

# Language technology, the next frontier of Artificial Intelligence

Annual Report  
2023



# INTRODUCTION



Generative Artificial Intelligence (AI) has advanced dramatically in recent years. Its rapid rise, exemplified most acutely by the advent of Large Language Models (LLMs) such as GPT, has raised undue expectations of super-human intelligence and spawned catastrophic predictions about the future that overshadow more subtle and mundane challenges. Fortunately, the initial tempest over the exceptional benefits or dangers of this novel technology has begun to subside in favor of a more down-to-earth assessment: LLMs are far from possessing human-level intelligence and pose no existential threat. They are nonetheless extraordinary tools that are helping many users with broad professional needs perform tasks better and more quickly. As with all disruptive technologies, there are also risks, including disinformation, harmful biases, high energy needs, and a widening of the digital divide between high- and low-resource languages.

In order to face the new challenges and opportunities engendered by artificial intelligence and the digital transformation, it is essential to foster open and public research in language technologies. Moreover, it is equally



language

hitzak

crucial to conduct research domestically, for without homegrown research we become mere consumers of technology created elsewhere. The benefits of such a stance are clear: research and results are made accessible to the local economy and society, while any associated risks may be easily audited.

We take this vision to heart at HiTZ. The year 2023 was a productive one that saw our center grow substantially around four key projects. The Basque Government funded the three-year project **IKER-GAITU** to research ways to close the digital divide for Basque. Through this initiative, we built and released Latxa, the largest and best LLM for Basque created to date. The required compute was delivered by our own servers and those made available by a competitive grant from **EuroHPC**. Additionally, the Spanish Government funded the three-year project **ILENIA** to produce the next generation of open language technologies for Basque, Catalan, and Galician in close collaboration with the main players in Galicia, Catalonia and Valencia. To these endeavors may be added the official launch of the distributed research infrastructure **CLARIAH-ES**. Coordinated by HiTZ,

CLARIAH-ES enables Basque and Spanish researchers and research institutions to participate in CLARIN and DARIAH, Europe's foremost digital infrastructures for research in the humanities and social sciences. Finally, the Spanish Secretariat for AI and the software company Avature co-funded an ambitious four-year **Chair in AI and Language Technology** that will reinforce technological transfer to industry, lifelong education in Language Technology and outreach activities, as well as focus on two key aspects of research with a high impact in society: green LLMs and fair LLMs.

We created HiTZ with a dual purpose: to become an international center of reference for language technology research and the computational processing of Basque. Our most recent achievements not only allow us to continue our cooperation with leading research and technological centers, companies, and institutions, but also to contribute to the country's increasing presence as an international hub for language technologies and artificial intelligence.

Eneko Agirre (Director of HiTZ) and German Rigau (Deputy Director of HiTZ)



# HiTZ IN NUMBERS

technology

## Research & Transfer

39

Projects

2

Transfer Contracts

hitzak







70

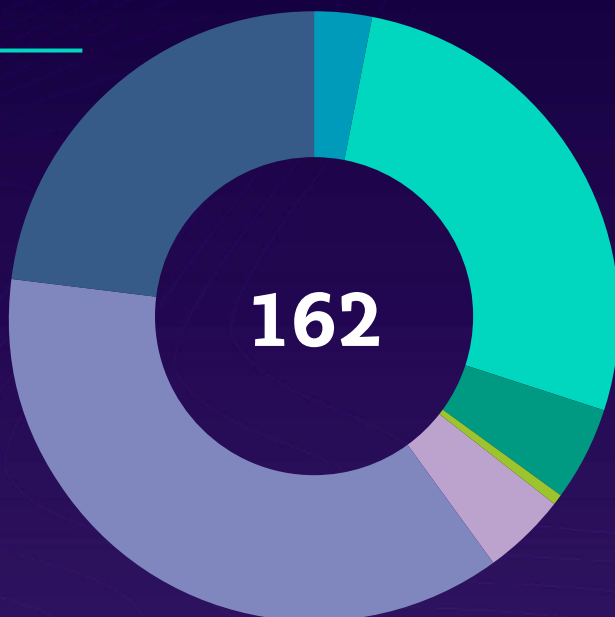
Publications





# People

## Members

	Administrative and technical staff	5
	Lecturers	44
	Postdoctoral researchers	8
	Ramón y Cajal	1
	Other researchers	7
	Funded predoctoral researchers	18

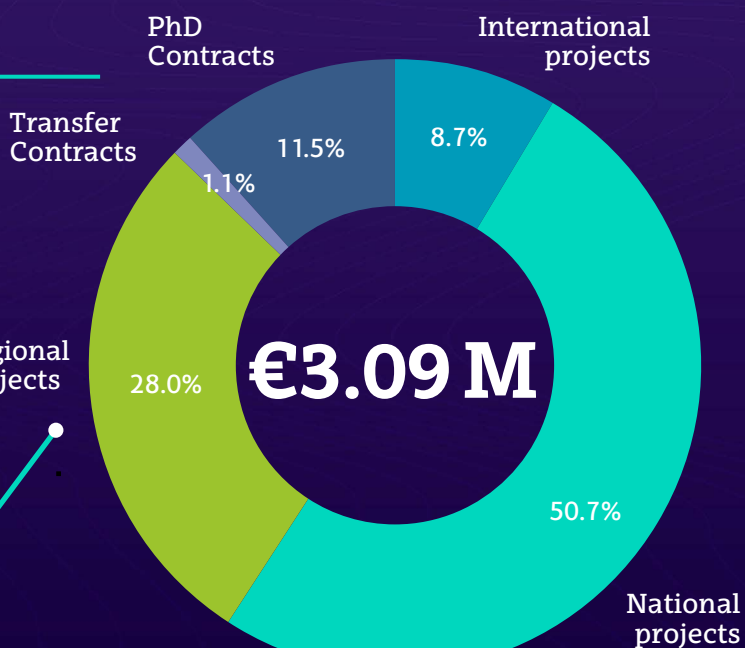


## Students

	Master's students	60
	PhD students	37



# Budget





# ORGANIZATION



HiTZ is a multidisciplinary research center on **Language-centric Artificial Intelligence** with members from seven departments of the University of the Basque Country. The objective of the center is to **investigate** language and speech technologies, with a significant effort towards the **transfer** of knowledge and technology to companies. It comprises two research groups Aholab and Ixa, both with extensive experience since 1993, performing basic research, creating linguistic resources and tools and launching different commercial products on the market. HiTZ is a member of CLAIRE and a full member of BDVA and DAIRO. Through CLAIRE and BDVA, we participate in the European Partnership on Artificial Intelligence, Data and Robotics. We are also a founding member of the Spanish CLARIN K-center.

Since September 2023, HiTZ coordinates the CLARIAH-ES distributed research infrastructure, which integrates two of the largest European scientific infrastructures in humanities and the social sciences into Spain: CLARIN (Common Language Resources and Technology Infrastructure), which offers data, tools, and linguistic technology services to facilitate research in the social sciences and humanities, and DARIAH (Digital Research Infrastructure for the Arts and Humanities), which promotes research and teaching for the arts and humanities based on digital resources. Both infrastructures are part of the European Strategy Forum

CLARIN



DARIAH-EU



CLARIAH-EUS



on Research Infrastructures (ESFRI), established in 2002 by the EU Member States and the European Commission for the purpose of coordinating a common strategy with respect to pan-European scientific facilities and research infrastructures. Our deputy director, German Rigau, is the national coordinator of both infrastructures and leads the CLARIAH-ES consortium, which includes a multidisciplinary group of experts belonging to ten leading research institutions in the areas of social sciences, arts and humanities, library science, linguistics, artificial intelligence, language technologies, computer science, and high-performance computing.



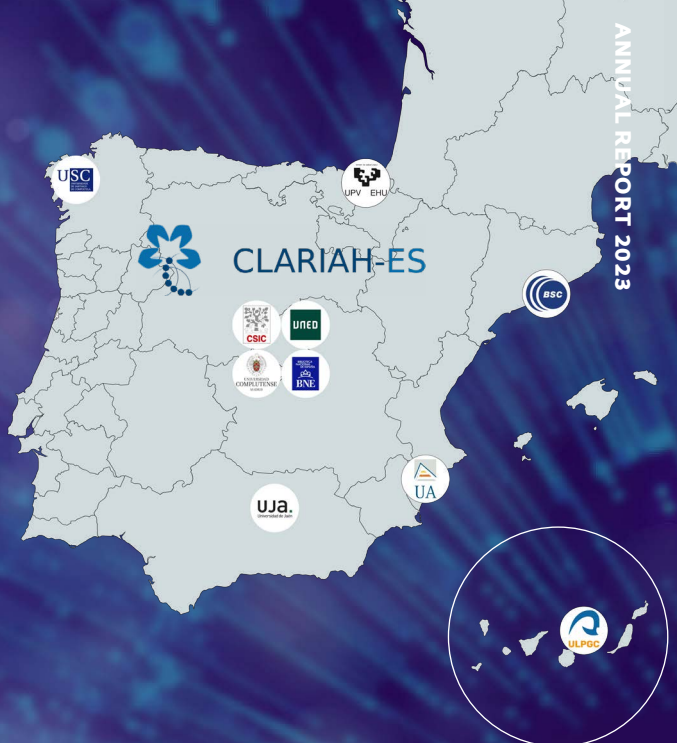
**Eneko  
Agirre**

Director



**German  
Rigau**

Deputy Director



**Suna Seyma  
Uçar**

Vocal



**Maite  
Oronoz**

Vocal



**Itziar  
Aldabe**

Vocal



**Inma  
Hernaez**

Vocal



**Esther  
Miranda**

Vocal



**Aitor  
Soroa**

Vocal

The members of the center are international referents in their scientific areas. At the moment, it is formed by **more than 80 members**, including computer scientists, linguists and 5 research technicians. In the last five years, the researchers now in the center have published more than 200 scientific publications. The group is a leader in applying deep learning techniques to language processing and its recent work in the area has been **cited more than 4,000 times** in the last two years. The members of the center have been **advisors** in the creation of the National Plan for Spanish Language Technologies and are currently advising the Basque Government's equivalent counterpart.

Both IXA and Aholab have been evaluated as high-performance research groups in the last research evaluation exercise by the science agency of the Basque Government. During their history, the groups have participated in more than 200 **research projects** ranging from regional to European projects. It has also participated in more

than 100 **industrial contracts** with the aim of transferring technology into the industry.

HiTZ is also a member of **Erasmus Mundus+ European Masters Program** in Language and Communication Technologies (LCT) **program**. It is designed to meet the demands of industry and research in the rapidly growing field of language technology. HiTZ also offers a **Doctoral Programme** in Language Analysis and Processing.

**The University of the Basque Country (UPV/EHU)** is the leading teaching and research institution in the Basque Country, a prosperous region stretching along the Atlantic coast of northern Spain. The UPV/EHU is among the best 400 universities in the world according to the Shanghai ranking, and has been recognized as an International Excellence Campus by the Spanish Government. The University of the Basque Country, a vibrant 30-year-old institution with 45,000 students, 5,000 world-class academic staff and state-of-the-art facilities distributed throughout 20 centers in its three campuses.



# 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

36

GPU cluster  
A100 80GB vram

48

GPU cluster  
A100 80 GB vram (shared @ DIPC)

26

GPU various models

1

HPC Cluster with 128 cores

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







## Text Analysis

Main Researcher:



Rodrigo  
Agerri

words



## Speech Technologies

Main Researcher:



Inma  
Hernaez



## Medical and Legal domains

Main Researcher:



Arantza  
Casillas



## Digital humanities and education

Main Researcher:



Mikel  
Iruskietea

techno

Over

450 TB

of raw Network storage capacity

1

Behringer  
4x4 audio/  
MIDI  
interface

1

Quiet PC  
Sentinel  
Fanless  
i10

1

acoustically isolated  
room with audio  
equipment for  
professional recordings

hitzak

# RESEARCH & TRANSFER

39

Research  
projects

2

Knowledge  
transference  
projects

4

Doctoral  
theses  
defended  
(2 International)

23

Journal  
papers  
(14 Q1)

36

Conference  
papers  
(9 A or A+)

11

Book  
chapters







## TRAINING

60

Students  
in masters

37

Students  
in doctoral  
program

4

EMLTC  
Master Thesis  
finalized

21

HAP/LAP  
Master Thesis  
Finalized

46

Students in 2 Deep  
Learning complementary  
courses

6

Ikasiker

17

Internal and external  
internships

# ACTIVITIES

# 24

Seminars



# 5

Webinars

# 2

Workshops

# 2

Awards

Order Carlos J. Finlay  
(2022)

Radio Bilbao Award  
for Excellence in Basque

