

## Euskararen eredutik hizkuntza-ereduen euskarara<sup>1</sup>

Izaskun Aldezabal Roteta\*, María Jesús Aranzabe Urruzola\*\*

HiTZ Hizkuntza Teknologiako Euskal Zentroa - Ixa ikerketa-taldea (UPV/EHU)

**Laburpena:** Lan honen helburua da euskararen ingurukoak azaltzea Euskal Filologia ikasketak abiapuntu hartuta eta ordenagailuek euskara ikasteko behar dituzten testu-corpus erraldoiak eta hizkuntza-ereduak lortzera iritsi direnera arte. Bide horretan bidelagun izan dugu Miren bai irakasle, bai gure tesi-lanetako epaimahaikide, bai sailkide modura. Hala, hizpide izan ditugu, alde batetik, hizkuntzalaritza konputazionalen murgilduta, morfologiari lotutako kontzeptu eta lan batzuk, bestetik, corpus-hizkuntzalaritza bideratzeko funtsezkoak diren zenbait corpus, eta, azkenik, erregistro akademikoaren garapena jomugan garatu ditugun datu-baseak, tresnak eta formakuntza-programa.

**Hitz gakoak:** euskara; morfologia; hizkuntzalaritza konputazionala; corpusak; corpus-hizkuntzalaritza; hizkuntza-ereduak.

### 1. SARRERA

Hizkuntzaren azterketarako interesa normalean berezkoa du filologia ikastea erabakitzen duen batek. Hizkuntzak dituen maila fonologiko, morfologiko, sintaktiko, semantiko eta pragmatikoak ulertzea, eta hizkuntza konkretu batek horretarako dituen baliabideetan arakatzea eta sorkuntzarako aukerak ikertzea, sinkronikoki edo diakronikoki, berez da interesgarria.

Ordea, horiek guztiak helburu jakin baterako aplikatu behar direnean, helburu horri begira landu behar dira horiek guztiak, eta horrek markatzen

\* Izaskun Aldezabal Roteta. Euskal Hizkuntza eta Komunikazioa Saila. Gipuzkoako Ingeniaritza Eskola (UPV/EHU). Europa plaza, 1 (20018 Donostia-San Sebastián). izaskun.aldezabal@ehu.eus. <http://orcid.org/0000-0001-7630-1406>.

\*\* María Jesús Aranzabe Urruzola. Euskal Hizkuntza eta Komunikazioa Saila. Kimika Fakultatea (UPV/EHU). Manuel Lardizabal pasealekua, 3 (20018 Donostia-San Sebastián). maxux.aranzabe@ehu.eus. <http://orcid.org/0000-0002-0401-1087>.

<sup>1</sup> Lan hau bi ikerketa-proiektu hauei atxikitzen zaie: Ixa Taldea. A motako Talde Finkatua (Eusko Jaurlaritza: IT-1570-22) eta HARTAvas (Zientzia eta Berrikuntza Ministerioa, MCIN: PID2019-109683GB-C22).



du azterketaren «baliagarritasuna». Hori gertatu zitzaigun guri Hizkuntzaren Prozesamendu Automatikoaren (HPA) munduan murgiltzen hasi ginenean, alegia, hizkuntzalaritza «konputazionalen»: ordenagailuari euskara «erakutsi» behar genion euskarazko testuak uler zitzan eta balizko akatsak identifikatu eta, ahal zela, zuzendu zitzan. Orduantxe hasi ginen jabetzen hizkuntzaren maila horiek nola osatzen ziren, edo osa zitezkeen, eta ikasitakoak «aplikatzen», hizkuntza konputazionalki lantzeari begira.

Gure hasierako hizkuntza-mailak hiztegia eta morfologia izan ziren, eta, jakina, hasieratik izan genituen eskuartean Miren Azkaratek irakasle gisa irakatsitakoak eta geroago ere euskaltzain nahiz sailkide gisa egindako ekarpen aberatsak.

Baina guretzat garrantzitsuena izan zen, hizkuntzaren esparru konputazional hau oraindik ezezaguna eta teoria mailan ez oso estimatua zenean, Miren Azkaratek beti erakutsi zuen gure lan-ildoarekiko interesa, begirunea eta eskuzabaltasuna, hizkuntzaren azterketa konputazionalak eskainitako aukerak eta horrek atzean izan duen talde-lana beti estimatu izan dituelako.

Hizkuntzalaritza konputazionalak orain arte, eta hemendik aurrera ingeniari-tza-hizkuntzalaritzaz lagunduta, hizkuntza aztertze eta ulertze bide zinez zirrargarriak eskaini ditu, eta horrekin batera, hizkuntzalaritza teorikoan ezezagun diren termino berri asko erabiltzea ere ekarri du. Hizkuntzalari konputazionalon lana, neurri handi batean, izan da terminologia berri hori hizkuntzalaritzako kontzeptuekin parekatzea eta diziplinartekotasun horretan bitartekari-lanak egitea. Termino berri horien aitzakiatan, hizkuntzalaritza konputazionalen barnako bidaia bat egingo dugu, hizkuntza-jakintza ordenagailuan nola kodetu dugun erakutsiz eta Miren Azkaratek jorratu dituen gai batzuekin lotura eginez.

## 2. HITZA, TESTU-HITZA ETA SARRERA-UNITATEA

*Hitza* terminoa hiztegiari lotuta ulertu izan da ia beti; hiztegia hitz-zerrenda bat da, hitz bakoitza bere forma kanonikoan, alegia, inolako flexiorik gabekoan adierazita.

Hiztegietatik haratago, hitzok testuetan aztertzean, zuriunetik zuriunerako unitate grafikoak ditugu, bere gauzapean sintaktikoan ageri direnak. Beraz, ez hiztegiko hitz baizik eta testuko hitzari erreferentzia egiteko ere baliagarri izan dadin, hitzaren beste definizio bat behar da: zuriunetik zuriunerako letra-segida (Fontenelle *et al.* 1994), izatez beti flexionatuta dagoena. Hiztegiko hitzak eta testuetako hitzak bereizteko *hitz* eta *testu-hitz* ere erabili izan da (Ezeiza 2002; Aldezabal 2004). Eta jakina, hitz bat baino gehiagoz osatutako hitz konplexuak edo *unitateak* ere baditugu, hitz horien elkarketak, osorik, balio sintaktiko-semantiko bakarra dutenak. Horregatik *hitz anitzeko unitateak* (esaterako, *hala eta guztiz ere*) eta *hitz bakunak* (esaterako, aurreko

adibideko hitz berak, honakoan hitz beregain gisa: *hala, eta, guztiz, ere*) bereizi beharra dago.

Testuko hitz horiek automatikoki ezagutzeko eta analizatzeko, baliabide eta tresna konputazionalak behar dira, eta hiztegiez haratago, datu-baseak erabiltzen dira (orain arte), euskararen moduko hizkuntza eranskarien ezau-garriak ondo kodetzeko. Datu-baseak «sarrerez» osatuta daude (hiztegiak dauden bezalaxe), baina datu-base horien helburuaren arabera, sarrera horien izaera ez da hiztegi-tan agertu ohi diren hitz beregainen modukoa bai-zik eta beharrak markatutakoen arabera. Adibidez, helburu morfosintakti-koa duen datu-base batek, hiztegi-sarrerez gain, morfema-sarrerak ere izango ditu; baita horien arteko lotura (morfofonologikoa) bideratzen duen aparteko modulua ere. Horiek guztiak ondorengo atalean azaltzen ditugu xehekiago, beste hainbat kontzepturekin batera.

### 3. MORFOLOGIA, SEGMENTAZIOA, MORFOTAKTIKA, LEMAK, ERROAK ETA MORFEMAK

Testu-hitzen azterketarako, euskaraz, hitzen osaera morfologikoa zehaz-tuta izan behar dugu, alegia, hitzen *segmentazio morfologikoa* egin beharra dago. Eta osaera morfologiko horretan hitza ez, baizik eta lemak, erroak eta morfemak bereizi behar ditugu. Baina garrantzitsuen dena: zein lemak zein morfema-segida onartzen dituen jakin behar dugu, alegia, *morfotaktika*, eta lotura horiek bideratu behar ditugu. Hau da, izen/adjektibo batek deklinabi-de-atzizkiak (Euskararen Gramatika (Euskaltzaindia 2021) *egitura-ko ka-suak* eta *adposizioak*) mugagabeen, singularrean eta pluralean onar ditzakeela aurreikusi behar dugu, besteak beste; orobat, aditzek aspektu burutu, burutugabe eta etorkizuneko morfemak onar ditzaketela, beste batzuen ar-tean; eta abar. Eta morfema-kateaketa horretan gertatzen diren aldaketa morfofonologikoak erregelen bidez deskribatu eta gauzatu behar dira. Hori guz-tia, hasteko, euskara batuan. Zeregin horretan, zalantzarik gabe, garai bateko EGLU I eta EGLU IIko deskribapenak (Euskaltzaindia 1991, 1997) ezinbes-teko oinarriak izan dira. Liburu horietan dagoen deskribapen osoa Euskara-ren Datu Base Lexikalean (EDBL) (Aldezabal *et al.* 1999, 2001) kodetu ge-nuen. Horretarako:

- a) EDBLko sarrera-unitate guztiak kategoria jakin baten arabera sail-katu genituen (ikus kategoria guztiak A eranskinean).
- b) Morfotaktikaren berri ematen zuten Jarraitze Klaseak (JK) definitu genituen; guztira 168. Izenen I JKaren barruan dagoen II JKaren adi-bidea dugu 1. taulan, mugagabe (IMG), pluraleko (IMP), singu-larreko (IMS) eta bizidunen kokapenezko deklinabidea (-gan, -gan-dik...) gauzatzeko.

**1. taula**

## I1 Jarraitze-klasearen adibidea

JK	Osagaiak (lexikoi nahiz JKak) eta adibideak	
I1	I0	etxe bat
	IMG	ardo, -en, -ik, -tan
	IMP	ardoek, -en, -etako, -on, -otako
	IMS	ardoa, -aren, -ko
	a_ms	ardoagan...

- c) Eta sarrera-unitateek JK horietako morfemak hartzerakoan gertatzen ziren aldaketa morfofonologikoak deskribatu genituen, erregelen birtartez. Erregela baten adibidea dugu 2. taulan, kontsonante afrikatuai (adibidean, zehazki, -ts- afrikatuari) leherkari gor bat lotzean gertatzen den aldaketa morfofonologikoa adierazten duena.

**2. taula**

## Aldaketa morfofonologiko baten adibidea

## 1. Txistukariak

## 1.1. Afrikariak + konts. =&gt; t = ø

Afrik + LhGor -ts + -t = -st-

adib.: *erakuts + -ten = erakusten*

Izen, adjektibo, aditz eta adberbio kategoriako sarrerei dagokienez, hitz-zerrendak UZEIk lantzen zituen testuen hustuketatik atera genituen, eta, erreferentziazko hiztegietan (garai haietan *Hauta-Lanerako Euskal Hiztegia*<sup>2</sup> eta *Elhuyar Hiztegia*<sup>3</sup> batez ere) ondo begiratu eta aztertu ondoren, datu-baseratu edota kanpoan uzten genituen.

Datu-basea etengabe hornitzen goaz, gaur egungo *Euskaltzaindiaren Hiztegia*<sup>4</sup> oinarri hartuta eta beste hiztegietako sarrerak (Elhuyar, Euskal-term,<sup>5</sup> batez ere) nahiz Xuxen ortografia-zuzentzailearen erabiltzaileen irado-kizunak kontuan hartuta.

<sup>2</sup> <http://www.euskara.euskadi.net/r59-sarasola/eu/sarasola/sarasola.apl>.

<sup>3</sup> <https://hiztegiak.elhuyar.eus/>.

<sup>4</sup> [https://www.euskaltzaindia.eus/index.php?option=com\\_hiztegiabildatu&Itemid=410](https://www.euskaltzaindia.eus/index.php?option=com_hiztegiabildatu&Itemid=410).

<sup>5</sup> <https://www.euskadi.eus/app/euskal-terminologia-banku-publikoa/viruela/kontsultaterrmino/viruela/non-du/hizk-es/ter-on>.

Gaur egun, 126.355 sarrera-unitate ditugu; horietatik 105.394 hiztegi-sarrerak dira (hitz beregainak, alegia), horien artean 2.219 hitz anitzeko unitateak izanik, eta gainerako guztiak, morfema ez-independenteak nahiz forma flexionatuak dira (aditz laguntzaile eta trinkoak barne).

#### 4. ERATORPEN-MORFOLOGIA, HITZ-ELKARKETA ETA MORFOTAKTIKAREN MUGAK

Morfema-kateaketa horretan, ordea, EGLUetan aurkitzen dugun morfologia flexibotik harantz, hitz berriak sortzeko mekanismo emankor diren eratorpen-morfologia eta hitz-elkarketa ere aintzat hartu behar ziren, jakina. Eta horretan ageri zaigu Miren Azkarateren lana guztiz erreferentzial eta beharrezko, hasita Euskal Filologiako ikasketak sortu zirenetik eman dituen ikasgaietan jorratutakotik, UZEIn eta Euskaltzaindiko LEF (Lexiko Ebazpenen Finkapena) Batzordean sortzen ziren txostenetara (Euskaltzaindia 1992, 1994).

Garai hartan dena paperean izaki, gure lana izan zen paperean zegoen informazio guztia formatu eskuragarriagoan eta baliagarriagoan jartzea, hau da, «datu-baseratzea», eratorpenez eta elkarketaz sortuak zeuden testu-hitzak ordenagailuak arazorik gabe ezagutu ahal izateko.

Txosten horietan azaltzen zirenak konputazionalki jartzea, ordea, ez zen sinplea. Morfema-kateaketaren bidez erraza zen *-tasun* atzizkia, oro har, adjektiboek hartzen dutela definitzea; baina, ez, aldiz, *-garri* eta *-kor* atzizkiak oro har nor-nork motako aditzek bakarrik onartzen zituztela zehaztea, horretarako lehenik eta behin, sailkapen hori egin izan beharko genukeelako eta ondoren atzizkien kateaketa jakinak aditzen azpimultzo jakinei lotu. Are gehiago, nahiz eta *-kor* atzizkiak nor-nork motako aditzak hartu, ikuspegi semantikoan sartuta, zail gertatzen zen zehaztea, sortzen dituen adjektiboak beti balio kausatibo/inkoatiboko (alegia, *da/du* esaten zaien aditzetako) *da* baliotik ulertu behar direla (hautsi > hauskor: zerbait hauskorra da (adib., *material hauskorra*) eta ez zerbaitek hausten du), eta, aldiz, aditz transsitibo edota *du/dio* motakoa denean *du* balioko subjektutik ulertu behar dela (ulertu > ulerkor: norbaitek zerbait ulertu (adib. *pertsona ulerkorra*)) (Euskaltzaindia 2021). Atzizkien bereizketa semantiko horiek egiteko, aditzen bereizketa xehegoa egin behar litzateke, eta ohiko mota sintaktikoetatik harantz, aditz kausatibo/inkoatiboak eta aditz transsitiboak bereizi.

Horiek horrela, oinarri-oinarrizko morfotaktika lantzeko helburuaz, lan horietan ageri ziren atzizki eta aurrizki lexikal emankorrenak hautatu eta horiek ere EDBL datu-basean kodetu genituen; alegia, atzizki horiek JKen barruan txertatu genituen, eta beharrezko izan ziren erregela morfologikoko berriak sortu genituen. Zehazki, 22 atzizki eta 2 aurrizki landu genituen.

Esaterako, 3. taulan, *-tasun* atzizkia ikus daiteke, adjektiboentzat sortutako ADJK JKaren barruan txertatuta, beste atzizki batzuekin batera lexikoi edo etiketa/multzo berean definitua (erat\_adj) eta I1\_MARRA JKa duela, izen-sortzaile izanik, ondoren onartzen duen morfema-multzoa hartzeko. *-garri* atzizkia, berriz, aditzek normalean hartzen duten JKetako baten barruan (A3) txertatuta dago (aditza nor-nork motakoa den kontuan izan gabe), eta adjektiboeti normalean eransten zaien JKa (ADJK5) du morfotaktika gauzatzeko.

### 3. taula

*-tasun* eta *-garri* atzizkiak JKetan

Atzizkia	Zein JKtan txertatuta	Zein lexikoitan	Zein JKrekin
-tasun	ADJK	erat_adj	I1_MARRA
-garri	A3	erat_adj	ADJK5

Hitz-elkarketaren kasuan, jakinik ere hainbat mota daudela (Euskaltzaindia 1992, Euskaltzaindiaren 25. araua<sup>6</sup>), marratxoa aukeran duten mendekotasunezko hitz elkartuak baino ez genituen landu morfotaktikaren bidez, bideragarria izateaz gain, motarik emankorrena zelako (beste mota guztietako hitz konposatuak hiztegi-sarrera beregain gisa lantzen joan gara). Horretarako, izenak hartzen duten JKan ELK\_MARRA JKaren bitartez gidoitxoa gehitu eta gidoitxoari aldi berean beste JK bat definitu genion ondoren izen guztiak har zitzan. 4. taulan ikus daitezke izenak hartzen duten I JKaren osagaiak, eta, 5. taulan, ELK\_MARRA JKaren osaera.

### 4. taula

Izenak hartzen duten I JKa,  
hitz elkartuen marratxoa txertatua duena

JK	Osagaiak	Adibideak
I	ELK_MARRA	botoi-zulo...
	I1	ardo, -a, -ek, -agan
	adj_ago	gizonago, basakristauago...
	dun_atz	pisudun...
	erat_ize	pisutsu, pisuka, mahaigile...
	grad1	etxetxo, etxetzar, etxeño...
	ko_ban	begiraleko (hamarna haur)
	ko_desk	bost urteko ardoa
	pe	giltzapean...

<sup>6</sup> <https://www.euskaltzaindia.eus/hizkuntza-baliabideak/baliabide-orokorrak/arauak>.

**5. taula**

Mendekotasunezko hitz elkartuak sortzeko ELK\_MARRA JKaren osaera

JK	Osagaiak	Adibidea
ELK_MARRA	elk_marra	botoi-zulo

**5. OINARRIZKO TRESNAK**

Testu-hitzen analisi morfologikoa automatikoki tratatzeko, beraz, ezin besteko urratsa da hitzak nola osatuta dauden deskribatzea. Osaketa horretan oinarrituta jakingo du analizatzaile morfologikoak nola lot daitezkeen lemak eta morfema-segidak, eta zein diren egin beharreko aldaketa morfofonologikoak horien arteko lotura gauzatzean. Horrela lortuko du, adibidez, *Mireni* ezagutzea eta sortzea, eta ez *Mirenei*, edo *bulegotik* eta ez *bulegotikan*. Analizatzaile horiek gai izaten dira emandako hitzaren analisia lortzeko edo hitz baten lema emanda dagozkion analisi morfologikoak eskaintzeko. Esaterako, 1. irudian ikusten den «Amagoiak Zuberoan lan egiten du» egituraren analisi morfologikoa da euskararako garatu den Morfeus analizatzaile morfologikoak (Aduriz *et al.* 1992) eskaintzen duena.

Amagoiak	Zuberoan	lan	egiten	du
<i>Amagoia</i> + <i>k</i> IZEIZB+ERG	<i>Zuberoa</i> + <i>0</i> + <i>n</i> IZELIB+ <i>Sar</i> +INE	<i>landu</i> + <i>0</i> ADISIN+AMM	<i>egin</i> + <i>te</i> + <i>n</i> ADISIN+AMM+INE	<i>du</i> ADL
<i>amagoi</i> + <i>ak</i> IZEARR+ABS		<i>lan</i> IZEARR	<i>egin</i> + <i>0</i> + <i>ten</i> ADISIN+AMM+ASP	<i>du</i> ADT
<i>amagoi</i> + <i>ak</i> IZEARR+ERG		<i>lan</i> + <i>0</i> IZEARR+ABS		

**1. irudia**

*Amagoiak Zuberoan lan egiten du* esaldiaren analisi morfologikoa

Hain zuzen ere, analizatzaile morfologiko hori da Xuxen ortografia-zuzentzailearen oinarria (Agirre *et al.* 1992). Xuxenek, testuko hitz baten forma ortografikoki zuzentzat edo okertzat hartzeko, hitz horren lema (eta ez forma) datu-basean dagoela egiaztatzeaz gain, lema horri lotzen zaizkion atzizkiak dagozkionak diren ala ez egiaztatzen du.

Bi horiek, analizatzaile morfologikoa eta Xuxen ortografia-zuzentzailea, izan dira euskararen tratamendu automatikoa gauzatzeko abiapuntuak.

Morfologiatik haratagoko maila linguistikoen azterketa (sintaxiarena eta semantikarena, esaterako) bideratzeko, ordea, nahitaezkoa da lematizatzaile/etiketatzaileak garatzea. Lematizatzaile/etiketatzaile horien bitartez lortzen da, adibidez, 1. irudiko analisisian ikusten den hitz-formetako bakoitzak analisi morfologiko bakarra izatea (2. irudia). Lematizatzaileek testuko hitz bakoitzak izan ditzakeen lema posibleen artetik dagokiona aukeratzeko dute eta etiketatzaileek, berriz, izan ditzakeen analisisetatik zuzena dena; hau da, tresna horien helburu nagusia desanbiguatzea da. Horrela, bada, lematizatzaile/etiketatzaileak testuinguruaren arabera erabaki behar du zein den, adibidez, 1. irudiko *Amagoia* hitzari dagokion lema (*Amagoia* ala *amagoi*) eta etiketa zuzena (izen berezia ala arrunta). Euskararako garatu den Eustagger lematizatzaile/etiketatzailearen (Alegria *et al.* 2002) emaitza da 2. irudikoa.

Amagoiak	Zuberoan	lan	egiten	du
<i>Amagoia</i>	<i>Zuberoa</i>	<i>lan</i>	<i>egin</i>	<i>*edun</i>
IZEIZB	IZELIB	IZEARR	ADI	ADL

## 2. irudia

*Amagoiak Zuberoan lan egiten du*  
esaldiaren analisi morfologikoa desanbiguatuta

Alegria *et al.*-ek (2017) dioten moduan, Eustagger oso baliagarria da terminologia/lexikografia lanetarako. Testu batetik terminologia erraz erauzteko aukera ematen du, baldin eta automatikoki lemak ondo identifikatzen badira eta dagozkien etiketak egokitzen bazaizkie.

## 6. CORPUS HIZKUNTZALARITZA

Testuak, corpusak alegia, funtsezko baliabideak dira hizkuntzaren inguruko ikerketa-lanetarako. Corpus horiek era, mota eta tamaina guztietakoak izan daitezke (Alegria *et al.* 2005; Aldezabal *et al.* 2009); hau da, bakoitzaren ezaugarriak corpora osatzerakoan jarritako helburuen eta erabilera-aren arabera izaten dira. Corpus horien erabilera ere nabarmen handitu da azken urteotan HPAn eta hiztegi-gintzan ez ezik, hizkuntza-i(r)a-kaskuntzan (Iruskieta *et al.* 2018). Ikasketa sakona (*deep learning*) nagusitu denetik (Agirre *et al.* 2022) are garrantzi handiagoa hartu dute hizkuntzaren erabilera *errealaren* erakusleio diren corpus erraldoiek.



Ixa ikerketa-taldean ez dira gutxi sortu ditugun corpusak. Horien artetik hiru hauek aipatu nahi ditugu, esanguratsuak direlako deskribatzen ari garen ibilbide honetan:

- a) EPEC corpora (Euskararen Prozesamendurako Erreferentzia Corpora) (Aduriz *et al.* 2006): gramatika-erregelatan oinarritutako tresnak garatzeko sortu dugun linguistikoki etiketatutako 300.000 testu-hitzeako corpora.
- b) EusCrawl corpora (Artetxe *et al.* 2022): datu ugari izatea hain garrantzitsua den garai honetan, denon eskuragai dagoen corpusik erraldoiena (423 milioi token), euskararako hizkuntza-eredurik ahalik eta egokiena sortu ahal izateko funtsezkoa.
- c) Garaterm corpora (Zabala *et al.* 2013): unibertsitateko jakintza-alor desberdinetako irakasmaterialez eta ikerketa-lan akademikoez osatutako 25.179.342 testu-hitzeako corpora, monitorizatuia izanik, etengabe elikatzen ari garena.

Euskarazko testuez osatutako hiru corpus horiek eredu dira corpus horietan oinarrituta egiten diren azterketetarako edo ikerketa-lanetarako.

Alderatzen hasita, bi muturretan kokatuko ginituzke lehen biak, EPEC eta EusCrawl, horietako bakoitzak garaian nagusi zen eta den sistema edo teknologia erakusten baitute; lehenak maila linguistiko desberdinetako etiketatze egokian oinarritzen zuen kalitatea (horregatik kantitatean mugatua) eta bigarrenak, aldiz, testu zuzen eta egoki ugari izatean (ereduek hortik ikasteari begira). Taldean osatu dugun EPEC corpusaren osaerak erakutsi digu euskara bezalako baliabide urriko hizkuntzetan zein zaila den corpus bat osatzea, bai diruaren aldetik, bai denboraren aldetik; baina era berean zein baliagarria den halako baliabideak izatea hizkuntzari/euskarari buruzko azterketa xeheak egiteko. Bigarren corpusak, EusCrawlek, teknologiak zein abailatan egiten duen aurrera erakutsi digu, egun nagusi diren hizkuntza-ereduak halako corpus erraldoiez baliatzen direlako hizkuntzaren nondik norakoak ikasteko. Hirugarren corpora, Garaterm, guk dakigula euskarazko corpus akademiko bakarra da, eta harreman zuzena du gure sailaren, Euskal Hizkuntza eta Komunikazioa Sailaren zerreginekin eta honenbestez sailkide dugun Mirenenekin partekatutako lanekin.

Garaterm corpora Garaterm proiektuari (Zabala *et al.* 2008) eta proiektu hori gauzatzen lagundu duen Terminologia Sareak Ehunduz (TSE) (San Martín 2013; Aldezabal *et al.* 2017) programari esker garatu den baliabidea da. Euskal Hizkuntza eta Komunikazioa Saileko zenbait irakasleren eskutik (Miren horietako bat) eta abiapuntu hartuta jakintza-alor desberdinetako irakasleen irakasmateriala eta ikerketa-lan akademikoak, ikusgai jarri nahi izan ditugu komunikazio espezializatuetan erabiltzen diren terminologia eta fraseologia errealak. Hori gauzatzeko diseinatu den Garaterm lan-inguruneari (Zabala *et al.* 2013) eta TSEn jarraitu den metodologiari esker, irakasleak berak dira testuak prestatu eta lan-ingurune horretara igotzen dituztenak. On-

doren, Eustagger lematizatzailer/etiketatzailer baliatuta, testuak linguistikoki prozesatzen dira eta kontsultagai jartzen dira Garaterm corpusean.<sup>7</sup> Behin testuak prozesatuta eta testu horietako termino hautagaiak Erauzterm termino-erauzlearen bitartez (Alegria *et al.* 2004) erauzita (ikus 3. irudia), testu horien egileak, irakasleak berak, dira testu horietako terminologia lantzen dutenak «deskripzio aktiboa» deritzon metodologiari jarraituz.

**NErauzterm\_594** (1178 termino hautagai)

Bistaratuak: 1178 | 1178 | Esportatu | Hiztegiak | Erlazioak atara

Lema	Bal.	Alorra	Hiztegiak	Eredus	Maiz.	Kop.
56 datu experimental	TI	Kim.	NApos	12	2	
datu experimentalen doiketa	TI	23	NAprepN	2		
datu experimental berdin	TI	23	NAposApos	1		
datu experimentalen tratadumendu	TI	23	NAprepN	1		
57 absorbantzia	TI	Kim.	TZOS ZTH EUS	11	1	
58 altuera	TI	Kim.	TZOS ZTH MAT EUS	11	1	

Lema: Lema | Italoa: Denak | Domenua: Denak | Irudua: Denak | Maiz.: 1 | Kop.: 4

1 2 3 4 5 6 7 8 9 10 > >> 1 / 12

**Testuinguruak**

DATU-TRATAMENDUAK LANTZEKO NEURRI BOLUMETRIKOAK Helburuak **datu experimentalak** neurtses inurikapen grafikoak lantzeko.

Tratamendu matematikoa: **datu experimentalak** doiketak.

...amendua Datu-tratamenduari buruzko mintegian erabiltzeko **datu experimentalak** itauek.

Sar itzazu 1. orrian (1 ura Sar itzazu 2. orrian (2 ura **datu experimentalak** doiketa egokiena aukeratzeko.

Uraren itzaz lortutako **datu experimentalak** eta informazio bibliografiko konbinatuz, tresnerien ber...

...ra, °C - tan, denboraren funtzioan, s - tan, eta doitu d **datu experimentalak** 2. graduako polinomio batera, Luzatu taula... Lehenengo urr...

Lortutako **datu experimentalak** gas baten isotermak erak. **datu experimentalak** gas baten isotermak erak.

... eta gailuri buruzkoak Zeigatik da gomendagarria erabiltzea **datu experimentalak** berdinak presio-bolumen-temperatura erlazioan dituzten...

### 3. irudia

Erauzitako terminoak balioztatzeko Erauztermen interfazearen adibidea

Lan horren ondotik osatzen diren glosarioak edo hitz-zerrenda eleaniztunak sarean kontsulta daitezke horretarako garatu den TZOS<sup>8</sup> (Terminologia Zerbitzurako Online Sistema) (Arregi *et al.* 2013; Aldezabal *et al.* 2022) baliabidearen bitartez. TZOSek helburu du adituek erabiltzen dituzten termino «natural»ak, aldakiak barne, deskribatzea, aldakiak diren horiek noizbait etorkizunean harmonizatu ahal izateko (Azkarate 2017). Egun, 136.540 termino daude kontsultagai. Termino horietako baten adibidea ikusten da 4. irudian. TZOS Garaterm corpusarekin lotuta dago, beraz, terminoak testuinguru diskurtsiboan azter daitezke.

<sup>7</sup> <http://garaterm-corpusa.ixax.es/>.

<sup>8</sup> <http://tzos.ehu.es/>.

<b>eredu</b>	
Zientzia Teknologikoak » Ingeniaritza eta Teknologia Elektrokoak	
<b>ald.</b>	<b>modelo</b>
<b>fr</b>	<b>modèle</b>
<b>en</b>	<b>model</b>
<b>es</b>	<b>modelo</b>
Saillkapena	
<b>Jakintza-arloak</b>	Zientzia Teknologikoak » Ingeniaritza eta Teknologia Elektrokoak
<b>Jatorria</b>	Bilboko Ingeniaritza Eskola » Industria Ingeniaritza Teknikoa - Elektrizitatea espezialitatea » Makina Elektrokoak
<b>Erabiltzaile-kopurua</b>	1
Informazio linguistikoa	
<b>Estatua</b>	Erabilia (TZOS komunitatea)

#### 4. irudia

*eredu* terminoaren sarrera TZOSen

Corpus Hizkuntzalaritzaren ildoan sakonduz eta HPAn oinarrituta, beraz, adituek egiten duten euskararen erabilera akademikoa monitorizatzeko eta ikertzeko baliabide nahiz tresnak garatu eta aberasten ari gara (Aranzabe *et al.* 2022a), erabilera horiek egonkortzen lagunduko dutelako, besteak beste.

Eredu dira corpusak bai gizakiontzat, bai tresna konputazionalentzat. Ezin erakusleihu hobeak dira egiten den hizkuntzaren erabilera aztertzeko edota hizkuntzaren inguruko aplikazio gehienenen oinarri diren **hizkuntza-ereduak** «entrenatzeko». Gizakiak, esaterako, TZOS osatzen ari dira corpusek gidatutako terminologia-lana eginda; euskararako garatu den iXamBERT hizkuntza-eredua (Agerri *et al.* 2020), berriz, HARTAes-vas proiektu koordinatuan (Alonso & Zabala 2022) erabiltzen ari gara bi corpusetako formulak edo *lexical bundles*ak (adibidez, *laburbilduz*, *azpimarratu beharra dago*) etiketatzeko (ikus 5. irudia).

Muxia Aranzabe Urruzola: eus-corpus/fold\_0002\_eus-corpus.conll.v

1-10 / 5000 sentences [doc 3 / 23]

Annotation: Delete X Clear

Layer: Named entity

Text: agerikoa da

No links or relations connect to this annotation.

value: READER\_INDICERT x ▼

1 Baina hitz hurrenkera hori ez da sakoneko egituran topatzen duguna , mugimenduen beharra agerikoa da .

2 ( 1a - b ) perpausel gainbegiratu bat emanaz gero oharikotu gara egitura desberdinak direla . ( 1a ) - n

3 Baina ideia hori ez da zuzena . Izan ere , Beaudricq kontraste fokua deituriokak landu zituen Ortiz de Urbinak (

## 5. irudia

iXambert hizkuntza-ereduarekin etiketatutako HARTAEus corpuseko testu zatia

dagozkien funtzio diskurtsiboekin (6. taula). Bi corpus horiek dira gaztelaniako HARTA-de-noveles-para-el-español (Villayandre 2018; García-Salido *et al.* 2018) eta euskarako HARTAEus (Aranzabe *et al.* 2022b).

Hau da, konparagarriak diren bi corpus horietan formula monolexikoak (adibidez, *gutxienez*) eta bi hizkuntzen arteko kokapen- eta formula-balio-kideak hobeto identifikatzen saiatzeko, hizkuntza-ereduetan oinarritutako *transformer* ereduak (Vaswani *et al.* 2017; Devlin *et al.* 2018) erabiltzen ari gara. Hala, gaztelaniarekin eta euskararekin lan egiteko *transformer* eredu elebakarrak erabili ditugu (gaztelaniarako BERTIN (De la Rosa *et al.* 2022; Cañete 2020) eta euskararako iXamBERT (Agerri *et al.* 2020)), baina baita eredu eleanitzak ere (Otegi *et al.* 2020; Lample 2019).

## 6. taula

Funtzio diskurtsiboak eta dagozkien etiketak

1 Estructurar el texto	: EST
1_1- Añadir información	: EST_ADDINFO
1_2- Comparar	: EST_COMP
1_3- Delimitar	: EST_DELIM
1_4- Ejemplificar	: EST_EXEMP
1_5- Expresar causa	: EST_EXPCAU
1_6- Expresar condición	: EST_EXPCOND
1_7- Expresar consecuencia	: EST_EXPCONS
1_8- Expresar finalidad	: EST_EXPPURP
1_9- Expresar oposición	: EST_EXPOPOS
1_10- Hacer referencia al propio trabajo	: EST_AUTOREF
1_11- Introducir un tema	: EST_INTTOPI
1_12- Introducir una alternativa	: EST_INTALT
1_13- Introducir una excepción	: EST_INTXEC
1_14- Ordenar	: EST_ORD
1_15- Reenviar	: EST_RESEND
1_16- Reformular	: EST_REFORM
1_17- Resumir	: EST_SUMM
2 Referirse al contenido de la investigación	: REF
2_1- Definir y describir	: REF_DEFDESC
2_2- Denominar	: REF_DENOM
2_3- Establecer grupos	: REF_SETGROUPS
2_4- Expresar cantidad	: REF_EXPAMOUNT
2_5- Expresar correlación	: REF_EXPCOR
2_6- Expresar frecuencia	: REF_EXPFREQ
2_7- Expresar progresión	: REF_EXPPROG
2_8- Expresar tiempo	: REF_EXPTIME
2_9- Presentar datos	: REF_PRESDATA
2_10- Presentar el objeto de estudio	: REF_PRESRES
2_11- Presentar hipótesis	: REF_PRESHIP
2_12- Presentar la metodología	: REF_PRESMETH
2_13- Presentar las conclusiones	: REF_PRESCONC
2_14- Presentar los objetivos	: REF_PRESOBJ
3 Posicionarse y dirigirse al lector	: READER
3_1- Atenuar	: READER_ATEN
3_2- Expresar necesidad	: READER_EXPNEED
3_3- Expresar una evaluación	: READER_EXPEVAL
3_4- Generalizar	: READER_GENERAL
3_5- Hacer hincapié	: READER_EMPH
3_6- Indicar certeza	: READER_INDNCERT
3_7- Indicar la fuente	: READER_INDNSOUR
3_8- Indicar posibilidad	: READER_INDPOS

## 7. ETA ORAIN?

Zalantzarik gabe, azken bospasei urteetan, hizkuntza-teknologiek garai batean pentsaezinak ziruditen helburuetara eramán gaituzte. Adimen artifizialeko aurrerakuntzek itzulpen automatikoan, adibidez, izugarrizko kalitate-jauzia ekarri dute, zenbaitetan gizakiarena baino hobea izateraino. Orain dena corpusetik «ikasten» dute «makinek», eta azken emaitza horiek lortzeko ez dira tarteko hizkuntza-baliabide «sofistikatuak» behar (orotariko datu-baseak, erregeletan oinarritutako analizatzaile morfologiko, sintaktiko nahiz semantikoak, etab.).

Berrikuntza horiek guztiek hizkuntzalariok apur bat noreaezean utzi gaituzte eta orain arteko jardute-modua asko aldarazi digute bai ikerkuntza-arloan, bai irakaskuntza-eremuan.

Horiek horrela, gaur egun bi jarduteko moduak, bata, esan dezagun «tradizionala», eta bestea, oraingoa, adimen artifizialak gidatutakoa, paraleloki garatzen dihardugu. Oraingo ildo berriak, hasierako hizkuntzalaritza konputazionalak bezalaxe, kontzeptu berri asko ekarri dizkigu: hizkuntza-ereduak (corpus batetik ikasitakotik sortutakoak; zenbat corpus hainbat eredu); ataza jakinetara egokitzeko moldaketak (*fine-tuning* delakoa), corpus hutsa erabiltita (*zero-shot* delako moduan, alegia, inolako hiztegiaren beharrik gabe), eta halakoak. Denak ere oso berriak hizkuntzalarion belarrietara, baina poliki-poliki geure egiten ari garenak.

Horixe da, hortaz, oraingo gure erronka: ikuspegi tradizionaleko lan-ildoak alde batera utzi gabe, ildo berri honetan gure alea jartzea; besteak beste, ereduak ebaluatzea edo/eta hobetzea, horretarako beharrezko corpus edota datutegi zehatzak (*data-setak*) diseinatu eta sortzea, eta makinek «zer» ikasten duten eta «zer ez» aztertzea.

Eta seguru gaude bide horretan ere Miren Azkarate bidelagun izango genukeela, *garaian garaiko metodologiaz, deskribamoldeez baliatuz kontatzen baitira hizkuntzen gertakariak* (Azkarate 2020).

## 8. BIBLIOGRAFIA

- Aduriz, Itziar, Eneko Agirre, Iñaki Alegria, Xabier Arregi, Jose Maria Arriola, Xabier Artola, Arantza Diaz de Ilarraza, Nerea Ezeiza, Montse Maritxalar, Kepa Sarasola & Miriam Urkia. 1992. A Morphological Analyzer for Basque based on Two-level Morphology. *Proceedings of the 5th International Morphology Meeting*. Austria: Universidad de Krems. <http://ixa.ehu.eus/node/3363>.
- Aduriz, Itziar, María Jesús Aranzabe, Jose Maria Arriola, Aitziber Atutxa, Arantza Diaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa & Ruben Urizar. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the

- automatic processing. In Andrew Wilson, Paul Rayson & Dawn Archer (arg.), *Corpus Linguistics Around the World. Book series: Language and Computers* 56, 1-15. Netherlands: Rodopi. <http://ixa.ehu.eus/node/3294>.
- Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa & Eneko Agirre. 2020. *Give your Text Representations Models some Love: the Case for Basque*. <https://doi.org/10.48550/arXiv.2004.00033>.
- Agirre, Eneko, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Montse Maritxalar, Kepa Sarasola & Miriam Urkia. 1992. Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. In *Third Conference on Applied Natural Language Processing*, 119-125. Italia: Association for Computational Linguistics. <http://doi.org/10.3115/974499.974520>.
- Agirre, Eneko, Marianna Apidianaki & Iva Vulić. 2022. Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. USA: Association for Computational Linguistics. <https://aclanthology.org/2022.deelio-1>.
- Aldezabal, Izaskun, Olatz Ansa, Xabier Artola, Aitzol Ezeiza, Koldo Gojenola, Jon M. Insausti & Mikel Lersundi. 1999. *Euskararen Datu-Base Lexikala (EDBL): eskema berriaren proposamena*. Donostia: UPV/EHU Informatika Fakultateko barne-txostena. <http://ixa.ehu.eus/node/4060>.
- Aldezabal, Izaskun, Olatz Ansa, Bertol Arrieta, Xabier Artola, Aitzol Ezeiza, Gregorio Hernández & Mikel Lersundi. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. In Steven Bird, Mark Liberman & Peter Buneman (arg.), *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia: University of Pennsylvania. <http://ixa.ehu.eus/node/3301>.
- Aldezabal, Izaskun. 2004. Aditz-azpikategorizazioaren azterketa sintaxi partziale-tik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz. Leioa: UPV/EHUko doktoretza-tesia. <http://hdl.handle.net/10810/13974>.
- Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Ainara Estarrrona, Nerea Ezeiza & Larraitz Uriá. 2009. Corpusen etiketatze linguistikoa. In Ricardo Etxepare, Ricardo Gómez & Joseba A. Lakarra (arg.), *A Festschrift for Bernard Oyharçabal. Volumen especial del Anuario del Seminario de Filología Vasca Julio de Urquijo (ASJU)* 43(1-2). 37-50, Bilbo: UPV/EHU Argitaipen Zerbitzua. <http://ixa.ehu.eus/node/3290>.
- Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Ilarraza & Igone Zabala. 2017. Terminologia lantzeko baliabideak EHUn. *Senez* 48. 211-216. <http://ixa.ehu.eus/node/11368>.
- Aldezabal, Izaskun, Jose Mari Arriola & Arantxa Otegi. 2022. TZOS: an Online Terminology Database Aimed at Working on Basque Academic Terminology Collaboratively. In Nicoletta Calzolari (arg.), *Proceedings of the 13th Language Resources and Evaluation Conference*, 1353-1359. Marseille: European Language Resources Association. <https://aclanthology.org/2022.lrec-1.144>.
- Alegria, Iñaki, Xabier Artola & Kepa Sarasola. 1997. Hizkuntzaren tratamendu automatikoa. Helburuak eta abiaburuak. *Jakin* 102. 61-82. <http://ixa.ehu.eus/node/3922>.
- Alegria, Iñaki, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza & Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *Third International Conference on Language Resources and Evaluation*



- (LREC'02): *Customizing knowledge in NLP applications workshop*, 1-6. Las Palmas de Gran Canaria: European Language Resources Association (ELRA). <http://ixa.ehu.es/node/3340>.
- Alegria, Iñaki, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea & Ruben Urizar. 2004. An Xml-Based Term Extraction Tool for Basque. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, 1733-1736. Lisboa: European Language Resources Association. <https://aclanthology.org/L04-1161/>.
- Alegria, Iñaki, Xabier Artola, Nerea Ezeiza, Kike Fernandez, Antton Gurrutxaga, Igor Leturia, Aitor Sologaitoa, Aitor Soroa, Andoni Valverde, Nerea Areta, Ziortza Polin & Rafa Saiz. 2005. Zientzia eta Teknologiaren Corpusa. Diseinua eta metodologia. In *Espezialitate hizkerak eta terminologia II. Euskara estandarra eta espezialitate hizkerak*, 1-23. Leioa: UPV/EHU Euskara Institutua. <http://ixa.ehu.es/node/3297>.
- Alonso-Ramos, Margarita & Igone Zabala. 2022. HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations, SEPLN*. A Coruña. <https://ceur-ws.org/Vol-3224/paper06.pdf>.
- Aranzabe, María Jesús, Antton Gurrutxaga & Igone Zabala. 2022a. Compilación del corpus académico de noveles en euskera HARTAeus y su explotación para el estudio de la fraseología académica. *Procesamiento del Lenguaje Natural* 69. 95-103. <http://hdl.handle.net/10045/127394>.
- Aranzabe, María Jesús, Izaskun Aldezabal & Igone Zabala. 2022b. Recursos y Herramientas de Lingüística de Corpus y PLN para la Monitorización e Investigación de los Usos Académicos del Euskera. In *III. Workshop de INTELE (Infraestructura de Tecnologías del Lenguaje)*. <https://ixa.si.ehu.es/node/13591>.
- Arregi, Xabier, Ana Arruarte, Xabier Artola, Igone Zabala & Mikel Lersundi. 2013. TZOS: An On-Line System for Terminology Service. In *Actualizaciones en Comunicación Social. Actas XIII Simposio Internacional de Comunicación Social*, 400-404. Santiago de Cuba: Centro de Lingüística Aplicada. <http://ixa.si.ehu.es/node/3988>.
- Artetxe, Mikel, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre & Aitor Soroa. 2022. *Does Corpus Quality Really Matter for Low-Resource Languages?* <https://doi.org/10.48550/arXiv.2203.08111>.
- Azkarate, Miren. 2017. La terminología vasca en el siglo XXI. *Terminàlia* 12. 60-62. <http://revistes.iec.cat/index.php/Terminalia/article/view/139441>.
- Azkarate, Miren. 2020. Gramatiken auzi batzuk. *Euskera* 65(2), 257-281. <https://dialnet.unirioja.es/servlet/articulo?codigo=7994419>.
- Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang & Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. *Practical ML for Developing Countries Workshop*. Eighth International Conference on Learning Representations (ICLR). [https://pml4dc.github.io/iclr2020/papers/PML4DC2020\\_10.pdf](https://pml4dc.github.io/iclr2020/papers/PML4DC2020_10.pdf).
- De la Rosa, Javier, Eduardo G. Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero & María Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural* 68. 13-23. <https://doi.org/10.48550/arXiv.2207.06814>.

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <https://doi.org/10.48550/arXiv.1810.04805>.
- Euskaltzaindia. 1991. *Euskal Gramatika. Lehen Urratsak I (EGLU-I)*. Bilbo: Euskaltzaindia. [https://www.euskaltzaindia.eus/dok/iker\\_jagon\\_tegiak/6844.pdf](https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/6844.pdf).
- Euskaltzaindia. 1992. *Hitz-elkarketa / 4*. LEF batzordea. Bilbo: Euskaltzaindia. [https://www.euskaltzaindia.eus/dok/iker\\_jagon\\_tegiak/795.pdf](https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/795.pdf).
- Euskaltzaindia. 1994. *Eratorpenaz*. LEF batzordea. Bilbo: Euskaltzaindia. <https://www.euskaltzaindia.eus/dok/euskera/49996.pdf>.
- Euskaltzaindia. 1997. *Euskal Gramatika. Lehen Urratsak II (EGLU-II)*. Bilbo: Euskaltzaindia. [https://www.euskaltzaindia.eus/dok/iker\\_jagon\\_tegiak/24569.pdf](https://www.euskaltzaindia.eus/dok/iker_jagon_tegiak/24569.pdf).
- Euskaltzaindia. 2021. *Euskararen Gramatika*. I. liburukia. Bilbo: Euskaltzaindia. <https://worldcat.org/oclc/1350149936>.
- Ezeiza, Nerea. 2002. Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzailer morfositaktiko sendo eta malgua. Doktoretza-tesia. Donostia: UPV/EHU. <http://ixa.ehu.es/node/4123>.
- Fontenelle, Thierry, Geert Adriaens & Gert de Braekeleer. 1994. The Lexical Unit in the Metal® MT System. *Machine Translation* 9. 1-19. <https://doi.org/10.1007/BF00980197>.
- García-Salido, Marcos, Marcos García, Milka Villayandre & Margarita Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In Calzolari N. *et al.* (arg.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 260-265. Japan: European Language Resources Association. <https://aclanthology.org/L18-1039/>.
- Iruskieta, Mikel, Arantxa Otegi, Larraitx Uria, Arantza Díaz de Ilarraza & Amaia Artolazabal. 2018. Zer i(r)akas dezakegu geure corpusekin «jolastuz»? In Aintzane Etxebarria, Aitor Iglesias, Hiart Legarra & Asier Romero (arg.), *Traineru bete lagun: Iñaki Gaminde omenduz*, 35-66. UPV/EHU Argitalpen Zerbitzua. <http://ixa.ehu.es/node/12749>.
- Lample, Guillaume & Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. <https://doi.org/10.48550/arXiv.1901.07291>.
- Otegi, Arantxa, Aitor Agirre, Jon A. Campos, Aitor Soroa & Eneko Agirre. 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*, 436-442. <https://aclanthology.org/2020.lrec-1.55>.
- San Martin, Itziar. 2013. Terminologia sareak Ehunduz: unibertsitateko ikasgeletan erabiltzen den terminologia ikusgai egin nahi duen programa. In Xabier Alberdi & Pello Aalaburu (arg.), *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*. 20-32. Bilbo: UPV/EHU Argitalpen Zerbitzua. <https://www.ehu.es/documents/2430735/2730483/LIBURUAehuei13-03.pdf>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention Is All Your Need. *Advances in Neural Information Processing Systems* 30. 1-15. <https://dl.acm.org/doi/10.5555/3295222.3295349>.
- Villayandre, Milka 2018. «HARTA» de novels: un corpus de español académico. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos* 5(1). 131-140. <https://doi.org/10.15366/chimera2018.5.1.011>.



- Zabala, Igone, Axun Aierbe, Izaskun Aldezabal, María Jesús Aranzabe, Xabier Arregi, Jose Mari Arriola, Agurtzane Elordui, Antton Elozegi, Kristina Elozegi, Joseba Ezeiza, Julio Garcia, Ines Garcia, Mikel Lersundi, Itziar San Martin & Iñaki Ugarteburu. 2008. GARATERM: Diskurtso akademiko-profesionalaren didaktika eta garapena uztartzeko tresna informatikoen diseinu eta integrazioa helburu duen proiektua. In Pello Salaburu & Iñaki Ugarteburu (arg.), *Espezialitate Hizkerak eta Terminologia III: Espezialitate Hizkeren Didaktika eta Komunikazioa*, 211-219. Bilbo: UPV/EHU Argitalpen Zerbitzua. <http://ixa.ehu.eus/node/4020>.
- Zabala, Igone, Mikel Lersundi, Igor Leturia, Iker Manterola & Gotzon Santander. 2013. GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. In Xabier Alberdi & Pello Salaburu (arg.), *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*, 98-114. Bilbo: UPV/EHU Argitalpen Zerbitzua. <http://ixa.ehu.eus/node/3991>.

## A. eranskina. KATEGORIA-SISTEMA EDBLn

Kategoria Lexikalak (15)								
Kategoria nagusiak								
IZE	ARR		IZENAK	ARRUNTAK ( <i>zuhaitz</i> )				
	IZB			PERTSONA-IZEN BEREZIAK ( <i>Mikel</i> )				
	LIB			LEKU-IZEN BEREZIAK ( <i>Donostia</i> )				
	ZKI			ZENBAKIA ( <i>bat</i> )				
ADJ	ARR		ADJEKTIBOAK	ARRUNTAK ( <i>handi, benetako</i> )				
	GAL			GALDETZAILEAK ( <i>nongo</i> )				
ADI	SIN		ADITZAK	SINPLEAK ( <i>ekarri</i> )				
	ADK			KONPOSATUAK ( <i>lo egin</i> )				
	ADP			PERIFRASTIKOAK ( <i>ahal izan</i> )				
	FAK			FAKTITIBOAK ( <i>etorrarazi</i> )				
ADB	ARR		ADBERBIOAK	ARRUNTAK ( <i>gaur, negarrez</i> )				
	GAL			GALDETZAILEAK ( <i>noiz</i> )				
DET			DETERMINATZAILEAK	Erakusleak				
	ERK	ERKARR			ARRUNTAK ( <i>hau</i> )			
		ERKIND			INDARTUAK ( <i>berori</i> )			
	NOL			Nolakotzaileak				
		NOLARR			ARRUNTAK ( <i>edozein</i> )			
		NOLGAL			GALDETZAILEAK ( <i>zein</i> )			
	ZNB				Zenbatzaileak			
		DZH		BAN		Zehaztuak ( <i>bi</i> )	Banatzaileak ( <i>bina</i> )	
				ORD			Ordinalak ( <i>bigarren</i> )	
				DZG		Zehaztugabeak ( <i>zenbait</i> )		
		ORO		Orokorrak ( <i>guzti</i> )				
	IOR	PER			IZENORDAINAK	Pertsonalak		
PERARR			ARRUNTAK ( <i>ni</i> )					
PERIND			INDARTUAK ( <i>neu</i> )					
IZG			Zehaztugabeak					
		IZGMGB		MUGAGABEAK ( <i>norbait</i> )				
		IZGGAL	GALDETZAILEAK ( <i>nor</i> )					
BIH			Bihurkariak ( <i>-@en burua</i> )					
ELK		Elkarkariak ( <i>elkar</i> )						

Kategoria Lexikalak (15)				
LOT	LOK		LOTURAZKOAK	LOKAILUAK ( <i>hala ere</i> )
	JNT			JUNTAGAILUAK ( <i>edo</i> )
PRT			PARTIKULAK ( <i>omen, ote...</i> )	
ITJ			INTERJEKZIOAK ( <i>alajaina!</i> )	
BST			BESTELAKOAK ( <i>baldin</i> )	
Kategoria lagungarriak				
ADL			ADITZ LAGUNTZAILEAK ( <i>du</i> )	
ADT			ADITZ SINTETIKOAK ( <i>dator</i> )	
SIG			SIGLAK ( <i>EHU</i> )	
SNB			SINBOLOAK ( <i>km, cm, g...</i> )	
LAB			LABURDURAK ( <i>etab.</i> )	
Kategoria morfologikoak (9)				
AMM			ADITZ-MOTA MORFEMAK ( <i>-tu, t(z)e...</i> )	
ASP			ASPEKTU-MORFEMAK ( $\emptyset$ , <i>-ko...</i> )	
ATZ			ATZIZKIAK ( <i>-pe</i> )	
AUR			AURRIZKIAK ( <i>ber-</i> )	
DEK			DEKLINABIDE-MORFEMAK ( <i>-aren</i> )	
ELI			ELIPSIA ( $\emptyset$ )	
ERL			ERLAZIO ATZIZKIAK ( <i>(-e)la</i> )	
GRA			GRADUATAILEAK ( <i>-ago</i> )	
MAR			MARRA ( <i>-</i> )	
Puntuazio-zeinuak (3)				
PNT			PUNTUA	
BMP			BESTE PUNTUAZIO ZEINUAK ( <i>puntuaren pareko izan daitezkeenak</i> )	
PSB			PUNTUAZIO SINBOLOAK ( <i>parentesia, marra luzea, komatxoak...</i> )	

