



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

HiTZ

Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

technology

hiTZ

language

Adimen Artifizial Generatiboari buruzko ikerketa aurreratua

Urteko Txostena
2024

ANTOLAKUNTZA

HiTZ diziplina anitzeko ikerketa-zentroa da. Hizkuntza Teknologian eta Adimen Artifizial Sortzailean goi-mailako zentroa da, eta Euskal Herriko Unibertsitateko zazpi sailetako adituak biltzen ditu. Zentroaren egiteko nagusia hizkuntza- eta hizketa-teknologiak ikertzea da, enpresa-sektorearekiko ezagutzaren eta teknologiaren transferentzian enfasi berezia jarritz. HiTZek 1993tik arlo horietan esperientzia zabala duten bi ikerketa-talde biltzen ditu, IXA eta Aholab. Bi taldeak errendimendu handiko ikerketa-talde gisa aitortu zituen Eusko Jaurlaritzaren zientzia-agentziak bere azken ebaluazioan. Haien historian zehar, bi taldeek 200 ikerketa-proiektu baino gehiagotan parte hartu dute, eskualdeko ekimenetatik Europako ekimenetara, eta 100 kontratu baino gehiago egin dituzte enpresa desberdinekin. HiTZ BDVAko kide da, eta elkarre horren bidez, zentroak Adimen Artifizialaren, Datuen eta Robotikaren Europako Partzuergoan parte hartzen du.



Eneko Agirre

Zuzendaria



German Rigau

Zuzendariordea



Suna Şeyma Uçar

Batzordeko kidea



Aritz Farwell

Batzordeko kidea



Maite Oronoz

Batzordeko kidea



Inma Hernaez

Batzordeko kidea



Esther Miranda

Batzordeko kidea



Aitor Soroa

Batzordeko kidea

HiTZek 100 kide baino gehiago ditu, informatikariak, hizkuntzalariak eta bost ikerketa-teknikari. Nazioartean aitoren handia duten ikertzaileak dira, eta horren erakusgarri, azken bost urteetan zentroko ikertzaileek 200 argitalpen zientifiko baino gehiago argitaratu izana. Zentroa liderra da ikasketa sakonean eta adimen artifizial sortzailearen teknikak testuari, hizketari eta ikusmen-hizkuntza ereduari aplikatzen. Arlo horretan egindako lan berriek 4.000 aipu baino gehiago izan dituzte azken bi urteetan.

HiTZek CLARIAH-ES ikerketa-azpiegitura banatuaren bulego tekniko kudeatzen du. Azpiegitura horrek Espainiaren parte-hartzea kudeatzen du ESFRiko (European Strategy Forum on Research Infrastructures) parte diren CLARIN eta DARIAH ikerketa-azpiegiturretan. HiTZeko zuzendariordeak, German Rigauk, koordinatzen du CLARIAH-ES partzuergoa, eta era berean, CLARINeko eta DARIAHko Espainiako koordinatzaile nazionala da.

Zentroaren gizarte-eragina Adimen Artifizialaren eta Hizkuntza Teknologiaren Katedraren (AA&HT) eta zentroak koordinatutako bi masterren bitartez indartzen da: Hizkuntza eta Komunikazio Teknologien Europako Erasmus Mundus+ Master Programa (LCT), eta Hizkuntzaren Analisia eta Prozesamendua Master Programa. Master programa horiek, graduondoko ikastaroa eta profesionalentzako lau etengabeko prestakuntza-ikastaro industriaren eta ikerketaren eskakizunak asetzeko diseinatuta daude. Gainera, HiTZek Hizkuntzaren Analisia eta Prozesamendua Doktorego Programa ere eskaintzen du.

ZUZENDARIAREN MEZUA

Adimen Artifizial Sortzaileak (GenAI) aurrerapen nabarmenak izan ditu azken urteotan. Haren gorakada azkarra Hizkuntza Eredu Handiek (LLM) erakusten dute, hala nola GPTk. Testu eta kode datu-base erraldoiekin entrenatutako eredu horiek gaitasun harrigarria erakusten dute giza hizkuntza ulertzeko, sortzeko eta moldatzeko, eta emaitza oso onak lortzen dituzte itzulpen automatikoan eta hizketaren prozesamenduan. Gauzak horrela, Adimen Artifizialeko (AA) tresnek eragin handia dute gaur egun teknologiaren eta gizartearen alderdi askotan. Aro eraldatzaile horrek aukera berriak eskaintzen dizkie punta-puntako AA teknologia aprobetxatzeko gai izango direnei. Hala ere, AAren bat-bateko agerpenak eztabaida garrantzitsuak eragin ditu garapenaren eta hedapen arduratsuen inguruan, hala nola, desinformazioa, alborapen kaltegarriak, energia-behar handiak eta hizkuntza gutxituen eta hizkuntza handien arteko arrakala digitala zabaltzea.

Aro berri horri aurre egiteko, eta eskaintzen dituen aukerak aprobetxatzeko, funtsezkoa da Hizkuntza Teknologiaren ikerketa irekia eta publikoa sustatzea, bertako ikerketarik gabe, gure herria beste nonbait garatutako teknologiaren kontsumitzaile huts bihurtu baitaiteke. Hori ez da horrela soilik Hizkuntza Teknologiarako, baizik eta beste AAeko teknologia batzuetarako ere, hala nola, ikusmen artifiziala edo robotika. GenAIk beste arlo eta diziplina batzuetara ere ekarri du iraultza, esaterako, 2024ko Kimikako Nobel saria teknologia hori oinarri duen AlphaFold-en sortzaileei eman zitzaizen.

HiTZen erabat geure egiten dugu ikuspegi berri hori. HiTZ sortu zuten aitzindariak ia sinbolikoaren ikerketari esker Xuxen garatu zuten berberak dira. Xuxen izan zen euskararako lehen analizatzaile morfologiko handia, 1994an euskarazko zuzentzaile ortografiko komertzial zabalduenaren oinarri bihurtu zena. Hogeita hamar urte geroago, Espainiako GenAI eta

Hizkuntza Teknologiako aditu nagusienetako asko HiTZ zentroko kideak dira, eta ikerketaren abangoardian jarraitzen dugu arlo horretan.

2024an, Latxa kaleratu genuen, euskaraz espezializatutako lehen eta onena den 70B parametroko LLM irekia. Lorpen horrek ACLn, arloko konferentziarik entzutetsuenean, artikulua onenaren saria lortu zuen. Urtearen amaieran, Latxaren instrukzio-doikuntza hobetu genuen, eta gizakiokin hitz egiteko gai izan zedin egokitu genuen, GPTren mailara iritsiz gizakiok egindako ebaluazio publikoetan. Aldi berean, gure diziplina anitzeko zentroa etengabe hazten ari da. Horren ondorioz, goi-mailako konferentzia eta aldizkarietan artikulua kopuru esanguratsua argitaratu dugu, eta inoiz baino doktorego-ikasle gehiago ditugu.

Azkenik, Adimen Artifizialaren eta Hizkuntza Teknologiaren HiTZ Katedrak profesionalak prestatzeko eta ezagutza eta teknologia gizarteari transferitzeko aukera ematen digu. Era berean, HiTZek koordinatzen duen CLARIAH-ES ikerketa-azpiegiturak gure lankidetza-sare nazionalak eta nazioartekoak sendotu ditu gizarte-zientzietarako eta humanitateetarako tresna eta baliabide digitalak eskainiz. Bukatzeko, GenAIren onurak domeinu berrietara zabaltzeko aktiboki lan eginez, hizkuntza bisualen eredu aitzindariak sortu ditugu, eta neurozientzialariekin batera asmo handiko proiektu batean lanean ari gara, hizketaren eta garunaren arteko elkargunea aztertuz. Egitasmo arrakastatsu horiek guztiak etorkizuneko erronkei ilusioz aurre egiteko oinarri sendoa dira HiTZentzat, eta era berean, gure herria Hizkuntza Teknologiaren eta Adimen Artifizial Sortzailearen ikerketan erreferente gisa kokatzen dute.



Eneko Agirre
HiTZeko zuzendaria

CLARIAH

HiTZek CLARIAH-ES ikerketa-azpiegitura banatua koordinatzen du. Azpiegitura horrek bideratzen du Espainiaren parte-hartzea Europako CLARIN (*Common Language Resources and Technology Infrastructure*) eta DARIAH (*Digital Research Infrastructure for the Arts and Humanities*) ikerketa-azpiegiturretan. CLARIAH-ESek ikerketa digitala hobetzea, eta ikertzaileei azken belaunaldiko baliabide, tresna eta zerbitzu digitalak eskaintzea du helburu. Diziplina anitzeko eta diziplinarteko ikuspegia du, eta Espainiako gizarte-zientzietako eta humanitateetako ikerketa goi mailara eramateko asmoa du. Espainiako humanitate digitalen komunitatea sustatu nahi du, eta era berean, Espainiako ikertzaileak nazioarteko proiektu eta programetan estrategikoki kokatzen lagundu nahi du, bereziki Europako Ikerketa Esparruan. German Rigau, HiTZeko zuzendariordea, da CLARIAH-ES, CLARIN eta DARIAHko Espainiako koordinatzaile nazionala.

2024an, CLARIAH-ESek bi kide berri hartu zituen: Dialnet, Espainiako argitalpen akademikoentzako datu-base bibliografiko garrantzitsua, eta SCAYLE, Gaztela eta Leongo Superkonputazio Zentroa. Bi kide berri horiekin, CLARIAH-ESek gaur egun Espainiako hamabi erakunde eta ikerketa-talde nagusietako bazkideak biltzen ditu. Erakunde horiek guztiek gizarte-zientzietan eta humanitateetan gaitasun digitalean aurrera egiteko asmoarekin lan egiten dute.

Testuinguru horretan, HiTZek CLARIAH-EUS nodoa abian jarri du, eta haren koordinazioaren ardura ere hartu du. Nodo horrek euskararen ikerketan eta gizarte-zientzietan zein humanitateetan diharduten hainbat erakunde eta talde biltzen ditu. CLARIAH-EUSEk Eusko Jaurlaritzaren, Gipuzkoako Foru Aldundiaren, Euskal Herriko Unibertsitatearen (EHU) eta HiTZ zentroaren finantziazioa eta laguntza jasotzen du.



German Rigau

HiTZeko Zuzendariordea eta CLARIAH-ES, CLARIN eta DARIAH-en Espainiako Koordinatzaile Nazionala



Xabier Arregi

CLARIAH-EUS-en Koordinatzailea

CLARIAH-ES



CLARIAH-EUS



KATEDRA

2024an, Adimen Artifizialaren Estrategia Nazionalaren (ENIA) baitan, Adimen Artifizialaren eta Hizkuntza Teknologiaren HiTZ Katedra (AA&HT) abiarazi genuen. Katedra hori Eraldaketa Digitalerako Ministerioak finantzatzen du, Adimen Artifizialean ikerketa, zabalkundea, irakaskuntza eta berrikuntza bultzatzen duten unibertsitate-katedrak babesteko ekimenaren bitartez. Euskal Herrian finantzatutako katedra bakarra da, eta baita, Espainian Hizkuntza Teknologiari eskainitako katedra bakarra ere.

HiTZ Katedraren helburu nagusiak honako hauek dira: AA&HTk gizartean, eta bereziki sektore produktiboan, duen eragin positiboa areagotzea; AA&HTn ikerketa-jarduerak indartzea; eta teknologia hori erabiltzen duten hainbat arlotako hezkuntza-programak babesteko ezagutza sendoa sortzea. Horretarako, hiru lan-ildo jorratzen ditu: hezkuntza, ikerketa eta ezagutzaren transferentzia. Ildo horiek Katedrari atxikitako enpresekin lankidetzan garatzen dira, Avature SL, Elhuyar Fundazioa, Ikerlan SCL, Euskaltel-MasOrange, Multiverse Computing eta Tecnalía Ikerketa eta Berrikuntza Fundazioa. Erakunde horien kopuruak, garrantziak eta bikaintasunak erakusten dute sektore produktiboak Katedrak gizartean izan dezakeen eraginean interesa duela.

emari ta zabal zazu
Universidad
del País Vasco
Euskal Herriko
Unibertsitatea

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



Katedrak Adimen Artifizial Sortzaileari, Ikasketa Sakonari eta Hizkuntza Teknologiari buruzko lau espezializazio-ikastaro eskaintzen ditu, profesionalak, ikertzaileak eta ikasleak arlo zirrargarri horretan azkar sartzeko diseinatuta daudenak. Ikastaro horiek arloko aditu nagusiek ematen dituzte (hitz.eus/training).



Aitor Soroa

Adimen Artifiziala eta Hizkuntza Teknologien (AA&HT) HiTZ Katedraren Zuzendaria

Financiado por la Unión Europea
NextGenerationEU



HiTZ ZENBAKITAN

Ikerkuntza eta transferentzia

34

ikerkuntza
proiektuak

5

ezagutza
transferitzeko
kontratuak

7

doktoretza-tesiak
defendaturik
(7 nazioarteko)

18

aldizkari-
artikuluak
(11 Q1)

56

artikuluak
biltzarretan
(18 A or A+)

3

liburu-
kapituluak

1

liburua



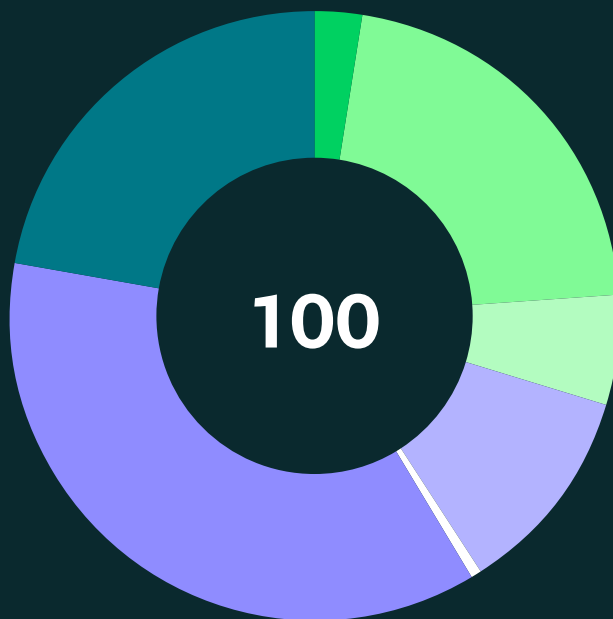
hitzak

language

Pertsonak

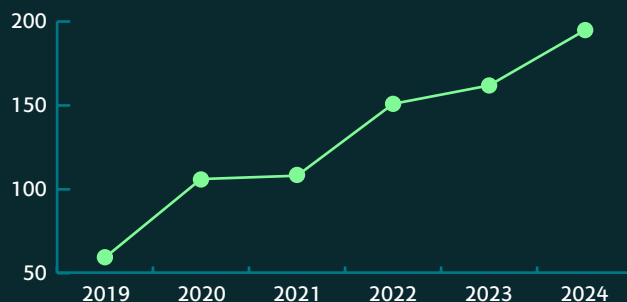
Kideak

	Kudeaketa eta administrazioa	5
	Irakasleak	42
	Doktoratu ondoko ikertzaileak	11
	Beste ikertzaileak	22
	Finantzatutako aurredoktoregoko ikertzaileak	19
	Kolaboratzaile akademikoa	1

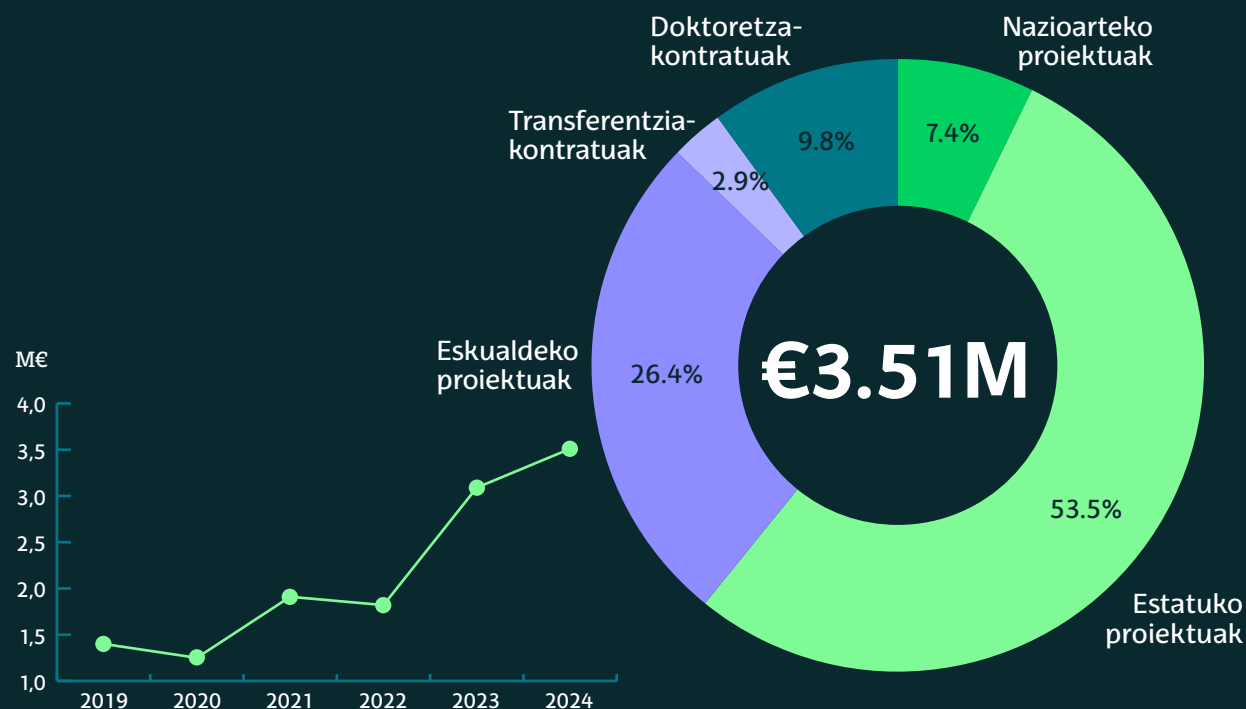


Ikasleak

	Master ikasleak	71
	Doktoretza ikasleak	43



Aurrekontua



IKERKETA-ARLOAK



**Informa-
zioaren
Erauzketa
eta Berres-
kurapena**

Ikertzaile nagusia:



**Aitor
Soroa**



**Itzulpen
Automatikoa**

Ikertzaile nagusia:



**Gorka
Labaka**



**Gizakiaren eta
Ordenagailua-
ren arteko
elkarreragina**

Ikertzaile nagusia:



**Gorka
Azkune**



**Hizketa- eta
Hizkuntza-
Baliabideak**

Ikertzaile nagusia:



**Ainara
Estarrona**

AZPIEGITURAK

18

Multiprocessor
GNU/Linux

1

SPARC
Solaris
zerbitzaria

1

HPC
Clusterra,
128
nukleoduna

14

zerbitzari
(71 GPU)

36

GPU cluster
A100 80GB
vram

98

GPU cluster
A100 80 GB vram
(partekatua / DIPC)

14

L40 GPU
(48 GBko
vram)

26

GPU
modelo
desberdinak

1.2 PB

sareko biltegitratze
ahalmen gordina



Testuen Analisia

Ikertzaile nagusia:



Rodrigo
Agerri



Hizketa- teknologiak

Ikertzaile nagusia:



Inma
Hernaez



Hizkuntzaren Prozesamen- dua Medi- kuntzan eta Zuzenbidean

Ikertzaile nagusia:



Arantza
Casillas



Humanitate Digitalak eta Hezkuntza

Ikertzaile nagusia:



Mikel
Iruskietza

1

Akustikoki
isolaturiko
gela

1

Behringer 4x4
audio/MIDI
interfazea

1

Quiet PC
Sentinel
Fanless i10

1

Delsys Trigno™
Wireless EMG
System

EuroHPC SuperComputer-en 1,5 milioi GPU orduko sarbidea (4,2 milioi euroko baliokoa), baliabide gutxiko Europako hizkuntzetarako eredu handiak ikertzeko.



PRAKTIKAK

71

masterretako
ikasle

43

doktorego
programako
ikasle

6

MALak
amaiturik
EMLCTn

21

MALak
amaiturik
HAP/LAPen

116

ikasle HEHrekin lotutako
4 etengabeko hezkuntza
ikastarotan

5

Ikasikerrak

17

barne- eta
kanpo-praktikak





27

mintegiak

hitzak

5

web-
mintegiak

5

tailer
antolatutak

4

sariak

words

hizkuntza

Association for Computational Linguistics-en
Baliabide Artikulu Onenerako saria: "Latxa: An Open
Language Model and Evaluation Suite for Basque"

SomosNLP #Somos600 hackathoneko proiektu
irabazlea: "Noticia: Resumen de Noticias Clickbait"

Albayzin-Iberspeech-RTVE 2024 Speaker Diarization
and Identity Assignment Challenge-ko irabazlea

Iberspeech 2024ko Ikasleen Artikulu Onena: "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