Euskal Herriko Unibertsitatea

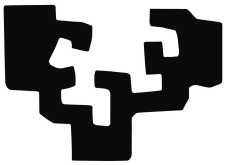
Baliabide Urriko Hizkuntzetarako Hizkuntza-Eredu Neuronalak

Gorka Urbizu Garmendia

Doktorego Tesia

2025

Informatika Fakultatea
eman ta zabal zazu



Euskal Herriko Unibertsitatea

Konputazio Zientziak eta Adimen Artifiziala Saila

Baliabide Urriko Hizkuntzetarako Hizkuntza-Eredu Neuronalak

Gorka Urbizu Garmendiak Xabier Saralegi Urizar eta Aitor Soroa Etxaberen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostian, 2025eko ekainean.

Doktorego tesi hau doktoregaiak Elhuyar eta Orai NLP Teknologiaken lan egin bitartean gauzatu da, nagusiki DeepText (KK-2020/00088) eta ICL4LANG (KK-2023/00094) Elkartek projektuen barnean. Bestalde, Googleren TRC programaren bidez kalkulu-ahalmen handiko TPU makinetarako sarbidea ere izan du.

Eskerrak

Hasteko ta behin, nire zuzendariak eskertu nahi nituzke, Xabi ta Aitor, eguneroko lanean iparra galtzen ez uztearren. Mila esker baita ere, Olatz eta Ander, nire hastapenetan NLPn ikertzeak zertan datzan erakutsi eta bide honi ekiteko grina transmititzearren.

Elhuyarren lantalde osoari, ta bereziki Oraiko lankideei, lan-giro atsegin eta eroso bat eskaintzearren, zuekin asko ikasi dudalako, platanoak nola zuritu, adibidez. Ta milesker Eneko, tesi-txostenean terminologia eta euskara orrazten laguntzeagatik.

Beste horrenbeste, Ixakideei, bertatik pasa nintzenean bat gehiago nintzela sentitzearren, ta ordutik gurutzatzen garen kongresuero adoptatzearren.

Milesker etxekoi re, ne uneik erretxiñenetan e aguantatzearren, ta behar non guztin launtzearren. Ta launei desertu erdin azaltzean limoizko helatue izatearren. Og tusen takk til dere også.

Esker mila Idoia, Naroa, Aitor, Mireia, Oscar, Xuban ta Maddalen beldurrek uxatzen laundu ta biziraun littekeela frogatzearren.

Ta bukatzeko, eskerrik asko, gràcies, gracias, grazas, merci, grazie, takk, hvala and thanks to everyone I've met at conferences and had the chance to share ideas* and beers* with.

*equal contribution

"Understand well as I may,
my comprehension can only be
an infinitesimal fraction of all I want to understand."
Ada Lovelace

"Gauzak ez dira horrela,
gauzak horrelaxe daude"
Gorka Urbizu

Laburpena

Doktorego tesi honek baliabide urriko hizkuntzetarako hizkuntza-eredu neuronalak ditu aztergai, bereziki, euskaran arreta jarriz. Hiru ikerketa-galdera nagusi jorratu dira: euskararako hizkuntza-ereduak nola ebaluatu modu estrintsekoan, datu-eskasiak ereduaren errendimenduan duen eragina, eta euskararen ezaugarri linguistiko bereizgarriek aurrentrenamenduan duten eragina. Galdera horiei erantzuteko, ebaluazio-proba berriak garatu dira, besteak beste, euskararako hizkuntza ulermen orokorrerako BasqueGLUE probabankua eta gaitasun gramatikala ebaluatzeko BL2MP datu-multzoa. Bestalde, baliabide urriko hizkuntzetan BERT ereduek zer-nolako eskala-legeak jarraitzen dituzten eta euskara bezalako morfologia aberats eta hitz-ordena malguko hizkuntzetan gramatika ikasteko duten gaitasuna ikertu dira. Halaber, aurrentrenamendurako automatikoki itzulitako datu sintetikoak erabiltzearen bideragarritasuna aztertu da. Lan honek utzitako ondorioek eta baliabideek euskararako eta baliabide urriko beste hizkuntzetarako hizkuntza-ereduen ikerketarako oinarri sendoa ezartzen dute.

Abstract

This PhD thesis focuses on neural language models for low-resource languages, with a particular emphasis on Basque. It addresses three main research questions: how to evaluate language models for Basque extrinsically, how data scarcity affects model performance, and how Basque's distinctive linguistic features influence pretraining. To answer these questions, new evaluation benchmarks have been developed, including BasqueGLUE for general language understanding in Basque and BL2MP, a dataset for evaluating Basque grammatical competence. Additionally, the study examines the scaling laws observed in BERT models in low-resource settings and their capacity to learn grammar in morphologically rich and syntactically flexible languages, such as Basque. The feasibility of using automatically translated synthetic data for pretraining is also explored. The findings and resources from this work provide a strong foundation for future research on language models for Basque and other low-resource languages.

Gaien Aurkibidea

SARRERA

1. KAPITULUA

Sarrera

Adimen artifiziala azkenaldian gizartean gori-gori[1] (Addley, 2023) dagoen gaia dugu, eta hizkuntzaren prozesamendua honen adar nagusietako bat da. Hizkuntzaren prozesamenduak (*Natural Language Processing* edo NLP) giza hizkuntzaren ulermena eta sorkuntza ditu helburu. Hizkuntzaren konplexutasun semantiko eta sintaktikoak, anbiguotasunak eta testuinguruaren interpretazioa automatikoki prozesatzeak erronka teknologiko eta zientifiko sakonak mahaigaineratzen ditu.

Azken urteotan, ikasketa sakoneko teknologietan oinarritutako hizkuntza-eredu neuronalek sekulako iraultza ekarri dute giza hizkuntza ulertzeko eta sortzeko gaitasunetan, teknologia hauek geroz eta esparru eta erabiltzaile gehiagorengana hurbilduz. Iraultza teknologiko hori gehienbat ingelesa eta beste hizkuntza gutxi batzuen bueltan eraikia izan da, eta munduko gainontzeko hizkuntzetako hiztunok tren hori ez galtzeko ahaleginetan gabiltza. Baliabide urriko hizkuntzen kasuan, hizkuntza-eredu horiek entrenatzeko funtsezko diren corpus erraldoiak eta kalkulu-ahalmen handia erronka gehigarriak dira bere horretan.

Doktorego tesi txosten honetan aurkeztutako lana hizkuntza-ereduak eta baliabide urriko hizkuntzak gurutzatzen diren ikerketa esparru horretan kokatzen da.

[1] `https://www.theguardian.com/technology/2023/nov/01/ai-named-most-notable-word-of-2023-by-collins-dictionary`

1.1 Motibazioa

Hizkuntzaren prozesamenduaren mamia ordenagailuek giza hizkuntza ulertzean dago. Gizakiok darabilzkigun hizkuntzen eta ordenagailuen arteko zubilana egin nahian, hizkuntzaren prozesamenduaren arloko ikerlari eta profesionalok giza hizkuntza idatzizko edo ahozko sarrera gisa maneiatzeko gai diren teknologiak ikertzen eta garatzen dihardugu.

Gaur egungo adimen artifizialeko laguntzaileek hizkuntza gaitasun sendoak eta erabiltzailearen eskaerari kasuen gehiengoan egoki erantzungo dion testua sortzeko gaitasuna dituzte (Dubey et al., 2024; DeepSeek-AI et al., 2024), mugak muga (Asher et al., 2023; Lu et al., 2023; Hossain et al., 2023; Shen et al., 2024; Wu et al., 2024). Adimen artifizialaren hastapenetan makinen adimena neurtzeko sortutako Turingen test klasikoa (Turing, 1950) ere gainditzear direla dirudi (Jones and Bergen, 2024). Hala ere, adimen artifizialeko laguntzaileen atzean dauden hizkuntza-eredu handiek gure hizkuntza benetan ulertzen duten eztabaida irekia da oraindik (Bender and Koller, 2020; Bender et al., 2021; Li et al., 2022; Bubeck et al., 2023; Fierro et al., 2024; Bender et al., 2025).

Orain arteko euskarazko paragrafoetako bat adimen artifizialak idatzia dela baieztatuko bagenu, nahiz eta orain urtebete ezinezkotzat joko genukeen, gaur egun zuhurrago jokatzea tokatzen zaigu, dagoeneko euskara itxuran ulertu eta idazteko teknologia errealitate bat baita: Adimen artifizialeko laguntzaile komertzialek ez ezik, hizkuntza-eredu irekiek (DeepSeek-AI et al., 2024; Etxaniz et al., 2024; Corral et al., 2024) ere euskaraz aritzeko gai direla erakutsi dute. Tesi honi ekin zitzaionean, ordea, teknologia hauen egoera oso bestelakoa zen; urte gutxitan arloak izan duen bilakaera azkarraren erakusle.

Aurreko hamarkadan zehar izandako kalkulu-ahalmen eta datu kopuru eskuragarrien gorakada handiak hizkuntzaren prozesamenduan aurrerapauso esanguratsuak ematea ahalbidetu du, ikasketa automatikoko teknikekin lehenik eta berrikiago ikasketa sakonarekin, neurona-sareen bitartez, alegia (Deng and Liu, 2018).

Neurona-sareetan oinarritutako ikasketa sakoneko teknikek hizkuntzaren prozesamenduko ataza askotan aurrerapen handiak ekarri dituzte. Besteak beste, hitzen esanahi semantikoa kodetzeko *Word2Vec* hitz-bektoreak (Mikolov et al., 2013) eta itzulpen automatikoa bezalako sekuentzia mailako atazetan hobekuntza nabarmenak ekarri dituzten neurona-sare errepikako-

rrak (RNN) (Bahdanau et al., 2015) edo Transformer arkitektura (Vaswani et al., 2017) aipa genitzake.

Azken horrek, Transformer arkitekturak (Vaswani et al., 2017), sekuentziatik sekuentziarako ataza gainbegiratuetan ez ezik, testu hutsaren gainean hizkuntza-eredu sendoak entrenatzeko bidea ere ireki du. Nagusiki, bi estrategia jorratu dira testu hutsetik, eskuzko anotaziorik gabe, ikasketa autogainbegiratu edo erdigainbegiratuaren bitartez, hizkuntza-ereduak sortzeko: alde batetik, helburu-funtzio gisa maskaratzea darabilten BERT (Devlin et al., 2019) bezalako kodetzaile motako hizkuntza-eredu diskriminatzaileak, eta bestetik GPT (Radford, 2018) gisako deskodetzaile motako hizkuntza-eredu autorregresiboak, helburu-funtzio gisa hurrengo tokena aurreikustea dutenak.

Horrelako hizkuntza-ereduak sentimenduen analisia, gaien sailkapena edo galdera-erantzuna bezalako atazak ebazteko funtsezko bilakatu dira, aukera ematen baitute modu erdigainbegiratuan aurrentrenatutako neurona-sarea datu anotatu gutxiagorekin birdoitzeko, alegia, ikasketa gainbegiratua egiteko. Hurbilpen berri honek atazen datu eskakizunak arintzen dituen arren, aurrentrenamendurako testu kopuru izugarriak eta kalkulu-ahalmen handiagoak eskatzen ditu.

Hiztun kopuruari eta baliabideei dagokienez hizkuntza handiagoak direnetan hizkuntzaren ulermenean emandako aurrerapausoei euskaratik ere ekitea erabaki zen doktorego tesi honekin, hizkuntza-ereduak funtsezko bilakatu baitira hizkuntza teknologietan.

Bide horri ez zaio hutsetik ekin ordea. Doktorego tesiaren abiapuntua "*Give your Text Representation Models some Love: the Case for Basque.*" (Agerri et al., 2020) lanean aurkeztutako BERTeus eredua da, Transformerretan oinarritutako euskararako lehen hizkuntza-eredua.

Behin euskararako hizkuntza-eredu bat eskura izanik, hainbat zalantza sortzen zaizkio bati. Eredua erabilgarria izan dakiguke, baina eredua ona ote da? Zer esan nahi du eredu ona izateak? Nola neurtu genezake hori euskararako? Beste eredu berri bat sortzean nola konparatu genezake aurrekoekin? Eredu hobea sortzeko, zertan jarri behar dugu arreta? Euskararako dugun datu kopurua nahikoa da? Edo gutxitxo? Zer eragin du horrek? Hizkuntza-eredu arkitekturen gehiengoa ingelesa lantzeko diseinatuak izan dira, euskararen kasuan bere bereizgarri linguistikoak hobeto tratatzea posible da? Eta merezi du esfortzu horrek? Galdera horiek erantzutea izan da doktorego tesi honen erronka.

Hizkuntza-eredu aurrentrenatuek ikasketa erdigainbegiratu sendo batekin, eta ikasketa gainbegiratu pisutsuen beharrik gabe, askotariko atazak jorratzeko bidea irekitzen dute. Baliabide urriko hizkuntzei dagokienez, teknologia honek aukera itxaropentsuak dakartza, aurrentrenatutako hizkuntza-eredu sendoekin hizkuntzaren prozesamenduko ia ataza oro ebazteko eskuzko anotazio eskakizunak egingarriagoak baitira. Horretarako, baina, testu eta kalkulu-ahalmen kopuru handiak behar dira, eta horiek ez daude hizkuntza komunitate guztien eskura, besteak beste, hiztun kopuruak, egoera soziolinguistikoak edo baliabide ekonomikoek eraginda.

Hizkuntza-ereduak entrenatzeko datu gutxi izateak zer eragin duen ez dago behar adina aztertua, eta hizkuntza-eredu sendoak aurrentrenatzeko gutxieneko datu eskakizunak ere zehazteke daude. Datu eta eredu tamainaren arteko erlazioa landu den kasuetan, bien arteko eskala-legeak hizkuntza-eredu handiak optimizatzeko helburuarekin sortu dira. Horrenbestez, eskala-legeak eskala txikian aztertzeko beharra identifikatu da, baliabide urriko hizkuntzen beharrei erantzuteko.

Datu eskasiaren hariari tiraka, hizkuntzaren prozesamenduan eta adimen artifizialean ohikoa da landu nahi den ataza zehatzerako entrenatzeko adina datu ez dugun testuinguruetan, datu gehikuntzako estrategiak eta datu sintetikoen sorkuntza aztertzea. Hizkuntza-ereduen aurrentrenamenduan, aldiz, hizkuntza nagusiek nahi adina ez bada, behar adina testu behintzat izan ohi dutenez, datu sintetikoen bidea askorik jorratu gabekoa da oraindik, ereduen destilaziorako ez bada. Baliabide urriko hizkuntzen kasuan, ordea, eskuragarri dagoen testuari ahal bezainbeste zuku ateratzeko ahalegina egitea tokatzen zaigu, alternatiba guztiak probestuz.

Euskarako hizkuntza-ereduak eraikitzerakoan, aurrentrenamendurako testu-bilduma mugatuak izateaz gain, nagusiki ingeleserako diseinatutako eta hizkuntza gutxi batzuetan balioztatutako tokenizatzaileak eta sare arkitekturak euskararen ezaugarri linguistikoetarako aproposak diren ere ebatzi beharko litzateke; euskararen morfologia aberatsak eta hitzen ordena malguak hizkuntza-eredugintzan duten eragina aztertzeke dago eta.

Ikerlerro guzti hauek egikaritzeko eta hizkuntza-ereduen ikerketan eta garapenean aurrerapausoak emateko ezinbestekoa da hizkuntza bakoitzeko ebaluazio-marko propioak garatzea. Ikerlariek aurkeztutako arkitektura neuronal eta hizkuntza-eredu jakinek aurrekoekiko hobekuntza dakartela baieztatzeko ebaluazio-marko beraren gainean ebaluatu eta konparatu behar dira, eta horretarako, lehenik, ebaluazio-marko bateratuak diseinatu eta eraiki.

Euskararen kasuan, badira ereduak ataza jakinetan ebaluatzeko datu-multzo irekiak, baina, hizkuntza-ereduak hizkuntza ulermenean ebaluatzeko proba-banku sendoen gabezia identifikatu da.

1.2 Helburuak eta Ikerketa-Galderak

Doktorego tesi lan honen helburu nagusia gaur egun hizkuntzaren prozesamenduko ataza gehienak ebazteko abiapuntu gisa erabiltzen diren hizkuntza-eredu neuronalak baliabide urriko testuinguruan ikertzea da, bereziki euskaran zentratuz.

Hizkuntzaren prozesamenduaren alorrean bi esparru bereizi genitzake: hizkuntzaren ulermena (*Natural Language Understanding* edo NLU) eta hizkuntzaren sorkuntza (*Natural Language Generation* edo NLG). Lan honi ekitean, NLUrako kodetzaile motako hizkuntza-eredu diskrimanatzaileak NLGrako deskodetzaile motako ereduak baino garatuago zeudenez, doktorego tesi honetan hizkuntzaren ulermenean jarri da fokua abiapuntu gisa eta kodetzaile motako hizkuntza-ereduetara mugatu da ikerketaren irismena.

Tesi honetan, euskaraz hizkuntzaren ulermena lantzeko eredu edo arkitektura onena topatu edo ahalik eta hizkuntza-eredu onena aurre-entrenatzea baino, euskararako hizkuntza-eredugintzan oinarri sendoak finkatzean jarri da lehentasuna, egungo zein etorkizuneko hizkuntza-ereduen garapena bideratu eta azkartuko duten esperantzan.

Horretarako ondorengo ikerketa-galderak (IG) definitu dira eta hauek erantzutea helburu duten ikerketa lerroak landu dira tesian:

IG1: Nola ebaluatu genezake euskararako hizkuntza-ereduen kalitatea modu estrintsekoan?
Hizkuntza-ereduak testuko hutsuneak betetzen edo hurrengo hitza aurreikusten entrenatu ohi dira, eta entrenamenduan zehar ataza berean ebaluatzen ditugu, modu intrintsekoan, eredua ikasten ari den ikusteko. Hala ere, ataza horietan lortutako emaitzek ez dute egoki adierazten hizkuntza-eredu horrek hizkuntza ulertzeko duen gaitasuna. Hizkuntza ulermeneko atazez osatutako ebaluazio-markoekin ebaluatu ohi dira horretarako; baina euskararen kasuan, tesi honen aurretik ez zegoen hizkuntza ulermenerako proba-banku landu finkorik.

IG2: Zein eragin du baliabide gutxi izateak hizkuntza-eredua entrenatzeko orduan?
Datu, parametro eta kalkulu-ahalmenen arteko oreka landu izan da, baina eskala-lege horiek aztertzean milaka miloi parametroko ereduak milaka milioi tokenekin entrenatzen dituzten inguruneetan arreta jarriaz aztertu dira nagusiki. Badira, bai datuei, zein kalkulu-ahalmenari dagokionez, baliabide murritzekin sortu diren hizkuntza-ereduak, baina oraindik aztertzeke daude baliabide gutxi izateak zer eragin duen entrenamenduan eta baliabide urriko erregimen horretan datu eta parametro kopuruen arteko erlazioak eta oreka zein diren.

IG3: Zein eragin dute euskararen ezaugarri linguistikoek hizkuntza-eredua entrenatzerakoan?
Gaur egun hizkuntza-eredugintzan funtsezko diren tokenizatzaile, atentzio mekanismo, Transformer arkitektura eta abarrak, nagusiki ingelesaren ezaugarriei erantzunez sortuak izan dira. Euskarak baditu morfologia aberatsa eta hitzen ordena malgua bezalako bereizgarriak ingelesarekiko. Ontzat jotzen diren diseinu erabaki horiek ba al dute eraginik hizkuntza-ereduek euskara barneratzeko orduan?

Ikerketa-galdera hauetako bakoitza, eta bidean sortutako beste hainbat erantzuteko burututako esperimentuak biltzen ditu tesi-txosten honek. IG1 3. eta 5. Kapituluetan landu da, IG2 4., 5. eta 6. Kapituluetan, eta IG3 5. Kapituluan.

1.3 Ekarpen Nagusiak

Doktorego tesi honek askotariko ekarpenak utzi ditu lauzpabost urteko ikerketaren emaitza gisa: nazioarteko kongresuetan argitaratutako eta aurkeztutako artikuluak, corpusak, ebaluaziorako datu-multzoak eta aurrentrenatutako hizkuntza-ereduak.

1.3.1 Doktorego Tesiaren Parte diren Argitalpenak

Doktorego tesi honen ekarpen zientifiko nagusiak ondorengo lau artikulu hauek izan dira, zeinak LREC, ACL (2x Findings) eta LREC-COLING nazioarteko kongresuetan argitaratu eta aurkeztu diren. Tesi-txosteneko 3-6. kapituluetako bakoitza artikuluetako bati dagokio.

- Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., and Soroa, A. (2022). **BasqueGLUE: A Natural Language Understanding Benchmark for Basque.** In Proceedings of the Language Resources and Evaluation Conference: LREC22 (pp. 1603–1612). Marseille, France.

- Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., and Soroa, A. (2023a). **Scaling Laws for BERT in Low-Resource Settings.** In Findings of the Association for Computational Linguistics: ACL23 (pp 7771–7789). Toronto, Canada.

- Urbizu, G., San Vicente, I., Saralegi, X., and Corral, A. (2023b). **Not Enough Data to Pre-train Your Language Model? MT to the Rescue!** In Findings of the Association for Computational Linguistics: ACL23 (pp. 3826-3836). Toronto, Canada.

- Urbizu, G., Zulaika, M., Saralegi, X., and Corral, A. (2024). **How Well Can BERT Learn the Grammar of an Agglutinative and Flexible-Order Language? The Case of Basque.** In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation: LREC-COLING24 (pp. 8334-8348). Torino, Italy.

1.3.2 Baliabideak

Argitalpenez gain, euskaraz hizkuntza-ereduak garatzen lagunduko duten bestelako ekarpenak ere utzi ditu tesiak. Horietako asko baliabideak dira, hizkuntza-ereduak entrenatzeko corpusak, hizkuntza-ereduak ebaluatzeko datu-multzoak eta aurrentrenatutako hizkuntza-ereduak.

1.3.2.1 Corpusak

Euskararako hizkuntza-ereduak aurrentrenatzeko xedearekin honako bi corpus elebakar bildu dira:

- **DeepText** euskarazko corpus elebakarra: 351M hitz ditu eta ElhBERTeu eredua entrenatzeko erabili da. 2020an bildua. Xehetasun gehiago Apendizeko A.3. Atalean.

- **ZelaiHandi**[2] lizentzia libreko euskarazko corpus elebakarra: 660M hitz ditu eta gaur gaurkoz ezaugarri horiek dituen corpusik handiena da (San Vicente et al., 2024). 2025eko otsailean eguneratu zen azkenekoz. 5K deskargatik gora izan ditu dagoeneko. Xehetasun gehiago Apendizeko A.4. Atalean.

1.3.2.2 Datu-Multzoak

Euskararako hizkuntza-ereduak euskararen ulermenean eta euskarazko gramatika gaitasunean ebaluatzeko datu-multzoak ere sortu ditugu:

- **BasqueGLUE** proba-bankua[3]: euskararako hizkuntza-ereduak hizkuntza ulermenean ebaluatzeko bederatzi ataza biltzen dituen proba-bankua da. 3.2. Atalean proba-bankuaren diseinua eta berau osatzen duten atazak aurkezten dira, besteak beste, WiC$_{eu}$, EpecKorrefBin eta QNLI$_{eu}$ datu-multzo berriak. Egun, 21K deskarga izan ditu.

- **BL2MP** datu-multzoa[4][5]: euskararen ezagutza gramatikala ebaluatzeko datu-multzo bat da, ingeleserako BLiMP datu-multzoan inspiratua. 5.2. Atalak datu-multzoaren xehetasun gehiago dakartza. 500 deskarga izan ditu dagoeneko.

1.3.2.3 Hizkuntza-Ereduak

Lan honetan dozenaka hizkuntza-eredu aurrentrenatu dira, horietako gehienak euskararako. Horien artean, euskararako hizkuntza-eredu erabilgarri gisa ekarpen nabarmenena ElhBERTeu da.

- **ElhBERTeu**[6]: DeepText corpusarekin aurrentrenatutako euskararako BERT eredua. Publikoki eskuragarri dago CC-BY-4.0 lizentziapean, eta 14K deskarga izan ditu dagoeneko.

[2] `https://huggingface.co/datasets/orai-nlp/ZelaiHandi`
[3] `https://huggingface.co/datasets/orai-nlp/basqueGLUE`
[4] `https://huggingface.co/datasets/orai-nlp/bl2mp`
[5] Baliabide honen sorkuntzan ekarpen-egile nagusia Muitze Zulaika Gallastegi izan da.
[6] `https://huggingface.co/orai-nlp/ElhBERTeu`

1.3.3 Tesiaren Zeharkako Ekarpenak

Euskaraz gain, tesian zehar beste hizkuntza batzuk ere landu ditugu. Horren ondorioz, BasqueGLUEn euskararako egin moduan, swahilirako, suomierarako eta gaztelaniarako ere QNLI datu-multzoak sortu ditugu. Xehetasun gehiago 4.3.5. Atalean.

Hizkuntza horietarako, hizkuntza-ereduak ere aurrentrenatu ditugu, eta horien artean, swahilirako entrenatutako BERT ereduak lehiakorrak direla ikusi dugunez, eskuragarri jarri ditugu *base*[7] eta *medium*[8] tamainako swahilirako gure BERT onenak, eta orotara 700 deskargatik gora izan dituzte.

Euskararako ElhBERTeuren aldaera arinagoa (*medium* tamainakoa) ere eskuragarri daukagu[9].

Azkenik, tesi-txosten hau bere osotasunean euskaraz idaztearen eraginez, berariaz landu den euskarazko terminologia teknikoa ere ekarpen interesgarritzat jotzen dugu. Glosategia txostenaren amaieran dago eskuragarri.

1.3.4 Tesitik Kanpoko Ekarpenak

Doktorego tesian jardun bitartean, tesi honen ikerlerroekin lotura zuzenik ez duten bestelako ekarpen batzuk ere gauzatu dira, hala nola, tesiko ikerlerro nagusitik kanpo aurrentrenatutako euskararako hizkuntza-ereduak eta tesiko ikerlerroekin lotura zuzenik ez duten argitalpenak.

Hizkuntza-ereduei dagokienean, tesian zehar BERT arkitekturan zentratu garen arren, kodetzaile motako beste arkitektura batzuk ere jorratu dira, besteak beste, Electra (Clark et al., 2020b) eta RoBERTa (Liu et al., 2019), baina ez da lortu ElhBERTeu ereduaren emaitzak hobetzerik. Bestalde, tesia kodetzaile motako ereduetan zentratu den arren, eredu sortzaileak ere landu dira bidenabar. Alde batetik, sekuentziatik sekuentziarako atazak lantzeko kodetzaile-deskodetzaile motako hainbat BART eredu (Lewis et al., 2020) entrenatu dira, eta parafrasien sorkuntzan (datu gehikuntzarako), laburpen automatikoan, galdera-erantzunean eta zuzentzaile gramatikalaren ikerketan eta garapenean lagungarriak izan dira. Beste alde batetik, deskodetzaile motako hizkuntza-eredu autorregresiboak ere landu dira, eta GPT2 (Radford

[7] `https://huggingface.co/orai-nlp/bert-base-sw`
[8] `https://huggingface.co/orai-nlp/bert-medium-sw`
[9] `https://huggingface.co/orai-nlp/ElhBERTeu-medium`

et al., 2019), OpenELM (Mehta et al., 2024), Llama3 (Dubey et al., 2024) eta SmolLM2 (Allal et al., 2025) hizkuntza-eredu sortzaileak aurre-entrenatu ditugu. Azken eredu sortzaile horietako batzuei lotuta, argitalpen berri bat bidean da.

Argitalpenei dagokienean, doktorego tesiarekin loturarik ez duten beste bi artikulu zientifikotan hartu dut parte epe honetan:

- Urbizu, G., Soraluze, A., and Arregi, O. (2020). **Sequence to Sequence Coreference Resolution.** In Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference: CRAC20 (pp. 39-46). Barcelona (online).
- López de Lacalle, M., Saralegi, X., Saizar, A., Urbizu, G., and Corral, A. (2023). **Strategies for Bilingual Intent Classification for Small Datasets Scenarios.** Procesamiento del Lenguaje Natural: SEPLN23 (pp. 137-147). Jaén.

1.4 Tesi-Txostenaren Antolaketa

Bukatzeko, tesi-txosten honek ondorengo egitura jarraitzen du:

Lehendabizi, sarreraren kapitulu honek doktorego tesiaren ikerketa esparrua, motibazioa, helburuak, ikerketa-galderak eta egindako ekarpenak biltzen ditu.

2. Kapituluak, tesi-txostena ulertzeko baliagarriak izan litezkeen aurrekariak eta ondorengo kapituluetako lana sostengatzen duten artearen egoera osatzen duten literaturako lanak ditu hizpide.

3. Kapituluan euskarazko hizkuntza-ereduak hizkuntza ulermenean ebaluatzeko BasqueGLUE proba-bankua aurkezten da, ElhBERTeu ereduarekin batera.

4. Kapituluan hizkuntza-eredu diskriminatzaileak entrenatzean baliabide urriko inguruneetan BERT ereduek jarraitzen dituzten eskala-legeak aztertu dira 4 hizkuntzatan.

5. Kapituluan euskararen hitz-ordena malguak eta morfologia aberatsak hizkuntza-ereduek euskararen gramatika barneratzeko duten gaitasunean zer eragin duten ikertu da, BL2MP datu-multzoa aurkeztearekin batera.

6. Kapituluan hizkuntza-ereduak aurrentrenatzeko datu eskasiari aurre egiteko itzulpen automatiko bidez sortutako datu sintetikoak erabiltzearen bideragarritasuna aztertu da.

Tesi-txostenari itxiera emateko, aurrez aurkeztutako ikerketa-galderak erantzun eta doktorego tesiaren ondorioak plazaratzen ditugu eta etorkizuneko lanei buruz hausnartzen da 7. Kapituluan.

Amaitzeko, bibliografia, corpusei buruzko eranskin mardula eta glosategia ditugu txostenaren bukaeran. Glosategia txostenaren amaieran agertzen den arren, tesian zehar landutako terminologia topa dezakezu bertan eta sarreraren ondoren irakurtzea gomendatzen da, txosteneko kapitulu mamitsuenen digestioa errazte aldera.

2. KAPITULUA

Artearen Egoera

Kapitulu honetan, lehenik eta behin, tesiaren ardatz izango diren hizkuntza-eredu neuronaletan sakonduko dugu: hitz-bektoreak eta Transformer arkitektura azaltzen hasi, eta kodetzaile motako hizkuntza-eredu maskaratuekin jarraituz.

Ondoren, hizkuntza-ereduen inguruan tesian jorratutako ikerlerroei dagokien artearen egoera izango dugu hizpide: hizkuntza-ereduak NLUn ebaluatzeko proba-bankuak, hizkuntza-ereduak jarraitzen dituzten eskala-legeak, hizkuntza-ereduen gramatika gaitasunak eta datu sintetikoen erabilera, besteak beste.

2.1 Hizkuntza-Eredu Neuronalak

Hizkuntzaren prozesamenduaren hastapenetan maiz erabiltzen ziren maiztasunetan oinarritutako hitz-zaku (*bag-of-words*) eta TF-IDF (*Term Frequency - Inverse Document Frequency*) testuaren errepresentazio metodo estatistikoek hitzen ordena ez dute kontuan hartzen. Gabezia horri erantzuteko, hitzen ordenari dagokion garrantzia ematen dioten hizkuntza-ereduak lantzeari ekin zitzaion.

Hizkuntza-eredua (*language model*) giza hizkuntzako (edo hizkuntza naturaleko) hitzen[1] (w) sekuentzien gaineko probabilitate-banaketa bat da (Ikus 2.1 ekuazioa). Edozein luzerako hitzen ausazko sekuentzia bati probabilitate bat esleitzeko, hizkuntza bateko (edo hainbatetako) testu corpusetan entrenatzen dira. Dena den, ikasketa-datuetan aurkitzen ez diren hizkuntza-sekuentzia baliozkoei probabilitate ez-nuluak esleitzea beharrezkoa dute.

$$P(w_1, \ldots, w_n) = \prod_{t=1}^{n} P(w_t \mid w_1, \ldots, w_{t-1}) \tag{2.1}$$

Hizkuntza-ereduak askotariko atazak jorratzeko erabili daitezke, besteak beste, eskuzko idazkeraren ezagutzan (Tensmeyer et al., 2018), teklatuetako testu prediktiboan (Hard et al., 2018), zuzentzaile ortografikoetan (Hu et al., 2020c), hizketaren ezagutzan (Gulati et al., 2020) eta hizkuntzaren sorkuntzako hainbat atazatan (Brown et al., 2020).

Horrelako atazak lantzeko, hizkuntza-eredu estatistiko ez neuronalen artean zabalduenak n-gramak (2-gram, 3-gram, etab.) izan dira (Jurafsky and Martin, 2019), zeinak hitz baten agerpena aurreko hitzek (n hitzeko leihoan) baldintzatua egotean oinarritzen diren (Ikus 2.2 ekuazioa).

$$P(w_1, \ldots, w_n) \approx \prod_{t=1}^{n} P(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1}) \tag{2.2}$$

Neurona-sareen etorrerarekin, corpus erraldoiekin elikatuta hizkuntza modelatzen eraginkorragoak diren hizkuntza-eredu neuronalekin ordezkatu dira n-gramak: neurona-sare errepikakorrekin (*recurrent neural network* edo RNN (Bahdanau et al., 2015)) lehenik, eta Transformer arkitekturarekin gerora (Vaswani et al., 2017).

2.1.1 Hitz-Bektoreak

Hizkuntzaren modelatzean hitzen errepresentazio eraginkorrak lortzea funtsezkoa da. Horretarako, hitzak ordenagailuak prozesatzeko moduan bektore espazio batean mapatzeko, hitz-bektoreak darabilzkigu.

[1] Edo hizpeko tokenen (*subword tokens*).

Hitz-bektoreek (*word embedding*) hitzen esanahiak dimentsio baxuko bektoretan biltzen dituzte abstraktuki. Bektore espazio horretan gertu dauden hitzek antzeko esanahia izatea bilatzen da hitzen errepresentazioak bektoreetan kodetzen diren moduekin. Horrek, hitz-bektoreen arteko distantzia, kosinu-antzekotasuna eta bestelako eragiketa aljebraikoen bitartez hitzen arteko erlazio semantikoak lantzea ahalbidetzen du.

Hitz-bektoreak hizkuntza-ereduetatik edo ezaugarri-ikasketa tekniken bitartez lor litezke, besteak beste. Hitz-bektoreen artean estatikoak eta testuingurudunak ere bereizten ditugu. Hitz-bektore estatikoen artean, *Word2vec* (Mikolov, 2013), *GloVe* (Pennington et al., 2014) eta *FastText* (Bojanowski et al., 2017) azpimarra genitzake.

Word2vec (Mikolov, 2013) hitz-bektoreak corpus batetik ikasten dira, hitzen ordena kontuan hartuta, sakonera txikiko (2 geruzako) neurona-sare arinak entrenatuz *CBOW* edo *Skipgram* algoritmoetako bat erabiliz. *GloVe* (Pennington et al., 2014) hitz-bektoreek, berriz, corpus bateko hitzen agerkidetza matrizeak faktorizatuz lortzen du hitzen erlazio semantikoak biltzea. Azkenik, *FastText* (Bojanowski et al., 2017) hitz-bektoreek, *Word2vec* teknikako hitzei karaktere mailako n-gramak gehituz, hitz-bektoreak hobeak lortzen dituzte.

Hitz-bektore horiek estatikoak dira, ordea, eta hitzaren adiera orori bektore berbera esleitzen diete, hitza ageri den esaldiaren testuingurua kontuan hartu gabe. Arazo horri aurre egin nahian, testuingurudun hitz-bektoreak garatu ziren: lehenetakoa ELMo (Peters et al., 2018) izan zen, neurona-sare errepikakor bidirekzional bat corpus handietan hizkuntza modelatzean entrenatuz; Geroztik, Transformer arkitektura oinarri duten BERT bezalako kodetzaile motako maskaratutako hizkuntza-ereduen (Devlin et al., 2019) edo GPT (Radford, 2018; Radford et al., 2019) familiako deskodetzaile motako hizkuntza-eredu autorregresiboen geruzetatik erauzten dira testuingurudun hitz-bektore aberatsenak (Arora et al., 2020), testuinguruaren arabera hitzen esanahia desanbiguatzeko eta ñabardurak gehitzeko.

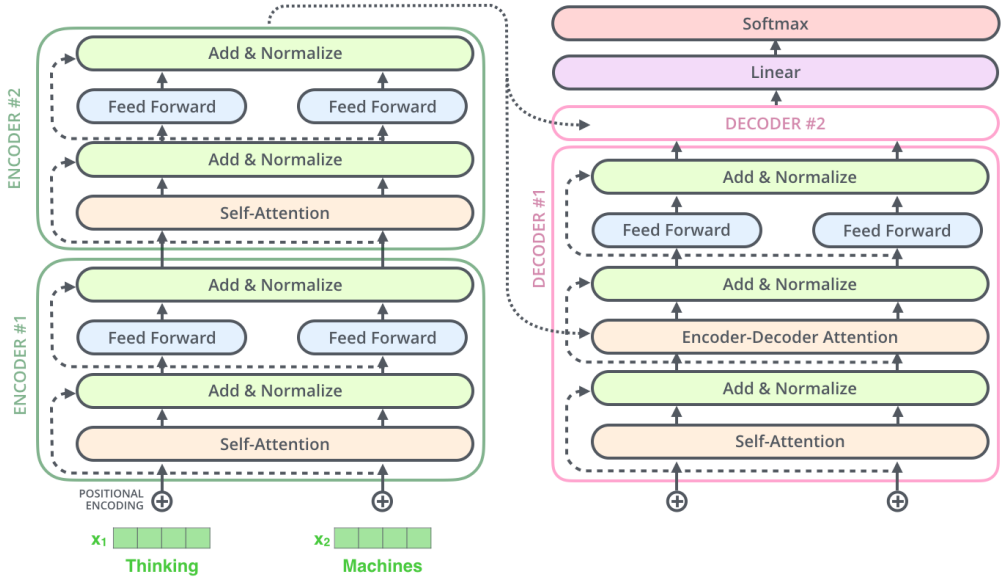
2.1.2 Transformer Arkitektura: Atentzioa eta Kitto!

"*Atentzioa eta Kitto!*"[2]. Horrelaxe zioen Transformer arkitektura neuronala aurkeztu zuen artikuluaren izenburuak (Vaswani et al., 2017). Ordura arte,

[2] *Attention Is All You Need!*

itzulpen automatikoan jaun da jabe zen neurona-sare errepikakorrak atentzio mekanismoarekin konbinatzen zituen hurbilpenari (Bahdanau et al., 2015) erreleboa eman zion sekuentziatik sekuentziarako atazak ebazteko, itzulpen automatikoan kasu. Transformer arkitekturak, Bahdanau et al. (2015) lanean proposatzen den atentzio mekanismoa optimizatu eta ustiatzen du, saren neuronal errepikakorrak bidean lagaz, euren desabantailekin batera: paralelizagaitza izatea eta urruneko dependentziekin zailtasunak.

Transformerrak, bi zati ditu, kodetzailea eta deskodetzailea, elkarri lotuak (Ikus 2.1. Irudia). Transformer originalak, adibidez, seina kodetzaile eta deskodetzaile geruza ditu, 12 guztira.



2.1. Irudia: Transformer arkitekturaren eskema. "Transformer Ilustratua" blog sarreratik (Alammar, 2018b) hartutako irudia. CC BY-NC-SA 4.0 lizentzia.

Kodetzaile geruza guztiek bi azpigeruzez osatutako egitura bera dute. Lehenik, Transformer arkitekturan elementu giltzarria den autoatentzio (*self-attention*) geruza dago, eta honek kodetzaileari hitz bakoitza kodetzean, sarrerako esaldiko gainontzeko hitz guztietan aldi berean arreta jartzeko aukera ematen dio, urruneko dependentziak modu eraginkorrean prozesatuz. Jarraian, autoatentzio geruzaren irteera aurrerantz elikatutako (*feedforward*) neurona-sareari pasatzen zaio.

Deskodetzaile geruza bakoitzak bi azpigeruza horien artean kodetzailearen azken geruzari kasu egiten dion atentzio geruza gehigarria du, kodetzaile-deskodetzaile atentzioa deritzona, sarrerako hitz esanguratsuetan arreta jar dezan.

Transformerra sarrerako esaldiko hitzekin elikatzeko, lehenik, hitz horiek zenbakizko balioetara bihurtu behar ditugu. Horretarako, hitz bakoitzari neurona-sarearen entrenamenduan zehar ikasiko den hitz-bektore bat esleitzen zaio. Praktikan, hiztegi (*vocabulary*) erraldoiekin jardutea saihesteko, tokenizatzaile batekin testua tokenizatu egin ohi da, hitzak hizpeko (*sub-word*) tokenetan zatikatuz, hizpeko tokenen hiztegi mugatua izateko helburuarekin, adibidez, 32K tokenekoa, bakoitzari hitz-bektore bat dagokiola.

Behin, sarrerako sekuentzia kodetuta, deskodetzaileak irteerako sekuentziaren hitzak sortzen ditu, iteratiboki. Horretarako, aurreko hitzen[3] autoatentzioa eta kodetzaile-deskodetzaile atentzioa uztartzen ditu deskodetzaileak. Deskodetzaileak irteeran itzulitako zenbaki bektoreak hitzetara bihurtzeaz geruza lineala eta *softmax* geruza arduratzen dira.

Transformerraren kodetzaile eta deskodetzaile geruzek paralelizazio maila altua ahalbideratzen dute. Honek entrenamendu garaian –hau da, galerafuntzio baten bidez ereduak ekoizten dituen irteerak datu-multzoko urrepatroiarekin alderatuz eta gradiente-jaitsiera bezalako optimizazio teknikak erabiliz neurona-sarearen pisuak edo parametroak eguneratzen diren prozesuan– GPUen kalkulu-ahalmena bere osotasunean aprobetxatzeko aukera emateaz aparte, neurona-sarearen eskalagarritasuna bideragarri egiten du.

2.1. Irudiak hemen azaldutako sekuentziatik sekuentziarako Transformer arkitekturaren eskema orokorra biltzen du. Transformer arkitektura sakonago eta hobeki ulertu nahi duenari, erreferentziazko bilakatu den "*Transformer ilustratua*" blog sarrerara[4] (Alammar, 2018b) jotzea gomendatzen diogu.

2.1.3 Maskaratutako Hizkuntza-Ereduak

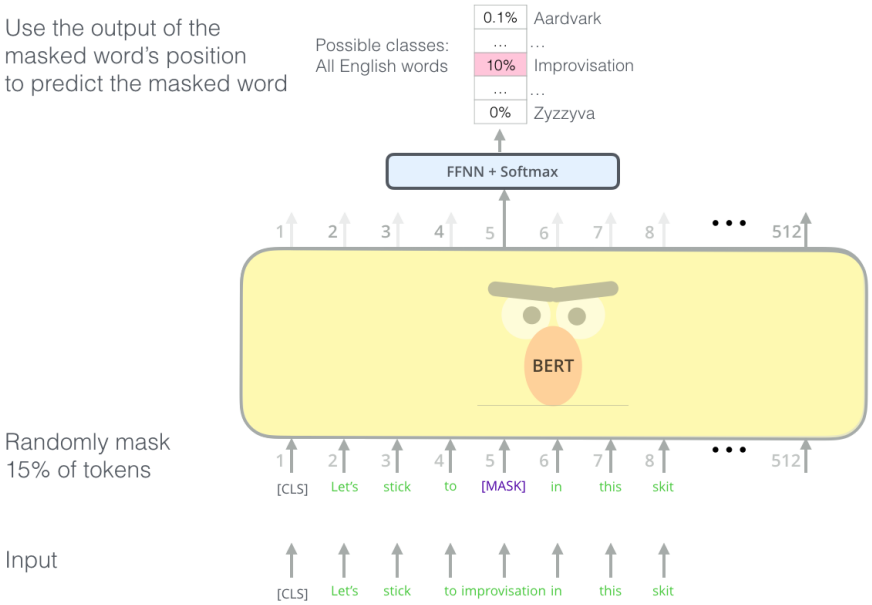
Transformerra modu gainbegiratuan entrenatzeko sekuentziatik sekuentziarako atazetako hamarnaka milaka sarrera-irteera bikote behar dira, eta horiek lortzea lan nekeza izan ohi da. Hizkuntza modelatze atazan, bestalde, ez dago inongo eskuzko anotazioen beharrik, eta hizkuntza-ereduak testu gordinaren

[3] Entrenamendu garaian urre-patroia, inferentzian aurreko posizioaren irteera.

[4] `https://jalammar.github.io/illustrated-transformer/`

gainean entrenatzen dira, ikasketa autogainbegiratua edo erdigainbegiratua deritzonean.

Testu hutsaren gainean Transformerrak hizkuntza modelatzen entrenatzeko, ataza erdigainbegiratu ezberdinak proposatu dira, eta nagusiki, bi bide jorratu dira. Alde batetik, BERTen (Devlin et al., 2019) antzerako kodetzaile motako eredu diskriminatzaileak maskaratutako hizkuntza-modelatzean (*Masked Language Modeling* edo MLM) entrenatzen dira, hau da, helburu-funtzio gisa sarrerako testuko hitzetako batzuk maskaratuz edo ezkutatuz, ereduak hutsune horiek betetzen ikas dezan. Kodetzaile-deskodetzaile motako hizkuntza-ereduek (Lewis et al., 2020; Raffel et al., 2020) ere hurbilpen hori bera darabilte. Beste aldetik, GPT (Radford, 2018) gisako deskodetzailez soilik osatutako eredu autorregresiboek hurrengo tokena aurreikusteko (*Next Token Prediction*) helburu-funtzioa dute, hots, esaldiarekin jarraitzen entrenatu dira.

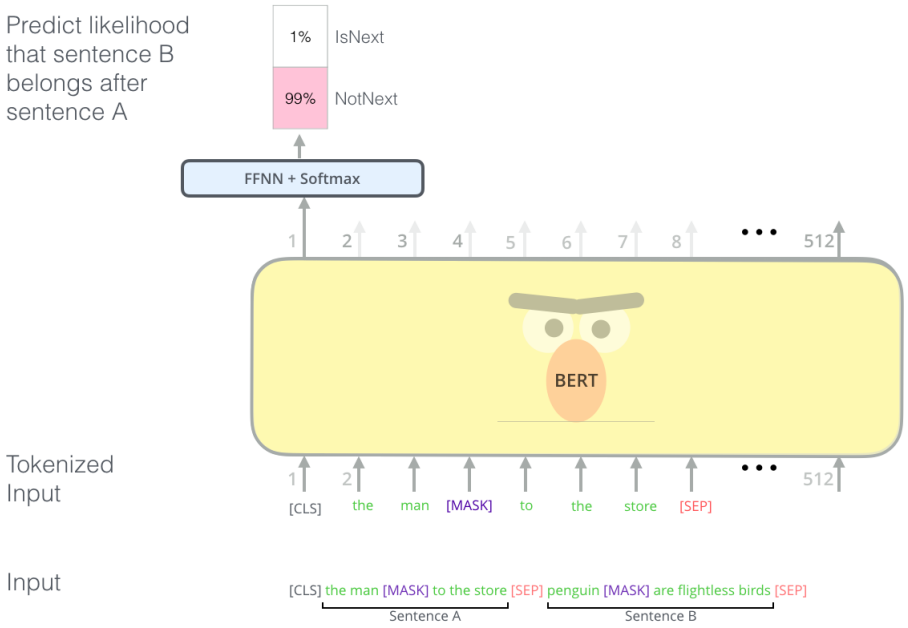


2.2. Irudia: BERT ereduaren aurrentrenamenduko maskaratze helburu funtzioa. "BERT Ilustratua" blog sarreratik hartutako irudia (Alammar, 2018a). CC BY-NC-SA 4.0 lizentzia.

BERT (Devlin et al., 2019) ereduak bi ataza erdigainbegiratu konbinatzen ditu helburu-funtzio gisa aurrentrenamenduan. Lehena, dagoeneko aipatutako MLM ataza da, non sarrerako tokenen %15a maskaratzen edo ezkutatzen

diren[5] token horiek testuingurutik ondorioztatu ditzan (ikus 2.2. Irudia). Bigarrena, berriz, hurrengo esaldia aurreikustearen *(Next Sentence Prediction* edo NSP) ataza[6] da, zeina CLS sailkapen tokenaren bitartez lantzen den (ikus 2.3. Irudia).

Ingeleserako aurkeztutako lehen BERTa liburuez eta Wikipediaz osatutako 3,3B hitzeko corpusarekin aurrentrenatua da. Ereduaren pisuak MLM eta NSP helburu funtzioetara doitzen diren heinean, eredua ezagutza linguistikoa (semantiko zein gramatikala) eta corpusak bildutako ezagutza orokorra barneratzen doa.



2.3. Irudia: BERT ereduaren aurrentrenamenduko hurrengo esaldia aurreikustearen ataza erdigainbegiratua, CLS sailkapen tokenaren irteera gisa ikasten dena. "BERT Ilustratua" blog sarreratik hartutako irudia (Alammar, 2018a). CC BY-NC-SA 4.0 lizentzia.

Orduz geroztik, maskaratze ataza erdigainbegiratuan oinarritutako hainbat MLM aldaera aireratu dira, kasu gehienetan NSP ataza erdigainbegiratua

[5]Token horien %80a maskaratzen da, %10a ausazko beste hitz batez ordezkatzen da, eta %10a bere horretan uzten da.

[6]NSP ataza baztertu egin dute ondorengo kodetzaile arkitektura berri gehienek.

baztertuz. Kodetzaile motako MLMen artean, RoBERTa (Liu et al., 2019) deritzon BERTaren bertsio optimizatua, maskaratze metodoarekin jolasten duten Electra (Clark et al., 2020b) eta SpanBERT (Joshi et al., 2020a), hizkuntza-eredu eleaniztun sendoagoak lortzeko testu paraleloetatik ere ikasteko helburuarekin maskaratze ataza moldatzen duen XLM (Conneau and Lample, 2019), geruzen artean pisuak elkarbanatuz ereduaren tamaina arintzen duen ALBERT (Lan et al., 2019), deskorapilatutako atentzioa gehitzea eta deskodeketako maskaratzean hobekuntzak proposatzen dituen DeBERTa (He et al., 2020) eredua edo egungo eredu sortzaileen aurrerapen berriak txertatzen dituen ModernBERT (Warner et al., 2024) eredua ditugu, besteak beste.

2.1.4 Transferentzia-Ikaskuntza

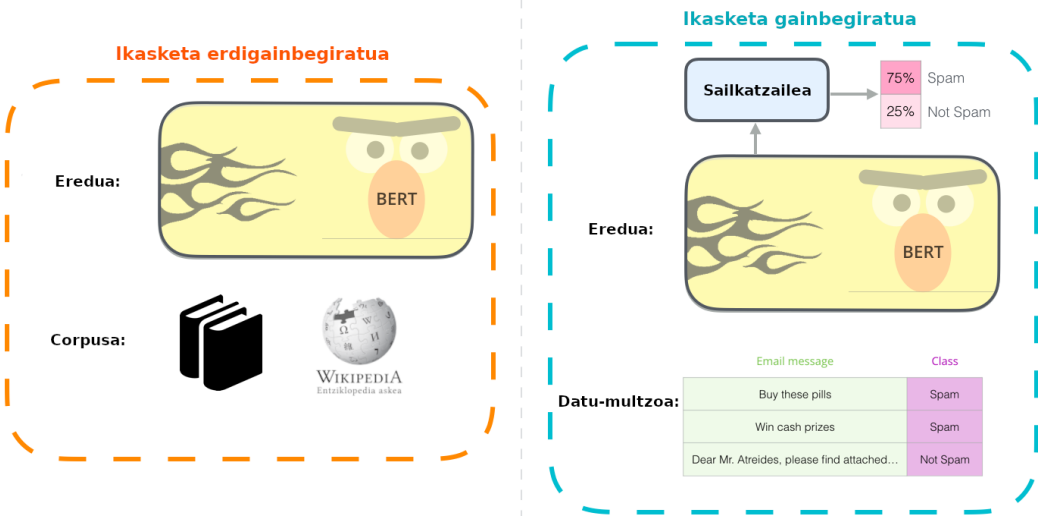
MLM atazan aurrentrenatutako kodetzaile motako hizkuntza-ereduetatik erauzitako testuingurudun hitz-bektoreek eta eredu horiek sekuentziei esleitutako probabilitateek jada aplikazio eremu ugari dituzte hizkuntzaren prozesamenduko hainbat atazatan (Arora et al., 2020; Fang et al., 2024; Salazar et al., 2020; Kauf and Ivanova, 2023).

Baina eredu hauek hizkuntzaren prozesamenduaren esparruan nonahiko bilakatu dituen estrategia bestelakoa izan da. Kodetzaileek aurrentrenamenduko corpusetik barneratutako ezagutza linguistiko eta mundu ezagutza guzti hori helburu-atazetara barreiatzeko transferentzia-ikaskuntza (*transfer-learning*) deritzon estrategia jorratzen da nagusiki (Arase and Tsujii, 2019; Fabien et al., 2020; Sun et al., 2019).

BERT bezalako kodetzaile motako hizkuntza-eredu diskriminatzaileek funtsezko aurrentrenamendu fase bat dute lehenik, eta bertan testu bilduma erraldoietatik hitzen eta esaldien arteko erlazio sakonak eta hizkuntzaren errepresentazio orokorra gordetzen dituzten testuingurudun hitz-bektoreak ikasten dituzte modu erdigainbegiratuan. Bigarren fase batean, sailkapen[7] ataza jakin bateko datu-multzo etiketatuekin, modu gainbegiratuan, entrenamenduarekin jarraitzen da, aurrentrenamenduan ikasitako pisuak atazara birdoituz. Bi fase hauek 2.4. Irudian ditugu grafikoki azalduak.

Transferentzia-ikaskuntzak helburu-atazarako anotatutako corpus eskakizunak murrizten ditu (Grießhaber et al., 2020), ataza eta hizkuntza berrieta-

[7]Token sailkapena edo testu sailkapena izan liteke.



2.4. Irudia: BERT ereduaren entrenamenduko bi faseak. Lehenik aurrentrenamendua corpusetatik ikasketa erdigainbegiratuaren bidez ezagutza orokorra barneratzeko. Bigarren, ataza jakin bateko datu-multzo anotatuetan birdoitzen da, ikasketa gainbegiratuan. "*BERT Ilustratua*" blog sarrerako iruditik moldatua (Alammar, 2018a). CC BY-NC-SA 4.0 lizentzia.

rako aplikazioak garatzearen kostua nabarmen murriztuz. Gainera, emaitzetan salto kualitatiboa ere lortzen da aurrentrenamendu fase horretan corpus erraldoietatik ikasitako ezagutza guztiaren ondorioz (Devlin et al., 2019; Tenney et al., 2019a).

Bestalde, BERT (Devlin et al., 2019) ereduaren pisuak (ingeleserako elebakarra, zein eleaniztuna), eta aurrentrenamenduko eta birdoitzeko kodea eskuragarri jarri izanak ere, teknologia hau ataza, hizkuntza (Antoun et al., 2020; Cañete et al., 2020) eta aplikazio eremu askotara (medikuntzara (Huang et al., 2019) eta zuzenbidera (Chalkidis et al., 2020), besteak beste) hedatzea bultzatu du. BERTen Github biltegiak[8] 39K izar eta 9700 adarkatze ditu, eta Hugginfaceen[9] 70K bert eredutik gora daude eskuragarri. Aurrentrenatutako hizkuntza-ereduen pisuak eta kodea elkarbanatzearen kulturak ikerketa-arloaren aurrerapena azkartu du, ikerlarion eta garatzaileon kalkuluahalmen, datu eta denbora kostuak nabarmen arinduz.

[8] `https://github.com/google-research/bert`
[9] `https://huggingface.co/models?other=bert`

2.2 Hizkuntza-Eredu Eleaniztunak eta Euskararako Hizkuntza-Ereduak

Ingelesa eta baliabide ugariko beste hizkuntza batzuetarako hizkuntzaren prozesamenduak sekulako aurrerapausoak eman ditu azken urteotan. Baina, tesi honen hastapenetan, baliabide urriko –datu edo kalkulu-ahalmenari dagokionez– hizkuntza askoren egoera bestelakoa zen. Gabezia horri erantzuteko proposatutako aterabideetako bat eredu eleaniztunen bidea jorratzea izan da.

Hizkuntza-eredu eleaniztunak eraikitzeko egin beharreko bakarra hizkuntza ezberdinetako corpusak elkartu eta aurrentrenamendua elkarrekin nahasian egitea da (Devlin et al., 2019). Hala ere, badira zenbait alor kontuan hartu beharrekoak.

Eredu eleaniztunetan hizkuntza kopurua igo ahala tokenizatzailearen tamaina handitzea onuragarria dela frogatu da (Devlin et al., 2019; Conneau, 2019; Virtanen et al., 2019), datu gutxieneko hizkuntzetan ere gutxieneko estaldura izan dezan. Horretaz gain, jorratu nahi diren hizkuntzen idazkera sistemak kontuan izatea ere garrantzitsua da, karaktere kopurua, idazkeraren noranzkoa eta zuriuneak egoki tratatzeko (Devlin et al., 2019).

Gainera, hainbat hizkuntza aldi berean lantzean, hizkuntza, hizkuntza familia edo idazkera sistema ezberdinen corpus tamainen arteko desoreka handiak suertatzen badira, baliabide urriko hizkuntzek gainontzekoen artean diluituegi geratzeko arriskua dute (Virtanen et al., 2019; Pires et al., 2019). Horri aurre egin asmoz, ohikoa da datu gutxien dituzten hizkuntzetako corpusari gainlaginketa aplikatzea (Devlin et al., 2019; Conneau, 2019; Le Scao et al., 2023).

Kodetzaile motako hizkuntza-eredu diskriminatzaile masiboki eleaniztunen artean mBERT (Devlin et al., 2019) eta XLM-RoBERTa (Conneau, 2019) nabarmenduko genituzke. mBERT Transformerretan oinarritutako lehen hizkuntza-eredu eleaniztuna da, eta BERT arkitektura (Devlin et al., 2019) du oinarri, MLM eta NSP helburu-funtzioak konbinatuz. 104 hizkuntzetako Wikipediekin entrenatua dago, testu gutxieneko hizkuntzei gainlaginketa aplikatuz, eta 110Kko hiztegi tamaina du. XLM-RoBERTak (Conneau, 2019), berriz, RoBERTa arkitektura (Liu et al., 2019) du oinarri, MLM helburu-funtzioa bakarrak darabilena. CommomCrawleko 100 hizkuntzetako testuekin dago aurrentrenatua, mBERTen corpusa hamarkoiztera iritsiz,

eta mBERTen egin bezala, testu gutxieneko hizkuntzei gainlaginketa aplikatzen zaie. Hiztegi tamaina ere 250Kra handituz, NLUko atazetan hainbat puntutako hobekuntzak lortzen ditu eta hizkuntza arteko transferentzia-ikaskuntzarako gaitasunak erakusten ditu. Sekuentziatik sekuentziarako ereduen eta eredu autorregresiboen alorrean ere eleaniztunerako joera bera behatu da: hala nola mBART (Liu et al., 2020), mT5 (Xue et al., 2021b), XGLM (Lin et al., 2022) eta BLOOM (Le Scao et al., 2023). Eredu eleaniztunek anotatutako datuen gabezia sakona pairatzen duten baliabide urriko hizkuntzetan aplikazioak garatzea erraztu du; hizkuntza arteko transferentzia-ikaskuntza deritzonak helburu hizkuntzan adibide anotaturik gabe jardutea ahalbidetzen baitu.

Urtez urte, hizkuntza-eredu txikiak entrenatzeko hardware eskakizunak txikitzen doaz inplementazio eraginkorragoak kodetu ahala (Izsak et al., 2021; Geiping and Goldstein, 2023), eta hiztun komunitate txikietako ikerlariek testu corpusak eskuragarri jartzeak ere hizkuntza horietarako hizkuntza-ereduen garapena azkartu du.

2.2.1 Euskarazko Hizkuntza-Ereduak

Euskarazko testua prozesatzeko gai izan zen Transformerretan oinarritutako lehen hizkuntza-eredu neuronala mBERT (Devlin et al., 2019) izan zen, aurrentrenamendurako erabilitako 104 hizkuntzetako testu-bilduma eleaniztunaren artean, euskarazko Wikipedia ere barne hartzen duena. XLM-RoBERTa (Conneau, 2019) eredu eleaniztunak ere euskara dakar 100 hizkuntzako eskaintzaren artean, euskarako erabilitako corpusa nabarmen handituz.

Alabaina, Agerri et al. (2020) lanak eredu eleaniztun hauen mugak erakusten ditu, eta BERTeus aurkezten du, Transformerretan oinarritutako euskararako lehen hizkuntza-eredu elebakarra. BERTeus 225M hitzeko *Basque Media Corpus* (BMC) corpusean dago entrenatua; zeina euskarazko albisteek eta Wikipediak osatzen duten. Tokenizatzaileari dagokionez, mBERTen tokenizatzaileak (Devlin et al., 2019) euskara bezalako baliabide urriko hizkuntzetan dituen gabeziak azaleratzen dituzte, eta 50Kko hiztegidun tokenizatzaile elebakar berri bat entrenatzen dute BERTeuserako. 2.1. Taulan ikus daitekeenez, BERTeusek artearen egoera berria ezartzen du mBERT eredu eleaniztunari (Devlin et al., 2019) eta aurreko artearen egoerako sistemei alde handiarekin gailenduz ebaluaziorako darabiltzaten NLUko euskarazko lau

atazatan: gaien sailkapena (Agerri et al., 2020), sentimenduen analisia (Agerri et al., 2020), kategoria gramatikalen etiketatzea (*POS tagging*) (Aranzabe et al., 2015) eta entitate izendunen ezagutza (*name entity recognition* edo NER) (Alegria et al., 2006).

Ataza	BERTeus	mBERT	Aurreko SOTA
Gai sailkapena	**76.77**	68.42	63.00
Sentimenduen analisia	**78.10**	71.02	74.02
POS	**97.76**	96.37	96.10
NER	**87.06**	81.52	76.72
b.b.	**84.92**	79.33	77.46

2.1. Taula: BERTeus ereduaren emaitzak NLUko lau atazetan, mBERT eredu eleaniztuna eta aurreko artearen egoerako sistema onenekin alderatuz.

IXAmBERT (Otegi et al., 2020), eredu eleaniztunak, berriz, aurrentrenamenduan euskara, gaztelania eta ingelesa konbinatuz, hizkuntzen arteko transferentzia-ikaskuntza ahalbidetzen du, eredu masiboki eleaniztunek baino emaitza hobeak lortuz.

Aipatutako euskararako hizkuntza-eredu elebakarra eta hiru hizkuntza-eredu eleaniztunak 2.2. Taulan bildu ditugu, bakoitzaren parametro, hizkuntza kopuru, erabilitako corpus eta hiztegi tamainari buruzko xehetasun gehiagorekin batera. Badira euskararako hizkuntza-eredu gehiago[10] ere, baina guk dakigunez, ereduen gehiengoak ebaluazio egokia falta du.

Eredua	Param	Hizk.	Corpusa (euskaraz)	Hiztegia
mBERT	179M	104	∼7B (∼40M)	110K
XLM-RoBERTa	279M	100	∼400B (270M)	250K
BERTeus	125M	1	225M (225M)	50K
IXAmBERT	177M	3	3B (225M)	119K

2.2. Taula: Lan honen aurretik euskararako eskuragarri zeuden kodetzaile motako hizkuntza-eredu diskriminatzaileak (elebakarrak zein eleaniztunak) eta horien aurrentrenamendu corpusak eta hiztegi tamainak.

Doktorego tesi honen aurretik euskararako berariaz sortutako hizkuntza-eredu neuronalek, aurrentrenamendurako euskarazko corpusaren aniztasuna-

[10] `https://huggingface.co/models?language=eu&sort=downloads`

ren aldetik gabeziak dituzte. BERTeus eta IXAmBERT albisteekin eta Wikipediako testuekin soilik aurrentrenatu dira, domeinuari, estiloari eta erregistroari dagokionean, literatura, zientzia argitalpenak edo ahozko elkarrizketak bezalako iturri aberasgarriak baztertuz. Horri erantzunez, 3. Kapituluan ElhBERTeu aurkeztu dugu. Bestalde, euskararen ezaugarriek, baliabide kopuruari eta ezaugarri linguistikoei dagokienez, maskaratutako hizkuntza-ereduak entrenatzean duten eraginari buruzko azterketa baten hutsunea identifikatu dugu. Gabezia horiei erantzuteko, tesi honetan IG2 eta IG3 ikerketa-galderak mahaigaineratu ditugu, 4. eta 5. Kapituluetan jorratuko direnak.

2.3 NLUrako Ebaluazio Proba-Bankuak

Hizkuntza-eredu sendo batek atazarekiko edo datu-multzoarekiko modu independentean prozesatu behar luke testua eta ulermen gaitasun eta ezagutza orokor horiek hizkuntzaren prozesamenduko ataza ezberdinetan aplikatzeko gai behar luke (Wang et al., 2018). Aurrentrenatutako hizkuntza-eredu neuronalak ebaluatu ahal izateko, hizkuntza ulermen orokorreko proba-bankuak sortu dira. Proba-banku hauek hainbat ataza biltzen dituzte, guztiak ataza formatu bateratuetara ekarriz.

GLUE (Wang et al., 2018) proba-bankuak ingeleserako 9 ataza biltzen ditu, bi esaldi sailkapenekoak (onargarritasuna eta sentimenduen analisia), hiru antzekotasunari edo parafrasiei lotuak eta azken lauak hizkuntza inferentziako atazak. Atazak aukeratzeko orduan, dagoeneko aurrez sortuak dauden datu-multzoen artean askotariko domeinuak, zailtasun maila ezberdinak eta entrenamendu datu kopuru ezberdinak[11] barne hartzea bilatu da. Aukeratutako atazek hizkuntza-ereduen hizkuntza (sintaxia eta semantika) ulertzeko gaitasunaz gain, sen ona (*commonsense*) eta mundu ezagutza ere neurtzen dituzte. Gainera, ataza batzuetan entrenamendu eta ebaluaziorako multzoek distribuzio ezberdina dute, ereduen orokortzeko gaitasuna saritzeko.

Hala eta guztiz ere, hizkuntza-ereduek gizakion maila gainditu dute ingeleserako proba-banku horretan argitaratu eta urtebetera (Kiela et al., 2021), nahiz eta oraindik NLU ebaztetik oso urrun gaudela jakina den (Kiela et al., 2021). Erantzun gisa, sen oneko arrazoiketa eta mundu ezagutza sakonagoa eskatzen duten ataza zailagoz osatutako banku-probak biltzen eta sortzen

[11] Datu-multzo handienak 393K adibide ditu eta txikienak 634.

jarri da arreta: adibidez, SuperGLUE (Wang et al., 2019), GEM (Gehrmann et al., 2021) eta BIG-Bench (BIG-bench, 2021). Berriki, hizkuntza-eredu sortzaileek NLUn ekarritako hobekuntzarekin, eta ikasketa gainbegiratuaren fasea saltatzeko aukerarekin (Brown et al., 2020; Wei et al., 2021), proba-bankuei ikasketako daturik gabeko ataza zailagoak gehitzen joan dira: adibidez arrazoiketa lantzen duten ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), HellaSwag (Zellers et al., 2019), CommonsenseQA (Talmor et al., 2019) eta WinoGrande (Sakaguchi et al., 2020); eta ezagutza faktuala neurtzeko TriviaQA (Joshi et al., 2017), MMLU (Hendrycks et al., 2020) eta MMLU-pro (Wang et al., 2024).

Orain arte atal honetan aipatutako proba-banku guztiak ingeleserako dira, hots, ezin dira erabili gainontzeko hizkuntzetan NLU ebaluatzeko. Proba-banku eleaniztunak badiren arren, XGLUE (Liang et al., 2020) edo XTREME (Hu et al., 2020b) adibidez, ataza horietan entrenatzeko banatutako datu-multzoak ingelesez bakarrik datoz. Hizkuntza-eredu handiek ikasketa gainbegiraturik gabe helburu-atazetan jarduteko duten gaitasunak, datu-multzo eleaniztunak sortzea errazten du, birdoitzeko entrenamendu datuen beharra saihestuz, eta hasiak dira horretarako datu-multzo eleaniztunak agertzen, adibidez MMMLU (Hendrycks et al., 2021), X-StoryCloze (Lin et al., 2021a), Belebele (Bandarkar et al., 2024) edo INCLUDE (Romanou et al., 2024).

Arazo horiek kontuan izanik, baliabide oparoko beste hizkuntza batzuetarako proba-banku elebakarrak ere ari dira ernetzen: CLUE txinerarako (Xu et al., 2020), FLUE frantseserako (Le et al., 2020), SuperGLUEren errusierarako aldaera (Shavrina et al., 2020), ALUE arabierarako (Seelawi et al., 2021), KLUE koreerarako (Park et al., 2021) edo JGLUE japonierarako (Kurihara et al., 2022). Baliabide urriko eta ertainetako hizkuntzetarako horrelako euskarriak oso bakanak dira, baina egon badira: indonesierak badu berea (Wilie et al., 2020), indo-ariar hizkuntzek IndicGLUE (Kakwani et al., 2020) dute, nederlanderak DUMB (de Vries et al., 2023) eta katalanak CLUB (Rodriguez-Penagos et al., 2021) dauka.

Euskarari dagokionez, tesi honen aurretik ez da izan berariaz hizkuntza-ereduak euskararen ulermenean ebaluatzeko sortutako proba-bankurik. Dagoeneko sortutako hizkuntza-ereduak ebaluatu dira ordea, horretarako, hainbat ataza ezberdin baliatuz. BERTeus ebaluatzeko kategoria gramatikalen etiketatzea (*POS tagging*), entitate izendunen ezagutza (*name entity recognition* edo NER), sentimenduen analisia eta gaien sailkapena darabilte (Agerri et al., 2020); IXAmBERT eredu eleaniztuna ebaluatzeko, berriz, ElkarHiz-

ketak elkarrizketazko galdera-erantzunen (*question anwsering edo QA*) datumultzoa (Otegi et al., 2020) darabilte. Tesi honen baitan, hutsune horri erantzuteko helburuarekin, eta etorkizunean euskararako eraikitako hizkuntza-eredu neuronalak NLUn ebaluatu eta eredu ezberdinak sendotasunez konparatu ahal izateko 3. Kapituluan BasqueGLUE proba-bankua aurkezten dugu.

2.4 Eskala-Legeak

Atentzio mekanismoan oinarritutako Transformer arkitekturaren (Vaswani et al., 2017) eta BERTek (Devlin et al., 2019) aurrentrenamendurako proposatutako maskaratze estrategiaren agerpenetik, hizkuntza modelatzeko hainbat aurrentrenamendu estrategia plazaratu dira.

Neurona-sareen arkitekturetan eta entrenamendu estrategietan egindako aurrerapenek ekarritako hobekuntzak alde batera utziz, hizkuntzaren ulermenean lortutako hobekuntza kualitatiboa, nagusiki ereduaren tamaina eta entrenatzean erabilitako testu kopurua izugarri haztearen ondorio zuzena da: Chinchilla (70B parametro) (Hoffmann et al., 2022), LaMDA (137B) (Thoppilan et al., 2022), GPT-3 (175B) (Brown et al., 2020), Gopher (280B) (Rae et al., 2021), Megatron-Turing NLG (530B) (Smith et al., 2022), PaLM (540B) (Chowdhery et al., 2022), GShard (600B) (Lepikhin et al., 2020) eta Switch-C (1,57T) (Fedus et al., 2021).

Eredu tamaina eta datu kopuruan hazkunde azkar honek hizkuntza-eredu handiengan gaitasun berriak azaleratu dituzte (Wei et al., 2022), eredu txikiagoekin aurrez bideragarriak ez ziren atazak lantzeko aukera emanez. Fenomeno hori eredu handiagoen arrazoiketarako gaitasun sakonagoarekin dago lotua (Wei et al., 2022).

Aurrentrenamendurako corpusen tamainek eta hizkuntza-ereduak ulermeneko atazetan lortutako emaitzen arteko erlazioa, aurretiaz ikergai izan dute hainbatek dagoeneko. Emaitzak hobetzera jotzen duten entrenamenduko datu kopurua handitzean (Zhang et al., 2021; Hu et al., 2020a; Micheli et al., 2020; Raffel et al., 2020), nahiz eta puntu batetik aurrera hobekuntza hori moteltzen doan, ereduaren tamaina finko mantentzen bada (Inoue et al., 2021; Martin et al., 2020; Micheli et al., 2020; Raffel et al., 2020; Liu et al., 2021). Gainera, komenigarriagoa da entrenamendu corpusen aniztasunean zentratzea gure ahaleginak, domeinu bereko testu gehiago biltzen aritzea baino (Inoue et al., 2021; Martin et al., 2020; Liu et al., 2021), hizkuntza-ereduak

orokortzeko gaitasun sendoagoa izan dezan, hau da, domeinu jakin batean ebatziko lukeen ataza, domeinuz kanpo ere ebazteko gai izan dadin.

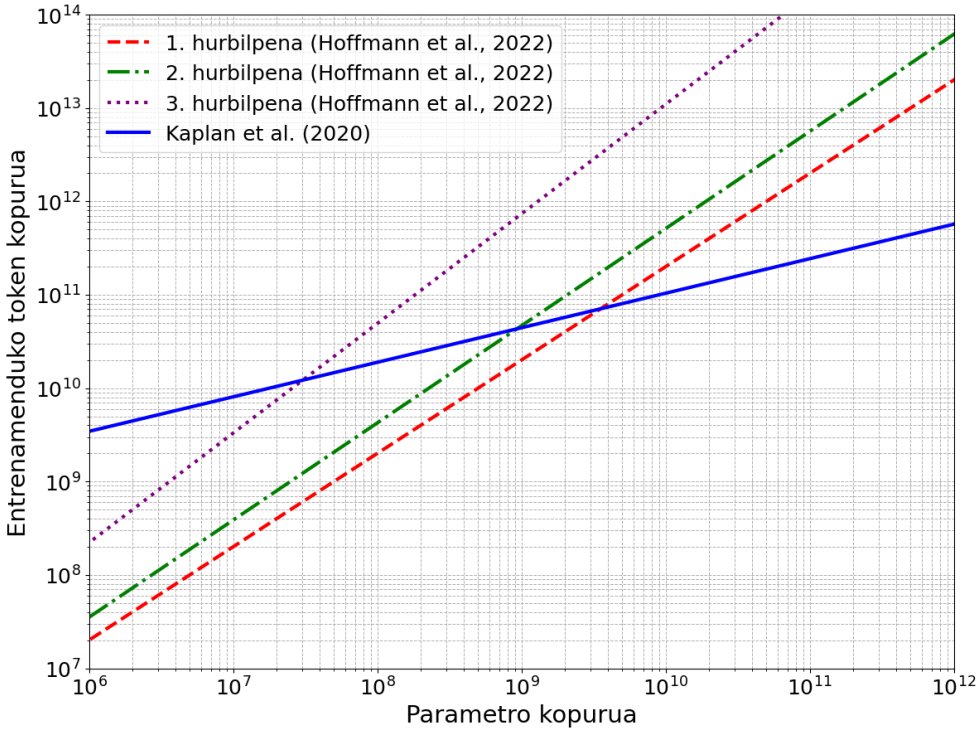
Ereduaren tamainaren eta hizkuntza-ereduak NLUn lortutako emaitzen arteko korrelazioa ere aztertu da. Ereduaren emaitzak hobetzen doaz ereduaren tamaina –eta kalkulu-ahalmena– eskalatu ahala (Turc et al., 2019; Raffel et al., 2020; Xia et al., 2022; Clark et al., 2022). Halaber, lan horietan guztietan aurrentrenamendurako datu-multzo oso handiak erabili dituzte dagoeneko. Beraz, ez dute aztertzen ea emaitzen hobekuntza hori mugatua ote dagoen aurrentrenamendu datu kopuruaren menpe, eta honenbestez ezin daiteke ondorioztatu eredu tamaina handitzea datu urriko inguruneetan ere lagungarria izango denik.

Kaplan et al. (2020) eta Hoffmann et al. (2022) lanek, enpirikoki aztertu dituzte aurrentrenatutako hizkuntza-eredu sortzaileen parametro kopuruaren, entrenamenduko tokenen eta kalkulu-ahalmenaren arteko ratio optimoak eta eskala-legeak ondorioztatu.

Kaplan et al. (2020) lanean 768tik 1.5B parametrora arteko ereduak entrenatzen dituzte, 22M eta 23B arteko token kopuruak erabiliz. Esperimentu sorta horretatik, ereduaren tamaina, corpusaren tamaina eta kalkulu-ahalmena handitu ahala, hizkuntza-ereduak leunki hobetzen doazela ondorioztatzen dute. Datu eta parametroen arteko erlazioari dagokionez, ereduaren tamaina zortzikoizten den aldiro datu kopurua boskoiztu behar litzatekeela diote, 2.5. Irudiko grafikoan lerro jarraituak (urdinez) adierazi bezala.

Hoffmann et al. (2022) lanean, kalkulu-ahalmen aurrekontu finkoa emanik, hizkuntza-ereduaren tamaina eta aurrentrenamenduko token kopuru optimoak topatzen dituzte. Horretarako, eskala-lege propioa eratortzen dute, 400 eredu baino gehiagoren galera kurbetan oinarrituz, 70Mtik 16B parametrora artekoak, eta 5Btik 400B token arteko corpusekin. Ereduaren tamainak eta token kopuruak elkarrekin eskalatu behar luketela ondorioztatzen dute hiru metodo ezberdinen bitartez, 2.5. Irudiko grafikoan lerro etenez adieraziak. Beraz, ordura arteko hizkuntza-eredu handi oro datu gosez geratu dela ondorioztatu ondoren, Chinchilla eredua entrenatzen dute, aurreko hizkuntza-eredu handiak baino nabarmen entrenamendu datu gehiagorekin.

Kaplan et al. (2020) eta Hoffmann et al. (2022) lanetan ingeleserako hizkuntza-eredu autorregresiboetarako eratorritako eskala-legeek ez dute zertan bete, ordea, hizkuntza-eredu diskriminatzaileen kasuan, edo gainerako hizkuntzetan.



2.5. Irudia: Entrenamendu tokenen eta ereduaren parametro kopuruaren arteko eskala-legeak (Kaplan et al., 2020; Hoffmann et al., 2022), Hoffmann et al. (2022) artikuluko 3. Taulako balioetatik eratorria. CC BY-SA 4.0 lizentzia.

Bitartean, baliabide urriko hizkuntzetarako hizkuntza-eredu diskriminatzaileak (100M parametro ingurudunak) eskura dituzten corpus tamainekin entrenatu dira (Joshi et al., 2020b): kinyarwandak 390M hitz (Nzeyimana and Rubungo, 2022), Irlandako gaelikoak 161M (Barry et al., 2021), luxenburgerak 130M (Lothritz et al., 2022), galizierak 45M (Vilares et al., 2021), swahiliak 16M (Martin et al., 2022) eta kitxuak 4,4M (Zevallos et al., 2022). Hala ere, aurrentrenamenduko datu kopurua hainbat magnitude ordenatan handitzea onuragarria dela frogatu da *base* eredu tamaina ohikoenean (Liu et al., 2019).

Bukatzeko, Kaplan et al. (2020) eta Hoffmann et al. (2022) lanetako eskala-lege optimo horiek ereduak corpusean epoka bakar batean entrenatuz ondorioztatu dira, baliabide urri eta ertaineko testuinguruan ereduak hainbat epokaz aurrentrenatu ohi direnean, adibidez, 10 epokaz LuxemBERTen

(Lothritz et al., 2022) eta 400 epokaz CamemBERTen[12] (Martin et al., 2020).

Orain arte hizkuntza-ereduen autorregresiboen eskala-legeek aztertzean baliabide urriko testuinguruetan ez da arretarik jarri zehazki nola eragiten dieten jakiteko, eta hizkuntza-eredu diskriminatzaileen eskala-legeak ikertu gabe daude. Gainera, eskala-legeak ingeleserako soilik aztertu dira, eta hizkuntzen artean estrapolagarriak ote diren ere bermatzeke dago. Hutsune horiek ikusirik, tesi honetako 4. Kapituluan gaiari heldu diogu, baliabide urriko hizkuntzen kasuan hizkuntza-eredu diskriminatzaileei dagozkien corpus tamaina eta parametro kopuruaren arteko erlazio optimokoak aztertuz.

2.5 Hizkuntza-Ereduak eta Gramatika

Aurrentrenatutako hizkuntza-eredu neuronalen zabalkundetik, hainbat autorek Transformerretan oinarritutako hizkuntza-ereduak sintaxiaren, morfologiaren eta semantikaren ezaugarri linguistikoak ikasteko gai direla frogatu dute (Tenney et al., 2019a; Jawahar et al., 2019; Niu et al., 2022). Zhang et al. (2021) lanaren autoreen arabera 10M eta 100M hitz arteko aurrentrenamendu corpusa nahikoa da hizkuntza-ereduak ingelesaren sintaxia, morfologia eta semantika barnera ditzan, baina sen oneko arrazoiketa eta ezagutza faktuala barneratzeko behar den datu kopurua askoz handiagoa da. Ildo beretik, Huebner et al. (2021) eta Cagatan (2023) lanek 5M eta 10M hitzeko haurren mailako testuekin aurrentrenatutako tamaina txikiko RoBERTa (Liu et al., 2019) ereduen ezagutza gramatikala aztertzen dute. Biak bat datoz hizkuntza-eredu neuronal horien ezagutza gramatikalaren jabetza RoBERTa-base eredu ofizialaren parekoa dela baieztatzean, nabarmen parametro kopuru gutxiago eta askoz ere aurrentrenamendu datu gutxiago izanagatik.

Gainera, hizkuntza-eredu neuronalak hizkuntza bakarreko gramatika ikasteko gai izateaz gain, mBERT eredu eleaniztunak dependentzia sintaktikoen errepresentazio unibertsal bat ikasten duenaren ebidentziak aurkezten dituzte Chi et al. (2020) lanean. Era berean, hizkuntza-eredu neuronal eleaniztun horiek informazio morfologikoa aurrentrenamenduan zehar nola ikasten duten erakusten dute Acs et al. (2023) lanean.

[12] 4GBeko corpuseko eredurako.

Sinha et al. (2021) lanaren autoreek ingeleserako hizkuntza-ereduak harrigarriro hitz ordenarekiko aldagaitzak direla ondorioztatzen dute, eta helburuatazetan duten arrakasta batez ere hitzen goi-mailako agerkidetza estatistikoak modelatzeko duten gaitasunagatik dela. Aitzitik, Papadimitriou et al. (2022) lanean hizkuntza-ereduek hitzen ordenari jaramon egiten diotela frogatzen dute gako lexikalak nahikoa ez diren kasuetan, eta horretarako beharrezko diren ezaugarri gramatikalak ikasten dituztela.

2.5.1 Hizkuntza-Ereduen Ezagutza Gramatikala Ebaluatzen

Hainbat hurbilpen ezberdin proposatu dira hizkuntza-ereduen ezagutza gramatikala ebaluatzeko, eta horrek datu-multzo ugariren sorrera ekarri du, bereziki ingeleserako.

Horietako bat, CoLA (*The Corpus of Linguistic Acceptability*) (Warstadt et al., 2019) da, gramatikalki onargarriak ote diren eskuz anotatutako esaldiak biltzen dituen datu-multzoa, GLUE proba-bankuan (Wang et al., 2018) sailkapen ataza gisa datorkiguna. Beste antzeko datu-multzo bat, 20 sailkapen bitarreko ataza anbiguo dakartzan MSGS (*Mixed Signals Generalization Set*) dugu (Warstadt et al., 2020b), hizkuntza-ereduak ezaugarri linguistuko zehatzetan ebaluatzeko aproposa (Tenney et al., 2019b).

Hala ere, bi datu-multzo horiek hizkuntza-ereduak sailkapen ataza bitarretan birdoitzeko beharra dute, eta honek, hizkuntza-ereduak aurrentrenamendu garaian barneratutako ezagutza gramatikala neurtzetik urruntzen gaitu. Honenbestez, hizkuntza-ereduak birdoitzeko beharrik gabe ebaluatzeko (*0-shot* deritzona) datu-multzo berriak proposatu dira, pare-minimoen (*minimal-pair*) kontzeptua oinarri hartuta. Pare-minimoak aspektu gramatikal bakarrean ezberdintzen diren esaldi bikoteak dira, gainontzeko ezaugarri linguistiko oro konstante mantenduz:

Ingeleserako pare-minimo linguistikoen lehen proba-bankua BLiMP (*Benchmark of Linguistic Minimal Pairs*, (Warstadt et al., 2020a)) izan zen. Hizkuntza-ereduak BLiMP osatzen duten pare-minimoak puntuatzen ditu (perplexitatea, galera, ...), eta jarraian bi balioak konparatzen dira, 12 fenomeno gramatikal zehatzetan ereduaren gaitasuna neurtzeko. Berriki, *babyLM* (Warstadt et al., 2023) ataza elkarbanatuaren testuinguruan, BLiMP probabankua ataza gehigarriekin osatu da, beste bost fenomeno gramatikal gehituz.

Onargarriak	Ez onargarriak
*These casseroles **disgust** Kayla.*	*These casseroles **disgusts** Kayla.*
*Aaron **broke** the unicycle.*	*Aaron **broken** the unicycle.*
*Rachelle had bought that **chair**.*	*Rachelle had bought that **chairs**.*
*Which **bikes** is John fixing?*	*Which is John fixing **bikes**?*
*Many girls insulted **themselves**.*	*Many girls insulted **herself**.*

2.3. Taula: BLiMP proba-bankuko (Warstadt et al., 2020a) pare-minimoen adibideak, aspektu gramatikalean ezberdintasuna letra lodiz adierazita.

BLiMPen izaera bera duen *Grammar Test Suite* (Huebner et al., 2021) proba bankua ere, sintaxiko eta morfologiako fenomeno zehatz bertsuak isolatzen dituzten pare-minimoek osatzen dute. *Grammar Test Suite* proba bankuak, esaldi bikoteetako esaldiak hitz errazez[13] eraikia izatea du BLiMPekiko (Warstadt et al., 2020a) bereizgarri, aurrentrenamendu corpusaren hiztegi aberastasunak emaitzetan izan dezakeen eragina ekidin eta hizkuntza-ereduaren gramatika ezagutza modu isolatuan ebaluatu ahal izateko helburuz.

2.5.2 Ingelesa Ez Besteko Hizkuntzen Gramatikak eta Hizkuntza-Ereduak

Gai honen bueltan argitaratutako artikulu eta datu-multzoen gehiengoek ingelesa bakarrik lantzen dute, baina hizkuntza-ereduen gramatika-ezagutza ebaluatzea hizkuntzarekiko mendekoa da, eta hizkuntza bakoitzak erronka propioak dakartza berarekin. Hortaz, beste hizkuntza batzuetarako BLiMPen antzeko pare-minimoen datu-multzoen beharra asetzeko helburuarekin, hainbat lan izan dira: CLiMP (Xiang et al., 2021), SLING (Song et al., 2022) eta ZhoBLiMP (Liu et al., 2024) txinerarako, BLiMP-Arabic (Alrajhi et al., 2022) arabierarako, JBLiMP (Someya and Oseki, 2023) japonierarako eta LINDSEA indonesiera eta tamil hizkuntzetarako (Leong et al., 2023).

Gainera, badira beste datu-multzo batzuk, CoLAren hurbilpena beste hizkuntza batzuetara zabaldu dutenak: RuCoLA (Mikhailov et al., 2022) errusierarako, ItaCoLA (Trotta et al., 2021) italierarako, BLiMP-NL (Suijkerbuijk et al., 2024) nederlanderarako, ScandEval (Nielsen, 2023) daniera,

[13]Haurren mailako testuetan agertzen diren hitz ulerterrazak.

suediera, norvegiera[14], islandiera eta faroerarako, eta CatCola katalanerako (Bel et al., 2024).

Azkenik, euskararako erlazionatutako lan bakarra, guk dakigula, Ravfogel et al. (2018) lana da, non neurona-sare errepikakorrak erabiltzen dituzten euskarazko komunztadura antzemateko. Ondorioen artean, ataza hori euskaraz askoz ere erronka konplexuagoa dela azpimarratzen dute, ingelesarekin alderatuz.

Hizkuntza-ereduen gramatika gaitasunak modu isolatuan ebaluatzeko sortutako datu-multzoak erabilgarriak dira hizkuntza-ereduen gaitasun linguistikoak eta mugak aztertzeko, baina hizkuntza gutxi batzuetarako soilik daude eskuragarri. Orain arte ez da egon euskararako horrelako baliabiderik, eta hutsune hori betetzeko BL2MP datu-multzoa sortu dugu tesi honen testuinguruan. 5. Kapituluan BL2MP datu multzoa nola sortu den eta datu-multzo hori erabiliz hizkuntza-ereduen euskarazko gramatika gaitasunen eta ordena malgua edo morfologia aberatsa bezalako bereizgarri linguistikoen eraginaren azterketa aurkezten dugu.

2.6 Entrenamendurako Datu Sintetikoak

Hizkuntzaren prozesamenduko ikertzaile eta garatzaile ororen buruhauste izan da noizbait datu eskasia. Zer esanik ez, baliabide urriko hizkuntzekin lanean dihardugunontzat. Erabilera kasu batzuetarako hizkuntza eta domeinuan behar adina testu lortzea nekeza izan daiteke (Joshi et al., 2020b). Zer esanik ez, testu anotatua behar dugun kasuetan (Joshi et al., 2020b), datu-multzo egokia sortzeak eskulan handia dakarrenez, anotazio prozesua garestia baita (Snow et al., 2008; Bai et al., 2021).

Hori dela eta, datu eskasia dagoen egoera askotan datu sintetikoak sortzearen edo datu-gehikuntzako tekniken alde egiten da (Sennrich, 2015; Rei et al., 2017; Zhao et al., 2018; Abonizio et al., 2021; Taori et al., 2023). Datu sintetikoak automatikoki sortutako datu artifizialak dira, balizko benetako datuak ordezkatzeko edo osatzeko erabiltzen direnak, eta ikasketa automatikoan erabiltzen dira entrenamendu datu-multzoak handitzeko (Sennrich, 2015; Rei et al., 2017), klase-desorekak zuzentzeko (Rei et al., 2017; Abo-

[14]Bokmål eta nynorsk aldaerak.

nizio et al., 2021) eta alborapenak murrizteko (Zhao et al., 2018) helburua izaten dute. Datu sintetiko horiek sortzeko hainbat modu daude.

Lehenik erregelatan oinarritutako hurbilpenak ditugu. Eskulan asko eskatzen dute, eta ataza bakoitzerako landu behar izaten dira. Adibidez, testu sailkapen atazetan, datu anotatuetatik abiatuz entrenamendu datu gehigarriak sortu daitezke, adibide errealetan hitz ordezkapenak eginez: sinonimoen hiztegiak edo Wordnet bezalako baliabideak erabiliz (Wei and Zou, 2019), kategoria bereko hitzengatik aldatuz (Zhao et al., 2018; Louvan and Magnini, 2020) edo txantiloiak erabiliz (Kafle et al., 2017). Token sailkapen ataza jakin batzuetan, hizlarien disfluentzien detekzioan (Passali et al., 2022) esate baterako, posible da testu hutsetik datu anotatu sintetikoak sortzea; horretarako, esaldi zuzenei zarata sartu dakieke erregela bidez, eta aldatutako tokenei disfluentzia etiketak esleitu. Sekuentziatik sekuentziarako atazetan ere erregelatan oinarritutako datu sintetikoak erabili daitezke helburutzat testu zuzena duten atazetarako adibide bikoteak sortzeko. Horretarako testu zuzenean akatsak eta zarata sartzeko erregelak lantzen dira, adibidez, zuzentzaile gramatikalerako (Yuan and Felice, 2013; Rei et al., 2017).

Itzulpen automatikoan bertan ere atzeranzko itzulpena (*backtranslation*) baliatzen da, helburu hizkuntzako testu elebakarra automatikoki itzuli eta datu sintetikoak sortzeko (Sennrich, 2015). Modu berean, pibote bidezko itzulpena ere erabili daiteke, landu nahi diren bi hizkuntzen arteko zubi gisa hirugarren bat (pibotea) baliatuz, datu sintetikoak sortzeko (Ahmed and Buys, 2024).

Testu sailkapeneko atazetan ikasketa gainbegiraturako datu gutxi ditugunean eskura, datu gutxi horietatik abiatuta parafrasiak sortu daitezke datu gehikuntzarako. Horretarako, parafrasiak sortzeko eredu propioak (Abonizio et al., 2021) edo itzulpen automatikoko sistemak erabili daitezke atzeranzko itzulpena erabiliz (Abonizio et al., 2021). Hala ere, datu sintetikoen kalitate kaxkarragoa medio, horrela sortutako adibide sintetikoak kaltegarri izan daitezke askotan datu erreal kopurua oso oso txikia ez denean (Meyer et al., 2022).

Itzulpen automatikoaren hurbilpenari dagokionez, zuzenean beste hizkuntza batzuetatik datu-multzoak itzultzeko aukera ere baliatu daiteke sailkapenen atazetan (Sun et al., 2015; Amjad et al., 2020) zein sorkuntzako atazetan (Zhu et al., 2019; Bao et al., 2023). Hurbilpen hau hizkuntza-ereduak ebaluatzeko proba-bankuen kasuan ere erabili izan da (Lai et al., 2023; Achiam et al., 2023).

Berriki, hasiak dira datu anotatu sintetikoen sorkuntzan hizkuntza-eredu handiak ere erabiltzen (Ubani et al., 2023; Whitehouse et al., 2023). Bide hori oso emankorra da, nahi dugun hizkuntzan eta domeinuan, nahi dugun atazarako nahi adina datu sortzeko bide ematen duelako. Metodo hau, ataza jakin baterako hizkuntza-eredu txikiak birdoitzeko datu-multzoak sortzeko erabiltzeaz gain, hizkuntza-eredu handien orotarako atazetan entrenatzeko instrukzioak[15] sortzeko ere erabili daiteke, eta nagusiki, eredu handiago edo hobeago batetik txikiago batek ikas dezan erabiltzen da (Taori et al., 2023; Wang et al., 2023).

Baliabide urriko hizkuntzen kasuan, datu sintetikoak aurrentrenamendurako ere erabiltzeko interesa izan genezake. Baliabide urriko hizkuntzan hizkuntza-eredu elebakarraren aurrentrenamendurako datu sintetikoak sortzeko estrategietako bat itzulpen automatikoarena da. Hala ere, bide hori oraindik askorik landu gabe zegoen tesi honetan landu aurretik, lan aipagarri bakarra Lothritz et al. (2022) izanik, non luxenburgerako hizkuntza-eredua entrenatzeko alemanetik itzulitako testua darabilten.

Hizkuntza-ereduak aurrentrenatzeko testu sintetikoa sortzeko itzulpen automatikoaren aukera sakontasunez aztertu gabe zegoelako, 6. Kapituluan hurbilpen horren bideragarritasuna aztertu dugu, euskara helburu-hizkuntzatzat hartuta. Tesi honetan landu ondoren, hurbilpen hau aztertzen duten lan gehiago ere argitaratu dira (Doshi et al., 2024; Wang et al., 2025).

Tesian aurrentrenamendurako datu sintetikoak sortzeko itzulpenaren hurbilpena aztertu geroztik, tesiaren marko tenporalaren ondoren kokatu arren, aipatzea merezi duen beste estrategia bat ere aztertu dute hainbatek (Gunasekar et al., 2023; Abdin et al., 2024a, b; Ben Allal et al., 2024): hizkuntza-ereduek sortutako testu sintetikoa erabiltzea. Hurbilpen hori normalean ereduak destilatzeko[16] helburuarekin garatu ohi da, eta datu sintetikoen aniztasuna (edukietan eta estiloan) behartzean arreta berezia jartzen den arren, betiere eredu berriaren ezagutza datu sintetikoak sortzeko erabilitako ereduen hasierako corpus originalak mugatuko du. Gainera, hizkuntza-ereduek sortutako testua iteratiboki ereduak entrenatzeko erabiltzeak ereduen kolapsora garamatza (Shumailov et al., 2024). Hau benetako arazo bilakatu daiteke etorkizun laburrean, sarean adimen artifizialek idatzitako albiste, blog argitalpen eta elkarrizketak gailentzen badira (Feng et al., 2024).

[15]Ereduak ataza jakin bat burutzera bideratutako giza hizkuntzan egindako eskaerak.

[16]Eredua txikiago bati (ikaslea) handiago baten (irakaslea) irteeretatik haren jokabidea imitatzen irakastea.

Hala ere, baliabide urriko hizkuntzen kasuan, baliabide ugariko hizkuntzetako testuan gordetzen den ezagutza aberatsa hizkuntza txikira ekartzean ezagutza berria sartuko genuke, kolapso arriskurik gabe. Baina, gaur gaurkoz, baliabide urriko hizkuntzan hizkuntza-eredu elebakarrak entrenatzeko datu sintetikoak sortzeko hizkuntza-eredu eleaniztunak erabiltzea aztertzeke dago oraindik. Orain arte, hizkuntzen arteko transferentzia-ikaskuntzarako jorratu den bide arrakastatsuena hizkuntza-ereduak hizkuntzak nahasian ikastea (Lin et al., 2022; Le Scao et al., 2023; Üstün et al., 2024) eta landu nahi den hizkuntzako testuari gainlaginketa aplikatzea (Choi et al., 2024; Etxaniz et al., 2024; Corral et al., 2024) izan da.

BALIABIDE URRIKO HIZKUNTZETARAKO HIZKUNTZA-EREDU NEURONALAK

3. KAPITULUA

BasqueGLUE: Euskarazko Hizkuntza Ulermena Ebaluatzeko Proba-Bankua

Laburpena

Hizkuntzaren Ulermenerako (NLU) teknologiek hobekuntza nabarmenak izan dituzte azken urteotan, eta GLUE bezalako ataza-anitzeko proba-bankuak giltzarri izan dira lortutako hobekuntzak modu sendo eta orokor batean ebaluatzeko. Horrelako proba-bankuak NLUko ataza multzo anitz eta zabalek osatzen dituzte, azaleko sailkapenez haratago, nolabaiteko hizkuntzaren ulermen sakonagoa eskatzen dutenak. Haatik, sortzeko neketsuak eta garestiak dira eta hizkuntzari lotuak, horregatik, hizkuntza gutxi batzuetarako soilik daude eskuragarri. Kapitulu honetan, BasqueGLUE dakargu, euskararako lehen NLU proba-bankua, dagoeneko sortuta zeuden datu-multzoak probestuz sortua, betiere, GLUE eta SuperGLUE eraikitzeko erabilitako antzeko irizpideak jarraituz. Gainera, euskararako artearen egoerako bi hizkuntza-eredu ebaluatu ditugu BasqueGLUEn, oinarri-lerro sendoa sortuz eredu berriak konparatzeko. BasqueGLUE eskuragarri dago lizentzia librepean.

3.1 Sarrera

Transformer arkitekturak (Vaswani et al., 2017) milaka milioi parametrotaraino eskala dezaketen hizkuntza-eredu ahaltsuen familia bati bide eman dio, testu kopuru itzelekin aurrentrenatzen direnak (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). Eredu hauek sekulako emaitzak lortu dituzte hizkuntzaren ulermeneko (NLU) hainbat atazetan transferentzia-ikaskuntzaren bitartez, non lehenik hizkuntza-eredua anotazio gabeko testu huts kopuru erraldoiekin aurrentrenatzen den, eta gero, helburu-ataza jakin batean birdoitzen den, datu anotatu kopuru txikiagoa erabilita.

GLUE (Wang et al., 2018) eta SuperGLUE (Wang et al., 2019) bezalako proba-bankuek asko erraztu dute hizkuntza-eredu neuronalen garapen eta ebaluazioa NLUari dagokionez, ulermen sakonagoa eskatzen duten gaitasunak egokiago neurtzeko aukera emanez. Proba-banku horiek ebaluazio marko bateratu eta sendoa eskaintzen dute, NLUrako sistema orokorrak ebaluatzeko. Hasiera batean, ulermeneko gaitasuna neurtzeko proba-banku gehienak ingeleserako soilik zeuden eskura, baina pixkanaka hasi dira beste hizkuntzetarako horrelako baliabideak plazaratzen (Xu et al., 2020; Le et al., 2020; Wilie et al., 2020; Kakwani et al., 2020; Shavrina et al., 2020; Park et al., 2021).

Kapitulu honetan BasqueGLUE[1], aurkezten da, Euskararako lehen NLU proba-bankua. BasqueGLUE euskarazko NLU tresnen garapenean ekarpen esanguratsua izango delakoan gaude, eta euskararen aurrerapen teknologikoa erraztuko duela uste dugu. BasqueGLUE sortzeko, ingeleserako GLUE eta SuperGLUE proba-bankuak hartu ditugu eredu gisa. Posible den kasuetan, euskararako dagoeneko sortuta diren datu-multzoak erabili ditugu, hauek dagokien ataza formatu zehatzetara egokituz beharrezkoa den kasuetan. Gainera, BasqueGLUEn aurrez publikatu gabeko sei datu-multzo berri plazaratu ditugu. Guztira, euskararen ulermenerako bederatzi ataza biltzen ditu BasqueGLUEk, zailtasun eta domeinu ezberdinetako ataza sorta zabala. GLUE proba-banku originalaren kasuan bezala, entrenamenduko datu-multzoen tamaina aldakorra da, atazen artean ezagutza zenbateraino transferitzen den neurtzeko.

Proba-bankuaz gain, euskarazko artearen egoera finkatzen duen BERTeus (Agerri et al., 2020) aurrentrenatutako hizkuntza-eredua eta datu gehiago-

[1] `https://github.com/Elhuyar/BasqueGLUE`

rekin lan honetan espresuki garatu den ElhBERTeu[2] BERT eredu berria ere ebaluatu ditugu.

Hurrengo atalean, BasqueGLUE nola eraiki den azaltzen da. 3.3. Atalean euskararako bi hizkuntza-eredu elebakar BasqueGLUEn ebaluatu dira, horietako bat lan honetan aurrentrenatu dena. Bukatzeko, ondorioak eta etorkizuneko lana aurkeztu ditugu.

3.2 BasqueGLUEren Diseinua

BasqueGLUE sortzeko ingeleserako GLUE (Wang et al., 2018) eta SuperGLUE (Wang et al., 2019) proba-bankuen diseinua jarraitu dugu. BasqueGLUE NLUko bederatzi atazen inguruan eraiki da, datu-multzoen tamaina, zailtasun eta domeinu ezberdinetako ataza sorta zabala bilduz. BasqueGLUEn ebaluatzean, ereduari metrika automatiko bakar bat esleitzen zaio. Datu-multzoak, test-multzoak barne, publikoki eskuragarri daude, ebaluazio bateratua egiteko kodearekin batera.

Dagoeneko eskuragarri dauden euskararen ulermenerako datu-multzoen artetik hautatutako atazak, posible den heinean, SuperGLUE eraikitzeko jarraitutako irizpideak betetzen dituzte (Wang et al., 2019):

- Atazaren mamia: atazak euskarazko testuak ulertzeko eta arrazoitzeko gaitasuna neurtu behar lukete.
- Atazaren zailtasuna: atazek gaur egungo artearen egoerako sistemen gaitasunen gainetik egon behar lukete, baina betiere unibertsitate ikasketadun euskal hiztun natiboek ebazteko modukoak izan behar dute.
- Ebaluagarritasuna: atazek irteeraren kalitatea neurtzeko, gizakien epaiarekin bat etorriko den metrika automatiko bat izan behar dute.
- Datu publikoak: SuperGLUEn, atazek dagoeneko publikoki eskuragarri izan behar dute entrenamenduko datu-multzoek, berriki sortutako datu-multzoekin egon daitezkeen arriskuak saihesteko. Hala ere, euskararako proba-banku berri honetan datu-multzo berri batzuk sartzea erabaki da, berriki beste helburu eta erabilera batzuetarako pentsatuak,

[2] `https://huggingface.co/elh-eus/ElhBERTeu`

edo dagoeneko anotatuta dauden eta ezagunak diren datu-multzoetatik eratorrita eraiki dira, euskararako anotatutako baliabide eskuragarriak mugatuak direlako. Erabaki hau ataza askotarikoagoak sartzeko xedearekin hartu da, ereduen NLU gaitasunak hobeto ebaluatzeko gai izateko.

- Atazaren formatua: sarrera-irteera formatu sinpleak dituzten atazen alde egin da, ikertzaileek ataza bakoitzerako berariazko arkitektura konplexuak garatzea saihesteko.
- Lizentzia: datu-multzoek publikoki eskuragarri egon behar dute, ikerketa helburuetarako erabilera eta zabalpena baimenduko duen lizentziapean.

3.2.1 Hautatutako Atazak

Atal honetan BasqueGLUEn dauden atazak eta dagozkien datu-multzoak bildu ditugu: 3.1. Taulak horien laburpen bat dakar, datu-multzoen tamaina, ebaluazio metrikak eta testuaren domeinua barne. 3.2. Taulan BasqueGLUEko ataza bakoitzeko adibide esanguratsu bana bildu dugu.

3.2.1.1 Entitate Izendunen Ezagutza

BasqueGLUEn sartzea erabaki dugun lehen ataza, entitate izendunen ezagutza da (*Name Entity Recognition* (NER)), hizkuntzaren prozesamenduko token sailkapen ataza ondo ezaguna.

NERerako datu-multzoa bi azpiatazetan banatuta dago: domeinu barneko NER (*in-domain*) eta domeinuz kanpoko NER (*out-of-domain*). Ebaluazio metrikari dagokionez, bi azpiatazen F1 balioen batez bestekoa hartu da puntuazio bakar modura.

Domeinuko NER azpiatazarako (NER_{id}) EIEC (Alegria et al., 2004) izan da orain arte NER euskaraz lantzeko lehenetsitako datu-multzoa. Besteak beste BERTeus (Agerri et al., 2020) eredua ebaluatzeko erabili zen. EIEC egunkarietako albiste testuek osatzen dute, eta BIO anotazio eskema jarraituz anotatua dator, honako lau kategoriekin: *person* (pertsona), *organization* (erakundea), *location* (lekua), eta *miscellaneous* (bestelakoak). Datu-multzo handiagoa izateko helburuarekin, EIEC anotazio eskema berbera jarraitzen

Datu-multzoa	Train	Dev	Test	Ataza	Metrika	Domeinua
NER_{id}	51.539	12.936	35.855	NER	F1	Albisteak
NER_{ood}	64.475	14.945	14.462			Albiste+Wiki
$FMTODeu_{intent}$	3.418	1.904	1.087	Asmoen sailkapena	F1	Elkarrizketa sistemak
$FMTODeu_{slot}$	19.652	10.791	5.633	Hutsune betetzea	F1	Elkarrizketa sistemak
BHTCv2	8.585	1.857	1.854	Gaien sailkapena	F1	Albisteak
BEC2016eu	6.078	1.302	1.302	Sentimenduen analisia	F1	Twitter
VaxxStance	864	206	312	Jarrera detekzioa	MF1*	Twitter
$QNLI_{eu}$	1.764	230	238	QA/NLI	Acc	Wikipedia
WiC_{eu}	408.559	600	1.400	Hitzen adiera desanbiguazioa	Acc	Wordnet
EpecKorrefBin	986	320	587	Korreferentzia ebazpena	Acc	Albisteak

3.1. Taula: BasqueGLUEn sartutako 9 atazak. NER_{id} domeinuko NERi dagokio, eta NER_{ood} domeinuz kanpoko NER azpiatazari. Acc-k zehaztasuna adierazten du; F1ek micro F1 puntuazioa; MF1*ek, FAVOR eta AGAINST kategorien batezbesteko macro F1 puntuazioa da.

duen Naiz[3] euskarazko egunkariko testua duen datu-multzo berri batekin elkartu dugu. Bi datu-multzoak nahasi, eta entrenamendu, garapen eta ebaluaziorako datu-multzo sorta berriak sortu ditugu bi iturrien arteko oreka mantenduz.

Domeinuz kanpoko NER azpiatazarako (NER_{ood}) entrenamendurako datu-multzoak albisteen domeinuko datuak biltzen ditu. Garapeneko eta ebaluaziorako datu-multzoak, berriz, Wikipediako artikuluez osatuta daude. Zehazki, entrenamenduko datu-multzorako, domeinuko NER azpiatazako entrenamendu eta garapen datu-multzoak elkartu ditugu. Domeinuz kanpoko NER ebaluatzeko, Wikipediako artikulu multzo bat anotatu da EIECen jarraibide berberak erabiliz. Wikipediako datu-multzo berri honetatik, BasqueGLUEko NER_{ood} atazaren garapen eta ebaluazio azpimultzoak sortu ditugu.

[3] `https://naiz.eus`

3.2.1.2 Asmoen Sailkapena (FMTODeu$_{intent}$)

BasqueGLUEn sartu dugun hurrengo ataza asmoen sailkapena (*intent classification*) da, elkarrizketa sistemen alorreko ulermenen ataza bat, eta bertan, erabiltzailearen asmoa asmatu behar da aurrez ezarritako kategoria jakin batzuen artean. Beraz, kategoria-anitzeko testu sailkapen moduan lantzen da. Atazarako aukeratutako datu-multzoa *Facebook Multilingual Task Oriented Dataset for Basque* edo FMTODeu (López de Lacalle et al., 2020) litzateke. Adibideak alarmari, gogorarazpenei eta eguraldiari lotutako 12 asmo kategoria ezberdinetan daude banatuta. Datu-multzoa FMTODeu$_{intent}$ izendatuko dugu proba-banku honetan, datu-multzo berdina hutsune betetze ataza lantzeko ere erabiltzen baita (ikus (3.2.1.3. Atala). Jatorrizko entrenamendu/garapen/ebaluazio banaketa mantendu dugu eta Micro-F1ekin ebaluatu.

3.2.1.3 Hutsune Betetzea (FMTODeu$_{slot}$)

Hirugarren ataza ere elkarrizketa sistemen alorretik datorkigu, eta asmoen sailkapenarekin batera landu ohi den hutsune betetzea (*slot filling*) da. Honen xedeak, erabiltzaileek esaldietan adierazitako asmoei dagozkien entitateak identifikatzean datza. FMTODeu (López de Lacalle et al., 2020) datu-multzoak dagoeneko entitate horiek anotatuta dakartza, beraz, zuzenean erabiltzeko moduan dago. Ataza token sailkapen ataza bat da, NER atazaren antzekoa egituraz, eta 11 kategoria dakartza, BIO etiketatze eskema jarraituz: *alarm/alarm_modifier*, *alarm/recurring_period*, *datetime*, *location*, *negation*, *reminder/noun*, *reminder/recurring_period*, *reminder/reference*, *reminder/todo*, *weather/attribute* and *weather/noun*. FMTODeu$_{slot}$ izendatu dugu datu-multzoa, eta asmoen sailkapenean egindako moduan, jatorrizko entrenamendu/garapen/ebaluazio banaketa mantendu dugu. Ebaluazioan darabilgun metrika Micro-F1 da.

3.2.1.4 Gaien Sailkapena (BHTCv2)

Gaien sailkapena kategoria-anitzeko testu sailkapen ataza bat da. BasqueGLUErekin zabaldu dugun datu-multzoa BHTC datu-multzotik (Agerri et al., 2020) dator. Argia[4] euskarazko astekariko albisteen izenburuak dakartza, hamabi kategoria tematikotan sailkatuak: Ekonomia, Euskal_Herria, Euskara,

[4] `https://www.argia.eus`

Gizartea, Historia, Ingurumena, Iritzia, Komunikazioa, Kultura, Nazioartea, Politika eta Zientzia. Datu-multzoa hau, jada, erabili zen BERTeus (Agerri et al., 2020) ebaluatzerakoan. Jatorrizko BHTCko adibideak letra xehez eta puntuazioa kenduta zetozen, eta testu originala berreskuratzea erabaki dugu BasqueGLUErako. Gainera, errepikatutako adibideak eta euskaraz ez zeudenak baztertu ditugu. Hemendik aurrera BHTCv2 deituko diogu datumultzoaren bertsio berri honi. Asmoen sailkapenean bezala, ereduak Micro-F1 metrikarekin ebaluatuko ditugu.

3.2.1.5 Sentimentuen Analisia (BEC)

Sentimenduen analisia NLUko proba-banku gehienetan agertu ohi den ataza ezaguna da. Testu bat emanik, honen polaritatea ebaztea du xede, gure kasuan aldekoa, neutrala edo kontrakoa kategorien artean.

The Basque Election Campaign 2016 Opinion Dataset (BEC2016eu) edo 2016ko Euskal Hauteskunde Kanpainako Iritzi Datu-Multzoa, euskarazko sentimenduen analisia atazarako datu-multzo berri bat da. Testu sailkapeneko ataza honek, 2016ko EAEko hauteskundeetako kanpainari buruzko txioak biltzen ditu. Txio hauen bilketa, hauteskunde kanpainan zehar[5] burutu zen, alderdi nagusiei eta horien hautagaiei segimendua eginez. Txioak eskuz anotatuak izan ziren alderdien edo hautagaien aldeko, kontrako edo neutral gisa. Datu-multzo hau ebaluatzeko, micro-F1 metrikaren alde egin dugu.

3.2.1.6 Jarrera Detekzioa (VaxxStance)

Jarrera detekzioa (*Stance Detection*) testu sailkapen gisa lantzen den *Fake News*en detekzioaren arloko ataza bat da. Sare sozialetan harrabotsa sortzen duen gai polemiko batekiko erabiltzaileen jarrera antzematea du helburu. Txio batek, gaiarekiko aldeko, kontrako edo jarrera neutroa azaltzen duen erabakitzea, alegia.

VaxxStance (Agerri et al., 2021) datu-multzoak txertoen aurkako mugimenduari lotutako euskarazko txioak dakartza. Datu-multzoak ez zekarren garapenerako sortarik, eta beraz, jatorrizko entrenamendurako sorta banatu dugu gure entrenamendurako eta garapenerako datu-multzo berriak sortzeko. Banaketa hori ausaz egin da, eta ereduen ebaluazioa eta konparaketa

[5] 2016/09/09-2016/09/23

bidezkoagoa eta sendoagoa izatea du jomuga. VaxxStance datu-multzoaren sortzaileek darabilten jarrera detekzioaren literaturan proposatutako metrika bera erabili dugu hemen, bi klaseren macro F1 (MF1) balioa: alde (FAVOR) eta kontra (AGAINST)[6].

3.2.1.7 QNLI

Eraiki dugun proba-bankuan galdera-erantzunen (QA) ataza bat sartzearren, euskal hiztun natibo boluntarioek sortutako galdera-erantzunen ElkarHizketak (Otegi et al., 2020) datu-multzoa egokitu dugu. Datu-multzoa pertsona eta erakunde ezagunen Wikipediako atalen gainean eraiki da, eta 400 elkarrizketa eta 1600 galdera-erantzun bikote ditu.

Datu-multzo hori sekuentzia-bikote sailkapen bitar batera egokitu dugu, ingeleserako QNLI (Wang et al., 2019) datu-multzoaren diseinua jarraituz. Bikoteak sortu ditugu galdera bakoitza eta testuinguru horretako esaldiak elkartuz; galderaren erantzuna esaldian badator, adibide positiboa da, bestela negatiboa. Ondoren, adibide negatiboak galderaren eta esaldiaren arteko gainezartze lexikal baxuena dutenak baztertzen joan gara datu-multzo orekatua lortu arte. Ebaluazio metrika gisa, zehaztasuna erabili dugu, ingeleseko QNLIn egin bezala.

3.2.1.8 WiC

Hitza testuinguruan edo WiC (*Word in Context*) (Pilehvar and Camacho-Collados, 2019) Ingeleseko SuperGLUE proba-bankuan dagoen hitzen adiera-desanbiguazioko (*word sense disambiguation*) ataza bat da. Esaldi-bikote sailkapen bitar modura dago diseinatua. Bi esaldi eta bietan ageri den hitz polisemiko bat emanik (hitzaren mugak markatuta datoz), hitzak bi kasuetan esanahi edo adiera berbera duen erabakitzea izango litzateke ataza. Metrika gisa, zehaztasuna hartu dugu.

EPEC-EuSemcor (Pociello et al., 2011) corpusetik abiatuz euskarazko WiC datu-multzo bat sortu dugu. EPEC-EuSemcor adierak anotatuta dituen euskarazko corpus bat da. Corpuseko izenak euskarazko WordNeteko (v1.6) adierekin (Pociello et al., 2011) daude aberastuak.

[6]Nahiz eta entrenamenduan eta iragarpenean hiru kategoriak erabili.

42.615 izenen agerpenen adierak ditu eskuz anotatuak. Hauek euskarazko 407 izen maizenei dagozkie. 10 eta 50 hitz arteko testuingurua duten izenen agerpenak soilik baliatu ditugu.

Euskarazko WiC datu-multzoak ingeleseko WiC datu-multzoaren (Pilehvar and Camacho-Collados, 2019) diseinua jarraitzen du, eta izen bakoitzaren testuinguru esaldiak binaka elkartu ditugu sailkapen atazarako adibideak sortzeko.

Ingeleserako egin bezala, WordNet grafo semantikoan lehen graduko ahaidetasuna duten adierak (ahizpa-adierak barne) dituzten bikoteak eta superadiera berari dagozkion adierak dituztenak baztertu ditugu datu-multzoaren argitasuna bultzatzeko. Kimatze honetarako, Wordnet v1.6tik v3.0rako mapaketa bat erabili dugu, adiera-desanbiguazioaren arloko datu-multzoetan ohikoa den bezala (Raganato et al., 2017).

Horretaz gain, garapenerako eta ebaluaziorako azpimultzoetan testuinguru esaldiak ez errepikatzea bermatu dugu. Honenbestez, 600 eta 1.400 adibide bereizi ditugu garapen eta ebaluazio multzorako hurrenez hurren.

Gainerako bikoteen artean, testuinguru esaldiak garapenen eta ebaluaziorako multzoetakoekin gainjartzen ez direnekin osatu dugu entrenamendurako datu-multzoa. Bertan, testuinguru-esaldiak bikote ezberdinen artean errepikatzea baimendu dugu, entrenamendu multzoa puzteko. Azkenik, azpimultzo guztiak adibide positibo eta negatiboen artean orekatuta daudela bermatu dugu.

3.2.1.9 Korreferentzia Ebazpena (EpecKorrefBin)

BasqueGLUE proba-bankurako aukeratu dugun azken ataza korreferentzia ebazpena da, eta euskararako EPEC-KORREF (Soraluze et al., 2012) datu-multzoa dugu jada eskura. Hala ere, korreferentzia ebazpenak aipamenak entitateetan multzokatzea dakar, eta hau ez dago sarrera-irteera formatu sinple batekin bideratzerik. Honenbestez, ataza testu sailkapen bitarreko ariketa bihurtu dugu, *Winograd Schema Challenge* (WSC) (Levesque et al., 2012) atazaren formatura moldatuz.

Ataza berri honetan, izenordainak, izenak edo izen-sintagmak izan daitezkeen testu bateko bi aipamen emanik, entitate berari ote dagozkion ebatzi behar du ereduak. EPEC-KORREF moldatu dugu ataza honetara, eta datu-multzo berri hau EpecKorrefBin izenez izendatu dugu. Adibide bakoitzeko

NER

Tokenak:	McLareni	eta	Ferrariri	aurre	egitea	izango	du	Helburu
Etiketak:	B-ORG	O	B-ORG	O	O	O	O	O

Asmoen Sailkapena (FMTODeu$_{intent}$)

Testua: Alarma ezarri gaurko 6:00etan
Set the alarm today at 6:00am
Asmoa: `alarm/set_alarm`

Hutsune Betetzea (FMTODeu$_{slot}$)

Tokenak:	Euria	egingo	du	gaur	?
Etiketak:	B-weather/attribute	O	O	B-datetime	O

Gaien Sailkapena (BHTCv2)

Testua: Gurasotasun baimena eta seme-alabak zaintzeko baimena lau hilabetera luzatzeko proposamena egitea onartu du Europako Batzordeak. Proposamenak aldaketa handia ekarriko luke Hego Euskal Herrian, lau asteetara luzatu berri baita baimen hori.
Gaia: `Gizartea` (*society*)

Sentimenduen Analisia (BEC)

Testua: Mezu txoro, patetiko eta lotsagarri hori ongi hartuko duenik badela uste du PSEk.
Polaritatea: `Negative`

Jarrera Detekzioa (VaxxStance)

Testua: Gure nagusiak babestuko dituen txertoa martxan da.
Zor genien. Gaur mundua apur bat hobeagoa da.
#OsasunPublikoarenGaraipena #GureGaraipena
Jarrera: `FAVOR`

QNLI

Galdera: "Irrintziaren oihartzunak" dokumentalaz gain, zein best lan egin ditu zinema arloan?
Esaldia: "Irrintziaren oihartzuna" du lehen filma zuzendari eta gidoilari gisa.
NLI: `not_entailment`

WiC

Esaldi1: Asterix, zazpi egunen segida asmatu zuen galiarra.
Esaldi2: Etxeko landareek sasoi aktiboan tenperatura epelak behar dituzte: egunez 25C inguru.
Adiera_bera: `False`

Korreferentzia Ebazpena (EpecKorrefBin)

Testua: Birmoldaketan daudenen artean Katalunia , Madril , Hego Euskal Herria , Aragoi , Balear irlak eta Errioxa aurkitzen dira . Horien artean , Hego Euskal Herriak 47.870 milioi pezeta jasoko ditu .
Korreferentzia: `True`

3.2. Taula: BasqueGLUEko ataza bakoitzeko adibideak. Azpimarratutako testua markatuta dator datu-multzoan. Monospace letra-tipodun testuak ereduaren irteeran esperotako emaitza adierazten du.

aipamenen agerpenak esaldi berera edo ondoz ondoko esaldietara mugatu ditugu.

Adibide negatiboak sortzeko, aipamen mota bera (adib. biak leku izenbereziak izatea) dituzten aipamen bikoteak edo bietako bat izenordaina dutenak aukeratu ditugu. Jarraian, atazaren erronka zaildu asmotan, aipamen bikote positiboen artean forma antzekoena zutenak, eta adibide negatiboen artean formaz ezberdinak zirenak baztertu ditugu. Horretarako, stringen antzekotasuna neurtzeko *Levenshtein* distantzia erabili dugu esaldi positiboetarako, eta *token-set-ratio* negatiboen kasuan. Azkenik, EpecKorrefBinen hiru azpimultzoak orekatu ditugu adibide positibo eta negatiboei dagokienez.

3.3 Ebaluazioa

3.3.1 Oinarri-Lerroak

Oinarri-lerro gisa bi eredu konparatu ditugu. Bata BERTeus (Agerri et al., 2020) da, euskarazko albisteekin eta Wikipediako testuekin (224M hitz) aurrentrenatutako euskararako lehen BERT elebakarra. Bestea, berriz, lan honetan aurrentrenatu dugun BERT eredu bat da, aurrekoak darabilena baino corpus elebakar handiago eta eguneratuago batekin entrenatu da, eta ElhBERTeu deitu diogu.

ElhBERTeu entrenatzeko, BERTeus entrenatzeko erabilitako BMC corpusa (Agerri et al., 2020) handitu dugu. Zehazki, 2020 eta 2021 arteko edukiekin eguneratu ditugu dagoeneko BMCn dauden iturrietako edukiak, eta, gainera, albiste iturri berriak eta bestelako domeinuetako testuak bildu ditugu, besteak beste, zientzia (akademikoa zein dibulgazioa), literatura eta azpidatziak. Corpusaren osaketa eta iturri bakoitzetik bildutako testuaren tamaina 3.3. Taulan daude ikusgai. Albisteetako testuei gainlaginketa aplikatu zaie (bikoiztuz), BERTeus entrenatzean egin bezala. Guztira, 575M hitz erabili dira ElhBERTeu aurrentrenatzeko.

ElhBERTeu BERTeusen (Agerri et al., 2020) entrenamendu diseinua jarraituz aurrentrenatu da. Tokenizatzailean eta hiperparametroetan hartutako erabakiak mantendu dira, ezberdintasun batekin, ereduaren aurrentrenamendu osoa (1M pausu) 512ko sekuentzia luzerarekin egin da TPU v3-8 batean.

Domeinua	Tamaina
Albisteak	2 * 224M
Wikipedia	40M
Zientzia	58M
Literatura	24M
Bestelakoak	7M
Guztira	575M

3.3. Taula: ElhBERTeu aurrentrenatzeko erabilitako corpusa. Tamainak tokenetan. Albiste iturrietako testuak gainlaginketa dute (bikoiztuta).

Oinarri-lerroak inplementatzeko ereduetako bakoitza ataza bakoitzean independenteki birdoitu dugu, hasierako 3e-5 ikasketa-tasarekin, 10 epokaz asko jota, garapeneko datu-multzoan kontrol-punturik onena aukeratuz. WiC eta korreferentzia ez beste ataza guztietarako, *Transformers* liburutegia (Wolf et al., 2020b) erabili dugu ereduak birdoitzeko, eta token sailkapen eta testu sailkapenerako prozedura estandarrak segi ditugu. Esaldiak [SEP] tokena tartekatuta kateatu dira sarrera esaldi bikote formatuan jasotzen duten atazen kasuan.

WiC eta korreferentziarako, berriz, *jiants* tresna (Phang et al., 2020) baliatu dugu ereduak birdoitu eta ebaluatzeko, SuperGLUEn egin bezala. WiC atazari dagokionez, azpimarratutako hitzaren errepresentazioa [CLS] tokenaren errepresentazioari kateatu zaio. Korreferentzia ebazpenaren kasuan, ataza leihoetan oinarritutako ataza gisa landu dugu eta horretarako *jiants*ek WSC atazarako dakarren inplementazioa hartu dugu.

3.3.2 Emaitzak

Bi ereduen emaitzak 3.4. Taulan daude ikusgai. NER atazaren kasuan, NER_{id} eta NER_{ood} azpiatazen batez besteko F1 balioa bakarrik erakusten du, 3.2.1.1. Atalean aipatu bezala[7]. Hori horrela egitea erabaki da, azken batez besteko puntuazioa kalkulatzean ataza bati besteei baino pisu gehiago ematea saihesteko heburuarekin.

[7]BERTeusek 81,92ko F1a du NERen, NER_{id} azpiatazako 86,40 F1a eta NER_{ood}ko 77,44 F1a. ElhBERTeuk 82,30eko F1a NERen, NER_{id}eko 86,81 F1 eta NER_{ood}eko 77,78 F1en batez bestekoa.

Eredua	AVG	NER F1	F_{intent} F1	F_{slot} F1	BHTC F1	BEC F1	Vaxx MF1	QNLI acc	WiC acc	coref acc
BERTeus	73,23	81,92	**82,52**	74,34	**78,26**	69,43	59,30	**74,26**	70,71	**68,31**
ElhBERTeu	**73,71**	**82,30**	82,24	**75,64**	78,05	**69,89**	**63,81**	73,84	**71,71**	65,93

3.4. Taula: Euskararako BasqueGLUE hizkuntza ulermeneko proba-bankuan ebaluatu ditugun BERTeus (Agerri et al., 2020) eta lan honetan entrenatutako ElhBERTeu ereduen emaitzak. NERen lortutako puntuazioa domeinuko eta domeinuz kanpoko azpiatazen batez bestekoa da. F_{intent} eta F_{slot}, $FMTODeu_{intent}$ eta $FMTODeu_{slot}$ datu-multzoei dagozkie, hurrenez hurren; BHTC, BEC, Vaxx eta coref, berriz, BHTCv2, BEC2016eu, VaxxStance, eta EpecKorrefBin datu-multzoei.

Atazei banaka begiratuz gero, emaitza kontraesankorrak ditugu. BERTeus korreferentzian gailentzen da, eta ElhBERTeuk F_{slot}, WiC eta bereziki VaxxStance atazetan irabazten du. Gainerako datu-multzoetan bi ereduen portaerak antzekoak dira, BERTeusek $FMOTD_intent$, BHTCv2 eta QNLIn irabazten du eta ElhBERTeuk NER and BECen aurrea hartzen dio. Oro har, proba-bankuaren batez besteko azken puntuazioan, puntu erdiko aldea erdietsiz, BasqueGLUEren arabera eredurik onena ElhBERTeu dela esan genezake euskarazko NLU atazetan.

Eraiki dugun proba-bankuaren egokitasunari dagokionez, emaitzek argi uzten dute oraindik ere hobekuntzarako tartea badagoela ataza guztietan, BasqueGLUE artearen egoerako hizkuntza-ereduentzat erronka bat dela frogatuz. GLUE argitaratu zenean oinarri-lerro ereduek 70 puntuko batez bestekoa eskuratzen zuten, SuperGLUEn, berriz, 74,6koa, ElhBERTeuk BasqueGLUEn lortutako 73,71 puntutik hurbil.

3.4 Ondorioak

Lan honetan BasqueGLUE aurkeztu dugu, hizkuntzaren ulermena ebaluatzeko euskararako lehen proba-bankua. Baliabide hau aurrentrenatutako hizkuntza-eredu neuronalen euskararen ulermen gaitasunak osorik eta sendo ebaluatzeko ezinbesteko tresna izango da. Proba-bankuak nolabaiteko hizkuntzaren ulermena eskatzen duten ataza sorta zabal eta askotarikoa biltzen du. BasqueGLUEko datu-multzoak aurrez existitzen ziren baliabideetatik eraikiak izan dira, batzuen kasuan atazen formatua aldatu edo sinplifikatzeko beharra ikusi delarik, GLUE eta SuperGLUEren diseinua segiz.

Bestalde, sortu berri den proba-bankuan euskararako entrenatutako bi BERT hizkuntza-eredu ebaluatu ditugu. Azkenik gai gara ereduak elkarrekin sakontasunez konparatzeko euskararen ulermenean, eta bi eredu horien artean, BasqueGLUEren arabera, onena ElhBERTeu dela ondorioztatu dugu.

BasqueGLUE publikoki eskuragarri dago, lizentzia irekipean zabaldua, ebaluazio bateratua aurrera eramateko kodea barne. ElhBERTeu eredua ere eskuragarri dago *Huggingface*en.

4. KAPITULUA

BERTen Eskala-Legeak Baliabide Urriko Inguruneetan

Laburpena

Hizkuntza-eredu handiek baliabide eskakizun ikaragarriak eta entrenamendu datu eskakizun aseezinak dituzte, arloko ikertzaile eta garatzaile gehienen eskueratik kanpo. Aurreko lanek, horrelako ereduen eskala-legeak ikertu dituzte, baina ereduaren parametroen, datu-multzoen tamainaren eta kostu konputazionalaren arteko ratio optimoek eskalaren goiko ertzean bakarrik dute jarria arreta. Aitzitik, aldagai horiek ingurune mugatuetan hizkuntza-ereduengan duten efektua aztertu dugu guk, horretarako hiru BERT eredu arin (16M/51M/124M parametro) entrenatuz, corpus txikiekin (5M/25M/125M hitz). Ezaugarri linguistiko ezberdinetako lau hizkuntzekin (euskara, gaztelania, swahilia eta suomiera) jardun dugu gure esperimentuetan eta ereduak MLMan eta NLUko hainbat atazetan ebaluatu ditugu. Baliabide urriko inguruneetan kodetzaile motako hizkuntza-eredu diskriminatzaileen parametroen, datuen eta kalkulu-ahalmenaren arteko erlazio optimoak zehaztu ditugu. Gure aurkikuntzak sendoak dira aztertu diren hizkuntza guztietan zehar, eta MLMn eta helburu-atazetan zehar ere bai. Gainera, esperimentalki finkatu dugu ea noiz merezi duen Transformerren alde egiteak, kalkulu-ahalmen arinagoko soluzioak bazter utziz.

4.1 Sarrera

Transformer arkitekturan oinarritutako aurrentrenatutako hizkuntza-eredu neuronalek emaitza ikusgarriak erakutsi dituzte hizkuntzaren prozesamenduko hainbat atazetan, ataza hauek ebazteko bide ohikoena bilakatzeraino. Eredu horien gaitasuna hobetzen doa, arkitekturaren konplexutasuna (parametro kopuruan) (Wei et al., 2022) eta aurrentrenamenduko corpusaren tamaina igo ahala (Zhang et al., 2021). Arrazoi horrengatik, inoizko datu gehienetan entrenatutako inoizko eredu handienak sortzeko joera dago gaur egun. Lasterketa espazial tankera duen joera honek, amaituko ez dela dirudien kalkulu-ahalmenaren igoera dakar berarekin, eta ez soilik ereduen aurrentrenamenduan, ondorengo birdoitzean eta produkzio garaiko inferentzian ere bai. Gainera, eredu oso handiak sortzeak, izugarrizko entrenamendu corpusen beharra du, eta hau, baliabide oparoko hizkuntza gutxi batzuen esku soilik dago.

Kaplan et al. (2020) eta Hoffmann et al. (2022) lanek modeloaren tamaina, corpusaren tamaina eta kalkulu-ahalmena erlazionatzen dituzten berreturа-legeen formulak proposatzen dituzte, kalkulu-ahalmen aurrekontu jakin bat emanik, konfigurazio optimoak topatzen laguntzeko.

Hala ere, euren analisia entrenamendurako corpus erraldoiak darabilten tamaina handi samarreko eredu autorregresiboetan zentratzen da. Kapitulu honetan, hizkuntza-eredu diskriminatzaileen (kodetzaileen) eskala-legeetan jarri dugu fokua, betiere baliabide urriko inguruneetan, non, datu eskuragarriak eta kalkulu-ahalmenerako aurrekontua mugatuak diren. Hainbat eredu eta datu tamainaren arteko konbinazioen emaitzak aztertu ditugu, eta horretarako, hizkuntza familia bereizietako ezaugarri linguistiko ezberdinetako lau hizkuntza (euskara, gaztelania, swahilia eta suomiera) aukeratu ditugu, baliabide-urritasuna simulatuz.

Egindako esperimentuen bidez, datuen eta ereduaren tamainen arteko oreka aztertu eta erlazio optimoak (4.5. Atalean laburbilduak) zehaztu dira, 16M eta 124M parametro arteko eredu txikiak optimoki entrenatzeko. Halaber, ikusi dugunez, kalkulu-ahalmen finko bat emanik, hobe da eredu handiago bat entrenatzea, eredu txikiago bat luzaroago entrenatzea baino.

Gainera, oinarri-lerro neuronal arinagoei Transformerretan oinarritutako hurbilpenak gailentzeko beharrezko gutxieneko kalkulu-ahalmen, eredu tamaina eta datu kopuruak finkatu ditugu, CO_2 isurketak kontuan izanik.

4.2 Esperimentuen Diseinua

Lan honetako esperimentuen bitartez, ereduaren tamaina, datu-multzoaren tamaina eta kalkulu-ahalmenaren arteko konbinazio optimoa topatu nahi da baliabide urriko testuingururako. Horrez gain, oinarri-lerro neuronal arinagoak gainditzeko eskakizun minimoak zein diren aztertu nahi dugu. Honenbestez, gure esperimentuak hiru corpus tamainarekin, hiru eredu tamainarekin eta lau hizkuntzatan burutu ditugu, guztira 36 eredu ezberdin entrenatuz.

4.2.1 Hizkuntzen Hautapena

Hizkuntza familia ezberdinetako lau hizkuntzetan gauzatu ditugu esperimentuak, hizkuntza-ereduak aurrentrenatzeko nahiko corpus elebakar dutenen artean, eta aldi berean NLUko atazetan ebaluatzeko nahikoa datu-multzo eskuragarri dutenen artean aukeratuz. Ondorioz, baliabide-urriko ingurunea simulatua izan da kasu batzuetan. Irizpide horiek betetzen dituztenen artetik, euskara (eu), gaztelania (es), swahilia (sw) eta suomiera (fi) hautatu ditugu. Hizkuntza familia disjuntuen parte izateaz gain, hizkuntza hauek linguistikoki askotarikoak dira (Coloma, 2015) konplexutasun maila ezberdinekin morfologia, sintaxia, aditz sistema eta hiztegiari dagokionez (ikus 4.1. Taula).

Hizkuntza	**Familia**	**Morf.**	**Sint.**	**Aditz**	**Hiztegi**
Euskara	bakartua	0.73	0.58	0.77	0.62
Gaztelania	erromantzea(indoeu)	0.64	0.42	0.62	0.69
Swahilia	bantu (niger-kongo)	0.64	0.42	0.54	0.31
Suomiera	uraldarra	0.82	0.42	0.46	0.31

4.1. Taula: Aukeratutako lau hizkuntzak, dagokien hizkuntza familia eta bakoitzaren morfologia, sintaxia, aditz sistema eta hiztegiaren konplexutasun maila.

4.2.2 Corpusak

Hizkuntza bakoitzerako hiruna corpus ezberdin sortu ditugu, 5M, 25M eta 125M hitzetakoa bakoitza. Hiru tamaina ezberdinetara mugatu da, esperimentuen kopurua eta kalkulu-ahalmen eskakizunak izugarri igo ez zitezen.

Burutu ditugun atariko esperimentuek erakutsi dute nola aurrentrenamenduko datu kopurua 1M hitzetara mugatuz, emaitzetan galera handia hautematen den, Zhang et al. (2021) lanaren emaitzek erakusten duten moduan. 5 miloi hitzeko corpusa lortzea anotatutako datu-multzoak dituzten hizkuntza gehienen eskura badaudenez dagoeneko (Joshi et al., 2020b), behe bornea 5Metan finkatu dugu. Zhang et al. (2021) lanak erakutsi duenez, 10M eta 100M hitz artean nahikoa dira hizkuntza-eredu batek ingelesezko sintaxia eta semantikaren gaitasun linguistikoak barneratzeko. Beraz, beste bi corpus tamainak 25M eta 125M hitzetan finkatu ditugu, igoera ratio konstantea mantenduz hiruren artean.

Testuen jatorriei dagokionez, euskarazko eta gaztelaniazko corpusak %75-ean albisteez eta %25ean Wikipediako testuez daude osatuak. Berria[1] egunkaria aukeratu dugu euskararako, eta El Pais[2] gaztelaniarako. Swahilirako eta suomierarako corpusak, berriz, *cc100* (Conneau et al., 2020; Wenzek et al., 2020) sareko corpusetik dokumentuak ausaz[3] hautatuz eraiki ditugu.

4.2.3 Ereduak

Turc et al. (2019) lanean egindakoa jarraituz, BERTaren hiru aldaera ezberdin erabili ditugu, BERT_{124M}[4], BERT_{51M} eta BERT_{16M} deritzogunak.

Ereduek 12, 8 eta 4 geruza dituzte, gainerako parametroak ere proportzionalki uzkurtuz (ezkutuko dimentsioak, atentzio-buru kopurua, etab.), ereduaren formak ez baitu emaitzetan eragin nabarmenik (Kaplan et al., 2020). 4.2. Taulan eredu bakoitzaren parametroak zehatzago nola antolatu diren ikus daiteke. Hiztegiaren tamaina ere 30K hizpeko tokenetik 50Kra igo dugu, hizkuntza eranskarietan lagungarria delako (Agerri et al., 2020).

4.2.3.1 Aurrentrenamenduko Xehetasunak

Tokenizatzailerako 50K hizpeko tokeneko hiztegi bat erabili da, letra larri eta xeheak bereiziz (*cased*) eta Kudok (2018) proposatutako unigrametan oinarritutako hizpeko segmentazio algoritmoa erabiliz entrenatua. Hiztegiak

[1] `https://www.berria.eus`
[2] `https://elpais.com`
[3] 10 esaldi baino luzeagoen artean.
[4] Hau Devlin et al. (2019) laneko BERT_{base}ari dagokio.

	Geruzak	HH	INT	Buru	NEP	Parametroak
BERT_{124M}	12	768	3072	12	86M	124M
BERT_{51M}	8	512	2048	8	25M	51M
BERT_{16M}	4	256	1024	4	3M	16M

4.2. Taula: Gure esperimentuetako eredu tamainak: geruza (*layers*). HH: ezkutuko dimentsio (*hidden dimensions*). INT: tarteko geruza dimentsio (*intermediate layer dimensions*). Buru: atentzio-buru (*attention heads*). NEP: hitz-bektore ez besteko parametroak (*non-embedding parameters*).

aurrentrenamenduko corpus bakoitzarekin ikasi dira, kasuan kasu, %99,95eko karaktere estaldurarekin, karaktere arraroak baztertze aldera. Honenbestez, hiru hiztegi sortu dira hizkuntza bakoitzerako, aurrentrenamenduko corpus tamaina bakoitzarekin bana (5M, 25M eta 125M). Tokenizatzaile edo hiztegi hauek elkarbanatu egiten dira esperimentu ezberdinetan entrenatutako tamaina ezberdineko hizkuntza-ereduen artean.

Sortutako ereduek 124M, 51M eta 16M parametro dituzte hurrenez hurren. Hitz osoen maskaratzea aplikatu da maskaratutako hizkuntza-eredua (*MLM*) entrenatzean (10eko dup-faktorea erabiliz[5]). Tamaina orotako ereduak defektuzko (Devlin et al., 2019) hiperparametroak erabiliz entrenatu dira: $1e^{-4}$ ikasketa-tasa, 0,9ko $\beta 1$, 0,999ko $\beta 2$, 0,01eko L2 pisuen gainbehera, 10K urratseko beroketa eta eredu bakoitza 500K urratsez entrenatu dugu, 256ko labekada tamainarekin eta 512ko sekuentzia luzerarekin. Honenbestez, datu-multzo beraren gainean epoka asko egingo ditugu (500/2500/12500), testu bera behin eta berriro ikusiz. Datuak birritan ikustearena ohiko praktika da aurrentrenamendu corpusa motz geratzen den kasuetan; adibidez, ingeleserako BERT originala (Devlin et al., 2019) 40 epokaz aurrentrenatu zen.

Eredu guztiak urrats kopuru eta labekada tamaina berdinaz entrenatu arren, eredua handitu ahala entrenamendu denbora luzeagoa behar du. Gure eredu guztiak TPUv3-8 makinetan entrenatu dira: BERT_{124M} ereduek 76 ordu behar izan dituzte, BERT_{51M} ereduek 32 ordu eta BERT_{16M} eredu txikienek, berriz, 10 bat ordu.

[5]Sarrerako datuen zenbat aldaera sortzen diren maskaratze ezberdinekin.

4.3 Ebaluazio Ingurunea

Eredu guztiak modu intrintsekoan eta estrintsekoan ebaluatu dira. Ebaluazio intrintsekorako, maskaratutako hizkuntza-modelatzean ebaluatu dira; ebaluazio estrintsekorako, aukeratutako hizkuntza guztietan eskuragarri dauden NLUko atazak aukeratu dira: entitate izendunen ezagutza, gaien sailkapena, sentimenduen analisia eta galdera-erantzunen hizkuntza inferentzia. Gure aukeraketaren barnean, token sailkapeneko ataza bat eta testu sailkapeneko hiru ataza daude, azken horien artean, sentimenduen sailkapena eta galdera-erantzunen hizkuntza inferentzia sartu ditugu, zeinek, entitate izendunen ezagutza eta gaien sailkapena bezalako azaleko ulermeneko atazek baino ulermen sakonagoa eskatzen duten (Zhang et al., 2021). 4.3. Taulak datu-multzoei buruzko xehetasun gehiago biltzen ditu.

	EU			ES			
Ataza	train	dev	test	train	dev	test	Metrika
NER	52K	13K	36K	265K	53K	52K	F1
Gai	9K	2K	2K	9K	1K	4K	F1
SA	6K	1K	1K	5K	2K	1K	F1
QNLI	2K	230	238	30K	4K	4K	acc.
MLM			1M			1M	acc.

	SW			FI			
Ataza	train	dev	test	train	dev	test	Metrika
NER	175K	25K	51K	180K	14K	46K	F1
Gai	10K	3K	7K	10K	10K	10K	F1
SA	6K	782	1K	4K	633	1K	F1
QNLI	4K	624	1K	7K	1K	1K	acc.
MLM			1M			1M	acc.

4.3. Taula: Ebaluazioan erabilitako datu-multzoak. NER eta MLMko tamainak tokenetan daude emanda, gainontzekoetan adibideetan. F1 micro-F1 puntuazioari deritzo, eta acc, berriz, zehaztasunari.

4.3.1 MLM

Maskaratutako hizkuntza-modelatzea (*Masked Language Modeling* edo MLM) BERTaren aurrentrenamenduko defektuzko helburu-funtzioetako bat da. MLMaren galera eta zehaztasuna, biak emango ditugu.

Horretarako, ereduen aurrentrenamenduan erabili ez diren albiste iturrietako testuak baliatuz ebaluaziorako datu-multzo bat sortu dugu. Euskararako Argia[6] aldizkaritik hartu dugu testu hori. Gaztelaniarako El Mundo[7] egunkaria aukeratu dugu. Swahilirako, aurrentrenamenduan erabili ez den testu iturri gisa, SwahBERT eredua (Martin et al., 2022) entrenatzeko erabilitako corpusaren azpimultzo bat erabili dugu, %80a albisteez osatua. Suomierarako, berriz, aurrentrenamenduan erabili ez den *cc100*en azpimultzo bat erabili dugu, lizentzia libreko dokumentu-mailako albiste corpusik ez baitago eskuragarri.

4.3.2 Entitate Izendunen Ezagutza

Entitate izendunen ezagutza (*Name Entity Recognition* edo NER) token sailkapeneko ataza bat da. Euskararako BasqueGLUE proba-bankuko (Urbizu et al., 2022) *domeinuko NER* datu-multzoa erabili da. Gaztelaniarako, *Conll2002* datu-multzoa (Sang, 2002) aukeratu dugu. Swahilirako berriz, *Masakhaner* (Adelani et al., 2021) datu-multzoa erabili da. Azkenik, Suomierarako, *FiNER* (Ruokolainen et al., 2019) hautatu dugu. Datu-multzo guztiek launa kategoria dituzte, eta doitasunaren eta estalduraren batezbesteko harmonikoa den F1a erabili dugu ebaluazio metrikatzat.

4.3.3 Gaien Sailkapena

Gaien sailkapena (*Topic Classification*) testu sailkapeneko kategoria anitzeko ataza bat da. Euskararako 12 kategoria dituen *BHTCv2* aukeratu da, (Urbizu et al., 2022). Gaztelaniarako, aldiz, 4 kategoria dituen *MLdoc* (Schwenk and Li, 2018) hautatu da. Swahilirako ere 4 kategoria dituen datu multzo bat aukeratu da: *Swahili News Classification Dataset* (David, 2020). Azken honek garapenerako azpimultzoa falta zuenez, %20 bat bereizi dugu ausaz

[6] `www.argia.eus`
[7] `www.elmundo.es`

entrenamendurako datu-azpimultzotik, berau sortzeko. Gainera, birdoitzeko datu-multzoa aurrentrenamenduko datu-multzo txikiena (5M) baino handiagoa denez, 10K adibidetara txikitu dugu entrenamendurako datu-multzoa. Eta Suomierarako berriz, *Yle*[8] corpusaren %10eko bertsioa aukeratu dugu, 10 kategoria ezberdin dituena. Emaitzak F1 metrika erabiliz eman ditugu.

4.3.4 Sentimenduen Analisia

Sentimenduen analisia (*Sentiment Analysis* edo SA) testu sailkapen ataza bat da. Euskararako *BEC2016eu* (Urbizu et al., 2022) erabili dugu, positibo, negatibo eta neutro kategoriekin. *InterTass2020* (Cumbreras et al., 2016) hautatu dugu gaztelaniarako sentimenduen analisirako datu-multzo gisa, eta honek ere positibo, negatibo eta neutro kategoriak ditu anotatuta. Swahilirako, Martin et al. (2022) lanean aurkeztutako datu multzoa erabili da, eta polaritate mapaketa bat egin da artikuluko gidalerroak jarraituz: *joy* ([1]) = positiboa, *disgust* ([4]) = negatiboa, *neutral* ([0]) eta *surprise* ([5]) = neutroa. Etiketa bakarreko adibideak soilik erabili ditugu eta jatorrizko entrenamendu/garapen/ebaluazio banaketa mantendu dugu. Azkenik, Suomierarako *Finnish sentiment*[9] hautatu dugu, kategoria positiboa eta negatiboa biltzen dituena. Emaitzak F1 metrika erabiliz eman ditugu.

4.3.5 QNLI

Galdera-erantzunen hizkuntza inferentzia (*Question-Answering NLI* edo QNLI) ere testu sailkapeneko ataza bat da. Euskararako $QNLI_{eu}$ (Urbizu et al., 2022) baliatu da. Gaztelaniarako, swahilirako eta suomierarako, berriz, dagoeneko eskuragarri zeuden galdera-erantzuneko (QA) datu-multzoak egokitu dira, sekuentzia-pare sailkapen bitar batera moldatuz, ingeleserako QNLIaren diseinua jarraituz (Wang et al., 2019). Hautatutako galdera-erantzun datu-multzoak $SQAD_{es}$ (Carrino et al., 2020), $Tydiqa_{sw}$ eta $Tydiqa_{fi}$ (Clark et al., 2020a) izan dira.

[8] `www.github.com/spyysalo/yle-corpus`
[9] `www.huggingface.co/datasets/sepidmnorozy/Finnish_sentiment`

QNLIko datu-multzo bakoitzak galdera-sekuentzia pareak ditu[10]. *Tydiqa* datu-multzoak entrenamenduko eta garapeneko zatiak bakarrik eskaintzen ditu. Horrenbestez, garapeneko zatia zena ebaluaziorako erabili dugu, eta entrenamenduko zatitik adibide batzuk ausaz bereizi ditugu, gure garapenerako datu-multzoa sortzeko. Ingeleseko QNLIren diseinua jarraituz, zehaztasuna erabili da ebaluazio metrika gisa.

4.3.6 Sistemak eta Oinarri-Lerroak

Ebaluazio estrintsekorako, 36 BERT ereduetako bakoitza Transformers liburutegiaren (Wolf et al., 2020a) bitartez birdoitu dugu, $3e^{-5}$ko ikasketa-tasarekin, 32ko labekada tamainarekin, eta 10 epokaraino entrenatuz[11]; balio lehenetsiak dira oro har (Devlin et al., 2019; Mosbach et al., 2020). Ataza eta hizkuntza bakoitzeko 5 exekuzioren batezbesteko balioa eman dugu.

Transformerretan oinarritutako ereduen emaitzak hurbilpen arinagoenekin konparatzeko, testuingurudun hitz-bektoreetan oinarritutako oinarri-lerro neuronal lehiakorra inplementatu dugu Flair (Akbik et al., 2019) erabiliz. Token sailkapeneko atazetan, hitz-bektoreak BiLSTM-CRF sistema batean sartzen dira, Huang et al. (2015) lanean proposatutako arkitektura baliatuz. Testu sailkapeneko atazetan, berriz, kalkulatutako Flair hitz-bektoreak BiLSTM bati ematen zaizkio, dokumentu mailako bektore bat sortzeko, eta hau geruza lineal bati pasatzen zaio ondoren, kategoria aurreikusteko.

Hizkuntza bakoitzerako testuingurudun Flair hitz-bektoreak 125Meko corpusekin aurrentrenatu ditugu, ondorengo hiperparametro hauekin: 2048ko ezkutuko geruza, 250eko sekuentzia luzera, 100eko labekada tamaina eta 10 epoka. Gainontzean, lehenetsitako hiperparametroak erabili ditugu entrenamendurako[12].

LSTMetan oinarritutako oinarri-lerro neuronalaz gain, mBERT_{base} eredu eleaniztuna (Devlin et al., 2019) ere sartu dugu konparaketan. Horrelako hizkuntza-eredu eleaniztunak beti eskuragarri egongo direla eta, baliabide urriko testuinguruetan nolabaiteko alternatibatzat jo ditugu. Adibidez,

[10]Euskararen eta gaztelaniaren kasuan sekuentzia horiek esaldi bakarra dira, swahiliaren eta suomieraren kasuan, aldiz, paragrafo laburrak, jatorrizko QA datu-multzoek metodologia ezberdinak jarraitzen baitituzte.

[11]Garapeneko datu-multzoan emaitzarik onenak ematen dituen epokarekin geratuz.

[12]Ereduek 80na ordu behar izan dituzte entrenatzen, Nvidia TitanV GPU batetan.

kalkulu-ahalmen mugatua dugunean edota aurrentrenamendurako eskuragarri dagoen testu kopurua mugatua denean hizkuntza jakin batean.

Ildo horretatik, konparaketa euskaraz eta swahiliz bakarrik egin dugu, gaztelaniarako eta suomierarako ez bezala, mBERTen aurrentrenamenduan erabilitako corpusak gure esperimentuetan erabilitako corpus tamainen bueltakoak[13] direlako.

4.4 Emaitzak

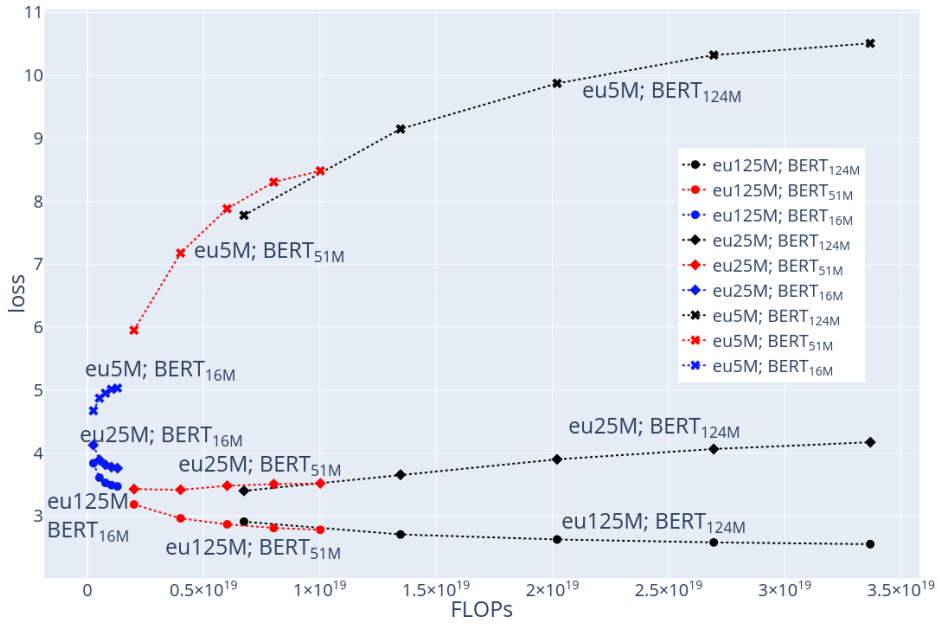
4.4.1 Eskala Txikiko Eskala-Legeak

4.1- 4.4 Irudiek lau hizkuntzetarako ereduek garapeneko datu-multzoetan lortutako MLMko galeraren eta FLOPSen arteko erlazioa erakusten dute. Kolore bakoitzak eredu tamaina ezberdin bat adierazten du (16M/51M/125M), eta ikur bakoitzak aurrentrenamenduko datu kopurua (5M/25M/125M). Lerro bakoitza ereduaren 5 kontrol-puntuk osatzen dute (100K urratsero).

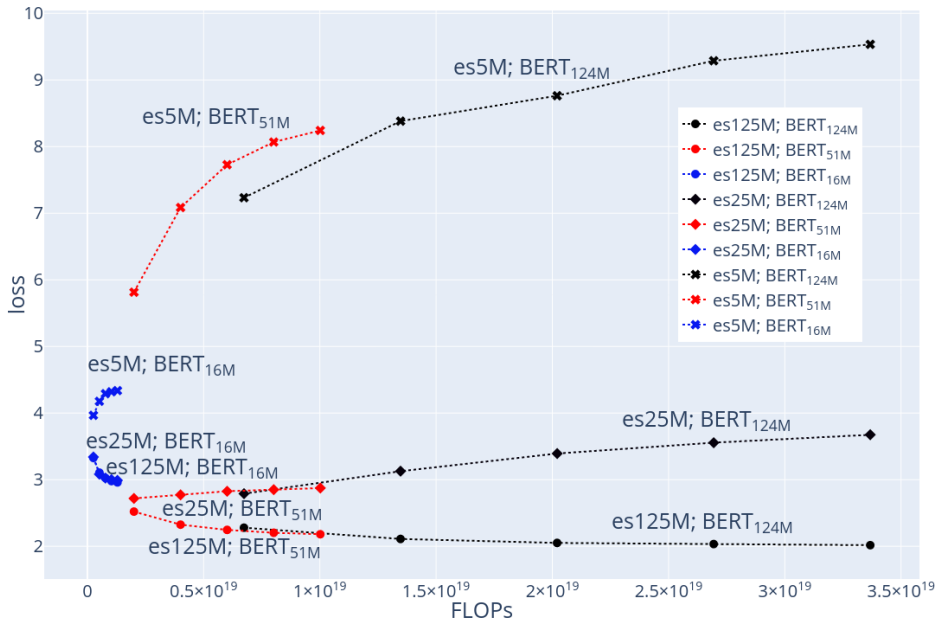
Ikus dezakegunez, galera txikiena aurrentrenamenduan datu gehien dituen eredu handienak lortzen du, eskala-legeen arabera esperotako moduan, eredu tamaina, datu-multzoen tamaina eta kalkulu ahalmena batera handitzean emaitzetan hobekuntza iradokitzen baitute. Bestalde, grafikoek erakusten dute aurrentrenamenduko datu kopurua igotzea ereduaren tamaina igotzea baino onuragarriagoa dela baliabide urriko ingurune honetan. Aurrentrenamenduko datu kopurua handituta lortzen den galeraren jaitsiera ereduen tamaina edo entrenamenduko urrats kopurua handituta baino handiagoa da.

Grafikoek, gainera, erakusten dute corpus txikiekin entrenatutako ereduek galera handiagoa lortzen dutela garapeneko MLM atazan, entrenamenduan aurrera jarraitu ahala ($\times$ kurbak), eta hau gainegokitzapenari leporatu geniezaioke, entrenamenduko MLM galerak txikitzen jarraitzen du eta. Datu-multzo ertaineko eredu handiek ($\diamond$ kurba gorri eta beltzak) ere joera bera erakusten dute, neurri txikiago batean bada ere. Grafikoen arabera, BERT_{51M} eta BERT_{124M}en kasuan 125M hitzeko datu-multzo bat erabili ezean, edota BERT_{16}en kasuan 25Meko datu-multzo bat erabili ezean, gelditze goiztiarra aplikatzeko erabakia hartu beharko genuke, gainegokitzapena saihesteko.

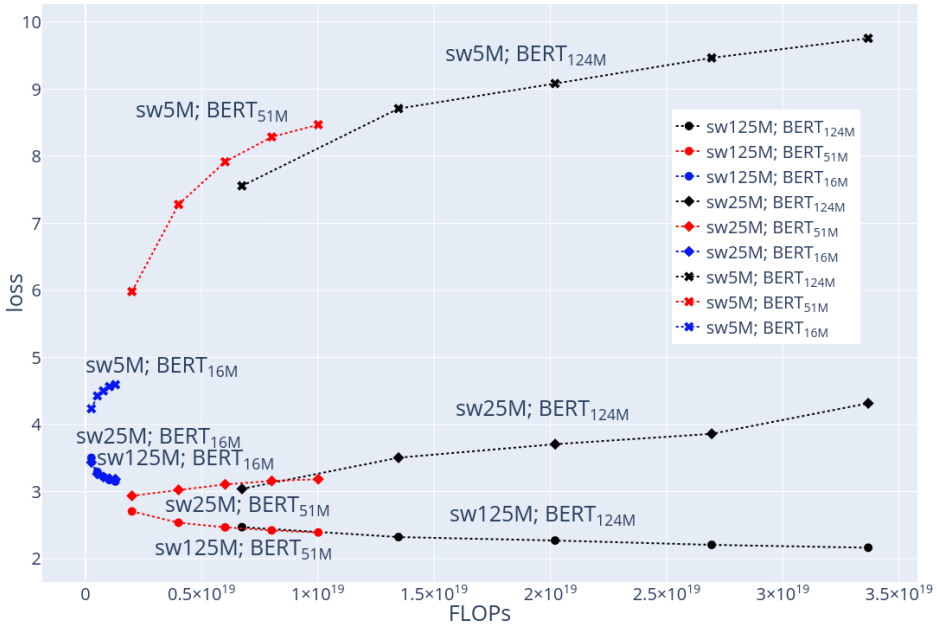
[13] Gutxi asko, euskararako 35M eta swahilirako 11M inguru.



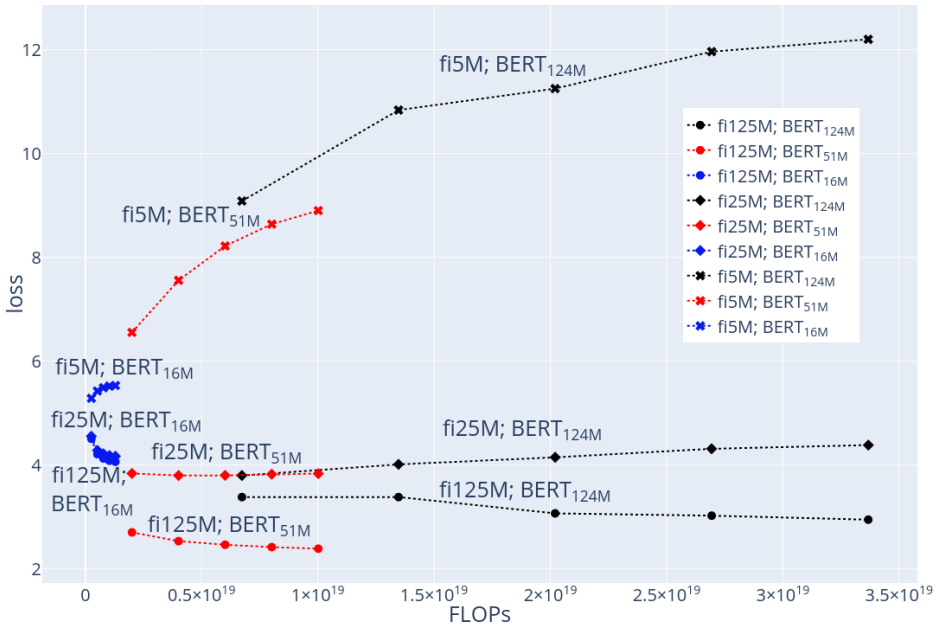
4.1. Irudia: Euskararako ereduen MLM galera eta FLOPSak.



4.2. Irudia: Gaztelaniarako ereduen MLM galera eta FLOPSak.



4.3. Irudia: Swahilirako ereduen MLM galera eta FLOPSak.



4.4. Irudia: Suomierarako ereduen MLM galera eta FLOPSak.

Aztertu ditugun kasu gehienetan, 100K. urratseko kontrol-puntuan edo lehenago izan beharko luke gelditze goiztiar horrek.

Hala eta guztiz ere, aurrentrenamenduko gainegokitzapen arazo horiek, ez dute eragin zuzenik ez MLMko zehaztasun emaitzetan, ezta helburu-atazetan ere 4.4.1.1. Atalak erakusten duenez. Honenbestez, 500K. urratseko azken kontrol-puntua erabili da ereduak MLMko atazan (4.4.2. Atala) eta NLUko helburu atazetan (4.4.3. Atala) ebaluatzeko.

Hizkuntzei dagokienez, lau irudiak konparatuz (4.1-4.4 Irudiak), argi dago, MLMko galeraren, eredu tamainaren, datu-multzoen tamainaren eta FLOPSen arteko korrelazioak bat datozela hizkuntzen artean, MLMn lortutako zehaztasunen kasuan bezala (ikus 4.4.2. Atala).

Parametro kopurua eta datu kopuruaren orekari helduz, gure emaitzen arabera, ereduak optimoki entrenatzeko datu eskakizunak gutxienez 25M hitz lirateke BERT_{16M} tamainaren kasuan, eta 125M hitz BERT_{51} eta BERT_{124}en kasuan. 4.5. Atalak analisi sakonagoa biltzen du, emaitza eta xehetasun gehiagorekin.

4.4.1.1 Aurrentrenamenduko Gainegokitzapena Helburu-Atazetara Barreiatzen ote da?

4.4.1. Ataleko galera kurbek iradokitzen dute eredu eta datu-multzo tamainen konbinaketa batzuk, datu gutxien eta parametro gehien dituztenak hain zuzen, luzaroegi entrenatu direla, galera berriro nabarmen igotzen hasten den punturaino, gainegokitzapen zantzuak, alegia.

Aurrentrenamenduko datu berberak behin eta berriz ikusten jardun izanak sortutako gainegokitzapen arazo horiek helburu-atazetara barreiatzen diren aztertzeko, hemen ereduaren 100K. eta 500K. urratseko kontrol-puntuak konparatu ditugu. Konparaketa BERT_{16M} eta BERT_{51M} tamainarekin burutu dugu, biak ala biak 5Mko corpusarekin aurrentrenatuak, galera kurbetan gainegokitzapen gehien erakusten duten bi ereduak baitira, hain zuzen.

Bi ereduen kontrol-puntuetako bakoitza atazetan birdoituta lortutako emaitzak 4.4. Taulan daude ikusgai. Laburbilduz, 500K. urratseko azken kontrol-puntuak 100K. urratsekoaren pare dabiltza, garaile argirik gabe. Zalantzarik gabe, MLMko galeran ikusitako diferentzia berdindu egin da.

Beraz, aurrentrenamenduan ikusitako galera helburu-atazetara ez dela barreiatzen ondorioztatu dezakegu. Honenbestez, azken kontrol-puntuak (500-

K. urratsekoak) aukeratu dira ereduak ebaluatu eta konparatzeko, 4.4.2. eta 4.4.3. Ataletan, beste aldagai bat gehitzea saihestearren.

		BERT_{16M}		BERT_{51M}	
	Ataza	100K urrats	500K urrats	100K urrats	500K urrats
eu	MLM loss	**4.6724**	5.0323	**5.9511**	8.4830
	MLM acc	31.56	**32.08**	**33.60**	32.42
	NER	**64.75**±0.6	63.90±0.5	70.13±0.4	**70.14**±0.4
	Topic	**68.56**±0.4	68.00±0.6	**70.18**±0.6	69.98±0.6
	SA	67.16±0.5	**67.80**±0.5	**67.32**±0.4	67.00±1.0
	QNLI	**68.20**±1.2	68.10±2.3	64.64±1.6	**65.06**±0.8
es	MLM loss	**3.9658**	4.3367	**5.8144**	8.2425
	MLM acc	**40.15**	39.91	**39.27**	38.93
	NER	75.84±0.4	**76.57**±0.3	**81.11**±0.3	80.43±0.4
	Topic	**94.69**±0.4	94.54±0.3	**94.89**±0.3	**94.89**±0.3
	SA	37.26±0.8	**37.67**±1.5	35.68±3.9	**36.05**±2.1
	QNLI	**65.38**±1.0	65.01±0.6	66.34±1.6	**67.00**±1.8
sw	MLM loss	**4.2363**	4.5952	**5.9836**	8.4719
	MLM acc	37.64	**38.03**	**38.70**	38.08
	NER	85.92±0.5	**86.36**±0.2	**89.05**±0.2	88.74±0.2
	Topic	91.28±0.3	**91.64**±0.3	91.85±0.4	**92.12**±0.2
	SA	71.31±0.5	**71.52**±0.6	**71.01**±0.6	70.49±0.7
	QNLI	60.29±0.9	**62.80**±1.1	**62.88**±0.9	62.27±1.7
fi	MLM loss	**5.2866**	5.5322	**6.5559**	8.9016
	MLM acc	28.76	**29.43**	**29.52**	28.18
	NER	76.47±0.3	**76.82**±0.3	**80.19**±0.2	79.73±0.2
	Topic	**88.53**±0.1	88.15±0.1	88.45±0.1	**88.53**±0.3
	SA	**89.95**±0.1	89.69±0.2	**90.50**±0.2	89.61±0.6
	QNLI	**51.51**±1.1	51.49±0.9	52.37±0.7	**54.28**±2.5

4.4. Taula: Gainparametrizatzearen ondorioz gainegokitzapen zantzuak erakutsi dituzten ereduen kontrol-puntu ezberdinek (100K eta 500K urratsetakoek) helburu atazetan lortutako emaitzen konparaketa.

4.4.2 MLM Ebaluazioa

4.5. Taulak eredu tamainaren eta corpus tamainaren konbinazio bakoitzeko ereduek hizkuntza bakoitzeko (euskara, gaztelania, swahilia eta suomiera) MLM atazan lortutako zehaztasunak biltzen ditu. Esperotako moduan, corpus handienekin entrenatutako eredu handienek erdiesten dituzte emaitzarik onenak, eta korrelazio positiboa garbia dago ereduaren edo corpusen tamainaren eta lortutako zehaztasunaren artean hizkuntza guztietan zehar. Horrez gain, emaitzek erakusten dute nola baliabide urriko ingurune hauetan, orokorrean, hobe dela aurrentrenamenduko datu kopurua handitzea ereduaren tamainaren aldean. Aurrentrenamenduko datuak gehitzeak MLMko emaitzak hobetu ditu hizkuntza guztietan, 125M tokeneko datu-multzoarekin entrenatutako BERT_{16M} ereduaren salbuespenarekin. Datuak gehitu ahala BERT_{16M} ereduaren emaitzetan lortzen den hobekuntza txikitzen doa, eredu tamaina horren goi langara iristen ari garela iradokiz.

MLM_{eu}	5M	25M	125M
BERT_{16M}	32.08	38.68	41.56
BERT_{51M}	32.42	44.29	50.07
BERT_{124M}	34.50	43.46	**53.19**
MLM_{es}	5M	25M	125M
BERT_{16M}	39.09	49.06	48.31
BERT_{51M}	39.24	53.49	59.04
BERT_{124M}	42.45	52.58	**62.00**
MLM_{sw}	5M	25M	125M
BERT_{16M}	38.03	45.71	44.98
BERT_{51M}	38.08	50.12	55.27
BERT_{124M}	40.43	49.06	**58.82**
MLM_{fi}	5M	25M	125M
BERT_{16M}	29.43	37.03	37.73
BERT_{51M}	28.18	42.07	45.86
BERT_{124M}	29.30	41.75	**49.88**

4.5. Taula: Lau hizkuntzetako MLMko zehaztasunak. Lerro bakoitza hizkuntza-eredu tamaina ezberdin bati dagokio; zutabe bakoitza aurrentrenamendu corpus tamaina bati.

Bestalde, ereduaren tamaina handitzeak, aurrentrenamendu datu kopuru

jakin batetik aurrera bakarrik laguntzen du. Ereduaren tamaina BERT_{16M}tik BERT_{51M}era handitzeak ez dakar hobekuntzarik MLMko zehaztasunean, 5Mko corpusaren kasuan, 3M hitz-bektore ez besteko parametro horrelako datumultzo baten ezagutzaren gehiengoa barneratzeko nahikoa direla iradokiz.

Hala eta guztiz ere, ereduaren tamaina BERT_{51M}etik BERT_{124M}era handitzeak 5Mko corpusarentzat hobekuntzak dakartza suomiera ez beste hiru hizkuntzetan. Azken hau, eredu handiagoak lagin-eraginkorragoak (*sample-efficient*) izateari leporatu dakioke (Kaplan et al., 2020).

Harrigarriro, BERT_{51M} eredua BERT_{124M}i gailentzen zaio hizkuntza guztietan zehar 25Mko corpusarekin aurrentrenatzean; hau eredu handiak lagin-eraginkorrak diren senaren aurka doa. Ostera, taulek erakusten dute datu gehiagodun eredu pitin bat txikiagoa datu gutxiagodun eredu handiago bati gailendu dakiokeela: 125M tokeneko corpusekin entrenatutako BERT_{51M} eredu guztiak 25M tokenekin entrenatutako BERT_{124M} ereduen gainetik dabiltza.

4.4.3 NLU Atazen Ebaluazioa

Atal honek, gure ereduek 4.3. Atalean zerrendatutako NLU atazetan lortutako emaitzen analisia dakar, ereduen tamainaren eta aurrentrenamenduko datu-multzoaren tamainaren efektua neurtzeko, behin helburu-atazetan birdoitu ondoren.

4.6. Taulak NER atazarako emaitzak biltzen ditu. Espero bezala, korrelazio positiboa dago ebaluazio metriken eta ereduaren eta corpusaren tamainen artean, baina zertxobait onuragarriagoa da corpus tamaina handitzea, ereduaren tamainarekin alteratuz.

Aurrentrenatutako gure ereduek gaien sailkapenean lau hizkuntzetan lortutako emaitzak 4.7. Taulan daude bilduta, eta joera berdina erakusten dute, aldeak leunagoak diren arren.

4.8. Taulan ikus daitezke sentimenduen analisiko emaitzak, hemen ere joera orokor berdina errepikatzen da, baina badira azpimarratzeko moduko muturreko datu apur batzuk.

Bukatzeko, QNLIn (ikus 4.9. Taula), emaitzek goranzko joera dutela ikus daiteke corpus eta ereduak handitu ahala, baina balioetako askok desbidera-

NER_{eu}	5M	25M	125M
$BERT_{16M}$	63.90±0.5	72.23±0.6	74.12±0.3
$BERT_{51M}$	70.14±0.4	79.07±0.4	82.98±0.1
$BERT_{124M}$	73.14±0.5	79.09±0.8	**84.58**±0.2
NER_{es}	5M	25M	125M
$BERT_{16M}$	76.57±0.3	81.56±0.5	81.70±0.5
$BERT_{51M}$	80.43±0.4	85.11±0.8	86.34±0.7
$BERT_{124M}$	81.75±0.4	84.99±0.8	**87.28**±0.3
NER_{sw}	5M	25M	125M
$BERT_{16M}$	86.36±0.2	88.62±0.2	88.63±0.4
$BERT_{51M}$	88.74±0.2	90.68±0.2	91.63±0.1
$BERT_{124M}$	88.93±0.4	90.97±0.2	**92.09**±0.2
NER_{fi}	5M	25M	125M
$BERT_{16M}$	76.82±0.3	81.48±0.4	81.83±0.3
$BERT_{51M}$	79.73±0.2	85.27±0.4	87.02±0.2
$BERT_{124M}$	80.56±0.7	85.77±0.3	**88.99**±0.2

4.6. Taula: NER atazan euskararako, gaztelaniarako, swahilirako eta suomierarako erdietsitako emaitzak (F1).

tze estandar altuak dituzte, eta joera orokorretik aldentzen diren muturreko datu[14] ugari ere badaude.

Ereduaren tamaina handitzean emaitzetan lortutako igoera handiagoa da helburu-atazetan MLMko ataza intrintsekoan baino, batez ere ereduaren tamaina $BERT_{16M}$ txikitik $BERT_{51M}$ ertainera handitzen dugunean. Horrek, eredu handiagoak birdoitze prozesuan eraginkorragoak direla adierazten du, entrenagarriak diren parametro kopuru handiagoa baitute.

Emaitzak eta eskalatze joerak sendoak dira aztertu diren hizkuntzetan zehar. Lortutako emaitza eta joerak ere bat datoz atazen artean, QNLIren salbuespenarekin, non emaitza ezegonkorrak lortu diren[15] eta muturreko balio asko dituzten.

[14]Muturreko datuak gorriz.
[15]1,8ko desbideratze estandarrarekin.

Topic_{eu}	5M	25M	125M
BERT_{16M}	68.00±0.6	71.81±0.3	72.49±0.4
BERT_{51M}	69.98±0.6	73.16±0.6	74.87±0.4
BERT_{124M}	71.70±0.7	74.61±0.3	**76.06**±0.4
Topic_{es}	5M	25M	125M
BERT_{16M}	94.54±0.3	95.86±0.3	95.42±0.4
BERT_{51M}	94.89±0.3	95.45±0.2	95.91±0.4
BERT_{124M}	95.32±0.4	95.82±0.3	**96.27**±0.3
Topic_{sw}	5M	25M	125M
BERT_{16M}	91.64±0.3	91.96±0.3	92.45±0.2
BERT_{51M}	92.12±0.2	92.39±0.1	92.88±0.2
BERT_{124M}	91.95±0.4	92.69±0.1	**93.07**±0.2
Topic_{fi}	5M	25M	125M
BERT_{16M}	88.15±0.1	88.94±0.2	89.16±0.2
BERT_{51M}	88.53±0.3	89.40±0.3	89.61±0.3
BERT_{124M}	88.41±0.2	89.72±0.2	**90.14**±0.1

4.7. Taula: Gaien sailkapenean euskararako, gaztelaniarako, swahilirako eta suomierarako erdietsitako emaitzak (F1).

4.4.4 Oinarri-Lerroekiko Konparaketa

4.10. Taulan, 125M hitzeko corpusarekin entrenatutako ereduen emaitzak, BiLSTM-CRF Flair oinarri-lerroarekin (125Meko corpusarekin entrenatua) eta mBERTekin[16] alderatu ditugu.

Bert ereduak Flair oinarri-lerro neuronalari gailentzen zaizkio, baina ebaluazio datu-multzoaren (ataza eta hizkuntzaren) arabera, batzuetan oinarri lerroari BERT_{124M} tamainako eredua bakarrik gailentzen zaio, eta beste batzuetan hiru eredu tamainek gainditzen dute. Horretaz gain, datu-multzo batzuen kasuan, corpus txikienarekin (5M) entrenatutako BERT_{16M} ereduak ere 125M hitzekin entrenatutako Flair oinarri-lerroaren emaitzak gainditzen ditu (taula horretan ez dira ageri). Bukatzeko, kalkulu-ahalmenaren kostuak (4.4.6. Atalean aztertua) kontuan hartu behar dira, eta kalitatean lortutako hobekuntzek kalkulu-ahalmen gastuaren igoera justifikatzen duten erabaki behar da erabilera kasu bakoitzean.

[16]Helburu-hizkuntzan aurrentrenamenduko corpusaren tamaina konparagarria izan duten hizkuntzen kasuan.

SA_{eu}	5M	25M	125M
$BERT_{16M}$	67.80±0.5	68.63±1.0	67.59±0.5
$BERT_{51M}$	67.00±1.0	68.54±0.5	**69.40**±0.9
$BERT_{124M}$	67.22±0.7	68.79±0.7	68.91±0.5
SA_{es}	5M	25M	125M
$BERT_{16M}$	37.67±1.5	37.62±0.7	37.51±1.4
$BERT_{51M}$	36.05±2.1	37.57±0.5	39.89±0.0
$BERT_{124M}$	36.37±2.2	37.17±1.1	**43.27**±1.1
SA_{sw}	5M	25M	125M
$BERT_{16M}$	71.52±0.6	75.56±0.3	74.84±0.6
$BERT_{51M}$	70.49±0.7	75.39±1.4	77.07±0.0
$BERT_{124M}$	69.60±1.3	75.54±0.9	**79.04**±0.7
SA_{fi}	5M	25M	125M
$BERT_{16M}$	89.69±0.2	90.96±0.2	91.14±0.4
$BERT_{51M}$	89.61±0.6	91.86±0.3	92.58±0.0
$BERT_{124M}$	90.32±0.5	91.55±0.3	**94.38**±0.3

4.8. Taula: Sentimenduen analisian euskararako, gaztelaniarako, swahilirako eta suomierarako lortutako emaitzak (F1).

4.4.5 Artearen Egoerarekiko Konparaketa Helburu-Atazetan

4.11. Taulan 125M hitzekin aurrentrenatutako gure $BERT_{124M}$ ereduen emaitzak, Flair oinarri lerroez eta mBERT ereduez gain, hizkuntza bakoitzeko ataza bakoitzerako artearen egoerako emaitzarik onenak dituzten sistemekin konparatu ditugu.

Lan honetan zehar aurrentrenatutako BERT ereduek emaitza lehiakorrak lortu dituzte. Esate baterako, swahilirako NER, gaien sailkapena[17] eta sentimenduen analisia atazetan artearen egoera berria ezartzen dute ebaluaziorako datu-multzo horietan, guk dakigunez. Suomierazko QNLIn ere, gure BERT ereduak lortzen ditu emaitza onena.

Gainera, 125M hitzeko corpusarekin entrenatutako $BERT_{124M}$ ereduek lortutako emaitzetako batzuk corpus erraldoiekin entrenatutako artearen ego-

[17]Entrenamendu datu-multzoaren azpimultzo bat bakarrik erabilita birdoitutako eredua.

$QNLI_{eu}$	5M	25M	125M
$BERT_{16M}$	68.19±2.3	68.95±2.0	71.22±5.0
$BERT_{51M}$	65.06±0.8	**76.37**±2.5	74.18±1.6
$BERT_{124M}$	67.43±2.7	72.66±2.0	74.09±1.7
$QNLI_{es}$	5M	25M	125M
$BERT_{16M}$	65.01±0.6	70.89±1.4	72.72±0.7
$BERT_{51M}$	67.00±1.8	74.11±1.1	78.00±0.0
$BERT_{124M}$	67.39±0.5	73.07±1.3	**81.10**±0.7
$QNLI_{sw}$	5M	25M	125M
$BERT_{16M}$	62.80±1.1	62.45±1.2	63.42±1.0
$BERT_{51M}$	62.27±1.7	63.83±1.4	**63.87**±1.5
$BERT_{124M}$	64.08±1.1	62.68±1.2	63.34±1.4
$QNLI_{fi}$	5M	25M	125M
$BERT_{16M}$	51.49±0.9	50.96±0.7	58.89±3.7
$BERT_{51M}$	54.28±2.5	54.58±3.2	57.30±4.3
$BERT_{124M}$	54.07±2.6	**59.89**±4.5	58.56±1.1

4.9. Taula: QNLIn lortutako zehaztasunak, euskararako, gaztelaniarako, swahilirako eta suomierarako.

erako ereduen parekoak dira edo ez dabiltza urruti, besteak beste, suomierarako gaien sailkapenean.

4.4.6 FLOPSak and CO_2 Isurketak

4.12. Taulak eredu bakoitzaren aurrentrenamendu, birdoitze[18] zein inferentziarako kalkulu-ahalmen kostuak eta CO_2 isurketak dakartza. FLOPSak Hoffmann et al. (2022) lanean azaltzen den metodoa jarraituz kalkulatu ditugu. Oinarri-lerroaren kasuak, FLOPSak Zhang et al. (2018) lanaren arabera kalkulatu dira. CO_2 isurketak *Machine-Learning Impact calculator*[19] (Lacoste et al., 2019) tresna erabiliz estimatu dira. Bi-LSTMetan oinarritutako oinarri-lerro neuronala arinagoa da FLOPSei dagokionez, bai aurrentrenamenduan eta baita birdoitzean eta inferentzia garaian, $BERT_{16M}$ eredu txikienarekin

[18]Birdoitzearen balioak gaztelaniako gaien sailkapeneko birdoitze exekuzio bakarretik kalkulatu dira.

[19]`https://mlco2.github.io/impact#compute`

		BERT_{16M}	BERT_{51M}	BERT_{124M}	Flair	mBERT
eu	NER	74.12±0.3	82.98±0.1	**84.58**±0.2	82.13±0.4	79.39±1.0
	Topic	72.49±0.4	74.87±0.4	**76.06**±0.4	67.89±0.3	70.57±0.5
	SA	67.59±0.5	**69.40**±0.9	68.91±0.5	68.17±0.3	67.34±0.7
	QNLI	71.22±5.0	74.18±1.6	74.09±1.7	48.66±5.2	**78.48**±1.9
es	NER	81.70±0.5	86.34±0.7	**87.28**±0.3	87.09±0.3	—
	Topic	95.42±0.4	95.91±0.4	**96.27**±0.3	94.08±0.4	—
	SA	37.51±1.4	39.89±0.0	**43.27**±1.1	34.73±3.0	—
	QNLI	72.72±0.7	78.00±0.0	**81.10**±0.7	56.42±0.6	—
sw	NER	88.63±0.4	91.63±0.1	**92.09**±0.2	92.04±0.1	91.17±0.1
	Topic	92.45±0.2	92.88±0.2	**93.07**±0.2	91.83±0.2	91.52±0.2
	SA	74.84±0.6	77.07±0.0	**79.04**±0.7	73.60±0.5	69.17±1.2
	QNLI	63.42±1.0	**63.87**±1.5	63.34±1.4	52.82±2.1	63.48±1.1
fi	NER	81.83±0.3	87.02±0.2	**88.99**±0.2	84.76±0.4	—
	Topic	89.16±0.2	89.61±0.3	**90.14**±0.1	86.58±0.7	—
	SA	91.14±0.4	92.58±0.0	**94.38**±0.3	89.74±0.5	—
	QNLI	**58.89**±3.7	57.30±4.3	58.56±1.1	51.54±1.2	—

4.10. Taula: Corpus handienarekin (125M) aurrentrenatutako BERT eta Flair ereduen emaitzen konparaketa. Konparaketara mBERT gehitu zaie, aurrentrenamenduan corpus konparagarria izan duten hizkuntzen kasuan.

alderatuta ere. Hala ere, Flair oinarri-lerroak CO_2 isurketa handiagoak dirtu birdoitze prozesuan, LSTMen izaera errepikaria dela eta, paralelizatzeko duen gaitasuna ezagatik.

MLM eta NLUko atazen emaitzak (4.4.1. eta 4.4.3. Atalak) berrazter ditzagun kalkulu-ahalmen kostuak kontuan izanik oraingoan. Corpus oso txiki bat (5M) bagenu, eredu txikiarekin (BERT_{16M}) lortutako emaitzak senide handiagoek lortutakoen parekoak dira, MLMn, gaien sailkapenean, sentimenduen analisian eta QNLIn, baina NERen kasuan, ordea, ereduaren tamaina handitu beharko litzateke (BERT_{51M}raino), emaitza lehiakorrak lortzeko. Corpus txiki bat (25M) izatearen testuinguruan, BERT_{16M} ereduak gaien sailkapenean eta sentimenduen analisian bakarrik da lehiakorra, baina BERT_{51M} ereduak BERT_{124M} handiagoaren pareko emaitzak edo hobeak lortzen ditu. Beraz, BERT_{51M}en alde egin genezake, eta kalkulu-ahalmenaren erdia aurreztu. Hala ere, 125M hitz baino gehiagoko aurrentrenamenduko corpusa dugun kasuan, BERT_{124M} ereduak emaitza nabarmen hobeak lortzen

		$BERT_{124M}$	Flair	mBERT	Artearen Egoera
eu	NER	84.58±0.2	82.13±0.4	79.39±1.0	**86.98**±0.4 roberta-euscrawl-l(Artetxe et al., 2022)
	Topic	76.06±0.4	67.89±0.3	70.57±0.5	**86.51**±0.4 ElhBERTeu (Urbizu et al., 2022)
	SA	68.91±0.5	68.17±0.3	67.34±0.7	**70.87**±0.5 Berteus (Agerri et al., 2020)
	QNLI	74.09±1.7	48.66±5.2	**78.48**±1.9	76.04±1.5 ElhBERTeu (Urbizu et al., 2022)
es	NER	87.28±0.3	87.09±0.3	87.21±0.4	**88.51** roberta-b(Gutiérrez Fandiño et al., 2022)
	Topic	96.27±0.3	94.08±0.4	95.92±0.6	**97.14** BETO(Gutiérrez Fandiño et al., 2022)
	SA	43.27±1.1	34.73±3.0	39.21±1.8	**49.80** Vega et al. (2020)
	QNLI	81.10±0.7	56.42±0.6	**83.92**±0.2	82.02 roberta-l(Gutiérrez Fandiño et al., 2022)
sw	NER	**92.09**±0.2	92.04±0.1	91.17±0.1	88.60 swahBERT (Martin et al., 2022)
	Topic	**93.07**±0.2	91.83±0.2	91.52±0.2	90.90 swahBERT (Martin et al., 2022)
	SA	**79.04**±0.7	73.60±0.5	69.17±1.2	71.12 swahBERT (Martin et al., 2022)
	QNLI	63.34±1.4	52.82±2.1	63.48±1.1	**64.72**±0.4 swahBERT(guk ebaluatua)
fi	NER	88.99±0.2	84.76±0.4	88.87±0.4	**92.40**±0.1 finBERT(Virtanen et al., 2019)
	Topic	90.14±0.1	86.58±0.7	88.16±0.3	**90.57**±0.2 finBERT(Virtanen et al., 2019)
	SA	94.38±0.3	89.74±0.5	88.05±0.7	**95.61**±0.3 finBERT(guk ebaluatua)
	QNLI	**58.56**±1.1	51.54±1.2	52.79±3.0	57.18±2.6 finBERT(guk ebaluatua)

4.11. Taula: Aurrentrenamenduko corpus handierarekin (125M hitz) entrenatutako $BERT_{124M}$ ereduaren, corpus berarekin entrenatutako Flair oinarri-lerroaren, mBERTen eta egungo artearen egoeraren emaitzak helburu-atazetan.

	Aurrentrenamendua		Birdoitzea		Inferentzia	
Ereduak	FLOPS	CO_2eq	FLOPS	CO_2eq	FLOPS	CO_2eq
$BERT_{124M}$	$4.9e{+}19$	98kg	$3.4e{+}16$	47g	$1.3e{+}11$	0.18mg
$BERT_{51M}$	$2.0e{+}19$	41kg	$1.4e{+}16$	23g	$5.3e{+}10$	0.07mg
$BERT_{16M}$	$6.3e{+}18$	13kg	$4.4e{+}15$	11g	$1.6e{+}10$	0.02mg
Flair	$1.4e{+}17$	4kg	$5.3e{+}14$	334g	$5.3e{+}09$	0.01mg

4.12. Taula: Kalkulu ahalmen kostuak FLOPetan aurrentrenamenduan, birdoitzean eta inferentzian eta bakoitzaren CO_2 isurketen estimazioak.

ditu, horrelako eredu handiago bat entrenatzeko behar den kalkulu-ahalmen gehigarriari etekina ateraz.

Haatik, hemen urrats kopuru berdinez entrenatutako tamaina ezberdinetako ereduak konparatzen ibili gara. Zein da entrenamendurako kalkulu-ahalmen finko batentzat eredurik onena lortzeko bidea? Galdera horri ondorengo atalean erantzun diogu.

4.4.7 Aurrekontu Finko baterako Optimizatzen

Ereduaren tamainari dagokionean, eskuragarri dugun corpus tamainarekin tamaina handiagoko eredu batekin hobekuntza lortuko litzatekeela iradoki arren, kasu askotan, kalkulu-ahalmen aurrekontu mugatuaren ari gintezke lanean. Beraz, kasu horietan, eredu tamaina aukeratzeko garaian gure kalkulu-ahalmen mugen barnean erabaki beharko genuke. Kaplan et al. (2020) lanaren arabera, konbergentzia ez da eraginkorra, eta kalkulu-ahalmen mugatua dugunean emaitzarik onenak lortzeko eredu handiago bat entrenatu eta hau konbergitzera heldu baino askoz lehenago moztu behar da.

		BERT_{51M}	BERT_{124M}
	Urrats	500K	200K
	FLOPSak	1.002E+19	1.347E+19
eu	NER	82.98±0.1	**83.67**±0.3
	Topic	74.87±0.4	**75.49**±0.3
	SA	**69.40**±0.9	68.43±0.8
	QNLI	**74.18**±1.6	72.49±2.9
es	NER	86.34±0.7	**86.47**±0.7
	Topic	95.91±0.4	**95.95**±0.1
	SA	39.89±0.0	**42.79**±1.1
	QNLI	78.00±0.6	**79.65**±0.4
sw	NER	91.63±0.1	**91.66**±0.2
	Topic	92.88±0.2	**93.07**±0.1
	SA	77.07±0.0	**77.60**±1.1
	QNLI	**63.87**±1.5	63.63±0.9
fi	NER	**87.02**±0.2	86.86±0.7
	Topic	89.61±0.3	**89.63**±0.2
	SA	92.58±0.0	**92.63**±0.5
	QNLI	**57.30**±4.3	56.35±7.6

4.13. Taula: BERT_{51M} eta BERT_{124M} ereduen emaitzak, kalkulu-ahalmen konparagarri batez (1E+19 FLOPS) entrenatuta, 500K eta 200K urratsez, hurrenez hurren.

Horretarako, kalkulu-ahalmen konparagarriarekin entrenatutako BERT_{51M} eta BERT_{124M} ereduak konparatu ditugu. BERT_{51M} 500K urratsez eta BERT_{124M} 200K urratsez aurrentrenatu ditugu 125M hitzeko corpusa baliatuz bi kasuetan. Bakoitzak lortutako emaitzak 4.13. Taulan daude ikusgai.

BERT_{124M}, parametro gehien dituen eredua, BERT_{51M} ereduari gailentzen zaio atazen gehiengoan: 4/4 espainieran, 3/4 swahilian, 2/4 euskaran eta 2/4 suomieran. Emaitza hauek bat datoz Kaplan et al. (2020) lanean egindako *"konbergentzia ez da eraginkorra"* baieztapenarekin. Hala ere, emaitzetan alde handirik ez dagoenez, bestelako faktore batzuk ere kontuan hartu genitzake. Adibidez, azpientrenatutako BERT_{124M} ereduak hobekuntzarako tarte handiagoa du aurrentrenamenduarekin jarraitzea erabakiko bagenu. BERT_{51M}, ostera, merkeagoa eta azkarragoa da birdoitzeari eta eskala handian zerbitzua emateari begira.

4.5 Eskala-Legeen Irakaspenak Baliabide Urriko Inguruneetan

Egindako esperimentu kopuruak mugatuta hizkuntza-eredu diskriminatzaileetarako eskala-lege propioak eratorri ez ditugun arren, 4.14. eta 4.15. Tauletan, baliabide urriko inguruneetan egindako esperimentuetatik ondorioztatutako datu, parametro eta kalkulu-ahalmenaren arteko erlazio optimoak bildu ditugu.

Datu-multzoa	FLOPS optimoak	Eredu tamaina optimoa
125M	$3.37\text{E}+19$	$> 8.56\text{E}+07$
25M	$5.39\text{E}+18$	$> 8.56\text{E}+07$
5M	$1.35\text{E}+18$	$8.56\text{E}+07$

4.14. Taula: Datu-multzo tamaina bakoitzarentzat FLOPS eta eredu tamaina optimoak.

Eredua	FLOPS optimoak	Datu-multzo tamaina optimoa
BERT_{124M}	$3.37\text{E}+19$	$> 1.25\text{E}+08$
BERT_{51M}	$1.00\text{E}+19$	$> 1.25\text{E}+08$
BERT_{16M}	$1.29\text{E}+18$	$1.25\text{E}+08$

4.15. Taula: Eredu tamainaren araberako bakoitzarentzat FLOPS eta datu-multzo tamaina optimoak.

Oro har, ondorengo irakaspenak azpimarratuko genituzke baliabide urriko inguruneetan hizkuntza-ereduak sortzen lanean ari direnentzat lagungarri izan daitezkeelakoan:

- Erabili eskura duzun bezainbeste datu, baliabide urriko inguruneetan aurrentrenamenduko datu kopurua igotzea ereduaren tamaina igotzea baino onuragarriagoa dela behatu baitugu.
- Hainbat epokaz (ehunka) aurrentrenatzea lagungarria da, zalantzarik gabe. Hala ere, kontuan izan ereduaren gainparametrizazioak gainegokitzapena eragin dezakeela; hori saihesteko gelditze goiztiarra erabili.
- Kalkulu-ahalmen aurrekontu finko bat emanik, hobe da eredu handixeago bat entrenatzea, kalkulu-ahalmen berarekin eredu txikiago bat urrats gehiagoz entrenatzea baino.
- 125M hitzeko datu-multzo batentzat: BERT_{124M} eredu tamaina, gutxienez 3.37E+19 FLOPSez[20] entrenatzea gomendatzen da.
- 25M hitzeko datu-multzoarentzat: BERT_{124M} eredu tamaina, 5.39E+18 FLOPSez entrenatua[21] gomendatzen da, baina 3.01E+18 FLOPSez[22] entrenatutako BERT_{51M} ereduak ere antzeko emaitzak lortzen ditu eta arinagoa da birdoitze eta inferentziari begira.
- 5M hitzeko datu-multzoarentzat: BERT_{124M} eredu tamaina entrenatzea gomendatzen da 1.35E+18 FLOPSez[23], baina 7.01E+17 FLOPSez[24] entrenatutako BERT_{51M} ereduak ere antzeko emaitzak erdiesten ditu eta arinagoa da birdoitze eta inferentziari dagokionez.

[20] 500K urrats, 256ko labekada eta 512ko sekuentzia-luzera.
[21] 80K urrats, 256ko labekada eta 512ko sekuentzia-luzera.
[22] 150K urrats, 256ko labekada eta 512ko sekuentzia-luzera.
[23] 20K urrats, 256ko labekada eta 512ko sekuentzia-luzera.
[24] 35K urrats, 256ko labekada eta 512ko sekuentzia-luzera.

4.6 Ondorioak

Hizkuntza-ereduen kalitatea baliabide urriko inguruneetan aztertzen duen azterlan bat aurkeztu dugu, hizkuntza-eredu handietan landutako eskala-legeen antzera, baliabide urriko inguruneetan datu eta parametroen arteko erlazio optimoak aztertuz. Laburbilduz, aurrentrenamenduko token kopuruak parametro kopuruak baino handiagoa izan beharko lukeela ondorioztatu dugu.

Gure esperimentuen bidez, esan daiteke, kalkulu-ahalmen jakin bat emanda, hobea dela eredu handiagoak ahalik eta datu gehienekin entrenatzea, eredu txikiagoak luzaroagoan entrenatzea baino. Gainegokitzapenerako joera nabarmena antzematen da eredu handiak corpus txikien gainean epoka ugariz, luzaroan aurrentrenatzean. Hala ere, aurrentrenamenduan gainegokitzapen zantzuak aurkezten dituzten eredu handi horiek, eredu txikiagoak gainditzen dituzte helburu atazetan birdoitu ondoren.

Emaitza esperimentalak sendoak dira hizkuntzen artean. Gainera, enpirikoki finkatu dugu Transformerretan oinarritutako hurbilpena erabiltzearen kalkulu-ahalmen kostuak noiz merezi duen.

Lan honetan zehar sortutako aurrentrenamenduko corpusak, ereduak eta datu-multzoak publikoki eskuragarri daude[25].

[25] `https://github.com/orai-nlp/low-scaling-laws`

5. KAPITULUA

Noraino Ikas dezake BERT batek Euskara bezalako Ordena-Malguko Hizkuntza Eranskari baten Gramatika?

Laburpena

Kapitulu honek hizkuntza-eredu neuronalek konpetentzia linguistiko formala eskuratzeko prozesua du aztergai, euskara bezalako gramatika konplexuko hizkuntzek aurrentrenamenduan zehar funtsezko erronkak dituzten hipotesitik abiatuz. Euskarak morfologia konplexua eta hitzen ordena malgua ditu bereizgarri eta ezaugarri horiek gramatika erauztea zaildu lezakete. Gure analisian, hainbat aurrentrenamendu konfiguraziotan entrenatutako BERT ereduen ezagutza gramatikala ebaluatu dugu, corpusaren tamaina, ereduaren tamaina, epoka kopurua eta lematizazioaren erabilera bezalako faktoreak kontuan hartuz. Ezagutza gramatikal hori neurtzeko, BL2MP (*Basque L2 student-based Minimal Pairs*) datu-multzoa eraiki dugu. BL2MPek esaldi bikoteak biltzen ditu, bata gramatikalki zuzena eta bestea okerra, gaitasun maila ezberdinetako euskara ikasleen idazlan zuzenduetatik erauziak. Bestalde, fenomeno gramatikal ezberdinen ikasketan dauden zailtasunak, hitzen ordena malguak dakartzan erronkak eta ikasleen mailaren eta BERTaren gaitasunaren arteko erlazioa esploratu dira.

5.1 Sarrera

Transformer arkitekturan oinarritutako hizkuntza-eredu neuronalek (*Neural Language Model*) giza hizkuntzarekin lotutako gaitasunak bereganatzeko gai direla erakutsi dute. Hizkuntza gaitasun horrek bi konpetentzia nagusi hartzen ditu barne: lexia eta gramatikan fokua jartzen duen gaitasun linguistiko formala eta gaitasun linguistiko funtzionala, zeinak arrazoiketa eta mundu ezagutza bezalako gaitasun kognitiboak dakartzan berarekin, mundu errealeko testuinguru batean hizkutza ulertzeko eta erabiltzeko ezinbestekoa. Hizkuntza-eredu neuronalek gaitasun hori bi arloetan erakutsi duten arren, ezagutza linguistiko formalean nabarmendu dira bereziki (Mahowald et al., 2023b).

Hizkuntza-eredu neuronalek askotariko gaitasun eta abilezia horiek aurrentrenamenduan zehar bereganatu ohi dituzte, entrenamendu corpusean topatutako patroiek eraginda. Gaitasun linguistiko formalaren ikasketan arreta jarriko bagenu, egitura gramatikal konplexuagoak dituen hizkuntza batek aurrentrenamendu garaian erronka gehigarriak izan ditzakeela pentsa genezake.

Badagokigu, orduan, corpusaren tamaina, ereduaren tamaina eta epoka kopurua bezalako aurrentrenamenduko aldagai ezberdinek, gramatikaren ikasketa prozesuan duten eragina aztertzea.

Ikerlan honetan, euskaran jarri dugu arreta, morfologia konplexua (ikus 5.3. Taula) eta hitzen ordena malgua dituelako: *"Katuak sagua jan zuen"* esaldia beste bost ordena baliokidetan berridatz daiteke: *"Katuak jan zuen sagua"*, *"Sagua jan zuen katuak"*, *"Sagua katuak jan zuen"*, *"Jan zuen sagua katuak"*, eta *"Jan zuen katuak sagua"*. Beste hainbat hizkuntzarekin elkarbanatzen dituen ezaugarri horiek, teorian, hizkuntza-ereduaren gramatikaren ikasketa prozesuan traba izan daitezke.

Gure analisian, hainbat aurrentrenamendu konfigurazio ezberdinen pean entrenatutako ereduen ezagutza gramatikala ebaluatu dugu, BERT arkitektura erabiliz. Ezagutza hori neurtzeko asmoz, BL2MP (*Basque L2 student-based Minimal Pairs* edo *Euskara L2 ikasleetan oinarritutako Pare Minimoak*) test multzoa sortu da. Datu-multzo hau euskara ikastaroetako ikasleek idatzitako idazlanetan oinarrituta eraiki dugu. Idazlan horiek akats gramatikalen zuzenketak (irakasleek eginak) dituzte osagarri, esaldi zuzen eta okerren pareak erauzteko baliagarriak zaizkigunak. Pare hauek profitatuz, BL2MP

datu-multzoa eraiki dugu. Honenbestez, datu-multzo horretan nabarmentzen diren fenomeno gramatikalek euskara ikasleek topatutako erronkak islatzen dituzte.

Aurrentrenamendu konfigurazio ezberdinen artean aztertutako aldagaiak honakoak izan dira:

- **Entrenamenduko corpus tamaina**: Zertan eragiten du entrenamenduko corpus tamainak gramatika ikaste prozesuan?
- **Ereduaren tamaina**: Ereduaren parametro kopurua handitzeak hobekuntzarik badakar gramatikaren barneratzean?
- **Epoka-anitzeko entrenamendua**: Gramatika barneratzea sendotu egiten ote da hainbat epokaz entrenatzean?
- **Tokenizazioa**: Euskara bezalako hizkuntza flexiboetan, lematizazioan oinarritutako tokenizazio batek lagun al lezake gramatika hobeto barneratzen?

Gainera, gramatika ikastearen prozesuaren eta honako aldagaien arteko erlazioa ere aztertu dugu:

- **Hitz-ordena**: Hitz ordena malgua izateak oztopatzen al du ikasketa prozesua?
- **Fenomeno gramatikala**: Bada ikasteko errazago edo zailago den fenomeno gramatikalik?
- **L2 ikasleen gaitasuna**: Bada loturarik L2 ikaslearen gaitasun mailaren eta hizkuntza-ereduak fenomeno gramatikal zehatz bat ikasteko dituen zailtasunen artean?

Lan honetan sortutako BL2MP datu-multzoa, aurrentrenamenduko corpusak[1], aurrentrenatutako hizkuntza-ereduak eta ebaluazio kodea, publikoki eskuragarri daude[2].

[1] Bere horretan eta lematizatua.

[2] `https://github.com/orai-nlp/bl2mp`

5.2 Datu-Multzoa

BLIMP edo *Benchmark of Linguistic Minimal Pairs* (Warstadt et al., 2020a) datu-multzoaren tankerakoak, baliagarri dira hizkuntza-ereduen gramatika eta sintaxi gaitasunak ebaluatzeko euskarri gisa. BLIMPen pare-minimo (*minimal-pair*) deituriko esaldi bikoteak daude, alderdi linguistiko bakar baten diferentzia dutelarik. Pare bakoitzak gramatikalki zuzena den esaldi bat eta akats bat duen beste bat ditu. Horrek, hizkuntza-eredu batek alde sotil bat bakarrik duten bi aldaerak bereizteko gaitasuna neurtzeko aukera ematen du.

Kapitulu honetan, BL2MP datu-multzoa aurkezten dugu, euskara hizkuntzaren ezagutza gramatikala ebaluatzeko diseinatua, ingeleserako BLIMP datu-multzoan inspiratuta. BL2MP datu-multzoak, bai&by[3] hizkuntza akademiatik bildutako adibideak biltzen ditu, akademiako ikasleek idatzitako idazlanetatik eta irakasleek egindako zuzenketetatik eratorriak. Adibide horiek benetako akats gramatikal aberatsak eskaintzen dituzte, ikasleek egiaz egindako akatsen erakusle, hortaz, mundu errealeko hizkuntza akatsen benetako isla.

Errore mota	Maila	esaldi #
E1: deklinabidea	A	200
	B	200
	C	200
E2: aditza	A	200
	B	200
	C	200
E3: egitura eta ordena	A	200
	B	200
	C	200
Guztira		1,800

5.1. Taula: BL2MP datu-multzoaren estatistikak, mailaren (A, B, C) eta errore motaren arabera sailkatuta. Mota bakoitzak 600na esaldi-bikote ditu, 3 mailatan banatuta, guztira 1800 esaldi-bikote.

bai&by akademiak hornitutako ikasleen idazlanetatik 1800 esaldi akastun ausaz aukeratu ditugu, pare-minimoen irizpidea hertsiki betez. Anizta-

[3] `https://www.baiby.com`

sun orekatu bat bermatzeko, hiru gaitasun mailen[4] (A: hasiberria, B: erdimailakoa, eta C: aurreratua) araberako eta hiru errore moten araberako (E1: deklinabidea, E2: aditza, E3: egitura eta ordena) distribuzio berdina mantendu da, 5.1. Taulak erakusten duen moduan. Hurbilpen honek, datu-multzoan gaitasun mailaren eta errore moten aniztasuna ordezkatzea izan du helburu. 5.2. Taulak errore mota bakoitzeko pare-minimoen adibide bana dakar.

E1: Deklinabidea

Euskararen morfologian ezaugarri bereizgarri bat deklinabidea deituriko fenomeno gramatikal baten bitartez agertzen da, esaldiko sintagma guztiei atzizkiak eranstean datzana.

Atzizki hauetako bakoitza kasu gramatikal jakin bati dagokio, eta beste hizkuntza batzuetan preposizioek duten rola betetzen dute. Kasu hauek nork, nor, nori, norekin, norentzat, nora edo nondik bezalako galdegaiei erantzuten diete, besteak beste. Gainera, kasu bakoitzak 4 forma bereizi har ditzake: mugagabea, mugatu singularra, mugatu plurala eta mugatu plural hurbila.

E2: Aditza

BL2MP datu-multzoko aditzekin lotutako akatsek bost alderdi hartzen dituzte barne:

Pertsona: aditza jokatzean, pertsona, akzioa burutzen ari den subjektuaren eta aditzaren arteko erlazioa adierazten duen kategoria gramatikalari deritzo. Hiru pertsona daude, lehen pertsona, bigarren pertsona eta hirugarren pertsona.

Numeroa: aditza jokatzean, numeroak subjektua singularra edo plurala den adierazten du.

Denbora: aditz denborek ekintza aldi zehatz batean kokatzen dute: orainaldia, lehenaldia eta geroaldia.

Modua: euskaraz, aditzak, hizlariak akzioarekiko duen jarrera edo asmoa ere adierazten ditu. Indikatiboak gertaerak eta adierazpenak modu inpartzial eta neutro batean adierazten ditu. Subjuntiboa hizlariaren desirak, helburuak edo asmoak aditzera emateko erabiltzen da. Agintera, berriz, aginduak emateko.

[4] `www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions`

Aspektua: aditzaren aspektuak ekintza amaitua dagoen edo martxan den definitzen du. Euskarak hiru aspektu nagusi ditu: burutua, burutu gabea eta gertakizuna.

Akats-mota	Esaldi okerrak	Esaldi zuzenak
E1: deklinabidea	*Nik oso pozik nago.* *Medikua gisa lan egiten dute.* *Zorte on eta igandean arte!*	*Ni oso pozik nago.* *Mediku gisa lan egiten dute.* *Zorte on eta igandera arte!*
E2: aditza	*Bai, nik daukat zure autoaren giltzak.* *Gero ondoko parkean jolastuko dugu.* *Zalantzan bazaude?*	*Bai, nik dauzkat zure autoaren giltzak.* *Gero ondoko parkean jolastuko gara.* *Zalantzan zaude?*
E3: egitura eta ordena	*Balkoitik oso ederra bista daukat.* *Oso pozik nago etxearekin, baina urduri apur bat nago dekorazioarekin.* *Ziur nago badakizula zer gertatu zela, baina kontatuko dizut irakurri nuen berria.*	*Balkoitik oso bista ederra daukat.* *Oso pozik nago etxearekin, baina apur bat urduri nago dekorazioarekin.* *Ziur nago badakizula zer gertatu zen, baina kontatuko dizut irakurri nuen berria.*

5.2. Taula: BL2MP datu-multzoko akats mota bakoitzeko pare minimoen adibideak, bata gramatikalki zuzena, eta bestea okerra. Akatsak eta zuzenketak azpimarratuta.

E3: Egitura eta Ordena

Akatsen hirugarren kategoria euskarazko esaldien egiturari eta ordenari dagokio. Kategoria honen barruan, akatsak bi mota nagusitan bana genitzake.

Lehen multzoan esaldien ordena gramatikala zuzen erabiltzeri lotutako akatsak daude. Azpimarratu beharrekoa da, hizkuntza eranskari gehienetan bezala (Mahowald et al., 2023a), euskarak ere esaldiaren ordena nahiko malgua duela (Laka, 1996), zenbait egitura sintaktikoren barruan malgutasun hori mugatua izan arren, adibidez, izen sintagmetan (Laka, 1996).

Bigarren multzoak, esaldi konposatuetan elementuen kokapenari edo loturari dagokien akats gramatikalak biltze ditu, adibidez, esaldi konpletiboak, zehar-galderak, esaldi erlatiboak eta beste hainbat.

Hona adibide batzuk:

- *Balkoitik oso **ederra bista** (→bista ederra) daukat.*
- *Oso pozik nago etxearekin, baina **urduri apur bat** (→apur bat urduri) nago dekorazioarekin.*
- *Ziur nago badakizula zer gertatu **zela** (→zen), baina kontatuko dizut irakurri nuen berria.*

5.3 Metodologia

5.3.1 Aurrentrenamendu Corpusak

Zhang et al. (2021) lanak erakutsi bezala, 10M eta 100M hitz arteko aurrentrenamendu corpusa nahikoa da hizkuntza-eredu batek sintaxia eta semantikaren gaitasunak ia osorik bereganatzeko ingelesez. Hortaz, euskararako eskala bereko hiru corpus aukeratu ditugu. Urbizu et al. (2023a) lanean definitutako aurrentrenamenduko corpusak erabili ditugu, 5M, 25M eta 125M hitz dituztenak hain zuzen, elkarren artean ratio konstantea mantenduz. Corpuseko testuen domeinuei dagokienez, %75a albisteek eta gainerako %25a Wikipediako testuek osatzen dute.

5.3.2 Eredu Arkitektura

Hiru tamainatako BERTak erabili ditugu: BERT_{12L}, BERT_{8L} eta BERT_{4L}, hamabi, zortzi eta lau ezkutuko geruza dituztelarik hurrenez hurren, gainontzeko parametroak[5] ere proportzionalki uzkurtuz.

50K tokeneko hizpeko hiztegi bat erabili da, letra-larri eta xeheak bereiziz, eta Kudok (2018) proposatutako unigrametan oinarritutako hizpeko segmentazio algoritmoa erabiliz entrenatua izan da[6]. Sortutako ereduek 124M,

[5]Ezkutuko dimentsioa, atentzio-buru kopurua, etab.

[6]Aurrentrenamenduko corpus bakoitzarekin berrentrenatua.

51M eta 16M parametro dituzte, hurrenez hurren. Ereduak defektuzko hiperparametroak erabiliz entrenatu dira[7], hitz osoen maskaratzea aplikatuz MLM helburu-funtzioan (10eko dup-faktorea[8] erabilita), 256ko labekada tamainarekin eta 512ko sekuentzia luzerarekin, 500K urrats inguruz, v3-8 TPU makinetan.

125M hitzeko corpusaren eredua 640.000 urratsez (512 epoka) entrenatu da, 25Mko hitzeko corpusaren eredua 512.000 urratsez (2048 epoka), eta azkenik, 5Mko corpusaren eredua 409.600 urratsez (8192 epoka). Kasu guztietan, beroketa fasea entrenamenduaren %6,25ean finkatu da, Izsak et al. (2021) lanaren gomendioak jarraituz. Hots, 125Mko corpusarentzat 40.000 urrats (32 epoka), 25Mkoarentzat 32.000 urrats (128 epoka) eta 5Mkoarentzat 25.600 urrats (512 epoka).

5.3.3 Ebaluazio Metodoa

BERT arkitektura erabiliz esaldiak ebaluatzeko Salazar et al. (2020) lanean proposatutako esaldiak osorik puntuatzeko metodologia jarraitzea erabaki dugu. Hurbilpen horrek, hizkuntza-eredu neuronalen konpetentzia linguistikoak neurtzeko bide ematen du, maskara altxatako tokenak iragartzen entrenatutako ereduak (BERT, RoBERTa) kaltetu gabe, beste metodo alternatibo batzuekin gertatzen ez den bezala, Zaczynska et al. (2020) laneko puntuatze holistikoa kasu.

Salazar et al. (2020) lanean aurkeztutako puntuatze metodologiak sasilogegiantza (*Pseudo-Log-Likelihood* edo PLL) kalkulatzen du hizkuntza-ereduetatik. Puntuazioak esaldiko token bakoitzaren $logP_{MLM}(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t})$ log probabilitate kondizionalen baturaren bitartez eratortzen dira. Prozesu honek, w_t token bakoitza, txandaka, BERTen [MASK] tokenarekin maskaratzea dakar, eta ondoren token originalaren probabilitatea kalkulatzea esaldiko gainerako testuingurua emanik $W_{\setminus t}$.

$\boldsymbol{W}$ esaldiaren PLL puntuazioa ondorengoa litzateke:

$$\mathrm{PLL}(\boldsymbol{W}) := \sum_{t=1}^{|\boldsymbol{W}|} \log P_{\mathrm{MLM}}(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t}; \Theta).$$

[7] Ikasketa-tasa(lr) = $1e^{-4}$, $\beta 1 = 0{,}9$, $\beta 2 = 0{,}999$, L2 w.d.=0,01

[8] Sarrerako datuak zenbat aldiz errepikatuko diren (maskara ezberdinekin).

Non, Θ ereduaren parametroei dagokien.

Esaldien sasilog-egiantz puntuazioak kalkulatzeko *minicons*[9] (Misra, 2022) tresna erabili dugu. Datu-multzoko esaldi bien sasilog-egiantzak kalkulatu ditugu, bai gramatikalki zuzenaren eta baita gramatikalki okerrarena ere. Pareminimo bakoitzarekin, lortutako bi sasilog-egiantzak konparatu ditugu, esaldi zuzena egoki identifikatu duela iritziz, berori izan bada bietatik puntuazio altuena jaso duena. Honenbestez, asmatze-tasa zuzenki identifikatutako esaldi zuzenen proportzio moduan kalkulatu dugu.

Tokenizazio osteko luzera ezberdina zuten bikoteak baztertu egin ditugu, Zaczynska et al. (2020) lanean alemanarekin egin moduan, esaldiak tokenizatu osteko luzeran dauden aldeek emaitzetan eraginik ez dutela bermatzeko.

5.4 Esperimentuak

5.4.1 Zein Eragin dute Corpus Tamainak, Ereduaren Tamainak eta Epokek Gramatika Ikastean?

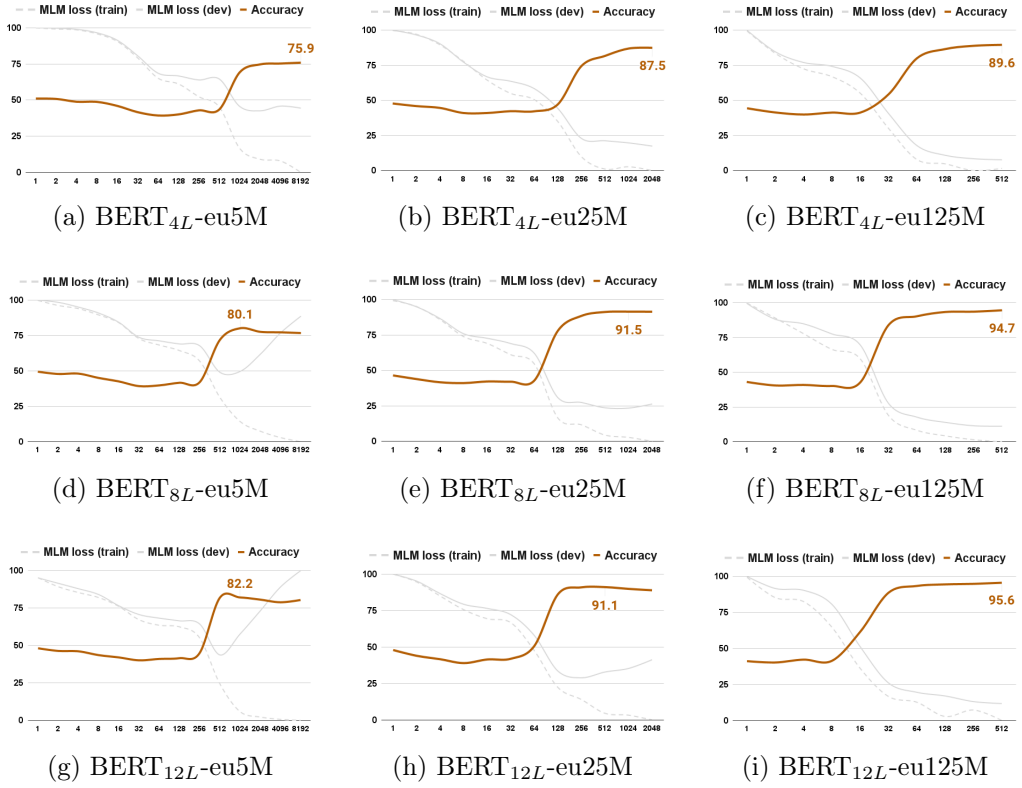
Lehen esperimentu honetan, konfigurazio ezberdinekin aurrentrenamenduetatik eratorritako zenbait eredu konparatu ditugu. Konfigurazio hauek corpus tamainan eta ereduaren parametro kopuruan aldatzen dira. Esperimentu honen helburuak aurrentrenamenduko corpusaren tamainak gramatikaren ezagutza barneratzean zein eragin duen ulertzea eta ea ereduen parametro kopurua handitzeak gramatikaren ezagutzan hobekuntzarik badakarren ebaztea da, eta hala balitz zenbatekoa zehaztea.

Gainera, ereduak gramatika barneratzen noiz hasten diren behatu dugu hainbat kontrol-puntu ebaluatuz.

Helburu hori lortzeko, 5.3.2. Atalean deskribatutako tamainetako bakoitzarekin (12, 8 eta 4na geruza) BERT eredu bana entrenatu dugu 5.3.1. Atalean aipatutako corpus bakoitzarekin (5M, 25M eta 125M hitzekoak). 5.1. Irudiak bederatzi ereduen aurrentrenamenduko kontrol-puntu ezberdinek BL2MP datu-multzoan lortutako zehaztasunak biltzen ditu.

Gure esperimentuetan, gutxi gorabehera, adibideen herena baztertuak izan dira ebaluazio garaian, esaldi pareak tokenizatzean luzera ezberdina izateagatik.

[9] `https://github.com/kanishkamisra/minicons`



(a) BERT$_{4L}$-eu5M (b) BERT$_{4L}$-eu25M (c) BERT$_{4L}$-eu125M

(d) BERT$_{8L}$-eu5M (e) BERT$_{8L}$-eu25M (f) BERT$_{8L}$-eu125M

(g) BERT$_{12L}$-eu5M (h) BERT$_{12L}$-eu25M (i) BERT$_{12L}$-eu125M

5.1. Irudia: 5M, 25M eta 125M hitzeko corpusekin aurrentrenatutako BERT$_{12L}$, BERT$_{8L}$ eta BERT$_{4L}$ ereduen zehatasunak BL2MPen, entrenamenduko epokatan zehar (eskala esponentzialean). Eredu bakoitza 500K urrats inguruz entrenatu da (5.3.2. Atala). Grisezko lerroek, entrenamenduko eta garapeneko multzoen galera normalizatua (min-max) adierazten dute.

Grafikoek erakusten duten legez, aurrentrenamenduko corpusaren tamaina handitzeak ereduen BL2MPeko emaitzak nabarmen hobetzen ditu. Bestalde, ereduaren tamaina handitzearekin emaitzetan lortutako hobekuntza apalagoa da. Igoera adierazgarria dago ereduaren tamaina BERT$_{4L}$etik BERT$_{8L}$-era igotzean. Haatik, ereduaren tamaina BERT$_{12L}$raino handitzeak, hobekuntza txikia dakar.

Eredu guztietan ikus dezakegu zehaztasuna %50aren ingurutik abiatzen dela (ausaz asmatzearen pare), jarraian, lehen epoketan, beherantz egiten duela zertxobait, eta ondoren, beroketa-fasearen amaiera aldera, igotzen has-

ten dela ereduak ikasten duen heinean, aurrentrenamenduko eta garapeneko galera kurbek adierazten duten bezala. Hobekuntza horrek jarraitu egiten du, eredu bakoitzaren zehaztasun maximora iritsi bitartean, zeina, eredu bakoitzaren arabera, epoka ezberdinetan gertatzen den. Zenbat eta entrenamenduko corpus handiagoa izan, orduan eta goizago ikus daiteke epoketan BL2MPen emaitzen hobekuntza. Ez da soilik onuragarria, baizik eta beharbeharrezkoa da hainbat epokaz entrenatzea eredua euskararen gramatika ikas dezan, lan honetan aztertu diren baliabide urriko corpusen testuinguruan, bederen. Eredu batzuen aurrentrenamenduan gaindoitze zantzuak hauteman daitezke, batez ere, 5M hitzeko corpusarekin entrenatutako ereduen artean, eta, maila apalagoan, 25M hitzeko corpusarekin entrenatutako eredu handienetan, garapeneko galera kurbek erakusten duten moduan. Hala eta guztiz ere, gaindoitze horrek ez dirudi gramatikaren ikasketan oztopoa denik.

Aurrez aipatu bezala, BERT_{8L} nabarmen gailentzen da BERT_{4L}en aurrean corpus tamaina guztietan zehar. Hala ere, ereduaren tamaina BERT_{8L}-etik BERT_{12L}era igotzeak ez du lortutako zehaztasunean hobekuntza esanguratsurik. Horrenbestez, egin beharreko esperimentuen kopurua neurripean mantentzeko, eredu tamaina bakarrarekin jarraitzea erabaki da, gainerako esperimentuetan BERT_{8L} erabiliz, konputazionalki arinagoa delako alde batetik, eta bestetik, ia BERT_{12L}ak bezain emaitza onak ematen dituelako.

Izan badira, BLIMPen eta BL2MPen pareko datu-multzoak beste hizkuntza batzuetarako, 2.5.2. azpiatalean azpimarratu bezala. Testerako datumultzo horietako bakoitza, ordea, metodologia ezberdinak jarraituz eraiki da eta fenomeno linguistiko ezberdinetan jartzen du arreta. Honenbestez, datumultzo hauek izan ditzaketen zailtasun mailak ez dira konparagarriak, eta euskararako lortutako emaitzak gainerako hizkuntzekin konparatzea galarazten du.

5.4.2 Bada Hizkuntza-Ereduentzat Ikasteko Zailagoa den Fenomeno Gramatikalik?

5.4.1. Atalean, BERT ereduak gramatika barneratzeko gai direla ikusi dugu, eta gaitasun hau nola ereduaren eta aurrentrenamendu corpusaren tamaina handitu ahala nola hobetzen den. Hala ere, ba ote da hizkuntza-ereduentzat ikasteko zailagoa den Fenomeno Gramatikalik?

Galdera hau erantzuteko, BERT_{8L} ereduaren emaitzak berraztertu ditugu, BL2MP datu-multzoko errore gramatikal mota ezberdinetan arreta jarriz. BL2MPen akats gramatikalak ondorengo hiru fenomeno gramatikaletan sailkatzen dira: deklinabidea (E1), aditza (E2) eta egitura eta ordena (E3). Errore mota bakoitzaren emaitzak 5.4. Irudiko a, b eta c grafikoetan daude ikusgai.

Eredu guztien grafikoek hiru motetako errore gramatikalak epoka berdinen inguruan ikasten direla erakusten dute, zehaztasun maila ezberdinetara iritsi arren. Zehazki, 5M hitzeko corpusarekin entrenatutako BERT_{8L} ereduak emaitza hobeak lortzen ditu egitura eta ordenan (E3) deklinabidearekin (E1) eta aditzarekin (E2) alderatuz. 83,3ko zehaztasunera iristen da E3n 1024. epokan, eta E1 eta E2n, ostera, 79,0 eta 77,4ko zehaztasunak lortzen ditu, hurrenez hurren.

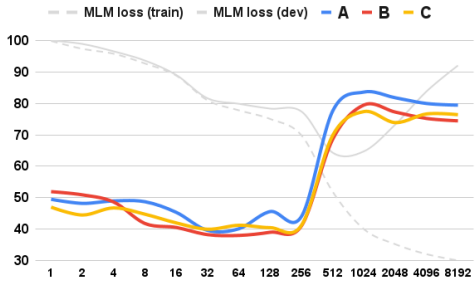
5.4.3 Bat Datoz L2 Ikasleen Eginahalak BERTarekin?

Ikasleen gaitasun mailaren eta hizkuntza-ereduak ikasle horiek egindako akatsetan izan dituen zailtasunaren artean ea korrelaziorik ba ote den zehaztea dugu helburu esperimentu honetan. Horretarako, 5M, 25M eta 125M hitzeko corpusekin entrenatutako 8 geruzako BERT ereduen (BERT_{8L}) zehaztasunak bildu ditugu 5.2. Irudian. Zehaztasun hauek, BL2MPeko adibideen jatorri diren L2 hizkuntza ikasleen gaitasun mailaren arabera daude multzokatuta.

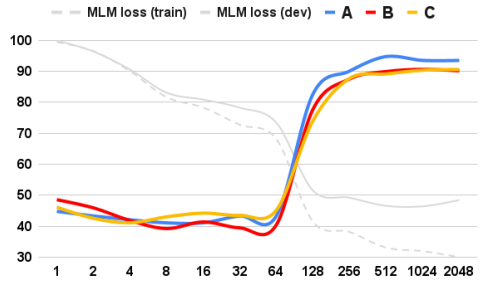
Grafikoek, hizkuntza-ereduek gaitasun maila guztietako ikasleen fenomeno gramatikal guztiak aldi berean ikasten dituztela erakusten dute, pareko ikasketa mugarrietara helduz epoka berdinaren inguruan, betiere aurrentrenamenduko corpus bakoitzaren menpe.

Maila baxueneko L2 ikasleek eginiko akatsak biltzen dituen A datu-multzoan hizkuntza-ereduek lortutako emaitzek aldiz, B eta C datu-multzoenak gainditzen ditu. Alde horiek murrizten doaz aurrentrenamenduko corpusa handitu ahala, eta 125 miloi hitzetara heltzean zailtasun maila guztietako pare-minimoetan emaitza konparagarriak lortzen dituela esan daiteke.

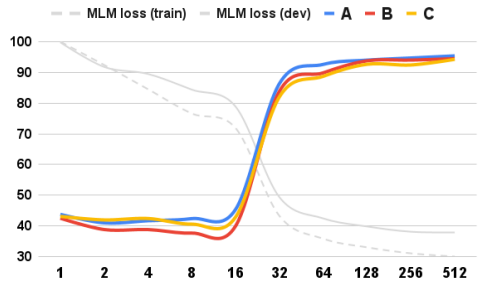
Ikasle aurreratuenen idazlanetako akatsak biltzen dituen C datu-multzoa B datu-multzoa baino zailagoa izan zitekeela espero arren, 8 geruzako BERT eredua (BERT_{8L}) C datu-multzoan B datu-multzoan bezain ondo dabil, aurrentrenamenduko corpus tamaina orotan.



(a) BERT$_{8L}$-eu5M



(b) BERT$_{8L}$-eu25M



(c) BERT$_{8L}$-eu125M

5.2. Irudia: 5M, 25M eta 125M hitzeko corpusekin entrenatutako BERT$_{8L}$ ereduaren zehaztasunak, adibideen jatorrizko L2 ikasleen mailaren arabera. Lerro grisek entrenamendu eta garapen multzoen gaineko galera normalizatua (min-max) adierazten dute.

5.4.4 Hitz-Ordena Malguak Gramatika Ikastea Oztopa dezake?

Esperimentu honekin, aurrentrenamendu garaian hitzen ordena malgua euskararen gramatika ikasteko prozesuan oztopo ote den argitu nahi da. Zehazki, aztertu nahi da ea hizkuntza-ereduek pare-minimoetan agertzen diren fenomeno gramatikalak ondo barneratzen dituzten, esaldiko hitzen ordena edozein izanik ere.

Kontu hau argitzeko, BL2MP datu-multzotik 200 pare-minimoko azpimultzo bat erauzi da, 200_BL2MP_1 aurreratzean.

Azpimultzo honek pare-minimo bakoitzarentzat hitz-ordena aldaera bakarra du. Jarraian, 200_BL2MP_1 datu-multzotik abiatuz, bigarren ebaluazio-

multzo bat eraiki dugu gramatikalki baliokidea den eta hitzen ordena ezberdina duen pare-minimo berri bat gehituz jatorrizko pare minimoari. Adibidez, [*Katuak sagua jan zuen. / Katuak sagua jan zen.*] paretik abiatuz, [*Sagua katuak jan zuen. / Sagua katuak jan zen.*] pare berria sortu dugu.

Bigarren ebaluazio-multzo honi, 200_BL2MP_2 deituko diogu, eta pare-minimo bakoitzeko bi hitz-ordena aldaera ditu: bata jatorrizkoa [*Katuak sagua jan zuen. | Katuak sagua jan zen.*] eta bestea eskuz sortutako aldaera berria [*Sagua katuak jan zuen. | Sagua katuak jan zen.*].

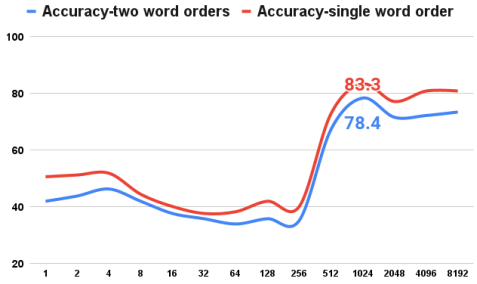
Ondorengo ebaluazioak 200_BL2MP_1 eta 200_BL2MP_2 ebaluazio-multzoetan burutu dira. 200_BL2MP_1 datu multzoaren ebaluazioan zehaztasuna pare bakoitzeko hitz-ordenaren aldaera bakarraren gainean kalkulatzen da (*Accuracy-single word order* 5.3. Irudian). Bestalde, 200_BL2MP_2 ebaluazio-multzoan hitz-ordenaren bi aldaerak hartzen dira kontuan zehaztasuna kalkulatzeko garaian *(Accuracy-two word orders* 5.3. Irudian). Hau da, pare-minimo bat asmatutzat emateko hizkuntza-eredua gai izan behar da gramatikalki zuzena den esaldia lehenesteko bi hitz-ordena aldaeratan.

5.3. Irudiko b eta c grafikoek erakusten diten moduan, BERT8L-eu25M eta BERT8L-eu125M gai dira errore gramatikal berbera zuzen identifikatzeko esaldi berrordenatuen aldaera ezberdinetan zehar galera txiki batekin (2,5 puntu ingurukoa), hitz-ordena bakarreko emaitzekin alderatuz.

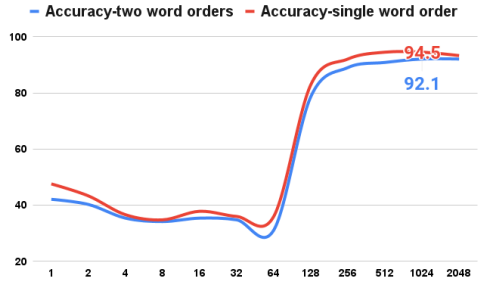
Galera hori nabarmenagoa da BERT$_{8L}$-eu5M ereduaren kasuan, non 5,1 puntuko galera antzematen den (5.3. Irudiko a grafikoan ikus daitekeenez), hitzen ordena malgua hizkuntza-ereduek euskararen gramatika barneratzeko orduan oztopo bat dela iradokiz, batez ere aurrentrenamenduko datuak urriak direnean.

5.4.5 Lematizatzeak Lagun lezake Hizkuntza Eranskarien Gramatika Ikasten?

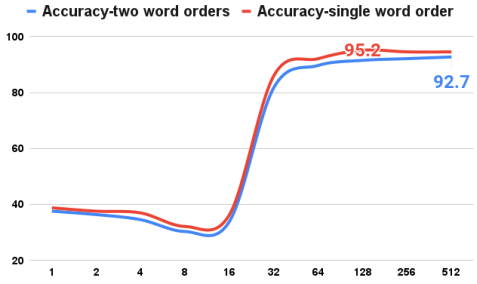
Morfologia konplexua ez duten hizkuntzetan defektuzko hizpeko hiztegi bat erabiltzea nahikoa izan ohi den arren, hizkuntza eranskarien kasuan, itzulpen automatikoan eta hizkuntza-eredu neuronaletan lematizazioa onuragarri izan daiteke (Pan et al., 2020; Mohseni and Tebbifakhr, 2019; Nzeyimana and Rubungo, 2022). Eredu neuronalari lagungarri zaio sarrerako testua lematizatua izatea, erregulartasun eta patroi gramatikalak hobeto ikasteko, batik bat datu baliabide urriko inguruneetan (Pan et al., 2020).



(a) BERT$_{8L}$-eu5M



(b) BERT$_{8L}$-eu25M



(c) BERT$_{8L}$-eu125M

5.3. Irudia: 5M, 25M eta 125M hitzeko corpusekin entrenatutako BERT$_{8L}$ ereduaren zehaztasunak 200 adibideko 200_BL2MP_1 eta 200_BL2MP_2 azpimultzoetarako. 200_BL2MP_1 azpimultzoan hitz-ordena aldaera bakarra (*Accuracy-single word order*, gorriz) aztertzen da; 200_BL2MP_2 azpimultzoan hitz-ordenaren bi aldaera (*Accuracy-two word orders*, urdinez).

Honenbestez, hurbilpen hau aztertu nahi izan dugu, euskararako ea lematizazioak BERT ereduen gramatikaren ikasketa prozesua areagotzen duen ebazteko.

Helburu horrekin, *Eustagger* euskararako analizatzaile morfosintaktikoa (Ezeiza et al., 1998) erabiliz aurrentrenamendu corpusa aurre-prozesatu da, hitzak euren lema eta deklinabide atzizkietan banatzeko (ikus adibidea 3. Taulan). Bereziki izen eta adjektiboetan zentratu den segmentazio honen bitartez, euskararen konplexutasun morfologikoa txikitu nahi izan da, ingelesa eta gaztelania bezalako morfologia konplexurik gabeko hizkuntzen maila batera murriztuz.

Kasu gramatikal berberari dagozkion atzizkiak beti ez datoz bat (adib.,

"ate**an** dago" -> "ate NUMS_MUGM_INE dago", "aldapa**n** dago -> aldapa NUMS_MUGM_INE dago"). Morfologia hau estandarizatzeko, atzizki ohikoena hartu dugu kategoria morfosintaktiko bakoitzaren ordezkari gisa. NUMS_MUGM_INEren (inesibo singularra) kasuan, adibidez, "an" atzizkia da ohikoena corpusean. Beraz, arestian aipatutako segmentazioa, honela geratuko litzateke: "ate**an** dago" -> "ate **an** dago" eta "aldapa**n** dago" -> "aldapa **an** dago".

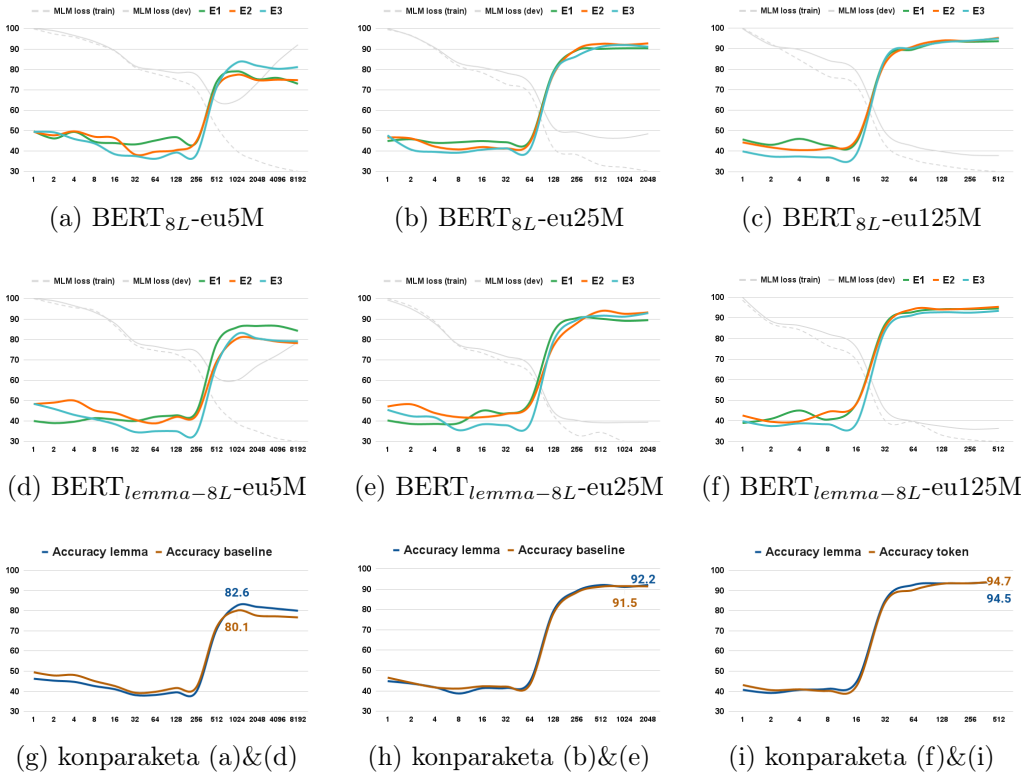
Jatorrizkoa	Etxe**ko** ate**an** dago
Lemma eta morfologia	Etxe NUMS_MUGM_GEL ate NUMS_MUGM_INE dago
Segmentazioa	Etxe ko ate an dago

5.3. Taula: Aurrentrenamendu corpusa tokenizatu aurretik aplikatutako segmentazio morfologikoaren adibidea. Etiketa morfosintaktikoak dagozkien atzizki esanguratsuenarekin ordezkatzen dira ("ko" atzizkia "NUMS_MUGM_GEL" etiketarentzat eta "an" atzizkia "NUMS_MUGM_INE" etiketarentzat). "NUMS_MUGM_GEL" etiketak lekuzko genitibo singularra adierazten du, eta "NUMS_MUGM_INE" etiketak inesibo singularra.

Esperimentu honetan, sei BERT_{8L} eredu ebaluatu dira, hiru corpus originalarekin aurrentrenatuak, 5M, 25M eta 125M hitzeko corpusekin, eta beste hiru eredu konparagarri, corpusaren aldaera lematizatua erabiliz. Emaitzak 5.4. Irudian bildu ditugu.

g h eta i grafikoek erakusten dute nola testu lematizatuarekin entrenatutako BERT ereduak, oro har, testu originalarekin entrenatutako ereduei gailentzen zaien, 5M hitzeko corpusarekin entrenatutako ereduen kasuan 2 puntutik gorako hobekuntza antzematen da. Hala eta guztiz ere, lematizazioaren abantaila murriztuz doa corpusaren tamaina areagotu ahala. 25Meko corpusaren kasuan, lematizazioak puntu erdiko aldeaz bakarrik hobetzen ditu emaitzak, eta 125Meko corpusean ez da hobekuntzarik sumatu testu originala darabilen eta %94,7ko zehaztasuna lortzen duen ereduarekin alderatuz.

5M hitzekin aurrentrenatutako ereduen emaitzak errore motaren arabera sakonago aztertzen baditugu (a&d grafikoek ilustratua), oinarrizko ereduak E3 (egitura eta ordena) erroreak ebazten, E1 (deklinabidea) eta E2 (aditza) erroreetan baino hobeto funtzionatzen duela ikus daiteke. Lematizazioa



(a) BERT$_{8L}$-eu5M (b) BERT$_{8L}$-eu25M (c) BERT$_{8L}$-eu125M

(d) BERT$_{lemma-8L}$-eu5M (e) BERT$_{lemma-8L}$-eu25M (f) BERT$_{lemma-8L}$-eu125M

(g) konparaketa (a)&(d) (h) konparaketa (b)&(e) (i) konparaketa (f)&(i)

5.4. Irudia: 5M, 25M eta 125M hitzeko corpusekin entrenatutako BERT$_{8L}$ ereduaren eta lematizatutako aldaeraren zehaztasunak BL2MP datu-multzoan, fenomeno gramatikalaren arabera. E1: deklinabidea, E2: aditza, E3: egitura eta ordena. g, h and i grafikoek a-d, b-e eta c-f grafiko bikoteen zehaztasun metatuak biltzen dituzte, hurrenez hurren. Azpialdeko lerro grisek entrenamendu eta garapen multzoen gaineko galera normalizatua (min-max) adierazten dute.

erabiliz, ordea, E3 erroreetan antzeko zehaztasuna lortzen da, E2 erroreetan hobekuntza txiki bat nabari da, E3ren zehaztasunera gerturatuz. Horrez gain, E1 motako erroreetan hobekuntza handi bat ikus daiteke lematizazioa aplikatzean, E2 eta E3 erroreen emaitzak gainditzen dituelarik.

Aurrentrenamenduko corpusaren tamaina 25M hitzetara handituz (b&e grafikoak), errore moten arteko aldea, eta lematizazioak dakarren abantaila txikitzen doa, eta erabat desagertzen dira corpusa 125M hitzetara handitzean (c&f grafikoak).

Eredua	Corpusa	Zehaztasuna
$BERT_{8L}$-eu25M	25M	91.5
$BERT_{8L}$-eu125M	125M	94.7
$BERT_{12L}$-eu125M	125M	95.6
BERTeus	225M	94.8
Roberta-euscrawl-B	289M	95.9
Roberta-euscrawl-L	289M	95.5
ElhBERTeu-medium	351M	95.4
ElhBERTeu-base	351M	**96.5**
mBERT	∼40M	58.9
XLM-RoBERTa base	∼270M	75.1
XLM-RoBERTa large	∼270M	82.4
IXAmBERT	225M	94.4

5.4. Taula: Euskararako MLM eredu elebakar eta eleaniztunen zehaztasun emaitzak BL2MP datu-multzoan. Lan honetako ereduak, eredu elebakarrak eta eredu eleaniztunak multzokatuta.

5.4.6 Kodetzaile Eredu Publikoekin Konparaketa

Azkenik, kapitulu honetan sortutako BL2MP ebaluaziorako datu-multzoan ebaluatu ditugu euskararako publikoki eskuragarri dauden kodetzaile motako eredu elebakar eta eleaniztun gehienak. 5.4. Taulak hainbat ereduren zehaztasunak biltzen ditu, besteak beste, lan honetako hiru eredurik onenak, BERTeus (Agerri et al., 2020), ElhBERTeu-[*base*|*medium*] (Urbizu et al., 2022) eta Roberta-euscrawl-[*base*|*large*] (Artetxe et al., 2022) euskararako hizkuntza-eredu elebakarrenak; eta mBERT (Devlin et al., 2019), XLM-RoBERTa [*base*|*large*] (Conneau et al., 2020) eta IXAmBERT (Otegi et al., 2020) hizkuntza-eredu eleaniztunenak.

5.4. Taulak erakusten duenez, eredu elebakar guztiek ($BERT_{8L}$-eu25Mek salbu) antzeko emaitzak lortzen dituzte, %95eko zehaztasunaren bueltan. Lan honetan entrenatutako $BERT_{8L}$-eu125M eta $BERT_{12L}$-eu125M ereduak artearen egoerako gainerako euskararako ereduekin lehiakorrak dira, 125M hitzeko corpus txikiago batekin entrenatuak izan diren arren. Guztien artetik, corpus handienarekin entrenatua izan den ElhBERTeu-base da emaitzarik onenak lortzen dituen eredua. Eredu eleaniztunen artean, hiru hizkuntzekin elikatu den IXAmBERT masiboki eleaniztunak diren ereduei gailentzen zaie.

5.5 Ondorioak

Laburbilduz, konfigurazio ezberdinetako hainbat BERT eredu aurrentrenatu, eta BL2MP datu-multzoan ebaluatu ostean, ondorioztatu dugu entrenamenduko corpus tamaina handitzeak inpaktu handiagoa duela gramatikaren ikasketan ereduaren tamaina handitzeak baino. Hemen aztertu diren 5M-125M hitz tarteko aurre-ikasketako corpusekin hainbat epoka egitea beharrezkoa dela azpimarratzen dute gure emaitzek.

Gure esperimentuek lematizazioan oinarritutako tokenizazioak, euskara bezalako hizkuntza flexionatuen gramatika ikasteko lagungarri izan daitekeela erakutsi du, datu kopuru oso txikia (5M hitz aurre-ikasketarako) dagoenean eskura. Hala ere, lematizazioaren onura lurrindu egiten da datumultzoaren tamaina 25M hitzetatik 125Metara handitzean. Halaber, antzeko joera ikusi daiteke hitz-ordena behatzean: hitz-ordena malguak erronka gehigarria dakar gramatika corpus txiki batetik (5M hitz) ikastean, baina entrenamendurako adibide nahikorekin, hizkuntza-ereduak gai dira fenomeno gramatikalak barneratzeko, hitz-ordena edozein izanik ere.

Gainera, BERT ereduek fenomeno gramatikal ezberdinetan duten doitasunean alde nahiko txikia dagoen arren, 5M hitzekin entrenatutako ereduak egitura eta ordenan (E3) indargunea duela erakusten du. Bestalde, 125M hitzeko eredua hobeto moldatzen da deklinabidean (E1), aditzetan (E2) eta egitura eta ordenan baino (E3). L2 ikasleen mailaren eta hizkuntza-ereduak dituzten zailtasunen arteko korrelazio nabarmen bakarra da hizkuntza-eredua apur bat hobeto dabilela maila baxueneko (A) adibideetan. Alde txiki hori berdindu egiten da aurrentrenamendurako corpusaren tamaina 125M hitzetara handitzean.

Azkenik, hizkuntza-ereduak euskararen gramatikan ebaluatzeko, zorroztasun handiarekin ondutako pare-minimodun BL2MP datu-multzoa ondasun erabakigarria izango dela uste dugu hizkuntza-ereduen ulermen linguistiko eta gramatikala aztertzea helburu duten etorkizuneko ikerketetan.

6. KAPITULUA

Zure Hizkuntzan Hizkuntza-Ereduak Aurrentrenatzeko Nahikoa Daturik Ez? Itzulpen Automatikoa Erreskatera!

Laburpena

Azken urteotan, Transformerretan oinarritutako aurrentrenatutako hizkuntza-ereduak hizkuntzaren prozesamenduko atazen gehiengoa lantzeko giltzarri izan dira. Horrelako ereduak aurrentrenatzeak izugarrizko testu bilduma eskakizunak ditu, ordea, hizkuntza gehienen kasuan eskuraezinak. Kapitulu honetan, hizkuntza-ereduak aurrentrenatzeko datu falta horri aurre egiteko automatikoki itzulitako corpusa erabiltzearen bideragarritasuna aztertu da. Ondorengo ikerketa-galderei erantzuten diegu: IG1: Itzulpen automatiko bidez sortutako datu-multzoak euskarazko testu errealaren alternatiba izan daitezke? IG2: Itzulitako datuak jatorrizko datuen osagarri izan daitezke hizkuntza-ereduen emaitzak hobetzeko? Bi galdera hauek erantzuteko prozesuan, hainbat BERT eredu entrenatu dira euskararako, euskarazko datu errealak gaztelaniatik itzulitako datu sintetikoekin osatuz. BasqueGLUEko NLUko 9 atazetan ebaluatu eta gero esan genezake, itzulitako datumultzoetan soilik entrenatutako ereduek emaitza lehiakorrak lortzen dituztela. Gainera, jatorrizko testua testu sintetikoarekin konbinatuta, ereduak hobetzea posible da, baina ez den berehalakoa, hainbat aldagaien menpe dago.

6.1 Sarrera

Atentzioan oinarritutako Transformer arkitekturaren (Vaswani et al., 2017) agerpenetik, eta BERTek (Devlin et al., 2019) aurrentrenamendurako maskaratze estrategiak mahaigaineratuz geroztik, Transformerretan oinarritutako hizkuntza-ereduak hurbilpen lehenetsiak bilakatu dira hizkuntzaren prozesamenduko ataza askotan, sekulako emaitzak erdietsiz baliabide oparoko hizkuntzetan, ingeleserako bereziki (Hoffmann et al., 2022; Thoppilan et al., 2022; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022).

Eskala-legeek (Kaplan et al., 2020; Hoffmann et al., 2022) iradoki bezala, horrelako eredu lehiakorrak kalkulu-ahalmen aurrekontu ikaragarriekin eta corpus erraldoiekin soilik lor litezke, hizkuntza gehienentzat eskuraezinak diren eskakizunak (Joshi et al., 2020b).

Zorionez, ari dira apurka baliabide-urriagoko hizkuntzetako ere hizkuntza-ereduak agertzen: KinyaBERT kinyaruandarako (390M hitz) (Nzeyimana and Rubungo, 2022), ElhBERTeu euskararako (351M) (Urbizu et al., 2022), gaBERT Irlandako gaelikorako (161M) (Barry et al., 2021), LuxemBERT luxenburgerarako (130M) (Lothritz et al., 2022), Bertinho 45M galegorako (Vilares et al., 2021), swahBERT swahilirako (16M) (Martin et al., 2022) eta QuBERT kitxuarako (4M) (Zevallos et al., 2022).

Zhang et al. (2021) lanak 10M-100M hitz artean estimatzen du sintaxiaren eta semantikaren gaitasun linguistikoak barneratzeko aurrentrenamenduan beharrezko diren gutxieneko datu kopurua. Ezagutza faktuala eta arrazoitzeko gaitasuna barneratzeko, baina, beharrezko datu kopurua askoz ere handiagoa da.

Kapitulu honetan, datu faltari aurre egiteko, beste hizkuntza batzuetan eskuragarri dagoen testua automatikoki itzultzea proposatzen dugu. Guk dakigunez, hurbilpen hau aurrez landua da (Lothritz et al., 2022), baina ez sakonki, eta lotura zuzena duten hizkuntza bikote bakarrarekin (alemana eta luxenburgera). Hizkuntza bakartua den euskara aukeratu dugu helburu-hizkuntza gisa, eta hizkuntza laguntzaile modura, berriz, gaztelania.

Ondorengo ikerketa-galderei (IG) erantzuten jardun da kapitulu honetan:

IG1: *Hizkuntza-eredu natibo baten pareko emaitzak lor litezke itzulpen automatiko bitartez sortutako datu sintetikoak bakarrik erabilita*?

IG2: *Posible da baliabide-urriko hizkuntzetarako hizkuntza-ereduak hobetzea, entrenamenduan automatikoki itzulitako datu sintetikoak gehituz*?

6.2 Metodologia

Gure ikerketa-galderak erantzun ahal izateko, ondorengo metodologia jarraitu dugu. Hasteko, bi oinarri-lerro proposatzen ditugu: i) ElhBERTeu (Urbizu et al., 2022) oinarri-lerro sendo gisa, 351M hitzeko corpusarekin entrenatua; eta ii) BERT_{125M} eredu bat, datu gutxiagorekin entrenatua. Jarraian, datu natibo eta sintetikoak hainbat modutara konbinatuz hainbat hizkuntza-eredu entrenatu ditugu. 6.3.1. eta 6.3.2. Atalek ereduen aurrentrenamenduaren xehetasun gehiago biltzen dituzte, oinarri-lerroko ereduenak barne. Lan honetan aurkeztutako eredu guztiek BERT *base* arkitektura dute (Devlin et al., 2019).

6.2.1 Hizkuntzak

Ikerketa hau burutzeko euskara hautatu da –hizkuntza bakartu bat– helburu hizkuntza gisa, eta gaztelania, hizkuntza erromantze bat, hizkuntza laguntzaile modura, testu corpus erraldoiak dituelako eskuragarri. Gainera, bi hizkuntzak eremu geografiko berean bizikide diren heinean, gaztelaniarekin partekatzen ditu euskarak testu paralelo gehien, eta hau erabakigarria da itzulpen automatikoko sistema sendoak entrenatzeko. Azkenik, aipatu, hau kasu erreal bat dela, euskaraz milioi erdi hitzetik gorako corpusak lortzeko zailtasunak baitaude.

6.2.2 Itzulpen Automatikoko Sistema

Esperimentuetan erabilitako gaztelaniatik euskararako itzulpen automatikoko sistema Transformer *base* arkitekturan (Vaswani et al., 2017) oinarritzen da, eta OpenNMT tresnaren (Klein et al., 2017) PyTorcherako bertsioa erabili da, 32K tokeneko hiztegi partekatuko BPE tokenizatzaile batekin (Sennrich et al., 2016).

8,6M esaldi paralelorekin entrenatu da sistema, eta FLORES-200 probabankuan (Costa-Jussà et al., 2022) ebaluatzean, 13,2ko BLEUa eta 47,4ko chrF++a eskuratzen ditu. Ikus 6.3.3.4. Atala, datu-paraleloen kopuruak duen eraginaren analisirako.

6.2.3 Corpusak

Jarraian, esperimentuetan zehar erabili diren corpusak aurkeztuko ditugu (6.1. Taulan laburbilduak):

N_ElhBERTeu, ElhBERTeu (Urbizu et al., 2022) entrenatzeko bildutako euskarazko corpusa da eta 351M hitz ditu.

N_small euskarazko corpus natibo txikiago bat da (125M hitz), baliabide urriko hizkuntza askoren eszenatokira gehiago hurbiltzen dena. Corpusa %75ean Berriako[1] albisteek eta %25 batean Wikipediako testuek osatzen dute.

S_beto2eu gaztelaniarako BETO (Cañete et al., 2020) hizkuntza-eredua entrenatzeko erabilitako 3B hitzeko *Spanish Unannotated Corpora*[2] da, 6.2.2. Atalean deskribatutako itzulpen automatikoko sistemarekin euskarara itzulia. Euskaraz lortutako corpus sintetikoak 2,2B hitz ditu.

S_loc2eu ere gaztelaniatik euskarara itzulitako corpusa da. Geografikoki Euskal Herrira mugatzen diren albiste iturrietatik gaztelaniazko albisteak bildu ditugu, 548M hitzetara iristeraino. Gure sistemarekin automatikoki itzuli ondoren, euskarazko 378M hitz ditu corpusak.

Corpusak	Hitzak	Domeinua
N_ElhBERTeu	351M	Nahasketa
N_small	125M	Albisteak+Wikipedia
S_beto2eu	2,17B	Nahasketa
S_loc2eu	378M	Albisteak

6.1. Taula: Gure ereduak entrenatzeko erabilitako corpusak. Testu sintetikoaren hitz kontaketa euskarara itzuli ondoren egin da.

6.2.4 Aurrentrenamenduko Xehetasunak

Lan honek fokua aurrentrenamenduko datuen eraginean jarrita duenez, gainerako hiper-parametroak finko utzi ditugu, balio lehenetsiekin. Eredu bakoitza, ElhBERTeu (Urbizu et al., 2022) entrenatzeko erabilitako prozedura

[1] www.berria.eus
[2] www.github.com/josecannete/spanish-corpora

berarekin aurrentrenatu da. Zehazki, letra larri eta xeheak bereizten dituen 50K hizpeko tokeneko hiztegi bat entrenatu dugu unigram hizpeko segmentazio algoritmoa erabiliz (Kudo, 2018). Hitz osoen maskaratzea erabili da, eta ereduak miloi bat urratsez entrenatu dira, 256ko labekada tamainarekin eta 512ko sekuentzia luzerarekin, TPUv3-8 bakarrean 6 egunez. Eredu bakoitza aurrentrenatzeak 9,8e+19 FLOPeko kalkulu-ahalmena behar izan du, eta 196kg-ko CO2 isuriak estimatu ditugu *Machine-Learning Impact calculator*[3] (Lacoste et al., 2019) tresnarekin.

6.2.5 Ebaluazioa

BasqueGLUE (Urbizu et al., 2022) euskararen hizkuntza ulermenerako probabankuaren bidez, gure ereduak helburu-atazetan ebaluatu ditugu. BasqueGLUEk ondorengo atazak biltzen ditu: entitate izendunen ezagutza (NER), asmoen sailkapena (*intent*)(López de Lacalle et al., 2020), hutsuneak betetzea (*slot*)(López de Lacalle et al., 2020), gaien sailkapena (BHTC)(Agerri et al., 2020), sentimenduen analisia (BEC), jarrera detekzioa (Vaxx)(Agerri et al., 2021), galdera-erantzunen hizkuntza inferentzia (QNLI), hitza testuinguruan (WiC) eta korreferentzia ebazpena (*coref*). Ebaluazioko metriken artean, zehaztasuna enplegatu da QNLI, WiC eta korreferentziaren kasuan, Macro F1 Vaxxen, eta Micro F1 gainerako atazetan.

Ereduetako bakoitza 10 epokaz birdoitu dugu, 32ko labekada tamaina eta 3e-5ko ikasketa-tasarekin, eta garapeneko datu-multzoaren gainean epoka onenaren kontrol-puntuarekin geratu gara ebaluaziorako. Datu-multzoaren gainean egindako bost birdoitze exekuzioren batez besteko emaitza aurkezten dugu. Birdoitze prozesua NVIDIA GeForce RTX 3090 GPUetan egin da.

6.3 Emaitzak eta Analisia

6.3.1 Datu Sintetikoekin Soilik Entrenatutako Eredua

Lehen ikerketa-galderaren xedea itzulpen automatikoarekin sortutako testu sintetiko hutsarekin hizkuntza-eredu lehiakorrik entrenatzerik badagoen aztertzea da. Horretarako S_beto2eu corpusarekin BERT eredu bat (S_BERT)

[3] `https://mlco2.github.io/impact#compute`

entrenatu da, eta datu sintetikoetan soilik entrenatutako ereduak ea datu errealekin entrenatutako ereduak bezain ondo dabiltzan ebaluatu da.

S_BERTek BasqueGLUE proba-bankuan lortutako emaitzak 6.2 Taulan daude ikusgai eta MLM atazan lortutakoak berriz 6.3. Taulan. S_BERT ereduak emaitza lehiakorrak lortzen ditu. Nahiz eta ElhBERTeu oinarri-lerro sendoaren parera ez iritsi, itzulitako testuan entrenatutako S_BERT eredua BERT_{125M}[4] eredu natiboaren parera heltzen da, atazaren arabera, kasu batzuetan honi gailenduz.

	avg	NER	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
ElhBERTeu	73.40	82.03	74.13	82.19	78.48	69.46	**76.04**	59.41	**72.27**	66.64
BERT_{125M}	71.98	79.51	75.18	80.83	76.94	70.15	73.76	58.32	70.09	63.10
S_BERT	71.40	79.72	74.03	80.94	73.32	68.83	73.93	58.92	70.06	62.83
SN_BERT	72.38	**82.33**	74.12	81.45	77.24	**70.32**	73.76	56.31	71.37	64.50
Sloc_BERT	72.21	79.74	74.75	**82.26**	75.91	69.80	72.66	61.08	70.37	63.27
SNloc_BERT	72.49	81.81	75.14	80.33	78.05	69.95	72.74	56.23	71.63	66.51
concat50-50	73.12	81.99	**75.29**	79.87	77.80	69.23	72.41	**62.80**	71.93	66.78
concat80-20	**73.47**	82.05	74.21	81.54	**78.57**	68.86	74.09	62.17	71.60	**68.11**
sequential	72.75	81.71	74.39	81.55	77.95	69.06	74.09	57.66	72.11	66.24

6.2. Taula: S_BERT (sintetikoa), SN_BERT (natiboa+sintetikoa), Sloc_BERT (sintetiko lokala) eta SNloc_BERTen (natiboa+sintetiko lokala) emaitzak ElhBERTeu eta BERT_{125M} eredu natiboekin alderatuta. concat50-50, concat80-20 eta sequential uztartze estrategia ezberdindun ereduak dira (Ikus 6.3.5. Atala).

6.3.2 Datu Natiboak eta Sintetikoak Elkartuz

Bigarren ikerketa galderaren bidez, ea corpus natibo bati automatikoki itzulitako testua gehitzeak helburu hizkuntzako ulermeneko atazetan hizkuntza-eredua hobetuko duen argitu nahi dugu.

Horretarako, hizkuntza-eredu berri bat entrenatu dugu, S_beto2eu eta N_ElhBERTeu corpusak uztartuz[5] (SN_BERT hemendik aurrea). 6.2. Taulak SN_BERTek BasqueGLUE proba-bankuan ateratako emaitzak daude ikusgai. SN_BERTek ataza pare batean (NER, BEC) ElhBERTeu gainditzen duen arren, batez beste azpitik geratzen da.

[4] 125M hitzeko N_small euskarazko corpus natiboarekin entrenatua.
[5] Uztarketa hori dokumentu-mailan nahastuz egin da.

	BasqueGLUE	MLM
ElhBERTeu	73.40	**61.07**
BERT_{125M}	71.98	58.97
S_BERT	71.40	49.67
SN_BERT	72.38	58.64
Sloc_BERT	72.21	53.66
SNloc_BERT	72.49	60.42
paral_{EU}	69.19	43.75
paral_{ES2EU}	68.65	41.98
concat50-50	73.12	59.84
concat80-20	**73.47**	61.05
sequential	72.75	58.61

6.3. Taula: MLM atazako zehaztasunak.

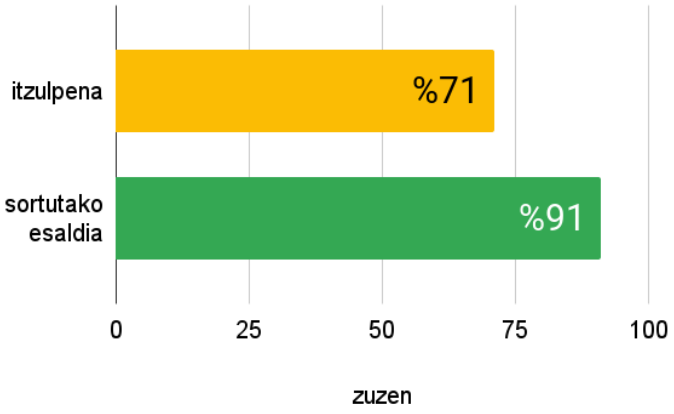
Datu sintetikoekin lortutako emaitzak hobetze aldera, jarraian datuen bi alderdi espezifiko aztertu ditugu: i) itzulpenen prozesuan gertatutako kalitate galera; ii) datu sintetikoen testuinguru kulturala.

6.3.3 Itzulpen Automatikoaren Kalitatea Neurtzen

Gaztelaniatik euskarara itzultzean gertatutako kalitatearen galera neurtzeko, itzulpen automatikoko (MT) sistemaren lagin batekin eskuzko analisi bat egin dugu, eta itzulpenean zehar izandako hiztegiaren aberastasun galera ere neurtu dugu. Bestalde, hizkuntza-ereduak itzulpen automatikoaren bitartez sortutako datu sintetikoekin entrenatzeak duen eragina aztertu dugu.

6.3.3.1 MT Sistemaren Eskuzko Ebaluazioa

Itzulpen sistemaren eskuzko ebaluazioa S_beto2eu corpusetik ausaz aukeratutako 100 esaldiko lagin batekin burutu da. Ebaluazioa bi hiztun elebidunek egin dute. Ebaluatzaileek gaztelaniazko esaldi originala eta itzulpen automatikoko sistemak sortutako itzulpena zituzten eskura. Alde batetik, sortutako esaldia esaldi originalaren itzulpen egokia den ebaluatzeko eskatu zitzaien. Beste aldetik, sortutako esaldia bere horretan linguistikoki zuzena den ebaluatu behar zuten. 6.1. Irudiak eskuzko ebaluazio horren emaitzak biltzen ditu.



6.1. Irudia: MT sistemaren eskuzko ebaluazioa. Itzulpena zuzena ez den kasuen gehiengoan, sortutako esaldia linguistikoki zuzena da bere horretan.

Esaldien %71 zuzen itzulita dagoela ebatzi da eta sortutako esaldien %91 linguistikoki zuzenak direla. Jomuga, testu honekin hizkuntza-ereduak entrenatzea izanik, haluzinazioa edo omisioa bezalako itzulpen akats batzuk, ez dute zertan larriak izan, testuaren esanahia aldatu arren.

6.3.3.2 Hiztegiaren Aberastasuna

Jarraian, itzulpenean zehar izandako hiztegiaren aberastasun galera neurtu dugu.

Horretarako, EITBcc (Etchegoyhen and Gete, 2020) eta EhuHac-ek (Tiedemann, 2012) osatzen duten euskara-gaztelania corpus paraleloa baliatu dugu. 6.4. Taulak euskarazko zatiko testuak eta gaztelaniako zatiko testuen itzulpen automatikoak (euskarara) dituzten hiztegiak konparatzen ditu. Itzulpen automatiko bidez sortutako euskarazko testuan gutxi gorabehera 90K hiztegi sarrera gutxiago ditu.

Automatikoki itzulitako testuaren lexikoa %16ean pobretu da, MT ereduek irteerako hiztegia orokortu eta sinplifikatzeko duten joeragatik.

Corpusa	Hitzak	Hiztegia
Natiboa (EU)	29M	549K
MT (ES2EU)	28M	462K

6.4. Taula: Itzulpen automatikoak eragindako hiztegiaren aberastasunaren galera.

6.3.3.3 Itzulitako Datu Sintetikoen Eragina Hizkuntza-Ereduak Entrenatzean

Bukatzeko, hizkuntza-ereduak entrenatzeko automatikoki itzulitako corpusak erabiltzearen inpaktua aztertu da, corpusaren tamaina eta testuen domeinua bezalako faktoreak alde batera utziz. Bi BERT eredu aurrentrenatu dira aurreko esperimentuan aipatutako corpus paraleloa erabiliz. Bata, corpus paraleloaren euskarazko zatia erabiliz; bestea, euskarara automatikoki itzulitako corpuseko gaztelaniazko zatiarekin. 6.5. Taulako emaitzek erakusten dute itzulitako datuetan entrenatutako ereduak eredu natiboa baino okerrago dabilela. Hau espero zitekeen arren MT sistemak dakarren kalitate galera eta lexikoia txirotzearen ondorioz, emaitzetan galera oso txikia izan da (%0,5 batez beste). Honenbestez, datu sintetikoen kalitatea balekoa dela ondorioztatu dugu hizkuntza-ereduak elikatzeko.

	avg	NER	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref
paral_{EU}	**69.19**	**73.88**	**74.04**	**81.82**	**72.56**	**68.88**	69.87	53.70	**68.67**	**59.28**
paral_{ES2EU}	68.65	72.42	72.60	78.12	71.09	68.10	**70.13**	**58.74**	68.13	58.50

6.5. Taula: Corpus natiboarekin entrenatutako ereduaren (paral_{EU} –29M–) eta jatorri bereko itzulitako corpusarekin (paral_{ES2EU} –28M–) entrenatutako ereduaren helburu-atazetako emaitzen konparaketa.

6.3.3.4 Datu Paralelo Kopuruaren Eragina

Gure hurbilpenak itzulpen automatikoko sistema sendoa behar du. Erabili dugun sistema 8,6M esaldi paralelorekin entrenatua dago. Tamaina horretako baliabide bat hizkuntza ororen eskura ez dagoela jakitun garenez, datu paralelo kopurua drastikoki jaisteak itzulpen automatikoko sistemaren kalitatean zer eragin duen aztertu nahi izan dugu.

Horretarako, itzulpen sistema berri bat entrenatu dugu datu kopuru erdiarekin. Esaldi paraleloen kopurua 8,6Mtik 4Mra jaisteak ez du eragin nega-

tibo nabarmenik izan FLORES-200 (Costa-Jussà et al., 2022) banku-proban 13,0 BLEU (-0,2) eta 47,2 chrF++ (-0,2) lortuz (ikus 6.6. Taula). Ondorioz, itzulpen automatikoan antzeko kalitatea lortuta, hizkuntza-ereduaren portaeran ere kalitatea mantentzea espero liteke.

Datu paraleloak	BLEU	charF++
8,6M esaldi	13.2	47.4
4M esaldi	13.0	47.2

6.6. Taula: Itzulpen automatiko datuen ablazio azterketaren emaitzak. Itzulpenak ebaluatzeko FLORES-200 proba-bankua eta BLEU eta charF++ metrikak erabili ditugu.

6.3.4 Domeinua eta Testuinguru Kulturala

Itzulitako corpusekin entrenatutako hizkuntza-ereduen kalitatean eragina izan dezakeen datuei lotutako beste faktoreetako bat hizkuntza laguntzailean aukeratutako testuen jatorria da. Gaztelaniazko *Spanish Unannotated Corpora* corpusa erraldoia da. Baina, domeinuari dagokionez ez da oso aberatsa, eta kulturalki euskarazko corpusetatik nahiko urrun dago, bereziki, Euskal Herriarekin lotura zuzena duten gaiek apenas dutelako tokirik. Gainera, corpus honetan entrenatutako tokenizatzaileek euskal komunitatearen testuinguruko hainbat hitz arrunt, izen bereziak adibidez (leku, pertsona edo erakundeenak), hiztegitik at uzten dituztela behatu dugu. 6.3.4.1. Atalean testerako datu-multzoko testuetan eredu bakoitzaren tokenizatzaileek duten hiztegi estaldura aztertu dugu.

Aurrentrenamenduen datuen alborapen kulturalen eta domeinuaren ezaugarriek helburu-atazetako emaitzetan duten inpaktua neurtzeko, S_loc2eu corpusa bildu dugu, dagoeneko 6.2.3. atalean aurkeztua. Corpus hau, Euskal Herriari geografikoki eta kulturalki lotutako egunkarietatik erauzitako gaztelaniazko testuek osatzek dute. Itzulitako albiste lokaletan entrenatutako ereduak (Sloc_BERT eta SNloc_BERT), edonolako testuarekin entrenatutako ereduak (S_BERT and SN_BERT) baino hobeto dabiltza erakusten dute 6.2. Taulako emaitzek, corpusaren tamaina txikiagoa izan arren. Joera berdina dago hiztegiaren estalduran ere, tokiko albiste itzulien kasuan estaldura handia lortzen da, 6.3.4.1. Atalean ikus daitekeen moduan.

6.3.4.1 Ebaluazio Multzoko Hiztegi Estaldura

Helburu-atazetan azaltzen den antzeko hizkuntzarekin (dialekto, domeinu edo idazkera) aurrentrenatutako hizkuntza-ereduek emaitza hobeak eman ohi dituzte (Lee et al., 2020; Inoue et al., 2021). Hiztegitik at (*out-of-vocabulary* edo OOV) dauden hitzen analisi bat burutu dugu, eredu bakoitzaren aurrentrenamenduko datuen eta BasqueGLUEko atazetako ebaluazio-multzoen arteko antzekotasuna neurtzeko.

Hiztegitik ateko analisirako, BasqueGLUEko ebaluazio-multzo bakoitzeko hiztegia hizkuntza-ereduen tokenizatzailearekin konparatu da, bakoitza bere corpusarekin entrenatua. Konparaketa hitz osoen mailan egin da sinplifikatzearren, hizpeko hainbat tokenetan banatzen diren hitzak kanpo utziz hiztegi gainjarrietatik. 6.3.5. Atalean aipatutako eredu sekuentziala ez da konparaketan agertzen, ElhBERTeuren tokenizatzaile bera baitarabil.

6.7. Taulak hiztegitik at daudenen analisia biltzen du. Emaitzen arabera, ebaluazio-multzoen antzekotasun handiena duen corpusa N_ElhBERTeu da, lan honetan aztertu den corpus natibo handiena da berau, hiztegitik ateko %42,82ko ratio batekin. N_ElhBERTeu da hiztegitik ateko ratio txikiena duena ataza gehienetan, zeinari QNLIn concat50-50 gailentzen zaion, eta Vaxx eta coref atazetan SN_BERTloc. Bestalde, hiztegiaren estaldura-ren analisi honek ebaluazio-multzotik S_loc2eu[6] S_beto2eu[7] baino gertuago dagoela ere erakusten du.

6.3.5 Uztartzeko Estrategiak

Datu natibo eta sintetikoak bateratuz ereduak emaitza apalagoak lortzea azal lezakeen faktoreetako bat datuak uztartzeko modua da. Azken esperimentu honen xedea datuak uztartzeko estrategia ezberdinak ikertzea da. SN_BERTren kasuan, sinpleki, N_ElhBERTeu eta S_beto2eu corpusak kateatu ditugu. Baina horrela, kalitate altuagoko corpus natiboa, kalitate xumeagoko, baina tamainaz handiagoa den (x4) itzulitako corpusaren artean galtzen da. Hori dela eta, corpus natibo eta sintetikoak uztartzeko beste hiru alternatiba proposatu ditugu, bi corpusen orekarekin jolastuz:

[6] S_BERTloc entrenatzeko erabilia.
[7] S_BERT entrenatzeko erabilia.

Test	NERC	slot	intent	BHTC	BEC	QNLI	Vaxx	WiC	coref	avg
ElhBERTeu	**41.34**	**30.61**	**30.61**	**54.26**	**61.25**	48.84	51.29	**40.94**	26.28	**42.82**
BERT_{125M}	41.77	33.42	33.42	55.71	62.89	49.81	53.07	42.67	26.28	44.34
S_BERT	50.35	35.33	35.33	63.64	67.80	51.81	56.58	48.56	34.66	49.34
SN_BERT	45.05	33.42	33.42	58.96	64.22	49.44	53.45	44.18	29.61	45.75
Sloc_BERT	45.53	34.31	34.31	57.96	63.67	50.88	52.50	44.67	29.04	45.88
SNloc_BERT	41.59	30.87	30.87	54.62	61.47	48.98	**50.56**	40.96	**26.08**	42.89
concat50-50	42.66	30.99	30.99	56.36	62.44	**48.56**	52.07	42.21	27.46	43.75
concat80-20	41.98	32.40	32.40	55.83	62.37	48.79	52.03	42.23	27.12	43.91
paral_{EU}	53.05	41.45	41.45	64.46	67.62	55.80	60.09	53.26	39.18	52.93
paral_{ES2EU}	57.55	47.83	47.83	67.88	71.19	58.64	64.57	58.85	45.02	57.71

6.7. Taula: Ebaluaziorako datu-multzoan hitz oso gisa tokenizatzailearen hiztegitik at dauden tokenen ehunekoa eredu bakoitzaren kasuan.

concat$_{20-80}$ (SN_BERT): N_ElhBERTeu eta S_beto2eu corpusen kateatzea, gutxi-asko aurrentrenamenduko corpusaren %20 eta %80 osatzen dutelarik hurrenez hurren. Aipatu bezala, konfigurazio honetan datu sintetikoek rol nagusia hartzen dute.

concat$_{50-50}$: N_ElhBERTeu corpus natiboari gainlaginketa aplikatu diogu, S_beto2eu corpusaren tamaina berdinduz. Konfigurazio honek datu natiboen eta sintetikoei pisu bera ematen die.

concat$_{80-20}$: N_ElhBERTeuri gainlaginketa aplikatu diogu %80ra iristeraino, beraz, aurrentrenamenduko datuen gehiengoa testu natiboak osatzen du, sintetikoak baino pisu gehiagorekin.

sequential[8]**:** Hizkuntza-eredua 750K urratsez S_beto2eu corpusean entrenatzen da, eta ondoren beste 250K urratsez N_ElhBERTeu corpusean[9].

Eredua	**Datuak**	**Uztartzea**	**BasqueGLUE**
ElhBERTeu	350M	-	73.40
BERT$_{\text{concat}_{20-80}}$	350M+3B*	concat$_{20-80}$	72.38
BERT$_{\text{concat}_{50-50}}$	350M+3B*	concat$_{50-50}$	73.12
BERT$_{\text{concat}_{80-20}}$	350M+3B*	concat$_{80-20}$	**73.47**
BERT$_{\text{sequential}}$	350M+3B*	sequential	72.75

6.8. Taula: Datu erreal eta datu sintetikoen (*) uztartze estrategia ezberdinekin entrenatutako ereduen emaitzak.

Uztartze estrategia ezberdinen emaitzak 6.8. Taulan daude laburbilduak (Ikus 6.2. Taula atazaz atazako emaitzetarako). Aurrentrenamenduko corpusean N_ElhBERTeuren pisua handitzean gure ereduen emaitzak hobetu dira, concat$_{80-20}$en kasuan euskarazko testu natiboa soilik ikusi duen ElhBERTeu gainditzeraino. Aurrentrenamendu sekuentzialak zertxobait hobetu du defektuzko SN_BERT konfigurazioa, baina bien artean, pisuz orekatutako kateatzea uztartze estrategia hobea dela ikusi da.

[8]Entrenamendu sekuentzialaren beste aldaera batzuk ere aztertu dira, eta zerrendako '*sequential*' aldaera (750K+250K) izan da emaitzarik onenak erdietsi dituena.

[9]bi faseek N_ElhBERTeu corpusean entrenatutako tokenizatzailea darabilte.

6.4 Ondorioak

Kapitulu honetako lehen ikerketa-galderari erantzunez, gure esperimentuen bitartez ondorioztatu ahal izan dugu automatikoki itzulitako datu sintetikoekin soilik entrenatutako hizkuntza-ereduek hizkuntza-eredu natibo baten pareko emaitzak lor ditzakeela. Azterketa sakonago batean, itzulpen automatikoko sistemaren kalitateaz gain, hizkuntza laguntzailean aukeratutako testuen testuinguru kulturalak eta domeinuak ere eragin zuzena dutela ikusi da. Hots, helburu-hizkuntzan landu nahi diren testuen antzeko izaerako testuak biltzea lehenetsi behar dela ondorioztatu dugu, hizkuntza laguntzaileko datu kopuru itzelak kontrolik gabe bildu beharrean.

Gainera, bigarren ikerketa-galderari dagokionez, gure esperimentuek erakutsi dute posible dela artearen egoerako ereduak gainditzea aurrentrenamenduan itzulitako testua gehituz, hobekuntza txikia izan arren. Datuak uztartzean testu natiboei, testu sintetikoei baino pisu gehiago ematea giltzarri dela behatu da.

Laburbilduz, hurbilpen honek badu potentziala baliabide-urriko hizkuntzentzat, behin itzulpen automatikoko sistema sendo bat izanik, ez baitago corpus handiagoak dituen hizkuntza batetatik testua itzultzeko mugarik.

Datu-multzoak eta ereduak publikoki eskuragarri daude[10].

[10] `https://github.com/orai-nlp/mt-bert`

ONDORIOAK, EKARPENAK ETA ETORKIZUNEKO LANAK

7. KAPITULUA

Ondorioak, Ekarpenak eta Etorkizuneko Lanak

Tesi txosten honetan euskararako –eta orokorrean, baliabide urriko testuinguruetarako– kodetzaile motako hizkuntza-eredu neuronalen hizkuntzaren ulermenerako gaitasuna aztertu da. Euskararako hizkuntza-ereduen egungo zein etorkizuneko garapena azkartzeko asmoz, giltzarri diren zenbait ertz jorratu dira, nagusiki, aurrentrenamendu datu mugatuei, euskararen ezaugarri linguistikoei eta hizkuntza ulermenaren ebaluazioari lotuak.

3. Kapituluan hizkuntza-ereduak euskararen hizkuntza ulermenean ebaluatzeko BasqueGLUE proba-bankua eta ElhBERTeu hizkuntza-eredua aurkeztu ditugu. 4. Kapituluan baliabide urriko inguruneetan BERT ereduek jarraitzen dituzten eskala-legeak aztertu dira 4 hizkuntzatan. 5. Kapituluan hizkuntza-ereduen euskarazko gramatika gaitasuna eta euskararen ordena malguaren eta morfologia aberatsaren eragina izan ditugu hizpide. Bestalde, 6. Kapituluan hizkuntza-ereduak aurrentrenatzeko automatikoki itzulitako datu sintetikoak erabiltzearen bideragarritasuna aztertu dugu. Bukatzeko, A Eranskinean euskararako hizkuntza-ereduak elikatzeko bildutako euskarazko bi corpus elebakar aurkeztu ditugu.

Kapitulu honetan tesitik ateratako ondorioak eta hausnarketak, tesian zehar sortutako baliabideen laburpen bat, eta tesitik eratorri daitezkeen ildoak ditugu mintzagai.

7.1 Ondorioak

Doktorego tesi honi itxiera emate aldera, berrar ditzagun ostera sarreran finkatutako ikerketa-galderak:

- IG1: Nola ebaluatu genezake euskararako hizkuntza-ereduen kalitatea modu estrintsekoan?
- IG2: Zein eragin du baliabide gutxi izateak hizkuntza-eredua entrenatzeko orduan?
- IG3: Zein eragin dute euskararen ezaugarri linguistikoek hizkuntza-eredua entrenatzerakoan?

Ikerketa-galdera horiek landu diren kapituluetako bakoitzean egindako esperimentuen ondorioak aurkeztu ditugu jada dagozkien kapituluen amaieran, baina jarraian, ikerketa-galderetako bakoitzari erantzungo diogu tesian zehar landu diren ertz guztiak kontutan izanda.

7.1.1 Nola Ebaluatu genezake Euskararako Hizkuntza-Ereduen Kalitatea modu Estrintsekoan?

Lehen ikerketa-galderari helduz, euskararako hizkuntza-ereduak modu estrintsekoan ebaluatzeko 3. Kapituluan BasqueGLUE eraiki dugu, euskararen hizkuntza ulermenean ebaluatzeko lehen proba-bankua; 5. Kapituluan, berriz, hizkuntza-ereduak euskararen gramatika gaitasunetan ebaluatzeko BL2MP datu-multzoa sortu da.

GLUE eta SuperGLUE bezalako proba-bankuak hizkuntzaren ulermenerako teknologietan erdietsitako aurrerapenak modu sendoan ebaluatzeko ezinbestekoak dira. Hutsune horri erantzun zaio euskararako hizkuntzaren ulermeneko ataza sorta zabal eta askotarikoa –domeinuari, zailtasunari eta datu-multzoen tamainari dagokionez– biltzen dituen BasqueGLUE eraikita. Proba-banku honek hizkuntza-ereduak atazetako datu-multzoetan birdoituta edo birdoitzeko beharrik gabe[1] NLUn ebaluatzeko aukera ematen du. Honenbestez, dagoeneko gai gara hizkuntza-ereduak sendotasunez konparatzeko

[1] 0-shot edo few-shot tekniken bitartez.

euskararen ulermenean. 7.1. Taulako emaitzek, ElhBERTeu BERTeusi gailentzen zaiola erakusteaz gain, BasqueGLUE osatzen duten datu-multzoen zailtasuna egokia dela ikus daiteke, euskararako hizkuntza-eredu diskriminatzaileak NLUn ebaluatzeko.

Eredua	AVG	NER F1	F_{intent} F1	F_{slot} F1	BHTC F1	BEC F1	Vaxx MF1	QNLI acc	WiC acc	coref acc
BERTeus	73,23	81,92	**82,52**	74,34	**78,26**	69,43	59,30	**74,26**	70,71	**68,31**
ElhBERTeu	**73,71**	**82,30**	82,24	**75,64**	78,05	**69,89**	**63,81**	73,84	**71,71**	65,93

7.1. Taula: BERTeus (Agerri et al., 2020) eta 3. Kapituluan entrenatutako ElhBERTeu ereduen emaitzak BasqueGLUE hizkuntza ulermeneko proba-bankuan.

Bestalde, BL2MP datu-multzoa ere sortu da, hizkuntza-ereduen gramatika gaitasunak euskaraz ebaluatzeko. Euskara ikasleen idazlan zuzenduetatik eratorria, benetako akats gramatikal aberatsak biltzen ditu, hiruna zailtasun mailatan eta errore motetan sailkatutako pare-minimoetan. Hizkuntza ereduak birdoitzeko beharrik gabe ebaluatzeko eta konparatzeko baliagarria da, 7.2. Taulak erakusten duen moduan.

Eredua	BL2MP
$BERT_{8L}$-eu25M	91,5
$BERT_{8L}$-eu125M	94,7
$BERT_{12L}$-eu125M	95,6
BERTeus	94,8
Roberta-euscrawl-B	95,9
Roberta-euscrawl-L	95,5
ElhBERTeu-medium	95,4
ElhBERTeu-base	**96,5**
mBERT	58,9
XLM-RoBERTa base	75,1
XLM-RoBERTa large	82,4
IXAmBERT	94,4

7.2. Taula: Euskararako MLM eredu elebakar eta eleaniztunek lortutako zehaztasunak BL2MP datu-multzoan.

Berriki sortu diren beste baliabide batzuekin batera (Lin et al., 2021b; Bandarkar et al., 2024; Etxaniz et al., 2024; Corral et al., 2024; Zulaika and

Saralegi, 2025), BasqueGLUE eta BL2MP euskararako hizkuntza-ereduak konparatu eta hauen garapena azkartzeko baliagarriak izango direla aurreikusten dugu.

7.1.2 Zein Eragin du Baliabide Gutxi izateak Hizkuntza-Eredua Entrenatzeko orduan?

Bigarren ikerketa-galderari dagokionez, aurrentrenamendurako nahikoa datu izatea giltzarria dela ikusi dugu 4. eta 5. Kapituluetan, eta datu eskasiari aurre egiteko alternatiba bat proposatu dugu 6. Kapituluan, datu sintetikoak ardatz hartuta.

Ezaugarri linguistiko ezberdinetako lau hizkuntzetan (euskara, gaztelania, swahilia eta suomiera) burututako esperimentuetan, aurrentrenamenduko galera kurbak eta helburu-atazetan birdoitu ondoren lortutako emaitzak kontuan izanda, hizkuntza-ereduen hizkuntza ulermeneko gaitasunak datu kopuruarekiko lotura zuzena duela ikusi dugu: aurrentrenamenduko corpusa 125M hitzetara handitu bitartean, hobekuntza nabarmena behatu dugu atazen gehiengoan; corpusa 5Mtik 25Mra handitzean, hobekuntza handia lortu da ia ataza guztietan, eta 25Mtik 125Mrako igoerak ere hobekuntza ekarri du atazen gehiengoan. Lortutako emaitzak sendoak dira aztertu diren hizkuntza ezberdinen artean. Honenbestez, aurrentrenamendurako datu-baliabide gutxi izateak (125Mtik hitzetik behera) eragin negatibo zuzena du hizkuntza-eredua ataza zehatzetan birdoitzean lortuko diren emaitzetan.

Joera berbera behatu dugu BL2MP datu-multzoan hizkuntza-ereduek euskararen gramatikaren ezagutzan ebaluatzean ere. Deklinabideari, aditzei zein egitura eta ordenari dagozkien ezagutza gramatikalean jauzi handia dago (hamar punturen bueltan) corpusa 5M hitzetik 25M hitzetara handitzean, eta hobekuntza apaltzen den arren, oraindik ere 25Mtik 125M hitzetara handitzean hainbat puntutako hobekuntza behatu da. Gure hizkuntza-ereduek euskararen morfologia aberatsa eta hitz-ordena malgua egoki barneratzeko, merezi du hizkuntza-ereduak gutxienez 125M hitzeko corpusarekin aurrentrenatzeak.

Beraz, baliabide urriko hizkuntzetan hizkuntza-ereduak lantzean, ahal adina corpus biltzeak duen berebiziko garrantzia azpimarratzen dugu, gutxienez 25M hitzeko corpusak erabiliz eta ahal dela 125M hitzetik gorako

corpusak bilduz, betiere, corpusa osatzen duten testu iturrien kalitatea eta aniztasuna bazter utzi gabe.

Bestalde, aurrentrenamendu corpusaren eta hizkuntza-ereduen tamainaren arteko orekari dagokionez, 5M eta 25M hitz arteko corpusa bakarrik dugunean eskura, gutxienez BERT_{51M} (*medium*) tamainako eredua gomendatzen dugu. Corpusa 125M hitzera handitzean, berriz, beharrezkoa litzateke BERT_{124M} (*base*) tamaina corpusari zukua ateratzeko.

Gainera, corpus txikietan hainbat (dozenaka eta ehunka) epokaz aurrentrenatzea mesedegarri dela ikusi dugu orokorrean. Hala ere, corpus tamaina oso txikiekin hizkuntza-ereduak aurrentrenatzean, datu-parametro kopuruen oreka horretan gainparametrizazioak gainegokitzapen arazoak dakartzala behatu dugu. Beraz, 25Mtik beherako corpusarekin BERT_{124M} tamainako eredua entrenatu nahiko bagenu entrenamendua garaiz mozteaz arduratu beharko genuke.

Horretaz gain, kalkulu-ahalmen jakin bat emanda, parametro-epoka oreka horretan, oro har, hobe da eredu handiagoak epoka mugatuagoz entrenatzea, eredu txikiagoak luzaroago entrenatzea baino.

7.1.2.1 Datu Sintetikoak Alternatiba Gisa

Mundu honetako hizkuntzen gehiengo zabalarentzat, ordea, lan nekeza edo ezinezkoa da milioika hitzeko testu-bildumarik eskuratzea. Datu eskasi horri aurre egiteko itzulpen automatiko bitartez sortutako datu sintetikoen erabileraren bideragarritasuna aztertu dugu 7. kapituluan (emaitzak 7.3. Taulan laburbilduak). Gure esperimentuetan itzulitako datu sintetikoekin bakarrik entrenatutako hizkuntza-ereduek hizkuntza-eredu natiboen pareko emaitzak lor ditzakeela ondorioztatu dugu.

Oro ez da urre, ordea. Euskaraz eskuragarri ditugun corpus tamainekin, testu natiboari itzulitako milaka miloi (3B gure esperimentuan) hitzeko corpus sintetikoa gehitzeak ez dakar berehalako hobekuntzarik. Emaitzen analisian sakontzean, itzulpen automatikoko sistemaren kalitateaz gain, hizkuntza laguntzailean aukeratutako testuen testuinguru kulturalak eta domeinuak ere eragin zuzena dutela ikusi da. Hizkuntza-eredu natiboen emaitzetan hobekuntza xumea lortzeko, helburu-hizkuntzan –gure kasuan, euskara– landu nahi diren testuen antzeko izaerako testuak biltzea lehenetsi behar da itzultzeko corpusa aukeratzean eta kalitate ezberdineko corpusak uztartzeko or-

duan testu natiboei, testu sintetikoei baino pisu gehiago ematea giltzarri dela ondorioztatu dugu.

Ereduak	**Datuak**	**BasqueGLUE**
$\text{BERT}_{\text{DatuErreal}}$	350M	73,40
$\text{BERT}_{125M\ \text{DatuErreal}}$	125M	71,98
$\text{BERT}_{\text{DatuSintetiko}}$	3B*	71,40
$\text{BERT}_{\text{Erreal+Sintetiko}}$	350M+3B*	72,38
$\text{BERT}_{\text{DatuSintetikoLokal}}$	378M*	72,21
$\text{BERT}_{\text{Erreal+SintetikoLokal}}$	350M+378M*	72,49
$\text{BERT}_{\text{Erreal(\%80)+Sintetiko(20\%)}}$	350M+3B*	**73,47**

7.3. Taula: Itzulpen automatiko bidez sortutako datu sintetikoekin (*) aurrentrenatutako hizkuntza-ereduekin lortutako emaitzak BasqueGLUE proba-bankuan.

Hurbilpen honen muga nagusiak hizkuntza bikotearen arteko corpus paralelo kopurua eta honekin lor genezakeen itzulpen kalitatea dira. Behin muga hori gainditzeko gai diren baliabide-urriko hizkuntzentzat, hau da, behin itzulpen automatikoko sistema sendoa izanik, potentzial handia duen hurbilpena da, ez baitago corpus handiagoak dituen hizkuntza batetatik testua itzultzeko inolako mugarik. Hala ere, kontutan izan behar da euskararen kasuan emaitzetan hobekuntza marjinal bat lortzearen truke, inongo kontrolik gabe erderetatik masiboki euskarara testua itzultzeak gure ereduen alborapen kulturala indartu lezakeela.

7.1.3 Zein Eragin dute Euskararen Ezaugarri Linguistikoek Hizkuntza-Eredua Entrenatzerakoan?

Bukatzeko, hel diezaiogun hirugarren ikerketa-galderari. Aurrez aipatu bezala, euskarak morfologia konplexua eta hitzen ordena malgua ditu bereizgarri. 5. Kapituluan ezaugarri horiek datu-baliabide gutxiko inguruneetan hizkuntza-ereduek gramatika barneratzea zaildu dezaketela ondorioztatu dugu, baina arazo horiek gainditzeko euskarak, gaur egun, behar adina datu badituela.

Corpusei (5M/25M/125M hitz) eta ereduaren tamainari (16M/51M/124M parametro) dagokionez, konfigurazio ezberdinetako hainbat BERT eredu aurrentrenatu eta BL2MP datu-multzoan ebaluatu ostean, gramatika ikastean

entrenamenduko corpus tamaina handitzeak ereduaren tamaina handitzeak baino inpaktu handiagoa duela eta hainbat epoka egitea beharrezkoa dela ondorioztatu dugu.

BERT ereduek BL2MP datu-multzoko fenomeno gramatikal ezberdinetan (deklinabidean, aditzetan eta egitura eta ordenan) duten doitasunean alde nahiko txikia erakusten dute orokorrean. Hala ere, badira azpimarratzeko ñabardurak: 5M hitzekin soilik entrenatutako eredua hobeto moldatzen da egitura eta ordenan, eta 125M hitzeko eredua deklinabidean gailentzen da. Gainera, hiru motetako fenomeno gramatikalak epoka berdinen inguruan barneratzen direla behatu dugu.

Eranskaria izateak –edota hobeto esanda, morfologia ez horren konplexuetarako diseinatutako tokenizatzaile eta arkitektura neuronalen bitartez euskara bezalako hizkuntza eranskari bat lantzeak– duen eraginari dagokionez morfologia egoki ikastea erronka gehigarria dela ikusi dugu. Morfologiaren araberako tokenizazio propioa ez izateak, eta tokenizazioa hitzen erroen eta morfemen banaketa estatistiko hutsean oinarritzeak, tokenizazio ez koherenteetara garamatza. Eta honek, hitz eta morfema bakoitzaren eta hauen arteko erlazioen agerpen kopurua zatikatuz, patroiak ikastea zailtzen du.

Horri aurre egiteko lematizazioan oinarritutako tokenizazioa erabiltzeko beharrik ba ote dagoen aztertu dugu. Esperimentuen emaitzek erakusten dute ereduek euskara bezalako hizkuntza flexionatuen gramatika hobeto ikas dezaten lematizazioa lagungarria izan daitekeela, esku artean datu kopuru oso txikia dagoenean eskura (25M hitzetik behera). Hala ere, lematizazioaren onura lurrindu egiten da patroiak barneratzeko nahikoa datu dagoen unetik aurrera. Beraz, gaur egun euskararako eskuragarri ditugun corpus baliabideen tamainak izanda, ez da beharrezkoa lematizazioaren urrats gehigarria txertatzea tokenizazioan, eta esfortzu gehigarri hori aurreztu genezake erabilera orokorreko aurrentrenatutako hizkuntza-ereduetan.

Halaber, hitzen ordenari dagokionez, joera bera aztertu dugu: hitz-ordena malguak erronka gehigarria dakar gramatika ikastean. Baina, hori ere, entrenamendurako adibide nahikoarekin gainditu daitekeen arazoa dela behatu da, eta 25M hitzetik gora hizkuntza-ereduak gai dira fenomeno gramatikalak barneratzeko, hitz-ordena ezohikoetan ere.

Azkenik, korrelazio handiegirik ez da topatu L2 ikasleen mailari dagokion zailtasunaren eta hizkuntza-ereduak dituzten zailtasunen artean. Hori gizakiok eta makinek hizkuntzak barneratzeko ditugun gaitasun eta hurbilpen

ezberdinek eragindakoa izan daiteke, edo gizakiok hizkuntza berri bat ikastean abiapuntutzat dagoeneko gutxienez hizkuntza bat badakigulako, eta hizkuntza berriaren ikasketa prozesua badakizkigun hizkuntzek baldintzatua dagoelako.

7.2 Euskararako Baliabide Andana

Ikerketa-galderak erantzuteaz gain, doktorego tesi honetan ekarpen mardula egin da hizkuntzaren prozesamendurako baliabideen sorkuntzan ere, nagusiki euskararako (ikus 7.4. Taula).

Hasteko, hizkuntza-ereduak aurrentrenatzeko helburua duten DeepText eta ZelaiHandi euskarazko corpus elebakarrak bildu ditugu, azken hau, lizentzia librez osatutako euskararako kalitatezko corpus elebakar handiena da gaur gaurkoz.

Jarraian, hizkuntza-ereduen ebaluazioari dagokionez, euskararako askotariko ataza sorta, besteak beste WiC$_{eu}$, EpecKorrefBin eta QNLI$_{eu}$ datumultzo berriak, biltzen dituen BasqueGLUE hizkuntzaren ulermena ebaluatzeko euskararako lehen proba-bankua eta euskararen gramatikan ebaluatzeko zorroztasun handiarekin ondutako pare-minimodun BL2MP (*Basque L2 student-based Minimal Pairs*) datu-multzoa sortu dira. Biak ala biak, ondasun baliotsuak izango direla uste dugu euskarazko hizkuntza-ereduen ulermen eta gramatika-gaitasunak ebaluatzeko etorkizuneko ikerketetan eta garapenetan.

Hizkuntza-ereduei dagokionez, lan honetan sortutako ElhBERTeu eredua ere ekarpen baliagarria izan da. Halaber, tesi-txosten honetan landu den euskarazko terminologia teknikoa ere ekarpen bat da bere horretan.

Euskararako baliabideez gain, swahilirako, suomierarako eta gaztelaniarako QNLI datu-multzo berriak ere sortu ditugu. Eta swahilirako aurrentrenatu diren BERT eredu lehiakorrak ere publikoki eskuragarri jarri ditugu.

7.3 Etorkizuneko Lanak

Tesi-txosten honetan landutako ikerlerroek, euskararako hizkuntza-ereduen ikerketan bidea ireki besterik ez dute egin. Aztertu diren gaietan sakontzeko

	Baliabidea	Tamaina	Hizkuntza
Corpusak	Deeptext	351M hitz	eu
	ZelaiHandi	660M hitz	eu
Proba-bankua	BasqueGLUE	9 ataza	eu
Datu-multzoak	WiC_{eu}	410K adibide	eu
	EpecKorrefBin	2K adibide	eu
	QNLI_{eu}	2K adibide	eu
	QNLI_{es}	38K adibide	es
	QNLI_{sw}	6K adibide	sw
	QNLI_{fi}	9K adibide	fi
	BL2MP	2K adibide	eu
Hizkuntza-ereduak	ElhBERTeu_{base}	124M param.	eu
	$\text{ElhBERTeu}_{medium}$	51M param.	eu
	bert-base-sw	124M param.	sw
	bert-medium-sw	51M param.	sw

7.4. Taula: Tesian zehar sortutako baliabide esanguratsuenen laburpena.

aukeraz gain, landu gabe utzi diren ertz ugariri hel geniezaioke. Hari horretatik segika, tesi honen irismenetik at utzitako sorkuntzaren lerroa ere hor dugu, egun puri-purian dagoena.

Lehen ikerketa-galderaren ildoari segika, ingelesez urtez urte proba-banku zorrotzagoak landu diren moduan, BasqueGLUE berrikusi eta eguneratzeko beharrik dagoen aztertzea komeni da. BasqueGLUE osatzen duten atazen egokitasuna berrikusi behar da, bereziki zailtasunari dagokionean, egungo euskarazko hizkuntza-ereduen NLU gaitasunetan aurrerapauso handiak eman baitira Llama-eus (Corral et al., 2024) eta Latxa-Llama (Etxaniz et al., 2024) bezalako hizkuntza-eredu handien etorrerarekin. NLUko ataza zailagoak ere gehitzea aurreikusi liteke, sen oneko arrazoiketa eta mundu ezagutza neurtuko duten ataza berriak gehituz. Bestalde, euskararako hizkuntza-eredu onenek ingeleserako ereduren bat dutela oinarrian ikusirik, ereduen euskarazko gaitasun linguistikoak bermatzen direla ziurtatzeko, BL2MP datu multzotik eratorritako CoLA[2] datu-multzo bat gehitu liteke gaitasun linguistikoak ebaluatzeko ataza gisa.

Bestalde, euskarazko sorkuntza ataza multzo zabala biltzen duen proba-banku bat sortzeko beharra ere identifikatu da, hizkuntza-eredu handien uler-

[2]Zuzentasun linguistikoko sailkapen ataza bitarra.

men gaitasunez gain sorkuntza gaitasunak ere ebaluatzen hasteko. Euskarazko sorkuntzarako proba-bankua laburpena, itzulpena, berridazketak, arrazoiketa matematikoa edo zuzentzaile gramatikala bezalako atazez osatu liteke.

Bigarren ikerketa-galderaren ildotik, baliabide gutxi izateak hizkuntza-eredua entrenatzeko orduan duen eragina aztertu ondoren, ZelaiHandi bezalako euskararako corpus elebakarrak handitzen eta kalitatezko iturri berriekin elikatzen jarraitzeaz gain, testu sintetikoaren bidean gehiago sakondu litekeela uste dugu.

Horretarako, itzulpen automatikoaren lerroa berraztertu daiteke hizkuntza-eredu handien bidezko itzulpen automatikoaren bidetik, dokumentu mailako itzulpenaren kalitateak gora egin baitu nabarmen. Gaztelania-euskara norabideaz gain, frantsesa-euskara edo ingelesa-euskara norabideak ere landu daitezke kalitate oso altuko edukietan zentratuz: euskaraz hutsuneak izan ditzakegun domeinuetan (elkarrizketa, medikuntza, mezuak, etab.) edo lizentzia librerik gabekoetan (ikas-materiala, fikzioa, ipuinak, etab.).

Euskarazko testua sortzeko jorratu daitekeen beste bide bat, itzuli ordez, euskaradun hizkuntza-eredu eleaniztunak erabiltzea izan liteke, batez ere euskarazko corpusak kamuts dabiltzan domeinu eta arloetan erderetako edukia eta ezagutza euskarara ekartzeko.

Hirugarren ikerketa-galderari dagokionez, tesian tokenizatzaileak eraikitzerakoan euskararen morfologia berariaz lantzea beharrezkoa ez dela ondorioztatu dugun arren, egungo tokenizatzaile sorta zabalaren aurrean, euskararen morfologia aberatsa egokiago tratatzen duenik dagoen aztertzeak ere mereziko luke. Adibidez, tokenizatzaile eraikitzean morfemen mugak kontutan izaten dituen MorphBPE (Asgari et al., 2025), eta *token-free* deritzen ereduak –hizpeko token mailan ordez byte mailan dihardutenak– aztertzea ere interesgarria izan liteke, esate baterako, byT5 (Xue et al., 2021a) edo *Byte Latent Transformer* (Pagnoni et al., 2024).

Tesi honen helburua euskararako hizkuntza-ereduetan oinarri sendo batzuk finkatzea izan da. Behin aurrentrenamenduko eta ebaluaziorako datumultzoak izanik, eta baliabide gutxi izatearen eta euskaren ezaugarri linguistikoen eraginak aztertuta, eredu berriak aurrentrenatzea tokatzen da. Alde batetik, lan honetan bazterrean utzitako kodetzaile motako eredu berriagoak entrenatzea dagokigu: Roberta (Liu et al., 2019), Deberta-V3 (He et al., 2020) eta ModernBERT (Warner et al., 2024) adibidez.

Kodetzailea soilik darabilten eredu diskriminatzaileen ostean, eredu sortzaileetara ere jauzi egitea ere badagokigu, hizkuntzaren prozesamenduko hainbat atazetarako ezinbestekoa. Sekuentziatik sekuentziarako hainbat atazatarako oso aproposak diren kodetzaile-deskodetzaile eredu arinek (BART (Lewis et al., 2020), T5(Raffel et al., 2020)) euskaraz ez dute dagokien arretarik jaso oraindik.

Deskodetzaile motako eredu sortzaileetan, berriz, eredu fundazional sendoak lortzeko bidean, estrategia berriak proposatzea ingelesezko sekulako ezagutza duten pisu irekiko Llama3 (Dubey et al., 2024), Gemma3 (Kamath et al., 2025), Qwen3 (Qwen Team, 2025) eta enparauak euskarara nola egokitu eta hauen tokenizatzaile sarraskitzaileak[3] euskararako abegikorragoak nola egin ikertzeko, tokenizatzaile arteko distilazioarekin (Minixhofer et al., 2025) adibidez.

Bukatzeko, hizkuntza-eredu handi horiek euskararako egokitu ondoren tresna erabilgarriak izan daitezen, instrukzioetan birdoitu (Ouyang et al., 2022) beharko dira eta gizakion xedeekin lerrokatzeko RLHF (*Reinforcement Learning from Human Feedback*, (Ouyang et al., 2022)) edo DPO (*Direct Preference Optimization*, (Rafailov et al., 2023)) teknikak aplikatu beharko zaizkie. Jarraian, inferentzia garaian hausnarketa-kate (*Chain-of-Thought*) luzeagoko arrazoiketa prozesuekin (DeepSeek-AI et al., 2025; Qwen Team, 2025) ereduak sendotzeko bidea ere jorratu daiteke. Instrukzio, lerrokatze eta arrazoiketan aurrerapauso esanguratsuak emateko, ordea, euskararako datu mamitsu sorta zabala sortzea izango dugu erronka nagusia ondorengo urteetan.

[3]Tokenizatzaile oldarkorraren eraginez eredu hauek euskaraz erabiltzea bizpahiru bider motelagoa eta garestiagoa da.

Bibliografia

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al. (2024a). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024b). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Abonizio, H. Q., Paraiso, E. C., and Barbon, S. (2021). Toward text data augmentation for sentiment analysis. *IEEE Transactions on Artificial Intelligence*, 3(5):657–668.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Acs, J., Hamerlik, E., Schwartz, R., Smith, N. A., and Kornai, A. (2023). Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, pages 1–40.

Addley, E. (2023). 'AI' named most notable word of 2023 by Collins dictionary. *The Guardian*. Accessed: 2025-05-23.

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al. (2021). Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Agerri, R., Centeno, R., Espinosa, M., de Landa, J. F., and Rodrigo, A. (2021). Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Ahmed, K. and Buys, J. (2024). Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 724–728.

Alammar, J. (2018a). The illustrated bert, elmo, and co. (how nlp cracked transfer learning) [blog post]. `https://jalammar.github.io/illustrated-bert/`.

Alammar, J. (2018b). The illustrated transformer [blog post]. `https://jalammar.github.io/illustrated-transformer/`.

Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2004). Design and development of a named entity recognizer for an agglutinative language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.*

Alegria, I., Arregi, O., Ezeiza, N., and Fernández, I. (2006). Lessons from the development of a named entity recognizer for Basque. *Procesamiento del lenguaje natural*, 36:25–37.

Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A. P., Srivastav, V., et al. (2025). Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Alrajhi, W. A., Al-Khalifa, H., and AlSalman, A. (2022). Assessing the linguistic knowledge in arabic pre-trained language models using minimal pairs. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193.

Amjad, M., Sidorov, G., and Zhila, A. (2020). Data augmentation using machine translation for fake news detection in the urdu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2537–2542.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Aranzabe, M. J., Atutxa, A., Bengoetxea, K., de Ilarraza, A. D., Goenaga, I., Gojenola, K., and Uria, L. (2015). Automatic conversion of the basque dependency treebank to universal dependencies. In *Proceedings of the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241.

Arase, Y. and Tsujii, J. (2019). Transfer fine-tuning: A bert case study. 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)*, pages 5393–5404.

Arora, S., May, A., Zhang, J., and Ré, C. (2020). Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663.

Artetxe, M., Aldabe, I., Agerri, R., Perez-De-Viñaspre, O., and Soroa, A. (2022). Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390.

Asgari, E., Kheir, Y. E., and Javaheri, M. A. S. (2025). Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *arXiv preprint arXiv:2502.00894*.

Asher, N., Bhar, S., Chaturvedi, A., Hunter, J., and Paul, S. (2023). Limits for learning with language models. In *Proceedings of the 12th Joint*

Conference on Lexical and Computational Semantics (SEM 2023)*, pages 236–248.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bai, F., Ritter, A., and Xu, W. (2021). Pre-train or annotate? domain adaptation with a constrained budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5002–5015.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S.Ñ., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. (2024). The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Bao, E., Pérez, A., and Parapar, J. (2023). Conversations in galician: a large language model for an underrepresented language. *arXiv preprint arXiv:2311.03812*.

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Meachair, M. J. , and Foster, J. (2021). gabert–an irish language model. *arXiv preprint arXiv:2107.12930*.

Bel, N., Punsola, M., and Ruiz-Fernández, V. (2024). Catcola, catalan corpus of linguistic acceptability. *Procesamiento del Lenguaje Natural*, 73:177–190.

Ben Allal, L., Lozhkov, A., Penedo, G., Wolf, T., and von Werra, L. (2024). Cosmopedia.

Bender, E. M., Bubeck, S., and Strickland, E. (2025). Chm live | the great chatbot debate: Do llms really understand? Recorded March 25, 2025. Accessed April 24, 2025.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

BIG-bench (2021). Beyond the imitation game benchmark (big-bench). `https://github.com/google/BIG-bench`.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cagatan, O. V. (2023). Toddlerberta: Exploiting babyberta for grammar learning and language understanding. *arXiv preprint arXiv:2308.16336*.

Carrino, C. P., Costa-jussa, M. R., and Fonollosa, J. A. (2020). Automatic spanish translation of squad dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Choi, C., Jeong, Y., Park, S., Won, I., Lim, H., Kim, S., Kang, Y., Yoon, C., Park, J., Lee, Y., et al. (2024). Optimizing language augmentation for multilingual large language models: A case study on korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12514–12526.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Clark, A., de Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. (2022). Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pages 4057–4086. PMLR.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020a). Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020b). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Coloma, G. (2015). Efectos de compensación entre indicadores de la complejidad de los idiomas. Technical report, Serie Documentos de Trabajo.

Conneau, A. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Corral, A., Sarasua, I., and Saralegi, X. (2024). Pipeline analysis for developing instruct llms in low-resource languages: A case study on basque.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cumbreras, M. G., Cámara, E. M., Román, J. V., and Morera, J. G. (2016). Tass 2015–the evolution of the spanish opinion mining systems. *Procesamiento del Lenguaje Natural*, I(56):33–40.

David, D. (2020). Swahili : News classification dataset. The news version contains both train and test sets.

De Gibert, O., Nail, G., Arefyev, N., Bañón, M., Van Der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., et al. (2024). A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.

de Vries, W., Wieling, M., and Nissim, M. (2023). Dumb: A benchmark for smart evaluation of dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang,

H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X.,

Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. (2024). Deepseek-v3 technical report.

Deng, L. and Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doshi, M., Dabre, R., and Bhattacharyya, P. (2024). Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Etchegoyhen, T. and Gete, H. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3792–3800. European Language Resources Association.

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., and Soroa, A. (2024). Latxa: An open language model and evaluation suite for basque. *arXiv preprint arXiv:2403.20266*.

Ezeiza, N., Alegria, I., Arriola, J. M., Urizar, R., and Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Fabien, M., Villatoro-Tello, E., Motlicek, P., and Parida, S. (2020). Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Fang, Z., He, Y., and Procter, R. (2024). Cwtm: Leveraging contextualized word embeddings from bert for neural topic modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4273–4286.

Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Feng, Y., Dohmatob, E., Yang, P., Charton, F., and Kempe, J. (2024). A tale of tails: Model collapse as a change of scaling laws. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Fierro, C., Dhar, R., Stamatiou, F., Garneau, N., and Søgaard, A. (2024). Defining knowledge: Bridging epistemology and large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16096–16111.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezudo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). The

GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Geiping, J. and Goldstein, T. (2023). Cramming: training a language model on a single gpu in one day. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11117–11143.

Grießhaber, D., Maucher, J., and Vu, N. T. (2020). Fine-tuning bert for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. *Interspeech 2020*.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Gutiérrez Fandiño, A., Armengol Estapé, J., Pámies, M., Llop Palao, J., Silveira Ocampo, J., Pio Carrino, C., Armentano Oller, C., Rodriguez Penagos, C., Gonzalez Agirre, A., and Villegas, M. (2022). Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hossain, T., Dev, S., and Singh, S. (2023). Misgendered: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020a). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020b). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Hu, Y., Jing, X., Ko, Y., and Rayz, J. T. (2020c). Misspelling correction with pre-trained contextual language model. In *2020 ieee 19th international conference on cognitive informatics & cognitive computing (icci* cc)*, pages 144–149. IEEE.

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging.

Huebner, P. A., Sulem, E., Cynthia, F., and Roth, D. (2021). Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Izsak, P., Berchansky, M., and Levy, O. (2021). How to train bert with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652.

Jansen, T., Tong, Y., Zevallos, V., and Suarez, P. O. (2022). Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Jones, C. R. and Bergen, B. K. (2024). People cannot distinguish gpt-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020a). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020b). The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Jurafsky, D. and Martin, J. H. (2019). Speech and language processing 3rd edition draft.

Kafle, K., Yousefhussien, M., and Kanan, C. (2017). Data augmentation for visual question answering. In *Proceedings of the 10th international conference on natural language generation*, pages 198–202.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for

Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al. (2025). Gemma 3 technical report. *CoRR*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kauf, C. and Ivanova, A. (2023). A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. (2021). Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Kurihara, K., Kawahara, D., and Shibata, T. (2022). Jglue: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lai, V., Nguyen, C., Ngo, N., Nguyn, T., Dernoncourt, F., Rossi, R., and Nguyen, T. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Laka, I. (1996). *A brief grammar of Euskara, the Basque language*. Universidad del País Vasco, Euskal Herriko Unibertsitatea.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Leong, W. Q., Ngui, J. G., Susanto, Y., Rengarajan, H., Sarveswaran, K., and Tjhi, W. C. (2023). Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. (2020). Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2021a). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2022). Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M. T., Stoyanov, V., and Li, X. (2021b). Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Shen, Y., Zhu, H., Xu, L., Qian, Z., Song, S., Zhang, K., Tang, J., Zhang, P., Yang, B., et al. (2024). Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *arXiv preprint arXiv:2411.06096*.

Liu, Z., Wang, Y., Kasai, J., Hajishirzi, H., and Smith, N. A. (2021). Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842.

López de Lacalle, M., Saralegi Urizar, X., Saizar, A., Urbizu, G., and Corral, A. (2023). Strategies for bilingual intent classification for small datasets scenarios.

Lothritz, C., Lebichot, B., Allix, K., Veiber, L., Bissyande, T. F. D. A., Klein, J., Boytsov, A., Goujon, A., and Lefebvre, C. (2022). Luxembert: Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference, 2022*, pages 5080–5089.

Louvan, S. and Magnini, B. (2020). Simple is better! lightweight data augmentation for low resource slot filling and intent classification. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 167–177.

Lu, A., Zhang, H., Zhang, Y., Wang, X., and Yang, D. (2023). Bounding the capabilities of large language models in open text generation with prompt constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1982–2008.

López de Lacalle, M., Saralegi, X., and San Vicente, I. (2020). Building a task-oriented dialog system for languages with no training data: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2796–2802.

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., and Futrell, R. (2023a). Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, 241:105543–105543.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2023b). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Martin, G., Mswahili, M. E., Jeong, Y.-S., and Young-Seob, J. (2022). Swahbert: Language model of swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., De La Clergerie, V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Mehta, S., Sekhavat, M. H., Cao, Q., Horton, M., Jin, Y., Sun, C., Mirzadeh, I., Najibi, M., Belenko, D., Zatloukal, P., et al. (2024). Openelm: An efficient language model family with open training and inference framework. *CoRR*.

Meyer, S., Elsweiler, D., Ludwig, B., Fernandez-Pichel, M., and Losada, D. E. (2022). Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6.

Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (2020). On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Mikhailov, V., Shamardina, T., Ryabinin, M., Pestova, A., Smurov, I., and Artemova, E. (2022). Rucola: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Minixhofer, B., Vulić, I., and Ponti, E. M. (2025). Cross-tokenizer distillation via approximate likelihood matching. *arXiv preprint arXiv:2503.20083*.

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Mohseni, M. and Tebbifakhr, A. (2019). Morphobert: A persian ner system with bert and morphological analysis. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 23–30.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Nguyn, T., VanÑguyen, C., Lai, V. D., Mn, H., Ngo, N. T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2024). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.

Nielsen, D. (2023). Scandeval: A benchmark for scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201.

Niu, J., Lu, W., and Penn, G. (2022). Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Nzeyimana, A. and Rubungo, A.Ñ. (2022). Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pagnoni, A., Pasunuru, R., Rodriguez, P., Nguyen, J., Muller, B., Li, M., Zhou, C., Yu, L., Weston, J., Zettlemoyer, L., et al. (2024). Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*.

Pan, Y., Li, X., Yang, Y., and Dong, R. (2020). Morphological word segmentation on agglutinative languages for neural machine translation. *arXiv preprint arXiv:2001.01589*.

Papadimitriou, I., Futrell, R., and Mahowald, K. (2022). When classifying grammatical role, bert doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J. Y., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.-W., and Cho, K. (2021). KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Passali, T., Mavropoulos, T., Tsoumakas, G., Meditskos, G., and Vrochidis, S. (2022). Lard: Large-scale artificial disfluency generation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2327–2336, Marseille, France. European Language Resources Association.

Penedo, G., Kydlíček, H., Sabolčec, V., Messmer, B., Foroutan, N., Jaggi, M., von Werra, L., and Wolf, T. (2024). Fineweb2: A sparkling update with 1000s of languages.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Phang, J., Yeres, P., Swanson, J., Liu, H., Tenney, I. F., Htut, P. M., Vania, C., Wang, A., and Bowman, S. R. (2020). `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Pilehvar, M. T. and Camacho-Collados, J. (2019). Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language Resources and Evaluation. Springer. Volume 45, Issue 2, pp 121-142. ISSN 1574-020X. DOI 10.1007/s10579-010-9131-y. official.*

Qwen Team, Q. (2025). Qwen3: Think deeper, act faster.

Radford, A. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110. Association for Computational Linguistics.

Ravfogel, S., Goldberg, Y., and Tyers, F. (2018). Can lstm learn to capture agreement? the case of basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107.

Rei, M., Felice, M., Yuan, Z., and Briscoe, T. (2017). Artificial error generation with machine translation and syntactic patterns. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 287–292.

Rodriguez-Penagos, C., Armentano-Oller, C., Villegas, M., Melero, M., Gonzalez, A., Bonet, O. d. G., and Pio, C. C. (2021). The catalan language club. *arXiv preprint arXiv:2112.01894*.

Romanou, A., Foroutan, N., Sotnikova, A., Chen, Z., Nelaturu, S. H., Singh, S., Maheshwary, R., Altomare, M., Haggag, M. A., A, S., Amayuelas, A., Amirudin, A. H., Aryabumi, V., Boiko, D., Chang, M., Chim, J., Cohen, G., Dalmia, A. K., Diress, A., Duwal, S., Dzenhaliou, D., Florez, D. F. E., Farestam, F., Imperial, J. M., Islam, S. B., Isotalo, P., Jabbarishiviari, M., Karlsson, B. F., Khalilov, E., Klamm, C., Koto, F., Krzemiński, D., de Melo, G. A., Montariol, S., Nan, Y., Niklaus, J., Novikova, J., Ceron, J. S. O., Paul, D., Ploeger, E., Purbey, J., Rajwal, S., Ravi, S. S., Rydell, S., Santhosh, R., Sharma, D., Skenduli, M. P., Moakhar, A. S., Moakhar, B. S., Tamir, R., Tarun, A. K., Wasi, A. T., Weerasinghe, T. O.,

Yilmaz, S., Zhang, M., Schlag, I., Fadaee, M., Hooker, S., and Bosselut, A. (2024). Include: Evaluating multilingual language understanding with regional knowledge.

Ruokolainen, T., Kauppinen, P., Silfverberg, M., and Lindén, K. (2019). A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. (2020). Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

San Vicente, I., Urbizu, G., Corral, A., Beloki, Z., and Saralegi, X. (2024). Zelaihandi: A large collection of basque texts.

Sang, E. F. (2002). Tjong kim (2002).“introduction to the conll-2002 shared task: Language-independent named entity recognition”. In *COLING-02: The 6th Conference on Natural Language Learning*.

Schwenk, H. and Li, X. (2018). A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Seelawi, H., Tuffaha, I., Gzawi, M., Farhan, W., Talafha, B., Badawi, R., Sober, Z., Al-Dweik, O., Freihat, A. A., and Al-Natsheh, H. (2021). ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sennrich, R. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., and Mihalcea, R. (2024). Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Someya, T. and Oseki, Y. (2023). Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1536–1549.

Song, Y., Krishna, K., Bhatt, R., and Iyyer, M. (2022). Sling: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634.

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and De Ilarraza, A. D. (2012). Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Suijkerbuijk, M., Prins, Z., de Heer Kloots, M., Zuidema, J., and Frank, S. L. (2024). Blimp-nl: A corpus of dutch minimal pairs and grammaticality judgements for language model evaluation.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206.

Sun, Q., Amin, M., Yan, B., Martell, C., Markman, V., Bhasin, A., and Ye, J. (2015). Transfer learning for bilingual content classification. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2147–2156.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Tenney, I., Das, D., and Pavlick, E. (2019a). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Tensmeyer, C., Wigington, C., Davis, B., Stewart, S., Martinez, T., and Barrett, W. (2018). Language model supervision for handwriting recognition model adaptation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 133–138. IEEE.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Trotta, D., Guarasci, R., Leonardelli, E., and Tonelli, S. (2021). Monolingual and cross-lingual acceptability judgments with the italian cola corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Turing, A. (1950). I.—computing machinery and intelligence. *Mind*, LIX(236):433–460.

Ubani, S., Polat, S. O., and Nielsen, R. (2023). Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., and Soroa, A. (2022). BasqueGLUE: A Natural Language Understanding Benchmark for Basque. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., and Soroa, A. (2023a). Scaling laws for bert in low-resource settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7771–7789.

Urbizu, G., San Vicente, I., Saralegi, X., and Corral, A. (2023b). Not enough data to pre-train your language model? mt to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836.

Urbizu, G., Soraluze, A., and Arregi, O. (2020). Sequence to sequence coreference resolution. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 39–46.

Urbizu, G., Zulaika, M., Saralegi, X., and Corral, A. (2024). How well can bert learn the grammar of an agglutinative and flexible-order language? the case of basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8334–8348.

Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., et al. (2024). Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Van Esch, D., Lucassen, T., Ruder, S., Caswell, I., and Rivera, C. (2022). Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.Ñ., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vega, M. G., Díaz-Galiano, M. C., Cumbreras, M. G., del Arco, F. M. P., Montejo-Ráez, A., Zafra, S. M. J., Cámara, E. M., Aguilar, C. A., Cabezudo, M. A. S., Chiruzzo, L., et al. (2020). In *IberLEF@ SEPLN*.

Vilares, D., Garcia, M., and Gómez-Rodríguez, C. (2021). Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018*, page 353.

Wang, J., Lu, Y., Weber, M., Ryabinin, M., Adelani, D., Chen, Y., Tang, R., and Stenetorp, P. (2025). Multilingual language model pretraining using machine-translated data. *arXiv preprint arXiv:2502.13252*.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., and Zhuang, C. (2023). Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Computing Research Repository*, arXiv:2301.11796.

Warstadt, A., Parrish, A., Liu, H., Monananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020a). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Warstadt, A., Zhang, Y., Li, X., Liu, H., and Bowman, S. (2020b). Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Whitehouse, C., Choudhury, M., and Aji, A. (2023). Llm-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., and Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020a). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020b). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. (2024). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862.

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., and Stoyanov, V. (2022). Training trajectories of language models across scales.

Xiang, B., Yang, C., Li, Y., Warstadt, A., and Kann, K. (2021). Climp: A benchmark for chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., and Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2021a). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021b). 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*, pages 483–498.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Yuan, Z. and Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61.

Zaczynska, K., Feldhus, N., Schwarzenberg, R., Gabryszak, A., and Möller, S. (2020). Evaluating german transformer language models with syntactic agreement tests. *arXiv preprint arXiv:2007.03765*.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zevallos, R., Ortega, J., Chen, W., Castro, R., Bel, N., Toshio, C., Venturas, R., Aradiel, H., and Melgarejo, N. (2022). Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.

Zhang, M., Wang, W., Liu, X., Gao, J., and He, Y. (2018). Navigating with graph representations for fast and scalable decoding of neural language models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhang, Y., Warstadt, A., Li, X., and Bowman, S. (2021). When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1112–1125.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., and Zong, C. (2019). Ncls: Neural cross-lingual summarization. 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)*, pages 3054–3064.

Zulaika, M. and Saralegi, X. (2025). Basqbbq: A qa benchmark for assessing social biases in llms for basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767.

A. ERANSKINA

Corpusgintza: Bildu Guztiak!

Laburpena

Eranskin honetan, euskararako hizkuntza-eredu neuronalak aurrentrenatzeko bildu ditugun DeepText eta ZelaiHandi kalitatezko euskarazko corpus elebakarrak aurkezten dira. DeepTextek 351M hitz ditu eta 2020an bildu zen. Azkenekoz 2025ean eguneratutako ZelaiHandik, berriz, 660M hitz ditu, eta gaur gaurkoz, euskarazko kalitatezko testuen lizentzia libreko corpusik handiena da. Bien kasuan, aurrez eskuragarri zeuden euskarazko kalitatezko corpusen tamaina eta aniztasuna handitzea izan da helburu nagusia.

A.1 Sarrera

Hizkuntzalaritzaren arloan, corpusak testu-bilduma egituratu handiak dira, gaur egun, euskarri elektronikoetan gordeak normalean, zorionez. Testuak hizkuntza, dialekto, garai, erregistro edo domeinu jakin batzuen erakusgarri izateko aukeratu ohi dira.

Hizkuntzalaritza konputazionalaren alorrean, corpusak ezinbestekoak dira metodo estatistikoen bitartez analisiak egiteko, besteak beste, hitz edo fenomeno jakinen agerpenak zenbatzeko edota aldaera linguistikoen ezaugarri teorikoak erabilera errealean balioztatzeko.

Hizkuntzaren prozesamendua ere ezin da ulertu corpusik gabe; ezinbestekoak zaizkigu algoritmoak eta eredu estatistikoak garatu eta entrenatzeko. Testuzko corpus horiek eskuz anotatzeko lana hartzen denean, hainbat atazetarako etiketatzaileak edo sailkatzaileak garatzeko ere baliagarriak zaizkigu. Hizkuntza-ereduak aurrentrenatzeko, ordea, ez da anotaziorik behar, testu gordina soilik, eta horregatik, corpus hauek eratzeko testu kopuru itzelak biltzen dira.

Hizkuntza-eredu neuronalak aurrentrenatzeko sekulako corpusak darabilzkigu gaur egun. Edo darabiltzate: BERT (Devlin et al., 2019) hizkuntza-eredua, Transformerretan oinarritutako lehena, 3,3B hitzeko[1] ingeleseko corpus batekin entrenatu zen 2018an, eta berriki Llama3 (Dubey et al., 2024) familiako hizkuntza-ereduak 11T hitzeko[2] corpus erraldoi batekin entrenatu dira, nagusiki ingelesezko testu bildumez osatua, baina beste hizkuntzetako testuak eta kodea ere baditu tartean.

A.2 Hizkuntza-Ereduak Elikatzeko Euskarazko Corpusak

Munduko hizkuntzen gehiengoaren errealitatea, euskararena barne, bestelakoa da, ordea (Joshi et al., 2020b). Euskarak duen egungo hiztun kopuruarekin, miloi bat pasatxo optimistak izanda, nekez iritsiko gara ingeleserako eskuartean ditugun corpusen tamainako euskarazko corpusik izatera epe labur zein ertainean. Hala ere, euskararen egoera osasuntsua da hizkuntzaren prozesamendurako baliabideei dagokionez, hizkuntzen gehiengoaren gainetik (Joshi et al., 2020b). Hau, hein batean, euskal herrian hizkuntza teknologien alorrean urte luzez egindako lanaren uzta da, horren erakusle, hiztun kopuruarekiko hizkuntzaren prozesamenduko publikazioei dagokionez, euskara gailurrean egotea (Van Esch et al., 2022).

Tesi honetan lanean hasi aurretik, euskararako kalitatezko corpus eskuragarririk handiena BERTeus (Agerri et al., 2020) eredua aurre-entrenatzeko erabilitako *Basque Media Corpus* (BMC) izan da. BMC corpusa eraikitzeko

[1] Oharra: tesi-txosten honetan zehar, tokenizatu gabeko corpusen tamainak zehazteko hitzak erabili dira tokenen beharrean, tokenizatu osteko hizpeko tokenekin ez nahasteko.

[2] 15,6T hizpeko token.

testu garbia eta egoki egituratutakoa[3] bildu zuten sareko albiste iturrietatik eta Wikipediatik.

BMCk orotara 224,6M hitz ditu, iturri hauetatik bilduak: Wikipedia (35M), Berria egunkaria (81M), EiTB albisteak (28M), Argia aldizkaria (16M) eta tokiko albiste iturriak (64,6M).

Gerora, EusCrawl (Artetxe et al., 2022) ere argitaratu da, kalitate handiko euskal corpus librea (289M hitz). Garai bertsuan sortutako DeepText corpusa baino txikiagoa izan arren, euscrawl osatzen duten euskarazko testuen lizentziek erabat baimentzen dute nahieran zabaltzea eta hizkuntza-ereduak entrenatzeko erabiltzea; corpusari emandako egitura eta formatua ere lagungarriak dira horretan. JSONL formatuan dator, dokumentutan banatuta, eta metadatuz aberastuta (besteak beste, iturria, urla, egilea eta data). Corpus hau, *Roberta-eus-euscrawl* (Artetxe et al., 2022) eta Latxa (Etxaniz et al., 2024) hizkuntza-ereduak aurrentrenatzeko erabili da adibidez.

Badira beste hainbat corpus eleaniztun euskarazko hizkuntza-ereduak entrenatzeko baliagarriak, euskararako berariaz landu ez diren arren, saretik masiboki crawleatutako testuez osatuak: CC-100, 270M hitz (Conneau et al., 2020); mC4, 576M hitz (Raffel et al., 2020); Colossal OSCAR, 283M hitz (Jansen et al., 2022); CulturaX, 600M hitz (Nguyn et al., 2024); HPLTv2, 777M hitz (De Gibert et al., 2024); eta FineWeb2, 711M hitz (Penedo et al., 2024).

Zerrendatutako corpus guzti horiek Common Crawleko uneko argazki ezberdinetatik bildu dira, eta ondoren, garbiketak eta hizkuntza filtroa aplikatu zaizkie. Prozesu horretan posible da euskarazko dokumenturen bat baztertu izana, eta aldi berean, erdarazko testuren bat euskara gisa hartu izana, corpusari zarata gehituz.

Saretik arrantzatutako testuen artean, posible da euskara izan arren kalitate handirik gabeko testua ere bildu izana, baina corpus tamaina handietan zarata apur bat izateak eragin handirik ez duela dirudi (Artetxe et al., 2022).

Corpus hauek OCD-by lizentziapean argitaratzen dira, hau da, testubildumak CC-BY 4.0 lizentzia du, baina corpusa osatzen duten testuetako bakoitzak jatorrizko lizentziapean jarraitzen du. Hala ere, lizentzia hori duten corpusekin aurrentrenatutako ereduak lizentzia librean zabaldu ohi dira maiz (Apache-2.0 edo MIT lizentziapean adibidez).

[3]Paragrafo eta dokumentu mugak izatea, ordura arteko esaldi mailako corpusak ez bezala.

Aurrekari hauekin, DeepText eta ZelaiHandi euskarazko corpus elebakarrak sortu ditugu doktorego tesi honetan zehar, lehena tesiaren hasieran, tesian zehar sortu diren ereduetako batzuk entrenatzeko, eta bigarrena amaieran, etorkizuneko euskarazko hizkuntza-ereduak elikatzeko. Corpus hauek hizkuntza-eredu neuronalak entrenatzeko datu beharrei erantzutea dute xede: tamaina handikoak, dokumentu mailan antolatuak, kalitatezkoak (egiturari eta edukiari dagokionez), domeinu aniztasuna bermatuz eta posible den heinean corpus eguneratuak izatea.

A.3 DeepText

Doktoretza tesi honen hastapenetan, ElhBERTeu hizkuntza-eredua entrenatu aurretik, euskarazko kalitatezko corpus elebakar handiena eraikitzeari ekin zitzaion. Ordura arte euskararako corpusik handienak, BMCk (224M hitz), albisteez eta Wikipediako testuez zeuden osatuak nagusiki, eta DeepText corpusean bestelako domeinuetako euskarazko testua ere biltzen saiatu gara. Bestalde, hizkuntza-ereduen ezagutza eguneratua izateko, corpusak eguneratzeko beharra identifikatu da; adibidez, BMCn entrenatutako ereduek COVID-19aren pandemiari lotutako hiztegia egoki tokenizatzeko arazoak dituztela detektatu da, pandemiaz aurreko testuek osatzen baitute.

DeepText corpusa eraikitzeko, BERTeus (Agerri et al., 2020) entrenatzeko erabilitako BMC corpusaren iturrietatik abiatu gara. Zehazki, 2020ko azarora arteko edukiekin eguneratu ditugu dagoeneko BMCn dauden iturri gehienetako edukiak, eta gainera, albiste iturri berriak eta bestelako domeinuetako testuak bildu ditugu, besteak beste, zientzia, literatura eta azpidatziak, corpusaren aniztasuna handitze aldera. Zientziaren barruan, testu akademikoak eta zientzia dibulgazioko testuak bildu dira. Literaturaren barruan, nobelak, poesia, antzerki obrak eta saiakerak aurki daitezke. Iturri batzuetako testuak (zientzia eta literaturako iturriei dagokien gehiengoa) PDF eta EPUB formatuetan eskuratu dira saretik, eta horietatik testu garbia erauzi eta lerro jauziak konpondu behar izan dira. Corpusaren osaketa eta domeinu bakoitzetik bildutako testuaren tamaina A.1. Taulan daude ikusgai.

DeeepText corpusa *txt* formatuan dator, eta dokumentuka banatuta dago. Dokumentu bakoitzaren inongo metadaturik ez du (ez iturririk, ez URLrik, ez datarik), eta dokumentu mailan nahastu da. Interneten eskura dauden edukiak biltzen dituen arren, libreki zabaltzen uzten ez duten lizentzia duten

Domeinua	Hitz	Iturriak
Albisteak	224M	Berria, Argia, tokiko albisteak, Egunkaria, EiTB, Gara
Zientzia	58M	ADDI, Ikergazte, Ekaia, EHU argitalpenak, zientzia.eus, zientzia kaiera
Entziklopedikoa	40M	Wikipedia
Literatura	24M	Susa, Booktegi, Erein, klasikoak, antzerkiak
Bestelakoak	7M	Azpidatziak
Guztira	**351M**	

A.1. Taula: DeepText corpusaren osaketa.

eduki batzuk[4] biltzen ditu tartean corpus honek. Horregatik, ez da corpusaren zabalkunde handirik egin, ElhBERTeuren aurrentrenamenduaren erreproduzigarritasuna bermatzeko corpusa publikoki eskuragarri jarri den arren.

A.4 Zelai Handi

DeepText corpusak tesian zehar izandako corpus beharrak soberan ase ditu, baina biltzen dituen testuetako batzuen lizentzia itxiek ez[5] dute baimentzen euskararako hizkuntza-eredugintzan diardugun ikertzaile eta garatzaile ororen eskura jartzerik. Horretaz gain, 2020ko azarora arteko testuak soilik biltzen dituenez, gaur arteko testuekin eguneratzeko beharra ere bazegoen.

Behar horiek asetzeko sortu da ZelaiHandi[6], 660M hitz (5,8GB) dituen euskararako corpus elebakarra (San Vicente et al., 2024). Guk dakigunez, une honetan euskararako eskuragarri dagoen lizentzia libreko kalitatezko testu bilduma eguneratu handiena da, eskuz aukeratutako iturrietatik bildua.

ZelaiHandi euskararako hizkuntza-ereduei jaten emateko aproposa izan dadin, dokumentuka antolatuta dago, paragrafoen mugak mantenduz. Lizentzia libreko eskuz aukeratutako kalitatezko iturrietatik bildu da, eta aldi berean, domeinu aniztasuna bermatzen saiatu gara ahal den neurrian.

[4]EITBko albisteak, Egunkaria, Gara eta ADDIko argitalpen zientifiko, literatura obra eta azpitituluetako batzuk.

[5]Praktikan iturri horietako testuak Common Crawlen bitartez bildutako corpusetan nahieran zabaldu eta erabiltzen diren arren.

[6]`https://huggingface.co/datasets/orai-nlp/ZelaiHandi`

Iturria	**Domeinua**	**Hitz**	**Lizentzia**
Tokikom	albisteak	187,6M	cc-by / cc-by-sa
Berria	albisteak	148,6M	cc-by-sa 4.0
Hitzak	albisteak	83,2M	cc-by-sa 4.0
Eusko Legebiltzarra	administratiboa	83,0M	domeinu publikoa
Wikipedia	entziklopedikoa	72,8M	cc-by-sa 4.0
Argia	albisteak	18,4M	cc-by-sa 4.0
ADDI	zientzia	16,4M	hainbat cc aldaera
Opendata Euskadi subs	azpidatziak	16,4M	cc-by-sa
GFA	administratiboa	8,7M	berezia
Zientzia.eus	zientzia	8,6M	cc-by-sa
Susa	literatura	5,8M	cc-by
BFA	administratiboa	4,3M	berezia
Zientzia Kaiera	zientzia	2,0M	cc-by-sa 3.0
Ekaia	zientzia	1,8M	cc-by-nc-nd 4.0
Ikergazte	zientzia	1,6M	cc-by-sa 3.0
Game-Erauntsia	bideojokoak	0,4M	cc-by-sa 4.0
Guztira		**659,6M**	

A.2. Taula: ZelaiHandi corpusa osatzen duten testuen iturriak, domeinua, tamaina eta lizentziarekin batera. Tokikom, euskarazko toki komunikabide taldetik, ondorengoak bildu dira: 28kanala, Aiaraldea, Aikor, Aiurri, Alea, Amezti, Anboto, Antxetamedia, Ataria, Baleike, Barren, Begitu, Berriketan, Drogetenitturri, Erran, Euskalerriairratia, Gitb, Goiena, Guaixe, Hiruka, Karkara, Kronika, Mailope, Maxixatzen, Noaua, Plaentxia, Txintxarri, Uriola, Urolakostakohitza, Uztarria eta Zarauzkohitza. Hitzakek ondorengo iturriak hartzen ditu barne: Hitza.eus, Irutxuloko Hitza, Bidasoako Hitza, Busturialdeko Hitza, Goierriko Hitza, Lea-Artibai eta Mutrikuko Hitza, Oarsoaldeko Hitza, Otamotz eta Goiberri.

Corpusak guztira 659.609.830 hitz[7] ditu eta 2,26M dokumentu[8]. ZelaiHandi corpusa osatzen duten testuen iturriak biltzen ditu A.2. Taulak, bakoitzari dagokion domeinua, hitz kopurua eta lizentziarekin batera.

ZelaiHandi osatzen duten dokumentuek cc-by, cc-by-sa, cc-by-nc edo lizentzia baliokideak dituzte. Aldundietako (BFA eta GFA) web orrietatik bil-

[7]"text" atributuaren gainean $wc - w$ komandoaren bidez zenbatuta.

[8]2025eko otsailaren 13ra arteko testuak biltzen ditu, baina aldizkako eguneraketak izatea aurreikusten da.

dutako testuek neurrira egindako lizentziak[9][10] dituzte, baina funtsean, ez dute eragozpenik jartzen testuak biltzeko, zabaltzeko eta hizkuntza-ereduak entrenatzeko erabiltzeko, erabilera komertzialerako barne. DeepText-en bildutako testu iturri batzuk kanpoan utzi behar izan ditugu, lizentzia itxia zutelako, besteak beste, Egunkaria, EiTBko albisteak, Gara, literatur testuetako batzuk (Susako itzulpenak, Erein, etab.), ADDIko dokumentuetako batzuk edota azpidatzien azpimultzo bat.

DeepTextekin ez bezala, ZelaiHandi metadatuekin aberastutako JSONL formatuan zabaldu da, besteak beste, euscrawl edo HPLTv2 corpusak darabilten JSONL formatuen parekoa. JSONL formatua erabilerraza da epe edo iturri jakinak iragazteko, ebaluazioko kutsatzeak ekiditeko, corpusetik dokumentu zehatzak ezabatzeko eskaerei erantzuteko eta bidenabar Europar legedia berrien eskakizunetara gerturatzen da. Corpusa osatzen duten dokumentuetako bakoitzak ondorengo atributuak ditu (*title*, *author* eta *date* hutsik egon daitezke):

```
{
    "id": "dokumentu bakoitzaren id-a",
    "source": "iturria",
    "license": "dokumentuaren lizentzia",
    "lang": "eu",
    "url": "dokumentuaren url-a",
    "title": "dokumentuaren izenburua",
    "author": "idazlea",
    "date": "publikazio data yyyy[-mm-dd[Thh:mm:ssZ]]",
    "text": "Dokumentuaren edukia testu hutsean,
        erabiltzeko prest. Izenbururik balu (albiste
        eta artikuluetan), hemen sartuta dago jada.",
    "domain": "testuaren domeinua: ['news',
        'administrative', 'wikipedia', 'science',
        'subtitles', 'literature', 'videogames']"
}
```

[9] https://jjggbizkaia.eus/eu/pribadutasun-politikea.
[10] https://www.bngipuzkoa.eus/WAS/CORP/DJGPortalWEB/aviso_legal.jsp.

Iturri batzuetako dokumentuak PDF eta EPUB formatuetan eskuratu dira (zientzia argitalpenak, testu administratiboak eta literatura lanak batik bat), eta hauetatik testu hutsa erauzteko *PyMuPDF*[11] eta *epub2txt*[12] tresnak erabili dira. Behin testu hutsa lortuta, lerro jauziak konpontzeko *dehyphen*[13] tresna erabili da. Iturrietako batzuetan (ADDI, Eusko Jaurlaritza, GFA eta BFA) hizkuntza filtroa aplikatu da *fasttext-langdetect*[14] tresnarekin, erdarazko dokumentuak baztertu eta dokumentu elebidunetatik euskarazko testua bereizteko. Luzera oso laburra zuten dokumentuak ere baztertu egin dira corpusetik. Bukatzeko testuko kodeketa arazoak konpontzeko *ftfy*[15] tresna erabili dugu.

[11] `https://github.com/pymupdf/PyMuPDF`
[12] `https://github.com/ffreemt/epub2txt`
[13] `https://github.com/pd3f/dehyphen`
[14] `https://github.com/zafercavdar/fasttext-langdetect`
[15] `https://github.com/rspeer/python-ftfy`

mila (K) *thousand*
miloi (M) *million*
mila milioi (B) *billion*
bilioi (T) *trillion*

alborapen *bias*
arrazoiketa *reasoning*
asmoen sailkapen *intent classification*
atalase *treshold*
atentzio *attention*
atentzio-buru *attention head*
aurrentrenamendu *pretraining*
aurrerantz elikatutako neurona-sarea *feedforward neural network*
ausaz *randomly*
azpilaginketa *downsample*
baliabide-urriko *low-resource*
baliabide-gutxiagoko *less-resouce*
baliabide-oparoko *high-resource*
beroketa *warm-up*
berretura-lege *power law*
birdoitu *finetune*
datu-multzo *dataset*
deskorapilatutako atentzio *disentangled attention*
distrakzio ataza *distractor task*
domeinuko *in-domain*
domeinuz kanpoko *out-of-domain*
ebaluazio multzo *test set*

entitate izendunen ezagutza *name entity recognition (NER)*

entrenamendu/garapen/ebaluazio banaketa
train/development/test split

epoka *epoch*

eredu autorregresibo *autoreresive model*

eredu diskriminatzaile *discriminative model*

erregelatan oinarritutako *rule-based*

eskalatze-joera *scaling trends*

eskala-lege *scaling law*

estaldura *coverage*

exekuzio *run*

ezagutza faktuala *factual knowledge*

ezkutuko geruza *hidden-layer*

ezkutuko dimentsio *hidden-dimention*

F puntuazio *F-score (F1)*

gaien sailkapena *topic classification*

gainegokitzapen *overfitting*

gainlaginketa *oversampling*

gaitasun linguistiko formala *formal linguistic competence*

galdera-erantzunen hizkuntza inferentzia
question-answering NLI (QNLI)

galera *loss*

gelditze goiztiarra *early-stopping*

geruza lineal *linear layer*

ausnarketa-kate *chain-of-thought (CoT)*

helburu-ataza *downstream-task*

helburu-funtzio *objective function*

helburu-hizkuntza *target language*

hiperparametro *hyperparameter*

hitza testuinguruan *word in context (WiC)*
hitz-bektore *word embedding*
hitz-bektore ez besteko parametroak *non-embedding parameters*
hitz ordena malgua *flexible word order*
hitz osoen maskaratze *whole-word masking*
hizkuntza bakartu *language isolate*
hizkuntza-eredu *language models (LM)*
hizkuntza-eredu neuronal *neural language models (NLM)*
hizkuntza-eredu handi *large language models (LLM)*
hizkuntza inferentzia *natural language inference (NLI)*
hizkuntzaren ulermena *natural language understanding (NLU)*
hizkuntzaren sorkuntza *natural language generation (NLG)*
hizpeko token *subword token*
hiztegi *vocabulary*
hiztegi partekatu *joint vocabulary*
hiztegitik at *out of vocabulary*
hutsuneak betetze *slot filling*
ikasketa-tasa *learning rate*
iragarri *predict*
iragarpen *prediction*
instrukzio *instruct*
itzulpenera *translationese*
jarrera detekzio *stance detection*
kalkulu-ahalmen *computation*
kateatu *concatenate*
kodetzaile-deskodetzaile *encoder-decoder*
kontrol-puntu *checkpoint*
labekada tamaina *batch size*
lagin-eraginkorra *sample-efficient*

lehio *span*
lerrokatze *alignment*
letra larri eta xeheak bereiziz *cased*
neurona-sare *neural network*
neurona-sare errepikakor *recurrent neural network (RNN)*
neurona-sare trinko *fully connected neural network*
maskaratze *masking*
maskaratutako hizkuntza-eredu *masked language model (MLM)*
maskaratutako hizkuntza-modelatze *masked language modeling (MLM)*
muturreko datu *outlier*
kategoria anitzeko *multi-class*
oinarri-lerro *baseline*
pare-minimo *minimal pair*
patroi *pattern*
pendiz-metaketa *gradient-descent*
perplexitate *perplexity*
pisuen gainbehera *weight decay*
proba-banku *benchmark*
sasilog-egiantz *pseudo-log-likelihood*
sekuentzia luzera *sequence length*
sekuentzia-pare sailkapen bitar *sequence-pair binary classification*
sen oneko arrazoiketa *commonsense reasoning*
sentimenduen analisia *sentiment analysis*
testuingurudun hitz-bektore *contextual word embeddings*
tokenizatzaile *tokenizer*
transferentzia-ikaskuntza *transfer-learning*
urrats *step*
uztartze estrategia *merging strategy*
zehaztasun *accuracy*