eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

www.hitz.eus
@hitz_zentroa

Les technologies du langage, la prochaine frontière de l'intelligence artificielle

Rapport annuel 2023

INTRODUCTION

L'intelligence artificielle générative (IA) a connu des progrès spectaculaires ces dernières années. Son essor rapide, dont les grands modèles de langage (LLM) tels que GPT sont un excellent exemple, a suscité des attentes excessives en matière d'intelligence surhumaine et conduit à des prédictions apocalyptiques sur l'avenir qui ont éclipsé d'autres défis plus subtils et banals. Heureusement, la tempête initiale autour des avantages ou des dangers exceptionnels inhérents à cette nouvelle technologie a commencé à s'apaiser en faveur d'une évaluation plus réaliste: les LLM sont loin de posséder une intelligence de niveau humain et ne constituent pas une menace existentielle. Néanmoins, il s'agit d'outils extraordinaires qui aident un grand nombre d'utilisateurs ayant des besoins professionnels très variés à améliorer leur travail et à le rendre plus agile. Comme toutes les technologies de rupture, elles comportent également des risques, notamment la désinformation, des préjugés nuisibles, une forte consommation d'énergie et une fracture numérique croissante entre les langues à faibles ressources et celles à ressources élevées.

Afin de relever les défis et de saisir les opportunités générées par l'intelligence artificielle et la transformation numérique, il est essentiel de promouvoir une recherche ouverte et publique dans le domaine des technologies de la langue. En outre, il est tout aussi crucial de développer la recherche au niveau local, car sans

cette recherche locale, nous devenons de simples consommateurs de technologies créées par d'autres ailleurs. Les avantages inhérents à cette approche sont évidents: elle permet à l'économie et à la société locales d'accéder facilement à la recherche et à ses résultats, et facilite l'intervention contre les risques qui y sont associés.

Au centre HiTZ, nous prenons cette approche au sérieux. 2023 a été une année très productive au cours de laquelle nous avons vu notre centre se développer considérablement autour de quatre projets clés. Tout d'abord, le Gouvernement Basque a décidé de financer pour trois ans le projet IKER-GAITU qui étudiera comment réduire la fracture numérique en langue basque. Dans le cadre de cette initiative, nous avons construit et lancé Latxa, le plus grand et le meilleur LLM pour la langue basque à ce jour. Les ressources informatiques nécessaires ont été fournies par nos propres serveurs et par ceux qui ont été mis à notre disposition grâce à une subvention concurrentielle accordée par EuroHPC. De son côté, le Gouvernement Espagnol a accepté de financer le projet ILENIA qui, pendant trois ans, produira la prochaine génération de technologies linguistiques ouvertes pour le basque, le catalan et le galicien, en étroite collaboration avec les principaux acteurs de la recherche dans ce secteur en Galice, en Catalogne et à Valence. A ces initiatives s'ajoute le lancement officiel de l'infrastructure de recherche distribuée CLARIAH-ES. Sous la coordination de HiTZ, CLARIAH-ES permettra aux chercheurs et aux institutions de recherche basques et espagnoles de participer à CLARIN et DARIAH, les principales infrastructures numériques pour la recherche en sciences humaines et sociales. Enfin, le secrétaire d'État espagnol à la numérisation et à l'intelligence artificielle et la compagnie de logiciels Avature ont décidé de cofinancer pendant quatre ans l'ambitieuse chaire d'intelligence artificielle et de technologies du langage, qui renforcera le transfert de technologies vers l'industrie, la formation continue et les activités de diffusion en matière de technologies du langage, en se concentrant sur deux aspects clés de la recherche à fort impact social : les LLM verts et les LLM équitables.

Lorsque nous avons créé HiTZ, nous l'avons fait avec un double objectif: devenir un centre international de référence pour la recherche en technologie linguistique et le traitement automatique du basque. Nos dernières réalisations nous permettront non seulement de poursuivre notre coopération avec des centres, des entreprises et des institutions de premier plan dans les domaines de la technologie et de la recherche, mais aussi de contribuer à consolider la présence croissante de ce pays dans la mesure où nous sommes un point focal international pour les technologies du langage et l'intelligence artificielle.

Eneko Agirre (Directeur de HiTZ) et German Rigau (Directeur adjoint de HiTZ)

HITZ EN CHIFFRES

technology

Recherche et transfert

39
Projets

2
Contrats de transfert

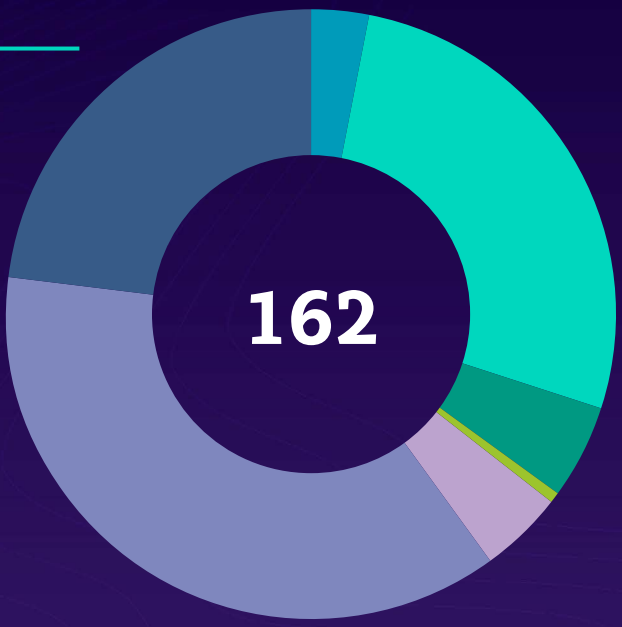
hitzak

39
Publications

Personnes

Membres

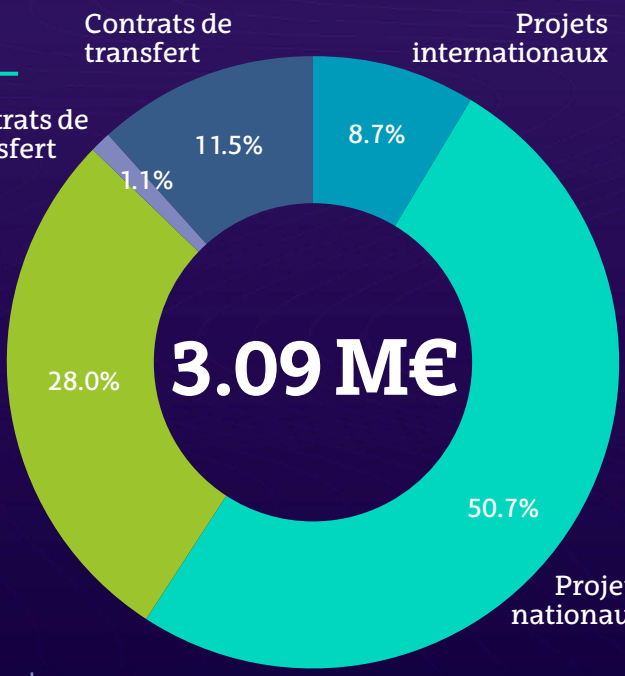
Gestion et administration	5
Enseignants	44
Chercheurs postdoctoraux	8
Ramón y Cajal	1
Autres chercheurs	7
Bourses de doctorat	18



Étudiants

Étudiants en master	60
Étudiants en doctorat	37

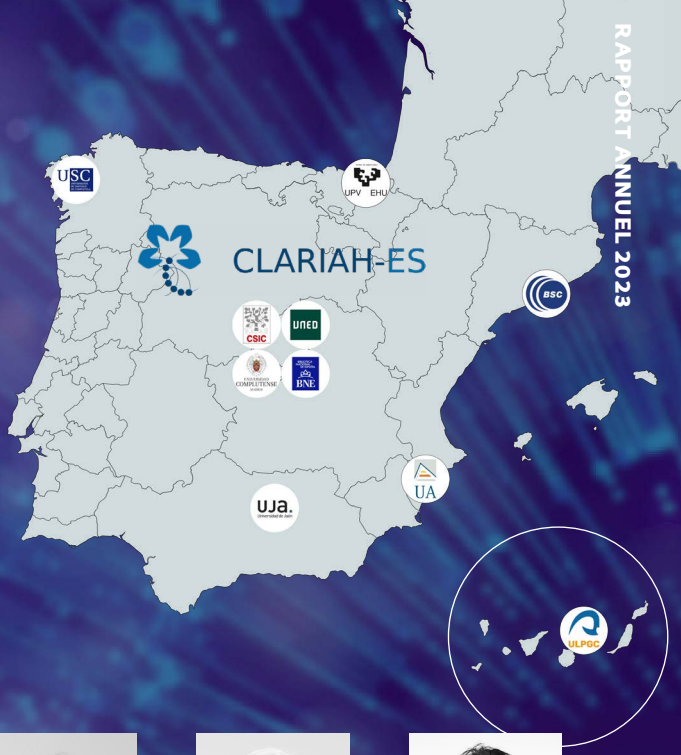
Budget



ORGANISATION

HiTZ est un centre de recherche multidisciplinaire sur **l'intelligence artificielle centrée sur la langue**, dont les membres proviennent de sept départements de l'Université du Pays Basque. L'objectif du centre est **d'étudier** les technologies du langage et de la parole, avec un effort important pour le **transfert** de connaissances et de technologies aux entreprises. Il comprend deux groupes de recherche, Aholab et Ixa, tous les deux dotés d'une grande expérience depuis 1993, qui effectuent de la recherche fondamentale, créent des ressources et des outils langagiers et lancent différents produits commerciaux sur le marché. HiTZ est membre de CLAIRE et aussi membre à part entière de BDVA et de DAIRO. A travers CLAIRE et BDVA, nous participons au Partenariat Européen sur l'Intelligence Artificielle, les Données et la Robotique. Nous sommes également un membre fondateur du centre Spanish CLARIN K

Depuis septembre 2023, HiTZ coordonne l'infrastructure de recherche distribuée CLARIAH-ES qui intègre en Espagne deux des plus grandes infrastructures scientifiques européennes en sciences humaines et sociales: CLARIN (Common Language Resources and Technology Infrastructure) qui offre des données, des outils et des services de technologie linguistique pour faciliter la recherche en sciences humaines et sociales, et DARIAH (Digital Research Infrastructure for the Arts and Humanities) qui promeut la recherche et l'enseignement dans les arts et les sciences humaines en s'appuyant sur des ressources numériques. Ces deux infrastructures font partie du Forum Stratégique Européen pour les Infrastructures de Recherche (ESFRI les sigles en anglais), formé par les États membres de l'UE et la Commission Européenne, qui a été créé en 2002 dans le but de coordonner une stratégie commune pour les installations scientifiques et les infrastructures de recherche paneuropéennes. Notre directeur adjoint, German Rigau, est le coordinateur national des deux infrastructures et dirige le consortium CLARIAH-ES qui intègre un groupe multidisciplinaire d'experts appartenant à dix institutions de recherche de premier plan dans les domaines des sciences sociales, des arts et des lettres, de la bibliothéconomie, de la linguistique, de l'intelligence artificielle, des technologies du langage, de l'informatique et du calcul à haute performance.



Eneko Agirre

Directeur

German Rigau

Sous-directeur

Suna Seyma Uçar

Membre du conseil

Maite Oronoz

Membre du conseil

Itziar Aldabe

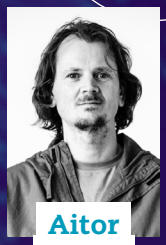
Membre du conseil

Inma Hernaez

Membre du conseil

Esther Miranda

Membre du conseil



Aitor Soroa

Membre du conseil

Les membres du centre sont des référents internationaux dans leurs domaines scientifiques. Il est actuellement composé de plus de 80 membres, dont des informaticiens, des linguistes et 5 techniciens de recherche. Au cours des cinq dernières années, les chercheurs du centre ont publié plus de 200 publications scientifiques. Le groupe est un leader dans l'application des techniques d'apprentissage approfondi au traitement des langues, et ses travaux récents dans ce domaine ont été cités plus de 4 000 fois au cours des deux dernières années. Les membres du centre ont été conseillers dans la création du plan national pour les technologies de la langue espagnole et conseillent actuellement l'homologue du gouvernement basque.

Tant Ixa comme Aholab ont été évalués comme des groupes de recherche performants lors du dernier exercice d'évaluation de la recherche par l'agence scientifique du gouvernement basque. Au cours de leur histoire, les groupes ont participé à plus de 200 **projets** de recherche allant de projets régionaux à des projets européens. Ils ont également participé à plus de 100 **contrats** industriels dans le but de transférer des technologies dans l'industrie.

HiTZ est également membre du programme Erasmus Mundus+ European Masters Program in Language and Communication Technologies (LCT). Il est conçu pour répondre aux demandes de l'industrie et de la recherche dans le domaine en pleine expansion des technologies linguistiques. HiTZ propose également un programme doctoral en analyse et traitement des langues, et est membre à part entière de l'Académie Internationale des Études Doctorales en Intelligence Artificielle (AIDA).

L'Université du Pays basque (UPV/EHU) est le principal établissement d'enseignement et de recherche du Pays Basque, une région prospère qui s'étend le long de la côte atlantique du nord de l'Espagne. L'UPV/EHU fait partie des 400 meilleures universités du monde selon le classement de Shanghai, et a été reconnue comme un campus d'excellence international par le gouvernement espagnol. L'Université du Pays basque, une institution dynamique de 30 ans avec 45 000 étudiants, 5 000 membres du personnel académique de renommée mondiale et des installations de pointe réparties dans 20 centres sur ses trois campus.

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

Chercheuse principale :

Ainara Estarrona

INFRASTRUCTURE

36
GPU cluster A100 80GB vram

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

26
GPU divers modèles

1
HPC Cluster with 128 cores

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

Analyse de textes

Chercheur principal :

Rodrigo Agerri

Technologies de la parole

Chercheuse principale :

Inma Hernaez

Domaines médical et juridique

Chercheuse principale :

Arantza Casillas

Sciences humaines numériques et éducation

Chercheur principal :

Mikel Iruskieta

Plus de

450 TB

de capacité de stockage brute en réseau

1

Behringer 4x4 audio/ MIDI interface

1

Quiet PC Sentinel Fanless i10

1

salle acoustiquement isolée avec équipement audio pour des enregistrements professionnels

RECHERCHE ET TRANSFERT

39
Projets de recherche

2
Contrats de transfert de connaissances

4
Thèses de doctorat soutenues
(2 internationales)

23
Articles de revue
(14 Q1)

36
Communications de congrés
(9 A ou A+)

11
Chapitres de livre

FORMATION

60
Étudiants en master

37
Étudiants en doctorat

4
Thèses du master EMLTC soutenues

21
Thèses du master HAP/LAP soutenues

46
Étudiants dans 2 cours complémentaires d'apprentissage profond

6
Ikasiker

17
Stages internes et externes

ACTIVITÉS

24
Séminaires

5
Webinaires

2
Workshops

2
Prix

Ordre Carlos J. Finlay (2022)

Prix d'excellence Radio Bilbao dans la catégorie Basque