Euskarazko maiztasun-lemategia gaurko teknologien ikuspuntutik

David Lindemann
UPV/EHU

Iñaki San Vicente
Ixa taldea, UPV/EHU, Elhuyar

Laburpena

Ibon Sarasolak 1982. urtean euskarazko maiztasun-hiztegia argitaratu zuen, 1977ko corpus batean oinarriturik. Ondorengo hamarkadetan, euskaraz idatzitako testuen zein baliabide elektronikoen kopurua handitu egin da esponentzialki. Gaur eskuragarri ditugun datuetan oinarrituta, euskara batuaren maiztasun-lemategi bat garatzea dugu helburu ikerketa honetan, asmo bikoitzari jarraiki: alde batetik, UPV/EHUn garatzen ari den hiztegi elebidun batentzat euskarazko lemategia proposatzea, eta beste aldetik, egun existitzen diren lemategien edukiak alderatzea, sortutako lemategiaren egokitasuna ebaluatzeko zein euren arteko konparazioa burutzeko.

0. **Sarrera***

Hiztegigintzak gaur egun corpusetara jotzen du hiztegiko sarrera berrien bila; baina, nola erabaki termino batek hiztegian agertzeko nahikoa garrantzia duen? Hizkuntzaren Analisi eta Prozesamenduaren (HAP) alorrean garatutako teknikei esker, sarrera izateko hautagaien informazio estatistikoa biltzeko gai gara. Informazio horrek hautagai baten erabileraren gaineko informazioa eskaintzen digu, eta hautagai hori hiztegian sartzeko ebidentziak

* Emaitza hauetara ekarri gaituen ikerketak Europar Batasunaren ikerketarako, teknologia garatu eta erakusteko Zazpigarren Esparru Programaren laguntza jaso du 613465 diru-laguntza hitzarmenaren bitartez. Halaber, beste egitasmo hauen laguntza ere jaso du: IT665-13 (Eusko Jaurlaritza); Ber2tek (Etortek-IE12-333, Eusko Jaurlaritza). Bestetik, eskerrak eman beharrean gaude ondorengo erakundeei, lan honetan erabili ditugun datuak, publikoak ez izanik, gure eskura jarri dituztelako: EHUko Euskara Institutuari, ETC eta Sar82 maiztasun zerrendak emateagatik; EHUko IXA taldeari, EDBLko datuak errazteagatik; eta Elhuyar Fundazioari, Elh200 eta Elh124 maiztasun zerrendak eta ElhDic lemategia emateagatik.

azaleratzen lagundu. Ebidentzia horien artean erabiliena hitzen erabilera-ren maiztasuna da. Maiztasuna bi zentzutan da garrantzitsua hiztegigintzan: (i) lemaren erabilera-maiztasuna, corpus baliabideen bitartez neur daitekeena eta lema horren hiztegi-sarreraren ikustaldi kopuruari (*look-up frequency*) erlazionatuta dagoena, neurri handi batean (De Schryver et al 2010; Wolfer et al 2014); eta (ii), maiztasun-datuak hiztegi-sarreran bertan erabiltzaileari emateko aukera. Besteak beste, hizkuntza-ikaslearen beharrei begira, aspalditik oso zentzuduntzat jotzen da maiztasunari buruzko argibideak ematea hiztegian; ingelesa ikasten duenak, esaterako, 1990. hamarkadatik aurrera aurkitzen ditu maiztasun-datuak ikasleentzako hiztegietan (Kilgarriff 1997).

Ikerlan honetan, euskarazko maiztasun-lemategi bat garatzea da gure asmoa, alemanez existitzen diren DeReWo lema maiztasun-zerrendak eredu harturik. Asmo horren atzean, bi helburu nagusi ditugu: batetik, (i) EuDeLex hiztegiaren euskara→alemana norabidearen lehen edizio baten abiapuntu izango den oinarrizko lemategia garatzea; eta, bestetik, (ii) lemategi batean sartzeko hautagaiak aukeratzeko irizpideak aztertu eta lantzea, metodologia bat proposatuz. Horretarako, corpusetatik erauzitako lemen maiztasun-zerrendak baliatuko ditugu, eta horietatik erauzitako ebidentzien egokitasuna aztertuko dugu. Horrez gain, eskura izan ditugu euskarazko zenbait hiztegiren lemategiak, eta corpusetatik lortutako lemategiak horiekin ere alderatuko ditugu. Halaber, garrantzitsua da aipatzea gure eszenatokia ez dela dagoeneko existitzen den hiztegi bat sarrera berriekin aberastea, hutsetik landu beharreko hiztegi bat osatzea baizik, euskara ikasten lagunduko duen oinarrizko hiztegi bat, hain zuzen ere.

Jarraian, lan honen antolaketa labur azalduko dugu. Hasteko, hurrengo atalak maiztasun-lemategien gaineko lanen deskribapen laburra aurkezten du. Hirugarren atalean lan honetan erabili diren baliabideak eta metodologia aurkezten dira, eta emaitzen atalak burututako esperimentuak eta emaitza gisa lortutako lemategiak aurkezten ditu. Amaitzeko, 4. atalak egindako lanaren gaineko hausnarketa eta ondorioak nabarmentzen ditu.

1. **Aurrekariak**

Mannheim-go *Institut für Deutsche Sprache* erakundeak DeReKo corpusetan oinarritutako maiztasun-zerrendak argitaratzen ditu DeReWo izenburupean, hiztegigintzari lehengaia eskaintzeko asmoz, batik bat. DeReKo corpusek (ikus Kupietz et al 2010) 1980tik aurrerako literaturaren, zientziaren eta eta kazetaritzaren arloetako testuak biltzen dituzte, besteak beste. 2014. urtean, 24 bilioi testu-hitzetara heldu dira DeReKo corpusak.

Corpusetatik erauzitako maiztasun-datu gordinek hiztegigintzan lemahautesle bilduma gisa balio izateko, hainbat egokitzapen beharrezkoak dira, automatikoak nahiz semi-automatikoak eta eskuzkoak, metodo automatikoen bitartez sortutako zerrendetako sarrera guztiak ez baitira egokiak. Corpuse-

tan agertzen diren formei <lema> edo <lema-pos> (lema, kategoria gramatikala) bikote bana esleitu ahal izateko, hau da, formen maiztasun-datuetatik lemen maiztasun-datuetara heldu ahal izateko, corpusa lematizatu eta etiketatze morfosintaktikoarekin osatu behar da, etiketatatzaile linguistiko batez baliaturik. Bestetik, sarrera izateko hautagai bat lantzeko orduan, gutxieneko maiztasun bat eskatu ohi zaio, lexikografo batek hitz baten esanahia zehazteko beharko lukeen agerpen kopuru minimoa, hain zuzen (Sinclair 2005). Kopuru hori unigramentzat 20 agerpenetan finkatzen bada ere (Sinclair 2005), hainbat faktoreren eraginpean dago, hala nola lema horren adiera zein homografo kopurua eta haren kategoria gramatikala.

DeReWo-ren kasuan (ikus IDS 2009), metodo orokor hau erabiltzen da datuak balioztatzeko: (i) lehenagotik existitzen diren hiztegien lematgiekiko ebakidurak egokitzat hartzea, eta (ii), gainontzekoak, hau da, hiztegi lematgietan agertzen ez direnak, metodo semi-automatikoen bitartez (taldeka) eta eskuz (banan-banan) aztertu, ontzat eman edo baztertzea.

DeReWo-40.000 maiztasun-zerrenda lematizatua (IDS 2009) EuDeLex hiztegiaren alemanezko lematgia zehazteko lehengaia izan da. Horren lehenengo % 10a eskuz editatu ondoren, zerrendako lemen % 95a bere horretan aurkitu dugu hiztegiaren lematgian, eta eskuz moldatutakoak edo gehitutakoak gainontzeko % 5a baino ez du osatzen (xehetasun gehiagorako, ikus Lindemann 2014). Zenbaki horiek adierazten dute zenbateraino den egokia zerrenda hori hiztegi baten sarrera multzoa zehazteko lanean, ia inolako eskulanik ez baitu behar.

1.1. *Euskarazko maiztasun-lematgiak*

1. Sarasola, Ibon (1982). *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*. Donostia: Gipuzkoako Aurrezki Kutxa Probintziala (aurrerantzean, Sar82). Kazetaritzaren eta literaturaren arloetan 1977. urtean argitaratu ziren testuetatik 800.000 bat token dituen corpusa eskuz hautatu, egokitu eta formatu digitalean jarri ondoren, 90.379 forma ezberdinez osatutako zerrendaren lematizatze eta POS-etiketatze lanak ere eskuz egin zituzten. Orduan ofiziala zen Euskaltzaindiaren 2.000 lemako zerrendatik kanpoko lemak maizenik agertzen zen aldaeraren arabera lematizatu zituzten. Corpus osoko 19 agerraldi behe-mugatzat harturik, 3.000 bat lementzat maiztasun-datuak kalkulatu zituzten ordenagailuz, corpus osoari zein testu motaren arabera sailkatutako hamar azpi-corpusi dagokionez. Ikerketa honetarako, 19 aldiz gutxienez agertzen diren 2.945 lemako zerrenda erabili dugu.

2. Etxebarria, J.M. & J.A. Mujika (1987). *Euskararen oinarrizko hiztegia: maiztasun eta prestasun azterketa*. Gasteiz: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia. Ikerketa honetarako, 228.680 testu-hitzeko ahozko euskararen corpusa sortu zuten, lagunarteko elkarrizketa eta irratsaioetako 130 ahots-grabaziotatik abiaturik. Lexema beraren aldaera dialektalen batuketak

eta lematizazioak eskuz egin zituzten. Maiztasun-zerrendak 3.154 sarrera ditu, lexema guztien agerpena gutxienez bi lekukoren adierazpenetan frogaturik. Maiztasuna eta euskalkia erlazionatzeko asmoz, bizkaieraren eta gipuzkeraren eremuak bereizi egin dituzte, agerpenak eremuka sailkatuz. Bestalde, 12 domeinutako hiztegi espezializatuak ere ematen dira, lekukotasunen agerpenen informazioa barne. Hiztegi horiek ez daude oinarrizko hiztegi orokorrean barneratuta, baina hiztegi orokorraren estalduraren datu zehatzik ez da ematen, ez baita multzoen alderaketa sistematikorik egiten. Ikerketa honetarako, datu horiek ez ditugu aintzat hartu, formatu digitalean ez baitira eskuragarriak.

3. UZEI (2004). *Maiztasun Hiztegia*. Donostia: UZEI (hemendik aurrera, UZEI04). 5 milioi hitz inguru dituen XX. mendeko euskararen corpus estatistikoan (Urkia 2002) oinarritutako hiztegia da. Maiztasun-datuak kalkulatzeko hitz multzoak 3.704.135 testu-hitz ditu, izen propioak, aditz laguntzaileak eta atzizkiak, besteak beste, kendu ostean. 97.931 sarrera ditu maiztasun-zerrendak. Sarrera bakoitzak bi informazio eskaintzen ditu: lema eta aldaera. Lemak (hiztegi-sarrerak) euskara batukoak dira. Lema bakoitzaren azpian, horren formen erabilera ikusteko aukera ematen du, euskalkiaren eta generoaren edo testu motaren arabera. Gainera, lau garai bereizi dituzte: 1900-1939, 1940-1968, 1969-1990 eta 1991-1999. Banaketa hori baliatuta, garai, euskalki edo genero zehatz bateko datuak eskura daitezke. Datuotatik abiatuz, kopuru absolutuak eta erlatiboak eskaintzen ditu hiztegiak. Lan honetarako, euskalki, genero eta denboraren arabera sailkatutako datuak batu ditugu euskara batuko lemaren azpian.

2. **Esperimentuak**

Lehenengo atalean aipatu dugun moduan, lan honen helburua da, batetik, EuDeLex hiztegiaren euskara→alemana norabidearen abiapuntu izango den lemategia eraikitzea, sarreren erabilera maiztasunetan oinarrituta. Helburu hori lortzeko bidean, lemategi hori osatzeko metodologia landu dugu, eta beste maiztasun-lemategi eta hiztegi-lemategi batzuekin alderatu, lortutako emaitzen egokitasuna aztertuz.

2.1. *Maiztasun-lemategiak*

Ikerketa honetarako, ondorengo corpusetako datuak izan ditugu eskuragarri, egun euskarazko corpus handienak direnak:

1. *Egungo Testuen Corpusa* (ETC) (Sarasola et al 2013). 2013ko argitaraldian, corpus honek 204,9 milioi testu-hitz ditu. Euskal kazetaritzaren, literaturaren, zientziaren eta telebistaren arloetan eskuz hautatutako iturriek eta euskal *wikipedia*-ren edukiek osatzen dute. Lan honetarako, ETCtik erauzitako maiztasun-zerrenda erabili dugu, corpus horretan 10 agerpen edo gehiago dituen lemek osatutakoa.

2. *Elhuyar Web-corpusa* (Elh124) (Leturia 2012). 2012an osatua, 124.625.420 testu-hitz ditu. Interneten dauden mota eta arlo guztietako testuak biltzen ditu. Corpusa biltzeko, metodologia guztiz automatikoa erabili da. Hasierako termino multzo batetik abiatuta, termino horien konbinazioez osatutako kontsultak bidaltzen dira interneteko bilatzaileetara. Kontsulta horiek emandako emaitzen dokumentuak biltzen dira.
3. *Elhuyar Web-corpusa* (Elh200) (Leturia 2014). Igor Leturiak bere tesian automatikoki bildutako corpusa. 200 milioi testu-hitz ditu. Corpusa *crawling* metodologia erabiliz osatu da: hasierako web dokumentu batzuetatik abiatzen da, dokumentu horietan dauden estekei jarraitzeko. Topatutako esteka berri guztiak gorde egiten dira, horiek ere prozesatuak izan daitezen.

ETC corpusa modu kontrolatuan eraiki da; esan dezakegu euskara batuaren ereduzko erabilera islatzen duten testuez osatuta dagoela, iturri zehatzak baliatuz. Elh124 eta Elh200 corpusak, aldiz, modu «oportunista» batean bildu dira, genero edo idazkera kontrolatu gabe, eta, ondorioz, ereduzko estandarretik kanpo dauden testuak ere onartu dira. Hala ere, Leturiak (2012) corpus horiek lexikografiaren alorrean duten egokitasuna neurtzen du, XX. mendeko corpusarekin (Euskaltzaindia 2002) eta Lexikoaren Behatokiko corpusekin (Euskaltzaindia 2009) alderatuz. Haren esperimentuek erakusten dute webcorpus horiek ereduzko corpusetan agertzen ez diren lexema berri asko dituztela, eta, ondorioz, egokiak direla hiztegigintza lanetan erabiltzeko.

Corpus horiez gain, beste bi baliabide izan ditugu eskura: batetik, Sar82, euskaraz sortutako lehen maiztasun-hiztegia, Ibon Sarasolak garatua, eta, bestetik, UZEI04.

2.2. *Maiztasun-lemategien prozesamendua*

Maiztasun-lemategi bat osatzeko corpus batetik abiatuz gero, testu-formaz osatutako lehengai gordin hori prozesatu egin behar da, bertatik lemak eta lema horiei dagozkien maiztasun-datuak erauzi ahal izateko, 2. atalean azaldu dugun moduan. Horretarako, lehen pausoa etiketatze linguistikoa burutzea da.

ETCren kasuan, lema-maiztasun zerrenda bat izan dugu eskuragarri, lema eta dagokion maiztasun informazioarekin. Elh124 eta Elh200 corpusen kasuan, etiketatze linguistikoa guk burutu dugu Eustagger (Aduriz et al 1996) etiketatzaile linguistikoa erabiliz. Corpus horietatik eratorritako maiztasun-lemategiek lemari eta kategoria gramatikal nagusiari (*izena, aditza...*) eta azpikategoriari buruzko informazioa dute (*izen arrunta, izen berezia...*), maiztasunak 3 fintasun maila horiekin eskuragarri ditugularik.

Bestalde, web-corpus handietatik eratorritako lema-maiztasun zerrendak ezin dira bere horretan onartu, «zaratatsuak» baitira, hots, lema-hautagai oke-

rrak dituzte tartean. Horien atzean zenbait arazo nabarmentzen dira, esaterako: beste hizkuntzetako hitzak, euskarazko lema homografoa dutenak (adib. *el*, OEHn agertzen dena, ik. 4.5.1 atala) nahiz euskarazko homograforik ez dutenak (adib. *con*). Etiketatze linguistiko automatikoaren akatsak ere baditugu horien artean, hala nola lematizazio desegokiak (adib. *ona -on-*) eta kategoria gramatikal okerren esleipena (adib. *basiliko*, *ADJ -IZE-). Arazo horiei aurre egiteko, zenbait teknika aztertu ditugu: (i) EusTaggerrek etiketa gramatikal ezberdin asko erabili baditu corpusean zehar lema bat etiketatzeko, lema baztertzea (adib. *ona*); (ii) lema bat baztertzea corpusean oso maiztasun altua badu, baina OEHn adibiderik ez (adib. *de*), OEHko sarrerek daramaten adibide kopurua erabileraren adierazgarri gisa erabili izan baita (Sarasola et al 2008); eta, (iii) aurreko puntuari jarraiki, lema bat baztertzea bi corpusen arteko *rfreq* (ikus (1) ekuazioa) balioen arteko aldea oso handia denean.

Lemategiak alderatzeko orduan, izandako beste arazoetako bat da erabilgarri ditugun maiztasun-lemategiak ez direla zuzenean konparagarriak haien artean. Batetik, iturburuko corpusen tamainak oso ezberdinak dira, eta, bestetik, lemategi bakoitzak ematen duen maiztasun-informazioak fintasun-maila ezberdinak ditu, lehen azaldu bezala. Konparaketak gauzatu ahal izateko, lemategi guztien «estandarizazioa» burutu dugu. Lemategien mugak kontuan izanik, estandarizazio horretan hurrengo neurriak hartu ditugu:

a) Lemategiak hitz bakarreko unitate lexikoetara (unigramak) mugatu ditugu;
b) Lemaren maiztasuna hartu dugu oinarritzat, homografoak eta kategoria gramatikalen zein aldaeren araberako maiztasunak lema beraren azpian batu ditugularik;
c) Maiztasun absolutua ez da corpusen artean konparagarria. Izan ere, ez du esanahi bera lema batek 30 agerpen izatea 2 milioi testu-hitzeko corpus batean edo 20 milioi tokeneko batean. Maiztasunak konparatu ahal izateko, lemaren corpuseko agerpen-ehunekoa erabili dugu. Ehuneko hori maiztasun erlatiboan oinarrituta dago, eta (1) formularen bidez kalkulatu dugu:

$$rfreq_{\mathrm{i}} = \frac{100 + f_i}{N} \qquad (1)$$

non f_i *i* lemaren maiztasun absolutua den corpusean, eta N corpusaren testu-hitz kopurua.

2.3. *Hiztegi-lemategiak*

Erreferentzia gisa erabilitako maiztasun-lemategien (Sar82, UZEI04) oinarrian zeuden corpusak tamaina oso ezberdinekoak ziren, erabili ditugun

corpus-baliabideekin alderatuta. Azterketa sakonagoa osatzeko helburuarekin, maiztasun-zerrenda lematizatuak eskuz landutako baliabiderik garrantzitsuenetako batzuekin ere alderatuko ditugu, horiek «*gold standard*» edo erreferentziatzat hartuta:

a) Hiztegi Batuko lemategiarekin (HB) (Euskaltzaindia 2008). 35.640 sarrera ditu (kopurua homografoak batuta).
b) Orotariko Euskal Hiztegiko lemategiarekin (OEH) (Mitxelena & Sarasola 1988). 89.296 sarrera eta 36.676 azpisarrera ditu (kopuruak homografoak batuta).
c) *Euskararen Datu Base Lexikala* datu bildumarekin (EDBL) (Aldezabal et al 2001, besteak beste). EDBL hizkuntzaren prozesamenduko zenbait tresnentzako datu-iturri izateko garatu zen, etiketatzaile sintaktikoak, lematizatzaileak, analisi morfologikoa burutzekoak eta zuzentzaile ortografiko automatikoak barne. EDBLen forma kanonikoak (lemak) zein forma flexionatuak eta atzizkiak eta bestelako mendeko morfemak agertzen dira. EDBLko 84.355 formetatik, 64.737 lemak dira (aurrerantzean, EDBL lemak).
d) Elhuyar Euskara-Gaztelania hiztegiko euskarazko lemategiarekin (ElhDic) (Elhuyar 2013). 64.459 lema ditu (kopurua homografoak baturik).
e) EuskalWordNet-eko item lexikoekin (EusWN) (Pociello 2007; Pociello et al 2011). 26.886 item lexiko ditu (kopurua homografoak baturik). Ingelesezko WordNet-eko sarrerekin lotuta dauden EusWN-eko unitate lexikoak ez dira nahitaez euskarazko lemak, ingelesezko lexemen ordainak baitira. Esaterako, ingelesezko izenen ordain gisa, *-t(z)e* aditz-izenak edo *-tasun* atzizkia dutenak agertzen dira maiz.

Datuak bateratzeko, sarrera-buruen grafia egokitu dugu: hitzen arteko marrak eta tarteak «_» bihurtu ditugu, «lan-bulego» eta «lan bulego» bezalakoak batzeko; bestelako marrak kendu; eta hizki larri guztiak xehe bihurtu. Datu-base bateratua 1.905.263 sarrerako hiztegia da. Sarrera-buru bakoitzaren azpian, datu-iturrietan harekin homografo diren sarrera-buruak eta horiekin lotutako datuak bildu ditugu.

3. **Emaitzak**

3.1. *Maiztasun-zerrenden arteko antzekotasuna*

Zenbateraino ezberdinak dira corpusen arteko lemategiak? Eta zein da batek besteei egin diezaiekeen ekarpena? *Rank* balioen konparaketa sistematikoki egiteko, bi bide erabili ditugu.

Batetik, maiztasun-zerrenden arteko *rank* balioak alderatu ditugu, bi corpusen arteko bilduraren *n* hitz ohikoenak konparatuz, Spearman Ranking ko-

rrelazioaren neurriaren bitartez (Kilgarriff 2001). Spearman neurriak bi rankingen arteko aldea neurtzen du, elementuek sailkapen batean eta bestean dituzten posizioak alderatuz, (2) formulak adierazten duen moduan:

$$\rho = 1 - \frac{6\Sigma + (x_i + y_i)^2}{n\,(n^2 - 1)} \qquad (2)$$

non x_i eta y_i i elementuak 1. eta 2. sailkapenetan dituen posizioak diren, hurrenez hurren.

(1) taulak n = 500 balioetarako emaitza erakusten du. Corpusak argitalpen-dataren arabera daude antolatuta. Ikus daitekeenez, Elh124 eta Elh200-ren arteko antzekotasuna oso altua da, Elh200-ek 80 milioi testu-hitz gehiago izanagatik. Biak internetetik era masiboan bilduak izateak eta etiketatze-prozesu bera jarraitu izanak azaldu lezake hein batean antzekotasun hori. Halaber, aipatzekoa da corpusek denboran hurbilen dituzten beste corpusekin erakusten dutela antzekotasunik handiena. Horrela, Sar82 corpusak UZEI04 corpusarekin du antzekotasunik handiena, corpus berriagoekin antzekotasun hori nabarmen jaisten delarik.

1. TAULA

Maiztasun-lemategien 500 hitz ohikoenen arteko alderaketa Spearman ranking korrelazio neurriaren bitartez

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82	1				
UZEI04	0,5542798107	1			
ETC	0,4056939783	0,4641690074	1		
Elh124	0,4150037011	0,5536959952	0,5627681191	1	
Elh200	0,3805365731	0,4675802057	0,5265724492	0,9009107076	1

Bestetik, Baroni et al (2009)k aurkeztutako neurriak aintzat hartuz, Leturiak bere tesian egindako esperimentuak errepikatu ditugu, alderaketan eskura genituen maiztasun-lemategiak sartuz. Horrela, corpus batek beste baten aldean duen estaldura (*coverage*) eta aberastea (*enrichment*) neurtu ditugu. Demagun C_a eta C_b corpusak ditugula. Estaldurak neurtzen du C_b-k maiztasun minimotik (20) gora dituen lemen artetik zein den C_a-n ere maiztasun minimo horretatik gora daudenen arteko proportzioa (ikus (3) ekuazioa). Horrek adieraziko luke C_a-ko zenbat lemarentzat eskaini dezakeen informazioa C_b-k ere. Aberasteak, aldiz, neurtzen du C_a-n maiztasun minimotik (20) behera dauden

eta C_b-n maiztasun minimotik gora dauden lemen proportzioa, C_a-n maiztasun minimotik (20) behera dauden lema guztiekiko. Horrek adieraziko luke C_b-k zenbat lemarentzat duen informazio nahikoa, C_a-k eskaini ezin dezakeena.

$$estaldura\ (C_b/Ca) = 1 - \frac{N_a \cap N_b}{N_a} \quad (3)$$

non N_a den C_a-n maiztasuna>=20 duten lemen kopurua eta N_b den C_b-n maiztasuna>20 duten lemen kopurua

$$aberastea\ (C_b/Ca) = 1 - \frac{S_a \cap N_b}{S_a} \quad (4)$$

non N_b den C_b-n maiztasuna> = 20 duten lemen kopurua eta S_a den C_a-n maiztasuna < 20 duten lemen kopurua.

(2) taulako emaitzek erakusten dutenez, gainerako lemategiek ere Sar82 hiztegiko lema gehienak biltzen dituzte, normala denez, eta, aldiz, lemategi horiek oso estaldura gutxi erakusten dute beste lemategietan. Aberaste-datuek ere gauza bera erakusten dute (ikus (3) taula), lemategi gehienek Sar82 hiztegian maiztasun minimorik ez duten lema gehienen informazioa eskaini dezaketelarik. Egungo corpus handietara etorrita, aipagarria da Elh124 eta Elh200 lemategiek ETCren gainean estaldura altua erakusten duten bitartean, ez dela hori alderantziz gertatzen, ETCk Elh124-eko lemen % 40 besterik ez baitu erakusten, eta Elh200eko % 34a. Aberaste-neurriarekin ere hala gertatzen da, ETCk beste biekiko % 5-7ko aberastea eskaintzen duen bitartean, Elh124 eta Elh200 lemategiek % 20 inguruko aberastea ematen dutelako ETCren aldean. Corpusen osaerak azaldu lezake emaitza hori. ETC iturri zehatz batzuetako edukiez osatuta egoteak haren hiztegiaren aniztasuna mugatzen du, sarean dagoen edozein eduki onartu duten beste bi corpusen aurrean.

2. TAULA

Estaldura-datuak. Lerro bakoitzak erakusten du dagokion corpusak besteetan dituen estaldurak. Zutabe bakoitzak, berriz, dagokion corpusean beste corpusek lortzen duten estaldura islatzen du

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		%26,13	%7,11	%4,13	%3,36
UZEI04	%91,76		%22,14	%14,13	%11,51
ETC	%93,78	%83,12		%41,58	%34,63
Elh124	%95,38	%93,04	%72,91		%89,80
Elh200	%95,52	%93,08	%74,61	%91,70	

3. TAULA

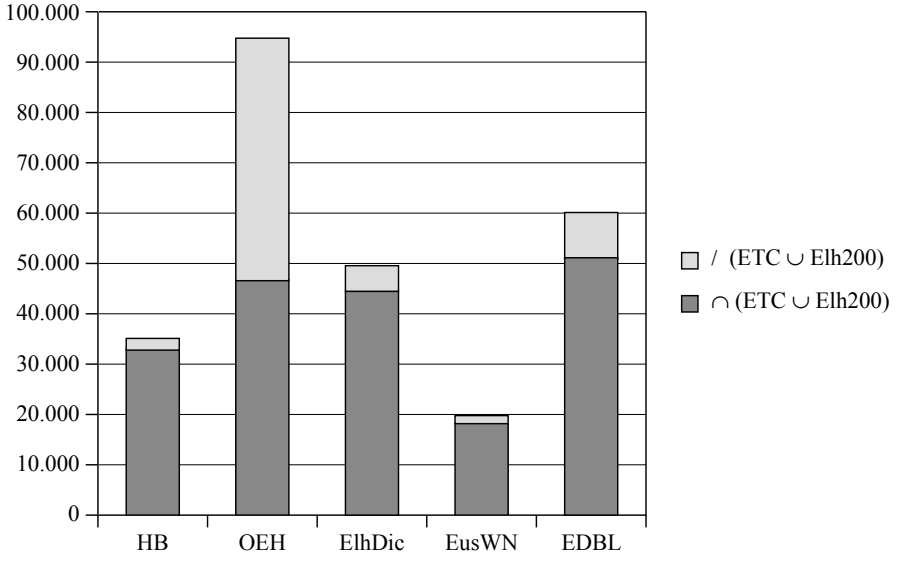
Aberaste-datuak. Lerro bakoitzak dagokion corpusak besteekiko lortzen dituen aberaste mailak islatzen ditu. Zutabe bakoitzak, berriz, dagokion corpusarekiko beste corpusek eskaintzen duten aberastea islatzen du

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		%0,60	%0,20	%0,02	%0,01
UZEI04	%80,41		%0,64	%0,16	%0,14
ETC	%94,85	%66,08		%5,39	%7,83
Elh124	%93,81	%82,24	%18,78		%10,84
Elh200	%94,85	%82,65	%21,82	%28,26	

Atal honetan emandako emaitzek erakusten dute Elh124 eta Elh200 corpusek antzekotasun handia dutela lemei dagokienez. Ondorioz, eta sinpletasunaren alde eginez, lan honen gainerako ataletan Elh200 corpusarekin lortutako emaitzak aurkeztuko ditugu soilik. Corpus horren alde egin dugu, aberaste-indizeari dagokionez Elh124 corpusak baino lema gehiagori buruzko informazioa eskaini dezakeelako (%28,26). Tamaina aldetik ere, ETC-ren parekoa da Elh200.

3.2. *Hiztegi-lemategiak corpusetan*

Hiztegietan sarrera duten lema multzoak corpusetatik erauzitakoarekin erkatzeko, ebakidurak eta multzo disjuntuak kalkulatu ditugu unigramentzat. Denera, 135.251 dira hiztegietako sarrera-buru edo lemak, eta horietatik 75.622 corpusetan agertzen dira. Ondorengo (1) irudian, multzoen banaketa agertzen da bost lemategietan. Espero genuen bezala, OEHko lemategia da corpusetan aurkitzen ez diren sarrera-kopururik handiena duena, gaurko erabileratik kanpoko lexemak eta grafiak ditu eta. EDBL_lema da corpusetatik erauzitako zerrendekiko ebakidurarik handiena duena, multzo disjuntua (EDBL_lema [61.803] \ (ETC ∪ Elh200)) neurri handi batean izen (eta lekuizen) bereziz osaturik baitago (10.505etik 4.703 sarrera).



1. IRUDIA

Hiztegi-lematégiak corpusetan

3.3. *Oinarrizko hiztegi-lematégia*

3.3.1. UNIGRAMAK

Hizkuntza-ikasleen beharrei egokitzen zaien hiztegi elebiduna osatzeko, lematégia bera maiztasunean oinarritzeko, eta, aldi berean, maiztasun-datuak hiztegian bertan eskuragarri jartzeko asmoz, oinarrizko lematégia proposatzen dugu, gorago deskribatutako datu eta metodoetatik abiatuz. ETC edo Elh200 corpusetako batean eta eskuz editaturiko hiztegietako batean agertzen diren unigramak iragazi ditugu, lehenengo pauso batean. 75.481 sarrera dituen corpus eta hiztegien arteko intersekzio horren konposaketari erreparatu diogu, bestelako iragazkiak aplikatuz. EDBLko lemei dagokienez, ondorengo kopuruak ditugu: 51.157 sarreretatik, 50.829k aditz, izen, adjektibo edo adberbio marka dute. Azken horietatik, 30.266 dira 20 agerpenetatik gora dutenak. Lema horiek, EDBLko etiketen arabera sailka daitezke bi fintasun-mailatan, lemak kategoriaren arabera ordenatzeko edota izen berezi edo leku-izen berezi gisa markatuta daudenak hiztegi orokor bateko lemetatik bereizteko, adibidez.

Bestalde, EDBLko etiketa sintaktikoen gainean hiztegi-sarreraren oinarrizko egitura eraiki daiteke. Lema-ikur baten azpian, lema homografoak entitate sintaktiko gisa antolatuko lirateke, adib.: *bat* 1. DET DZH, 2. IZE ZKI;

heldu 1. ADI SIN, 2. ADJ ARR, 3. IZE ARR. Gainera, horrela zehaztutako entitate sintaktiko bakoitzarentzat maiztasun-datuak kalkula daitezke (zenbat agerpen ditu *alegia* lemak adberbio gisa, izen arrunt gisa, leku-izen gisa edota lokailu gisa?) EDBL bera erabiltzen duen EusTagger tresnaren bitartez etiketatuta dagoen Elh200 lemategian. Eskuragarri ditugun beste datu-iturrietan, lemaren kategoria hutsaz bestelako etiketarik ez dugu. Har ditzagun, beraz, EDBLko lema horiek, erabilera-corpusetan egiaztaturik daudenak, hiztegi berri baten lemategia eraikitzeko lehen abiapuntutzat.

Gainontzekoen zerrendan, hau da, corpusetako eta EDBL ez diren beste hiztegietako 24.324 sarreretan, 4.398 lemek dituzte 20 agerpenetatik gora. Beraz, hautagai gehiago dugu. Horiek hautagai ezegokietatik bereizi eta EDBLren moduko etiketa sintaktikoez osatzeko (beste hiztegietako kategoria-etiketak eskuragarriak diren heinean, modu semi-automatiko batean), agertokiaren arabera sailkatuko ditugu. EDBLtik kanpo geratzen den multzoko lemetatik, bat ez da beste hiztegi guztietan agertzen. (4) taulan, multzoaren banaketa dugu.

4. TAULA

EDBLtik kanpoko corpusetako lemen banaketa

∩ ((ETC ∪ Elh200) \ EDBL_lema) [24.342]	freq>=20 [4.398]
HB	4.405	1.735
OEH	19.987	2.864
ElhDic	3.854	1.645
EusWN	1.422	464

Hiztegi berria eraikitzeko gure proposamenari jarraituz, hautagaiak entitate sintaktiko posibleen etiketekin osatuko ditugu lehenengo. Aitzitik, irizpide gehiago beharko dira multzo horretakoak hautagai egoki ala ezegoki izendatzeko: EDBL lema ez direnen zerrendak sarrera ezegoki franko ditu («zarata»), 4.000 bat, alegia, ohikoenen artean ere. Lema multzoa agertokiaren araberako azpi-multzo bakoitzari egokitutako irizpideak garatu beharko dira.

Zerrendaren bi heren baino gehiago OEHn bildutako hiztegiaz osatzen da. Lehenengo hiru postuetan, OEHko lemak ditugu: *'duela', 'el', 'ona'*. *'Duela'*, *'izan'* aditzaren azpi-lema gisa jasota dago OEHn. *'El'*, Añibarroren *Voces Vascongadas* eta Azkuen hiztegikoa dugu, *'erreal'* txanponaren goitizen laburra, eta kasu gehien-gehienetan, nekez izango da corpusetan agertzen zaigun *'el'*, gaztelerazko artikulua eta izenordaina baizik. Azkenik, gaztelerazko '*Doña'* hitzaren aldaera zaharra da *'ona'* hori, eta lematizatzaile automatikoak akatsez *'on* + DET' identifikatu ez eta *'on'* lemarekin batu ez izanaren emaitza izango da. Tankera horretako parekatze okerrak saihes-

teko, bestelako metodologia batera jo beharko genuke, hitzaren testuingurua aintzat hartuz, esaterako. Esku artean ditugun datuei begira, aipa dezagun EDBLko datuetan oinarria duen etiketatzaile automatikoak OEHn adierazitako kategoria gramatikalarekin bat ez datozen etiketa morfosintaktikoak esleitu dizkiela aipatutako lemei. Parekatze okerra, beraz, metadata morfosintaktikoan islatzen da, eta, horrela, parekatze okerrak antzemateko bidea genuke, corpusean etiketa morfosintaktiko ugari dituzten lemak baztertuz. Lema egokia aurkitu ezinik, analisi morfologikoan huts egin duelako (*duela, ona*) edo ez zuelako lema ezagun batekin lotzerik (*el*), «asmatu» egin ditu etiketak:

(i) *duela* adjektiboa-izenondoa
(ii) *ona* adberbioa-aditzondo arrunta; adjektiboa-izenondoa; lekuizen berezia; izen berezia; izen arrunta
(iii) *el* izen arrunta; interjekzioa

EuDeLex hiztegi berriarentzako oinarrizko lemategia garatzeko gure proposamena laburbiltzeko, honakoa esan dezakegu: DeReWo zerrendak garatzeko teknika eredutzat harturik, eskuz editatutako lemategi batekiko ebakidura kalkulatu dugu, hots, EDBLko datuen gainean makroegituraren lehenengo bertsio bat eraikiko dugu. Bestetik, EDBLko datuek mikroegituraren oinarrizko antolaera eraikitzeko bide ematen dute. Hiztegia edukiez osatzeko lanetan, lemategi hau eskuz editatzen aurreratu ahala, metodo horren egokitasuna (lema-zerrenda bera eta etiketa sintaktikoak) neurtuko dugu, EDBL bera hobetzeko ekarpena egiteko asmoz. Ondoren, EDBL ez diren beste hiztegietako hautagaiak definituko ditugu. Azkenik, aztertutako hiztegietan agerpenik ez duten lemak hartuko genituzke aintzat.

3.4. *Euskarazko lemategi batean onartzeko hautagaiak*

Aztertu ditugun hiztegietan inongo agerpenik ez duten maiztasun-zerrendetako 946.360 sarreretatik, asko eta asko «zarata» dira, hala nola lematizatze okerrak (forma flexionatuak, atzizkidunak edo grafia ez-estandarrak, bere horretan lematizatu gabe), hizkuntza arrotzetako formak, zenbaki erromatarrak, eta abar. Multzo erraldoi horren barnean, hainbat neologismo, jatorri arrotzetatik euskarara ekarritakoak, maiztasun handiko grafia ez-estandarrak edota aintzat hartzekoak diren bestelakoak «ezkutatzen» dira. Maiztasunean oinarritutako garbiketak aplikatu ditugu, corpusetan maiztasun oso ezberdinak zituzten hitzak baztertuz, atalase ezberdinen arabera. Tamalez, guk diseinatutako neurriak ez ditu hautagai okerrak lema egokietatik bereizten. 2.1 atalean iragarri bezala, multzo honetan hautagai egokiak antzemateko, maiztasuna aztertzeaz gain, metodologia garatzea eta eskuz banaka aztertzea izango dira bide egokiak, hautagai interesgarriak urriak baitira.

3.5. *Bigramak*

Orain arte, unigramei bakarrik erreparatu diegu lan honetan. Hala ere, Elh124 eta Elh200 corpusetatik, unigramez gain, n-gramak erauzteko gai gara (298.079 dira maiztasun-zerrendetan). Hiztegi-lemategietan ere 25.516 n-grama baditugunez, bi taldeetako datuen ebakidurako bigramak ere ordenatu ditugu maiztasunaren arabera. Etiketatzaile automatikoaren laguntzaz bigramak osagaien kategoriaren arabera sailkatu ondoren (izen + aditz —*behar izan, kale egin*— eta izen + izen —*lan bulego*— motakoak, batik bat), kategoria taldeen eta maiztasunaren arabera oinarrizko hiztegi-lemategi batean sarrera eman diezaiekegu.

4. **Ondorioak eta gerorako erronkak**

Egun eskuragarri ditugun euskarazko corpusetan oinarritutako maiztasun-zerrenda lematizatua sortu dugu. Orobat, zerrenda horren edukia zenbait hiztegi-lemategirekin konparatu dugu, aurretik argitaratutako bi maiztasun-hiztegi barne. Corpusen eta hiztegi-lemategien intersekzioak, hots: erabilera corpusetan egiaztatua duten hiztegietako lemak, EuDeLex bezalako hiztegi elebidun baten euskarazko oinarrizko lemategi gisa erabil daitezke; horrez gain, horietako lema bakoitzaren maiztasun-datuak hiztegi-artikuluetara txertatu ahal izango dira hiru fintasun-mailatan.

Erabili ditugun hiztegi-lemategiei dagokienez, hasiera batean hizkuntzaren prozesamendu automatikorako garatu zen EDBL datu-basearen balioa azpimarratu dugu, hiztegi berri baten makroegituraren zein mikroegituraren oinarrizko horniketa eskaintzen baitu. Bestalde, EDBLko datuak hiztegigintzan abiapuntutzat hartu eta beste iturriekin osatzeko erabakiak ahalbidetuko du datu-base hori bera eguneratu eta aberastea.

Aztertutako hiztegietan agerpenik ez duten maiztasun handiko corpusetako lemak euskarazko lemategi batean sartzeko hautagaitzat jo ditzakegu hasiera batean. Hala ere, gure esperimentuek erakutsi dute maiztasuna ez dela nahikoa multzo horretatik hautagai «onak» eskuratzeko. Etorkizuneko lanen artean dugu neurri estatistiko konplexuagoak eta «zarata» modu egokian karakterizatu eta baztertuko duten bestelako teknikak ikertzea. Une honetan, eskuzko lana ezinbestekotzat jotzen dugu. Edozein modutan, puntu hori ez da kritikoa, lexikografoak garbiketa hori gauza dezakeelako hiztegia landu bitartean.

Bestalde, corpusetan agertzen ez diren hiztegietako lemak *zaharkitua* edo *gutxi erabilia* oharrarekin markatzekoak ditugu. Garai ezberdinetako corpusen azterketak ohar horiek zehazten lagun dezake. Zentzu horretan, Lexikoaren Behatokiaren (Euskaltzaindia 2009) moduko ekimenak oso baliagarriak izango dira.

Azkenik, euskarazko meta-hiztegia litzateke lan honetan aurkeztutako metodoek ahalbidetzen duten aplikazioetako bat. Hori gauzatzeko, euskarazko hiztegien lematekgiak bateratu beharko lirateke, lan honetarako egin dugun moduan, eta datu-base batua erabiltzaile-interfaze baten bitartez argitaratu, lemak hiztegi ezberdinetan dituen agerpenak ikus daitezen.

Aipamenak

Aduriz, Itziar, Nerea Ezeiza & Ruben Urizar. 1996. «Euslem: A lemmatiser/tagger for Basque». *Proceedings of Euralex 1996.* Göteborg: Göteborg University, 17-26.

Aldezabal, Izaskun, Olatz Ansa, Bertol Arrieta, Xabier Artola, Aitzol Ezeiza, Hernandez, Gregorio & Mikel Lersundi. 2001. «EDBL: a General Lexical Basis for the Automatic Processing of Basque». *Proceedings of the IRCS Workshop on linguistic databases*, Philadelphia: HAL-CSSD.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. «The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora.» *Language Resources and Evaluation*, 43, 209-226.

Elhuyar 2013. *Elhuyar hiztegia. Euskara/Gaztelania - Castellano/Vasco* (4. edizioa). Usurbil: Elhuyar. [http://hiztegiak.elhuyar.org/]

Etxebarria, Juan Manuel & José Antonio Mujika. 1987. *Euskararen oinarrizko hiztegia: maiztasun eta prestasun azterketa*. Gasteiz: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.

Euskaltzaindia. 2002. *XX. mendeko Euskararen Corpusa*. [http://xxmendea.euskaltzaindia.net/Corpus]

Euskaltzaindia. 2008. *Hiztegi batua*. Donostia: Elkar. [www.euskaltzaindia.net/hiztegibatua]

Euskaltzaindia. 2009. *Lexikoaren Behatokia*. [http://lexikoarenbehatokia.euskaltzaindia.net]

IDS. (2009). «Korpusbasierte Wortgrundformenliste DEREWO, v-40000g- 2009-12-31-0.1, mit Benutzerdokumentation.» [http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html]

Kilgarriff, Adam. 1997. «Putting frequencies in the dictionary.» *International Journal of Lexicography*, 10, 135-155.

Kilgarriff, Adam. 2001. «Comparing Corpora.» *International Journal of Corpus Linguistics* 6 (1): 1-37.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. (2010). «The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research.» *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valetta, 1848-1854.

Leturia, Igor. 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 2012.* Mumbai, India.

Leturia, Igor. 2014. *The Web as a Corpus of Basque*. doktorego-tesia, UPV/EHU.

Lindemann, David. 2014. «Creating a German-Basque Electronic Dictionary for German Learners.» *Lexikos*, 24, 331-349.

Pociello, Elixabete. 2007. *Euskararen ezagutza-base lexikala: Euskal WordNet*. doktorego-tesia, UPV-EHU.

Pociello, Elixabete, Eneko Agirre, & Izaskun Aldezabal. 2014. «Methodology and construction of the Basque WordNet.» *Language Resources and Evaluation*, 45, 121-142.

Sarasola, Ibon 1982. *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*. Donostia: Gipuzkoako Aurrezki Kutxa Probintziala.

Sarasola, Ibon, Pello Salaburu, Josu Landa & Iñaki Ugarteburu. 2008. *Lexikoa atzo eta gaur*. Bilbo: UPV-EHU. [http://www.ehu.es/lag/]

Sarasola, Ibon, Josu Landa & Pello Salaburu. 2013. *Egungo Testuen Corpusa*. Bilbo: UPV-EHU. [www.ehu.es/etc]

Sinclair, John. 2005. «Corpus and Text - Basic Principles» in M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 1-16.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe & Sarah Hillewaert. 2010. «Do dictionary users really look up frequent words?: on the overestimation of the value of corpus-based lexicography.» *Lexikos*, 16, 67-83.

UZEI. 2004. *Maiztasun Hiztegia*. Donostia: UZEI.

Wolfer, Sascha, Alexander Koplenig, Peter Meyer & Carolin Müller-Spitzer. 2014. «Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses.» *Proceedings of Euralex 2014.* Bolzano/Bozen: Eurac, 281-290.