Weba euskarazko corpus gisa

*Igor Leturia**[1], *Xabier Arregi*[2], *Kepa Sarasola*[2]

[1] Hizkuntza eta Teknologia Saila,
Elhuyar Fundazioa

[2] IXA Taldea, Informatika Fakultatea
(UPV/EHU)

* i.leturia@elhuyar.com

Jasoa: 2014-05-30 Onartua: 2014-09-12

Laburpena: Euskarak, beste edozein hizkuntzak bezala, testu-corpusak behar ditu mundu modernoan bizirauteko eta normalki erabiltzeko. Alabaina, euskarazko corpusak gutxi eta txikiak dira, beste hizkuntza handiagoenekin konparatuz gero. Hori horrela da beste hizkuntzek «Web-as-Corpus» izeneko planteamendua baliatu dutelako, hau da, weba erabili dutelako corpus gisa edo corpusak osatzeko testu-iturritzat. Artikulu honetan azaltzen dira bere doktorego-tesian lehenengo autoreak, beste bi autoreen zuzendaritzapean, euskarazko corpusgintzarako weba eta metodo automatikoak baliatzeko egindako ikerketak, garatutako tresnak eta lortutako emaitzak. Horietatik ondorioztatu daiteke «Web-as-Corpus» planteamendua baliagarria dela euskarazko corpusen egoera hobetzeko, garatu diren tresna informatikoen bidez weba corpus gisa kontsultatzeko tresna bat eraiki baita eta mota askotako eta kalitatezko corpusak lortu ahal izan baitira (corpus orokor oso handiak, corpus espezializatuak, corpus konparagarriak,..). Horietako asko jada online gizartearen eskura jarri dira.

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

Abstract: The Basque language, just as any other, needs text corpora to survive in the modern world and to be used normally. But Basque corpora are few and small compared to those in other major languages. This is so because other languages have made use of the «Web-as-Corpus» approach, which consists of using the web as a corpus or as a source of texts for corpora. In this paper, we describe the research carried out in his PhD thesis by the first author, under the supervision of the other two authors, to use the web and automatic methods for Basque corpus building, and also the tools developed and the results obtained. Out of them we can conclude that the «Web-as-Corpus» approach is valid to improve the state of Basque corpora, since with the developed tools we have collected quality corpora of different types (very large general corpora, specialized corpora, comparable corpora...) and built a service to query the web as a Basque corpus. Many of these tools and services have already been placed online for their public use.

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

1. **SARRERA**

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 handiagoan gorde daitezke eta azkarrago eta modu fidagarriagoan kontsultatu, adituen burmuinetan baino. Hizkuntzalaritzarekin erlazionatutako azterketa eta lanetarako erabiltzen diren idatzizko hizkuntzaren laginei deritze testu-corpusak, eta hizkuntzalaritza-lanok corpus hauen lekukotasunean oinarrituta egiteko diziplinari corpus hizkuntzalaritza.

Mota askotako corpusak daude, eman nahi zaien erabileraren arabera: orokorrak edo espezializatuak; elebakarrak edo eleaniztunak; azken kasu honetan konparagarriak edo paraleloak... Egunetik egunera mota guztietako corpus gehiago eta handiagoak jartzen dira eskuragarri. Corpusak kontsultatzeko, bilaketa-aukera linguistikoak eskaintzen dituzten kontsulta-tresnak erabiltzen dira, galdetutako hitzen erabilera-testuinguruak eta kopuruak erakusten dituztenak.

Corpusen garrantziaren erakusgarri da ingelesezkoen tamaina periodikoki magnitude-ordena bat handitzea: milioi bat hitzeko Brown Corpusa [1] izan zen lehena 60ko hamarkadan, geroago 100 milioi hitzeko BNC edo British National Corpus izenekoa [2] eratu zen 90eko hamarkadan, eta gaur egun 70 mila milioi hitzeko ingelesezko corpusak daude [3].

Beraz, argi dago komunikabideetan, hezkuntzan eta eguneroko bizitzan normaltasunez erabili nahi den edozein hizkuntzak corpusen beharra duela gaur egun, baliabide oso garrantzitsuak baitira hizkuntza baten garapenaren arlo askotarako.

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. Baina, hala ere, euskaraz sei corpus orokor besterik ez daude eskuragai[1], eta handienak 26,5 milioi hitz besterik ez du. Corpus espezializatuei dagokienez, bakarra dago (Zientzia eta Teknologiaren Corpusa [10], 8,5 milioi hitzekoa). Eta bost corpus paralelo ere badaude[2], hainbat elkarte

[1] Orotariko Euskal Hiztegiaren Testu-Corpusa [4], XX. mendeko Euskararen Corpusa [5], Ereduzko Prosa Gaur [6], Klasikoen Gordailua [7], Euskararen Prozesamendurako Erreferentziazko Corpusa [8] eta Lexikoaren Behatokiko Corpusa [9].

[2] EIZIEren[11], Gipuzkoako [12] eta Bizkaiko[13] Foru Aldundien eta IZOren [14] itzulpen memoriak eta Consumer Corpusa [15].

edo erakunde publikok eskuragarri jarritako itzulpen memoriak, hain zuzen ere.

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

1.3. **«Web-as-Corpus» planteamendua**

Ingelesak eta beste hizkuntza batzuek 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 [16], non hizkuntzalaritza lanetarako weba erabiltzearen aldeko lehenengoetariko apologia eginik, diziplina oso bat abiarazi zuen.

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 [17]; azkenik, ia edozein hizkuntza, erregistro edo domeinu dago gaur egun webean.

Web-as-Corpus planteamenduari arazoa ikusten dionik ere bada. Horien arabera, webaren tamaina ez da ezagutzen, ez da ikuspegi linguistikotik diseinatu, bertatik ateratzen diren emaitzak ezin dira erreproduzitu [18], bertako testuen kalitatea ez da ona [19], ez da benetako hizkuntzaren adierazgarria [20]... 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 [21]; kalitateari dagokionez, webekoa hizkuntzaren erabilera erreala dela diote, eta berau aztertzeko webera jo behar dela halabeharrez [22]; eta weba beste edozein corpus bezain adierazgarria dela aldarrikatzen dute [23].

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

1.4. **Weba euskarazko corpus gisa**

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

Hala ere, planteamendu honen arrakasta ez da segurua euskararen kasuan. Batetik, euskarazko weba ez da inondik ere beste hizkuntza handi horietakoa bezain handia, eta 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. Arazo hauek Web-as-Corpus planteamenduan 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.

Web-as-Corpus planteamendua lau modalitatetan probatu dugu: weba euskarazko corpus bat bailitzan zuzenean kontsultatzea eta hiru corpus-mota biltzea: euskarazko corpus orokor oso handi bat, euskarazko domeinu-corpus espezializatuak eta euskara eta beste hizkuntza bateko domeinu-corpus konparagarriak.

2. WEBA EUSKARAZKO CORPUS BAT BAILITZAN KONTSULTATZEA

Weba zuzenean corpus gisa kontsultatzeko dagoen arazo nagusia da derrigorrez bilaketa-motorrak erabili behar direla, baina hauek ez daude diseinatuta helburu linguistikoekin egindako kontsultei erantzuteko, eta horrek hainbat desabantaila dakartza [24-26]: ez dituzte lema baten inflexio eta deklinazioen agerpen denak itzultzen, bilaketa-sintaxia mugatua dute, itzultzen dituzten kopuruak oso arbitrarioak eta aldakorrak dira, emaitzen ordena ez da irizpide linguistikoen araberakoa, ezin izaten dira emaitza guztiak ikusi...

Hala ere, kasu batzuetan zilegi (eta are beharrezko) da bilatzaileak erabiltzailea kontsulta linguistikoetarako [24], 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 aukera bakarra izan daiteke hori.

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 [27-32].

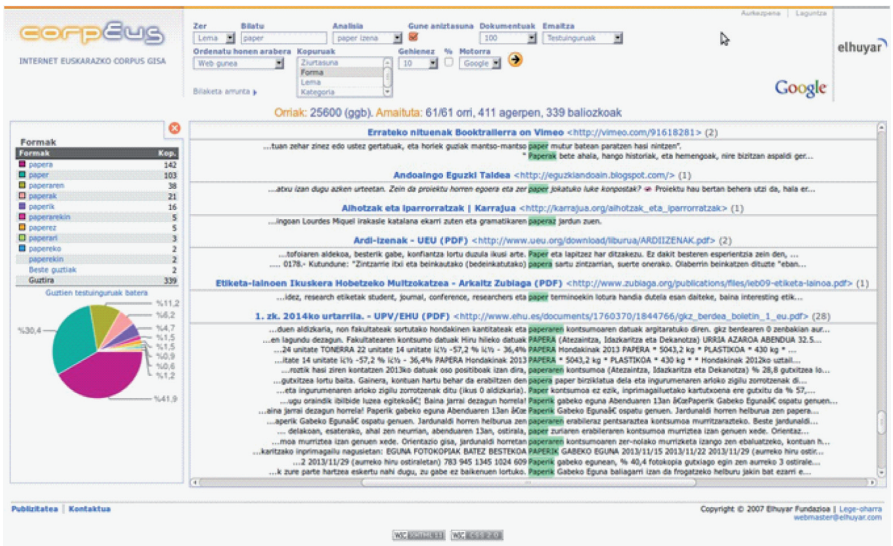
Edonola ere, zerbitzu eta tresna hauek ez dabiltza ongi euskararen kasuan, morfologiagatik batetik eta bilatzaileek euskarari ematen dioten trata-

menduagatik (edo, hobeto, tratamendu ezagatik) bestetik. Horregatik, tesiaren helburuetako bat izan da tresna bat eraikitzea, ahalbidetuko duena weba kontsultatzea euskarazko corpus bat bailitzan.

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 horientzako 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). Frogatuta gelditu da hobekuntza nabaria dela eta teknikak baliagarriak direla [33, 34] weba corpus gisa kontsultatzeko tresna bat eraikitzeko, euskararentzat ongi funtzionatuko duena.

Gainera, weba euskarazko corpus gisa kontsultatu ahal izateko tresna eraiki da eta online jarri da, euskal gizartearen eskura: CorpEus (http://www.corpeus.org) [35].



1. irudia. CorpEus-en pantaila-irudia.

Horrez gain, esperimentuen bidez lortutako datuak (hedapenerako kasuak, hizkuntzaren filtro-hitzak, zehaztasuna, estaldura...) oso baliagarriak izango dira euskararentzat etorkizunean egingo diren informazioa bilatzeko tresnetarako. Eta erabili ere erabili dira Elebila euskarazko online bilatzailean (http://www.elebila.eu) [36].

3. 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), URL horiei dagozkien orriak jaisten dira, eta orri horietan aurkitutako estekak URL zerrendari gehitzen zaizkie, berriro prozedura bera egiteko; hau errekurtsiboki aplikatzen da zerrenda amaitu arte edo behar den tamaina lortu arte. Corpus orokor handiak biltzeko proiektuetan, crawling metodoa da erabiliena [37-41].

Bilatzaileen metodoa, gehien bat corpus espezializatuak lortzeko erabiltzen bada ere, corpus orokor handiak biltzeko ere erabili izan da [42-46]. «Hazi» hitzen zerrenda bat erabiltzen da, horien konbinazioak bilatzaileetara bidaltzen dira eta itzulitako orriak jaitsi eta corpuseratzen dira, helburu -tamaina iritsi arte edo konbinazioak agortu arte.

Euskarazko corpus orokor handi bat osatzea helburu dela, bi metodoak probatu eta ebaluatu dira, crawling-arena eta bilatzaileena, ikusteko zein den onena euskararentzat, abiadura, kostua, tamaina edo kalitateari dagokionez [47].

Bilatzaileen metodoari dagokionez, parametro ezberdinak erabilita probatu da («hazi» hitzen luzera eta konbinazioen luzera ezberdinak). Bilaketek euskararentzat emaitza optimoa eman dezaten, gorago deskribatutako morfologia bidezko galderaren hedapena eta hizkuntza iragazteko hitzen teknikak erabiltzen dira berriz ere. Hala ere, kasurik onenetan 125 milioi hitz inguruko corpusak lortu dira, eta hazte-abiadura oso mantsoa da amaieran; beraz, ez da espero bide honetatik corpus askoz handiagoak lor daitezkeenik.

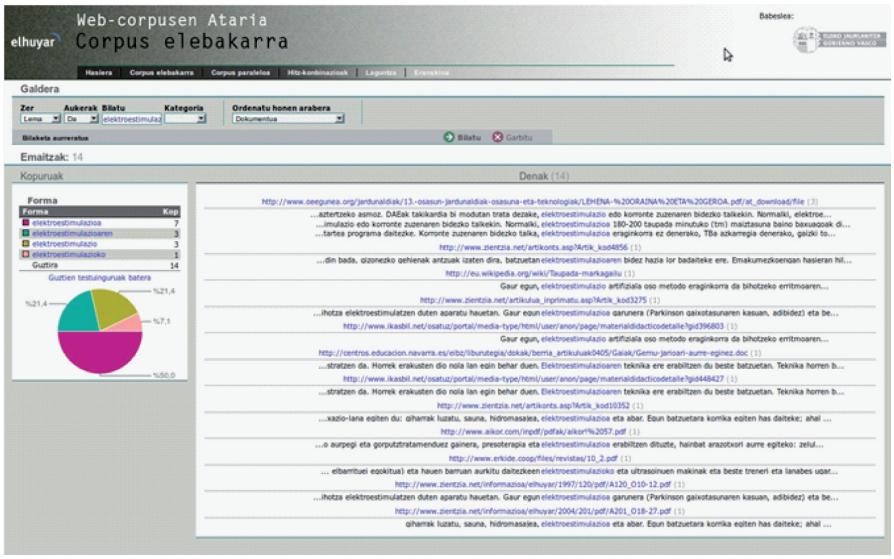
Crawling metodoarekin, 200 milioi hitzetik gorako corpusa eskuratu da, eta hazte-erritmoa ez da asko jaitsi; beraz, gehiago hazteko potentziala dauka.

Corpusen kalitatea ebaluatzeko, bilatzaileen bidez lortu den corpus handiena eta crawling bidez lortutako corpusa XX. mendeko Euskararen Corpusarekin eta Lexikoaren Behatokiko Corpusarekin konparatu dira,

hiru ezaugarriri begiratuta: corpus bakoitzeko hitz «erabilgarrien» kopurua (20 baino maiztasun handiagokoena), corpus baten estaldura besteekiko eta corpus baten ekarpena besteekiko [37].

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 bidez eta bai bilatzaileen bitartez 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 (http://webcorpusak.elhuyar.org/cgi-bin/kontsulta.py?mota=arrunta)[3].



2. irudia. Web-Corpusen Atariaren pantaila-irudia.

[3] Guk hau online jarri eta gutxira, Egungo Testuen Corpusa aurkeztu zen, 200 milioi hitzekoa.

4. EUSKARAZKO DOMEINU-CORPUS ESPEZIALIZATUAK OSATZEA WEBETIK

Webetik automatikoki corpus espezializatuak biltzeko metodo tradizionala crawling fokatua (focused crawling) [49] izan da. Crawling fokatua, finean, crawling egitean datza, baina nolabait bilketa gure helburu izango diren orrietara bideratzeko sistemaren bat ezarriz. «Hazi» URLak domeinuko orri asko dituzten webguneak izaten dira, eta crawling-ean webgune berri bakoitzeko ahalik eta orri gehien jaistea hobesten da beste webguneetara joan aurretik. Edonola ere, sistema honek amaieran iragazkiren bat pasatzea eskatzen du.

Baina 2004an BootCaT-ek [52] metodologia berri bat aurkeztu zuen, segituan estandar bihurtu zena: domeinuko «hazi» hitzen zerrenda batetik abiatuz, bilatzaileei horien konbinazioak bidaltzen zaizkie eta itzulitako orriak corpuseratzen dira. «Hazi» hitzek egiten duten iragazki lanari nahikoa iritzirik, jaitsi osteko iragazkirik ez da pasatzen. Baina honelako iragazkirik gabe lortzen den domeinu-zehaztasuna ez da hain ona, %33 inguru baztergarria izan daitekeela frogatu baita. Eta euskararekin proba egiteko egin ditugun esperimentuen arabera, lortzen den domeinu-zehaztasuna askoz txarragoa da, onartezin bihurtzeraino, berriz ere euskararen morfologiaren izaera aglutinatiboagatik eta bilatzaileek hizkuntza txikiekin duten utzikeriagatik.

Horregatik, bi metodoen konbinazio antzeko bat probatzea erabaki zen. BootCaT-en antzera, gure metodoak domeinuko hitzak eta bilatzaileak erabiliko ditu orriak jaisteko (baina euskararentzat errendimendu hobearekin), eta ondoren domeinu-iragazki sinpleren bat pasatuko du (baina sortzeko corpus handirik beharko ez duena).

Lehenengo zatiko helburua lortzeko, aurreko kapituluetan bezala, morfologia bidezko galderaren hedapena eta hizkuntza iragazteko hitzen metodoak erabili eta ebaluatu dira, eta ikusi da horiekin euskararentzat ere beste hizkuntzen antzeko domeinu-zehaztasuna lortzen dela, hau da, %66koa.

Bigarren zatia lortzeko, honako metodologia hau diseinatu da. Hasteko, helburu-domeinuaren erakusgarri izango den corpus txiki bat lortzen da hasieran (10-20 dokumentu labur nahikoa izan daitezke, domeinuaren espezializazio mailaren arabera). Lagin corpus horretatik automatikoki «hazi» hitzak erauzten dira [53] (ondoren, zerrendari eskuz egin dakioke orrazketa, nahi bada). Hitz horien konbinazioak bidaltzen zaizkie bilatzaileei (morfologia bidezko galderaren hedapena eta hizkuntza iragazteko hitzak aplikatuz) eta itzulitako orriak jaisten dira. Eta azkenik, orri hauei domeinu-iragazki bat pasatzen zaie, honetan datzana: orri bakoitza lagin corpuseko dokumentuekin konparatzen da banan-banan, dokumentu-antzekotasun teknikak [54] erabiliz, eta horietako bat edo batekin gutxieneko antzekotasun bat ez dutenak kanpo uzten dira.

Egin ditugun esperimentuetan [57], frogatu da metodologia hau erabilita eta antzekotasun atalasea nahikoa igota, zehaztasun oso ona lor daitekeela, %90 ingurukoa. Baina atalasea igotzeak estaldura jaisten du (%60 eta %40 arte ere jaits daiteke), euskara bezalako hizkuntza batean agian onargarria ez den neurrira (berez tamaina nahikoko domeinu-corpusak lortzea zaila izan daiteke kasu askotan, eta are gehiago estaldura hainbeste jaisten badugu). Beraz, emaitza oso onak lor ditzakegu domeinu-zehaztasunari dagokionez, baina ez dago argi domeinu guztietan beharrezko tamaina lortzerik ote dagoen exijentzia maila horrekin jokatuz gero.

Corpus baten kalitatea neurtzeko modurik onena ataza jakin batean duen emaitza neurtzea denez [23, 46], gure metodologia aplikatuta hainbat corpus jaitsi ditugu (informatika, bioteknologia eta partikulen fisika domeinuetakoak) eta terminologia-erauzketa automatikoko lanetan nola moldatzen diren aztertu dugu. Erauzterm [58, 59] tresnaren bidez corpus hauetatik terminoak erauzi ditugu, eta automatikoki erauzitako termino hauek Zientzia eta Teknologiaren Hiztegi Entziklopedikoarekin [60] balidatu dira. Bertan ez zeudenetako lehen hautagaiak eskuz ere balidatu dira, eta emaitzak eskuz bildutako antzeko corpus batzuenekin erkatu dira.

Webeko corpus bakoitzetik erauzi eta balidatu diren terminoen gehiengehienak dagokien domeinuko terminoak dira. Eta lortzen diren terminoen domeinu-zehaztasuna, estaldura eta termino berrien kopurua eskuzko corpusenen antzekoak edo hobeak dira. Beraz, frogatuta gelditu da automatikoki bildutako corpusak domeinukoak direla eta terminologia erauzketako lanetarako baliagarriak direla [61].

5. EUSKARA ETA BESTE HIZKUNTZA BATEKO DOMEINU-CORPUS KONPARAGARRIAK ERAUZTEA WEBETIK

Hizkuntza arteko corpus konparagarriak oro har eskuz bildu izan dira: berri-agentzietatik [62, 63], ikerketarako eskuragarri dauden corpusetatik edo webgune jakinak eskuz aukeratu eta jaitsita. Honela jokatzearen arazoa da corpusok ez direla askotarikoak (iturri gutxitatik datoz) eta, euskararen kasuan, ezin direla identifikatu domeinu bateko testu nahikoa duten iturri egokiak.

Guk baliatu nahi izan dugun metodoa [64] aurreko kapituluan azaldutakoa da, baina bi hizkuntzari aplikatuta. Hau da, bi hizkuntzatan domeinu-corpusen lagin txikiak eraikitzen dira, ahalik eta antzekoenak, eta horietatik aurreko kapituluko prozesua abiarazten da. Hizkuntza bakoitzarentzat lortzen den corpusa domeinu horretakoa bada, biek domeinu-corpus konparagarri bat osatuko dute.

Metodo hau ebaluatzeko, bera erabiliz bildutako bi corpusek, bata informatikakoa eta bestea fisikakoa [65], berriz ere terminologia erauzketa lan batean (kasu honetan, terminologia elebiduna) lortzen duten errendimendua neurtu dugu. Aurrekoan bezala, eskuz bildutako antzeko corpus batzuen errendimenduarekin konparatu ditugu emaitzok. Eta berriz ere, Zientzia eta Teknologia Hiztegiaren aurka balidatu dira erauzitako terminopareak, eta aurkitu ez direnetako batzuk eskuz. Oraingoan ere, erauzitako terminoen domeinu-banaketa, domeinu-zehaztasuna, estaldura eta termino berriak begiratu dira. Eta ikusi da erauzitako termino-pareak dagozkien domeinuetakoak direla eta webetik bildutako corpusek eskuz eraikitakoak baino emaitza hobeak lortzen dituztela.

Beraz, esan dezakegu webetik domeinu-corpus konparagarriak biltzeko sortu dugun metodoak benetan lortzen dituela corpus konparagarriak, terminologia erauzketarako baliagarriak direnak, eskuz bildutako corpusen antzeko emaitzekin. Hala ere, baliteke domeinu guztietarako ezin lortzea corpus konparagarrietatiko terminologia-erauzketak eskatzen duen tamainako corpusik.

6. **EMAITZAK ETA ONDORIOAK**

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

Helburu horren bila, lehenbizi 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 dira horiek, baina baita tesian bilatzaileen bidez corpusak biltzeko garatu diren beste tresna denetan eta euskarazko bilatzaile batean (Elebila) ere.

Ondoren, 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.

Geroago, tresna bat eraiki genuen, zeinak, bilatzaileak erabilita, corpus espezializatuak biltzen zituen oso lan gutxi eskatuta (soilik aurrez helburu-

domeinuko dokumentu labur gutxi batzuk biltzea). Tresnak oso domeinuzehaztasun ona lor zezakeela frogatu zen eta arrakastaz erabili zen terminologia-erauzketa automatikoko ataza batean. Hainbat corpus espezializatu bildu dira tresna horren bidez, terminologia eta ikerketa lanetan erabili direnak.

Azkenik, corpus espezializatuak biltzeko metodologian oinarrituta, domeinu-corpus eleaniztun konparagarriak biltzeko teknika bat asmatu eta garatu da. Hori erabiliz hainbat corpus konparagarri eratu dira, terminologia-erauzketa lan batean arrakastaz erabili direnak.

Areago, goian aipatutako corpusak biltzeko tresnen garapenak eskatu digu corpus garbiketako hainbat tresna garatzea (testuaren ingurukoak garbitzeko [66], 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 lezakeela, eta aldaketa hori tesian egindako lanarekin etorri dela. Beste hizkuntza handiagoen egoerarekin ezin dezakegu konparatu euskararena, baina metodologia eta tresna batzuk garatu ditugu dagoena biltzeko, eta asko bildu da, euskarazko corpusen kantitatea, aniztasuna eta tamaina modu esanguratsuan handitzeraino.

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 haien egoera ere hobetzeko.

7. **BIBLIOGRAFIA**

[1] KUČERA, H. eta FRANCIS, W.N. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, USA.

[2] ASTON, G. eta BURNARD, L. 1998. The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh, U.K.

[3] POMIKÁLEK, J., JAKUBÍCEK, M. eta RYCHLỲ, P. 2012. «Building a 70 billion word corpus of English from ClueWeb». Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 502-506.

[4] EUSKALTZAINDIA 1984. Orotariko Euskal Hiztegiaren Testu-Corpusa: http://www.euskaltzaindia.net/oeh. Atzipen-data: 2014/05/20.
[5] EUSKALTZAINDIA 2002. XX. mendeko Euskararen Corpusa: http://xxmendea.euskaltzaindia.net/Corpus/. Atzipen-data: 2014/05/20.
[6] UNIVERSITY OF THE BASQUE COUNTRY 2006. Ereduzko Prosa Gaur: http://www.ehu.es/euskara-orria/euskara/ereduzkoa/. Atzipen-data: 2014/05/20.
[7] SUSA 2005. Klasikoen Gordailua: http://klasikoak.armiarma.com/corpus.htm.
[8] ADURIZ, I., ARANZABE, M., ARRIOLA, J.M., ATUTXA, A., DIAZ DE ILARRAZA, A., EZEIZA, N., GOJENOLA, K., ORONOZ, M., SOROA, A. eta 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.
[9] EUSKALTZAINDIA 2009. Lexikoaren Behatokia: http://lexikoarenbehatokia.euskaltzaindia.net. Atzipen-data: 2014/05/20.
[10] ARETA, N., GURRUTXAGA, A., LETURIA, I., ALEGRIA, I., ARTOLA, X., DIAZ DE ILARRAZA, A., EZEIZA, N. eta SOLOGAISTOA, A. 2007. «ZT corpus: Annotation and Tools for Basque Corpora». *Proceedings of Corpus Linguistics 2007,* Birmingham, UK.
[11] EIZIE 2002. EIZIEren itzulpen memoriak: http://www.eizie.org/Tresnak/Memoriak. Atzipen-data: 2014/05/20.
[12] GIPUZKOAKO FORU ALDUNDIA 2011. Gipuzkoako Foru Aldundiaren itzulpen memoriak: http://www.gipuzkoa.net/imemoriak/. Atzipen-data: 2014/05/20.
[13] BIZKAIKO FORU ALDUNDIA 2011. Bizkaiko Foru Aldundiaren itzulpen memoriak: http://www.bizkaia.net/home2/temas/detalletema.asp?tem_codigo=6130. Atzipen-data: 2014/05/20.
[14] EUSKO JAURLARITZA 2010. Eusko Jaurlaritzako Itzulpen Zerbitzu Ofizialaren itzulpen memoriak: http://www.ivap.euskadi.net/r61-vedorok/eu/contenidos/ds_recursos_linguisticos/memorias_traduccion/eu_izo/memorias_traduccion_izo.html. Atzipen-data: 2014/05/20.
[15] EROSKI FUNDAZIOA 2010. Consumer Corpusa: http://corpus.consumer.es. Atzipen-data: 2014/05/20.
[16] KILGARRIFF, A. 2001. «Web as corpus». *Proceedings of Corpus Linguistics 2001,* Lancaster, UK, 342-344.
[17] FETTERLY, D., MANASSE, M., NAJORK, M. eta WIENER, J.L. 2004. «A large-scale study of the evolution of Web pages». *Software: Practice and Experience,* **34**, 213-237.
[18] SINCLAIR, J.M. 2005. «Corpus and text: Basic principles». *Developing linguistic corpora: A guide to good practice,* Wynne, M. (arg.). Oxbow Books, Oxford, U.K.
[19] THELWALL, M., TANG, R. eta PRICE, L. 2003. «Linguistic patterns of academic Web use in Western Europe». *Scientometrics,* **56 (3)**, 417-432.

[20] THELWALL, M. 2005. «Creating and using web corpora». *International Journal of Corpus Linguistics,* **10 (4)**, 517-541.
[21] FERRARESI, A. 2007. *Building a very large corpus of English obtained by Web crawling: ukWaC*. University of Bologna.
[22] SCHÄFER, R. eta BILDHAUER, F. 2013. *Web Corpus Construction*. Morgan & Claypool, San Rafael, USA.
[23] KILGARRIFF, A. eta GREFENSTETTE, G. 2003. «Introduction to the Special Issue on Web as Corpus». *Computational Linguistics,* **29 (3)**, 333-347.
[24] KILGARRIFF, A. 2006. «Googleology is bad science». *Computational Linguistics,* **33 (1)**, 147-151.
[25] VOLK, M. 2002. «Using the web as corpus for linguistic research». Tähendusepüüdja. Hatcher of the Meaning. A Festschrift for Professor Haldur Õim, Pajusalu, R. eta Hennoste, T. (arg.). University of Tartu, Tartu, Estonia.
[26] LÜDELING, A., EVERT, S. eta BARONI, M. 2007. «Using the web for linguistic purposes». *Corpus Linguistics and the Web,* Hundt, M., Nesselhauf, N. eta Biewer, C. (arg.), 7-24. Rodopi, Amsterdam, The Netherlands.
[27] HÜNING, M. 2001. WebCONC: http://gandalf.uib.no/lingkurs/templates_c/%25%256E%5E6ED%5E6EDF58DD%25%25labor.html.php. Atzipen-data: 2014/05/20.
[28] RESNIK, P., ELKISS, A., LAU, E. eta TAYLOR, H. 2005. «The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine». Proceedings of the 31st Meeting of the Berkeley Linguistics Society, Berkeley, USA, 265-276.
[29] RENOUF, A., KEHOE, A. eta BANERJEE, J. 2007. «WebCorp: an Integrated System for Web Text Search». Corpus Linguistics and the Web, Hundt, M., Nesselhauf, N. eta Biewer, C. (arg.), 47-67. Rodopi, Amsterdam, The Netherlands.
[30] FLETCHER, W.H. 2007. Web Concordancer: http://webascorpus.org/searchwac.html. Atzipen-data: 2014/05/20.
[31] FLETCHER, W.H. 2006. «Concordancing the Web: Promise and Problems, Tools and Techniques». Corpus Linguistics and the Web, Hundt, M., Nesselhauf, N. eta Biewer, C. (arg.), 25-46. Rodopi, Amsterdam, The Netherlands.
[32] VAN NOORD, G. 1997. NetKwic: http://www.hit.uib.no/corpora/ 2000-2/0135.html. Atzipen-data: 2014/05/20.
[33] LETURIA, I., GURRUTXAGA, A., ARETA, N. eta POCIELLO, E. 2008. «Analysis and performance of morphological query expansion and language-filtering words on Basque web searching». Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC), Marrakech, Morocco.
[34] LETURIA, I., GURRUTXAGA, A., ARETA, N., ALEGRIA, I. eta EZEIZA, A. 2013. «Morphological query expansion and language-filtering words for improving Basque web retrieval». *Language Resources and Evaluation,* **47 (2)**, 425-448.

[35] LETURIA, I., GURRUTXAGA, A., ALEGRIA, I. eta EZEIZA, A. 2007. «CorpEus, a «web as corpus» tool designed for the agglutinative nature of Basque». Proceedings of the 3rd International Workshop on the Web As Corpus (WAC), Louvain-la-Neuve, Belgium, 69-81.

[36] LETURIA, I., GURRUTXAGA, A., ARETA, N., ALEGRIA, I. eta EZEIZA, A. 2007. «EusBila, a search service designed for the agglutinative nature of Basque». Proceedings of Improving non-English web searching (iNEWS'07) workshop, Amsterdam, The Netherlands, 47-54.

[37] BARONI, M., BERNARDINI, S., FERRARESI, A. eta ZANCHETTA, E. 2009. «The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora». *Language Resources and Evaluation,* **43**, 209-226.

[38] KEHOE, A. eta GEE, M. 2007. «New corpora from the web: making web text more "text-like"». *Towards Multimedia in Corpus Studies,* Pahta, P., Taavitsainen, I., Nevalainen, T. eta Tyrkkö, J. (arg.). University of Helsinki, Helsinki, Finland.

[39] SCHÄFER, R. eta BILDHAUER, F. 2012. «Building large corpora from the web using a new efficient tool chain». Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 486-493.

[40] CALLAN, J., HOY, M., YOO, C. eta ZHAO, L. 2009. The ClueWeb09 Dataset.

[41] POMIKÁLEK, J., RYCHLÝ, P. eta KILGARRIFF, A. 2009. «Scaling to billion-plus word corpora». *Advances in Computational Linguistics,* **41**, 3-13.

[42] SHAROFF, S. 2006. «Creating General-Purpose Corpora Using Automated Search Engine Queries». *WaCky! Working Papers on the Web as Corpus,* Baroni, M. eta Bernardini, S. (arg.), 63-98. Gedit Edizioni, Bologna, Italy.

[43] BARONI, M. eta UEYAMA, M. 2004. «Retrieving Japanese specialized terms and corpora from the World Wide Web». Proceedings of KONVENS 2004, Vienna, Austria, 13-16.

[44] UEYAMA, M. eta BARONI, M. 2005. «Automated construction and evaluation of a Japanese web-based reference corpus». *Proceedings of Corpus Linguistics 2005,* Birmingham, UK.

[45] UEYAMA, M. 2006. «Evaluation of Web-based Japanese reference corpora: effects of seed selection and time interval». *WaCky! Working Papers on the Web as Corpus,* Baroni, M. eta Bernardini, S. (arg.), 99-126. Gedit Edizioni, Bologna, Italy.

[46] KILGARRIFF, A., REDDY, S., POMIKÁLEK, J. eta PVS, A. 2010. «A corpus factory for many languages». Proceedings of the 7th International Conference on Language Resources and Evaluations (LREC), Valletta, Malta, 904-910.

[47] LETURIA, I. 2012. «Evaluating different methods for automatically collecting large general corpora for Basque from the web». Proceedings of the 24th International Conference on Computational Linguistics (COLING), Mumbai, India.

[48] DUNNING, T. 1994. «Accurate Methods for the Statistics of Surprise and Coincidence». *Computational Linguistics,* **19 (1)**, 61-74.

[49] CHAKRABARTI, S., VAN DER BERG, M. eta DOM, B. 1999. «Focused crawling: a new approach to topic-specific web resource discovery». Proceedings of the 8th International World Wide Web Conference (WWW), Toronto, Canada, 545-562.

[50] BAYKAN, E., HENZINGER, M. eta WEBER, I. 2008. «Web page language identification based on URLs». Proceedings of the VLDB Endowment, Auckland, New Zealand, 176-187.

[51] BAYKAN, E., HENZINGER, M., MARIAN, L. eta WEBER, I. 2009. «Purely URL-based topic classification». Proceedings of the 18th International World Wide Web Conference (WWW), Madrid, Spain, 1109-1110.

[52] BARONI, M. eta BERNARDINI, S. 2004. «BootCaT: Bootstrapping corpora and terms from the web». Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC), Lisbon, Portugal, 1313-1316.

[53] SARALEGI, X. eta ALEGRIA, I. 2007. «Similitud entre documentos multilingües de carácter científico-técnico en un entorno web». *Procesamiento del Lenguaje Natural,* **(39)**, 71-78.

[54] LEE, M.D., PINCOMBE, B. eta WELSH, M. 2005. «An empirical evaluation of models of text document similarity». Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci), Stresa, Italy, 1254-1259.

[55] SEBASTIANI, F. 2002. «Machine learning in automated text categorization». *ACM Computing Surveys,* **34 (1)**, 1-47.

[56] SHAROFF, S. 2007. «Classifying web corpora into domain and genre using automatic feature identification». Proceedings of the 3rd International Workshop on the Web As Corpus (WAC), Louvain-la-Neuve, Belgium, 83-94.

[57] LETURIA, I., SAN VICENTE, I., SARALEGI, X. eta LOPEZ DE LACALLE, M. 2008. «Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision». Proceedings of the 4th International Workshop on the Web As Corpus (WAC), Marrakech, Morocco, 40-46.

[58] ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. eta URIZAR, R. 2004. «Linguistic and Statistical Approaches to Basque Term Extraction». Proceedings of GLAT 2004, Barcelona, Spain.

[59] ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. eta URIZAR, R. 2004. «An Xml-Based Term Extraction Tool for Basque». Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC), Lisbon, Portugal, 1733-1736.

[60] ELHUYAR FUNDAZIOA 2009. Zientzia eta Teknologiaren Hiztegi Entziklopedikoa: http://zthiztegia.elhuyar.org. Atzipen-data: 2014/05/20.

[61] GURRUTXAGA, A., LETURIA, I., SARALEGI, X. eta SAN VICENTE, I. 2009. «Evaluation of an automatic process for specialized web corpora col-

lection and term extraction for Basque». *Proceedings of eLexicography Conference 2009,* Louvain-la-Neuve, Belgium.

[62] BARZILAY, R. eta LEE, L. 2003. «Learning to paraphrase: An unsupervised approach using multiple-sequence alignment». Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT/NAACL), Edmonton, USA, 16-23.

[63] MUNTEANU, D.S. eta MARCU, D. 2005. «Improving machine translation performance by exploiting non-parallel corpora». *Computational Linguistics,* **31 (4)**, 477-504.

[64] LETURIA, I., SAN VICENTE, I. eta SARALEGI, X. 2009. «Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet». Proceedings of the 5th International Workshop on the Web As Corpus (WAC), Donostia/San Sebastian, Spain, 53-61.

[65] GURRUTXAGA, A., LETURIA, I., SAN VICENTE, I. eta SARALEGI, X. 2013. «Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making». BUCC-Building and Using Comparable Corpora, Sharoff, S., Rapp, R., Zweigenbaum, P. eta Fung, P. (arg.), 51-75. Springer, Dordrecht, The Netherlands.

[66] SARALEGI, X. eta LETURIA, I. 2007. «Kimatu, a tool for cleaning non-content text parts from html docs». *Proceedings of the 3rd International Workshop on the Web As Corpus (WAC)*, Louvain-la-Neuve, Belgium, 163-167.