LINGUATEC-IA: Research and development of Artificial Intelligence for the low-resourced languages of the Pyrenees

Itziar Aldabe[1], Itziar Aduriz[1], Xabier Arregi[1], Aitzol Astigarraga[2,*], Myriam Bras[3], Pauline Charrier[4], Itziar Cortes[2], Urtzi Etxeberria[5], Igor Leturia[2], Matthieu Martel[6], Miguel Angel Pelles[7], Aure Séguier[4] and Jordi Suïls[8]

[1]*HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia*
[2]*Orai NLP Technologies - Elhuyar, Usurbil*
[3]*Université Toulouse-Jean Jaurès, Toulouse*
[4]*Lo Congrès permanent de la lenga occitana, Pau*
[5]*IKER-UMR5478, Bayonne*
[6]*LAMPS - Université Perpignan Via Domitia, Perpignan*
[7]*Patrimonio Cultural, The Directorate-General of Cultural Heritage of the Government of Aragon*
[8]*SoGEL - University of Lleida (UdL), Lleida*

Abstract

This paper presents the LINGUATEC-IA project, an ongoing initiative (2024–2026) aimed at advancing artificial intelligence applications for low-resource languages in the POCTEFA region (Aragonese, Catalan, Basque, and Occitan). The project focuses on developing specialized language models that can operate effectively with limited linguistic resources. The key objectives are to enhance transcription, machine translation, and speech synthesis systems for these languages in conjunction with French and Spanish; creating an automatic subtitling and dubbing platform; establishing an online repository for Pyrenean language resources; and strengthening a cross-border network for language technology excellence.

Keywords

Low resource languages, Large Language Models, Speech Synthesis, Machine Translation, Cross-border linguistic infrastructure

1. Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized various fields, particularly natural language processing (NLP). However, most research and technological development focus on widely spoken languages, which has left many regional and minority languages underrepresented. The LINGUATEC-IA project[1] aims to address this gap by developing neural language models tailored for low-resource languages in the POCTEFA region, specifically Aragonese, Catalan, Basque, and Occitan. The primary goal is to improve digital accessibility and establish a cross-border language technology infrastructure to support multilingual communication and information access. The LINGUATEC-IA project is part of the Interreg VI-A Spain-France-Andorra Programme (POCTEFA 2021-2027), which aims to strengthen the economic and social integration of the Spain-France-Andorra border region.

Accordingly, the LINGUATEC-IA project addresses three key challenges of POCTEFA: 1) Innovation Challenge 1 by increasing territorial innovation through AI and language technologies; 2) Territorial Challenge 1 by supporting cultural integration via digitalization of minority languages (Aragonese, Catalan, Basque, and Occitan), preserving cultural heritage while enabling communication across six

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.
*Corresponding author.
†These authors contributed equally.
a.astigarraga@orai.eus (A. Astigarraga)

[1]https://linguatec-ia.eu/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-04

languages; and 3) Social Challenge 3 by strengthening cross-border labor markets through improved language learning tools that facilitate worker mobility. The POCTEFA Programme recognizes languages as a key territorial strength, and this project aims to develop these resources through targeted technological innovation.

2. Background: LINGUATEC

The European project EFA 227/16/LINGUATEC "Development of cross-border cooperation and knowledge transfer in language technologies" (2018-2020)[2] [1] established a consortium composed of 6 entities working together in the development and dissemination of new innovative language resources, tools and applications to improve the digitalization of Aragonese, Basque and Occitan.

As a result, a series of tools, applications, and linguistic resources were developed designed to facilitate communication and break language barriers across various languages.[3] For Basque, technologies for speech recognition and improvements in Spanish-Basque automatic translation were created, in addition to Euskara Eskuz Esku, a tool for consulting linguistic norms. In Aragonese, notable developments include voice synthesis, TRADUZE to improve automatic translation, the online dictionary ARAGONARIO, and various multilingual applications oriented toward tourism and the Camino de Santiago. In Occitan, monolingual and bilingual lexicons were developed, along with morphosyntactic and syntactic analysis tools, as well as advanced voice synthesis systems (VOTZ) and speech recognition (ReVOc), together with improvements in French-Occitan automatic translation. Finally, in the multilingual domain, automatic translation applications between the languages of the Pyrenees were implemented, available on Google Play and AppStore, as well as browser extensions and translation tools for websites and content management systems (CMS). These developments represented a significant advance in cooperation between languages and access to information in a more inclusive digital environment in 2020.

The level of development achieved in the project encouraged the institutions belonging to the consortium to take a strategic step: creating a network of excellence in AI, to create a cross-border linguistic infrastructure. The LINGUATEC-IA project is a result of this network. This new initiative builds upon the success of the previous project, aiming to develop more sophisticated natural language processing capabilities. Thus, cooperation has been extended to new entities, languages, and territories, and new goals have been set in terms of innovation.

3. Consortium

The LINGUATEC-IA project brings together a diverse consortium of academic institutions and organizations focused on developing language technologies for the mentioned languages. The LINGUATEC-IA consortium is composed of 8 partners, with ELHUYAR serving as the lead coordinator and LO CONGRÈS PERMANENT DE LA LENGA OCCITANA as co-coordinator. The partnership includes five academic institutions (University of the Basque Country's HITZ center, Université Toulouse Jean Jaurès, Université Perpignan Via Domitia, IKER-CNRS, and Universitat de Lleida) and one government entity (Government of Aragon).

The project requires these 3 profiles, because it is not only about generating knowledge, but as stated in POCTEFA, this knowledge must be transferred and applied to address the challenges of the cross-border region. One of these challenges is the need to combine the preservation of all the languages of the POCTEFA space with the improvement of the intercommunication between all of them, taking advantage of the opportunities offered by digitalization.

The roles and specializations of the consortium members are outlined below:

ELHUYAR[4]: The primary beneficiary and project coordinator, handling administrative and

[2] https://linguatec-poctefa.eu

[3] https://linguatec-poctefa.eu/recursos/

[4] https://www.orai.eus/

financial management. Its AI center, Orai, specializes in machine translation, linguistic resources, text mining, and speech technologies, with a focus on Basque and customised projects for companies and organisations.

LO CONGRÈS PERMANENT DE LA LENGA OCCITANA[5]: Co-coordinator and the interregional regulatory body for Occitan, focused on strengthening knowledge and codification of the language through various tools.

HITZ (University of the Basque Country)[6]: A reference center for language technologies with 80+ specialists focusing on AI for language and speech, particularly for Basque and other low-resource languages.

Université Toulouse Jean Jaurès[7]: A major center for Occitan language research and teaching, contributing expertise through its OCRE research group.

Université Perpignan Via Domitia[8]: Represented by LAMPS, contributing research in automatic language processing for minority languages, especially Catalan.

IKER-UMR5478[9]: The only research center specializing in Basque studies in France, bringing expertise in Basque linguistics and language processing.

Universitat de Lleida[10]: The reference university for western Catalan territory, contributing research on geographical and social language dynamics through its SoGeL research group.

Government of Aragon[11]: Participating for the first time through its General Directorate of Cultural Heritage, promoting the inclusion of Aragonese in digital language technologies.

This consortium represents a cross-border collaboration aimed at advancing language technologies for the preservation and development of Pyrenean languages.

4. Objectives and Scope

The project focuses on multiple key objectives that will drive technological progress and linguistic preservation that builds upon the work started in LINGUATEC:

1. Research on the development of neural language models to suit low-resource language settings.
2. Improve transcription systems, neural machine translation, and speech synthesis for Aragonese, Catalan, Basque, and Occitan. Similarly, to develop multilingual models that facilitate communication between these languages and major languages such as French and Spanish.
3. Develop a prototype of an automatic subtitling and dubbing platform between the project's languages.
4. Create an online repository of resources, technologies, and applications for the languages of the Pyrenees.
5. Consolidate the 'Cross-Border Network of Excellence in Language Technologies.'

5. Methodology

To achieve these ambitious goals, the project employs a structured methodology integrating collaborative research, iterative development, and coordinated implementation. The common methodology for all activities is based on a structured and collaborative approach, beginning with the establishment of a Working Team formed by technicians from partners and external experts. This team developed a

[5] https://locongres.org/
[6] https://www.hitz.eus/
[7] https://www.univ-tlse2.fr/
[8] https://lamps.univ-perp.fr/
[9] https://iker.cnrs.fr/iker/
[10] https://www.sogel.udl.cat
[11] https://www.aragon.es/organismos/departamento-de-educacion-cultura-y-deporte/direccion-general-de-patrimonio-cultural

detailed Work Plan that defined the phases, schedule, and responsibilities, holding regular meetings for follow-up and preparing semi-annual Progress Reports that have to be validated by the Technical Working Group. The tools and applications are developed iteratively, with continuous feedback cycles to refine outputs, generating progressively improved versions of each tool.

To implement this methodology effectively, five strategic actions have been defined that cover all aspects of the project: from management, communication, and dissemination, to the development and implementation of the systems. These actions are designed to ensure efficient coordination between the different components of the project, ensure fluid communication among all participants, maximize the dissemination of results, facilitate the technical development of the systems, and allow for successful implementation that meets the established objectives.

ELHUYAR acts as the leading partner and leads Actions 1 (Management) and 5 (Dissemination). LO CONGRES leads Actions 2 (Communication) and 4 (Development) and HiTZ leads Action 3 (Research). The other partner entities lead some of the Activities within the actions and participates in cooperation in all actions and activities. The actions are detailed in greater detail below.

5.1. Action 1: Project Management

Project Management is structured in three core activities. The administrative management establishes the necessary structure through committees and working groups, secures technical assistance, and develops essential planning documentation. The financial component maintains fiscal discipline through regular monitoring, standardized reporting procedures, and careful tracking of ERDF funds while ensuring regulatory compliance. Finally, the project monitoring keeps everything on track through systematic progress reports, a comprehensive dashboard of indicators, and regular quality checks on deliverables. This integrated framework allows the consortium to effectively manage all administrative, legal, and financial aspects while remaining in accordance with the POCTEFA requirements.

5.2. Action 2: Communication

The objective of the action is to attract the interest of key target groups by effectively promoting the solutions developed within the project, as well as sharing its activities and results. The aim is to ensure a broad dissemination and visibility in a realistic, achievable, and measurable way. The primary target audience includes researchers in linguistic technologies, individuals involved in regional language development, local media, and the general public.

5.3. Action 3: Research

This action focuses on researching and developing language models for low-resource languages such as Basque, Catalan, Occitan, and Aragonese. A key objective is to explore new AI techniques that reduce the need for large datasets and computational resources. This is especially important given that not all these languages have the same amount of available data. Some languages, such as Aragonese and Occitan, are significantly more under-resourced, requiring tailored strategies and adapted tools. The work is structured into four main activities: **1)** gathering and preparing as much quality text as possible in each language, including cleaning and organizing the data for training and evaluation; **2)** developing generative language models tailored to each language's specific context, with a strong experimental component to account for data scarcity; **3)** evaluating these models using language-specific benchmarks to measure accuracy, usefulness, and linguistic quality—focusing on automated evaluation wherever possible; and **4)**, applying the models in real-world scenarios by building optimized, user-friendly demonstrators that show their practical value and make them accessible to end users.

5.4. Action 4: Development

This action aims to develop innovative resources and applications that support the use and preservation of the POCTEFA region's languages by leveraging cutting-edge neural models. The goal is to improve

translation, transcription, and speech synthesis, fostering multilingualism and digital inclusion. The work is organized into four main activities. First, the digital roadmaps for each language have been updated using the European Language Equality [2] framework, identifying current resources and outlining priorities for the next three years. Second, the project will create multimodal linguistic resources—text and speech data—especially for the most under-resourced languages, ensuring models have the material needed for quality training. Third, new tools and technologies will be developed or improved to enhance digitalization and interoperability, such as better machine translation systems, speech technologies, and neural models tailored to each language. Finally, practical applications will be built to boost the real-world use of these languages, including translation and speech tools, and platforms for subtitling and dubbing—contributing to a dynamic, multilingual digital ecosystem across the region.

5.5. Action 5: Dissemination and Consolidation

This Action aims to share the project's key results and establish a "Center of Excellence in Language Technologies of the Pyrenees" by turning linguistic research into practical multilingual tools for the POCTEFA region. It targets researchers, institutions, media, and the general public interested in technologies like machine translation, speech synthesis, and multilingual communication. Building on the LINGUATEC project, it expands collaboration among regional universities and seeks broader international engagement. The Action includes developing a resource platform, hosting dissemination events, and strengthening the already established cross-border excellence network.

Although each Action is led by one entity, at least two partner entities participate in all Activities. Additionally, the actions and the activities follow a logical process in which all are interrelated and necessary to achieve the expected results.

6. Preliminary Analysis

At the beginning of the project, it was considered appropriate to carry out an analysis of the availability of data, resources, and tools that would enable research and development in AI focused on these languages. The digital roadmap described for each language is the result of this need. The preliminary analysis revealed that the amount of text corpora available is limited for the development of robust large language models (LLMs), especially for Occitan and Aragonese. In contrast, speech and parallel corpora are comparatively more abundant, enabling the development of more advanced speech and machine translation (MT) technologies. Based on this analysis, actions 3 and 4 have been undertaken, whose status and objectives are described below.

6.1. Action 3: Research

Regarding the development of generative language models, the initial analysis confirmed the original hypothesis: the situation of Catalan and Basque is not comparable to that of Occitan and Aragonese. In the cases of Catalan and Basque, not only are data and tools available, but there are also research and technological centers specialized in the field of language-centered AI. These centers have already developed generative language models for Catalan and Basque and are actively collaborating on various projects [3, 4, 5].

In contrast, Occitan and Aragonese lack comparable support. With regard to data availability, it became evident that the size of the existing text corpora is extremely limited. Table 1, Table 2 and Table 3 show the volume of accessible textual corpora for Aragonese and Occitan. In the case of Occitan, the wide dialectal variety is worth highlighting, as it constitutes an additional difficulty.

These corpora contain significantly fewer tokens than typically required for training robust language models, even in the context of low-resource languages. This situation raises complex and thought-provoking challenges for the project's advancement.

Table 1
Size of the Aragonese text corpus

	Sentences (#)	**Words (#)**
Transcriptions	93,400	456,100
Student and other texts	57,900	887,400
Biquipedia	452,300	8,750,000
Entries from the Aragonario dictionary		
Aragonese (total)	603,600	10,093,500

Table 2
Size of the monolingual corpus in Occitan by variety

	Words (#)
Gascon	7,284,487
Languedocien	5,837,066
Limousin	904,763
Provençal	392,209
Vivaro-Alpine	356,722
Auvergnat	72,305
Aranese	60,745
Occitan monolingual (total)	**15,200,616**

Table 3
Size of the bilingual corpus in Occitan by variety

	Words (#)
Languedocien–French	1,120,569
Gascon–French	926,711
Limousin–French	257,062
Vivaro-Alpine–French	68,517
Provençal–French	57,628
Auvergnat–French	5,552
Aranese–French	979
Bilingual Occitan–French (total)	**2,468,664**

Therefore, with respect to the development of generative language models for Occitan and Aragonese, several issues arise. Continued pre-training based on a multilingual foundational model might be an adequate way to address this challenge, but it is essential to increase the size of the textual corpus to enable this approach. One technique to augment the available data would be the generation of synthetic data through machine translation. With regard to the training process, it initially seems advantageous to start from models that perform well in French and Catalan, given the linguistic proximity of these languages to Occitan and Aragonese.

The evaluation of the models developed in the project is another significant challenge. The scarcity of data, specifically the lack of evaluation datasets, is once again a crucial factor. Moreover, given the limitations, scenarios will need to be defined to evaluate the performance of the models on specific tasks, such as translation or text summarization.

6.2. Action 4: Development

This action focuses on the development of speech and MT technologies, leveraging the comparatively greater availability of such corpora across the languages studied. The digital roadmaps for each language

have already been updated.

Work is currently underway to collect and expand speech corpora, which will be used for training and improving speech technologies. In the case of Basque, there are already substantial resources available to support high-quality ASR (Automatic Speech Recognition) and TTS (Text-to-Speech) systems. Current efforts aim to further enhance these systems, particularly for informal and dialectal speech, and to create emotional and dialectal speech corpora for TTS applications. For Occitan and Aragonese, the focus is on collecting new audio recordings along with transcriptions to support ASR development. This involves both manual transcription of media content and crowdsourced contributions. Special attention is being given to the Aranese variant of Occitan, for which new transcribed audio data is being gathered. Additional TTS recordings using new voices, particularly in Aranese, are also being planned.

Technology development efforts include improvements to the existing ASR system for Basque, especially in handling informal and dialectal speech. This includes the integration of new corpora and a migration from Kaldi[6] to Whisper[7]. For TTS, work is focused on emotional and dialectal speech synthesis, as well as voice cloning. For Occitan and Aragonese, existing ASR systems[8] will be upgraded by incorporating additional speech corpora and migrating to Whisper technology. A dedicated ASR system will also be developed for the Aranese variant. Given Whisper's robust multilingual pretraining, it offers excellent performance for low-resource languages when fine-tuned with relatively small datasets. Existing TTS systems for Occitan[9] and Aragonese will be migrated to the FastPitch-HiFiGAN framework, and support for the Aranese variant will be added. The project also aims to improve voice quality and diversity across all Occitan variants by building a single, multispeaker, multilingual TTS model.

Regarding MT system development, current models already demonstrate solid performance, but future work will focus on further enhancing their quality. This will be achieved by expanding training data through a strategic combination of authentic and synthetic resources, leveraging automatic translation of high-resource corpora. Continued training using transformer architectures and systematic evaluation on established benchmarks will guide ongoing improvements.

7. Expected Impact

The implementation of this project is expected to bring benefits to both linguistic communities and the field of technological research. One of the key impacts will be to help on the digital preservation of regional languages, as the development of AI-driven language tools will help preserve the relevance and usability of Aragonese, Catalan, Basque, and Occitan in the digital era. This contributes to the digital preservation of cultural heritage of these languages, making them accessible to future generations. Additionally, the enhanced multilingual communication enabled by neural translation, speech synthesis, and other technologies will foster smoother interactions among different linguistic groups, promoting greater inclusivity and understanding.

From a technological point of view, the project will develop new data and tools for low-resource languages, where there is often a lack of such resources. The algorithms and models developed through this initiative will not only serve the POCTEFA region but will also contribute to the global AI research community, offering new insights into the challenges and solutions for under-resourced languages.

On a broader level, the project is expected to have a positive economic and social impact. By improving linguistic accessibility, it will support cross-border collaboration, foster cultural exchange, and create new economic opportunities within the POCTEFA region. This could ultimately help bridge gaps between communities and enhance regional cohesion, contributing to a more integrated and dynamic multilingual environment.

8. Conclusion

This project represents an effort to bridge the digital divide for low-resource languages using cutting-edge AI technologies. By developing language models, enhancing translation and speech synthesis systems, and fostering a cross-border linguistic infrastructure, this initiative plays a crucial role in the preservation and modernization of the languages spoken across the POCTEFA region. As AI continues to evolve, projects like this try to ensure that all languages, regardless of their resource availability, have a place in the digital world.

Acknowledgments

The LINGUATEC-IA project has been 65% co-financed by the European Union through the Interreg VI-A Spain-France-Andorra Programme (POCTEFA 2021-2027). The objective of POCTEFA is to strengthen the economic and social integration of the Spain-France-Andorra border region.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. Aldabe, J. Aztiria, F. Beltrán, M. Bras, K. Ceberio, I. Cortes, J.-B. Coyos, B. Dazeas, L. Esher, G. Labaka, I. Leturia, K. Sarasola, A. Séguier, J. Sibille, LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los pirineos, Procesamiento del lenguaje natural 63 (2019) 159–162.

[2] I. Aldabe, J. Dunne, A. Farwell, O. Gallagher, F. Gaspari, M. Giagkou, J. Hajic, J. P. Kückens, T. Lynn, G. Rehm, G. Rigau, K. Marheinecke, S. Piperidis, N. Resende, T. Vojtěchová, A. Way, Overview of the ELE project, in: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, p. 353–354, 2022. URL: https://aclanthology.org/2022.eamt-1.66/.

[3] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra technical report, 2025. URL: https://arxiv.org/abs/2502.08489. arXiv:2502.08489.

[4] A. Corral, I. S. Antero, X. Saralegi, Pipeline analysis for developing instruct LLMs in low-resource languages: A case study on Basque, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 12636–12655. URL: https://aclanthology.org/2025.naacl-long.629/.

[5] J. Etxaniz, O. Sainz, N. Miguel, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for Basque, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14952–14972. URL: https://aclanthology.org/2024.acl-long.799/. doi:10.18653/v1/2024.acl-long.799.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi speech recognition toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of

Proceedings of Machine Learning Research, PMLR, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[8] I. Morcillo, I. Leturia, A. Corral, X. Sarasola, M. Barret, A. Séguier, B. Dazéas, Automatic speech recognition for gascon and languedocian variants of occitan, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 1969–1978.

[9] A. Corral, I. Leturia, A. Séguier, M. Barret, B. Dazéas, P. B. de Mareüil, N. Quint, Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of gascon occitan, in: Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop «Language Resources and Evaluation Conference–Marseille–11–16 May 2020», European Language Resources Association (ELRA), 2020.