



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

HiTZ

Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

technology

HiTZ

language

Investigación Avanzada en Inteligencia Artificial Generativa

Informe anual
2024

ORGANIZACIÓN

HiTZ es un centro de investigación multidisciplinar líder en Tecnología del Lenguaje e Inteligencia Artificial Generativa, que reúne a expertos de siete departamentos de la Universidad del País Vasco. Su misión principal es investigar en tecnologías del lenguaje y del habla, con un fuerte enfoque hacia la transferencia de conocimiento y tecnología al sector empresarial. HiTZ está compuesto por dos grupos de investigación, IXA y Aholab, los cuales cuentan con una amplia experiencia que se remonta a 1993. Ambos han sido reconocidos como grupos de investigación de alto rendimiento por la agencia científica del Gobierno Vasco en su última evaluación. A lo largo de su trayectoria, los dos grupos han participado en más de 200 proyectos de investigación, desde iniciativas regionales hasta europeas, y han realizado más de 100 contratos con empresas de diversos sectores. HiTZ es miembro de BDVA y, a través de esa asociación, participa en la Alianza Europea de Inteligencia Artificial, Datos y Robótica.



**Eneko
Agirre**

Director



**German
Rigau**

Subdirector



**Suna Şeyma
Uçar**

Miembro de
la Junta



**Aritz
Farwell**

Miembro de
la Junta



**Maite
Oronoz**

Miembro de
la Junta



**Inma
Hernaez**

Miembro de
la Junta



**Esther
Miranda**

Miembro de
la Junta



**Aitor
Soroa**

Miembro de
la Junta

HiTZ cuenta con más de 100 miembros, incluyendo informáticos, lingüistas y cinco técnicos de investigación, reconocidos internacionalmente en sus respectivas áreas científicas. En los últimos cinco años, los investigadores que actualmente forman parte del centro han publicado más de 200 artículos científicos. El grupo es líder en la aplicación de técnicas de aprendizaje profundo e inteligencia artificial generativa a texto, habla y modelos de lenguaje-visual. Su trabajo reciente en esta área ha sido citado más de 4.000 veces en los últimos dos años.

HiTZ también coordina la infraestructura de investigación distribuida CLARIAH-ES, el consorcio que administra la participación de España en CLARIN y DARIAH, ambas parte del Foro Estratégico Europeo sobre Infraestructuras de Investigación (ESFRI). Nuestro subdirector, German Rigau, lidera el consorcio CLARIAH-ES y es el coordinador nacional de España para CLARIN y DARIAH.

El impacto del centro en la sociedad se refuerza a través de la Cátedra de Inteligencia Artificial y Tecnología del Lenguaje (IA&TL), así como de dos másteres coordinados por el centro: el Máster Europeo Erasmus Mundus+ en Tecnologías del Lenguaje y la Comunicación (LCT) y el Máster en Análisis y Procesamiento del Lenguaje. Estos programas de máster, junto con un grado en IA y cuatro cursos de formación continua para profesionales, están diseñados para responder a las demandas de la industria y la investigación en este campo en rápida expansión. Además, HiTZ ofrece un Programa de Doctorado en Análisis y Procesamiento del Lenguaje.

MENSAJE DEL DIRECTOR

La Inteligencia Artificial Generativa (GenAI) ha avanzado de forma espectacular en los últimos años. Los Grandes Modelos de Lenguaje (LLMs, por sus siglas en inglés) como GPT son un ejemplo de ese rápido auge. Esos modelos, entrenados con enormes conjuntos de datos de texto y código, muestran una notable capacidad para comprender, generar y manipular el lenguaje humano, liderando los rankings de la traducción automática y el procesamiento del habla. Así, las herramientas de IA están teniendo un impacto profundo en diversos aspectos de la tecnología y la sociedad. Esta era transformadora ofrece nuevas oportunidades para aquellos que son capaces de aprovechar la tecnología de IA más avanzada. Sin embargo, al mismo tiempo, su aparición repentina también está generando debates cruciales sobre su desarrollo y despliegue responsables, entre los que se incluyen los relativos a riesgos como la desinformación, los sesgos dañinos, el alto consumo energético y la creciente brecha digital entre lenguas con muchos y pocos recursos, como el euskera.

Para afrontar esta nueva era y aprovechar las oportunidades que ofrece, es esencial fomentar la investigación abierta y pública en Tecnología del Lenguaje. Sin investigación propia, nuestro país corre el riesgo de convertirse en un mero consumidor de tecnología desarrollada en otros lugares. Esto es aplicable no solo a la Tecnología del Lenguaje, sino también a otras tecnologías clave de la IA, como la visión artificial y la robótica. A su vez, la GenAI está revolucionando otras ciencias y disciplinas, como lo demuestra la concesión del Premio Nobel de Química 2024 a los creadores de AlphaFold, cuyo trabajo se basa precisamente en dicha tecnología subyacente.

En HiTZ adoptamos plenamente esta visión. Los pioneros que fundaron HiTZ son los mismos que, mediante la investigación en IA simbólica, desarrollaron Xuxen, el primer analizador morfológico a gran escala para el euskera, que en 1994 se convirtió en la base del ampliamente

utilizado corrector ortográfico comercial en euskera. Treinta años después, HiTZ alberga a muchos de los principales expertos de España en GenAI y Tecnología del Lenguaje y continúa a la vanguardia de la investigación en este ámbito.

En 2024 lanzamos Latxa, el primer y mejor LLM abierto de 70B especializado en euskera. Este logro fue reconocido con el premio al mejor artículo en ACL, la conferencia más prestigiosa del área. Más adelante, mejoramos la adaptación de Latxa a instrucciones, haciéndolo accesible para el público general, alcanzando resultados comparables a los de GPT en ejercicios públicos de evaluación. Al mismo tiempo, nuestro centro multidisciplinar ha seguido creciendo. Hemos publicado un número significativo de artículos en congresos y revistas de primer nivel, y hemos alcanzado el mayor número de estudiantes de doctorado hasta la fecha.

La Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje ha afianzado nuestros esfuerzos para formar profesionales y transferir conocimiento y tecnología a la sociedad. La infraestructura de investigación CLARIAH-ES, liderada por HiTZ, ha consolidado aún más nuestras redes de colaboración nacionales e internacionales al proporcionar herramientas y recursos digitales esenciales para las ciencias sociales y las humanidades. Finalmente, en nuestra labor por extender los beneficios de la GenAI a nuevos ámbitos, hemos desarrollado modelos innovadores de lenguaje-visual y hemos llevado a cabo un ambicioso proyecto con neurocientíficos orientado a la exploración de la intersección entre el habla y los datos cerebrales. Sobre la base de estos éxitos, en HiTZ afrontamos con entusiasmo los retos futuros y la consolidación de la posición de nuestro país como referente de la investigación en Tecnología del Lenguaje e Inteligencia Artificial Generativa.



Eneko Agirre
Director de HiTZ

CLARIAH

HiTZ coordina CLARIAH-ES, la infraestructura digital distribuida que lidera la participación de España en los Consorcios Europeos de Infraestructuras de Investigación CLARIN (Common Language Resources and Technology Infrastructure) y DARIAH (Digital Research Infrastructure for the Arts and Humanities). CLARIAH-ES busca fomentar la investigación digitalmente habilitada y ofrecer a los investigadores acceso a recursos, herramientas y servicios digitales de vanguardia. Su enfoque multidisciplinar e interdisciplinar está diseñado para avanzar en la investigación española en ciencias sociales y humanidades, fomentar la comunidad de humanidades digitales en España y ayudar a posicionar estratégicamente a los investigadores españoles en proyectos y programas internacionales, especialmente dentro del Espacio Europeo de Investigación. German Rigau, subdirector de HiTZ, es el Coordinador Nacional de España para CLARIAH-ES, CLARIN y DARIAH.

En 2024 CLARIAH-ES dio la bienvenida a dos nuevos miembros: Dialnet, una importante base de datos bibliográfica de publicaciones académicas en español, y SCAVLE, el Centro de Supercomputación de Castilla y León. Con estas nuevas incorporaciones, CLARIAH-ES cuenta actualmente con socios de doce instituciones y grupos de investigación líderes en toda España, dedicados al desarrollo de capacidades digitales en las ciencias sociales y humanidades.

En este contexto, HiTZ también lanzó y coordina el nodo CLARIAH-EUS, que reúne a varias instituciones y grupos dedicados a la investigación en y para la lengua vasca en las ciencias sociales y humanidades. CLARIAH-EUS recibe financiación y apoyo del Gobierno Vasco, la Diputación Foral de Gipuzkoa, la Universidad del País Vasco (EHU) e HiTZ.



German Rigau

Subdirector de HiTZ
Coordinador Nacional en España
para CLARIAH-ES, CLARIN y DARIAH



Xabier Arregi

Coordinador de CLARIAH-EUS

CLARIAH-ES



CLARIAH-EUS



En 2024 lanzamos la Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje (IA&TL) como parte de la Estrategia Nacional de Inteligencia Artificial (ENIA). La Cátedra está financiada por el Ministerio de Transformación Digital dentro de una iniciativa para apoyar cátedras universitarias comprometidas con la investigación, la divulgación, la docencia y la innovación en Inteligencia Artificial. Es la única cátedra financiada en el País Vasco y la única en España dedicada a la Tecnología del Lenguaje.

Los principales objetivos de la Cátedra son potenciar el impacto positivo de la IA&TL en la sociedad en general y en el sector productivo en particular, fortalecer las actividades de investigación en IA&TL y crear un cuerpo sólido de conocimiento que sirva de apoyo para programas educativos de diversos ámbitos que utilizan esta tecnología. Para ello, cuenta con tres líneas de trabajo: formación, investigación y transferencia de conocimiento. Estas líneas se desarrollan en colaboración con las empresas afiliadas a la Cátedra, entre las que se incluyen Avature SL, Elhuyar Fundazioa, Ikerlan SCL, Euskaltel-MásOrange, Multiverse Computing y Fundación Tecnalía Research & Innovation. El número de organizaciones colaboradoras, así como su relevancia y excelencia, demuestran el interés del sector productivo en el potencial de la Cátedra para impactar positivamente en la sociedad.



Aitor Soroa

Director de la Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje



Financiado por la Unión Europea
NextGenerationEU



HiTZ EN NÚMEROS

Investigación y transferencia

34

proyectos de
investigación

5

proyectos de
transferencia de
conocimiento

7

tesis doctorales
defendidas
(7 internacionales)

18

artículos de
revistas
(11 Q1)

56

artículos en
congresos
(18 A o A+)

3

capítulos
de libro

1

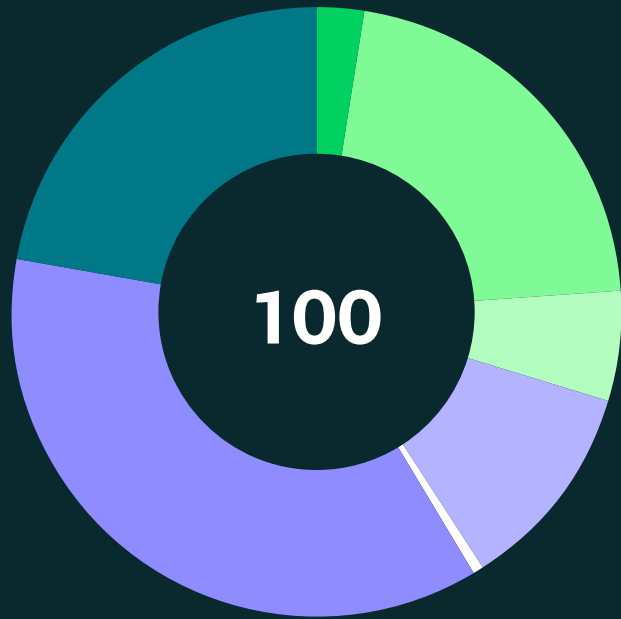
libro



Personas

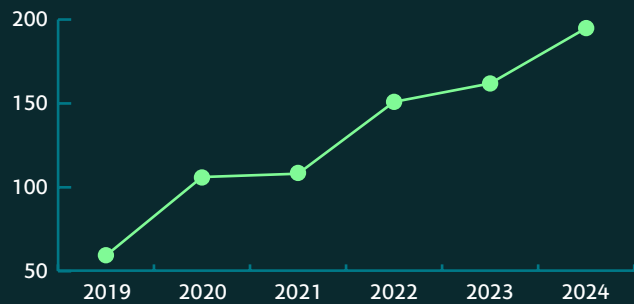
Miembros

	Gestión y administración	5
	Profesores	42
	Investigadores postdoctorales	11
	Otras investigadoras	22
	Investigadores con becas predoctorales	19
	Colaborador académico	1

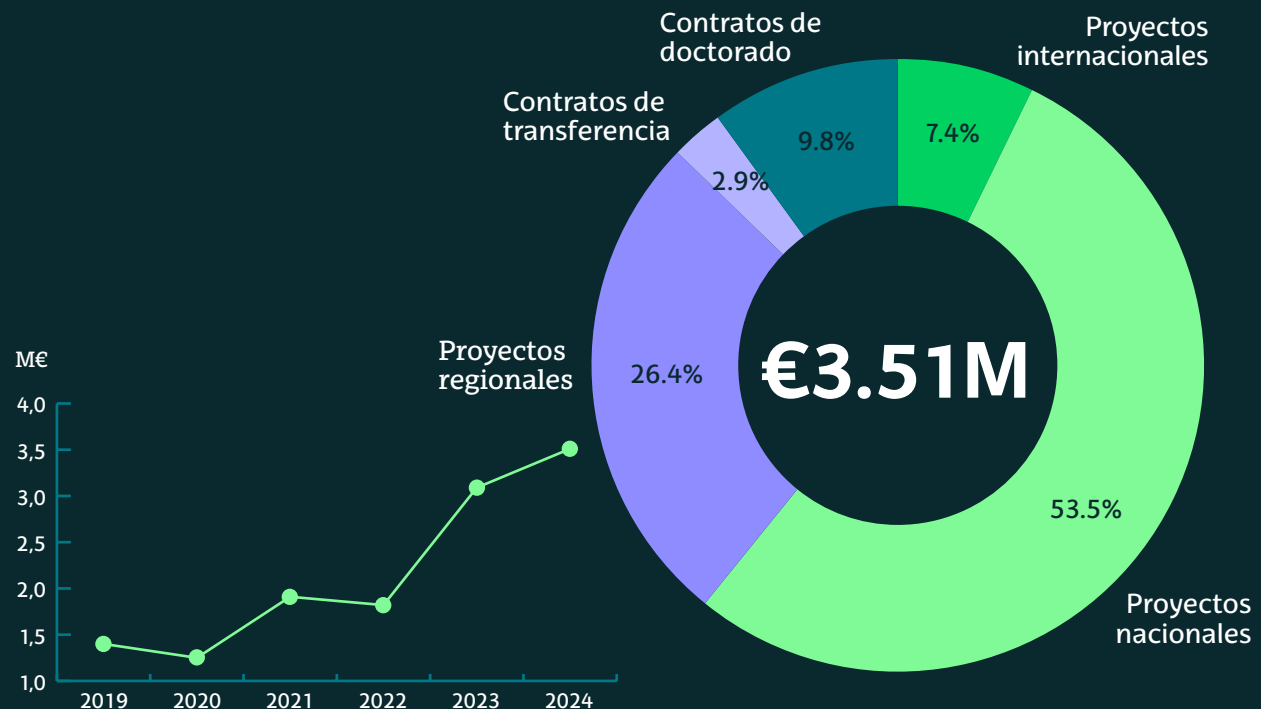


Alumnado

	Estudiantes de master	71
	Estudiantes de doctorado	43



Presupuesto



ÁREAS DE INVESTIGACIÓN



Extracción y recuperación de información

Investigador principal:



Aitor Soroa



Traducción automática

Investigador principal:



Gorka Labaka



Interacción persona-computadora

Investigador principal:



Gorka Azkune



Recursos del habla y el lenguaje

Investigadora principal:



Ainara Estarrona

INFRAESTRUCTURA

18

multiprocesador
GNU/Linux

1

servidor
SPARC
Solaris

1

HPC Cluster
con 128
cores

14

servidores
con 71
GPUs

36

GPU cluster
A100 80GB
vram

98

GPU cluster
A100 80 GB vram
(compartido /
DIPC)

14

L40 GPUs con
48GB vram

26

GPU
varios
modelos

1.2 PB

de capacidad de
almacenamiento bruto
en red



Análisis de texto

Investigador principal:



Rodrigo Agerri



Tecnologías de Voz

Investigadora principal:



Inma Hernaez



Dominios médico y legal

Investigadora principal:



Arantza Casillas



Humanidades digitales y educación

Investigador principal:



Mikel Iruskietia

1

sala aislada acústicamente

1

Interfaz de audio Midi 4x4 Behringer

1

Quiet PC Sentinel Fanless i10

1

Delsys Trigno™ Wireless EMG System

Acceso a 1,5 millones de horas de GPU (valorado en 4,2 millones de euros) en el Superordenador EuroHPC para la investigación de grandes modelos para lenguas europeas con pocos recursos.



PRÁCTICAS

71

estudiantes
en másters

43

estudiantes en
el programa de
doctorado

6

TFMs de EMLCT
finalizadas

21

TFMs de
HAP/LAP
finalizadas

116

estudiantes en cursos
de formación continua
sobre LLM

5

Ikasiker

17

prácticas internas
y externas





27

seminarios

center

5

webinarios

hitzak

5

talleres

4

premios

words

hizkuntza

Premio al Mejor Artículo de Recursos de la Association for Computational Linguistics para "Latxa: An Open Language Model and Evaluation Suite for Basque"

Proyecto ganador en el hackathon SomosNLP #Somos600: "Noticia: Resumen de Noticias Clickbait"

Ganador del Albayzín-Iberspeech-RTVE 2024 Speaker Diarization and Identity Assignment Challenge

Mejor Artículo de Estudiante en Iberspeech 2024 por "Phone Pair Classification During Speech Production Using MEG Recordings"



www.hitz.eus

✉ hitz@ehu.eus

🦋 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