I. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Euskarazko corpus orokorrak osatzeko weba ustiatzen

I. Leturia

573-579 or.
https://dx.doi.org/10.26876/ikergazte.i.79

ANTOLATZAILEA: udako euskal unibertsitatea
BABESLEAK: Eusko Jaurlaritza Gobierno Vasco; BFA DFB Bizkaiko Foru Aldundia Diputación Foral de Bizkaia; eman ta zabal zazu UPV EHU
LAGUNTZAILEAK: Deusto Universidad de Deusto Deustuko Unibertsitatea; Euskal Herria Udalbiltza; Mondragon Unibertsitatea; upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Euskarazko corpus orokorrak osatzeko weba ustiatzen

Leturia, I.
Elhuyar I+G
i.leturia@elhuyar.com

Laburpena

Testu-corpusak behar-beharrezkoak dira egoera normalean bizi nahi duen hizkuntza batentzat, eta hala da euskararentzat ere. Baina euskarazko corpus orokorren tamaina oso txikia da beste hizkuntza handiagoenekin konparatzen badugu. Horregatik, logikoa da, beste hizkuntza horiek egin duten bezala, euskarak ere "Web-as-Corpus" planteamendua (hau da, weba eta metodo automatikoak) baliatzea. Artikulu honetan azaltzen dira egileak bere doktore-tesian euskarazko corpus orokor handi bat biltzeko eta weba euskarazko corpus gisa kontsultatzeko egindako ikerketak, garatutako tresnak eta lortutako emaitzak. Lan horretan lehenbiziko aldiz lortu da 100 milioi hitzetik gorako euskarazko corpus bat osatzea, eta online jarri dira gizartearen eskura corpus hori eta weba corpus bat bailitzan kontsultatzeko tresna.

Hitz gakoak: euskara, corpusak, weba, web-as-corpus

Abstract

Text-corpora are essential for any language that aims to live in a normal situation, and so it is for the Basque language too. But the sizes of Basque general corpora are small if we compare them to those in other major languages. Therefore it is logical that the Basque language, just as any of those other languages, makes use of the "Web-as-Corpus" approach (that is, making use of the web and automatic methods). In this paper we describe the research carried out, the tools developed and the results obtained in his PhD thesis by the author to collect a large general corpus of Basque and to build a service to query the web as a Basque corpus. This work has allowed to build a 100 million word corpus for the first time, which has been , and this corpus and a tool to query the web as a Basque corpus have already been put online for their public use.

Keywords: Basque, corpora, web, web-as-corpus

1. Sarrera eta motibazioa

1.1 Testu-corpusen gero eta garrantzi handiagoa

Gaur egun, hizkuntzalaritza eta berarekin zerikusia duten lanak (lexikografia, terminologia, hizkuntza-normalizazioa...) ez dira egiten adituen memoria eta intuizioan oinarrituta soilik. Ordenagailuei esker, testuak kopuru askoz handiagotan gorde daitezke eta azkarrago eta modu fidagarriagoan kontsultatu, adituen burmuinetan baino. Hala ere, ordenagailuok testuen sorkuntza erraztu eta orokortu ere egin dute, eta gaur egun dagoen testu-kopurua ikaragarria da, ezinezkoa bihurtuz ikertzaileek hizkuntza baten idatzizko produkzio osoa eskura izatea. Horregatik, oraindik ere guztiaren lagin bat aztertzearekin konformatu behar dugu, baina eskura dauden lagin hauek handiagoak eta fidagarriagoak dira eta errazago eta azkarrago kontsultatu daitezke ordenagailuen aurreko garaietan baino. Idatzizko hizkuntzaren lagin hauek testu-corpusak dira, eta hizkuntzalaritzarekin erlazionatutako lanak corpus hauek eskainitako lekukotasunetan oinarrituta egitearen diziplinari corpus hizkuntzalaritza.

Corpusen garrantziaren erakusgarri da ingelesezkoen tamaina periodikoki magnitude ordena bat handitzea: milioi bat hitzeko Brown Corpusa (Kučera eta Francis, 1967) izan zen lehena 60ko hamarkadan, geroago 100 milioi hitzeko BNC edo British National Corpus izenekoa (Aston eta Burnard, 1998) eratu zen 90eko hamarkadan, eta gaur egun 70 mila milioi hitzeko ingelesezko corpusak daude (Pomikálek et al., 2012).

1.2 Euskarazko corpusak

Euskarak ere corpusen beharra du, beste edozein hizkuntzak bezala, eta ziurrenik beste hizkuntza handiago batzuek baino gehiago, hainbat arrazoi dela medio: estandarizazioa duela urte gutxi hasi izana eta oraindik ere martxan egotea, irakaskuntzako eremu askotan oso berriki arte sartu ez izana, arlo askotako hiztegi terminologikoen falta, hizkuntza-teknologiak behar bezainbeste garatuak ez egotea...

Eta hala ere, euskarazko corpusak ez dira beharko luketen adina edo beharko luketen bezain handiak, euskarak, edozein hizkuntza txikik bezala, ez baititu nahi beste baliabide (giza-baliabideak zein ekonomikoak) eta corpusak modu klasikoan egitea (hau da, inprimatutako testuetatik erauzita) oso garestia eta mantsoa baita. Euskaraz sei corpus orokor besterik ez daude eskuragai: Orotariko Euskal Hiztegiaren Testu-Corpusa[1], XX. mendeko Euskararen Corpusa[2], Ereduzko Prosa Gaur[3], Klasikoen Gordailua[4], Euskararen Prozesamendurako Erreferentziazko Corpusa (Aduriz et al., 2006) eta Lexikoaren Behatokiko Corpusa[5].

Ikusten denez, euskarazko corpusak gutxi dira, gehienbat txikiak (beste hizkuntza handiagoetakoekin konparatuz behintzat) eta ez eguneratuak, euskarak, edozein hizkuntza txikik bezala, ez baititu nahi beste baliabide (giza-baliabideak zein ekonomikoak) eta corpusak modu klasikoan egitea (hau da, inprimatutako testuetatik erauzita) oso garestia eta mantsoa baita.

1.3 “Web-as-Corpus” planteamendua

Euskara eta ingelesezko corpusen arteko tamaina-ezberdintasun itzela ez da esplikatzen soilik beraiei dedikatutako baliabideen ezberdintasunarekin. Ingelesak eta beste hizkuntza batzuk milaka milioiko corpusak eskuratu badituzte, duela urte gutxi hasitako planteamendu berri bati esker da: “Web-as-Corpus” planteamendua, lekukotasun linguistikoen iturri gisa weba erabiltzean datzana. “Web-as-Corpus” terminoa ziurrenik Adam Kilgarriff-ek sortu zuen *Web as corpus* izenburudun 2001eko bere artikuluan (Kilgarriff, 2001), non hizkuntzalaritza lanetarako weba erabiltzearen aldeko lehenengoetariko apologia egin zuen, diziplina oso bat abiaraziz.

Planteamendu honek abantaila asko eskaintzen ditu: batetik, testu kopuru ikaragarri handia dago webean, eta bertako testuekin corpus oso handiak osa daitezke; bestetik, testu horiek formatu digital publiko eta maneiagarri batean (HTML) egoten dira; gainera, weba beti ari da handitzen eta eguneratzen, ez corpus tradizionalak bezala (zeinak ez diren eguneratzen edo testuak sortu diren datatik denbora luzera sartzen diren), eta egokiagoa da hizkuntza-fenomeno berriak aztertzeko; azkenik, ia edozein hizkuntza, erregistro edo domeinu dago gaur egun webean, beraz modurik merkeena eta egokiena da hainbat baliabide urriko hizkuntzetako corpusak osatzeko.

Web-as-Corpus planteamenduak bere aurkakoak ere baditu. Batzuentzat, weba ez da corpus bat, ez dituelako corpus baten definizioko baldintzak betetzen: ez da bere tamaina ezagutzen, beti aldatzen ari da eta bertatik ateratzen diren emaitzak ezin dira erreproduzitu, ez da ikuspegi linguistikotik diseinatu... Beste batzuentzat, bertako testuen kalitatea ez da ona testu asko, inprimatuak ez bezala, ez baitaude errebisatuak. Eta iritziez gain, ezin ukatuzko desabantaila batzuk ere baditu: zarata asko dago (spama, menu eta abarretako eduki errepikatua, toki ezberdin askotan kopiatutako eduki errepikatua...), copyright baldintzen ezinjakintasuna eta metadatuen falta.ri arazoak ikusten dionik ere bada. Horien arabera, webaren tamaina ez da ezagutzen, ez da ikuspegi linguistikotik diseinatu, bertatik ateratzen diren emaitzak ezin dira erreproduzitu, bertako testuen kalitatea ez da ona, ez da benetako hizkuntzaren adierazgarria... Baina beste egile batzuek eragozpen hauei aurre egiten diete, esanez eragozpen horietako asko weba zuzenean kontsultatzen denean soilik direla egia eta ez corpusak osatzeko iturritzat hartzen bada; kalitateari dagokionez, webekoa hizkuntzaren erabilera erreala dela diote, eta

1 http://www.euskaltzaindia.net/oeh
2 http://xxmendea.euskaltzaindia.net/Corpus/
3 http://www.ehu.es/euskara-orria/euskara/ereduzkoa/
4 http://klasikoak.armiarma.com/corpus.htm
5 http://lexikoarenbehatokia.euskaltzaindia.net

berau aztertzeko webera jo behar dela halabeharrez; eta weba beste edozein corpus bezain adierazgarria dela aldarrikatzen dute.

Edozein modutan, weba gero eta gehiago erabiltzen dela hizkuntzaren ikerketarako edo corpusak egiteko ezin ukatuzko errealitatea da.

1.4 Weba euskarazko corpus gisa

Aipatutako guztiak kontuan izanik, atera daitekeen ondorioa argia da: euskarak ere Web-as-Corpus planteamendua baliatu behar du corpusak egiteko.

Hala ere, hurbilpen honen arrakasta euskararentzat ez da segurua. Batetik, euskarazko weba ez da inondik ere beste hizkuntza handi horietakoa bezain handia. Beste horiek nahikoa dute webean dagoenaren zati bat bakarrik lortzea edozein tamaina, domeinu eta abarreko corpusak lortzeko. Baina ikusteko dago euskarazko webaren zati bat lortzea nahikoa izango ote den corpus handi eta denetarikoak osatzeko.

Bestetik, webeko bilaketa motorrek ez dute euskara kontuan hartzen, ez morfologiari dagokionean (euskarazko deklinazio eta inflexioak ez dituzte bueltatzen) ezta euskarazko emaitzak soilik bueltatzeko aukera ematean (hori 40 hizkuntzentzat besterik ez dute egiten). Arazo hauek Web-as-Corpus hurbilpenean erabiltzen diren teknikak ezin baliatu ahal izatea ekar dezakete.

Edonola ere, gure hipotesia da Web-as-Corpus planteamendua egokia izan daitekeela euskarazko corpusen egoeran hobekuntza esanguratsua lortzeko. Artikulu honetan laburtzen den tesiak hipotesi honen zuzentasuna ebaluatu nahi zuen eta, hori eginez, euskarazko corpusen egoera hobetu.

Lan honetan, Web-as-Corpus planteamendua erabilita euskarazko corpus orokor oso handi bat biltzen eta weba euskarazko corpus bat bailitzan zuzenean kontsultatzen saiatu gara.

2. Arloko egoera

2.1 Euskarazko corpus orokor handi bat osatzea weba testuen iturburutzat hartuta

Webetik corpus orokor handiak biltzeko metodoei dagokienez, bi metodo erabili ohi dira: crawling metodoa eta bilatzaileen metodoa.

Crawling metodoan, hasierako URL zerrenda batetik abiatuta (“hazi” URLak deritzenak), horiei dagozkien orriak jaisten dira, eta orri horietan aurkitutako estekak URL zerrendari gehitzen zaizkie, beraiekin gauza bera egiteko; hau errekurtsiboki aplikatzen da zerrenda amaitu arte edo behar den tamaina lortu arte. Metodo honen errendimendua optimoa izan dadin, garrantzitsua da hainbat parametro ongi doitzea: “hazi” URLen zerrenda lortzeko metodoa (batzuek hitz orokorren konbinazioak eta bilatzaileak darabiltzate hasierako URLak lortzeko, beste batzuek Open Directory Project bezalako webeko direktorioak), “hazi” URLen zerrendaren luzera (luzera ezberdinak erabiltzen dira), crawling estrategia (detektatzen den webgune bakoitzetik ahalik eta gehien jaistea beste webgune batzuetara jo aurretik, edo, gehien erabiltzen dena, ahalik eta webgune ezberdin gehienetatik jaistea), goi-mailako domeinu nazionalen batera mugatu nahi dugun, orrien jaistea paralelizatzea azkarragoa izan dadin, formatu ez egokiak edo URL berdinak jaitsi aurretik detektatzea... Corpus orokor handiak biltzeko proiektuetan, crawling metodoa da erabiliena (Baroni et al., 2009).

Bilatzaileen metodoa, orokorrean corpus espezializatuak lortzeko erabiltzen bada ere, corpus orokor handiak biltzeko ere erabili izan da (Sharoff, 2006). “Hazi” hitzen zerrenda bat erabiltzen da, horien konbinazioak bilatzaileetara bidaltzen dira eta bueltatutako orriak jaitsi eta corpuseratzen dira, helburu-tamaina iritsi arte edo konbinazioak agortu arte. Kasu honetan ere garrantzitsua da prozesua ongi doitzea: “hazi” hitzak nondik lortu (normalean maiztasun altuko hitz orokorrak dira, baina funtzio-hitzak izan gabe eta beste hizkuntzetan esanahirik ez dutenak), bilatzaileek behar den hizkuntzako testuak soilik itzultzea nola lortu (bilatzaileen

aukera baliatu ohi da, baina hizkuntza batentzat ez dagoenean hizkuntza-iragazteko hitzak gehitzen zaizkio hitzen konbinazioari), "hazi" hitzen zerrendaren luzera (luzera oso ezberdinak erabiltzen dira lan ezberdinetan), konbinazioen luzera (2 eta 4 artean erabili ohi da), zenbat bilaketa egin (5.000 eta 30.000 artean egin izan dira), emaitzetatik zenbat jaitsi (lehen 10ak normalean)...

2.2 Weba euskarazko corpus bat bailitzan kontsultatzea

Weba zuzenean corpus gisa kontsultatzeko dagoen arazo nagusia da derrigorrez bilaketa-motorrak erabili behar direla, eta hauek ez daude diseinatuta helburu linguistikoekin egindako kontsultei erantzuteko, informazio-bilaketarako baizik. Hauek dira zehazki bilaketa-motorrek kontsulta linguistikoetarako ematen dituzten arazoak: ez dituztenez orriak linguistikoki etiketatzen, ez dituzte lema baten inflexio eta deklinazioen agerpen denak itzultzen; bilaketa-sintaxia mugatua dute (ezin dira hitzen arteko distantzia, kategoria edo zernahitarako karaktereak erabili); bueltatzen dituzten kopuruak oso arbitrarioak eta aldakorrak dira, ikerketak erreproduzitzea ezinezkoa eginez; emaitzen ordena ez da irizpide linguistikoen araberakoa; eta ezin izaten dira emaitza guztiak ikusi, batzuk soilik.

Hala ere, kasu batzuetan zilegi (eta are beharrezko) da bilatzaileak erabiltzailea kontsulta linguistikoetarako, adibidez inongo corpusetan aurkitzen ez diren hitzen ebidentziak bilatzea, neologismo oso berrien erabilera ikustea, zenbait hitzen maiztasun erlatiboak konparatzea... Corpusik ez duten, edo corpus gutxi eta txikiak dituzten, hizkuntzentzat ere (euskara bezalako hizkuntzentzat, alegia) aukera bakarra izan daiteke.

Bilatzaileak zuzenean erabiltzeak are arazo gehiago ditu. Batez ere, orriak itzultzen dituztela eta ez bilatutako hitzaren agerpenak. Horregatik, haien gainean lan egiten duten tresnak eta web zerbitzuak eraiki izan dira, bilatzaileek itzulitako orriak deskargatu eta bilatutako hitzaren agerpenak corpus tresnen gisara erakusten dituztenak.

Edonola ere, zerbitzu eta tresna hauek ez dabiltza ongi euskararen kasuan, morfologiagatik batetik eta bilatzaileek euskarari ematen dioten tratamenduagatik (edo, hobeto, tratamendu ezagatik) bestetik. Horregatik, tesiaren helburuetako bat izan da tresna bat eraikitzea, ahalbidetuko duena weba kontsultatzea euskarazko corpus bat bailitzan.

3. Ikerketaren muina

3.1 Euskarazko corpus orokor handi bat osatzea weba testuen iturburutzat hartuta

Euskarazko corpus orokor handi bat lortzeko helburuarekin, bi metodoak probatu eta ebaluatu dira, crawling-arena eta bilatzaileena, ikusteko zein den onena euskararentzat, abiadura, kostua, tamaina edo kalitateari dagokionez (Leturia, 2012).

Bilatzaileen metodoari dagokionez, parametro ezberdinekin probatu da. "Hazi" hitzen luzerarentzat, 500, 1.000, 2.000, 5.000 eta 10.000 probatu dira eta konbinazioen luzerari dagokionez, 1, 2, 3, 4 eta 5. "Hazi" hitzentzat XX. mendeko Euskararen Corpuseko hitz maizenak erabili dira, funtzio-hitzak kenduta. Eta bilaketek euskararentzat emaitza optimoa eman dezaten, gorago deskribatutako morfologia bidezko galderaren hedapena eta hizkuntza iragazteko hitzen teknikak erabiltzen dira berriz ere. Eta kasu bakoitzean 12.000 galdera egin zaizkie bilatzaileei.

Ikusi da hainbat parametrorekin ez dela lortzen corpus nahiko handirik. 2.000 edo 5.000 "hazi" hitzeko zerrenda eta 2 edo 3 hitzeko konbinazioak erabilita lortzen dira emaitzarik onenak. Hala ere, beti 125 milioi hitz inguruko corpusak lortu dira, eta hazte-abiadura oso mantsoa da amaieran, beraz ez da ikusten bide honetatik corpus askoz handiagoak lor daitezkeenik.

Crawling metodoan, Open Directory Project-eko "Euskara" ataleko helbideak erabili dira "hazi" URLtzat eta orriak paraleloan jaitsi dira webgune aniztasuna hobesteko estrategiarekin.

Bide hori jarraituz 200 milioi hitzetik gorako corpusa eskuratu da, eta hazte-erritmoa ez da asko jaitsi, beraz gehiago hazteko potentziala dauka.

Corpusen kalitatea ebaluatzeko, bilatzaileen bidez lortu den corpus handiena eta crawling bidez lortutako corpusa XX. mendeko Euskararen Corpusa eta Lexikoaren Behatokiko Corpusarekin konparatu dira, lau ezaugarriri begiratuta: zeintzuk diren corpus bakoitzean besteekiko gehien nabarmentzen diren hitzak (LLR elkartze-neurriaren bidez kalkulatuta), corpus bakoitzeko hitz erabilgarrien kopurua (20 baino maiztasun handiagokoena), corpus baten estaldura besteekiko eta corpus baten ekarpena besteekiko.

Corpus bakoitzean besteekiko gehien nabarmentzen diren hitzei dagokienez, emaitzetan ez dago espero ez zitekeen gauzarik edo corpusak ezegokiak direla adierazten duen gauzarik. Bilatzaileen corpusean eremu administratiboko hitzak nabarmentzen dira eta crawling-ekoan web orrietako tipikoak diren hitzak, baina XX. mendekoan hitz zaharrak eta Lexikoaren Behatokian komunikabideetakoak nabarmendu diren gisan. Hitz erabilgarrien kopuruari dagokionez, web corpusetakoak askoz gehiago dira corpus klasikoetakoak baino. Estaldurari dagokionez, web corpusek klasikoen hitzen %95 inguru dituzte, baina alderantzizko norabidean kopuru hori %35 inguru edo txikiagoa da. Eta azkenik, ekarpenari dagokionez, web corpusek klasikoei egiten dieten hitz berrien ekarpena ia %85 ingurukoa da, alderantzizko norabidean %1era iristen ez den bitartean.

Ondorioz, esan dezakegu bai crawling eta bai bilatzaileen bidez lor daitezkeela euskarazko corpus handiak, baina handiagoak crawling bidez (bilatzaileen bidez lor daitekeen tamaina mugatuta dago). Kalitate aldetik corpus egokiak dira, corpus klasikoen hitzak ia osorik barne hartzen dituztenak eta besteek ez dituzten hitzen ekarpen handia egiten dutenak. Beraz, weba iturburu egokia da euskarazko corpus orokorren egoera nabarmen hobetzeko, eta hobekuntza hau gauzatu egin da, 100 milioi hitzetik gorako corpus handi horietako bat Web-Corpusen Atarian jarri baita jendearen eskuragarri[6].

3.2 Weba euskarazko corpus bat bailitzan kontsultatzea

Weba euskarazko corpus gisa kontsultatu ahal izateko tresna bat eraiki ahal izateko, bi teknika garatu dira: morfologia bidezko galderaren hedapena eta hizkuntza filtratzeko hitzak. Lehenengoa honetan datza: bilatu nahi den hitzaren deklinazio edo inflexioak sortzen dira lengoaia naturalaren prozesamenduko teknikak erabiliz, eta horiek guztiak bidaliko zaizkio bilatzaileari OR (EDO) operatzaile batean. Eta bigarrena, euskarazko hitzik ohikoen eta bereizgarrienak gehitzea galderari, horrela euskarazko emaitzak soilik lortzeko.

Teknika horiek doitu ahal izateko, hainbat esperimentu eta neurketa egin dira. Horien bidez, teknika horientzat parametro egokienak aurkitu dira (zein kasu diren produktiboenak morfologia bidezko galderaren hedapenerako, zenbat eta zeintzuk diren filtro-hitz egokienak...) eta horien bidez lortzen den errendimendu hobekuntza neurtu da (zenbatean handitzen diren estaldura, morfologia bidezko galderaren hedapenari esker. eta hizkuntz-zehaztasuna, hizkuntza filtratzeko hitzen bidez). Eta frogatuta gelditu da hobekuntza nabaria dela eta teknikak baliagarriak direla weba corpus gisa kontsultatzeko euskararentzat ongi funtzionatuko duen tresna bat eraikitzeko (Leturia et al., 2008, 2013).

Horrez gain, weba euskarazko corpus gisa kontsultatu ahal izateko tresna eraiki egin da eta euskal gizartearen eskura online jarri: CorpEus[7] (Leturia, Gurrutxaga, Alegria, et al., 2007).

Gainera, esperimentuen bidez lortutako datuak (hedapenerako kasuak, hizkuntzaren filtro-hitzak, zehaztasuna, estaldura...) euskararentzat etorkizunean egingo diren informazioa bilatzeko tresnetarako oso baliagarriak izango dira. Eta erabili ere erabili dira Elebila[8] euskarazko online bilatzailean (Leturia, Gurrutxaga, Areta, et al., 2007).

6 http://webcorpusak.elhuyar.org/cgi-bin/kontsulta.py?mota=arrunta. Guk hau online jarri eta gutxira, Egungo Testuen Corpusa aurkeztu zen, 200 milioi hitzekoa.

7 http://www.corpeus.org

8 http://www.elebila.eu

4. Ondorioak

Artikuluaren sarreran, corpusei dagokienez euskararen egoera txarra azpimarratzen genuen. Gure hipotesia zen Web-as-Corpus planteamendua baliozkoa zela euskarazko corpusen egoeran hobekuntza esanguratsua lortzeko, eta lan hau hipotesi horren zuzentasuna frogatzera eta, aldi berean, euskarazko corpusen egoera hobetzera bideratu da.

Helburu horren bila, lehenbizi bi tresna garatu genituen, bata bilatzaileetan oinarritua eta bestea crawling metodoan, euskarazko corpus orokor handiak lortzeko. Horien bidez, ordura arteko corpusen tamaina 8 aldiz gainditzen zuten corpusak osatu dira, 200 milioi hitzera ailegatuz, eta etorkizunean are handiagoak lortzea espero dugu crawling metodoaren bidez. Corpus horietako bat, 125 milioi hitzekoa (ordura arte bildu genuen handiena), online jarri da kontsulta publikorako Web-Corpusen Atarian.

Ondoren, web zerbitzu bat osatzea lortu genuen (CorpEus) weba euskarazko corpus gisa kontsultatzea ahalbidetzen duena, horrelako beste zerbitzu batzuek euskararekin dituzten arazoak gainditzen zituena. Horretarako, morfologia bidezko galderaren hedapena eta hizkuntza iragazteko hitzen teknikak asmatu, inplementatu eta optimizatu genituen, tresna honetan erabili dena baina baita tesian bilatzaileen bidez corpusak biltzeko garatu diren beste tresna denetan eta euskarazko bilatzaile batean (Elebila) ere.

Areago, goian aipatutako corpusak biltzeko tresnen garapenak eskatu digu corpus garbiketako hainbat tresna garatzea (testuaren ingurukoak garbitzeko (Saralegi eta Leturia, 2007), errepikatuen eta barne-hartzeen detekziorako...), euskarara eta gure beharretara egokituak eta corpus bilketa prozesuaren funtzionamendu optimoan laguntzen dutenak, baina beste lanetan ere erabil daitezkeenak.

Honegatik guztiagatik, ondoriozta dezakegu gure hasierako hipotesia egiaztatu dela, hau da, Web-as-Corpus planteamenduak euskararen corpusen egoeran aldaketa ekar zezakeela, eta aldaketa hori etorri etorri dela hemen deskribatzen den lanarekin. Beste hizkuntza handiagoen egoerarekin ezin dezakegu konparatu euskararena, baina metodologia eta tresna batzuk garatu ditugu dagoena biltzeko eta asko bildu da, euskarazko corpusen kantitatea eta tamaina modu esanguratsuan handituz.

Gainera, tesian eraikitako web-as-corpus eta corpus-bilketa tresnek behar dituzten baliabide eta tresna linguistikoak nahikoa oinarrizkoak direnez (N-grametan oinarritutako hizkuntza detekzioa eta analisi eta sorkuntza morfologikoa, besterik ez), uste dugu euskarazko tresnak eraiki eta corpusak biltzeko aplikatu dugun metodologia euskararenaren antzeko egoeran dauden beste hizkuntza batzuekin ere –hizkuntza gutxitu eta morfologikoki konplexuekin ere– aplika daitekeela beraien egoera ere hobetzeko.

5. Etorkizunerako planteatzen den norabidea

Lan honen barruan, etorkizunean bide hauek landu nahiko genituzke:

- Corpus orokor handiagoak lortu eta publikoaren eskura jarri
- Generoetako corpus espezializatuak lortzeko tresna bat garatzea
- Landutako metodologia eta tresnak beste hizkuntza gutxiagotu batzuentzat egokitu eta probatu

6. Erreferentziak

Aduriz, I., Aranzabe, M., Arriola, J. M., Atutxa, A., Diaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., & Urizar, R. (2006). Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World*, 56, 1–15.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh, U.K.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43, 209–226.

Kilgarriff, A. (2001). Web as corpus. In *Proceedings of Corpus Linguistics 2001*, Lancaster, UK, 342–344.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*, Brown University Press, Providence, USA.

Leturia, I. (2012). Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India.

Leturia, I. (2014). *The Web as a Corpus of Basque*, Euskal Heriko Unibertsitatea, Donostia.

Leturia, I., Gurrutxaga, A., Alegria, I., & Ezeiza, A. (2007). CorpEus, a «web as corpus» tool designed for the agglutinative nature of Basque. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, Louvain-la-Neuve, Belgium, 69–81.

Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I., & Ezeiza, A. (2007). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*, Amsterdam, The Netherlands, 47–54.

Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I., & Ezeiza, A. (2013). Morphological query expansion and language-filtering words for improving Basque web retrieval. *Language Resources and Evaluation*, 47(2), 425–448.

Leturia, I., Gurrutxaga, A., Areta, N., & Pociello, E. (2008). Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC)*, Marrakech, Morocco.

Pomikálek, J., Jakubícek, M., & Rychlỳ, P. (2012). Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 502–506.

Saralegi, X., & Leturia, I. (2007). Kimatu, a tool for cleaning non-content text parts from html docs. In *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, Louvain-la-Neuve, Belgium, 163–167.

Sharoff, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Queries. In M. Baroni & S. Bernardini (arg.), *WaCky! Working Papers on the Web as Corpus*, Gedit Edizioni, Bologna, Italy (or. 63–98).

7. Eskerrak eta oharrak

Egileak eskertu nahi du batez ere Elhuyar Fundazioa, bertako I+G sailean garatu baita hemen aipatzen den lan guztia.

EHUko IXA Taldeak ere merezi ditu esker onak, lan honetan parte handia izan baitu aholkulari gisa.

Lan honek hainbat publikazio eman ditu, erreferentzietan aipatzen direnak. Horietan denetan parte hartutako jendea ere eskertu beharrean gaude, denek hartu baitute parte.

Eusko Jaurlaritzako Industria eta Kultura sailei ere eskerrak eman nahi dizkiegu proiektuaren hainbat zatitan eta hainbat urtetan Etortek, Saiotek eta IKT programen barruan emandako diru-laguntzengatik.

Artikulu honetan deskribatzen den lana egilearen doktore-tesia izan da, *The Web as a Corpus of Basque* izenburuduna (Leturia, 2014).