VI. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

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

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Euskara eta gaztelaniazko kontra-narratiben sorkuntza: datuen sorrera eta ebaluazioa

Jaione Bengoetxea,
Itziar Gonzalez-Dios
eta *Rodrigo Agerri*

133-140 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.16

Euskara eta gaztelaniazko kontra-narratiben sorkuntza: datuen sorrera eta ebaluazioa

Jaione Bengoetxea, Itziar Gonzalez-Dios, Rodrigo Agerri

HiTZ Hizkuntza Teknologiako Euskal Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU

jaione.bengoetxea@ehu.eus

Laburpena

Kontra-narratibak (KN) gorroto-diskurtsoen (GD) erantzun ez-negatiboak dira, online gorrotoa eta haren hedapena murrizten laguntzen dutenak. Sare sozialetan GDaren presentzia hazi den arren, KNen sorkuntza automatikoari buruzko ikerketa mugatua da oraindik, eta egindako lan gehienak ingeleserako izan dira. Artikulu honek CONAN-EUS aurkezten du, euskarazko eta gaztelaniazko KNen sorkuntzarako datu-multzoa, Itzulpen Automatikoa (IA) eta post-edizio profesionala erabiliz sortua. Ingelesezko CONAN datu-multzoaren corpus paralelo gisa, hizkuntza anitzeko eta hizkuntza arteko KNen sorkuntzan ikerketa ahalbidetzen du. mT5 hizkuntza ereduarekin egindako esperimentuetan ikusten da post-editatutako datuek KNen sorkuntza-kalitatea hobetzen dutela. Eskuzko ebaluazioak baieztatu du datuen eskuzko berrikusketa beharrezkoa dela. Hizkuntza anitzeko datu-gehikuntzak gaztelaniari mesede egiten dio, baina ez euskarari, hizkuntza anitzeko eredu sortzaileen erronkak agerian utziz.
Eduki-oharra: Artikulu honek hizkera iraingarriaren adibideak ditu, ez datozenak bat egileen ikuspuntuekin.

Hitz gakoak: kontra-narratibak, gorroto-diskurtsoa, eleaniztasuna, testu sorkuntza

Abstract

Counter Narratives (CNs) are non-negative responses to Hate Speech (HS) that help reduce online hatred and its spread. Despite the rise in HS online, research on automatic CN generation remains limited, most works being done in English. This paper introduces CONAN-EUS, a Basque and Spanish dataset for CN generation, created using Machine Translation (MT) and professional post-editing. As a parallel corpus to the English CONAN dataset, it enables research on multilingual and crosslingual CN generation. Experiments with the language model mT5 show that training on post-edited data improves CN generation quality compared to using only MT data. Manual evaluation confirms that manually revising data remains crucial. Multilingual augmentation benefits Spanish but not Basque, highlighting challenges in multilingual generative models.
***Content Warning**: This paper contains examples of offensive language that do not reflect the authors' views.*

Keywords: *Counter Narratives, Hate Speech, Multilinguality, Text Generation*

1 Sarrera eta motibazioa

Azken urteetan gorroto-diskurtsoaren (GD) presentzia areagotu egin da komunikabideetan, sare sozialek sortzen duten anonimizazioak bultzatuta. GDa "talde jakin baten aurkako gorrotoa adierazteko erabiltzen den hizkera"gisa definitu da, "talde horretako kideak iraindu, umiliatu edo gutxiesteko asmoa duen hizkera"(Davidson *et al.*, 2017).

Gaur egun, sare sozialek etengabe eguneratzen dituzte gorroto-mezuei aurre egiteko politikak, mezuak blokeatuz edo ezabatuz. Horrek dakarren lan-karga arintzeko, GDren detekzio automatikoan interesa piztu da, datu-multzoak garatuz (Basile *et al.*, 2019) eta ikasketa automatiko eta sakoneko teknikak erabiliz (Davidson *et al.*, 2017).

Hala ere, GD moderatzeko metodo hauek eztabaida sortu dute, askok argudiatu baitute adierazpen-askatasuna mugatu dezaketela (Schieb eta Preuss, 2016). Alternatiba gisa, kontra-narratibak (KN) proposatu dira, GDren hedapenari aurre egin eta hura arintzeko neurri eraginkor gisa (Benesch, 2014). KNak iruzkin gorrotogarri batentzako erantzun ez-oldarkorrak dira, argudioetan eta gertaeretan oinarritutako *feedback* ez-negatiboak barnean hartzen dituztenak (Schieb eta Preuss, 2016). Hemen ikus daiteke adibide bat (Chung *et al.*, 2019):

Gorroto-diskurtsoa

Musulmanek ez dute gure kultura aberastu dezakeen ezer erabilgarririk.

Kontra-narratiba

Zer egin dute guretzat musulmanek? Beno, kafea, erlojuak, unibertsitateak, tresna kirurgikoak, mapak, musika, aljebra...

Azken urteotan, KNen inguruko ikerketa nabarmen hazi da, batez ere datu-bilketa (Chung *et al.*, 2019; Fanton *et al.*, 2021), detekzioa (Chung *et al.*, 2021a) eta sorkuntza (Chung *et al.*, 2021b; Tekiroğlu *et al.*, 2022) arloetan. Hala ere, KNen sorkuntza automatikoari buruzko lanak nagusiki ingelesez egin dira, eskuz bildutako datuen eskasia eta hizkuntza-eredu sortzaileen urritasuna dela eta (Chung *et al.*, 2020).

Arazo horri aurre egiteko nahiak motibatuta, CONAN-EUS datu-multzoa aurkezten dugu, KNen sorkuntza automatikorako baliagarria den euskarazko eta gaztelaniazko lehenengo datu-multzo paraleloa. CONAN-EUS sortzeko, ingelesezko CONAN corpusaren itzulpen automatikoa eta eskuzko post-edizio profesionala egin dugu euskara eta gaztelaniara, hizkuntza bakoitzean 6654 GD-KN bikote bilduz.

Ikerketan honetan, CONAN-EUS erabili dugu itzulpen automatikoarekin eta post-editatutako datuekin sortutako kontra-narratibak alderatzeko, eta horrela post-edizioaren garrantzia aztertzeko. Gainera, xede-hizkuntzara itzulitako datuekin entrenatutako eredu sortzaileak aztertu dira (datu-transferentzia), baita hizkuntza batean entrenatutako eredu eleaniztunak beste hizkuntza batean probatu ere (eredu-transferentzia). Muinean, lan honek KNen sorkuntzarako eta hizkuntza arteko transferentzietarako bide berriak irekitzen ditu, hizkuntza gutxituen datu eskasia arintzen lagunduz. Datuak, kodea eta ereduak publikoki eskuragarri daude[1].

2 Arloko egoera eta ikerketaren helburuak

Atal honetan azken urteetako KNen datu-multzoak eta sorkuntza-metodoak aurkeztuko ditugu, batez ere ikuspegi eleaniztun batetik, baita gure ikerketaren helburu eta ekarpen nagusiak ere.

Kontra-narratiben datu-multzoak Azken urteetan hainbat GD eta KN datu-multzo argitaratu dira hala nola Qian *et al.* (2019), ingeleserako, edota Chung *et al.* (2019), ingelesa, frantsesa eta italierarako. Gure ikerketa bigarren artikulu honetan aurkeztu zuten corpusean zentratu dugu, CONAN (*COunter NArratives through Nichesourcing*) izenekoa. Chung *et al.* (2019) autoreek islamofobiaren inguruko GD-KN bikoteak bildu zituzten, hiru hizkuntzetan. Honen bitartez, GD-KN bikote corpus sendo bat eskaini zuten, ez-iragankorra, adituetan-oinarritua eta eleaniztuna. *Nichesourcing* metodoaren bitartez, arloan adituak diren langileak kontratatu ziren datu-bilketa prozesuan, KN kalitatea lehenetsiz eta datu ez-estereotipatu eta anitzak bilduz.

Zoritxarrez, *nichesourcing* metodoak denbora asko behar du eta garestia da. Ondorioz, datu-handitze prozesu bat egin zen: hiru anotatzaile ez-adituk jatorrizko KN bakoitzaren bi parafrasi sortu zituzten, baita jatorrizko frantsesezko eta italierazko GD-KN bikoteak ingelesera itzuli ere. Horrela, 6654 GD-KN bikoteko behin betiko corpusa sortu zen ingelesez. CONAN datu-multzoaren xehetasunak 1. taulan ikus daitezke.

1. Taula: Jatorrizko CONAN datu-multzoan dauden GD-KN bikote kopurua. **Parafrasiak**: jatorrizko datuen bi parafrasi. **Itzulpena**: jatorrizko italieratik eta frantsesetik ingelesera egindako eskuzko itzulpenak.

	En	Fr	It	Guztira
Jatorrizkoa	1288	1719	1071	4079
Parafrasiak	2576	3438	2142	8159
Itzulpena	2790	-	-	2790
Guztira	6654	5157	3257	15025

Kontra-narratiben sorkuntza automatikoa Ikerketa ugari argitaratu dira KN sorkuntzaren inguruan, hala nola kodetzaile-dekodetzaile ereduak landuz (Qian *et al.*, 2019) edo ebaluazio automatiko eta eskuzkoaren arteko korrelazioa falta azpimarratuz (Qian *et al.*, 2019). Beste lan batzuk sortutako testuen aniztasun falta eta GDrekiko esangura falta nabarmendu dute (Fanton *et al.*, 2021; Qian *et al.*, 2019) baita honi aurre egiteko metodologia berritzaileak aurkeztu ere (Zhu eta Bhat, 2021).

Lan horiek badute ezaugarri bat amankomunean: ingelesarekin lan egin zuten. Guk dakigunaren arabera, lan gutxi argitaratu dira KNen sorkuntzaren inguruan, ingelesa ez den hizkuntzetan. Adibidez, Möhle *et al.* (2023)

[1] https://huggingface.co/datasets/HiTZ/CONAN-EUS

autoreek alemanierazko KNen sorkuntza jorratu dute, edo Chung *et al.* (2020) autoreek hainbat esperimentu aurkeztu zituzten KNen sorkuntzarako italieraz, automatikoki itzulitako zilarrezko datuak eta eskuz sortutako datuen erabileraren ondorioak aztertuz. Gainera, Vallecillo-Rodríguez *et al.* (2023) autoreek CONAN-KN datu-multzoa (Chung *et al.*, 2021b) gaztelaniara itzuli zuten, 238 GD-KN bikotez osatua. Datu tamaina txikiak bultzatuta, ikerketa-adibide gutxiko eta *prompt* metodoetan zentratutako esperimentuak egin zituzten.

Helburuak Hauek izan dira gure ikerketaren helburu nagusiak:

1. **Euskarazko lehen KN datu-multzoa kaleratzea:** CONAN-EUS datu-multzoa aurkezten dugu, datu urritasunaren arazoa arintzeko ahaleginean. KNen sorkuntzarako baliagarria den euskarazko eta gaztelaniazko datu-multzo paraleloa da, ingelesezko CONAN corpusaren itzulpen automatikoa eginez eta eskuzko post-edizio profesionala eginez sortua.
2. **Post-edizioaren garrantzia aztertzea sorkuntza automatikoan**: Automatikoki itzulitako eta post-editatutako datuekin lortutako emaitzak alderatu ditugu, entrenatzen erabilitako datuen kalitateak emaitzetan duen eragina aztertzeko.
3. **Datu transferentzia eta eredu-transferentzia metodologiak aztertzea**: Datu-transferentzian, xede-hizkuntzara itzulitako datuekin entrenatutako ereduak erabiltzen dira (hau da, eredu elebakarrak). Eredu-transferentzian, hizkuntza batean entrenatutako eredu eleaniztunak beste hizkuntza batean probatzen dira (hots, hizkuntza anitzeko eta hizkuntza arteko ereduak).

3 Ikerketaren muina

Gure ikerketak bi atal nagusi ditu: (i) euskara eta gazteleraz KN sorkuntzarako datu-multzo bat biltzea (3.1. atalean azaldua); eta (ii) datu berri hauekin esperimentazio fasea bat garatzea (3.2. atala), KNen sorkuntza automatikoa landuz. Bertan, automatikoki itzulitako eta post-editatutako datuekin lortutako emaitzak alderatu ditugu, baita hizkuntza anitzeko eta hizkuntza arteko emaitzen ondorioak atera ere (3.3. eta 3.4. atalak).

3.1 Datuak

CONAN-EUS jatorrizko CONANen ingelesezko 6654 GD-KN bikoteetan oinarrituta dago. Horretarako, bikote hauek Google[2] APIren bidez automatikoki itzuli ziren gaztelaniara eta euskarara. Ondoren, automatikoki itzulitako gaztelaniazko eta euskarazko datu hauek 3 itzultzaile profesionalek post-editatu zituzten. Eskuzko post-edizioa egitea erabaki zen, nahiz eta garestia izan, antzeko baliabide bat zerotik eraikitzea baino azkarragoa eta errazagoa baitzen. Gainera, jatorrizko ingelesezko CONANekiko datu-multzo paralelo bat izateak ikerketa berritzaileak egiteko aukera ematen du hizkuntza anitzeko eta hizkuntza arteko transferentzia metodoak erabiliz. Hurrengo ataletan CONAN-EUSen post-edizio prozesua azalduko da.

3.1.1 Post-edizioa euskaraz

Post-edizioaren estatistikak 2. taulan ikus daitezke. Euskararen kasuan, post-edizioaren ehunekoa oso altua dela ikus dezakegu: % 51,05 GDen kasuan, % 89,21 KNen kasuan. Hori itzulpen automatikoaren kalitate kaxkarraren ondorioa izan daitekeenaren hipotesia daukagu. Gainera, KN gehiago post-editatu behar izan ziren GDak baino, ziurrenik KNak orokorrean luzeagoak eta konplexuagoak direlako eta beraz post-edizio gehiago behar izan zuten.

2. Taula: Post-edizioaren estatistikak. **Guztira**: GD eta KN instantzia totalak. Ez dator bat corpus osoaren zenbatekoarekin, datu-gehikuntzan sortu diren errepikapenak kendu baititugu; **PE**: post-editatutako instantzia kopurua; **%**: post-editatutako KNen ehunekoa

		Gaztelera		**Euskara**	
	Guztira	**PE**	**%**	**PE**	**%**
GD	523	114	21,75	267	51,05
KN	4041	630	15,59	3605	89,21

Oro har, euskarazko datu-multzoa post-editatzeko prozesuan, hiru post-edizio joera identifikatu ziren. Alde batetik, esaldi batzuetan ez zen aldaketarik behar izan, itzulpen automatikoa zuzena zelako. Bigarren kasuan, aldaketa txikiak behar izan ziren (hitz bat, hitz-amaiera edo puntuazioa). Adibidez, "islam" izena maiz "islam" gisa

[2] https://pypi.org/project/google-trans-new/

itzuli zen, baina euskaraz, testuinguru gehienetan hitz honek "-a" kasu marka behar zuela ikusi genuen: "islama". Azkenik, hirugarren kasuan, aldaketa sakonagoak egin behar izan ziren, hala nola esamolde osoak, hitzen ordena, edo esaldi osoen gehiketa/ezabaketa.

Errore ohikoenak akats gramatikalak ziren, hala nola, aditz denborak edo txarto erabilitako izenordainak. 1. adibidean ikus dezakegun bezala, itzulpen automatikoak aditza orainaldian erabili zuen iraganaren ordez. Gainera, polisemiarekin eta esamoldeen itzulpen literalekin lotutako akats lexikoak ere hauteman ziren. Adibidez, 2. adibidean, "*free from conflict*" esamoldea "doan gatazkatik" gisa itzuli zen. Hau da, itzulpen automatikoak "*free*" hitza "ordaindu gabe" esanahiarekin itzuli zuen, baina testuinguru honetan behar den esanahia "ez mugatua" da. Hau bai, "libre" gisa itzuli beharko litzateke, eta beraz post-edizioan zuzendu zen.

3.1.2 Post-edizioa gazteleraz

Gaztelaniazko post-edizioaren estatistikei begiratuz gero, 2. taulan ikus dezakegu GD eta KNen post-editatutako ehunekoak euskarazko zenbatekoak baino askoz ere baxugoak direla, hala nola % 21,75 eta % 15,59. Hori baliteke ingelesa-gazteleraren arteko itzulpenaren kalitatea altuagoa zelako gertatu izana.

Gaztelaniazko post-edizio prozesuan, euskaraz aurkitu ziren antzeko akats batzuk aurkitu ziren, baina baita gaztelaniarentzat hizkuntza-espezifikoak diren beste akats batzuk ere. Adibidez, genero konkordantzia akatsak aurkitu ziren, hots, izen batzuen eta haiei erreferentzia egiten dieten hitzen arteko genero konkordantzia falta. Horrela, 3. adibidean, lodiz idatzitako bi hitzak "o" genero marka maskulinoarekin idatzita daude, baina "mujeres"-i (emakumeak) erreferentzia egiten diotenez, forma zuzena "a"-rekin litzateke, hau da, *verlas* eta *orgullosas*.

1. adibidea

Jatorrizkoa	Itzulpen automatikoa	Post-edizioa
Then why **did** we ask them to come in the first place?	Orduan, zergatik eskatu **diegu** lehenbailehen etortzeko?	Orduan, zergatik eskatu **genien** etortzeko, lehenik eta behin?

2. adibidea

Jatorrizkoa	Automatikoki itzulia	Post-editatua
... only [some countries] are **free from conflict.**	... [herrialde batzuk] bakarrik daude **doan gatazkatik.**	... [herrialde batzuk] bakarrik daude **gatazkatik libre.**

3. adibidea

Jatorrizkoa	Automatikoki itzulia	Post-editatua
Women of our culture that decide to become Islamic are selfish. They are so happy and **proud** to join this religion (...) I would like to **see them** in Pakistan (...).	Las mujeres de nuestra cultura que deciden volverse islámicas son egoístas. Están muy felices y **orgullosos** de unirse a esta religión (...) me gustaría **verlos** en Pakistán (...).	Las mujeres de nuestra cultura que deciden volverse islámicas son egoístas. Están muy felices y **orgullosas** de unirse a esta religión (...) me gustaría **verlas** en Pakistán (...).

3.2 Esperimentuak

Gure esperimentuetan CONAN-EUS erabili nahi dugu automatikoki sortutako KNetan post-edizioak duen eragina ikusteko, baita datu-transferentzia eta eredu-transferentzia teknikak esploratzeko ere. Horretarako, hainbat datu-multzo mota erabili ditugu: ingelesezko jatorrizkoa (CONAN), baita automatikoki itzulitako eta post-editatutako euskarazko eta gaztelerazko datuak ere (CONAN-EUS). Corpus hauek hiru datu-banaketetan zatitu ditugu: 4.833 bikote entrenamendurako, 537 baliozkotzeko eta 1.278 ebaluatzeko. Banaketen arteko datu gainjartzea saihestu dugu. Horrela, ikerketan hiru esperimentazio-metodologia probatu ditu, guztiak mT5 (Xue *et al.*, 2021) hizkuntza-eredu eleanitza erabiliz:

- **Elebakarrak:** mT5 hizkuntza bakoitzerako findu dugu. Hasteko, ingelesezko datu-multzo originala erabili da (*en-jatorrizkoa*). Gainera, gaztelaniaz eta euskaraz ere findu dugu, automatikoki itzulitako (*es-MT, eu-MT*)[3] eta post-editatutako (*es-PE, eu-PE*) datuak erabiliz. Post-editatutako datuetan ebaluatu ditugu guztiak.
- **Eleaniztunak (hizkuntza anitzak)**: Hiru hizkuntzetarako datuak batu eta mT5 eredua findu dugu, irteerako datuak ingelesez (*all2en*), gazteleraz (*all2es*) eta euskaraz (*all2eu*) sortuz. Honela, datu-handitzearen onurak ebaluatu ditugu.
- **Hizkuntza artekoak**: mT5 ingelesez findu dugu eta gazteleraz (*en2es*) edota euskaraz (*en2eu*) ebaluatu dugu, hizkuntza arteko ezagutza transferentzia aztertzeko.

Ereduak HuggingFace-eko mT5-base bertsioa erabiliz entrenatu ditugu, lauki-sare bilaketa bidez optimizatuta. Azken konfigurazioan 50 *epoch*, 1e-3ko ikaskuntza-tasa, 4ko *batch* tamaina, eta 6ko *beam search* tamaina erabili dira. Esperimentuak erreproduzigarriak izateko, hazi finko bat (42) ezarri dugu.

3.3 Emaitzak: ebaluazio automatikoa

Aurreko lanetan oinarrituta (Tekiroglu *et al.*, 2020; Chung *et al.*, 2020), honako ebaluazio-metrika automatikoak erabili ditugu: BLEU, sarrera eta irteera testuaren arteko n-gramen korrelazioa neurtzen duena (Papineni *et al.*, 2002); ROUGE-L, erreferentzia eta irteera-testuaren arteko Azpisekuentzia Komun Luzeena (LCS) kalkulatzen duena, doitasuna eta estaldura kalkulatuz (Lin, 2004); eta Errepikapen Tasa (RR, ingeleseko *Repetition Rate*), sortutako testuan errepikatzen diren n-gramak kalkulatuz lortzen dena (Bertoldi *et al.*, 2013). Metrika guztietan emaitza altuenak lortu nahi ditugu, RRen izan ezik, emaitza baxuagoek errepikapen baxuagoa adierazten dute eta.

Elebakarrak 3a taulan euskara eta gaztelaniaren emaitza elebakarrak aurkezten ditugu. Emaitza nabarmenena automatikoki itzulitako datuen post-edizioak KNen sorkuntza hobetzen duela da, bai euskaraz eta bai gaztelaniaz. Gainera, es-PEek BLEU eta ROUGE-L neurketetan emaitza hobeak lortu arren, RR emaitzak hobeak dira ingelesezko datuekin (7,79 gazteleraz, 5,73 ingelesez). Honek adierazten digu, gazteleraz n-grama gainjartze metriketan lortu diren emaitza altuak errepikapenek sortu ahal izan dituztela. Euskararen emaitzetan, bestalde, eu-PEen emaitzak eu-MTrenak baino hobeak diren arren, ingelesa eta gazteleraren emaitzekin alderatuta nabarmen baxuagoak dira. Gure hipotesia da euskara ez dagoela ingelesa edo gaztelania bezain presente mT5 ereduaren hiztegian.

Laburrean, ebaluazio automatikoko emaitzek erakusten dute datu-transferentzia metodoek (hots, xede-hizkuntzara itzulitako datu-multzoekin findutako ereduak) post-edizio pauso bat behar dutela emaitza optimoak lortzeko, automatikoki itzulitako datuekin findutako ereduek ez baitute errendimendu maila bera lortzen.

3. Taula: Ebaluazio automatikoaren emaitzak

(a) Eredu elebakarren emaitzak, post-editatutako datuetan ebaluatuta. MT: automatikoki itzulita; PE: post-editatuta. Letra **lodiz** emaitza onenak, hizkuntza eta metriken arabera.

Eredua	BLEU	Rouge-L	RR
en-jatorrizkoa	**9,81**	**16,82**	**5,73**
es-MT	7,94	16,18	7,59
es-PE	**11,23**	**18,99**	**6,32**
eu-MT	4,58	9,93	9,87
eu-PE	**6,49**	**11,60**	**7,79**

(b) Eredu eleaniztunen eta hizkuntza arteko ereduan emaitzak, post-editatutako datuetan ebaluatua. Letra **lodiz** emaitza onenak, hizkuntza eta metriken arabera.

Eredua	BLEU	Rouge-L	RR
all2en	**10,79**	**17,19**	8,94
en2es	10,03	17,78	**5,15**
all2es	**11,36**	18,87	6,27
en2eu	2,81	7,26	12,40
all2eu	6,45	11,08	11,38

[3] *MT* laburpena ingeleseko *Machine Translated* hitzetik dator, automatikoki itzulitako datuak direla adieraziz

Eleaniztunak eta hizkuntza artekoak Hizkuntza anitzeko eta hizkuntza arteko esperimentuen emaitzak 3b taulan aurkezten ditugu. Eredu eleaniztunen eredu-transferentzian, ingelesa eta gazteleraren kasuan (all2en eta all2es), hiru hizkuntzetan fintzeak emaitza hobeak lortu dituela ikusi da, bereziki ingelesaren kasuan. Aldiz, hau ez da gertatzen euskararen kasuan (all2eu), eta horrek iradoki dezake hizkuntza anitzeko datu-handitzeak hobeto funtziona dezakeela egitura aldetik antzekoak diren hizkuntzetan, gure kasuan, ingelesean eta gaztelanian.

Hizkuntza arteko eredu-transferentzia inguruneari dagokionez, euskara eta gaztelerazko esperimentuek (en2es eta en2eu) oso emaitza desberdinak lortu dituzte. Gaztelaniarako hizkuntza arteko transferentziak (en2es) nahiko ondo funtzionatu du, datu-transferentzia ereduaren emaitzak (es-MT) gaindituz. Euskaraz, ordea, hizkuntza arteko eredu-transferentziak (en2eu) guztiz huts egiten du, eu-MTren nabarmen azpitik geratuz.

Laburbilduz, euskaraz datu-transferentzia (eu-MT) metodoak eredu-transferentzia (en2eu) baino emaitza hobeak ematen ditu, baina gaztelaniaz kontrakoa gertatzen da. Aldiz, eredu eleaniztunen datu-handitzeak ingeles eta gaztelaniaren eredu elebakarren emaitzak hobetzen ditu, baina euskararenak okertu. Badirudi euskararako estrategiarik onena entrenamendu datuen post-edizio osoa egitea dela; gaztelaniarako, emaitza lehiakorrak lor daitezke datu-transferentzia (es-MT) edo hizkuntza arteko transferentzia (en2es) metodoak erabiliz, post-edizio osoa egin gabe. Orokorrean, ebaluazio automatiko honetatik ondoriozta dezakegu mT5 bezalako eredu sortzaileek errendimendu eskasagoa dutela baliabide gutxiko eta hizkuntza-egitura desberdina duten euskara bezalako hizkuntzetan.

3.4 Emaitzak: eskuzko ebaluazioa

KN sorkuntzari buruzko hainbat ikerketek adierazi duten bezala, aurreko atalean erabilitako ebaluazio metrika automatikoek sarritan ez dute korrelazio onik izaten giza iritziarekin (Qian *et al.*, 2019). Hori dela eta, eskuzko ebaluazio kualitatiboa egin dugu, euskara eta gaztelania ama hizkuntza gisa edo maila altuan hitz egiten zuten bi hizkuntzalari aditu kontratatuz. Eskuzko ebaluazioan, bost irizpide hartu ditugu kontuan:

1. **Erlazioa**: KNak GDrekiko duen erlazioa neurtzen da, hau da, GDari erantzun egokia ematen dion KNak.
2. **Espezifikotasuna**: KNa orokorra edo espezifikoa den adierazten du emandako GDrako, hots, “beste GD guztiz desberdin baterako erabil daiteke ala ez?”galderari erantzuten dio.
3. **Aberastasuna**: Hizkuntza eta hiztegi aldetik, KN sinpleak edo konplexuak diren neurtzen du.
4. **Koherentzia**: Esaldiek elkarrekin zentzua duten neurtzen du, ea ideiak argiak diren eta erraz uler daitezkeen.
5. **Gramatikaltasuna**: KNen zuzentasun gramatikala neurtzen du.

Ebaluazio honetarako, sei eredu aukeratu ziren: es-MT, es-PE, en2es, eu-MT, eu-PE eta en2eu. Datuei dagokienez, eredu bakoitzak sortutako 20 GD-KN bikote hautatu ziren ausaz. Bi ebaluatzaileek sei ereduen emaitzak ebaluatu zituzten itsuan (zein ereduren emaitzak ziren jakin gabe), bost irizpideen arabera, 1etik 5erako eskalan. Irizpideen hautaketa Chung *et al.* (2019) eta Chung *et al.* (2020) autoreen eskuzko ebaluazioetan oinarritu zen.

Etiketatze prozesuaren kalitatea ziurtatzeko, ebaluatzaileen arteko adostasun-maila (IAA, ingelesezko *Inter Annotator Agreement*) kalkulatu genuen. Ikusi genuen, eredu guztientzat, IAA nahiko altua zela, bataz besteko 0.84 adostasun-maila lortuz. Horrek ebaluazio kualitatibo honen fidagarritasuna bermatzen du.

Ondorioz, adibide guztiak etiketatu ondoren, irizpide bakoitzaren batezbestekoa eta eredu bakoitzaren puntuazio orokorra kalkulatu ziren, baita desbideratze estandarra ere. Ebaluazio kualitatiboaren emaitzak 4. taulan aurkezten ditugu. Eredu elebakarretan ikus daiteke post-editatutako emaitzak direla onenak oro har bi hizkuntzentzat. Horrela, automatikoki itzulitako eta post-editatutako emaitzak alderatuz, ikus daiteke desberdintasun handienak Erlazioan eta Espezifikotasunean gertatzen direla. Hau da, automatikoki itzulitako datuekin entrenatutako ereduek KN errepikakorragoak eta orokorragoak sortzen dituzte post-editatutako bertsioekin findutakoek baino.

Hizkuntza arteko ereduen emaitzak, aldiz, desberdinak dira euskara eta gaztelaniarentzat. Gaztelaniarako, hizkuntza arteko transferentziako ereduak (en2es) puntuazio altua lortu du Aberastasuna, Koherentzia eta Gramatikaltasunean, eta es-MT bertsioa bost irizpideetan gainditu du. Euskararako, berriz, eredu-transferentziaren (en2eu) emaitzak baxuak dira Erlazio (1,85) eta Espezifikotasunean (1,53), baina hala ere eu-MT baino puntuazio orokor altuago lortu du. Euskaraz, hizkuntza arteko ereduaren emaitzak eredu elebakarrenak baino altuagoa izatea ez dator bat ebaluazio automatikoarekin.

Laburbilduz, gaztelaniarako emaitza kuantitatibo eta kualitatiboak korrelazionatuta daude, eredu eta metodologia guztietan. Bestalde, hau ez da gertatzen euskararen kasuan, hizkuntza arteko transferentzia eredua (en2eu) hobea baita eskuzko ebaluazioaren arabera, baina nabarmen okerragoa metrika automatikoetan.

4. Taula: Eskuzko ebaluazioaren emaitzak (anotatzaileen arteko bataz-bestekoa). Azpimarratuta: emaitza onenak kategoria eta hizkuntzako; **Lodiz:** emaitza onena orokorrean, hizkuntzako.

	Erlazioa	**Espezifikotasuna**	**Aberastasuna**	**Koherentzia**	**Gramatika**	**Orokorra**
es-MT	2,30	2,50	3,33	3,58	3,78	3,10 ± 0,66
es-PE	3,61	3,31	3,72	4,25	4,33	**3,84 ± 0,43**
en2es	3,20	2,90	3,83	4,00	4,08	3,60 ± 0,52
eu-MT	2,33	2,20	2,75	3,25	3,75	3,01 ± 0,65
eu-PE	3,13	3,08	3,15	3,43	4,03	**3,85 ± 0,39**
en2eu	1,85	1,53	3,70	3,98	4,43	3,10 ± 1,31

4 Ondorioak

Artikulu honetan CONAN-EUS aurkeztu dugu, KN sorkuntzarako euskara eta gaztelaniazko datu-multzo paralelo berri bat, jatorrizko ingelesezko 6654 GD-KN bikoteak automatikoki itzuliz eta profesionalen bitartez post-editatuz sortua. Baliabide honek hizkuntza anitzeko eta hizkuntza arteko ikuspegitik ikerketa berritzaileak egitea ahalbidetu du, KNen sorkuntzaren arloan.

Esperimentuetan ikusi dugu KNen sorkuntza hobea dela mT5 post-editatutako datuekin fintzean, automatikoki itzulitako datuekin baino. Esperimentu eleaniztunek erakutsi dute hizkuntza arteko transferentzia ona dela ingelesez eta gazteleraz, baina ez euskaraz. Hori hainbat arrazoirengatik izan daiteke: (i) ingelesa eta gaztelania egitura aldetik antzekoagoak direlako euskara baino, eta (ii) mT5en hiztegian gaztelera eta ingelesaren presentzia handiagoa dagoelako euskararekin alderatuta, hau baliabide gutxiagoko hizkuntza izanik. Hizkuntza arteko esperimentuetan antzeko joera antzeman da. Hala ere, hizkuntza arteko transferentzia metodoa ikerketa-arazo ireki eta konplexua izaten jarraitzen du eredu sortzaileetan (Lin *et al.*, 2022).

5 Etorkizuneko lanak

Gure lanak etorkizuneko ikerketetarako hainbat aukera azpimarratzen ditu, batez ere euskara eta gaztelaniako GD arintzeko eta hizkuntza arteko transferentzia teknikak hobetzeko asmoarekin. Esperimentuetan ikusi dugu post-editatutako datuek errendimendua hobetzen dutela, baina ez dago argi datu guzti-guztiak eskuz berrikustea beharrezkoa den. Etorkizunean, automatikoki itzulitako eta post-editatutako datu-konbinazio ezberdinak aztertu ahal dira, honen eragina behatzeko. Horrez gain, lan honetan bi hizkuntza soilik landu dira, baina etorkizunean beste hizkuntzetarako aplikagarritasuna eta baliabide gutxiko testuinguruetarako estrategiak ikertzea beharrezkoa ikusten dugu. Azkenik, KN sorkuntzako ebaluazio automatikoa oraindik ikerkuntzarako galdera irekia da. Adibidez, Zubiaga *et al.* (2024) autoreek Hizkuntza Eredu Handiak (HEH) erabili dituzte KNen ebaluazio automatikorako.

Erreferentziak

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, eta Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Benesch, Susan. 2014. Countering dangerous speech: New ideas for genocide prevention. *In SSRN 3686876* .

Bertoldi, Nicola, Mauro Cettolo, eta Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, 35–42.

Chung, Yi-Ling, Marco Guerini, eta Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, 125–132. Association for Computational Linguistics.

——, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, eta Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829, Florence, Italy. Association for Computational Linguistics.

——, Serra Sinem Tekiroglu, eta Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *CLiC-it*.

——, Serra Sinem Tekiroğlu, eta Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914, Online. Association for Computational Linguistics.

Davidson, Thomas, Dana Warmsley, Michael Macy, eta Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the international AAAI conference on web and social media* 11.512–515.

Fanton, Margherita, Helena Bonaldi, Serra Sinem Tekiroğlu, eta Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3226–3240, Online. Association for Computational Linguistics.

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, eta Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Möhle, Pauline, Matthias Orlikowski, eta Philipp Cimiano. 2023. Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, ed. by Yi-Ling Chung, Helena Bonaldi, Gavin Abercrombie, eta Marco Guerini, 44–61, Prague, Czechia. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, eta Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding, eta William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Schieb, Carla, eta Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Tekiroğlu, Serra Sinem, Helena Bonaldi, Margherita Fanton, eta Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3099–3114. Association for Computational Linguistics.

Tekiroglu, Serra Sinem, Yi-Ling Chung, eta Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *ACL*.

Vallecillo-Rodríguez, Maria Estrella, Arturo Montejo-Raéz, eta Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural* 71.227–245.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, eta Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498, Online. Association for Computational Linguistics.

Zhu, Wanzheng, eta Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 134–149, Online. Association for Computational Linguistics.

Zubiaga, Irune, Aitor Soroa, eta Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, ed. by Yaser Al-Onaizan, Mohit Bansal, eta Yun-Nung Chen, 9572–9585, Miami, Florida, USA.

6 Eskerrak eta oharrak

Jaione Bengoetxeak Eusko Jaurlaritzako doktoretza-aurreko bekaren finantzaketa du (PRE_2024_1_0028). Gainera, eskerrak eman nahi dizkiegu MCIN/AEI/10.13039/501100011033-ko proiektu hauei: (i) DeepKnowledge (PID2021-127777OB-C21) eta FEDER, EU-ko; (ii) Disargue (TED2021-130810B-C21) eta Europar Batasunaren NextGenerationEU/PRTR-eko; (iii) DeepMinor (CNS2023-144375) eta Europar Batasunaren NextGenerationEU/PRTR-eko.