Fontes

Linguae Vasconum

50 urte. Ekarpen berriak euskararen ikerketari ¶ Nuevas aportaciones al estudio de la lengua vasca

Sintaktikoki etiketatutako euskarazko corpus historikoa eraikitzen

AINARA ESTARRONA, IZASKUN ETXEBERRIA, RICARDO ETXEPARE, MANUEL PADILLA-MOYANO, ANDER SORALUZE

Argitaratzaileak / Editores

Ekaitz Santazilia

Dorota Krajewska

Eneko Zuloaga

Borja Ariztimuño

Gobierno de Navarra
Nafarroako Gobernua

Sintaktikoki etiketatutako euskarazko corpus historikoa eraikitzen

Building a syntactically annotated historical corpus of Basque

AINARA ESTARRONA
HiTZ zentroa (Ixa taldea), Euskal Herriko Unibertsitatea UPV/EHU
ainara.estarrona@ehu.eus

IZASKUN ETXEBERRIA
HiTZ zentroa (Ixa taldea), Euskal Herriko Unibertsitatea UPV/EHU
izaskun.etxeberria@ehu.eus

RICARDO ETXEPARE
Centre national de la recherche scientifique (CNRS) – IKER (UMR 5478)
r.etxepare@iker.cnrs.fr

MANUEL PADILLA-MOYANO
Centre national de la recherche scientifique (CNRS) – IKER (UMR 5478)
manuel.padilla@iker.cnrs.fr

ANDER SORALUZE
HiTZ zentroa (Ixa taldea), Euskal Herriko Unibertsitatea UPV/EHU
ander.soraluze@ehu.eus

JASOA: 2019/07/31 | BEHIN-BEHINEAN ONARTUA: 2019/09/20 | BEHIN BETIKO ONARTUA: 2019/12/17

Artikulu hau Iñaki Alegriari eskaini nahi diogu, bere karrera euskal hizkuntzalaritza konputazionalaren garapenaren zerbitzuan emateagatik. Iñakiren ibilbide oparoari esker, hain zuzen ere, aritzen ahal gara proiektu honetan. Gainerakoan, lan hau Frantziako *Agence nationale de la recherch*ek (ANR) eta Espainiako Zientzia, Berrikuntza eta Unibertsitate Ministerioak (MICINN) finantzatutako BIM (*Basque In the Making: a historical look at a European language isolate*) eta SAHCOBA (*Syntactically Annotated Historical Corpus in Basque*, RTI2018-098082-J-I00) proiektuen barruan kokatzen da.

Santazilia, E. *et al.* (arg./eds.), *Fontes Linguae Vasconum 50 urte,* pp. 237-251. or.
DOI: 10.35462/fontes50urte.15

LABURPENA

Lan honetan morfologikoki eta sintaktikoki etiketatutako euskararen corpus historikoaren proiektua aurkezten dugu. Corpusak XV-XVIII. mende bitarteko euskalki guztietako ekoizpen idatzi esanguratsuena besarkatuko du, haren tamaina milioi bat hitz ingurukoa izanik. Etiketatze morfosintaktikoak hainbat konplexutasun-mailatako bilaketa sistematikoak ahalbidetuko ditu: lemaren, formaren, kategoria gramatikalaren, tasun morfosintaktikoaren, bai eta zenbait egitura sintaktikoren arabera ere. Halaber, corpusa metadatu-sorta batekin hornituko da, irizpide sozio-historikoen araberako bilaketak ere posible egiteko. Euskararen lehen corpus historiko anotatua sortzeaz gainera, proiektu honi esker hizkuntzaren prozesamenduko tresnak gaur egungo euskara batutik urruntzen diren barietateekin trebatuko dira. Harago, proiektu honek etorkizuneko euskara historikoaren corpusgintzarako oinarriak finkatu nahi lituzke.

Gako-hitzak: Humanitate Digitalak; corpus historikoa; Hizkuntzaren Prozesamendua (HP); sintaxi diakronikoa.

ABSTRACT

In this paper we present an ongoing project to build a morphosyntactically annotated historical corpus of Basque. The corpus will have around one million words, encompassing the most significant written production of Basque between the 15th and 18th centuries. Morphosyntactic tagging will allow for systematic searches at different levels of complexity: lemma, form, part of speech, morphosyntactic feature, and also a number of syntactic constructions. In addition, a set of metadata will enable searches based on socio-historical criteria too. Beyond being the first annotated historical corpus of Basque, through this project tools for language processing will be improved by analysing Basque historical varieties more or less distant from present-day standard Basque. Moreover, this project aims to establish a model for further works in historical corpora of Basque.

Keywords: Digital Humanities; historical corpus; Natural Language Processing (NLP); diachronic syntax.

1. Sarrera

Beharbada hizkuntza ez-indoeuroparra eta Europako esparruan hizkuntza isolatua izateagatik, euskarari buruzko ikerketa asko eta asko jatorriaren, antzinakotasunaren edo ahaidegoaren ingurukoak izan dira. Paradigma horretan sartzen dira, besteak beste, eusko-iberismoaren teoria (Schuchardt, 1908), hipotesi afro-asiarra (Schuchardt, 1914) edo teoria kaukasiarra (Uhlenbeck, 1924; Lafon, 1951, 1952).

XX. mendearen bigarren erdialdetik aurrera euskal hizkuntzalaritza historikoa aztertzeko ikuspuntua zehaztu eta zorroztu egin zen. Paradigma berria testuen azterketa arretatsuan eta aldakortasun dialektalari aplikatutako metodo konparatiboan oinarritu zen, ondorengo konparazioetarako oinarri sendoa ezarriz (ik. Trask, 1997). Barne-konparazio horren lorpenik aipagarrienak aitzineuskararen sistema fonologikoaren irudikapen zehatza (Mitxelena, 1977 [1961]), eta horretatik gaur egungo sistemara iristeko gertatutako aldaketa fonologikoen deskribapena dira. Euskara arkaikoaren aditz-sistemaren ikerketa beste mugarri bat izan zen (Lafon, 1944).

Mitxelenaren ondorengo lanik garrantzitsuenen helburua haren sistemari esker berreraikitako hizkuntzaren faseetatik harago joatea izan da (Lakarra, 1995, 2005, 2006). Orain arte euskararen munduan *hizkuntzalaritza historikoa* deitu izan dena mota honetako ikerketan oinarritua izan da (Martínez-Areta, 2013). Berreraikitze-lanari esker, badakigu aitzineuskarak seguruenik egungo euskaratik aldentzen diren ezaugarri tipologikoak zituela, esaterako, VSO hitz-ordena (Gómez & Sainz, 1995). Halaber, baliteke ergatibotasuna berrikuntza izatea: haren arrasto morfologikoa partikula lokatibo batean ikus daiteke (Lakarra, 2006). Aditzaren forma jokatu konplexuak ere nahiko berriak direla esan daiteke; numero- eta datibo-komunztaduraren sorrera, adibidez, testuetan araka daitezke. Orobat gertatzen da determinatzaile-sistemarekin (erakusleetatik eratortzen da [Manterola, 2015]), edo numero morfologikoaren kategoriarekin.

Beraz, euskara inguruko hizkuntzetatik bereizten duten ezaugarri tipologiko gehienak idatzizko lekukotasunetan iker daitezke, hizkuntzalaritza diakroniko modernoaren metodo eta baliabideak erabilita. Honek guztiak barne hartzen du inguruko hizkuntza erromantzeekiko ukipenagatik gertatutako aldaketak sakon aztertzea, bai eta aldakortasun dialektala ere. Izan ere, dialektologia hizkuntzalaritza diakronikoaren ikergai naturala da, eta euskararen hizkuntzalaritza historikoarentzat balio erantsia hartzen du, lekukotasun idatzirik ez duten hizkeretan gramatika-ezaugarri jakin bat egiaztatzeko aukera ematen baitigu (barne-berreraiketa). Harago, ebidentzia dialektalak argia eman dezake aldaketen kronologia erlatiboan.

Beste hizkuntzetan badira hainbat baliabide testu zaharretan lexikoaren, morfologiaren edota sintaxiaren inguruko bilaketak egiteko. Horien artean *Penn Parsed Corpora of Historical English* (Kroch & Taylor, 2000; Kroch, Santorini & Delfs, 2001; Kroch, Santorini & Diertani, 2016); *Tycho Brahe Corpus* portugesaren corpus historikoa (Galves, Andrade & Faria, 2017); *Modéliser le changement : les voies du français* frantsesaren corpus historikoa (Martineau, 2005-2010); IcePaHC islandieraren corpus historikoa (Wallenberg, Ingason, Sigurðsson & Rögnvaldsson, 2011), edota POMIC irlandera zahar eta erdiko corpusa (Lash, 2014) aipa daitezke. Denak dira sintaktikoki etiketatuko corpusak.

Euskaraz, aldiz, ez da sintaxi diakronikoa modu sistematikoan aztertzeko ekimenik. Aurkezten dugun proiektuak, Frantziako *Agence nationale de la recherch*e (ANR) finantzatuak, erronka horri aurre egin nahi dio. *Basque in the Making: a historical look at a European language isolate* (BIM) proiektuak bi helburu nagusi ditu: 1) euskararen gramatika-ezaugarri zenbaiten azterketa diakroniko sistematikoa egitea[1]; eta 2) morfosintaktikoki etiketatuta egongo den euskararen corpus historiko zabala eratzea. Corpusak, era berean, metadatu-egitura aberatsa izango du, euskalkiaren, garaiaren edo informazio sozio-historikoaren araberako bilaketak ahalbidetuko dituena. Horrek guztiak corpusaren gaineko bilaketa konplexu eta aberatsak egiteko aukera emango du, besteak beste, hitzak, lemak, kategoria gramatikalak, egitura sintaktiko jakinak (erlatibozkoak, korrelatiboak, mendekoak, etab.) eta horien guztien arteko konbinaketak. Lan honetan batez ere bigarren helburu horri lotuko gatzaizkio. Bestalde, BIM diziplinarteko proiektua dela esan dezakegu, non hizkuntzalaritzaren eta hizkuntzaren prozesamenduaren (aurrerantzean HP) alorreko adituek parte hartzen duten, HiTZ eta IKER UMR 5478 zentroen arteko elkarlanean.

Euskararen inguruko ikerketa historikoaren egoera labur aurkeztu ondoren, 2. atalean corpus historikoa sortzeko jarraituko dugun metodologia azalduko dugu urratsez urrats. Jarraian, 3. atalean metodo konputazionalei esker orain arte lortutako emaitzak aurkeztuko ditugu. 4. atalean garatu nahi dugun bilaketa-interfazearen inguruko zehaztasun zenbait emango dugu, eta bukatzeko, 5. atala ateratako ondorioak eta etorkizunerako aurreikusten ditugun lanak aurkezteko erabiliko dugu.

2. Metodologia

Morfosintaktikoki etiketatutako corpus historikoa sortzeko hiru urrats nagusi aurreikusten ditugu. Lehenik eta behin, corpusa diseinatu, bildu eta prestatu behar da. Ondoren, HPko tresna estandarrek corpusa analizatu ahal izateko, testuak normalizatu behar ditugu, hau da, testu historikoak euskara estandarrera bihurtu behar ditugu. Bukatzeko, behin normalizazio-lan hori eginda, Ixa taldean ditugun analizatzaileak erabiliz testuak morfosintaktikoki analizatuko eta etiketatuko ditugu. Jarraian, urrats horietako bakoitza zehazkiago azalduko dugu.

[1] Ondorengo hauek aurreikusi dira: determinatzaileak, aditz-egitura perifrastikoak, numero morfologikoaren jatorria eta bilakaera, kasu morfologikoaren eta funtzio gramatikalen arteko komunztadura, forma pronominal zehaztugabeen sorrera, postposizio-egituren bilakaera, aditz laguntzaile zehaztuen egoera eta galdegaiaren eta aditzaren arteko hurrentasun-hertsidura.

2.1. Corpusa bildu eta prestatu

Lehenengo zeregina euskara historikoaren corpus esanguratsua biltzea izan da. Esanguratsua diogu, corpusak garai eta euskalki desberdinen adierazgarri izan behar duelako. XV. mendetik XVIII.era bitarteko idatzizko testu garrantzitsuenak bildu ditugu, denbora-tarte horretan euskararen dialekto historiko guztiak jasota baitaude. Testuak hiru irizpideren arabera aukeratu ditugu: i) adierazgarritasuna, hizkuntzaren barietate eta garai baten ordezkari izateko; ii) edizio fidagarriak izatea; eta iii) alderdi soziolinguistikoak.

Euskal klasikoen corpusa (EKC hemendik aurrera [Euskara Institutua, 2013]) izan dugu abiapuntu. Corpus horrek hamabi milioi hitz inguru ditu; beraz, oinarri sendoa da euskalkien bilakaera aztertzeko. EKCk hainbat literatura-generotako testuak jasotzen ditu; halere, gure proiektuaren aztergaia egitura sintaktikoak diren heinean, prosazko testuak hobetsi ditugu corpus historikoa osatzerakoan (garaiaren arabera posible izan denean, behintzat). Bestalde, EKCn testuak garaiaren, hizkeraren eta generoaren arabera sailkatuta daude, eta horrek halako parametroen araberako bilaketak ahalbidetzen ditu. Gainera, EKCk testuen transkripzioak eskaintzen ditu .pdf eta .rtf formatuetan, baita faksimileak ere. Hala ere, testuen transkripzioak kalitate diferenteko edizioetan oinarrituak dira, eta guztiek ez dute fidagarritasun-maila bera. Horregatik, gure proiektuko lehen eginkizuna testuen transkripzioen errebisioa izan da. Transkripzioak euren faksimileekin konparatu ditugu, eta kasuan kasu bi aukera izan ditugu: i) transkripzioa zuzentzea ala ii) transkripzio berria egitea. Lan filologiko horren irizpide nagusia grafia eguneratzea izan da, baina betiere ezaugarri fonologikoak mantenduz. Hau da, EKCko transkripzioetan hitzak euskara batuaren ortografiaren arabera daude, horrek dakartzan ondorio guztiekin; BIM proiektuaren corpusean, aldiz, grafiaren eguneratze hutsa egin dugu, HPko tresnek horiekin lan egin ahal izan dezaten, baina testuetako ezaugarri fonologikoak desitxuratu gabe[2].

Corpusaren tamainaz den bezainbatean, gure helburua ahalik eta testu gehien lantzea da, ikerketaren emaitzak adierazgarriak izan daitezen corpus sendo baten beharra dugulako. Oraingoz, 1.000.000 hitz inguruko erreferentziazko corpusa sortzen ari gara, Euskara Arkaikoa eta Zaharra barne hartzen dituena (XV-XVIII. mende bitarteko testuak). Hizkuntzaren iraganaren iturriei lotutako arazoak eta mugak aintzat hartuta (eta are gehiago euskara bezalako hizkuntza baten kasuan) hitz-kopuru hori onargarritzat jotzen da corpus historiko baterako (*cfr*. Claridge, 2009).

[2] Transkripzio-irizpide guztien berri zehatza ematea artikulu honen helmenetik kanpo gelditzen da. Bihoaz, halere, gutxieneko adigarri batzuk: i) Mitxelenak *hache parásita* deitzen zuena ezabatzea; ii) herskari hasperendunak mantentzea; iii) hots sabaikarituak grafikoki islatuak direnean ematea, bai eta grafiatuak ez direnean ere, delako testuan horren aldeko nolabaiteko segurantza baldin bada; iv) txistukarien neutraltzeari dagokionez, oro har lan filologikoek deskribatu duten egoera islatzea (preseski azpimarratzen dugu proiektu honen eginkizunen artean ez dagoela filologiak argitu ez dituen arazoak konpontzea); v) ozen ondoko afrikatuak grafikoki markatuak ez direnean ez ezartzea, salbu eta afrikazioaren aldeko zantzu argiak ditugunean.

2.2. Corpusaren normalizazioa

HPko tresna gehienak gaur egungo hizkeran idatzitako kazetaritza-testuak prozesatzeko daude diseinatuta. Jakina, estandarizatutako hizkuntza modernoen ezaugarriak eta testu historiko edo dialektalenak ez dira inondik ere berdinak. Hasteko, barietate estandarrek i) isla ona dute hiztegi eta gramatiketan; ii) ortografia estandarra dute, zeinaren arauei lotzen zaizkien argitaratutako testu gehientsuenak; eta iii) HPko tresnak garatzeko erabil daitekeen testu kopuru handia dute formatu elektronikoan eskuragarri. Aldiz, testu historiko eta dialektalekin kontrakoa gertatzen da, eta beraz, HPko tresna estandarrak ezin zaizkie zuzenean aplikatu. Ondorioz, corpusaren normalizazioa ezinbesteko urratsa da testu historikoen eta dialektalen analisi morfosintaktikoa lortzeko bidean.

Hainbat garai eta lekutako testuak biltzean oso corpus zabala osatzen ari gara, hizkuntzalaritzaren eta filologiaren ikuspegitik aniztasun handikoa. Jakina, aniztasun horrek zailtasunak gehitzen dizkio normalizazio-lanari. Horregatik, testuen normalizazio-lan hori bi pausotan egingo dugu. Lehen urrats batean testu bakoitzaren eskuzko normalizazioa egingo dugu, eta bigarren urratsean teknika konputazionalak erabiliz testuaren normalizazio automatikoa egikarituko da. Bestela esanda, sistema automatiko batek eskuz normalizatutako eta etiketatutako laginetik ikasi egingo du, eta hortik ikasitakoarekin testuaren gainerako guztia automatikoki normalizatuko du.

2.2.1. Eskuzko normalizazioa

Testu historikoen normalizazioari ekiteko Ixa taldeak aurretik garatutako estandarizazio-lanetan oinarrituko gara. Etxeberriak (2016) bere tesian Axularren *Gero* eta Mogelen *Peru Abarka* erabili zituen normalizazio-esperimentuak egiteko, eta haren ondorioetako bat izan zen testuaren % 10 eskuz normalizatzea nahikoa dela normalizazio automatikoan artearen egoera berdintzen duten emaitzak lortzeko (Etxeberria, Alegria & Uria, 2019)[3]. Horri jarraikiz, gure corpusean testu bakoitzaren % 10 biltzen duen ausaz sortutako lagina erabiliko dugu eskuzko normalizaziorako. Halere, testuen tamainaren arabera laginaren hautuak salbuespenak ditu; adibidez, Euskara Arkaikoaren testu gehienak eskuz etiketatuko ditugu, oso laburrak izanda ez lukeelako zentzurik % 10eko lagina erabiltzeak. Testua eskuz lantzen hasi baino lehen aurreprozesu bat egiten da hiru urratsetan:

1. Tokenizazioa: testua hitzetan banatzea.
2. Izen berezien ezagutzea (*named entity recognition*).
3. Lexikoaren ezagutzea.

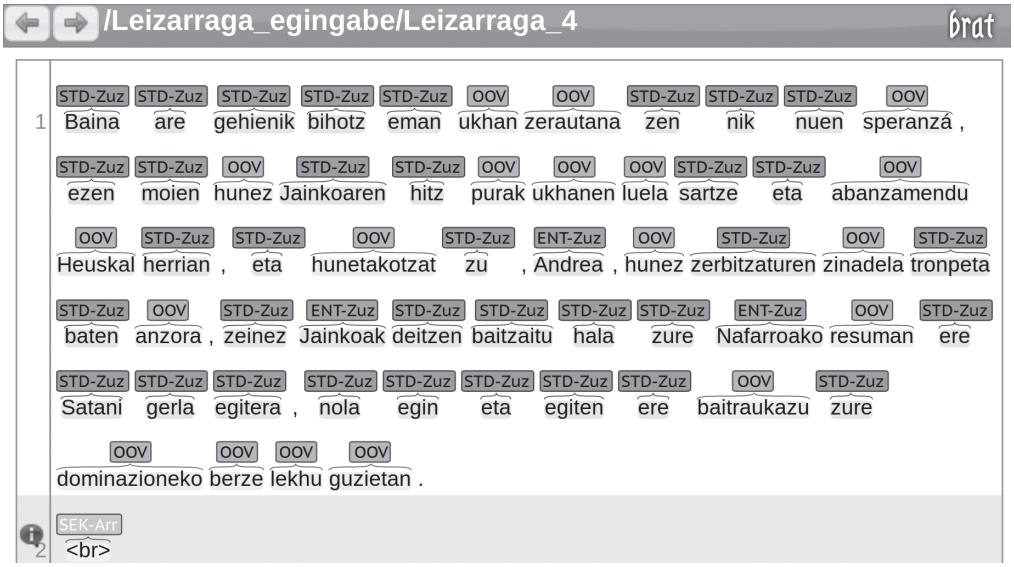
Hiru urrats horiek bukatutakoan, testuko hitz bakoitzari etiketa bat esleituko zaio (ik. 1. irudia):

- ENT-Zuz: *Entitate zuzena*, izen berezien ezagutzaileak (Alegria, Arregi, Ezeiza & Fernández, 2006) izen berezi bat identifikatzean etiketa hau jarriko dio hitzari.

[3] Etxeberria *et al.* (2019) lanean ikusten den bezala, arlo honetan % 75-80tik gorako asmatze-tasa lortzen duten emaitzak onargarritzat hartzen dira.

- STD-Zuz: *Estandar zuzena*, *Euskararen Datu Base Lexikala*n (Aldezabal *et al.*, 2001) oinarritutako euskara estandarraren analizatzaile morfologikoak (Alegria, Artola, Sarasola & Urkia, 1996) hitza edo lema identifikatzen duenean.
- OOV: *Out of vocabulary*, analizatzaile morfologikoak ez duenean hitza edo lema identifikatzen.

Eskuzko etiketatzea BRAT tresnarekin (Stenetorp, Pyysalo & Topić, 2012) gauzatuko da (ik. 1. irudia). Atazaren zailtasuna ikusita, ezinbestekoa da etiketatzailea hizkuntzalaria izatea, eta testu historikoekin lan egiten ohituta egotea. 1. irudiak eskuzko etiketatze-lana egiteko interfazea erakusten du.



1. irudia. Leizarragaren *Testamentu berria*ren zati bat erakusten duen corpusaren eskuzko normalizazioa egiteko interfazea.

Etiketatzailearen betebeharra testuko hitz bakoitzari *Euskaltzaindiaren hiztegia*ren (hemendik aurrera EH [Euskaltzaindia, d. g.]) arabera dagokion forma estandarra (edo aldaera hobetsia) esleitzea da. Irizpide nagusi bi jarraitu ditugu eskuzko normalizazioa egiterakoan:

- Hitzak jatorri bereko aldaerak direnean, EHk hobetsitakoak normalizatuko ditugu: *andra*, *bertze*, *guzi*, *saindu* > *andre*, *beste*, *guzti*, *santu*.
- Hitzek ez dutenean jatorri bera, eta ondorioz ez direnean forma bereko aldaerak, EHk gomendatzen duena normalizatuko dugu, aldaera fonetikoak izango balira bezala: *entrepresa*, *xipi* > *enpresa*, *txiki*.

Horrela eginda, aldaerak lotuta daude, eta etorkizunean corpusaren erabiltzaile batek interfazean lema bati buruzko galdera egiten badu, bilaketa-tresnak lema horren aldaera guztien adibideak eskainiko dizkio. Halaber, testuko hitza ez badago EHn, gure

bigarren iturri nagusia Mitxelenaren eta Sarasolaren (1987-2005) *Orotariko euskal hiztegia* da (OEH). Horren arabera, OEHk ematen duen aldaera nagusia hartuko dugu forma *estandar*tzat (uler bedi, analizatzaile automatikoari begira egiten dugun normalizazioa).

Normalizaziorako hartutako erabakien artean, ondoko hauek azpimarratzen ditugu: i) izenaren morfologian postposizio-kasuen forma estandarrak hobestea: *-rano*, *-rekilako*, *-akgatik* > *-raino*, *-rekiko*, *-engatik*; ii) euskara batuaz kanpoko aditz laguntzaileetako erroaren forma gordetzea, analizatzaileari *egin,* **eradun*, *eutsi* eta **iron* aditz laguntzaileen paradigmak gehitzea ekarri duena (horri esker, erro horien araberako bilaketak egin ahalko dira bilaketa-interfazean).

Eskuzko etiketatze honen azken helburua OOV etiketa guztiak kentzea da. Aldaeren sailkapena ondorengo etiketa hauek erabilita egingo da:

- OOV-Ald: *Aldaera*, etiketatzaileak forma estandarra esleitzen dionean aldaera ez-estandarrari.
- OOV-AldADIJ: *Aldaera – Adizki jokatua*, aurreko etiketaren kasu bera, baina adizki jokatu bati buruz ari garenean.
- OOV-Zuz: *Zuzena*, analizatzaile morfologikoak ez du hitza ezagutzen, baina OEHn sarrera du eta etiketatzaileak markatu nahi du forma zuzena dela.
- STD-Ald: *Aldaera*, testuko hitzaren forma gaur egungo euskara batuko hitz batekin bat datorrenean (*false friend*).
- STD-AldADIJ: *Aldaera – Adizki jokatua*, aurreko etiketaren kasu bera, baina adizki jokatu bati buruz ari garenean.

OOV-AldADIJ eta STD-AldADIJ etiketak gerora gehitutako etiketak dira. Aditz laguntzaileak eta adizki trinkoak normalizatzeko zailak izango direla jakinda, ondorengo normalizazio-lanetarako, bai eta analisi morfosintaktikorako ere, komenigarria iruditzen zaigu horiek identifikatuta izatea.

Aldaeren sailkapen horretaz gain, badugu SEK-Ber (*Sekzio berezia*) izeneko etiketa bat. Etiketatzaileak marka hori erabiliko du lantzen ari den paragrafoan historikoki interesgarria den fenomeno morfosintaktikoren bat aurkitzen duenean. Horrela, ondorengo analisi morfosintaktiko automatikoa egitean, identifikatuta izango ditugu analizatzailearentzat zailagoak izango diren zatiak edo esaldiak, eta horietan arreta berezia jartzeko gai izango gara. Horrela markatuko ditugu, esaterako, kasu prosekutiboa, egungo euskaran existitzen ez diren aditz-erroak (**-idi-* adibidez), aoristoak (*mana zezan* 'manatu zuen'), subjuntibo zaharrak (*garean* 'gaitezen'), forma preskriptiboak (*egin albaiteza*), etab.

Amaitzeko, aipa dezagun corpus osoa eskuz normalizatzeko beharko litzatekeen denboraren estimazioa egin dugula. Horren arabera, orduko 143 hitz etiketatzen baditu, pertsona batek hiru urte inguru beharko luke corpus osoa etiketatzeko. Ditugun baliabideekin hori ezinezkoa izanik, proiektu honetan corpusa normalizatzeko teknika konputazionalak erabiltzea proposatzen dugu. Eta arazoia ez da bakarrik ekonomikoa: etiketatze automatikoaren aldeko hautuak euskararen analizatzaile automatikoaren hobekuntza ekarriko du, tresna hori euskararen forma historiko eta dialektalekin trebatuko delako.

2.2.2. Normalizazio automatikoa

Gure corpusaren normalizazio automatikoa lortzeko Etxeberria *et al.* (2019) lanean aurkezten den normalizazio-metodoa jarraitu dugu. Metodo hori ikasketa estatistikoan oinarritzen da, eta eskuz etiketatutako lagina erabiltzen du ikasketarako. Eskuzko laginetik ditugun hitz-pare guztiak hartuko ditugu gogoan ikasketa-prozesuan, batzuek *Zuz* ('zuzena') etiketa izango dute eta beste batzuek *Ald* ('aldaera'), baina denak erabiliko ditugu ikasketa automatikorako. Metodo horrek egoera finituko transduktoreek (WFST) gidatutako eta grafemetan oinarritutako *Phonetisaurus* tresna fonologikoa erabiltzen du (Novak, Minematsu & Hirose, 2012, 2016). Aldaera/estandar eta estandar/estandar bikoteak pasatzen zaizkio tresnari grafemen sekuentziak ikas ditzan. Beraz, *Phonetisaurus* tresna erabiltzen dugu ikasketa-lagineko hitz-pareetan gertatzen diren aldaketak ikasteko, eta grafematik grafemarako modeloak sortzeko. Behin sistema entrenatuta dagoenean, aurretik ikusi gabeko aldaerentzat forma estandarrak proposatzeko gai izango da.

Metodoa prest dugunean, normalizazio automatikoaren kalitatea ebaluatu beharra dago. Ebaluazioa egiteko *10-fold cross-validation* izeneko ebaluazio-teknika erabili dugu, eta horretarako eskuz etiketatutako lagina hamar zatitan banatu dugu. Hamar zati horietatik bederatzi fitxategi erabili ditugu ikasketarako, eta geratzen den hamargarrena emaitzak ebaluatzeko. Esperimentu bera hamar aldiz errepikatu dugu, aldi bakoitzean ebaluaziorako erabiltzen den fitxategia aldatuz. Obra bakoitzaren normalizazio-kalitatea jakiteko, hamar esperimenturen batezbesteko aritmetikoa kalkulatu da. Orain arte landutako obretan lortutako emaitzak 3. atalean azalduko ditugu zehatz.

2.3. Corpusaren etiketatze morfosintaktikoa

Corpusa normalizatuta izango dugunean, HPko analizatzaileak gai izango dira testuak analizatzeko. Gure kasuan Ixa taldean garatutako *Eustagger* analizatzaile morfosintaktikoa (Ezeiza, Alegria, Arriola, Urizar & Aduriz, 1998) erabiliz etiketatuko dugu corpusa. Analizatzaile horrek hiru urratsetan prozesatzen du testua: i) tokenizazioa; ii) lematizazioa; eta iii) analisi morfosintaktikoa (segmentazioa gehi funtzio sintaktikoak).

Analizatzaileak egiten duen lehenengo gauza testua hitzetan (tokenetan) banatzea da. Ondoren, testuko hitz bakoitzaren lema identifikatzen du eta, bukatzeko, analisi morfosintaktikoa proposatzen du. Alde batetik, hitzak segmentatzen ditu, hau da, hitzen morfemak banatzen ditu bakoitzari bere balioa esleituz eta, bestetik, funtzio sintaktikoak markatzen ditu (subjektua, objektu zuzena, zeharkako objektua, adizlaguna…).

Tresna automatikoak egiten duen analisi morfosintaktikoa eskuz berrikusiko da, SEK-Ber (ik. 2.2.1 atala) etiketa duten paragrafoetan arreta berezia jarriz. Tresnak gaizki analizatzen dituen egiturak eskuz zuzenduko dira, eta gaur egungo euskaran existitzen ez diren fenomeno morfosintaktiko zenbait ondo analizatzeko hainbat erregela berri idatzi behar izatea aurreikusten dugu. Eskuz zuzendutakoarekin eta sortutakoarekin analizatzaileak ikasketa berria egingo du emaitzak hobetze aldera. Prozesu hori bukatuta, testu historikoen analizatzaile morfosintaktikoa prest izango dugu.

3. Emaitzak

Atal honetan testuak normalizatzeko proposatzen dugun metodoak lortutako emaitzak erakutsiko ditugu. Orain artean XVI. mendeko hiru obraren gaineko esperimentuak egin ditugu:

- Leizarragaren *Iesus krist gure iaunaren testamentu berria* (hemendik aurrera TB) (Arroxela, 1571).
- *Refranes y sentencias* (hemendik aurrera RS) (Iruñea, 1596).
- Etxepareren *Linguae vasconum primitiae* (hemendik aurrera LVP) (Bordele, 1545).

2.2.1 atalean azaldu bezala, testu bakoitzaren % 10 aukeratu dugu ausaz eskuz normalizatzeko. Halere, esan beharra dago RS osorik etiketatu dugula eskuz: batetik, obraren ezaugarri bereziengatik (errefrauak izatea, oso hizkera arkaikoa…), eta bestetik tamaina handiegia ez izateagatik. Dena dela, testu osoa eskuz normalizatuta izan arren, interesgarria iruditu zaigu gure normalizazio-metodoak RS bezalako testu batean izango lukeen emaitza ebaluatzea, eta horregatik obra honen kasuan ere *10-fold cross validation* ebaluazio-metodoa aplikatu dugu. 1. taulan ikus daiteke obra bakoitzean lortu dugun asmatze-tasa. Alde batetik, testuko hitz guztiak kontuan hartuta zein asmatze-tasa lortzen dugun neurtu dugu (hau izango da egoera erreala); bestetik, estandarrak ez diren hitzak bakarrik kontuan hartuta zenbatetan asmatzen dugun ere neurtu dugu.

1. taula. TB, LVP eta RSn lortutako asmatze-tasak

	TB	LVP	RS
Aldaerak	% 87,68	% 70,38	% 66,83
Hitz guztiak	% 94,24	% 86,46	% 80,66

Ebaluatu genuen lehenengo obra Leizarragarena izan zen, eta ikusi genuen emaitzak oso onak izan zirela, testuko hitz guztiak kontuan hartuta % 94ra iristen baikinen. Aztertutako bigarren obra Etxeparerena izan zen, eta hor hainbat esperimentu egin genuen. Lehenik eta behin, eskuzko lana aurreztearren Leizarragarengandik ikasitakoarekin Etxepareren obra normalizatzen saiatu ginen baina, espero bezala, Leizarragaren eta Etxepareren hizkerak nahiko desberdinak direnez (batez ere estandarrak ez diren formetan) emaitza kaskarrak lortu genituen: hitz guztiak kontuan hartuta % 75, baina aldaeren kasuan % 46 baino ez. Hori ikusita, pentsatu genuen Leizarragarekin egin bezala, Etxepareren eskuzko lagina bakarrik erabiltzea ikasketa prozesuan, eta hori eginda emaitzak hobetu ziren, baina ez ziren oraindik normalizazio prozesu arrakastatsua bermatzeko modukoak (% 59,58 aldaeren kasuan). Egoera horren aurrean bi irtenbide posible genituen: i) testu gehiago eskuz etiketatzea edo ii) Leizarragarengandik ikasitakoa ikasketa prozesuan gehitzea. Eskuz gehiago etiketatzen hasi baino lehen, bigarren aukerarekin saiatzea erabaki genuen, eta emaitzek gora egin zuten nabarmen: % 86,46 hitz guztiak kontuan hartuta eta % 70,38 aldaerak kontuan hartuta. Azken esperimentu horrek erakusten digu etorkizunean obra bakoitzerako normalizazio-sistema bat izan

beharrean, beharbada irtenbide egokia izan daitekeela eremu dialektal bakoitzeko sistema bat izatea.

Azken obra RS izan zen, eta hemen emaitza baxuagoak lortu ditugu: % 80,66 hitz guztiak kontuan hartuta, baina % 66,83 aldaeren kasuan. Begi-bistakoa da RSko hizkera arkaikoak zailtasun nabarmena gehitzen diola normalizazio-prozesuari. Beraz, emaitza horietatik ateratzen dugun ondorio nagusiena hauxe da: zenbat eta obra euskara batutik aldenduago, orduan eta zailagoa izango da testua normalizatzea. 2. taulan aztertutako obra bakoitzean dagoen hitz estandarren, aldaeren eta entitateen kopuruak eta ehunekoak ikus daitezke.

2. taula. Obra bakoitzean dauden hitz kopuru orokorrak eta STD-Zuz, OOV eta ENT-Zuz etiketa duten hitzen kopuruak eta ehunekoak

	TB	LVP	RS
Hitzak	73.610	6.860	3.083
STD-Zuz	50.321	4.416	1.709
%	% 68,36	% 64,37	% 55,43
OOV	20.924	2.403	1.352
%	% 28,42	% 35,03	% 43,85
ENT-Zuz	2.365	41	22
%	% 3,21	% 0,60	% 0,71
Eskuz etiketatuta	7.548	694	3.083

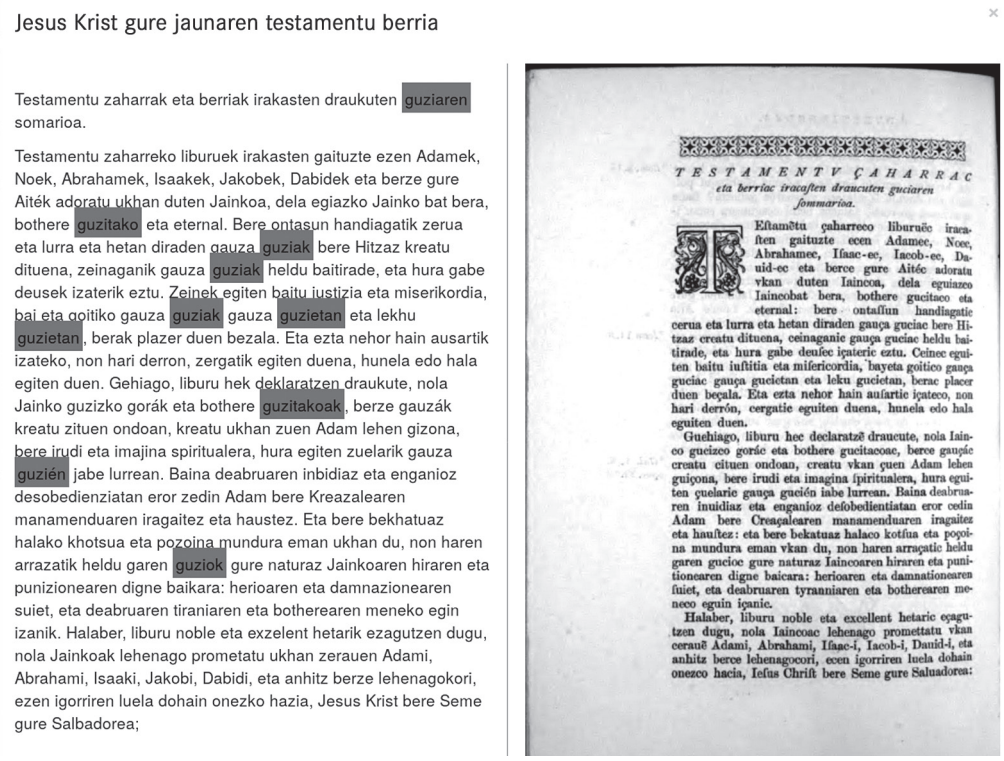
2. taulan argi ikusten den bezala, ebaluazioan lortutako emaitzak erabat datoz bat obra bakoitzean dagoen aldaera-kopuruarekin. RSren kasuan hitzen erdia ia estandarretik aldentzen diren formak dira, eta horrek, nabaria denez, normalizazio-prozesua zaildu egiten du. RSn lortutako emaitzek erakusten digute estandarretik asko aldentzen diren obretan merezi duela eskuzko lanean ahalegin berezia egitea.

4. Bilaketa-interfazea

Proiektu honen azken emaitza corpusa arakatu ahal izateko bilaketa-interfazea izango da. Esan bezala, aurreikusten dugun interfazeak sintaxi diakronikoa aztertzeko erabilgarria izan behar du; beraz, bilaketa konplexuak egiteko gai izan beharko du, bai metadatuen arabera (garaia, euskalkia, egilea, generoa…), bai eta ezaugarri morfosintaktikoen arabera ere (lema, jatorrizko forma, kategoria gramatikala, postposizio-kasua, aditz laguntzailearen erroa, denbora, aspektua, modua…) eta horien guztien arteko konbinazioak. Corpusa morfosintaktikoki etiketatuta egoteak hizkuntzalaritza historikoaren ikuspegitik interesgarriak izan daitezkeen egiturak bilatzeko aukera emango digu, esaterako:

- Erlatibo postnominala aztertzeko: [izena (IZE) + erlatibozkoa] egitura bilatu.
- Aoristoak bilatzeko: [aditzoina + subjuntibozko lehenaldia + perpaus nagusian] egitura bilatu, edo [*EDIN lema + lehenaldia + perpaus nagusian].
- Ezeztapenaren ordena berezia aztertzeko: [aditz nagusia + EZ + aditz laguntzailea] egitura bilatu.
- Ohiturazko perifrasi zaharrak aztertzeko: [aditz izena + *eroan*/*joan*/*eraman* trinkoan] egitura bilatu.

Bukatzeko, ezin aipatu gabe utzi bilaketa-interfazearen beste ezaugarri garrantzitsu bat. Erabiltzaileak corpusean bilaketak egin eta adibideen zerrenda jasotzen duenean, adibide baten gainean klik eginda ikusgai izango du adibide hori agertzen den testuko orrialdea grafia eguneratuan, bai eta faksimilean dagokion orrialdea ere, 2. irudiak erakusten duen bezala.



2. irudia. Leizarragaren *Testamentu berria*ren orrialde bat, ezkerrean grafia eguneratuan, eta eskuinean faksimilean.

5. Ondorioak eta etorkizuneko lanak

Lan honetan morfosintaktikoki etiketatutako euskarazko corpus historikoa sortzeko abian den proiektua aurkeztu dugu. Erakutsitakoaren arabera, begi-bistakoa da horrelako corpus bat osatzeko teknika konputazionalen beharra dagoela, lan hori guztia eskuz egitea ez baita bideragarria. Horregatik diziplinarteko proiektua da, non hizkuntzalaritza teorikoak, hizkuntzalaritza konputazionalak eta informatikak bat egiten duten.

Metodologiari dagokionez, lehenik eta behin corpusa bildu eta prestatu dugu. Ondoren, tresna automatikoak analisia egiteko gai izan daitezen, corpusa normalizatu egin behar izan dugu, zati bat eskuz, eta eskuzko zati horretan oinarrituta eta ikasketa automatikoko teknikak aplikatuz, gainerako testu guztia automatikoki normalizatu dugu. Behin testua normalizatuta, *Eustagger* analizatzaile morfosintaktikoa gai izango da testua morfosintaktikoki etiketatzeko. Bukatzeko, corpusean gordetako informazio morfosintaktiko guztia arakatzeko bilaketa-interfaze bat garatuko dugu.

Testuaren normalizazioa funtsezkoa da proiektuaren arrakasta bermatzeko, eta horren harira, hasierako hiru obretan (Leizarragaren TB, Etxepareren LVP eta RS) lortu ditugun emaitzek proposatutako metodologia balidatzen dutela erakutsi dugu. Hala ere, argi ikusi dugu testua zenbat eta estandarretik urrunago, orduan eta normalizatzen zailagoa dela. Hori dela eta, emaitzak eta bilaketa-tresnaren kalitatea hobetze aldera, zenbait testurekin eskuzko lanean esfortzu berezia egitea erabaki dugu (RS da horren adibide).

Etorkizun hurbilean dugun erronkarik garrantzitsuena analisi morfosintaktikoa testu historikoetara egokitzea da. Bestalde, interfazearen diseinuan eta funtzionalitateen inplementazioan aurrera egiteko hizkuntzalari teorikoen iradokizunak eta proposamenak gogoan hartuta lanean dihardugu. Epe luzera erdietsi nahi dugun helbururik garrantzitsuena, aldiz, corpusa zabaltzea da, XX. mendera arteko testuak bilduz SAHCOBA[4] proiektuaren baitan.

Bukatzeko, ezin aipatu gabe utzi garatzen ari garen bilaketa-tresna ezinbestekoa dela euskararen sintaxi diakronikoa aztertzeko, eta orain arte ez dela horrelako baliabiderik izan ikertzaileen eskura. Beraz, uste dugu eskainiko dugun corpus historikoa ekarpen garrantzitsua izango dela euskal hizkuntzalari-komunitatearentzat. Era berean, proiektu honek hizkuntzaren prozesamendurako tresnak hobetzeko parada ematen du, etorkizuneko beste proiektuen oinarriak finkatuz.

Erreferentziak

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G. & Lersundi, M. (2001). EDBL: A General Lexical Basis for the Automatic Processing of Basque. In S. Bird, M. Liberman & P. Buneman (arg.), *Proceedings of the IRCS workshop on linguistic databases*. Philadelphia: University of Pennsylvania, Institute for Research in Cognitive Science.

[4] Espainiako Zientzia, Berrikuntza eta Unibertsitate Ministerioak (MICINN) finantzatutako *Syntactically Annotated Historical Corpus in Basque* (SAHCOBA) hiru urteko proiektua.

Alegria, I., Arregi, O., Ezeiza, N. & Fernández, I. (2006). Lessons from the development of a named entity recognizer for Basque. *Procesamiento del Lenguaje Natural*, *36*, 25-37.

Alegria, I., Artola, X., Sarasola, K. & Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, *11*(4), 193-203.

Claridge, C. (2009). Historical corpora. In A. Lüdeling & M. Kytö (arg.), *Corpus linguistics. An international handbook* (242-259. or.). Berlin: Mouton de Gruyter.

Etxeberria, I. (2016). *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz* (doktore-tesia). UPV/EHU, Donostia.

Etxeberria, I., Alegria, I. & Uria, L. (2019). Weighted finite-state transducers for normalization of historical texts. *Natural Language Engineering*, *25*(2), 307-321.

Euskaltzaindia. (d. g). *Euskaltzaindiaren hiztegia*. https://www.euskaltzaindia.eus/index.php?option=com_hiztegianbilatu&view=frontpage&Itemid=410&lang=eu helbidetik eskuratua.

Euskara Istitutua. (2013). Euskal klasikoen corpusa (EKC) [datu-basea]. http://www.ehu.eus/ehg/kc/ helbidetik eskuratua.

Ezeiza, N., Alegria, I., Arriola, J. M., Urizar, R. & Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (I, 380-384. or.). Montreal: Association for Computational Linguistics.

Galves, Ch., Andrade, A. L. de & Faria, P. (2017). Tycho Brahe Parsed Corpus of Historical Portuguese [datu-basea]. http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip helbidetik eskuratua.

Gómez, R. & Sainz, K. (1995). On the origin of the finite forms of the Basque verb. In J. I. Hualde, J. A. Lakarra, & R. L. Trask (arg.), *Towards a history of the Basque language* (235-274 or.). Amsterdam-Philadephia: John Benjamins.

Kroch, A., Santorini, B. & Delfs, L. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) [datu-basea]. https://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3/ helbidetik eskuratua.

Kroch, A., Santorini, B. & Diertani, A. (2016). The Penn Parsed Corpus of Modern British English (PPCMBE2) [datu-basea]. https://www.ling.upenn.edu/hist-corpora/PPCMBE2-RELEASE-1 helbidetik eskuratua.

Kroch, A. & Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCMBE2) [datu-basea]. https://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/ helbidetik eskuratua.

Lafon, R. (1944). *Le système du verbe basque au XVI^e siècle*. Bordele: Éditions Delmas.

Lafon, R. (1951). Concordances morphologiques entre le basque et les langues caucasiques (I). *Word*, *7*, 227-244.

Lafon, R. (1952). Concordances morphologiques entre le basque et les langues caucasiques (II). *Word*, *8*, 80-94.

Lakarra, J. A. (1995). Reconstructing the root in Pre-Proto-Basque. In J. I. Hualde, J. A. Lakarra & R. L. Trask (arg.), *Towards a history of the Basque language* (189-206. or.). Amsterdam: John Benjamins.

Lakarra, J. A. (2005). Prolegómenos a la reconstrucción de segundo grado y análisis del cambio tipológico en (proto) vasco. *Palaeohispanica*, *5*, 407-459.

Lakarra, J. A. (2006). Protovasco, munda y otros: reconstrucción interna y tipología holística diacrónica. *Oihenart*, *21*, 229-322.

Lash, E. (2014). The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1 [datu-basea]. https://www.dias.ie/index.php?option=com_content&view=article&id=6586&Itemid=224&lang=en helbidetik eskuratua.

Manterola, J. (2015). *Euskararen morfologia historikorako: artikuluak eta erakusleak* (doktore-tesia). UPV/EHU, Vitoria-Gasteiz.

Martineau, F. (2005-2010). Modéliser le changement : les voies du français (MFVF) [datu-basea]. Université d'Ottawa.

Martínez-Areta, M. (arg.). (2013). *Basque and Proto-Basque*. Frankfurt am Main: Peter Lang.

Mitxelena, K. (1977 [1961]). *Fonética histórica vasca*. Donostia: Diputación Provincial de Guipúzcoa.

Mitxelena, K. & Sarasola, I. (1987-2005). *Orotariko euskal hiztegia*. Bilbo: Euskaltzaindia. https://www.euskaltzaindia.eus/index.php?option=com_oehberria&task=bilaketa&Itemid=413&lang=eu helbidetik eskuratua.

Novak, J. R., Minematsu, N. & Hirose, K. (2012). WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In I. Alegria & M. Hulden (arg.), *Proceedings of the 10th international workshop on finite state methods and natural language processing* (45-49. or.). Donostia: Association for Computational Linguistics.

Novak, J. R., Minematsu, N. & Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, *22*(6), 907-938.

Schuchardt, H. (1908). *The Iberische Deklination*. Viena: Holder.

Schuchardt, H. (1914). Baskisch und Hamitisch. *RIEV*, *8*(1), 76.

Stenetorp, P., Pyysalo, S. & Topić, G. (2012). BRAT rapid annotation tool (1.3.) [softwarea]. https://brat.nlplab.org/ helbidetik eskuratua.

Trask, R. L. (1997). *The history of Basque*. Londres-New York: Routledge.

Uhlenbeck, C. C. (1924). De la possibilité d'une parenté entre le basque et les langues caucasiques. *RIEV*, *15*, 565-588.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F. & Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9 [datu-basea]. http://www.linguist.is/icelandic_treebank helbidetik eskuratua.

Aurkibidea / Índice

11 Aitzinsolasa

13 Prólogo

15 Testu-corpusen informazio morfosintaktikoaren etiketatze automatikoa hizkuntz ezagutzan oinarrituz: zenbait arazo, hainbat erronka
ITZIAR ADURIZ, JOSE MARI ARRIOLA

31 Bertsolaritzaren genealogia subalternoak
LUZIA ALBERRO, K. JOSU BIJUESCA

47 Euskal hiztun berri gazteak muda-prozesuan: ekintza-ikerketa baten behinbehineko emaitzak
ESTIBALIZ AMORRORTU, ANE ORTEGA, JONE GOIRIGOLZARRI

63 Garaziko euskararen azterketa diafasikoa
ALEXANDER ARTZELUS MUXIKA

81 Berridazketak Sarrionandiaren poesian eta Saizarbitoriaren *Egunero hasten delako* nobelan: hurbilpen genetiko bat
MIKEL AYERBE SUDUPE

95 Somorrostro, mendebaldeko euskararen muga
GOIO BAÑALES, MIKEL GORROTXATEGI

113 Euskara sasoian? Euskararen sozializazioa Gasteizko pilota-elkarte batean
IÑIGO BEITIA

127 Euskal hiztun natiboak, ez-natiboak eta hitanoa
GARBIÑE BEREZIARTUA ETXEBERRIA, BEÑAT MUGURUZA ASEGINOLAZA

141 Lingüística Histórica: estado actual
LYLE RICHARD CAMPBELL

163 Lenguas y sociolingüística en el interior del País Vasco en el siglo XIX: testimonios del Archivo Zavala
BRUNO CAMUS BERGARECHE, SARA GÓMEZ SEIBANE

177 Les verbes psychologiques du basque : typologie et diachronie
DENIS CREISSELS, CÉLINE MOUNOLE

195 Testu-generoaren irudikapena eta erabilera ahozkoaren didaktikan
LEIRE DIAZ DE GEREÑU LASAGA, ITZIAR IDIAZABAL GORROTXATEGI, LUIS MARI LARRINGAN ARANZABAL

209 Zentsura-ikasketak gaur egun: aplikazioa euskal literaturan
AMAIA ELIZALDE ESTENAGA

223 Azentuazioaren eremu sintaktikoak mugatzen
ARANTZAZU ELORDIETA

237 Sintaktikoki etiketatutako euskarazko corpus historikoa eraikitzen
AINARA ESTARRONA, IZASKUN ETXEBERRIA, RICARDO ETXEPARE, MANUEL PADILLA-MOYANO, ANDER SORALUZE

253 Ahozkotasunaren didaktika ikuspegi dialektikotik abiatuta
AINHOA EZEIZA, JAVIER ENCINA

267 Ahozko euskararen erabilera eskolan: gako zenbait irakasleen prestakuntzarako
INES M. GARCIA-AZKOAGA, OLATZ BENGOETXEA, JOSUNE ZABALA

283 El corónimo navarro *Salazar* / *Zaraitzu*: origen y desarrollo de su doble denominación
ROBERTO GONZÁLEZ DE VIÑASPRE

295 *Fontes Linguae Vasconum*: orígenes y documentos para una Historia del Euskara
JOAQUÍN GORROCHATEGUI

315 Euskal literatura itzuliaren historiografia bateraturantz
MIREN IBARLUZEA SANTISTEBAN

329 Basque among the world's languages: a typological approach
IVÁN IGARTUA

351 Familias vascohablantes: propuesta de definición desde la socialización lingüística
PAULA KASARES

363 Formation linguistique de basque aux enseignants, pour un enseignement bilingue à parité horaire au Pays Basque Nord
BEÑAT LASCANO

375 Eñaut Etxamendiren obra narratiboaren ekarpena euskal poetika erruralari
ITZIAR MADINA

391 Erdaretarako literatur itzulpena: zeharkako eta zuzeneko itzulpenaren arteko muga lausoa
ELIZABETE MANTEROLA AGIRREZABALAGA

405 Ahozko euskararen irakaskuntzarako irakasleen prestakuntza: berrikuntza didaktikoa eta soziala?
IBON MANTEROLA

421 Enkarterriko PI-*(h)aran*/*(h)uri* motako euskal toponimoak
MIKEL MARTÍNEZ ARETA

437 Diachronical hypotheses accounting for synchronic variation: the case of the Basque particle *ote*
SERGIO MONFORTE

453 Jardueraren azterketa irakasleak prestatzeko bide: debatearen ikas-irakaskuntzaren adibidea
AROA MURCIANO EIZAGUIRRE, ARANTZA OZAETA ELORTZA

467 Hausnarketa zenbait euskal literatura-ikerketez
MARI JOSE OLAZIREGI

485 Ahozko euskara *Kolegioko ikastresna* ikasmaterialean
ARGIA OLÇOMENDY

501 Euskararen postposizioak
JAVIER ORMAZABAL

517 Hitz-ordenaren eragina zenbait ezaugarri gramatikalen erabilera-maiztasunean
LUIS PASTOR

533 Differential D-marking on proper names? A cross-linguistic study
IKER SALABERRI

547 Externalization and morphosyntactic parameters in Basque
HISAO TOKIZAKI

561 XIX. mendeko Debagoieneko testuez zenbait argitasun: egiletasuna eta iturriak
OXEL URIBE-ETXEBARRIA

579 Euskarazko perpausik gabeko azpikonparazioak
LAURA VELA-PLO

595 Latinaren aurreko osagai indoeuroparra Euskal Herriko toponimian: bukaeran *-(iz)amo* duten leku-izenak
LUIS MARI ZALDUA

Izenburua/Título:
Fontes Linguae Vasconum 50 urte. Ekarpen berriak euskararen ikerketari/Nuevas aportaciones al estudio de la lengua vasca

© Argitaratzailea/Edita:
Nafarroako Gobernua/Gobierno de Navarra
Kultura eta Kirol Departamentua/Departamento de Cultura y Deporte
Vianako Printzea Erakundea-Kultura Zuzendaritza Nagusia/Dirección General de Cultura-Institución Príncipe de Viana

Lanak adituek berrikusi dituzte, itsu bikoitzeko sistemaren bidez/Los trabajos han sido revisados por pares doble ciego.

Diseinua eta maketazioa/Diseño y maquetación:
Kö estudio

Imprimaketa/Impresión:
Linegrafic

ISBN: 978-84-235-3561-3
LG/DL: NA 1438-2020

Sustapena eta banaketa/Promoción y distribución:
Nafarroako Gobernuaren Argitalpen Funtsa/Fondo de Publicaciones del Gobierno de Navarra
Navas de Tolosa, 21
31002 Iruña/Pamplona
Tel.: 848 427 121
fondo.publicaciones@navarra.es
https://publicaciones.navarra.es

Fontes Linguae Vasconum 50 urte.

2019an 50 urte egin zituen Nafarroako Gobernuaren Vianako Printzea Erakundeak argitaratzen duen *Fontes Linguae Vasconum: studia et documenta* euskal hizkuntzalaritzako aldizkariak.

Horren gorazarre, liburu honek gaur egungo euskal hizkuntzalaritza- eta literatura-ikerketa zertan den erakutsi nahi du. Eskarmentu handiko ikertzaileek eta belaunaldi berriek bat egin dute argitalpen honetan, besteak beste, dialektologia, hizkuntzaren didaktika, filologia, gramatika teorikoa, hizkuntz tipologia, hizkuntzalaritza historikoa, itzulpengintza, literatura, onomastika eta soziolinguistika hizpide dituztela.

La revista de lingüística vasca *Fontes Linguae Vasconum: studia et documenta*, publicada por la Institución Príncipe de Viana del Gobierno de Navarra, cumplió 50 años en 2019.

En homenaje de la efemérides, este libro pretende dar cuenta del estado actual de la investigación en lingüística y literatura vascas. Investigadores de gran trayectoria y nuevas generaciones se aúnan en esta publicación para tratar, entre otros temas, sobre dialectología, didáctica de la lengua, filología, gramática teórica, tipología lingüística, lingüística histórica, traducción, literatura, onomástica y sociolingüística.

ISBN: 978-84-235-3561-3