Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

www.hitz.eus
@hitz_zentroa

La tecnología del lenguaje, la próxima frontera de la Inteligencia Artificial

Informe anual
2023

INTRODUCCIÓN

La Inteligencia Artificial Generativa (IA) ha experimentado un avance dramático en los últimos años. Su rápido ascenso, del que han sido máximo exponente los Modelos de Lenguaje de Gran Tamaño (LLMs) como GPT, ha generado unas indebidas expectativas de inteligencia sobrehumana y ha hecho que surjan predicciones catastrofistas acerca del futuro que han eclipsado otros retos más sutiles y mundanos. Afortunadamente, la tormenta inicial en torno a los excepcionales beneficios o peligros inherentes a esta nueva tecnología ha comenzado a amainar en favor de una evaluación más realista: los LLMs están muy lejos de poseer inteligencia al nivel de los humanos y no suponen ninguna amenaza existencial. No obstante, son herramientas extraordinarias que están sirviendo de ayuda a un gran número de usuarios con necesidades profesionales de muy diverso tipo a la hora de acometer su labor mejor y más ágilmente. Como sucede con todas las tecnologías disruptivas, ésta también conlleva sus riesgos, entre los que se encuentran la desinformación, los sesgos dañinos, el elevado consumo energético y el crecimiento de la brecha digital entre los idiomas de bajos recursos y los de recursos elevados.

Con el fin de afrontar los retos y las oportunidades generados por la inteligencia artificial y la transformación digital, es esencial promover la investigación abierta y pública en el ámbito de las tecnologías del lenguaje. Más aún, es igualmente crucial desarrollar

investigación a nivel local, ya que sin esa investigación propia nos convertimos en meros consumidores de la tecnología creada por otros desde otros lugares. Los beneficios inherentes a este planteamiento son claros: permite un fácil acceso de la economía local y de la sociedad local a la investigación, así como a sus resultados, y facilita, a su vez, la intervención frente a los riesgos asociados a estos.

En HiTZ nos tomamos en serio este enfoque. 2023 ha sido un año muy productivo en el que hemos visto crecer sustancialmente nuestro centro en torno a cuatro proyectos clave. Así, el Gobierno Vasco ha decidido financiar el proyecto **IKER-GAITU** que durante tres años investigará el modo de cerrar la brecha digital del euskera. En el marco de dicha iniciativa, construimos y lanzamos Latxa, el mayor y mejor de los LLM para el euskera creados hasta la fecha. El cálculo necesario para ello fue proporcionado por nuestros propios servidores y por aquellos puestos a nuestra disposición mediante una subvención en régimen de concurrencia competitiva otorgada por **EuroHPC**. A su vez, el Gobierno de España ha acordado financiar el proyecto **ILENIA** que a lo largo de tres años producirá la próxima generación de tecnologías abiertas del lenguaje para el euskera, el catalán y el gallego en estrecha colaboración con los principales agentes de la investigación en dicho sector en Galicia, Cataluña y Valencia. A estas iniciativas debe añadirse el lanzamiento oficial de la infraestructura de investigación distribuida **CLARIAH-ES**. Bajo la coordinación de HiTZ, CLARIAH-ES permitirá que investigadores e instituciones de investigación vascos y españoles participen en CLARIN y DARIAH, principales infraestructuras digitales para la investigación en el área de las humanidades y la ciencias sociales. Finalmente, la Secretaría de Estado de Digitalización e Inteligencia Artificial del Gobierno de España y la compañía de software Avature han decidido cofinanciar durante cuatro años la ambiciosa **Cátedra en Inteligencia Artificial y Tecnología del Lenguaje** que reforzará la transferencia tecnológica a la industria, la educación permanente y las actividades divulgativas sobre Tecnología del Lenguaje, centrándose, a su vez, en dos aspectos clave de la investigación con un gran impacto social: los LLM verdes y los LLM justos.

Cuando creamos HiTZ lo hicimos con un doble propósito: convertirnos en un centro de referencia a nivel internacional para la investigación en tecnología del lenguaje y en el procesamiento computacional del euskera. Nuestros logros más recientes no sólo van a hacer posible que continuemos nuestra cooperación con centros, compañías e instituciones líderes en los campos de la tecnología y la investigación, sino que también nos permitirán contribuir a consolidar la creciente presencia de este país en la medida en que somos un punto de enlace internacional para las tecnologías del lenguaje y la inteligencia artificial.

Eneko Agirre (Director del HiTZ) y German Rigau (Subdirector del HiTZ)

HITZ EN CIFRAS

technology

Investigación y transferencia

39
proyectos

2
contratos de transferencia

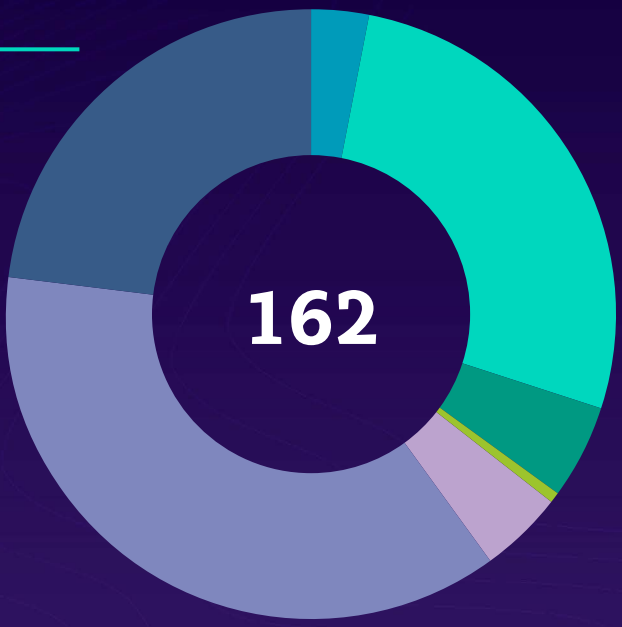
hitzak

70
publicaciones

Personas

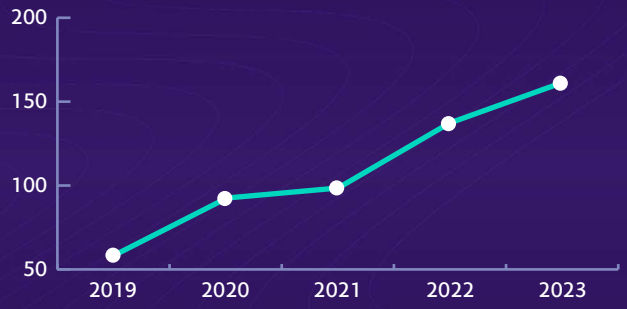
Miembros

Gestión y administración	5
Profesores	44
Investigadores postdoctorales	8
Ramón y Cajal	1
Otras investigadoras	7
Becas de doctorado	18

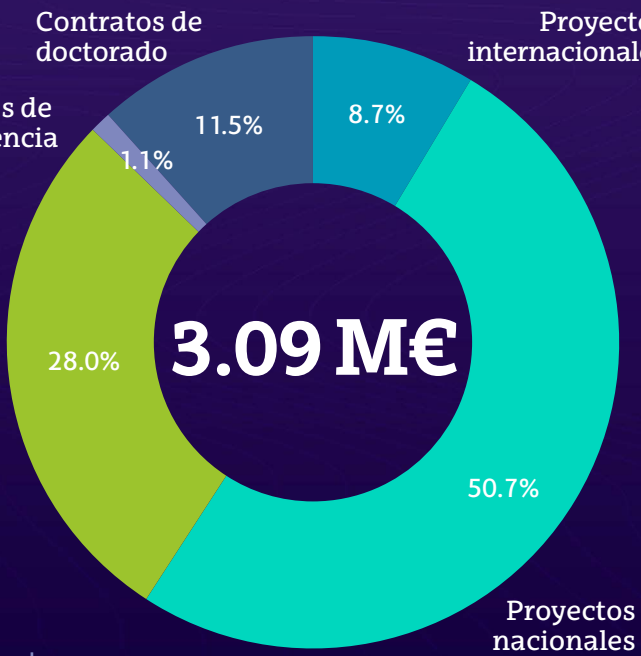


Alumnado

Estudiantes de master	60
Estudiantes de doctorado	37



Presupuesto



ORGANIZACIÓN

HiTZ es un centro de investigación multidisciplinar especializado en la **Inteligencia Artificial centrada en el lenguaje**, formado por miembros de 7 departamentos de la Universidad del País Vasco/Euskal Herriko Unibertsitatea. El objetivo del centro es **investigar** en el lenguage y las tecnologías del habla, con un significativo esfuerzo en la **transferencia** de conocimiento y tecnología a las empresas. HiTZ está compuesto por dos grupos de investigación, Aholab e Ixa. Ambos poseen una amplia experiencia en investigación de base, creando recursos y herramientas lingüísticas y habiendo lanzado varios productos al mercado. HiTZ es miembro de CLAIRE y miembro de pleno derecho de BDVA y de DAIRO. A través de CLAIRE y BDVA, participamos en el European Partnership on Artificial Intelligence, Data and Robotics. Somos a su vez miembros fundadores del Spanish CLARIN K-centre.

Desde septiembre de 2023 HiTZ coordina la infraestructura de investigación distribuida CLARIAH-ES que integra en España dos de las mayores infraestructuras científicas europeas en Humanidades y Ciencias Sociales: CLARIN (Common Language Resources and Technology Infrastructure) que ofrece datos, herramientas y servicios de tecnología lingüística para facilitar la investigación en Ciencias Sociales y Humanidades, y DARIAH (Digital Research Infrastructure for the Arts and Humanities) que impulsa la investigación y la enseñanza dirigida a las Artes y las Humanidades basada en recursos digitales. Ambas infraestructuras forman parte del Foro Estratégico para las Infraestructuras de Investigación (ESFRI por sus siglas en inglés), formado por los Estados Miembro de la UE y la Comisión Europea, que se constituyó en 2002 con el objetivo de coordinar una estrategia común en materia de instalaciones científicas e infraestructuras de investigación de carácter paneuropeo. Nuestro subdirector, German Rigau, es el coordinador nacional de ambas infraestructuras y lidera el consorcio CLARIAH-ES que integra a un grupo multidisciplinar de expertos pertenecientes a diez instituciones líderes de la investigación en las áreas de las Ciencias Sociales, Artes y Humanidades, biblioteconomía, lingüística, inteligencia artificial, tecnologías del lenguaje, informática y computación de alto rendimiento.

Eneko Agirre

Director

German Rigau

Subdirector

Suna Seyma Uçar

Miembro de la Junta

Maite Oronoz

Miembro de la Junta

Itziar Aldabe

Miembro de la Junta

Inma Hernaez

Miembro de la Junta

Esther Miranda

Miembro de la Junta

Aitor Soroa

Miembro de la Junta

Los miembros del centro son referentes internacionales en sus áreas científicas. En este momento HiTZ está formado por más de 80 miembros que incluyen informáticos, lingüistas y 5 técnicos de investigación. En el último lustro, los investigadores que forman el centro han publicado más de 200 publicaciones. El grupo es líder aplicando técnicas de aprendizaje profundo y, en los dos últimos años los trabajos realizados en éste área han sido citados en más de 4000 ocasiones. Los miembros del centro han realizado labores de asesoramiento en Plan Nacional de Tecnologías del Lenguaje y actualmente están asesorando al Gobierno Vasco en labores similares.

Tanto Ixa como Aholab han sido denominados grupos de alto rendimiento en la última evaluación realizada por el Consejo Científico del Gobierno Vasco. A lo largo de su historia, los grupos han participado en más de 200 **proyectos** de investigación, desde regionales hasta europeos. Además de participar en más 100 **contratos** industriales con el objetivo de transferir tecnología a la industria.

HiTZ es miembro de **Erasmus Mundus+ European Masters Program in Language and Communication Technologies (LCT)**, un programa diseñado para aunar las demandas tanto de la industria como de la investigación en un campo de rápido crecimiento como es el de la tecnología del lenguaje. HiTZ ofrece además un programa Doctoral en Procesamiento del Lenguaje Natural, y es miembro de pleno derecho de la Academia Internacional de Doctorado en Inteligencia Artificial (AIDA).

La Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) es la principal entidad para la enseñanza e investigación en el País Vasco. La UPV/EHU está entre las 400 mejores universidades del mundo según el ranking Shangai, y el Gobierno de España la ha reconocido como Campus de Excelencia. La Universidad del País Vasco es una institución de más de 30 años de antigüedad, con 45.000 alumnos, 5.000 académicos de nivel mundial e instalaciones de vanguardia, distribuidas en 20 centros en sus 3 campus.

ÁREAS DE INVESTIGACIÓN

Extracción y recuperación de información

Investigador/a principal:

Aitor Soroa

Traducción automática

Investigador/a principal:

Gorka Labaka

Interacción persona-computadora

Investigador/a principal:

Gorka Azkune

Recursos del habla y el lenguaje

Investigador/a principal:

Ainara Estarrona

INFRAESTRUCTURA

36 GPU cluster A100 80GB vram

48 GPU cluster A100 80 GB vram (compartido / DIPC)

26 GPU varios modelos

1 HPC Cluster con 128 núcleos

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

EuroHPC Joint Undertaking

Análisis de texto

Investigador/a principal:

Rodrigo Agerri

Tecnologías de Voz

Investigador/a principal:

Inma Hernaez

Dominios médico y legal

Investigador/a principal:

Arantza Casillas

Humanidades digitales y educación

Investigador/a principal:

Mikel Iruskieta

Más de

450 TB

de capacidad de almacenamiento bruto en red

1

Interfaz de audio/ MIDI 4x4 Behringer

1

Quiet PC Sentinel Fanless i10

1

sala aislada acústicamente con equipo de audio para grabaciones profesionales

INVESTIGACIÓN Y TRANSFERENCIA

39 **proyectos de investigación**

2 **proyectos de transferencia de conocimiento**

4 **tesis doctorales defendidas** (2 internacionales)

23 **artículos de revistas** (14 Q1)

36 **artículos en congresos** (9 A o A+)

11 **capítulos de libro**

PRÁCTICAS

60
estudiantes en el Máster

37
estudiantes en el programa de doctorado

4
TFMs de EMLCT finalizadas

21
TFMs de HAP/LAP finalizadas

46
estudiantes en 2 cursos complementarios de Aprendizaje Profundo

6
Ikasiker

17
prácticas internas y externas

ACTIVIDADES

24
seminarios

5
webinarios

2
talleres

2
premios

Orden Carlos J. Finlay (2022)

Premio Radio Bilbao a la Excelencia en la categoría de Euskera