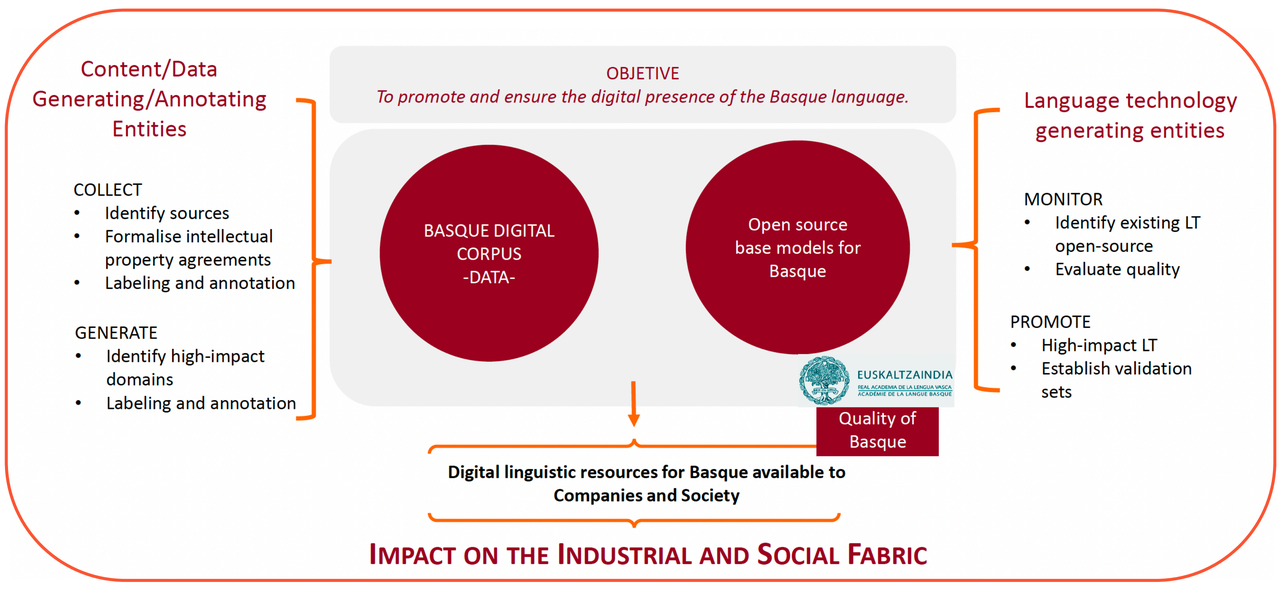
EUSKORPORA: A STRATEGIC FRAMEWORK FOR DIGITAL SOVEREIGNTY AND LINGUISTIC INCLUSION OF BASQUE IN THE ERA OF AI

Victoria Arranz, Sara Arregi, Leire Barañano, Aitor García-Pablos
Euskorpora
{victoria.arranz, sarregi, lbaranano, agarcia}@euskorpora.eus

INTRODUCTION

- Basque is a non-Indo-European language spoken by ~800,000 people.
- Despite strong social presence, it remains under-resourced in digital environments.
- Languages without sufficient technological support risk digital extinction.
- Current AI systems require large-scale, high-quality linguistic data **(Euskorpus project)**.

Euskorpora addresses this gap by building a comprehensive infrastructure for Basque language technologies, combining corpora, models, and data governance.



EUSKORPORA´S INFRASTRUCTURE

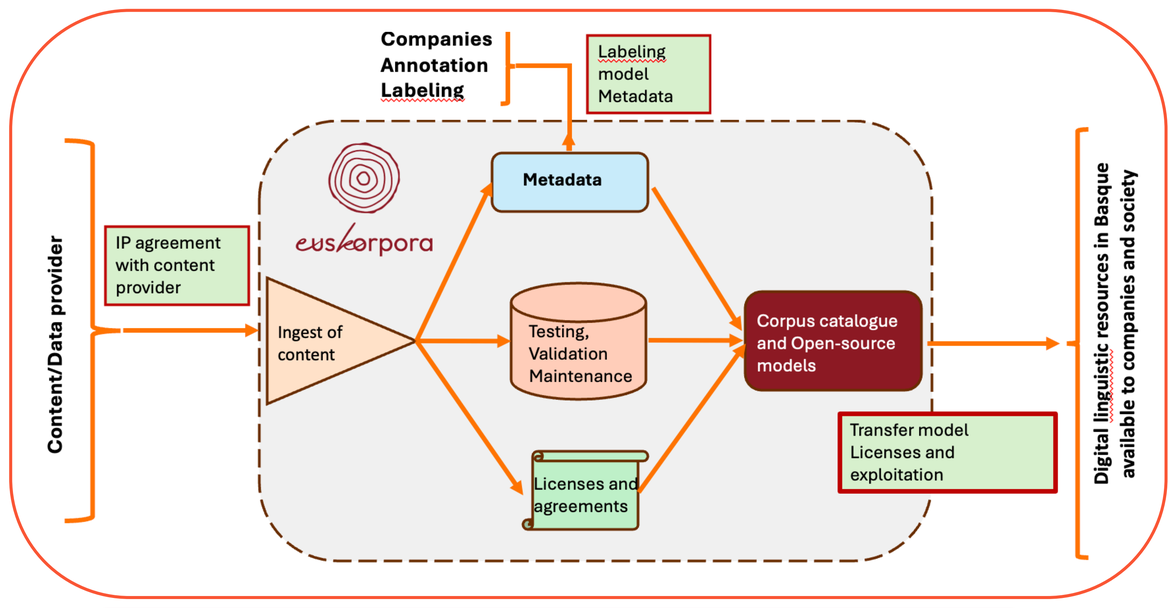
- **Strategic framework:** European, Spanish, and Basque language technology initiatives.
- **Collaborative ecosystem:** public institutions, academia, industry; shared governance and resource contribution.
- **Impact:** pioneering model advancing technologies for low-resource languages.

PARTNERS

COLLABORATORS

CORPUS PRODUCTION

- Data is collected through collaboration with public institutions, academia, and industry.
- Contributions include speech, video, and text data across multiple domains and dialect variations.
- A dedicated platform (“Bildaltzaile”) enables secure data submission.
- Data acquisition follows legal agreements and licensing frameworks ensuring compliance and trust.
- Full corpus production workflow, comprising data processing, annotation, validation, sharing and language model development.



EUSKORPUS STATISTICS

CONTENT TYPE	CONTENTS	SIZE
Speech from EiTB (EUS)	**Collected** (TV and radio)	~2,000 Hours (TV)
		~4,000 Hours (radio)
	Transcribed (TV)	1,396 Hours (TV)
Text (EUS, SPA, ENG) Aim: feed Euskorpus Produce data for MT training and test	**Crawled and processed** data from 3 webs Source: Mondragon Corporation	TULANKIDE: 4739749 words ORAIN: 22834014 words MONDRAGON HEALTH: 87557 words
	ADDI (scientific documents repository) Source: EHU	**Collected**: ~2,000 scientific documents
		Converted, cleaned, categorised, segmented: ~700 documents

CONTENT TYPE	CONTENTS	SIZE
Speech from EiTB (EUS)	**Transcription** (radio)	~900 Hours
Text (EUS, SPA, ENG)	**Crawled and processed** data from 3 webs Source: Mondragon Corporation	Under categorisation (per domain and language) and segmentation Aim: prepare data for translation
Text (SPA) **Collected, converted, cleaned, segmented** **Aim: to be translated into Basque to produce SPA – EUS training data and benchmarks (2 references)**	*Cultural Studies and Anthropology (ADDI)*	Source for train: 336,966 words
		Source for benchmarks: 4,000 segments (126,599 words)
	Education and Pedagogy (ADDI)	Source for train: 83,259 words
		Source for benchmarks: 2,000 segments (75,198 words)
	History and Archaeology (ADDI)	Source for train: 239,635 words
		Source for benchmarks: 4,000 segments (145,853 words)
	Public Health and Epidemiology (ADDI)	Source for train: 147,289 words
		Source for benchmarks: 2,000 segments (66,678 words)

CONCLUSIONS

- Comprehensive infrastructure for Basque language technologies
- End-to-end pipeline: from data collection to AI models
- Aligned with European frameworks (LDS, FAIR)
- Supports digital sovereignty, inclusion, and innovation
- Scalable and transferable to low-resource languages

FUTURE WORK

- Expand both speech & text corpora
- Address new domains according to user needs
- Develop LLMs and LT services for pilots
- Automate data processing and annotation
- Integrate European infrastructures (LDS)
- Deploy real-world AI applications
- Focus on scalability, usability, and impact

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.

euskorpora