ALCESTE euskaraz: egokitzapena eta ebaluazioa

Juan Abasolo, Naia Eguskiza
Bilboko Hezkuntza Fakultatea, EUDIA ikerketa-taldea, UPV/EHU

Artikulu honek ALCESTE metodoaren eta berori erabiltzeko egun baliatzen den Iramuteq softwarearen egokitzapena eta ebaluazioa aurkezten ditu euskarazko testuetarako. Testu-bolumen luzeen sailkapen automatizaturako metodologiak eremu semantikoak identifikatzeko baliabideak eskaintzen ditu. Artikulua lexikoaren egokitzapen-prozesuan eta barne- eta kanpo-ebaluazioetan oinarritzen da. Barne-ebaluaziorako, UPV/EHUko Hezkuntzako fakultateetako irakasleen autodeskribapenak erabili dira; kanpo-ebaluaziorako, Itun Berriko San Pauloren gutunen corpus paralelo eleaniztuna aztertu da euskaraz, gaztelaniaz, ingelesez eta frantsesez.

GAKO-HITZAK: Reinert metodoa · Iramuteq · Euskarazko lexikoa · Ikerketa-tresna · Eremu semantikoak.

ALCESTE in Basque: adaptation and evaluation

This article presents the adaptation and evaluation of the ALCESTE method and the Iramuteq software, which is currently used for this method, for Basque texts. The methodology for the automated classification of large text volumes offers tools to identify semantic domains. The article focuses on the lexicon adaptation process and both internal and external evaluations. For the internal evaluation, self-descriptions of faculty members from the UPV/EHU education faculties were used; for the external evaluation, a multilingual parallel corpus of Saint Paul's letters from the New Testament was analyzed in Basque, Spanish, English, and French.

KEY WORDS: Reinert method · Iramuteq · Basque lexicon · Research tool · Semantic areas.

https://doi.org/10.26876/uztaro.5073 ***Jasotze-data:*** 2023-08-19 ***Onartze-data:*** 2024-02-09

1. Sarrera[1]

Azterlan honetan ikertzaile talde bat Iramuteq testuak eta galdetegiak aztertzeko tresna (Ratinaud eta Déjean, 2009) euskarazko datuekin erabili ahal izateko egiten ari garen egokitzapena aurkezten da labur-labur. Gure asmoa da euskarara egokitutako lexiko bat sortzeko prozesua garden bihurtu eta partekatzea, eta, aldi berean, testu-analisian erabilitako oinarri teorikoen ikuspegia argi ematea. Era berean, metodologia honen bidez lortutako emaitzetako batzuk ere azalduko ditugu eta, azkenik, lexiko hau eskaintzen diogu euskal ikertzaile-komunitateari, horren doitasunaz hausnartzeko gonbitarekin batera. Gure ustez, azterlan honek interesa pitz lezake testuen eta galdetegien analisia euskarazko datuetan oinarritzeko bidea ezartzeko.

2. Aurrekariak

Corpusaren hizkuntzalaritzaren esparruan, azkenaldiko aurrerapenek metodologia berriak sortzea ahalbidetu dute, hala nola informatikan izandako aurrerapenek, estatistika-teknika sofistikatuagoen bilakaerak edota datu-bolumen handien biltegiratze- eta prozesatze-ahalmenaren hobekuntzak. Horien artean, lexikometria eta testumetria nabarmentzen ditugu. Beaudouinek nabarmentzen duen moduan, aipatu teknikok xumetzat har daitezke Machine Learning edota Topic Modeling modukoen aldean, baina gizarte-zientzialariari corpusean oinarritutako hipotesi esplizituak eraikitzeko bidea ere eskaintzen diote, besteek ez moduan (Beaudouin, 2016); azken ezaugarri horrek eta askoz corpus xumeagoekin lan egiteko bidea eskaintzeak gizarte-zientzietako tresnerian tokia eman diote lexikometriari.

Corpuseko hizkuntzalaritzaren alorreko praktikek jatorri goiztiarra izan zutela esan daiteke; Sancta Vulgataren (latinezko Bibliaren) indexazioarekin gertatu zen hori; XIII. mendearen hasieran Hugo de San Carok eta haren zuzendaritzapeko 500 fraidez osatutako talde batek egina, *Correctio Biblie*[2] (Hanon, 1991) lanean. Hala eta guztiz ere, 80ko hamarkadan semantika interpretatiboa garatu zen arte, ez dago aurrerapen esanguratsuegirik aipatzeko ikuspegi horretan. Françoos Rastierren lana (Rastier, 1987) har daiteke diziplinako bultzadaren abiapuntutzat, diskurtsoa modelatzeko teknika eta tresna konputazionalak eman zituena.

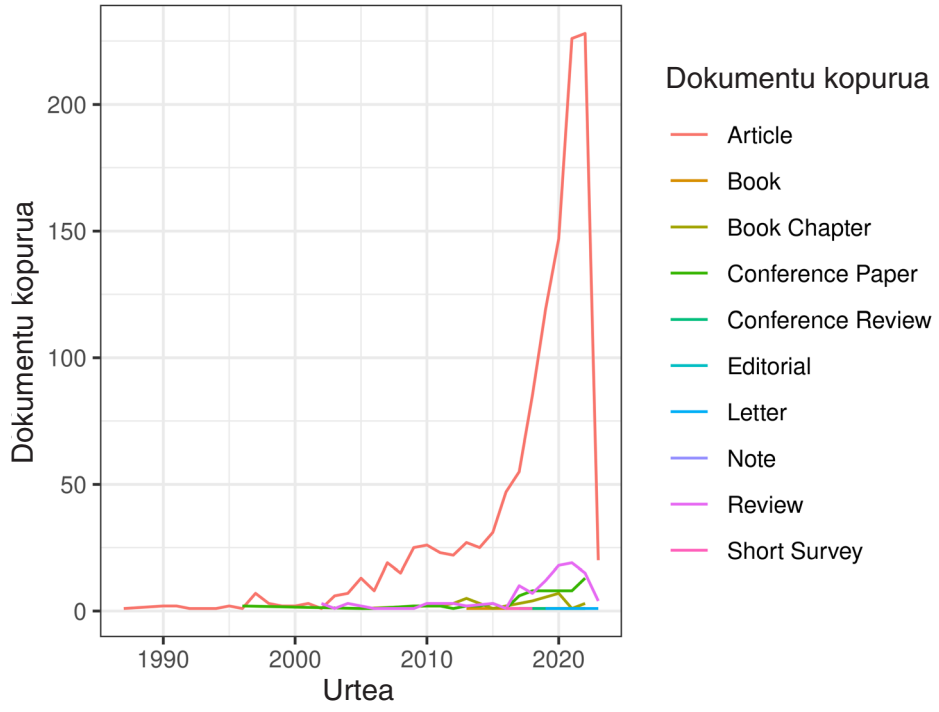
Lexikometriaren helburua da corpus baten barruan eremu semantikoak aurkitzea teknika kuantitatiboen bidez. Metodologia lexikometrikoaren berezko ezaugarri aipagarria da interpretazio kualitatiboaren beharra, corpus-analisiak teknika kuantitatiboetan oinarritzen badira ere, horixe baita emaitzak esanahiz janzteko unea.

Izan ere, teknika lexikometrikoak analizatu beharreko corpusean patroiak eta pareko gertaldiak hautematean oinarritzen dira. Ezaugarri jakin batzuen garrantzi estatistikoa ezarri ondoren, funtsezkoa da matematikaren eta estatistikaren ikuspegi-

1. GIU21/016 proiektuaren finantzaketa jaso du lan honek.

2. Online kontsultagai dago 1498ko argitalpena: *Inkunabeln / Biblia latina: Cum postillis Hugonis de Sancto Caro/ : 1. [Basel]. [nach 29. X. 1498 u. nicht nach 1499].* (1498). http://digital.ub.uni-duesseldorf.de/ink/7856461

tik identifikatutako eta nabarmendutako elementuak berrinterpretatzea eta esanahia ematea teoriaren eta aurreko ezagutzen argitara.



Scopuseko datuak (2023/02/28)

1. irudia. Alceste, Reinert eta Iramuteq Scopus datu-baseko dokumentuetan.

Lexikometriak ahalbidetzen dituen analisi-tekniken artean, litekeena da Reinertek 1983an proposatutako ikuspegia (Reinert, 1983), esperientzia akademikoari dagokionez, esanguratsuena izatea. Horren arrazoia, besteak beste, ikuskera kuantitatiboaren eta kualitatiboaren arteko konbergentzia da; adierazgarritasun estatistikoaz informatzeaz gain, eraikuntza estatistiko horren interpretazio kualitatiboa ere erabili beharra baita. Ikuspegi hori oso preziatua izan da argitalpen akademikoetan, teknika hori erabiltzen duten dokumentu zientifikoen ugaritzea erakusten duen 1. irudian ikus daitekeenez datu-base erabilienetariko bateko agerreren argitalpen moten sailkapenean. Bereziki nabarmena da Reinertek diseinatutako metodologiari nolako argitaratzaileen zabalak ematen dion garrantzia, moneta-politiketatik (Schonhardt-Bailey eta Bailey, 2013) osasun- (Lelorain *et al.,* 2012) edo hezkuntza-alorretaraino (Sobczak *et al.,* 2006).

2.1. Reinerten metodoa: ALCESTE

1980ko hamarkadaren hasieran, Reinertek (Reinert, 1983) algoritmo iraultzailea argitaratu zuen, testua bera inguratzen duen testuaren, literalki *testuinguruaren*, arabera sailkatzeko aukera ematen zuena, dimentsio anitzeko analisi baten bidez.

Hasiera batean, gizarte-psikologiaren esparruan, erantzun irekiak galdera-sortetan kategorizatzeko tresna gisa aurkeztu zen, analisi erreproduzigarriak ezartzeko. Hiru urte geroago, proposatutako algoritmoaren lehen inplementazio informatikoa abiarazi zen, ALCESTE[3] programa gisa ezagutu zena (Reinert, 1986). Horrekin batera, lan-fluxu bat argitaratu zen, ondoren gizarte-zientzien esparruan artikulu batean zehatz-mehatz azaldu zuena (Reinert, 1990).

Nahiz eta artikulu honetan metodologia horren alderdi tekniko guztiak jorratu nahi ez, ikuspegi teknikoa oinarri hartuta, gure asmoa da teoria interpretatu eta ulertzea.

Benzecrirekin elkarlanean (1981), Reinertek proposatu zuen testu batean hitz bakoitzaren testuingurua testu osoak osatzen duela, hitza bera alde batera utzita, «testuinguru» hitzaren jatorrizko definizioaren ildotik dio. Ikuspegi teoriko batetik, horrek hitzen testuinguruen modelazioa errazten du, inguruko hitzen arabera. Irudikapen hori lortzeko, Reinertek garatutako ALCESTE izeneko teknikak testuinguruak deskribatzea proposatzen du, testuinguru-hitzen elkarketaren bidez.

Emandako corpus bat aztertzeko, aztergaia testu-unitatetan banatu behar da; idealki, esanahi-unitate ere izan behar dutenak. Testu-unitateoi etiketak ere eslei dakizkieke, geroagoko azterketarako. Testu-unitate horiek hierarkikoki sailkatzen ditu ALCESTE teknikak goranzko klusterrak kalkulatuta (Ward, 1963), testu-unitateetako lexikoaren agerreraren arabera. Teknika horren bidez, lexiko-eredu erregularrak dituzten testu-unitateak identifika daitezke *klaseka*, kluster edo talde horiek «klase» izendatzen baitira metodologia honetan; azkenik, definitutako klase guztietan termino bakoitzaren presentzia edo absentzia espezifikoa kalkula daiteke, $\chi^{\wedge 2}$.

Hala ere, proposamen teoriko honek aurrez aurre jartzen du hizkuntzaren benetako erabileran aplikatzeko erronka; izan ere, hizkuntza-formen azpian dauden kontzeptuzko elementuak bereziak edo polimorfikoak izan daitezke, lemek hainbat forma har ditzaketelako testuan. Zailtasun hori areagotu egiten da bereziki hizkuntza eranskarietan, hala nola euskaran.

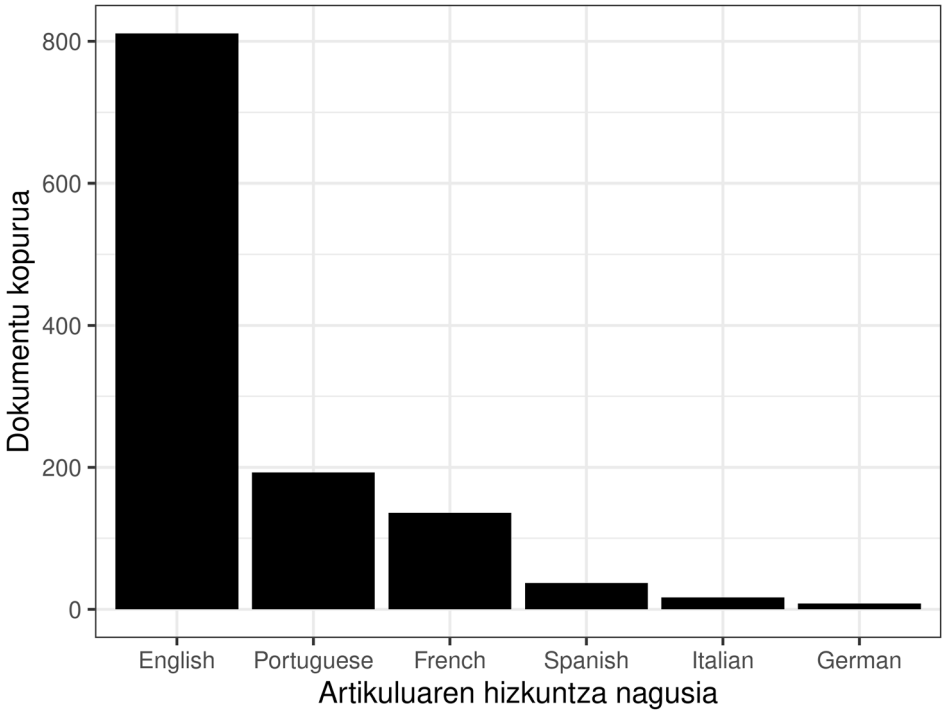
Reinertek asmatutako irtenbideak, oraingo honetan euskarari aplikatzen zaionak, beharleku galanta eskatzen du: aldez aurretik hizkuntza-forma erabilgaien zerrenda antolatua izatea, zein bere lemarekin eta balio sintaktikoarekin etiketatuta, analisia egin aurretik; 6. atalean azaltzen dena. Estrategia horrek, aurretiazko lexiko lematizatua izateak, besteak beste, erraztu egiten du elementuak esanahiz janzteko zeregina, teknika lexikometrikoen bidez identifikatu eta nabarmendutakoak.

2.2. Iramuteq

Reinertek (1983) proposatutako algoritmoan oinarritzen da Iramuteq, R programazio-lengoaia erabiliz inplementatua (Ihaka eta Gentleman, 1996). Bai iturburu-kodea bai Iramuteq-en aplikazioa doan eta libre daude eskuragarri (Ratinaud, 2014; Ratinaud eta Déjean, 2009). Izan ere, Iramuteq R interfaze bat bezala aurkezten da testu eta galdeketen dimentsio anitzeko analisirako (*Interface*

3. *Analyse des Lexèmes Coocurrents dans les Enoncés Simples d'un Texte.*

de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires), deskargako webgunean deskribatzen denez. Dimentsio anitzeko analisiak eskaintzeaz gain, Iramuteq-ek testu-analisien beste mota batzuk ere egiteko aukera ematen du, dimentsio anitzekoak zein dimentsio bakarrekoak. Horren barruan sartzen dira faktore-analisia (Borko, 1965), hitz-hodeiak, bat-etortzeen bilaketa eta antzekotasun-analisia (Baril & Garnier, 2015). Artikulu honetan analisi gehigarri horiei buruzko xehetasunetan murgilduko ez bagara ere, ikertzailearentzat interesgarria bada, bihoakio horren aipamena berak sakontzeko.



Scopuseko datuak (2023/02/28)

2. irudia. ALCESTE, Reinert eta Iramuteq Scopus datu-baseko dokumentuetan, hizkuntzen arabera.

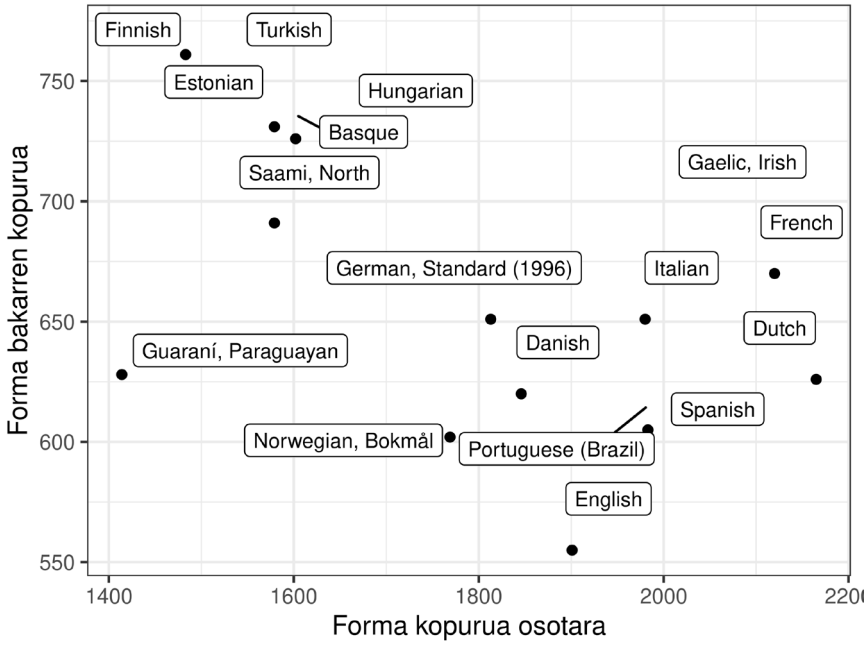
2. irudian nabarmentzen den bezala, orain arte tresna horren eraginkortasuna berretsi da. Scopus datu-basean, hainbat hizkuntzatako argitalpenen garrantzia ikus daiteke, baita hasieran espero ziren frantsesetik eta ingelesetik haraindi ere. Argitalpen zientifikoen datu-baseak berrikustean, agerian geratzen da, halaber, hainbat kasutan ikerketek Reinerten metodoa (edo ALCESTE edo Iramuteq) aipatzen dutela. Hala ere, ikerketaren komunikazioan erabilitako hizkuntza ez dator beti bat metodoa garatu zen hizkuntzarekin. Azterketa horien arabera, teknika, metodo eta tresna horiek hainbat gizarte-irudikapen deskribatzeko erabiltzen dira (Navarro & Idoiaga, 2021; Trigo *et al.,* 2021).

2.2.1. Hizkuntzak

Lehen aipatu bezala, Iramuteq tresnak lexiko antolatua eskatzen du, aztertu beharreko testuetan agertzekoak diren hizkuntza-formak bilduko dituena. Garrantzitsua da tresna zein hizkuntzatan erabiltzen ari den eta haren ezaugarri bereizgarriak zein diren ikertzea, baldin badaude.

Gaur egun, Iramuteq-en banaketan alemanezko, ingelesezko, frantsesezko, galizierazko, grezierazko, italierazko, latinezko, portugesezko, errumanierazko, espainierazko eta suedierazko lexikoak daude. Hala ere, eskuragarri dauden lexiko guztietatik, batzuk soilik daude automatikoki eskuragarri programaren interfazearen bidez, eta horietariko gehienek tamaina egokia dute analisi eraginkorrak egiteko. Hautaketa automatikorako aukerarik eskaintzen ez duten lexikoen kasuan, hiztegi gisa fitxategi espezifiko bat esleitzeko aukera ematen duen elkarrizketa-koadro baten bidez eskura daitezke.

Aukera aipatu berri horri esker, hainbat hizkuntzatan erabiltzeko prestatutako lexiko ugari daude, termino gehiago dituztenak. Aurretik aipatutako banaketa ofizialeko lexikoez gain, beste batzuk ere badaude eskuragarri. Adibidez, gaztelaniazko lexiko bat dago, X-ko (lehengo Twitter) esteka baten bidez partekatzen dena (Ideia [@ ideiainova], 2017), baita alemanezko lexiko bat ere, posta publikoko zerrendetan banatzen dena (Loubere, 2023).



Iturria: UNESCO (Giza Eskubideen Aldarrikapen Unibertsala zenbait hizkuntzatan)

3. irudia. Zenbait hizkuntzatako itzulpenen hitz kopuru eta forma bakarren arteko alderaketa.

ALCESTE metodorako diseinatutako lexiko gehienak indoeuropar hizkuntzetan daude. Latina izan ezik, guztiak hizkuntza prepositiboak.

Ikerlan honetan, metodo hau euskarara nola egokitu den aurkezten dugu. Ez da, gainerako hizkuntzak bezala, hizkuntza indoeuroparra, eta egitura oso eranskaria du euskarak. Alde horren erakusgarri 3. irudia dugu; UNESCOren Giza Eskubideen Adierazpen Unibertsala ofizialki itzultzeko behar diren formen kopurua erakusten du, formen guztizko kopurua eta erabilitako forma desberdinen kopurua alderatuaz.

3. Helburuak eta metodoa

Azterlan honen helburua da euskarazko lexiko bat egitea eta ebaluatzea, Reinertek proposatutako metodologiaren bidez euskaraz idatzitako testuak aztertzeko. Atal honetan, corpusak eraikitzeko erabilitako metodoa eta, ondoren, prozesatzeko eta aztertzeko erabili diren prozedurak zehazten dira, lexikoa eraikitzea helburua dela.

3.1. Lehen helburua: euskarazko lexikoa sortzea

Ikerketa honen lehen helburua euskarazko lexiko bat sortzea izan da. Horretarako, hainbat urrats egin dira. Lehenik eta behin, corpus zabal bat garatu behar izan da, hainbat testuingurutako hizkuntzaren erabilera islatzeko. Ondoren, corpusaren azterketa sakona egin da, erabilitako hizkuntza-formen, horiei dagozkien lemen eta horiek kokatzen diren kategoria gramatikalen artean dauden hartu-emanak identifikatzeko. Azkenik, identifikatutako erlazio horiek multzo bakarrera murriztu dira, erabilera-maiztasuna oinarritzat hartuta.

4. irudia. Iramuteq-erako lexiko baten zenbait lerro.

Aurreko 4. irudiak irudikatzen du lexikoak behar duen egitura: lehenengo zutabean formak edo *tokenak* dira, lemen flexioak barne. Bigarren zutabeak dagozkion lemak erakusten ditu. Hirugarren, berriz, esleitutako kategoria gramatikala adierazten duen frantsesezko laburdura ikusten da.

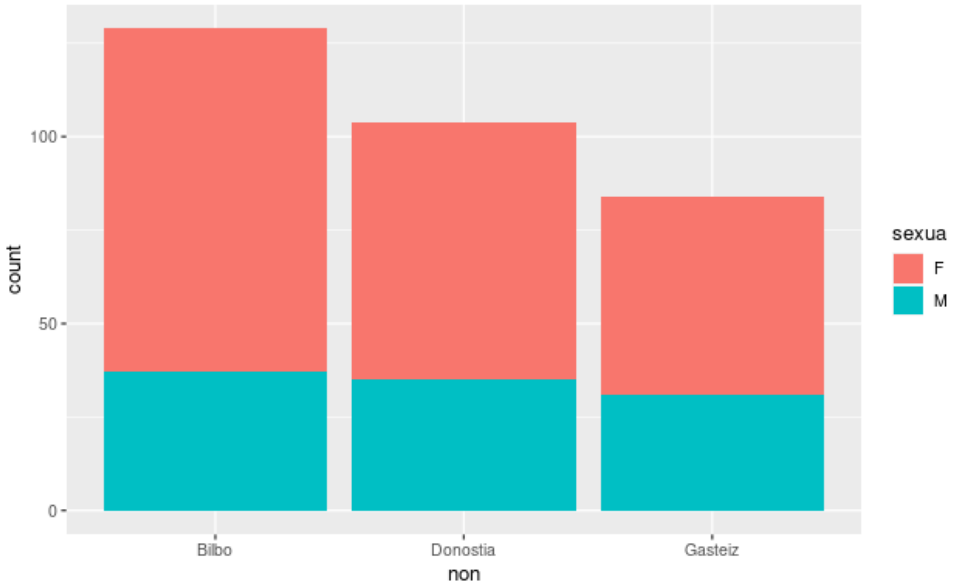
Analisian erabilitako corpusak hiru iturri nagusi izan ditu: *Argia* aldizkariko formatu digitaleko artikulu eta gutunek osatu dute, 1.082.787 esaldirekin (21.448.752 token), corpusa batzeko momenturako eskuragai zeudenak. Horrez gain, Wikipediako euskarazko artikuluen 61.567 esaldi gehitu ziren (1.156.358 token), Common Voice proiekturako aukeratutakoak, bai eta mendebaldeko euskalkian idatzitako *Bizkaie* aldizkariaren 2020. urteko argitalpenetatik ateratako 11.379 esaldi ere (203.284 token). Oro har, milioi eta ehun mila pasa esaldi (22 milioi eta erdi token) bildu ziren aztertzeko, hainbat hizkuntza-erregistro eta -forma biltzen dituztenak.

Hizkuntza-formak aztertzeko eta sailkatzeko, R programazio-lengoaia erabili da (R Core Team, 2020). Zehazki, «udpipe» paketearen 0.8.5 bertsioa erabili da (Straka *et al.,* 2016) (Wijffels *et al.,* 2020), «Universal Dependencies» ereduak aplikatzeko (Nivre *et al.,* 2016) interfaze moldagarria ematen duena. Aranzabek eta kolaboratzaileek garatutako euskara-eredua erabili da (Aranzabe *et al.,* 2015), lemak eta gramatika-kategoriak modelatu eta sailkatzeko.

3.2. Bigarren helburua: landutako lexikoaren hizkuntza barneko sendotasunaren ebaluazioa

Eraikitako lexikoaren eraginkortasuna aztertzeko asmoz, ikuspegi bi hartu dira kontuan. Lehenik eta behin, barne-ikuspegitik, euskaratik bertatik, ebaluazio bat egin da, ea lortutako emaitzek koherentzia erakusten duten egiaztatzeko asmoz, segidan azalduko dena. Eta bigarren ikuspegia, aurrerago landuko dena, balioa erakutsia duten beste hizkuntza batzuen emaitzarekiko alderaketa. Lehenengo azterketarako, UPV/EHUko irakasle-ikasketetako irakasleek idatzitako autodeskribapenen corpusa eratu da. Unibertsitateko irakasleek ikasketetako deskribapenean erakusten duten euren buruaren deskribapen-fitxak dira, euskaraz idatzitakoak, zehazki Haur Hezkuntzako edo Lehen Hezkuntzako graduak egiten diren hiru fakultateetako webguneetatik bildutakoak, 5. irudian irudikatzen da autodeskribapen horien egileen generoa eta zein campusi atxikitzen zaizkion. Testuok Iramuteq erabilita aztertu dira (ikus *Emaitzak* atala), euskaraz eraikitako lexikoa erabiliz, Reinertek proposatutako ALCESTE metodologiaren araberako sailkapen hierarkiko goranzkoa egiteko.

Corpus aipatu berria osatzen dute euskaraz idatzitako 129 autodeskribapenek, irakasle-ikasketak egiten diren UPV/EHUko hiru fakultateetako 317 irakasleen artean 2021eko azarorako idatzita zituzten guztiek. Testu-multzo horren bidez, guztira 1.279 esanahi- eta etiketa-unitate pare sortu dira; horretarako, idatziak puntuazioaren arabera zatitu dira eta bakoitzaren egilearen campusaren, irakasleen sailaren eta haien generoaren arabera sailkatuta ezarri zaizkie etiketak.



5. irudia. Corpusaren banaketa campus eta generoaren arabera.

3.3. Hirugarren helburua: landutako lexikoaren hizkuntza arteko trinkotasunaren ebaluazioa

Bestalde, euskaraz lortutako emaitzen koherentzia baliozkotzeko, corpus paralelo eleaniztuna sortu da. Euskarazko emaitzak Iramuteq erabiltzen ibilbide luzeagoa duten beste hizkuntza batzuetako emaitzekin alderatzea izan da horren helburua, hala nola frantsesarekin, ingelesarekin eta gaztelaniarekin. Instantzia honetan, corpusa Itun Berriko San Pauloren gutun guztiek osatzen dute, denak teknika eta prozedura bera erabilita segmentatu eta lematizatu dira. Iramuteq-ek hizkuntza bakoitzean emandako lexikoak erabili dira eta euskarazko testuetan 2. atalean azaldutako lexiko hori. Hizkuntza bakoitzeko testuak banaka ALCESTE metodoaren arabera sailkatu dira, klaseak identifikatzeko.

Hainbat hizkuntzatan identifikatutako klaseen arteko koherentzia aztertzeko, ALCESTE analisien eraikuntzan erabilitako testuinguru-unitateen arabera bat etortzea aztertu da; hau da, hizkuntza bakoitzeko klase bakoitzaren definigaitzat hartu dira testu-unitate osagaien kodeak. Hori oinarri hartuta, hizkuntza guztietako klaseen arteko antzekotasun-matrize bat sortu eta horretan oinarrituta goranzko sailkapen hierarkikoa eregi da. Reinertek proposatzen duen sailkapen hierarkikoa goitik beherakoa da, osotasun baten unitateak identifikatzea helburu duelako, hala eraikitzen dira azpicorpus bakoitzeko klaseak. Klase horiek denak bat direnetz ez dago ezarrita, horregatik ea taldekatzerik eta antzekotasunik dagoen ezartzeko orduan, behetik gorako taldekatze hierarkikoa erabili da; zehazki, Ward-ek (Ward, 1963) proposatutako ereduari jarraikiz eraiki dira klusterrak.

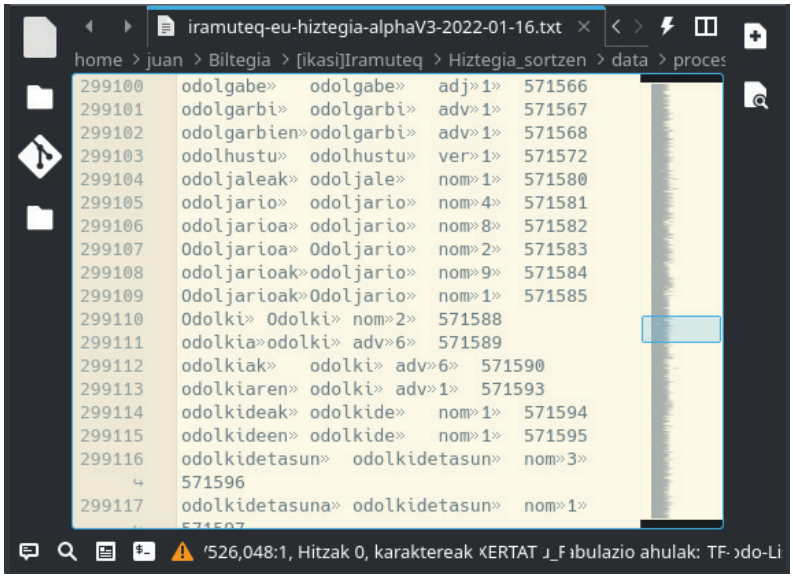
Corpus paralelo eleaniztuna eraikitzeko, Esloveniako Biblia Elkartearen webgunean dauden baliabideak erabili dira[4]. Elkarte horrek hainbat itzulpen pareka aurkeztea errazten du. Testuinguru horretan, gaur egungo euskarazko itzulpen bakarra hartu da kontuan, 1994ko *Elizen Arteko Biblia* edizio ekumenikoa. Horrekin batera, ingelesezko 1611ko King James-en bertsioa, frantsesezko 1910eko Louis Segond-en itzulpena eta gaztelaniazko 2002ko *Dios Habla Hoy* bertsio ekumenikoa. Itzulpen horiek osatzen dute corpus eleaniztuna, eta hizkuntzen arteko funtsezko alderaketak egitea ahalbidetu dute, txatalez txatal, kapituluz kapitulu eta gutunez gutun parekatuta.

4. Emaitzak

Atal honetan, hurrenez hurren, lexikoa sortzeko prozesuan lortutako emaitzak azaltzen dira, lexikoaren errendimenduaren inguruko barne-azterketa eta hizkuntzen arteko kontrasteak, azkenik.

4.1. Lexikoaren garapena

Lexikoa sortzeko corpusaren hasierako analisiak 13.065.741 forma-lema-kategoria gramatikaleko konbinazioa izan du hasierako analisietan. Lehen nabarmendu den bezala, funtsezkoa da zifra hori formen eta lemen arteko eta lemen eta kategoria gramatikalen arteko erlazio bakanetara murriztea. Murrizteko prozedura hori erabileraren maiztasunean oinarritu da; lehentasuna harreman kopuruei eman zaie. Murrizketa hori aplikatu ondoren, lexiko bat osatzea lortu da, 627.479 forma sailkaturekin.



6. irudia. Iramuteq-erako euskarazko lexikoaren zenbait lerro.

4. https://biblija.net

Egitura prepositiboa duten hizkuntzetarako egindako beste lexiko batzuetan ez bezala, euskararen propietate eranskariengatik, lexiko honetan izen propio gisa sailkatu diren 20.093 forma sartu dira. Forma horiei pro_per etiketa esleitu zaien arren, izen propioak dira. Aipatu etiketa frantsesezko lexikoan pertsona-izenordainei dagokie. Euskararako lexikoan, izenordain horiek ez dira bereizten, gaztelaniarako, ingeleserako, portugeserako eta italierarako lexikoak egiteko hartutako erabakiekin bat etorriz.

4.2. Euskarazko erabilera

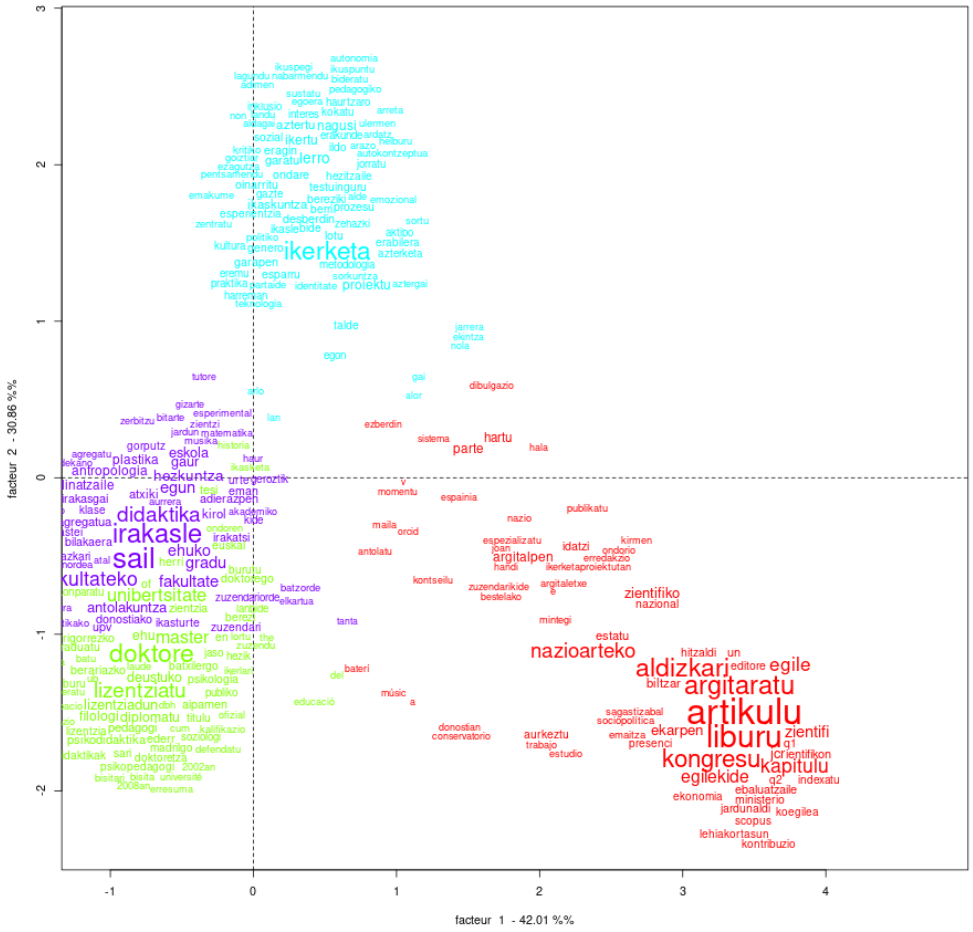
Euskarazko lexikoaren erabileraren hasierako azterketa batean argi ikusten da nolako aldea dagoen sortutako lexikoa lematizatzeko erabili ala ez, 5.2. ataleko irakasleen autodeskribapenen corpusaren azterketan.

(a) Lematizatu gabekoa (b) Lexikoarekin lematizatutakoa

7. irudia. Corpusaren ikuskera bi.

Goiko lehenengo 7(a) hitz-lainoa sortzeko corpuseko formen maiztasuna erabili da, eta eskuinaldeko 7(b) irudiko hitz-lainoa sortzeko, berriz, corpus horretako lemen maiztasuna erabili da. Garrantziari dagokionez, *eta* juntagailua nabarmentzen da 5(a) irudian, *hezkuntza* lema, berriz, askoz garrantzi baxuagoz agertzen da. Lematizatutako analisian, 7 (b) irudian, berriz, *hezkuntza* lema nabarmentzen da nagusi; guztira 487 aldiz erabiltzen da, hamabi formatan, besteak beste: *hezkuntza*, *hezkuntzan*, *hezkuntzako*, *hezkuntzarentzat* eta abar.

Reinerten metodologiaren araberako testuinguru lexikoen sailkapena egitean lau klase identifika daitezke; 8. irudian proiektatzen dira klase horiekin χ^2 altuena duten lemak. Bertan, lau eremu semantiko identifikatzen dira, horietako bik elkarrekin lotura estua dutela.



8. irudia. Lau klaseetako lemen proiekzioa plano batean.

Gorrizko klasean argitalpen eta komunikazioen eremu semantikoa bereizten da, «artikulu», «argitaratu», «liburu», «kongresu»,«aldizkari», besteak beste, garrantzi berezikotzat hartuta.

Bigarren klaseak, urdinez irudikatuak, ikerketa zientifikoaren jarduera hartzen du barne, non «ikerketa», «metodologia», «lerro» eta «proiektu» erako lemak nabarmentzen diren, besteak beste.

Azken bi klaseei dagozkien eremu semantikoen arteko lotura zein bereizketa argi ikusten da irakasleen autodeskribapenarekin lotuta. Irakaskuntza-jardueraren esparrua kolore morez identifikatzen da, eta ibilbide akademikoaren eremua, berriz, berdez. «Irakasle», «didaktika», «sail», «atxiki», «eskola», «hezkuntza», «plastika» lemak, besteak beste, irakaskuntza-ekintzari buruzkoak dira nagusiki, eta ibilbide akademikoarekin lotzen dira, «doktore», «lizentziatu», «diplomatu» eta «master», beste hainbat lemarekin.

4.3. Alderaketa paraleloa

Gorago azaldu den moduan, alderatu nahi izan da euskararako sortutako lexikoa beste lexiko batzuekin. Horretarako, egitura bakarreko lau azpicorpus baliokide paraleloki batuz corpus bat eraiki da, San Pauloren gutunen itzulpenak erabilita. Horrela, lehenengo pausoan azpicorpus bakoitza sailkatu da ALCESTE metodoa erabilita. Gorago azaldu moduan, Iramuteq-ek hizkuntza bakoitzeko lexiko bat erabili behar du ALCESTE metodoaren araberako sailkapenak egiteko. Pauso honetan azpicorpus bakoitzerako erabili dira Iramuteq-ek banatzen dituen frantsesezko, inegelesezko eta gaztelaniazko lexikoak eta, euskararako, hemen aurkezten ari garena. Azken pausoa, atal honen amaieran erakutsiko dena, sailkapenen arteko alderaketa da.

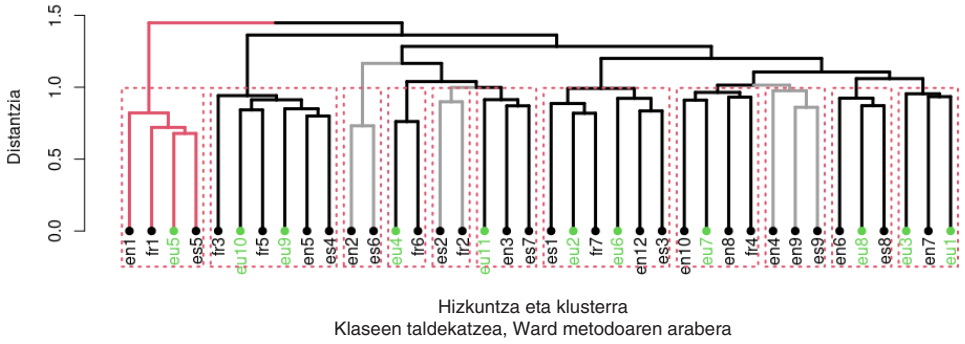
San Pauloren testuen ALCESTE metodoaren araberako sailkapenak klase kopuru desberdinak sortu ditu azpicorpus bakoitzeko, erabilitako lexikoen arabera betiere. Segidako 1. taulan aurkezten dira horien xehetasunak, zein itzulpen erabili den, zein lexiko, azpicorpusaren zein ratio sailkatu duen eta zenbat klase sortu dituen sailkapenak.

1. taula. Lau hizkuntzetako sailkapenen laburpena.

Itzulpena	Lexikoa	Sailkatutakoaren ratioa	Klase kopurua
SEG	frantsesa	0,9751	7
KJB	ingelesa	0,952	12
DHH	gaztelania	0,950	9
EAB	euskara	0,942	11

Lehen 3.3. azpiatalean azaldu den moduan, klaseen arteko antzekotasunak aztertzeko, klase guztien arteko antzekotasun-matrizea sortu da, antzekoagotzat hartuta testu txatal baliokideak erabiltzen dituzten klaseak eta desberdinagotzat testu txatal ez-baliokideak erabiltzen dituztenak. Matrize hori oinarritzat hartuta taldekatze edo kluster-sailkapena eraiki da Ward-en algoritmoa (1963) erabilita.

Dendrograma batez irudikatzen da kluster-sailkapen hori, 9. irudian. Y ardatzak adierazten du antzekotasun-neurria, horregatik, klaseak zenbat eta beherago batu, antzekoagotzat hartu behar dira eta zenbat eta altuago batu, orduan eta desberdinagoak euren artean. Irudian kutxa gorrien bitartez, hamaika kluster markatzen dira, euskarazko lexikoarekin euskarazko azpicorpusak hamaika klase sailkatu baititu, gorago esan moduan.



9. irudia. Hizkuntza arteko alderaketako klaseen taldekatzearen dendrograma. Klase bakoitza etiketatzeko, dagokion azpicorpusaren hizkuntza-kodea erabili da (eu, euskara; en, ingelesa; fr, frantsesa; es, gaztelania) eta zenbaki bat.

Klaseetako taldeek antzekotasun handiagoa dute unitateen edo taldeen artean, *y* ardatzean 0 baliotik zenbat eta gertuago egon. Horrela, antzekotasun handiena erakusten duten bi klaseak eu5 eta es5 bezala etiketatutakoak direla ikus dezakegu, euskarazko eta gaztelaniazko azpicorpusaren analisien bosgarren klaseei (eu5 eta es5) dagozkienak, hurrenez hurren. Bi klase horiek unitate gisa integratzen dira fr1 eta en1 klaseekin, eta horiek, aldi berean, azpicorpusaren analisien lehen klaseak dira frantsesez eta ingelesez, 7. irudiko dendrograman talde horren adarra gorriz nabarmenduta agertzen da.

Sortzen den lehenengo taldea osatzen dute itzulpen ekumenikoetako testuen analisitik jasotako klase bik, euskarazkoa bata eta gaztelaniazkoa bestea. Ondoren, bi klase horiek bat egiten dute frantsesezko klase batekin eta ingelesezko beste batekin segidan. Talde hau osatzen lehena da, baita gainerako sailkapenekin bat egiten azkena ere. Ezaugarri horrek agerian uzten du klase-multzo hori dela kalkulatutako gainerako talde guztietatik urrunen dagoena.

Jarraian, lehen taldeko hitz-lainoak aurkezten dira 10. irudian, hizkuntza bakoitzeko azpicorpusen jatorrizko sailkapenetan oinarrituta bereizita. Hitz-laino horietan, aipatutako lema asko hizkuntzen artean bat datozela ikus daiteke, horietako batzuk 2. taulan laburbiltzen dira.

(a) Euskarazkoa

(b) Gaztelaniazkoa

(c) Frantsesezkoa

(d) Ingelesezkoa

10. irudia. Lehenengo klasean sailkatutako klaseen hitz-lainoak.

Lehen talde honen kasuan, garrantzitsua da azpimarratzea sailkapenetan erabilitako lema gehienek esanahi baliokideak dituztela gainontzeko hizkuntzetan, 2. taulan aurkezten dira horiek . Horrek iradokitzen du corpus paralelo horien bitartez aurkitutako klaseak antzekoak direla, lema baliokideak aurkitu dituelako lau lexikoen erabileran.

2. taula. Lehenengo klusterreko klaseei dagozkien lema baliokideen aurkezpena.

eu	es	fr	en
edan	beber	boire	drink
jan	comer	manger	eat
sasijainko(ei)	idolo	idole	idol
haragi	carne	viande	meat
gose	hambre	faim	hungry
ogi	pan	pain	bread
opari	sacrificio	sacrifier	sacrifice
eskaini	ofrecer	-offre (hori ez dago)	offer

Euskarazko klasean, «sasijankoei» forma nabarmentzen da, «idolo» (gaztelania), «idol» (ingelesa) eta «idole» (frantsesa) lemekin parekatzen dena. Euskal forma ez dago lematizatuta, baliokidea *sasijainko* izan behar baitzukeen; hori gertatzen da «sasijainko» forma ez delako egon euskarazko lexikoan, horretarako erabilitako corpusak ez duelako jaso eta corpusa lematizatzeko erabilitako udpipe-k ez duelako lematizatu.

Atal honetan aurkeztutako emaitzen laginean oinarrituta, jarraian hausnarketa merezi duten ezaugarri batzuk nabarmentzen dira.

5. Eztabaida

Atal honetan, artikuluan planteatutako hiru helburu nagusiei heltzen zaie eta horietako bakoitzean lortutako emaitzak aztertzen dira.

5.1. Lexikoa euskaraz eraikitzea

Euskarazko lexiko bat eraikitzeko lehen helburuari dagokionez, emaitzek erakusten dute lexiko bat sortzea lortu dela, hizkuntza-formen maila zabala eta forma-lema-kategoria gramatikaleko erlazio unibokoak barne hartzen dituena. Lexikoak 473.330 forma ezberdin ditu guztira, eta horrek oinarri sendoa ematen du euskaraz idatzitako hainbat diskurtso Iramuteq-en bitartez lematizatu eta ALCESTE metodoaren arabera sailkatzeko, gaur gizarte-zientzietan zabalduta dauden baliabide teknologikoekin. Hala ere, garrantzitsua da aitortzea lexiko horrek hainbat forma eta hainbat lema ez dituela izan kontuan, nekez ordezka baititzake lagin bakar batek hizkuntzaren idatzizko forma guztiak. Besteak beste, izen propioekin eta eremu semantiko bereizietan antzeman daiteke arazo hori, lexikoa eraikitzeko erabili den corpusak estaltzen ez dituen errealitateak, hain zuzen.

5.2. Euskarazko autodeskribapenen azterketa

Bigarren helburuari dagokionez, euskarazko autodeskribapenen azterketan emaitza koherenteak batu dira. Analisiari esker, Haur eta Lehen Hezkuntzako graduetako irakasleen deskribapenarekin lotutako lau eremu tematiko identifikatu ahal izan dira. Eremu horiek ikerketa zientifikoaren jardueratik hasi eta irakaskuntza-ekintzaren eta ibilbide akademikoaren deskribapenera jariatzen dira. Gaikako eremu horiek hautemateak emaitzen koherentzia eta egokitasuna iradokitzen ditu.

5.3. Hizkuntzen arteko alderaketa

Hirugarren helburuari dagokionez, euskaraz lortutako emaitzak beste hizkuntza batzuekin erkatzean, antzekotasun eta ezberdintasun interesgarriak ikusi dira. Kontuan hartu diren hizkuntzetako testuen sailkapenek antzeko multzo batzuk erakutsi izanak iradokitzen du badaudela eremu semantiko partekatuak. Hala ere, hizkuntza bakoitzaren berezitasun linguistiko eta semantikoei egotz dakizkiekeen aldeak ere sumatzen dira. Desberdintasun horiek nabarmendu egiten dute analisien emaitzak interpretatzean hizkuntza bakoitzaren ezaugarri espezifikoak kontuan hartzearen garrantzia. Zehazki, 1. taulan sailkatutakoaren ratioari dagokion zutabean ikusten da azpicorpus bakoitzeko zer ratio sailkatu ahal izan den dagokion lexikoaren bitartez, ALCESTE metodoa erabilita. Hor, nabarmentzen da euskarazko lexikoak izan duela sailkatzeko gaitasunik baxuena; 0,95; hau da, azpicorpusaren % 5 ezin izan du sailkatu, nahiz eta lexikorik handiena izan; hizkuntza eranskaria izatearen ezaugarria, inondik ere. Nabarmendu dezagun, gaur-gaurkoz, aztertutako datu-baseetan (Scopus, Clarivate, Scielo) ALCESTE metodologiarekin erabili den hizkuntza eranskari bakarra euskara dela.

Laburbilduz, azterlan honetan lortutako emaitzek lexikoak eraikitzeko eta euskarazko testuak aztertzeko erabilitako metodologiaren eraginkortasuna babesten dute. Proposatutako helburuak arrakastaz betetzat hartu behar ditugu, eta aurkikuntzek eremu semantiko batzuk sakonago ulertzen laguntzen dutela dirudi. Hizkuntzen arteko konparazioek antzekotasunak eta aldeak erakutsi dituzten arren, azken horiek analisi lexikometrikoetan hizkuntza bakoitzaren berezitasunak kontuan hartzearen garrantzia nabarmentzen dute. Oro har, lexiko honek euskarazko testuak aztertzen eta ulertzen laguntzen du eta lan honek etorkizunean alor horretan egingo diren ikerketen oinarriak ezartzen ditu.

6. Mugak eta aukerak

Euskarazko ikerketa testuala egiten duten zientzialarientzat eskuragai eskaintzen da Iramuteq erabiltzeko eraikitako lexikoa zein bera eraikitzeko erabilitako R lengoaiako gidoi edo *script*ak Open Science Framework barruko proiektu batean (Abasolo eta Eguskiza, 2022).

Betiere, erabili behar duen ikertzaileak tentuz erabili behar du, amaitu gabeko prozesu bateko urratsa baino ez baita euskarazko lexiko hori. Euskara zein beste hizkuntza eranskarietan ere antzera gertatuko litzateke, Iramuteq erabiltzeko

protokolo bereizia zehaztea beharrezkoa da, hizkuntza prepositiboetarako idatzitakoek gure hizkuntzaren portaerako hainbat ezaugarriri ez baitiete heldu ere egiten. Azken hori bera ere, ikerketagai izan daiteke hurrengo lanetan, euskararako propio garatutako softwarea sortzera bidean.

Kontuan hartu behar da ikuspegi hau euskarazko hainbat genero eta testu motatara hedatu behar dela, eta, hartara, testuinguru mota gehiagotan ebaluatu ahal izango da horren eraginkortasuna. Literatura-testuak, diskurtso politikoak, dokumentu historikoak eta idatzizko bestelako komunikazio mota batzuk barne har litzake hurrengo lanak. Genero bakoitzak erronka eta ezaugarri bereziak izan ditzake, eta horiek doitu egin beharko dira lexikoan eta analisi-prozesuan.

Azkenik, garrantzitsua da hizkuntzalarien, gizarte-zientzietako ikertzaileen eta informazioaren teknologiako adituen arteko diziplinarteko lankidetza sustatzen jarraitzea. Lankidetza horrek metodologia eta aplikazioa hobetzen lagun dezake, bai eta euskarazko eta beste hizkuntza batzuetako testuak aztertzeko prozesuan sor daitezkeen erronka espezifikoei heltzen ere.

Laburbilduz, nahiz eta euskarazko lexikoaren eraikuntzan eta aplikazioan aurrerapen esanguratsuak lortu diren, testuen azterketan lexikoaren erabilera zabaltzeko eta hobetzeko hainbat eremuk lantzeke segitzen du. Aukera horien arteko batzuk dira: batetik, lexikoaren etengabeko hobekuntza, hainbat genero eta testuingurutara egokitzea, eta, bestetik, lexikoa beste hizkuntza eta gramatika-egitura batzuetan aplika daitekeen aztertzea. Diziplinarteko ikuspegiaren eta etengabeko lankidetzaren bidez, metodologia hori modu eraginkorrean erabil daiteke euskaraz eta haraindi idatzitako testuak aztertzeko.

7. Erreferentziak

Abasolo, J., eta Eguskiza, N. (2022). Euskarazko lexikoa iramuteqerako. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/T8JEVX

Aranzabe, M.J., Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., Goenaga, I., Gojenola, K., eta Uria, L. (2015). Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In M. Dickinsons, E. Hinrichs, A. Patejuk eta A. Przepiórkowski (argtz.), *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)* (233-241. or.). Institute of Computer Science of the Polish Academy of Sciences.

Baril, E., & Garnier, B. (2015). *Utilisation d'un outil de statistiques textuelles*. Institut National d'Etudes Démographiques. http://iramuteq.org/documentation/fichiers/Pas%20a%20Pas%20IRAMUTEQ_0.7alpha2.pdf

Beaudouin, V. (2016). Retour aux origines de la statistique textuelle: Benzécri et l'école française d'analyse des données. In D. Mayaffre, C. Poudat, & L. Vanni (Arg.), *JADT 2016* (17-27. or.). al-01376938. https://hal.science/hal-01376938v1

Benzécri, J.-P. (1981). *Pratique de l'analyse des donnees: Linguistique et lexicologie*. Dunod.

Borko, H. (1965). A Factor Analytically Derived Classification System for Psychological Reports. *Perceptual and Motor Skills*, 20(2), 393-406. https://doi.org/10.2466/pms.1965.20.2.393

Hanon, S. (1991). 165. La concordance. *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, *2*, 1.562-1.567. https://doi.org/10.1515/9783110124200.2X

Ideia [@ideiainova]. (2017). Sharing a new version of the Spanish dictionary for #Iramuteq (+500k entries) [Tweet [Link a Archivo]]. In *Twitter*.

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299-314. https://doi.org/gddc3nX

Lelorain, S., Tessier, P., Florin, A., & Bonnaud-Antignac, A. (2012). Posttraumatic growth in long term breast cancer survivors: Relation to coping, social support and cognitive processing. *Journal of Health Psychology*, 17(5), 627-639. https://doi.org/10.1177/1359105311427475

Loubere, L. (2023). *Re: [Iramuteq-users] Dictionary in german? | iramuteq*.

Navarro, G., & Idoiaga, N. (2021). Bertso-eskolak, nerabezaroan hezteko espazio gisa. *Uztaro: giza eta gizarte-zientzien aldizkaria*, *Uztaro,* 117, 75-90. https://doi.org/10.26876/uztaro.117.2021.4

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (1.659-1.666. or.).

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing.

Rastier, F. (1987). Représentation Du Contenu Lexical Et Formalismes De L'intelligence Artificielle. *Langages*, 87, 79-102. https://doi.org/10.3406/lgge.1987.1964X

Ratinaud, P. (2014). *IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*.

Ratinaud, P., & Déjean, S. (2009). IRaMuTeQ: Implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. *Modélisation Appliquée Aux Sciences Humaines Et Sociales MASHS* (8-9. or.).

Reinert, A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.

Reinert, M. (1986). Un logiciel d'analyse lexicale. *Les Cahiers de l'analyse Des Données*, 11(4), 471-481.

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin of Sociological Methodology/ Bulletin de Méthodologie Sociologique*, 26(1), 24-54. https://doi.org/cbhfwpX

Schonhardt-Bailey, C., & Bailey, A. (2013). *Deliberating American Monetary Policy: A Textual Analysis*. The MIT Press. https://www.jstor.org/stable/j.ctt9qf5r7X

Sobczak, A., Debucquet, G., & Havard, C. (2006). The impact of higher education on students' and young managers' perception of companies and CSR: An exploratory analysis. *Corporate Governance*, 6(4), 463-474. https://doi.org/10.1108/14720700610689577

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 4.290-4.297.

Trigo, A., Marta-Costa, A., & Fragoso, R. (2021). Principles of sustainable agriculture: Defining standardized reference points. *Sustainability (Switzerland)*, 13(8). Scopus. https://doi.org/10.3390/su13084086

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244. https://doi.org/fz95kgX

Wijffels, J. (2019). *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package wervion 0.8.2. https://doi.org/10.32614/CRAN.package.udpipe