Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Recherche de Pointe sur l'Intelligence Artificielle Générative

Rapport annuel 2024

ORGANISATION

HiTZ est un centre de recherche multidisciplinaire de premier plan en technologies du langage et en intelligence artificielle générative (GenAI), rassemblant des experts de sept départements de l'Université du Pays Basque. La mission principale du centre est d'étudier les technologies du langage et de la parole, avec une forte orientation vers le transfert de savoir et de technologie vers le secteur industriel. HiTZ regroupe deux groupes de recherche, IXA et Aholab, qui possèdent une vaste expérience depuis 1993. Ces deux groupes ont été reconnus comme groupes de recherche de haut niveau par l'agence scientifique du Gouvernement Basque lors de sa dernière évaluation. Au fil des ans, ils ont participé à plus de 200 projets de recherche à l'échelle régionale, nationale et européenne, et réalisé plus de 100 contrats avec diverses entreprises. HiTZ est membre de la BDVA, et via cette association, il participe au Partenariat européen pour l'intelligence artificielle, les données et la robotique.

HiTZ rassemble plus de 100 membres, dont des informaticiens, des linguistes et cinq techniciens de recherche. Ils sont internationalement reconnus dans leurs domaines scientifiques. Au cours des cinq dernières années, les chercheurs du centre ont publié plus de 200 articles scientifiques. Le groupe est à la pointe de l'application des techniques d'apprentissage profond et de GenAI au texte, à la parole et aux modèles de langage visuel. Leurs travaux récents dans ce domaine ont été cités plus de 4.000 fois au cours des deux dernières années.

HiTZ coordonne également l'infrastructure de recherche distribuée CLARIAH-ES, le consortium qui supervise la participation de l'Espagne aux infrastructures de recherche européennes CLARIN et DARIAH, toutes les deux inscrites sur la feuille de route de l'ESFRI *(European Strategy Forum on Research Infrastructures)*. Notre directeur adjoint, German Rigau, dirige le consortium CLARIAH-ES et est le coordinateur national de l'Espagne pour CLARIN et DARIAH.

L'impact du centre sur la société est renforcé par la Chaire en Intelligence Artificielle et Technologies du Langage (IA&TL), ainsi que par deux masters qu'il coordonne : le Master Erasmus Mundus+ européen en Technologies du Langage et de la Communication (LCT), et le Master en Analyse et Traitement du Langage. Ces programmes, accompagnés d'une licence en IA et de quatre formations continues à destination des professionnels, visent à répondre aux besoins de l'industrie et de la recherche dans ce domaine en pleine croissance. HiTZ propose également un programme doctoral en Analyse et Traitement du Langage.

Eneko Agirre
Directeur

German Rigau
Sous-directeur

Suna Şeyma Uçar
Membre du conseil

Aritz Farwell
Membre du conseil

Maite Oronoz
Membre du conseil

Inma Hernaez
Membre du conseil

Esther Miranda
Membre du conseil

Aitor Soroa
Membre du conseil

MESSAGE DU DIRECTEUR

L'intelligence artificielle générative (GenAI) a connu des avancées spectaculaires ces dernières années. Son essor rapide est illustré par les grands modèles de langage (LLM) tels que GPT. Entraînés sur d'immenses ensembles de données textuelles et de code, ces modèles démontrent une remarquable capacité à comprendre, générer et manipuler le langage humain, atteignant des performances de pointe en traduction automatique et en traitement de la parole. Ces outils d'IA ont ainsi un impact profond sur divers aspects de la technologie et de la société. Cette ère de transformation offre de nouvelles opportunités à ceux qui sont capables de tirer parti des dernières technologies de l'IA. Dans le même temps, son émergence soudaine suscite des débats cruciaux sur son développement et son déploiement responsables, notamment sur les risques de désinformation, les préjugés nuisibles, la forte consommation d'énergie et la fracture numérique croissante entre les langues riches en ressources et celles qui en sont dépourvues, comme le basque.

Pour affronter cette nouvelle ère et exploiter ses opportunités, il est essentiel de promouvoir une recherche ouverte et publique dans le domaine des technologies du langage. Sans une recherche locale, notre pays risque de devenir un simple consommateur de technologies développées ailleurs. Cela est vrai non seulement pour les technologies du langage, mais aussi pour d'autres technologies clés de l'IA telles que la vision par ordinateur et la robotique. La GenAI révolutionne également d'autres sciences et disciplines, comme en témoigne le prix Nobel de chimie 2024 décerné aux créateurs d'AlphaFold, dont les travaux reposent sur la même technologie sous-jacente.

Chez HiTZ, nous partageons pleinement cette vision. Les pionniers qui ont fondé HiTZ sont les mêmes qui, par le biais de la recherche en IA symbolique, ont développé Xuxen, le premier analyseur morphologique à grande échelle pour le basque, devenu en 1994 la base du correcteur orthographique commercial le plus utilisé en langue basque. Trente ans plus tard, HiTZ regroupe de nombreux experts de premier plan en GenAI et en technologies du langage, et reste à la pointe de la recherche dans ce domaine.

En 2024, nous avons lancé Latxa, le premier et meilleur LLM ouvert de 70 milliards de paramètres spécialisé dans la langue basque. Cet accomplissement a été récompensé par le prix du meilleur article lors de la conférence ACL, la plus prestigieuse du domaine. Par la suite, nous avons amélioré l'adaptation de Latxa aux instructions, le rendant accessible au grand public, obtenant des résultats comparables à ceux de GPT dans les exercices d'évaluation publique. En parallèle, notre centre pluridisciplinaire a continué à croître. Nous avons publié un nombre important d'articles dans des conférences et revues de premier plan et nous accueillons plus d'étudiants en doctorat que jamais.

La Chaire HiTZ en Intelligence Artificielle et Technologies du Langage a renforcé nos efforts de formation et de transfert de savoir-faire vers la société. L'infrastructure de recherche CLARIAH-ES, coordonée par HiTZ, a consolidé nos réseaux de collaboration nationaux et internationaux en fournissant des outils et ressources numériques essentiels pour les sciences humaines et sociales. Enfin, en étendant les bénéfices de la GenAI à de nouveaux domaines, nous avons produit des modèles innovants mêlant langage et image et lancé un projet ambitieux avec des neuroscientifiques, explorant l'intersection entre la parole et les données cérébrales. Forts de ces succès, HiTZ aborde avec confiance les défis futurs pour consolider la position de notre pays en tant que pôle de recherche majeur dans les domaines des technologies du langage et de l'intelligence artificielle générative.

Eneko Agirre
Directeur de HiTZ

CLARIAH

HiTZ coordonne CLARIAH-ES, l'infrastructure numérique distribuée qui pilote la participation de l'Espagne aux consortiums européens de recherche CLARIN (Common Language Resources and Technology Infrastructure) et DARIAH (Digital Research Infrastructure for the Arts and Humanities). CLARIAH-ES vise à promouvoir la recherche numérique et à fournir aux chercheurs des outils, services et ressources numériques de pointe. Son approche pluri- et interdisciplinaire vise à renforcer la recherche espagnole en sciences humaines et sociales, à soutenir la communauté espagnole des humanités numériques, et à positionner stratégiquement les chercheurs espagnols dans les projets et programmes internationaux, notamment dans l'Espace européen de la recherche. German Rigau, directeur adjoint de HiTZ, est le coordinateur national pour CLARIAH-ES, CLARIN et DARIAH.

En 2024, CLARIAH-ES a accueilli deux nouveaux membres : Dialnet, une importante base de données bibliographiques pour les publications universitaires en espagnol, et SCAYLE, le centre de supercalcul de Castille-et-León. Avec ces nouvelles adhésions, CLARIAH-ES regroupe aujourd'hui des partenaires issus de douze institutions et groupes de recherche espagnols de premier plan, engagés dans le développement des capacités numériques dans les sciences humaines et sociales.

Dans ce cadre, HiTZ a également lancé et coordonne le nœud CLARIAH-EUS, qui regroupe plusieurs institutions et groupes de recherche dédiés à la langue basque dans le champ des sciences humaines et sociales. CLARIAH-EUS bénéficie du soutien financier du Gouvernement Basque, du Conseil Provincial de Gipuzkoa, de l'Université du Pays Basque (EHU) et de HiTZ.

German Rigau
Directeur adjoint de HiTZ
Coordinateur national a l'Espagne de CLARIAH-ES, CLARIN et DARIAH

Xabier Arregi
Coordinateur de CLARIAH-EUS

CLARIAH-ES

CLARIAH-EUS

CHAIRE

Universidad del País Vasco / Euskal Herriko Unibertsitatea

Adimen Artifizial eta Hizkuntza Teknologiako HiTZ Katedra
Cátedra HiTZ de Inteligencia Artificial y Tecnología de la Lengua

Avature

En 2024, nous avons lancé la Chaire HiTZ en Intelligence Artificielle et Technologies du Langage (IA&TL) dans le cadre de la Stratégie nationale pour l'intelligence artificielle (ENIA). La Chaire est financée par le Ministère de la Transformation Numérique dans le cadre d'une initiative visant à soutenir les chaires universitaires dédiées à la recherche, à la diffusion, à l'enseignement et à l'innovation en IA. Elle est la seule chaire financée au Pays Basque et la seule en Espagne spécifiquement consacrée aux technologies du langage.

Les objectifs principaux de la Chaire sont d'accroître l'impact positif de l'IA&TL sur la société en général et sur le secteur productif en particulier, de renforcer la recherche dans ce domaine et de créer une base solide de connaissances pour soutenir les programmes éducatifs dans les disciplines utilisant cette technologie. Pour ce faire, elle se concentre sur trois axes : la formation, la recherche et le transfert de connaissances. Ces axes sont développés en collaboration avec les entreprises partenaires de la Chaire, parmi lesquelles Avature SL, Elhuyar Fundazioa, Ikerlan SCL, Euskaltel-MasOrange, Multiverse Computing et la Fondation Tecnalia pour la recherche et l'innovation. Le nombre d'entreprises collaboratrices, ainsi que leur pertinence et leur excellence, démontrent l'intérêt du secteur productif pour le potentiel de la Chaire à avoir un impact positif sur la société.

La Chaire propose quatre cours de spécialisation sur l'IA générative, l'apprentissage profond et les technologies du langage, conçus pour initier rapidement professionnels, chercheurs et étudiants à ce domaine passionnant. Ces formations sont dispensées par des experts de haut niveau (hitz.eus/training).

Aitor Soroa
Directeur de la Chaire HiTZ d'Intelligence Artificielle et de Technologie Linguistique

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Vicepresidencia Primera del Gobierno · Ministerio de Asuntos Económicos y Transformación Digital · Secretaría de Estado de Digitalización e Inteligencia Artificial
Plan de Recuperación, Transformación y Resiliencia
ENIA Inteligencia Artificial
España | digital 2026
orai NLP Teknologiak
euskaltel
ikerlan Member of Basque Research & Technology Alliance
Multiverse Computing
tecnalia Member of Basque Research & Technology Alliance

HiTZ EN CHIFFRES

Recherche et transfert

34
Projets de recherche

5
Contrats de transfert de connaissances

7
Thèses de doctorat soutenues
(7 internationales)

18
Articles de revue
(11 Q1)

56
Communications de congrés
(18 A ou A+)

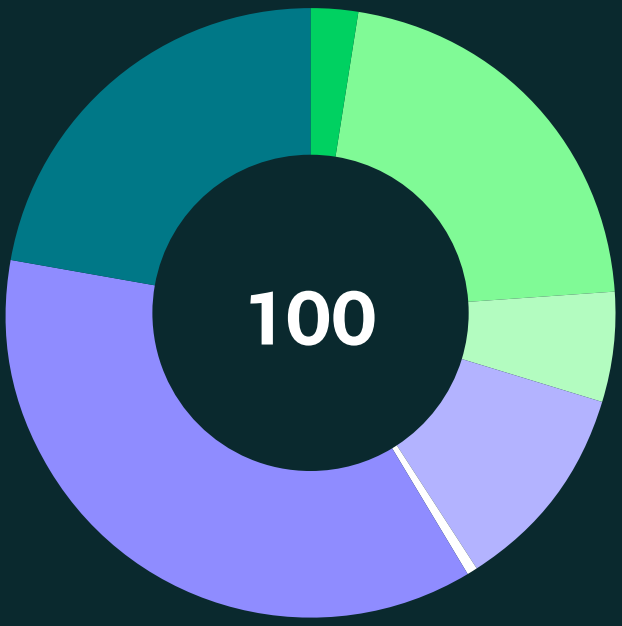
3
Chapitres de livre

1
Livre

Personnes

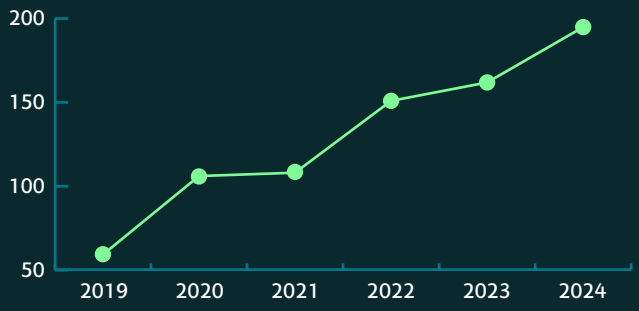
Membres

Gestion et administration	5
Enseignants	42
Chercheurs postdoctoraux	11
Autres chercheurs	22
Chercheurs titulaires d'une bourse prédoctorale	19
Collaborateur académique	1

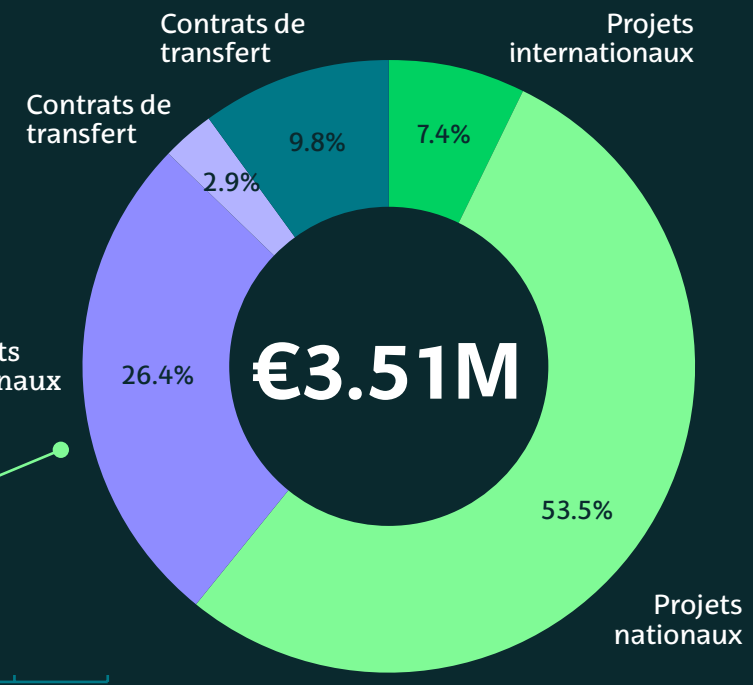


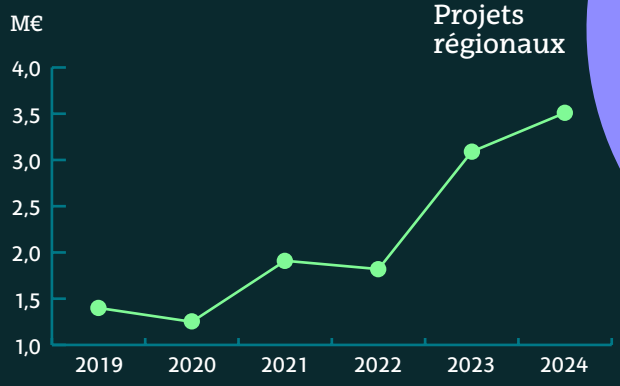
Étudiants

Étudiants en master	71
Étudiants en doctorat	43



Budget





DOMAINES DE RECHERCHE

Extraction et recherche d'information

Chercheur principal :

Aitor Soroa

Traduction automatique

Chercheur principal :

Gorka Labaka

Interaction homme-machine

Chercheur principal :

Gorka Azkune

Ressources vocales et langagières

Chercheur principal :

Ainara Estarrona

INFRASTRUCTURE

18 multiprocesseur GNU/Linux

1 serveur SPARC Solaris

1 HPC Cluster avec 128 cores

14 serveurs avec 71 GPUs

36 GPU cluster A100 80GB vram

98 GPU cluster A100 80 GB vram (partagé / DIPC)

14 L40 GPUs avec 48GB vram

26 GPU divers modèles

1.2 PB de capacité de stockage brute en réseau

Analyse de textes

Chercheur principal :

Rodrigo Agerri

Technologies de la parole

Chercheur principal :

Inma Hernaez

Domaines médical et juridique

Chercheur principal :

Arantza Casillas

Sciences humaines numériques et éducation

Chercheur principal :

Mikel Iruskieta

1 salle acoustiquement isolée

1 Behringer 4x4 audio/MIDI interface

1 Quiet PC Sentinel Fanless i10

1 Delsys Trigno™ Wireless EMG System

Accès à 1,5 million d'heures de GPU (d'une valeur de 4,2 millions d'euros) au Superordinateur EuroHPC pour la recherche de grands modèles pour les langues européennes avec peu de ressources

EuroHPC Joint Undertaking

FORMATION

71
Étudiants en masters

43
Étudiants en doctorat

6
Thèses du master EMLTC soutenues

21
Thèses du master HAP/LAP soutenues

116
Étudiants de 4 cours de formation continue liés au LLM

5
Ikasiker

17
Stages internes et externes

ACTIVITÉS

27
Séminaires

5
Webinaires

5
Workshops

4
Prix

Prix du Meilleur Article de Ressources de l'Association for Computational linguistics pour "Latxa: An Open Language Model and Evaluation Suite for Basque"

Projet gagnant au hackathon SomosNLP #Somos600: "NoticIA: Resumen de Noticias Clickbait"

Vainqueur du Albayzín-Iberspeech-RTVE 2024 Speaker Diarization and Identity Assignment Challenge

Meilleur Article Étudiant à Iberspeech 2024 pour "Phone Pair Classification During Speech Production Using MEG Recordings"

www.hitz.eus

hitz@ehu.eus

@hitz_zentroa

hitz-zentroa.bsky.social

hitz-zentroa

hitz_zentroa | hitz-zentroa

@Hitz_zentroa | @hitz.zentroa

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology