e-HIZPIDE
HELDUEN EUSKALDUNTZEAREN ALDIZKARIA

Ikerketak

101. zenb.
XLI. URTEA
2023 URTARRILA-EKAINA

ISSN 2530-3287

HABE-IXA euskarazko idazmen-proben corpuseko idazlanen mailakatze automatikoa

Ekain Arrieta

HiTZ Ixa (UPV/EHU. Lengoaiak eta Sistema Informatikoak)

ekain.arrieta@ehu.eus

Igor Odriozola

HABE (Eusko Jaurlaritza, Kultura eta Hizkuntza Politika Saila)

Xabier Arregi

HiTZ Ixa (UPV/EHU. Lengoaiak eta Sistema Informatikoak)

Mikel Iruskieta

HiTZ Ixa (UPV/EHU. Lengoaiak eta Sistema Informatikoak)

Gero eta euskarazko testu gehiago idazten da ordenagailuz eta hainbat erabileratarako interesgarria litzateke Helduen Euskalduntzearen Oinarrizko Curriculumeko (HEOC) komunikagaitasun-mailetan oinarrituta testuok automatikoki mailakatzea. Artikulu honetan azalduko den lanaren helburua honako hau da: HABE-IXA euskarazko idazmen-proben corpusa aurkeztea eta, Europako Erreferentzia Marko Bateratuko (EEMB) B1, B2, C1 eta C2 mailen arabera sailkatzeko tresna automatikoekin lortutako emaitzak azaltzea. HABE-IXA corpusa HABE erakundeak egiaztatzegintza-prozesuetan jasotako 480 idazlanez eta horien ebaluazioez osaturik dago. Testu-sailkapenean, Ixa taldeak (UPV/EHU) hizkuntza-prozesamendurako sorturiko analisi-tresnak eta ikasketa automatikoko teknikak erabiliz, zenbait sailkapen-ataza garatu dira eta emaitzarik onena (% 97ko zehaztasuna) idazlanei dagokien maila esleitzean lortu da. Etorkizuneko asmoa da corpus handiagoak osatu eta euskararen irakaskuntzarako baliagarriak izan daitezkeen sailkapen-tresnak garatzea. Corpusa eskura dago CC BY-NC 4.0 lizentziapean.

Gako-hitzak
corpusa, HEOC, EEMB, idazlanen kalifikazio automatikoa, idazmen-probak, ikasketa automatikoa, sailkatzaileak

1. SARRERA

Gero eta aukera gehiago dago euskaltegietan, hizkuntza-eskoletan eta hezkuntza formalean sortzen diren hizkuntza-probetan testuak formatu digitalean eskuratzeko eta datuok ikerketan erabiltzeko. Ikasleak ohituta daude Ikaskuntza Kudeatzeko Sistemetan idazten (Camacho eta Iruskieta, 2021), eta paperean idazten diren testuak digitalizatzea ere gero eta errazagoa da (Ibarra eta Iruskieta, 2022). Bestalde, ikasleentzako autoikaskuntza-aukera gehiago dago (Unibaso, 2004), eta ohituago daude informazioaren eta komunikazioaren teknologiak (IKTak) erabiltzen. Gainera, datuak publiko egiteko eta corpusak ugaritzeko eta handiagotzeko politikak gero eta nabarmenagoak dira (ikus, adibidez, 2016an Euskararen Aholku Batzordearen IKT Batzorde-Atalak egindako *Euskarazko IKTak: gomendioak herri-aginteentzat* dokumentua). Argi dago ezen, euskarak aro digitalean biziraungo badu, behar-beharrezkoa dela euskarazko corpusak ugaritzea eta handiagotzea, eta euskara aztertzeko baliabideak sortzea, baita helduen hizkuntza i(ra)kaskuntzan lagunduko dutenak ere.

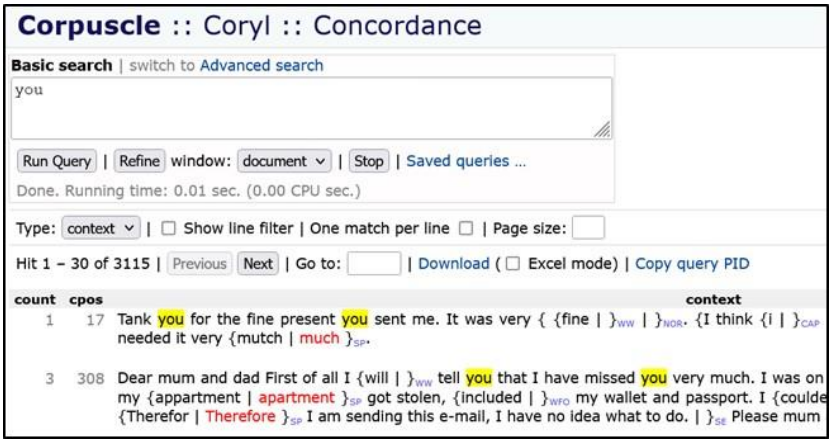
Testuak sortzeko eta gordetzeko gaitasunean aurrerapauso handiak eman dira; halaber, egin dute aurrera testu horiek aztertzeko baliabideek ere aurrera egin dute. Beraz, oso testu-bilduma handiak prozesatzeko eta analizatzeko aukera berriak daude. Hala, euskara-ikasleek ekoitzitako testuen maila neurtzeko probak egin eta zuzendu litezke hizkuntza-prozesamendurako tresnak erabiliz (Correnti et al., 2019; Maamuujav et al., 2021).

Horren erabilera praktiko bat da euskara ikasten ari direnen komunikagaitasun-maila zehaztea. Gure ustez, oso baliagarria izan daiteke ikasleek darabilten hizkuntza aztertzea, haien komunikagaitasun-mailaren ikuspegi globala emango duten heinean, bai ikasleentzat beraientzat, bai irakasleentzat, bai euskaltegientzat edota erakundeentzat oro har. Ikasleentzat, zer maila duten jakiteko; irakasleentzat, ikertzeko edo ebaluazioan laguntzeko; erakundeentzat, berriz, beste ebaluazio neurri bat gehiago edukitzeko. Testuak batuta ikasleen corpus bat —eta horiek aztertzeko tresnak— edukitzea oso interesgarria litzateke pedagogian eta hizkuntzaren ikerketan; izan ere, tresna eta metodo ezberdinak erabiliz, ikasleen hizkuntza-ikasprozesua deskribatu ahal izango litzateke, hainbat mailatan: morfologikoan, lexikoan, sintaktikoan eta diskurtsiboan.

Ikerketari dagokionez, hala sortutako corpusek daukaten bereizgarri nagusia ikasleen tarteko hizkuntza da; alegia, ikasleek, hizkuntza-ikasprozesuan zehar, zer akats egiten duten eta zer hizkuntza-forma erabiltzen dituzten (Osinalde eta Iruskieta, 2022), beren H1aren (lehen hizkuntzaren) eraginez. Halako hainbat corpus daude CLARIN (https://www.clarin.eu/) Europako ikerketarako azpiegitura digitalean eskuragarri, baita corpusetan bilaketak egiteko analizatzaileak ere. Adibidez, CLARINek 74 ikasle-corpus eskaintzen ditu, horietatik 11 eleanitzak dira, eta hainbat formatutan eskaintzen dira: idatzizkoak, ahozkoak eta multimodalak. Guztira, 13 hizkuntza daude: arabiera, txekiera, ingelesa, finlandiera, frantsesa, alemana, hungariera, islandiera, italiera, txinera, norvegiera, gaztelania eta suediera. Gehienek lizentzia publikoa daukate, eta, aipatu dugunez, ikasleen tarteko hizkuntza aztertzeko diseinatu dira. Corpusak erabiltzaileentzat baliagarriak izan daitezen, ezaugarri batzuk eskuz txertatzen dira; beste batzuk, ordea, prozesu automatikoen bidez, hizkuntza-teknologiak erabiliz. Horri esker, corpus horietan, testu hutsa ez ezik, akatsak ere nabarmentzen dira, edo informazio morfosintaktikoa eta forma zein lema erauzketaren informazioa erakusten da (https://www.clarin.eu/resource-families/L2-corpora).

CLARIN azpiegiturako corpus interesgarri bat *Coryl* corpusa da (Hasselgreen eta Sundet, 2017), ikasle norvegiar gazteen ingelesezko testuz osaturikoa. 2004-2005 ikasturtean jaso ziren idazlanak, estatu mailako ingelesezko idazketa-proba batean, Bergeneko unibertsitatean, eta hainbat mailatako ikasleen testuak dauzka. Behin testu horiek anonimatuta, hainbat aztertzailek EEMBren araberako komunikagaitasun-maila esleitu zieten. Ondoren, testuen erroreak eskuz etiketatu ziren, eta informazio morfosintaktikoa automatikoki etiketatu zen, corpuseko testuetan erroreak (ikus 1. irudia) eta *Key Word in Context* (KWIC) estiloko informazioa bilatu eta azpimarratu ahal izateko (ikus 2. irudia).

1. irudia. Coryl corpusaren eskuz etiketatutako zatian, "you" hitzaren bilaketaren emaitza: "you" guztiak horiz nabarmenduta agertzen dira; ikasleen erroreak, berriz, giltzen artean, gorriz zuzenduta. Iturria: Coryl corpusa.



***2. irudia**: Coryl corpusaren automatikoki etiketatutako zatian, "you" hitzaren bilaketaren emaitza eta hitzaren testuinguru laburra, KWIC gisa erakutsita. Iturria: Coryl corpusa.*

CLARIN azpiegituran ez dago mota horretako euskarazko corpusik; hala ere, badaude bestelakoak eta aipagarriak diren euskarari buruzko lanak. Adibidez, ikuspegi konputazionaletik, Aldabek et al.ek (2005) euskarazko errore-motak eta desbideratzeak aztertu zituzten, eta Larreak (2009) errore horiek zuzentzeko proposamen didaktikoa egin zuen. Nabarmentzekoa da Pérezek (2014) proposamen hori HABEren 2. mailako (gaur egungo B2ko) azterketetan erabili izana eta Belokik et al.ek (2020) errore gramatikalak zuzentzeko corpusa sortu izana, erroreak modu automatikoan sortuz eta ikasketa sakoneko teknikak erabiliz, hizkuntza erroreduna zuzentzeko. Horrez gain, Osinaldek eta Iruskietak (2022) B2 eta C1 mailako azterketetan ekoitzitako testuak aztertu dituzte erroreak *Markin* tresnarekin etiketatuz eta corpusean maila horietako akats nabarmenak zein diren erakutsiz.

Gure lana arestian aipatutako lan horietatik hurbil badago ere, bestelako bi helburu ditugu: batetik, HABE-IXA euskarazko idazmen-proben corpusa aurkeztea; bestetik, corpus hori oinarri hartuta, testuak EEMB/HEOCeko mailetan sailkatzeko egindako proben emaitzak aurkeztea. Lan honetan HABEk komunikagaitasun-mailak egiaztatzeko probetan batutako idazlanen artetik zenbait hautatu dira, haiei dagozkien metadatuekin batera, eta corpus hori, ondoren, prozesatu egin da, Ixa Taldean sorturiko hizkuntza-prozesamendurako tresnak baliatuz. Gainera, ikasketa automatikoa erabiliz, aztertu dugu

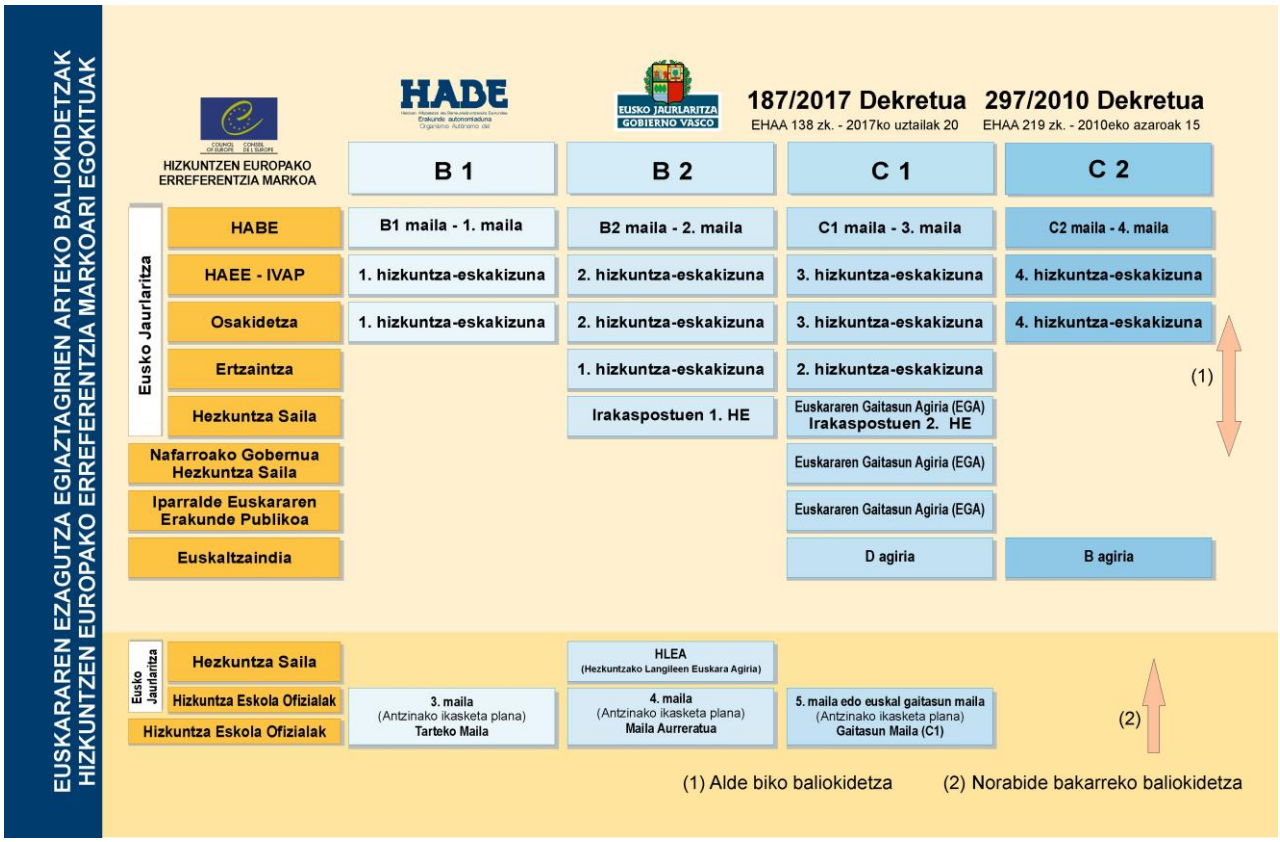
ea testu bat maila jakin baterako egokia den, eta sailkatu ere egin dugu, koherentzian, kohesioan, zuzentasunean eta aberastasunean duten puntuazioaren arabera.

Gure ustez, lan honek hizkuntzen irakaskuntzan dauden bi beharri erantzuten die:

1. Testuen kalitatea zein den jakiteko beharra oso garrantzitsua da euskal jendartean, helduen alfabetatze- eta berreuskalduntze-prozesuan, Eusko Jaurlaritza lantzen ari den Hezkuntza Legean eta baita hainbat lanpostutan. Lanpostuetan, oro har, erabiltzaile aurreratuak (B2) edota erabiltzaile gaituak (C1) eskatzen dira (ikus 3. irudia), baita, zenbait kasutan, erabiltzaile adituak (C2) ere. Esaterako, testu idatzi gehien sortzen diren esparruan, euskal hezkuntza-sisteman alegia, behar hori badelakoan gaude; izan ere, ikasleek, Lehen Hezkuntzaren amaieran, EEMB/HEOCeko euskarazko B1 maila lortu behar dute. Derrigorrezko Bigarren Hezkuntzan, berriz, euskarazko B2 maila lortu behar dute (ikus 1. taula).

2. Corpusak eta tresnak irakasleei eta ikertzaileei eskaintzeko beharra. Erakundeek, ikertzaileek eta irakasleek badute beharra jakiteko nolakoak diren, bai tarteko hizkuntza, bai ikasleek idatzitako testuen kalitatea, bai testu horien ezaugarriak. Tresna automatikoak erabiliz, testuak adierazle desberdinen arabera sailkatu, intereseko ezaugarrien arabera multzokatu, edo estandarretik urruntzen diren testuak identifika daitezke. Ikasleek ere balia ditzakete halako tresnak sortze-prozesuetan haien lanak maila jakin baterako egokiak diren aztertzeko edo zer alderdi hobetu ditzaketen lantzeko.

3. irudia. Egiaztagirien arteko baliokidetzak, EEMBari egokituak. Iturria:HABEren webgunea (https://www.habe.euskadi.eus/helduen-euskalduntzearen-oinarrizko-curriculuma-heoc/webhabe00-edukiak/eu/).



1. taula. EAEn hezkuntza-sistemako ikasleek lortu beharreko mailak, ikasketa-etapa bakoitzaren amaieran. Berezko ekoizpena. Iturria: EAEko hezkuntza-sistema hobetzeko plana. Eusko Jaurlaritza (2016).

Ikasketa-etapak	Lortu beharreko euskara maila
Lehen Hezkuntzan (LH6 bukatzean)	B1
DBH eta Batxilerra bukatzean	B2
Unibertsitateko lizentziatura-ikasketak bukatzean	C1 (144 ECTS euskaraz edo kreditu gutxiago zenbait baldintza betez gero)
Unibertsitateko tesia edo Euskal Filologia gradua bukatzean	C2 (doktorego-tesia euskaraz idatzi eta defendatzen bada)

Lan honen mugei dagokienez, ikusi dugu esku artean dugun corpusa handiagotuz joan ahala, garatutako tresnen emaitzak ere hobetuz doazela. Nahiko argi dago ezen, corpusa handiagotuz eta testu gehiago lortuz gero, tresna zehatzagoak lortuko direla. Bestalde, hizkuntza-ezaugarri sakonagoak aztertzeko eta erauzteko metodoak aztertzea ere komeni da, emaitzak hobetzeko ez ezik, maila bakoitzaren ezaugarriak ulertzeko ere, eta, orobat, koherentziak, kohesioak, aberastasunak eta zuzentasunak ezaugarri linguistikoekin korrelaziorik baduten aztertzeko ere. Bukatzeko, komenigarria litzateke testu-motak, generoak eta arloak ere kontuan izatea, iragarpen eta tresna egokiak sortzeko.

2. METODOLOGIA

Atal honetan, *HABE-IXA euskarazko idazmen-proben corpusa* eta berau erabilgarri izateko egin diren zenbait aurreprozesamendu aurkeztuko dugu lehendabizi. Ondoren, informazio linguistikoa erauzteko zenbait tresna aurkeztuko dira, baita lan horretan erabili diren ezaugarri linguistikoak ere. Azkenik, egindako sailkapen-esperimentuak azalduko dira.

2.1. HABE-IXA EUSKARAZKO IDAZMEN-PROBEN CORPUSA

HABE-IXA corpusa HABEren B1, B2, C1 eta C2 komunikagaitasun-mailak egiaztatzeko probetan jasotzen diren idazlanen lagin batekin eta azterketari buruzko hainbat metadaturekin osatu da. Idazmen-proban, azterketariei gai bat eta egiteko bat aurkezten zaie, eta, horri erantzunez, denbora-muga batean burutu behar izaten dute idazlana. Corpusaren deskribapen orokorra 2. taulan ageri da.

2. taula. HABE-IXA euskarazko idazmen-proben corpusaren ezaugarriak. Iturria: berezko ekoizpena.

HABE-IXA euskarazko idazmen-proben corpusa	Hizkuntza	Deskribapena
Tamaina: 480 testu, 146465 hitz	Euskara	HABEren azterketa ofizialetako EEMB/HEOCeko B1, B2, C1 eta C2 mailetako idazlanekin osaturik dago. Maila jakin bateko probara aurkeztutako ekoizpen idatziak daude, gaindituak nahiz ez-gaindituak, eta 5 ebaluazio-irizpidetan eskuz jarritako kalifikazioa daukate: i) egokitasuna, ii) koherentzia, iii) kohesioa, iv) aberastasuna eta v) zuzentasuna. Corpusa orekatua da, maila bakoitzean testu kopuru bera baitago.
Hemen eskuragarri: https://doi.org/10.23728/b2share.81433fddcd06405f8505c7606b29ff99 CC BY-NC 4.0 lizentziapean		

Corpuseko idazlan bakoitza gutxienez bina aztertzailek kalifikatu dute, itsu bikoitzeko sistemaren bidez. Desadostasuna egonez gero, tutore batek erabaki du azken kalifikazioa, bi aztertzaileen proposamenak kontuan izanda. Idazlana honako bost ebaluazio-irizpide hauen arabera kalifikatu da: i) egokitasuna;[1] ii) koherentzia; iii) kohesioa; iv) aberastasuna; eta v) zuzentasuna. Horietako irizpide bakoitzak puntuazio bat jaso du (A, B, C, D, E) eta, horien balioak haztatuz, nota bat kalkulatu da, idazlanak proba gainditu duen ("gai") ala ez ("ez gai") erabakitzeko erabili dena.

Ebaluazio-irizpide bakoitzaren balio haztatua edo pisua komunikagaitasun-mailaren araberakoa da, 3. taulan ikus daitekeenez. Horrez gain, puntuazio bakoitzak (Atik Era) irizpide bakoitzarentzat duen pisua ere aldatu egiten da maila batetik bestera, maila horretan eskatzen denari egokitzeko.

3. taula. Kalifikazio-irizpide bakoitzaren balio haztatuak HABEren kalifikazio-sisteman. Iturria: berezko ekoizpena.

	B1	B2	C1	C2
Egokitasuna	% 10	% 13	% 13	% 20
Koherentzia	% 25	% 24	% 24	% 20
Kohesioa	% 15	% 17	% 17	% 20
Aberastasuna	% 20	% 20	% 20	% 20
Zuzentasuna	% 30	% 26	% 26	% 20

[1] *Egokitasuna* ebaluazio-irizpideak neurtzen du ea azterketariak modu egokian erantzun dion proban eskatu zaizkion egoerari eta egitekoari, eta, beraz, idazlanari berari buruzko informazio gutxi ematen du. Testuinguru horren faltan, egokitasunari buruzko datuak corpusean kontuan ez hartzea erabaki dugu.

Ebaluazio-irizpide guztiak kontuan hartuta kalkulatzen den emaitzak 0 eta 30 arteko balioa du. Idazmen-proban 15 puntu baino gutxiago lortzen dituen idazlanaren kalifikazioa "ez-gai" izango da; 15 puntu edo gehiago lortuz gero, berriz, "gai". Maila jakin bateko egiaztagiria eskuratu ahal izateko, azterketariak, oro har, trebetasun guztietako irizpideak gainditu behar ditu[2].

Corpusean, testu bakoitzak metadatu hauek ditu: i) probari dagokion maila, ii) idazlanari maila horretarako eman zaion kalifikazio orokorra ("gai" / "ez-gai") eta iii) ebaluazio-irizpideen balioak. Corpuseko hainbat azterketa mugakoak direnez, alegia, "gai" izateko kalifikazio minimoa dutenez, bereizketa gehigarria egin dugu "Gai nahikoa" eta "Gai ondo" azterketen artean, puntuazio zehatzaren arabera. 4. taulan, maila bakoitzeko testu- eta *token*-kopuruak (hitzak eta puntuazio-ikurrak) ageri dira.

4. taula. Idazmen-proba gainditu duten eta gainditu ez duten testuen kopuruak eta token kopuruak HABE-IXA euskarazko idazmen-proben corpusean, maila bakoitzeko. Iturria: berezko ekoizpena.

Maila	Testuak	"GAI nahikoa" testuak	"GAI ondo" testuak	"EZ GAI" testuak	Tokenak orotara
B1	120	40	40	20	21570
B2	120	40	40	20	28319
C1	120	40	40	20	40305
C2	120	30	17	73	56271
Orotara	480	150	137	133	**146465**

2.2. AURREPROZESAKETA

HABE-IXA euskarazko idazmen-proben corpusa osatzeko, honako urratsak eman ditugu:

1. **Testuak digitalizatzea**. HABEren egiaztatzegintzako idazmen-probetako PDF fitxategietan gordetako irudi multzoak testu digital bihurtu dira, eta bai testua, bai kalifikazioak, datu egituratu gisa gorde ditugu.
2. **Karaktere-kodeketa**. ISO-8859-1 kodeketara pasatu ditugu idazlan guztiak, eta testu fitxategi sinple formatuan gorde.
3. **Garbiketa**. Transkripzioan eta erauzketan gerta zitezkeen erroreak eskuz identifikatu eta konpondu ditugu. Gainera, idazlanetan agertzen diren datu pertsonalak ezabatu edo aldatu ditugu, testuen anonimotasuna ziurtatzeko.
4. **Datu-basea.** Datu-basea sortu da testu garbiekin, kalifikazioekin eta azterketen metadatuekin.

2.3. INFORMAZIO LINGUISTIKOA ERAUZTEKO TRESNAK

Testu idatzietatik ezaugarri linguistikoak automatikoki erauzteko eta testu horiei informazio linguistikoa edo zenbakizko balioak esleitzeko euskararako hizkuntza-prozesamendurako tresnak baliatu ditugu. Guztira lau kategoria hauetako 63 ezaugarri[3] hautatu ditugu.

[2] Mugako kalifikazioak konpentsatzeko aukera egon daiteke.
[3] Ezaugarrien zerrenda osoa material osagarrian ikus daiteke.

1. **Hizkuntza konplexutasun ezaugarriak**. Hizkuntzaren jabekuntza neurtzeko erabiltzen diren zenbait ezaugarri linguistiko eta paragrafoak, esaldiak eta karaktereak zenbatzeko, Pythonen NLTK liburutegia (Bird, 2009) eta Ixa Taldean garaturiko euskarazko CTAP tresnaren bertsioa (Chen eta Meurers, 2016) erabili ditugu. Adibidez, hitz, esaldi eta testuen luzera; token-, silaba- eta karaktere-kopuruen arabera eta horiekin lotutako irakurgarritasun-formulak.
2. **Maila lexikaleko informazio linguistikoa**. Informazio hori gehitzeko *ANALHITZA* (Otegi et al., 2017) eta *IxaKat* (Otegi, 2016) hizkuntza-prozesatzaileen kate modularrak erabili dira: a) Testuak hitzetan tokenizatzeko ("tokenizazioa"), b) token bakoitzetik lema erauzteko (lematizazioa), eta c) lema bakoitzari kategoria gramatikala esleitzeko. Esaterako, testuen aniztasun lexikala, eduki lexikala duten hitzen kopurua, eta aberastasun- eta maiztasun-neurriak.
3. **Maila morfosintaktikoko informazio linguistikoa**. Dependentzia sintaktikoak analizatzeko, *UDPipe 2* (Straka, 2018) erabili dugu. Esaterako, ezaugarri morfosintaktikoen artean, aditzen modu-denbora, kasu gramatikalen banaketa eta aditz-formak eta ezaugarri sintaktikoen artean: sintagma-kopuruak eta zuhaitz sintaktikoen ezaugarriak.
4. **Ortografia-erroreak**. *Xuxen* zuzentzaile ortografikoa (Agirre et al., 1992) pasatu zaie testuei, akats ortografikoak etiketatu eta zenbatzeko.

2.4. EZAUGARRI-BEKTOREAK

Lan honetan erabili diren ikasketa automatikoko sailkatzaileek zenbakizko balioak interpreta ditzakete. Hortaz, testu arruntak zenbakizko bektoreen bitartez adierazi ditugu. Gure kasuan, automatikoki erauzitako ezaugarri linguistiko denak bektore batean bildu eta bektore horiek idazlanen zenbakizko adierazpen gisa erabiliko ditugu.

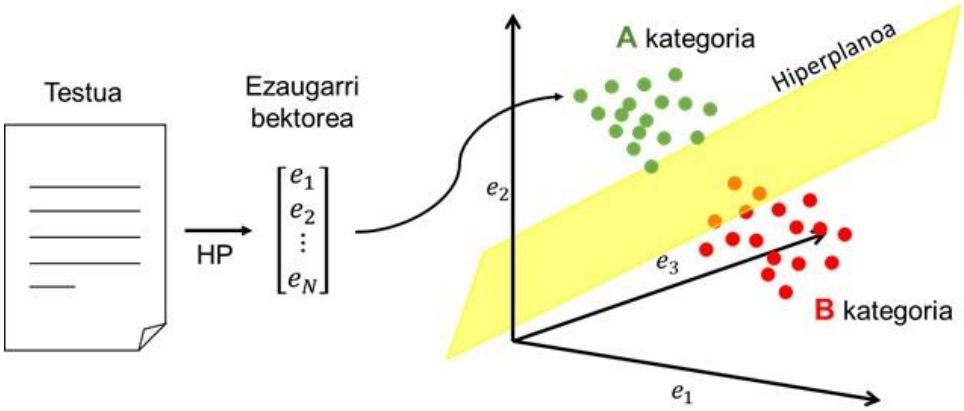
Badakigu hizkuntza-ezaugarrien kopuru mugatu horrekin idazlana ez dela bere osotasunean errepresentatzen eta, beraz, etorkizunean, ezaugarrien zerrenda handitzen eta hobetzen jarraitzeko asmoa dugu. Izan ere, hizkuntza-maila hautemateko ezaugarri esanguratsuak antzematea da HABE eta Ixa Taldearen arteko lankidetzaren helburuetako bat.

Ezaugarri horiek guztien balioak proportzio moduan edo testuaren luzeraren arabera normalizatuta erabili dira, eskala edo magnitude handiagoa duten ezaugarrien aldeko alborapenak saihesteko. Horrela, bada, estandarizatu egin dira neurri guztiak, batezbesteko aritmetikoa kenduz eta unitateko bariantzara eskalatuz. Estandarizatu ondoren, neurri denek batezbesteko ($\mu = 0$) eta desbideratze estandar ($\sigma = 1$) bera dute.

2.5. IKASKETA AUTOMATIKORAKO EREDUA

Lan honetan, Euskarri Bektoredun Makina (EBM; ingelesez, *Support Vector Machine* edo SVM) ereduak erabili ditugu. EBMak ikasketa gainbegiratua erabiltzen duten ereduak dira: algoritmoaren entrenamendu fasean, kategoria jakin bateko etiketa duten datuak emanda (lortu nahi dugun emaitza), kategoria horiek banatuko dituen hiperplanoa sortzen du. Ostean, etiketa gabeko datuak jasotzean, datu horiek kategorietako batean sailkatuko ditu ezaugarri-bektoreak erabiliz. Lan honetan, datuak 63 dimentsioko ezaugarri espazioan adieraz daitezkeen ezaugarri-bektoreak dira eta kategoriak idazlanaren maila edota kalifikazioa izan dira. EBMak maiz erabiltzen dira testuen sailkapena egiteko (Schwarm 2005), eta beste sailkapen-metodo batzuk baino emaitza hobeak ematen dituzte, bereziki datu gutxi dagoen kasuetan.

4. irudia. Ezaugarri linguistikoen (e) erauzketa, testuen bektorizazioa eta EBM algoritmoak sortzen duen hiperplano sailkatzailea (horiz). Iturria: berezko ekoizpena.



Ikasketa automatikoko ereduak *Python*en *Sci-Kit Learn* liburutegia (Pedregosa et al., 2011) erabiliz entrenatu eta ebaluatu ditugu. EBM sailkatzailerako, *kernel* funtzio lineala eta *C=1* erregularizazio-parametroa erabili ditugu.

Ezaugarri linguistiko denak ez dira baliagarriak sailkapen-atazak entrenatzeko unean (Bahassine *et al.*, 2018). Atazarako esanguratsuak diren ezaugarriek eragin handiagoa izango dute hiperplanoaren kokapenean esanguratsuak ez direnek baino. Eragin hori ezaugarri bakoitzak hiperplanoa sortzeko duen pisua edo koefizientea erabiliz neurtu dezakegu. Sailkapenean esanguratsuak ez diren ezaugarriak baztertzeko, aukeraketa bat egin dugu *Recursive Feature Elimination* (RFE) algoritmoa erabiliz. Prozesu hori honela deskribatzen da:

1. Sailkatzailea entrenatu egiten da, ezaugarri guztiak erabiliz.
2. Ezaugarri linguistiko bakoitzak EBMren hiperplanoa sortzeko izan duen garrantzia kalkulatzen da koefizientea erabiliz.
3. Sailkapenean garrantzia gutxien duen ezaugarri linguistikoa baztertzen da.
4. Prozesu hori errepikatu egiten da, sailkapenean esanguratsuenak diren *N* ezaugarri linguistiko lortu arte.

Orain artekoa kontuan izanda, lan honetan azaltzen ditugun esperimentuetan bi sailkatzaileren emaitzak aurkeztuko ditugu: i) EBM, ezaugarri linguistiko guztiak erabiliz eta ii) EBM, 10 ezaugarri linguistiko esanguratsuenak erabiliz (N=10).

2.6. SAILKAPEN-ATAZAK

Lan honetan, lau sailkapen-ataza zehaztu ditugu, eta sailkatzaileak ataza bakoitzaren helburuetara doitu dira:

- **1. sailkapen-ataza.** Idazlan batek komunikagaitasun-maila jakin bat gaindituko lukeen iragartzea. Sailkapen-ataza horretarako, sailkatzaile bitar bat entrenatu dugu maila bakoitzean: ea testu bat "gai" edo "ez gai" den.
- **2. sailkapen-ataza.** Idazlan bat EEMB/HEOCeko mailen arabera sailkatzea. Sailkapen-ataza honetan, testu bati dagokion komunikagaitasun-maila iragarriko da. Entrenamendurako, maila guztietan "gai" kalifikazioa lortu duten testuak soilik erabili dira, maila jakin bat gainditu ez duten testuak ezin baitira maila horretakotzat hartu. Horretarako, klase anitzeko sailkatzaile bat entrenatu da.
- **3. sailkapen-ataza.** Idazlan bat *B1-B2* eta *C1-C2* multzoetan sailkatzea. Sailkapen-ataza honetan, B2 eta C1 mailen arteko banaketa soilik hartu da kontuan,

HABEko azterketa gehienak B2 eta C1 mailakoak baitira. Sailkapen-ataza honetan ere, "gai" kalifikazioa lortu duten idazlanak erabilita, "B2 edo beheragoko mailen" edota "C1 edo goragoko mailen" artean sailkatuko duen sistema bitarra entrenatu dugu.

- **4. sailkapen-ataza**. Idazlanak HABEko ebaluazio-irizpideen arabera sailkatzea. Sailkapen-ataza horretan, HABEko bi zuzentzaileren arteko kalifikazio-irizpideen balioak (koherentzia, kohesioa, aberastasuna eta zuzentasuna) sailkatzeko entrenatu dugu eredua (ikus 2. taula). Berez, irizpide bakoitzak 5 puntuazio posible baditu ere, 3 puntuaziora murriztu dugu, (*ona, nahikoa, ez-nahikoa*), sailkapen-ataza sinplifikatzearren.

3. EMAITZAK

Atal honetan, aipatutako 4 sailkapen-atazetako emaitzak aurkeztuko ditugu. Sailkapen-ataza guztietan –2. sailkapen-atazeko kasu batean izan ezik, 3.2 atalean azalduta dagoen moduan– emaitzak lortzeko 5 iteraziodun balidazio gurutzatua erabili dugu. Sailkatzaileen emaitza guztiak *zehaztasun* gisa erakutsiko dira.

3.1. IDAZLAN BATEK KOMUNIKAGAITASUN-MAILA JAKIN BAT GAINDITUKO LUKEEN IRAGARTZEA

Sailkapen-ataza honetan, maila bakoitzeko bi sailkatzaile entrenatu dira, bat ezaugarri guztiekin eta bestea RFEren bidez lortutako 10 ezaugarri esanguratsuenekin. Emaitzarik egokienak C1 mailan lortu dira.

5. taula. Testuen GAI / EZ GAI sailkapena (1. sailkapen-ataza), maila jakin bakoitzerako. Emaitza guztiak zehaztasun gisa adierazirik daude. Iturria: berezko ekoizpena.

Sailkapen bitarra (GAI / EZ GAI)	B1	B2	C1	C2
EBM, ezaugarri denak	0,64 (±0,05)	0,71 (±0,08)	0,84 (±0,07)	0,71 (±0,11)
EBM, 10 ezaugarri	0,66 (±0,04)	0,75 (±0,08)	0,92 (±0,05)	0,79 (±0,08)

EBMak ezaugarri bakoitzari ematen dion pisua ezaugarriaren garrantziaren adierazle modura erabili izan dira (Guyon et al., 2002), eta testu berri bat kategoria batean edo bestean sailkatzeko ezaugarri linguistikoek zenbateko pisua duten adierazten du. Sailkapen-ataza honetan iragarpena egiteko garrantzia handiena duten ezaugarriak honakoak dira: i) lexikoaren aberastasunaren markatzaileak; ii) adjektiboen dentsitatea; iii) subjuntiboaren erabilera; eta iv) postposizioen eta kasu-marken erabilera (zehazki, *instrumentala, adlatiboa* eta *ergatiboa*).

Ezaugarri horiek automatikoki idazlan bat maila jakin baterako "gai" ala "ez gai" izango den erabakitzeko garrantzitsuak diren arren, kontuan izan behar dugu maila bakoitzeko 120 testu besterik ez daudela corpusean eta, hain multzo txikiarekin, alborapenaren aukerak handitu egin daitezkeela. Emaitza hobeak lortzeko eta ezaugarrien interpretazioa fidagarriagoa izateko, beharrezkoa litzateke corpusa handitzea.

3.2. IDAZLAN BAT EEMB/HEOC-EKO MAILEN ARABERA SAILKATZEA

Sailkapen-ataza honetan, entrenamenduko testu bakoitzak EEMB/HEOCeko maila bat izango du esleituta. Idazlan batek aurkeztu den mailako proba gainditzen ez badu, ezin da maila horretako testutzat hartu . Hori horrela, "gai" kalifikazioa duten testuak bakarrik erabili dira sailkatzailea entrenatzeko. Metodo honen gabezia da "ez gai" testuak

entrenamendurako erabili ez direnez, sailkatzaileak "ez gai" testuen bestelako ezaugarriak ez dituela "ikusi". Ataza honetan, "ez gai" testuei maila baxuago bat iragarri bazaie ontzat hartu dira eta maila bera edo altuagoa iragarri bazaie, berriz, okertzat.

Entrenatutako sailkatzailearen portaera interpretatzeko, bi ebaluazio-metodo erabili ditugu: alde batetik, entrenatzeko erabili ez diren "gai" idazlanen kontra ebaluatu dugu; beste aldetik, "gai" idazlanez gain "ez gai" kalifikazioa duten beste horrenbeste idazlan gehituz ebaluatu dugu. Hartara, sailkatzaileak "ez gai" testuak nola sailkatzen dituen egiaztatuko dugu. Emaitza horiek lortzeko, balidazio gurutzatua erabili beharrean, ausaz aukeratu dira entrenatzeko lagina eta ebaluatzeko lagina 80-20 proportzioan, eta prozesua bost aldiz errepikatu da, laginak aldatuz (ikus 6. taula).

6. taula. Klase anitzeko sailkapenaren (B1, B2, C1 edo C2) emaitzak (2. sailkapen-ataza). Iturria: berezko ekoizpena.

Sailkapen Anitza (B1, B2, C1 edo C2)	GAI entrenatzeko eta GAI soilik ebaluatzeko	GAI entrenatzeko eta EZ GAI soilik ebaluatzeko	GAI entrenatzeko eta idazlan denak ebaluatzeko
EBM, ezaugarri guztiak	0,95 (±0,02)	0,84 (±0.05)	0,87 (±0,02)
EBM, 10 ezaugarri	**0,99 (±0,01)**	0,87 (±0.03)	0,91 (±0,02)

Emaitzetan ikus dezakegunez, gainditutako idazlanekin entrenatutako sailkatzaileak asmatze-tasa oso altuarekin sailkatu ditu mota horretako testuak; batez ere, 10 ezaugarri esanguratsuenak erabilita. Bestalde, "ez gai" testuak ebaluaziorako erabili ditugunean, emaitzak jaitsi egin dira. Horrek esan nahi du "ez gai" testu guztiak ez dituela dagokion multzopean sailkatzen, testuak maila baxuago batean sailkatzeko joera duela (ikus 7. taula) eta sailkatzaileak, egoki mailakatzeaz gain, idazlan gaindituen ezaugarriak ere kontuan hartzen dituela.

7. taula. Idazmen-proba gainditu duten testuen sailkapenaren kontingentzia-taula: ebaluatzeko erabili diren testuen benetako etiketa errenkadetan, eta sailkatzailearen iragarpena, berriz, zutabeetan. Sailkapena zuzena da, biak bat datozenean. Iturria: berezko ekoizpena.

GAI testuak		Iragarritako etiketa			
		B1	B2	C1	C2
Benetako etiketa	B1	**13**	1	0	0
	B2	0	**19**	0	0
	C1	0	0	**11**	0
	C2	0	0	1	**13**

8. taula. Idazmen-proba gainditu ez duten testuen sailkapenaren kontingentzia-taula. Iragarritako etiketa baxuagoa bada, ontzat hartu dugu, eta altuagoa bada, okertzat. Iturria: berezko ekoizpena

EZ GAI testuak		Iragarritako etiketa			
		B1	B2	C1	C2
Benetako etiketa	B1	**13**	1	0	0
	B2	3	**13**	0	0
	C1	0	3	**12**	0
	C2	0	0	2	**13**

8. taula sakonago begiratuta, aurkeztu den azterketaren mailan sailkatzen du ez-gaindituen % 84; hau da, modu okerrean. Bestalde, modu egokiagoan, maila bat beheragoko testutzat hartzen ditu testuen % 14, beharbada maila horretako testuen ezaugarri sinpleagoak dituelako. Azkenik, falta den % 2 maila bat gorago jartzen ditu modu okerrean, beharbada akatsak konplexutasun-ezaugarritzat hartuz.

Argi dago ezen, corpusa handituta eta ezaugarri gehiago kontuan hartuta, kalifikazio okerra jaso duten idazlan horien sailkapena hobetu litekeela. Halaber, aipatu nahi dugu posible dela maila jakin bateko testuen gaiek edota luzerak lekarketen ezaugarriek eta ezaugarrien banaketek nolabaiteko alborapenak sortzea; izan ere, bi ebaluazio-metodo horietan erdietsiriko emaitza onak baitira.

3.3. IDAZLANAK *B1-B2* ETA *C1-C2* MULTZOETAN SAILKATZEA

Gure gizarte-testuinguruan B2 eta C1 mailak bereizteak duen garrantzia kontuan hartuta, sailkapena sinplifikatu dugu: *B1-B2* etiketa B1 eta B2 probetako idazlanen multzoari esleitu zaio; *C1-C2* etiketa, berriz, C1 eta C2 probetako idazlanen multzoari. Hala entrenatu den sailkatzaile bitarraren emaitzak zertxobait hobetu dira, 9. taulan ikus daitekeenez; batez ere, entrenatzeko GAI idazlanak soilik erabiltzean.

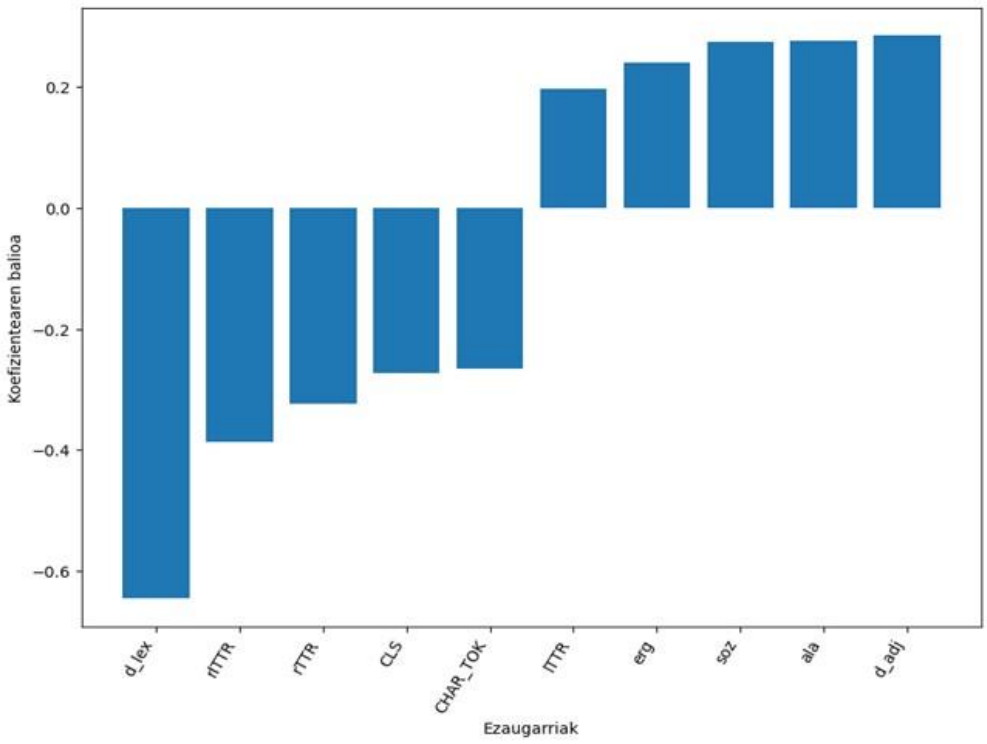
9. taula. Sailkapen bitarraren (B1-B2 edo C1-C2) emaitzak (3. sailkapen-ataza). Iturria: berezko ekoizpena.

Sailkapen bitarra (B1-B2 / C1-C2)	GAI idazlanak	Idazlan guztiak
SVM, ezaugarri-denak	0,942 (± 0,020)	0,91 (± 0,02)
SVM, 10-ezaugarri	**0,988 (± 0,025)**	0,92 (± 0,03)

Sailkatzaileen emaitzez gain, ezaugarri linguistiko bakoitzari zer pisu eman zaion aztertu dugu (6. irudia). Ezaugarrien garrantziaren interpretazioa sinpleagoa da sailkapen bitarrean: hiperplanoa eraikitzeko koefizienteen balio absolutu handiagoak garrantzia

handiagoa adierazten du; koefizientearen ikurrak, berriz, zer kategoriaren aldera sailkatuko duen: negatiboa baldin bada, lehen kategorian sailkatzeko joera izango du (*B1-B2*); positiboa baldin bada, ordea, bigarren kategorian sailkatzeko joera (*C1-C2*).[4]

6. irudia. Komunikagaitasun-mailaren sailkapen bitarrean, balio absolutuan pisu handiena duten 10 ezaugarriek dituzten pisuak, hiperplano sailkatzailea zehazteko. Iturria: berezko ekoizpena



Emaitzen arabera, testuak *B1-B2* gisa sailkatzeko ezaugarri linguistiko garrantzitsuenak honako hauek dira: i) dentsitate lexikala (izen, aditz, adjektibo eta adberbioen proportzioa), ii) aberastasun lexikalaren adierazgarriak (*Type-Token Ratio*, *root Type-Token Ratio*, *Lema Type-Token Ratio*), iii) *Coleman-Liau* indizea, iv) hitz bakoitzeko batez besteko karaktere kopurua, v) adjektibo dentsitatea eta vi) postposizioen eta kasu-marken banaketa (bereziki ergatiboa, soziatiboa eta adlatiboa).

3.4. IDAZLANAK HABE-KO EBALUAZIO-IRIZPIDEEN ARABERA SAILKATZEA

Laugarren sailkapen-atazari dagokionez, HABEren egiaztatze-prozesuan parte hartzen duten aztertzaileek ebaluazio-irizpide bakoitzerako jarritako puntuazioak iragarri ditugu. Sailkapen-ataza sinplifikatze aldera —datu gehiago behar baitira halako sailkapen-ataza konplexu bat aurrera eramateko—, irizpide bakoitzaren puntuazioak hiru multzotan bildu ditugu: “ona” (A, B), “nahikoa” (C) eta “ez nahikoa” (D, E) (10. taula).

[4] Esan behar da klase anitzeko sailkatzailean pisu horien interpretazioa zailagoa dela, hiperplano bakar bat egon beharrean, kategoria bikote bakoitzeko hiperplano bat baitago eta, beraz, ezaugarrien garrantzia ikertzeko beste metodo batzuk erabiltzea komeni da.

10. taula. Ebaluazio-irizpideen kalifikazioen banaketa, EEMB/HEOCeko maila bakoitzean. Zenbakiek kalifikazio hori jaso duten testu kopurua adierazten dute (gelaxka bakoitzean maila horretako testu guztiak daude: 120 testu). Iturria: berezko ekoizpena.

		Koherentzia	Kohesioa	Aberastasuna	Zuzentasuna
B1	ona:	53	48	43	40
	nahikoa:	56	51	67	41
	ez nahikoa:	11	21	10	39
B2	ona:	52	41	40	39
	nahikoa:	57	66	58	42
	ez nahikoa:	11	13	22	39
C1	ona:	46	30	30	23
	nahikoa:	59	72	65	55
	ez nahikoa:	15	18	25	42
C2	ona:	36	32	29	22
	nahikoa:	66	59	67	53
	ez nahikoa:	18	29	24	45

Aipatu beharrekoa da ezen, sailkatzaile horiek entrenatzeko, 2.4 atalean deskribatu dugun ezaugarri multzo berbera erabili dela. Hori horrela, koherentzia eta kohesioa testuen ebaluazio-irizpideak ezaugarri linguistiko-multzo berberarekin iragartzeak okerra dirudi, baina sailkapen-ataza horien helburua artikulu honetan ezarritako oinarrizko prozesamenduaren gaitasuna eta mugak aztertzea denez, zilegi da horretarako doitu ez den sailkapen-ataza honetako emaitzak erakustea. Beraz, garatu diren sailkatzaileen emaitzetan ikus daitekeenez (11. taula), emaitzak ez dira aurreko atazen puntuazio arrakastatsuetara heltzen, erabilitako ezaugarrien zerrendak ez baitaude doituta ebaluazio-irizpide horientzat; koherentzia eta kohesioa, esate baterako.

11. taula. Garatutako sailkatzaileen zehaztasuna (4. sailkapen-ataza), maila bakoitzean (B1, B2, C1 eta C2) eta irizpide jakin batekiko (koherentzia, kohesioa, aberastasuna eta zuzentasuna). Iturria: berezko ekoizpena

	Koherentzia	Kohesioa	Aberastasuna	Zuzentasuna
B1	0,37	0,34	**0,52**	0,43
B2	0,50	0,35	**0,58**	0,43
C1	0,49	0,50	**0,57**	0,51
C2	0,33	0,30	**0,43**	0,40

Zentzuzkoa da batezbesteko zehaztasun-puntuaziorik altuena lortu duen irizpidea aberastasuna izatea; izan ere, sailkatzaileak darabiltzan ezaugarri linguistikoen artean lexikoaren aberastasunarekin lotutako 20 ezaugarri baitaude, eta sailkatzaileak horien informazioa erabil baitezake entrenatzeko eta modu egokian sailkatzeko. Zuzentasunaren kasuan, ezaugarri bakarra (*akats ortografikoen tasa*) da lotura zuzena duena. Zuzenta-

sunaren sailkatzaileko ezaugarrien garrantziak aztertuz gero, akats-tasak du pisu handiena, batez beste, eta horrek erakusten du guk erabilitako zuzentzaile automatikoaren eta ebaluatzaileen emaitzak bat datozela, neurri batean. Koherentziarekin eta kohesioarekin lotura zuzena duen ezaugarririk ez dugu erabili, horiek ez baitira hain erraz lortzen eta euskarazko testuetatik informazio hori erauzteko tresnak, edo ez dira existitzen, edo ez dira nahikoa fidagarriak.

Sailkapena sinplifikatu arren, sailkapen-atazik konplexuena izan da azken hau. Ezaugarri automatikoak ditugun kasuan ere (aberastasuna) emaitzak ez dira onak. Oro har, sailkapen-ataza horren emaitzak hobetzeko, ezaugarri linguistiko konplexuagoak aztertu behar ditugu, eta ezaugarri linguistiko horiek testutik erauzteko metodoak garatu. Horrez gain, sailkapen-ataza horretan emaitzak hobetzeko corpusaren tamaina handitu beharko litzateke.

4. EZTABAIDA

Lortu ditugun emaitzak kontuan izanik, gure lanak hiru motatako erronka hauei aurre egin beharko die etorkizunean:

- **Corpusaren tamaina**. Lan honetan erabili den corpusaren kalitatea eta tamaina orain arte euskarazko lanetan erabili den egokiena eta handiena bada ere, askoz datu gehiago behar dira, sailkapen automatikoan emaitza erabilgarriak lortuko badira. Hori bereziki nabaria da sailkapen-ataza konplexuagoa denean (3.4 atalean ikusi den moduan). Gainera, lan honen garapenean zehar nabarmen ikusi dugu corpusa handitzeak, neurri batean bada ere, emaitzak hobetzen dituela.

- **Ezaugarri linguistiko orokorrak**. Oraingoz, erabili ditugun ezaugarri linguistiko guztiak ez dira testuak zuzendu behar dituzten irakasleentzat esanguratsuak edo, bestela esanda, ezin dira HEOCarekin konparatu. Argi ikusi dugu sailkatzaile automatiko batek eta zuzentzaile batek irizpide ezberdinak erabiltzen dituztela testuak EEMB/HEOCeko maila batekoa den ebazteko edo sailkatzeko. Edonola ere, ezaugarri ezberdin horien ikerketa eta garapena ezinbestekoak dira sailkatzaile fidagarriak eta adierazgarriak garatzeko.

- **HEOCean zehaztutako adierazpideak**. Lan hau asko aberastuko litzateke HEOCean deskribatzen diren hizkuntza-egituren gramatika aztertzerik balego eta testuak ezaugarri horien arabera mailakatuko balira. Hori gure etorkizun hurbileko asmoa bada ere, guk dakigula ez da horrelakorik modu sistematikoan garatu, ezta beste hizkuntzetarako ere. Gramatika horri esker, maila batetakoa den jakiteaz gain, testu horrek zer hizkuntza-baliabide erabiltzen dituen jakingo genuke eta, beraz, sailkapena esplikagarriagoa litzateke.

5. ONDORIOAK

Lan honen abiapuntu gisa, euskara ikasten ari den ikasleen euskarazko komunikagaitasun-maila zehazteko beharretik abiatuta, oso interesgarria litzateke jakitea, adibidez, euskara ikasten duten ikasleek zer maila duten euskaltegira sartzean, urtearen epe ezberdinetan hizkuntza-konplexutasuna garatzen ari diren gainbegiratzea eta ikasturtea bukatutakoan zer mailatan dauden jakitea. Horretarako, oso erabilgarriak izan litezke HABEk egiaztatze-prozesuetan, azterketaldiz azterketaldi, jasotzen dituen milaka idazmen-probak, bada, mailakatuak eta kalifikatuak baitaude, nahiz eta digitalizatu gabe egon. Lan honetan, idazmen-proba horietako zenbaitekin osatutako corpusa aurkeztu dugu, baita corpus horren gainean egindako sailkapen-probak ere. Corpus hori ikerketarako eskura dago helbide honetan, CC BY-NC 4.0 lizentziapean: https://doi.org/10.23728/b2share.81433fddcd06405f8505c7606b29ff99

Oro har, emaitza interesgarriak lortu dira hainbat teknika baliatuz eta, etorkizunean, corpus handiagoak lortzean eta informazio linguistikoa erauzteko tresna hobeak erdiestean, emaitza hobeak lortzea espero da. Hala, euskara deskribatzeko balio duten tresna horiekin, ikertzaileek eta irakasleek testuei buruzko informazioa lortzeko eta datuetan oinarritutako erabakiak hartzeko aukera izango dute, hainbat alorretan: testu mailakatuak idazteko ezaugarriak bistaratuz, testuen egokitasuna modu masiboan aztertuz eta ikasleak berarentzat egokia den testua aztertuz, besteak beste.

Emaitzei dagokienez, artikulu honetan erakutsi dugu ezaugarri bakoitzak bere garrantzia duela mailaren arabera eta halaber ikusi dugu *garrantzi* horiek kuantitatiboki neurtzeko metodoak ere badirela. Adibidez, B1-B2 testuak C1-C2koetatik banatzeko ezaugarri linguistiko garrantzitsuenak honako hauek dira: aberastasun lexikalaren adierazgarriak, dentsitate lexikala, *Coleman-Liau* indizea, hitz bakoitzaren batezbesteko karaktere-kopurua, adjektiboaren erabilera eta postposizioen eta kasu-marken aberastasuna eta banaketa.

Etorkizunean, testu-multzo handiagoekin lan egingo dugu, eta ezaugarri linguistiko aberatsagoak erabiliko ditugu. Horrez gain, HABEren eta EHUren arteko hitzarmenari esker, ikerketa honetan erabilitako corpusa eta tresnak CLARIN ikerketarako azpiegitura digitalaren webgunean baliatu ahalko dira, euskarazko ikerketa sustatzeko eta hezkuntzarako aplikazioak garatzeko.

Bibliografia eta erreferentziak

AGINDUA, 2022ko uztailaren 22koa, Hezkuntza Sailburuarena, Euskal Autonomia Erkidegoko Hezkuntza Legearen Aurreproiektua Aldez Aurretik onartzen duena. https://www.euskadi.eus/contenidos/proyecto_ley/21_pley_xileg/eu_def/adjuntos/2.-Aurretiazko-onarpena-eta-lege-aurreproiektu-testua.pdf

Agirre, E., Alegria, I., Arregi, X., Artola, X., Diaz de Ilarraza, A., Maritxalar, M., Sarasola, K. eta Urkia, M. (1992). Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. *Third Conference on Applied Natural Language Processing*, 119-125.

Aldabe, I., Arrieta, B., Díaz de Ilarraza, A., Maritxalar, M., Oronoz, M. eta Uria, L. (2005). Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, *10*(2), 47-59.

Bahassine, S., Madani, A., Al-Sarem, M. eta Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University – Computer and Information Sciences, 32*(2), 225-231. https://doi.org/10.1016/j.jksuci.2018.05.010

Beloki, Z., Saralegi, X., Ceberio, K. eta Corral, A. (2020). Grammatical Error Correction for Basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, *65*, 13-20. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6271

Bird, S., Klein, E. eta Loper, E. (2009). Natural Language Processing with Python. *O'Reilly Media Inc*.

Camacho, A. eta Iruskieta, M. (2021). Euskararen i(ra)kaskuntza-prozesuak: hezkuntza eta hizkuntza teknologiak. *Tantak, 32*(2), 9-31. https://doi.org/10.1387/tantak.21654

Chen, X. eta Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, 113-119. Osaka, Japan. https://aclanthology.org/W16-4113/

Correnti, R., Matsumura, L., Wang, E., Litman, D., Rahimi, Z. eta Kisa, Z. (2019). Automated Scoring of Students' Use of Text Evidence in Writing. *Reading Research Quarterly, 55*(3), 493-520. https://doi.org/10.1002/rrq.281

Guyon, I., Weston, J., Barnhill, S. *et al.* (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning 46,* 389–422. https://doi.org/10.1023/A:1012487302797

Eusko Jaurlaritza. Hezkuntza, Hizkuntza Politika eta Kultura Saila. (2016). *EAEko Hezkuntza-Sistema Hobetzeko Plana: Bikaintasunerantz ekitatean oinarrituta*. https://www.euskadi.eus/contenidos/informacion/innheziberri_hobekuntza_plana/eu_def/adjuntos/Hobetzeko_Plana_2016_martxo_e.pdf

Hasselgreen, A., eta Sundet, K. T. (2017). Introducing the CORYL Corpus: What it is and how we can use it to shed light on learner language. *Bergen language and linguistics studies*, *7*. https://doi.org/10.15845/bells.v7i0.1107

Ibarra, I., eta Iruskieta, M. (2022). Disgrafia hobetzeko eskuhartzea idazkailu digitala erabiliz. *Uztaro, 121*, 155-178. https://doi.org/10.26876/uztaro.121.2022.8

Euskararen Aholku Batzordea. IKT Batzorde-atala. (2016). *Euskarazko IKTak: gomendioak herri-aginteentzat*. Eusko Jaurlaritza, Hezkuntza, Hizkuntza-Politika eta Kultura Saila. https://www.irekia.euskadi.eus/uploads/attachments/8107/Euskarazko_IKT_Gomendioak_herri-aginteentzat.pdf?1463400638.

Maamuujav, U., Booth Olson, C. eta Chung, H. (2021). Syntactic and lexical features of adolescent L2 students' academic writing . *Journal of Second Language Writing, 53.* https://doi.org/10.1016/j.jslw.2021.100822

Larrea, K. (2009). Ikasleen hizkuntza erroreak eta hutsegiteak: zuzentzeko proposamena. *Revista de psicodidáctica*, *14*(1), 63-78. http://hdl.handle.net/10810/6475

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. eta Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *The Journal of the Maching Learning Research, 12,* 2825-2830. https://doi.org/10.5555/1953048.2078195

Osinalde, M. eta Iruskieta, M., (2022). Hizkuntza-ikasleen testu-corpus etiketatuaren analisia eta interpretazioa B2 eta C1 mailetan. *e-Hizpide, 100.* https://doi.org/10.54512/HLFA9295

Otegi, A., Imaz, O. Díaz de Ilarraza, A. Iruskieta, M. eta Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural, 58*, 77-84. http://hdl.handle.net/10045/64033

Otegi, A., Ezeiza, N., Goenaga, I. eta Labaka, G. (2016). A Modular Chain of NLP Tools for Basque. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue - TSD 2016, Brno, Czech Republic, volume 9924 of Lecture Notes in Artificial Intelligence,* 93-100*.*

Pérez, N. (2014). Erroreen analisia HABEren 2. mailaren azterketetan. *Hizpide: helduen euskalduntzearen aldizkaria, 84,* 3-23. https://www.ikasten.ikasbil.eus/mod/habecms/view.php/irakasbil/argitalpenak/erroreen-analisia-haberen-bigarren-mailaren-azterketetan

Unibaso, I. (2004). Autoikaskuntza-zerbitzua euskara ikas-irakasteko (I). *Hizpide: helduen euskalduntzearen aldizkaria, 55*, 88-104. https://www.ikasten.ikasbil.eus/mod/habecms/view.php/irakasbil/didakteka/autoikaskuntza-zerbitzua-euskara-ikas-irakasteko-ii

Schwarm, S. eta Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *ACL'05. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics June 2005,* 523–530*.* https://doi.org/10.3115/1219840.1219905

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning,* 197-207. https://doi.org/10.18653/v1/K18-2020

ERANSKINA

MATERIAL OSAGARRIA
Ezaugarri linguistikoen zerrenda osoa

Hizkuntza konplexutasun ezaugarriak (7)

- Esaldiko token-kopurua
- Esaldiko karaktere-kopurua
- Esaldiko silaba-kopurua
- Tokeneko karaktere-kopurua
- Tokeneko silaba-kopurua
- *Coleman-Liau Score*
- *Flesch-Kincaid Score*

Maila lexikaleko informazioa (20)

- Aniztasun lexikala: *type-token* ratioa eta *root type-token* ratioa testu osoan eta 100 tokeneko laginean (4)
- Aniztasun lexikala: *lema type-token* ratioa eta *root type-token* ratioa testu osoan eta 100 tokeneko laginean (4)
- Dentsitate lexikala: token lexikalen guztien (ad+iz+adj+adb) proportzioa token guztien artean.
- Token lexikal-mota bakoitzaren proportzioak token guztien artean (4)
- Bariazio lexikala: token lexikalen *type-token* ratioa
- Token lexikal mota bakoitzaren *type-token* ratioa (4)
- Tokenen maiztasunaren *log*a corpusean, batez beste
- Lemen maiztasunaren *log*a corpusean, batez beste

Ezaugarri morfo-sintaktikoak (32)

- Modu denboren (ind, subj, ahal, bald, orain, lehen) banaketa aditz guztien proportzio gisa (6)
- Kasu gramatikalen banaketa (abs, erg, dat, gen, des, mot, soz, ins, ine, abl, ala, abz, abu, gel, pro, par) kasu guztien proportzio gisa (16)
- Testuko menpekotasun-zuhaitz sakonenaren luzera
- Sintagmen batez besteko luzera

Erroreak (2)

- Akats ortografiko eta gramatikal-kopurua eta akats-kopurua tokeneko (2)