



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

**HiTZ**

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

technology

HiTZ

language

# Frontier research on Generative Artificial Intelligence

Annual Report  
2024

# ORGANIZATION

HiTZ is a leading multidisciplinary research center for Language Technology and Generative Artificial Intelligence, bringing together experts from seven departments at the University of the Basque Country. The center's core mission is to investigate language and speech technologies, with a strong emphasis on transferring knowledge and technology to the business sector. HiTZ comprises two research groups, IXA and Aholab, which boast extensive experience dating back to 1993. Both IXA and Aholab were recognized as high-performance research groups by the Basque Government's science agency in its latest research evaluation. Throughout their history, the groups have participated in over 200 research projects, ranging from regional to European initiatives, and have carried out more than 100 contracts with diverse companies. HiTZ is a member of BDVA and through this association, the center participates in the European Partnership on Artificial Intelligence, Data and Robotics.



**Eneko  
Agirre**

Director



**German  
Rigau**

Deputy Director



**Suna Şeyma  
Uçar**

Board Member



**Aritz  
Farwell**

Board Member



**Maite  
Oronoz**

Board Member



**Inma  
Hernaez**

Board Member



**Esther  
Miranda**

Board Member



**Aitor  
Soroa**

Board Member

HiTZ consists of more than 100 members, including computer scientists, linguists, and five research technicians. They are internationally recognized in their scientific areas. In the past five years, researchers currently at the center have published more than 200 scientific publications. The group is a leader in applying deep learning and generative AI techniques to text, speech, and visual-language models. Its recent work in this area has been cited more than 4,000 times in the last two years.

HiTZ also coordinates the CLARIAH-ES distributed research infrastructure, the consortium that administers Spain's participation in CLARIN and DARIAH, both part of the European Strategy Forum on Research Infrastructures (ESFRI). Our deputy director, German Rigau, leads the CLARIAH-ES consortium and is Spain's national coordinator for CLARIN and DARIAH.

The center's impact on society is boosted by the Chair of Artificial Intelligence and Language Technology (AI&LT) as well as two masters coordinated by the center: the Erasmus Mundus+ European Master Program in Language and Communication Technologies (LCT) and the Master's program Language Analysis and Processing. The Master's programs, together with a bachelor degree in AI and four continuing education courses for professionals, are designed to meet the demands of industry and research in this rapidly growing field. In addition, HiTZ also offers a Doctoral Program in Language Analysis and Processing.



# WORDS FROM THE DIRECTOR

Generative Artificial Intelligence (GenAI) has advanced dramatically in recent years. Its rapid rise is exemplified by Large Language Models (LLMs) such as GPT. These models, trained on vast datasets of text and code, exhibit a remarkable ability to understand, generate, and manipulate human language, topping the charts in machine translation and speech processing. As such, AI tools are profoundly impacting various aspects of technology and society. This transformative era offers new opportunities for those who are able to harness cutting-edge AI technology. At the same time, however, its sudden emergence is prompting crucial discussions about its responsible development and deployment, including risks such as disinformation, harmful biases, high energy needs, and the widening digital divide between high- and low-resource languages like Basque.

In order to face this new era and exploit its opportunities, it is essential to foster open and public research in Language Technology. Without homegrown research, our country might become a mere consumer of technology that is developed elsewhere. This is true not only of Language Technology and other AI core technologies such as vision and robotics. GenAI is also revolutionizing other sciences and disciplines, as attested to by the 2024 Chemistry Nobel prize awarded to the AlphaFold creators, whose work is based on the same underlying technology.

At HiTZ we take this vision to heart. The pioneers that founded HiTZ are the same who, through research on symbolic AI, built Xuxen, the first large-scale morphological analyser for Basque, which in 1994 became the basis for the widely used commercial Basque spellchecker. Thirty years later, HiTZ is home to many of Spain's leading experts in GenAI and Language Technology and continues to remain at the forefront of research in this area.

In 2024, we released Latxa, the first and best open 70B LLM specialized in Basque. This accomplishment was recognized with the best paper award at ACL, the field's most prestigious conference. Later in the year, we improved Latxa instruction-tuning and made it amenable to chat with the general public, matching GPT in public evaluation exercises. At the same time, our multidisciplinary center continued to grow. We published a significant number of top-tier conference and journal papers and have more students pursuing doctorates than ever before.

The HiTZ Chair of Artificial Intelligence and Language Technology reinforced our efforts to train professionals and transfer knowledge and technology to society. The CLARIAH-ES research infrastructure, led by HiTZ, further consolidated our national and international collaborative networks by providing essential digital tools and resources for the social sciences and humanities.

Finally, in actively extending the benefits of GenAI into new domains, we produced ground-breaking visual-language models and pursued an ambitious project with neuroscientists, exploring the intersection of speech and brain data. Building on these successes, HiTZ looks forward to future challenges and to solidifying our country's position as a leading research hub for Language Technology and Generative AI.

**Eneko Agirre**  
Director of HiTZ



# CLARIAH

HiTZ coordinates CLARIAH-ES, the distributed digital research infrastructure that steers Spain's participation in the European Research Infrastructure Consortia CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities). CLARIAH-ES strives to enhance digitally-enabled research and provide researchers with access to state-of-the-art digital resources, tools, and services. Its multi- and interdisciplinary approach is designed to advance Spanish research in the social sciences and humanities, foster Spain's digital humanities community, and help strategically position Spain's researchers in international projects and programs, particularly within the European Research Area. German Rigau, Deputy Director of HiTZ, is Spain's National Coordinator for CLARIAH-ES, CLARIN, and DARIAH.

In 2024, CLARIAH-ES welcomed two new members: Dialnet, a major bibliographic database for Spanish academic publications, and SCAYLE, the Supercomputing Center of Castile and León. With these new additions, CLARIAH-ES currently includes partners from twelve leading institutions and research groups across Spain that are dedicated to advancing digital capabilities within the social sciences and humanities.

In this framework, HiTZ also launched and coordinates CLARIAH-EUS node, which encompasses several institutions and groups dedicated to research in and for the Basque language in the social sciences and humanities. CLARIAH-EUS receives funding and support from the Basque Government, the Provincial Council of Gipuzkoa, the University of the Basque Country (EHU), and HiTZ.



**German Rigau**

Deputy Director of HiTZ  
Spain's National Coordinator for  
CLARIAH-ES, CLARIN, and DARIAH



**Xabier Arregi**

Coordinator of CLARIAH-EUS

# CLARIAH-ES



CLARIAH-EUS



In 2024 we launched the HiTZ Chair of Artificial Intelligence and Language Technology (AI&LT) as part of the National Artificial Intelligence Strategy (ENIA). The Chair is funded by the Ministry of Digital Transformation under an initiative to support university chairs that are committed to research, dissemination, teaching, and innovation in Artificial Intelligence. It is the only funded chair in the Basque Country and the only chair in Spain dedicated to Language Technology.

The HiTZ Chair's main objective is to enhance the positive impact of AI&LT on society in general and the productive sector in particular, strengthen research activities in AI&LT, and create a solid body of knowledge to support educational programs in various fields that use this technology. To achieve this, it focuses on three lines of work: education, research, and knowledge transfer. These lines are developed in collaboration with the companies that are affiliated with the Chair, which include Avature SL, Elhuyar Fundazioa, Ikerlan SCL, Euskaltel-MasOrange, Multiverse Computing, and Tecnalia Foundation for Research and Innovation. The relevance, number, and excellence of these organizations demonstrate the productive sector's interest in the Chair's potential to positively impact society.



The Chair offers four specialization courses in Generative AI, Deep Learning, and Language Technology, designed to quickly introduce professionals, researchers, and students to this exciting field. These courses are taught by leading experts in the field ([hitz.eus/training](https://hitz.eus/training)).



**Aitor Soroa**

Director of the HiTZ Chair of Artificial Intelligence and Language Technology



# HiTZ IN NUMBERS

## Research & Transfer

34

Research  
projects

5

Knowledge  
transference  
projects

7

Doctoral  
theses  
defended  
(7 International)

18

Journal  
papers  
(11 Q1)

56

Conference  
papers  
(18 A or A+)

3

Book  
chapters

1

Book









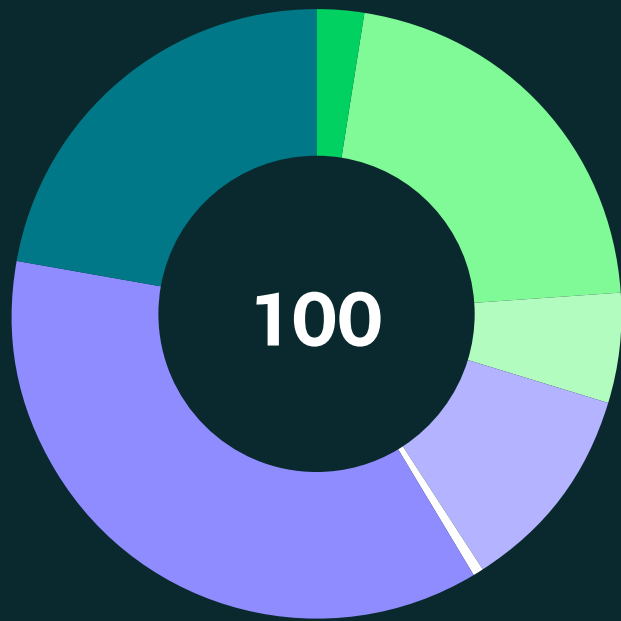
hitzak

language



# People

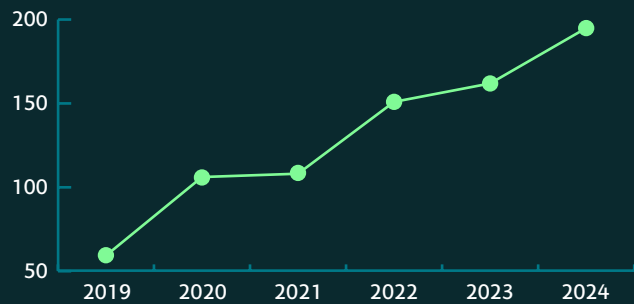
## Members

	Administrative and technical staff	5
	Lecturers	42
	Postdoctoral researchers	11
	Other researchers	22
	Funded predoctoral researchers	19
	Academic collaborator	1

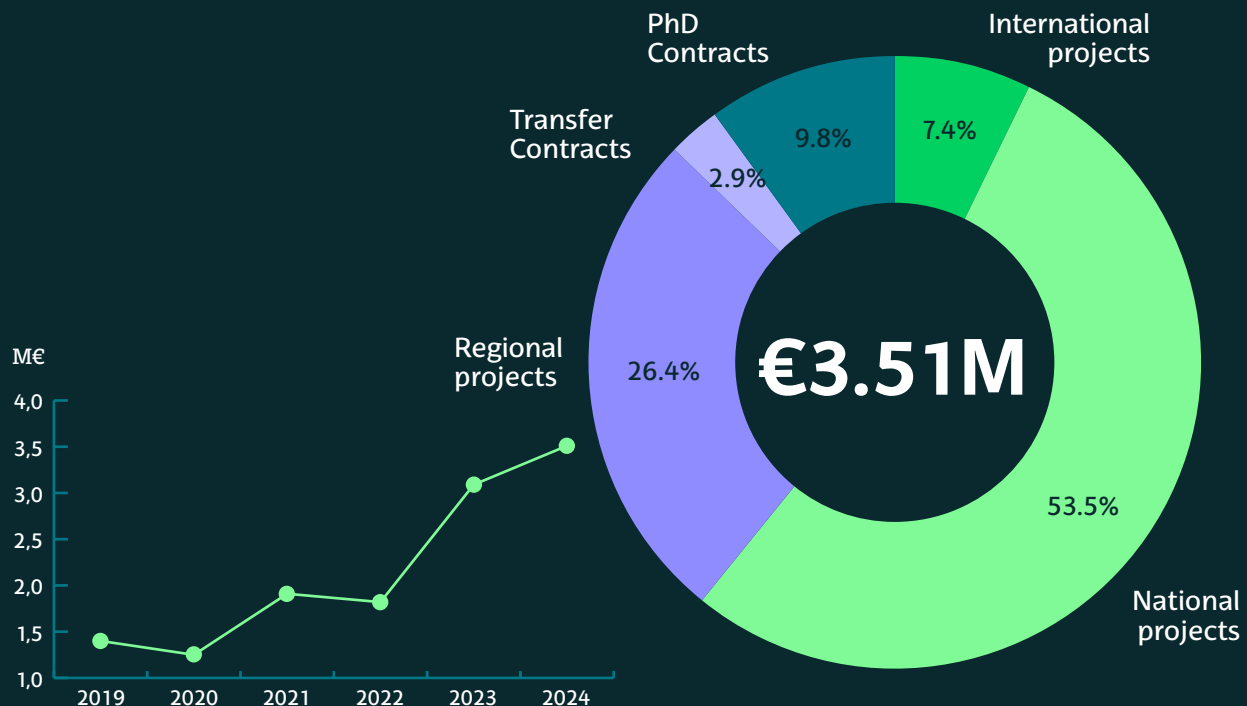


## Students

	Master's students	71
	PhD students	43



# Budget



## RESEARCH AREAS



**Information  
Extraction  
and  
Information  
Retrieval**

Main Researcher:



**Aitor  
Soroa**



**Machine  
Translation**

Main Researcher:



**Gorka  
Labaka**



**Human-  
Computer  
Interaction**

Main Researcher:



**Gorka  
Azkune**



**Speech and  
Language  
Resources**

Main Researcher:



**Ainara  
Estarrona**

## INFRASTRUCTURE

**18**

Multiprocessor  
GNU/Linux

**1**

SPARC  
Solaris  
servers

**1**

HPC Cluster  
with 128  
cores

**14**

servers  
with 71  
GPUs

**36**

GPU cluster  
A100 80GB  
vram

**98**

GPU cluster  
A100 80 GB vram  
(shared with DIPC)

**14**

L40s GPUs  
with 48GB  
vram

**26**

GPU  
various  
models

**1.2** PB

of raw Network storage  
capacity





## Text Analysis

Main Researcher:



Rodrigo Agerri



## Speech Technologies

Main Researcher:



Inma Hernaez



## Medical and Legal domains

Main Researcher:



Arantza Casillas



## Digital humanities and education

Main Researcher:



Mikel Iruskietia

1

acoustically isolated room

1

Interface de audio Midi 4x4 Behringer

1

Quiet PC Sentinel Fanless i10

1

Delsys Trigno™ Wireless EMG System

Access to 1.5 million GPU hours (valued at 4.2 million euros) at the EuroHPC SuperComputer to research large models for European languages with few resources.



## TRAINING

71

Students  
in masters

43

Students  
in doctoral  
program

6

EMLTC  
Master Thesis  
finalized

21

HAP/LAP  
Master Thesis  
Finalized

116

Students in 4  
LLM-related continuing  
education courses

5

Ikasiker

17

Internal and external  
internships



# ACTIVITIES



# 27

Seminars

# 5

Webinars

# 5

Workshops

# 4

Awards

Association for Computational Linguistics Best Resource Paper Award for "Latxa: An Open Language Model and Evaluation Suite for Basque"

Winning project at the SomosNLP #Somos600 hackathon: "Noticia: Resumen de Noticias Clickbait"

Winner of the Albayzin-Iberspeech-RTVE 2024 Speaker Diarization and Identity Assignment Challenge

Best Student Paper at Iberspeech 2024 for "Phone Pair Classification During Speech Production Using MEG Recordings"





# www.hitz.eus

✉ [hitz@ehu.eus](mailto:hitz@ehu.eus)

🦋 [hitz-zentroa.bsky.social](https://hitz-zentroa.bsky.social)

📺 [hitz\\_zentroa](#) | [hitz-zentroa](#)

✂ [@hitz\\_zentroa](#)

in [hitz-zentroa](#)

hitzak  
📧 [@Hitz\\_zentroa](#) | [@hitz.zentroa](#)

technology

language



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

## HiTZ

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology