Zientzia Sozialak eta Humanitate Digitalak gaur egun

CLARIAH-EUS

Begoña Altuna | Jon Alkorta | Xabier Arregi | Jose Mari Arriola
Ainara Estarrona | Aritz Farwell | Joseba Fernandez de Landa
Xabier Goenaga | Mikel Iruskieta

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

CIP. Unibertsitateko Biblioteka

CLARIAH-EUS [Recurso electrónico]: Zientzia Sozialak eta Humanitate Digitalak gaur egun / [editoreak, Begoña Altuna, ...(et al.)] – Datos. – [Leioa] : Universidad del País Vasco / Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua = Servicio Editorial, [2025]. – 1 recurso en línea : PDF (142 p.).

Modo de acceso: World Wide Web.

Trabajos presentados al CLARIAH-EUS 2. Whorshop (Donostia, 2023.)

ISBN: 978-84-9082-949-3

Euskara (lengua) – Escritura. 2. Ciencias sociales – Investigación. 3. Humanidades – Investigación. 4. Tecnología educativa. I. Altuna, Begoña, coed.

(0.034)809.169

Eskerrak eman nahi dizkiegu Eusko Jaurlaritzari eta bere Kultura eta Hizkuntza Politika Sailari, Gipuzkoako Foru Aldundiari, Euskal Herriko Unibertsitateko (UPV/EHU) Euskara, Kultura eta Nazioartekotze Errektoreordetzari, eta HiTZ zentroari haien laguntza eskuzabalagatik. San Telmo Museoari bereziki eskertu nahi diogu workshopa antolatzeko lankidetza.

ISBN: 978-84-9082-949-3

Aurkibidea

Ikerketa-taldeen eta proiektuen deskribapenak

Hitzaurrea

CLARIAH-EUS Humanitateetan eta Gizarte Zientzietan Europako ikerketa-azpiegiturekin lankidetzan euskara eta euskarari buruzko ikerketa bultzatzeko azpiegitura da. Hain zuzen ere, Europa mailako CLARIN eta DARIAH azpiegiturek eskaintzen dituzten ikerkuntzarako baliabideak eta laguntza euskara eta euskarari buruzko ikerketa egiteko eskuragarri jarri nahi ditu CLARIAH-EUSek. Horretarako, hainbat ekintzaren artean, ikertzaileen lanen berri izateko eta elkartrukea sustatzeko CLARIAH-EUS workshopak antolatzen ditugu.

Bigarren CLARIAH-EUS workshop honetan, gure helburua euskal ikerketa-komunitatea biltzea eta ikertzaileen eta bestelako eragileen arteko elkar ezagutza sustatzea izan zen, baita ikerketa-azpiegiturak eta -tresnak hobetzea eta euskal kulturaren eta hizkuntzaren ikerketa sustatzea ere. Workshopak euskal ikerketa-esparruko 50 aditu eta ikertzailetik gora batu zituen Donostiako San Telmo Museoan. Maria Cristina Marinescu (BSC-CNS) eta Gustavo Candela (Alacanteko Unibertsitatea) izan genituen hizlari gonbidatu eta 21 poster aurkeztu ziren. Arratsaldean CLARIAH-EUSen lehen batzar orokorrean azpiegituraren lehen urratsei buruz hitz egin genuen. Laburbilduz, CLARIAH-EUS azpiegiturak martxan jartzeko behar zituen osagaiak batu genituen: komunitatearen indarra, adituen jakinduria eta antolakuntzaren egitura.

Hemen bildutako ikerketa-ekarpenak CLARIAH-EUS azpiegituran jaso nahi diren baliabideen isla dira. CLARIAH-EUS komunitateak asko du eskaintzeko eta handia da baliabide horietatik atera daitekeen etekina. Ondoko bilduma CLARIAH-EUS komunitatea osatzen duten ikertzaileek sortutako eta prozesu itsu bikoitz baten bidez hautatutako lanek osatzen dute.

Azkenik, eskerrak eman nahi genizkieke parte hartzaile guztiei, haien ekarpenengatik eta CLARIAH-EUS azpiegituraren alde egiten duten lanagatik. Gure asmoa da CLARIAH-EUS workshopetan hasitako elkarlana eta ezagutzaren trukaketa etengabea izatea, eta hurrengo CLARIAH-EUS workshopetan elkarlan horren emaitzak ikustea espero dugu.

CLARIAH-EUSeko bulego teknikoa

Ikerketa-lanak

Corpusetik abiatutako idazketa programa baten sorkuntza

Developing a Writing Program Utilizing Corpus Analysis

Irune Ibarra[1], Mikel Iruskieta[2]

[1] Hezkuntza, Filosofia eta Antropologia Fakultatea, Euskal Herriko Unibertsitatea UPV/EHU
irune.ibarra@ehu.eus

[2] HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU
mikel.iruskieta@ehu.eus

Laburpena

Hezkuntzan eskuz idaztea garrantzitsua da oraindik ere, baina horren ohikoa den trebetasuna lantzeko baliabideak urriak dira edota intuizioan oinarrituta egiten dira. Lan honetan aurkezten dugun idazketa programaren helburu nagusia da eskuzko transkripzioa garatzea, corpusetako informazioan oinarriturik. Idazketa programa honetan esku-idazketa azkarra eta ortografia lantzen dira, azken hau modu inplizituan.

«Azkar idatzi eta ortografia onarekin» programa portugesez idatzitako programa arrakastatsu honetan oinarritzen da: «Clube dos Escritores: Escrevo depressa e sem erros!». Lehen Hezkuntzako 2. mailakoentzat moldatua badago ere, transkripzio geldia edota ortografia zailtasunak dituztenentzat aproposa izan daiteke. Berrikuntza bikoitza dakar: batetik, hizkuntza-corpuseko datutan oinarrituta dago eta bestetik, esku-idazketa azkarra eta ortografia uztartzen dituen programa da. Programak bi atal ditu: i) hezitzailearen gida eta ii) ariketa koadernoa. Programak 10 asteko iraupena du eta 15 minutuko saioetan antolatuta dago, eskolan talde handian nahiz bakarka lantzeko edo etxean lantzeko ere baliagarria da.

Gako hitzak: idazketa, corpusa, ortografia, idazketa programa.

Abstract

Handwriting remains a vital skill in education, yet resources for its practice are often limited or based on intuitive methods. This work introduces a novel writing program designed to enhance manual transcription through corpus-based information. The program, titled «Azkar idatzi eta ortografía onarekin [Quick Writing and Good Spelling]» focuses on fostering rapid handwriting and implicit spelling improvement. This program follows the successful Portuguese program «Clube dos Escritores: Escrevo depressa e sem erros!» adapted for 2nd grade of Primary Education, this program offers potential benefits for individuals struggling with slow transcription or spelling challenges. It introduces a dual innovation by leveraging linguistic corpus data and integrating quick handwriting and spelling exercises. Comprising a writing guide and an exercise notebook, the program has a duration of 10 weeks, with sessions of 15 minutes. Its flexible structure accommodates both classroom settings for large groups and individual or home-based learning environments.

Keywords: writing, corpus, spelling, writing program.

1. Sarrera

Eskolan eskuz idaztea garrantzitsua da oraindik orain, baina horren ohikoa den trebetasuna lantzeko baliabideak urriak dira edota intuizioan oinarrituta egiten dira. UNESCOren arabera (2017), idazketa trebea izatea gakoa da arrakasta lortzeko, bai eskoletan, bai eskolatik kanpo. Besteak beste, transkripzioa da idazketa trebearen funtsezko osagaietako bat eta transkripzioan esku-idazketa (jarioa eta ulergarritasuna) eta ortografia dira atal nagusiak. Atal horiek oso garrantzitsuak dira haurrak idazten hasten diren lehen urteetan idazketa modu egokian hasteko eta garatzeko.

Berningerrek eta Grahamek (1998) eta Berningerrek eta Amtmannek (2003) transkripzioak testu-sorkuntzan dituen eraginak azaleratu zituzten; izan ere, letrak nola idatzi edo ortografia zuzena zein den pentsatzen badabil idazle hasiberria, orduan eta arreta gutxiago jarriko die mezuaren edukiari edo idatzi nahi duen eta idazten ari den testuari. Karga Kognitiboaren teorian (Australiako Gobernua, 2017) ere ildo hori azpimarratzen da: giza garunak aldi berean prozesa dezakeen informazio berria mugatua da eta idazketaren irakaskuntza egituratzea komeni da. Bide horri jarraituz eta transkripzioaren eraginkortasuna erakutsi nahian, ebidentzietan oinarritutako estrategiak sortu dira (Limpo eta Graham, 2020) bai eta esku-idazketaren jarioa hobetzeko programak (bereziki ingelesez), baina momentuz euskaraz ez da bat bera ere sortu. Guk dakigula, programa gutxi daude nazioartean esku-idazketa jarioarekin egiteko eta ortografia modu inplizituan hobetzeko asmoa dutenak. Aurkitutako programa arrakastatsu bakanetarikoa corpus batetik sortuta dago (Soares *et al.,* 2014) eta euskarara moldatzea erabaki da, eskura dauden euskarazko corpusekin. Programa hori hauxe da: «Clube dos Escritores: Escrevo depressa e sem erros!» (Limpo eta Alves, 2020).

Artikulu honen helburua da Lehen Hezkuntzako 2. mailako ikasleen eskuzko transkripzioa lantzeko programa bat nola sortu den erakustea, corpusak eta nazioarteko transkripzio-irizpideak erabiliz. Programa horren izena «Azkar idatzi eta ortografia onarekin» da eta aipatutako Limpo eta Alvesen (2020) programaren egokitzapena da. Horretarako, lehenik transkripzioa hobetzeko eta handitzeko programak erakutsiko dira eta gero, corpusetatik ateratako emaitzak azalduko dira. Gero, aipaturiko programa nola moldatu den deskribatuko da. Azkenik, zenbait ondorio azaleratuko dira.

2. Aurreko programak

Eskuzko transkripzioaren osagaiak bi dira: batetik, esku-idazketa (letra-jarioa eta letraren ulergarritasuna) eta, bestetik, ortografia. Hainbat ikerketak konfirmatu dute eskuz minutuko letra ulergarria eta kopuru handia dutenek testu luzeak eta kalitatezkoak idazten dituztela, hots, korrelazioa dagoela letra-kopuru ulergarri handiaren eta idatzi den testu luzearen eta kalitatearen artean (Berninger *et al.,* 1992; Graham *et al.,* 1997; Alves *et al.,* 2012; Yan *et al.,* 2012). Euskaraz ere horixe gertatu da eta, gainera, genero asimetriak egon dira esku-idazketaren jarioan ikasturte bukaeran, bai Alfabetoaren proban, bai Kopiaren proban (Ibarra, 2016; Ibarra *et al.,* 2017). Hau da, batez besteko minutuko letra-kopuruak desberdinak izan dira neskenak eta mutilenak eta honek desberdintasun esanguratsuak sortu ditu testu-sorkuntzan: mutilek emaitza baxuagoak lortu dituzte. Lehen Hezkuntzako 2. mailan Alfabetoaren proban nesken batez bestekoa 15 letra/minutuko izan da eta mutilena 13,10 letra/minutuko. Horrez gain, neskek idatzitako testuak esanguratsuki hobeak izan dira. Era bertsuan, Kopiaren probari dagokionez, neskek 6 hitz ulergarri kopiatu zituzten azkar minutu batean eta testuak esanguratsuki hobeak izan dira, mutilek, ordea, 4 hitz kopiatu dituzte azkar minutu batean.

Era berean, Abbott *et al.*ek (2010) erakutsi zuten ortografia trebetasun kritikoa zela testua osatzerakoan. Garrantzitsua da pixkanaka ikasleak idazteko modu arautura gerturatzea, hots, Euskaltzaindiaren arau ortografikoetara. Euskaltzaindiaren arauak hainbat erabakitan oinarritzen badira ere: arau batzuk hitzen etimologiarekin lotuta daude; hau da, latinetik edo grezieratik, besteak beste, etorri bada hitza, nolabait errespetatu egiten da idazteko modua (*curriculum* hitza adibidez); beste arau batzuk euskalkien ahoskerarekin lotuta daude *(h*-aren erabilera iparraldean adibidez), etab. Antzaka *et al.*en (2018) arabera, euskaraz, fonema-grafema loturak, hau da, entzuten ditugun soinuen eta idazten ditugun letren arteko loturak, bat datoz gehienbat (ikus Euskara Batuaren Ahoskera Zaindua)[1] eta hurbilago daude ingelesarekin edo frantsesarekin alderatuz. Galuschka *et al.*en (2020) arabera, historikoak edo ohikoak dira fonema-grafema eta grafema-fonema korrespondentziak egitera bideratzen diren argitalpenak, esku-hartze fonikoak deitzen direnak. Egile horiek diotenez, fonema-grafema korrespondentziak ez dira nahikoak ortografia onarekin idazteko. Izan ere, korrespondentzia horiek euskarara ekarriz, fonema bat bi grafemekin idatzita egon daiteke (*tx* adibidez). Horregatik, Galuschka *et al.*ek (2020) eta Pujolek (2000) diotenez, nazioartean ezagutza ortografikoa edota ezagutza morfologikoa modu esplizituan erakusteko estrategiak erabiltzen dira.

Idazten hasten diren ikasleek ez dute zertan hizkuntza bateko letra guztiak idazten jakin behar, ezta ere letren arteko lotura guztiak egiten jakin behar. Hori horrela, ikasle batek idazten dakien letra-kopurua txikia izan daiteke; gainera, letra ez oso ulergarriak edota ortografia akatsak ugariak izan daitezke. Esku-idazketa edo ortografia (edota bi arloak) egon daitezke erasanda garapeneko disgrafia izanez gero. Gainera, esku-idazketan soilik ulergarritasuna edo soilik jarioa edo biak batera egon daitezke erasanda (Berninger, 2004). Egia da ere ortografiarekin zailtasunak izatea (adibidez, letrak edo grafemak gehitzea, omisioak egitea edo ordezkatzea) ohikoa dela ikasleetan eta bereziki agertzen direla dislexia duten ikasleetan (Galuschka *et al.,* 2020). Berningerrek (2004) dioenez, ortografia arazoak errazagoak dira konpontzen soilik ortografia arazoak egonda, ortografia gehi esku-idazketako arazoak egonda baino, eta presente eduki behar dira baita ideia horiek esku-hartzea egin aurretik. Aipatutako akats horiek identifikatu ahal izateko Ibarrak *et al.* ek (2021) txantiloi bat prestatu dute.

Nazioartean badira transkripzioaren atala den esku-idazketa hobetzeko curriculumean oinarritutako esku-hartze programa eraginkorrak, bereziki 2011tik aurrera sortu direnak (Engel *et al.,* 2018). Egile horiek diotenez, curriculumean oinarrituta egoteak esan nahi du programa horiek eskoletan irakasten direla eta haur guztien esku-idazketa hobetzera zuzenduta daudela; alegia, ez direla soilik zailtasunak dituztenentzako programak. Programa horietako ideia nagusia hauxe da: esku-idazketa curriculumarekin lerrokatzeak trebezien orokortze handiagoa sustatzen duela. Honela, esku-idazketa lantzeko Haur Hezkuntzako bigarren ziklotik Lehen Hezkuntzako 2. mailara 13 programa aztertu dituzte aipaturiko egileek. Programa horiek azkartasuna handitzeko (9 programa) edo letra ulergarriak egiteko (12 programa) helburua dute. Badirudi identifikatu behar dela gela jakin baten ulergarritasunari ala azkartasunari eragiten dion.

Engel *et al.*en (2018) aipatzen dituen esku-idazketako hiru programak artikulu hauekin daude lotuta: *The effectiveness of the size matters handwriting program* (Moskowitz, 2017), *Eficacy of an explicit handwriting program* (Kaiser *et al.,* 2011) eta *Handwriting club* (Howe *et al.,* 2013). Bestalde, Size *matters handwriting program* delako programari dagokionez,[2] esku-idazketaren ulergarritasuna da helburu nagusia. Lehen Hezkuntzako 1. eta 2. mailarako esku-hartzea da. Instrukzio zuzena, mnemotekniak, motibazio pizgarriak, gurasoen parte-hartzea, seinale bisualak, autokritika eta autokontrola lantzen ditu. Aipatutako programa horietan guztietan, esku-hartzeak izan zuen

[1] https://www.euskaltzaindia.eus/dok/arauak/Araua_0087.pdf

[2] Ikus webgunea: https://realotsolutions.com/blogs/news/welcome-size-matters-handwriting-program

eraginik handiena (0,80ko Hedges'en *g* neurria). Bestalde, *Explicit handwriting program* esku-hartzeari dagokionez, esku-idazketaren ulergarritasunean eta azkartasunean zentratzen da. Lehen Hezkuntzako 1. mailako ikasleentzako programa da eta 6 aste irauten du. Bertan lantzen diren jarduerak honelakoxeak dira: atzamarren trebezia, idazkera kurtsiboa, zeregin metakognitiboa, eskuz idaztearen praktika eta taldean aipatzea. Talde esperimentalak kontrol taldeak baino emaitza hobeak izan ditu, bai azkartasunean, bai irakurgarritasunean. Azkenik, *Handwriting club* esku-hartzeari dagokionez, irakurgarritasunean eta abiaduran zentratzen da. 12 saio dira, bakoitza 40-45 minutukoak. Eskolan klubak egiten dira. Emaitzarik onenak irakurgarritasunean lortu dira, aldiz, abiadurari buruzko emaitzak ez dira izan esanguratsuak.

Era berean, badira ortografia programen berri ematen dituzten lanak (Galuschka *et al.,* 2020). Egile horiek, ortografiaren tratamendu edo ikuspegi ezberdinen eraginkortasuna kuantifikatzen dute, 28 ikerketa kontuan hartuta: ortografia gabeziak erakusten dituzten ikasleengan eta dislexia-dunengan. Lan horretako emaitzen arabera osagai eraginkorrena dira: esku-hartze fonikoak, ortografikoak eta morfologikoak.[3] Halaber, baieztatzen dutenez, esku-hartze morfologikoek ezgaitasun desberdinak dituzten haurrengan eragin positiboak dituzte (adierazpen hizkuntzaren arazo espezifikoa dutenengan, irakurketa gabezia gehigarriak dituztenengan, etab.).

Galuschka *et al.* (2020) meta-analisiaren egileek diotenez, esku-hartze ortografikoen ezarpen goiztiarrari buruzko lanetan ikerketa gutxi dagoela adierazten dute. Gainera, ikasgelako esku-hartzeak ez dira oso espezifikoak izaten eta ezinbestekoa egiten da esku-hartzeak talde txikietan egitea. Meta-analisi horretan ortografia arauak erakusten dituzten programak 7 dira. Horietatik LHko 2. mailako ikasleei zuzendutako programak ondoko egileek sorturikoak dira: Schulte-Körne *et al.* (2001), Darch *et al.* (2006), Ehri *et al.* (2009) eta Kirk eta Gillon (2009).

Hori horrela, gure argitalpenak Soares *et al.*en (2014) proposaturiko metodologia jarraitzen du eta corpusetan oinarritzen da. Limporen eta Alvesen (2020) programaren egituran oinarrituz sortu dugu euskarazko lehen programa.

3. Corpusak eta idazketa programak

Testu-corpusak edo corpusak egituratutako eta hizkuntza erabilera errealeko testu-bilduma digitalak dira. Egiazko hizkuntza adibideak biltzen dituztelako garrantzitsu bihurtu dira hizkuntzaren erabilera aztertzeko eta hizkuntzalaritzan. Corpusak hizkuntza-prozesatzeko tresna ezberdinak sortzeko ere baliagarriak izan dira: adimen artifizialean oinarritzen diren zerbitzuak, zuzentzaile ortografikoak, itzulpen automatikoa eta elkarrizketa-sistemak, besteak beste. Modu askotako corpusak daude, adibidez, ahoz esandakoak bildu daitezke edo testu idatzietan oinarritutakoak izan daitezke. Corpus bakoitza helburu jakin batekin egindakoa izan daiteke (Salaburu, 2024).

Corpusetan ageri diren hitzak eta maiztasun handieneko silabak garrantzitsuak izan daitezke transkripzioa (esku-idazketarena edota ortografiarena) lantzeko, detekzio probak edo atazak egiteko eta idazketa programak sortzeko. Horretarako badira zenbait corpus interesgarri, ikasle gaztetxoentzat egokitutakoak. Bata, Euskal Herriko Ikastolen Elkartearen, Haur Hezkuntzako ipuin-bilduma,[4] 19.917 hitz inguru ditu. Corpus horretan Haur Hezkuntzako umeei kontatutako

[3] Ikus hemen 28 ikerketa horiek https://www-tandfonline-com.ehu.idm.oclc.org/action/downloadSupplement?doi=10.1080%2F00461520.2019.1659794&file=hedp_a_1659794_sm5371.pdf

[4] Bilatzailea: http://ixa2.si.ehu.eus/clarink/corpusak/ipuinak eta corpusa http://hdl.handle.net/11304/f27f5e92-af01-4a37-a6d9-82cf14afa160

ipuinen testu hutsak bildu dira. Bestea, Txikipediakoa.[5] Corpus horrek 260.000 hitz inguru biltzen ditu eta 8-13 urte bitarteko umeentzako euskarazko entziklopedia txiki eta askea da. Erabiltzen den hizkuntza ere adin tarte horretako umeentzako egokitua izan da.

Baliabide horiek erabilita Ibarra *et al.*ek (2020) letren loturen konplexutasuna kontuan hartuz eta Kandel *et al.*en (2011) ikerketak aintzat hartuz, letrak lotzen erakusteko gida proposatu dute. Argitalpenak bi silabako edo gehiagoko hitzak erabiltzen ditu letren arteko elkarketak lantzeko eta erabiltzen dituen hitzak corpusetatik atera dira, maiztasunak kontuan izanik. Txikipediaz gain Vikidia corpusa[6] ere erabili da, gaztelaniazko hitzak bertan sartzeko (azken hau 500.000 hitz ingurukoa).

1. taula. Kopiaren proba egiteko hitz zerrenda, maiztasunak eta EGWAko hitzak gaztelaniaz

Moldatutako hitzak	Zenbat aldiz corpusean	EGWA probako hitzak
ate	45	*niña*
lagun	56	*marido*
mendi	31	*casa*
egun	52	*luna*
alde	20	*codo*
otso	45	*foca*
txantxa	4	*dedo*
baserri	10	*camino*
kopeta	8	*tejado*
sekretu	21	*cometa*
herritar	10	*verano*
gaizto	14	*salado*

Horrez gain, Haur Hezkuntzako ipuin-bildumatik, aipatutako ipuin-bildumatik, hiru proba edo ataza sortu dira, oraindik baliozkotu gabe daudenak (Ibarra, *et al.*, 2022).[7] Transkripzioa ebaluatzeko proba asko diktaketa edo kopia bidezkoak izaten direnez, bi proba sortu dira diktaketa bidez egiteko eta bat kopia bidez egiteko. Honela, proba bat, ortografiaren barruan kokatzen den hitzen segmentazioa aztertzeko egin da eta esaldi bat diktatzen zaio haurrari silaba maiztasunean oinarrituta: 'Etxetik etorri zen' ('txe' silabaren corpuseko maiztasuna 1237, 'tik' 1298, 'rri' 3161 eta 'zen' 2260). Bigarren proba, silaba maiztasun handiko sasi-hitzak edo existitzen ez diren hitzak idaztearen proba da, hau da, ortografiaren bide fonologikoa jarraituz nola idazten den ikusteko. 2 sasi-hitz aukeratu ziren: 'mentik' ('men' 815 aldiz agertzen da corpusean eta 'tik' 1298) eta 'kintzen' ('kin' 1073 aldiz eta ''tzen' 1479 aldiz). Sasi-hitzak ezagutzen ez diren hitzak idazteko eta entzuten diren soinuak nola idazten diren jakiteko baliabide interesgarriak dira. Hirugarren proba, haurrek hitzak nola kopiatzen dituzten ikusteko proba da, era berean letren arteko lotura motak argi ikusteko balio du. Hitzak aukeratu ondoren (ikus 1. taula), EGWA (*Early Grade Writing Assessment*) probaren (Jiménez, 2018) 3. zeregina moldatu da, hots, hitzak kopiatzea. Bi eta hiru silabakoak aukeratu dira, silaben maiztasuna kontuan hartuta eta hitzen silaben nolakotasuna kontuan izanik (silaba zuzena, trabatua, e.a.). Minutu batez hitz horiek kopiatzea eskatzen da eta,

[5] Euskarazko Txikipediako azpicorpuseko bilatzailea: http://ixa2.si.ehu.eus/clarink/corpusak/txikipedia/

[6] Gaztelaniazko Vikidia: http://ixa2.si.ehu.eus/clarink/corpusak/vikidia/

[7] Ikus CLARIAH-EUSen lehen topaketetan aurkeztutako posterra: https://www.clariah.eus/sites/default/files/posterrak/POSTERRA-Ibarra%2C%20Iruskieta%20eta%20Mart%C3<%ADnez-Arbelaiz.pdf

ondoren, letra kopuru ulergarria zenbatzen da: *ate* eta *lagun* (zer egin behar duten erakusteko adibideak dira eta idatzizko beroketa edo frogak egiteko hitzak). Ondoren, kronometroa martxan jartzen da eta hitz hauek kopiatzea eskatzen da: *mendi, egun, alde, otso, txantxa, baserri, kopeta, sekretu, herritar, gaizto* (ikus 1. taula).

4. Metodologia

Idazketa programa egiteko metodologiari dagokionez, honako urratsak jarraitu dira:[8]

4.1. Ortografia arauen aukeraketa

Lehen urratsari dagokionez, zenbait argitaletxek erabilitako ortografia arauak aukeratu dira. Honela, Lehen Hezkuntzako 2. mailarako zenbait argitaletxeren lanetan oinarritu gara: Ikasmina (SM taldea), Anaya/Haritza, Elkar, Ibaizabal eta Erein argitaletxeak hain zuzen ere. Ortografia arau horiek, portugesez egindako programaren proposamenarekin konparatu ostean, honako arau ortografikoak lantzea erabaki dugu.

— 1. astean, soinu berbera duten 'r' letradunak lantzea erabaki da. Hitz batzuk 'rr' idazten dira eta beste batzuk 'r' bakarrarekin.

— 2. astean, 'nt', 'nd', 'np' eta 'nb'.

— 3. astean, 'h' hasieran daramaten zenbakiak jartzea erabaki da (*ehun* salbuespena da). Gainera, bokalez hasten diren hitzak baina 'h'rik ez daramatenak erabiltzea pentsatu da.

— 4. astean, 'in' hitzaren erdian duten eta 'ñ' hitzaren erdian duten hitzak jartzea erabaki da.

— 5. astean, errepasoa egiten da, aurreko lau asteetako arauak erabiliz.

— 6. astean 'ts', 'tx' eta 'tz' daramaten hitzak lantzea erabaki da.

— 7. astean: 'z' eta 's' duten hitzak jartzea erabaki da.

— 8. astean, gidoia duten hitz-elkartuak eta gidoirik ez duten hitz bikoteak jarri dira.

— 9. astean, 'ge', 'gi', 'je' eta 'ji' letrekin idazten diren hitzak aukeratu dira.

— 10. astean errepasoa egiten da. Aurreko lau asteetako (6-9) arauak eta hitzak erabili dira.

4.2. Corpusa eta bilaketa sistema

Bigarren urratsari dagokionez, haurrentzat egokia izan daitekeen corpusa erabili da, ohikoenak diren hitzak zehazteko, LHko ikasleentzat egokituriko Txikipediako testuetan duten maiztasunaren arabera. Txikipediako testuetan bilaketak egiteko IXA-CLARIN-K azpiegiturari laguntza eskatu eta azpiegitura horretan sortutako web interfazea erabili da.[9]

---

[8] Poster CLARIAH-EUS 2021: https://www.clariah.eus/sites/default/files/posterrak/POSTERRA-Ibarra%2C%20Iruskieta%20eta%20Mart%C3<%ADnez-Arbelaiz.pdf

[9] http://ixa2.si.ehu.es/clarink/corpusak/txikipedia helbidean kontsulta daiteke.

4.3. Hitzen aukeraketa

Hirugarren urratsari dagokionez, karaktere-konbinazio zehatzak bilatu dira corpus horretan, esaterako, 'np', 'nb', 'nd' eta 'nt' karaktere-konbinazioak dituzten hitzak. Arauak jarraituz corpuseko hitz hauek aukeratu dira:

1. olinpiar	7. denbora	3. mundu	8. inperio	5. lagundu
4. elementu	2. inprimatu	9. zerrenda	6. txantiloia	10. Kolonbia

4.4. Esaldien aukeraketa

Laugarren urratsari dagokionez, karaktere-konbinazio zehatzak bilatu dira corpus horretan eta, ondoren, aurkitutako hitz (edo lema) bakoitza interfazean behatu da eta interfazeak hitz horrekin corpuseko esaldi guztiak eskaini ditu (1. irudia):

	Dokumentua	Sent Id	Hitza(k)	Esaldia
1	Ainhoa_Murua.txt.lemlnk.xml	sent5	Olinpiar	Olinpiar Jokoetan lau aldiz hartu du parte :
2	Bartzelona.txt.lemlnk.xml	sent16	Olinpiar	1992ko Udako Olinpiar Jokoak Bartzelonan izan ziren .
3	Hipika.txt.lemlnk.xml	sent2	Olimpiar	Olimpiar joko modernoetan ia hasieratik hipika egon da (zehazki bigarrenetatik) .
4	Igeriketa.txt.lemlnk.xml	sent8	Olinpiar	Baina igerilari onenek eta Olinpiar jokoetan parte hartzen dutenek lau estilotan ibilt
5	Los_Angeles.txt.lemlnk.xml	sent30	Olinpiar	1984an , Udako Olinpiar Jokoak bigarren aldiz antolatu ziren Los Angelesen .
6	Maialen_Chourraut.txt.lemlnk.xml	sent8	Olinpiar	2009ko Munduko Txapelketan bigarren eta 2011koan hirugarren gelditu ondoren , 2 Jokoak bere arrakasta handiena izan ziren , brontzezko domina lortuz .

1. irudia. Txikipediako datuekin sorturiko datu-baseko bilaketaren emaitza

4.5. Esaldien egokitzapena

Bosgarren urratsari dagokionez, corpuseko esaldi guztiak lortuta, sintagma edo perpaus egokienak hautatu dira eta esaldi horiek egokitu egin dira; idazketa ariketetarako baliagarriak izateko, batzuetan luzeegiak izanik informazio osagarria kendu da. Sintagma egokirik aurkitu ez den zenbait kasutan adibideak asmatu dira, betiere maiztasun handieneko hitzetatik abiatuta.

4.6. Programa sortzea

Seigarren eta azken urratsari dagokionez, jarraibideak eta esaldiak portugesezko txantiloian sartu dira eta euskarazko programa sortu da.

5. Emaitzak

Lan honetako emaitzei dagokionez, 'Azkar idatzi eta ortografia onarekin' (ikus 2. irudia) bi atal dituen programa sortu:[10] a) Hezitzailearen gida eta b) Ariketa koadernoa. Hezitzailearen gidan programa martxan jartzeko irizpideak eta jarraipenerako kontrol-zerrendak proposatu dira. Bestetik, Ariketa koadernoan ikasleentzako 10 astetan zehar egiteko saioak bildu dira. Aste bakoitzerako 3 saio daude, bakoitza 15 minutuz lantzeko. Saioetako bi, gela barruan lantzeko dira eta beste bat etxean. Saioetan, bi motatako jarduerak daude: i) alfabetoarekin edo ortografiarekin zerikusia duten jarduerak eta ii) hitzen edo esaldien kopia azkarreko jarduerak. Orobat, 5. eta 10. asteko saioak errepasorako diseinatu dira.

2. irudia. 'Azkar idatzi eta ortografia onarekin' programa: https://www.booktegi.eus/liburuak/azkar-idatzi

6. Ondorioak eta etorkizuneko lana

Programa honen berrikuntza bikoitza da: batetik, corpusetako datuetan oinarrituta dago eta, bestetik, nazioarteko irizpideak kontuan hartuta eskuzko transkripzioaren azkartasuna eta ortografia lantzeko proposamena da.

[10] https://www.booktegi.eus/liburuak/azkar-idatzi

LHko 2. mailan programa hau erabili ondoren, haurrek lortu duten eskuzko transkripzioaren egoera hobetu den aztertu beharko litzateke, programan bertan dauden probetan oinarrituta, haurraren esku-idazketaren jarioa zein den jakiteko. Jario horren neurrirako, bi erakusle ditugu: i) alfabetoa buruz minutu batean argi eta azkar idaztea eta ii) kopia minutu batean argi eta azkar egitea. Horrela, alfabetoari dagokionez, erabili daiteke programaren azkeneko asteko 'Alfabeto lasterketa', zenbat letra dakizkien buruz eta ordenan ebaluatzeko. Kopiari dagokionez, erabili daiteke programa bereko azkeneko asteko 'Esprint' ariketaren bukaerako esaldia: 'Min hartu duzu?', ikasleak minutuko zenbat letra/hitz kopiatu dituen jakiteko. Kopia ebaluatzeko erabili daiteke baita minutu bateko proba (1. taula) (Ibarra *et al.,* 2022). Ikaslea ikasturte bukaeran kokatzeko eta ikasle horri zein esku-hartze proposatu behar zaion jakiteko; horretarako, baliatu ditzakegu Ibarra (2016) eta Ibarra *et al.*en (2017) lanak. Honela, ikasturte bukaerako 'Alfabeto lasterketa' proban minutuko 15 letrara hurbilduz eta esku-hartzea 14 letra ulergarri edo gutxiago egin dituztenen egoera hobetuz, inor atzean gelditu ez dadin.[11] Kopiari dagokionez, ikasturte bukaerarako minutuko 6 hitz kopiatzera hurbiltzea komeni da, hau da, 1. taulan agertzen den 'baserri' hitzera arte edo 'Min hartu duzu?' esaldia bi aldiz idatzi arte. Beraz, proposatu dugun programa baliatu beharko litzateke minutu bateko proban 5 hitz ulergarri edo gutxiago kopiatu duten ikasleekin.

Programa hori erabili ondoren, haurrak duen transkripzioaren beste osagaia, ortografiarena, aztertu beharko litzateke. Batetik, ikusi behar da inplizituki landuta lortu diren hitzen arauak beste hitz batzuetan aplikatzen badira eta transferentzia positiboa izan bada; hau da, *hamar* hitza landu bada, ikusi behar da *hamabi h* letrarekin idazten den. Bestetik, ortografiako arau zailenak esplizituki erakutsi daitezke, estrategia ezberdinak erabiliz. Modu berean jokatu daiteke erabiltzen diren 'ts' digrafoarekin edo 'tt' digrafoarekin. Halako estrategiak erabiliz, haurrak zentzua hartzeko aukera izango du ortografia landuz. Beste estrategia bat izan daiteke modu esplizituan gaztelaniarekin konparatzea adibidez 'np/nb araua', ezberdintasunak nabarmenduz. Horrez gain, hitzen estruktura morfologikoa ezagutzeak ortografia errazten du, Pujol (2000) eta Galuscka *et al.*ek (2020) adierazi bezala. Hortaz, estrategia garrantzitsua izan daiteke ere lexema berdinetik sortutakoak erabiltzea (*itsas, itsas-izar, itsasgizon, itsasertz*, e.a.) eta horretan oinarrituriko ariketak prestatzeko corpusak baliagarriak dira.

Bukatzeko, etorkizunean komenigarria litzateke corpusetan oinarrituriko datuak baliatuz eskuzko transkripzio-erroreak detektatzeko probak sortzea.

Eskertza

Lan hau egiteko 'Etorkizuna Eraikiz' (17. proiektua 24-25) Gipuzkoako Foru Aldundiko dirulaguntza izan dugu. Horrez gain, IXA-CLARIN-K azpiegiturari ere eskerrak eman nahi dizkiogu.

Bibliografia

Abbott, R., Berninger, D., Virginia W. eta Fayol, Michel. (2010): Longitudinal relationships of levels of language in writing and between writing and reading in grades 1-7, *Journal of Educational Psychology*, 102, 281-298.

[11] Erabili daiteke 'Letra xeheak bai!' doako deskarga duen argitalpena (Ibarra *et al.,* 2020). Esku-idazketaren jarioa lortzen joateko lan horretako 23tik 29rako orriak aproposak izan daitezke.

Antzaka, A., Martin, C., Caffarra, S., Schlöffel, S., Carreiras, M. eta Lallier, M. (2018). The effect of orthographic depth on letter string processing: the case of visual attention span and rapid automatized naming. *Reading and Writing*, 31, 583-605.

Alves, R. A., Branco, M., Castro, S. L. eta Olive, T. (2012). Effects of handwriting skill, output modes and gender of fourth graders on pauses, written language bursts, fluency and quality. In V. W. Berninger (ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (389-402). New York: Psychology Press.

Australiako gobernua (2017). Cognitive load theory: Research that teachers really need to understand. https://education.nsw.gov.au/content/dam/main-education/about-us/educational-data/cese/2017-cognitive-load-theory.pdf

Berninger, V. W. (2004). Understanding the «Graphia» in Developmental Dysgraphia: A Developmental Neuropsychological Perspective for Disorders in Producing Written Language. In D. Dewey & D. E. Tupper (Eds.), *Developmental motor disorders: A neuropsychological perspective* (pp. 328-350). The Guilford Press.

Berninger, V.W., Yates, C., Cartwright, A., Rutberg, J., Remy, E. eta Abbott, R. (1992). Lower-level developmental skills in beginning writing, Reading and Writing: An *Interdisciplinary Journal*, 4, 257-280.

Berninger, V.W., eta Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In *Handbook of Learning Disabilities*, Swanson, H.L., Harris, K.R., and Graham, S. (Eds.) (pp. 345-363). New York: Guilford Press.

Berninger, V.W. eta Graham, S. (1998). Language by hand: A synthesis of a decade of research on handwriting, *Handwriting Review,* 12, 11-25.

Berninger, V., Abbott, R., Rogan, L., Reed, E., Abbott, S., Brooks, A., ... & Graham, S. (1998). Teaching spelling to children with specific learning disabilities: The mind's ear and eye beat the computer or pencil. *Learning disability quarterly*, *21*(2), 106-122.

Ehri, L. C., Satlow, E., & Gaskins, I. (2009). Grapho-phonemic enrichment strengthens keyword analogy instruction for struggling young readers. *Reading & Writing Quarterly*, *25*(2-3), 162-191.

Engel, C., Lillie, K., Zurawski, S. eta Travers, B. G. (2018). Curriculum-based handwriting programs: A systematic review with effect sizes. *American Journal of Occupational Therapy*, 72(3), 7203205010p1-7203205010p8.

Galuschka, K., Görgen, R., Kalmar, J., Haberstroh, S., X. eta Schulte-Körne, G. (2020) Effectiveness of spelling interventions for learners with dyslexia: A meta-analysis and systematic review. *Educational Psychologist*, 55:1, 1-20, DOI: 10.1080/00461520.2019.1659794

Graham, S., Berninger, V.W., Abbott, R., Abbott, S. eta Whitaker, D. (1997). The role of mechanics in composing of elementary school students: A new methodological approach, *Journal of Educational Psychology,* 89, 170-182.

Graham, S. (2009-2010). Want to improve children's writing? *American Educator*, 33, 20-40.

Howe, T. H., Roston, K. L., Sheu, C. F., & Hinojosa, J. (2013). Assessing handwriting intervention effectiveness in elementary school students: A two-group controlled study. *The American Journal of Occupational Therapy*, 67(1), 19-26.

Ibarra, I. (2016). Esku-idazketa eta testu eleanitzen arteko loturak. Doktoregotesia. Eskuragarri: https://dialnet.unirioja.es/servlet/articulo?codigo=6187245

Ibarra, I., Atutxa, U. eta Iruskieta, M. (2021). Idazteko zailtasunak detektatzeko eta esku hartzeko proposamena Lehen Hezkuntzan. IKASTORRATZA. e-Revista de Didáctica, 27, 1-29. DOI: 10.37261/27_alea/1

Ibarra, I. Iruskieta M. & Martínez-Arbelaiz, A (2022). Hizkuntzen corpusen bidez 2. mailako ikasleen transkripzioa ebaluatzeko proben euskaratzea. «Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen jardunaldia». http://ixa2.si.ehu.eus/clariah-eus/node/17

Ibarra, I., Etxague, X. eta Etxeberria, J. (2017). Letren abiadura Lehen Hezkuntzan: euskarazko batez bestekoak ikasmailaren eta generoaren arabera, *Gogoa*, 16, 3-23. doi: 10.1387/gogoa.17910

Ibarra, I.; Ortube, M. eta Iruskieta, M. (2020). Loturak landuz: idazketa errazeko programa. Hemendik berreskuratua https://www.booktegi.eus/liburuak/loturak-landuz/

Ibarra, I., Ortube, M., eta Muxika, I. (2020). Letra xeheak, bai! Booktegi. Hemendik berreskuratua https://www.booktegi.eus/liburuak/letra-xeheakbai/

Jiménez, J. E. (2018). *Early Grade Writing Assessment: a report on development of an instrument*. Unesco.

Kaiser, M. L., Albaret, J. M., & Doudin, P. A. (2011). Efficacy of an explicit handwriting program. *Perceptual and Motor Skills*, *112*(2), 610-618.

Kandel, S., Peereman, R., Grosjacques, G. eta Fayol, M. (2011). For a psycholinguistic model of handwriting production: Testing the syllable-bigram controversy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1310-1322. 10.1037/a0023094

Kirk, C., & Gillon, G. T. (2009). Integrated morphological awareness intervention as a tool for improving literacy. *Language, Speech, and Hearing Services in Schools*, 40(3),341-351.

Moskowitz, B., Carswell, B., Kitzmiller, J., Bushell, M., Neikrug, L., Gottesman, C., & Murray, T. (2017). The effectiveness of the size matters handwriting program. *The American Journal of Occupational Therapy*, *71*(4_Supplement_1), 7111520304p1-7111520304p1.

Limpo, T. eta Alves, R. (2020). Clube dos escritores: Escrevo depressa e sem erros!. University of Porto. Porto, Portugal

Limpo, T. eta Graham, S. (2020). The role of handwriting instruction in writer's education. *British Journal of Educational Studies. 68*(3), 311-329.

Pujol, M. (2000). https://uvadoc.uva.es/handle/10324/8821

Salaburu, P. (2024). «Corpusak eta hiztegiak», *Sareko Euskal Gramatika (SEG)*, www.ehu.eus/seg. ISBN: 978-84-693-9891-3

Schulte-Körne, G., Deimel, W., Hülsmann, J., Seidler, T., & Remschmidt, H. (2001). Das Marburger Rechtschreib-Training–Ergebnisse einer Kurzzeit-Intervention. *Zeitschrift Für Kinder-und Jugendpsychiatrie und Psychotherapie*, 29(1), 7-15.

Soares A. P., Medeiros J. C., Simões A., Machado J., Costa A., Iriarte A., Almeida, J.J., Pinheiro, A.P., Comesaña M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese Elementary to Middle School textbooks. *Behavior Research Methods*, 46, 240-253.

UNESCO, E. S. (2017). Reading the past, writing the future: Fifty years of promoting Literacy. https://unesdoc.unesco.org/ark:/48223/pf0000247563

Yan, C. M., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A.M. eta Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing*, 25, 1499-1521.

Eranskina. Booktegiko argitalpeneko hitzak

1. astea	r		rr	
	Sortu	Serbia	Orri	Inurriak
	Urdin	Iturburu	Barruan	Erresuma
	Artikulu	Bakarka	Orokorra	Txakurra
	Urdaila	Arpa	Urtarrila	Sagarra
	Karga	Aurten	Zerrenda	Arraina
2. astea	np- nb		nt-nd	
	Kanpo	ezinbesteko	Kontu	Mundua
	Inprimatu	Kolonbia	Indentifikatu	Abendua
	Olinpiar	denbora	Zerrenda	Internazional
	Inperio	zenbait	Elementu	Handitu
	enpresa	hainbat	txantiloila	lagundu
3.astea	h		h ez	
	Hiru	Ehun	Adierazi	Aske
	Hogei	Hamar	Aipatu	Arteko
	Hamazazpi	Hamaika	Atal	Aurkitu
	Hamabi	Hogeita bost	Azken	Aditu
	Hamahiru	hamasei	Aurrean	aldaketa
4.astea	in		ñ	
	Ekainaren	Ezina	Ikurriña	Ariño
	Espainiako	Gainean	Iruñea	Iñigo
	Baina	Laino	Beñat	Ñimiño
	Mina	Inor	Andereño	Txanpiñoi
	Oinez	Zotina	Goñi	Piraña
5. astea – Errepasoa				
6. astea	ts		tx	
	Jaitsi	Itsatsi	Txinera	Kontxako
	Pertsona	Frantses	Txiki	Txirrindulari
	Otsaila	Bertsio	Txertatzea	Altxatu
	Unibertsitatera	Otsoa	Martxoaren	Txerriak
	Itsasoa	Amets	Fitxategia	Litxarreriak
7.astea	s		z	
	Beste	Ausaz	Berezi	Uztailaren
	Saiatu	Hostoak	Izan	Idazlea
	Estatu	Serbia	Azken	Guztiak
	Hasi	Deus	Buruz	Eztabaida
	Sinple	Salto	Azala	Zabalik
8. astea	(-)		(-) ez	
	Seme- alaba	Poliki- poliki	Hitz egin	Diru eske
	Eguzki- sistema	Eduki- taula	Etxez etxe	Itsasgizon
	Anai- arreba	Erabilera- baldintzak	Hasiberriak	Aurreikusi
	e-mail	Urdin- urdina	Bidasoa ibaia	Sinestezina
	dela-eta	Hizkuntza- eskola	Euskal idazlea	Egongela
9. astea	ge-gi		je-ji	
	Jorge	Egutegia	Jende	Erlijio
	Errege	Teknologi	Objektu	Kalejira
	Geometria	Liburutegira	Garajean	Mujika
	Agenda	Pedagogia	Makillaje	Emoji
	Gepardoa	Musikagile	Jeltzale	imajinazio
10. astea – Errepasoa				

Eskolako laburpen-testuak biltzeko baliabideak eta euskarazko laburpenen-corpusa

Resources for the collection of school summary texts and corpus of summaries in Basque

Unai Atutxa-Barrenetxea[1], Mikel Iruskieta[2], Olatz Ansa[2]

[1] HiTZ Zentroa
unai.atutxa@ehu.eus

[2] Ixa, Euskal Herriko Unibertsitatea UPV/EHU
mikel.iruskieta@ehu.eus olatz.ansa@ehu.eus

Laburpena

Tesi-lanetan egiten diren datu-bilketak eta datu horiek aztertzeko sortzen diren tresnak sakabanatuta egoten dira unibertsitateetako biltegietan. Dispertsio horrek oztopatu egiten du datu horien edo tresnen berrerabilera. CLARIAH-EUS bezalako ikerketa-azpiegitura digitalen helburuetako bat da tesietan jasotako datuak erraz bilatzeko tresnak sortzea eta beste ikertzaileentzat ikusgarri eta berrerabilgarri jartzea. Esaterako, CLARIN-ERIC azpiegiturak, *Virtual Language Observatory* (VLO) tresnarekin, biltegi ezberdinetan dauden datuak eta tresnak bilatzeko modua ematen du. Lan honetan, eskolako laburpen-testuen corpusa bildu dugu, eta datuak EuDatera igo ditugu; ondoren, CLARINen VLOn eskuragarri jarri ditugu, beste ikertzaileek erabil ditzaten, CC BY-NC 4.0 lizentziapean. Bestalde, laburpenak lortzeko tresnak eta laburpenak eskolan lantzeko ebaluazio-metodoak proposatu ditugu.

Gako hitzak: laburpen-corpusa, euskara, CLARIAH-EUS, eskola.

Abstract

The data collected in PhD theses and the tools created to analyze these data are currently dispersed across university warehouses, hindering their potential reuse. One of the primary objectives of digital research infrastructures such as CLARIAH-EUS is to mitigate this issue by facilitating easy searchability of thesis-produced data and enhancing their visibility and reusability for other researchers. For instance, the CLARIN-ERIC infrastructure offers the Virtual Language Observatory (VLO) tool, which enables users to search for data and tools across various repositories. This paper aims to contribute to the research community by uploading the school summary text data collected onto EuDat, subsequently making it accessible through CLARIN's VLO for wider utilization by fellow researchers, under the CC BY-NC 4.0 license. Additionally, we propose novel methods and tools aimed at generating summaries and evaluating methods for working with school summaries.

Keywords: sumary corpora, Basque, CLARIAH-EUS, classroom.

1. Sarrera

Laburpenaren inguruko ikerkuntza benetan da garrantzitsua, laburpena, besteak beste, erabakigarria delako ikaskuntzan zein hezkuntzan. Autore ugarik laburpenaren garrantzia ulermenarekin duen lotura estuan sostengatzen du. Andersonek eta Hidik (1988) testuen ulermena hobetzeko tresnatzat jotzen dute, eta, Khoshsimaren eta Rezaeian-Tiyarren (2014) hitzetan, ikasleak ulermen-prozesuaz jabetzeko estrategiarik esanguratsuena eta osoena da. Hala ere, ikaskuntza-prozesuari begira, ulermenaz gain, laburpenaren beste alderdi batzuk ere nabarmentzen dituzte beste hainbat autorek; esaterako, badira fokoa ideien garrantzian jartzen dutenak. Khoshsimaren eta Rezaeian-Tiyarren (2014) hitzetan, sintesi-prozesuak ideia garrantzitsuenetan arreta jartzea eskatzen du, ulermenarekin batera memoria lantzea ere badakarrelako. Ildo beretik, Seidlhoferrek (1991) esaten du edozein ikasketa-prozesutan ideia garrantzitsuenak bereizi behar direla, ideia horietako informazioa ezagutzari gehitzeko. Gainera, bada ikaskuntzaren eta laburpenaren arteko harremana mutur-muturrera daramanik ere; izan ere, Van Dijkek (1979) adierazten du funtzio bera dutela laburpenak eta ikaskuntzak, ataza biek egitura garrantzitsuak zeintzuk diren erabakitzea eskatzen digutelako.

Horiek horrela, gure aburuz, laburtzeko gaitasuna funtsezkoa da ikasleek curriculumean zehaztutako helburu asko eta asko lortu ahal izateko; alabaina, hainbat dira laburpena ikaskuntza-irakaskuntza prozesuan eraginkortasunez txertatzeko erronkak. Alvarez Angulok (2014) adierazi legez, curriculum ofizialetan zein testuliburuetan intuitiboki lantzen da laburpena. Horrekin lotuta, euskal hezkuntza-sisteman erabiltzen diren testuliburuetan laburpena lantzeko jarduerek ez dutela jarraikortasunik izaten dio Atutxak (2022), eta, EAEko Oinarrizko Hezkuntza Curriculuma (77/2023 Dekretua) eta Batxilergokoa (76/2023 Dekretua) aztertzen baditugu, laburpena garrantzitsua dela antzeman dezakegun arren, ez du lotura edo zehaztapen sendorik curriculumak biltzen dituen oinarrizko jakintzekin, helburuekin eta ebaluazio-irizpideekin. Egoera horren aurrean, gure iritziz, ikasleek egindako eskola-testuen laburpen-corpusa biltzen joatea oso garrantzitsua da. Eskolan, ikaslea Vygotskyk (1978) proposatzen duen garapen potentzialera laguntzeko, oso lagungarria da irakaslearentzat ikasketa-maila bakoitzean lortu beharreko trebeziak ondo jasota izatea, eta hori erdietsi ahal izateko, funtsezkoa da laburpen-corpus egokia biltzea, bai kalitate, bai kantitate aldetik.

Ikusiko dugun moduan, aipatu berri dugun beharrizanari aurre egiteko, guk lan honetan 1.676 laburpenez osatutako laburpen-corpusa bildu dugu, eta corpus hori komunitatean handitzen joateak aukerak asko handituko lituzke. Aspaldidanik mahaigaineratua dagoen afera da datuak elkarbanatzearena, baina azken urteetan izandako aurrerapen teknologikoek datuen garrantzia ikaragarri handitu dute, bereziki adimen artifizialaren eraginez. Izan ere, ikasketa sakona (ingelesez, *deep learning*) bezalako teknikak oso ohiko bilakatu dira ataza konplexuei aterabidea emateko, eta, Rouhiainenek (2018) adierazten duen legez, teknikok datu-kantitate oso handia behar izaten dute.

Testuinguru horrek askoz garrantzitsuago egiten du datuak eta horien bilketarako erabilitako tresnak komunitate osoari eskaintzea. Laburpena eta beronekin lotutako ikerketa-alorrak ez dira salbuespen, are gutxiago euskaraz egindako laburpenez ari bagara; izan ere, euskara bezalako hizkuntza batek duen komunitatea txikia da gaztelania edo ingelesa bezalako hizkuntzekin alderatuz gero, eta horrek balio erantsia ematen dio datuak jaso eta elkarbanatzeko dugun edozein abagune probesteari. Laburpenaren zenbait ikerketa-alor Hizkuntzaren Prozesamenduaren baitan daude; laburpen automatikoa eta laburpenaren ebaluazio automatikoa bezalako atazak, esaterako. Ataza horietan aurrerapauso nabarmenak eman ahal izateko, jakina da kalitatezko datuak eskura izateak berebiziko garrantzia duela; baina kalitatea ez ezik, kantitatea ere funtsezkoa da, adimen ar-

tifizialaren aurrerakuntzek datu-kopuru handien beharrizana nabarmendu baitute. Hori dela eta, ezaugarri ezberdinak dituzten laburpen-corpusak biltzen joatea funtsezkoa izango da, baita corpus horiek biltzea erraztuko duten baliabideak sortzen eta elkarbanatzen joatea ere. Hezkuntzari begira, Hizkuntzaren Prozesamenduan ematen diren aurrerapausoek ikastetxe eta unibertsitateetan ekarpen esanguratsua ekar liezaiokete laburpenaren aferari. Izan ere, aipatu ditugun oztopoez gain, kontuan izan behar dugu irakasle batek 20 ikasle baino gehiago izaten dituela oinarrizko hezkuntzan, batxilergoan eta zer esanik ez unibertsitatean; eta, ondorioz, lan oso nekeza da ikasle guztien laburpenak zuzentzea eta ikasle bakoitzari dagokion atzeraelikadura eskaintzea. Laburpen baten atzeraelikadura emateko, hainbat alderdi ebaluatu behar dira, besteak beste: ideien hierarkia, objektibitatea, erregistroa, kohesioa, koherentzia, zuzentasuna eta luzera. Alderdi horien guztien ebaluazio eta atzeraelikadura automatikoa diseinatu eta eskaini ahal izateko, esku artean laburpen-corpus esanguratsua izatea nahitaezkoa izango da, ikasleen ikasketa-prozesuan lagunduko duen edozein baliabidek kalitate onekoa izan beharko duelako, eta, horretarako, ikasketa automatikoak datu-multzo handia eskatuko digu.

Ikerkuntza-esparruan, beste lan batzuetan burutu diren datu-bilketak eta bilketa horretarako sortutako tresnak maiz berrerabili ohi dira, aberasgarria eta lagungarria izan baitaiteke euskarri sendodun ikerketa bat bideratzeko. Hala ere, tamalez, askotan zaildu egiten zaigu baliabide horiek berrerabiltzea, datu-bilketak eta datu horiek aztertzeko sortzen diren tresnak sakabanatuta egoten dira-eta; tesi-lanetan egiten diren ekarpenak kasu, arrunta da unibertsitateetako biltegietan dispertsaturik egotea. Baina argi izan behar dugu datu eta baliabideen elkar trukea benetan dela interesgarria, abantaila handia baitakarkigu. Arzberger *et al.*en (2004) eta Vickersen (2006) esanetan, beharrezkoa da zientzian inbertitzen den diruari ahalik eta etekin handiena ateratzea, bereziki diru publikoaz finantzatutako proiektuak baldin badira. Gainera, datu horiek eskuragarri izateak ikerkuntza berriei ateak irekiko dizkie (Ramasamy *et al.,* 2008). Halaber, datuak partekatzeak gardentasun handiagoa emango lieke ikertzaileei; batez ere, iruzurraren aurkako borrokan, esperimentuei eta hipotesien egiaztapenari erraz erantzuteko aukera emango bailuke (Rennolls, 1997); eta, Piwowar *et al.*ek (2007) azpimarratu legez, interesgarria da egilearen interes pertsonalei begira, aipuak areagotzeko abagunea ematen duelako. Ildo beretik mintzo dira Aleixandre-Benavent *et al.* (2021), eta ikerkuntza elkarbanatzeak ekar litzakeen onurak hiru bloketan banatzen dituzte: gizarteari egiten zaion ekarpena, datuak elkarbanatzen dituenari egiten zaiona eta ikertzaileen komunitateari egiten zaiona. Ikus 1. taula.

Hainbesteko onurak ekar litzakeen elkarbanatzeak oinarri sendoak izan ditzan, funtsezkoa da ikertzaileek beharrezko azpiegiturak eskuragarri izatea eta erabiltzea. Horretarako bide ematen duten azpiegituren artean, EUDAT (Lecarpentier *et al.,* 2013) eta CLARIN nabarmenduko ditugu, horiexek baitira lan honetan izango ditugun euskarriak. EUDAT Europar Batasunak finantzatutako ekimena da, eta datuen elkarlanerako azpiegitura-eredu bat garatu nahi du (*Collaborative Data Infrastructure*, CDI), tartean dauden eragile guztiak barne hartuko dituena (datu-gordailu nazionalen kudeatzaileak; biltegiratze- eta konputazio-zentroak; sareak; eta datu-azpiegitura horiek erabiltzen dituzten ikerketa-zentroak eta unibertsitateak). Lan honetan, EUDAT baliatzea erabaki dugu, datu-azpiegitura horren helburuak eta gureak bat datozelako; hau da, elkarrekintzaren alde egiten duelako eta Europa mailan diharduten hainbat diziplinatako ikertzaileen beharrei zein gizartearen onurari begira antolatua dagoen testuinguru publikoa delako. EUDATek *B2 Service Suite* izeneko zerbitzuak eskaintzen ditu, eta zerbitzu horiek erabiltzaileen komunitateekin batera diseinatu, eraiki eta ezartzen dira; artean, CLARIN. CLARIN hizkuntzarekin lotutako humanitateen eta gizarte-zientzien ikerketa sustatzeko Europako e-azpiegitura da, erronka nagusitzat honako hau duena: hizkuntza-teknologia eta hizkuntza-baliabideak partekatzen dituzten zentroen sare gisa hedatzea, humanitateen eta gizarte-zientzien arloetan testuak (idatziak edo ahozkoak) prozesatzen eta ustiatzen lan egiten duten ikertzaileen eskura jartzeko. Helburutzat du hizkuntza-

datu asko (edo hizkuntzekin lotutakoak) eta teknologia-baliabideak eskuratzeko, integratzeko eta ustiatzeko aukera bermatzea (Bel *et al.*, 2016).

1. taula. Ikerketa-datuak partekatzearen abantailak (Aleixandre-Benavent *et al.*, 2021)

Gizartearentzako onurak
— Aurkikuntzak baliozkotzeko aukera ematen du.
— Zientziaren sinesgarritasuna handitzen du, agerian uzten baitu inbertsio publikoa ikerketan.
— Aukera ekonomikoak eskaintzen ditu.
— Funtsezkoa da lehiakortasuna bermatzeko.
Datuen jabearentzako onurak
— Lanen ikusgarritasuna eta irisgarritasuna hobetzen ditu; aipuak eta inpaktua areagotzen ditu.
— Lankideekin harreman berriak ezartzeko aukera ematen du; ondorioz, lankidetza sustatzen da.
— Prozesuak, hasieran, ikertzaileek beren datuak etiketatzeko ahalegin handiagoa egitea eskatzen badu ere, azken emaitza onuragarria da, informazioa eta datuak hobeto antolatuta dituztelako.
Ikertzaileen komunitatearentzako onurak
— Ezagutza berriak aurkitzeko aukera gehiago daude, berrikuntza sustatzen baita.
— Datu-kantitate handiekin lan egiteko eta estatistika-analisi itsuak egiteko aukerak handitzen dira.
— Lanak erreproduzitzeko aukera ematen du.
— Kostuak murrizten dira, proiektu garestiak errepikatzea saihesten baita.
— Iruzurra eta jardunbide txarrak antzeman daitezke.
— Emaitzak nola lortu diren hobeto ulertzeko aukera ematen du.
— Proiektu askok gutxieneko kostua izatea ahalbidetzen du, dauden datuak aprobetxatzen direlako, baliabideak hobeto erabilitzen direlako eta eraginkortasuna hobetzen delako.
— Datuak erabiliko dituztenen eta datu horiek sortu zituztenen artean lankidetza berriak ezartzeko aukera ematen du.
— Gardentasuna hobetzen du.
— Ikerketa-metodoen hobekuntza eta baliozkotzea sustatzen du.
— Ikerketaren ikusgarritasuna eta inpaktua handitzen ditu.
— Kreditu handiagoa eman diezaioke datuak sortu dituen ikertzaileari.
— Hezkuntzarako eta prestakuntzarako baliabideak ematen ditu.

Beraz, orain arteko guztia kontuan izanda, funtsezkoa deritzogunez ikasketa-maila guztietako ikasleen laburpenak batzeari zein aztertzeari, eta, horrekin batera, corpusa komunitatean handitzen joateko aukera zabaltzeari, lan honen helburu nagusiak hori lortzera bideratu ditugu: i) ikasketa-maila ezberdinetako ikasleek eginiko eskolako laburpen-testuen corpusa euskaraz jasotzea eta ii) corpusa biltzeko erabilitako baliabideak aurkeztea eta erabilgarri jartzea. Horretarako, Unai Atutxaren tesian (Atutxa, 2022) lortutako eskolako laburpen-testuen datuak EuDatera igo ditugu eta, ondoren, CLARINen VLOn eskuragarri jarri ditugu, beste ikertzaileek erabil ditzaten, CC BY-NC 4.0 lizentziapean. Bestalde, laburpenak lortzeko tresnak eta laburpenak eskolan lantzeko ebaluazio-metodoak proposatu ditugu, eta horiek ere eskuragarri jarri ditugu.

2. Metodologia

Atal honetan, gure helburuak lortzeko eman ditugun urratsak azalduko ditugu; bi izan dira. Lehendabizi, laburpenak egiteko erabili diren jatorrizko testuak zeintzuk izan diren zehaztuko dugu, eta horien ezaugarri nagusiak jakitera emango ditugu. Ondoren, jatorrizko testu horien laburpenak egiteko erabilitako baliabideak aurkeztuko ditugu.

2.1. Laburtu diren jatorrizko testuak

Ikasleek laburtzeko aukeratu ditugun testuak eskolan ikasleekin curriculuma lantzeko erabiltzen diren testuak dira, horrek aukera ematen digulako testuinguru erreal batean aritzeko. Hauek dira laburtu beharreko jatorrizko testuen ezaugarri nagusiak:

— 13 dira laburtzeko aukeratu ditugun testuak.

— Azalpen-testuak dira guztiak. Ikasleak azalpen-testuekin aritzen direnez curriculumeko oinarrizko jakintzak lantzeko, testu tipologia hori hautatu dugu.

— LHko 5. eta 6. mailako ikasleek eskolan erabilitako testuak dira. Adin tarte horretan da egokia laburpena lantzen hastea (Sanz Moreno, 2005); beraz, egokiak begitandu zaizkigu. Unibertsitateko ikasleen laburpenak ere bilduko ditugu, eta, horretarako, testu berberak laburtuko dituzte Unibertsitatekoek testu horiek laburtzea egokia iruditu zaigu; izan ere, Hezkuntza Fakultate batekoak dira ikasleak, etorkizunean irakasle izango direnak, eta horrelako testuekin aritu beharko dutenak.

— Batzuk monodokumentuak dira eta besteak multidokumentuak. Ikasleek testuak banan-banan laburtu dituztenean, monodokumentuen laburpenez ariko gara. Baina, kasu batzuetan, 4 monodokumentu batu egin ditugu, testu bakarra bailiran; kasu horietan, multidokumentuen laburpentzat joko ditugu. Multidokumentua sortzeko, bildu diren monodokumentuen batuketa erabat koherentea dela bermatu dugu.

- Monodokumentuak: 11 azalpen-testu ditugu; 2. taulako lehen 11 testuak.
- Multidokumentuak: 4 azalpen-testu batuz, 2 multidokumentu sortu ditugu; 2. taulako 12- eta 13-Testuak.

Laburtu beharreko testuek dituzten ezaugarri orokorren berri emateko, Analhitza[1] testu-analizatzailea erabili dugu (Otegi *et al.,* 2017), eta hauek dira ezaugarri nabarmenak (ikus 2. taula):[2]

[1] Analhitza: http://ixa2.si.ehu.eus/clarink/analhitza.php?lang=eu

[2] Analhitza erabilita 2. taulan bildutako ezaugarri orokorren adibidea: 1-Testuan, testuaren izenburua «Giza gorputzak eta elikagaiak» da. Testu horrek 5 paragrafo, 11 esaldi eta 121 hitz ditu orotara. Hitz horien artean, testuak 23 motatako 42 izen ditu.

2. taula. Laburtu diren jatorrizko testuen ezaugarriak Analhitzarekin

Testua	Izenburua	Paragrafoak	Esaldiak	Hitzak	Izenak
1	Giza gorputza eta elikagaiak	5	11	121	23/42
2	Elikagai eraikitzaileak	4	11	131	39/55
3	Elikagai energetikoak	10	17	218	65/107
4	Elikagai erregulatzaileak	11	25	289	69/128
5	Eragile geologikoak	3	3	49	13/18
6	Inguruneko kalteen erantzule	4	6	69	25/27
7	Itsasoaren eragina	4	7	111	18/24
8	Lurraren eraketa	7	15	154	42/75
9	Lurrikarak	5	13	155	43/67
10	Uholdeak	3	6	125	29/51
11	Uraren erabilera desegokia	7	6	102	25/45
12	Giza gorputza eta elikagaiak	4	64	759	158/332
13	Eragile geologikoak	12	30	440	105/174

2.2. Corpusa biltzeko erabilitako baliabideak

Jatorrizko testuak laburtzeko eta sortutako laburpenak bildu ahal izateko, zenbait baliabide sortu ditugu. Hasteko, Compress-eus (Atutxa *et al.*, 2017) aurkeztuko dugu, laburpenak digitalki biltzea ahalbidetuko digun tresna. Ondoren, laburpena egin bitartean ikasleak laguntzeko proposatu dugun atzeraelikadura automatikoa azalduko dugu. Atzeraelikadura hori ideien hierarkiarena da, eta bi proposamen egin ditugu, bi metodotan oinarrituta: i) Hierarkia Mailen Arteko Metodoan (HIMAM) eta ii) Galderetan Oinarritutako Metodoan (GOM). Azkenik, baliabide horiekin laburpena lantzeko tailerrak egin ditugunez, tailer horietako bat burutzeko sortu dugun webgunea aurkeztuko dugu; bertan, ikasleak tailerra burutzeko behar dituen azalpen, argibide eta materialak biltzen dira.

2.2.1. Compress-eus: laburpena lantzeko eta biltzeko tresna

Euskarazko laburpen-corpusa batzeko, laburpenak digitalki biltzeko aukera ematen duen tresna erabili dugu: Compress-eus (Atutxa *et al.*, 2017). Jarraian, corpusa biltzeko nahitaezkoa izan den tresna horren ezaugarri nagusiak, erabilera-urratsak eta ematen duen informazioa azalduko ditugu.

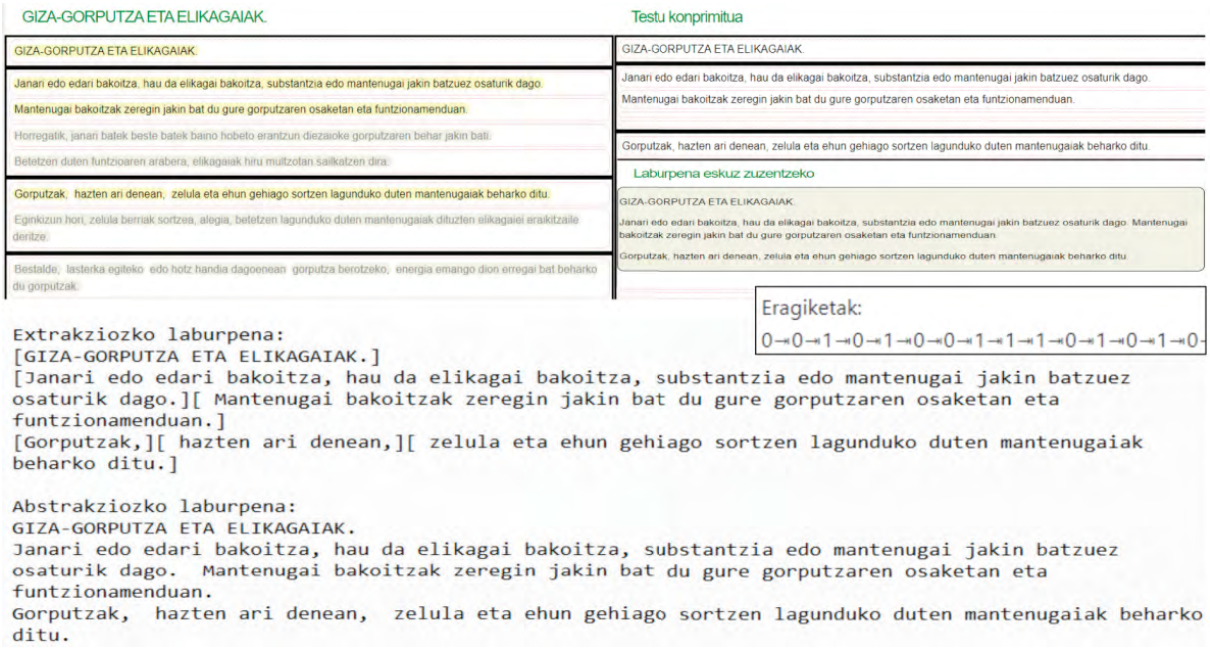
Compress-eus tresnaren ezaugarri nagusiak honako hauek dira:

— Irakasleen zein ikasleen laburpen-kopuru handia digitalki eta automatikoki jasotzeko ahalmena du.
— Laburpenak estrategia jakin bati jarraituz biltzea ahalbidetzen du: estrakzio-laburpenetik abiatuz, abstrakzio-laburpena sortzea[3].
— Ikasleak estrakzio-laburpenean egiten duena ondo ulertzeko eta aztertzeko aukera ematen du.

[3] Estrakzio-laburpena testuko zati garrantzitsuenak aukeratzean datza; beraz, testuak ez du inolako aldaketarik. Aldiz, abstrakzio-laburpena egitean, testua berregin egiten dugu; edukia berbera da, baina erabilitako hitzek eta esaldiek aldaketak izan ditzakete.

Compres-eus erabiltzeko urratsak honako hauek dira[4]:

1. Sisteman erregistratu: erabiltzailea erregistratu egin beharko da, bere erabiltzaile-izena eta pasahitza eskuratzeko. Erregistro horri esker, erabiltzailearen inguruko informazio garrantzitsua geureganatuko dugu.
2. Estrakzio-laburpena egin: laburtu beharreko testua agertuko zaio erabiltzaileari, ODUetan (Oinarrizko diskurtso-unitateetan) segmentatuta. Erabiltzaileak segmenturik garrantzitsuenak mantenduko ditu, eta beharrezkoak ez direnak ezabatu. Estrakzio-laburpena egiten denean, erabiltzaileak ez dio erreparatu beharko testuaren zuzentasun gramatikalari. 1. irudian, goialdean, ezkerretara, estrakzio-laburpena non egiten den Compress-eusen.
3. Abstrakzio-laburpena egin: estrakzio-laburpenean egindakoa abiapuntu izanda, erabiltzaileak abstrakzio-laburpena burutuko du; hau da, testua berregingo du. Litekeena da estrakzio-laburpena gramatikalki eta kohesio aldetik erabat zuzena ez izatea; horregatik, erabiltzailea testuari zuzentasuna ematen ahaleginduko da, estrakzio-laburpenean aukeratutako segmentuetako ideiak bere hitzekin idatziz. 1. irudian, eskuinetara, «Laburpena eskuz zuzentzeko» jartzen duen tokian, abstrakzio-laburpena non egiten den Compress-eus tresnan.



1. irudia. COMPRESS-EUS tresna:
lortutako estrakzio-laburpena, estrakzioaren kodea eta abstrakzio-laburpena

Compress-eus tresnak ematen duen informazioa honako hau da:

— Ikasleari:

- Egindako bi laburpen-motak (estrakzioa eta abstrakzioa) jasoko ditu erabiltzaileak, TXT fitxategi batean. Ikus 1. irudiaren behealdea.
- Estrakzio-laburpenean egin dituen eragiketak[5]. Ikus 1. irudiaren erdialdean.

[4] Compress-eus hemen proba daiteke: http://ixa2.si.ehu.eus/compress-eus/

[5] Testu-zati (ODU) bakoitzean egindakoa agertuko zaio. «0» agertzen bazaio, esan nahiko du estrakzioan ODU hori mantendu egin duela; eta «1» agertzen bazaio, ezabatu egin duela.

— Irakasleari:

- Laburtu beharreko testuaren informazioa:
  - — Dokumentuaren izena.
  - — Testuak duen paragrafo-kopurua.
  - — Testuak duen esaldi-kopurua.
  - — Testuak duen UZ-kopurua (Unitate Zentrala).
  - — Testuko ODUak: ODU bakoitza zein esaldi eta paragraforen parte den zehaztuz.
  - — ODU-kopurua: testuarena, esaldiena eta paragrafoena.
  - — TOKEN-kopurua: testuarena, esaldiena eta paragrafoena.
- Ikasle bakoitzak estrakzio-laburpenean egindakoa:
  - — Ikaslearen izena.
  - — Ikaslearen kodea.
  - — Testuko zenbat UZ kendu dituen.
  - — Kendutako ODU-kopurua: testu osoan, esaldietan eta paragrafoetan.
  - — Kendutako ODUen ehunekoa: testu osoan.
  - — Zein ODU mantendu eta ezabatu dituen.
  - — Kendutako TOKEN-kopurua: testu osoan, esaldietan eta paragrafoetan.
  - — Kendutako TOKENen ehunekoa: testu osoan.

2.2.2. Laburpenen hierarkiaren atzeraelikadura automatikoa

Ideien hierarkiaren atzeraelikadura automatikoa diseinatu dugu, ikasleek (beraien burua) zein irakasleek ebalua dezaten zer moduz aritu den ikaslea testuko ideiarik garrantzitsuenak mantentzen eta garrantzi gutxikoak ezabatzen. Bi metodo aurkeztuko ditugu: i) Hierarkia Mailen Arteko Metodoa (HIMAM) eta ii) Galderetan Oinarritutako Metodoa (GOM): aurreko metodoaren hobekuntza da.

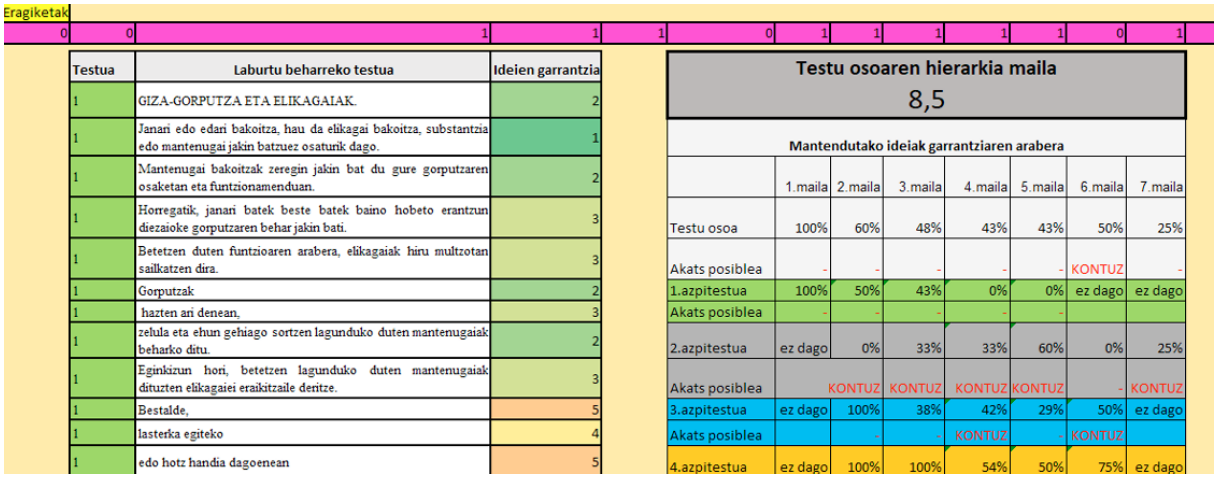
HIMAM metodoan oinarritutako atzeraelikadura automatikoa

Atutxak (2022) proposatzen duen metodologiari jarraituta, ikaslea hierarkian nola aritu den kalkulatutzen da, hiru urrats nagusitan datzan metodologiarekin: i) jatorrizko testuko ideia (ODU) bakoitzaren garrantzi-maila zehaztu, ii) testuak dituen garrantzi-mailak multzokatu eta iii) ikaslea garrantzi-maila bakoitzean nola aritu den aztertu eta ponderatu, garrantzi-maila bakoitza osatzen duten testu-zatien (ODUen) ehunekoak alderatuz. Gero, informazio hori ikasleari aurkezten diogu, ikaskuntza-prozesurako lagungarri izango zaion atzeraelikadura gisa eskainiz, kalkulu-orri batean. HIMAM metodoan oinarrituta, bi testuren atzeraelikadura prestatu dugu kalkulu-orrian: 12-Testuarena eta 13-Testuarena[6]. Honatx kalkulu-orrian prestatutako atzeraelikadurak eskaintzen duen informazioa (ikus 2. irudian 12-Testuari dagokion atzeraelikaduraren adibidea):

— Eragiketak itsasteko lekua: ikasleak Compress-eus tresnak ematen dizkion eragiketak itsatsi behar ditu arrosa koloreko laukietan; goialdean.

[6] HIMAM metodoan oinarritutako atzeraelikadura automatikoaren kalkulu-orria, 12-Testuarena eta 13-Testuarena: https://bit.ly/3TLT1tj

— Laburtu behar izan duen testua, ODUetan zatituta.
— ODU bakoitzaren garrantzi-maila: ODU bakoitzaren eskuinean.
— Hierarkian lortu duen emaitza: kalkulu-orriaren eskuin aldean.
— Garrantzi-maila bakoitzean nola aritu den eta arazoa non izan dezakeen: hierarkia-emaitzaren azpian.



Eragiketak

0	0	1	1
Testua	Laburtu beharreko testua	Ideien garrantzia	
1	GIZA-GORPUTZA ETA ELIKAGAIAK.	2	
1	Janari edo edari bakoitza, hau da elikagai bakoitza, substantzia edo mantenugai jakin batzuez osaturik dago.	1	
1	Mantenugai bakoitzak zeregin jakin bat du gure gorputzaren osaketan eta funtzionamenduan.	2	
1	Horregatik, janari batek beste batek baino hobeto erantzun diezaioke gorputzaren behar jakin bati.	3	
1	Betetzen duten funtzioaren arabera, elikagaiak hiru multzotan sailkatzen dira.	3	
1	Gorputzak	2	
1	hazten ari denean,	3	
1	zelula eta ehun gehiago sortzen lagunduko duten mantenugaiak beharko ditu.	2	
1	Eginkizun hori, betetzen lagunduko duten mantenugaiak dituzten elikagaiei eraikitzaile deritze.	3	
1	Bestalde,	5	
1	lasterka egiteko	4	
1	edo hotz handia dagoenean	5	

0	1	1	1	1	1	0	1
Testu osoaren hierarkia maila							
8,5							
Mantendutako ideiak garrantziaren arabera							
	1.maila	2.maila	3.maila	4.maila	5.maila	6.maila	7.maila
Testu osoa	100%	60%	48%	43%	43%	50%	25%
Akats posiblea	-	-	-	-	-	KONTUZ	-
1.azpitestua	100%	50%	43%	0%	0%	ez dago	ez dago
Akats posiblea	-	-	-	-	-		
2.azpitestua	ez dago	0%	33%	33%	60%	0%	25%
Akats posiblea		KONTUZ	KONTUZ	KONTUZ	KONTUZ	-	KONTUZ
3.azpitestua	ez dago	100%	38%	42%	29%	50%	ez dago
Akats posiblea		-	-	KONTUZ	-	KONTUZ	
4.azpitestua	ez dago	100%	100%	54%	50%	75%	ez dago

2. irudia. HIMAM metodoan oinarrituta, ikasleek jasotako atzeraelikadura automatikoa

GOM metodoan oinarritutako atzeraelikadura automatikoa

Atutxak (2022) HIMAM metodoari zenbait hobekuntza iradokitzen dizkio. Hobekuntza horien helburuetako bat da ikaslea laburpen-prozesuaren une ezberdinetan lagunduko duen proposamena egitea, atzeraelikadura prozesuko momentu gehiagotan egon dadin eta irakasleak aukera izan dezan prozesua ere ebaluatzeko. Horretarako, galderen bidez gidatutako laburpena egitea hartzen du aukeratzat, eta GOM metodoa proposatzen du. Hauek dira GOM metodoan ematen diren urratsak: i) ikasleak gidatzeko galderak sortzea, *Rhetorical Structure Theory*rekin (RST) (Mann eta Thompson, 1987) uztartuz, ii) jatorrizko testuko ideia (ODU) bakoitzaren garrantzi-maila zehaztea (ez da HIMAM metodoan erabilitako modu bera) eta iii) garrantzi-mailan oinarrituz, ikaslea ideiak mantentzen eta kentzen nola aritu den aztertu eta ponderatzea.

GOM metodoaren kasuan, HIMAM metodorekin alderatuz, atzeraelikadura diseinatzeko bi informazio-iturri daudela izan behar dugu kontuan: i) galderen inguruko informazioa eta ii) ideiak mantenduz eta kenduz lortutako emaitza; beraz, ikasleari eman beharreko galderez gain, bi atzeraelikadura sortu ditugu kalkulu-orri batean, eta 12-Testuari dagokion atzeraelikadura da kalkulu-orri horretan burutu duguna. Honatx sortutako galderen eta bi atzeraelikadura moten azalpen laburra:

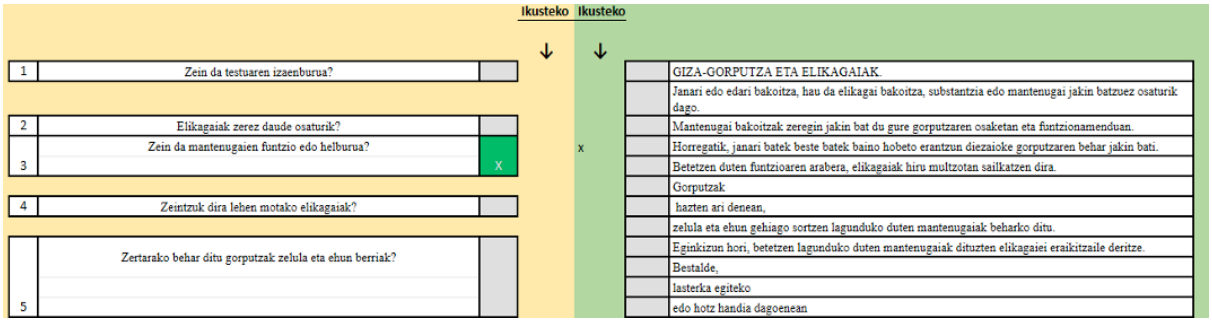
I. Ikasleak gidatzeko galderak:

Laburpenak zeri erantzun behar dion jakin dezan ikasleak, irakasleak galderak sortuko ditu. Atutxaren (2022) tesian azaltzen den moduan, galderak RST-zuhaitzen adarrei jarraituz sortzen dira; izan ere, Compress-eus tresna eta diseinatutako ideien garrantzi-mailaketa ere RSTn oinarrituta daude, eta atzeraelikadura emateko informazioak ere RST du abiapuntu. Hala ere, irakasleak ez du RST-zuhaitzarekin aritu beharko; testu-zatien aukeraketa eta gainerako ekintza guztiak testuan bertan egingo ditu. Irakasleak ikaslearen mailara egokitu ahalko ditu galderak, erabaki ahalko

baitu ikasleak aukeratu beharreko testu-zatientzat zenbat galdera erabili, eta galdera errazago edo konplexuagoak sor ditzake, testu-zati horiek identifikatzeko egoki deritzon zailtasun-mailaren arabera. Eranskinean, ikasleek 12-Testuaren laburpena egin dezaten sortu ditugun galderak daude ikusgai.

II. Galderak erantzuten laguntzeko atzeraelikadura:[7]

Compress-eus erabilita, ikaslea estrakzio-laburpena egiten hasiko da, galderak erantzunez. Horretan laguntzeko, atzeraelikadura automatikoa jasoko du. 3. irudian ageri den kalkulu-orriari esker, ikasleak ODU bakoitzak zein galdera erantzuten laguntzen duen ikusi ahalko du[8]. Prozesu hori alderantziz ere egin daiteke; hau da, posible da ikasleak galdera bat aukeratzea, galdera horri erantzuteko behar dituen ODUak zeintzuk diren ikusteko[9].



3. irudia. Galderak erantzuten laguntzeko atzeraelikadura: 4. ODUa x batekin markatuta ageri da eskuman, eta ezkerrean adierazten da ODU hori 3. galdera erantzuteko behar dela

III. Galderen eta hierarkiaren atzeraelikadura automatikoa[10]:

Estrakzio-laburpena egiten bukatu eta gero, ikasleak estrakzio-laburpenean egindako ideien aukeraketen atzeraelikadura automatikoa jasoko du. Atzeraelikadura kalkulu-orri batean garatu dugu; hau da bertan ikasleak izango duena (ikus 4. irudia):

— Eragiketak itsasteko lekua: ikasleak Compress-eus tresnak ematen dizkion eragiketak itsatsi behar ditu arrosa koloreko laukietan; goialdean.
— Laburtu behar izan duen testua, ODUetan zatituta (ez da ageri 4. irudian; 2. irudikoa bezalakoa da).
— ODU bakoitzaren garrantzi-maila (ez da ageri 4. irudian; 2. irudikoa bezalakoa da).

---

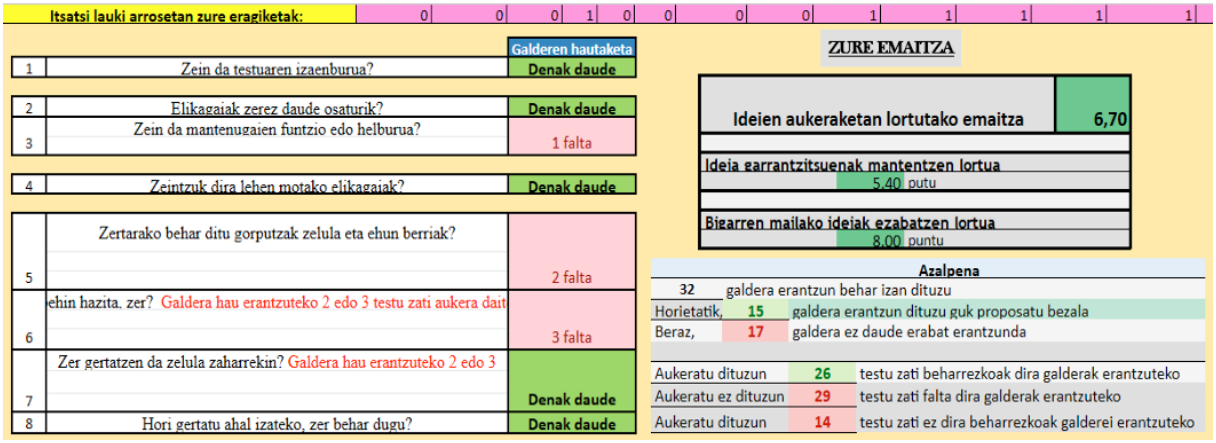
[7] Galderak erantzuten laguntzeko atzeraelikadura automatikoa, 12-Testuarena: https://bit.ly/3EIu8uq

[8] Irudiko adibidean, ikasleak 4. ODUak zein galdera erantzuten duen jakin nahi izan du; horretarako, eskuineko aldean, testua segmentatuta du, eta 4. ODUaren alboan «x» bat jarri du. Hori egindakoan, kalkulu-orriak adierazi dio ODU hori 3. galdera erantzuteko behar dela. Hala ere, posible da ikasleak kontsultatutako ODUa beharrezkoa ez izatea galderaren bat erantzuteko; hori gertatzen denean, kalkulu-orriak adieraziko dio hautatu duen testu-zatia ez dela beharrezkoa galderei erantzuteko.

[9] Ikasleari asko lagun diezaiokeen atzeraelikadura izan daiteke, baina erabilera mugatzea beharrezkoa da. Atzeraelikadura horrekin ikasleari nahi beste kontsulta egiten uzten baldin badiogu, atzeraelikadurak esana itsuitsuan jarraitu besterik ez du, eta horrek ez luke ekarriko ikaskuntzarik. Muga non jarri zehazteko, zenbait faktore izan beharko ditugu kontuan: ikaslearen maila, testuaren zailtasuna, testuaren ODU-kopurua, galdera-kopurua eta galderen zailtasuna.

[10] Galderen eta hierarkiaren atzeraelikadura automatikoa 12-Testuarena: https://bit.ly/3Epup3R

— Galderak nola erantzun dituen: galdera guztiak agertuko zaizkio, eta, galdera bakoitzaren alboan, galdera nola erantzun duen (4. irudiko erdi aldean)[11].
— Ideien hierarkiarekin nola aritu den: eskuinean, lortu duen emaitza orokorra agertuko zaio. Emaitza horren azpian, mantendu beharreko ideiak mantentzen eta kendu beharrekoak kentzen nola aritu den[12].
— Galderen eta aukeratutako testu-zatien informazioa: hierarkian lortutako emaitzen azpian, galderen eta aukeratutako testu-zatien informazioa emango diogu ikasleari:
— Zenbat galdera erantzun behar izan dituen; horietatik zenbat erantzun dituen osorik; eta zenbat utzi dituen guztiz erantzun gabe.
— Aukeratu dituen zenbat testu-zati diren beharrezkoak galderak erantzuteko; aukeratu ez dituen zenbat testu-zati diren beharrezko galderak erantzuteko; eta aukeratu dituen zenbat testu-zati ez diren beharrezko.



4. irudia. Galderen eta hierarkiaren atzeraelikadura automatikoa

2.2.3. Laburpen-gaitasuna garatzeko tailerra

Aurreko azpiataletan aurkeztutako baliabideak erabilita, laburpen-gaitasuna garatzeko tailerrak burutu ditugu. Horietako bat GOM metodoan oinarritu dugu, eta, ikasleak ondo bideratu ahal izateko, webgune bat sortu dugu[13]. Bertan, ikasleak urrats guztiak ditu azalduta, eta, horrekin batera, behar dituen baliabideak eta horien erabileraren azalpena. 5. irudian, webgunearen ataletako bat ageri da, laburpena galderak erantzunez egiteko atazari dagokiona, hain zuzen ere. Irudi horretan, esaterako, egin behar duenaren azalpena idatzita dauka ikasleak, eta, ezkerreko aldean, beharko duen material guztia gehi material hori erabiltzeko laguntza: laburpenak egiteko jarraibideen bideoa, Compress-eus tresna erabiltzeko sarbidea, laburpena egiteko irakasleak emandako galderak, galdera horiek erantzuten laguntzeko atzeraelikadura eta ataza igotzeko esteka (azken hau irudian agertu ez arren).

[11] Galdera bat erantzuteko behar diren ODU guztiak hautatuta daudenean, «denak daude» agertuko da. Ostera, galdera erantzuteko behar diren ODU guztiak ez baldin baditu hautatu, zehaztuko zaio galderaren alboan zenbat ODU falta diren galdera guztiz erantzuteko.

[12] Ikasleak ez badu 5 baino gehiago erdietsi mantendu eta kendu beharreko atalen emaitzetan, emaitza orokorrean ez zaio notarik agertuko (bataz bestekoa 5 baino altuagoa denean ere ez). Horren ordez, ikasleari adieraziko zaio zertan hobetu behar duen: ideiak mantentzen, kentzen edo bietan.

[13] Laburpen-gaitasuna garatzeko tailerraren webgunea: https://atutxaunai.wixsite.com/lab-tailerra2022

5. irudia. Laburpen-gaitasuna garatzeko sortu dugun tailerraren webgunea

Jarraian, labur-labur, webgunearen atal guztiak azalduko ditugu, ikasleak bertan aurkituko duenaren berri emanez:

I. Hasiera: atal honetan, tailerra aurkezten diegu ikasleeri. Tailerraren helburua zein den zehazten diegu, eta jarraibide orokorrak ematen dizkiegu, gainontzeko atalak aurkeztuz: i) Compress-eus, ii) Laburpen-irizpideak, iii) Laburpena galderekin, iv) Nola aritu naiz? eta v) Ni ebaluatzaile.
II. Compress-eus: bertan estrakzio- eta abstrakzio-laburpenak egin beharko dituztela esaten diegu, eta, horiek Compress-eus tresnan nola egin jakiteko, bideo bat dute ikusgai. Horiez gain, Compress-eus tresnaren sarbidea ere badute.
III. Laburpenak egin eta ebaluatzeko irizpideak: irizpide horiek talde handian lantzeko, irakasleak erabiliko duen aurkezpena dago eskuragarri atal honetan; gainera, aurkezpenean egingo diren ariketa batzuk Wooclap (ariketa dinamikoak sortzeko tresna) erabilita egingo direnez, bertara jotzeko sarbidea prest dute ikasleek.
IV. Laburpena galderei erantzunez: ikasleei laburpen bat egin beharko dutela adierazten diegu. Horretarako, lehendabizi Compress-eus tresnara jo beharko dute, bertan izango dutelako laburtu beharreko testua; arestian esan bezala, Compress-eus tresnarako sarbidea izango dute bertan. Ondoren, irakaslearen galderak erantzunez, estrakzio- eta abstrakzio-laburpenak egin beharko dituzte, eta, estrakzioarekin laguntzeko, momentuko atzeraelikadura erabilgarri izango dute. Laburpena egiteko prozesu osoa ondo uler dezaten, dituzten baliabideak nola erabili jakin dezaten bereziki, adibidetzat bideo bat dute, laguntzeko. Bukatzeko, ataza bidaltzeko esteka ere bertan dago.
V. *Nola aritu naiz?* atalean, egindako laburpenaren ideien hierarkia ebaluatu ahalko dute, hots, zer-nola aukeratu dituzten laburpena eratzeko ideiak. Horretarako, ideien hierarkia izeneko atzeraelikadura erabili ahalko dute, hots, 4. irudian azaldu duguna. Gainera, bertako informazioa ulertzen laguntzeko bideo bat dute ikusgai. Bestalde, aurreko atazan erabili ahal izan

duten momentuko ataza ere badago; izan ere, laburpena egitean, 3 kontsulta egitera mugatu dugu, ez dezaten kopia hutsa egin, baina, orain, behin laburpena eginda, zalantza guztiak argitzeko aukera izango dute, nahi beste kontsulta eginez. Horrez gain, atal honetan, webgunean bertan, beheko aldean, testuaren egitura orokorra ikusi ahalko dute, baita testuaren hierarkiaren eta irakaslearen galderen arteko lotura ere. Bukatzeko, ikasleek galdetegi bat erantzun behar izan dute, burututako prozesuaz hausnar dezaten; galdetegi horretara daraman sarbidea webgunean bertan dute.

VI. Ebaluatzera: atal honetan, azken egitekoa burutzeko argibideak jasoko dituzte. Kasu honetan, ikasleen eta irakaslearen laburpenak ebaluatu beharko dituzte. Webgunean, laburpenak ebaluatzeko errubrika eta ebaluatu behar dituzten laburpenetara bideratuko dituen esteka dituzte. Azkenik, galdetegi bati erantzun beharko diote, euren laburpen-gaitasunaz hausnartzeko; galdetegi horren esteka ere badute bertan.

3. Emaitzak

Emaitzen atalean, lehendabizi, ikasleek burututako laburpen-corpusa deskribatuko dugu, orotara zenbat laburpen bildu zehaztuz diren eta laburpen horien ezaugarri nagusiak azalduz. Ondoren, bildutako laburpenak EuDat eta CLARINen bidez nola jarri ditugun eskuragarri azalduko dugu.

3.1. Bildutako laburpen-corpusa

Orotara 1.676na estrakzio- eta abstrakzio-laburpen bildu ditugu; horietatik, 1.380 (% 82,33) monodokumentuen laburpenak dira, eta 296 (% 17,66), aldiz, multidokumentuenak. Laburpen batzuk Lehen Hezkuntzako ikasleek burutu dituzte, eta beste batzuk unibertsitateko ikasleek, etorkizunean irakasle izango direnek, alegia. Gainera, laburpen guztiak ez dira baliabide berberak erabiliz egin. Jarraian, laburpenak ezaugarrien arabera multzokatuko eta azalduko ditugu; horretarako, 3. taulaz baliatuko gara. Lehen Hezkuntzako 352 (guztien % 21) laburpen batu ditugu, 2017. urtean batuak, hain zuzen ere; denak dira monodokumentuak, 1-, 2-, 3- eta 4-Testuei dagozkienak. Laburpenik gehientsuenak, 1.324 (guztien % 79), unibertsitatekoak dira; hala ere, bada ezberdintasunik euren artean. Monodokumentuak dira nagusi, 1.028 orotara (unibertsitatekoen % 77,64), eta horiek guztiak Compress-eus erabilita egin dituzte ikasleek. 2017an bildu ziren horietako 303 (unibertsitatekoen % 22,88); 5-, 6-, 7-, 8-, 9, 10- eta 11-Testuaren laburpenak dira. 370 laburpen (unibertsitatekoen % 27,94) 2018. urtean bildutakoak dira, eta 2019. urtekoak beste 355 (unibertsitatekoen % 26,81); urte bi horietakoak 1-, 2-, 3-, eta 4-Testuaren laburpenak dira. Bukatzeko, multidokumentuen laburpenak lortu ditugu, 296 (laburpen guztien % 17,66), guztiak unibertsitateko ikasleek eginak. Bi urtetan jaso ditugu, 2020 eta 2021, batean zein bestean Compress-eus erabilita. 2020. urtean 12- eta 13-Testuaren 230 laburpen bildu ditugu, eta, laburpenok egiteko, ikasleek HIMAM metodoan oinarritutako atzeraelikadura erabili dute. 2021. urtean, ostera, 12-Testuaren laburpenak dira jasotako guztiak; 66 orotara. Urte horretako laburpen guztiak GOM metodoan oinarrituz burutu ziren; hau da, ikasleek honako atzeraelikaduraz baliatu ahal izan zuten laburpena egiteko: laburpena egiten gidatzeko galderak; galderak erantzuten laguntzeko atzeraelikadura automatikoa; eta galderen eta hierarkiaren atzeraelikadura automatikoa.

3. taula. EuDat eta CLARINen eskuragarri jarritako laburpen-corpusaren ezaugarri nagusiak

EUDATen izendatua	Testua	Laburpen kantitatea	EUDATen izendatua	Testua	Laburpen kantitatea
Compress-eus soilik erabilita					
1mono_LH17	1	102	11mono_uni17	11	28
2mono_LH17	2	97	1mono_uni18	1	97
3mono_LH17	3	82	2mono_uni18	2	94
4mono_LH17	4	71	3mono_uni18	3	91
5mono_uni17	5	49	4mono_uni18	4	88
6mono_uni17	6	31	1mono_uni19	1	98
7mono_uni17	7	48	2mono_uni19	2	91
8mono_uni17	8	30	3mono_uni19	3	85
9mono_uni17	9	44	4mono_uni19	4	81
10mono_uni17	10	73			
Compress-eus eta HIMAM atzeraelikadura erabilita					
12multi_uni20	12	151	13multi_uni20	13	79
Compress-eus, GOM atzeraelikadura eta laburpen-tailerraren webgunea erabilita					
12multi_uni21	12	66			

3.2. Laburpen-corpusa elkarbanatzea: EUDAT eta Clarin

Deskribatu berri dugun corpusa eskuragarri uzteko, EUDAT azpiegitura erabili dugu. EUDATek *B2 Service Suite* izeneko zerbitzuak eskaintzen ditu, eta zerbitzu horiek erabiltzaileen komunitateekin batera diseinatu, eraiki eta ezartzen dira, artean CLARINeko komunitatea. Gure datuak CLARIN komunitatearekin elkarbanatzeko, EUDATek eskaintzen duen B2SHARE zerbitzua erabili dugu, Virtual Language Observatory (VLO) izeneko biltegian gure datuak partekatuz. Datuak Dataset kategorian gorde dira, Eskola-testuen laburpenak izenarekin eta honako lizentzia honekin: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Laburpenak 3. taulako lehen zutabean dauden izenekin antolatuta aurki genitzake. Corpusa honako helbide honetan dago eskuragarri: https://doi.org/10.23728/B2SHARE.65867F59BFAE498B89376033275F24A8

4. Ondorioak eta etorkizuneko lanak

Lan honetan, 1.676 estrakzio- eta abstrakzio-laburpenez osatutako laburpen-corpusa aurkeztu dugu. Gure helburua ikasketa-maila ezberdinetako ikasleek eginiko eskolako laburpen-testuen corpusa euskaraz jasotzea izan da, izan ere; ikasketa-maila bakoitzean ikasleek laburpenak biltzen dituen trebetasunetan izan beharreko maila zehazteko, lehenik laburpen kopuru handi bat aztertzeari beharrezkoa deritzogu. Bada, lan honetan, Lehen Hezkuntza eta unibertsitateko ikasleen laburpenak bildu ditugu, monodokumentuen eta multidokumentuen laburpenak. Corpusa aurkezteaz gain, CLARINen VLOn eskuragarri jarri ditugu, beste ikertzaileek erabil ditzaten, CC BY-NC 4.0 lizentziapean, eta, horretarako, EUDAT azpiegitura erabili dugu. Horrela, gure asmoa da beste ikertzaile batzuek corpus hori erabilgarri izatea eta eurekin bat corpusa handitzeko abagunea ematea.

Horrez gain, guk laburpenak biltzeko sortu ditugun baliabideak ere aurkeztu ditugu, nahi duenak erabilgarri izan ditzan. Alde batetik, estrakzio- eta abstrakzio-laburpenak egin eta laburpenok digitalki biltzeko aukera ematen duen tresna: Compress-eus. Eta, bestetik, ikasleei laburpena egin bitartean eta bukaeran eskaintzen diegun atzeraelikadura automatikoa. Atzeraelikadura ho-

rrek ideien hierarkiari buruzko informazioa eskaintzen du, bi metodotan oinarrituta: i) Hierarkia Mailen Arteko Metodoa (HIMAM) eta ii) Galderetan Oinarritutako Metodoa (GOM).

Corpus eta baliabide horiek eskuragarri jartzea lagungarria izango da corpusa handitzen joateko; izan ere, etorkizuneko lanei begira erabat funtsezkoa izango da. Esan bezala, ikasketa-maila guztietako ikasleak bildu nahi ditugunez, Batxilergoko ikasleen laburpenak ere bildu beharko dira, eta, gainera, laburtzeko erabili ditugun 13 azalpen-testuez gain, beste testu-mota batzuen laburpenak ere jaso beharko dira, zailtasun-maila ezberdinetakoak. Corpusa horrela handitzeak beste ate batzuk ere ireki ditzake, esaterako Hizkuntzaren Prozesamenduarekin zerikusia dutenak, datu asko lortzea nahitaezkoa baita laburpen automatikoa edo laburpenaren ebaluazio automatikoa bezalako atazei heltzeko, baita, ideien hierarkiarekin egin antzera, laburpenak biltzen dituen gainontzeko trebetasunen atzeraelikadura automatikoa eskaintzeko ere.

Bibliografia

76/2023 Dekretua, maiatzaren 30ekoa, Batxilergoaren curriculuma zehaztu eta Euskal Autonomia Erkidegoan ezartzekoa (EHAA, 2023/06/09).

77/2023 Dekretua, maiatzaren 30ekoa, Oinarrizko Hezkuntzaren curriculuma zehaztu eta Euskal Autonomia Erkidegoan ezartzekoa (EHAA, 2023/06/09).

Aleixandre-Benavent, R., Sapena, A. F., eta Peset, F. (2021). Compartir los recursos útiles para la investigación: datos abiertos (open data). Educación Médica, 22, 208-215.

Alvarez Angulo, T. (2014). A vueltas con el resumen escolar. Esta vez en la pizarra, con tiza y borrador. *Revista digital de la Asociación de Profesores de Español «Francisco de Quevedo» de Madrid*, *19*, 11-30.

Anderson, V., eta Hidi, S. (1988). Teaching students to summarize. *Educational leadership*, *46*(4), 26-28 pp.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., eta Wouters, P., (2004). An international framework to promote access to data. *Science*, v. 303, n. 5665, pp. 1777-1778.

Atutxa, U., Iruskieta M., Ansa O., eta Molina A. 2017. Compress-eus: I(ra)kasleen laburpenak lortzeko tresna. *EUDIA: Euskararen bariazioa eta bariazioaren irakaskuntza-III*, 87-98.

Atutxa-Barrenetxea, U. (2022). *Laburpen-gaitasunaren garapena eta eskolako laburpen-testuen prozesamendua*. Doktorego tesia. Euskal Herriko Unibertsitatea (UPV/EHU). https://addi.ehu.es/handle/10810/59346

Bel, N., Gonzalez-Blanco, E., eta Iruskieta, M. (2016). CLARIN centro-k-español. *Procesamiento del Lenguaje Natural*, (57), 151-154.

Khoshsima, H., eta Rezaeian, F. (2014). The effect of summarizing strategy on reading comprehension of Iranian intermediate EFL learners. *International Journal of Language and Linguistics*, *2*(3), 134-139.

Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R., Coveney, P., eta Baxter, R. (2013). EUDAT: a new cross-disciplinary data infrastructure for science. *International Journal of Digital Curation*, 8(1), 279-287.

Mann, W. C., eta Thompson, S. A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. *Text*, 8(3), 243-281.

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.

Piwowar, H. A., Day, R. S., eta Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3), e308.

Ramasamy, A., Mondry, A., Holmes, C. C., eta Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, 5(9), e184.

Rennolls, K. (1997). Intersalt data. Science demands data sharing. *BMJ: British Medical Journal*, 315(7106), 486.

Rouhiainen, L. (2018). *Inteligencia artificial. Madrid: Alienta Editorial*, 20-21.

Sanz Moreno, Á. M. (2005). Irakurmena lantzeko jarduerak nola prestatu: lehen hezkuntzako 3. Zikloa eta DBHko 1. Zikloa.'Cómo diseñar actividades de comprensión lectora: tercer ciclo de Primaria y primer ciclo de la ESO'. *BLITZ*. Nafarroako Gobernua Hezkuntza Departamentua.

Seidlhofer, B. (1991). *Discourse analysis for summarization*. Doktoregoko tesia., University of London. https://discovery.ucl.ac.uk/id/eprint/10018780

Van Dijk, T. A. (1979). Relevance assignment in discourse comprehension. *Discourse processes*, 2(2), 113-126.

Vickers, A. J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials*, 7(1), 1-6.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Eranskina. GOM metodoan oinarrituta, ikasleei 12-Testua laburtzeko emandako galderak

1. Zein da testuaren izenburua?
2. Elikagaiak zerez daude osaturik?
3. Zein da mantenugaien funtzio edo helburua?
4. Zeintzuk dira lehen motako elikagaiak?
5. Zertarako behar ditu gorputzak zelula eta ehun berriak?
6. Eta behin hazita, zer?
7. Zer gertatzen da zelula zaharrekin?
8. Hori gertatu ahal izateko, zer behar dugu?
9. Baina zer gertatzen da neurriz kanpo jaten badugu behar dugun hori?
10. Mantenugai horiek dituzten zein jaki-motatan aurki genitzake?
11. Zeintzuk dira bigarren motako elikagaiak?
12. Elikagai horietan bi motatako mantenugaiak bereizten dira; zeintzuk dira lehenak?
13. Guretzat beharrezko den zer ematen digute?
14. Zertarako da beharrezkoa?
15. Gorputzeko zein organotan batzen dira mantenugai horiek?
16. Eta bertatik nora jotzen dute ondoren?
17. Baina berehala erabiltzen ez ditugunean, non geratzen dira?
18. Ondorioz, glukogenoari zer gertatzen zaio?
19. Zehazki, noiz?
20. Eta non pilatzen da?
21. Horrek guztiak zer eragin du azkenean?
22. Elikagai horietan bi motatako mantenugaiak bereizten dira; zeintzuk dira bigarrenak?
23. Mantenugaiok, bereziki, zer ematen diote gorputzari?
24. Batzuetan gorputzak zer egiten du berarekin?
25. Zein kasutan, zehazki?
26. Zein motatako jakietan aurki genitzake mantenugaiok?
27. Bereizten diren moten artean, zein da onuragarriagoa gorputzarentzat?
28. Zeintzuk dira hirugarren motako elikagaiak?
29. Zein da elikagai horien garrantzia guretzat?
30. Zeintzuk dira bitaminen ezaugarriak?
31. Zeintzuk dira gatz-mineralen ezaugarriak?
32. Zeintzuk dira uraren ezaugarriak?

Goi-mailako testu akademikoak lantzeko baliabideak eta tresnak

Resources and tools for the development of high level academic texts

Maria Jesus Aranzabe, Igone Zabala, Izaskun Aldezabal

Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU),
Euskal Hizkuntza eta Komunikazioa Saila; Ixa Taldea, HiTZ zentroa
maxux.aranzabe@ehu.eus igone.zabala@ehu.eus izaskun.aldezabal@ehu.eus

Laburpena

Ezinbestekoa da hizkuntza bat erabiltzea eremu akademikoetan, eremu horretako adierazpide-baliabide bereizgarriak garatu ahal izateko. Izan ere, hizkuntza-komunitate batek baliabide horiek garatzeari uzten dionean, erabilera-eremu espezializatuak galtzeko arriskua du. Nahiz eta eremu akademikoen konkista erabakigarria izan den euskara biziberritzeko, erregistro akademikoak oraindik ez daude erabat garatuta, ezta egonkortuta ere. Artikulu honetan, euskararen erregistro akademikoak garatzen laguntzeko eta garapena bera ikertzeko sortu ditugun baliabide eta tresnak deskribatu ditugu, gaur egun ere aberasten dihardugunak: Garaterm corpusa, TZOS (Terminologia Zerbitzurako Online Sistema) terminologia-datubasea eta Testu Akademikoen Idazketarako Laguntza Tresna (HARTA/TAILA) elebiduna. Garapen-lan horietaz gain, helburuen artean ditugu ere alor akademikoan erreferentziazko lan-ingurunea izatea eta lankidetzak bilatzea, datuak partekatuz sare dinamiko nahiz kolaboratiboak sortze aldera.

Gako hitzak: corpus espezializatua, terminologiarako datubasea, konbinazio lexiko akademikoak (KLA).

Abstract

The use of a language in academic context is crucial for the development of the characteristic expressions of specialized fields, since if a linguistic community stops developing such expressions, the risk of losing the use in specialized areas grows dramatically. Although the conquest of academic fields has been decisive for the revitalization of the Basque language, academic registers are not yet fully developed or stabilized. In this paper we describe the resources and tools we have created and are continuously enriching to contribute to the development and research of the academic registers in Basque: The Garaterm corpus, the TZOS (Online System for the Service of Terminology) terminological database and the bilingual Academic Text Writing Support Tool (HARTA/TAILA). In addition to these development tasks, we aim to be a referential work environment in the academic field as well as to search for collaboration, having as objective the creation of dynamic expert nets and interoperable data sharing.

Keywords: specialized corpus, terminological database, academic lexical combinations (ALC).

1. Sarrera

Ezinbestekoa da hizkuntza bat erabiltzea eremu akademikoetan, eremu horretako adierazpide-baliabide bereizgarriak garatu ahal izateko. Izan ere, hizkuntza-komunitate batek baliabide horiek garatzeari uzten dionean, erabilera-eremu espezializatuak galtzeko arriskua du (Lauren *et al.,* 2002). Zer esanik ez hizkuntza gutxituen biziberritzearen ikuspegitik. Hizkuntza alor guztietan erabiltzeak berebiziko garrantzia du, bere biziraupena bermatu nahi bada. Hori adierazten dute hizkuntza gutxituen inguruan azken urteotan egin diren hainbat biltzarretan, hala nola 2023ko azaroaren 29an Zientzia eta Teknologia Fakultatean egin zen *Lekuan Lekuko hizkuntzak unibertsitateko instrukzio-hizkuntza modura* jardunaldian, eta 2021eko martxoaren 24tik 26era izan zen *Hizkuntza Gutxituen XVIII. Nazioarteko Biltzar*rean.

Jakina denez, euskararen kasuan, ezagutza handitzen doa; baina ezin da halakorik esan erabilerari dagokionez, kale-eremuetan eta, zehazkiago, hiriguneetan bederen (Altuna *et al.,* 2021). Unibertsitate-eremura iraganda, ordea, euskararen erabilera handiagoa da ikastetxeak kokatuta dauden hiriburuetako erabilerarekin alderatuta, Euskal Herriko Unibertsitateko (UPV/EHU) hainbat ikastetxetako neurketek adierazi dutenez.[1] Gainera, euskararen erabilera oso mugatua den eremuetako hiztun askok unibertsitatean aurkitzen dute euskaraz murgiltzeko aukera. Datu hauek agerian jartzen dute unibertsitatea funtsezko eragilea dela euskara biziberritzeko prozesuan.

Izatez, goi-mailako ikasketetan, 1970eko hamarkadaren bukaeran eta laurogeiko hamarkadan hasi zen euskara sartzen, eta ordutik beti egin dira ahaleginak hezkuntza-maila hau ahalik eta normalizatuena eta euskaldunena izateko. Prozesu horretan, alorretako aditu euskaldunez osatutako komunitateak sortzen joan dira, eta horiek bermatu dute gaur egungo euskararen garapen funtzionala.

Hala ere, nahiz eta eremu akademikoen konkista erabakigarria izan den euskara biziberritzeko (Zabala, 2019), erregistro akademikoak oraindik ez daude erabat garatuta, ezta egonkortuta ere (Zabala *et al.,* 2011; Zabala *et al.,* 2021), eta adituek erregistro hau garatzen lagunduko dieten laguntza-tresnak behar dituzte (Frankenberg *et al.,* 2018; García Salido *et al.,* 2018; Granger y Paquot, 2015).

Behar horiei erantzuteko helburuaz sortu zen Garaterm lan-ingurunea[2] duela hamabost urte (Zabala *et al.,* 2013), euskararen erregistro akademikoen garapenean laguntzeko eta garapena bera ikertzeko egokiak diren tresnak eta baliabideak lan-ingurune berean integratzea zuena helburu. Baliabide eta tresna horien abiapuntua unibertsitateko komunikazio-egoera espezializatuetan erabiltzen diren idatzizko eta ahozko ekoizpenak/testuak dira, terminologia eta fraseologia modu naturalean sortzen eta garatzen den testuinguru naturala. Testu horiek aukera ematen digute, batetik, adituek eta aditugaiek egiten duten erregistro formaletako euskararen erabilera deskribatzeko eta, bestetik, sortu ditugun baliabide eta tresna horiek bai irakaskuntzan, bai ikerkuntzan erabiltzeko. Nahiz eta garatutako tresna eta baliabide horiek ez dauden oraindik erabat integratuta Garaterm lan-ingurunean, uneotan bilduta ditugunak deskribatuko ditugu lan honetan.

Hala, 2. atalean Garaterm eta HARTAeus corpusak deskribatuko ditugu. 3. atalean, Garaterm corpusean oinarrituta sortu dugun TZOS (Terminologia Zerbitzurako Online Sistema) terminolo-

---

[1] UPV/EHUren III. Euskararen Plan Gidariko (2018-2022) helburuetako bat izan zen neurketa horiek egitea. IV. Plan Gidaria uneotan garatzen ari denez, emaitzok ez daude oraindik argitaratuta (Eskola/Fakultate bakoitzak du horren berri).

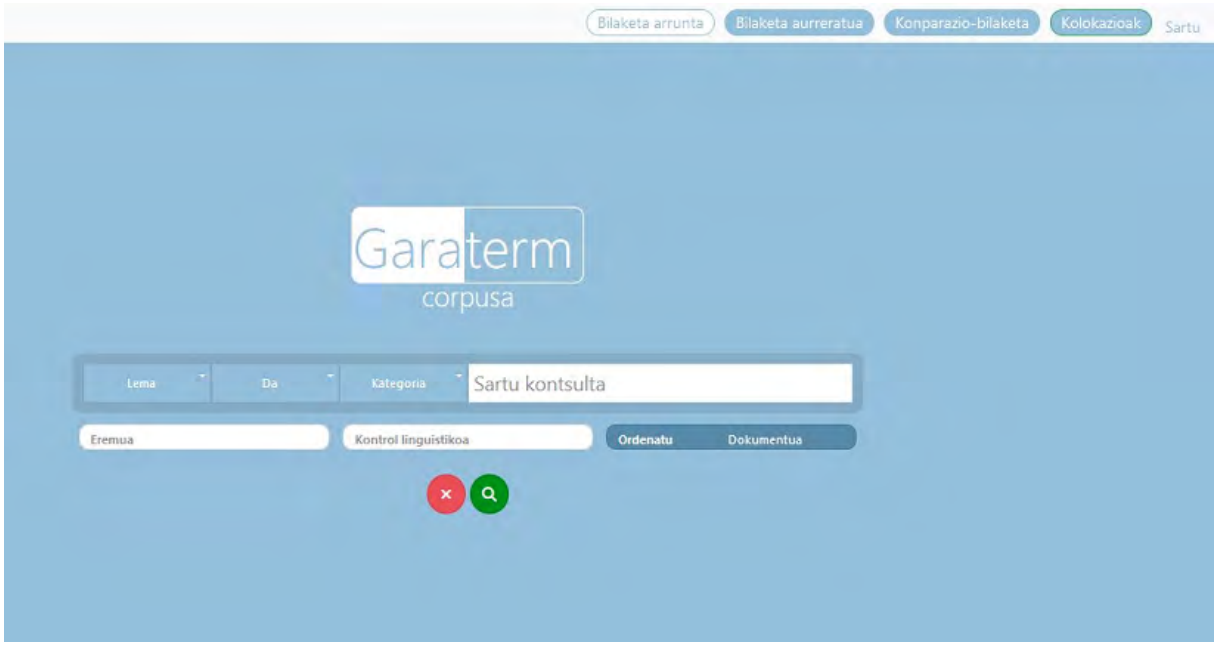
[2] http://garaterm.ehu.es/garaterm_ataria/

giarako datubasea azalduko dugu, eta horri lotuta ireki ditugun bi lan-ildo ere aurkeztuko ditugu: bata, Ingeniaritza Zibila alorrera egokitutako IZIBI-TZOS datubasea, eta, bestea, hodeian datu ireki gisa (LLOD, Linguistic Linked Open Data) sortu dugun prototipoa, beste hizkuntzekin lotura egin ahal izateko. 4. atalean, HARTAeus corpusean oinarrituta garatutako HARTA/TAILA (Testu Akademikoen Idazketarako Laguntza Tresna) laguntza-tresna elebiduna azalduko dugu. Bukatzeko, 5. atalean, ondorioen eta etorkizuneko lanen berri emango dugu.

2. GARATERM corpusa

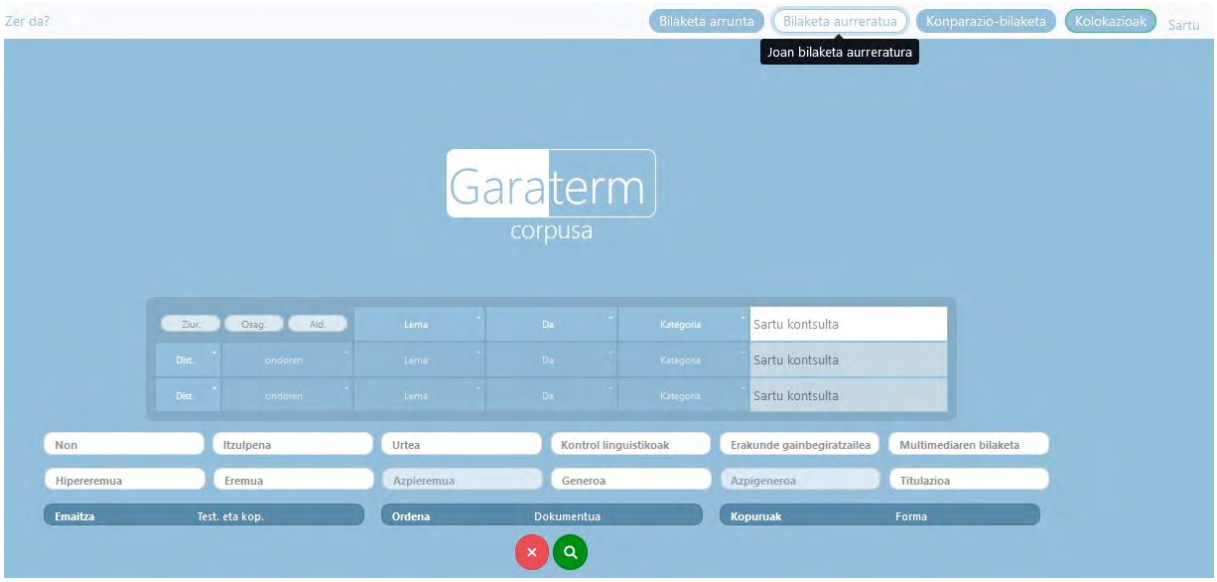
Garaterm corpusa (Zabala *et al.*, 2013) unibertsitateko irakaskuntza-testuez eta ikerketa-lan akademikoez (GrALak, MALak eta doktorego-tesiak) osatutako euskarazko corpus akademiko bakarra da. Erregistro akademikoen garapena aztertzeko aukera eskaintzen duen corpus monitore hau etengabe elikatzen ari gara. Elikatze hori testuen egileek egin ohi dute testuak kargatzeko intuitiboa eta atsegina den datubasearen interfazearen bitartez. Orain artean, 1.131 irakasle-ikertzailek hartu dute parte.

Corpuseko testuak eta metadatuak erregistratzeko datubase horrek eremu hauek ditu: jakintza-alorra eta azpialorra (Unescoren Zientzia eta Teknologiarako sailkapenean oinarritutako 24 jakintza-alor eta 287 azpialor); egilearen rola (egilea, gainbegiratzailea, itzultzailea, lan akademikoen tutorea/zuzendaria) eta espezializazioa (aditu irakaslea, aditu profesionala, ikertzailea, kazetaria); testu-generoa (artikulua, hitzaldia, ikerketa-lan akademikoa, ikerketa-txostena, jarduera profesionala, komunikabideetako testua, kongresuetako akta, liburua, material didaktikoa edo material lexikografikoa) eta azpigeneroa (material didaktikoan, adibidez, apunteak, ariketak, ariketa zuzenduak, azterketa, irakaskuntza-gida, PowerPoint aurkezpenak, praktika-protokoloa…); testuaren datu zehatzagoak (adibidez, irakasgaiaren izenburua, erakundea, ikastegia, titulazioa, gaia, irakasmaila eta ikasturtea); testuaren jatorria (itzulpena den ala ez) eta kontrol linguistikoa (zuzendua izan den ala ez). Linguistikoki kontrolatutako (zuzendutako) testuak eta kontrolatu gabeko testuak (espontaneoak) bereizten dira, irakasle-ikertzaileek espontaneoki erabiltzen dituzten hizkuntza-ekoizpenak, zuzentzaileen interbentzioaren ondorio direnetatik bereizi ahal izateko.



1. irudia. Bilaketa arrunta egiteko kontsulta-interfazea Garatermen

Garaterm corpusak 25.322.231 hitz ditu gaur egun. Corpuseko testuak linguistikoki prozesatu ondoren (Zabala *et al.*, 2013), ohiko kontsulta linguistikoak egiteko aukera dago Garaterm corpusaren bi kontsulta-interfaze hauetan: bilaketa arrunta (1. irudia) eta bilaketa aurreratua (2. irudia).



2. irudia. Bilaketa aurreratua egiteko kontsulta-interfazea Garatermen

Kontsulta-interfaze arruntean, erabiltzaileak bilaketak egin ditzake aukera hauek eginda: lema edo forma, lema- edo forma-hasierak eta -bukaerak, gramatika-kategoria, jakintza-alorra eta kontrol linguistikoa (zuzendua, ez-zuzendua edo biak). Bilaketa horien emaitzek corpusean duten agerraldi kopurua eta zenbat egilek erabili duten erakusten dute, besteak beste. Gainera, emaitza horien testuinguru zabalagoak, hau da, testu-zatiak ere eskaintzen ditu, eta testu motari buruzko informazio zehatza ere erakusten du: zein urtetako testua den, zein hipereremu, eremu eta azpieremutakoa den, eta zein genero eta azpigenerotako testua den.

Kontsulta-interfaze aurreratuan, arruntean egin daitezkeen bilaketez gain, hainbat lema- edo forma-segidaz galdetuta egin daitezke bilaketak. Are gehiago, kontsulta-interfaze arruntean ez bezala, bilaketa zehatzagoak egin daitezke dauden hainbat parametro erabilita, adibidez, bilaketa Garaterm corpus osoan ala azpicorpus jakin batean egin daiteke; bilaketa hipereremu batera muga daiteke eta, ondoren, hipereremu horretako eremua eta azpieremuak zehaztu; eta bilaketa testugenero eta azpigenero jakin batean egin daiteke. Gainera, aukera dago bilaketari dagokion audioa entzuteko. Behin bilaketaren emaitza izanda, emaitzak hainbat parametroren arabera ordena daitezke: lema, kategoria, forma, generoa…

Bi kontsulta-interfaze horietaz gain, konparazio-bilaketak ere egin daitezke corpusean. Horrek aukera ematen du, esaterako, aztertzeko zein den adituek gehien erabili duten lema edo forma. Bilaketa horien emaitzek bakoitza zein ehunekotan erabilia izan den erakusteaz gain, eremuen eta urteen araberako konparazio-datuak ere erakusten dituzte.

Garaterm corpusa kontsultagai dago helbide honetan: http://garatermcorpusa.ixa.eus/

2.1. HARTAeus azpicorpusa

HARTAeus (Aranzabe *et al.*, 2022a) Garaterm corpuseko azpicorpusetako bat da. Unibertsitateko ikasleen idazkera akademikoaren lagin adierazgarriak biltzen ditu; zehazki, Gradu eta Master Amaierako Lanak (GrAL eta MAL). Corpus hau gaztelaniako HARTAnoveles corpusarekin (Villayandre, 2018) konparagarria izateko sortu da; beraz, haren ereduari jarraituz osatu da. Osaera horretan, testu motez gain, kontuan hartu dira jakintza-alorrak eta azpialorrak, eta hitz kopurua.

Bi hizkuntzetako corpusak konparagarriak izateko, corpuseko testuak lau hipereremu nagusitan (Artea eta Humanitateak, Biologia eta Osasun Zientziak, Zientzia Fisikoak eta Gizarte Zientziak) banatu dira, eta horietako bakoitza eremuetan; adibidez, Zientzia Fisikoak hipereremuan Lurraren eta Espazioaren Zientziak, Fisika, Ingeniaritza, Informatika eta Kimika eremuak daude. Tamainari dagokionez, 3.285.098 hitzeko euskarazko corpus akademikoa osatu da bildu diren 398 lan akademikoekin (% 71 GrAL eta % 29 MAL). Gaztelaniakoarekin alderatuta, euskarazkoan MAL gutxiago bildu dira, euskaraz egiten diren MALen kopurua oso txikia delako, lan asko konfidentzialtasun-klausulek babesten dituztelako eta horietako asko ez direlako ADDI plataforman (UPV/EHUren Irakaskuntza eta Ikerketarako Artxibo Digitala) argitaratzen; dena den, bi corpusen arteko alderaketa egiteko adinako kopurua izatea lortu da.

HARTAeus, Garaterm corpuseko azpicorpusetako bat izanik, kontsultagai dago Garaterm corpusaren helbide berean. Helbide horretan «bilaketa aurreratua» kontsulta-interfazea aukeratu ondoren, Non leihotxoan HARTAeus aukeratu behar da (3. irudia).

3. irudia. HARTAeus euskarazko corpusaren kokalekua Garaterm corpusean

Gaztelaniarekiko konparagarria den corpusa Universidade da Coruñakoekin batera UPV/EHUkook izan dugun HARTAesvas proiektu koordinatuaren barruan (Alonso-Ramos y Zabala, 2022) osatu

dugu. Proiektuaren helburu nagusia da erregistro akademikoen bereizgarriak diren bi hizkuntzetako konbinazio lexiko akademikoak (KLA) aztertzea eta alderatzea testu akademikoen idazketarako laguntza-tresna garatzeko.

KLAk hitz-segida errepikariak dira, semantikoki konposizionalak edo ez-konposizionalak izan daitezkeenak eta testuan diskurtso-funtzio jakin batzuk (informazioa gehitzea, adibideak ematea, laburbiltzea…) betetzen dituztenak. HARTAeus corpusetik bi KLA mota erauzi ditugu: kolokazioak eta diskurtso-formulak.

Kolokazioak bi osagai lexikoz osatutako konbinazio konposizionalak (*helburuak bete*, *helburu nagusi*, *ondorioak atera*) dira. Erlazio sintaktikoa duten bi osagai horietan, batak (oinarria) bestea (kolokatua) mugatzen du (Mel'čuk, 2015), eta ez dute zertan nahitaez elkarren ondoan egon. HARTAeus corpuseko kolokazio akademiko hautagaiak erauzi ditugu lexiko akademikoaren zerrendako hitzak (Aranzabe *et al.*, 2002a) abiapuntutzat hartuta eta Gurrutxaga *et al.*en (2011, 2018) garatutako elkartze-neurriak erabilita. Erauzketaren emaitzak eskuz balioztatu ondoren, konbinazio lexiko edo patroi sintaktiko hauek dituzten 1005 kolokazio akademiko lortu ditugu: izena-adjektiboa (*helburu nagusi*), izena(postposizioa)-izena (*lagin tamaina*; *laginaren tamaina*), subjektua-aditza (*emaitzek erakutsi*) eta aditza-objektua (*helburuak lortu*).

Diskurtso-formulak, *lexical bundles* ere esaten zaienak, luzera aldakorreko (2 eta 5 osagai artekoa) sekuentzia errepikari jarraituak (*helburu nagusia da*, *azpimarratzekoa da*) dira, diskurtso-funtzio batekin lotzen direnak. Diskurtso-formulen artean diskurtso-markatzaileak ere (*ondorioz*, *hala eta guztiz ere*) sartzen dira (Biber *et al.*, 2004). Euskara hizkuntza eranskaria izanik, diskurtso-formula polilexikoez gain, diskurtso-formula monolexiko polimorfemikoak, *morphemic bundles* ere deritzenak, (*laburbilduz*, *ondorioz*) ere hartu ditugu kontuan azterketa honetan. HARTAeus corpusean diskurtso-formulak identifikatu ditugu irizpide hauei jarraituta: 2 eta 5 hitz bitarteko n-gramak izatea; 4 hipereremu nagusietan (Artea eta Humanitateak, Biologia eta Osasun Zientziak, Zientzia Fisikoak eta Gizarte Zientziak) agertzea eta maiztasunari dagokionez, milioi bat hitzeko 10 agerraldi edo gehiagokoa izatea. Identifikatutako n-grama horiek eskuz balioztatu eta bakoitzari diskurtso-funtzioa esleitzeko prozesuan, balio semantikorik eransten ez dion elementuren bat ezabatu dugu zenbaitetan. Adibidez, identifikatutako *eta hala ere* diskurtso-formulan, *eta* juntagailua ezabatu eta bi hitzeko formula (*hala ere*) mantendu dugu. Prozesu horretan, diskurtso-formula monolexiko polimorfemikoak identifikatu ditugu; adibidez, *eta ondorioz* n-grama *ondorioz* diskurtso-formula modura balioztatu dugu. Behin prozesu hori amaituta, 644 diskurtso-formula dituen zerrenda, HARTA/TAILA tresnan txertatu dena, osatu dugu. Zerrenda horretan, bi hitzeko formulak dira ugarienak.

Hain zuzen ere, HARTAeus corpusean identifikatutako kolokazio eta diskurtso-formula horietan oinarrituta diseinatu dugu HARTA/TAILA (Testu Akademikoak Idazketarako Laguntza tresna), lan akademikoak idazteko laguntza-tresna elebiduna, artikulu honen 4. atalean azaltzen duguna.

3. TZOS terminologiarako datubasea

TZOS (Terminologia Zerbitzurako Online Sistema)[3] euskarazko terminologia lantzeko eta biltzeko datubasea da, elkarlanean aritzeko eta adituen artean sareak sortzeko diseinatuta dagoena (Arregi *et al.*, 2013; Zabala *et al.*, 2019).

[3] http://tzos.ehu.eus/

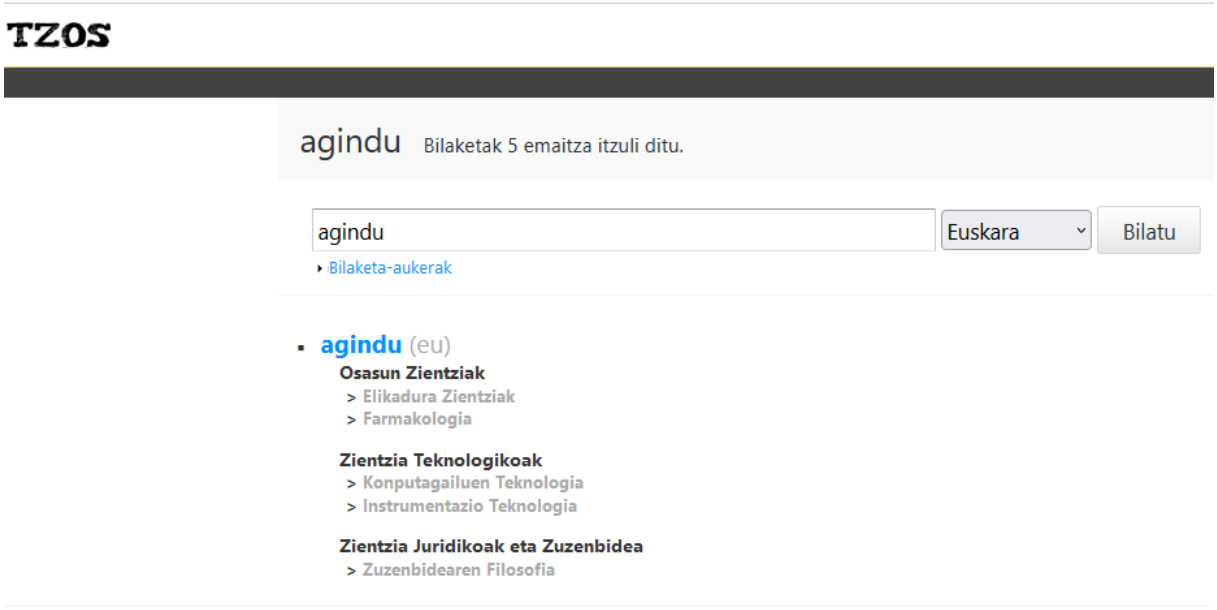
Esan bezala, terminoak lantzeko abiapuntua GARATERM corpusa da, adituen testuak, ezagutza eta erabilera direlako hitz baten terminotasuna erabakitzeko gakoak. Euskaraz alor espezializatuetako komunitateen hizkera garapenean dagoenez, ezinbestekotzat jotzen dugu lehenik adituen erabilera horren deskribapena egitea eta, ondoren, beharrezkoak diren azterketak egin ostean, egon litekeen aldakortasuna harmonizatzea. Horregatik, TZOSek duen ikuspuntua deskriptiboa da, teoria tradizionalean (Wüster, 1968) izan ohi den ikuspuntu preskriptiboaren aldean (zerrenda itxi batetik abiatuta, terminoak aztertu eta ezarri beharreko termino-proposamena egin, adituen eginkizuna termino-hornitzaile huts izatera mugatuta). Hain zuzen, espezialitate-hizkerak hazten hasi eta Eusko Jaurlaritza sortu zenetik, terminologia-lana geroz eta instituzionalagoa bilakatu da, eta aipatutako ikuspuntu preskriptiboa izan da terminologia lantzeko metodologiarako oinarri. Horrelako ekimenen beharra ukatu gabe, garapenerako gune natural diren eremu akademikoak aintzat ez hartzea arriskutsua ikusten dugu, naturalki sortzen diren esapide, termino eta balizko aldaera asko gera baitaitezke bazterrean (Zabala *et al.*, 2018).

Hori dela eta, deskripzio hori bermatuko duen eta adituari behar besteko protagonismoa emango dion metodologia darabilgu: *aktiboa* eta *in vivo* deritzona. Aktiboa, adituak terminoaren lanketa-prozesuan parte hartzen duelako, eta *in vivo,* terminoak testuetako agerraldietan, adibide errealei lotuta alegia, erakusten direlako; hau da, TZOSeko termino-sarrera bakoitza Garaterm corpuseko agerraldiekin konektatuta dago. Metodologiaren xehetasun guztiak Zabala (2019) lanean deskribatuta daude eta alderdi teknologikoari dagozkionak Arregi *et al.* (2013) lanean. Labur esanda, Garaterm corpusetik termino-hautagaiak automatikoki erauzten dira Erauzterm erauzle automatikoaren bidez (Gurrutxaga *et al.*, 2005), irakasle adituei termino-hautagaiak erakusten zaizkie Erauzterm plataforman, eskuz balioztatzen dituzte, eta, termino-zerrenda osatutzat ematen dutenean, beste hizkuntza bateko ordaina ematen diete, horretarako kalkulu-orri eredu jakin bat erabilita.

Horrela egindako lanek beste hizkuntzekiko mendekotasun gutxiago dute eta terminoen orotariko aldaerak detektatzeko oso aproposak dira. Izatez, 2023ko urriaren 9ko AETEReko XXII. Jardunaldian, Elea Giménez CSIC ES-Ciencia plataforma tematikoaren eta TERESIA atariaren[4] (Portal de acceso a TERminologías en ESpaña y servicios de Inteligencia Artificial) koordinatzaileak metodologia bera aurkeztu zuen, Espainia mailan eta Europari begira egin beharreko terminologia-lanerako bide apropos gisa (Martín-Chozas *et al.*, 2022).

TZOSen terminoak bi alderdiren arabera deskribatzen dira: batetik, termino-sarrera definitzen duten ezaugarriak daude, eta, bestetik, termino-sarrera hori erabiltzen duten autoreak eta beren irakasgaien jatorria (zein eskola/fakultate eta zein ikasgai) definitzen dituztenak. Termino-sarrera definitzen duen ezaugarri nagusia 2. atalean aipatutako Unescoren Zientzia eta Teknologiarako sailkapeneko azpialorra da (irakasgaiari dagokiona, berez), eta sarrera horretatik abiatuta, dagozkion aldaerak (euskaraz nahiz beste hizkuntzetan), definizioa(k) eta adibideak ditu. Beste hainbat eremu ere badaude, baina une honetan ez ditugu lantzen; esaterako, hiperonimoa edota kategoria semantikoa. Erabileraren zehaztapenari dagokionez, EHUko eskola/fakultateetako gradu eta irakasgai guztiak daude zerrendatuta, eta irakasleak terminoa horri lotuta lantzen du. Adibidez, "agindu" lemak 5 sarrera ditu, 4. irudian ageri den bezala. Alde batetik, izen gisa lau sarrera ditu: (1) *agindu* (Elikadura Zientziak); (2) *agindu* (Farmakologia); (3) *agindu* (Konputazio Teknologia) eta (4) *agindu* (Instrumentazio Teknologia). Bestetik, aditz gisa sarrera bat du: *agindu* (Zuzenbide Filosofia).

[4] https://pti-esciencia.csic.es/project/teresia-portal-de-acceso-a-terminologias-en-espana-y-servicios-de-inteligencia-artificial/



4. irudia *agindu* lemaren sarrera TZOSen

Erabilerari dagokionez, adibidez (1) terminoa, «Farmazia Fakultatea "Giza Elikadura eta Dietetikako Gradua" Deontologia, Legeria eta Kudeaketa» irakasgaia ematen duen irakasle batena dela adierazten da, datuen babesa bermatze aldera «zenbaki» modura erakusten dena (kopurua, alegia). Horretaz gain, (3)ko sarrerak, adibidez, bi aldaera ditu: *instrukzio* eta *sententzia*. Aldaera horiek autore berak edo desberdinak proposa ditzake. Horregatik, termino osoaren deskribapena hainbat autoreren ekarpenez osatuta egon daiteke. Termino-sarreran klik eginez gero (*agindu* lemak dituen bost termino-sarreretan kasu honetan), adibideak ikusi ahal izango ditugu.

Termino bat datubaseratzeko garaian aldez aurretik termino bera datu-basean baldin badago, *talka* bat gertatzen da eta egileari hiru aukera ematen zaizkio: a) sarrera berri gisa sartu, b) sarrera bera izan arren, informazio berria sartu (beste adibide, aldaera, definizio... bat/batzuk…), edo c) termino beratzat hartu eta bere horretan utzi; kasu honetan, sistemak automatikoki autorea beste erabiltzaile bat bezala gehitzen du termino horretan.

Amaitzeko, termino baten *lantze-prozesua*n, hiru *estatus* edo egoera bereizten dira, *hiru erabiltzaile profil*ekin lotuta daudenak eta batzorde baten modura jokatzeko aukera ematen dutenak, horrela elkarlana erraztuz: 1) Irakasle aditua: terminoa sartu eta dagokion informazioa deskribatzen du; une horretan terminoa «hasieran» egoeran jartzen da. 2) Alorrean esperientzia handia duena eta hausnarketa terminologikoetan murgildu izan dena: terminoa gainbegiratzen du eta egin beharreko oharrak egiten ditu, behar denetan; une horretan, terminoa «lantzen» egoerara pasatzen da. Eta 3) Hizkuntzalari zuzentzailea: terminoa zuzen idatzita eta ondo osatuta dagoen aztertzen du; honek lana bukatutzat ematen duenean, terminoa «kontsolidatu» egoerara pasatzen da. Hala ere, beharrezkoa ikusiz gero, terminoa berriz itzul daiteke nahi den egoerara.

Lan hau guztia posible izan da 2008tik martxan dagoen TSE programari esker (San Martin, 2013; Zabala *et al.*, 2018); alegia, alde batetik, adituen lan eskergari esker, bestetik, EHUko Euskara Errektoreordetzaren finantziazioari esker,[5] eta, azkenik, Euskal Hizkuntza eta Komunikazioa

[5] Errektoreordetza honen izena aldatzen joan da; gaur egun, Euskara, Kultura eta Nazioartekotzearen arloko errektoreordetza da.

Saileko hainbat irakasleren parte hartzeari esker. Urte hauetan guztietan, TZOS terminologiarako datubasea osatzen joan gara, horretarako beharrezkoak izan diren egokitzapenak eta hobekuntzak eginez, bai TZOS datubaseari, bai GARATERM corpusa kudeatzeko sistemari (Aldezabal *et al.,* 2022). Hala, orain, 1) ingurune berean daude corpusa biltegiratzeko modulua, testuen prozesaketarako tresnak, eta Erauzterm balioztatze-gunea, 2) erauzle automatikoa hiztegi eta termino gehiagorekin aberastu da, eta 3) TZOSeko termino guztiak bere corpuseko agerpenekin lotuta daude. Funtzionalitate horiek guztiak ez zeuden hasieran, eta horrek asko erraztu du lan-prozedura.

Zoritxarrez, azken bi urteetan aipatutako errektoreordetzaren finantziazioa etenda dago, eta horrek kolokan jarri ditu proiektuaren helburuak betetzeko aukera guztiak.

Azken batean, TZOSek helburu du adituek erabiltzen duten terminologiaren (erabilera errealaren) erakusleiho izatea, eta erabiltzaileen artean sareak sortzea, terminologia lanetan elkarrekin jarduteko. Egun, 136.520 sarrera ditu eta 180 adituk hartu dute parte (Aranzabe *et al.,* 2022b). Termino horiek esteka honetan kontsulta daitezke: http://tzos.ehu.eus/.

3.1. IZIBI-TZOS

IZIBI-TZOS[6] Ingeniaritza Zibilaren alorrera egokitutako TZOS da, eta mundu akademikoa eta profesionala baliabide berean elkartzen ditu. Ekimen honek UPV/EHUren eta Euskadiko Bide, Ubide eta Portuetako Elkargoaren arteko akordio batean du abiapuntu. Batetik, Ingeniaritza Zibilaren alorra euskaraz oso mugatua izaki, eta bestetik, Elkargoa euskararen erabilera sustatzeko eta areagotzeko plan batean murgildurik, EHUko Ingeniaritza Eskolako irakasle eta elkargoko kide batzuk, eta Euskal Hizkuntza eta Komunikazioa Saileko irakasle batzuk harremanetan jarri ziren eta lan-esparru bat ireki zuten, TZOSen ikuspegi deskriptiboa egokiena zelakoan alor hau lantzeko.

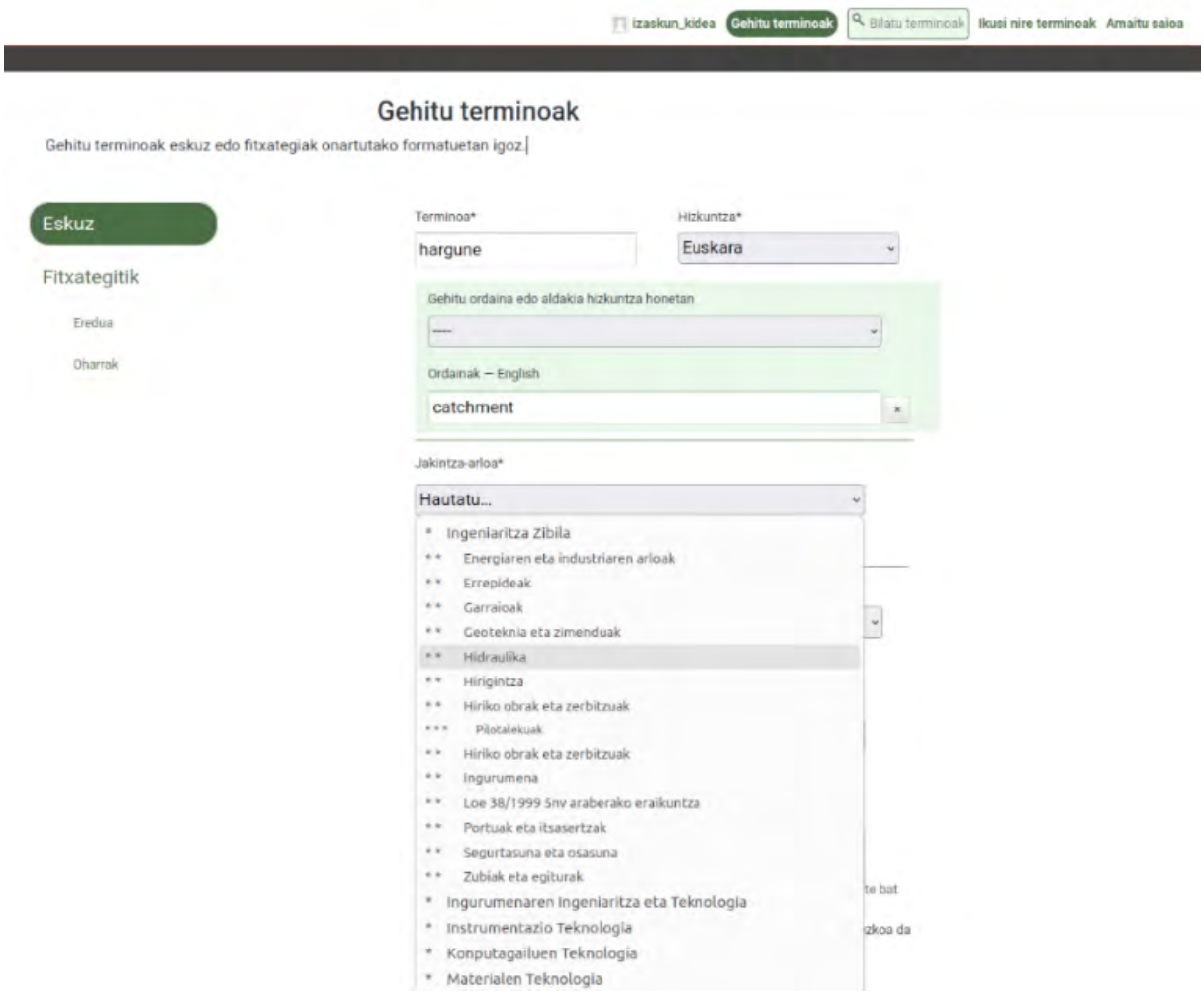
Nahiz eta TZOS batez ere lan akademikoak lantzeko erabili izan den orain arte, duen egitura malguak aukera ematen du lan-mundura ere egokitzeko, horretarako azpian duen eremu-zuhaitza baliatuz. Hain zuzen, eremu-zuhaitzean Ingeniaritza Zibila azpialor bat da, eta azpialor horren azpian, Ingeniaritza Zibilaren lanbideari dagozkion alorrak gehitu ditugu, azpialorraren azpialor gisa (Hirigintza; Ingurumena; Garraioak; Errepideak; Hidraulika; Portuak eta itsasertzak; Zubiak eta egiturak; Geoteknia eta zimenduak; Energiaren eta industriaren arloak; Loe 38/1999 5nv araberako eraikuntza; Segurtasuna eta osasuna). Horretaz gain, erabiltzaile/autoreen artean, alor profesionalekoak gehitu ditugu, oraingoz Elkargoa bakarrik badugu ere. 5. irudian ikusten dira Ingeniaritza Zibila azpialorreko azpialorrak.

IZIBI-TZOSen hastapeneko bertsioa TZOSen kopia bat eginez eraiki zen; zehazki, TZOSen dagoeneko sartuak zeuden Ingeniaritza Zibila azpialorreko termino guztiak eta Ingeniaritza Zibila graduko ikasgaietako termino guztiak kopiatu ziren. Hala, egun IZIBI-TZOSen 82.359 termino daude eta alorreko irakasle eta adituekin jarraitzen dugu datubasea aberasten, 3. atalean deskribatu den metodologia aplikatuz. Hala ere, datubaseak aukera ematen du terminoak banan-banan sartzeko, aldez aurretik eginda dagoen zerrenda batetik abiatuta edo egunerokotasunean sortutako behar batetik, baina betiere erabileraren berri emango duen adibideaz hornitua.

Ingeniaritza Zibilaren alorra aberastea bada ere helburu behinena, orobat oso aukera ona da lan-ildo berriak jorratzeko, hala nola unibertsitatean eta lanean erabiltzen den terminologia konparatzeko, eta Ingeniaritza Zibilaren alorrean terminoak sortzeko dauden joera eta zailtasunak aztertzeko.

---

[6] http://izibi-tzos.ehu.eus/



5. irudia. IZIBI-TZOSerako landu diren azpialorrak Ingeniaritza Zibila azpialorrean

3.2. TZOS LLOD gisa

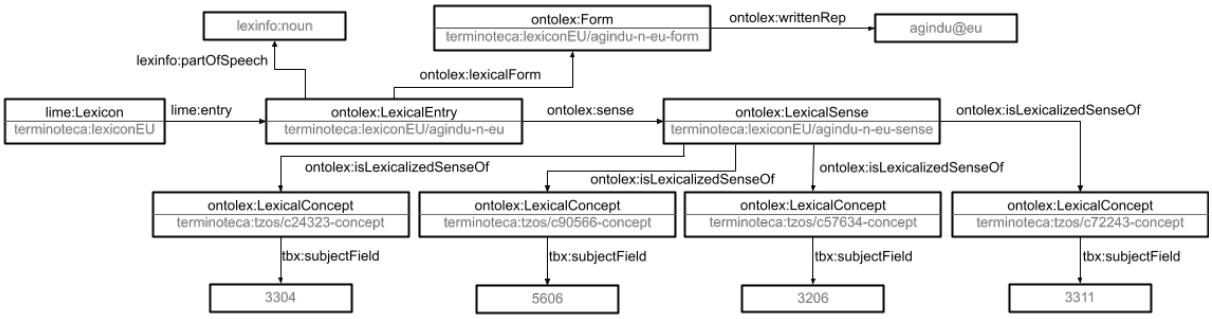
Euskaratik sortutako datubasea izatea funtsezkoa dela azpimarratzen dugun modu berean, beste hizkuntzetan sortu diren edo sor daitezkeen datubase terminologikoekiko lotura egitea ere garrantzitsua dela uste dugu. Lehendik ere aipatu den AETEReko XXII. Jardunaldian, datubase terminologiko eleaniztun irekiak izatearen garrantzia agerian jarri zuten, eta datu-kodeketa amankomuna erabiltzeko deia egin zuten, horretarako erabili ohi den *Ontolex-lemon* eredua (McCrae *et al.*, 2017) bultzatuz. Chiarcos *et al.* (2013) lanean ere formalismo berean kodetzeak (formatuak, galdetzeko lengoaiak...) abantaila ugari ekartzen dituela azpimarratzen da, iturri eta mota desberdineko baliabideak elkarreragile, eskuragarri eta berrerabilgarri jartzen baititu hainbat aplikaziotarako.

Hain zuzen, horretarako urratsak emanak ditugu TZOSen, Aldezabal *et al.* lanean (2022) xeheki azaldu bezala. Dagoeneko garatuta dugu prototipo bat RDF (*Resource Description Framework*) errepresentazioari (Klyne y Carroll, 2004) jarraiki eta *Ontolex-lemon* eredura egokituta.

Ontolex-lemon ereduak nahikoa ezaugarri ditu terminologia eleaniztuna egoki lantzeko: sarrera lexikalak, euren adierak, euren forma flexionatuak, eta kanpo-ontologia orokorrekiko loturak, besteak beste. Eta bestalde, hainbat moduluren bitartez, terminoen arteko aldaera-erlazioak eta beste hizkuntzetako ordain-erlazioak deskriba daitezke.

Laburki esanda, TZOSeko hizkuntza bakoitzeko termino-zerrenda *lime:Lexicon* lexikoian biltzen da. Hala, 5 lexikoi ditugu, euskarari, gaztelaniari, ingelesari, frantsesari eta latinari dagozkienak hain zuzen. Horretaz gain, sinboloentzako aparteko lexikoi bat dugu. Lexikoi hauetako bakoitzean termino (edo sinbolo) bakoitza *ontolex:LexicalEntry* sarrera gisa errepresentatzen da. Termino-sarrera bakoitzak adierazten duen kontzeptuari erreferentzia egiteko, *ontolex:LexicalConcept* deritzon klasea erabiltzen da. Sarrera lexikala eta kontzeptu hau tarteko klase baten bidez lotzen dira, *ontolex:LexicalSense* deiturikoaren bidez alegia, eta honek terminoaren esanahi zehatza adierazten du. Kontzeptu horretan ageri den ezaugarri bat, *tbx:subjectField,* oso garrantzitsua da bertan adierazten baita zein azpialorretan erabiltzen den terminoa.

TZOSeko termino-sarrera edo kontzeptu batek identifikatzaile numeriko bakarra du (URI, *unique numeric identifier*). Guk URI hori egokitu egin dugu, Terminotecako RDF modura izendatzeko (Bosque-Gil *et al.,* 2016). Hala, URI egitura halakoa da: lexikoiaren izena, forma bat (idatzizko errepresentazioa), adiera-zenbaki bat, kategoria gramatikala eta hizkuntza-identifikatzailea. Adibidez, *agindu* izenaren URIa *lexiconEU/agindu-n-eu* da eta ingelesezko *instruction* izenarena *LexiconEN/instruction-n-en.* Horiei lotutako kontzeptuak iturriko zenbakia gordetzen du (adibidez, *tzos/c24323-concept*), eta kontzeptu horrek era berean azpialor bat du lotuta; adibidez, 3 atalean aipatutako *agindu* izenaren kasuan, *tzos/c24323-concept* kontzeptua 3304 gisa sailkatuta dagoen azpialorrarekin lotuta dago. 6. irudian ikus daiteke nola errepresentatzen diren *agindu* lemaren izen kategoriako lau sarrerak *ontolex-lemo*n ereduari jarraituta.



6. irudia. TZOSeko *agindu* sarrera Ontolex-lemon ereduaren arabera landuta

Orain gure helburua da, ildo honetatik jarraituz, elkarlana bideratzea CSICen lan-esparrura lotzeko.

4. HARTA/TAILA: Testu Akademikoen Idazketarako Laguntza tresna elebiduna

Goi-mailako ikasketak dituen edozein ikaslek garatu behar du testu akademikoak idazteko gaitasuna. Gaitasun hori lortzeko gakoetako bat da konbinazio lexiko akademikoak menderatzea. Orain arteko irakaskuntza-esperientziak, ordea, erakutsi digu testu akademikoak idazteko orduan ikasleek zailtasunak dituztela horiek ondo erabiltzeko. Zailtasun horien zergatiak izan daitezke euskararen erregistro akademikoak garapenean egotea, ingeleseko baliabideak nagusitzea kontsultarako edota jasotzen duten irakasleen inputa askotan egokia ez izatea. Hain zuzen ere, zailtasun horiei aurre egiten laguntzeko eta fraseologia akademikoaren azterketak euskararen erregistro akademikoen garapenari buruzko datu garrantzitsuak eman ditzakeela kontuan hartuta garatu dugu HARTA/TAILA,[7]

[7] http://harta.ixa.eus/

lan akademikoak idazteko laguntza-tresna elebiduna. Tresna horretan kontsultagai daude HARTAeus corpusetik erauzi ditugun KLAk. Adibidez, 1. taulan ikus daitezke *helburu* izenak dituen kolokazioak HARTAeus corpusean eta 2. taulan, berriz, *Datuak aurkeztea* diskurtso-funtzioan sailkatu diren diskurtso-formulak.

1. taula. *Helburu* izenaren kolokazioak maiztasunaren arabera

Oinarria		**Kolokazio-lagunak**
Izena		**Aditzak**
helburu	ABS	bete, lortu, zehaztu (*helburuak bete*, *helburuak lortu*, *helburuak zehaztu*)
	DAT	erantzun (*helburuari erantzun*)
		Adjektiboak (eskuineko eta ezkerreko hedapenak)
helburu		nagusi, espezifiko, ekonomiko (*helburu nagusi*, *helburu espezifiko*, *helburu ekonomiko*) honako, hurrengo (*honako helburu*, *hurrengo helburu*)
		Izenak (GEN)
helburu		proiektu, lan, ikerketa (*proiektuaren helburu*, *lanaren helburu*, *ikerketaren helburu*)

2. taula. *Datuak aurkeztea* diskurtso-funtzioko formulak maiztasunaren arabera antolatuta

Diskurtso-funtzioa	Diskurtso-formulak
Datuak aurkeztea	ikusi da, lortutako emaitzak, behatu da, lortutako datuak, emaitzen arabera, lorturiko emaitzak, jasotako informazioa, emaitzak ez dira, emaitza da, ateratako emaitzak, jasotako datuak, bildutako informazioa, emaitzek erakusten dute, eskuratutako informazioa, lortutiko datuak, bildutako datuak, datuei erreparatuta, egindako azterketak agerian utzi du, eginiko azterketak agerian utzi du, egindako azterketak erakusten du, eginiko azterketak erakusten du.

HARTA/TAILA tresnan, erabiltzaileak aukera du diskurtso-funtzioen araberako bilaketa onomasiologikoa egiteko edo esamolde jakinaren araberako bilaketa semasiologikoa egiteko (ikus 7. irudia). Bilaketa hori euskaraz zein gaztelaniaz egin dezake.

7. irudia. HARTA/TAILA tresnaren interfazea: bilaketa onomasiologikoak ezker aldean eta semasiologikoak eskuin aldean

Diskurtso-funtzioak hiru multzotan sailkatuta daude HARTA/TAILAn, 3. taulan ikus daitekeen bezala. Testuaren egiturara, ikerketaren edukira eta igorlearen eta hartzailearen arteko harremanera bideratutako sailkapen hori García-Salido *et al.*ek (2019) eginda duten diskurtso-funtzioen sailkapen bera da, zeina, era berean, Biber *et al.*en (2004) eta Hyland-en (2008) sailkapenean oinarrituta dagoen. Guztira 39 diskurtso-funtzio dira.

3. taula. Diskurtso-funtzioen sailkapena

Testua egituratzea	Ikerketaren edukiari erreferentzia egitea	Iritzia eta ikuspegia ematea eta irakurleari zuzentzea
Informazioa gehitzea	Definitzea eta deskribatzea	Baieztapenak moteltzea
Mugatzea	Denominatzea	Beharra adieraztea
Adibideak ematea	Konparatzea	Ebaluazio bat adieraztea
Kausa adieraztea	Taldeak ezartzea	Zerbait azpimarratzea edo nabarmentzea
Baldintza adieraztea	Kantitatea adieraztea	
Ondorioa adieraztea	Datuen arteko korrelazioa adieraztea	Ziurtasuna adieraztea
Xedea adieraztea	Maiztasuna adieraztea	Iturria aipatzea
Aurkaritza adieraztea	Datuen balioen progresioa adieraztea	Posibilitatea adieraztea
Kontzesioa adieraztea	Denbora adieraztea	
Lanari erreferentzia egitea	Datuak aurkeztea	
Gai bat sartzea	Ikergaia aurkeztea	
Alternatiba bat aurkeztea	Hipotesiak aurkeztea	
Salbuespen bat sartzea	Metodologia aurkeztea	
Ordenatzea	Ondorioak aurkeztea	
Diskurtsoaren beste atal batera bideratzea	Helburuak aurkeztea	
Birformulatzea		
Laburbiltzea		

Hiru multzo horietako bakoitzean ageri diren diskurtso-funtzioak dira esleitu zaizkienak HARTAeus corpuseko diskurtso-formulei, eta baiezta dezakegu diskurtso-funtzio guztietarako formulak erauzi direla. Jarraian, diskurtso-funtzio horiek azalduko ditugu labur; azalpen hori egiteko, azpimultzotan bilduko ditugu eta diskurtso-funtzio bakoitzeko diskurtso-formula baten adibidea emango dugu parentesien artean idatzita.

— Testuaren egituratzean parte hartzen duten formulei esleitu zaizkien diskurtso-funtzioak 17 dira eta bost azpimultzo hauetan bil daitezke:

- Informazioa gehitu edo argitzen dutenak: informazioa gehitzea (*horrekin batera*), adibideak ematea (*esaterako*), birformulatzea (*hau da*) eta laburbiltzea (*hitz gutxitan esanda*).
- Aztergaiak kokatzeko helburua dutenak: batetik, gaia sartzea *(-z jardungo dugu*) eta alternatiba bat aurkeztea (*beste aukera bat da*) eta, bestetik, mugatzea (*barne hartzen ditu*) eta salbuespena sartzea (*alde batera utzita*).
- Ideien arteko erlazio logikoak adierazten dituztenak: kausa adieraztea (*horren arrazoia da*), baldintza adieraztea (*kontuan izanda*), ondorioa adieraztea (*ondorioa da*), xedea adieraztea (*hobetze aldera*), aurkaritza adieraztea (*baizik eta*) eta kontzesioa adieraztea (*izanda ere*).
- Antolaketarekin lotura dutenak: ordenatzea (*hurrenez hurren*).
- Erreferentziak egiteko erabiltzen direnak. Horien artean ditugu, alde batetik, lanari berari erreferentzia egitea (*lan hau*) eta, bestetik, diskurtsoaren beste atal batera bideratzea (*hurrengo taulan*).

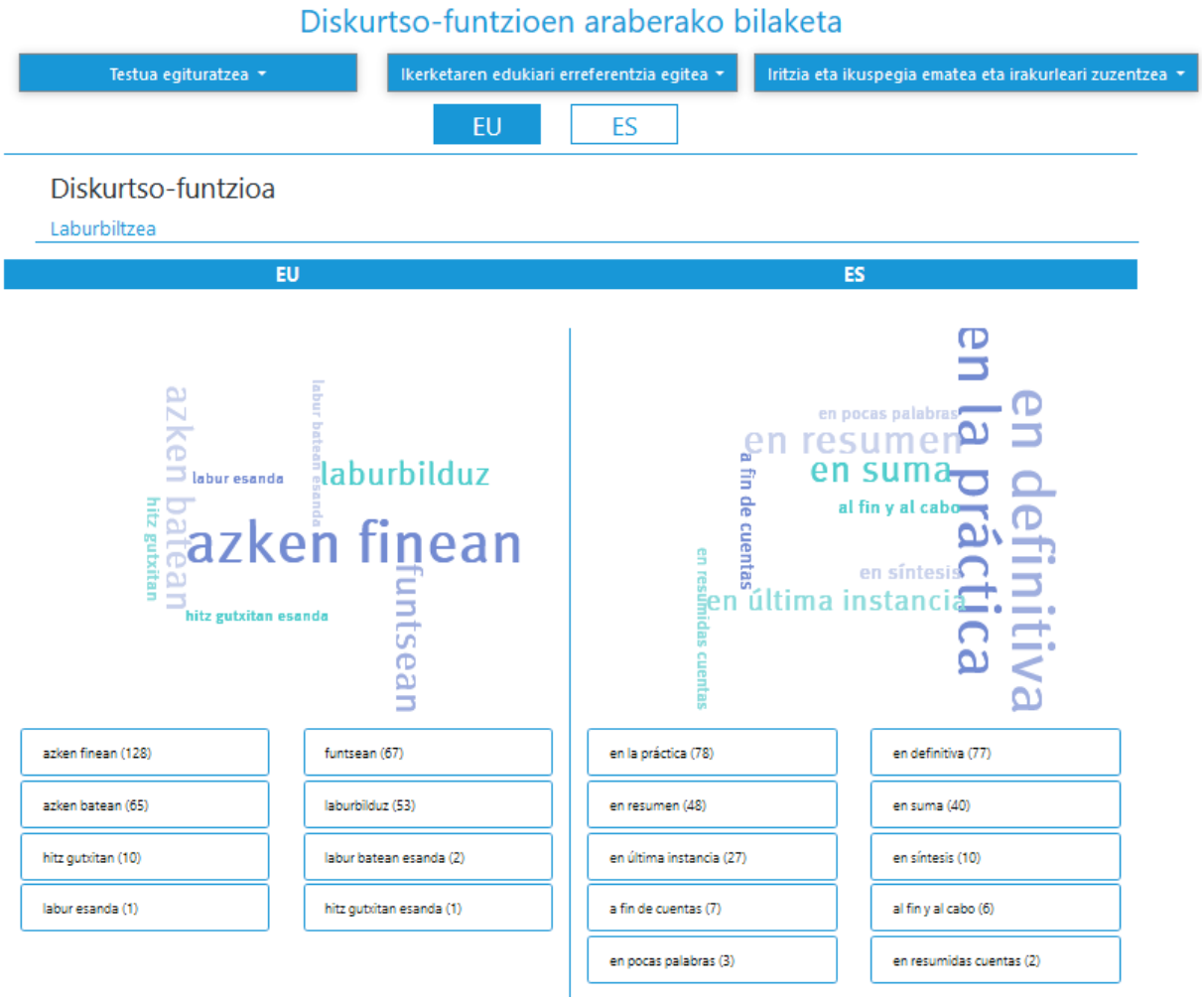
— Ikerketaren edukiari erreferentzia egiten dioten formulei dagozkien diskurtso-funtzioak 15 dira. Labur azaltzeko, honela multzokatu ditugu:

- Ikerketa-prozesuarekin lotutakoak: datuak aurkeztea (*emaitza da*), ikergaia aurkeztea (*zentratuko gara*), hipotesiak aurkeztea (*espero da*), metodologia aurkeztea (*metodo hau*), ondorioak aurkeztea (*agerian utzi du*) eta helburuak aurkeztea (*helburua da*).
- Deskribatzea helburu dutenak: definitzea eta deskribatzea (*ezaugarri moduan*), denominatzea (*esaten zaio*), konparatzea (*bestelakoa da*) eta taldeak ezartzea (*bat egiten dute*).
- Kopuruarekin zerikusia dutenak: kantitatea adieraztea (*batez beste*), maiztasuna adieraztea (*berriro ere*) eta denbora adieraztea (*lehen aldiz*).
- Datuen arteko erlazioarekin lotutakoak: datuen arteko korrelazioa adieraztea (*zenbat eta*) eta datuen balioen progresioa adieraztea (*gero eta*).

— Igorlearen eta hartzailearen arteko harremanarekin lotura duten diskurtso-funtzioak 7 dira. Multzo hauetan sartzen dira:

- Modalizazioarekin lotura dutenak: baieztapenak moteltzea (*gehiago edo gutxiago*), zerbait azpimarratzea edo nabarmentzea (*esan beharra dago*), ziurtasuna adieraztea (*jakina da*) eta posibilitatea adieraztea (*posible da*).
- Beharrarekin zerikusia dutenak: beharra adieraztea (*beharrezkoa da*).
- Ebaluazioarekin loturikoak: ebaluazio bat adieraztea (*baliagarria da*).
- Informazio-iturrien berri ematen dutenak: iturria aipatzea (*autore horiek*).

Diskurtso-funtzioen araberako bilaketan, bi hizkuntzak konpara daitezke, *Konparatu hizkuntzak* edo *Comparar idiomas* botoiari esker bi hizkuntzetako formula-zerrendak eta hitz-hodeiak erakusten direlako. Formula horiek maiztasun-erabilera handienetik txikienera ordenatuta ageri dira hizkuntza bakoitzean; adibidez, 8. irudian euskarazko *laburbiltzea* eta gaztelaniazko *resumir* diskurtso-funtzioan bildu diren formulak ageri dira.



8. irudia. *Laburbiltzea* diskurtso-funtzioko diskurtso-formulak euskaraz ezkerreko aldean eta gaztelaniaz eskuineko aldean

Gainera, formula bat aukeratuz gero, tresnak HARTAeus corpusetik ateratako adibide batzuk erakusten dizkio erabiltzaileari, baita datu kuantitatibo hauek ere: zenbatekoa den formularen banaketa lau hipereremu nagusietako bakoitzean eta testuetako ataletako bakoitzean (laburpena, sarrera, metodologia, gorputza, emaitzak, ondorioak). Oker erabili diren edota egokiagoak diren formulen berri ere ematen du. Adibidez, *adierazi* aditza oker erabilita dago *datuek adierazten dute* diskurtso-formulan, *adierazi* aditzak gertaera kontrolatzen duen subjektua behar duelako; beraz, zuzena da *erakutsi* aditza erabiltzea diskurtso-formula horretan.

Esamoldearen bidezko bilaketan, formula edo kolokazio jakin batez egin daiteke kontsulta. Esamoldearen lema idazten hasi orduko eskaintzen ditu tresnak lema hori duten esamolde guztiak; adibidez, *emaitza* hitza idatzita agertzen dira *emaitza* kolokazioa, eta *ateratako emaitzak*, *emaitza da*, *emaitzak ez dira*, *lortutako emaitzak*, *lorturiko emaitzak*, *emaitzek erakusten dute*, *emaitzek adierazten dute* eta *emaitzen arabera* formulak. Kolokazioa aukeratuz gero, kolokazio-interfazean kolokazio ohikoenak ikus daitezke beren maiztasunekin (adibidez, *emaitzak izan* (322 agerraldi), *emaitzarik lortu* (192 agerraldi), *emaitza on* (131 agerraldi) eta hitz-hodeian hitz bikoteek duten tamaina kolokazioaren maiztasunaren isla da. Horretaz gain, kolokazioaren patroi sintaktikoak ere eskaintzen ditu, hau da, oinarri hori duten kolokazio akademiko guztiak beren analisi sintaktikoarekin, agerraldi kopuruarekin eta corpuseko adibide batzuekin, adibidez, *emaitzetan ikusi* (inesiboa+aditza, 29 agerraldi).

Une honetan, HARTAeus corpusean maiz oker edo desegoki erabili diren formulen erabilera zuzenak edo egokiak zein diren azaltzen duen informazioa gehitzen ari gara HARTA/TAILAn. Adibidez, «aditza+bitartean» segidaz osatutako diskurtso-formula ez da zuzena aurkaritza adierazteko, denbora adierazten baitu.

5. Ondorioak eta etorkizuneko lanak

Garaterm egitasmoaren barruan hainbat tresna eta baliabide garatu ditugu alor akademikoetan erabiltzen diren testuak eta terminologia landu eta ikusgai egiten joateko (ikus 4. taula). Orain arte modu independentean erabiltzen ziren modulu batzuk (corpus-biltegia, testuen prozesaketa, termino-hautagaien balioztatze-ingurunea) Garaterm ingurune arin eta atseginean integratu dira. Hala ere, oraindik beste modulu batzuk integratzea falta da, eta aztertzeko dugu, dena integratuta dagoenean, zeintzuk izango diren erabiltzaile-profilak eta bakoitzari zer eta nola eskainiko zaion. Propotipo bat egina dugu, hala ere, HARTA-TAILA laguntza-tresnarekin (http://harta.ixa.eus/box), eta ideia da hori garatzen joatea.

4. taula. Garaterm lan-inguruneko tresnak eta baliabideak

Tresna eta baliabideak		Kopurua
Garaterm corpusa		25.322,23 hitz
HARTAeus azpicorpusa		3.285.098,00 hitz
TZOS		136.520 termino
IZIBI-TZOS		82.359
HARTA/TAILA	kolokazioak	1.005
	diskurtso-formulak	644

Bestalde, egungo tresna automatikoek, eta, zehazkiago, itzultzaile neuronalek, zalantzan jarri dute garai batean ezinbestean sortu behar ziren baliabide egituratu eta sofistikatuen beharra (hiztegiak, kontsulta-corpusak…). Hala eta guztiz ere, gure irakasle-esperientziaren eta eguneroko erabileraren arabera, justu alor akademikoetan egokiak diren hautu lexikalak egiterakoan, edota testu akademikoen eremuan dauden orotariko estilo eta erregistroak hautatzerakoan, emaitzak oraindik ez dira hain finak.

Horregatik, hemen aurkeztu ditugun baliabideekin ateratako emaitzek eta aurrerantzean aterako ditugunek, bi helburu dituzte: alde batetik, erabiltzaile adituei modu tradizionalean baliabide interesgarriak eskaintzea, eta, bestetik, tresneria automatikoa hornitzeko baliagarri izango diren datutegiak osatzea, horretan ere ikerketa-lerroak jorratuz; esate baterako, termino polisemikoen detekzio eta ebazpenerako, edota ohiko erroreen detekziorako eta zuzentzaile automatikoetan integratzeko.

Horiekin batera, helburuen artean ditugu alor akademikoan erreferentziazko lan-ingurunea izatea eta lankidetzak bilatzea, sare dinamiko nahiz kolaboratiboak sortzeko eta datuak elkarrera-gintasun-printzipioetan oinarrituta partekatzeko.

Eskertzak

Lan hau gauzatu da, alde batetik, Eusko Jaurlaritzak «Ixa Taldea. A motako Talde Finkatua» proiektuari emandako dirulaguntzari (IT-1570-22) esker eta, bestetik, Ministerio de Cien-

cia e Investigación erakundetik HARTAvas proiektuak jaso duen dirulaguntzari (PID2019-109683GB-C22) esker.

Bibliografia

Aldezabal, I., Arriola, J.M., y Otegi, A. (2022). TZOS: an Online Terminology Database Aimed at Working on Basque Academic Terminology Collaboratively. En N. Calzolari (Conference chair), F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, y S. Piperidis (Eds.). *Proceedings of the 13th Language Resources and Evaluation Conference* (pp. 1353–1359). European Language Resources Association (ELRA).

Alonso-Ramos, M., y Zabala, I. (2022). HARTAes-vas: Lexical combinations for an academic writing aid tool in Spanish and Basque. En M.A. Alonso, M. Alonso-Ramos, C. Gómez, D. Vilares,, y J. Vilares (Eds.). *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations, SEPLN* (pp. 22–25). CEUR Workshop Proceedings.

Altuna, O., Iñarra, M., Basurto, A., y Uranga, B. (2021). Hizkuntzen erabileraren kale-neurketa. Euskal Herria, 2021. Emaitza nagusiak. *BAT Soziolinguistika Aldizkaria*, 125 (4), 11–52.

Aranzabe, M.J., Gurrutxaga, A., y Zabala, I. (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. https://doi.org/10.26342/2022-69-8

Aranzabe, M.J., Aldezabal, I., y Zabala, I. (2022b). Recursos y Herramientas de Lingüística de Corpus y PLN para la Monitorización e Investigación de los Usos Académicos del Euskera.. [Presentación de poster]. III. workshop de INTELE (Infraestructura de Tecnologías del Lenguaje), Madrid, Spain.

Arregi, X., Arruarte, A., Artola, X., Zabala, I., y Lersundi, M. (2013). TZOS: An On-Line System for Terminology Service. En *Actualizaciones en Comunicación Social. Actas XIII Simposio Internacional. de Comunicación Social* (pp. 400–404). Centro de Lingüística Aplicada.

Biber, D., Conrad, S., y Viviana, C. (2004). If you look at....: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371–405.

Bosque-Gil, J., Montiel-Ponsoda, E., Gracia, J., y Aguado-de-Cea, G. (2016). Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain. En H. Erdman Thomsen, A. Pareja-Lora, y B. Nistrup Madsen (Eds.). *Proceedings of TKE 2016 the 12th International conference on Terminology and Knowledge Engineering* (pp. 136–146). Copenhagen Business School, CBS.

Chiarcos, C., McCrae, J., Cimiano, P., y Fellbaum, C. (2013). Towards open data for linguistics: Lexical Linked Data. En A. Oltramari, P. Vossen, L. Qin, y E. Hovy (Eds.). *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing* (pp.7–26). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31782-8_2

Frankenberg-Garcia, A, Lew, R., Roberts, J.C., Rees, G. P., y Sharma. N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31 (1), 23-39.

García Salido, M., García, M., Villayandre, M., y Alonso-Ramos, M. (2018). A Lexical Tool for Academic Writing in Spanish Based on Expert and Novice Corpora. En N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H.Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 260–265). European Language Resources Association (ELRA).

García-Salido, M., Garcia, M., y Alonso-Ramos, M. (2019). Identifying Lexical Bundles for an Academic Writing Assistant in Spanish. En G. Corpas Pastor, y R. Mitkov, R. (Eds). *Computational and Corpus-Based Phraseology. Lecture Notes in Computer Science*, vol 11755 (pp. 144–158). Springer, Cham. https://doi.org/10.1007/978-3-030-30135-4_11

Granger, S., y Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a need-driven online academic writing aid. *Lexicographica: International Annual for Lexicography*. 31 (1), 118–141.

Gurrutxaga, A., Saralegi, X,, Ugartetxea, S., y Alegria, I. (2005). Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa. *Euskera zientifiko-teknikoa: normalizazinotik homologazinora* (pp. 371–411). Mendebalde Kultur Alkartea.

Gurrutxaga, A., y Alegria, I. (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. En V. Kordoni, C. Ramisch, y A. Villavicencio (Eds.). *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World* (pp. 2–7). Association for Computational Linguistics.

Gurrutxaga, A., Alegria, I., y Artola, X. (2018). Caracterización computacional de la idiomaticidad: aplicación a la combinación nombre+verbo en euskera. En L. Ruiz Miyares (Ed.). *Estudios de Lexicología y Lexicografía. Homenaje a Eloína Miyares Bermúdez*. Ediciones Centro de Lingüística Aplicada.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Klyne, G., y Carroll, J. (2004). Concepts and Abstract Syntax. En B. McBride (Ed.). *Resource Description Framework (RDF)*. https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ World Wide Web Consortium (W3C).

Laurén, Ch., Myking, J., y Picht, H. (2002). Language and domains: a proposal for a domain dynamics taxonomy. *LSP and Professional Communication*, 2(2), 23–30.

Martín-Chozas, P., Vázquez, K., Calleja, P., Montiel, E., y Rodríguez, V. (2022).TermitUp: generation and enrichment of linked terminologies. *Semantic Web*, 13(6), 967–986.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., y Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. En I. Kosem, C. Tiberius, M. Jacubícek, J. Kallas, S. Krek, yV. Baisa (Eds.). *Proceedings of the 5th Biennial Conference on Electronic Lexicography (eLex) in the 21st century* (pp. 587–597). Lexical Computing CZ s.r.o.

Mel'cŭk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. En M. Everaert, Erik-Jan van der Linden, A. Schenk, y R. Schreu (Eds). *Idioms: Structural and Psychological Perspectives* (pp. 167–232). Lawrence Erlbaum Associates.

San Martin, I. (2013). Terminologia Sareak Ehunduz: unibertsitateko ikasgeletan erabiltzen den terminologia ikusgai egin nahi duen programa. En X. Alberdi, y P. Salaburu (Eds.). *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira* (pp. 20–32). UPV/EHUko Argitalpen Zerbitzua.

Universidad del País Vasco/Euskal Herriko Unibertsitateko Euskara Errektoreordetza. (2018). III Euskararen Plan Gidaria (2018-2022). Universidad del País Vasco/Euskal Herriko Unibertsitateko Argitalpen Zerbitzua. https://www.ehu.eus/documents/2660428/2968265/Euskararen-III-plan-gidaria.pdf/c1e54c15-f34e-0e1c-963a-22400f52a0ba?t=1552036765000

Villayandre, M. (2018). «HARTA» de noveles: un corpus de español académico. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, 5(1), 131–140.

Wüster, E. (1968). *The Machine Tool. An Interlingual dictionary of basic concepts*. Technical Press.

Zabala, I., San Martin, I., Lersundi, M., y Elordui, A. (2011). Graduate teaching of specialized registers in a language in the normalization process: Towards a comprehensive and interdisciplinary treatment of academic Basque. En S. Maruenda-Bataller, y B. Clavel-Arroita (Eds.). *Multiple voices in academic and professional discourse* (pp. 208–218). Cambridge Scholars.

Zabala, I., Lersundi, M., Leturia, I., Manterola, I., y Santander, G. (2013). GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. En X. Alberdi, y P. Salaburu (Eds). *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira* (pp. 98–114). Servicio Editorial de la UPV/EHU.

Zabala, I., Aldezabal, I., Aranzabe, M.J., Arriola, J.M., Gonzalez-Dios, I, y Lersundi, M. (2018). Corpus-driven Terminology Work for Describing Basque Academic Terminology: the Weaving Terminology Networks programme (TSE programme). [Presentación de poster]. EAFT Terminology Summit, Donostia-San Sebastián, Spain.

Zabala, I. (2019). The elaboration of Basque in Academic and Professional Domains. En L. Grenoble, P. Lane, y U. Røyneland (Eds.), I. Igartua, y L. Oñederra (Basque Eds.). *Linguistic Minorities in Europe Online*. De Gruyter Mouton. https://doi.org/10.1515/lme.9612443

Zabala, I., Aranzabe, M.J. y Aldezabal, I. (2021). Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera. *Círculo de Lingüística Aplicada a la Comunicación*, 88, 31–50.

Adimen Artifiziala Ikerketa Sozialerako: euskal hurbilpena

Artificial Intelligence for Social Research: the Basque Approach

Joseba Fernandez de Landa, Rodrigo Agerri

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU

joseba.fernandezdelanda@ehu.eus rodrigo.agerri@ehu.eus

Laburpena

Teknologia berrien etengabeko garapenak aldaketak eragin ditu gizakion arteko komunikazio moduetan. Horrela, geroz eta ohikoagoa da sare sozialak eguneroko bizitzan erabiltzea, inolako mugarik gabeko komunikazioa ahalbidetuz. Komunikazio-esparru birtual horrek hartu-eman publiko horiek jasotzeko aukera ematen du. Twitterren adibidez, testuan oinarritutako informazio mordoa dago eskuragarri. Komunikazio-esparru berri horren sorrerak eta berau ustiatzeko aukerek, teknika berriak beharko dituen ikerketa-eremu berri baterako bidea irekitzen dute. Aukera horri esker, euskararen etorkizunarekin erlazionatutako ikerketa egitea izango da asmoa, hizkuntza horren erabilera Twitterren aztertuz. Ikerketa hau gazteengan oinarrituko da, esparru berri horietan nagusi izateaz gain, etorkizuna baitira. Horretarako euskal erabiltzaile gazteengan zentratuko da ikerketa, horiek erabiltzaile guztien artetik automatikoki erauziz. Behin banaketa eginda, erabiltzaile horiek harremanak zeinekin dituzten azaleratzea izango da asmoa. Datu mordo horiek kudeatu eta aipatutako helburuak lortzeko ikasketa sakoneko metodo aurreratuak erabiliko dira, ikerketa sozialerako adimen artifiziala erabiliz.

Gako hitzak: sare sozialak, ikasketa sakona, gizarte zientziak, euskara.

Abstract

The continuous advance of new technologies has generated changes in the way of relating between humans. Therefore, it is increasingly common to use social networks in our day to day, allowing communication without limits. This new virtual space of communication allows to collect the ways of relating, on Twitter for example, we will have access to a lot of information based on text. The creation of this new communication space and the possibility of mining it, allows the creation of a new field of research that needs new research techniques. Thanks to this new opportunity, the intention will be to carry out an investigation related to the future of basque language, analyzing the use of this language on Twitter. This research is focused on young people, since they are the majority in these new spaces as well as being the future. The research will focus on Basque young users, automatically extracting them from the entire user base. Once the extraction is made, we will inquire about how these users relate among them. To manage all the massive data and conduct our experiments, we will utilize deep learning approaches, applying artificial intelligence to inquire into social research.

Keywords: social networks, deep learning, social sciences, basque language.

1. Sarrera

Sare sozialek sinesmenak, sentimenduak edo iritziak hainbat formatutan adierazteko aukera eskaintzen dute, testua, irudia, audioa eta bideo formatuak erabiliz. Gainera, gure argitalpenek norberaren egoera sozial, emozional eta arrazionalaren adierazpen kontziente edota inkontzienteak adierazten dituzte. Hori guztia gutxi balitz, edukia konpartitu, jendea jarraitu edo atsegindako argitalpenak bezalako ekintza sinpleek elkarrekin harremanetan jartzeko moduei buruzko informazio mordoa eskaintzen dute. Horregatik guztiagatik, gizarte ikerkuntzarako datu-iturri emankorra iruditzen zaizkigu sare sozialak. Datu-iturri erraldoi horietatik ganorazko informazioa lortzea ere ez da erronka erraza, egitura gabeko datu kantitate itzelak baitira.

Sare sozialak adierazpen librerako esparru bilakatu dira, nahi dena nahi den momentuan esateko aukera zabalduz eta hargatik gazteen esparru gisa kontsideratuz (Fernandez de Landa, 2017). Horrek aukera ematen du euskararen inguruko hausnarketa egiteko, eta batez ere, gazteen errealitatea ezagutzeko. Euskarak ere badauka bere tokia sare sozialetan, Twitterren (gaur eguneko X) bereziki eduki handia topatu dezakegu (Fernandez de Landa *et al.,* 2019).

Datu mordo hauek kudeatu eta ikerketa soziala egiteko asmoarekin ikasketa sakoneko metodo aurreratuak erabiliko ditugu, ikerketa sozialerako adimen artifiziala aplikatuz. Metodo horiek garatu eta portaera aztertzeko euskal gazteen komunitatea identifikatu eta aztertzea erabaki dugu. Horretarako, testu eta elkarrekintzetan oinarritutako datuak erabiliko ditugu, adimen artifiziala erabilita, erabiltzaileen informazio soziala erauzteko asmoz. Horrela, ikasketa sakoneko teknika aurreratuak aplikatuko ditugu eskala handiko datuetatik abiatuta, ezaugarri demografikoak iradoki eta komunitate-azterketa egiteko. Alde batetik, testuetan oinarritutako datuen bitartez erabiltzaileen bizitza-etapa zehaztea bilatuko dugu, aurre-entrenatutako hizkuntza-ereduak erabiliz. Bestalde, erabiltzaileen birtxioak erabiliko ditugu, ikasketa sakoneko hurbilpen ez-gainbegiratuen bitartez erabiltzaileen harremanak aztertzeko.

2. Aurrekariak

Soziolinguistikako lanen aburuz, ahoz edo idatziz komunikatzeko darabilgun estiloak norberaren ezaugarri demografikoak azalaraz ditzake, besteak beste, adin talde batekiko partaidetza (Nguyen *et al.,* 2016). Hizkuntza fenomeno soziala dela kontuan hartuta eta hizkuntzaren prozesamenduaren eremuan testu-kopuru handiak biltzeko eta prozesatzeko gero eta ahalmen handiagoari esker, gero eta egingarriagoa izango da soziolinguistika konputazionala. Twitterren erabilera hedatuak, hain zuzen, horrelako planteamenduei mesede egin die, orain posible baita testu-kopuru handiak lortu eta automatikoki aztertzea hizkuntzaren prozesamendua baliatuz.

Sare sozialetarako berariaz egokitutako hizkuntzaren prozesamenduko teknikak erabili izan dira sexua, adina edo kokapen geografikoa bezalako ezaugarri demografikoak iragartzeko (Cesare *et al.,* 2017; Morgan-Lopez *et al.,* 2017). Guretzat bereziki interesgarriak dira Twitterreko erabiltzaileen adinaren edo bizitzako etapen detekziorako egindako lanak. Helburu horrekin egindako lan horiek testuetan oinarritzen diren datu-multzo propioak sortzen dituzte 300-3.000 erabiltzaile artean eskuz etiketatuz nederlandera, ingelesa edo gaztelania bezalako hizkuntzetarako (Rao *et al.,* 2010; Al Zamal *et al.,* 2012; Nguyen *et al.,* 2013; Marquardt *et al.,* 2014; Morgan-Lopez *et al.,* 2017; Zaghouani & Charfi, 2018).

Bestalde, hainbat dira Twitter sare sozialeko edukiak konpartitzeko ekintzetan (birtxioetan) oinarrituta, komunitateak antzeman dituen ikerketak. Horien artean polarizazio politikoa azter-

tzeko (Conover *et al.,* 2021) eta afiliazio politikoa identifikatzeko (Pennacchiotti & Popescu, 2011) egindakoak nabarmendu genitzake. Ikerketa horietan erakusten da, erabiltzaileek egindako birtxioetan oinarrituz, badagoela komunitateak edo taldeak iragartzeko aukera. Beraz, lan zehatz honetan, metodologia antzekoak erabiliko ditugu, euskal komunitateak aurreikusteko, hau da, elkarren arteko harremanak nola gertatzen diren ikusteko.

Sare sozialetan, era berean, baliabide gutxiko hizkuntzen erabilera ikertzen duten ikerketalan desberdinak daude, hala nola galesera (Jones *et al.,* 2013), irlandera (Mhichíl *et al.,* 2018) eta frisiera (McMonagle *et al.,* 2018). Euskaraz era badira hainbat lan, jarreren hautematea (Agerri *et al.,* 2021) edo komunitate-azterketak (Fernandez de Landa *et al.,* 2019) egiteko baliatu izan direnak. Aipatutako lan horiek erakusten dute Twitterrek baliabide gutxiko hizkuntzetarako ere baduela testu-datuak eskaintzeko ahalmena, hizkuntza eta kultura ugari aurkitu eta aztertzeko aukera emanez.

3. Adin Tartearen Sailkapena

Euskal erabiltzaileen testuetatik beren bizitza-etapa automatikoki nola iradoki izango da atal honen aztergaia. Horretarako, ikasketa automatikoan oinarritutako hurbilpen gainbegiratua jorratuko dugu, aurrez etiketatutako datu-multzo batekin sailkatzaile bat entrenatu eta ebaluatzean oinarrituko dena. Helburu horrekin, lehenik eta behin, datu-multzo propio bat nola sortu zehaztuko dugu, erabiltzaile euskaldunen eduki testualen bizitza-etapa zehaztuko duena. Horretarako, idazketa estiloan oinarrituz, erabiltzaile baten eduki zehatza helduena edo gazteena den etiketatuko dugu metodo erdi-automatiko bat erabiliz. Euskararako propio sortutako datu-multzo horrekin esperimentu ezberdinak egingo ditugu, testu sekuentziak gazte eta heldu artean automatikoki sailkatzeko intentzioarekin. Horretarako, besteak beste, euskara barne daukaten aurre-entrenatutako hizkuntza-eredu elebakar eta eleaniztunak erabili eta ebaluatuko ditugu. Azkeneko pauso gisa, sailkatzaile arrakastatsuena erabiliko dugu euskal erabiltzaileak etiketatzeko. Horretarako, testu-sekuentzia mailatik erabiltzaile mailarako sailkapena egingo dugu.

3.1. Datuak

Txioak gazteenak edo helduenak diren ezberdinduko dituen gazte-heldu sailkatzaileak entrenatu eta ebaluatzeko, datu-multzo propioa sortuko dugu. Horretarako, idazkera estiloa aintzat hartzen duen metodo erdi-automatiko hau proposatzen dugu:

1. Lehenik eta behin, *Heldugazte-osoa* (Fernandez de Landa *et al.,* 2019) corpuseko erabiltzaileen 6M txioak automatikoki etiketatu ditugu idazkera estiloaren arabera informal-formal sailkatzailea erabilita (Fernandez de Landa *et al.,* 2019).
2. Bigarrenik, erabiltzaileak euren denbora-lerroko txio informalen proportzioaren arabera ordenatu ditugu. Mutur bateko erabiltzaileek txio informalak izango lituzkete batez ere, eta beste muturreko erabiltzaileek txio formalak.
3. Hirugarrenik, muturreko 100 erabiltzaileen (50 informalenak eta 50 formalenak) denbora-lerroen eskuzko ikuskapena egin dugu. Urrats horretan bereziki lagungarria izan da ikuskapena erabiltzaile mailan egitea, denbora-lerroak erabiltzailea ezaugarritzeko testuinguruko informazio gehiago eskaintzen duelako. Eskuzko azterketa horrek egiaztatzen du erabiltzaile gazte eta helduen etiketatze erdi-automatikoaren emaitzak onargarriak direla (1. taulako adibidea).

4. Laugarrenik, sailkapenaren mutur informalenean dauden 500 erabiltzaileak erabiltzaile gazte kontsideratuko ditugu eta mutur formalenean dauden 500 erabiltzaileak heldu kontsideratuko ditugu. Proposatutako metodo berri horri esker, gazte eta heldu gisa anotatutako 1.000 erabiltzaile lortu ditugu.

Erabiltzaile mailako anotazio erdi-automatikoa egin ostean, txio mailako datuak lortzera igaroko gara berriz ere. Erabiltzaile mailatik txio mailara igaroz, erabiltzaileak datu kopuru berdinarekin ordezkatu nahi ditugu, datu-multzo orekatu bat lortuz. Horretarako, erabiltzaile bakoitzeko, ausazko 80 txio aukeratu ditugu, txio bakoitza erabiltzaileari egotzitako heldu edo gazte etiketarekin anotatuz. Era horretan datu-multzo esanguratsu eta heterogeneo bat sortu dugu, txio indibidualak gazte edo heldu gisa anotatuta dituena.

1. taula. Heldugazte-Age datu-multzoko adibideak

Etiketa	Edukia (txioa)
heldu	— Taldeak mikel laboaren lanean oinarritu du bere hurrengo diskoa. — Gure herriko ateak zabalik dituzu.
gazte	— Buaa q follaa eun guztia eon zea ikasi ordez jolasateeenn jajaja. — Batzutan ze gutxi aguantatze zaituten.

Emaitza gisa Heldugazte-Age datu-multzoa edukiko genuke. Datu-multzo handi eta orekatu horrek 80K txio dauzka gazte edo heldu gisa etiketatuta (ikusi 2. taula). Datu-multzoa osatzen duten txioen adibide bat 1. taulan ikus daiteke. Datuak entrenamendu, garapen eta ebaluazio multzoen arabera banatu ditugu esperimentuetarako. Horrela, klase bakoitzeko 24K txio daude eskuragarri entrenamendurako eta 8K txio garapen zein ebaluaziorako, hurrenez hurren.

2. taula. Heldugazte-Age datu-multzoaren ezaugarriak.
Erabiltzaile mailan etiketatutako txioak bizitza-etaparen arabera

	Entrenamendua	Garapena	Ebaluazioa	totala	Erabiltzailea
Gazte	24.000	8.000	8.000	40.000	500
Heldu	24.000	8.000	8.000	40.000	500
Totala	24.000	16.000	16.000	80.000	1.000

3.2. Metodologia

Atal honetan bizitza-etapa identifikatzeko erabiltzen diren ikasketa automatikoko bi arkitektura nagusiak aurkeztetuko ditugu: (i) ezaugarri testualetan oinarritutako hurbilpen lineala (Agerri *et al.*, 2014), eta (ii) aurre-entrenatutako hizkuntza-ereduak (Devlin *et al.*, 2019; Agerri *et al.*, 2020).

Ezaugarri testualetan oinarritutako hurbilpen lineala: IXA pipes

IXA pipes (Agerri *et al.*, 2014) metodoak datuetatik ezaugarri testualak erauzi eta errepresentazio lokalak aberastean oinarritzen da. Sistema horrek entrenamenduko datuetatik eratorritako informazio lokala konbinatzen du etiketatu gabeko testuetatik induzitutako ezaugarrien klusterrekin. Era horretan, entrenamendu datuetako hitzak, hiru errepresentazio modu ezberdinen

konbinaketarekin ordezkatuko dira: Brown (Brown *et al.,* 1992) klusterrak, Clark (Clark, 2003) klusterrak eta word2vec (Mikolov *et al.,* 2013) klusterrak. Hitz bakoitza, aipatutako tekniketatik eratorritako klusterren konbinaketarekin ordezkatzen da. Horretarako, sekuentziako hitzak klusterretako lexiko bakoitzean dauden hitzekin mapatzen dira. Hitzen errepresentazioetarako aipatutako hiru teknika ezberdinetatik (Brown, Clark eta word2vec) eratorritako klusterrak pilatu eta konbinatzen dira. Klusterren ezaugarri horiek, hitz bakoitzari talde batekiko kidetasuna ematen diote, entrenamenduan ikusi gabeko hitzak, ikusitakoekin erlazionatzen dira kluster berdinean azalduz gero. Horrela, eskuz etiketatu beharreko datu kopuru handiekiko dependentzia arindu egiten da, etiketatutako datu-multzo txikiekin ere sailkapen egoki bat egitea ahalbidetuz (Agerri *et al.,* 2014). Gerora, klusterretatik erauzitako ezaugarriak, pertzeptroi (Collins, 2002) sailkatzaile baten sarrera datu bezala erabiltzen dira, ikasketa automatiko bitartez klaseak iragartzeko. Metodo horrek emaitza onak lortu ditu hainbat atazatan, hala nola izendun entitateen identifikazioan (Agerri & Rigau, 2016) zein iritzi erauzketan (Agerri & Rigau, 2019) hainbat hizkuntzatarako, euskara barne.

Aurre-entrenatutako hizkuntza-ereduak: mBERT eta BERTeus

Hizkuntzaren prozesamenduko beste zeregin askotan bezala, testu sailkapeneko atazetan ere errendimendurik onena erakusten duten sistemak aurre-entrenatutako hizkuntza-ereduak dira (Devlin *et al.,* 2019; Liu *et al.,* 2019). Hizkuntza-eredu hauek hitzen errepresentazio aberatsak sortzeko ahalmena daukate, horretarako testuingurudun hitz-bektoreak erabiliz. Hizkuntzaren prozesamenduko metodoak hitzen edo hitz-sekuentzien errepresentazioak egitean oinarritzen dira, hitzetatik zenbakizko errepresentazioetara joz, gerora sailkatzaileak zenbakizko datuekin elikatzeko asmoarekin. Orain arteko planteamenduek hitz-bektore estatikoak proposatzen dituzte (Mikolov *et al.,* 2013; Bojanowski *et al.,* 2017), hots, hitz jakin baterako bektoreetan oinarritutako errepresentazioak eskaintzen dira, baina, hitzaren errepresentazioa gertatzen den testuingurutik independentea da. Horrek esan nahi du ezin dela polisemia adierazi. Beraz, «banku» hitza kontuan hartzen badugu, hitz-bektore estatikoek errepresentazio bakarra sortuko dute, nahiz eta hitz horrek adiera desberdinak izan, hots, «finantza erakunde», «eserleku», etab.

Arazo horri aurre egiteko, testuingurudun hitz-bektoreak proposatzen dira, Flair, ELMO edo BERT adibidez (Akbik *et al.,* 2018; Peters *et al.,* 2018; Devlin *et al.,* 2019). Era horretan, hitz-bektore errepresentazioak sortzen dira baina hitza sortu den testuingurua ere barneratuz. Egun, testuingurudun hitz-bektore errepresentazioak sortzeko planteamendu ezberdinak daude, baina testuen sailkapenean eragin zuzena dutenetan zentratuko gara, hots, Transformer arkitekturan (Vaswani *et al.,* 2017) oinarritutako ereduetan, eta bereziki, ezagunenetakoa den BERT (Devlin *et al.,* 2019) ereduan.

Transformer arkitektura baliatuta, sekuentziako hurrengo hitza iragarri ordez, BERT ereduak sekuentziako hitz guztiak hartzen ditu kontuan, eta horrela testuinguruaren ulermen sakonagoa garatzen da. BERT-ek etiketarik gabeko testuaren noranzko biko irudikapenak aurre-entrenatzen ditu geruza guztietan ezkerreko zein eskuineko testuingurua kontuan hartuta, hurrengo esaldien iragarpena (*Next Sentence Prediction*) eta hizkuntza maskaratuaren modelizazioa (*Masked Language Modeling*) erabiliz. Era horretan testuingurua hitzen eta esaldien arteko asoziazioak barneratuz ikasten da, informazio gehiago txertatuz.

Gure ataza zehatza euskarazko testu sekuentziak sailkatzean oinarritzen denez, euskara barne hartzen duten BERT ereduak erabiliko ditugu. Bi izango dira erabili eta alderatuko ditugun ereduak: (a) mBERT eredu eleaniztuna (Devlin *et al.,* 2019) eta (b) BERTeus (Agerri *et al.,* 2020) euskarazko eredu elebakarra.

mBERT eredua, BERT ereduaren bertsio eleaniztuna da, Wikipediako 104 hizkuntza handienekin aurre-entrenatua dagoena. Eredu eleaniztun hauek oso ondo funtzionatzen dute baliabide handiko hizkuntzekin erlazionatutako zereginetan, esate baterako, ingelesarekin edo gaztelaniarekin. Hala ere, baliabide gutxiko hizkuntzak ez daude behar bezala ordezkatuta hizkuntza-eredu erraldoi hauetan (Agerri *et al.,* 2020). Besteak beste, entrenamendurako corpusean hizkuntza txikiek ingelesa edo gaztelania bezalako hizkuntzek baino datu gutxiago dauzkate (Devlin *et al.,* 2019; Conneau *et al.,* 2019). Horrez gain, badirudi eredu eleaniztunek emaitza hobeak dituztela antzeko egitura duten hizkuntzekin sortuak direnean (Karthikeyan *et al.,* 2020).

BERTeus: eredua, BERT arkitekturan oinarritutako euskarazko hizkuntza eredu elebakarra da. Euskara barne hartzen duen eredu baten ordez, euskararako berariaz prestatutako eredu propioa da. Agerri *et al.* (2020) lanean erakusten dutenez, euskarazko BERT eredu elebakarra entrenatzeak emaitza hobeak lortzen ditu bertsio eleaniztunak baino. Hortaz, euskarazko testua sailkatzea helburu duen gure atazarako ere, eredu honekin egingo ditugu esperimentuak.

3.3. Esperimentuen ezarpenak

Ataza zehatza testu sekuentzia baten egilearen bizitza-etapa (gazte/heldu) iradokitzean oinarrituko da. Fernandez de Landa *et al.* (2019) lanean ez bezala, oraingo honetan ataza ez da idazkera estiloan oinarrituta egongo. Hala ere, testu-sailkapen ataza berdina da: sarrera datu gisa testu sekuentzia (txio bakarra) bat emanda, *gazte* edo *heldu* etiketa iragartzea da asmoa. Gure esperimentuak egiteko, *Heldugazte-Age* entrenamendu-multzoa entrenamendurako erabili dugu eta test-multzoa ebaluatzeko. Sarrera datu diren txioetan gutxieneko aurreprozesaketa egiten dugu; URLak, hashtag-ak eta erabiltzaile-izenak kentzen ditugu, etiketa-txio bikoteak utziz, 1. taulan azaltzen diren adibideetan erakusten den bezala.

3.1. atalean sortutako *Heldugazte-Age* datu-multzoa, beraz, *3.2.* atalean aurkeztutako hiru testu sailkatzaile ezberdin trebatzeko erabiliko dugu: (i) IXA pipes (Agerri *et al.,* 2014), (ii) mBERT (Devlin *et al.,* 2019) eta, (iii) BERTeus (Agerri *et al.,* 2020). IXA pipes aukeratu dugu oinarri-lerro gisa txio informal eta formalak sailkatzeko lanean lortutako emaitza altuengatik (Fernandez de Landa *et al.,* 2019). Ataza zehatz horretan sistema entrenatzeko, aipatutako lanean erabilitako ezarpen berdinak erabili ditugu. Horrez gain, mBERT eta BERTeus-en errendimenduak alderatuko ditugu bizitzako etapa detektatzeko atazan, eredu eleaniztun eta elebakarren portaerak neurtzeko asmoa baitugu hizkuntza gutxituen arloan. Bi ereduentzat oinarrizko birdoitze hiperparametro berberak erabili ditugu (Agerri *et al.,* 2020).

3.4. Ebaluazio emaitzak

3. taulan aurreko atalean deskribatutako sistemak ebaluazio multzoaren gainean erabilita lortzen diren emaitzen berri ematen dugu. Lehenik eta behin, azpimarratu beharra dago aukeratutako metodo guztiek emaitza altuak lortzen dituztela, 0,95-etik gorako asmatze-tasa zein F1 balioak erdietsiz. Gainera, sistemen arteko aldeak ez dira horren handiak, nahiz eta BERTeus-ek emaitza onenak lortu dituen. Oinarri-lerro gisa hautatutako IXA pipes metodoak mBERT-ek bezain emaitza onak lortu ditu, metodo horren fidagarritasuna frogatuz beste behin ere. Bestalde, aurre-entrenatutako hizkuntza-ereduei erreparatuz, BERTeus eredu elebakarrak mBERT eredu eleaniztunak baino emaitza hobea lortu duela ikusi dugu. Horrek erakusten du, hizkuntza zehatzetan oinarritzen diren lanabesak garatzea beharrezkoa dela, batez ere hizkuntza gutxituetarako (Agerri *et al.,* 2020).

3. taula. Ebaluazio-emaitzak Heldugazte-Age (gazte-heldu) test-multzoan

Sistema	Asmatzea	Doitasuna	Estaldura	F1 Score
IXA pipes (Agerri *et al.*, 2014)	0,956	0,977	0,935	0,955
mBERT (Devlin *et al.*, 2019)	0,955	0,972	0,936	0,954
BERTeus (Agerri *et al.*, 2020)	0,963	0,968	0,958	0,963

Sailkatzaileek lortutako emaitza altuen aurrean, giza-ebaluazio baten beharra ikusi dugu, lagin baten eskuzko anotazio baten bitartez egingo dena. Eskuzko azterketa horretarako, datu-multzoko test-sortatik ausaz aukeratutako 200 txio eskuz etiketatzea erabaki dugu. Bi giza anotatzailek 200 txioak etiketatu dituzte, 0,78 puntuko anotatzaileen arteko adostasuna eta 0,55ko Kappa-scorea lortuz. Zenbaki hauek erakusten dute anotatzaileen arteko adostasuna moderatua izan dela, atazaren zailtasuna agerian utziz. Gainera, bi anotatzaileen asmatze-tasa 0,795 eta 0,775 puntukoa izan da hurrenez hurren. Puntuazio hauek 3. taulan jasotako sistema automatikoen emaitzekin alderatzean, argi uzten dute txio mailako gazte edo heldu etiketak eskuz esleitzea oso lan zaila dela. Anotatzaileen arteko desadostasun eta asmatze-tasa baxuek erakusten dute Heldugazte-Age datu-multzoa lortzeko proposatutako metodoaren (3.1. atala) eraginkortasuna, gizakiek kostata egin dezaketena eta gure metodoarekin zehaztasun eta erraztasun handiagoz egiten dena.

3.5. Aplikazioa

Orain arte, erabiltzaile gazteak identifikatzeko helburua duen atal honetan, testu-sekuentziak sailkatzeko baliagarria den gazte-heldu sailkatzailea proposatu dugu. Hala ere, sistema hori tamaina txikiko (txioak: 240 karaktere baino gutxiago, bat edo bi esaldi) testu-sekuentziak sailkatzeko entrenatuta dago. Erabiltzaileen adierazpen zehatz edo txioak sailkatzeko pentsatuta dago, hau da, soilik erabiltzaile batek sortzen duen edukiaren zati txiki bat sailkatzeko. Hortaz, esaldi edo txio mailatik, dokumentu edo erabiltzaile mailako sailkapen bat ematera igaro beharko gara. Hau da, erabiltzaile baten publikazio zehatzak sailkatuta, erabiltzaile mailako sailkapen orokorra lortu beharko da. Txio mailako sailkapenetik, erabiltzaile mailara igarotzeko bi pauso hauek emango ditugu: (i) txioak banan-banan sailkatuko ditugu gazte-heldu sailkatzailea erabilita; (ii) erabiltzailearen txio guztien etiketak kontuan hartuta erabiltzailea gazte edo heldu gisa etiketatuko dugu.

Erabiltzaile bakoitzaren txioak sailkatzeko, gazte-heldu testu-sekuentzia sailkatzailea erabiliko dugu. Hau da, Heldugazte-age datu-multzoarekin birdoitutako BERTeus sailkatzailea erabiliko dugu erabiltzaileen txioak banan-banan automatikoki etiketatzeko. Horretarako, gutxienez euskaraz idatzitako 10 txio pertsonal dauzkaten erabiltzaileen denbora-lerroak erabili ditugu *Heldugazte-osoa* corpusetik arakatu ditugun 7.980 erabiltzaileetatik 7.087 erabiltzaile aukeratuz.

Erabiltzaile bakoitzaren txioak banaka etiketatu ostean txioak informal/formal edo gazte/heldu gisa etiketatuta egongo dira. Txio mailatik erabiltzaile mailara etiketak proiektatzeko etiketen kontzentrazioan oinarrituko gara. Erabiltzaile mailako etiketatze automatikoa txioen etiketen kontzentrazioan oinarriturik egingo denez, kontzentrazioaren balioa ezarri beharko dugu atalase moduan. Era horretan, denbora-lerro zehatz batean txioen %60 *gazte* gisa etiketatuta badago, erabiltzailea gaztetzat joko dugu. Bestalde, denbora-lerro baten txioen %40 soilik *gazte* gisa etiketatuta badago, erabiltzailea heldutzat joko dugu.

Teknika hori aplikatuta, 1.635 erabiltzaile gazte identifikatu ditugu. Bestalde 4.472 erabiltzaile heldu gisa etiketatu ditugu.

Azkenik 980 erabiltzaile etiketa gabe utzi ditugu, txio gazte eta helduen kontzentrazioa antzekoa baitzen. Emandako emaitzen azterketa kualitatiboa egin dugu ausazko lagin baten eskuzko azterketa egiteko. Jarraian adibide bat ikus daiteke, gazte gisa etiketatutako bi erabiltzaileri dagokiena. Euren txioak ikusita, badirudi erabiltzaileak gazteak direla idazketa estiloan oinarrituta, baina baita azterketei buruz hitz egiten dutelako ere, oro har gazteei lotutako jarduera baita.

— @1erabiltzailea:

- Horrelakoekin gustua ta guzti hartzen zaio ikasteari.
- Buenobueno ba ikasiko dut gehio jaja ta ikusikozu gaindituko dutt jaja.

— @2erabiltzailea:

- Ze txupi txatxi no me da la nota.
- Ai naiz rayatzen pixkat asko con la mierda de la uni.

4. Komunitateen identifikazioa

Atal honetan sareko euskal gazte erabiltzaileen artean gertatzen diren harremanak aztertuko ditugu. Abiapuntua aurreko atalean *gazte* gisa sailkatutako 1.635 erabiltzaileek egindako euskarazko birtxioak izango dira. Horrela, identifikatutako euskal erabiltzaile gazteek beren denboralerroan konpartitutako edukiak jasoko ditugu. Birtxioak aukeratu ditugu erabiltzaileen arteko interakzio ekintzak direlako eta aipamenak bezalako elkarrekintzek baino hobeto erakusten dutelako erabiltzaileen arteko korrelazioa (Conover *et al.,* 2021). Azterketa egiteko, birtxio bakoitzetik ateratako bi ezaugarri erabiliko ditugu: (i) publikazioa konpartitzen duen erabiltzailea edo birtxiokatzailea (iturburu) eta (ii) publikazioa sortu duen erabiltzailea (helburu) edo birtxiotua. Aukeratutako ezaugarri horiek erabiltzaileen arteko harremana edo erlazioa zehazten dute. Hau da, erabiltzaile zehatzen interakzio konkretuak kontuan hartuta, lagin guztiaren harreman dinamikak azaleratzea izango da asmoa.

Zehazki, 1.635 erabiltzaile gazteren 418.903 birtxioetatik 24.837 nodo eta 148.304 konexio atera ditugu. Nodoak birtxioak egiten dituzten erabiltzaileei dagozkie (gure 1.635 erabiltzaileko lagina) baina baita horiek jasotzen dituzten erabiltzaile ezberdinei ere (gure laginekoak izan edo ez). Bestalde, konexioek adierazten dute iturburu-erabiltzaile batek beste helburu-erabiltzaile bat behin edo gehiagotan birtxiokatu duen.

Erabiltzaileen harremanak sakonean aztertzeko interakzioak baliatuko ditugu, erabiltzaileak harremanen arabera kokatu eta antolatzeko. Antolaketa horretarako gainbegiratu gabeko metodoak erabiliko ditugu, interakzioak baliatuta erabiltzaileen errepresentazio dentsoak sortzeko. Errepresentazio horiek lortzeko hurbilpen ezagun eta eraginkorrenen artean, DeepWalk (Perozzi *et al.,* 2014) eta node2vec (Grover & Leskovec, 2016) dauzkagu (Jusup *et al.,* 2022; Ma *et al.,* 2023). Metodo horiek erabiltzaileak ordezkatzen dituzten dimentsio baxuko ezaugarriak sortzen dituzte, horretarako etiketatu gabeko datu kopuru handiak baliatuz. Node2vec (N2V) algoritmoak erabiltzaileen errepresentazioak ikasten ditu sareko nodoen artean ausazko ibilaldi uniformeak simulatuz. Skip-gram algoritmoan oinarrituta (Mikolov *et al.,* 2013b), instantzia bat emanez inguruko nodoak aurreikusten dira. Instantzia bakoitzarentzat aurreikusi beharreko nodoak ausazko-ibilaldiek markatzen dituzte. Horrela, testuingurua sortzeko, instantzia bakoitzeko ibilaldien luzerak eta kopuruak zehaztuko du instantzia horren testuingurua. Horrez gain, DeepWalk (Perozzi *et al.,* 2014) metodoak ez bezala, sarearen egitura kontrolatzeko itzulera (p) eta sarrera-irteera (q) parametroak gehitzen dira. Itzulera parametroak (p) ausazko-ibilaldietan bisitatutako puntuetara itzul-

tzeko probabilitatea kontrolatzen du, balio altuagoetan nodo bat berriro bisitatzeko probabilitatea gutxitzen da. Sarrera-irteera parametroak (q) nodo urrunak arakatzeko probabilitatea kontrolatzen du, balio altuak nodo urrunekin erlazionatzen dira.

Beraz, euskal erabiltzaileak ordezkatuko dituen eredua sortzeko, erabiltzaile gazteek konpartitutako edukiak erabili ditugu, hots, birtxioak. Eredua sortzeko birtxiokatzaile-birtxiotu pareak erabili ditugu sarrera-datu gisa. Horrela, N2V gure datuetan aplikatzeko, hiperparametro balio lehenetsiak ezarri ditugu: *walks_per_node = 10* (ibilaldi kopurua), *walk_length = 80* (ibilaldi luzera), *window or context_size = 10* (testuinguru tamaina), eta optimizazioa *epoc* bakarrean exekutatzen dugu (Perozzi *et al.,* 2014; Grover & Leskovec, 2016). Bestalde, itzulera eta sarrera-irteera parametroak, azpi-komunitateen inguruko informazio zehatzagoa lortzeko finkatu ditugu $p = 1$ eta $q = 0{,}5$ balioak aukeratuz (Grover & Leskovec, 2016).

Euskal erabiltzaileek konpartitutako edukia erabilita N2V eredu bat entrenatu dugu, interakzioetan oinarriturik, erabiltzaile bakoitza dimentsio anitzeko espazioko puntu batean kokatuz. N2V eredua sortu ondoren azpitaldetan banatu dugu, euskal erabiltzaile gazteen baitako azpikomunitateak edo azpitaldeak nola eratzen diren aztertzeko. Horretarako, lortutako eredua lau kluster ezberdinetan zatitu dugu, azpitalde kopurua modu kualitatiboan aukeratuz. Kluster bidezko zatiketak, modularitatean oinarritutako algoritmoak (Blondel *et al.,* 2008) ez bezala, atera beharreko komunitate kopuru zehatza hautatzeko aukera ematen du. Ereduaren bistaraketarako erabilitako 1. irudiak erakusten du N2Vek argi eta garbi bereizten diren komunitateak sortzen dituela, eta horrek, aldi berean, interpretagarriak egiten ditu, erabiltzaileen artean dauden harremanak ulertzea erraztuz.

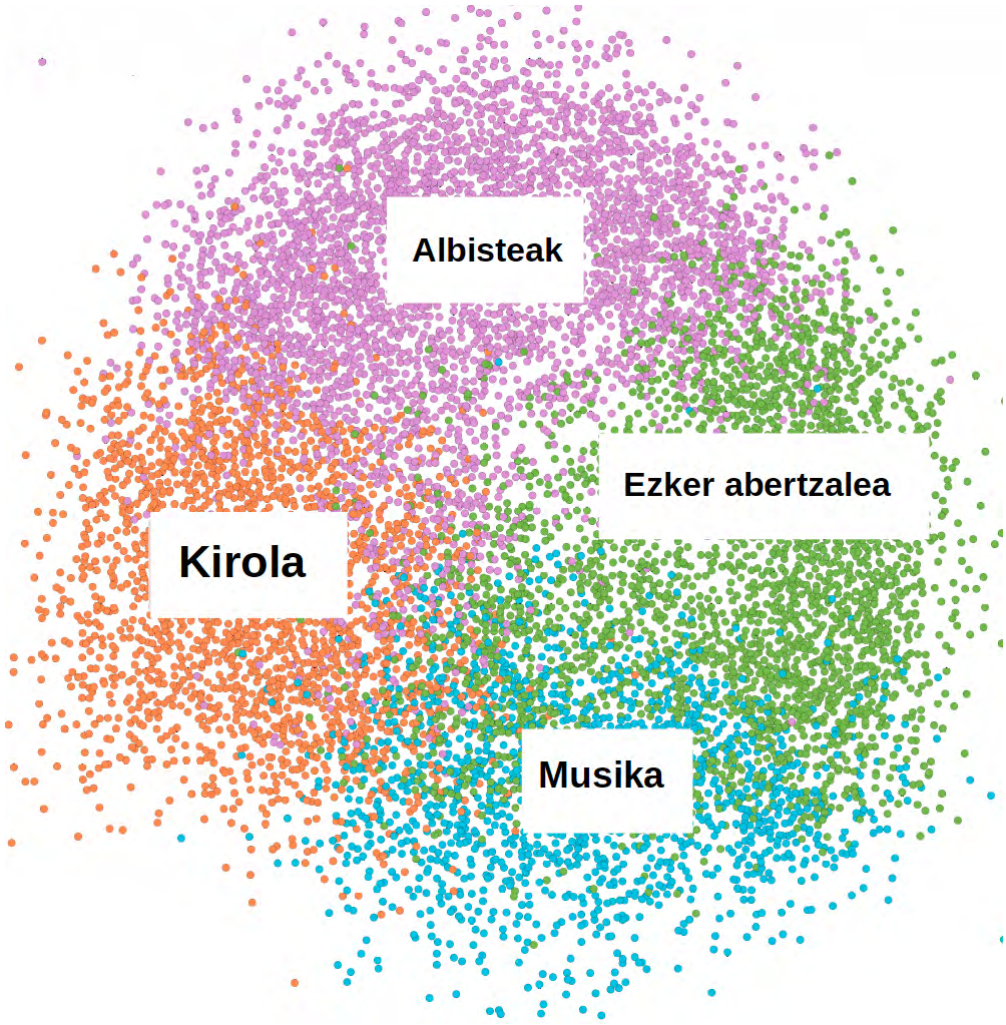
Erabiltzaileak irudikatzen dituen N2V errepresentazioa lau komunitatetan banatu ondoren, azpitalde bakoitzaren ezaugarri nagusiak ondorioztatu ditugu. Prozesu horretarako, komunitate bakoitza ordezkatzen duten nodo garrantzitsuenetan oinarritu gara, hau da, gehien konpartituak izan diren erabiltzaileetan jarriko dugu fokua. Gerora, erreferentzialak diren erabiltzaile horien nolakotasuna aztertuko dugu modu kualitatiboan, gaien arabera ordenatuz. Erabiltzaileak aztertuta, komunitatearen identitatea markatzen duen ezaugarri orokor bat esleituko zaio azpitalde bakoitzari. Gai horiek desberdinak dira grafikoko azpitalde bakoitzean, komunitate bakoitzaren ezaugarriak edo desberdintasunak erakutsiz. Jarraian, grafikoan jasotako lau azpitaldeetako bakoitzaren ezaugarri nagusiak deskribatuko ditugu.

— **Albisteak** (% 29,96): Komunitate honetan, sailkapenaren buruan aurkitzen diren nodoak aurkitzen dira, Euskal Herriko komunikabide eta gaurkotasunarekin zuzenean erlazionatutako erabiltzaileek osatuta. Horrela, Euskal Herriko komunikabideak (@berria, @argia, @HamaikaTb, @eitbAlbisteak, @euskaltelebista, @zuzeu, @euskadi_irratia, @Gaztezulo, @Sustatu, @eitbeus...) eta euskal kazetariak (@MaddalenIriarte, @boligorria, @urtziurkizu, @zaldieroa, @bzarrabeitia, @AneIrazabal...) dira komunitate horretako nodo erreferenteak. Horrez gain, euskara edo Euskal Herriarekin erlazionatutako edukia sortzen duten erabiltzaile aktiboak ere badira (@ielortza, @kalaportu, @KikeAmonarriz, @maia_jon).
— **Ezker abertzalea** (% 26,98): Azpitalde zehatz hau Ezker Abertzale edo independentistarekin erlazionatutako erabiltzaileez osatuta dago. Erabiltzaileak erakunde politiko eta sozialekin (@ernaigazte, @GureEskuDago, @AskeGunea, @ehbildu, @sortuEH, @EtxeratElkartea...) zein mugimendu politiko horretako pertsona erreferenteekin (@ArnaldoOtegi, @jpermach...) erlazionatu ditzakegu. Azpitalde tematiko horretan ere komunikabideen presentzia edukiko genuke, aukera politiko horrekin erlazionatutakoak (@naiz_info, @topatu_eus, @info7irratia, @AhotsaInfo...).
— **Kirolak** (% 22,58): Kirol azpitaldean nodo garrantzitsuenak kazetari (@iBROKI, @XabierEuzkitze, @Imagreto, @TxetxuUrbieta, @jontolest, @unaizubeldia...) zein albis-

tegiak (@eitbkirolak, @ukHitza, @3ErregeenMahaia...) dira, bereziki kirol arloan espezializatuta daudenak. Azpitalde zehatz horretan ere, erabiltzaile konpartituenak egunkari eta telebista kateei egiten diete erreferentzia, beste behin ere komunikabideekin zuzenean erlazionatuta egonik. Komunitate horretako beste nodo garrantzitsu batzuk kirol taldeekin lotutakoak dira, hala nola, futbol taldeak (@RealSociedad, @RealSociedadEUS, @SDEibar, @AthleticClub...) edo jokalariak (@InigoMartinez, @mikelsanjo6, @ilarra4...), txirrindulari ezagunak (@AmetsTxurruka, @mikelastarloza, @Markelirizar...) zein euskupilota klubak (@ASPEpelota...).

— **Musika** (% 20,49): Musika azpitaldean leku nabarmenetan agertzen dira euskaraz abesten duten musika taldeak edo abeslariak (@ZuriHidalgo, @vendettaska, @hesian-taldea, @EsneBeltza, @gatibu, @ZeEsatek...) eta baita musikaren munduarekin lotutako beste hainbat erabiltzaile ere (@GustokoMusika, @euskalkantak5, @KantuBatGara...).

Aztertutako azpitaldeek erakusten dute guztiek harreman zuzena dutela Euskal Herriarekin lotutako gai edo kontuekin. Hala, ikusten da euskara erabiltzen dela euskal gaurkotasunarekin (albisteak) eta politikarekin (ezker abertzalea) erlazionatutako edukiak partekatzeko. Gainera, ikus daiteke aisialdiarekin erlazionatutako euskal musika eta kirol edukiak ere asko konpartitzen direla gazteen artean. Hau da, badirudi Twitterreko elkarrekintzen asmo nagusia politika eta gizarte gaiei buruzko edukiak partekatzea dela, baina euskal komunitateari eta hizkuntzari arreta argia emanez.



1. irudia. Erabiltzaile gazteen sarea komunitateen arabera zatikatua

1. irudiak erakusten du euskal erabiltzaile gazteen errepresentazioaren bistaratzea, erabiltzaileak puntuen bidez adierazita daude eta bakoitzaren kolorea azpiatalei dagokie. Ikusi daiteke, gazteak oro har, gizarte gaiekin (politika eta albisteak) zein aisialdiarekin (musika eta kirola) zerikusia duten gaien inguruan erlazionatzen direla. Komunitateak detektatzeko aplikatutako metodologia dela eta, azpitaldeak modu koherentean mapatzeko gai gara, komunitate bakoitza gaien arabera antolatuz. Hau da, komunitate bakoitzak grafikoan duen posizioak eta komunitateen arteko hurbiltasunak erakusten dute gaiek haien artean zein erlazio duten. Modu horretan gizarte gaiekin (politika eta albisteak) lotutako komunitateak elkarren ondoan daudela ikus dezakegu, aisialdiarekin (musika eta kirola) lotutako komunitateekin gauza bera gertatzen den bitartean. Politikarekin (Ezker Abertzalea) lotutako komunitatea albisteetatik eta musikatik gertu dago, sare sozialetatik mugitzen diren euskarazko albiste zein musika talde batzuen jarrera politikoa erakutsiz. Bestalde, kirolarekin erlazionatutako azpitaldea politikarekiko urrutien dagoena da, musika eta albisteekin gertatzen den moduan. Gainera, lau azpitaldeetatik hirutan (Albisteak, Ezker Abertzalea eta Kirolak) hedabideak eta kazetariak dira erabiltzaile erreferenteak, berriro ere frogatuz komunikabideak garrantzitsuak direla gazteen artean euskarazko edukiak zabaltzeko.

5. Ondorioak

Sare sozialetara konektatuta dauden euskal hiztun gazteen errealitatera hurbilpen bat egitea lortu dugu. Ikasketa sakoneko teknika aurreratuak aplikatu ditugu eskala handiko datuetatik abiatuta, ezaugarri demografikoak iradoki eta komunitate-azterketa egiteko. Horretarako, lehenik eta behin, Twitter sare sozialeko euskal erabiltzaileak sailkatu ditugu gazte eta heldu artean, testua oinarri duen metodologia berri bat proposatuta. Bigarrenik eta azkenik, gazteen komunitateak zeintzuk diren ikusi dugu, konpartitzen dituzten edukietatik erlazionatzeko moduak erauziz. Lan horrekin gizarte ikerkuntzarako adimen artifiziala erabiltzea aberasgarria dela frogatu dugu.

Sare sozialetako erabiltzaileen testua baliatuta bizitza-etapa automatikoki iradokitzeko hurbilpen bat aurkeztu dugu. Hurbilpen horrek euskal errealitatea barneratzen du, ikasketa gainbegiratua ahalbidetuko duten euskarazko datu-multzo eta aurre-entrenatutako hizkuntza-ereduak erabiltzen dituena. Sare sozialetako edukian oinarritutako datu-multzo handiak hizkuntza gutxituetarako nola eskuratu erakutsi dugu, esaldi edo sekuentzia mailatik erabiltzaile mailara igaroz, berriz ere sekuentzia mailako datuak jasotzeko hau da, zehatzetik hasi eta orokorrera salto eginez datu aberatsagoak eskuratzeko. Horrez gain, ikerketa horretan beste behin ere, ikusi ahal izan dugu aurre-entrenatutako hizkuntza-ereduek errendimendu hobea daukatela ereduak elebakarrak baldin badira. Horrela, hizkuntza bakoitzera, gure kasuan hizkuntza gutxitu batera, egokitutako datuak eta tresnak sortzea beharrezkoa dela azpimarratu nahi da lan horretan.

Horrez gain, ikusi dugu ikasketa sakoneko metodo ez-gainbegiratuen aplikazioa eraginkorra dela erabiltzaileen komunitateak zeintzuk diren iragartzeko. Aplikatutako teknikak, azpitalde kopurua erabakitzeko aukera emateaz gain, azpitaldeen erlazioa zein den erakusteko gai ere bada. Gainera, azpitaldeko erabiltzaile arrakastatsuenak baliatuta, komunitate ezberdinen identitatea modu intuitiboan definitzea lortu dugu.

Euskal komunitatearen baitako erakunde eta norbanako erreferentzial gehienak komunikabideekin eta lerrokatze politiko zehatzekin zerikusia daukatela ikusi dugu. Horrela, euskal komunikabideen papera azpimarratu behar da sareetan euskal edukiak hedatzeko duten ahalmenagatik. Laburbilduz, esan daiteke, euskara testuinguru berrietara moldatzeko gai dela, betiere hiztunen komunitatearen eguneroko errealitatearekin modu estuan lotuta. Horrek erakusten digu, globalizatutako eta etengabe konektatutako mundu honetan ere, euskaldunek badutela gaitasun berezia beren lekua bilatu eta bertan finkatzeko.

Bibliografia

Agerri, R., Bermúdez, J., & Rigau, G. (2014). IXA Pipeline: efficient and ready to use multilingual NLP tools. *Proceedings Of The Ninth International Conference On Language Resources And Evaluation*, 3823-3828.

Agerri, R., Centeno, R., Espinosa, M. S., Fernandez de Landa, J., & Rodrigo, Á. (2021). VaxxStance@IberLEF 2021: Overview of the task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento Del Lenguaje Natural*, 67(67), 173-181.

Agerri, R., & Rigau, G. (2016). Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence*, *238*, 63-82.

Agerri, R., & Rigau, G. (2019). Language independent sequence labeling for Opinion Target Extraction. *Artificial Intelligence*, *268*, 85-95.

Agerri, R., Vicente, I. S., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., & Agirre, E. (2020). Give your Text Representation Models some Love: the Case for Basque. *Proceedings of The 12th Language Resources and Evaluation Conference*, 4781-4788.

Al Zamal F., Liu W., and Ruths D. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media*, 270:2012.

Akbik, A., Blythe, D. A. J., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *Proceedings Of The 27th International Conference On Computational Linguistics*, 1638-1649.

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal Of Statistical Mechanics: Theory And Experiment*, *2008*(10), P10008.

Bojanowski, P., Grave, É., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions Of The Association For Computational Linguistics*, 5, 135-146.

Brown, P. F., deSouza, P., Mercer, R., Della Pietra, V. J., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467-479.

Cesare, N., Grant, C., & Nsoesie, E. O. (2017). Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices. *arXiv*.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. *10th Conference Of The European Chapter Of The Association For Computational Linguistics*.

Collins, M. (2002). Discriminative training methods for hidden Markov models. *Proceedings Of The Conference On Empirical Methods In Natural Language Processing*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Advances In Neural Information Processing Systems*.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2021). Political Polarization on Twitter. *Proceedings Of The International AAAI Conference On Web And Social Media*, *5*(1), 89-96.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019,* 4171-4186.

Fernandez de Landa, J. (2017). Sare sozialen erabilera moduak eta maiztasunak Gasteizko nerabeen kolektiboaren baitan. *II. Ikergazte 2017*.

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6), 212.

Fernandez de Landa, J., & Agerri, R. (2021). Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 1-15.

Grover, A., & Leskovec, J. (2016). node2vec. *Proceedings Of KDD*.

Jones, R. J., Cunliffe, D., & Honeycutt, Z. R. (2013). Twitter and the Welsh language. *Journal of Multilingual and Multicultural Development*, *34*(7), 653-671.

Jusup, M., Holme, P., Kanazawa, K., Takayasu, M., Romić, I., Wang, Z., Geček, S., Lipić, T., Podobnik, B., Wang, L., Luo, W., Klanjšček, T., Fan, J., Boccaletti, S., & Perc, M. (2022). Social physics. *Physics Reports*, *948*, 1-148.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *International Conference On Learning Representations*.

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

Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., & Akoglu, L. (2023). A Comprehensive Survey on Graph Anomaly Detection With Deep Learning. *IEEE Transactions On Knowledge And Data Engineering*, *35*(12), 12012-12038.

Marquardt, J. J., Farnadi, G., Vasudevan, G., Moens, M., Davalos, S., Teredesai, A., & De Cock, M. (2014). Age and gender identification in social media. *Proceedings Of CLEF 2014 Evaluation Labs*, 1129-1136.

McMonagle, S., Cunliffe, D., Jongbloed-Faber, L., & Jarvis, P. (2018). What can hashtags tell us about minority languages on Twitter? A comparison of #Cymraeg, #Frysk, and #Gaeilge. J*ournal of Multilingual and Multicultural Development*, 40(1), 32-49.

Mhichíl, M. N. G., Lynn, T., & Rosati, P. (2018). Twitter and the Irish language, #Gaeilge – Agents and Activities: exploring a data set with micro-implementers in social media. *Journal of Multilingual and Multicultural Development*, *39*(10), 868-881.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. M. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems*, 26, 3111-3119.

Morgan-Lopez A.A., Kim A.E., Chew R.F., and Ruddle P. (2017). Predicting age groups of Twitter users based on language and metadata features. *PloS one*, 12(8): e0183537.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & De Jong, F. (2016). Computational Sociolinguistics: a survey. *Computational Linguistics*, *42*(3), 537-593.

Nguyen D., Gravel R., Trieschnigg D., and Meder T. (2013). «How Old Do You Think I Am?»A Study of Language and Age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 7 lib., 439-448.

Pennacchiotti, M., & Popescu, A. (2011). Democrats, republicans and starbucks afficionados. *Association For Computing Machinery*.

Peters, M. E., Neumann, M. E., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics,* 2227-2237.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk. *Association For Computing Machinery*.

Rao D., Yarowsky D., Shreevats A., and Gupta M. (2010). Classifying latent user attributes in Twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37-44.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Zaghouani, W., & Charfi, A. (2018). Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Eusko Legebiltzarreko eztabaida saioak ParlaMint 4.0 proiektuan txertatzen

Incorporating the debate sessions of the Basque Parliament into the ParlaMint 4.0 project

Jon Alkorta[1], Mikel Iruskieta[1], Kike Fernandez[1], Ekain Arrieta[1], Rodrigo Agerri[1], Manex Agirrezabal[2]

[1] HiTZ-Hizkuntza Teknologiako Euskal Zentroa, Ixa taldea, Euskal Herriko Unibertsitatea (UPV/EHU)
jon.alkorta@ehu.eus, mikel.iruskieta@ehu.eus, kike.fernandez@ehu.eus, ekain.arrieta@ehu.eus, rodrigo.agerri@ehu.eus

[2] Centre for Language Technology (CST), University of Copenhagen (KU)
manex.aguirrezabal@hum.ku.dk

Laburpena

Lan honetan, Eusko Legebiltzarreko eztabaida saioen corpusa aurkezten dugu. Corpusa ParlaMint 4.0 proiektuaren parte da, eta 2015etik 2022ra bitarteko datuak biltzen ditu. Corpusa sortu ahal izateko, eztabaida saioen transkripzio dokumentuak laga dizkigu Eusko Legebiltzarreko mahaiak, eta Legebiltzarreko mahaiaren oniritziarekin ParlaMinten eskatzen ziren ataza guztiak egin ditugu. Lehendabizi, transkripzio edo datuak ParlaMinteko XML TEI formatuan jarri dira. Ondoren, transkripzio horiek hainbat informazioz aberastu ditugu: hizketaldien egileak, esaldi bakoitzaren hizkuntza, besteak beste. Horrez gain, metadatuen beste dokumentu bat ere eratu dugu; bertan, Eusko Legebiltzarraren (sorrera urtea, webgunea, helbidea, izena hiru hizkuntzetan...), alderdi politikoen (sorrera urtea, sigla, webgunea, orientazio politikoa...) eta legebiltzarkideen datuekin (izen-abizenak, jaiotza urtea, afiliazio politikoa, generoa, hiria...). Bukatzeko, osatu dugun corpusaren analisi linguistikoa egin dugu. Corpusak 13.321.393 hitz ditu, eta ikertzaileek aztertzeko prest dago KonText, TEITOK eta NoSketch Engine webguneetan.

Gako-hitzak: Eusko Legebiltzarra, ParlaMint 4.0, XML TEI, Humanitate Digitalak.

Abstract

In this work, we present the corpus of debate sessions of the Basque Parliament. The corpus is part of the ParlaMint 4.0 project and includes textual data from 2015 to 2022. In order to create the corpus, the transcripts of the debate sessions have been handed over to us by the Basque Parliament, and with the approval of the parliament, we have done all the tasks required of ParlaMint. First of all, the transcripts are converted into ParlaMint's XML TEI format. Then, these transcripts are enriched with extra information, such as, the author names and the language of each sentence. In addition, another metadata document is created with the data of the Basque Parliament (year of establishment, website, address, name in three languages, etc.), political parties (year of establishment, acronym, website, political orientation, etc.) and members of parliament (surnames, year of birth, political affiliation, gender, city, etc.). Finally, we conducted a linguistic analysis of the corpus. The corpus consists of 13,321,393 words, and it is available for further research on the KonText, TEITOK and NoSketch Engine websites.

Keywords: Basque Parliament, ParlaMint 4.0, XML TEI, Digital Humanities.

1. Sarrera

Azken bi hamarkadetan, eta batez ere, azken urteetan Humanitateetako eta Gizarte Zientzietako ikertzeko moldea aldatzen ari da, eta horretan, Humanitate Digitalek bere eragina izan dute.

Terras-ek (2011) eta Drucker-ek (2013) diotenez, Humanitate Digitala informatika edo teknologia digitalen eta humanitateen diziplinen arteko eremuan diharduen ikerketa arloa da. Hark, giza zientzietan, tresna eta baliabide digitalen erabilera sistematikoa eta erregularra bultzatzen du, baita haien aplikazioaren analisia ere. Ildo beretik, Burdick *et al.*-en (2013) arabera, Humanitate Digitalak ikerketa eta lankidetza (esaterako, irakaskuntzan eta argitalpenetan) bultzatzen du, diziplinaz gaindi eta konputazionalki. Horretarako, teknika eta aplikazio berriak sortzen eta erabiltzen dira; eta horrek ikerketan bi norabidetako onurak ekartzen ditu (onura konputazionalak humanitateetan dihardutenentzat eta humanitate arloko onurak konputazioaren arloan dabiltzanentzat).

Testuinguru horretakoa da aurkezten dugun Eusko Legebiltzarreko eztabaida saioen corpus hau. Corpusa ParlaMint 4.0 proiektuaren parte da eta honen sorkuntzak harreman zuzena du CLARIN europar azpiegiturarekin eta hark bultzatzen duen Humanitate Digitalaren arloarekin. ParlaMint CLARINen proiektu garrantzitsu bat da, izan ere, Europako eztabaida parlamentarioen corpus konparagarriak egitea du helburu, betire corpus guztiek estandar bateratuak edukiz. Orain arte, ParlaMint proiektuak hainbat etapa izan ditu.

Parlamentuetako eztabaidak ikertzeak hainbat motibazio izan ditu Erjavec *et al.* (2023) lanean aipatzen den moduan: Zientzia Politikoak aztertzeko (van Dijk, 2010), Soziologia aztertzeko (Cheng, 2015), Historia aztertzeko (Pančur & Šorn, 2016), Diskurtsoaren Analisirako (Hirst *et al.,* 2014), Soziolinguistikarako (Rheault *et al.,* 2016), hizkuntza aniztasuna eta kultur artekotasuna aztertzeko (Bayley, 2014) erabili dira, besteak beste.

Lehen etapan (2020-2021), ParlaMint 1.0 (Erjavec *et al.,* 2023) jarri zen martxan, eta bertan zenbait helburu finkatu ziren: (i) 2019ko azarotik 2020ko uztailera bitarteko parlamentu-prozeduren corpus eleaniztun bat sortzea (gaitzat Covid-19 pandemia hartuz), (ii) 2015etik 2019ko urrira arteko datu parlamentarioen erreferentzia eleaniztun konparagarrien multzoa sortzea, (iii) corpusak linguistikoki prozesatzea *Universal Dependencies* (UD) formalismoaren egitura sintaktikoak erabiliz eta izen berezien (*Named Entities*, NE) informazioa gehituz, iv) Konkordatzaile eta Parlameter moduko corpusak eskuragarri jartzea, eta azkenik, v) Zientzia Politikoetan eta Humanitate Digitaletan erabilera kasuak eraikitzea corpuseko datuetan oinarrituta.

Helburu horiek betetzeko, lehendabizi, proba (erreferentziazko corpus parlamentarioak prestatzea, COVID-19ren corpus parlamentarioak sortzea, corpusak muntatzea NoSketch Engine eta KonText konkordatzaileetan, jarraibideen eta diru-laguntza txikien prozedura prestatzea) lau hizkuntzatan egin zuten, eta ondoren, corpusak zabaldu eta erakusgarri jarri zituzten (azpiegiturari corpus osagarriak gehitzea, erakusleihoak prestatzea eta dokumentazioa prestatzea interesdunek erabiltzeko).

Proiektua bi fasetan egin zen, lehena 2020ko uztailetik 2020ko irailera bitartean, eta bigarrena 2020ko urritik 2021eko maiatzera bitartean. Ondorioz, zenbait hizkuntzaren corpusak eskuragarri jarri ziren CLARIN.SI biltegian: alde batetik, corpus garbiak jarri ziren, eta beste aldetik, linguistikoki etiketatutako corpusak. Halaber, corpusak konkordatzaileetan eskuragarri jarri ziren ikertzaileen esku uzteko (NoSketch eta KonText tresnen bitartez) eta hainbat hizkuntzatako corpusak paraleloki aztertzeko aukera ere jarri zen. Guztira 17 parlamentutako corpusak bildu ziren.

ParlaMint 2.0k (Ogrodniczuk *et al.,* 2022), berriz, 2022tik 2023ra arteko iraupena izan zuen. Bere helburuak, besteak beste, honakoak izan ziren: (i) corpusen denbora-tartea (urte kopurua)

handitzea, (ii) herrialde eta autonomia-erkidego gehiagotako corpusak gehitzea (tartean, Eusko Legebiltzarrekoa), (iii) corpusak itzulpen automatikoaren bitartez ingelesez eskaintzea, iv) corpusak metadatu gehigarriekin areagotzea eta v) corpusaren erabilgarritasuna hobetzea.

Hori dela eta, ParlaMint 2.0 proiektuak 5 lan-pakete izan zituen: (i) dokumentazioa, elkarreragingarritasuna eta metadatuak (kodeketaren harmonizazioa, Git kudeaketa, dauden corpusetan metadatuak gehitzea), (ii) corpusaren hedapena (corpus berriak gehitzea, dauden corpusak zabaltzea, datuen banaketa), (iii) corpusa aberastea (itzulpen automatikoa eta etiketatze semantikoa, multimodalitatea), (iv) parte hartzeko jarduerak (tutorialak, hackathoiak, partekatutako zeregina eta erakusleihoak), eta azkenik, (v) koordinazioa (kudeaketa, zabalkundea, kanpoko jarraipena). ParlaMint-en fase hau indarrean zegoenean, Eusko Legebiltzarreko eztabaida saioen transkripzioak ParlaMint proiektuan txertatzeko bidean zeuden (Alkorta & Iruskieta, 2022).

Lan-pakete horien emaitza ParlaMint 3.0 izan zen (Kuzman *et al.*, 2023). 26 parlamentutako corpus parlamentario berriak eskuragarri jarri ziren CLARIN.SI biltegian. Corpusak testu arrunteko bertsioan (hau da, etiketatzerik gabe) eta linguistikoki etiketatutako bertsioan daude eskuragarri. Linguistikoki etiketatutako corpusak konkordatzaileen bidez kontsulta daitezke (NoSketch Engine eta KonText tresnen bidez). Bertsio honetan, Europako 26 corpus parlamentario berberak daude, eta 1.100 milioi hitz baino gehiago dituzte.

Azkenik, 2023an bertan, ParlaMint 4.0 argitaratu zen. Kasu honetan, Europako 29 herrialde eta eskualde autonomoetako eztabaida parlamentarioen transkripzioak dituen corpus konparagarri bat da, eta gehienbat 2015ean hasi eta 2018 erdialdera arte luzatzen da. Parlamentu ezberdinetako corpusek 9 milioi eta 126 milioi hitzen artean dituzte. Transkripzioak egunez banatzen dira. Epeari, saioari eta bilerari buruzko informazioa dute, eta hizlariak eta bere eginkizunak (adibidez, hizlaria lehendakaria edo hizlari arrunta den) markatutako hitzaldiak jasotzen dituzte. Hitzaldiek transkripzio-iruzkin markatuak ere badituzte, hala nola transkripzioaren hutsuneak, etenak, txaloak, etab. Corpusak metadatu zabalak ditu, batez ere, hiztunen (izena, generoa, diputatu eta ministro estatusa, alderdiko kidetasuna), alderdi politiko eta talde parlamentarioei buruzkoa (izena, koalizio/oposizio egoera, Wikipediaren araberako ezkerretik eskuineko orientazio politikoa eta Chapel Hill Expert Survey, CHES, aldagaiak). Proiektuaren bertsio honetan, Eusko Legebiltzarreko eztabaida saioen transkripzioak barneratuta daude.

Lan hau honela banatuta dago: bigarren zatian, artearen egoera deskribatzen da; eta bertan, parlamentuetako transkripzioekin egin diren lanak azaltzen dira (lehendabizi, nazioartean eta ondoren, Eusko Legebiltzarrarekin egin direnak). Azken honetan, eskuragarri dauden corpus motak eta egindako ikerketa lanak aipatzen dira. Hirugarren atalean, eztabaida saioetako transkripzioak ParlaMint 4.0-n integratzeko egin ditugun urratsak azaltzen dira: transkripzioak lortu, transkripzioak formatuz aldatu, metadatuak erantsi, eta corpusa baliozkotu. Laugarren atalean, berriz, lanaren emaitzak eta zabalkundea aipatzen dira. Bertan, eratu dugun corpusak dituen ezaugarriak deskribatzen ditugu hainbat ikuspegitatik (hitz kopurua, metadatuei buruzko datuak, eta abar). Halaber, ikertzaileek ikerketa egiteko corpusa nola erabil dezaketen azaltzen da; horretarako, NoSketch Engine, KonText eta TEITOK tresnei buruzko azalpenak emanez. Bukatzeko, azken eta bosgarren atalean, lanaren ondorioak eta etorkizuneko lanak aipatzen dira.

2. Artearen egoera

Artearen egoera hiru ataletan bana daiteke: (i) parlamentuetako eztabaida saioak eta lengoaia naturalaren prozesamendua, (ii) parlamentuetako eztabaida saioak eta hizkuntzaren azterketa, eta (iii) Eusko Legebiltzarreko eztabaida saioekin egin diren askotariko lanak.

Lan honen motibazioan aipatu ditugun aztergaietarako erabiltzeaz gain, parlamentuetako eztabaida saioak nazioartean askotan erabili dira, batez ere, lengoaia naturalaren prozesamenduaren arloan. Bertan, erreferentziazko corpusa Europarl (Kohen, 2005) da. Europako Parlamentuko transkripzioak biltzen ditu corpusak. Transkripzio horiek 11 hizkuntzatan daude, eta corpus guztia lerrokatuta dago: hau da, ingelesezko lerro baten parekideak dira frantsesekoa, alemanerakoa eta beste hizkuntzetakoak. Horretaz gain, itzulpen automatiko estatistikorako (*Statistical Machine Translation*, SMT) prestakuntza-datu gisa erabiltzen dute corpusa, izan ere, SMT sistemak trebatu egin dituzte 110 hizkuntza-bikoteentzat.

Lan honen atzetik, beste zenbait lan etorri dira baina denak ez dira lengoaia naturalaren prozesamenduaren arlokoak. Hor daude besteak beste: hizkuntzen arteko harremana aztertzea (Rama, T., & Borin, 2011), diskurtsoaren arloan hizkuntza markatzaileak ikertzea (Van Halteren, 2008), eta hizkuntza arteko ikerketak egitea (Cartoni *et al.,* 2013), lanetako batzuk esatearren. Baina horietako bakoitzean ikerketa galderei erantzuteko, corpus bera abiapuntutzat izan arren, oso metodologia ezberdinak garatu izan behar dituzte. ParlaMint proiektuak hori eragotzi nahi du, eta parlamentu askotako corpusak formatu berean jarri nahi ditu ikertzaileei lana errazteko.

Euskal Herrira itzuliz, Eusko Legebiltzarraren kasuan, badaude zenbait lan bertako eztabaida saioen transkripzioak erabili dituztenak. Gure antzeko helburuekin egin den corpus bat ere badago, eta Eusko Legebiltzarreko eztabaida saioekin sortu dute. Corpus horren izena Basqueparl (Escribano *et al.,* 2022a) da. Bertan, 2012tik 2020ra bitarteko eztabaiden transkripzioak hartu dituzte, eta transkripzioetako hizkuntzak identifikatzeaz gain, legebiltzarkideen zenbait datu ere bildu dituzte (jaiotza-urtea, afiliatutako alderdi politikoa, generoa, eta abar). Transkripzioei dagokienez, zenbatgarren eztabaida saioa den, hori egin den eguna, eta zenbatgarren paragrafoa den ere adierazita daude. Bukatzeko, testuak lematizatuta daude, eta entitate izendunak ere identifikatuta daude. Corpusa baliatuz, bi ikerlerro garatu dituzte: hizkuntzaren erabilera eta generoa (urtez urte, alderdiz alderdi, eta abar). Lan hau gure lanetik ezberdintzen da, alde batetik, eztabaidetako testuak ez daudelako XML TEI moduan kodetuta, eta beste aldetik, ez dituztelako ParlaMint proiektuko irizpideak edo baldintzak jarraitzen.

Aurretik aipatu dugun Basqueparl corpusean oinarrituta, Escribano et. al.ek (2022b) euskararen erabilera aztertu dute Eusko Legebiltzarrean 2012tik 2020ra bitartean. Emaitzen arabera, Eusko Legebiltzarreko euskararen erabilera % 18,4koa da (ganberako lehendakaria aintzat hartuz, % 21,2koa) eta hori erabilera sinbolikoaren erakusgarri da. Bestetik, euskararen presentzia hitzaldi gehienetan agertu ohi dela antzeman dute, baina hitzaldia zenbat eta luzeagoa izan, euskararen presentzia txikiagoa da. Azkenik, alderdi politikoen artean ere, erabilerari dagokionez ezberdintasunak antzeman dituzte.

Eusko Legebiltzarreko eztabaida saioen transkripzioak erabiliz, bestelako ezaugarriak dituen corpus bat ere badago: Mintzai-ST (Etchegoyhen *et al.,* 2021). Mintzai-ST corpusa euskara-gaztelania hizkera itzultzeko corpus paralelo bat da, eta 2011tik 2018ra bitarteko Eusko Legebiltzarreko bilkuretan egindako aktak ditu abiapuntu. Corpusa audio-fitxategiek, transkripzioek eta itzulpenek osatzen dute, eta euskara-gaztelaniarako muturreko edo kaskadako ahozko itzulpen-sistemak trebatzeko erabil daitezke bi noranzkoetan.

Aurretik aipatutako lan guztien helburua corpus bat sortzea izan da, baina eztabaida saioak erabiliz, bestelako ikerketak ere egin dira. Esaterako, «aldibereko interpretazioa testuarekin» izeneko modalitatea aztertzen du Torralbak (2021) Eusko Legebiltzarrean. Kasu honetan, gure lanaren abiapuntu diren benetako hitzaldiak eta horien itzulpenak aztertzen ditu. Zehazki, itzulpenak zergatik egiten diren euskaratik gaztelaniara eta ez gaztelaniatik euskarara erantzuten saiatzen

da, eta horretarako Euskal Autonomia Erkidegoko hizkuntza-araudia eta egoera soziolinguistikoa kontuan hartzen ditu.

Eusko Legebiltzarreko eztabaida saioekin beste ikerketa lerro bat ere badago, eta hori informazio multimodalaren tratamenduarena da. Legebiltzarreko saioen bideoan ahotsa eta horien idatzizko transkripzioa uztartzean datza. Bordel-García *et al.*ek (2013) OBAM-PV izeneko software tresna bat garatu dute, eta bideoentzako azpitituluak sortzen dituzte eskura dauden eztabaida saioen transkripzioekin. Eleaniztasuna tratatu egin dute bertan, eta alderdirik zailenak aurreprozesamendua, eta testuaren eta audioaren arteko sinkronizazioa izan dira. Aipaturiko lana, aurreko lan baten ondorio da (Bordel *et al.,* 2011) eta geroztik informazio multimodalaren tratamenduan lan gehiago ere egin dituzte (Bordel *et al.,* 2012, Peñagarikano *et al.,* 2023).

Laburbilduz, ikus daitekeen moduan, Eusko Legebiltzarreko saioetako hizketaldiak oso interesgarriak dira ikerketarako, azken batean, jendartean kezka sortzen duten gaiak jorratzen direlako. Ikerketa ikuspegi ezberdinetatik egin daiteke, baina atal honetan ikusi dugun moduan, eskuragarri dauden datuak tratatu egin behar dira eta horrek zailtasunak ekar ditzake. Gure xedea Eusko Legebiltzarreko eztabaida saioen testuak ParlaMint proiektuan txertatzea da, datuak jada landuta dituen tresna bat jendartearen eta ikertzaileen esku jarriz. Modu honetan, helburuak Humanitate Digitalarekin bat egiten du.

3. Metodologia

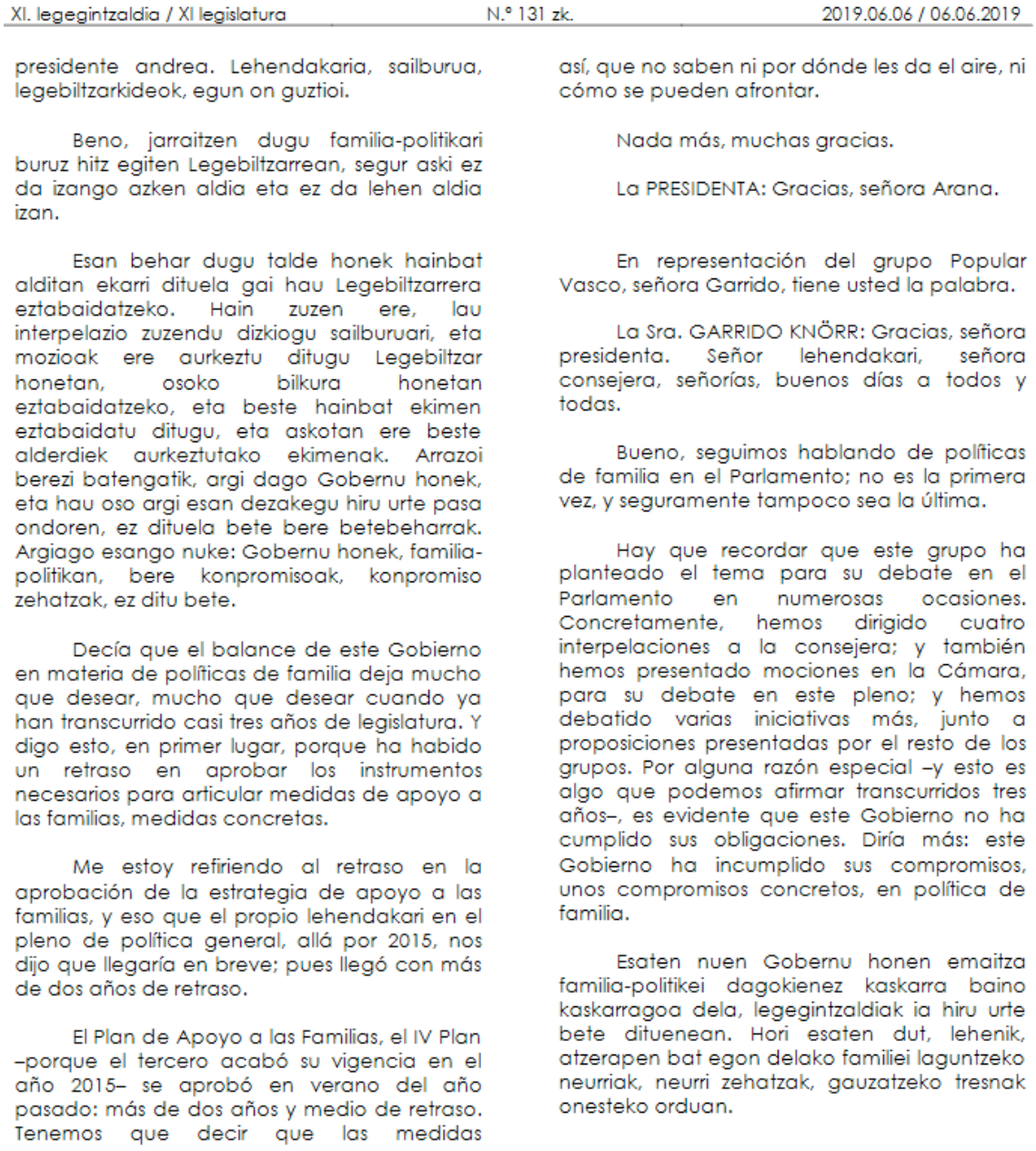
Eusko Legebiltzarreko testuetan oinarrituta, ParlaMint proiekturako corpusa sortzeko guztira bost urrats egin ditugu: (i) datuak eskuratu, (ii) eztabaida saioetako testuak egituratu eta metadatuak sortu, (iii) corpusaren analisi linguistikoa egin, (iv) corpusa baliozkotu eta bihurketa egin, eta (v) corpusa ikusgai eta erabilgarri jarri. Jarraian, urrats horiek banan-banan deskribatuko ditugu.

3.1. Datuak eskuratu

Lehen faseko helburua corpusa osatzeko datuak eskuratzea izan da. Eusko Legebiltzarreko eztabaidak ParlaMint proiektuak eskatzen dituen baldintzetara egokitzeko, hurrengo urratsak jarraitu behar izan ditugu.

Lehendabizi, Eusko Legebiltzarreko eztabaiden transkripzioen testuak lortu behar izan ditugu. Horretarako, 2020ko azaroan, hitzarmen bat egin genuen Eusko Legebiltzarreko arduradunekin eta Eusko Legebiltzarrean 2015etik 2022ra bitartean egin ziren eztabaiden transkripzioen testu guztiak lortu genituen. Testu horiek *word* formatuan jaso genituen, eta urtearen eta legegintzaldiaren arabera, *word* horietan zegoen testuen egituraketa ezberdina zen.

1. irudian ikusten den moduan, eztabaida saioen transkripzioak biltzen dituzten dokumentuek egitura jakin bat dute. Alde batetik, eztabaida eguna noiz izan den, zer legegintzaldi den eta zenbatgarren eztabaida saioa den zehazten da (orrialdearen goiko zatian). Bestetik, ezkerreko zutabean benetako hizketaldia dago, eta eskuineko zutabean haren itzulpena. Originala euskaraz denean, itzulpena gaztelaniaz dago, eta originala gaztelaniaz dagoenean, euskarazko itzulpena. Bukatzeko, hizketaldiaren egilea edo legebiltzarkidea letra larriz zehazten da (kasu honetan, Garrido Knörr anderea da).



XI. legegintzaldia / XI legislatura N.º 131 zk. 2019.06.06 / 06.06.2019

presidente andrea. Lehendakaria, sailburua, legebiltzarkideok, egun on guztioi.

Beno, jarraitzen dugu familia-politikari buruz hitz egiten Legebiltzarrean, segur aski ez da izango azken aldia eta ez da lehen aldia izan.

Esan behar dugu talde honek hainbat alditan ekarri dituela gai hau Legebiltzarrera eztabaidatzeko. Hain zuzen ere, lau interpelazio zuzendu dizkiogu sailburuari, eta mozioak ere aurkeztu ditugu Legebiltzar honetan, osoko bilkura honetan eztabaidatzeko, eta beste hainbat ekimen eztabaidatu ditugu, eta askotan ere beste alderdiek aurkeztutako ekimenak. Arrazoi berezi batengatik, argi dago Gobernu honek, eta hau oso argi esan dezakegu hiru urte pasa ondoren, ez dituela bete bere betebeharrak. Argiago esango nuke: Gobernu honek, familia-politikan, bere konpromisoak, konpromiso zehatzak, ez ditu bete.

Decía que el balance de este Gobierno en materia de políticas de familia deja mucho que desear, mucho que desear cuando ya han transcurrido casi tres años de legislatura. Y digo esto, en primer lugar, porque ha habido un retraso en aprobar los instrumentos necesarios para articular medidas de apoyo a las familias, medidas concretas.

Me estoy refiriendo al retraso en la aprobación de la estrategia de apoyo a las familias, y eso que el propio lehendakari en el pleno de política general, allá por 2015, nos dijo que llegaría en breve; pues llegó con más de dos años de retraso.

El Plan de Apoyo a las Familias, el IV Plan –porque el tercero acabó su vigencia en el año 2015– se aprobó en verano del año pasado: más de dos años y medio de retraso. Tenemos que decir que las medidas así, que no saben ni por dónde les da el aire, ni cómo se pueden afrontar.

Nada más, muchas gracias.

La PRESIDENTA: Gracias, señora Arana.

En representación del grupo Popular Vasco, señora Garrido, tiene usted la palabra.

La Sra. GARRIDO KNÖRR: Gracias, señora presidenta. Señor lehendakari, señora consejera, señorías, buenos días a todos y todas.

Bueno, seguimos hablando de políticas de familia en el Parlamento; no es la primera vez, y seguramente tampoco sea la última.

Hay que recordar que este grupo ha planteado el tema para su debate en el Parlamento en numerosas ocasiones. Concretamente, hemos dirigido cuatro interpelaciones a la consejera; y también hemos presentado mociones en la Cámara, para su debate en este pleno; y hemos debatido varias iniciativas más, junto a proposiciones presentadas por el resto de los grupos. Por alguna razón especial –y esto es algo que podemos afirmar transcurridos tres años–, es evidente que este Gobierno no ha cumplido sus obligaciones. Diría más: este Gobierno ha incumplido sus compromisos, unos compromisos concretos, en política de familia.

Esaten nuen Gobernu honen emaitza familia-politikei dagokienez kaskarra baino kaskarragoa dela, legegintzaldiak ia hiru urte bete dituenean. Hori esaten dut, lehenik, atzerapen bat egon delako familiei laguntzeko neurriak, neurri zehatzak, gauzatzeko tresnak onesteko orduan.

1. irudia. Eusko Legebiltzarretik jasotako *word* dokumentu baten irudia

3.2. Eztabaida saioetako testuak egituratu eta metadatuak sortu

Bigarren fasean, lorturiko transkripzioen dokumentuak eta horietako testuak egituratu eta, horietan oinarrituz, metadatuak sortu behar izan ditugu. Horrela, *word* batzuek euskarazko edo gaztelaniazko testu hutsa zuten. Kasu horietan, testu horietako zati batzuk benetako hizketaldiak ziren, eta beste batzuk, berriz, itzulpen bidez lortutako testuak. Beste batzuetan, ordea, *word* dokumentuak bi zutabe zituen. Ezkerreko zutabean, benetako hizketaldia agertzen zen (euskaraz edo gaztelaniaz), eta eskuineko zutabean, benetako hizketaldiaren itzulpena (hau ere, euskaraz edo gaztelaniaz).

Hurrengo urratsean, *word* dokumentuetan zeuden testu horiek formatuz aldatu behar izan ditugu. ParlaMint proiektuak dokumentuak XML-TEI formatuan egon daitezela eskatzen du. Hori

dela eta, dokumentuak formatuz aldatzeko *OxGarage* webgunea baliatu dugu. Bertan, dokumentuen sarrerako formatua *Microsoft Word* formatua dela zehaztu dugu, eta irteerako formatua, aldiz, TEI Simple XML Document. Webgune honen bidez, banan-banan dokumentu guztiak formatuz aldatu ditugu.

Ondoren, jadanik XML formatua duten dokumentuetan, metadatuak sartzen hasi gara. Metadatu batzuk eskuz sartu ditugu eta beste batzuk erdiautomatikoki egin dira. Lehendabizi, legebiltzarrean egindako eztabaida horri buruzko datuak sartu ditugu. Datu horiek, besteak beste hauexek dira: eztabaidari dagokion corpusaren izenburua, eztabaida egin zen eguna, eztabaida hori legegintzaldiko zenbatgarren eztabaida den, eztabaida hori zenbatgarren legegintzaldikoa den, dokumentuaren egileak, dokumentuaren hizkuntza, dokumentuak duen karaktere eta hitz kopurua, lizentzia kontuak, dokumentua nondik eta noiz eskuratu den, eta Eusko Legebiltzarrari buruzko zenbait datu (izena, helbidea, hiria, eskualdea eta data).

Aipaturiko metadatu horiek guztiak dokumentu bakoitzaren hasieran doaz eta datu batzuk eskuz sartu ditugu (izenburua, eguna, egilea edota Eusko Legebiltzarrari buruzko zenbait datu). Aldiz, beste datu batzuk automatikoki lortu ditugu (karaktere eta hitz kopurua, esaterako).

Dokumentuaren goiko atala osatu ondoren, dokumentuaren zatirik garrantzitsuena bera, hau da, eztabaida testuak metadatuz elikatzen jarraitu dugu. Kasu honetan ere, metadatuak hainbat dira eta bi modutan jarri ditugu: eskuz eta automatikoki. Lehenik eta behin, hizketaldi bakoitzaren egilea eta hizketaldiaren hasiera eta amaiera zehaztu behar izan ditugu. Hizketaldiaren egilea zehazteko, ParlaMinten gidalerroei jarraituz, <note> (hasieran) </note> (amaieran) etiketak erabili ditugu. Bertan, egilearen izen-abizenak eta parlamentuan duen eginkizuna (esaterako, parlamentuko lehendakaria) zehazteaz gain, eztabaida horretan zenbatgarren hizketaldi den zehaztu dugu kode batzuen bitartez.

```
<text ana="#covid #war">
   <body>
      <div type="debateSection">
         <note type="speaker" xml:id="ParlaMint-ES-PV_2022-10-14.note1">LEHENDAKARIAK (Tejeria Otermin):</note>
         <u who="#TejeriaOtermin" xml:id="ParlaMint-ES-PV_2022-10-14.u0" ana="#chair">
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg2" xml:lang="eu">Egun on guztioi. Osoko bilkurari hasiera emango diogu.</seg>
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg3" xml:lang="eu">Gai-zerrendako lehenengo puntua: "Galdera, Jon Andoni Atutxa Sainz Euzko Abertzaleak taldeko legebiltz
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg4" xml:lang="eu">Atutxa jauna, zurea da hitza.</seg>
         </u>
         <note type="speaker" xml:id="ParlaMint-ES-PV_2022-10-14.note2">ATUTXA SAINZ jaunak:</note>
         <u who="#AtutxaSainz" xml:id="ParlaMint-ES-PV_2022-10-14.u1" ana="#regular">
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg5" xml:lang="eu">Eskerrik asko, legebiltzarburu andrea. Lehendakari, sailburuak, legebiltzarkideok, egun on.</seg>
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg6" xml:lang="es">Señor Erkoreka, en su comparecencia el pasado mes de marzo nos aportó un dato que desde nuestro grupo
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg7" xml:lang="eu">Delitu informatikoak ez dira berriak, baina azken urteotan etengabe ari dira hazten.</seg>
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg8" xml:lang="eu">Erkoreka jauna, zuk zeuk adierazi zenuen bezala, gero eta ingurune digitalizatuagoan gaude, gero eta i
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg9" xml:lang="eu">Erkoreka jauna, egoera horren aurrean, zer jarduketa egiteko asmoa du Ertzaintzak ziberdelinkuentziare
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg10" xml:lang="eu">Eskerrik asko.</seg>
         </u>
         <gap reason="editorial">
            <desc xml:lang="en">SAMPLING</desc>
         </gap>
         <u who="#ZupiriaGorostidi" xml:id="ParlaMint-ES-PV_2022-10-14.u135" ana="#regular">
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg690" xml:lang="es">Presidente andrea.</seg>
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg691" xml:lang="es">Barrio jauna, mi ignorancia no llega hasta el punto de no saber que Orduña es Bizkaia. Pero es que e
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg692" xml:lang="es">Mire, mi relación con usted en estos seis años en los que me ha correspondido estar al frente del De
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg693" xml:lang="es">Yo creo que…</seg>
            <note xml:id="ParlaMint-ES-PV_2022-10-14.note198">32. zintaren amaiera</note>
            <note xml:id="ParlaMint-ES-PV_2022-10-14.note199">33. zintaren hasiera</note>
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg694" xml:lang="es">… clarísimas de nuestro departamento en el territorio alavés. A veces hace falta que todos nos creamo
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg695" xml:lang="es">Yo creo que todo el entorno que tiene que ver con Añana, Iruña-Veleia, es una zona que si nos la cre
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg696" xml:lang="es">En cualquier caso, señor Barrio, no voy a dejar de citar a tres proyectos más que estamos desarrolla
            <seg xml:id="ParlaMint-ES-PV_2022-10-14.seg697" xml:lang="eu">Eskerrik asko.</seg>
         </u>
```

2. irudia. Eztabaida saio bateko transkripzioak metadatuz elikatuta

Ondoren, hau ere eskuz, hizketaldi bakoitza non hasi eta non amaitzen den zehaztu dugu. Horretarako, XML formatuak ahalbidetzen dituen <u> eta </u> etiketak baliatu ditugu. Jarraian

aipatzen diren metadatuak automatikoki egin ditugu. Alde batetik, hizketaldia segmentuetan banatu dugu. Segmentu bakoitza paragrafo bati dagokio; beraz, segmentu bakoitzean esaldi bat baino gehiago egin daitezke. Horiek etiketatzeko <seg> eta </seg> etiketak erabili dira.

Bestetik, segmentu bakoitzaren hizkuntza zein den zehaztu dugu. Segmentua euskaraz baldin badago «eu» etiketa jarri diogu, eta gaztelaniaz baldin badago «es» etiketa. Etiketa hori segmentu bakoitzaren hasieran ageri da, testua bera baino lehenago. Aipatu behar da etiketa hau dela Eusko Legebiltzarreko corpusak duen berezitasuneko bat. Izan ere, Belgikako parlamentuko corpusean kenduta (bertan, frantsesa, nederlandera eta alemana baitira ofizialak), beste corpusetan hizkuntza bakarra egon ohi da. Bukatzeko, segmentu bakoitzari kode bat ere esleitu zaio. Adibidez, 64. segmentuari jarritako etiketa hau da: «ParlaMint-ES-PV_2020-10-15.seg64».

2. irudian, metadatuz elikatutako eztabaida saio baten transkripzioa ikus daiteke. <note> eta </note> etiketen barruan, hitzen egilea adierazten da (kasu horretan, legebiltzarkideko lehendakaria eta Atutxa Sainz jauna). Ondoren, <u> eta </u> etiketen barruan, hizketaldiari kode bat jarri zaio, eta hitzen egilea ere zehaztu da haren kodearen bidez. Gero, esaldiak segmentatuta ageri dira, eta segmentu bakoitzak bere kodea du. Segmentu bakoitzean agertzen den hizkuntza ere zehaztuta ageri da.

Prozedura bera jarraitu da eztabaida bakoitzarekin. Eztabaida bakoitzak dokumentu bat du corpusean. Gainera, egun batzuetan, bi eztabaida egon daitezke. Beraz, posible da egun bateko bi dokumentu egotea. Urrats honekin, ParlaMinterako Eusko Legebiltzarreko corpusa osatzeko zati bat bukatu dugu. Hurrengo urratsean, berriz, beste dokumentu bat sortu behar izan dugu metadatuekin. Kasu honetan, metadatuak legebiltzarkide eta alderdi politikoei buruzkoak dira, besteak beste.

```
37      <org xml:id="EAJ-PNV" role="politicalParty">
38         <orgName full="yes" xml:lang="eu">Euzko Alderdi Jeltzalea</orgName>
39         <orgName full="abb">EAJ/PNV</orgName>
40         <event from="1895-07-31">
41            <label xml:lang="en">existence</label>
42         </event>
43         <idno type="URI" subtype="wikimedia">https://eu.wikipedia.org/wiki/Euzko_Alderdi_Jeltzalea</idno>
44         <idno type="URI" subtype="wikimedia">https://es.wikipedia.org/wiki/Partido_Nacionalista_Vasco</idno>
45         <idno type="URI" subtype="wikimedia">https://en.wikipedia.org/wiki/Basque_Nationalist_Party</idno>
46         <state type="politicalOrientation">
47            <state type="Wikipedia" source="https://en.wikipedia.org/wiki/Basque_Nationalist_Party" ana="#orientation.C"/>
48         </state>
49      </org>
50      <org xml:id="EHBildu" role="politicalParty">
51         <orgName full="yes" xml:lang="eu">Euskal Herria Bildu</orgName>
52         <orgName full="abb">EH Bildu</orgName>
53         <event from="2012-06-12">
54            <label xml:lang="en">existence</label>
55         </event>
56         <idno type="URI" subtype="wikimedia">https://eu.wikipedia.org/wiki/Euskal_Herria_Bildu</idno>
57         <idno type="URI" subtype="wikimedia">https://es.wikipedia.org/wiki/Euskal_Herria_Bildu</idno>
58         <idno type="URI" subtype="wikimedia">https://en.wikipedia.org/wiki/EH_Bildu</idno>
59         <state type="politicalOrientation">
60            <state type="Wikipedia" source="https://en.wikipedia.org/wiki/EH_Bildu" ana="#orientation.LLF"/>
61         </state>
62      </org>
```

3. irudia. Alderdi politikoei buruzko datuak

Metadatu horiek ere XML-TEI formatuan jaso ditugu. Alde batetik, zenbait erakunderi buruzko datuak bildu dira. Zehaztu egiten da Eusko Jaurlaritza Eusko Legebiltzarraren menpe dagoen gobernua dela. Bestetik, Eusko Jaurlaritzaren eta Eusko Legebiltzarraren izenak euskaraz,

gaztelaniaz eta ingelesez jarri dira. Bi erakundeen webguneak ere txertatuta daude metadatuetan, eta Eusko Legebiltzarraren kasuan, jaso ditugun transkripzioen legegintzaldia zein den aipatu dugu: hamargarren, hamaikagarren eta hamabigarren agintaldiak. Legegintzaldi horiek noiz hasi eta noiz bukatu diren ere adierazi dugu.

Ondoren, alderdi politikoei dagokien atala dator. 3. irudian ikus daitekeen moduan, alderdi politiko bakoitzaren izena, sigla eta webguneak (Wikipediako artikuluak hiru hizkuntzatan) gehitu ditugu dokumentuan. Erakunde politiko bakoitzaren sorrera zein egun edo urtetan izan zen ere aipatu dugu. Azkenik, alderdi politiko bakoitzak zer orientazio politiko duen (ezkerreskuin ardatza) ere erantsi dugu. Alderdi politikoei dagokienez, bukatzeko, agintaldi bakoitzean gobernuan eta oposizioan zer alderdi politiko egin diren ere gehitu dugu metadatuetan.

```
89      <person xml:id="AnduezaLorenzo">
90         <persName>
91            <surname>Andueza</surname>
92            <surname>Lorenzo</surname>
93            <forename>Eneko</forename>
94         </persName>
95         <sex value="M"/>
96         <birth when="1979-06-04">
97            <placeName>Eibar</placeName>
98         </birth>
99         <affiliation role="member" ref="#ES-PV" to="2022-07-01" from="2012-10-20"/>
100        <affiliation role="member" ref="#PSE-EE" from="2011-06-13"/>
101        <idno type="URI" subtype="wikimedia">https://eu.wikipedia.org/wiki/Eneko_Andueza</idno>
102     </person>
103     <person xml:id="AnguloGarcía">
104        <persName>
105           <surname>Angulo</surname>
106           <surname>García</surname>
107           <forename>Gustavo</forename>
108        </persName>
109        <sex value="M"/>
110        <birth when="1979">
111           <placeName>Gasteiz</placeName>
112        </birth>
113        <affiliation role="member" ref="#ES-PV" to="2022-07-01" from="2012-10-20"/>
114        <affiliation role="member" ref="#EP-IU" from="2020-07-12"/>
115        <idno type="URI" subtype="wikimedia">https://eu.wikipedia.org/wiki/Gustavo_Angulo_Garc%C3%ADa</idno>
116     </person>
```

4. irudia. Legebiltzarkideei buruzko datuak

Metadatuen dokumentuaren hurrengo atala legebiltzarkideek osatzen dute. 4. irudiak erakusten duenez, legebiltzarkideei buruzko hainbat datu bildu ditugu: izena, lehen bi abizenak, jaiotza-eguna (edo horren datua falta bada urtea), sexua, jaioterria, legebiltzarkidearen informazioa daraman esteka (normalean, Wikipediakoa edo Wikidatakoa), eta alderdi politikoetan izan duen afiliazioa. Kasu batzuetan, legebiltzarkide bat alderdi bat baino gehiagotan egon da. Beraz, ahal izan den kasuetan, alderdi politiko bakoitzean bere afiliazioa noiz hasi eta noiz bukatu zen zehaztu dugu. Bukatzeko, legebiltzarkide bakoitzari etiketa bat ere esleitu diogu. Etiketa hori bere bi abizenek osatzen dute. Etiketa honen bidez, metadatuen dokumentu hau eta eztabaida saioen beste dokumentuak elkarrekin lotzen dira.

Amaitzeko, badira beste zenbait dokumentu ParlaMinteko corpus guztietan berberak direnak. Horietan, besteak beste, eztabaida saioen, hizketaldiaren, alderdi politikoaren orientazioaren taxonomia biltzen da.

3.3. Corpusaren analisi linguistikoa egin

Hirugarren fasean, corpusaren analisi linguistikoa egin dugu. Hasteko eta behin, aurreko corpuseko dokumentuak bikoiztu egin ditugu. Ondoren, markatze linguistikoa egiten hasi gara, bigarren urratsean aipatu ditugun segmentuak baliatuz.

Lehendabiziko urratsa token bakoitza identifikatzea izan da. Hau da, hitzak, puntuazio markak eta hutsuneak elkarren artean bereiztu ditugu. Gero, esaldiak elkarren artean bereizi ditugu; eta horretarako, aurreko urratseko puntuazio markak erabili ditugu. Hirugarren urratsean, aldiz, segmentuetako hitz bakoitzaren lema identifikatu dugu. Lematizazioa deritzon prozesuaren bidez, hitz batek duen lema lortu dugu (adibidez, etxera -> etxe, naiz -> izan).

```
# newdoc id = ParlaMint-ES-PV_2019-12-20.u0
# newpar id = ParlaMint-ES-PV_2019-12-20.seg1
# lang = eu
# sent_id = ParlaMint-ES-PV_2019-12-20.seg1.s1
# text = Egun on guztioi.
1	Egun	egun	NOUN	NOUN	_	0	root	_	NER=O
2	on	on	ADJ	ADJ	_	1	amod	_	NER=O
3	guztioi	guzti	NUM	NUM	NumType=Card	1	det	_	NER=O|SpaceAfter=No
4	.	.	PUNCT	PUNCT	_	1	punct	_	NER=O

# sent_id = ParlaMint-ES-PV_2019-12-20.seg1.s2
# text = Osoko bilkurari hasiera emango diogu.
1	Osoko	osoko	ADJ	ADJ	_	2	amod	_	NER=O
2	bilkurari	bilkura	NOUN	NOUN	_	3	nmod	_	NER=O
3	hasiera	hasiera	NOUN	NOUN	Animacy=Inan|Case=Abs|Definite=Def|Number=Sing	4	obj	_	NER=O
4	emango	eman	VERB	VERB	Aspect=Prosp|VerbForm=Part	0	root	_	NER=O
5	diogu	*edun	AUX	AUX	Mood=Ind|Number[abs]=Sing|Number[dat]=Sing|Number[erg]=Plur|Person[abs]=3|Person[dat]=3|Person[erg]=1	4	aux	_
6	.	.	PUNCT	PUNCT	_	4	punct	_	NER=O

# newpar id = ParlaMint-ES-PV_2019-12-20.seg2
# lang = eu
# sent_id = ParlaMint-ES-PV_2019-12-20.seg2.s1
# text = Aldez aurreko gaia:
1	Aldez	alde	NOUN	NOUN	Case=Ins|Definite=Ind	3	nmod	_	NER=O
2	aurreko	aurre	ADP	ADP	Case=Loc|Definite=Def|Number=Sing	1	case	_	NER=O
3	gaia	gai	NOUN	NOUN	Animacy=Inan|Case=Abs|Definite=Def|Number=Sing	0	root	_	NER=O|SpaceAfter=No
4	:	:	PUNCT	PUNCT	_	3	punct	_	NER=O
```

5. irudia. Corpusaren analisi linguistikoa

Hitz guztiak lematizatu ondoren, Dependentzia Unibertsalak (*Universal Dependencies*, UD) baliatu ditugu analisi linguistikoaren zenbait urrats betearazteko. Besteak beste, gramatika-kategoriak (*Part-of-Speech*, PoS) eta ezaugarri morfologikoak etiketatu ditugu. Hurrengo urratsean, izendatutako entitateak (*Named Entities*, NE) etiketatu ditugu. Zehazki, izen bereziak izen arruntetatik bereizi ditugu, eta ondoren, izen berezi hori kategorizatu dugu (pertsona, lekua, erakundea edo produktua den zehaztuz). Corpusaren analisi linguistikoarekin amaitzeko, perpausen dependentzia sintaktikoak etiketatu ditugu, Dependentzia Unibertsalak erabiliz.

5. irudia corpusaren analisi linguistikoari dagokio. Bertan, analisi linguistikoaren emaitza ikus daiteke. Esaldiak segmentatuta daude (sent_id) eta segmentu bakoitzak kode bat du. Irudiko zutabeetan, hitz bakoitzari buruzko zenbait informazio biltzen da: bere lema, gramatika kategoria, dependentzia gramatikari buruzko informazioa eta izen berezi bat den edo ez.

3.4. Corpusa balioztatzea eta bihurketa

Laugarren fasean, corpusa baliozkotu eta formatu bihurketa egin dugu. Lehenengo urratsa corpusa baliozkotzea da, eta horretarako ParlaMint proiektuak eskuragarri jarri dituen RelaxNG eskemak baliatu ditugu. Honek ParlaMint corpus baterako izatez baliozkoak diren ele-

mentu, atributu eta eduki ereduak soilik onartzen ditu. Halako lau eskema daude batera, bata «testu arrunteko» corpus-errorako, bestea bere corpus-osagaietarako, bat linguistikoki etiketatutako corpus-errorako eta bestea bere osagaietarako. XML eskemekin balioztatzeak XML fitxategien egitura formala egiaztatzen du. Urrats honetan, badaude beste mota batzuetako zenbait baliozkotze ere.

3.5. Corpusa ikusgai eta erabilgarri jarri

Azken urratsean, Eusko Legebiltzarreko corpusa NoSketch Engine, KonText eta TEITOK tresnekin erabiltzeko moduan jarri da. Kasu honetan, tresna horiek CLARIN azpiegituraren parte direnez, eta ParlaMint proiektua bera ere azpiegituraren barnean dagoenez, CLARIN azpiegiturako zerbitzuek egin dute betebehar hori.

4. Emaitzak eta zabalkundea

Atal honetan, egin dugun lanaren emaitzak azalduko ditugu. Lehenik eta behin, 4.1 atalean, corpusa non jaitsi daitekeen, eta zer-nolako corpus motak jaitsi daitezkeen azalduko dugu. Ondoren, 4.2 atalean, eratu dugun corpusaren ezaugarriak deskribatuko ditugu, Besteak beste, bere tamaina, metadatuen zenbait ezaugarri eta azterketa linguistikoaren zenbait datu emango ditugu. Bukatzeko, 4.3 atalean, norbaitek corpusa erabili nahi badu ikertzeko, zer tresna dituen eskura eta horien ezaugarriak aipatuko ditugu.

4.1. Corpusa eskuratzeko moduak

Lan honen ondorioz, Eusko Legebiltzarreko corpusa ParlaMint proiektuan txertatzea lortu dugu. Corpusa eskuragarri dago CLARINen webgunean, eta bi modutan jaitsi daiteke: hizkuntza ezaugarriak etiketatuta edo hizkuntza ezaugarriak etiketatu gabe.

Alde batetik, ParlaMint v4.0 corpusa da. Bertan, Eusko Legebiltzarreko eta beste parlamentuetako corpusak «testu soila» deritzon formatuan daude. Hau da, bertako testuak jatorrizko hizkuntzan eta linguistikoki etiketatu gabe daude.

Beste alde batetik, ParlaMint.ana v4.0 motatako corpusa dago. Kasu honetan ere, corpusak jatorrizko hizkuntzan daude baina linguistikoki etiketatuta daude. Ezaugarri hauek daude linguistikoki etiketatuta: tokenizazioa; perpausen segmentazioa; lematizazioa; mendekotasun unibertsalak, gramatika-kategoriak (PoS), ezaugarri morfologikoak eta mendekotasun sintaktikoak; eta entitate izendunak (CoNLL-2003ren 4 klaseak aintzat hartuz).

Bukatzeko, ParlaMint-en.ana v4.0 motatako corpusa dago. Beste kasuetan ez bezala, corpusa ingelesera itzuli da itzulpen automatikoaren bidez, eta aurreko corpusa bezalaxe, linguistikoki etiketatuta dago.

4.2. Corpusaren ezaugarriak

Jarraian, corpusak dituen ezaugarri orokor batzuk azalduko ditugu. Hala ere, corpusaren inguruko ikerketa lan gehiago eta datu gehiago ere lor daitezke 4.3 atalean aipatu ditugun hiru tresna edo baliabideak erabiliz.

Corpusaren ezaugarriei dagokienez, corpusak guztira 15.591.266 token eta 13.321.393 hitz ditu. 657.729 esaldi, 278.211 paragrafo eta 39.144 dokumentu ere baditu corpusak. Corpusean ageri den lexikoaren ezaugarriei dagokienez, corpusak 209.062 hitz (hitz ez errepikatuak), 120.286 lema, 36 gramatika-kategoria (PoS), 883 dependentzia unibertsalen ezaugarri eta 202 dependentzia etiketa ditu.

Guztira, 190 eztabaida saioetako testuak daude jasota. Lehenengo eztabaida saioa 2015eko otsailaren 13koa da, eta azken eztabaida saioa; berriz, 2022ko urriaren 28koa. Horrenbestez, hiru legegintzaldietako testuak daude: Eusko Legebiltzarreko X. Legegintzaldia (2012-2016), Eusko Legebiltzarreko XI. Legegintzaldia (2016-2020), eta Eusko Legebiltzarreko XII. Legegintzaldia (2020-2024).

Izendatutako entitateei dagokienez, egiturak kontuan hartuz, 217.181 entitate erakundeak (ORG) dira (% 52), 80.394 pertsonak (PER) (guztiaren % 19,3), 71.876 lekuak (LOC) (% 17,2), eta azkenik, 48.146 denetik (MISC) (% 11,5). Datuak zerbait ezberdinak dira token kopurua neurtzen baldin bada: 377.194 token erakundeenak dira (% 55,5), 109.226 token pertsonenak dira (% 16,1), 105.393 token denetarikoak (% 15,5), eta azkenik, 87.655 token lekuzkoak dira (% 12,9).

Corpusean ageri den hizkuntzari dagokionez, tokenak aintzat hartzen baldin badira, 3.538.137 token euskaraz daude (guztiaren % 22,5) eta 12.053.129 token, aldiz, gaztelaniaz (% 77,3). Egituraren maiztasunari (hau da, esaldi eta dokumentu kopurua) begiratuz gero, 115.073 euskaraz daude (% 41,4) eta 163.138 gaztelaniaz (% 58,6). Datu hauek Escribano *et al.* (2022b) diotena berresten dute: oro har, euskara hizketaldi gehienetan agertzen da, baina hizketaldi luzeagoetan bere presentzia txikiagoa da. Horregatik dago hainbesteko ezberdintasuna ehunekoetan egituraren maiztasunaren eta tokenen artean.

Generoari dagokionez, gobernukideak ere kontuan hartuta, 29.592 egitura (hots, hizketaldi kasu honetan) (% 75,6) emakumezkoek egin dituzte. Aldiz, 9.552 hizketaldi (% 24,4) gizonezkoek egin dituzte. Datuak zeharo ezberdinak dira, tokenak aztertzen badira: 7.880.341 token (% 50,5) gizonezkoek egin dituzte, eta 7.710.925 token (% 49,5) emakumezkoek. Beraz, badirudi, emakumezkoek hizketaldi asko eta laburrak dituztela, eta gizonezkoek, ordea, hizketaldi gutxi baina luzeak. Hor Eusko Legebiltzarreko presidentearen eragina ere egon daiteke (presidentea izanik hizketaldi gehiago egiten dituelako).

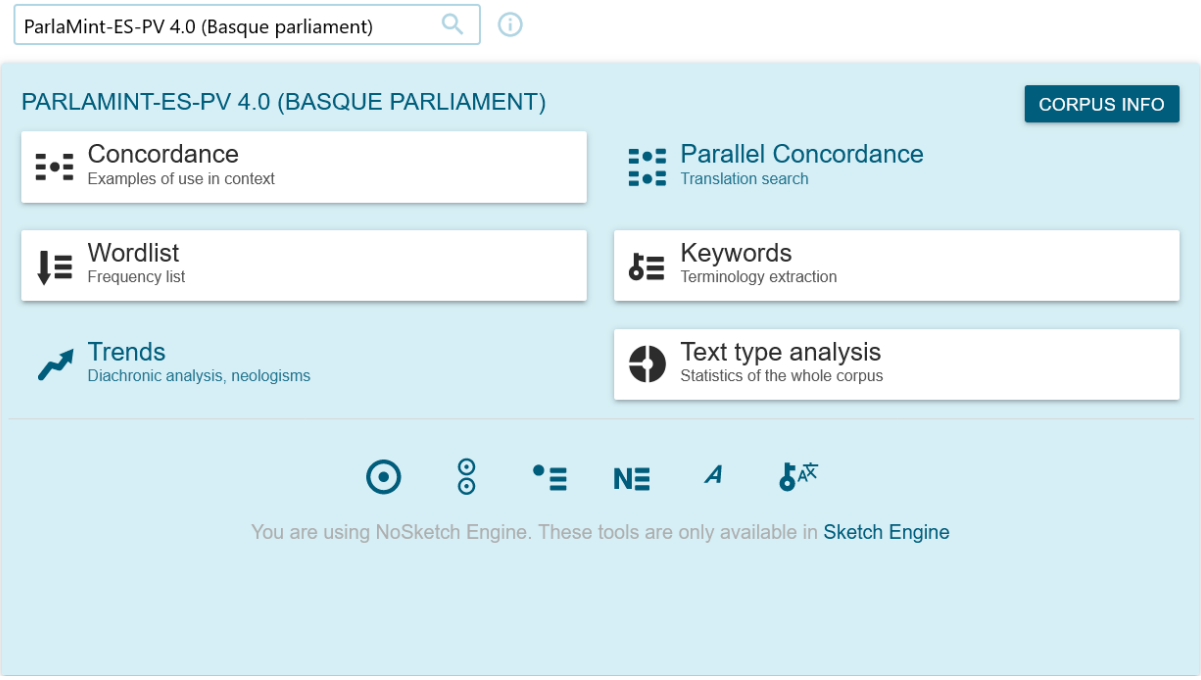
Alderdien orientazio politikoari dagokionez eta gobernukideak ere kontuan hartzen badira, hizketaldi kopuruari dagokionez, 26.572 hizketaldi (% 67,9) zentrokoak diren alderdienak dira. 3.747 hizketaldi (% 9,6) ezkerretik ezker muturrera bitarteko alderdienak dira, 3.291 hizketaldi (% 8,4) eskuinetik eskuin muturrera bitarteko alderdienak, eta 2.747 hizketaldi (% 7) zentro-ezkerreko alderdienak. Azkeneko lekuan, eskuin muturra dago 365 hizketaldirekin (% 0,9). Emaitzak zeharo ezberdinak dira, tokenak aztertzen baldin badira: 4.664.530 token (% 29,9) zentrokoak, 3.141.352 token (% 20,1) ezkerretik ezker muturrera bitartekoak, 2.513.395 token (% 16,1) zentro ezkerrekoak, 1.629.899 token (% 10,5) ezkerrekoak, eta berriz ere, gutxiena eskuin muturrekoak dira: 252.444 token (% 1,6). Kasu honetan, alde handia dago hizketaldien eta tokenen artean zentroan, eta hori ere Eusko Legebiltzarreko presidentearen eraginez izan daiteke.

4.3. Corpusa aztertzeko tresna eta baliabideak

Aurreko atalean aipatu ditugunak corpusak dituen ezaugarri orokorrak dira. Hala ere, CLARIAN azpiegiturak, ParlaMint proiektuaren bultzatzaile izanik, baditu zenbait tresna eta baliabide ikertzaileei corpus horien azterketa egitea ahalbidetzen dietena. Jarraian, NoSketch Engine,

KonText eta TEITOK tresnak Eusko Legebiltzarreko corpusean nola erabil daitezkeen ikusiko dugu.

NoSketch Engine tresnak (Rychlý, 2007) corpusa aztertzeko hainbat modu eta aukera eskaintzen ditu. 6. irudian, tresnaren hasierako gunea ikus daiteke.



6. irudia. NoSketch Engine tresnaren sarrerako orria

CORPUS INFO botoian klik eginez gero, 4.2 atalean erakutsi diren datuak eta gehiago ikusteko aukera ematen du. Bertan, corpusa metadatuz osatuta dagoenez, bilaketak eta estatistikak irizpide edo ezaugarri jakin baten bidez egiteko aukera ematen du. Bestetik, sarrerako orri honetan ikus daitekeen moduan, tresnak beste lau osagai ere baditu: konkordantziak bilatzea, hitz-zerrenda lortzea, gako-hitzak aurkitzea eta testu-motaren analisia.

Konkordantzia bilaketa-aukera ugari dituen tresna da. Hitzak, esaldiak, etiketak, dokumentuak, testu motak edo corpus-egiturak bilatzen ditu eta emaitzak testuinguruan bistaratzen ditu konkordantzia moduan. Tresnak hitz zerrenda bat lortzeko aukera ere ematen du, eta irizpide hauen arabera egin ditzake bilaketak: (i) izenak, aditzak, adjektiboak eta beste gramatika kategoria batzuen bidez, (ii) hasierako, amaierako, zenbait karaktere dituzten hitzen bidez, eta azkenik, (iii) hitz-formak, etiketak, lemak eta beste atributu batzuen bidez.

Gako-hitzak bilatzeko aukera ere ematen du NoSketch Engine tresnak. Kasu honetan, gako-hitzak eta terminoak ateratzeko hiru aukera ematen ditu: (i) itzulpenean eta interpretazioan erabilitako terminologia erauztea, (ii) corpus/dokumentu/testu baten ohikoak diren edo bere edukia edo gaia definitzen duten hitz bakarreko eta hitz anitzeko unitateak ateratzea eta (iii) bi corpus/dokumentu/testu alderatu, eta lehen corpusean bakarra dena eta bigarrenean ez dagoena identifikatzea.

Corpusa aztertzeko CLARIN azpiegiturak eskura jarri duen beste tresna KonText (Machálek, 2020) da. Aipaturiko tresna corpusaren kontsultarako da. Bi kontsulta mota egiteko aukera eskaintzen du: kontsulta sinplea edo kontsulta aurreratua.

Corpusa aztertzeko eskuragarri dagoen azken tresna TEITOK (Janssen, 2016) da. TEITOK tresna KonText tresnaren osagarria da. TEITOK-ek corpusean bilaketak egiteaz gain (KonText tresnak egiten duen moduan), dokumentu indibidualak bistaratzeko eta corpusak editatzeko aukera ere ematen du. Corpus estatikoak eta corpus biziak ere bereizten ditu TEITOK tresnak. Corpus estatikoak estilo tradizionaleko corpusak dira, baliabide finko gisa eskuragarri daudenak, eta bilaketa erreproduzigarriak egiteko aukera ematen dutenak. Corpus biziak, aldiz, garatzen ari diren eta denborarekin hobetzen ari diren corpusak dira.

5. Ondorioak eta etorkizuneko lana

Lan honen ondorioz, Eusko Legebiltzarreko eztabaida saioetako datuak corpus batean bildu ditugu, eta hori CLARIN azpiegiturak bultzaturiko ParlaMint 4.0 proiektuan txertatu dugu. Eusko Legebiltzarreko corpus hau euskaraz eta gaztelaniaz idatzita dago, eta guztira 15.591.266 token, 13.321.393 hitz, 657.729 esaldi, 278.211 paragrafo eta 39.144 dokumentu barne hartzen ditu. Corpusa ParlaMint 4.0 proiektuan txertatuta dagoenez, corpusa osatzeko zenbait jarraibide bete behar izan ditugu: hala nola, eztabaida saioen inguruko metadatuak sortu, paragrafoak eta esaldiak segmentatu, paragrafo eta esaldi horiek metadatuz elikatu (hizkuntza zehaztu, egilea identifikatu eta abar.). Halaber, legebiltzarkideei eta alderdi politikoei buruzko datuak ere bilduta daude. Corpusa askotariko arloei buruzko ikerketa lanak egiteko edonoren eskura dago, humanitate digitalen helburuekin bat eginez.

Etorkizunera begira, Eusko Legebiltzarreko corpus honekin lanean jarraitu nahiko genuke. Alde batetik, Eusko Legebiltzarra sortu zenetik eskuragarri dauden eztabaida saioen testuak dagoen corpusari gehitu nahiko genizkioke. Izan ere, orain dagoen corpusa 2015etik 2022ra artekoa da eta horrekin zenbait gairen inguruko ikerketa-lanak egin daitezkeen arren, muga batzuk ere sor daitezke. Eusko Legebiltzarra 1980an sortu zenez, urte horretatik aurrerako eztabaida saioen testu guztiak ParlaMinten gehitu nahiko genituzke.

Bestetik, Eusko Legebiltzarrreko transkripzioekin euskara hutsezko corpus bat ere sortu nahiko genuke. Orain eskuragarri dagoen corpusean, euskara eta gaztelania txandakatzen diren arren, gaztelania da nagusi. Horren ondorioz, Eusko Legebiltzarreko corpusa ezin esan liteke euskara hutsezko corpus bat denik. Kontuan hartuta hizkuntza gutxituen kasuan tresna eta baliabide falta dagoela, Eusko Legebiltzarreko eztabaida saioen euskara hutsezko corpus bat sortzea interesgarria izango litzateke.

Egun eskuragarri ditugun testu eta baliabideekin hori egin daiteke. Metodologia atalean aipatu dugun moduan, Eusko Legebiltzarrarekin egindako hitzarmenaren bidez, eztabaida saioen hainbat motatako dokumentuak jaso ditugu: batzuetan, euskara hutsezko dokumentuak, beste batzuetan gaztelania hutsezko dokumentuak, eta bukatzeko, bi hizkuntzak dituzten dokumentuak (ezkerreko zutabean, eztabaida saioko benetako hizketaldia euskaraz edo gaztelaniaz eta eskuineko zutabean, benetako hizketaldiaren itzulpena euskaraz edo gaztelaniaz).

Gure ustez, eztabaida saioen benetako testuak eta itzulitako testuak batuz Eusko Legebiltzarreko eztabaida saioen euskara hutsezko corpus bat sor genezake, eta hori interesgarria da politika arloko euskara hutsezko corpusik ez dagoela kontuan hartuta (hedabide eta unibertsitate eremuko euskara hutsezko corpusak badaude).

6. Eskertza

Lan hau egiteko CLARIN-ERIC eta CLARIAH-EUS erakundeen laguntza izan dugu.

Bibliografia

Alkorta, J., & Iruskieta, M. (2022). Adding the Basque Parliament Corpus to ParlaMint Project. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference* (107-110 orr.).

Bayley, P. (2014). Introduction: The whys and wherefores of analyzing parliamentary discourse. In P. Bayley (Ed.), *Cross-cultural perspectives on parliamentary discourse* (1-44 orr.). John Benjamins Publishing.

Bordel, G., Nieto, S., Penagarikano, M., Rodriguez-Fuentes, L. J., & Varona, A. (2011). Automatic subtitling of the basque parliament plenary sessions videos. In *Twelfth Annual Conference of the International Speech Communication Association*.

Bordel, G., Penagarikano, M., Rodríguez-Fuentes, L. J., & Fernández, M. A. V. (2012). Aligning very long speech signals to bilingual transcriptions of parliamentary sessions. In *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings* (69-78 or.). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bordel-García, G., Penagarikano-Badiola, M., Rodríguez-Fuentes, L. J., Varona-Fernández, A. (2013). OBAM-PV: una aplicación para el subtitulado de vídeos de Sesiones Plenarias del Parlamento Vasco. *Actas del XXIX Congreso de la Sociedad Espanola de Procesamiento de lenguaje natural, SEPLN*. 2013.

Burdick, A.; Drucker, J.; Lunenfeld, P.; Presner, T.; Schnapp, J. (2012). *Digital_Humanities (PDF)*. Open Access eBook: MIT Press. ISBN 9780262312097

Cartoni, B., Zufferey, S., & Meyer, T. (2013). Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, *27*(1), 23-42 orr.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. Discourse & Society, 26(5), 562-586 orr.

Drucker, J. (2013). «Intro to Digital Humanities: Introduction». UCLA Center for Digital Humanities.

Erjavec, T., Ogrodniczuk, M., Osenova, P. *et al.* (2023). The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation* **57**, 415-448 orr. https://doi.org/10.1007/s10579-021-09574-0

Escribano, N., González, J. A., Orbegozo-Terradillos, J., Larrondo-Ureta, A., Peña-Fernández, S., Perez-de-Viñaspre, O., & Agerri, R. (2022a). Basqueparl: A bilingual corpus of basque parliamentary transcriptions. *arXiv preprint arXiv:2205.01506*.

Escribano-García, N., González, J. A., Orbegozo-Terradillos, J., Larrondo-Ureta, A., Peña-Fernández, S., Perez-de-Viñaspre, O., & Agerri, R. (2022b). Euskararen erabilera Eusko Legebiltzarreko debateetan (2012-2020). *Mediatika. Cuadernos de Medios de Comunicación*, (19).

Etchegoyhen, T., Arzelus, H., Ugarte, H. G., Alvarez, A., González-Docasal, A., & Fernandez, E. B. mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation. In *Proceedings of IberSPEECH 2020*, 190-194 or., 2021.

Hirst, G., Wei Feng, V., Cochrane, C., & Naderi, N. (2014). Argumentation, ideology, and issue framing in parliamentary discourse. In *ArgNLP*. ftp://www.cs.toronto.edu/pub/gh/Hirst-etal-Bertinoro-2014.pdf

Janssen, M. (2016). TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 4037-4043 orr.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers* (79-86 orr.).

Kuzman, T. *et al.* (2023). *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1810

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 7003-7008 orr.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešic, N., Çagrı Çöltekin, Kopp, M., Meden, K. (2022). ParlaMint II: The show must go on, in: D. Fišer, M. Eskevich, J. Lenardič, F. de Jong (Ed.),

Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseilla, Frantzia, 1-6 orr., URL: https://aclanthology.org/2022.parlaclarin-1.1.pdf

Pančur, A., & Šorn, M. (2016). Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History. Prispevki za novejšo zgodovino **56**(3), 130-146 orr. https://ojs.inz.si/pnz/article/view/193

Penagarikano, M., Varona, A., Bordel, G., & Rodriguez-Fuentes, L. J. (2023). Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Applied Sciences*, *13*(14), 8492.

Rama, T., & Borin, L. (2011). Estimating language relationships from a parallel corpus. A study of the Europarl corpus. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011),* 161-167 orr.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLoS ONE, 11*(12), 1-18 orr. https://doi.org/10.1371/journal.pone.0168843

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. In *RASLAN,* 65-70 orr.

Terras, M. (2011). «Quantifying Digital Humanities» *(PDF). UCL Centre for Digital Humanities.*

Torralba-Rubinos, C.M. (2021). Itzulpena eta interpretazioa: zurekin baina zu gabe. In Arbelaitz, O., Latatu, A., Ormaetxebarria, M.J. & Urgell, B., *IV. Ikergazte nazioarteko ikerketa euskaraz, giza zientziak eta artea*, 35-46.

van Dijk, T. A. (2010). Political identities in parliamentary debates. In C. Ili (Ed.), *European parliaments under scrutiny: Discourse strategies and interaction practices,* 29-56 orr.. John Benjamins Publishing.

Van Halteren, H. (2008). Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* 937-944 orr.

Ziterauzi: euskarazko artikulu akademikoetatik zitazioak erauzteko tresna-katea

Ziterauzi: the Tool Chain for Citation Extraction from Basque Academic Texts

Aitzol Astigarraga[1], David Lindemann[2], Marije Bidaguren[2]

[1] Goimailako Online Institutua-Udako Euskal Unibertsitatea
a.astigarraga@ueu.eus

[2] Udako Euskal Unibertsitatea
david.lindemann@ehu.eus m.bidaguren@ueu.eus

Laburpena

Euskaraz argitaratutako artikulu zientifikoen kopurua gora egin ahala, euskararen erabilera akademikoan garrantzi nabarmena duten galderak sortzen zaizkigu: zeintzuk dira zientzialarien erreferentziazko lanak arlo bakoitzean eta zer pisu eta erabilera dauka euskarak mundu akademikoan, besteak beste.

Galdera horiei eta antzekoei erantzuteko azpiegitura sortu nahi du ZITERAUZI egitasmoak, euskaraz argitaratzen diren artikuluen zitazioak erauziz, grafo batean jasoz eta azterketa bibliometrikoak ahalbidetuz.

Aurkezten dugun proiektu hau Humanitate Digitalen esparruan kokatzen da. Euskarazko ekoizpen zientifikoa biltzen duen Inguma datu-basea abiapuntu gisa hartuta, argitalpen zientifikoen metadatuak (gaur egun Inguma-n jasota dauden argitalpen-metadatuak) testu osoak prozesatuz aberastea proposatzen dugu, zitazio-erlazioak erauziz eta erakutsiz.

Izan ere, ikerketa-kopurua gora egin ahala artikuluak bilatzeko metodo tradizionalak eraginkortasuna galtzen ari dira. Lotutako Datu Irekiak (LOD, *Linked Open Data*) deritzen web semantikoaren teknologiek argitalpen bildumak jaso, antolatu eta aztertzeko aukera berriei bide eman diete eta hor ere kokatzen da gure ekarpena.

Aurkezten dugun lan honek bi helburu zehatz dauzka: batetik, ikerkuntzaren testuinguruan, zitazio-erlazioen grafo libreak daukan garrantzia nabarmentzea, bai azterketa bibliometrikoetarako, bai eta ikerkuntzaren beraren ebaluaziorako ere; eta, bestetik, euskaraz argitaratzen diren artikuluetako aipuak erauzi eta sare zuzendu batean biltzeko tresna-katea erakustea: ZITERAUZI.

Eraiki nahi dugun zitazio-grafo horrek hainbat erabilera izan ditzake eta euskarazko ekoizpen akademiko-zientifikoaren azterketan, ebaluaketan eta bilaketan aurrerapausoak ekarriko ditu.

Errenteriako Humanitate Digitalen gunearekin lankidetzan (Astigarraga, Iñurrieta *et al.*, 2021), 2024. urtean Inguma datu-basean jasotako IKERGAZTE kongresuko artikuluen erauzketa burutuko dugu eta horrek balioko digu ZITERAUZIren ebaluazio zehatzagoa egiteko. Tresna-katea eta eskuzko lan-fluxua prest geratuko dira bilduma gehiago prozesatzeko. Eskuzko lanari dagokionez, proiektu honetan lortutako prozesu metadatuek lan-kargaren aurreikuspena ahalbidetuko dute, hau da, argitalpen baten bibliografiatik lortutako zitazio-erlazioak eskuz balidatzeko batez besteko denborak.

Artikulu honetan gaia behar bezala kokatzeko oinarri teorikoak azalduko ditugu, zitazio-erlazioak bezalako metadatu akademikoei formatu egituratua eta konputagailuetarako irakurgarria nola eman, eta hori egiteko egun eskura dauden tresnak erakutsiz.

Gako hitzak: Inguma, Zientzia Corpusa, Bibliometria, Zitazioen Erauzketa, Datu Lotuak.

Abstract

As the number of scientific articles published in Basque is constantly growing, important questions about the academic use of Basque arise: What are the reference works in each area and the use of Basque in the academic world, among others.

The ZITERAUZI project aims to create an infrastructure for addressing these and similar questions, extracting citation relations from and to scientific articles published in Basque, representing them as graphs, and allowing bibliometric studies.

The project we present is located in the field of digital humanities. Starting point is the Basque scientific production database Inguma. We propose to enrich the metadata of scientific publications currently part of Inguma by representing extracted citation relations.

In fact, traditional methods of article searches, as the number of research publications increases, are losing efficiency. Semantic web technologies (Linked Open Data), have given way to new opportunities for recording, exhibiting, and querying publication collections. This is where our contribution is aiming at.

This article follows two specific goals: on the one hand, to highlight the importance of open citation graphs for bibliometric studies and research evaluation; and, on the other hand, to discuss a tool chain for citation extraction from Basque scientific articles and their representations in a directed graph, that is, ZITERAUZI.

The graph that we intend to build serves as infrastructure for different use cases, which together allow progress in the study of the scientific production in Basque.

In collaboration with the Digital Humanities Centre in Errenteria (Astigarraga, Iñurrieta *et al.,* 2021), in 2024, we will carry out a pilot study on articles from the IkerGazte conference series in the Inguma database, enabling a more detailed evaluation of Ziterauzi. The tool chain and manual work flow will then be ready to process further collections. Regarding manual efforts, process metadata obtained in this first study will allow us to make predictions of the workload necessary to manually validate the citation relations extracted from the bibliography section of a publication.

In this article, we enquire into theoretical fundamentals, we delve into techniques for effectively representing structured metadata such as citation relations in a machine-readable format, and we discuss relevant tools available today.

Keywords: Inguma, Science Corpus, Bibliometry, Citation Extraction, Linked Data.

1. Sarrera

Gaur egun, argitalpen akademikoen kopurua ikaragarria da. 2014. urtean, 114 milioi artikulu zeuden sarean eskuragarri (Khabsa & Giles, 2014), 100 milioi ingelesez, eta 27 milioi atzipen librekoak izanik. Ordutik hona kopuru horrek gora egin du etengabe eta urtero 1,4 milioi artikulu akademiko inguru argitaratzen direla kalkulatzen da. Argitalpen horiek bildu, prozesatu, sailkatu eta modu egokian erakusteko ahalegina ikerketa-eremu aktibo bihurtu da, eta egun, Bibliotekonomia arloko ikerketa zena Informazio Zientzia (*Information Science*) diziplinaren parte da (Stock & Stock, 2013; Yan, 2011).

Argitaratutako ikerketen kopurua handitzearekin batera, artikuluak bilatzeko metodo tradizionalak eraginkortasuna galtzen ari dira gai jakin baterako artikulu garrantzitsuak zehazteko orduan, eta bilaketarako bide eta estrategia berriak beharrezkoak bihurtu dira. Alde horretatik, artikuluetako zitazio-erlazioen azterketa, lan akademikoen garrantzia neurtzeko metodo gisa, ikerketa askoren xede izan da azken urteotan (ikus Nasar *et al.* —2018— ikuspegi orokor baterako).

Gai honen garrantzia hobeto ulertzeko, beharrezkoa da jakitea lan akademiko batean (artikulu zientifikoa, liburua, txostena) sartzen diren aipamen edo erreferentzia bibliografikoak komunikazio zientifikoaren funtsezko osagai direla. Horien bidez, ikerketa zein testuingurutan kokatzen den jakin daiteke, eta argitalpenek ezagutza berriaren sorreran duten eragina ikus daiteke.

Gero eta argitalpen eta testu akademiko gehiago daudenez, ezinezkoa da literaturako datubaseak eskuz osatzea. Erreferentziak automatikoki erauziz eta prozesatuz ikerketa-komunitatea karga horretatik askatzen dugu eta literaturaren azterketa sakonagoa egiteko hainbat aukera sortzen ditugu. Horren ekarpenetako bat argitalpenaren historia ulertzea da: autoreek elkarri nola eragiten dioten, zein erakundek parte hartzen duten zenbait eremutan eta gaitan, etab. Hori lortzeko modu bat erreferentzia-grafo bat eraikitzea da artikuluetako zitazio-informazioa erabiliz. Sistema horri esker, erraz eskura daiteke zein diren autore jakin baten funtsezko informazio-iturriak edo erreferentziak, baita datu-iturriak ere, edota etorkizunerako lanak.

Elkarrekin erlazionatutako erreferentzien datu-baseak aipu-indize gisa ezagutzen dira. Web of Science[1] (WoS) eta Scopus,[2] estaldura orokorra dutenak, ikerketa bibliometrikoetarako ohiko iturriak dira. Izan ere, ebaluazio zientifikoa gero eta gehiago oinarritzen da aipu-adierazleetan (inpaktu-faktorea, h indizea, etab.), eta indize horiek funtsezko pieza bihurtzen ari dira komunikazio zientifikoaren sisteman. Hala ere, indize horiek enpresa pribatuek garatzen dituzte (Clarivate WoSen kasuan eta Elsevier Scopusen kasuan), eta beraien indizeetan dauden aldizkarietako aipuak soilik prozesatzen dituzte. Beraz, estaldura mugatua dute eta iturriak, gainera, itxiak.

Eredu horren alternatiba bat bilatzaile akademikoak dira, argitalpen akademikoen metadatuak bildu eta eskaintzen dituzten plataformak. Weba automatikoki arakatzen dute argitalpen zientifikoen bila, haietan sartutako zitazioak erauziz (Gusenbauer, 2019). CiteSeer (1998) aitzindariaz geroztik, Google Scholar (2004), Microsoft Academic (2011) eta SemanticScholar (2015) arte, produktu horiek aipu-indizeen lehiakide bihurtzen ari dira, aipuak azkarrago kontabilizatzen baitituzte (Singh *et al.,* 2022; Thelwall & Kousha, 2017). Desabantaila nagusiak hauek dira: prozesu automatikoa denez, errazago manipulatzen dira (Delgado López-Cózar *et al.,* 2014), eta akats eta tirabira gehiago eragiten dituzte aipuak identifikatzeko eta zenbatzeko orduan. Aldi berean, plataforma orokorretan, metadatu bibliografikoak ez ohi dira eskuz landu, eta, ondorioz, askotan ez dute zitazioak egiteko bezainbesteko kalitaterik.

---

[1] https://www.webofscience.com/wos/

[2] https://www.scopus.com/home.uri

Aipuen indize bat osatzeko oztopo nagusia ondokoa da: elkarren arteko aipuak dituzten dokumentu (corpus) asko behar direla. Beraz, zenbat eta dokumentu gehiago izan, orduan eta errazagoa izango da haien arteko aipuak egotea. Ez da harritzekoa WoSen edo Scopusen artikulu batek izan dezakeen batez besteko aipamen-kopurua Google Scholarren baino askoz txikiagoa izatea, Google Scholarren tamaina aurrekoak baino hainbat aldiz handiagoa baita (Martín-Martín *et al.,* 2021). Baina horrek konpartimentu estankoak sortzeko arazoa du, non dokumentu baten inpaktua datu-base jakin baten arabera behatu behar baita beti.

Azken urteotan, web semantikoak eta Lotutako Datu Irekien paradigmak eredu deszentralizatuak eraikitzeko modua erraztu dute, non dokumentu baterako aipua biltegi ezberdinetan egon baitaiteke, dokumentu batek hainbat iturriren arabera duen eragina ezagutzeko aukera emanez. Modu horretan dokumentu zientifikoetatik erreferentziak ateratzeko mekanismo sinpleagoak eta gardenagoak eraikitzea posible da, informazio bibliografikoaren bolumen handiak kudeatzeko eta, bibliometriak eta ebaluazio zientifikoak estaldura eta zehaztasun handiagoa izan dezaten lortuz.

2. Ezagutza-grafoak

Web semantikoa, Tim Berners-Leek proposatu zuen moduan, World Wide Webaren hobekuntza bat da (Berners-Lee *et al.,* 2023). Webguneen edukia software-agenteentzat ulergarri bihurtzean datza, datuak partekatu, aurkitu, integratu eta berrerabiltzeko[3]. Makinek mundu errealaren ezagutza uler dezaten, ontologien bidez kontzeptuak eta kontzeptuen arteko erlazioak definitzen dira eredu formal bati jarraituz, kontzeptuek eta erlazioek grafo bat osatzen dutelarik. Web Semantikoaren adierazpenak bi datu-elementuk eta batetik bestera bidaltzen duen erlazio batek osatzen dituzte.

Hiru osagai horiek URI[4] banaren bidez identifikatu egiten dira modu unibokoan, makinari subjekto-predikatu-objektu hirukote horiek ulertarazteko.

Datu Lotuak (Linked Data)[5] Interneten argitaratzeko molde bat da, datuak Web Semantikoaren parte izatea ahalbidetzen duena. Argitaratutako datuak Datu Lotutzat har daitezen, Tim Berners-Leek aipatutako printzipioei jarraituz[6] argitaratu behar dira. Lotutako Datu Irekiak *(LOD, Linked Open Data)* lizentzia irekia duten datu lotuak dira. Hainbat datu-bilduma LOD gisa argitaratu eta elkarrekin lotu dira dagoeneko[7]. Web Semantikoaren helburua litzateke, Datu Lotuen bidez, egungo webaren edukia makinek ulertzeko moduan birmoldatzea, iturri ezberdinetako datusortak elkarrekin lotuz.

Ezagutza-grafoek, Web Semantikoaren parte edo baliabide isolatu gisa ulertuta, kontzeptuen eta entitateen arteko loturak adierazten dituzte. Eredu ontologiko batean oinarrituta, arlokako informazio nahiz ezagutza orokorra eskaintzen da ezagutza-grafo eran.

Oro har, SPARQL (*Protocol and RDF Query Language*) izeneko hizkuntza espezifikoa erabiltzen da lotutako datuak kontsultatzeko. SQLren (*Structured Query Language*) antzeko logika bati jarraituz, lotutako datu-baseak kontsultatu daitezke beharrezko informazioa berreskuratzeko.

---

[3] Lau kontzeptu horiek FAIR principles gisa ezagunak dira, findable, accessible, interope ra ble, reusable, ikus https://www.go-fair.org/fair-principles/

[4] Ikus https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

[5] Ikus https://en.wikipedia.org/wiki/Linked_data

[6] Ikus https://5stardata.info eta https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

[7] Ikus https://lod-cloud.net/ eta https://www.lod-cloud.net/dataset/wikidata

Hala ere, biltegi askok aukera gehiago eskaintzen dituzte, hala nola REST APIak, zenbait parametroren arabera erregistroak deskargatzeko aukera ematen dutenak, edo zerbitzuko erregistro guztiak dituzten fitxategi gordinak (*dump files*).

3. Zitazio-grafoak

Zitazio-grafoa, edo zitazio-sarea, ikerketa-arloko elementuak (oro har artikuluak, aldizkariak eta liburuak) elkarren artean nola erlazionatzen diren deskribatzeko erabiltzen den grafo zuzendua da (Życzkowski, 2010).

Hainbat jarduera garrantzitsu egiteko aukera ematen du, hala nola (Buneman *et al.*, 2021):

— Grafoa arakatu argitalpen interesgarriak aurkitzeko.
— Artikulu egileen jarraipena: aipamenen bitartez egileen lana aitortzeaz gain, haien lanen jarraipena egiteko aukera ematen du.
— Ikerketa-arloko aurrerapenen zabalkundea: zitazioak eta autoreak aztertuz, ikertzaileen komunitate sakabanatuek beren aurkikuntzak partekatu eta eztabaidetan parte har dezakete.
— Elkar maiz aipatzen dutenen autore-klusterrak identifikatzea, gai berdintsuak lantzen dituzten komunitate zientifikoak bistaratzeko.
— Bibliometria konputatzea ikertzaile, kongresu edo argitalpenaren eragina eremu partikularretan aztertzeko. Zitazio-grafoa da gaur egun erabiltzen den ia bibliometria guztiaren oinarria, hala nola inpaktu-faktorea eta h-index.

Laburbilduz, batez ere zitazioen analisirako, bilaketa akademikorako eta ikerketaren ebaluaketarako tresna gisa erabiltzen dira.

Gaur egun plataforma ugari daude, ezagutza-grafo gisa, aipuak jaso eta publikoki eskaintzen dituztenak. Atal honetan argitalpen akademikoaren datu-multzo garrantzitsuenak berrikusiko ditugu, harpidetzarik behar ez dutenak, hau da, metadatu bibliografiko akademikoak dituzten datu-multzo publikoak.

3.1. Crossref

Irabazi asmorik gabeko elkartea da Crossref, argitaletxeen lanak jaso eta identifikatzaile uniboko bat (Digital Object Identifier, DOI) esleitzea helburu duena (Hendricks *et al.*, 2020).

Crossref 2000. urtearen hasieran sortu zen, aldizkarietako editoreen arteko lankidetza-ahalegin gisa eta, aldizkari akademikoetan, plataforma ezberdinen arteko zitazio-erlazioak ahalbidetzeko tresna gisa. 2023an, 150 milioi metadatu-erregistro identifikatu eta lotzen ditu mugarik gabe berrerabiltzeko eskuragarri dauden ikerketa-objektuen gainean. Hilean, batez beste, 1.000 milioi metadatu-kontsulta izaten dituzte[8].

Hala ere, Crossref-ek zitazioak irekian uzteak agerian utzi ditu arazo batzuk. Hainbat artikuluk ez dituzte beren metadatuetan erreferentziak jasotzen. Beste kasu batzuetan, zitazioaren kalitatea hain da eskasa, ezinezkoa baita aipatutako dokumentuarekin lotura ezartzea. Arazo horiek guztiak direla eta, Crossref-eko datu gordinak aurrez prozesatu behar dira, zitazio indizeetan erabili ahal izateko (Ortega, 2021).

[8] https://www.crossref.org/

3.2. OpenAlex

OpenAlex (Priem *et al.,* 2022), 2022an sortu zen *Microsoft Academic Graph* bertan behera geratu ostean, bere datu guztiak jasoz.

Bertan jasotako lanak dokumentu akademikoak dira, hala nola aldizkarietako artikuluak, liburuak, datu multzoak eta tesiak. Guztira 240 M dokumentu baino gehiago biltzen ditu.

OpenAlexek identifikatzaile iraunkorrak esleitzen dizkio baliabide bakoitzari. Datuak CC0 lizentziapean banatzen dira, eta API bidez edo web interfazearen bidez eskura daitezke.

3.3. Semantic Scholar

Semantic Scholar 2015. urtean sortu zen, eta bilatzaile akademikoen merkatuan lehiatzen da adimen artifizialean oinarritutako irtenbideak eskainiz. Allen Institute for Artificial Intelligence-k garatua, bilatzaile horrek, web akademikoa arakatzeaz gain, biltegi gisa ere jarduten du, emaitza zientifiko baten bertsio irekiak deskargatu eta gordetzen baititu. Semantic Scholar-ek automatikoki erauzi daitezkeen erreferentziak soilik eskaintzen ditu (Fricke, 2018).

3.4. Springer Nature SciGraph

Springer Nature argitalpen-taldearen eta Digital Science zientzia-informazioan espezializatutako enpresaren arteko ekimena da. Web semantikoaren teknologietan oinarrituta, argitalpen-taldeak sortutako argitalpenei balioa emango dien ezagutza-sistema sortzea du helburu (Hammond *et al.,* 2017). Oraingoz, zitazioak Springer Nature-n edukiekin bakarrik lotzen dira, baina datu lotuen ereduak beste corpus batzuekin integratzeko aukera emango du.

3.5. Open Citations

OpenCitation irabazi-asmorik gabeko erakunde independentea da, eta Semantic Web teknologietan oinarritutako datu bibliografikoak eta zitazio-datuak argitaratzen ditu. Zitazio akademikoei buruzko informazioa biltegiratu eta kudeatzen du OpenCitation-ek (Peroni & Shotton, 2020). 2010ean sortu zenetik, ondokoak dira abian jarri dituen zitazio-indize garrantzitsuenak:

— Open Citations Corpus (OCC).[9]
— OpenCitations Index of Crossref open DOI-to-DOI Citations.[10]
— Crowdsourced Open Citations Index (CROCI).[11]

3.6. Wikidata

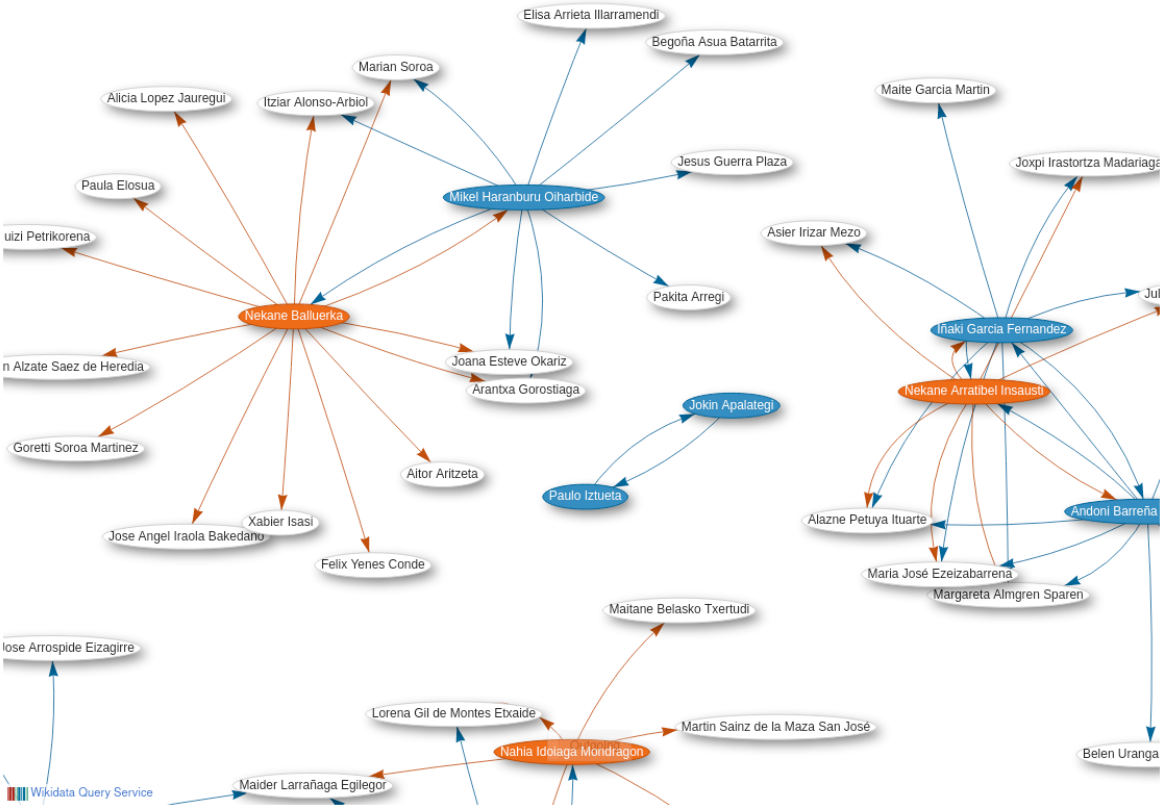
Wikidata auzolanean sortutako ezagutza-grafo eleaniztun bat da, Wikimedia Fundazioak 2012. urtean sortu zuena, erabiltzaileen komunitateak kudea zezan. Wikipedia guztietarako, hau

[9] https://opencitations.net/corpus
[10] https://opencitations.net/index/coci
[11] https://opencitations.net/index/croci

da, hizkuntzetan zehar baliagarri diren datuak iturri bateratu bakar batean eskaintzea da Wikidataren xede historikoa.

6.000 baino gehiago dira Wikidatan lotura esplizituak dituzten kanpoko datu-baseak. Ezaugarri horiek hainbat liburutegi eta biltegi digitalen interesa piztu dute, metadatu bibliografikoak gorde, ondu eta elkartrukatzeko plataforma egokia baita (Tharani, 2021).

2020ko otsailean, artikulu zientifikoak deskribatzen dituzten 22,5 milioi entitate zituen Wikidatak.[12] Haietatik mila bat baino ez dira euskarazkoak momentu honetan, eta mila bat horren gehien-gehiena Inguma lantaldearen ekimenez dago bertan jasota (Astigarraga, Bidaguren *et al.*, 2021). Artikulu zientifikoen metadatuak Wikidatan biltzeak edozeinentzat erabilgarri bihurtzen ditu, eta komunitateak hainbat aplikazio garatu ditu dagoeneko (Farda-Sarbas & Müller-Birn, 2019; Mora-Cantallops *et al.*, 2019). Hala nola Wikidataren grafoa erabiltzen duen Scholia tresnaren bitartez (Nielsen *et al.*, 2017)[13] azter daitezke item bibliografikoak eta haien arteko loturak, bibliometriako hainbat neurketa eginez. Ikusi, esaterako, 1. irudian Scholia bidez UZTARO aldizkariko ko-autoretza grafoa. Zitazio-erlazioak ezarrita dauden heinean, Wikidatan ez ezik, OpenCitations grafoan[14] ere agertuko dira.



1. irudia. Nork idazten duen norekin UZTARO aldizkarian (datuen iturria: Wikidata)

[12] Ikus https://www.wikidata.org/wiki/Wikidata:Statistics
[13] Ikus https://scholia.toolforge.org/
[14] Ikus http://opencitations.net/

4. Euskarazko produkzio akademikoa: Inguma

Orain artean, produkzio akademiko orokorrari buruz aritu gara. Baina euskarazko ekoizpenari dagokionez, euskaraz ekoitzitakoa biltegi orokor ezberdinetan barreiatuta agertzen da, eta horrek zaildu egiten du lan horiek eskuratzea eta euskarazko produkzio akademikoaren ikuspegi orokorra izatea. Hala ere, bada plataforma bat euskarazko produkzio akademiko osoaren datuak jasotzea helburu duena: Inguma.

4.1. Inguma

Udako Euskal Unibertsitateak sortu zuen 2000. urtean. Egun, 13.027 egile, 216 erakunde (130 aldizkari barne) eta 46.297 item bibliografiko daude Inguman bilduta. SQL datu-base batean honako datu hauek daude bilduta eta erabiltzaile-interfaze grafiko batean bila eta ikus daitezke: argitalpenen metadatuak, alde batetik, eta egileen, erakundeen eta aldizkarien metadatuak, bestetik.

Inguma eskuz eguneratzeko lan-fluxua aspaldi ezarrita dago, eta UEUko bi lankide arduratzen dira metadatu berriak jaso eta ontzeaz. Egileak eta erakundeak eskuz desanbiguatu eta zientzia-eremuen adierazleak eskuz ezartzen dituzte. SQL tankerako datu-base erlazionalean gordetzen dituzte datuak; datu-base hori *inguma.eus* web-atariko edukiak sortzeko erabiltzen da.

4.2. Wikibase

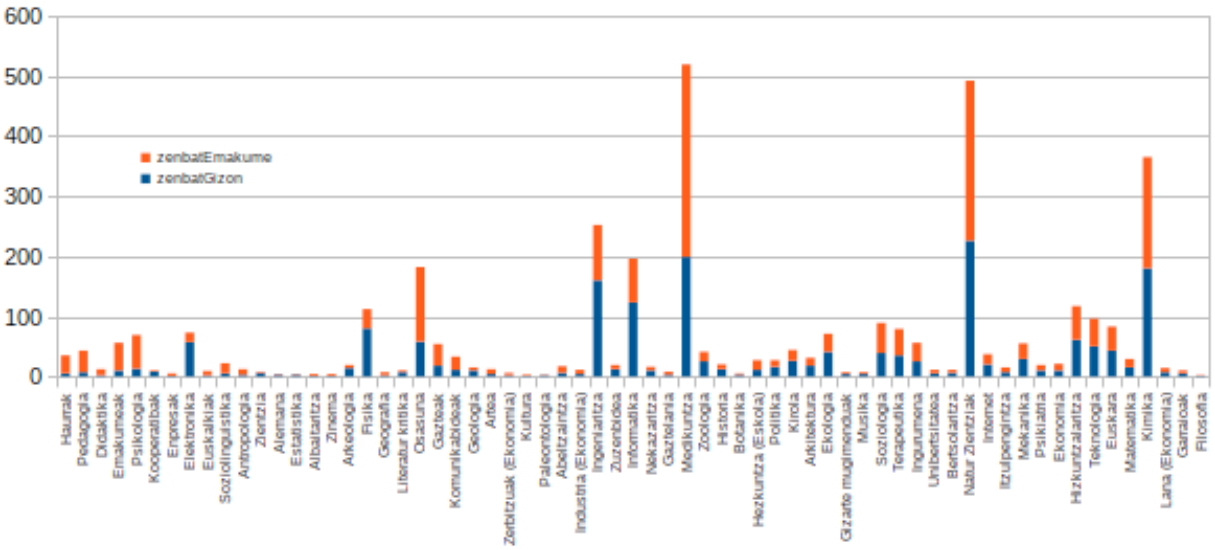
Wikidatak datuak bildu eta eskaintzeko, Wikibase izeneko kode ireki eta lizentzia libreko softwarea du azpiegitura. Wikidataren antolaketa bera duen datu-basea hutsetik sor eta kudea daiteke Wikibase-instantzia propio batean.

Ezagutza-grafoak sortzeko gainerako tresnekin alderatuz gero, Wikibase softwarea erabiltzearen abantaila nagusia berarekin dakartzan zerbitzu edo tresna gehigarriak erabili ahal izatean datza: SPARQL bidezko ohiko datuen bistaratzeak ez ezik, testu-bidezko bilaketak egiteko eta datuak bistaratzeko eta editatzeko interfaze erraza ere badu, adituak ez diren erabiltzaileek ere erabiltzeko modukoa.

Wikibase datu-gordailu batean, Wikidatako ereduarekin bateragarri, baina Wikidata nagusitik at, presta daitezke Datu Lotuak. Wikidata plataformak ere edonori ematen dio bide datuak bertan sartu eta moldatzeko, baina tarteko Wikibasea izateak abantaila ugari dauzka berekin, Lindemann *et al.* (2023) lanean zehazten direnak.

Inguma datu-baseko datuetan oinarrituta, euskal komunitate zientifikoaren azterketa sakona egin daiteke. Badira SQL tresnak Ingumaren datuak bere horretan analizatzeko, baina Ingumako datu horiek ezagutza-grafo ireki bihurtuz gero, analizatzeko beste tresna interesgarri batzuk ere eskura izango genituzke. Ikusi, esaterako, IKERGAZTE kongresuko datuen gainean egindako bistaratzeak (2. eta 3. irudiak).

2. irudia. IkerGazte I-IV: Jakintza-arloen banaketa



3. irudia. IkerGazte I-IV: Egileen genero desoreka, arloka

Euskal komunitate zientifikoaren datuak ezagutza-grafo ireki batera igortzea epe ertaineko helburu lorgarria da: 2022an, Ingumako artikulu zientifiko guztiak esportatu genituen Wikibase instantziara (Lindemann *et al.,* 2023), emaitza online eskuragarri jarriz[15]; ZITERAUZI egitasmo honekin urrats berri bat eman nahi dugu bide horretan.

5. ZITERAUZI

Azken atal honetan, une honetan garatze-lanetan gabiltzan ZITERAUZI tresna-katea aurkeztuko dugu: artikulu akademikoetako PDF fitxategietatik erreferentzia bibliografikoak erauzteko, tratatzeko eta konbinatzeko tresna-katea, hain zuzen ere. ZITERAUZIren epe motzeko helburuak dira IKERGAZTEko argitalpenetatik aipuak atera eta parekatzea, eta ikertzaileen eskura zitazio-datu gehiago jartzea. Ondoren, ZITERAUZI tresna-kateko urrats bakoitza eta lanak egiteko erabiltzen diren kode irekiko tresnak azaltzen dira.

IKERGAZTE kongresuko argitalpenetan oinarritzen den arren, ZITERAUZI edozein esparrutarako erabil daitekeen tresna-kate generikoa da. Tresna-katea eta balidazio-prozesu osoa ikerketa-komunitatearen eskura jartzea da gure asmoa.

ZITERAUZI tresna-katean, erreferentziak lau urratsetan ateratzen dira:

1. PDF artxiboetako testu-geruza egituratzea.
2. Erreferentzia-erlazioak identifikatzea eta formatu egituratu batean adieraztea, autorea, izenburua, etab. etiketatuz.
3. Erreferentziak datu-base bibliografikoekin erkatzea.
4. Zitazio-erlazioak argitalpen-metadatuei gehituta argitaratzea.

Hainbat tresna garatu dira PDF fitxategietako zitazioak erauzi eta egituratzeko, erabilera-modu eta arrakasta-maila desberdinekin (Cioffi & Peroni, 2022). Aurrerago beste tresna batzuekin proba egitea baztertu gabe, lan honetan Grobid (GeneRation Of Bibliographic Data) kode irekiko Java liburutegia erabiliko dugu, PDF fitxategietako informazio guztia ondoren automatikoki prozesatzeko aukera ematen duen formatu egituratu batera esportatzeko. Grobidek ikasketa automatikoa erabiltzen du PDF fitxategia egituratzeko. Bereziki egokia da argitalpen zientifikoetarako, eta ausazko eremu kondizionaletan (CRF) oinarritutako aurrez entrenatutako zenbait eredu ditu (Lopez, 2009; Romary & Lopez, 2015).

Prozesatu osteko formatua, TEI-XML[16], testuetan ohar zehatzak egiteko aukera ematen du irakurgarritasuna galdu gabe. Horrekin, artikulu guztien edukia modu egituratuan ateratzeaz gain, haien bibliografietako informazioa ere lor dezakegu. Bibliografia <listBibl> elementuan agertuko da eta zitazioak, berriz, <biblStruct> elementu barruan.

Adibide gisa, hona hemen 2023ko Ikergazte kongresuko *Gipuzkoa erdialdeko testu zaharrak erkatzen* lanetik erauzitako erreferentzia bat:

Werner, V. (2021), Text-linguistic analysis of performed language: revisiting and re-modeling Koch and Oesterreicher, *Linguistics*, 59(3), 541-575. https://doi.org/10.1515/ling-2021-0036

GROBIDek prozesatu ostean, ondoko irudiak erakusten du zitazioa \<biblStruct> elementuan nola jasotzen den:

---

[15] Ikus https://wikibase.inguma.eus

[16] Ikus https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

```
                    <biblStruct xml:id="b21">
                        <analytic>
                            <title level="a" type="main">Text-linguistic analysis of performed language: revisiting an
d re-modeling Koch and Oesterreicher</title>
                            <author>
                                <persName>
                                    <forename type="first">V</forename>
                                    <surname>Werner</surname>
                                </persName>
                            </author>
                            <idno type="DOI">10.1515/ling-2021-0036</idno>
                            <ptr target="https://doi.org/10.1515/ling-2021-0036" />
                        </analytic>
                        <monogr>
                            <title level="j">Linguistics</title>
                            <imprint>
                                <biblScope unit="volume">59</biblScope>
                                <biblScope unit="issue">3</biblScope>
                                <biblScope unit="page" from="541" to="575" />
                                <date type="published" when="2021">2021</date>
                            </imprint>
                        </monogr>
                    </biblStruct>
                    <biblStruct xml:id="b22">
```

4. irudia. GROBID tresnak PDFa prozesatu ostean, <biblStruct> elementuan jasotzen ditu erreferentziak

<listBibl> elementu bakoitzaren edukia, hau da, automatikoki egituratutako bibliografia-sarrera bakoitza, nazioarteko datu-baseetan agertzen diren argitalpen-metadatuetan konparatu egiten du LOC-DB softwareak, Mannheimeko Unibertsitate Liburutegian garatu zutena, edozein liburutegik zitazio-erlazioen grafo libreari (goian aipatutako OpenCitations grafoari) ekarpena egin diezaion (Lauscher *et al.*, 2018). Interfaze grafiko batean antzekotasun handieneko aurkikuntzak proposatzen zaizkio erabiltzaileari, dagokion argitalpena aukeratu ahal izateko. Argitalpena aurkituz gero, zitazio-erlazioa definiturik geratzen da, hau da, lantzen ari den argitalpenaren bibliografia-sarrera aurkitutakoaren identifikatzaile unibokoarekin (DOI) osatzen da. Aurkitutakoak DOI-rik izan ezean, barneko identifikatzaile bat esleitzen da. LOC-DB software librearen garatzaileekin harremanetan gaude (ikus Lindemann *et al.*, 2019), eta haiekin elkarlanean egokituko dugu softwarea, Inguma datu-basea ere erabil dezan zitazioei dagozkien argitalpenak aurkitzeko, eta Inguma Wikibase-ko identifikatzailea erabiltzeko.

Aipatzen den argitalpena aurkitzen ez bada, ez nazioarteko datu-baseetan eta ez Inguman, argitalpenari sarrera emango diogu Inguma Wikibase-n, eta horrela barneko identifikatzaile unibokoa esleitzen zaio. Euskarazko argitalpen askok ez dute DOI identifikatzailerik eta horregatik nazioarteko datu-baseetan ez zaizkigu kasu ugaritan agertuko. Proiektu honek, beraz, gure bildumetako argitalpenen ikusgarritasuna ez ezik, aipatzen diren lanen ikusgarritasuna ere hobetuko du; ikerketaren ebaluazio osatuago bati ekarpena egingo diogu horrela. Gainera, Inguma Wikibase-ko edukiak Wikidata plataformara pasatzeko asmoa dugu, UEU-ren beraren argitalpenekin egin genuen moduan (Lindemann *et al.*, 2023); horrela, nazioarteko identifikatzaile librea lotuko zaio argitalpen bakoitzari, Wikidata-ko datuen gainean eraikitako zitazio-grafoetan ere agertzeko, Scholia tresnaren barruan, esaterako.

6. Emaitzak

Lan honen emaitzak erabilera publikoa izango du eta zitazio-grafoa eta honen gaineko azterketak web bidez kontsultatu eta erabili ahal izango dira. Horrez gain, nazioarteko datu-baseetako

erreferentziekin egingo den erkaketak euskarazko ekoizpen akademiko-zientifikoaren ikusgarritasuna handitzen lagunduko du. Barrura begira, euskal eremuan egiten diren ikerketak hobeto ezagutzea ekarriko du.

Humanitate Digitalen diziplina *open access*-arekin eta Datu Lotuen paradigmarekin guztiz bat dator. Eduki, kode, formatu eta ikerketa-emaitzen zirkulazio librearekin eta berauek elkarlanean aberasteko aukerarekin. Hori horrela, esan dezakegu proposamen honek euskarazko ezagutza librea sustatzen duela eta ondoren etor daitezkeen bilduma berrien azterketarako azpiegitura prest utziko duela.

Eskerrak eta oharrak

Lan honek Humanitate Digitalak Ikertzeko Gunearen laguntza jaso du, Errenteriako udalak Udako Euskal Unibertsitatearen lankidetzarekin sortutako gunea (Torrekua eraikina).

Bibliografia

Astigarraga, A., Bidaguren, M., Delgado, E., Gonzalez, G., & Sarasola, K. (2021/, azaroa 26). *Cooperation between INGUMA and Wikidata databases. Exploring and analizing the production of Basque academic journals*. Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen. https://ixa2.si.ehu.eus/clariah-eus/eu/node/28

Astigarraga, A., Iñurrieta, U., Irasuegi, B., Landabidea, X., Gonzalez, G., & Sarasola, K. (2021/). *Torrekua HD Gunea: Humanitate digitalak ikergunea Errenterian. Ekonomia soziala, kultura, hezkuntza eta euskara | CLARIAH EUS*. http://www.clariah.eus/eu/node/20

Berners-Lee, T., Hendler, J., & Lassila, O. (2023/). The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. In O. Seneviratne & J. Hendler (arg.), *Linking the World's Information* (1. arg., or. 91-103). ACM. https://doi.org/10.1145/3591366.3591376

Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021/). Data citation and the citation graph. *Quantitative Science Studies*, *2*(4), 1399-1422. https://doi.org/10.1162/qss_a_00166

Cioffi, A., & Peroni, S. (2022/). Structured References from PDF Articles: Assessing the Tools for Bibliographic Reference Extraction and Parsing. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, & A. Poggi (Arg.), *Linking Theory and Practice of Digital Libraries* (Libk. 13541, or. 425-432). Springer International Publishing. https://doi.org/10.1007/978-3-031-16802-4_42

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014/). The G oogle scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446-454. https://doi.org/10.1002/asi.23056

Fricke, S. (2018/). Semantic Scholar. *Journal of the Medical Library Association*, *106*(1). https://doi.org/10.5195/jmla.2018.280

Gusenbauer, M. (2019/). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*(1), 177-214. https://doi.org/10.1007/s11192-018-2958-5

Hammond, T., Pasin, M., & Theodoridis, E. (2017/). *Data integration and disintegration: Managing springer nature scigraph with shacl and owl*. International Workshop on the Semantic Web. https://www.semanticscholar.org/paper/Data-integration-and-disintegration%3A-Managing-with-Hammond-Pasin/1eb9d54fac20963a6050db178865579ff4564833

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020/). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414-427. https://doi.org/10.1162/qss_a_00022

Khabsa, M., & Giles, C. L. (2014/). The Number of Scholarly Documents on the Public Web. *PLOS ONE*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Lauscher, A., Eckert, K., Galke, L., Scherp, A., Rizvi, S. T. R., Ahmed, S., Dengel, A., Zumstein, P., & Klein, A. (2018/). Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 109-118. https://doi.org/10.1145/3197026.3197050

Lindemann, D., Astigarraga, A., Bidaguren, M., Delgado, E., Gonzalez, G., & Sarasola, K. (2023/). Inguma eta Wikidata uztartuz, euskarazko zientziaren ezagutza-graforantz. In D. Lindemann (Arg.), *Miren Azkarateri esker onez* (or. 193-210). UPV/EHU Argitalpen Zerbitzua.

Lopez, P. (2009/). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Arg.), *Research and Advanced Technology for Digital Libraries* (or. 473-474). Springer. https://doi.org/10.1007/978-3-642-04346-8_62

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021/). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906. https://doi.org/10.1007/s11192-020-03690-4

Nasar, Z., Jaffry, S. W. eta Malik, M. K. (2018). Information extraction from scientific articles: A survey. Scientometrics, 117(3), 1931-1990. or. https://doi.org/10.1007/s11192-018-2921-5

Ortega, J. L. (2021/). El movimiento Open Citations y sus implicaciones en la transformación de la evaluación científica. *Arbor*, *197*(799), a592. https://doi.org/10.3989/arbor.2021.799007

Peroni, S., & Shotton, D. (2020/). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428-444. https://doi.org/10.1162/qss_a_00023

Priem, J., Piwowar, H., & Orr, R. (2022/). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. https://doi.org/10.48550/ARXIV.2205.01833

Romary, L., & Lopez, P. (2015/). Grobid—Information extraction from scientific publications. *ERCIM News*, *100*. https://inria.hal.science/hal-01673305

Singh, V. K., Srichandan, S. S., & Lathabai, H. H. (2022/). ResearchGate and Google Scholar: How much do they differ in publications, citations and different metrics and why? *Scientometrics*, *127*(3), 1515-1542. https://doi.org/10.1007/s11192-022-04264-2

Stock, W. G., & Stock, M. (2013/). *Handbook of Information Science*. De Gruyter Saur. https://www.degruyter.com/document/doi/10.1515/9783110235005/html

Tharani, K. (2021/). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, *47*(2), 102326. https://doi.org/10.1016/j.acalib.2021.102326

Thelwall, M., & Kousha, K. (2017/). ResearchGate versus Google Scholar: Which finds more early citations? *Scientometrics*, *112*(2), 1125-1131. https://doi.org/10.1007/s11192-017-2400-4

Yan, X.-S. (2011/). Information Science: Its Past, Present and Future. *Information*, *2*(3), 510-527. https://doi.org/10.3390/info2030510

Życzkowski, K. (2010/). Citation graph, weighted impact factors and performance indices. *Scientometrics*, *85*(1), 301-315. https://doi.org/10.1007/s11192-010-0208-6

Testu historikoak wiki-plataformetan, Datu Lotu gisa

Historical Texts in Wiki-platforms as Linked Data

David Lindemann, Mikel Alonso

Euskal Herriko Unibertsitatea (UPV/EHU)
david.lindemann@ehu.eus mikel.alonso@ehu.eus

Laburpena

Artikulu honetan proposatzen dugu euskarazko testu historikoak Datu Lotuen paradigmaren arabera datu-base batean jasotzeko eredua, Wikibase softwarearen instantzia bat azpiegitura moduan erabilita. Alde batetik, modelatzen ditugu software horrek berez dakarren oinarrizko datu-ereduaren gainean, tokenak eta token-multzoak deskribatzen dituzten entitateak. Bestetik, Ontolex-Lemon hiztegia erabiliz, Euskara Batuko hiztegia eraiki dugu Wikibase instantzia berean. Token edo token-multzoak hiztegi horren sarrerekin lotzeko honako aukerak ematen ditu gure ereduak: hiztegi-sarrera (lema) mailan, hiztegi-adiera mailan, eta forma flexionatuaren mailan. Halaber, aukera dago hiru maila horietako entitateek deskribapen eta lotura gehiago izateko. Horrek anotazio morfosintaktiko klasikoaren eredua anotazio literalez haratago hedatzen du, gure eredua hainbat anotazio entitate gisa lotzean datzalako. Corpus-hiztegi interfazeaz gain, ohiko corpus-anotazioak errepresentatzeko aukera ere badago gure ereduan, hala nola anotazio filologikoak eta anotazio semantikoak.

Gako hitzak: testu historikoen digitalizazioa, Datu Lotuak, Wikibase.

Abstract

In this article, we propose a data model for storing Basque historical texts in a database following the Linked Data paradigm, using an instance of the Wikibase software as infrastructure. On the one hand, we model entities describing corpus tokens and token spans on top of the pre-set Wikibase data model. On the other hand, we build a Standard Basque dictionary deploying the Ontolex-Lemon vocabulary, on the same Wikibase instance. Our model allows for linking tokens and token spans to entries of that dictionary, on lexical entry (lemma) level, lexical sense level, and inflected form level, while entities of these three levels may carry additional descriptions and further links. That extends a traditional way of morphosyntactic annotation with literal values towards linking dictionary elements as entities. In addition to this corpus-lexicon interface, other kinds of annotations such as philological and semantic annotations are also modeled.

Keywords: historical text digitization, Linked Data, Wikibase.

1. Sarrera

Euskarazko testu historikoen digitalizazioari dagokionez, ahalegin anitz ikusi ditugu azken urteotan; helburu eta metodologia ezberdinak darabiltzaten hainbat ekimen. Proiektu horien emaitza berrienen artean, EHUko Euskara Institutuak kudeatzen duen Corpus Historikoa osatzen duten testu digitalak ditugu, sareko interfaze batean esplora daitezkeenak; IXA taldean garatutako SAHCOBA (Estarrona *et al.,* 2022), anotazio morfosintaktikoez aberastua eta bilaketa finduak interfaze batetik eskaintzen dituena; eta, bestetik, anotazio filologikoak interfaze grafiko bidez eskaintzen dituzten testu-edizio digitalak, Lazarragaren eskuizkribuarena (Bilbao *et al.,* 2011) adibide.

Bestalde, euskarazko datu lexikografiko historikoak eta estandarrak Datu Lotu (Linked Data)[1] gisa errepresentatu eta elkarrekin lotzeko esperimentuak aurkeztu ditugu (Lindemann & San Vicente, 2020; Alonso & Lindemann, 2022); lan horietan, Wikiteka (euskarazko Wikisource) eta Wikidata plataformetan integratu ditugu erabilitako datu-multzoak. Dagoeneko, eskuz edo programa bidez, azter daitezke iturri historikoko edukiek egungo ezagutza-grafo librean dituzten loturak. Manuel Larramendiren euskarazko testuen eta egile beraren *Hiztegi Hirukoitza*ren argitalpen digitalak dira epe ertaineko helburua. Horretarako, jatorrizko eskuizkribu edo lehenengo argitalpen inprimatura jo eta edizio digitala eraiki nahi dugu.

Goian aipatutako proiektuetako helburu zehatz ezberdinen araberako metodologiak elkartu nahi ditugu egitasmo honetan, hau da, honako osagai hauek bateratu nahi ditugu elkarrekin lotutako datu-multzoan:

— Corpuseko tokenaren agertokia faksimile digitalean (Wikiteka plataforma).
— Faksimileko transkribapena, wikitext formatuan (Wikiteka platafoma).
— Corpuseko tokenaren inguruko anotazio morfosintaktikoak (Wikibase).
— Corpuseko tokenaren inguruko anotazio filologikoak (Wikibase).
— Tokenari lotzen zaion hiztegi-lema estandarra, haren adiera, eta haren forma flexionatua (Wikibase).
— Hiztegi-lema estandarrak beste hainbat baliabidetan duen deskribapena (Wikidata).
— Corpuseko tokenak entitate izendun bati egiten dion erreferentzia (Wikibase, Wikidatako entitateak erabilita).

Datu Lotu Irekien irizpideak aintzat hartu eta Wikimediak eskaintzen dituen plataformetan gorde, editatu eta argitaratuko ditugu datuak. Wikiteka plataforman jatorrizko argitalpenaren edo eskuizkribuaren faksimilea eta haren transkripzioa gordetzen ditugu. Wikibase instantzia batean, MLV Wikibasean,[2] transkripzioko tokenak jasotzen ditugu, bakoitzak faksimileko agertokira lotura eginez. Tokenak deskribatzeko eredua proposatzen dugu, Linguistic Linked Open Data[3] arloko estandarretan oinarrituta. Testu-tokenak deskribatzen dituzten entitateei anotazioak gehitzeko, NIF ontologia hartzen dugu oinarri (Hellmann *et al.,* 2013), eta hiztegi-sarrerak errepresentatzeko, berriz, Ontolex-Lemon (McCrae *et al.,* 2017). Horrela, testu-tokenak lexema mailan, adiera mailan nahiz forma mailan deskribatzeko moduan gaude, Orotariko Euskal Hiztegian[4],

[1] Ikus https://eu.wikipedia.org/wiki/Datu_estekatuak
[2] Ikus https://monumenta.wikibase.cloud
[3] Ikus https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data
[4] Ikus https://www.euskaltzaindia.eus/component/oeh

Egungo Testuen Corpusean[5], Elhuyar hiztegian[6], eta potentzialki beste hainbat baliabidetan dituzten deskribapenetara loturak ezarriz. Horretaz gain, token-multzo bati anotazioak gehi diezazkiokegu baldin eta anotazio hauek batera dagozkionean: entitate izendun baten erreferentzia, adituen anotazio filologikoa edo hiztegi-lerrokaketa.

Datu-eredua proposatzearekin batera, lehenengo testu bat hartu dugu adibide, eta haren edukiak ereduaren arabera modelatzen hasi: Larramendiren Azkoitiko Sermoia, 1737. urtean sorturiko eskuizkribua.

Azkenik, Wikimediako plataformek berez eskaintzen dituzten eta script propioek ematen dituzten zenbait funtzio erakusten ditugu, datuak editatu eta argitalpenerako berrerabiltzeko lan-fluxua irdudikatzeko. Erakutsiko dugunez, aipatutako datu linguistikoak wiki-plataformetan gordetzeak, editatzeak eta argitaratzeak abantaila nabarmenak dakartza.

2. Eskuizkribuaren transkripzioa Wikiteka plataforman

Aipatu bezala, Larramendiren Azkoitiko Sermoia hartu dugu adibide moduan, hainbat arrazoi direla eta. Batetik, eskuizkribuaren irudiak ditugu, eta horri esker, Wikiteka[7] aintzat hartzen duen lan-fluxuarekin esperimentatzeko aukera dugu, irudiak plataforma hartara igoz. Bestetik, Lakarrak (1985) egindako testu horren transkripzioa dugu, eta bertako anotazio filologikoak gure datu-ereduaren arabera jasotzea da egitasmoaren helburuetako bat. Gainera, 1990. urtean testu hori berriz argitaratu zen (Larramendi, 1990) eta argitalpen horretako oharrak ere gehitu daitezke.

Eskuizkribuaren irudiak Wikitekara igo ondoren, plataformak transkripzioa egiteko aukera ematen du, transkripzioa eta irudia elkarren ondoan izanda, 1 irudian ikusten den bezala. Transkripzioa hutsetik hasi beharrean, Lakarrak (1985) egindako transkripzioa hartu dugu oinarri, *ASJU* aldizkariaren webguneko pdf dokumentutik hartutakoa, eskuizkribuarekin konparatzeko. Abiapuntu izan dugun *ASJU*ko testuko PDFaren testu-geruzak akatsak zituenez, TXT akastun hori zuzendu behar izan dugu, eskuizkribuan agertzen diren testu-hitzekin (tokenekin) erkatzen hasi aurretik. Transkripzioaren bertsio erabat fidel, *ground truth* hori ekoitzi eta kontserbatzeak berebiziko balioa du, eta haren gordeleku egokia Wikiteka plataforma da. Bertsio horrek interes filologikoa du, bertsioak egiteko orduan edozein esku (edo hanka?) sartzearen gardentasuna manten daitekeelako horrela. Horretaz gain, inongo egokitzapenik ez duen bertsio hori da *hand-written text recognition* (HTR) tresna batek ikasteko behar duena, Larramendiren idazkera, kasu honetan. Helburu horrekin, eskuizkribuan agertzen diren hizki-segidak bere horretan islatzen ditugu bertsio honetan, testu-hitzen hurrenkera ere errespetatuz. Irizpide horrek salbuespen bakarra baimentzera eraman gaitu: Larramendik bigarren idatzaldian, «gero» erantsitako pasarteak, eskuizkribuan alboan agertzen direnak, letra lodiz idazten ditugu Wikitekako testuan; horrela, testu lodia kenduko genuke HTR entrenamenduetarako (eta irudian alboetan agertzen dena ere bai).

---

[5] Ikus https://www.ehu.eus/etc/

[6] Ikus https://hiztegiak.elhuyar.eus/

[7] Ikus https://eu.wikisource.org/wiki/Aurkibide:Larramendi_1737_Azkoitiko_Sermoia.pdf

1. irudia. Azkoitiko sermoia, Wikiteka plataformaren interfazean

Transkripzioa modu kolaboratiboan editatzeko, Wikitekak funtzio egokiak badakartza berez, baita testu osoak edo pasarteak gunera bidali edo hortik jasotzeko ere. Paragrafoen artean lerro-jauzi bikoitza txertatzen dugu; bestela ez dugu lerro-jauzirik sartzen. Egia da HTR helburuetarako paragrafo barruko lerro-jauziei begira ere zorrotzak izatea komeni dela, baina kasu honetan paragrafoaren edukia lerro-jauzirik jaso gabe jasotzearekin konformatu gara, momentu honetan Larramendirekin HTR lanik aurreikusten ez dugulako. HTR lana egin nahi izatekotan, eskuizkribuko lerro-jauziak bere horretan islatuko beharko lirateke. Beraz, bi lerrotan dauden hitzak transkripzioan ez ditugu zatitu (adibidez, *Justi-ciac*, *go-gaitu* edo *esatendi-gu*, eskuizkribuaren lehen orrialdean).

Wikitekako transkripzioa amaituta, paragrafo mailan lotura-aingurak jartzen ditugu. Hau da, lerro-jauzi bikoitz bakoitzaren ondoren, kode bat gehitzen diogu wikitext dokumentuan,[8] kanpotik paragrafo bakoitzaren hasierara web-lotura iraunkorra emango diguna. Paragrafoak zenbatu egiten ditugu horretarako, eta paragrafo-zenbakia loturaren parte izango da.

Tokenizazioa da hurrengo pausoa, hau da, testu-hitzak isolatzea. Horretarako erregelak idatzi ordez, *nltk toolkit*[9] moduluek dakarten tokenizatzailea erabiltzen dugu.

[8] *Wikitext* Wikimediako plataformetan erabiltzen den testu-formatua da. TXT (testu hutsa) aberastua da wikitext markdown formatuaren antzekoa; ikus https://en.wikipedia.org/wiki/Help:Wikitext

[9] Ikus https://www.nltk.org/

3. Corpus-tokenen errepresentazioa eta anotazioa Wikibase plataforman

Wikidatak[10], gaur egungo datu ontologiko nahiz lexikoak bateratzen dituen ezagutza-grafo librerik handienak, datuak bildu eta eskaintzeko, Wikibase[11] izeneko kode ireki eta lizentzia libreko softwarea du azpiegitura. Wikimedia Germany (WMDE) erakundeak garatu du softwarea, eta edonoren esku jarri. Hau da, edonork izan dezake orain «Wikidata huts bat» eduki propioez osatzeko, eta, nahi izanez gero, Wikibase instantzia hori Wikidata nagusiarekin edo Datu Lotuen paradigmari jarraitzen dion beste edozein datu-baserekin federatzeko. Hemen aurkezten ditugun eredu eta esperimentuak *Monumenta Linguae Vasconum* (MLV) Wikibase izendatu dugun instantzian gauzatzen ditugu. Instantzia horrek WMDEk eskainitako Wikibase Cloud hosting zerbitzuan[12] du ostatua. Datuak atzitzeko moduan edonor dago, interfaze grafikoaren edo programaren bitartez. Datuak editatzeko, aldiz, erabiltzaile onartuak soilik daude baimenduta.

3.1. Softwarearen ezaugarriak

Wikibase softwarea erabiltzearen abantaila nagusia (ikus Lindemann *et al.*, 2023), berarekin dakartzan zerbitzu edo tresna gehigarriak erabili ahal izatean datza: SPARQL bidezko datuen bistaratzeak ez ezik, testu-bidezko bilaketak egiteko eta datuak bistaratzeko eta editatzeko interfaze grafikoa ere badu. Lanerako komunitate propioa sortzeko aukera ematen du, eta, Wikibase instantzia bat Mediawiki sistemaren instantzia baten atala denez, Wikipediakoak bezalako testu-orriak ere eduki ditzake.

Datu Lotuen paradigmari jarraituz, deskribatzekoak diren entitateak grafo baten erpinak dira, haien arteko erlazioak eta datu-mota zehatz bateko propietateak grafoaren ertzak izanik. Xehetasunak jarraian azalduko ditugu.

3.2. Corpusaren edukia deskribatzen dituzten entitateak: Tokena eta token-multzoa

Testu-hitz edo token bakoitzari Wikibasean entitate edo erregistro bakar bat egokituko zaio, eta tokena deskribatzen duen entitateari honako adierazpen hauek gehitu zaizkio:

— Klaseko adierazpena (entitatea 'token' klasekoa dela, hots, token bat deskribatzen duela).
— Tokenaren zenbakia hurrenkeran.
— Tokenaren forma (hizki segida eskuizkribuan).
— Tokena agertzen den dokumentua deskribatzen duen entitatea.
— Tokenari dagokion paragrafoaren aingura Wikitekan.
— Ondorengo tokena deskribatzen duen entitatea.

Testu-hitza eskuz edo programa bidez egindako anotazioez hornitzeko, tokena deskribatzen duen Wikibaseko entitateari propietate-balio bikoteak erantsiko zaizkio. Esaterako, lehenengo erabilpen-kasuan, Larramendiren Azkoitiko sermoian, hizkuntzalaritzako irizpideen araberako anotazio hauek gehitu ditugu:

---

[10] Ikus http://www.wikidata.org
[11] Ikus https://wikiba.se
[12] Ikus https://wikibase.cloud

— Anotazio filologikoak: «gero erantsia», «lerro alboan»,«beste zerbaiten ordez», «zerbait borratu eta gainean idatzia», «tinta mantxa edo tatxatua», «hasiera ahula, ez da argi irakurtzen», eta abar.
— Forma historikoa, grafia egokituta edo eguneratuta.
— Formari hiztegi morfologikoan dagokion sarrera.
— Formari dagokion hiztegi-lema.
— Hitzari dagokion hiztegi-adiera.
— Kategoria gramatikala testuinguruan.
— Hitzak erreferentzia egiten dion entitate izenduna (pertsona, tokia, erakundea…).

Ikuspuntu tekniko batetik, aldiz, anotazioen beste sailkapen bat egin daiteke. Anotazio-mota horiek guztiak bi eratako propietateen bitartez jasoko dira datu-basean: batetik, anotazioaren edukia literalki jasotzen den hizki kate gisa eta, bestetik, anotazioaren edukia ezagutza-grafoko entitate baten bitartez. Ahal dugun guztietan *balio literal* gisa ditugun anotazioak entitate gisa jasoko ditugu, edo entitate batekin lotu. Lemaren ikurra baino hobea da dagokion hiztegi-sarreraren identifikatzailea; kategoria gramatikala adierazten duen testuzko kodea baino hobea, bokabulario baten sarrera; hiri baten izena baino hobea, hiri horrek Wikidatan duen entitate deskribatzailea. Elkarrekin lotutako entitateen eredua hobea da, elkarrekin zerikusia duten bi arrazoi nagusigatik: balio literalen desanbiguazioa (Durango Bizkaian vs. Mexikon) edo batuketa (Koldo Mitxelena *vs.* Luis Michelena) antolatzeko bidea delako; entitateak ezagutza-grafoan dituen harremanak ustiatzeko bidea delako, hau da, aurrean dugun adierazpena datu-base galdeketa konplexuetan erabilgarri izanez gero, balio literalekin ezinezko diren hainbat neurketa eta datu-ikuspegi lor ditzakegulako. Horregatik, anotazio semantikoa Wikidatako grafoa berrerabiliz egitea proposatzen dugu; entitate izendunak deskribatzen dituzten Wikidata entitateak tokenarekin lotuz, haien deskribapen eta loturak federazioaren bitartez gure Wikibase instantzian ere erabilgarri izango dira.[13]

Token-multzo batek errepresentatzen duen entitateak bere baitan hartzen dituen tokenen identifikatzaileak eta haien hurrenkera ditu erantsiak. Tokenari eransten zaizkion anotazioak token multzo bati modu berean eransten zaizkio, anotazioaren besarkadura multzoarekin bat etorriz, hau da, token multzoak talde gisa anotazioa daramala. Goian zerrendatutako anotazio mota guztiak izan daitezke, eta multzoa izanda, baita beste bat ere: aipua, testu-berrerabilera deskribatzen duen propietatea. Adibidez, Larramendiren sermoian, aipatutako Biblia pasarteak; eta kasu honetan ere, pasartearen jatorria erreferentzia bibliografiko literal batez edota item bibliografiko hori deskribatzen duen entitateaz adieraz daiteke, bigarren aukera hobea izanik.

Edozein anotazio, balio literal edo entitate motakoa izanda, gehiago deskriba daiteke Wikibasean, propietate *kualifikatzaileak*[14] erabiliz. Horrela, anotazio filologikoaren egilea eta haren jatorrizko argitalpena jasotzen dugu, egokitzapen grafikoa burutu duen arau-sortaren bertsioa, entitate izendunak anotatzen duen tresnaren izen eta saioa, eta antzekoak, hots, adierazpenaren iturriak. Bigarren mailako anotazio hauek ere bi eratakoak izan daitezke: balio literalak edo lotutako entitateak.

[13] SPARQL galdeketa *federatuen* bitartez, Wikibase instantzia ezberdinak aldi berean atzitzen dira, hau da, adibidez, entitate baten inguruan Wikidatan egiten diren adierazpenak gurean izango balira bezala erabil daitezke (esaterako, Durango Bizkaiko hiri bat dela, eta abar).

[14] Wikibasen, kualifikatzaileak bigarren mailako adierazpenak dira (adierazpen bati buruzko adierazpena).

3.3. Euskara batuko lema-bankua Wikibase plataforman

Datu lexikalak errepresentatzeko, Wikibasek berez dakarren oinarrizko datu-eredua Ontolex-Lemon (McCrae *et al.,* 2017) izeneko RDF bokabularioan oinarritzen da, zehazki, hark definitzen dituen oinarrizko hiru ontologia-klaseetan, batetik, eta haiek hirurak elkarrekin lotzen dituen eta klase bakoitzari dagozkion deskribapen linguistikoa daramaten zenbait propietatetan, bestetik. Bokabulario hori W3C konsortzioak gomendatzen duena da, eta *Linguistic Linked Data* eremuan datu lexikalak errepresentatzeko estandarra dela esan daiteke. Hura erabiltzeak beste hainbat baliabideekiko elkarreraginkortasuna dakar. Wikibase batean datu lexikalak jasotzeko oinarrizko hiru klase horiek jarraian deskribatzen ditugunak dira.

3.3.1. ontolex:LexicalEntry

Klase honetako entitateek hiru propietaterentzat dituzte balioak, derrigorrez: (a) lema edo sarrera-buru gisa erabiltzen den hitza edo hitz multzoa, (b) hizkuntza, Euskara Batua, gure kasuan, eta (c), kategoria gramatikala (POS, *part of speech*). Azken hori euskarazko lema-bankua osatzeko ez dugu zehazten, eta kategoria gramatikalaren adierazle gisa *lema bankuko sarrera* esleitzen diogu sarrerari, hau da, lema horri dagozkion formek corpusean zer POS izan dezaketen ez dugu maila honetan desanbiguatzen, nahiz eta Ontolex-Lemon bokabularioak gomendio hori egin, batu nahi ditugun iturriek ere ez dutelako maila honetan desanbiguazio koherenterik eskaintzen. POS balioa, beraz, adiera eta formei gehituko diegu, maila horietan kontraesan edo inkoherentzietan erortzeko arriskua saihesteko moduak baditugulako.

Euskara Batuko lemak Orotariko Euskal Hiztegiko (OEH) sarrerekin lerrokatzeko, adibidez, ez dugu POS desanbiguaziorik ezartzeko modurik, hiztegi horrek sarrera-buruarekin batera berez ez baitu halakorik jasotzen (sarrera barruan baizik); halaber, Euskaltzaindiko webguneko edizio elektronikoa atzitzeko loturek ere kategoriarik gabeko lema-ikurra baino ez dute. Adibidez, *aditu* lemari dagokion sarreran, aditza ez ezik, adjektiboa eta izena ere deskribatzen dira, hiztegiaren makroegituran (atzitzeko bide dugun lategian) hirutasun horren berririk eman gabe. Elhuyar Hiztegietako identifikatzailearekin lerrokatuta dagoen Wikidatan, aldiz, *LexicalEntry* klasekoek badute POS desanbiguazioa, eta horregatik maila honetako lerrokaketan POS desanbiguatzailea adierazten dugu lerrokaketaren *kualifikatzaile* gisa, 2 irudian ikusten den legez.[15]

Orotariko Euskal Hiztegiaren lategiko 86.871 hitz bakarreko sarrera-buruak hartu ditugu oinarri MLV Wikibaseko lema-bankua definitzeko. Lematizaziorako irizpide bera jarraitzen duten heinean, baliabide lexiko gehiago lot dakizkioke lema-banku honi, eta hori da gure asmoa. Wikidatako lexemak batu dizkiogu dagoeneko, eta, horren bitartez, Elhuyar hiztegietako identifikatzailea; esandakoaren harira, azken baliabide horietako sarrera bat baino gehiago egon daiteke lema-bankuko sarrera berarekin lotuta, Wikidatak hiztegi- sarrerak honela modelatzen dituelako, eta POS ezberdin nahiz POS berdineko homografoak sarrera banatan jasotzen dituelako. Aipatu beharra dago izenki ezberdinak Elhuyarrekoetan ere sarrera berean tratatzen direla, hots, adjektibo eta izenari aparteko sarrerarik esleitu gabe.[16]

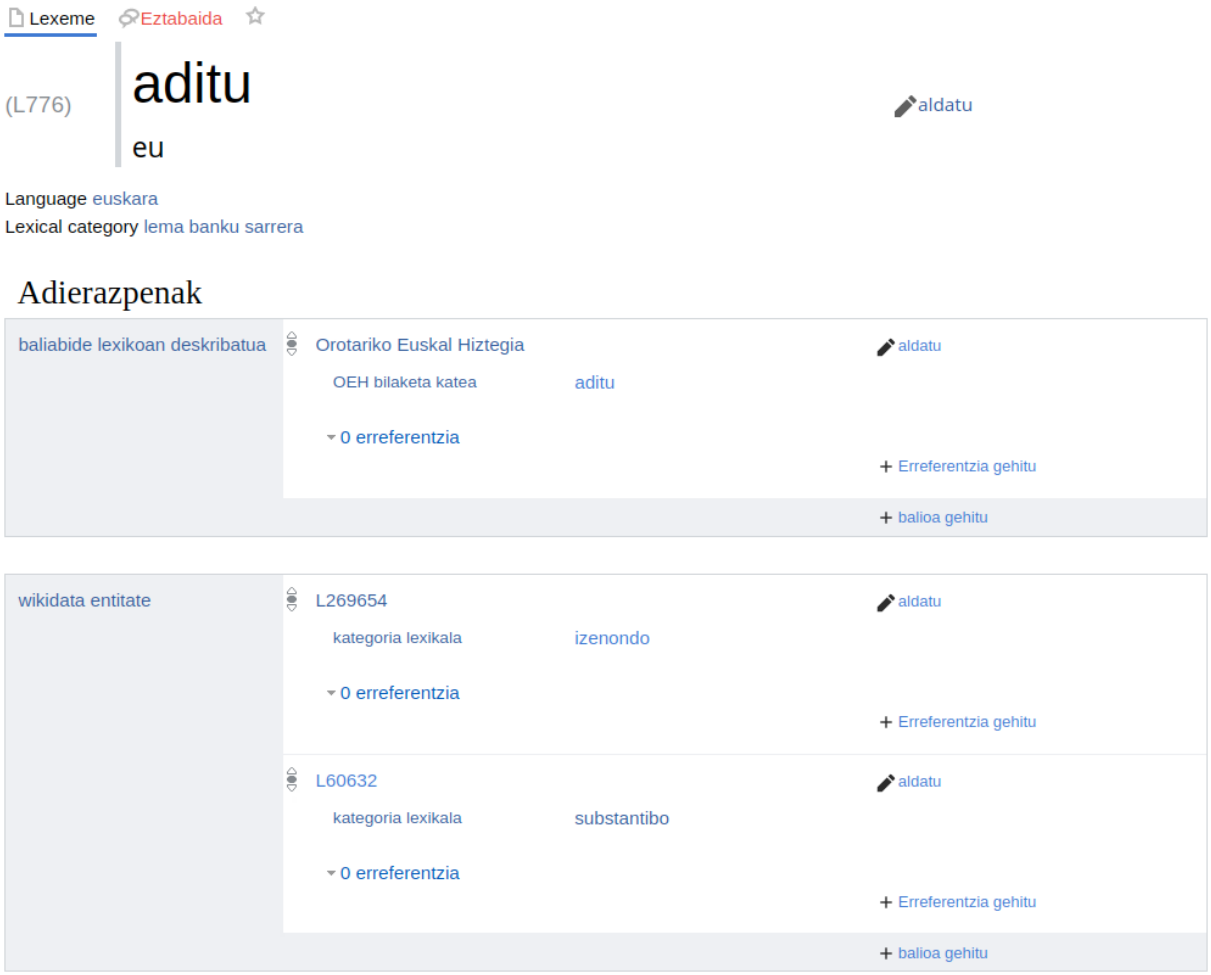
Kanpoko baliabide lexikoetara daramaten loturak adierazteaz gain, klase honi lotzeko beste edozein propietate definitu daiteke Wikibase batean, erlazio lexikalak adieraztekoak, esaterako.

---

[15] Artikulua idazteko unean, Wikidatak ez du *aditu* aditza jasotzen; aditzen estaldura murritza da oraindik Wikidatan.

[16] Ikus *aditu* aditzarena, https://hiztegiak.elhuyar.eus/eu/Heu003238_0, alde batetik, eta adjektibo eta izenarena, https://hiztegiak.elhuyar.eus/eu/Heu003239_0, bestalde.

Horrela, erroa edo hitz-eraketaren iturriak adieraz daitezke (Wikidatan, P5920 eta P5191, hurrenez hurren).

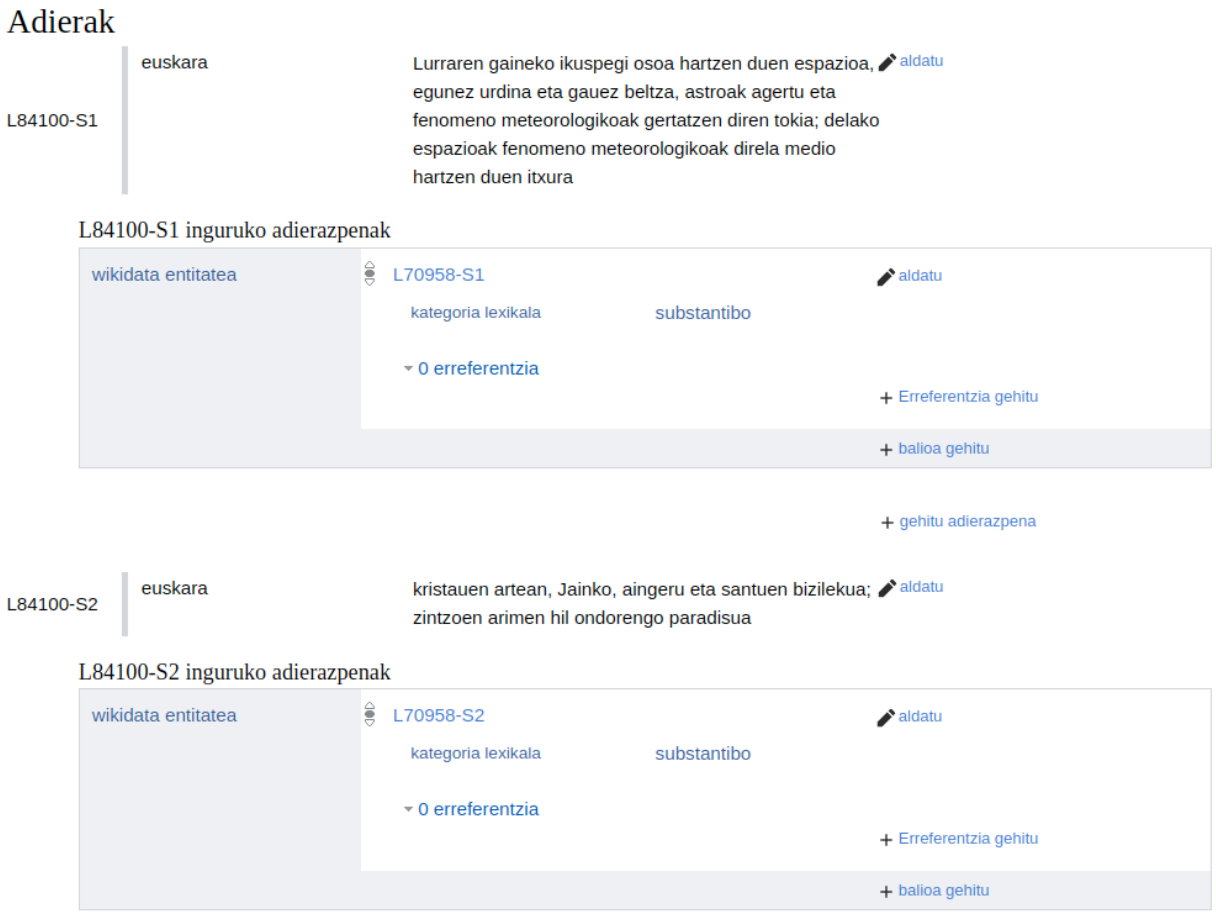


2. irudia. Hiztegi-sarrera Wikibaseko interfazean

3.3.2. ontolex:LexicalSense

Dagokion *LexicalEntry* entitatetik ontolex:sense propietatearen bitartez loturik, klase honetako entitateek lemaren adiera bana jasotzen dute. MLV Wikibaseko lema-bankua eraikitzeko, bi iturri ditugu, adiera-mailan lotura zuzena ezartzeko balio duen identifikatzaile unibokoa duten bakarrak, guk dakigula: Euskarazko WordNet (Pociello *et al.*, 2011) eta Wikidata, non Elhuyarreko Ikaslearen Hiztegitik erauzitako adierak agertzen diren. Bigarren iturri horretatik jaso ditugu adierak Wikibasean, POS desanbiguatzailearekin eta Wikidatako loturarekin batera, 3 irudiak erakusten duen moduan, *zeru* lemari lotutako adierak adibide.

Zeru sarrerako bi adiera horiek Wikibaseko beren identifikatzailea dute (baita Wikidatakoa ere, hortik datoz eta). Beste iturri batzuetatik adierak gehitzeko, haiei ere identifikatzaile bana esleituko diegu, gero erabakitzeko sarrera bereko beste adierekin zer erlazio duten. Wikidata eta Euskal Wordneteko adierak erlazionatzeko ahalegin batean (Ahmadi *et al.*, 2020 kolaborazioan), hiztegi adiera biren arteko erlazioak adierazteko balio hauek erabili ditugu: *exact match*, *broader sense*, *narrower sense*, *related sense*, *unrelated sense*. Zerrenda horri adieren arteko erlazio diakronikoak adierazteko balioak gehitu behar zaizkio.



3. irudia. Hiztegi-adierak Wikibaseko interfazean

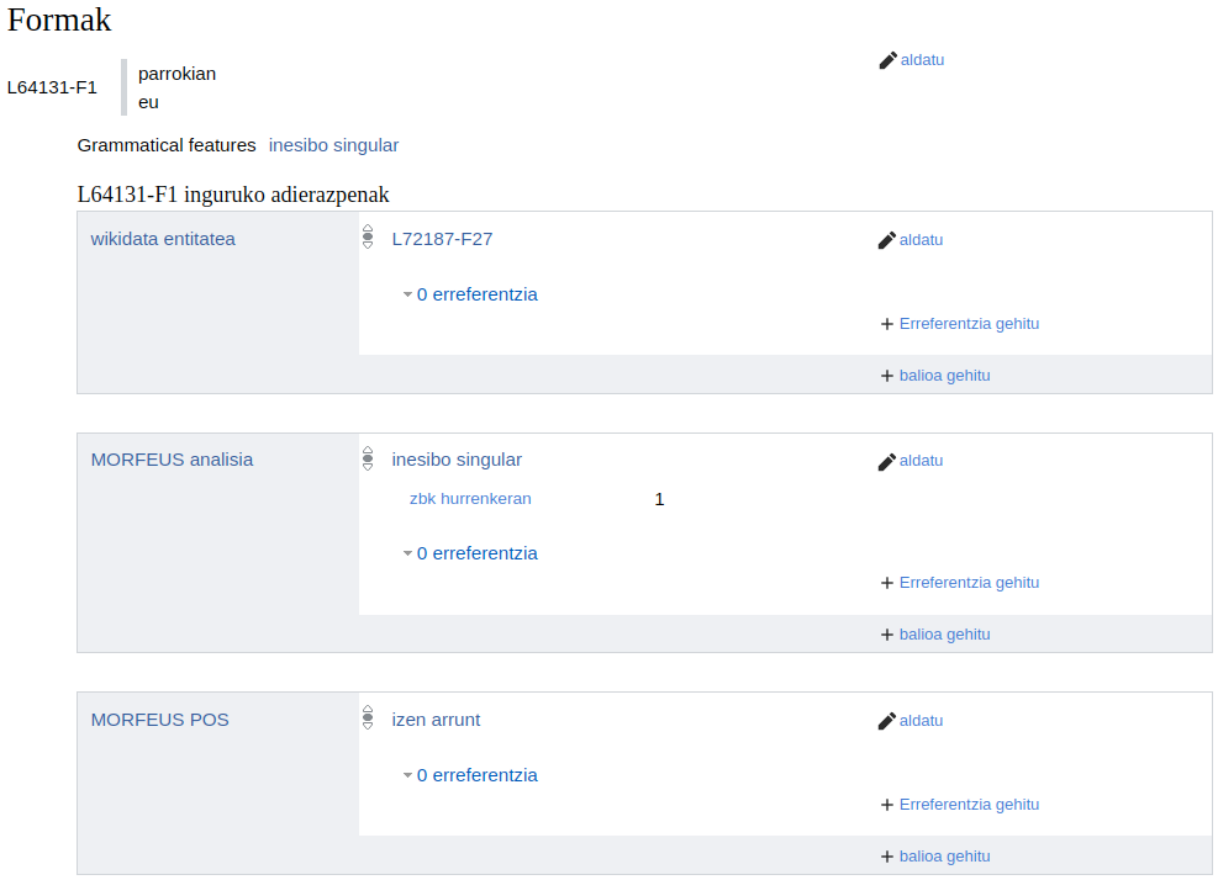
3.3.3. ontolex:Form

LexicalEntry bati dagokion forma bat ontolex:lexicalForm propietateaz lotzen zaio. Forma deskribatzen duten propietateak, ordea, hauek dira: ontolex:representation, forma bera errepresentatzen duen hizki-katea objektu duena, baita wikibase:grammaticalFeatures izeneko bat ere, Wikibase-ko oinarrizko ereduan formaren analisi morfologikotzat hartzen den balio-sorta duena; 4 irudian, inesibo singularreko forma bat deskribatzen da horrela. Euskarazko (eta, ustez, edozein hizkuntza eranskariren) morfologia deskribatzeko, oinarrizko eredu hori mugatua da, propietate honen balioa zerrenda ez ordenatua baita, hau da, *kaletiko* eta *kalekotik* moduko forma-bikoteek analisi bera izango lukete. Horregatik, berez sortutako propietatea gehitzen diogu ereduari, non MORFEUS moduko tresna batek (Aduriz *et al.*, 1999) erabiltzen duen gisako moldea erabiltzen baitugun: ordenatutako zerrenda bat, Wikibase batean ohikoa den eran antolatuta, hau da, zenbaki ordinal bat daraman propietate *kualifikatzailea* erabiliz.

Lema-bankuaren antzera, euskarazko formen bankua eraiki eta forma bakoitza deskribapen morfologiko posibleekin lotzeko, lehenengo saiakera bat egin dugu, ETC corpusean (Sarasola *et al.*, 2013) agertzen diren formak MORFEUSek POS eta analisi morfologikoari buruz ematen dizkigun datuak[17]

[17] MORFEUS tresnak datu-multzo handiak prozesatzeko interfaze publikorik oraindik ez duenez, CLARIAH-EUS bitartez eskatu dugu ETC-ko formak MORFEUS-ekin prozesatzea; prozesatutako datu-multzoa jaso genuen.

Wikibase batean jasoz.[18] Jokabide hori ebaluatu eta gero, MLV Wikibasean ere erabiltzeko asmoa dugu. ETC-ko formak (POS desanbiguatu gabe) milioi bateko multzoa direnez, ETC/MORFEUS hiztegi morfologikoa hirugarren datu-base batean gordetzea komenigarria dirudi, Corpus Historikoko formen baliokide estandarrak deskribatzeko beharrezkoak direnak MLV-n klonatuz, token-formaren bat baliokide horrekin lotzen den momentuan. Jokabide alternatiboa arauetan oinarrituta sortutako forma-bildumak erabiltzea litzateke (EDBL/XUXEN, Agirre *et al.*, 1992-tik), horrela arauen arabera posible den forma bakoitza deskribatzen duen hiztegi morfologikoa eraiki ahalko balitz.



4. irudia. Hiztegi-formak Wikibase interfazean

Hori esanda, argi dago hurbilpen honen atzean hipotesi bat erabiltzen dugula: Corpus Historikoko formek Euskara Batuko hiztegi morfologiko batean agertuko zaigun forma bana dutela lagun edo baliokide, hark duen analisi morfologikoa eurengana hedatzeko balioko duena. Adibideko *parroquian* eta estandarreko *parrokian* biek analisi morfologiko bera duten bitartean, izenkien zenbait atzizkiren kasuan, eta batez ere aditz morfologian kasuistika zailagoa dugu, morfema bakoitzak duen balio sintaktikoa desanbiguaturik adierazi nahi badugu, behintzat. Estandarrean subjunktiboari dagozkion morfema batzuk, adibidez, literatura historikoan aoristoari zegozkion. Kontu horiek nola modelatu etorkizun hurbileko erronka dugu; aditz morfologia modelatzeko Ariztimuño (2023) laneko proposamenetik abiatuta, Wikibaseari egokitutako eredua

[18] Ikus https://datuak.ahotsak.eus/wiki/Lexeme:L251 *agindu* lemari dagozkion formak, aditz nahiz izenkien morfologiari dagozkionak.

zehaztea da helburua.[19] Lema mailan gertatzen den bezalaxe (**edun*, **eradun*, **iron*, **erazan*, eta abar), dokumentutako errepresentaziorik ez duten formak hiztegian jaso behar izatea ezin dugu baztertu.

4. Loturak corpusetik hiztegira

Estarrona *et al.* (2022) lanak darabilen irizpideari jarraituz, uste dugu forma historikoen analisi konputazionala hasteko biderik egokiena forma historiko hori forma estandarrekin lotzea dela. Horren arrazoi nagusia zertan datza: LNPko (Lengoai Naturalaren Prozesamendua) tresnak Euskara Batuarentzat dira, esaterako, etiketatze morfosintaktikoak eta hortik abiatzen diren analisi aurreratuak; horietaz baliatzeko, beraz, forma ez-estandarra estandarrarekin lotu behar da lehendabizi. Horrela egingo dugu, berezko propietatea erabiliz.[20] Corpus tokena hiztegiko forma bati lotzeak, gure hiztegiaren antolamenduaren arabera, token hori lema-bankuko sarrera zehatz bati lotzea dakar inplizituki, baina hala ere, lotura hura errepresentatzen duen propietatea ere badugu, tokenetik lemara doana, alegia. Izan ere, baliteke forma estandarra hipotetikoa izatea, hau da, formak Euskara Batuan dokumentaturik ez egotea edo hiztegi morfologikoan hura deskribatuko lukeen sarrerarik ez egotea, eta hori modelatzeko irtenbidea definitu bitartean, lema mailako lotura behintzat adierazteko baliabidea dugu. Bestetik, tokenetik hiztegi-formara doan lotura ezartzeko, OEHn oinarritutako lemategia dugu, non hainbat aurrizkidun eta atzizkidun formak ez diren lemategian agertzen, esaterako baieztapeneko *bai-* eta ezeztapeneko *ez-* aurrizkiak *baitut* edo *eztaki* formetan kasu, edo *-lako* moduko atzizkiak. Partikularen presentzia adierazpen kualifikatzaile batez adieraziko dugu, tokenetik formara doan loturari erantsita.[21] Arazoak arazo, tokenetik hiztegi-formara daraman lotura definituz gero, tokenaren deskribapen morfologikoa dugu, forma estandarraren bitartez.

Halaber, adieraren lerrokaketa adierazteko, beste propietate berezitu bat erabiltzen dugu. Zehazki, lemari dagokion loturaren kualifikatzaileak dira formaren eta adieraren lerrokaketa proposatzen dugun ereduan; Wikibasearen interfaze grafikoan hierarkia hori 5 irudian bezala irudikatzen zaio erabiltzaileari.

5. irudia. Wikibase adierazpena kualifikatzaileekin

[19] Izan ere, euskarazko aditz morfologia Wikidatan ere ez dago modu egokian modelatuta, eta modelo ezak oztopatzen du momentu honetan adizkien formak bertan deskribatzea.

[20] Propietateak https://monumenta.wikibase.cloud/entity/P7 identifikatzailea du MLV Wikibasean.

[21] Horrela kualifikatutako adierazpena hizkuntza naturalera ekarrita: Token t-k duen forma historikoari Euskara Batuko x forma dagokio, eta forma historikoak y aurrizkia eta z atzizkia daramatza.

Corpus Historikoko tokenak edo token-multzoak semantikoki anotatzeko zabaltzen dugun aukera, beraz, hiztegiko adierekin lotzean datza. Honela zenbait neurketa ahalbidetzen ditugu. *Zeru* lemaren adibidean geratzeko, Larramendik bere sermoian bi adierak erabiltzen ditu (izarrak dituena eta kristau kontzeptua); erabilera kuantitatiboki konpara daiteke eta, dokumentuak gehitzen diren heinean, testu eta testuen metadatuen bitartez, garai eta kokapen geografiko desberdinetan zehar ere neurtu ahalko dugu.

Formei dagokienez, 3.3.3 atalean aipatutako modelatzeko zailtasunak baztertu gabe honako aurreikuspena argi dago: forma historikoa hiztegi estandarreko formen deskribapenekin lotzea lortuz gero, hainbat aplikazio posible bihurtzen dira. Esaterako, itzulpen edo erregeletan oinarritutako testu-sorkuntza diakroniko-dialektalak egitea, kontuan izanda forma historikoak agertzen diren dokumentuen metadatuekin lotuta daudela. Beste aplikazio bat, ikasketa automatikoak eszenatokia hartzen duen honetan, forma historikoak forma estandarrekin lotzeko prozesu automatizatuak ebaluatzeko balio duen *gold standard* gisa erabiltzea izan daiteke.

5. Lan-fluxuak

Gorago esan bezala, eskuizkribuko tokenak definitu eta Wikibasen jasotzeko helburuarekin, *script* automatizatua erabiltzen dugu. Puntu horretatik aurrera, hau da, datuak Wikibase plataforman jasotakoan, bi aukera daude beti ere datuak aberasteko, editatzeko edo datu-basetik ateratzeko: i) eskuz lan egitea, Wikibaseren interfaze grafikoaren bitartez, edo ii) programen bidez lan egitea, hots, *script* berezituak sortuz eta aplikatuz, ekintza bat baino gehiago aldi berean, ekintza bat entitate askotan edo askotan errepikatuko den ekintza bat gauzatuz. Azkoitiko Sermoiarekin lotutako gure esperimentuetako zenbait ataza honako hauek dira:

— Token-multzoak definitu eta sortzea.
— Anotazio filologikoak tokenari (edo token-multzoari) lotzea.
— Tokena (edo token-multzoa) hiztegiarekin lotzea: dagozkion lema, adiera eta forma zehaztuz.
— Tokena (edo token-multzoa) pertsona bat, leku bat, erakunde bat, eraikin bat, lan literario bat edo bestelako entitate izendun bat deskribatzen duen Wikidatako kontzeptuarekin lotzea.

Ataza horiek guztiak Wikibasearen interfazetik bertatik gauza daitezke eskuz, eta hark berez dakartzan funtzionalitateak lagungarriak dira zeregin horietan. Token bat deskribatzen duen entitate-orrialdean tokenari dagokion lema gehitzeko, adibidez, erlazio horri dagokion propietatearen izena edo zenbakia adierazi ondoren, eta propietate horren balioa daraman eremuan idazten hasiz gero, lematégiko proposamenak datozkigu, teklatuan sartzen dugunaren arabera (ikus 6 irudia, non «ceruan» forma duen token bat deskribatzen den). Lema egokia aukeratuta, adiera eta formarekin modu berdintsuan jarraituko dugu. *Zeru* lemak[22] bi adiera ditu hiztegian, bakoitza bere definizioarekin: «Lurraren gaineko ikuspegi osoa hartzen duen espazioa, egunez urdina eta gauez beltza, astroak agertu eta fenomeno meteorologikoak gertatzen diren tokia», alde batetik, eta «kristauen artean, Jainko, aingeru eta santuen bizilekua; zintzoen arimen hil ondorengo paradisua», bestetik. Adiera egokiaren identifikatzailea hautatuz gero, erlazio hori ere grabatuta geratuko da.

[22] Ikus https://monumenta.wikibase.cloud/wiki/Lexeme:L84100

6. irudia. Eskuz editatzea Wikibase interfazean

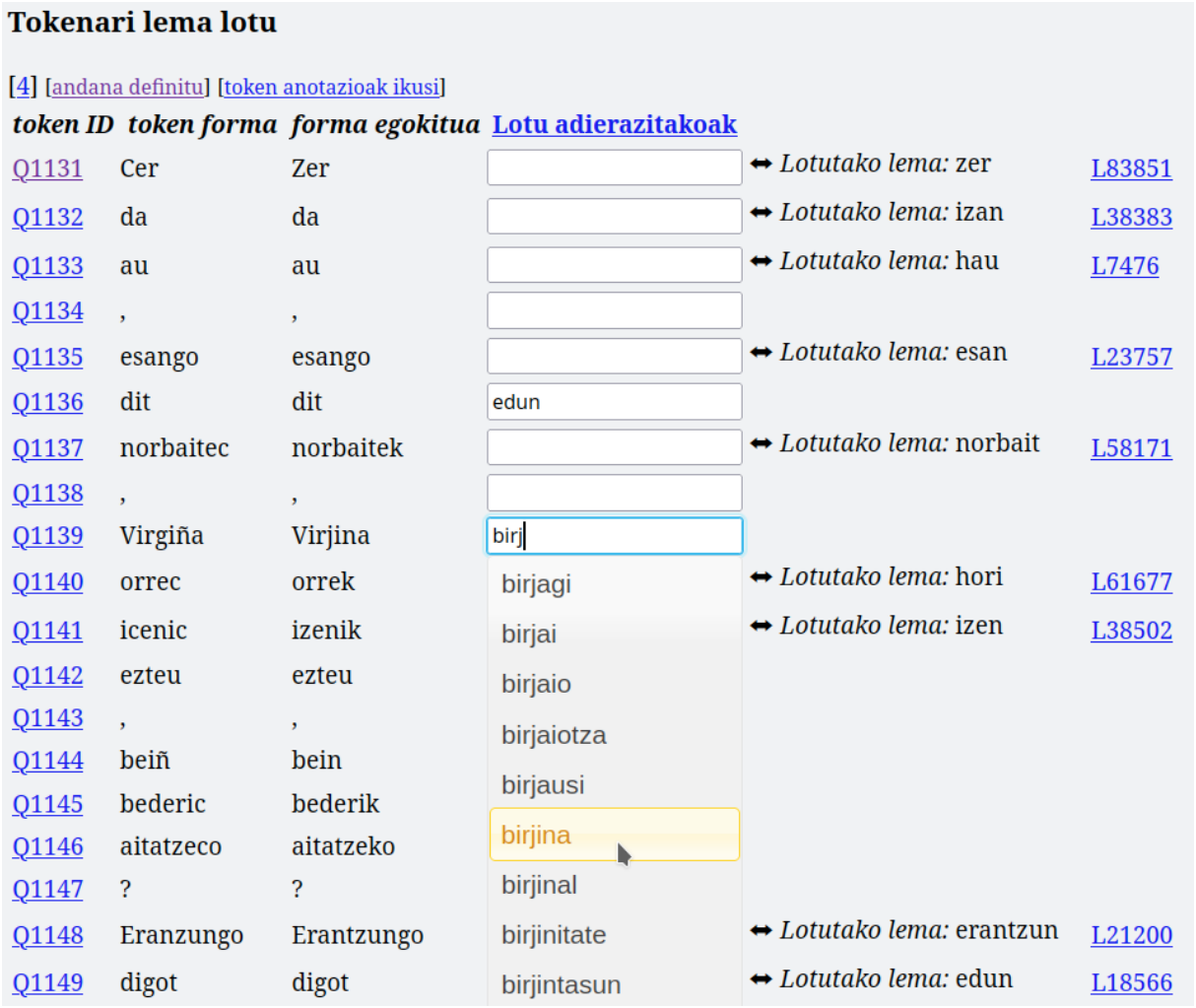
Metodo bat baliatu bagenu forma historikoei dagozkien euskara batuko formak edo lemak proposatzeko, Estarronak *et al.* (2022) erabilitakoa, ikasketa automatikoan oinarritzen dena, adibidez, eskuzko editatze-metodo hori automatikoki proposatutako balioak balidatzeko erabiliko genuke, hau da, beharrezkoa denean, zuzentzeko. Proposamenak egiten dizkigun tresnak ekoizten duen erlazio-sorta Wikibasean jasoko genuke ez eskuz, baizik eta era masiboan, horretarako idatzitako *script* baten bitartez, iturriko datuak irakurri eta Wikibasearen APIra bidaliko dituena, grafoan jaso daitezen.[23]

Horrelako *script* batek berez programatutako web-aplikazio baten barruan ere lan egin dezake. Horrela egiten dugu, adibidez, token-multzoak sortzeko: Aplikazioan,[24] erabiltzaileak testu baten paragrafo zehatza kargatu ondoren, multzoaren lehenengo eta azkenengo tokena hautatzen ditu saguarekin. Ondoren, hautatutako tarteko token guztiak sortzear dagoen multzoaren parte izango diren egiaztatuko du erabitzaileak, eta berehala izango du multzo sortu berria Wikibasean ikusi eta editatzeko aukera.

Lehenengo esperimentuetan beharrezko ikusi ditugun beste bi funtzio ere sortu ditugu: tokenaren grafiaren gaurkotze automatikoa zuzentzeko, alde batetik, eta tokenak hiztegiko lemekin lotzeko, bestalde. 7 irudiak azkenik aipatutako atazari dagokion lan-fluxuaren momentu bat jasotzen du, eskuzko lematizatzailearena, alegia. Ikusten denez, lema idaztean, aplikazioak OEHren lemategiko sarrerak proposatzen ditu.

[23] Wikibase entitate baten datuak JSON errepresentazio batean sortu edo aldatu egiten dira; horretarako, modulu laguntzaileak daude dagoeneko, adib. *python wikibaseintegrator* moduluak, ikus https://github.com/LeMyst/WikibaseIntegrator

[24] Aplikazioa sortzeko, aurretik existitzen zen ZotWb aplikazioa hartu dugu oinarri (ikus https://github.com/dlindem/zotwb). ZotWb Wikibase batekin komunikatzen duen kode irekiko python web aplikazio bat da, eta beste zeregin batzuetarako sortu zen arren, egokitzapen gutxirekin gure atazak gauzatzeko moldatzen dugu.



7. irudia. Ataza bati egokitutako web-aplikazioa

6. Ondorioak

Testu historikoen edukiak eta hainbat motatako anotazioak errepresentatzeko datu-eredua aurkeztu dugu, eta lehenengo testu bat, Larramendiren Azkoitiko Sermoia, prozesatzeko moldatu dugu. Argi utzi dugu hemen aurkezten duguna martxan dagoen prozesu baten argazkia besterik ez dela. Espero dugu CLARIAH-EUS 2. workshop honetan egin dugun ekarpen honek datu-eredua eta harekin lotutako lan-fluxu nahiz aplikazioak eztabaidatzeko balioko duela, urruneko gure helburuari begira, euskararen Corpus Historikoko testu guztien edukiak Datu Lotuen paradigmaren arabera jaso eta eskuragarri jartzeari begira, alegia, filologia konputazionalerako lanabesak osatzeko bidean.

Bibliografia

Aduriz, I., Agirre, E., Aldezabal, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Maritxalar, A., Sarasola, K. and Urkia, M. (1999). MORFEUS: Euskararako analizatzaile morfosintaktikoa. Barne-txostena, UPV/EHU/LSI/TR. http://ixa.si.ehu.es/node/4059

Agirre, E., Alegria, I., Arregi, X., Artola, X., Ilarraza, A. D. de, Maritxalar, M., Sarasola, K. and Urkia, M. (1992). XUXEN: a spelling checker/corrector for basque based on two-level morphology. Proceedings

of the Third Conference on Applied Natural Language Processing, 119-125. pp. ACM. http://portal.acm.org/citation.cfm?id=974499.974520

Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Gabrovsek, D. (2020). A multilingual evaluation dataset for monolingual word sense alignment. Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020). LREC 2020, 11-15 May, Marseille. https://doi.org/10.13025/v0mx-x745

Alonso, M. and Lindemann, D. (2022). Larramendiren Hiztegi Hirukoitzaren digitalizazioa. Karaktereen ezagutze optikoa eta Wikitekara igotzea. Uztaro. Giza eta gizarte-zientzien aldizkaria, 120, 83-93. pp. https://doi.org/10.26876/uztaro.120.2022.5

Ariztimuño, B. (2023). Euskara Arkaikoko adizki jokatuen gaineko ikergaiak. Corpusa eta azterketa filologiko-linguistikoak [PhD thesis, UPV/EHU]. https://addi.ehu.es/handle/10810/62152

Bilbao, G., Gómez, R., Lakarra, J. A., Manterola, J., Monoule, C. and Urgell, B. (2011). Lazarraga eskuizkribuaren edizioa eta azterketa, v.1.2, Vitoria Gasteiz: UPV/EHU. https://www.ehu.eus/monumenta/lazarraga/

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R. and Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. Digital Scholarship in the Humanities, 37(2), 391-404. pp. https://doi.org/10.1093/llc/fqab066

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating NLP Using Linked Data. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Eds.), The Semantic Web – ISWC 2013 (pp. 98-113). Springer. https://doi.org/10.1007/978-3-642-41338-4_7

Lakarra, J. A. (1985). Literatur gipuzkerarantz: Larramendiren Azkoitiko sermoia (1737). Anuario del Seminario de Filología Vasca «Julio de Urquijo», 19(1). https://doi.org/10.1387/asju.7685

Larramendi, M. de. (1990). Manuel Larramendi: Euskal testuak : (1690-1990 III. Mendeurrena) (P. Altuna eta J. A. Lakarra, Arg.). Andoaingo Udala.

Lindemann, D. and Alonso, M. (2023ko azaroak 23). Testu historikoak wiki-plataformetan, Datu Lotu gisa [Poster presentation]. CLARIAH-EUS 2. workshopa: azpiegitura eraikitzen, Donostia. https://doi.org/10.13140/RG.2.2.30500.86400

Lindemann, D., Astigarraga, A., Bidaguren, M., Delgado, E., Gonzalez, G. and Sarasola, K. (2023). Inguma eta Wikidata uztartuz, euskarazko zientziaren ezagutza-graforantz. In D. Lindemann (Ed.), Miren Azkarateri esker onez (193-210. pp.). UPV/EHU Argitalpen Zerbitzua. https://doi.org/10.5281/zenodo.7902516

Lindemann, D. and San Vicente, I. (2020). Baliabide lexikoen sarea: Baldintza filologiko eta tekniko zenbait. In Hitzak sarean: Pello Salaburuari esker onez (79-96. pp.). UPV/EHU Argitalpen Zerbitzua. http://www.ehu.eus/ehg/salaburu/liburua/HitzakSarean06.pdf

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek and V. Baisa (Eds.), Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017 (http://lexbib.elex.is/entity/Q6744; 587-597. pp.). Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Pedonese, G., Cecchini, F. M. and Passarotti, M. C. (2023). Linking the Computational Historical Semantics corpus to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch and A. Salgado (Eds.), Proceedings of the 4th Conference on Language, Data and Knowledge (74-85. pp.). NOVA CLUNL, Portugal. https://aclanthology.org/2023.ldk-1.7

Stanković, R., Chiarcos, C., Utvić, M. and Kitanović, O. (2023). Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. In S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch and A. Salgado (Eds.), Proceedings of the 4th Conference on Language, Data and Knowledge (180-191. pp.). NOVA CLUNL, Portugal. https://aclanthology.org/2023.ldk-1.16

XVIII. mendean Euskal Herrian inprimatu idazkiak WikiDatan

Writings published in the Basque Country in the 18th century in WikiData

Iñaki Lopez de Luzuriaga Martinez

Ikertzaile independentea
inaki@wikimedia.eus

Laburpena

Artikulu honetan, prozesuan den katalogazio bibliografiko digital irekiko proiektu bat, haren bilakaera eta erabilitako metodologia deskribatzen dira. Hau da, Bizkaiko Foru Liburutegiko *Lau Haizeetara* funts bibliografikotik abiatuz, Euskal Herriko inprimategietan XVIII. mendean editatu ziren idazkiak Wikidata datu-base librera igotzen ari gara eta, hartara, hura erabiltzaileen eskueran jarriko dugu. Digitalizazio soila gaindituz, katalogo bibliografiko honen sorrerak 5 izarreko datu ireki estekatuen ideia hartu du oinarri teorikotzat, Tim Berners-Leek definitu bezala. Prozesuan, bitarteko gisa datu bibliografikoen mordoiloa kudeatu eta Wikidatarekin bateratzeko OpenRefinek dituen ahalmen eta gabeziak egiaztatu dira. Orain arteko emaitza bibliografikoak SPARQL eskaeren bidez eskura daitezke, eta aukera ematen da, gainera, Wikipediako wikiproiektu-orri bateko interfazearen bidez horiek ikusteko, estatistika-grafikoekin eta Wikipediako artikulu-estekekin batera. Proiektua irekita dago bukatzeko.

Gako-hitzak: katalogo bibliografiko digital irekia, datu bibliografikoen trataera, datu bibliografikoen bistaratzea.

Abstract

This article describes a project to generate an open digital bibliographic catalogue, its current progress and the methodology used in its pursuit. The project aims to upload to the free Wikidata database a catalogue of 18th century bibliographic elements printed in Basque Country presses. The task goes beyond digitalisation, adopting as its theoretical mainstay the generation of 5 star data, as defined by Tim Bernes-Lee. Data wrangling and reconciliation of these bibliographic elements through the application OpenRefine has enabled us to test its capabilities and deficiencies. Bibliographic results so far may be accessed by SPARQL queries, and an interface has also been provided on a Wikiproject portal to visualize them alongside related statistical diagrams and links to Wikipedia articles. The project is still due for completion.

Keywords: open digital bibliographic catalogue, bibliographic data manipulation, bibliographic data visualisation.

1. Sarrera

1.1. Helburua

Katalogazio proiektu honen azken helburua da Euskal Herriko inprimategietan XVIII. mendean editatu zen guztia Wikidatara igotzea eta hura erabiltzaileen eskueran jartzea, dela norberak Wikidatan bilaketak eginez (elementu zehatzena edo datu serieena, SPARQL eskaerak edo PetScan baliatuz), dela horretarako berariaz jada sortu den Wikiproiektuko zerrenda-orria erabiliz.

Hain zuzen, une honetan, proiektuak EU Wikipedian du bere adierazpena nagusia: «Wikiproiektu:XVIII. mendean Euskal Herrian inprimatu idazkiak». Orri horren bidez, 339 elementu bistaratu daitezke kontsulta egiteko unean (2024ko maiatzaren 23an). Wikiproiektuko kontsulta horrek Wikidataren SPARQL eskaera bat du, etengabe eguneratzen da, eta eskaintzen dituen elementu bibliografiko gehienak artikulu honetan azaltzen den proiektuaren bidez igo dira Wikidatara, Bizkaiko Foru Liburutegiko *Lau Haizeetara* katalogo digitala arakatu ondoren.

1.2. Abiaburua

Proiektua abiatzeko, hainbat zirkunstantzia gurutzatu dira: UEUko Humanitate Digitalen graduondokoaren amaierako ikerlana (2021-2022), biltegi digitalen egungo garapena (Sancho el Sabio, Bilketa, Bizkaiko Foru Liburutegia, Euskariana...), datu irekietan oinarritzen den Wikipediako zein Wikidatako nire jarduera, baita OpenRefine datuak kudeatzeko aplikazioaren garapen askea ere.

Edukiari dagokionez, proiektuak Joxe Azumendi filosofoaren *Pentsamenduaren historia Euskal Herrian* idazlanean du sorburu, non adierazten baita ongi legokeela «XVIII. mendean Euskal Herrian editatu zen guztiaren zerrenda bat eskuratzea». Azken hamarkadetan, hainbat idazlan plazaratu dira XIX. mendeaz aurreko edizioaz, banaka euskal lurralde bakoitzean (ikus behean, bibliografia). Edizioari buruzko ikerlanak dira, baina inprimatu zenaren katalogoak ere eskaintzen dituzte.

Hutsune bat sumatzen zen: garaiko edizioaren irudi oso baten falta Euskal Herrirako. Bizkaiko Foru Liburutegiak Jon Bilbaoren Eusko Bibliographia bilduma du, Euskal Herriko edizio historikoan ezinbesteko erreferentzia. Proiektu hau hasteko unean, ordea, guztia digitalizatu gabe zegoen. Egun, UPV/EHUk (Addi) Eusko Bibliographiako 1981-1985 urteei dagokien euskal ikasketetako bibliografia argitaratu du digitalizaturik edonork eskuratzeko moduan. Hasteko, Araba, Bizkai eta Gipuzkoako elementu bibliografikoak lagin handi samarra ziruditen katalogazio lana egiteko, baina *Lau Haizeetara* katalogoko lagina txiki samarra zela ikusita, markotzat Euskal Herria hartzea erabaki zen, hartara Azurmendik adierazitako nahiarekin bat eginez.

2. Metodologia

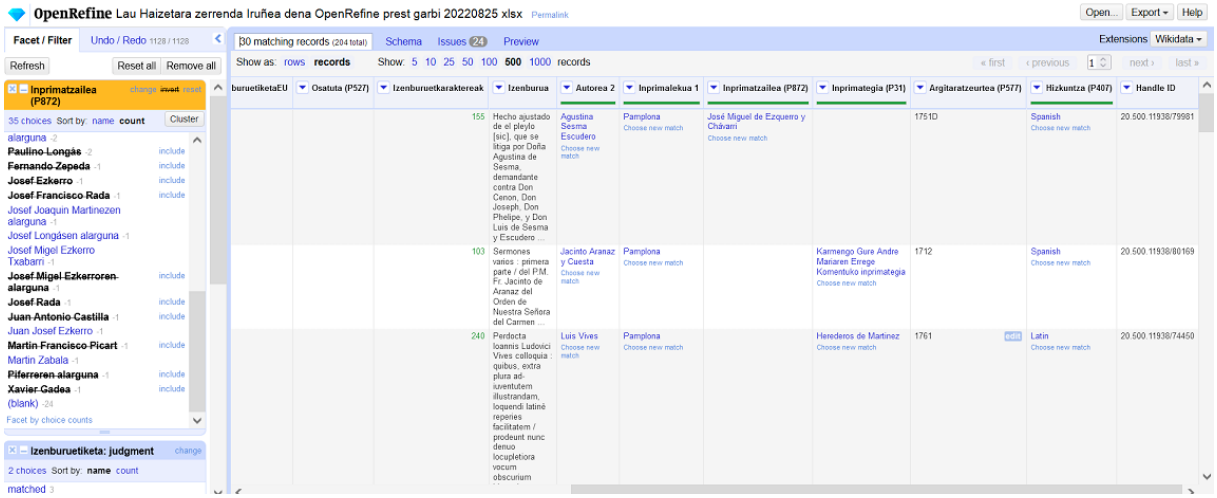
Metodologia aldetik, hainbat fase behar izan dira proiektuan, jatorrizko *Lau Haizeetara* katalogotik Wikidatara:

1. **Bilaketa eta ikerketa fasea.** Bilaketa aurreratuaren bidez, *Lau Haizeetara* biltegi digitalean iragazkiak ezarri dira: mende horretan (1700-1799), Euskal Herriko inprimalekuak zeintzuk ziren aurkitu dira, non inprimatzen zen ezagutzeko. Horretarako, Bizkaiko Foru Liburutegiko

«Biltegi osoa» aukeran, 1700-1799ko zerrenda osoa behatu da. Ariketa horrek eta Jose Azurmendiren *Pentsamenduaren historia Euskal Herrian* idazlana irakurtzeak berak dedukzio baterako bidea zelaitu du: XVIII. mendean inprimatu eta argitaratu ziren tokiak identifikatzea.

1. irudia. Gasteizko bilaketen pantaila-irudi bat *Lau Haizeetara* katalogoan

2. **Inprimalekuen eta argitalpenen zerrenda eskuratzea.** *Lau Haizeetara* katalogoan iragazki bidezko bilaketa eginez, zerrenda batzuk eskuratu dira 1700etik 1799ra, aurreko fasean ikertutako herrien eta hirien arabera. Gehienez 100 sarrerako zerrenda-orriak sor zitezkeen aldi berean, gure kasuan, inprimalekuen araberakoak.
3. **Zerrendak ordenagailuan.** Zerrenda horiek kopiatu eta Calc orrietan itsatsi dira. Zerrenda horietan, ordea, ez dago Handle IDrik, jatorrizko dokumentura zuzeneko esteka eskaintzen duena. Ctrl+U komandoaren bidez, zerrenda orrien iturburu kodea lortu da, eta bertan Handle IDa. Gero, identifikatzaile horiek sarrera bibliografikoetan ordenatuki txertatu dira. Markaketa lengoaia erantzi eta zerrenda gorde egin da.
4. **Informazio-antolaketa eta -kudeaketa eta datu-uztarketa (OpenRefine).** Oraindik ere datu-mordoilo bibliografikoak geratzen dira, eta ez dira datu egituratuak, ezin dira digitalki irakurri. Puntu honetan, Calc orritik OpenRefine aplikaziora pasa dira zerrendak, askotariko parametroak identifikatu eta antola ditzan. OpenRefine aplikazio lokala izan daiteke, edo Wikimediaren PAWS plataforma erabil daiteke, hodeian dagoena. OpenRefine aplikazioak datu-mordoiloak garbitu eta guk aukeratzen ditugun zutabeen arabera antolatu ditu, Wikidatan existitzen diren propietateekin bat etorri behar direnak: izenburua (Wikidatako etiketa), deskribapena, autorea (P50), argitaratze-lekua (P291), inprimatzailea (P872), inprimategia (P123), argitaratze-data (P577) lanaren edo izenaren hizkuntza (P407), Handle ID (P1184). OpenRefineko taulan datuak Wikidatarekin uztartu ondoren (*reconciliation*), Eskema fitxan (*Schema*), zutabeak eskema gisa egituratu dira Wikidatara igotzeko.



2. irudia. Iruñea inprimalekuaren sarrera bibliografikoetako datuen kudeaketa OpenRefinen, zatika igotzeko prestatzen

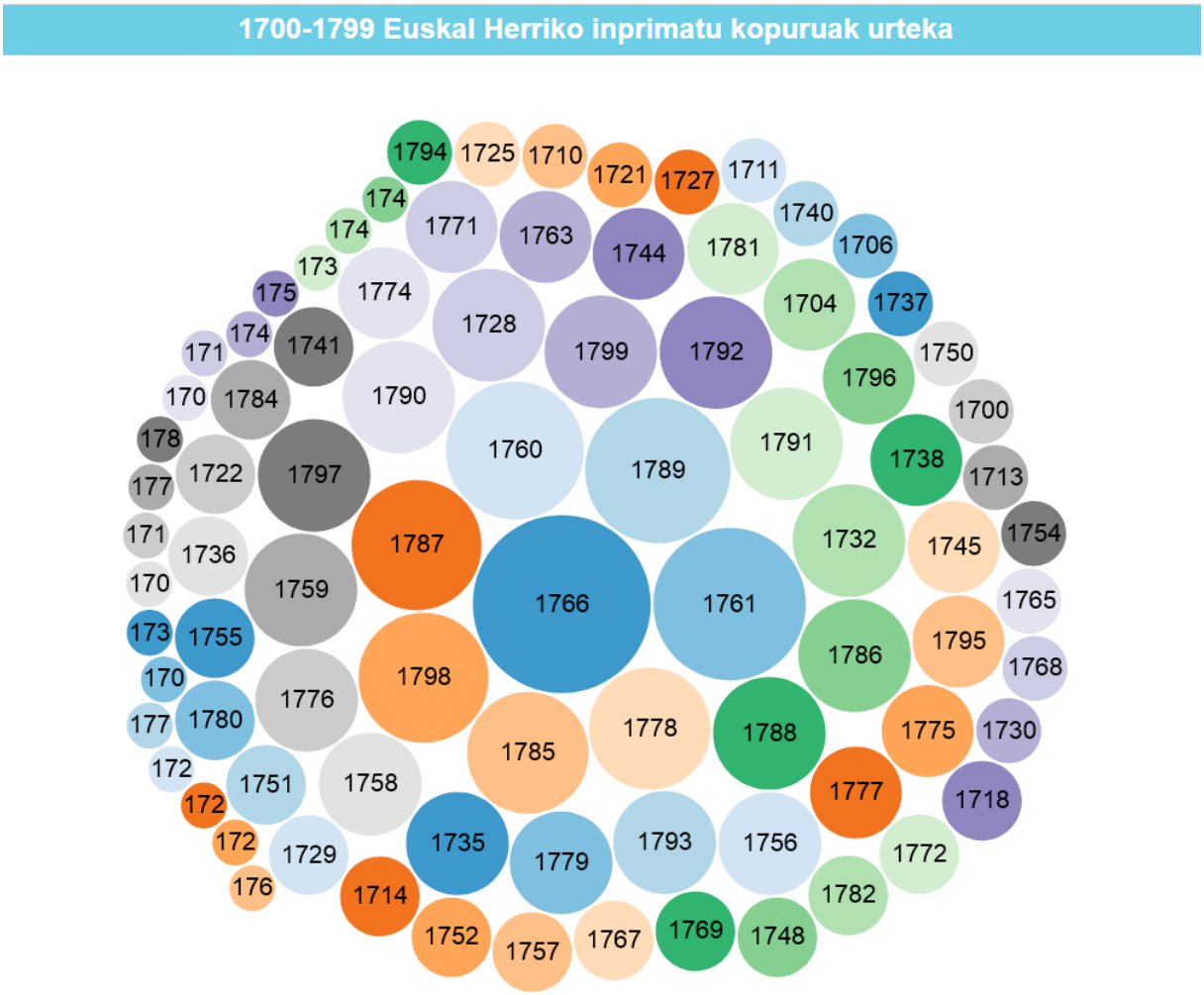
5. **Wikidatako igoera.** Azken fasean, OpenRefineko datu antolatuak Wikidatara igo dira (*Upload edits to Wikibase*). Hartara, QID elementu berriak sortu dira wikibase horretan. Igoera datu-serie laburretan egin da, erroreak eta ondorioak errazago identifikatzeko. Wikidatan bertan, batzuetan, elementu (QID) berriak sortu behar izan dira banaka, igotako elementu bibliografiko batzuek lehenago Wikidatan existitzen ez ziren balioak zituztelako, hala nola autore batzuk.
6. **Bistaratzea.** Elementu bibliografikoak Wikidatara igotzean, eskueran daude bistaratuak izateko. Horretarako, *Wikidatako Query Service* balia daiteke (https://query.wikidata.org/), galdera egokia eginez eta *Graphs* edo *Maps* aukeraren bitartez emaitzak bistaratuz. Datuak bistaratzeko, beste aplikazio baliagarri batzuk ere badaude *Wikidata:Tools* orrian bilduak (https://www.wikidata.org/wiki/Wikidata:Tools).

3. Ondorioak eta etorkizuneko lanak

Une honetan, *Lau Haizeetara* katalogoko zerrenda bibliografikoa igo da Wikidatara UEUren bekaren bitartez. Bertan, hainbat propietate edo metadatu adierazi dira: autorea (eta, beraz, haren generoa), idazkiaren (edo haren izenburuaren) hizkuntza, inprimatzailea, inprimategia, argitaratze-data, inprimalekua, eta azalaren irudia. Hala ere, inoiz baino gehiagotan, propietate batzuek ez dute balio bat esleiturik: hutsik egon daitezke, adibidez, irudia.

Wikidatako emaitzak, *Lau Haizeetara* katalogoko igoeratik eskuratuak, XVIII. mendean Euskal Herrian inprimatu zenaren zati txiki bat dira, azterketa bat eginez atzeman denez. Azken helburua partzialki baizik ez da bete. Izan ere, identifikatutako inprimaleku edo hiriburu batzuek oso emaitza urriak eman dituzte jatorrizko katalogoan eta Wikidatan. Baionan, zazpi elementu bibliografiko besterik ez dira ageri, jakina den arren, Bilketa atari gigitalaren kontsultaren bitartez, askoz gehiago argitaratu zirela. Hala izanagatik, XVIII. mendeko Euskal Herriko katalogo bibliografiko baterako oinarri bat finkatu da.

Bizkaiko Foru Liburutegian, *Lau Haizeetara* katalogotik at, Jon Bilbaoren Eusko Bibliographiak arreta handiagoa merezi du, bertatik XVIII. mendeko zerrenda bibliografiko bat eskura baitaiteke. Hargatik, hura ez dago guztiz digitalizatuta, eta hori eragozpen bat da Internet bidez zuzenean bilaketak egiteko eta datuak erauzteko.

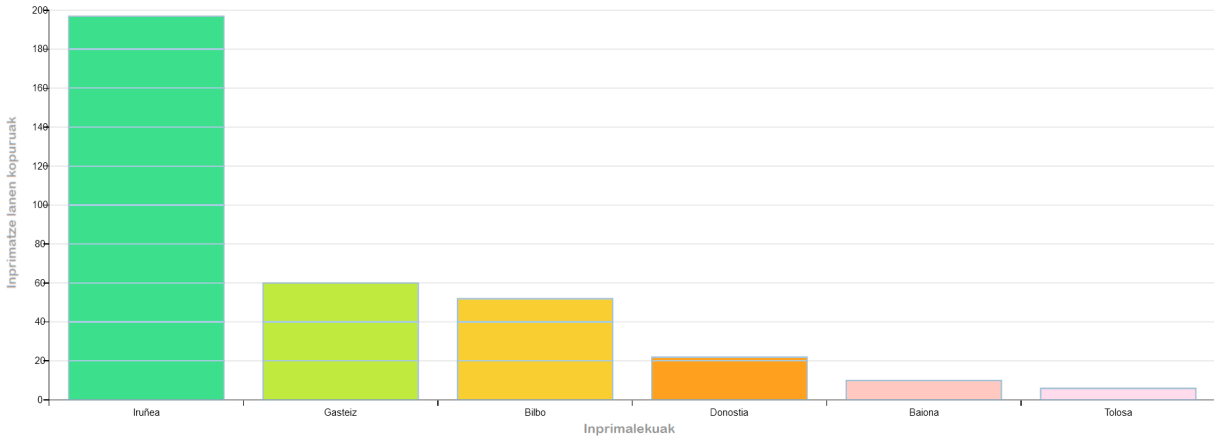


3. irudia. Urtekako edizio kopuruen bunbuilo-diagrama, bolumen handienetatik (erdialdean) txikienetara. Query Service bidez eskuratutako datuak

Proiektuan aurrera egiteko, nahitaezkoa da Euskal Herriko beste katalogo digitaletan arakatzea: Nafarroako Artxibo Nagusia, Bilketa, Koldo Mitxelena Liburutegia. Sancho el Sabio eta Euskariana liburutegi digitalak ere arakatu beharrekoak dira. Azken bi biltegi digital horiek arakatzean, zuhur erreparatu behar zaie elementu bibliografikoen iturriei. Hainbat sarrera bibliografiko ez dira bertakoak, beste katalogo digital batzuetatik edaten dute: ez daude beti biltegi horietan, eta izenburuak eta horien kopuruak errepikatzeko arriskua dugu. Bestalde, XVIII. mendeko edizioari buruz orain arte argitaratu ikerlanek informazio eta iturri bibliografiko osagarriak eskain ditzakete.

Hauxe lizentzia askeko proiektu bat da, edonor sar daiteke bertara, lankidetzara irekia da, eta Wikipediara bideratutako ekarpenak egin daitezke, elementu bibliografiko bakoitzeko datuak osatuz, hala nola haien azalak gehituz, falta baldin badira: nahikoa da Wikiproiektuko zerrendako dena delako elementu bibliografikoaren Handle IDtik irudia eskuratzea, Wikimedia Commonsera igotzea eta, gero, Wikidatako irudian (P18 propietatea) txertatzea. Halaber, jada Wikidatan dauden elementuen copyright estatusa edo orrialde kopuruak gehitu daitezke, orain arte egin ez dena.

Prozesuaren askotariko etapetan, aurrerantzean, lagungarria izan daiteke CLARIAH-EUS Giza eta Gizarte Zientzietan euskara eta euskaraz ikertzeko egitasmoa.



4. irudia. Euskal Herrian 1700etik 1799ra inprimatu idazkien kopuruak inprimalekuka (Wikidatako laginaren araberako barra-diagrama)

Zerrenda [aldatu iturburu kodea]

#n	item	Dokumentua	Urtea	Egilea	Tokia	Irudia	Hizkuntza	Inprimatzailea
1	Q30133560	Lecciones náuticas explicadas en el museo matemático del M. N. y M. L. Señorio de Vizcaya, noble villa de Bilbao y su ilustre casa de Contratacion	1756	Miguel Archer	Bilbo		gaztelania	Antonio Eguzkitza
2	Q51461531	Curiosidades de la naturaleza, y del arte (...) Tomo II. Segunda impression	1735		Iruñea		gaztelania	
3	Q57672560	Jesu-Christoren Imitacionea	1720	Michel Churio	Baiona		euskara	Paul Fauvet
4	Q65216433	Larramendiren Hiztegi Hirukoitza	1745	Manuel Larramendi	Donostia		euskara gaztelania latin	Bartholome Riesgo Montero
								

5. irudia. Pantaila irudia: 1700etik 1799ra Euskal Herrian inprimatu elementu bibliografikoen zerrenda (Wikidatako lagina)

Bibliografia

Arana Martija, Jose Antonio (1994): «Euskal Herriko inprentaren historia», *Uztaro*, 37-53. (Kontsulta: 2024-05-26) https://aldizkariak.ueu.eus/index.php/uztaro/article/view/3794

Azurmendi, Joxe (2020): *Pentsamenduaren historia Euskal Herrian*, Jakin - UPV/EHU, Andoain (Leioa).

Berners-Lee, Tim: *5 Star Open Data*. (Kontsulta: 2024-03-01) https://5stardata.info/en/

Bilbao, Jon (1996): *Eusko bibliographia (1981-1985). Vol. 1 (A-B)*, Servicio Editorial de la Universidad del País Vasco/Euskal Herriko Unibertsitatearen Argitalpen Zerbitzua (Kontsulta: 2024-05-24) https://addi.ehu.es/handle/10810/53051?show=full

Bilketa (Kontsulta: 2024-04-12) https://www.bilketa.eus/

Euskariana (Kontsulta: 2024-04-24) https://www.euskariana.euskadi.eus/euskadibib/eu/home/home.do

Fernández de Casadevante Romaní, Mª Dolores (2015): *La imprenta en Guipúzcoa (1585-1850)* (doktoregotesia), Aula Documental de Investigación (ADI).

Humanitate Digitalak: Aukera berriak ikertzen, UEU (jarduera akademikoa) (Kontsulta: 2024-04-12) https://www.ueu.eus/jarduera-akademikoa/1637-humanitate-digitalak-aukera-berriak-ikertzen

Itúrbide Díaz, Javier (2007): *Escribir e imprimir. El libro en el Reino de Navarra en el siglo XVIII*, Gobierno de Navarra / Nafarroako Gobernua, Iruñea/Pamplona.

Koldo Mitxelena Liburutegia (Kontsulta: 2024-05-24) https://www.kmliburutegia.eus/

Lau Haizeetara, Bizkaiko Foru Liburutegia (Kontsulta: 2024-04-12) https://liburutegibiltegi.bizkaia.eus/

Lopez de Luzuriaga Martinez, Iñaki (2023): «Edizioa XVIII. mendeko Euskal Herrian: datu kuantitatiboak, joerak eta inprimategien bilakaera Wikidataren argitan», *Uztaro*, 126, 35-60. (Kontsulta: 2024-05-24) https://doi.org/10.26876/uztaro.126.2023.2

Nafarroako Artxibo Nagusia (Kontsulta: 2024-05-24) https://www.navarra.es/eu/artxiboak

«OpenRefine History», OpenRefine (Kontsulta: 2024-04-12) https://openrefine.org/blog/2013/10/12/openrefine-history.html

«OpenRefine user manual», OpenRefine (Kontsulta: 2024-04-12) https://openrefine.org/docs

«PAWS», Wikitech (Kontsulta: 2024-04-24) https://wikitech.wikimedia.org/wiki/PAWS

Sancho el Sabio, Sancho el Sabio Fundazioa (Kontsulta: 2024-04-12) https://www.sanchoelsabio.eus/eu/

Santoyo, Julio Cesar (1995): *La imprenta en Álava. Historia, obras, documentos. Vol. I. El siglo XVIII* (1995), Fundación Sancho el Sabio, Vitoria.

«UEU-EWKE Humanitate Digitalen Beka 2021», wikimedia.eus (Kontsulta: 2024-04-12) https://wikimedia.eus/2022/02/ueu-ewke-humanitate-digitalen-beka-2021/

«Wikiproiektu: XVIII. mendean Euskal Herrian inprimatu idazkiak», *Euskarazko Wikipedia* (Kontsulta: 2024-05-24) https://eu.wikipedia.org/wiki/Wikiproiektu:XVIII._mendean_Euskal_Herrian_inprimatu_idazkiak

Ikerketa-taldeen eta proiektuen deskribapenak

Hizkuntzalaritza Teorikorako Taldea (HiTT)

Gorka Elordieta, Elena Castroviejo, Azler Garcia-Palomino

Hizkuntzalaritza Teorikorako Taldea (HiTT),
Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)
gorka.elordieta@ehu.eus elena.castroviejo@ehu.eus azler.garcia@ehu.eus

1. Aurkezpena

Hizkuntzalaritza Teorikorako Taldea (HiTT) 1990eko hamarkadako bukaeran sortua da, eta Euskal Unibertsitate Sistemako ikerketa talde egonkortua da 2007tik, betiere A balorazioarekin (2022-2025 eperako identifikazio kodea IT1537-22). Euskal Herriko Unibertsitateko Hizkuntzalaritza eta Euskal Ikasketak Sailari dago atxikia. Bere gunea Arabako Campuseko Micaela Portilla Ikergunean dago, eta helbidea hauxe da:

Micaela Portilla Ikergunea, 3.7
Justo Velez de Elorriaga, 1
01006 Gasteiz

Ikertzaile Nagusia Gorka Elordieta Alcibar da, eta kideak hurrengoak dira (hurrenkera alfabetikoan):

Zuriñe Ábalos Juez, Alejo Alcaraz Tricio, Natàlia Barbarroja Capdevila, Jelena Borise, Elena Castroviejo Miró, Arantzazu Elordieta Alcibar, Azler García Palomino, Mark Jary, Aitor Lizardi Ituarte, Irene Makazaga Núñez, Isabel Martín González, Jon Ander Mendia Aldama, Javier Ormazabal Zamakona, Marina Ortega Andrés, Dennis Ott, Sergio Parrillas Manchón, Valentina Petrolini, Marta Ponciano Lázaro, Antonio Scarafone, Myriam Uribe-Etxebarria Goti, Laura Vela Plo, Agustín Vicente Benito.

Webgunea: http://www.hittlinguistics.eus

HiTT taldeak giza hizkuntza gaitasuna aztertzen du, bi ikerketa-arlo nagusiren inguruan: (a) hizkuntzaren arkitektura eta bere ezaugarri formalak; eta (b) hizkuntzaren eta beste sistema kognitibo batzuen arteko harremana. Jatorriz hizkuntzalaritza formalean murgildu da, ikuspegi edo jarduera esperimentaletik ere (batez ere fonetika eta fonologian). Halaber, azken urteotan Hizkuntzaren Filosofiako ikertzaile talde aktibo bat batu da hizkuntzaren eta kognizioaren arteko harremanaren inguruan ikertzeko, eta Autismo Araba Elkartearekin hasitako lankidetzarekin, nabarmen zabaldu da taldearen diziplinartekotasuna eta ikerketa-lerro berrien garapena jakintza-arlo teoriko zein esperimentaletan.

2. Ikerketa lerroak

2.1. Populazio neurotipikoa

Hasieratik, HiTT hiztun neurotipikoen hizkuntza gaitasunaren ikerkuntzan erreferente izan da. Hizkuntzaren arkitekturarekin eta hizkuntzaren moduluen arteko interfazearekin lotutako gaiak landu ditu batez ere, besteren artean:

a) *(Morfo)Sintaxian:* argumentu egitura eta haren oinarri sintaktikoa (Kasu Teoria eta A-mugimendua); denbora-harremanen gramatika; ezeztapena; konparazio egiturak; A-marra mugimendua; perpaus nagusien eta mendeko perpausen hitz hurrenkeren arteko asimetria; elipsia; konplementatzaileak eta euskararen aditz jokatuaren lekua; ezker periferia eta diskurtsoaren egitura.
b) *Fonologian:* euskarazko, gaztelaniazko eta beste hizkuntza erromantze batzuetako intonazioaren ezaugarriak eta egitura prosodikoa; informazio-egituraren prosodia; hizkuntz-ukipena eta intonazioa.
c) *Semantikan.* Galde- eta harridura-perpausak; modalitatea, gradabilitatea eta ebaluatibitatea; informazio- eta diskurtso-egituren arteko korrespondentzia; esanahi motak; hizkera gutxiesgarria; polisemiaren azterketa eta harekin lotutako gaiena (ko-predikazioa, esaterako).

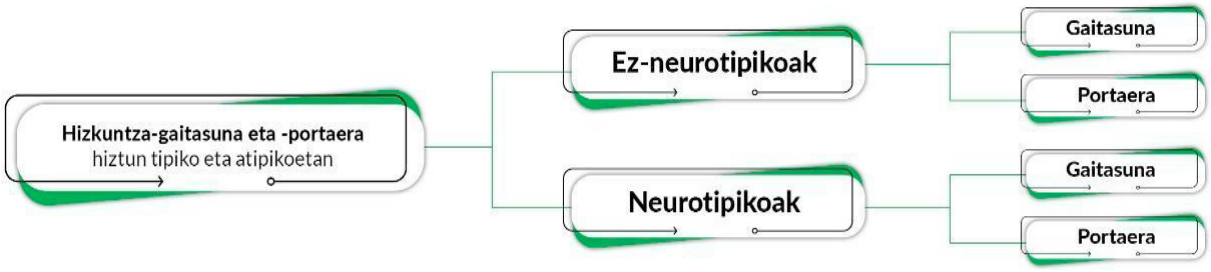
Azken urteotan, halaber, hiztun neurotipikoen hizkuntza portaeraren ikerkuntzan aritu gara HiTTen. Hizkuntza-aldaeren arteko ukipenezko fenomenoak eta beste hainbat faktore sozialen araberako aldakortasuna aztertzen dihardugu, orain arte fonetika-fonologian bereziki. Adibidez:

a) Hego Euskal Herriko gaztelanian dauden euskarazko intonazioaren ezaugarriei erreparatu diegu. Ezaugarri horiek zenbait faktore sozialekin lotuta daude, euskararekiko kontaktua zein estua den, esate baterako, edota hiztun horiek euskararekiko eta euskal talde etnolinguistikoarekiko dituzten jarrerekin.
b) Euskarazko hizkera ugariren gaineko lanak ere egin ditugu aldakortasun soziolinguistikoa zeren araberakoa den deskubritu nahian. Batetik, norbere herriarekiko atxikimenduak hizkera supralokaleko moldeak hartzean eragina daukala ikusi dugu. Eta bestetik, konturatu gara nagusien belaunaldian lanbide zehatz batzuetan zebiltzanen ahotan gehien entzuten ziren formek herriko nortasuna adierazteko funtzioa hartu ahal izan dutela gazteentzat.

2.2. Populazio ez-neurotipikoa

Language in Neurodiversity Lab (Lindy Lab: www.lindy-lab.eus) izenpean Autismoaren Espektroko Nahasmena (AEN) nahiz Mintzamenaren eta Hizkuntzaren Garapen-Nahasmendua (MHGN) daukaten hiztunengan zentratu gara. Lehen populazio mota Autismo Araba Elkartearekin elkarlanean ikertu dugu. Zehazki, gramatikaren eta hiztegiaren jabekuntzan zein faktore kognitibok eragiten duten erantzuten ahalegintzen gara. Ikertzen ditugun lerroen artean aipagarria da AEN daukaten ume eta helduen gaitasun pragmatikoen gainekoa, bereziki irudizko hizkuntza edo hizkuntza metaforikoa zelan interpretatzen den, berbaldia zelan egituratzen den eta inplikaturak zelan ateratzen diren.

Irudi honek gure ikertaldearen ikerlerroak erakusten ditu labur-laburrean.



1. irudia. HiTTen ikerlerroak

3. Hizkuntza-teknologiarekiko erlazioak

CLARIAH-EUS bezalako egitasmoak onura ekarriko dio HiTT-i, gure ikerketarako baliagarriak izan daitezkeen hizkuntza-teknologiak ezagutzeko eta erabiltzen ikasteko. Hizkuntza-teknologiako tresnek arlo askotan lagundu diezaguketelakoan gaude. Momentu honetan esku artean ditugun ikerketetako datuen analisiak automatizatzea beharko genuke, bi abantaila dira bistakoak: batetik, analisiak eta kalkuluak egiterakoan gizakiok egin ditzakegun akatsak ahalik eta gehien saihestea, eta bestetik, datuak analizatzeko denbora aurreztea. Hurrengo lerroetan, behar batzuk aipa ditzakegu, adibide zehatzak ematearren. Esan bezala, euskara da ikertzen dugun hizkuntza nagusietariko bat, eta badakigu euskara hainbat arlotan analizatzeko softwarea existitzen dela, baina guk dakigunez euskara estandarrerako ematen dituzte emaitza onak, ez horrenbeste beste barietate batzuetarako. Hor dago koska, ea nola moldatu edo egokitu software hori aldaera ez-estandarretarako.

a) Audiozko artxiboak era automatikoan ortografikoki transkribatzeko programak. Existitzen dira horretarako programak, Whisper esaterako[1], baina euskara estandarrerako funtzionatzen badute ere, euskara mota ez-estandarretarako arazoak aurkitzen dituzte.
b) Audiozko artxiboak era automatikoan fonetikoki transkribatzeko programak.
c) Idatzizko testuetako grafemen eta ikur fonologiko edo fonetikoen traskripzioa. Grapheme to phoneme.
d) Audiozko artxiboen segmentatze automatikoa (esalditan, hitzetan, silabatan, edo fonotan) egiteko programak: Montreal Forced Aligner (McAuliffe *et al.,* 2017), MAUS (Schiel, 2015), edo beste batzuk. Goian aipatu bezala, euskara estandarra ez den beste euskalki edo hizkera batekin erabili daitezkeenak behar ditugu.
e) Euskararen intonazioaren transkripzio automatikoa egiteko programak, ToBI sisteman. Gaztelaniarako edo katalanerako existitzen dira horrelako programak, adibidez, Eti_ToBI (Elvira-García *et al.,* 2016), baina agian hortik abiatu gaitezke euskarara egokitzeko.
f) Euskara estandarrekoak ez diren aldaeretako idatzizko/ahozko korpusetatik datuak automatikoki jasotzeko interfazeak (morfosintaktikoki eta informazio-egituraren arabera etiketatuak).
g) Euskarazko barietate ez-estandarretako korpusak anotatzeko programak, bilaketa automatikoak egin ahal izateko eta datuak estatistikoki aztertu ahal izateko.
h) Testua edo audioa duten *on line* esperimentu linguistikoak egiteko software intuitiboa (adib., aukera anitzetatik zuzena dirudiena hautatu, edo gramatikaltasunaren inguruko eskalazko ebaluazioak).

[1] https://huggingface.co/deepdml/whisper-small-eu

Dena den, hizkuntzalaritza eta hizkuntza-teknologiaren arteko erlazio onuraduna bi norabideetakoa izan beharko litzateke, hots, hizkuntzalaritzatik hizkuntza-teknologiara eta alderantziz. Gure iritziz, hainbat gertaera linguistikoren ezagutza enpirikoa sakonduz orokortze sendoak eskaini ditzakegu gero hizkuntza-teknologiako ikerketa taldeek aplikatu ahal izateko.

Ahoskerari dagokionez, hizkuntza-forma jakin batzuk nork, zelan eta zein testuingurutan esaten dituen jakiteak lagunduko luke transkribatzaileak ezaugarri sozialen arabera moldatzen eta ahotsen profila zehatzago identifikatzen. Kasurako, Euskal Herriko mendebaldeko txistukari-sistemaren berri daukagun arren, oraindik azterketa akustiko sakonak dauzkagu faltan, batez ere sistema arin aldatzen dabilelako zenbait zonaldetan, adinean behera egin ahala zein hiztunaren sexuaren arabera. Hiztuna zein sexutakoa edo EHko zein zonaldekoa den jakinez gero, transkribatzaile automatikoa entrenatzeko aukera legoke hizkuntza-forma jakin batzuk espero izateko transkribatzerako orduan. Eta alderantziz, hizkuntzalaritza forenserako adibidez, grabatutako hiztun baten profila identifikatu nahi izanez gero, hiztunaren informaziorik gabe, aldez aurretik hizkuntza-forma jakin batzuk nork esaten dituen jakiteak profila arinago eta zehatzago ateratzen lagunduko luke.

Beste horrenbeste egin daiteke AEN edo MHGN daukatenen hizkerarekin, sarritan hainbat egitura ez-kanoniko erabiltzen baitituzte. Hori guztia sistematizatzeko oinarri teorikoa sortzen lagundu diezaieke hizkuntzalaritzak hizkuntza-teknologiei.

Hizkuntza-teknologian adituak ez izanik, CLARIAH-EUSek aukera paregabea ematen digu HiTT taldeari gure ikerketarako mesedegarriak diren baliabideez ikasteko. Ikusi beharko litzateke laguntza nola bideratu daitekeen; agian HiTTekoen beharrak eztabaidatu ondoren ikerlari bat beren-beregi trebatu hizkuntza-teknologia zehatz horietan? Edo gure taldeko ikerlariak egokiak diren programak eta baliabideak erabiltzen ikastea, formakuntza saioetan? Ala laguntza eskaera puntualak adierazi? Edozein kasutan, esan bezala, guk kolaboraziorako aukerak ikusten ditugu. CLARIAH-EUSen barruan gure ezagutza zientifikoa arlo batzuetarako lagungarria izan daitekeela pentsatzen bada, gu prest gaude hori eskaintzeko.

Esker onak

Bi ebaluatzailek egindako oharrak eskertzen ditugu. Bestalde, lan hau hurrengo proiektuen eta dirulaguntzen barruan sartzen da:

— IT1537-22 (Eusko Jaurlaritza).
— PID2021-128511NB-I00, PID2021-122233OB-I00 (MICIU/AEI/10.13039 /501100011033 eta FEDER/UE).
— PES24/08 (UPV/EHU).

Bibliografia

Elvira-García, W., Roseano, P., Fernández Planas A. M., eta Martínez Celdrán, E. (2016), A tool for automatic transcription of intonation: Eti-ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation*, 50(4),767-792.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. eta Sonderegger, M. (2017), Montreal Forced Aligner. https://montreal-forced-aligner.readthedocs.io/en/latest/

Schiel, F. (2015). A statistical model for predicting pronunciation. In M. Wolters, J. Livingstone, B. Beattie, R.l. Smith, M. MacMahon, J. Stuart-Smith & J. Scobbie (Eds.,) *Proceedings of the ICPhS 2015*. University of Glasgow.

TRALIMA-ITZULIK[1]

Zuriñe Sanz-Villar, Elizabete Manterola

TRALIMA-ITZULIK, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

zurine.sanz@ehu.eus elizabete.manterola@ehu.eus

1. Aurkezpena

TRALIMA/ITZULIK ikerketa-taldearen egitekoa kulturen arteko itzulpena eta hartu-emana ikertzea da, nola literatura testuetan hala ikus-entzunezko produktuetan, ikuspegi historikoa eta deskribatzailea erabilita, eta teknologia berrietako tresnez baliatuz (Humanitate Digitalak). Hala, orain arte landu ditugun ikerketa-lerroen ardatza izan da idatzizko nahiz ikus-entzunezkoetako itzulpenek agertzen dituzten mekanismo linguistikoak, kulturalak eta sozialak aztertzea. Ikerketa horiek ikuspegi historikoa erabilita egin ditugu, metodologia deskribatzaile komun batez baliatuta, eta teknologia berriak garatu eta aplikatu ditugu. Hain zuzen ere, lan egiteko modu hori bihurtu da taldearen ezaugarri nagusi. Horiek horrela, ikerketa-taldearen ikerlerro nagusiak honako hauek dira: corpusetan oinarritutako itzulpen ikasketak, euskaratik eta euskararako itzulpena, literatura itzulpena, ikus-entzunezko itzulpena eta irisgarritasuna, itzulpengintzaren didaktika, eta itzulpengintza eta hizkuntza-teknologiak. Ekarpen honetan, ikerlerro horien testuinguruan ikerketa-taldeak burutu dituen ekimenak eta abian dituenak aurkeztuko dira, euskarazko datu digitalekin lan egiten dugun talde gisa ditugun beharrak edo gabeziak azpimarratuz.

2. Proiektuak

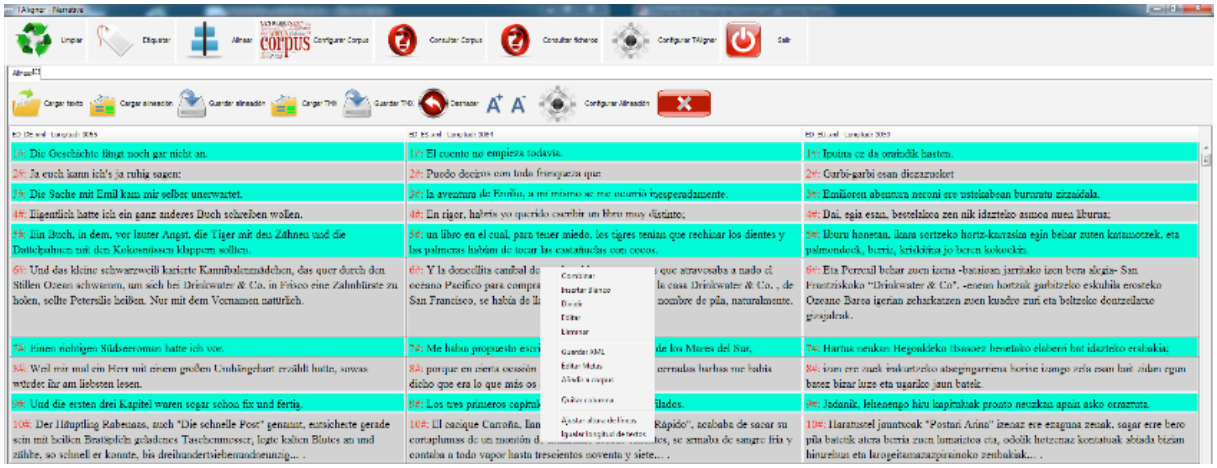
Corpusetan oinarritutako itzulpen ikasketa deskribatzaileak (Toury, 1995) izan dira taldean egindako ikerketa-lan askoren oinarri teoriko eta metodologikoa (Andaluz-Pinedo, 2022; Arrula, 2019; Manterola, 2012; Pérez López de Heredia, 2003; Ros Abaurrea, 2023; Sanz-Villar, 2015; Zubillaga 2014). Testuinguru horretan lan egiteak corpusak sortzea eskatzen du, corpus digital, eleaniztun eta paraleloak. Hutsetik sortu behar izan ditugu guztiak, hala nola: euskaratiko itzulitako literatura testuen corpusa (Manterola, 2012), alemanetik euskararako literatura testuak —helduei zein haurrei zuzenduak— biltzen dituen corpusa (Sanz-Villar, 2015; Zubillaga, 2014), euskaratik gaztelaniara eta frantsesera itzuliriko literatura testuen corpusa (Arrula, 2019), ingelesetik gaztelaniarako antzezlanen corpusa (Andaluz-Pinedo, 2022) eta Leonard Cohenen abestien gazte-

[1] **Taldekideen izen-abizenak, ordena alfabetikoan:** Marta Iravedra, Marlén Izquierdo, Elizabete Manterola Agirrezabalaga (IN), María Pérez López de Heredia, Alejandro Ros Abaurrea, Zuriñe Sanz Villar, Ana Tamayo, Naroa Zubillaga.

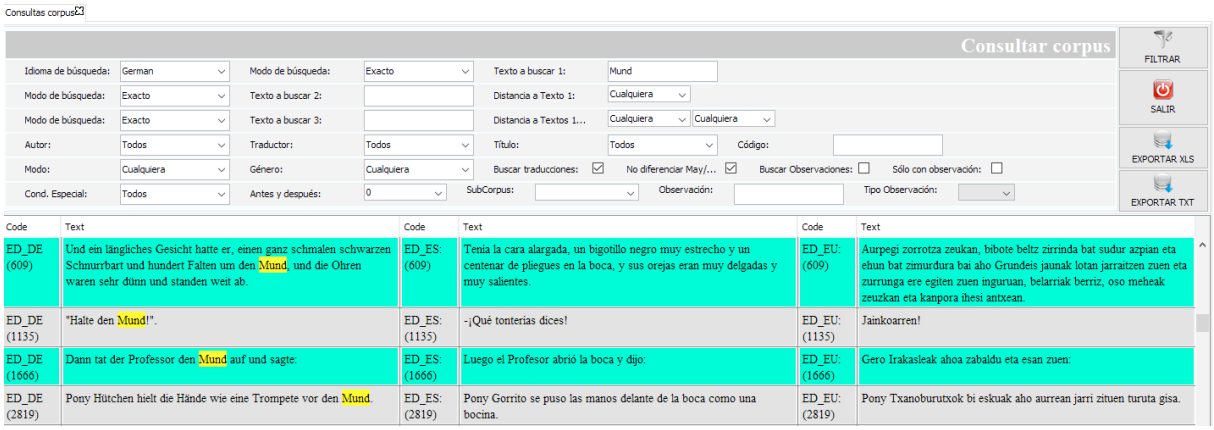
laniarako itzulpenen corpusa (Ros Abaurrea, 2023). Erabilitako tresna, TAligner, taldean bertan sortutakoa da, informatikari eta itzulpengintza alorreko ikertzaileen elkarlanari esker, eta corpus eleaniztun paraleloak sortzeko eta ustiatzeko aukera ematen du.

Honakoak ditu ezaugarri nagusi:

— Erraz abia daiteke, eta erabiltzaileek ez dute laguntza teknikorik behar.
— Gainera, hainbat motatako testuak biltzen dituzten corpusak sor daitezke.
— TAligner-en bitartez hainbat testu lerroka daitezke aldi berean, ikertzaileak bere lanerako behar dituen adina (1. irudian ikus daitekeen legez).
— 2. irudiak erakusten duen moduan, sortuko den corpusa tresnan bertan ustia daiteke; hau da, bilaketak tresnan bertan egin daitezke



1. irudia. TAligner-en hiru testu aldi berean lerrokatzen



2. irudia. Alemanezko *Mund* hitzaren bilaketa TAligner-en ta bilaketa horren emaitza batzuk

Bai ADDIn[2] bai CorpusNet izeneko webgunean[3] deskarga daiteke TAligner-en azken bertsioa.[4] Etorkizunera begira, alde batetik, tresna bera garatzen jarraitzeko, ezinbestekoa da aipatu

[2] https://addi.ehu.es/handle/10810/42445?locale-attribute=eu

[3] http://corpusnet.unileon.es/

[4] «CorpusNet – Baliabide eleaniztunak eta haien aplikazioak» izeneko estatu mailako bikaintasun sareko partaide izan zen TRALIMA/ITZULIK (MINECO proiektua, FFI2016-81934-RED).

elkarlana, informatikari eta itzultzaileon artekoa. Bestalde, dagoeneko sortuta dauden corpusak sareratzea (COVALT ikerketa-taldeak egin duenaren ildotik[5]) etorkizuneko beste helburu bat da.

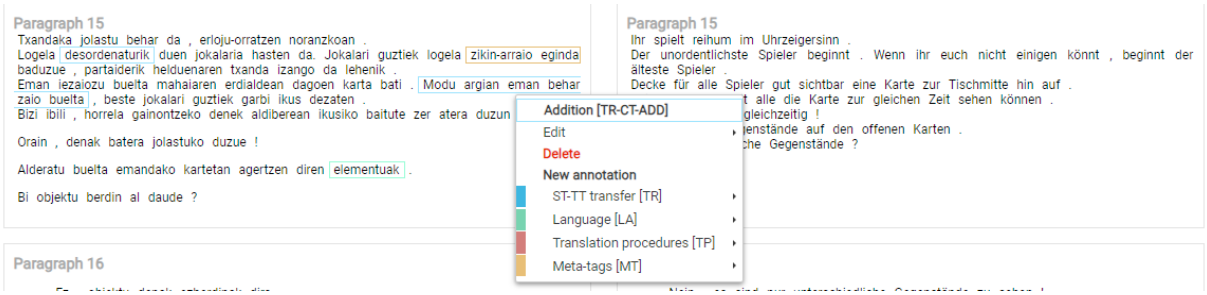
Itzulpengintzaren didaktikaren alorrean, Leuven-go Unibertsitateko Sylviane Grangerrek eta Marie-Aude Leferrek abiarazitako MUST (Multilingual Student Translation) nazioarteko proiektuko kide gara TRALIMA-ITZULIK ikerketa-taldeko hainbat kide. Hypal4MUST plataforma digitalean ikasleek sortutako itzulpenez osatutako corpus paraleloak sortzeko eta ustiatzeko aukera ematen digu proiektu honek, eta, gaur-gaurkoz, lau hizkuntza-konbinaziotan sortu ditugu corpusak: ingelesa-euskara, ingelesa-gaztelania, euskara-gaztelania eta alemana-euskara.

Language pair statistics for Universidad del Pais Vasco within MUST project

Pair	Total texts	NEW texts	PROGRESS texts	FINISHED texts	ST tokens	TT tokens	Active students	Annotations
en>es	65	48	0	17	22691	26307	0	399
eu>es	53	0	0	53	20275	29875	0	531
de>es	0	0	0	0	0	0	0	0
en>eu	122	38	0	84	49813	40942	16	1534
de>eu	104	36	0	68	38240	33692	9	1525

3. irudia. MUST proiektuaren baitan sortutako corpusak

Anotazio sistema (TAS 2.0[6]) ere sortu dute proiektuko zuzendari eta ikertzaileek, eta horrek esan nahi du corpusak sortzeaz gain ikasleen itzulpenak anotatzeko aukera ere badugula.



4. irudia. Anotazioa Hypal4MUST plataforman

Anotatu ditugu corpus batzuk eta anotazioaren emaitzak kongresuetan eta artikuluetan argitaratu ditugu (Sanz-Villar, 2024). Kongresu horietako bat PaCor 2021 nazioarteko sinposioa izan zen, corpus paraleloen ingurukoa eta TRALIMA-ITZULIKeko kideek 2021eko ekainean antolatua.[7]

Nazioarteko proiektu honek eskaintzen duen azpiegitura digitalak ikasleen itzulpen corpus (*Learner Translation Corpus*) anotatuak sortzeko aukera eskaintzen digu. Baina, aldi berean, ez denez guk geuk kudeatzen dugun egitasmo bat askatasuna mugatua da, hizkuntza-konbinazio jakin batek eskatzen dituen beharretara egokitu nahi badugu. Esaterako, hainbat hizkuntzatan testuak POS mailan etiketatzeko aukera eskaintzen du tresnak, baina euskararako ez dago aukera hori. Gainera, corpusa esportatu eta Sketch Engine-ra inportatzeko zailtasunak ditugu, tresna hori ordainpekoa izanik.

---

[5] Ikus http://cwbcovalt.xtrad.uji.es/cqpweb/

[6] https://cdn.uclouvain.be/groups/cms-editors-cecl/cecl-papers/TAS-2.0_annotation_manual_2021-10-26.pdf

[7] Kongresutik eratorritako argitalpena berriki argitaratu da (Izquierdo eta Sanz-Villar, 2023).

Hizkuntzen teknologien alorrean, Tamayok eta Ros Abaurreak (2024) Elhuyar Fundazioaren hizketa-ezagutzaileak, Aditu-k, sortzen dituen euskaratik euskararako azpidatzi automatikoak aztertu dituzte, eta, Elhuyar Fundazioarekin lankidetzan sortu den beste proiektu baten testuinguruan, Elia itzultzaile automatikoak euskara-gaztelania hizkuntza konbinazioan sortzen dituen emaitzak aztertzen ari gara generoaren ikuspegitik, genero alborapena baita tresna automatiko horiek piztu duten kezketako bat. Horretarako, testu-mota ezberdinez osatutako corpusak osatzen ari gara, eta literaturakoa izan da osatzen arazotsuena, egile-eskubideak medio. Gure ustez, alor horretan dauden mugak ondo ezagutzeko sare bat sortzeko aukera eskain dezake CLARIAH-EUS ekimenak.

3. Laburbilduz

Gako hitz bat aukeratu beharko bagenu, diziplinartekotasuna izango litzateke. Izan ere, diziplina arteko elkarlan hori gabe hemen aurkeztu ditugun proiektu asko ez ziratekeen bere horretan gauzatuko. Sinetsita gaudenez elkarlana dela bidea, CLARIAH-EUS ekimenaren baitan, arlo ezberdinetako kideekin lan egiteko prestutasuna da guk eskaintzeko daukaguna, bai guk proposatutako ekimenen testuinguruan (hala nola corpusak sareratzea, TAligner garatzen jarraitzea, egile-eskubideen auziari heltzea), bai itzulpengintzari eta humanitate digitalei lotutako beste proiektu batzuen baitan.

Bibliografia

Andaluz-Pinedo, O. (2022). *Traducciones teatrales (inglés-español) desde la censura franquista hasta el siglo xxi: análisis del corpus TEATRAD* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/56092

Arrula, G. (2019). *Autoitzulpenaren teoria eta praktika Euskal Herrian* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/27983

Izquierdo, M. eta Sanz-Villar, Z. (2023). *Corpus Use in Cross-linguistic Research*. John Benjamins.

Manterola, E. (2012). *Euskal literatura beste hizkuntza batzuetara itzulia. Bernardo Atxagaren lanen itzulpen moten arteko alderaketa* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/12382.

Pérez López de Heredia, M. (2003). *Traducciones censurales de teatro norteamericano en la España de Franco (1939-1963)*. [doktorego tesia, UPV/EHU].

Ros Abaurrea, A. (2023). *Música y traducción: Leonard Cohen en la cultura de España* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/64177

Sanz-Villar, Z. (2015). *Unitate fraseologikoen itzulpena: alemana-euskara. Literatur testuen corpusean oinarritutako analisia* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/15128

Sanz-Villar, Z. (2024). German-into-Basque Translation Analysis of Multiword Expressions in a Learner Translation Corpus. *Ikala*, *29*(1), 1-21.

Tamayo, A. eta Ros Abaurrea, A. (2024). Speech-to-text Recognition for the Creation of Subtitles in Basque: An Analysis of ADITU Based on the NER Model. *The Journal of Specialised Translation*, *41*, 48-73.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. John Benjamins.

Zubillaga, N. (2014). *Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez* [doktorego tesia, UPV/EHU]. ADDI. https://addi.ehu.es/handle/10810/12431

Behategia
Euskarazko komunikabideen audientzia azterketarako datu zientzia

Libe Mimenza Castillo, Naroa Burreso Pardo, Ane Martinez Juez, Hibai Castro Egia, Josu Amezaga Albizu

NOR Ikerketa Taldea, Euskal Herriko Unibertsitatea (UPV/EHU)
libe.mimenza@ehu.eus naroa.burreso@ehu.eus ane.martinezj@ehu.eus
hibai.castro@ehu.eus josu.amezaga@ehu.eus

Euskal Hedabideen Behatokia (*behategia.eus*) euskarazko komunikabideen ikerketarako espazioa da. Behategia 2016an sinatutako hitzarmen batetik abiatu zuten Hekimen elkarteak, Euskal Herriko Unibertsitateak, Deustuko Unibertsitateak, Mondragon Unibertsitateak eta Udako Euskal Unibertsitateak. Hedabideen egungo erronkei erantzuteko ikerketak koordinatzen dira bertan, eta horien artetik NOR Ikerketa Taldeak bideratzen dituen lanek ikerkuntza digitala dute ipar.

Euskarazko komunikabideak eta komunikazioa ikergai gisa hartuta, hedabideen kontsumoa eta erabiltzaileak ezagutzeko xedez planteatutako ikerketek osatzen dute NOR Ikerketa Taldearen Behategiko analisi eremua: gizarte zientzietako audientzia azterketa klasikoak datu zientziarekin gurutzatzeko urratsetan hasita honezkero, eta bide horretan gehiago sakontzea izanik ikertzaileen motibazio, diziplina arteko begiradak eta *big data* uztartuko dituen ikerketa corpusaren sorkuntza zein analisia dira helburu.

Jarraian azaltzen dira abian diren hiru proiekturen nondik norakoak:

1. BEHA

   Euroeskualdeko (2021, 2022) hiritartasuna deialdiari esker abiatutako **BEHA ikerketa-lerroak** euskarazko hedabideetan datuen kultura sustatzea du helburu. Marko orokor horri bultzada bat emateko, 2024tik aurrera azterketa kuantitatiboa eta kualitatiboa konbinatuta, datu kulturaren gako-adierazleak identifikatuko dira euskal hedabideen egoera ebaluatzeko; eta behin diagnostikoa ezagututa, **datu kultura bultzatzeko baliabideak** eskainiko zaizkie komunikabideei —hala formazioa nola datuen analisiak—.

   Azken hiru urteetan, halaber, datu analitikoen bilketarako tresnak diseinatu eta garatu ditugu. Lehenengo mugarri gisa, duela lau hilabete euskal hedabideen **web-analitika datuen jarraipena ahalbidetzen duen panel interaktiboa** (BEHA panela) jarri zen martxan —60 komunikabidek haien Google Analytics tresnak biltzen duen web trafikoa nahieran kontsultatzeko eskura duten panela, eta hilero-hilero modu estandarrean txosten forman jasotzen hasi direna— (Mimenza, 2023). Abiapuntua hori izanda, software librean oinarritutako garapen propioko **ingurune digital osoko panela** sortzea da hurrengo urratsa (Behategia Analytics), zeinetan, web-analitikaz gain, sare sozialen trafikoa ere neurtuko den —hedabideek haien ingurune digital osoaren jarraipena egiteko analitika tresna izango da, eta euskal komunikabideen audientzia azterketa

digitala egitea ahalbidetuko du—. Bi tresna horien garapen-, mantentze- eta eguneratze-lanez gain, Behategiaren erronketako bat da **denbora errealeko trafiko-datuen jarraipena** ahalbidetuko duen panela diseinatzea.

Halaber, datu kultura sustatzeko ikerketa-lerro honen indargune da komunikabideen arteko lankidetza espazioa, eta datuen partekatzean oinarritzen den ezagutza kolaboratiboa.

2. D^i—Datu integralak

   Jose Ignacio Ruiz Olabuenaga ikerketa-bekari esker (Burreso *et al.*, 2023), Behategian analisi matematikoa txertatzea lortu da **D^I—Datu integralak** proiektuaren bitartez. Ikerketa-lan horretan, datu-base ezberdinak **fusionatzeko metodologia** bat garatu da, eta metodologia horren aplikaziorako pilotu gisa *Estudio de audiencia de medios* (CIES) eta Ikusiker panelaren (Ikusentzunezkoen Behategia) datuak fusionatzea lortu da. Egun, bidean dago CIES eta Inkesta Soziolinguistikoaren datuak bilduko dituen fusioa eta aurrera begira Behategia Analytics eta CIES fusionatzea da erronka.

   Fusioez gain, Ikasketa Automatiko (*Machine Learning*) tekniken erabilpenak **audientzien azterketa analisi prediktibo** eta **preskriptiboaren** bitartez egitea ahalbidetu du. Horren adibide dira Hegoaldeko hedabide tradizionalen kontsumitzaileen profilen *clustering*-a, euskarazko hedabideen kontsumoa aurresaten duten modelo prediktiboen eraikuntza eta hedabide horien kontsumoan eragiten duten aldagaien identifikazioa (Burreso, 2023).

3. Komunikazioaren azterketarako Datutegia

   Datutegia (Behategia, d. g.) euskarazko komunikazioaren inguruko datuak batzen dituen **datu-bilduma** da: adierazle soziodemografikoak, ekoizpen datuak, audientzia datuak edota euskal hedabideen direktorioa ditu kontsultagai egun. Datu-bilduma hori osatzea eta eguneratzea da erronka, corpus hori analisi gurutzatuetarako baliatzea eta baita ezagutza-iturri izango den datu-base erabilgarri gisa konfiguratzea ere (Amezaga, 2022; Martinez, 2023).

Proiektu guztien atzean asmo argia dago: komunikazioari buruz modu irekian zein hedabideen eskura dauden datu-multzo handiak euskarazko produktuen kontsumoa sustatzeko baliatzea; bidean euskal hedabideen sektoreari baliabideak emanez, hizkuntza politiken diseinuan lagunduz, eta maila akademikoan ezagutzan sakonduz.

Behategiak CLARIAH-EUS azpiegituran parte hartzeko duen interes nagusia diziplina ezberdinetako eragileekin sare bat sortzea da, kideen artean baliabideak eta tresnak partekatzeko eta diziplinarteko lana bultzatzeko. Behategiak, bere aldetik, eskura dituen baliabideak eskaini nahi dizkie azpiegituraren parte diren gainerako kideei. Alde batetik, *www.behategia.eus* web orrian eskuragarri dauden komunikazioaren azterketarako Datutegia eta urtero argitaratzen duen Euskal Hedabideen Urtekaria partekatu nahi ditu. Bestalde, ikerketa-taldeak egindako proiektuetan lortutako emaitzen inguruko informazioa, txostenetan jasota edo aurrez aurre zein online aurkeztuta ere azpiegiturako kideen esku jarri nahi ditu. Azkenik, eskaintza eta eskari forma ere baduen helburua dugu: CLARIAH-EUS sarea aukera izan daiteke proiektu ezberdinetan —bai martxan direnetan, bai batuta abiatu ditzakegunetan— elkarrekin lan egiteko.

Bibliografia

Amezaga, J. (2022). Audientziak aztertzeko sistema: Burutzeke dugun erronka. *Euskal Hedabideen Urtekaria 2021*, 147-159 or.

Behategia (d. g.). *Euskal hedabideen datutegia* [dataset]. https://datutegia.behategia.eus/

Burreso, N. (2023). *Técnicas de ML para el análisis de consumo de medios de comunicación tradicionales en Euskadi y Navarra* [Trabajo Fin de Máster]. Universidad Internacional de Valencia.

Burreso, N., Amezaga, J., Mimenza, L., Arana, E., & Barrio, I. (2023). *Komunikabideen audientzia ikerkuntzarako metodologia berriak esploratzen: Datuen fusioa oinarri duen azterketa aplikatua / Explorando nuevas metodologías de investigación de audiencias: Un análisis aplicado basado en la fusión de datos.* Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.

Euroeskualdea (2021). «*Euroeskualdeko hiritartasuna» 2021ko proiektu deialdiaren ebazpena*. Eurorégion Nouvelle-Aquitaine, Euskadi, Navarre. https://www.euroregion-naen.eu/eu/artikuluak/euroeskualdeko-hiritartasuna-2021ko-proiektu-deialdiaren-ebazpena/

Euroeskualdea (2022). «*Euroeskualdeko Hiritartasuna» 2022 proiektu deialdiko emaitzak*. Eurorégion Nouvelle-Aquitaine, Euskadi, Navarre. https://www.euroregion-naen.eu/eu/artikuluak/euroeskualdeko-hiritartasuna-2022-proiektu-deialdiko-emaitzak/

Martinez Juez, A. (2023). *Euskarazko edukien kontsumoa eta gune soziolinguistikoak: Bi aldagaien arteko harremana fokuan* [Gradu Amaierako Lana]. Euskal Herriko Unibertsitatea/Universidad del País Vasco (UPV/EHU).

Mimenza, L. (2023). *Euskal hedabideen ingurune digitala neurtzeko analitika: Diagnostikoa eta proposamena / Data analytics to measure the digital environment of Basque media: Diagnosis and proposal* [Doktorego tesia]. Euskal Herriko Unibertsitatea/Universidad del País Vasco (UPV/EHU).

Gizapedia.org: Giza eta Gizarte Zientzien euskarazko entziklopedia

Josemari Sarasola Ledesma[1], Eneko Sarasola Telleria[2]

[1] EHUko Ekonomia eta Enpresa Fakultatea, Donostiako atala
josemari.sarasola@ehu.eus

[2] hirusta.io
info@gizapedia.org

Hasiera batean ikasleei euskarazko edukiak eskaintzeko asmoarekin, 2016ko bukaeran arlo akademikoan diharduen lagun-talde baten eskutik sortua, Gizapedia giza eta gizarte zientzien arloan kalitatezko artikuluak biltzen dituen entziklopedia irekia da. 2024ko urtarrilean, euskaraz 6.000 artikulu baino gehiago biltzen ditu hainbat alorretan, hala nola antropologian, soziologian, filosofian, ekonomian, estatistikan, informatikan eta hizkuntzalaritzan. Euskaraz kalitatezko eduki akademikoa sortzea da gure eguneroko erronka, Internetetik informazio guztiontzat irisgarri eta irekia, zehatza eta, behar denean, sakona ere ematea, bereziki ikasle unibertsitarioei begira. Wikipedia ez bezala, ez da edonork idatzi dezakeen entziklopedia. Lan-taldea irekia da baina parte hartu nahi duena aditua edo aritua bada izan behar da landu nahi duen arloan.

Irisgarritasunari begira eta bilaketak errazteko, terminologia bereziki lantzen dugu, kontzeptu bakoitza termino egoki baina baita ere erabilienen bitartez izendatzen. Artikulu bakoitzaren hasieran definizio zehatza eta ulergarria ematen dugu, gero kontzeptua sakonago aztertuz eta beste kontzeptuekiko loturak zehaztuz, testu barneko hiperesteken bitartez eta gainera artikulu bakoitzarekin loturiko artikuluen estekak proposatuz. Argigarriak direla kontsideratzen dugunean, testua laguntzen duten irudiak txertatzen ditugu. Artikuluak iturri akademikoak kontsultatuz idazten dira orokorrean, eta lan-taldearen baitan berrikuspenak egiten dira horien kalitatea bermatzeko.

Artikuluak jakintza-arloaren arabera sailkatzen dira, haiei dagozkien kategorietan banatuta. Halaber, erabiltzaile guztiek Gizapediako lankideekin interakzioa izateko mekanismoak dituzte webgunean: bilatzaile bat, iradokizunak egiteko esteka eta harremanetarako kontaktu zuzena. Eskaera gehienak artikulu berriak osatu eta jadanik zeuden artikuluak sakontzeko izaten dira. Ahaleginak egiten ditugu eskaera horiek lehenbailehen betetzeko.

Aurkezpen zabala behar duten gaietarako, ikasliburuak ditugu, sarrera baten ondoren, gaia ikasgai moduan jorratzen dutenak, azalpen zabalen bitartez, adibide eta ariketekin, PDF dokumentu batean bilduak. Ikasliburuak ikastaro direlakoetan biltzen dira. Adibidez, estatistika alorreko ikastaroek milaka jarraitzaile izan dituzte urteotan zehar (ikus, https://gizapedia.org/kategoria/estatistikako-ikasliburuak).

Egun, Gizapediak milaka erabiltzaile ditu egunero. 1. taulako datuek erakusten dutenez, ikasleak dira erabiltzaile nagusiak, ikusita uztailean eta abuztuan gertatzen den bisiten beherakada,

baina ikertzaileen aldetik gradu eta master amaierako lanak, tesiak eta monografien erreferentzia gisa ere erabili da *google.scholar* bilatzaile akademikoan ikus daitekeenez (https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=gizapedia&btnG=). Google, Bing eta beste bilatzaileetan ere puntako kokapenetan agertzen dira bere artikuluak.

1. taula. Gizapedian izandako bisitak 2023 urtean, hilabeteka

Hilabetea	Bisitak
Urtarrila	192.819
Otsaila	159.111
Martxoa	183.112
Apirila	177.309
Maiatza	182.439
Ekaina	157.753
Uztaila	63.808
Abuztua	83.335
Iraila	154.303
Urria	261.228
Azaroa	256.373
Abendua	115.057

Esan bezala ikasleak nagusi diren erabiltzaileengandik jasotako eskaerak direla eta, gaztelaniazko artikuluak sartzen ere hasi ginen 2020an, pandemia garaian: gaztelaniazko edukiek hartu duten bolumena eta bisita kopurua ikusita, beste webgune bat prestatu dugu (ikusmira.org) haiek guztiak hara eraman eta gaztelaniazko hiztunen eskura jartzeko, Gizapedia euskal entziklopedia gisa atxikitze aldera, egungo 6.425 artikulurekin (2024ko apirilak).

Gizapediako hiztegia ere osatu dugu, Hizkailua izenekoa (hiztegia.gizapedia.org/). Hiztegi egituratua da, hitzak eta lokuzioak biltzen dituena, eta definizioez gainera, kategoria gramatikala, etimologia, itzulpenak, audioak eta abar jasotzeko diseinatua.

2. taula. Baliabideen informazioa (2024ko apirilaren 7ko datuak)

Baliabidea	URLa	Testu-kopurua
Gizapedia	gizapedia.org	6.425 artikulu eta ikasliburu
Hizkailua	hiztegia.gizapedia.org	4.510 hitz
Ikusmira	ikusmira.org	1.890 artikulu

Etorkizunerako erronkak dira giza eta gizarte zientzien inguruan entziklopedia lantzen jarraitzea (ez da erronka makala!), baita hiztegia ere, bereziki giza eta gizarte zientzietako lokuzio bereziak txertatuz eta horrekin batera, ikastaroen eskaintza zabaltzea.

CLARIAH-EUS azpiegituran Gizapedia baliabidea hezkuntza atalean sartu dugu (ikus clariah.eus/hezkuntza) eta CLARIAH-EUS azpiegiturari gure edukiak eskaintzen dizkiogu, testu-masa moduan, baina halaber ezagutza datu-base moduan. Gizapediako edukiak sarbide askeko edukiak dira, egile-eskubide guztiak atxikitzen badituzte ere. Webguneak sortzen dituen datuak (bisitak, bilaketak…) partekatu litezke, baldintza batzuekin bada ere. CLARIAH-EUS azpiegituratik, berriz, testu eta datu horien azterketarako tresnak baliagarriak izango zaizkigulakoan gaude, barne mailan gure azterketa propioak egin, eta azterketa horien emaria ere partekatzeko. Esate baterako, Voyant Tools tresnako emaitzak artikulu jakinen bukaeran epe ertainera txertatzea oso interesgarri deritzogu.