bat Soziolinguistika aldizkaria
GUREAN ATALA

Kode-alternantzia aztertzeko hizkuntza-teknologien ekarpena

Larraitz Uria, Iñaki Alegria eta Ander Corral
IXA taldea (UPV/EHU)
larraitz.uria@ehu.es / i.alegria@ehu.eus / andercorral95@gmail.com

Sarrera-data: 2017-09-03 / Onartze-data: 2017-10-17

Laburpena. Kode-alternantzia (code-switching) ohiko fenomenoa izan ohi da hizkuntza bat baino gehiago hitz egiten diren komunitate linguistikoetan. Elkarrizketa berean hiztun elebidunek edo eleaniztunek hizkuntza bat baino gehiago erabiltzean datza kode-alternantzia. Bat-batekotasuna du ezaugarri nagusia eta ahozko hizkuntzan ez ezik, hizkuntza idatzian ere agertzen da egun hain erabiliak diren sare sozialetan. Hala, sare sozialek eta hizkuntza-teknologiek eskaintzen dizkiguten aukerak baliatu nahi ditugu idatzizko mezuetan aurkitzen dugun kode-alternantzia aztertzeko. Artikulu honetan, hain zuzen ere, Hizkuntzaren Prozesamenduaren (HP) alorrean fenomeno hau aztertzeko garatzen ari garen metodologia eta hizkuntza-teknologiak aurkezten ditugu. Corpusa biltzeko nahiz etiketatzeko baliatuko ditugun metodoak hizkuntzarekiko independenteak izango dira. • ***Hitz gakoak:*** *Kode-alternantzia, Hizkuntzalaritza Konputazionala, Soziolinguistika, Hizkuntzaren Prozesamendua (HP).*

Abstract. In language communities where more than one language is spoken, a commonplace phenomenon is code-switching, which consists of bilingual or multilingual people using more than one language in the same conversation. This is a characteristic of spontaneous language use which occurs not only in speech but also in the written language on social media, the use of which is so widespread today. Thus it was decided to make use of the opportunities provided by social media and language technology to study the use of code-switching in written messaging. This article presents methods and language technology that are being used to study this phenomenon in the field of Language Processing (LP). Language-independent methods are used to compile and tag the corpus. • ***Key words:*** *Code switching, Computational linguistics, Sociolinguistic, Linguistic Processing.*

1. SARRERA

IXA ikerketa-taldeak hogeita zortzi urteko ibilbidea du Hizkuntzaren Prozesamenduaren (HP)[1] alorrean lanean. 1988. urtean sortu zen euskararako oinarrizko baliabide informatikoak garatzeko helburuarekin, eta ekarpen interesgarriak egin ditu euskal hizkuntza komunitatera (hala nola, eta batzuk aipatzearren, Xuxen zuzentzaile ortografiko-gramatikala, itzulpen automatikoko hainbat sistema, Euskal WordNet (BasWN) datu-base lexikala, morfologikoki etiketatutako Zientzia eta Teknologia Corpusa, sintaktikoki etiketatutako EPEC corpusa, etab.[2]).

IXA ikerketa-taldeak hogeita zortzi urteko ibilbidea du Hizkuntzaren Prozesamenduaren (HP) alorrean lanean.

Lan honekin, ikerketa-lerro honi heldu nahi izan diogu kode-alternantzia aztertzeko baliabide informatiko lagungarriak garatuz.

Urte hauetan guztietan, batez ere hizkuntza estandarrean eta elebakarrez idatzitako testuen azterketa eta prozesamendua bideratzera mugatu garen arren, azken urteotan hasiak gara bai aldaera dialektalak erakusten dituzten testuekin eta baita testu eleaniztunekin ere lanean (Etxeberria, 2016; Etxeberria et al., 2016a; Etxeberria et al., 2016b; Etxeberria et al., 2015; Etxeberria et al., 2014; Uria eta Etxepare, 2012 eta 2011; Uria et al., 2011). Izan ere, teknologia berriek eta Internetek azken hamarkada hauetan gure gizartean izan duten eraginaren ondorioz, komunikazio-moduak etengabe aldatuz doaz, aldaketa horien ondorioz ezaugarri berriak dituzten testuak sortu direlarik: testu dialektalak, testu informalak eta testu eleaniztunak, besteak beste.

Bestalde, azken urteotan hizkuntza-ukipena eta eleaniztasunaren inguruko azterketek interes handia piztu dute hainbat ikerketa-eremutan: hizkuntzalaritza, soziolinguistika, neurolinguistika edota hizkuntzaren prozesamenduaren alorretan, besteak beste. Arlo bakoitzean ikuspuntu edota helburu desberdinekin aztertzen da kode-alternantzia, eta gaur egun, jada, (orain gutxi arte ez bezala) teknologia berriek eskaintzen dituzten laguntzak baliatzea da ohikoena horrelako analisiak bideratzeko garaian (batez ere azterketa kuantitatiboak direnean).

Hala, hainbat ikerketa egin dira gai honen inguruan (Nguyen et al., 2016), eta guk ere, lan honekin, ikerketa-lerro honi heldu nahi izan diogu kode-alternantzia aztertzeko baliabide informatiko lagungarriak garatuz. Izan ere, ezaugarri berriak dituzten testu horien azterketa eta prozesamendua bideratzeko teknologiak garatzeak gure hizkuntza-komunitatean ekarpen interesgarriak egin ditzakeela uste dugulako ekin diogu erronka berri honi.

Baliatuko ditugun metodoak hizkuntzarekiko independenteak izango dira, eta lizentzia librean eskuragarri jarri dira.

1.1 Kode-alternantzia. Zer da?

Kode-alternantziaz hitz egiterakoan nahiz idazterakoan bi hizkuntza edo gehiago batera erabiltzen direnean gertatzen da, hau da, hiztun elebidun edo eleaniztunak elkarrizketa berean bi hizkuntza, hizkera nahiz erregistro, edo gehiago, txandakatzen dituenean.

Kode-alternantzia komunikazio-ekintza gisa deskribatu ohi da: hitz egiterakoan nahiz idazterakoan bi hizkuntza edo gehiago batera erabiltzen direnean gertatzen da, hau da, hiztun elebidun edo eleaniztunak elkarrizketa berean bi hizkuntza, hizkera nahiz erregistro, edo gehiago, txandakatzen dituenean. Eta hizkuntza bat baino gehiago kontaktuan daudenean maiz gertatu ohi den fenomenoa da kode-alternantzia.

Normalean, hiztegi edo lexiko mailako aldaketak egiten hasten da prozesua, hau da, hitz solteak aldatzen, eta hizkuntzen arteko kontaktua indartu ahala, egitura morfologiko nahiz sintaktikoak ere aldatzera pasatzen gara.

Kode-alternantzia maila desberdinetan gerta daiteke:

- Esaldi arteko kode-alternantzia (*inter-sentential switching*) esaten zaio hiztun batek kodea esaldika aldatzen badu, hau da, esaldi bat hizkuntza batean esan eta hurrengo esaldia beste hizkuntza batean esaten badu.
- Esaldi barruko kode-alternantzia (*intra-sentential switching*) esaten zaio kodea esaldi beraren barruan aldatzeari, hau da, esaldi bera hizkuntza batean baino gehiagotan osatzeari.

80. hamarkadatik aurrera gai honen inguruko aditu gehienek hiztun elebidun nahiz eleaniztunen hizketaldietan gertatzen den fenomeno naturaltzat hartu izan dute kode-alternantzia, eta esan izan dute ez dela hizkuntzaren gabeziari lotuta dagoen fenomenoa; aitzitik, hiztun elebidunek edo eleaniztunek helburu komunikatibo zehatz batzuekin erabiltzen duten fenomenoa dela diote, besteak beste, identitatearen markatzaile edo markatzaile sozial gisa edota baliabide diskurtsibo bezala.

Kode-alternantzia fenomeno zaharra da gure hizkuntza-komunitatean ere, Euskal Herrian. Ez da ohikoa euskaldun berri edo umeetan bakarrik, ama-hizkuntzak euskara eta gaztelania nahiz euskara eta frantsesa (Iparraldean) dituzten euskaldunen artean ere asko gertatzen den fenomenoa da. Adibide gisa, euskara-gaztelania eta euskara-frantsesa kode-alternantziei dagozkien adibide hauek ekarri ditugu hona[3]:

> *Daude…, daude* ***cardiacos****… eh? …eta ziren* ***una cuadrilla****.*
>
> *Zu zer moduz* ***en la Uni****? Eta zer moduz daude* ***los de clase****?*
>
> *Maritxuk izan zuen problema bat* ***avec la clavicule*** *Gaxinto ateratzen da* ***au balcon de la mairie***
>
> *Pagatzen zuen mila bortz ehun libera,* ***quinze cents francs****.*

Kode-alternantzia duten adibideak izango dira, hain zuzen ere, gure corpusa osatuko dutenak. Izan ere, corpusa izango da kode-alternantzia aztertzeko oinarri eta abiapuntu nagusia.

1.2 Kode-alternantziaren inguruko ikerketak

Euskal komunitatean kode-alternantziaren inguruko hainbat azterketa burutu izan dira, besteak beste:

- Etxebarriak (1998, 2004), Muñoak (1997) eta Rotaetxek (1994) euskara-gaztelania alternantzia aztertu zuten pertsona heldu elebidunetan eta Barreña, Ezeizabarrena eta Garcíak (2008) eta Ezeizabarrena eta Manterolak (2004), berriz, umeetan ematen den euskara-gaztelania kode-alternantzia.
- 90. hamarkadan euskara-gaztelania alternantzia erakusten duten testu batzuk bildu zituzten Esnaolak (1999), Etxabek (1993), Fernándezek (1992) eta Sarrionandiak (1992), baina Esnaolarena izan ezik, gainontzekoak argitaratu gabe geratu ziren.
- Etxabek (2005) gazteen lagunarteko hizkeran egiten den kode-alternantzia aztertu zuen. Zehazki, Getariako ahozko hizkera jasotzen duen lagin bat osatu zuen kode-alternantziak zein funtzio betetzen dituen argitzeko, hau da, kode-alternantzia zergatik eta zertarako egiten duten jakiteko.
- Epelde eta Oyharçabalek (2009) Iparraldeko hizkeratan gertatzen den kode-alternantzia ikertu zuten eta ACOBA proiektua jarri zuten martxan. Euskara-gaztelania eta euskara-frantsesa hitz egiten dituzten hiztun elebidunek kode-alternantzia perpaus barruan egiten dutenean nola moldatzen diren ulertzeko eta azaltzeko helburuarekin garatu zen proiektu hau. 2011 eta 2013 urte bitartean grabatu eta transkribatutako adibideekin corpus adierazgarri bat osatu zuten. 2.000 audio-grabazio biltzen ditu zehazki corpusak, audio bakoitza bere transkripzioarekin. Bildutako adibide guztiak bat-bateko elkarrizketa libreetatik hartuak dira, Euskal Herri guztiko 150 bat hiztunei egindako elkarrizketetatik.
- Etxabek (2010) kode-alternantziak Euskal Herriko bizpahiru herritako gazteengan duen eragina aztertu zuen, gaia hobeto ulertzeko oinarri teorikoak finkatuz, eta gazte horien lagunarteko ahozko nahiz idatzizko diskurtsoan dauden alternantzia-mailak, eraginpean dauden eremuak (lexikoa, sintaxia, morfologia…), maiztasuna eta abar deskribatuz, eta Zaldibia eta Ordizia herrien artean dauden aldeak aztertuz, bietan bizi den egoera soziolinguistikoa desberdina izanik.

Kode-alternantzia ez da hizkuntzaren gabeziari lotuta dagoen fenomenoa; aitzitik, hiztun elebidunek edo eleaniztunek helburu komunikatibo zehatz batzuekin erabiltzen duten fenomenoa dela diote adituek.

• Orreaga Ibarrak (2013 eta 2011) gazteek esaldi atributiboetan egin ohi duten euskara-gaztelania alternantzia aztertu zuen.
• Igartzabalek (2014) Gabiria bezalako arnasgune batean zer nolako kode-alternantzia dagoen aztertu zuen, bi belaunaldi ezberdinetako bina talde grabatuaz; zehazki, 18 lagun elkarrizketatu zituen fenomeno hau noiz, zertarako eta zergatik gertatzen den jakin aztertzeko beste urrats bat eman nahian.
• Lanttok (2015) kode-alternantziak Bilboko hirigunean daukan rola aztertu zuen, euskara eta gaztelaniaren ezaugarri linguistikoak kontuan hartuz.
• Perez-Gaztelu (2017) eta Perez-Gaztelu eta Zulaika (2014) ere euskal gazte-nerabeek lagun artean nola idazten duten aztertzen ari dira. Oraingoz, txat generoko mezuetan (testu *mintzidatzi*etan[4]) testumailako loturak nola egiten dituzten izan dute aztergai nagusi, hau da, testuen diskurtso-antolamendua.

Hauek dira, beraz, euskara-gaztelania eta euskara-frantsesa alternantzien inguruan argitaratu diren lanetako batzuk.

Aipatu azterketak, baina, datu-multzo (*dataset*) txikiak baliatuz egin izan dira, ikertzaile bakoitzak bere azterketak burutzeko edota bere helburuak lortzeko osatutako datu-multzo zehatzekin, hain zuzen. Alegia, ez dira corpus handi edo masiboak baliatu, ez eta HPren alorreko baliabide konputazionalak ere datuen bilketak eta analisiak egiteko. Aitzitik, azterketa hauek burutzeko, datuak eskuz bildu izan dira, normalean elkarrizketak eginez.

Metodo tradizionaletan oinarritutako azterketek, ordea, desabantaila nagusi bat izan ohi dute: hiztun parte-hartzaileek beraien hizkuntza-erabilera egokitu egin dezakete datu-biltzailearen helburuetara, eta hain zuzen ere, ikerketa linguistiko nahiz soziolinguistikoen xedea izan ohi da hiztunek egoera naturaletan, errealetan, sistematikoki behatuak ez direnean, nola hitz egiten duten ikertzea. Oro har, soziolinguistikaren alorrean tradizionalki gehien erabili izan diren metodoak behaketa, inkestak eta elkarrizketak izan dira, alegia, aurrez ongi prestatutako metodoak. Horrek, ordea, lan handia eta neketsua suposatzen du, eta denbora asko eskatzen duen prestaketa. Grabaketen transkripzioa eta transkribatutako adibideen eskuzko etiketatzea ere lan astuna eta garestia izaten da.

Horiek horrela, bildutako datu-multzoak txikiak izan ohi dira Hizkuntzalaritza Konputazionalaren alorrean erabili ohi diren corpusekin al-

deratuta. Izan ere, egun, HPko tresnetan oinarrituta, gero eta corpus handiagoak nahiz baliabide teknologiko sofistikatuagoak daude eskura azterketa linguistikoak egiteko. Hain zuzen ere, Hizkuntzalaritza Konputazionala hizkuntza teknologien laguntzarekin aztertzean datza, eta egun, jada, Humanitate Digitalen alorrean hizkuntza teknologien bidez aztertzeko joera nagusitzen da. Izan ere, datuak biltzeko eta aztertzeko irizpideak eta metodoak aldatuz doaz teknologia berriei esker.

HPren alorrean, esaterako, aukera dago sare sozialak eta teknologiak konbinatuz corpus handiak modu bizkorrean biltzeko. Izan ere, Web 2.0ren etorrerak testuinguru sozialetan sortzen diren testuak modu erraz, libre eta azkarrean eskuratzeko aukera handiak ekarri ditu, hau da, sare sozialak adibideak biltzeko eta corpusak osatzeko bitarteko interesgarri bilakatu dira. Hain zuzen ere, Interneten erabilera hedatua duten eta populazio eleaniztunetan erabiltzen diren sare sozialek (Twitter, Instagram, Facebook, foroak, blogak, etab.) hizkuntza-erabileren analisi sakonak eta eskala handikoak egiteko aukerak eskaintzen dizkigute egun, eta elkarrekintza sozialetan gertatzen diren komunikaziomota eleaniztunak analizatzeko datu-multzo ikaragarriak sortzeko aukera daukagu. Adibide horiek informazio-iturri aberatsak dira kode-alternantzia aztertzeko. Gainera, testuinguruari buruzko datuak ere bil daitezke, metadatuak (erabiltzaileari, garaiari, kokapenari, hizkuntzari, erregistroari, etab.i buruzko informazioa). Metadatuak biltzea ere garrantzitsua izan daiteke ikerketa sakon eta aberatsagoak egiteko aukera emango digutelako (datu horiek beharrezkoak dira, esaterako, azterketa soziolinguistikoak egin ahal izateko). Beraz, corpusarekin batera metadatuak ere biltzeko metodoak garatu ditugu guk lan honetan.

Metodo honek bere eragozpenak ere baditu, noski; batetik, teknologia hauek erabiltzen dituztenen unibertsoa mugatuagoa da (gazteen kasuan zabalkundea oso handia da), eta bestetik, horien alde publikoa bakarrik atzi daiteke. Horren ondorioz, sor daitekeen alborapen efektua (*bias*) kontuan hartzea garrantzitsua da.

Beraz, sare sozialetan *mintzidatzi*tako mezuekin osatutako corpusak teknologiaren laguntzarekin bildu nahi ditugu guk, kode-alternantzia aztertzeko aukera emango diguten corpusak osatzeko. Ikerketa honen abiapuntu gisa egin dugun lanean, euskara-gaztelania alternantzia duten adibideak bildu ditugu. Hala ere, hizkuntzarekiko independenteak diren baliabideak garatu ditugu, edozein hizkuntza edota hizkeren arteko alternantzia erakusten duten adibideak bilatzeko eta aztertzeko aukera izan dezagun.

Sare sozialetan mintzidatzitako mezuekin osatutako corpusak teknologiaren laguntzarekin bildu nahi ditugu, kode-alternantzia aztertzeko aukera emango diguten corpusak osatzeko.

Kode-alternantzia aztertzeko corpusak biltzea da, beraz, lan honen helburu nagusia, eta xede hori lortzeko funtsezko urratsak dira iturrien bilaketa/identifikazioa, corpus gordinen bilketa, hizkuntzen edota erregistroen identifikazioa eta corpusaren biltegiratzea.

Artikuluari jarraipena emateko, bigarren atalean fenomeno hau aztertzeko baliatu edota garatu ditugun hizkuntza teknologiak aurkeztuko ditugu, eta hirugarren atalean, berriz, lan honekin egin ditugun ekarpenak. Laugarrenean, azkenik, ondorioak eta etorkizuneko lanak aipatuko ditugu.

2. KODE-ALTERNANTZIA ETA HIZKUNTZA TEKNOLOGIAK

Hizkuntzaren Prozesamenduak azken urteotan kode-alternantziaren arloan egindako ekarpena aztertzea da atal honen helburua. Nguyen et al. (2016) artikuluan atal bat eskaintzen zaio "*Multilingualism and Social Interaction*" gaiari, non kode-alternantziaren azterketari heltzen zaion. Bertan erabiltzen diren metodo konputazionalez gain, Twitterrek eskaintzen dituen aukerak ere aipatzen dira; adibidez, aipatutako esaldi barruko eta esaldi arteko kode-alternantziaz gain, erabiltzaileen arteko kode-alternantzia aztertzeko aukera, hau da, aipamenak eta birtxiokatzeak egitean gertatzen diren alternantziak.

Kode-alternantzia aztertzeko corpusak biltzea da, beraz, lan honen helburu nagusia, eta xede hori lortzeko funtsezko urratsak dira iturrien bilaketa, corpus gordinen bilketa, hizkuntzen edota erregistroen identifikazioa eta corpusaren biltegiratzea. Eta aipatu urrats horiek egiteko baliatu diren metodoak eta aurkitu diren mugak aurkeztuko ditugu jarraian.

2.1 Metodoak

Sare sozialetako APIak (Aplikazioak Programatzeko Interfazea, ingelesez *Application Programming Interface*) erabiliz hainbat tresna konputazional gauzatu dira, batez ere bi erronka hauekin: i) hizkuntzaren identifikazioa eta ii) erregistroaren identifikazioa. Horretan sakontzeko oso interesgarria da 2016an egin zen "Computational Approaches to Code Switching" *workshop*a. Bertan, Molina et el.ek (2016) aurkeztu zuten lanean deskribatzen dira arloko erronka nagusiak, eta horien artean aipatutako biak nagusitzen dira (batzuetan beste izen edo ikuspuntu batekin). Erronka horiei teknika laguntzaileak edo post-prozesuak gehitzen zaizkie (testu-normalizazioa, POS etiketatzea, itzulpen automatikoa edota analisi sintaktiko automatikoak).

Hizkuntzaren identifikazioari dagokionez, gaia sakon landu da eta eszenatoki arruntetan ebatzitako problematzat jotzen dira testu-zati luze samarrak, testu normalizatuak eta hizkuntza ezagunak izatea. Erronka

izaten jarraitzen dute, ordea, oso gertu dauden hizkuntzak/dialektoak bereizi beharrak, testu-zatia oso txikia izateak edota testua normalizatuta ez egoteak (testu ez-normalizatuak ohikoak direlarik erregistro ez-formalean).

Aipatutako *workshop*ean txapelketa moduko bat egin zen, kode-alternantzia gertatzen denean hizkuntzaren identifikazioa ebazteko teknikarik onenak zeintzuk diren identifikatzeko asmoz (Molina et al., 2016)[5].

Gehien aipatzen den teknika ikasketa automatikoan oinarritutako CRFak (*Conditional Random Fields*) dira. Aipatu teknikaren azaleko definizioa emateko, Agirrezabal et al. (2017:51) lanean esaten dena ekarriko dugu hona:

> CRFek aurreikuspena egiteko sekuentziako elementu bakoitzari buruzko informazioa jasotzen dute eta, informazio horren arabera, predikzio bat proposatuko dute. Predikzio hori, ordea, ez da beste proposamen guztiekiko independentea izango, Viterbi izeneko algoritmo bat erabilita irteerako sekuentzia optimoa proposatuko du ereduak.

CRFekin batera, gaur egun konputazioan gailentzen ari diren sare neuronalen teknikak ere agertzen dira, LSTM (*Long Short Term Memory*) motakoak.

Kontuan hartu behar da aipatutako bi konputazio-ereduak egokiak direla "sekuentzien etiketatzea" izeneko problemarako, eta testu batean hizkuntza desberdinetan dauden zatiak problema horren barruan sartzen dira. Argitaratutako emaitza onenetan %90 inguruko doitasuna lortzen da hizkuntza-mugak bereizten.

Azken artikulu horretan ere sare sozialetatik bildutako kode-alternantzia fenomenoa erakusten duten hainbat corpus aipatzen dira hizkuntza/dialekto hauek barne hartzen dituztenak: gaztelania-ingelesa, arabieraren dialektoak eta alemana-turkiera.

Twitter sare sozialera soilik mugatu gara, beste sare sozialak kontuan hartu gabe, eta erregistro formala eta ez-formalaren arteko bereizketa automatikoa bazterrean utzi dugu momentuz.

2.2 Mugak

Baina lehen aipatu bezala, ez da ahaztu behar proposatutako bideak muga batzuk ere badituela:

- Sare sozialetako informazioa kasu askotan pribatua izaten denez, ezin izaten da bildu. Hori da helbururako interesgarriak liratekeen Whatsapparen kasua edota Facebookeko atal nagusien kasua, eta baita SMS mezuena ere.

• Pribatua ez izan arren batzuetan ez da erraza informazio-iturria identifikatzea edota dagokion edukia eskuratzea. Horretarako funtsezkoa da sare sozialak API bat eskaintzea, datuak programa informatikoen bidez eskuratu ahal izateko.
• Aurreko guztia kontuan izanda, Twitter geratzen zaigu iturburu posible (ia) bakar gisa: mezuak publikoak dira eta API[6] bat eskaintzen du.
• Twitterreko erabiltzaileen soziologiak esaten digu erabiltzaile gehiago daudela helduak direnak oso gazteak direnak baino. Gainera, ikusteko dago zein punturaino erabiltzaile euskaldunek erabiltzen duten erregistro informala Twitterren. Horren aurrean, Instagram ere erabiltzea izan daiteke irtenbide egokia, baina eskaintzen dituen APIak dauzkan mugak aztertu behar dira bideragarria den ala ez erabaki baino lehen.

Beraz, jarraitzen dugun metodoak berak hainbat muga uzten ditu agerian: batetik, Twitter sare sozialera soilik mugatu gara, beste sare sozialak kontuan hartu gabe, eta erregistro formala eta ez-formalaren arteko bereizketa automatikoa bazterrean utzi dugu momentuz (horretarako beste metodo matematiko batzuk erabili behar baitira). Aipatu erronka horiei ekitea izango da hurrengo urratsa.

3. GURE LANAREN EKARPENAK

Aurretik aipatutako oinarriekin euskara-gaztelania kode-alternantzia isla dezakeen corpus bat biltzen hastea erabaki genuen. Bagenekien lehenengo hurbilpen bat izango zela, diseinuan bertan hainbat muga aurreikusi genituelako.

3.1 Helburuak

Oro har, gure lanaren helburu nagusiak hauek izan dira:

• Euskara-gaztelania alternantzia erakusten duten corpus elebidunak biltzea Twitter sare sozialeko txioetatik, HPko metodo automatikoak baliatuz. Twitter sarea da iturburua eta behatutako kontu zehatzak eta traol (#hashtag) zehatzak erabili dira lehen bilketarako.
• Corpusak eskuz etiketatzea, eta ondoren ikasketa automatikoko teknikak erabili ahal izatea, kode-aternantzia non dagoen automatikoki detektatu ahal izateko.
• CRF bidez tresna automatiko bat eraikitzea eta lehenengo corpusa osatzea.

3.2 Jarraitutako urratsak

Kode-alternantzia erakusten duten adibideak dituzten txioak iragazketa heuristiko bidez eta manualki eskuratu dira.

Testu elebidun/eleaniztunen nahiz kode-alternantziaren detekzioa, identifikazioa eta bilketa bideratu ahal izateko, urrats hauek jarraitu ditugu:

1. Corpusa biltegiratu, antolatu eta indexatu da eta datu-base bat osatu da, hainbat metadatu ere bilduz.
2. Euskara-gaztelania alternantzia erakusten duen corpus sendo bat osatu da metodo automatikoak baliatuz. Horretarako, Twitterreko APIa (aplikazioetarako interfazea) erabiltzea da usuena eta hori da gure helburua. Aurretik identifikatutako kontu batzuk eta traol batzuk erabili dira. Behin lehenengo bilduma osatuta, errazagoa izango da bilduma zabaltzea antzekotasun-irizpideak erabiliz.
3. Corpusa etiketatzeko irizpideak finkatu dira, Molina et al. (2016) lanean jarraitutako irizpideak kontuan hartuz.
4. Bigarren urratseko corpusa oinarri hartuta, kode-alternantziari dagozkion adibideak aukeratu eta etiketatu dira, eskuz, irizpideei jarraituz.
5. Corpusa modu automatikoan etiketatzeko sistema prestatu da. Horretarako, etiketatutako lagina eta ikasketa automatikoan oinarritutako CRF metodoa erabili dira.
6. Aurreko urratsean garatutako sistemak ebaluatu eta hainbat ondorio atera dira (Corral, 2017).
7. Tresna horren laguntzarekin corpusaren lehenengo bertsio bat sortu da.

Zehaztasun guztiak Corralen (2017) lanean kontsulta daitezke.

3.3 Lortutako emaitzak

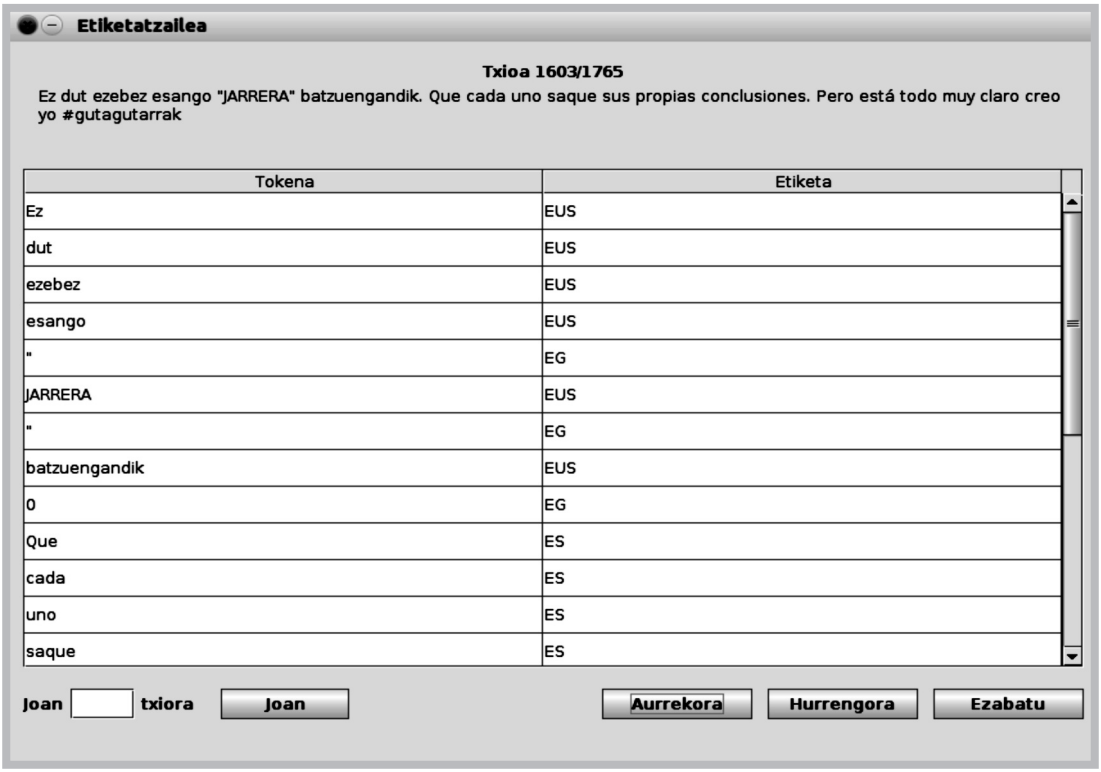
Egindako lanetik lau emaitza lortu dira, etorkizuneko lanetarako eta bestelako ikerlarientzat interesgarriak izan daitezkeenak: anotaziorako eskuliburua (3.3.1. puntua), eskuz anotatutako corpusa (3.3.2. puntua), txioak biltzeko tresna (3.3.3. puntua) eta automatikoki bildutako corpusa (3.3.4. puntua).

3.3.1 Anotaziorako eskuliburua

Anotazio-prozesua osorik azaltzen duen eskuliburu bat osatu da. Anotatzaileak interfaze sinple baten bidez irakurtzen ditu etiketatu beharreko adibideak (1. irudia), eta hitz bakoitzari dagokion etiketa esleitzen dio. Guztira, lehen fase honetan, 8 etiketa definitu dira:

- ES, gaztelaniazko hitza adierazteko.
- EUS, euskarazko hitza adierazteko.
- ID, Twitterreko erabiltzaile baten identifikatzailea adierazteko.
- URL, esteka bat adierazteko.
- IE, izen entitate bat adierazteko.
- NH, bi hizkuntzak token berdinean nahasita daudela adierazteko.
- ANB, testuinguruaren arabera hitza anbiguoa dela adierazteko.
- EG, etiketarik gabeko tokena adierazteko (hau da, aurreko etiketak jarri ezin diren kasuetarako).

1. irudia: Anotatzeko interfazea



3.3.2 Eskuz anotatuko corpusa

Kode-alternantzia fenomenoa aztertzeko, corpusa bildu eta eskuz anotatu da.

Sortutako corpusa errealitatetik ahalik eta hurbilen egotea garrantzitsua da, kode-alternantziaren identifikazioa eta analisia modu fidagarri eta eraginkor batean egin ahal izateko. Izan ere, gero eta corpus errepresentagarriagoa eskuratu, orduan eta emaitza hobeagoak lortuko dira.

Horregatik, kode-alternantzia erakusten duten adibideak dituzten txioak iragazketa heuristiko bidez eta manualki eskuratu dira.

Twitterren, normalean, bai euskaraz eta bai gaztelaniaz idazten duten hainbat erabiltzaile bilatu dira. Izan ere, erabiltzaile bakoitzak bere estilo propioa dauka idazteko orduan eta horien idazteko modua patroi ez bihurtzeko, aniztasuna ezinbestekoa da. Pertsonaia publikoen txioak, euskal telebistako saioei lotutako erabiltzaileen txioak eta momentuko euskarazko *trending topic* nagusiak analizatu dira, ahalik eta corpus errepresentagarri eta anitzena osatzeko.

Adibide egokiak bilatzeko, lagungarria izan zaigu *umap.eus* webgunea[7]. Webgune honek Twitterreko euskarazko jarduna monitorizatzen du eta rankingak bistaratzen ditu.

APIak murrizketak ditu zerbitzatu ditzakeen txio kopuruan. Ondorioz, erabilitako metodoarekin gehienez ere 3.200 txio eskura daitezke erabiltzaile batetik. Hala ere, kopuru hau nahikoa izan da gure proiekturako, erabiltzaileen aniztasuna lortu nahi izan baititugu. 1. taulak erakusten duen moduan, guztira 424 erabiltzaile desberdinen 1.765 txio bildu dira, eta adibide guzti horien artean 959 txiotan agertu da kode-alternantzia (kode-alternantziarik gabeko adibide batzuk ere, noski, ezinbestekoak izan dira ikasketa automatikoa modu egokian burutzeko).

1. taula: Eskuz anotatuko corpusaren ezaugarri orokorrak

Txioak guztira	1765
Gaztelanizko txioak	467
Euskarazko txioak	337
Code-switching txioak	959
Bestelako txioak	2
Erabiltzaileak guztira	424
Gaztelaniazko erabiltzaileak	249
Euskarazko erabiltzaileak	172
Code-switching erabiltzaileak	85

Eskuliburuan zehaztutako etiketekin anotatu da corpusa, eta 2. taulak laburbiltzen ditu etiketa bakoitzari dagozkion kopuruak.

Ikasketa automatikoko beste metodoekin egiten den bezala, sortu den tresna ebaluatu egin da garatu den tresnaren kalitatea zenbaterainokoa den jakiteko.

2. taula: Anotatzeko etiketa bakoitzari dagokion txio kopurua

Token kopurua	34771
ES token kopurua	13891
EUS token kopurua	7234
ID token kopurua	1399
URL token kopurua	740
EG token kopurua	7939
ANB token kopurua	112
NH token kopurua	82
IE token kopurua	3353

3.3.3 Txioak biltzeko tresna

Bildutako eta anotatutako corpusa oinarri hartuta, aipatu CRF algoritmoa aplikatu da txio berriak sailkatzeko eta etiketatzeko tresna bat sortzeko asmoz. Tresna horretan oinarrituta, geroxeago azalduko den web-aplikazioa eraiki da txioak monitorizatzeko asmoarekin.

Ikasketa automatikoko beste metodoekin egiten den bezala, sortu den tresna ebaluatu egin da garatu den tresnaren kalitatea zenbaterainokoa den jakiteko. Bi ebaluazio egiten dira: lehena, ikasten ari den datuekin egiten da "ebaluazio gurutzatu" (*cross-validation*) teknikaren bidez eta bigarren bat, zorrotzagoa dena, apartatutako datuen (test datuak) gainean egiten dena. Ikasketa modu egokian egin bada, bi emaitzak antzekoak izan behar dira.

3. taulan ikus daitekeenez, ikasitako tresnak hitzen %93-94ren artean ondo egokitzen die ondo dagokien etiketa. Txio osoari begira (txio baten barruko etiketa guztiei begira), berriz, emaitzak dibergenteagoak dira bi ebaluazioetan, asmatze-tasa %41-47ren artean dago eta ebaluazio osoa (txioa zein hizkuntzatan dagoen edo alternantzia duen) %87-92ren artean. Azken neurri hori hobetu liteke horretarako propio sortutako sailkatzaile gehigarri bat garatuz, baina hori ere etorkizunerako utzi da.

3. taula: Tresnaren ebaluazioari dagozkion emaitzak

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Ebaluazio fasea	0.94142	0.47727	0.92045
Test fasea	0.93072	0.41477	0.86932

Egindako aurreko lanean oinarrituta, txioak biltzeko web-aplikazio bat garatu da, proiektuaren produktu eta helburu nagusia izan dena. Aplikazioaren betekizuna txioak denbora errealean atzitzea da eta hauetan kode-alternantzia fenomenoa detektatzea. Txioak denbora errealean lor-

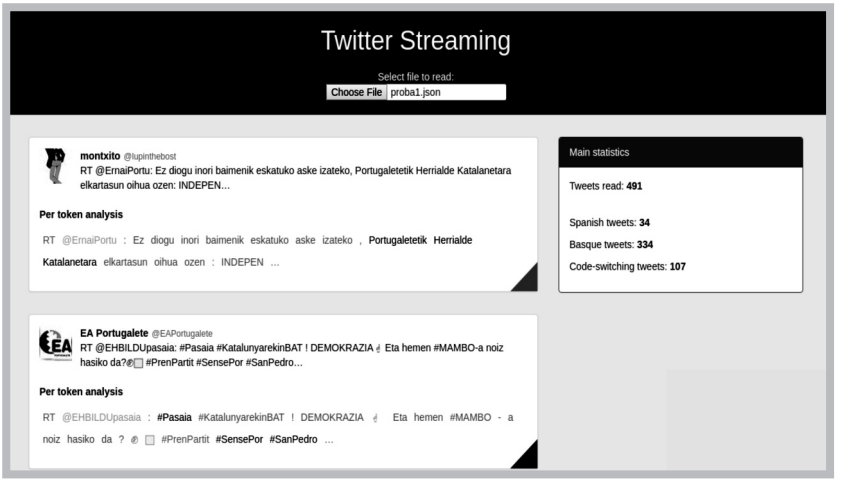
tzeko Twitterreko APIaren *streaming* funtzioa erabili da. Funtzio honek txioak erabiltzaile konkretu baten arabera edo hitz konkretu baten arabera iragazteko aukera ematen du. Aplikazioa *Python* programazio lengoaia erabiliz garatu da eta kodea GitHub bidez atzigarri dago[8].

Aplikazio honen interfazeak ematen diogun fitxategia irakurtzen du eta txioak banan-banan erakusten ditu Twitterreko itxura bera jarraituz. Txio bakoitzari bi analisi mota egiten zaizkio. Batetik, tokenak analizatzen dira eta token bakoitzari dagokion etiketa esleitzen zaio, kolore desberdinez adierazita; esaterako, gorria, ES etiketa duen tokenarentzat; berdea, EUS etiketa duen tokenarentzat; horia, ID etiketa duenarentzat; urdin argia, URL etiketa duenarentzat, etab. Bestetik, txio bakoitzari kode-alternantziari dagokion etiketa orokor bat esleitzen zaio. Etiketa orokor horretan hiru kolore bereizten dira, gorria, berdea eta morea. Kolore gorriak adierazten du gaztelaniazko txioa dela, hau da, hitz guztiek gaztelaniazko etiketa dutela, token bereziak (identifikatzaileak, zenbakiak, estekak…) salbu; berdeak euskarazko txioa dela adierazten du, hau da, hitz guztiei euskarazko etiketa esleitu zaiela, token bereziak salbu; eta azkenik, moreak markatutako txioan kode-alternantzia dagoela adierazten du. Horretarako, algoritmo sinple bat aplikatzen da: txio batean euskarazko eta gaztelaniazko etiketak baldin badaude, kode-alternantzia dago, nahiz eta hitz bakarra egon (Corral, 2017).

Tresna hau hobetzeko beharra aurreikusten dugu, eta orain egiteke utzi diren atazak bukatzeko eta integratzeko beharra ere bai (besteak beste, txioen sailkapen globala edota mezu formal/ez-formalen bereizketa).

Aplikazioak sarrerako fitxategia irakurtzen duenean, zenbait datu orokor ere ematen ditu; hala nola, irakurritako txio kopurua, euskarazko txio kopurua, gaztelaniazko txio kopurua eta kode-alternantzia duten txioen kopurua. 2. irudiak erakusten digu web-aplikazioaren interfazea.

2. irudia: Txioak biltzeko web-aplikazioaren interfazea



Hizkuntzaren Prozesamenduaren alorretik ekarpen interesgarriak egin ditzakegu kode-alternantziaren analisi sakonetan laguntzeko, baliabide informatikoak garatuz.

3.3.4 Automatikoki bildutako corpusa

Lehenengo fasean bildutako corpusak eskuz etiketatu dira eta ikasteko zein ebaluaziorako erabili dira. Behin lehenengo fase hori bukatuta, gai gara corpus berriak osatzen jarraitzeko. Ahalmen horren erakusgarri da 2017ko irailaren 26 eta 28 bitartean egindako bilketa berria. *umap.eu* guneko traolak kontsultatuta 12.535 txio bildu dira eta horietako 1.858ri esleitu zaie kode-alternantzia kategoria[9].

Erabili diren traolak hauek izan dira: #KataluniarekinBat, #BaimenikEz, #Demokrazia, #GukEreMambo, #65SSIFF, #Zinemaldia, #30egun 30doinu, #lekaixoka, #kmk17, #alderdieguna, #Eraldaketa, #GudariEguna17, #Altsasu.

Eta hauexek dira kode-alternantzia kategoria duten hainbat txioren adibideak:

> 1. "Mila esker Kilometroak #lekaixoka Muchas gracias a toda la peña! EGURRA#PresoakEtxera#ELKARTASUNTAUPADAK@MaukaMusik@…"
> 2. "Paz Vega,nerviosa e ilusionada con el Premio al Cine Latino que recibirá hoy en el #65SSIFF Zorionak!! https://t.co/iM885IrVOc"
> 3. "RT @NHEkhimena: Nabarra #KataluniarekinBat independentziaren alde! \nNabarra #ambCatalunya per la independència!…

Kontuan hartu behar da modu automatikoan sailkatu diren txioak direla, eta beraz, gerta liteke txio batzuk ez izatea kode-alternantziaren adibide adierazgarriak. Bigarren txioan, esaterako, euskarazko hitz bakar bat dago (*Zorionak*) eta gure programaren arabera, hitzez hitz ondo sailkatu den arren, hobe litzateke alternantziatzat ez hartzea.

Azkenik, eta proiektu handiago bati begira, tresna hau hobetzeko beharra aurreikusten dugu, eta orain egiteke utzi diren atazak bukatzeko eta integratzeko beharra ere bai (besteak beste, txioen sailkapen globala edota mezu formal/ez-formalen bereizketa).

4. ONDORIOAK ETA ETORKIZUNEKO LANA

Azken urteotan sare sozialen erabileraren hedadura ikaragarria izan da eta ikusi dugu kode-alternantzia hizkuntza mintzatuan ez ezik, idatzian ere asko gertatzen den fenomenoa dela, arlo desberdinetako ikertzaileon artean interesa piztu duen fenomenoa delarik hauxe.

Hizkuntzaren Prozesamenduaren alorretik ekarpen interesgarriak egin ditzakegu kode-alternantziaren analisi sakonetan laguntzeko, baliabide

informatikoak garatuz. Eta proiektu honen helburua, hain zuzen ere, kode-alternantziaren analisia eta prozesamendua bideratzeko, errazteko edota azkartzeko baliabide informatikoak garatzea izan da, hau da, corpus elebidun nahiz eleaniztunetan hizkuntza/hizkera batetik bestera noiz aldatzen den detektatzeko gai diren sistemak garatzea.

Online inguruneetatik testu eleaniztunak eskuratzeko, testuak multzo txikitan eta epe laburretan hartu izan dira. Datu edo testu horien azterketa automatikoa zaila izan da hizkuntza gehienetan, kontuan izanik orain artean oso baliabide informatiko gutxi egon direla eskura helburu horretarako. Hizkuntzalaritza Konputazionalaren alorrean, interes handia piztu da, egun, testu elebidun edota eleaniztunen analisi automatikoan, baina horretarako beharrezkoak dira corpus elebidun/eleaniztunak detektatzeko, ulertzeko eta aztertzeko gai diren HPko tresna bereziak.

Eta artikulu honen helburu nagusietako bat horixe izan da, hain zuzen ere. Hala, Twitterreko txioen bilketa bat egin eta eskuz etiketatu ondoren, txioak biltzeko tresna bat garatu da. Izan ere, gure ustea da corpus elebidunak/eleaniztunak bildu eta horiek prozesatzeko gai izateak aukera emango digula kode-alternantziaren fenomenoa ikerketa-lerro edota ikuspuntu desberdinetatik aztertzeko, eta ekarpenak hainbat diziplinatan egiteko. Horregatik, ikerketa-lerro berri bat abiarazi nahi izan dugu lan honekin.

Hala ere, konturatzen gara programak oraindik hainbat hobekuntza behar dituela, eta epe motzean hobekuntza horiek txertatzen saiatu nahi dugu. Hala, aurrera begira, egin beharreko lanen artean aurreikusten ditugu, jada, hainbat azterketa posible. Esaterako, Hizkuntzaren Prozesamenduaren ikuspuntutik, garatu dugun aplikazioa osatzea eta publiko jartzea da helburu nagusia, gai honetan interesa dutenek erabil dezaten. Hizkuntza desberdinak eta alternantzia identifikatzeaz gain, erregistroak (formala / ez-formala) ere bereiztea nahi dugu eta baita Twitter ez den beste sare sozial batzuetatik ere kode-alternantzia duten adibideak eskuratzea (Instagram izango da aztertuko dugun aukera bat, APIa eskaintzen duelako).

Hizkuntzalaritzaren ikuspuntutik, berriz, interesgarria izango da biltzen den corpusa gramatikalki aztertzea, hau da, hizkuntza matrizea eta hizkuntza txertatuak identifikatu eta kode-alternantziaren ezaugarri gramatikalak aztertzea, informazio lexikoaz gain morfosintaktikoari ere erreparatuz.

Azterketa honetan euskara-gaztelania alternantziara mugatu garen arren, erabili diren metodo guztiak hizkuntzarekiko independenteak dira, eta beraz, beste kode-alternantzia batzuk prozesatzeko gai dira.●

OHARRAK

1. Hizkuntzaren Prozesamenduak (HP) eta Lengoaia Naturalaren Prozesamenduak (LNP) ikerketa-alor berdinari egiten diote erreferentzia, sinonimoak dira.

2. http://ixa.eus/Produktuak

3. Euskara-gaztelania alternantzia erakusten duten adibideak Ibarra (2003) lanetik hartuak dira eta euskara-frantsesa alternantzia erakusten dutenak, berriz, ACOBA proiektutik (Epelde eta Oyharçabal, 2009).

4. Perez-Gazteluk *mintzidatzi* terminoa erabiltzen du, hain zuzen ere, sare sozialetan eta txat mezuetan hitz egiten dugun modu berean idazten diren mezuei erreferentzia egiteko. Mezu horiek ahozko hizketarik oso gertu daude eta ahozko hizkeran ohikoa da hizkera ez-formala. Adierazkortasuna, bat-batekotasuna, arrunkeriak esatea edota silabak jatea dira, besteak beste, hizkera informalaren ezaugarri batzuk. Gainera, dialektoen erabilera handia izan ohi da (gure kasuan, euskalkiena) eta kode-alternantzia ere askotan egiten da *mintzidatzi*etan. Ezaugarri horiek betetzen dituzten mezuak izango dira gure azterketaren oinarria.

5. *Workshop*ean erabiltako corpusak atzigarri daude helbide honetan: http://care4lang1.seas.gwu.edu/cs2/call.html

6. https://dev.twitter.com/rest/public

7. https://umap.eus/

8. https://github.com/anderleich/CodeSwitchingDetection

9. Bildutako txioak https://github.com/anderleich/CodeSwitchingDetection webgunean daude eskuragarri, *Results* izeneko karpetan.

ERREFERENTZIAK

Agirrezabal M., Alegria I., Hulden M. (2017). "Poesiaren eskantsio automatikoa: bi hizkuntzen azterketa", in II. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Giza Zientziak eta Artea. UEU. http://www.ueu.eus/download/liburua/IKERGAZTE.2017.GIZAZIENTZIAKetaARTEA.pdf#48

Barreña A., Ezeizabarrena, M., García, I. (2008). "Entzundako hizkuntzaren eragina haur euskaldun txikien gramatika-garapenean", in *Gramatika jaietan*. Patxi Goenagaren omenez, Artiagoitia X. y Lakarra J. A. (eds.). "Julio Urkixo" Euskal Filologi Mintegiaren Urtekariaren Gehigarriak, LI, 107- 127, Donostia-Bilbao: Diputación Foral de Gipuzkoa & EHU/UPV.

Corral, A. (2017). Twitter-eko txioen Code Switching detekzio automatikoa. Gradu Amaierako Proiektua (GAP). Informatika Fakultatea. (laster publikatzeko: https://addi.ehu.eus)

Epelde, I. eta Oyharçabal, B. (2009). "Code Switching en las variedades orientales del vasco". IKER–UMR 5478, Basque Text and Language Study Center, Baiona.

Esnaola, I. (1999). *Gazte euskaldunen lagun arteko hizkera hemen eta orain*. Bilbao: UEU.

Etxabe Azkarate, K. (2010). "Kode-alternantzia lagunarteko gazte-hizkeran. Zaldibia eta Ordiziako kasuak", in HIZNET Hizkuntza Plan-

gintza ikastaroa. http://hiznet.asmoz.org/ikerketa_lanak/images/stories/ikerketa_lanak/2010/Kode_alternantzia_lagun arteko_gazte_hizkeran.pdf

Etxabe Aramendi, M.A. (1993). *Hizkuntza aukeraketa edo "code switching"a euskaldunen artean*. Hizkuntza Politikarako Idazkaritza Nagusia (HABE liburutegia, Donostia. Argitaratu gabea).

Etxabe Iribar, A. (2005). "Kode-alternantzia lagun arteko gazte-hizkeran", in HIZNET Hizkuntza Plangintza ikastaroa.

http://hiznet.asmoz.org/ikerketa_lanak/images/stories/ikerketa_lanak/AmetsEtxabe.pdf

Etxebarria, M. (2004). "Bilingüismo y code-switching en el País Vasco", in Las lenguas en la Europa Comunitaria II, Amsterdam/Atlanta: Rodopi, 55-66.

Etxeberria, I. (2016). *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz*. Doktore-tesia. Informatika Fakultateko LSI saila. UPV/EHU. https://addi.ehu.es/bitstream/handle/10810/19492/TESIS_ETXEBERRIA_UZTARROZ_IZASKUN.pdf?sequence=1

Etxeberria, I., Alegria, I., Uria, L. and Hulden, M. (2016a). "Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene", in LREC 2016 conference. pp: 1064-1069.

Etxeberria, I., Alegria, I., Uria, L. and Hulden, M. (2016b). "Combining Phonology and Morphology for the Normalization of Historical Texts", in Latech 2016: Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. At ACL 2016. ISBN 978-1- 945626-09-8.

Etxeberria I., Alegria I. & Uria L. (2015). "Induction of Phonology and Morphology for the Normalization of Historical Texts", in the 22nd International Conference on Historical Linguistics, Naples, 27-31 July 2015.

Etxeberria I., Alegria I., Hulden M. & Uria L. (2014). "Learning to map variation-standard forms using a limited parallel corpus and the standard morphology", in the Procesamiento del Lenguaje Natural, revista num. 52, pp. 13-20 (ISSN edición impresa: 1135-5948) (ISSN edición digital: 1989- 7553).

Ezeizabarrena, MJ. & Manterola, J., (2004). "La mezcla de códigos (euskera-castellano) en el habla infantil", in EUROSLA2004.

Fernández Alcober I. (1992). "Kode aldaketa Durangoko gazteen hizkeran". Corpus inédito euskera/español.

Ibarra Murillo, O. (2013). "El code switching vasco-castellano en oraciones atributivas de hablantes jóvenes (The Basque-Castillian code

switching in attributive sentences by young speakers)", in Oihenart, 28 (ISSN: 1137-4454, eISSN: 2255-1050); pp. 115-130.

Ibarra Murillo, O. (2011). "Sobre estrategias discursivas de los jóvenes vascohablantes: causas que motivan el cambio de código", in Eusko Ikaskuntza Oihenart, Cuadernos de Lengua y Literatura nº 26; Terceras jornadas de Lingüística vasco-románica: teoría y análisis, pp. 277-298.

Igartzabal Bidegain, I. (2014). "Kode-alternantzia eta hizkuntza-ohiturak Gabiriako bi belaunalditan", in Bat: Soziolinguistika aldizkaria, ISSN 1130-8435, Nº. 92-93, 2014 (Ejemplar dedicado a: VII. Hausnartu euskal soziolinguistika sariak), pp. 229-251.

Lantto, H. (2015). *Code-switching in Greater Bilbao: A bilingual variety of colloquial Basque*. Doctoral dissertation. University of Helsinki, Faculty of Arts, Department of Modern Languages. ISBN:978-951-51-1184-5.

Molina, Giovanni, Nicolas Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, and Mona Diab. (2016). "Overview for the second shared task on language identification in code-switched data", in EMNLP 2016): 40.

Muñoa I. (1997). "Pragmatic Functions of Code-Switching among Basque-Spanish Bilinguals", in Actas del I Simposio Internacional sobre Bilingüismo, Universidad de Vigo, 528-541.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. (2016). "Computational sociolinguistics: A survey", in *Computational Linguistics*.

Perez-Gaztelu, E. (2017). "CS metaforikoen eginkizun pragmatiko-diskurtsiboak euskaldun "zahar" gazteen txat mintzidatzietan". Hizkuntzalari Euskaldunen III. Topaketa. UEU Baiona.

Perez-Gaztelu, E. eta Zulaika, E. (2014). "Gzteak lgnrtn idztn: Mintzidatzien antolamendua". I. Aduriz & R. Urizar (ed.), *Euskal hizkuntzalaritzaren egungo zenbait ikerlerro*. Hizkuntzalari euskaldunen I. topaketa. Bilbo: Udako Euskal Unibertsitatea, 111-133 or.

Rotaetxe, K. (1994). "Alternance codique et langue minoritaire", in Martel P., J. Maurais (eds.), Mélanges offerts à J.C Corbeil: *Langues et Sociétés en Contact*, Tübingen, Max Niemeyer Verlag, 395-408.

Sarrionandia, B. (1992). "Kode alternantzia edo 'Code-switching': hiztun euskaldunduen arteko komunikazio estrategien tipologia batetarantz". Corpus inédito euskera/español.

Uria, L. eta Etxepare, R. (2012). "Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: BASYQUE aplikazioa", in Lapurdum, Euskal ikerketen aldizkaria, nº XVI, pp. 117-135 (ISSN elektronikoa 1965-0655).

Uria, L., Hulden, M., Etxeberria, I. & Alegria, I. (2011). "Recursos y métodos de sustitución léxica en las variantes dialectales en euskera", in the Proceedings of the Workshop on Iberian Cross- Language Natural Language Processing Tasks (ICL 2011), pp. 70-76 (ISSN: 1135-5948). (Proceedings: http://ceur-ws.org/Vol-824).

Uria, L. eta Etxepare, R. (2011). "BASYQUE: Aplicación para el estudio de la variación sintáctica", in Revista Linguamática (ISSN: 1647-0818), V.3, nº 1, 35-44.