Euskararen erreferentzia-corpusaren beharraz

Miriam Urkia
UZEI

Azken urteetan corpusak bazterreko izatetik hizkuntzaren ikerketan oinarrizko tresna izatera pasa dira, baita gure artean ere.

Euskaltzaindiak aspaldi egin zuen corpusaren aldeko apustua: tradizioa jasotzen duen *Orotariko Euskal Hiztegia*-ren oinarrizko obrak batetik, eta XX. mendeko euskara biltzen duen EEBS *(Egungo Euskararen Bilketa-lan Sistematikoa) b*estetik. Hain zuzen, lan hauetan oinarritu dira *Orotariko Euskal Hiztegia* bera, EGLU liburukiak eta *Hiztegi Batua*, besteak beste. Gaur ez genituzke eskura izango halako tresnarik gabe.

1963an, *Brown* corpusa kaleratu zenetik, asko aldatu da corpusei buruzko ikuspegia, azken hogei urteetan batez ere. Eta bi aldaketa nabarmen islatzen ditu honek: batetik, hizkuntzaren ikerketa enpirikoak eta estatistikoak gora egin du; bestetik, teknologia-aurrerapenek prozesatzeko ahalmena ekarri dute, masa handiak modu erosoan ustiatzea ahalbidetuz.

Erabiltzaileak ere, orain arte hizkuntza naturalaren prozesamenduan aritzen zirenak eta lexikografoak ziren batez ere, baina egun erabileren eta erabiltzaileen dibertsifikatzea etorri da, corpusak edozeinen eskura baitaude eta, ahaltsuak izateaz gain, eskuragarriak ere badira.

Lexikografoen artean, esaterako, gaur susmagarria da erabilera dokumentatuan oinarritzen ez den lana: corpusetan frogatzen dira proposamenak. Ez da corpusaren beharra planteatzen; eztabaida tamainan eta edukian dago, corpus orekatua, handia eta ona izatea zaila baita.

Baina, zer da corpus bat? Hemendik abiatuta osatuko dugu txosten hau, izenburuari begiratuta. Hasteko, corpusa zer den argituko dugu, ondoren erreferentzia-corpusa azaltzeko. Eta, azken atalean, euskararen erreferentzia-corpusaren beharraz arituko gara, hori izango baita, gure ustez, XXI. mendeko ikerketen oinarria.

1. Corpusak

Definizio zabalena hartuta, corpusa *hizkuntzaren ikerketarako oinarrizko tresna* dela esan genezake, goian aipatu dugun bezala. Alegia, testu-bilduma da, hizkun-tzari buruzko azterketak egiteko eta hipotesiak frogatzeko erabiltzen dena. Testu-masa handia izan ohi da, datu-base egoki batean antolatua eta hizkuntzaren erakusgarri dena, benetako erabilerak biltzen dituena.

Baina, corpusa osatzen hasi aurretik hainbat galdera izan behar dira kontuan: nolako corpusa nahi dugu? zer islatu nahi dugu? zertarako, zein informazio eskaini behar du? Sartu dugun informazioaren eta honen kalitatearen araberakoa izango da eskuratuko dugun emaitza ere. Corpusak orekatua, erakusgarria behar du, hau da, baliagarria. Horrela, gauza izango da adierak, kategoria sintaktikoak, klase semantikoak, hitzen arteko kookurrentziak, unitateen agerpen-maiztasuna, unitateak bere testuinguruak, erabilera-adibideak, murriztapen selektiboak, hitz elkartuak, lexiak, lokuzioak, etab. eskaintzeko. Dudarik ez da, corpusa zenbateraino egituratu, etiketatu, lematizatu eta sailkatu den hartu beharko dela kontuan.

Beraz, nolakoak izan daitezke corpusak? Testualak edo ahozkoak, edo bietarikoak, gaur egin ohi den bezala. Baina, horren barruan, erreferentzia-corpusak, estatistikoak, paraleloak, konparatuak, bereziak, esperimentalak, literarioak edo bestelakoak izan daitezke. Eta, gehiago finduz, hauen barruko sailkapenak ere egin daitezke: hedadura, aldaera historikoa, geografikoa, dialektala... izan daiteke muga-irizpide.

Nolanahi ere, egonkorrak izan behar dute, eta horren arabera, irekiak ala itxiak izango dira, alegia, epe bat landu eta ez eguneratu (CTILCen[1] kasuan bezala) edo eguneratu eta osatu egiten direnak, beti ere diseinuaren aldetik egonkortasuna mantenduz (CREA[2] eta EEBS[3], esaterako).

Esan dugunez, lehen corpusa 60ko hamarkadan kaleratu zen, *Brown* corpus[4] izenez ezagutzen dena eta hogei urtean eredu izan dena. Baina milioi bat hitz besterik ez zuen, sailkapen oso orokorra, 2000 testu-hitzeko 500 lagin edo obra-zati eta iturri idatzi argitaratuetara mugatzen zen. Ordurako asko zen, baina berehala ohartu ziren mugatuegia zela eta 80ko hamarkadan, John Sinclair-en gidaritzapean, 7.3 milioi hitzeko ingeleseko corpusa osatu zuten *Birmingham Collection of English Texts* (BCET) taldekoek, nahiz COBUILD hiztegia osatzeko 20 milioitara zabaldu zuten 1987an. 90eko hamarkadan 320 milioi zituen honek berak, *COBUILD-Bank of English*[5] izena hartuta. *British National Corpus*-ek[6] ere, 1994an, 100 milioi zituen jasoak, berrikuntza nagusi batekin: testu idatziak eta ahozkoak biltzen zituen lehena zen.

Ingeleseko adibideak dira hauek, baina beste hizkuntzetan ere hasi ziren corpusak osatzen. Adibidez, *The Bank of Swedish*-ek 75 milioi testu-hitzeko corpusa du, CREAk 125ekoa eta CTILCek 52.3koa. Hala ere, badira txikiagoak eta, aldi berean, zehatzago etiketatuak ere.

Gaur egun EAGLES[7] da corpusen osaerarako irizpideak zehazten dituena, estandartzat hartzen dena. Corpusa *"a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language*" bezala definituz, irizpideok markatzen ditu: ahalik eta handiena izan behar du, lagin desberdin asko bildu erakusgarri izan dadin, erdi mailako sailka-

1 CTIL C (Corpus Textual Informatitzat de la llengua Catalana), Institut d'Estudis Catalans-ek egindakoa: 1833-1988 epea jasotzen du eta egungo katalaren hiztegi deskriptiboaren (DCC: Diccionari del Catalá Contemporani) oinarri da.
2 CREA (Corpus de Referencia del Español Actual), Real Academia Española: azken 25 urteetako produkzioa biltzen du. 1975eko testuekin hasi ziren eta, 25 urteak gaindutu ostean, CORDEra (CORpus Diacrónico del Español) pasatzen dira lehen urteetakoak. Horrela, beti dituzte azken 25 urteak. (www.rae.es).
3 EEBS (Egungo Euskararen Bilketa-lan Sistematikoa): 1900. urtetik gure egunetaraino iristen da, urtero eguneratuz eta handituz. Hala ere, XX. mendeko corpusa denez, urtearen bukaeran corpus itxi izatera pasako da, alegia, ez da eguneratzen jarraituko.
4 www.hit.uib.no/icame/brown/bcm.htlm
5 http://titania.cobuild.collins.co.uk/boe_info.html
6 www.info.ox.ac.uk/bnc/
7 EAGLES (Expert Advisory Group on Language Engineering Standards): http://ww.ilc.pi.cnr.it/EAGLES96/

pena, iturriak azaldu, testu idatziak eta ahozko transkripzioak. Bost irizpideotan oinarritzen dira fidagarri eta erakusgarri izan nahi duten corpusak, eta azken urteotan gorakada handia izan da, eskanerrak eta euskarri elektronikoan eskuratzeko erraztasunak lagunduta. Horrez gain, baliabide informatikoek lematizazioan eta etiketatzean laguntza handiak eskaintzen dituzte.

Euskararen corpusei begiratuta, OEH eta EEBS aipatu behar dira:

OEH nolabait erreferentzia-corpus historiko itxia dela esan dezakegu, 310 obra oso (edo ia oso) aukeratu biltzen baititu, 5.800.000 hitzez osatua; testu gordina da, kodetu gabea eta lematizatu gabe dago. Sailkapen zabala du: epea, euskalkia eta testumota zabala.

EEBS, berriz, XX. mendeko euskara jasotzen duen corpus estatistiko irekia da, oraingoz behintzat: 1998 arteko 6.047 obra-zatitatik jasotako 4.237.000 testu-hitz ditu (euskal argitalpenen inbentarioan oinarrituz, unibertso osoa proportzionalki ordezkatzen du zozketa bidez aukeratutako laginak), SGML[8] formatu estandarrean kodetuak, lematizatuak (98.800 lema desberdin), sailkapenaren arabera erakusgarria da (epea, euskalkia, testu-mota eta obraren tamaina) eta urtero eguneratzen da oraingoz, irizpideen oreka mantenduz.

2. ERREFERENTZIA-CORPUSAK

Corpusak era askotakoak izan daitezkeela ikusita, osoenak eta hizkuntzaren erakusgarrienak erreferentzia-corpusak direla ohartzen gara. Hala definitzen du EAGLESek: *"A reference corpus is one that is designed to provide comprehensive information about a language"*, alegia, hizkuntza, bere osotasunean hartuta, erakusteko diseinatua egon behar du corpusak: hizkuntzaren aldaera esanguratsuak adierazteko besteko tamaina eta kalitatea behar du.

Erreferentzia-corpusak hierarkikoki antolatuta daude, azpicorpusetan banatuta: hau da, corpus desberdinek osatuko dute nagusia. Honen arazoa adierazgarritasun-balantzea zehaztean datza, baina kontsultarako aukera desberdin asko eskaintzen ditu, kontsultak beharren arabera mugatuz. Dudarik ez da etorkizuneko lan gehienen oinarri izango dela.

Gorago aipatu dugun *Bank of English*[9] da adibideetako bat, azpicorpusak hala banatuz: egunkariak (% 25.7), liburuak (% 22.1), aldizkariak (% 22.7), irratia (% 23.3), *ephemera* (% 0.9) eta ahozkoa (informala, % 4.8). Hauek, aldi berean, beste azpicorpusetan eta osagaietan banatzen dira.

RAEko CREAk azken 25 urteetako 125 milioi hitz biltzen ditu, 100 gai desberdinetan banatuz, testu idatziak eta ahozko transkripzioak dituela. Idatzien artean liburuek, egunkariek eta aldizkariek osatzen dituzte azpicorpus nagusiak. Ahozkoen artean, berriz, elkarrizketak, berriak, magazinak, dokumentalak etab. jasotzen dira.

CTILC katalanak, 3299 obra osotako 52.3 milioi testu-hitz ditu: % 44 literarioa (narratiba, poesia, antzerkia eta saioa) eta % 56 ez-literarioa, hamar taldetan banatua (filosofia, erlijioa/teologia, prentsa, giza zientziak,...).

8 SGML (Standard Generalized Mark-up Language): http://www.uic.edu/orgs/tei/sgml/teip3sg/
9 Ikus 6. oin-oharra.

Azken urteotan Europako Batzordeak MLAP (Multilingual Action Plan) egitasmoaren barruan PAROLE proiektua[10] garatu du, idatzizko baliabide linguistikoak bildu nahian. Europako 14 hizkuntza jaso dira, bakoitzetik 20 milioiko corpusak bilduz, oinarrizko parametro hauen arabera diseinatuta: 1. liburuak; 2. egunkariak; 3. aldizkariak; 4. "miscellaneous" (korrespondentzia, elektronikoa, *ephemera*, eskuzkoa, makinakoa eta bestelakoak). EAGLESen oinarrituta osatu da, irizpide bateratuekin.

Txosten honetan aipatu ere egingo ez ditugun hainbat erreferentzia-corpus dago hizkuntza askotan[11].

Erreferentzia-corpusaren ezaugarrietako bat handia izatea dela dio EAGLESek, eta adibideek ere hala erakusten dute. Hala ere, horrekin batera ona izatea ere aipatzen du, kalitatea behar da.

Kalitateari ez bezala, mugarik jarri behar al zaio tamainari? Aipatu izan da puntu batetik aurrera erakusgarritasunaren proportzioa ez dela asko aldatzen: hitz gutxi batzuk ehunekoaren gehiena hartzen dute eta besteak maiztasun urrikoak dira (askotan hitz elkartuak). Hala ere, maiztasun handienekoetan ere, adieren araberako maiztasun-kurbak ageri dira. Hitz bat oso arrunta izan daiteke, baina ez adiera batean (eta hori corpusak ez jasotzea gerta daiteke, non eta ez den gaiari buruzko informazio zehatza jasotzen). Gauza bera gertatzen da hitzen konbinazioan ere. Beraz, honek justifikatzen du, neurri batean behintzat, corpusaren tamaina handia.

Baina edukia da irizpide nagusia, emaitzak ere horren araberakoak izango baitira. Dibertsitatea hartu behar da kontuan: generoa, dialektoak, diskurtso-mailak... Eta ez lexikoari begira bakarrik, baita gramatikari eta beste arloei ere.

Praktikan, baina, arazoak sortzen dira tamaina edo edukia osatzean, batez ere diruak eta epeek baldintzatzen dituztelako halako proiektu zabalak. Honen aurrean bi jarrera nabarmentzen dira: *oportunista* bezala ezagutzen dena, erraz eta azkar eskura daitekeen guztia jasoaz; ahozkoak eskuratzean areagotu egiten da korronte hau, hauek eskuratzea lan gaitza baita. Eta b) *printzipiozkoa* bezala ezagutzen dena, testu egokiei lehentasuna emanez. Bigarren honen erakusgarri nagusia *Brown* corpusa da. Eguneroko lanean, hala ere, bi korronteok nahasian erabiltzen dira.

Erreferentzia-corpusek aldaketa nagusia ekarri dute: azpicorpusen garrantzia dela-eta, etorkizunean, corpus bakarra beharrean, corpusak izango dira nagusi, pluralean. Eta, honen arabera, erabiltzaile-kopuruak ere gora egingo du, aldian-aldian behar duten azpicorpusa kontsultatu ahal izango du-eta.

3. EUSKARAREN ERREFERENTZIA-CORPUSAREN BEHARRA

Puntu honetara iritsita, ez dugu uste euskarak ere bere erreferentzia-corpusa behar duela zalantzan jar daitekeenik, XXI. mendeko gure hizkuntza lantzen, aztertzen eta hobetzen jarraitu nahi badugu behintzat.

UZEIko lexikografia sailean, Euskaltzaindiko Hiztegi Batuko Lantaldearen prestalana egiten dugunez, alegia, formen erabilerak corpusetan (OEH eta EEBS) eta bestelako iturrietan dokumentatu, corpus handiagoaren eta erakusgarriagoaren pre-

10 www.icp.inpg.fr/ELRA/cata/parole.html
11 www.ruf.rice.edu/~barlow/corpus.html

mia sentitzen dugu eguneroko lanean, are gehiago tradizio urriko formei buruzko txostenak prestatzean. Eta kezka bera azaltzen du Hiztegi Batuko Lantaldeak berak ere, forma bat proposatzeko nahiko daturik ez duenean.

Euskaltzaindiak iaz antolatutako Hiztegigintza Jardunaldian ere ikusi zen gizartearen esparru desberdinetako lexikoa bereiz landu beharra, eta horrek azpicorpusak eskatzen ditu: eguneroko hitz erabilienetakoa, administraziokoa, hezkuntzakoa eta ahozkoa behintzat azaldu ziren Bilboko jardunaldian. Hauek guztiak erreferentzia-corpusaren azpicorpus bezala bakarrik uler daitezke gure ustez, corpus oso eta orekatu baten barruan, alegia.

Beharra, beraz, argia da. Gaur, eta gure hizkuntzaren egoeran, ez dira lexikografoak bakarrik kontuan hartu behar, hiztegi batuak oinarrizko 40.000 formak laster izango baititu, baina eguneroko bizitzarako hori baino gehiago beharko dugu: hiztegi berezituak, terminologia, alegia, egunero sortzen eta osatzen ari da, gramatikako erabilera berriak ere ageri dira, ahozkoa eskura izatea komeni da. Erreferentzia-corpusak guztiei erantzun behar die, eta guztiok izan behar dugu eskura modu erosoan eta azkarrean.

Dudarik gabe, honen ardura Euskaltzaindiak izan behar luke, berari dagokio irizpideak zehaztea, sartu beharreko material erreferentea aukeratzea, kalitatea bermatzea, alegia.

Une egokian gaudela uste dugu gainera: mendea bukatu berri da eta EEBS corpus itxi izatera pasako da. Beraz, XXI. mendeko euskal produkzioa eskuratzen hasi beharra dago; baina ez eskuratzen bakarrik, baizik eta aukeratzen, kodetzen, etiketatzen eta modu erosoan eskaintzen. Gainera, ezin ahantz dezakegu Euskaltzaindiak egunen batean hiztegi arau-emaileari heldu beharko diola, eta hori pentsaezina da atzean erreferentzia-corpus modernorik gabe, OEH eta EEBSrekin osatuko dena.

Irizpideak zehaztean sailkapena izan beharko da kontuan, orokorra baina lagungarria, ahozko materiala zer eta nola bildu, corpusaren tamaina (bai abiapuntukoa, bai gerokoa ere) eta, batez ere, azpicorpusak zehaztu beharko dira. Hala ere, oreka mantenduz, edozein unetan gehi daitezke azpicorpus bereziak, irekia baita.

Egituraren aldetik, datu-base multifuntzionalaren beharra izango da: erabilera askotarikoa eta erraza. Tresna berrerabilgarriak behar dira, malguak eta elkartrukerako egokiak. Honek baliabideen bateratzea ekarriko du ezinbestean, elkarlana bideratuz: lexikografoak, gramatikariak, arlo desberdinetako adituak, informatikariak, hizkuntzalari konputazionalak...

Azkenean, Akademiak bere hiztegia behar duen bezala, corpusa ere ezinbesteko du, *euskararen erreferentzia-corpusa* izango dena.

Diru- eta giza baliabideak bideratzea ez da erraza izango, lantalde handia behar baitu atzetik, baina, euskarak aurrera egingo badu, behar-beharrezko du halako oinarri sendoa osatzea.

ERREFERENTZIAK

ABEILLÉ, A., CLÉMENT, L., KINYON, A. (2000). "Building a treebank for French", in: *Proceedings of the Second International Conference on Language Resources and Evaluation,* Atenas.

GELLERSTAM, M., CEDERHOLM, Y., RASMARK, T. (2000). "The Bank of Swedish", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

HATZIGEORGIU, N., GAVRILIDOU, M., PIPERIDIS, S., CARAYANNIS, G., PAPAKOSTOPOULOU, A., SPILITOPOULOU, A., VACALOPOULOU, A., LABROPOULOU, P., MANTZARI, E., PAPAGEORGIOU, H., DEMIROS, I. (2000). "Dessing and implementation of the online ILSP Greek Corpus", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

IDE, N., BONHOMME, P., ROMARY, L. (2000). "XCES: An XML-based Encoding Standard for Linguistic Corpora": in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

JOHANNESSEN, J.B., NOKLESTAD, A., HAGEN, K. (2000). "A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

MACLEOD, C., IDE, N., GRISHMAN, R. (2000). "The American National Corpus: A Standardized Resource for American English", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

MADROÑAL, A. (1998). "Corpus diacrónico del español (CORDE)", in: *Seminario de Industrias de la Lengua.* Soria. Fundación Duques de Soria.

MARTÍN MUNICIO, A., ROJO, G., SÁNCHEZ LEÓN, F, PINILLOS, O. (2000). "Language Resources Development at the Spanish Royal Academy", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

ORASAN, C., KRISHNAMURTHY, R. (2000). "An Open Architecture for the Construction and Administration of Corpora", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

RUNDELL, M. (1996). "The corpus of the future, and the future of the corpus", in www.ruf.rice.edu/~barlow/futcrp.html

SANTOS, D., BICK, E. (2000). "Providing Internet Access to Portuguese Corpora: the AC/DC Project", in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Atenas.

SINCLAIR, J. (1998). "Standards for Textual Representation and Integrity", in: *Seminario de Industrias de la Lengua.* Soria. Fundación Duques de Soria.

SOLER, J. (1998). "Los corpus textuales en lengua catalana", in: *Seminario de Industrias de la Lengua.*Soria. Fundación Duques de Soria.

_____ SOLER, J. (1998)."Written Linguistic Resources in Catalan: the DCC Project", in: *First International Conference on Language Resources and Evaluation* (Workshop on Language Resources for European Minority Languages), Granada.

URKIA, M. (1998). "Los corpus textuales en lengua vasca", in: *Seminario de Industrias de la Lengua.* Soria. Fundación Duques de Soria.