the Basque language and devising techniques to address data scarcity, such as adopting generic LLMs to verticals' needs where data availability may be lesser and may require adapted strategies.

Currently, MT system development and evaluation tasks are being defined in agreement with known needs of MT players working on Basque. Being a co-official language in Spain, both Basque Government and Osakidetza (Basque Public Health System) are providing MT solutions to the community for the Spanish-Basque language pair with the Itzuli[31] and Itzulbide[32] MT systems, respectively. The former addresses the general public and the latter is a domain-specific tool to help doctors in clinical test translation.

Therefore, it is foreseen that LLM development will explore the technological needs of the stakeholders providing data to Euskorpora and will implement LLM development and technology deployment in collaboration with different research and development players.

[31] https://www.euskadi.eus/itzuli/

[32] https://www.orai.eus/en/successful-cases/itzulbide

Examples of models to be explored and trained using newly gathered data might be base GPT models of different sizes, machine-translation models for language pairs such as Spanish-Basque, English-Basque and French-Basque. For instance, and in addition to the above-mentioned MT systems for Basque, the Latxa LLM family (derived from Llama 2) developed for Basque by the the Basque Center for Language Technology HiTZ (Etxaniz et al., 2024; Sainz et al., 2025) is also being explored, together with the work of the Vicomtech Language Technology experts on this area (Ponce et al., 2024; Ponce et al., 2024b).

Finally, regarding stakeholder-oriented actions, an event was organised by Euskorpora for its members and collaborators in November 2025. This event allowed to discuss data provision and needs as well as LT service needs and developments from the involved stakeholders. The conclusions from these exchanges have been drafted as part of Euskorpora's future roadmap for collaboration, service/technology development and data production.

5.3 Data Processing Platform

The data processing will be managed through a platform called "Gailu" ("Device" in Basque), which has released its version 1. This platform will allow to configure data processing pipelines, composed of a mixture of automatic and manual steps. Each pipeline starts from a collection of received data, and ends when the data is properly curated, annotated and validated, and ready to be part of a final data asset (a corpus or a dataset). The objective is to keep developing technology to automate, systematise as many steps as possible, to allow the data processing workflow scale and alleviate the need of human intervention.

5.4 Data Sharing and Monetising

Euskorpora aims to share all the resulting assets via a catalogue. This includes all the gathered corpora and annotated datasets, and any AI model derived from the data. This will contribute to the advancement of the multilingual European digital data market as well as support both public and private institutions to work on LTs and develop AI services for Basque.

For this purpose, Euskorpora has joined the **European Language Data Space** (LDS)[33] (Piperidis et al, 2026).

From a **technical point of view**, this will allow to leverage all the tools provided to make the resources findable, and to expose them in a controlled way, under certain terms of use and prices.

Most importantly, from a **strategical point of view**, Euskorpora endorses and ensures the digital sovereignty and inclusion of the Basque language within the European AI ecosystem. After formally being approved by the European Commission and the LDS, Euskorpora is currently working on the adoption and installation of an LDS connector following the recently launched LDS infrastructure v3.0.0. This is a key action to support linguistic diversity, technological innovation, and equitable digital transformation in multilingual Europe.

In parallel, Euskorpora is designing its own catalogue to serve as a portfolio of the generated resources, pointing to the LDS in an interoperable manner for any further interaction or purchase.

6. Conclusions

Euskorpora constitutes a landmark initiative in the advancement of digital infrastructures for minority and low-resourced languages, combining scientific rigour with strategic foresight. Conceived within the broader European and national frameworks for language technologies, it implements the objectives of linguistic equality, digital sovereignty, and technological innovation by providing the Basque language with the necessary data, models, and infrastructures to thrive in the age of artificial intelligence.

From a methodological perspective, Euskorpora exemplifies a comprehensive approach to language technology development, encompassing the full lifecycle of data management, going from collection, annotation, and validation to model training, evaluation, and deployment. Its alignment with European infrastructures such as the European Language Data Space (LDS) and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable)[34] ensures both scientific robustness and long-term sustainability.

Strategically, the initiative embodies the convergence of language policy and digital innovation, reflecting the commitment of the Basque Government (*ETDE2025*) and Spanish national frameworks (*Plan TL, PERTE Nueva Economía de la Lengua*) to inclusive digital transformation. Euskorpora's integrative design, connecting academia, industry, and public institutions, reinforces its potential as a model of governance and cooperation for the sustainable development of language resources.

In conclusion, the initiative goes beyond the mere production of technological assets: it establishes the foundations of a digital ecosystem that safeguards the linguistic and cultural identity of Basque while contributing to the European vision of a multilingual, AI-driven society. By doing so, Euskorpora not only enhances the technological resilience of Basque but also sets a replicable

[33] https://language-data-space.ec.europa.eu/index_en

[34] https://www.go-fair.org/fair-principles/

precedent for fostering linguistic diversity and equity in the digital era.

7. Acknowledgements

Euskorpora is funded by the Members of the Association (Baleuko, BBVA, CAF, Deustuko Unibertsitatea, Ereil, Euskal Herriko Unibertsitatea, Euskaltel Fundazioa, Euskaltzaindia, Eusko Jaurlaritza (Basque Government), Iberdrola, Gipuzkoa Foru Aldundia, Kutxabank, Laboral Kutxa, Logicaline, Mixer Servicios Audiovisuales, Mondragon Korporazioa, Mondragon Unibertsitatea, Petronor, PWC and Vicomtech) and is also supported by its Collaborators (Azkue Fundazioa, Basque Artificial Intelligence Center -BAIC-, Basque Research & Technology Alliance -BRTA, Elhuyar, Gureak, Innobasque, Langune and Trebe).

The Euskorpus project is funded by the Basque Government (Eusko Jaurlaritza) in Spain.

8. Bibliographical References

Barras, C., Geoffrois, E., Wu, Z., Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication 33* (Vol. 1-2): pages 5--22.

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M. and Soroa, A. (2024). Latxa: An Open Language Model and Evaluation Suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 14952--14972. Bangkok, Thailand.

Piperidis, S., Labropoulou, P., Galanis, D., Choukri, K., Vasiļjevs, A., Deligiannis, M., Gkirtzou, K., Gkoumas, D., Kolovou, A., Voukoutis, L., Pouli, K., Giagkou, M., Gavriilidou, M., Marheinecke, K., Leitner, E., Ostermann, S., Raccioppa, S., Talmoudi, K., Arranz, V., Mapelli, V., Mazo, H., González Campo, F., Yu, S., Bērziņš, A., Lagzdiņš, A. and Rehm, G. (2026). Common European Language Data Space: Development, Current Status, and Future Perspectives. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Ponce, D., Etchegoyhen, T., Calleja, J. and Gete, H. (2024). Split and Rephrase with Large Language Models. In In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 11588–11607. Bangkok, Thailand.

Ponce, D., Gete, H. and Etchegoyhen, T.. (2024b). Vicomtech@WMT 2024: Shared Task on Translation into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, pages 934–942, Miami, Florida, USA. Association for Computational Linguistics.

Rehm, G. and Way, A. (ed.) (2023). A Strategic Agenda for Digital Language Equality. Cognitive Technologies. Springer, Cham, Switzerland.

Sainz, O., Perez, N., Etxaniz, J., Fernandez de Landa, J., Aldabe, I., García-Ferrero, I., Zabala, A., Azurmendi, E., Rigau, G., Agirre, E., Artetxe, M. and Soroa, A. (2025). Instructing Large Language Models for Low-Resource Languages: A Systematic Study for Basque. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.

Sarasola, K., Aldabe, I., Díaz de Ilarraza, A., Estarrona, A., Farwell, A., Hernáez, I. and Navas, E. (2022). European Language Equality (ELE). *Deliverable D1.4: Report on the Basque Language*.