VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

EuMediCS – Euskarazko Medikuntzaren Domeinuko Corpus Sintetikoa Itzultzaile Automatikoen Ekarpena

Ane G. Domingo-Aldama, Irune Palacios, Maitane Urruela, Iker De la Iglesia, Ander Barrena eta *Josu Goikoetxea*

173-180 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.21

EuMediCS - Euskarazko Medikuntzaren Domeinuko Corpus Sintetikoa Itzultzaile Automatikoen Ekarpena

Ane G. Domingo-Aldama[1⋆], Irune Palacios[1⋆], Maitane Urruela[1],
Iker De la Iglesia[1], Ander Barrena[1], Josu Goikoetxea[1]

[1] HiTZ Zentroa - Ixa. Bilboko Ingeniaritza Eskola (UPV/EHU)

anegdaldama@gmail.com

Laburpena

Azken urteotan, Hizkuntza-Eredu Handiek (HEH) adimen artifizialaren alorra erabat irauli dute, itzulpena eta testu-sintesia bezalako zereginetan berebiziko arrakasta lortuz. Medikuntzaren domeinuan ere, eredu horiek errendimendu handia erakutsi dute, zenbait kasutan gizakion mailara hurbilduz, zeregin horietarako entrenatu direnean. Hala ere, HEH espezializatuen inguruko aurrerapen gehienak baliabide ugari dituzten hizkuntzetan egin dira, hala nola, ingelesean. Hori euskara bezalako hizkuntza gutxituen kalterako da, hizkuntza horietako HEHak urriak eta kalitate baxuagokoak baitira; domeinu espezializatuetan, medikuntzan esate baterako, sarritan existitu ere ez dira egiten. Gauzak horrela, proiektu honen helburua euskarazko medikuntzaren domeinuko lehenengo corpus sintetikoa sortzea da, etorkizunean domeinu horretako HEH bat elikatuko duena. Hori lortze aldera, lehenik, medikuntzaren domeinuko hiru itzulpen-eredu aurkeztuko dira, gaztelaniazko eta ingelesezko testuak euskarara itzultzen dituztenak; ondoren, hiru itzulpen-eredu horien kalitatea ebaluatuko da; azkenik, eredurik egokiena aukeratuko da, eta azken horrekin etorkizunean testuak masiboki euskarara itzuliko dira. Hala, lan hau domeinura egokitutako euskarazko HEHen sorkuntzan ekarpen esanguratsua da, zehazki, medikuntzaren domeinuan.

Hitz gakoak: Hizkuntza-Eredu Handiak, Itzulpen Automatiko Neuronala, Euskarazko Corpusa, Medikuntza

Abstract

In recent years, Large Language Models (LLM) have significantly transformed the field of artificial intelligence, achieving remarkable success in tasks such as translation and text synthesis. In the medical domain, these models have also demonstrated great performance, even reaching the human level in some cases, when they have been specifically trained for the task. However, most of the advancements using LLMs have been made in high-resource languages like English, which means a great disadvantage for low-resource languages like Basque, since the LLMs trained in these languages are few and of low quality in comparison, due to the lack of data. Moreover, in specific domains like medicine, these models do not even exist in most cases. To address this gap, this project aims to create the first synthetic medical-domain corpus in Basque, in order to train an LLM in that context. To achieve this, we propose three translation models capable of translating from Spanish and English to Basque. Then, the quality of these translation models will be evaluated, ultimately selecting the best of them for a large-scale translation of medical texts into Basque in the future. This work represents a notable contribution to the development of specialized LLMs for Basque, particularly in the medical domain.

Keywords: *Large Language Models, Neural Machine Translation, Basque Corpus, Medical-domain*

1 Sarrera eta motibazioa

Hizkuntza-Eredu Handiek (HEH) adimen artifizialaren alorra guztiz aldatu dute, arrazonamendu konplexuran, ulermenean eta hainbat datu-mota sortzeko dituzten gaitasunak direla-eta. Sistema horiek domeinuko hainbat ataza ebazteko ahalmena erakutsi dute; adibidez, chatbot-en garapenean (Kim *et al.*, 2023), arlo legalean (Lai *et al.*,

⋆ Pareko ekarpena.

2024), baita eremu medikoan ere (Thirunavukarasu *et al.*, 2023; Zhou *et al.*, 2023). Eredu eleaniztun irekiak sortzeko ahaleginak egin badira ere, eredu sendoenak ingelesarentzat eta baliabide handiko hizkuntzentzat zentratuta daude. Hala ere, Etxaniz *et al.* (2024) ikerketan, *Latxa* HEHa aurkeztu zuten, hainbat eredu ireki gainditzen dituena euskarazko zenbait atazetan.

Domeinu honetan medikuntzara egokitutako HEHek eredu orokorrek baino emaitza hobeak lortzen dituztela (Chen *et al.*, 2024) frogatu da, eta, gainera, eredu horietako batzuk euskara prozesatzeko ere gai dira. Hori kontuan hartuta, euskarazko medikuntzaren domeinuko HEH bat garatzeak interes handia du, euskal komunitatean hobekuntzak ekarriko baititu osasun-arretaren irisgarritasunean eta kalitatean. Zehazki, pazienteen eta osasungintzako profesionalen arteko komunikazioa hobetzeaz gain, osasungintzako profesionalen prestakuntzan laguntzeko eta hura pertsonalizatzeko tresna aurreratuak eskainiko ditu. Domeinuko HEH horien ekarpen nagusien artean, osasungintza bezalako testuinguru kritikoan pazienteen ongizatearen hobekuntza, eta, esan gabe doa, domeinu horretan euskararen erabilera sendotzea leudeke.

Domeinura egokitutako HEH bat garatzeko, eredua entrenatzeko medikuntza alorreko datuak izatea ezinbestekoa da. Hala ere, euskara baliabide urriko hizkuntza da eta kalitatezko datu medikoak lortzea konplexua (García-Ferrero *et al.*, 2024); beraz, kalitatezko corpus bat osatzea erronka handia da.

Gauzak horrela, lan honen helburu nagusia euskarazko medikuntzaren domeinuko aurreneko corpus sintetiko handia sortzea da, etorkizunean HEH espezializatu bat entrenatu eta domeinura egokitzeko, edota beste tresna batzuk sortzeko ezinbesteko baliabidea izango dena. Hori lortzeko domeinuko corpusa gaztelaniatik eta ingelesetik (iturri-hizkuntzak) masiboki itzuliko da euskarara (helburu-hizkuntza), publikoki atzigarri diren medikuntzaren domeinuko hainbat corpusetatik (ikusi 3. atalaren hasiera), eta lan honetan jatorrizko corpus horiek euskaratzeko itzultzaile egokienak aukeratzeko metodo bat proposatzen da. Ondorioz, metodo hori hizkuntza gutxituentzat ere ekarpen esanguratsua da, domeinura egokitutako HEH gabezia nabarmena baita hizkuntza nagusietatik at.

2 Arloko egoera eta ikerketaren helburuak

Medikuntzaren domeinuko esaldiak euskarara itzultzen trebea den eredu bat garatzea erronka handia da, euskara baliabide urriko hizkuntza gisa ezaugarritzen delako eta medikuntzaren domeinuko terminologia hizkuntza generikotik nabarmen aldentzen delako. Gainera, itzulpen horien ebaluazioa egitea ere zaila izan daiteke. Atal honetan, gure proiektuaren testuingurua argituko da, aipatutako konplexutasunetan arreta berezia jarriz.

2.1 Baliabide urriko hizkuntzak eta domeinuaren egokitzapena IAN sistemetan

Baliabide urriko hizkuntza-bikoteetarako Itzulpen Automatiko Neuronalen (IAN) tekniken inplementazioak arreta handia erakarri du azken urteetan. Eredu gehienak datu-multzo paralelo[1] handietan oinarritzen direnez, datu-eskasiaren erronkari aurre egiteko irtenbide berritzaileak behar dira.

Aldi berean, domeinu espezifikoetara egokitutako IANa hastapenetan dago, eta komunitate zientifikoaren barruan gero eta interes handiagoa pizten ari da. Esaterako, *Workshop on Machine Translation* (WMT) kongresuan medikuntzaren domeinuko azpiataza bat dago, itzulpen klinikoan espezializatutako IAN ereduen garapena sustatzen duena (Bawden *et al.*, 2020), eta euskara da sartutako hizkuntzetako bat. Domeinuaren egokitzapena sustatzen duten tekniken artean, ohikoa da terminoen hiztegiak erabiltzea itzulpenen kalitatea hobetzeko, askotan atzeranzko-itzulpenak[2] baino hobeto funtzionatuz (Liu *et al.*, 2021; Soto, 2021).

HiTZ Zentroak eta Osakidetzak gaztelaniatik euskarara egokitzeko itzultzaile automatikoen garapenean lan egin dute Itzulbide proiektuan 2019tik 2021era eta 2023tik 2025era. Gainera, Soto (2021) ikerketa baliabide urriko hizkuntzetarako domeinu espezifikoetan espezializatutako IAN ereduak garatzeko egindako ahaleginen adibide argia da. Haien IAN sistemak osasun-erregistro elektronikoen txantiloiak euskaratik gaztelaniara itzultzeko gaitasuna du, testu klinikoen edo osasun-erregistroen corpus elebidunik erabili gabe. Itzulpenen kalitatea ebaluatzeko metrika automatikoak eta bi medikuk egindako eskuzko rankinga erabili ziren. Lortutako emaitzek *Google Translate* sistemarenak gainditu zituzten.

Artikulu honetan aurkezten diren itzultzaileak Soto (2021)-k proposatutako metodologian oinarritzen dira, non domeinu orokorreko corpus elebidunak euskarazko terminologia espezifikoarekin uztartzen diren. Gure kasuan,

[1] Corpus paralelo bat esaldi mailan lerrokatuta dauden bi hizkuntza edo gehiagoko testuen bilduma da.

[2] Datu gehikuntza teknika bat: iturri-hizkuntzako testua helburu-hizkuntzara itzultzen da lehenengo, eta ondoren azken horretatik berriro iturri-hizkuntzara.

Transformer arkitekturan (Waswani *et al.*, 2017) oinarritutako itzultzaileak garatu ditugu, ez soilik txosten medikuen itzulpenerako, baizik eta 3. atalean aipatzen diren corpus kliniko guztietarako. Horrez gain, ingelesa ere iturri-hizkuntza gisa gehitu dugu.

2.2 Itzulpenen ebaluazioa

Itzulpenen ebaluazioa beti izan da intereseko gaia, itzulpen baten kalitatea neurtzea hainbat irizpidetan oinarrituta baitago eta atazaren arabera, horiek alda daitezke. Eskuzko ebaluazioa garestia izan ohi da denborari eta baliabide ekonomikoei dagokienez, eta horrek metrika automatikoen arrakasta areagotu du. Hala ere, metrika horiek ez datoz beti bat giza ebaluazioarekin, eta normalean corpus paraleloak behar dira itzulpenen kalitatea ezartzeko.

Gauzak horrela, aipatutako ebaluazioen inguruko Yang *et al.* (2007) lanak hurrengoa frogatu du: giza irizpideak fidagarriagoak dira hainbat itzulpen-hautagairen kalitatea sailkatzean itzulpen isolatuen kalitatea puntuatzean baino.

2.3 Helburuak

Proiektu honen helburu nagusia hurrengoa da:

- **Euskarazko medikuntzaren domeinuko lehenengo corpus sintetikoa** sortzea: Corpus sintetiko hori etorkizunean HEH bat entrenatzeko oinarrizko baliabidea izango da, besteak beste. Corpus sintetikoa sortzeko, gaztelaniazko eta ingelesezko corpusa itzuliko da. Hori dela eta, artikulu honetan bi azpi-helburu nabarmentzen dira:
  - **Itzultzaile automatikoak sortzea:** Euskara helburu-hizkuntza duten medikuntzaren domeinurako itzultzaile automatikoak sortzea, gaztelania eta ingelesa iturri-hizkuntzak izanda.
  - **Sortutako itzultzaile automatikoen artean egokiena aukeratzea:** Horretarako, aipatutako itzultzaileen kalitatea ebaluatuko da.

3 Ikerketaren muina

Euskarazko medikuntzaren domeinuko testuen kopurua oso urria da, eta haiek eskuratzea ez da erraza. Orain arte, 2 milioi hitz inguruko euskarazko corpus originala bildu dugu, hainbat iturri eta corpus publikoetatik ateratakoa. Hala ere, tamaina hori ez da nahikoa euskarazko medikuntzaren domeinurako kalitatezko HEH bat lortzeko. Ondorioz, ezinbestekoa da gaztelaniazko eta ingelesezko medikuntza-datuak euskarara itzultzea. Horretarako, medikuntzaren domeinuko corpusa bildu da bi hizkuntzetan (ikus 1. taula).

Itzulpenak lortzeko hiru IAN sistema berri garatu ditugu, medikuntzako testuak ingelesetik eta gaztelaniatik euskarara itzultzeko gai direnak. Gainera, itzultzailerik egokiena zein den zehazteko, beste bi eredurekin alderatu ditugu. Hurrengo azpiataletan sistemen garapena, ebaluazioa eta lortutako emaitzak aurkezten dira.

3.1 Erabilitako itzultzaile automatikoak

Euskarazko medikuntzaren domeinuko corpus sintetiko bat lortzeko, kalitatezko itzultzaileak izatea funtsezkoa da. Gaizki eraikitako euskarazko medikuntzako corpus sintetiko batek terminologia zehaztugabea heda lezake, gaixoen osasuna arriskuan jarriz. Euskarazko kalitate handiko baliabide medikoak jada urriak direnez, datu-multzo sintetikoetan egindako akatsek hutsune horiek areagotuko lituzkete, hizkuntza-ereduak eta tresna klinikoak kaltetuz. Itzulpen okerrek diagnostikoak, tratamenduak edo pazientearen orientazioa baldintzatu dezakete, osasun laguntza-sistemei kalte eginez. Ondorioz, medikuntzako terminologia menperatzen duten euskarazko itzultzaile trebeak ezinbestekoak dira.

Jarraian 5 itzultzaile automatiko konparatu ditugu:

- **en2eu**: lan honetan sortutako ingelesa iturri-hizkuntza eta euskara helburu-hizkuntza dituen eredu elebiduna.
- **es2eu**: lan honetan sortutako gaztelania iturri-hizkuntza eta euskara helburu-hizkuntza dituen eredu elebiduna.
- **enes2eu**: lan honetan sortutako helburu-hizkuntza euskara duen eta iturri-hizkuntza bezala ingelesa eta gaztelania dituen eredu eleaniztuna.

Hizkuntza	Iturria	Hitz kopurua	Deskribapena
Euskara	E3C	564,7K	Semantikoki anotatutako kasu klinikoak.
	PubMed	1,53M	Literatura medikoaren artikuluen bilduma, MEDLINE eta zientziari buruzko aldizkariak bezalako iturrietatik bildua (PubMed Central (PMC), 2024).
Ingelesa	ClinicalTrials	127,4M	Artikulu zientifikoak, editorialak, txosten monografikoak, komunikazio laburrak eta arlo kliniko eta antzekoetarako garrantzitsuak diren beste datu batzuk.
	EMEA	12,9M	Corpus paraleloa, Sendagaien Europako Agentziaren PDF dokumentuekin egina (Tiedemann, 2012).
	PubMed	26.000M	Ingelesezko PubMed corpusa.
Gaztelania	EMEA	14,7M	Gaztelaniazko EMEA corpusa.
	PubMed (Abstract + Articles)	5,4M	Gaztelaniazko PubMed corpusa.
	Medical Crawler	850,7M	Internetetik erauzitako testu medikoen bilduma (Carrino *et al.*, 2021).
	SPACC	350K	Gaztelaniazko 1.000 kasu kliniko biltzen ditu (Intxaurrondo, 2018).
	UFAL	10,5M	Medikuntzako testuen corpus paraleloa (ÚFAL, 2017).
	WikiMed	16,3M	Medikuntza eta erlazionatutako alorren Wikipediako artikuluak.
	SNOMED	7,2M	Terminologia klinikoaren kontzeptuak eta deskribapenak (Snomed, 2013).

1. Taula: **Hizkuntza-Eredu Handien entrenamenduan erabiliko diren datu-multzoak eta haien ezaugarriak.** Lehenengo biak euskarazkoak dira, eta ondorengoak ingelesetik eta gaztelaniatik euskarara itzuliko direnak. Datu-multzoaren izena, hitz kopurua eta deskribapen txiki bat erakusten da.

- **Google Translate**[3]: Googlek garatu eta eskaintzen duen itzultzaile eleaniztun automatikoa.
- **Latxa**: Etxaniz *et al.* (2024) artikuluan aurkezten den euskarazko HEHa, itzultzaile gisa erabiltzeko baldintzatuz.

Lehenengo 3 ereduak guk sortutakoak dira, terminologia medikoa erabiliz; *Google Translate*, berriz, erreferentzia-eredu bat da artearen egoeran, eta hizkuntza-aukera zabal baterako euskarria du. *Latxa* HEH bat da, eta itzulpengintzarako entrenatuta ez badago ere, itzulpenak egiteko gai da baldintzapen tekniken bidez.

3.1.1 en2eu, es2eu eta enes2eu itzultzaileak

Domeinuko hiru itzultzaileak sortzeko, terminologia espezifikoa daukaten datu-multzoak eta sintaxi askotariko esaldi orokorrak bildu ditugu. Modu honetan, domeinuko datu-multzoak eta orokorrak konbinatzean, terminologia medikoaren eta hizkuntza-aniztasunaren estaldura zabala bermatzen da, ereduen itzulpenen terminologia espezifikoaren kalitatea hobetuz. Itzultzaileak entrenatzeko erabilitako datu-multzoei buruzko informazio zehatza 2. taulan aurkezten da. Guztira 22 milioi esaldi pare baino gehiago lortu dira ereduak entrenatzeko.

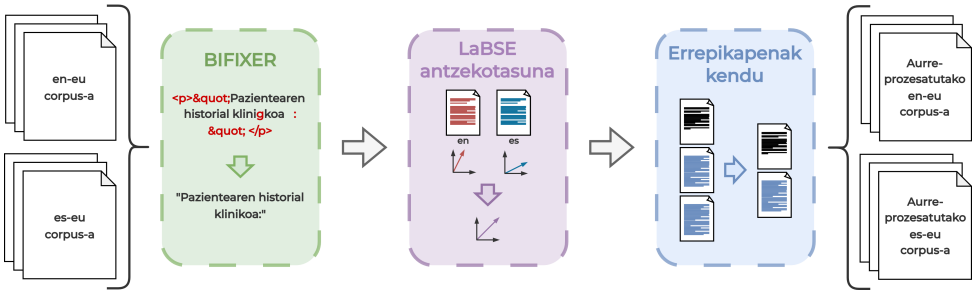
Datu mota	Datu-multzoa	Hizkuntzak	Esaldi pareak
Terminologia	ICD-10 [4]	es, eu	59.563
	snomed [5]	en, es, eu	1.282.698
Domeinuko testuak	elhuyar_med	en, eu	76.660
	elhuyar_kimika	en, eu	60.492
Esaldi orokorrak	General en-eu	en, eu	9.033.998
	General es-eu	es, eu	11.489.382

2. Taula: **Itzultzaileen entrenamenduan erabilitako corpusaren ezaugarriak.** Lehenengo zutabeak datu-multzo motari dagokion informazioa adierazten du, bigarrenak datu-multzoaren izena, eta hirugarrenak datu-multzo paralelo bakoitzaren hizkuntzak. Azkenik, datu-multzoaren tamaina aurkezten da, esaldietan neurtuta.

Sortutako corpus paralelo honen kalitatea hobetze aldera, aipatutako datu-multzoak aurreprozesatu eta garbitu egin dira (ikus 1. irudia) hurrengo pausoei jarraituz:

[3] `https://cloud.google.com/translate/docs/reference/rest`
[4] UZEI: `https://uzei.eus/en/`
[5] Pérez de Viñaspre Garralda (2017)

1. Corpusaren bilketa eta kateamendua. 2. taulan agertzen diren datu-multzoak batu eta ausaz ordenatu dira, sistema bakoitzerako datu-sorta bakar bat sortzeko.
2. *Bifixer*[6] tresna erabiliz, corpus paraleloaren garbiketa orokorra egin da: testu normalizatzea, HTML etiketak ezabatzea, ortografia eta erroreen iragazketa, etab.
3. Ondoren, lortutako corpus elebiduna hainbat iturritik hartu denez, esaldien iragazketa egin da horren kalitatea ziurtatzeko. Horretarako, *LaBSE* (Feng *et al.*, 2022) eredu eleaniztuna erabiliz, iturri- eta helburu-esaldien errepresentazio bektorialak lortu dira eta haien arteko antzekotasuna kalkulatu da kosinua erabiliz. Horri esker, kalitate baxuko itzulpenak edo gaizki parekatutako esaldiak baztertu daitezke.
4. Azkenik, esaldi errepikatuak kendu dira.



1. Irudia: **Itzultzaileak entrenatzeko datu-multzoei egindako aurreprozesua eta garbiketa.** Lehenik, hizkuntza bakoitzerako corpusak biltzen dira. Ondoren, testuaren garbiketa egiten da Bifixer tresna erabiliz. LaBSE antzekotasun metrika aplikatuz perpausen antzekotasuna kalkulatzen da, eta antzekotasun txikia duten esaldi paraleloak baztertzen dira. Azkenik, errepikapenak ezabatzen dira.

Ereduak entrenatzeko, MarianMT *framework*-a[7] erabili da, *Transformer* (Waswani *et al.*, 2017) arkitekturan oinarritua. Sortutako ereduak *PyTorch*-era moldatu dira haien erabilpena errazteko.

3.2 Itzultzaileen ebaluazioa

3.2.1 Eskuzko ebaluazioa

Proposatu berri diren *en2eu*, *es2eu* eta *enes2eu* itzultzaileen kalitatea ebaluatzeko eta *Google Translate* eta *Latxa* ereduek sortuko dituzten itzulpenekin alderatzeko, eskuzko ebaluazio bat planteatu da. Itzulpenen kalitatea ebaluatzeko medikuntzari aplikatutako adimen artifizialaren domeinuan lan egiten duten 5 aditu hirueledun bildu dira. Zehazki, adituek esaldi bakoitzaren 4 itzulpenak sailkatu behar izan dituzte 1etik 4ra ordenatuz. Itzulpenik hoberenari lehen postua esleituko zaio eta okerrenari laugarren postua. Gainera, postuak errepikatzeko aukera eduki dute.

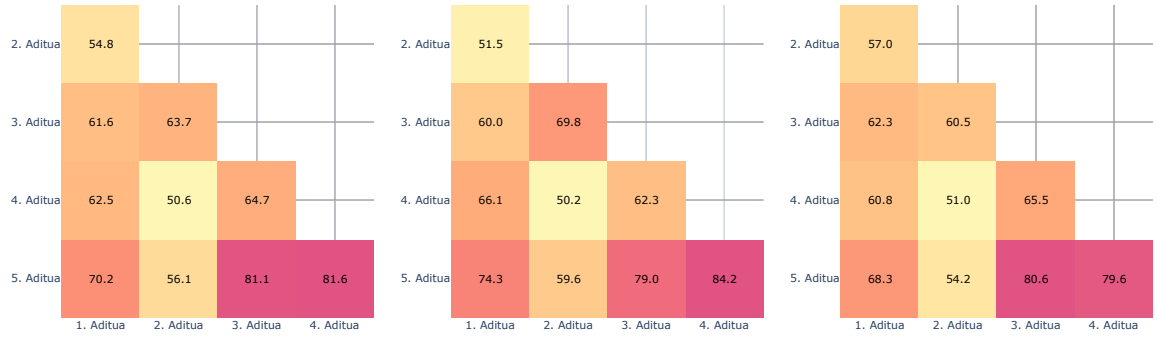
Itzulpenak ebaluatzeko irizpide berdinak aplikatzen direla ziurtatzeko, gida bat garatu da. Bertan, testuinguru honetan itzulpen on bat zer den ezarri da eta hurrengo lau irizpideak ditu oinarri:

- **Euskararen zuzentasuna**. Itzulpenak euskararen gramatika-arauak eta ortografia errespetatu behar ditu, hizkuntza egituraz zuzena eta koherentea izanik.
- **Naturaltasuna**. Jatorrizko testua zorrotz jarraitzeaz gain itzulpenak euskaraz modu jator eta ulergarrian idatzita egon behar du, itzulpen-zantzurik gabe.
- **Terminologia egokia**. Arlo medikoan bereziki garrantzitsua da terminologia teknikoaren erabilera egokia.
- **Osotasuna**: Itzulpenak jatorrizko testuak duen edukia mantendu behar du.

Aditu guztiek irizpide horiek modu egokian aplikatzen dituztela ziurtatzeko, bost adituek esaldi berdinak ebaluatu dituzte eta adostasuna ezartzeko *Inter Tagger Agreement* (ITA) kalkulatu da, zehazki, Krippendorff-en alfa metrika (Krippendorff, 2011) erabiliz.

[6] https://github.com/bitextor/bifixer
[7] https://huggingface.co/docs/transformers/model_doc/marian

ITA kalkulatzeko, 1. taulako iturri bakoitzetik ausaz 2 esaldi hartu dira, ingelesezko *Clinical Trials* eta gaztelerazko *PubMed* corpusetatik izan ezik, 4 hartu direla, horietan erabiltzen den terminologia bereziki konplexua delako. Gauzak horrela, guztira 24 esaldi bildu dira, 16 gazteleratik eta 8 ingelesetik.



(a) (b) (c)

2. Irudia: **Adituen artean lortutako ITAren emaitzak, Krippendorff-en alfa erabiliz;** (a) irudian ingelesezko eta gaztelerazko itzulpenen adostasunaren emaitza orokorrak aurkezten dira, (b) irudian ingelesezkoena eta (c) irudian gaztelerazkoena. ITA balioa zenbat eta altuagoa izan, orduan eta kolore ilunagoarekin adierazten da. ITA balio altuak adituen arteko adostasun handia islatzen du.

2. irudiaren ezkerraldean adituen arteko sailkapenen adostasun maila ikus daiteke 24 esaldientzat. Batez bestean % 64ko adostasuna lortu da, eta ingelesezko eta gaztelerazko azpimultzoetan adostasun maila mantentzen da. Hortaz, adostasun maila altua da ataza honetarako.

Behin anotatzaileen adostasuna baieztatuta, laginaren tamaina 100 esaldira handitzea erabaki da, ebaluazioan emaitza esanguratsuagoak lortzeko. Horretarako, 76 instantzia gehiago hartu dira 1. taulan aurkeztutako datumultzoetatik, eta bost anotatzaileen artean banatu dira ausaz. Beraz, anotatzaile bakoitzari beste 15 esaldi (bati 16) esleitu zaizkio. 3. taulan instantzia berri hauen banaketa azaltzen da. Hizkuntzen arabera ere orekatu egin dira kopuruak; hortaz, guztira, gaztelaniazko 50 instantziek eta ingelesezko beste 50ek osatzen dute orain lagina.

	Ingelesa			Gaztelania							Guztira
Iturria	ClinicalTr.	EMEA	PubMed	EMEA	PubMed	Crawler	SPACCC	UFAL	Wiki	SNOMED	–
ITA	4	2	2	2	4	2	2	2	2	2	24
Berriak	14	14	14	4	10	4	4	4	4	4	76

3. Taula: **Instantzia berrien banaketa, hizkuntzaren eta iturriaren arabera.** Lehenengo lerroak iturriak adierazten ditu, eta bigarrenak iturri bakoitzetik ITA egiteko hartutako instantzia kopurua. Hirugarren lerroak anotatzaileen arteko adostasuna baieztatu ostean iturri bakoitzetik hartutako instantzia kopurua islatzen du. Datu-multzoak hizkuntzaren arabera banatuta agertzen dira zutabeetan.

3.2.2 Ebaluazio automatikoa

Ingelesetik euskararako itzulpenen ebaluazio automatikorako, WMT21 kongresuan aurkeztutako Osagaiz corpusa (Yeganova *et al.*, 2021) erabili da. Bertan laburpen biomedikoen itzulpen-pareak batzen dira. Ez da gaztelaniatik euskarara itzulpenak ebaluatzeko corpusik aurkitu, eta eskuragarri dauden terminologia corpusak jada itzultzaileen entrenamendu corpusaren parte dira.

3.3 Sailkapenaren emaitzak

Batetik, 4. taulan ageri den moduan, gaztelera iturri-hizkuntza den itzulpenen artean, *es2eu* itzultzaile elebidunak egindakoak lehenengo postuan geratu dira, eta horien atzetik *enes2eu* itzultzaile eleaniztunak egindakoak. *Latxa* eta *Google Translate* ikerketa honetan garatu diren itzultzaileen atzetik sailkatu dira. Bestetik, ingelesa iturri-hizkuntza bezala duten itzulpenetan, beste behin itzultzaile elebidunak (*en2eu* oraingoan) puntuazio altuena lortu du. *enes2eu* itzultzaileak bigarren postua eskuratu du eta *Latxa* eta *Google Translate* azkenengo postuetan daude.

Beraz, argi dago 4. taulako emaitzei erreparatuz, *es2eu* eta *en2eu* izan direla adituek hoberen sailkatutako ereduak. Ondorioz, agerian geratzen da medikuntzako terminologiaz sortutako itzultzaileen beharra. Gainera, *Latxaren* itzulpenek lortutako 3. posizioak argi erakusten dute medikuntzaren domeinuan duen gabezia.

Era berean, ingelesetik euskararako itzulpen-sistemen ebaluazio automatikoaren emaitzak ere lortu dira. BLEU metrikak argi erakusten du gure ereduek besteek baino errendimendu hobea dutela eta horrek berretsi egiten du domeinu medikoko terminologia barneratzearen garrantzia, itzulpenen kalitatea hobetzeko. Eskuzko eta ebaluazio automatikoen artean desberdintasunak daude, hala ere, bietan argi geratzen da medikuntzako terminologiaren beharra.

		Itzultzaileen sailkapena			
Ebaluazio mota	**Iturri-hizkuntza**	1.	2.	3.	4.
Eskuzkoa	*Gaztelera*	es2eu (1.67±0.76)	enes2eu (2.04±0.89)	Latxa (2.61±1.04)	Google Tr. (3.25±0.91)
	Ingelesa	en2eu (2.03±1.04)	enes2eu (2.20±0.94)	Latxa (2.40±0.99)	Google Tr. (2.95±1.06)
Automatikoa	*Ingelesa*	enes2eu (11.57)	en2eu (11.30)	Latxa (10.67)	Google Tr. (9.01)

4. Taula: **100 instantzien ebaluazioan lortutako itzultzaileen sailkapena.** Lehenengo bi lerroetan eskuzko ebaluazioa eta azken lerroan ebaluazio automatikoa aurkezten dira. Posizio bakoitzean dagokion itzultzailearen izena, baita horrek lortutako puntuazioa agertzen da: eskuzko ebaluazioaren kasuan batezbesteko posizioa eta desbiderapena, eta ebaluazio automatikoan BLEU metrika.

4 Ondorioak eta etorkizuneko ildoak

Ikerketa hau euskarazko HEHak medikuntzaren domeinura egokitzeko aurrerapauso bat gehiago da, xede nagusia alor konkretu horretako aurreneko corpus sintetiko erraldoia sortzea izanik.

Hasteko, gaztelerazko eta ingelesezko iturri-hizkuntzen corpusen sorkuntza azaldu da, eta guztira 27.395 milioi hitz bildu dira hainbat iturritatik. Ondoren, IANak sortu dira aipatutako corpus hori euskarara itzultzeko. Gainera, itzultzaile horiek ebaluatzeko bi metodologia aurkeztu dira, eskuzkoa eta automatikoa. Eskuzko ebaluazioaren kasuan, adituen itzulpenen sailkapenetan oinarrituta itzultzaile onenak aukeratu dira, eta horretarako ebaluazio irizpide argiak ezarri dira. Adituek adostasun handia erakutsi dute itzulpenen sailkapenean, eta ikerketa honetan proposatutako *es2eu* eta *en2eu* itzultzaileek lortu dituzte emaitzarik hoberenak. Hala, domeinuko terminologian espezializatutako ereduak orokorrak baino egokiagoak direla frogatu da. Are gehiago, 4. taulan adierazten den moduan, bi eredu horiek *Google Translate* bezalako eredu komertzialek baino kalitate hobeagoko itzulpenak lortzeko gai direla frogatu dugu. Ingelesezko itzulpenen ebaluazio automatikoak emaitza hauek berretsi ditu.

Etorkizunean 1. taulako gaztelerazko eta ingelesezko corpusak masiboki itzuliko dira euskarara *es2eu* eta *en2eu* itzultzaileen bidez, eta lortutako euskarazko corpus sintetiko hori *Latxan* oinarritutako HEH bat medikuntzan trebatzeko erabiliko da.

Ikerketa honen berregingarritasuna bermatze aldera, garatutako itzultzaileak erabiltzeko kodea publikoki eskuragarri dago[8]. Etorkizunean, sortutako euskarazko corpus sintetikoa ere eskuragarri egongo da biltegi berean.

5 Eskerrak eta Oharrak

Lan hau HiTZ Zentroaren eta Eusko Jaurlaritzaren (IXA bikaintasun ikerketa taldearen IT-1570-22 finantzaketa) laguntzarekin egin da, baita Espainiako Unibertsitate, Zientzia eta Berrikuntza Ministerioaren (MCIN/AEI/10.13039/501100011033) Proyectos de Generación de Conocimiento 2022 EDHER-MED/EDHIA (PID2022-136522OB-C22) proiektuaren bidez ere. Halaber, lehenengo egileak Eusko Jaurlaritzaren Doktoratu Aurreko Programaren laguntza jaso du (PRE_2024_1_0224), eta laugarren egileari Espainiako Zientzia, Berrikuntza eta Unibertsitate Ministerioaren (MCIUren) FPU beka aitortu diote (FPU23/03347).

[8] https://github.com/hitz-zentroa/EuMediCS-IKERGAZTE-2025

Erreferentziak

Bawden, Rachel, Giorgio Maria DiÑunzio, Christian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, eta others. 2020. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Fifth Conference on Machine Translation*, 658–685. Association for Computational Linguistics (ACL).

Carrino, Casimiro Pio, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, eta Marta Villegas, 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Chen, Xi, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaoting Zhang, Gang Chen, Kang Li, eta Jian Li. 2024. Evaluating and enhancing large language models' performance in domain-specific medicine: Development and usability study with docoa. *Journal of Medical Internet Research* 26.e58158.

Etxaniz, Julen, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, eta Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *arXiv preprint arXiv:2403.20266* .

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, eta Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Smaranda Muresan, Preslav Nakov, eta Aline Villavicencio, 878–891, Dublin, Ireland. Association for Computational Linguistics.

García-Ferrero, Iker, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, eta others. 2024. Medical mt5: an open-source multilingual text-to-text llm for the medical domain. *arXiv preprint arXiv:2404.07613* .

Intxaurrondo, Ander, 2018. Spaccc. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Kim, Jin K, Michael Chua, Mandy Rickard, eta Armando Lorenzo. 2023. Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology* 19.598–604.

Krippendorff, Klaus, 2011. Computing krippendorff's alpha-reliability.

Lai, Jinqi, Wensheng Gan, Jiayang Wu, Zhenlian Qi, eta S Yu Philip. 2024. Large language models in law: A survey. *AI Open* .

Liu, Ling, Zach Ryan, eta Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, 44–50.

Pérez de Viñaspre Garralda, Olatz, 2017. *Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea*. Universidad del País Vasco-Euskal Herriko Unibertsitatea tesia.

PubMed Central (PMC), 2024. PubMed Central Corpus.

Snomed, CT, 2013. International health terminology standards development organisation.

Soto, Xabier, 2021. *Txosten klinikoak euskararen eta gazteleraren artean itzultzen laguntzeko corpusaren bilketa eta itzultzaile automatikoaren garapena*. Universidad del País Vasco-Euskal Herriko Unibertsitatea tesia.

Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, eta Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29.1930–1940.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, 2214–2218. Citeseer.

Waswani, A, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, eta I Polosukhin. 2017. Attention is all you need. In *NIPS*.

Yang, Y, M Zhou, eta CY Lin. 2007. Sentence level machine translation evaluation as a ranking problem: one step aside from bleu. In *Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic*, 240–247.

Yeganova, Lana, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, eta others. 2021. Findings of the wmt 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *EMNLP 2021-Sixth Conference on Machine Translation.*.

Zhou, Hongjian, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, eta others. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* .

ÚFAL, 2017. UFAL: Medical Corpus.