XXI. mendeko euskararen erreferentzia-corpusaz

Miriam Urkia

UZEI

1. SARRERA

Hizkuntzen industrian dihardutenak aspaldi ohartuak dira hizkuntzaren tratamendu automatikorako dagoen tresnen beharraz, oinarrizko materialen beharraz, alegia. Horietako bat da corpusa, azken urteetan bazterreko izatetik hizkuntzaren ikerketan oinarrizko tresna izatera pasa dena, baita gure artean ere.

Euskaltzaindiak aspaldi egin zuen corpusaren aldeko apustua: tradizioa jasotzen duen *Orotariko Euskal Hiztegia*ren corpusa osatu zuen batetik, eta *XX. mendeko euskararen corpus estatistikoa* osatzeko agindua eman zion UZEIri, bestetik. 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.

Corpusen historia ez da oso luzea; bai, ordea, emankorra. 1963an, lehen corpusa —*Brown*[1] corpusa, hain zuzen— kaleratu zenetik, asko aldatu da hauei 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.

1.1. Zer dira corpusak

Definizio zabalena hartuta, corpusa *hizkuntzaren ikerketarako oinarrizko tresna* dela esan dezakegu, edo, EAGLESen[2] definizioa baliatuz, bera baita corpusen osaerarako

[1] www.hit.uib.no/icame/brown/bcm.html

[2] EAGLES (Expert Advisory Group on Language Engineering Standards): http://www.ilc.pi.cnr.it/EAGLES96/

irizpideak zehazten dituena, estandartzat hartzen dena, ***"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"*** (John Sinclair, 1996). Alegia, testu-bilduma da: testu-masa handia, datu-base egoki batean antolatua, egituratua eta hizkuntzaren erakusgarri gisa erabiliko dena, benetako erabilerak bilduz. Zehatzago esanda:

1. Lagin erakusgarria (aztergai dugun hizkuntza-atalaren zati orekatu, zabal eta adierazgarria)
2. Mugatua (adierazgarria izan dadin, proportziotan orekatua eta mugatua)[3]
3. Informatizatua (bilaketa azkarrak eta orotarikoak bideratzeko)
4. Erreferentzia estandarra

McEneryk eta Wilsonek (2001) hala definitzen dute hizkuntza-corpusa: "*a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration*".

Hori guztia hizkuntzari buruzko azterketak egiteko eta hipotesiak frogatzeko erabiltzen da, besteak beste.

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. Datu enpirikoek garrantzia duten neurrian balio du corpusak ere. Azkenean, aukeratu den hizkuntzaren atalaren —hizkuntza bere osotasunean biltzea ezinezkoa baita— adierazle maximoa izan nahi duen testu-multzoa da, kontu handiz aukeratutako lagin erakusgarria. Informazio-bolumen handia biltzen duenez, eta sailkatzeko erraztasunak ematen, proba-banku erosoa da: teoriak edo/eta intuizioak frogatzeko erabiltzen da. Gainera, metodo estatistikoak erabiltzeko aukera ere ematen du.

Argi dago corpusek hizkuntzaren alor gehienetan dutela zerikusia: ahotsaren tratamenduan, lexikografian[4], gramatikan, semantikan, pragmatikan eta diskurtsoaren analisian, soziolinguistikan, estilistikan, hizkuntzen eta hizkuntzalaritzaren irakaskuntzan, hizkuntzalaritza historikoan, dialektologian, psikolinguistikan, azterketa kulturaletan eta psikologia sozialean, besteren artean.

Horrela, gauza izango da adierak, kategoria sintaktikoak, klase semantikoak, egitura argumentala, hitzen arteko kookurrentziak, unitateen agerpen-maiztasuna, unitateak bere testuinguruan, analisi probabilistikoa, erlazio lexikoak, 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. Eta, azkenean, abantaila horiek guztiek laginketa eta kuantifikazioa bideratzen lagunduko dute, eskuragai egotean erabilerraza izango baita eta datu aberastuak zein naturalak baliatu ahal izango baitira.

Modu askotakoak izan daitezke corpusak, betiere helburua kontuan hartzen bada: testualak edo ahozkoak, edo bietarikoak, gaur egin ohi den bezala. Baina, horren barruan, erreferentzia-corpusak, monitoreak, estatistikoak, paraleloak, konparatuak, bereziak, esperimentalak, literarioak edo bestelakoak izan daitezke.

---

[3] Besterik gabe "testu-bilduma" denean, etengabe testuak gehitzen direnean, ereduak proportzionalki errespetatu gabe, corpus monitore edo "monitor corpus" gisa izendatu ohi dira.

[4] Gogora dezagun corpus erraldoi nagusiak, *BNC*, *BoE* edo *FRANTEXT* modukoak, hiztegi handien oinarri gisa sortu direla. Hurrenez hurren, *BNC Longman*, *Larousse* eta *Oxford* hiztegien oinarri da, *COBUILD BoE*-n oinarritua da eta, azkenik, *Trésor de la Langue Française*-k *FRANTEXT* du atzean.

Eta, gehiago finduz, hauen barruko sailkapenak ere egin ohi dira: hedadura, aldaera historikoa, geografikoa, dialektala... izan daiteke muga-irizpide.
Nolanahi ere, egonkorrak izan behar dute, barne-osaera *orekatua* izatea ezinbestekoa da. Oreka mantentzen bada, irekiak ala itxiak izatea ez da oztopo izango, alegia, epe bat landu eta ez eguneratu (CTILCen[5] kasuan bezala) edo eguneratu eta osatu egiten direnak, betiere diseinuaren aldetik egonkortasuna mantenduz (CREA[6] , esaterako).

Bestalde, corpusean biltzen den informazioa kodetu eta markatu egin behar da, corpusa erabilgarri izateko garrantzitsua baita bilaketa-sistema erraza izatea. Baina ez tresna informatikoen aldetik soilik, baizik eta ezagutza linguistikoaren aldetik. Adibidez, deklinabideak, aditz-jokoak, aldaerek... testu-hitz desberdin asko sortzen dituzte. Corpusa kontsultatu nahi duenak modu erraza izan beharko du hitz zehatz bat berehala kontsultatzeko, zenbat forma desberdinetan ager daitekeen jakin gabe (inoiz gutxitan jakin edo gogoratuko ditu aldaera guztiak gainera): lematizazioa beharrezkoa da, helduleku ezinbestekoa da kontsultak azkartzeko. Baina, lemen barruan ere, homografoekin egin dezakegu topo: kasu horietan, gutxienez kategoria gramatikalak izatea komeniko zaigu. Erabiltzaileak datuok eskura dituen neurrian jarraituko du corpusaren beharra sentitzen, lana erraztuko dio-eta.

Egun, kodeketa bideratzeko *TEIk* (*Text Encoding Initiative*) proposatzen duen kodetze-sistema da estandartzat hartu ohi dena, eta, bereziki, dokumentuak markatzeko *TEIk* baliatzen duen *SGML (Standard Generalized Markup Language)*, edo are *XML* (*eXtended Markup Language)* eta *CES* (*Corpus Encoding Standards*), gero azalduko dugunez.

Esan dugunez, lehen corpusa 1960ko hamarkadan kaleratu zen, *Brown corpus* 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. Garai hartan asko zen, baina berehala ohartu ziren mugatuegia zela eta, 1980ko 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. 1990eko hamarkadan 320 milioi zituen honek berak, *COBUILD-Bank of English*[7] (BoE) izena hartuta. *British National Corpus*-ek[8] (BNC) ere, 1994an, 100 milioi zituen jasoak, berrikuntza nagusi batekin: testu idatziak eta ahozkoak biltzen zituen lehena zen. Azken hau da, hain zuzen, erreferentzia-corpusen eredu.

Ingeleseko adibideak dira hauek, baina beste hizkuntzetan ere berehala hasi ziren corpusak osatzen. Adibidez, *The Bank of Swedish*-ek 75 milioi testu-hitzeko corpusa du, CREAk 133koa eta CTILCek 52.3koa. Hala ere, badira txikiagoak eta, aldi berean, zehatzago etiketatuak daudenak.

1.2. Erreferentzia-corpusak

Corpusak era askotakoak izan daitezkeela ikusita, osoenak eta hizkuntzaren erakusgarrienak erreferentzia-corpusak direla ohartzen gara. Hala definitzen du EAGLESek (1996): ***"A reference corpus is one that is designed to provide comprehensive information about a language"***; alegia, hizkuntza, bere

[5] CTILC (*Corpus Textual Informatitzat de la Llengua Catalana*), Institut d'Estudis Catalans-ek egindakoa: 1832-1988 epea jasotzen du eta egungo katalaren hiztegi deskriptiboaren (DCC: *Diccionari del Català Contemporani*) oinarri da: http://pdl.iec.es.

[6] CREA (Corpus de Referencia del Español Actual), Real Academia Española: azken 25 urteetako produkzioa biltzen du. 1975eko testuekin hasi ziren eta, 25 urteak gainditu ostean, CORDEra (CORpus Diacrónico del Español) pasatzen dira lehen urteetakoak. Horrela, beti dituzte azken 25 urteak. (www.rae.es).

[7] http://titania.cobuild.collins.co.uk/boe_info.html

[8] www.info.ox.ac.uk/bnc/

osotasunean hartuta, erakusteko diseinatua egon behar du corpusak. Ondorioz, hizkuntzaren aldaera esanguratsuak adierazteko besteko tamaina eta kalitatea behar du.

Erreferentzia-corpusa elektronikoki kodetutako testu-multzoa da, informazioa berreskuratzeko ondo prestatua egongo dena: antolatua, kodetua, etiketatua, lematizatua, analizatua. Eta, informazioa bera hizkuntzaren erakusgarri zabala izango da. *Erakusgarria*, 'zer islatu nahi da?' galderari erantzuten dion neurrian, eta *baliagarria*, orekatua bada. Alegia, irizpideok bete beharko ditu: ahalik eta handiena izan, lagin desberdin asko bildu erakusgarri izan dadin, erdi mailako sailkapena, iturriak azaldu eta testu idatziak zein ahozko transkripzioak jaso. Bost irizpideotan oinarritzen dira fidagarri eta erakusgarri izan nahi duten corpusak. Gainera, *de facto* estandar bihurtzen direla erakutsi dute orain artekoek. Izan ere, azken urteotan gorakada handia izan da, eskanerrak eta euskarri elektronikoan eskuratzeko erraztasunak lagunduta. Horrez gain, baliabide informatikoek lematizazioan eta etiketatzean laguntza handiak eskaintzen dituzte.
Diseinuari dagokionez, EAGLESek oinarrizko zazpi irizpide markatzen ditu:

1. *Funtzioa*: Zein helbururekin sortu? Zertarako?
2. *Adierazgarritasuna*
3. *Lagina*: ahalik eta erakusgarriena
4. *Tamaina*: estatistikoki egokiak diren emaitzak emateko besteko tamaina behar du
5. *Hedadura*
6. *Eskuragarritasuna*
7. *Berrerabilgarritasuna*

Zazpi irizpide hauek baldintzatuko dute corpusaren egokitasuna edo egokitasun-eza.

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.

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

Kalitateari ez bezala, mugarik jarri behar al zaio tamainari? MacMullenek (2002) dioenez, "*as large as necessary, but as small as possible*" izan beharko luke corpusak. 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[9]—. 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.

[9] Real Academia Española-ko CREA corpusean, adibidez, 119 milioi hitzeko corpus-zatia kontuan izanda, 50.000 hitz desberdinek corpus osoaren % 96.33 hartzen dute. Are gehiago, maiztasun handieneko 10 hitzak corpus osoaren % 28.55 dira. Lemei erreparatuz, bestalde, milioi bateko corpus lematizatu bat hartuta (23.000 lema desberdin dituena, bestalde), maiztasun handieneko 69 lemek testuen erdia hartzen dute. Edo, bestela esanda, 40 milioiko corpusean, 70 lemak % 50 hartzen dute (Guillermo Rojo jaunak helarazitako datuak dira hauek). Euskararen kasuan ere datuak argigarriak dira: *XX. mendeko euskararen corpus estatistikoan*, maiztasun handieneko 10 lemek corpusaren % 14 hartzen dute, eta 20 lehenek % 18. Bestalde, agerraldi bakarreko lemak lema desberdinen % 52,6 dira. Laburbilduz, corpuseko forma "errealak" kontuan hartuta, maiztasun handieneko 100 lemak corpus osoaren % 38,4 dira.

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. Are gehiago, egituratzean ere "tipularen printzipioa" aplikatzen ari dira hainbat corpus, alegia, azpicorpus bat erabat etiketatu, lematizatu eta desanbiguatu, eta bestea modu "oportunistagoan" erabili, maila apalagoan landua.

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. Baina, aurrera egin aurretik, ikus ditzagun behar hori erakusten duten hainbat adibide.

1.3. Euskararako erreferentzia-corpusaren beharra

UZEIko lexikografia sailean, Euskaltzaindiko Hiztegi Batuko Lantaldearen prestalana egiten dugunez, hau da, formen erabilerak corpusetan (*OEH* eta *XX. mendeko euskararen corpus estatistikoa*) eta bestelako iturrietan dokumentatu, corpus handiagoaren eta erakusgarriagoaren premia 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. Kezka hau arazo larri izango da Euskaltzaindiak bere *Hiztegi Hiritar Arauemailea* (Euskera, 1986) bideratzen duenean, gaurko euskararen berri ere eman beharko baitu eta, horretarako, egungo erabilerak ere eskura beharko baititu.

Euskaltzaindiak 2000. urtean antolatutako Hiztegigintza Jardunaldian 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. Azpitalde bakoitzak bere irizpide propioak baliatzen baditu, hiztegi osora biltzean desoreka nabarmena sortuko delakoan gaude.

Bada, gainera, Euskaltzaindiak berak 2001. urteko XV. Biltzarrean, Bilbon, ondorioetan jasotako adierazpena: "*(...) Bestalde, hiztegigintzarekin eta Euskaltzaindiaren beste esparruekin ere zerikusi garbia dute Corpusek. Orain artekoak baliotsuak izan arren mugatuegiak dira etorkizuneko lanetarako.* ***Horregatik, gaurko euskara jasoko duen Corpus handia biltzeari ekin behar zaio****. Euskaltzaindiak lan horiek gidatu eta lagundu egin behar dituela uste du. Ahozko euskararen Corpusa ere bideratu behar duela argi dauka*"[10].

[10] Markatua gurea da, ez Euskaltzaindiarena.

Eusko Jaurlaritzako Hizkuntza Politikarako Sailordetzak bideratuta, *Euskara Biziberritzeko Plan Nagusiak* (*EBPN*), VI.3.2.6 atalean hala dio: "*(...) hizkuntza analisirako oinarrizko tresnen azpiegitura sortu. Azpiegitura honek bilduko lituzke datu-base lexikala, corpus-bilduma elebakarra eta eleanitza, analisi eta sintesi morfologiko eta sintaktikoa, analisi semantikorako oinarriak*" (1999: 55. or.).

Bukatzeko, azken datu gisa, UZEIk 2002. urteko urrian antolatutako *Hizkuntza-corpusak. Oraina eta geroa* jardunaldiak aipatu nahiko genituzke. Parte-hartzaileen ondorioa garbia izan zen: euskarak erreferentzia-corpusaren beharra du, beste hizkuntzek bezala. Jarduera desberdinetako adituak bertaratu ziren eta elkarlanerako borondatea ere erakutsi zuten.

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.

Une egokian gaudela uste dugu gainera: XX. mendea bukatu berri da eta, horrekin batera, mendea jasotzen duen corpus estatistikoa. Beraz, XXI.[11] mendeko euskal produkzioa eskuratzen hasi beharra dago; baina ez eskuratzen bakarrik, baizik eta aukeratzen, kodetzen, etiketatzen eta modu erosoan eskaintzen. Alegia, **kalitatezko corpus eskuragarri baten beharra** dugu. Azkenean, Akademiak bere hiztegia behar duen bezala, corpusa ere ezinbesteko du, *euskararen erreferentzia-corpusa* izango dena. Euskarak aurrera egingo badu, behar-beharrezko du halako oinarri sendoa osatzea.

2. AURREKARIAK: euskaraz dauden corpusak

Euskarak, Euskaltzaindiaren ekimenari esker, badu bere corpus-tradizioa, idatzia, mugatua bada ere. Tradizioa biltzen duen *Orotariko Euskal Hiztegia*ren corpusa du batetik, eta *XX. mendeko euskararen corpus* estatistikoa[12] bestetik:

Orotariko Euskal Hiztegia (OEH) corpus diakroniko itxia dela esan dezakegu, 1970. hamarkada arte iristen baita. 310 obra oso (edo ia oso) "aukeratu" biltzen ditu, 5.800.000 hitzek osatua; testu gordina da (akats eta guzti, bilaketa zailtzen dutenak), kodetu gabea eta lematizatu gabe dago. Hala ere, euskararen historiaren altxor ezinbestekoa da. Sailkapen orokorra du: epea, euskalkia eta testu-mota zabala. Corpusa ez dago erabiltzaile-multzo zabalaren esku; beraz, haientzat ez da "existitzen": zenbait ikertzailek bakarrik balia dezakete CD-ROMeko kopia.

XX. mendeko euskararen corpus estatistikoa, bestalde, XX. mendeko euskara jasotzen duen corpus estatistiko itxia[13] da, 1900-1999 urteetan argitaratutako 6.352 obra-zatitatik jasotako 4.657.165 testu-hitz dituena, SGML formatu estandarrean kodetua, lematizatua (104.817 lema desberdin), sailkapenaren arabera erakusgarria (epea, euskalkia, testu-mota eta obraren tamaina) eta

[11] XXI. mendea diogunean ez da XX. mendea baztertzen, alegia, ez da 2001. urtarrilaren 1etik aurrerakoa bakarrik kontuan hartzen. XXI. mendean erabiliko dugun corpusa esan nahi da, XX. mendearen azken urteak ere, aurrerago ikusiko dugunez, jasotzen dituena.

[12] Corpus hau, duela gutxi arte EEBS (Egungo Euskararen Bilketa-lan Sistematikoa) izendatzen zena, UZEIk Euskaltzaindiaren eskariz eta harentzat egina da.

[13] Corpus itxi izatera pasa da 2002. urtean, 1987an martxan jarri zenetik, irizpideak eta tamaina proportzionalki zainduz, urtero eguneratu baita mendea bukatu arte. Izan ere, euskal argitalpenen inbentarioan oinarrituz, unibertso osoa proportzionalki ordezkatzen du zozketa bidez aukeratutako laginak.

Internet bidez kontsultagai dagoena, www.euskaracorpusa.net helbidean. Corpusak izan duen harrera onak eta kontsulta-kopuruak[14] erakusten dute erabiltzaileek corpusaren *beharra* zutela.

Gaur egun badira beste corpus idatzi batzuk ere, baina helburu desberdinekin osatuak: itzulpen-memoriak, EIZIEk eskaintzen duena, baina ez du erreferentzia-corpusekin zerikusirik, elebiduna baita ezinbestean. Bestalde, Euskaltzaindia bera biltzen ari den *Lamiategi* proiektua aipa daiteke, testu-corpus oportunista baten oinarri gisa aurkezten dutena, baina ez dago horren erabilera publikorik. Txosten honetan ez dira aipatuko, besteak beste, eskura ez dauden corpusak ez baitira "existitzen" guretzat, ez baititugu ezagutzen.

Eskura ditugun bi corpusetara mugatuta, beraz, Euskaltzaindiaren lan-emaitzetan duten eta izan duten eragina ukaezina dela esan dezakegu. Haren "produktu" nagusien artean *Orotariko Euskal Hiztegia* bera, *Euskal Gramatika. Lehen Urratsak* (EGLU liburukiak) eta *Hiztegi Batua* daude. Gaur egun, eskura izango al genituzke euskararen normalizazioan oinarrizko osagai diren hiru emaitzok corpusik gabe? Erantzuna argia dela uste dugu: ez. Bestalde, ikertzaileek eta hiztegigileek ere maiz jotzen dute corpusetara, haien lana dokumentatu beharra ezinbestekoa baita.

3. PROPOSAMENA: nolakoa behar du XXI. mendeko euskararen erreferentzia-corpusak?

3.1. Diseinua

Euskara modernoaren erakusgarri ahalik eta zabalena izan behar du XXI. mendeko corpusak. Hori adierazten saiatuko gara ondoko orrietan.

Oinarrizko baldintza corpusa ***orekatua*** izatea da. Hortik abiatuko da beste guztia. Eta, orekatua diogunean, corpusaren atal bakoitza bere garrantziaren arabera ordezkatua egotea esan nahi da. Nola neurtu, baina, "garrantzia"? Sortzen den material-kopurua kontuan izan beharko da, baina baita zenbat sortzaile eta zenbat hartzaile dituen ere. Oreka hori diseinuaren atal guztietan izan beharko dugu gogoan.

Erreferentzia-corpusak osatzeko estandarrek markatzen dituzten oinarrizko ezaugarriak hartuko ditugu abiapuntu gisa gure proposamena egiteko:

1. elebakarra
2. sinkronikoa
3. orokorra (ereduak, aldaerak, idatzia eta ahozkoa)
4. lagina (baina erakusgarria)

[14] 2002. urtean, otsailean jarri zen kontsultagai eta lehen hilabeteetan, batez beste, eguneko 37 kontsulta izan zen. Ekaina-iraila arteko daturik ez dugu (Euskaltelen arazo bat dela-eta), baina urria-azaroan 53ko media da eguneroko kontsultena. Kontsultagileak ez dira Euskal Herrikoak bakarrik, ez behintzat hemen bizi direnak. Ameriketako Estatu Batuetatik hainbat kontsulta egiten dira, Washingtondik (eta, zehatzago, Seattletik) eta Californiatik (San José) batez ere, nahiz besteetatik ere hainbat sarrera dauden. Mexiko, Argentina eta Txile ere bisitarien artean ageri dira. Europa mailan, ia herrialde guztietatik (Britainia Handia, Suedia, Polonia (Lodz), Herbehereak, Alemania, Italia,...), Kanadatik. Bestalde, bakanago bada ere, Australia eta Japongo kontsultak detektatu dira, bai eta Afrikatik ere, Nigeriatik, hain zuzen. Dena den, hainbat bisitaren jatorria ezezaguna da eta ez dago horiei buruzko datuak ematerik.
Orain arteko datuok erakusten dute Euskal Herritik kanpo ere kontsultatzen duenik badela, eta diasporan dauden euskaldunak aipatu behar dira gainera: EEBBetan hainbat euskal ikertzaile dago, Hego Amerikako euskaldunak ere sartzen dira web gunean, eta Europan zehar dauden euskaldunak izango dira, seguruena, bisitari ugari horiek. Hori da, hain zuzen, interneti zor diogun gauzatako bat: corpusa kontsulta dezakete atzerrian diren euskaldunek ere.

eta, honen barruan, gehiago finduz:

5. handia: milioika testu-hitz
6. egungoa: azken urteak
7. sailkatua
8. idatzia + ahozkoa
9. egituratua[15]: kodetua; etiketatua; lematizatua

Datu argigarriak eskaintzen saiatuko gara, horretarako lau iturritatik edanez:

1. Corpusen osaerarako estandarrek (EAGLESek, nagusiki) markatzen dituzten ezaugarri nagusiak.
2. MLAPen[16] barruan PAROLE proiektuak landu eta proposatua.
3. *XX. mendeko euskararen corpus estatistikoa* osatzeko egindako euskal argitalpenen inbentarioaren azken epea (1991-1999: araugintza berriaren ondoko gisa izendatzen dena). Datuok pasa den mendearen azken bederatzi urteetako euskal produkzioaren berri eskaintzen dute eta erreferentzia gisa erabil daitezke.
4. Gaur egungo hainbat erreferentzia-corpusei buruzko datuak, haien nondik-norakoak lagungarri izan daitezke-eta.

Ikus ditzagun, bada, ezaugarriok banaka. Argi dezagun, halere, proposamen erabat zabala dela hau, hain zuzen, irizpideak zehaztea adituei baitagokie, ez guri. Proposamen-zertzelada batzuk besterik ez ditugu hona bilduko.

1. Elebakarra

Corpus-mota asko daude, txosten honen hasierako atalean azaldu dugunez, baina erreferentzia-corpusaren definizioan bertan dator elebakar izatea: "hizkuntza baten atal baten erreferentzia" izan nahi baitu, EAGLESek dioenez. Gure proposamena, beraz, euskara aztergai bakarra izatea da.

Honekin ez dugu esan nahi azpicorpus paraleloak —eleanitzak— sortzea, adibidez, ongi ikusten ez dugunik, baina hori beste kontu bat da, erreferentzia-corpusekin zerikusirik ez duena. Bestalde, euskararen barruan aukerak egin beharko dira: euskara batua eta euskalkiak. Baina arazo hau ondoren azalduko dugu.

2. Sinkronikoa

"Azken urteak" da honen definizioa, baina, nola ulertu dute corpus-sortzaileek hau?

PAROLE proiektuan, 1990. hamarkadan landuan, 1975ean jarri zuten abiapuntua, azken urteetako Europako hizkuntzak biltzea baitzuen helburu.

---

[15] Egituratze horretan ere mailak daude: a) gordina (ASCII hutsean, tipografia, orrialdeak eta halakoak soilik markatuz), b) etiketatua (kategoria sintaktikoa, semantikoa, desanbiguatua,...), c) parentizatua, treebank gisa, d) analizatua (sintaktikoki etiketatua).

[16] MLAP (MultiLingual Action Plan). Azken urteotan Europako Batzordeak MLAP egitasmoaren barruan PAROLE (www.icp.inpg.fr/ELRA/cata/doc/parole.html) proiektua garatu du, idatzizko baliabide linguistikoak bildu nahian. Europako 14 hizkuntza jaso dira, bakoitzetik 20 milioi hitzeko corpusak bilduz, oinarrizko parametro batzuen arabera diseinatua.

Egungo corpus gehienek ere hala jokatzen dute:

BNC (British National Corpus)	1975-1994[17]
CTILC (Corpus Textual Informatitzat de la Llengua Catalana)[18]	1832-1988
CREA (Corpus de Referencia del Español Actual)	1975etik hona[19]
CORGA (Corpus de Referencia do Galego Actual)	1975-2004[20]
CNC (Czech National Corpus)	"modern"[21]
HNK (Croatian National Corpus)	1990etik hona
HNC (Hellenic National Corpus)	1976etik hona
HNC (Hungarian National Corpus)	1995etik hona
FIDA (Corpus of Slovene Language)	XX. mendeko 2. erdia[22]
CNG (National Corpus of Irish)	daturik ez
CORIS (CORpus di Italiano Scritto)	1980. eta 1990. hamarkadak
CRPC (Corpus de Referência do Português Contemporâneo)	daturik ez
ANC (American National Corpus)[23]	1990etik hona

Azken 25 edo 10 urteak biltzen dituzte corpus gehienak, baina azalpena argia da: 1975 ingurua aukeratu duten gehienak garai hartan hasi ziren corpusa osatzen. Azken hamarkadan bideratutakoek hamar urtetik honakoa besterik ez dute jasotzen (HNK, HNC-hu eta ANC, esaterako). Arrazoia ere bistakoa da, eta hala aitortzen dute egileek: euskarri elektronikoan dagoena biltzen dute batez ere eta, ezinbestean, azken hamar urteetan jarri behar izan dute muga.

Euskararako proposamena ere bide horretatik egin daiteke eta 1990ean edo are 1991n jar daiteke corpus berriaren abiapuntua, baina ez eskuragarritasunaren irizpideagatik bakarrik. Izan ere, *XX. mendeko euskararen corpus estatistikoa*n lau epe bereizi ziren, azkena 1991-1999 zela: "araugintza berriaren ondokoa"[24], alegia. 1991n Ibon Sarasolaren *Hauta-Lanerako Euskal Hiztegia* kaleratu zen eta Euskaltzaindia lehen gomendioak kaleratzen hasi zen. Beraz, "egungo" euskararen abiapuntu gardenena —data bat ezarri behar bada behintzat— hori izan daiteke, 1968ko hura urruti samar geratu baita (eta hizkuntza bera asko aldatu da ordutik hona). Gainera, modu estatistikoan bada ere, *XX. mendeko euskararen corpus estatistikoa*k jasotzen du garai hura.

Honek ez du esan nahi, noski, 1991 baino lehenagoko obrek lekurik ez dutenik. Irizpide orokorra besterik ez da, malgutasunez hartu beharrekoa.

3. Orokorra

Erreferentzia-corpusa izatea "orokorra" izatea da, hau da, hizkuntza baten barruan dauden eredu desberdinak eta aldaerak jaso beharko ditu, guztiaren berri eman nahi badu behintzat. Murrizketak egin ohi dira, noski. Esaterako, ingelesaren

---

17 Hau ez da erabat zehatza. Izan ere, literaturako obra garrantzitsu batzuk (erreferentzia direnak, hain zuzen) sartu ahal izateko, literatura saila 1964tik honakoa da).

18 Katalanaren CTILC corpusa ez da erreferentzia-corpusa, ez gaur ulertzen dugun bezala behintzat. Alegia, 150 urte biltzen ditu eta duela 14 urte bukatua da, itxia, eta ez da eguneratzen. Baina hona ekarri dugu haren osaerarako erabili ziren irizpideak argigarriak izan daitezkeelako.

19 Irizpidea hau da: CREAn beti azken 25 urteak egongo dira. Egun 2000-2004 urte arteko obrak sartzen ari direnez, 1979 aurrekoak CORDEra (CORpus Diacrónico del Español) pasako dira.

20 Egun 2001 artekoa osatua dute eta 2004rako osorik izatea espero dute.

21 Berria da, "modern" bezala definitzen dute, baina ez dute datu zehatzagorik eskaintzen.

22 Modu orokorrean XX. mendeko bigarren erdia hartzen badute ere, egileek aitortzen dute 1990. hamarkadakoak direla baliatu dituzten obrarik gehienak.

23 ANC abian jarri berria da eta, oraingoz, oinarrizko lana besterik ez dute egina, baina *BNC*aren irizpide berak erabiltzen ari dira. Horrez gain, corpus landua osatuz, corpus monitore bat osatzeko asmoa ere badute, *COBUILD*en antzekoa.

24 Galegoaren corpusak, CORGAk ere, berez araugintza bideratua dutenetik honako obrak bildu nahi dituen arren, hau da, 1982tik honakoa, obrak biltzeko orduan 1975 urtea ezarri dute abiapuntu gisa, hainbat obra erreferente kanpo ez uzteko, besteak beste.

kasuan, britanikoa eta amerikarra bereizi egiten dira —BNC vs. ANC—, hizkuntzaren aldaerak oso markatuak direlako, besteak beste. Idatzia eta ahozkoa batu zein banatzea ere bada orokortasunean murriztapenak ezartzea, baina aukeratzen den unibertsoaren barruko orokortasuna da axola duena.

Euskararen kasuan, hiru atal hartu beharko ditugu kontuan:

1. **Hizkuntza-ereduak**: auzi honek lanak eman ditu azken urteotan euskalgintzan dihardutenen artean. Gure ustez, eredu horiek jaso beharko lirateke, guztiaren berri emanez. Euskaltzaindiko Jagon Sailak gai honen inguruan antolatu zituen Jardunaldiak (*Euskera*, 1996) kontuan izan beharko lirateke zer eta nola jaso erabakitzeko orduan.

2. **Euskara batua eta euskalkiak**: euskara batuarekin batera, euskalkien berri ere eman beharko litzateke, horrek guztiak osatzen baitu egungo euskara. Baina, zein proportziotan? *XX. mendeko euskararen corpus estatistikoa* osatzeko sortu zen datu-base bibliografikoko azken urteetako erabilerek hau erakusten digute[25]:

Euskalkia	*Unibertsoa dok. (testu-hitzak)*	*%*	*Corpusa dok. (testu-hitzak)*	*%*
Bizkaiera	1044 (6.895.000)	5	59 (67.544)	4.8
Gipuzkera	790 (2.871.000)	2.1	45 (41.647)	3
Zuberera	41 (214.500)	0.1	5 (2.975)	0.2
Lapurtera / Nafarrera	304 (1.292.500)	1	18 (17.044)	1.2
Euskara batua	18.738 (123.620.500)	91.3	996 (1.273.846)	90.4
'Nahasiak'[26]	66 (429.500)	0.3	6 (5.286)	0.7
GUZTIRA	**20.983 (135.323.000)**	**100**	**1129 (1.408.340)**	**100**

Datuok euskalkien erabilera urria erakusten dute, nahiz, esan dugun moduan, aldizkarietako datuak ez diren hemen jaso (eta horiek dira, hain zuzen, euskalkiei garrantzia ematen dietenak: herri-aldizkariak eta halakoak). Horrez gain, ezin dugu ahaztu unibertso honek XX. mendean euskaraz argitaratua besterik ez duela biltzen. Beraz, ahozkoa eta euskarri elektronikoan sortuaren daturik ez dugu. Bere muga guztiekin irakurri behar dira datuok.

Erreferentzia-corpusak euskalkiei zein leku eman erabaki beharko luke, lekua dutela ez baitago dudarik gure ustez. Erabaki horiek hartzeko, gero proposatuko dugun aholkulari-batzorde bat beharko litzateke.

Beste hizkuntzetan ere, aldaeren tratamenduak eragina du corpusaren osaeran. Ingelesaren kasuan, esaterako, modu desberdinak ikus ditzakegu. Batetik, muturreko bi aldaeraren berri ematearren, *BNC* eta *ANC* bereiz mantentzen dira, nahiz irizpide berak erabili, bata ingeles britanikoa baita (*BNC*) eta bestea amerikarra (*ANC*), gorago aipatu dugunez. Baina, *ICE*ren[27] kasuan, 20 *ICE* corpus daudela ere esan dezakegu, berez *ICE* bakarraren azpicorpusak badira ere: 20 herrialde —eta aldaera— desberdin jasotzen ditu bakoitzak, aldaera bakoitzetik milioi bat hitz bilduz. Helburua, noski, lekuan lekuko "ingelesak" aztertu eta erkatzea da.

---

25 Kontuan izan behar da banaketa nagusia obra sailkatuak *vs.* sailkatu gabeak dela eta, beraz, lehenengoak besterik ez ditugula jasoko: liburuak eta aldizkari nagusietako artikuluak. Egunkariak, astekariak eta bestelako aldizkariak masa oso bezala bilduta daude, euskalkiei erreferentziarik egin gabe. Hain zuzen ere, batez ere kazetaritza-lan gisa har ditzakegun horiek 1178 dokumentuk osatzen dituzte.

26 "Nahasiak" multzoan euskalki-marka berezirik ez duten dokumentuak bildu dira: euskalki bat baino gehiago nahasten dituzten elkarrizketak, bertsoak, etab.

27 ICE (International Corpus of English) http://www.ucl.ac.uk/english-usage/ice/. ICEk lantzen dituen hizkuntzen artean Ameriketako Estatu Batuak, Australia, Fiji, Ghana, Hong Kong, India, Irlanda, Kamerun, Kanada, Karibea, Zeelanda Berria, etab. aipa daitezke.

3. **Idatzia eta ahozkoa**: banaketa hau ez da ezinbesteko sailkapen orokor honetan (beherago finduko ditugu sailkapen-irizpideak), baina biak jasotzeak osoago egingo du corpusa.

4. Lagina

Hizkuntzak ezin dira corpus batean bere osotasunean jaso. Horregatik, lagin bat aukeratu behar da, baina erakusgarri izango dena: bai kalitatearen aldetik, bai kantitatearenetik ere. Edukiaren kalitateari ez zaio inongo mugarik ezarri behar, emaitzak ere horren araberakoak izango baitira. Bestalde, dibertsitatea hartu behar da kontuan: generoa, dialektoak, diskurtso-mailak... Eta ez lexikoari begira bakarrik, baita gramatikari eta beste arloei ere.

Gaur egun milioi askotako corpusak sortzea ez da arazoa, material gehiena elektronikoki eskura baitaiteke eta ondoko urratsak nahiko automatizatuak daude lehen hurbilpen batean behintzat.

PAROLE proiektuak Europako 14 hizkuntza jaso zituen, eta bakoitzak hogeina milioi hitz helarazi zituen proiektura.

Aipagai ditugun beste erreferentzia-corpusek ere kopuruok dituzte:

Corpusa	Tamaina (hitzak, milioitan)
BNC	100
CTILC	52,3
CREA	133
CORGA	17,5/25[28]
CNC	100[29]
HNK	30
HNC-gr	24
HNC-hu	100/150
FIDA	100
CNG	15
CORIS	100
CRPC	86
ANC	*100*

Egun, estandartzat 100 milioi hitz hartzen da, erreferentzia-corpusa osatzeko orduan.

Euskararako corpusaren tamaina proposatu aurretik euskal produkzioaren berri izatea ezinbesteko da, unibertsoarena, horren arabera erabaki beharko baita zer jaso ahal izango duen. UZEIko Lexikografia Sailak *XX. mendeko euskararen corpus estatistikoa* osatzeko euskal argitalpenen[30] inbentario sailkatua egin zuen, aipatua dugunez. Ondoko irudian ikus daiteke zenbateko unibertsotik abiatu zen eta zein neurritako corpusa bildu zen 1991-1999 epean, testu-hitzetan:

	Sailkatuak (liburuak eta artikuluak)	Sailkatu gabeak (egunkariak eta aldizkariak)	GUZTIRA
Unibertsoa	135.323.000	302.396.857	*437.719.857*
	% 1.04	% 0.14	% 0.42

[28] Egun 17.5 milioiko corpusa dute, baina 2004. urterako 25 milioitara iristea dute aurreikusia.

[29] 100 milioi hitzetik gora duela diote, baina etengabe ari omen dira eguneratzen. Orekatua eta erakusgarria dela diotenez, irizpideak errespetatuko dituztela pentsatu behar da.

[30] Lehen edizioa besterik ez da kontuan hartzen.

Lagina	1.408.340	421.190	*1.829.530*

XX. mendeko azken bederatzi urteetan (1991-1999), beraz, *euskal produkzio idatzia* ia 438.000.000 hitzekoa izan da: testu-masa handia, nolanahi ere. Eta produkzioa mantendu egin da urte horietan gainera: aurreko urte batzuetako gorakada izugarriaren ondoren, egonkortu egin dela dirudi. Hala ere, ez dugu ahaztu behar euskal produkzio idatzi *argitaratua* dela hemen bildu dena, alegia, elektronikoki sortutako materiala ez da kontuan hartzen, ez eta ahozkoa ere. Datu hauek, beraz, ez dute bere osotasunean balio, baina argigarri izan daitezke duten mugekin bada ere. Corpus osoa eratzeko idatzi argitaratua, elektronikoa eta ahozkoa hartu beharko da kontuan.

Horiek horrela, adierazgarritasun-maila tartekoa behintzat lortu ahal izateko, euskal argitalpenak ikusiz batetik —goian aipatutako mugekin—, eta beste hizkuntzetakoak bestetik, euskararen erreferentzia-corpusaren tamaina 25.000.000 / 50.000.000tara ekartzea proposa dezakegu, gaurko estandarrak 100.000.000tan ezartzera badoaz ere, gaur-gaurkoz egingarria —erreala— eta nahiko erakusgarria dela pentsa baitezakegu. Bederatzi urtetako produkzio idatzia 400 milioitik gorakoa bada, ez dirudi oso egingarria 100 milioi proposatzea, hori unibertso (idatzi argitaratu) osoaren laurdena izango bailitzateke. Corpus "oportunista" edo monitore bat sortzeko arazorik ez legoke kopuru hauek kontuan hartzea, baina erreferentzia-corpusa osatzeko egokiago ikusten dugu maila apalago batetik abiatzea, nahiz beti izango den osatzeko aukera, corpus irekia nahi bada behintzat. Hori bai, orekatua beharko du.

Horregatik, hasierako fase batean 25 milioi bilduz, maiztasuna eta prestasuna zenbateraino erakusten duten azter daiteke eta, urte batzuen buruan eguneratu (bestela ere, euskal produkzioak aurrera egiten duen neurrian, osatu beharko da).

Hala ere, liburu vs. egunkari/aldizkari banaketaren ehunekoak zehaztean, *erakusgarritasuna* kontuan hartzekoa izango da. Bai eta ahozkoari zenbat esleituko zaion ere.

Honaino estandarrek markatzen dituzten lau irizpide nagusiek eman digutena. Baina, hasieran aipatu dugun moduan, gehiago finduz, beste irizpide batzuk ere bazeudela azpimarratu dugu: handia, egungoa, sailkatua, idatzia + ahozkoa eta egituratua. Horietako batzuk argitu ditugu irizpide orokorretan: handia eta egungoa; baina besteak gainetik besterik ez ditugu aipatu. Ikus ditzagun banaka:

5. Handia

Esan dugunez, erakusgarri izan behar badu, ondorioz, handia izango da. 25 milioi hitzekin hastea da proposatu duguna.

6. Egungoa

Irizpide hau ere aipatua dugu: gaurko euskararen erreferente izan behar badu, azken urteak hartuko dira kontuan: 1991tik honakoa da proposatu duguna.

7. Sailkatua

Corpus orokorra izateak erakusgarri izatea ere badakar berarekin, baina erakusgarri izateko nolabait sailkatu beharko dugu, unibertso osoaren berri eman beharko du-eta. Hasteko, idatzi/ahozko banaketa egin beharko genuke, baina hori hurrengo atalean ikusiko dugu.

EAGLESek sailkapen orokorra proposatzen du:

1. **Ahozkoa**
2. **Idatzia**
   a. Argitaratua
      i. Liburuak
         1. Ez-fikzioa
         2. Fikzioa
      ii. Egunkariak
         1. Berri orokorrak
         2. Enpresa/ekonomia
         3. Kirolak
         4. Arteak
         5. Bitxikeriak
      iii. Aldizkariak
      iv. Argitalpen iragankorrak
      v. Korrespondentzia
   b. Makinaz idatzitako materiala
   c. Eskuizkribuak
3. **Elektronikoa**

Modu orokor batean, corpus gehienek jarraitu dute sailkapen hau.

PAROLEk, esaterako, EAGLESen oinarritu bazen ere, sailkapen askoz ere orokorragoa egin zuen, besteak beste, lehendik existitzen ziren corpusekin zihardutelako lanean eta helburu nagusia gutxieneko bateratze bat lortzea zelako. Hauek dira diseinurako oinarrian adostu zituzten parametroak, nahiz horren barruan finduagoak dauden hizkuntza batzuetan:

1. Liburuak
2. Egunkariak
3. Aldizkariak
4. Miszelanea (korrespondentzia, elektronikoa, pasakorrak, eskuzkoa, makinakoa eta bestelakoak)

Etxeko adibideei dagokienez, *XX. mendeko euskararen corpus estatistikoa*ren azken epeko sailkapena ikusiko dugu, bere bi azpicorpusak banaka hartuta.

a) Obra sailkatuen (liburuak eta artikuluka landutako aldizkari nagusiak) azpicorpusa[31]:

[31] OEHk, gaurko irizpideetatik urrun samar badago ere, datuok ditu: bertsoa (46 obra, corpusaren % 14,8), poesia (39, % 12,5), erlijioa (88, % 28,3), prosa literarioa (76, % 24,5), prosa ez-literarioa (33, % 10,6) eta antzerkia (28, % 9). Alegia, literario / ez-literario banaketa eginez, literarioa 189 obrak osatzen dute: % 61. Ez-literarioa, berriz, 121 obrak, % 39. Gaurko estandarretan kontrakoa izan ohi da, baina tradizioan obra literarioek dute lekurik handiena. Beraz, erreferentzia-corpusa sortzeko ez digu askorik balio.

Sailkatuak: 1991-1999	Testu-masa (dok.)	% (azpicorpusa)	% (totala)
Saio-artikuluak	165.837 (202)	11.8	9
Administrazio-idazkiak	24.053 (18)	1.7	1.3
Ikasliburuak	225.220 (149)	16	12.3
Saio-liburuak	558.193 (414)	39.6	30.5
Prosa literarioa	205.372 (145)	14.6	11.2
Poesia	8.888 (11)	0.6	0.5
Antzerkia	37.936 (28)	2.7	2.1
Bertsoak	12.229 (17)	0.9	0.7
Ikerketa-lanak	8.569 (5)	0.6	0.5
Haur- eta gazte-literatura	139.754 (126)	10	7.6
Ahozkoak: ahozko jardunen transkripzioak	4.920 (4)	0.3	0.3
Liturgia	17.369 (10)	1.2	1
GUZTIRA	**1.408.340 (1129)**	**100**	**% 77**

Sailkatuen artean, medioaren araberako banaketa hau da: literarioa % 28.8 eta ez-literarioa edo informatiboa % 71.2.

b) Sailkatu gabeen edo kazetaritza-lanen (egunkariak eta aldizkariak) azpicorpusa:

Sailkatu gabeak: 1991-1999	Testu-masa (dok.)	% (azpicorpusa)	% (totala)
Egunkariak	263.211 (1.467)	62.5	14.4
Astekariak + hamaboskariak	98.366 (1.002)	23.3	5.3
Aldizkariak	59.613 (416)	14.1	3.2
GUZTIRA	**421.190 (2885)**	**100**	**% 23**

Corpusa bere osotasunean hartuta, beraz, hala bana daiteke:

	Testu-masa	%
Sailkatuak	1.408.340 (1129 dok.) Literarioa: % 28.8 Informatiboa: % 71.2	77
Sailkatu gabeak	421.190 (1178 dok.) Egunkariak: % 62.5 Aldizkariak: % 37.5	23
GUZTIRA	1.829.530 (2307 dok.)	100

literarioa corpus osoaren % 22.1 da eta informatiboa % 77.9. Baina hau ez da egia osoa; izan ere, sailkatu gabeetan aldizkari literarioak ere ageri dira, nahiz ez diren bereiz markatzen. Bestalde, liburuek eta artikuluek corpusaren % 77 osatzen dute eta kazetaritza-lanek % 23. Orekari dagokionez, corpusa ondo eratua dagoela esan dezakegu; besteak beste, argitaratua proportzionalki jasotzen duelako. Baina, horrez gain, gaurko estandarrek markatzen dutenetik gertu dago.

Halere, corpusa mugatua da, batez ere tematikaren araberako bilaketak bideratzea ia ezinezkoa delako. Adibidez, informatikan forma bat nola erabili den jakin nahi badugu, saio-artikuluak, saio-liburuak eta ikerketa-lanak behintzat banaka begiratu beharko ditugu. Generoaz gain, edukiak izan beharko luke irizpide nagusi.

Gaur egun sailkapen-irizpide desberdin samarrak baliatzen dituzte corpusek. EAGLESek oinarrizko irizpide batzuk markatu baditu ere, ez da beheko mailatara jaitsi. Eta, gorago esan dugunez, normala ere bada neurri batean. Izan ere, *oreka* mantendu behar da, barruko oreka, eta hori hizkuntza bakoitzaren unibertsoaren, egoeraren eta baliabideen araberakoa izango da. Garaia, geografia, edukia

(argitalpen-mota, sailak)[32] eta abar hartzen dituzte hainbat corpus-sortzailek kontuan.

Aipagai ditugun corpusetan, edukiari dagozkion irizpideak jaso ditugu hemen, nahiz banaketa gardena egiterik ez dugun izan. Horregatik ezarri ditugu medioa vs. argitalpen-mota vs. tematika, ez baitira beti kontuan hartzen.

Corpusa	Medioa / Argitalpen-mota	Tematika / Argitalpen-mota
BNC	Literarioa (% 25) Informatiboa (% 75)	*Argitalpen-mota:* % 60: liburuak % 25: egunkariak, astekariak, aldizkariak % 5-10: denetik (panfletoak, iragarkiak,...) % 5-10: argitaragabea <% 5: ahozkorako prestatutako idatzia
CTILC	Literarioa (% 44) Ez-literarioa (% 56)	*Literarioa* (generoa): 1. Saioa 2. Narratiba 3. Poesia 4. Antzerkia *Ez-literarioa* (sailkapen tematikoa): 1. Filosofia 2. Erlijioa eta teologia 3. Giza zientziak 4. Prentsa 5. Natur zientziak eta puruak 6. Zientzia aplikatuak 7. Arte ederrak. Aisia. Kirolak. Jolasak 8. Hizkuntza eta literatura 9. Historia eta geografia. Biografia 10. Korrespondentzia
CREA	Fikzioa (% 44) Ez-fikzioa (% 56) % 10 ahozkoa % 90 idatzia: % 49 liburuak % 49 prentsa % 2 efemera	*Fikzioa:* 1. Bertsoa 2. Prosa 3. Drama *Ez-fikzioa*: % 9: didaktika % 15: zientzia eta teknika % 8: gizartea, prentsa eta publizitatea % 6: erlijioa % 13: historia % 5: legeak eta zientzia juridikoak
CORGA	Liburuak (% 66.8) Egunkariak (% 26.41) Aldizkariak (% 6.78)	1. Ekonomia eta politika (% 22.5) 2. Kultura eta arteak (% 6.80) 3. Gizarte-zientziak (% 13.32) 4. Zientzia eta teknologia (% 5.67) 5. Bestelakoak (% 11.96) 6. Fikzioa (% 39.79) Horrez gain, epekako sailkapena ere badute (bost urterokoa).
CNC	Daturik ez	Daturik ez

[32] BNCk, esaterako, laginaren tamaina, obraren gaia, autorea (izena, adina, sexua, nongoa den), "maila" (literarioa edo teknikoa bada, orduan maila "jasoa" izango du) hartzen ditu kontuan.

HNK	Informatiboa (% 74) Fikzioa (% 23) Nahasiak (% 3)	% 38: egunkariak % 18: aldizkariak % 20: testuliburuak % 22: prosa % 1: imajinazioa % 1: saioa (hitzaldiak...)
HNC-gr	Liburuak (% 15.75) Egunkariak (% 69.01) Aldizkariak (% 6.97)	
HNC-hu	Egunkariak/astekariak (% 50) Literatura (% 10) Zientzia (arrunta) (% 13.3) Ofiziala (% 13.3) Pertsonala (% 13.3)	
FIDA	Ahozkoa Elektronikoa Idatzia 1.argitaratua (liburuak, aldizkakoak (egunkariak, aldizkariak...)) 2. argitaragabea (publikoa, barne-erabilerakoa, pribatua)	*Literarioa:* 1. Poesia 2. Prosa 3. Drama *Ez-literarioa:* 1. Zientifikoa (giza- eta gizarte-zientziak; natur zientziak eta teknikoak) 2. Ez-zientifikoa
CNG	1. Liburuak 2. Egunkariak 3. Aldizkariak 4. Miszelanea	1. N/A 2. Komertzioa 3. Geografia 4. Osasuna 5. Historia 6. Humanitateak 7. Aisia 8. Zientzia 9. Gizartea eta, generoa: 1. N/A 2. Publizitatea 3. Eztabaida 4. Ezaugarria 5. Fikzioa 6. Informazioa 7. Instrukzioa 8. Ez-fikzioa 9. Testu ofiziala 10. Testu pribatua
CORIS	Prentsa (% 38)[33] Fikzioa (% 25) Prosa akademikoa (% 12) Prosa legala eta administratiboa (% 10) Miszelanea (% 10) Pasakorra (% 5)	

[33] Atal hauetako bakoitza azpicorpus bat da. Bakoitza ataletan banatzen da eta hauek, aldi berean, azpiataletan.

CRPC	Egunkariak (% 60.8) Liburuak (% 22.6) Aldizkariak (% 7.7) Legegintza (% 2) Miszelanea (% 4.3) Orri solteak (% 0.3) Korrespondentzia (% 0.1)	
ANC	*ikus BNC*	

Guztiek irizpide bateraturik ez dutela ikusi badugu ere, halako ondorio orokor bat atera dezakegu:

a) Medioari dagokionez, informatiboa eta literarioa da banaketa nagusia.

b) Argitalpen-mota —euskarria, hobe— kontuan hartzen badugu, banaketa nagusia idatzia (argitaratua eta elektronikoa) eta ahozkoa da. Argitaratuaren barruan liburuak, egunkariak / aldizkariak, argitaragabea, elektronikoa eta ahozkorako prestatutako idatzia biltzen dira. Ahozkoa, berriz, hurrengo atalean aipatuko dugu.

c) Tematikari begiratuta, ez dago irizpide bateraturik, gehienek sailkapen zabalak egin dituzte-eta. Hala ere, medio literarioaren barruan poesia, narratiba eta antzerkia jaso ohi dira. Informatiboan, berriz, zehaztasun-maila oso desberdinak ageri dira.

Bide horretatik, euskararako ere halako zerbait proposatzea egoki ikusten dugu. Medioari dagokionez, literarioa vs. informatiboa banaketa egin beharko da, % 25 literarioa / % 75 informatiboa banaketara hurbilduz, estandarren eta gaurko corpus nagusien bidetik, nahiz ez dugun modu zurrunean aplikatzea proposatzen. Argitalpen-motari dagokionez, berriz, atal honetan idatzia bakarrik kontuan hartuta, liburuak, egunkariak / aldizkariak, material soltea eta ahozkorako prestatutako idatzia proposa ditzakegu, betiere kontuan hartuta idatzian argitaratua eta elektronikoa, biak aintzat hartzen ditugula. Azkenik tematikari begiratzen badiogu, literarioaren barruan narratiba, poesia, antzerkia eta, gure kasuan, bertsoa ere gehitu beharko litzateke. Informatiboan, berriz, gaika eta adituekin zehazten joateko kontua litzateke, bai eta hizkuntzak ematen dizkigun aukerak gogoan hartzen ere, gai batzuek besteak baino landuago baitaude eta ez da erraza izango proportzioan guztiak eskuratzea. Unibertsoa polikiago eta sakonago aztertu beharko da eta adituen beharrak ezagutu.

Ez dugu deus aipatu obra originalei eta itzulpenei buruz. Dudarik ez da, euskaraz, biek dutela "original" izaera eta ez dugu banaketa hori egin beharrik ikusten, beste hizkuntza batzuetan egiten duten bezala. Esaterako, egungo galegoa biltzen duen CORGA corpusak ez du itzulpenik corpuseratzen, hala erabaki dutelako, nahiz ez den jokaera orokorra corpusgintzan dihardutenen artean. Ez dugu halako planteamenduen beharrik ikusten euskararen kasuan.

Obren aukeratze-prozesurako, irizpideak ezarriko dituen *aholkulari-batzorde* bat beharko da, Euskaltzaindia, aditu gisa, buru duela. Batzorde horrek erabakiko du corpusean "sartu beharreko" ezinbesteko erreferentzia-obrak zeintzuk diren eta, beste guztiei oniritzia eman beharko die, batzorde berrikusle gisa, nahiz maila horretan bestelako adituak ere beharko diren. Corpusa aukeratzea, ezinbestean, *subjektiboa* izango da. Baina corpus barruko objektibotasuna lortzen behintzat saiatu behar da, gerora ere balio dezan. Argi izan behar da corpusaren onespena edo gaitzespena, neurri handi batean, aukeratutako edo baztertutako obren baitakoa izango dela: alegia, corpusean sartzen denaren *kalitatea* dela corpusaren balioa erabakiko duena, oreka ahaztu gabe. Puntu honetan autoritatea dutenek

ontzat ematen badute, corpusa benetan izango da onartua eta, ondorioz, orduan esan ahal izango dugu orotarikoa dela.

8. Idatzia + ahozkoa

Corpus idatziak eta ahozkoak bereiz landu izan dira hainbat kasutan, batez ere aplikazio desberdinetarako pentsatuta zeudelako. Gaur egun, eta erreferentzia-corpusak nagusi direla gainera, biak elkartzea —betiere azpicorpusen askatasuna mantenduz— egoki (eta ia beharrezko) ikusten da. Egia da, gero, praktikan, ahozkoa "hurrengo faserako" uzten dutela gehienek, lan ikaragarria baita emaitza mugatuagoa lortzeko. Idatzizkoa, azken urteetakoa batik bat, modu elektronikoan eskuratu eta berehala landu daiteke. Hala ere, corpusaren orekari begiratzen bazaio, ahozkoaren garrantzia ukaezina da.

PAROLEk, hasieran esan dugunez, idatzia besterik ez zuen kontuan hartu, ahozkoa lantzeko SpeechDat[34] taldea sortu baitzen. Irizpideak, hemen ere, EAGLESen oinarrituak ziren.

XX. mendeko euskararen erreferentzia corpusa idatzia da eta, ahozkoa, transkribatu eta argitaratu den neurrian bakarrik jaso da. Badira, halere, euskaraz gai honen inguruan diharduten lantaldeak, beherago ikusi ahal izango dugunez.

Aipagai ditugun erreferentzia-corpusetan datuak argigarriak dira:

Corpusa	Idatzia	Ahozkoa
BNC	% 90	% 10
CTILC	% 100	
CREA	% 90	% 10
CORGA	% 100	
CNC	% 100[35]	
HNK	% 100	
HNC-gr	% 100	
HNC-hu	% 100	
FIDA	% 100[36]	
CNG	daturik ez	daturik ez
CORIS	% 100	
CRPC	daturik ez	daturik ez
ANC	*% 90*	*% 10*

Taula honek ondorio garbi batera garamatza: ahozkoa oso gutxik lantzen dute eta, jasotzen dutenen artean, gainera, leku gutxi dute —% 10 gehienez ere—. Arrazoiek ere ez dute azalpen askoren beharrik, batez ere azken-azkenik bideratu diren corpusei (salbu *ANC*, *BNC*aren irizpide berak hartzen ditu-eta) kasu egiten badiegu: eskuratzen den emaitza ez da adierazgarriegia eta lana itzela da.

Hobeto ulertzeko, ikus dezagun BNCeko ahozko atala zein irizpidetan oinarrituta egin den. Bi sailetan banatu dituzte bildu beharrekoak:

a) demografikoa: 124 boluntario: lau gizarte-mailatakoak, % 50 gizonezkoak / % 50 emakumeak, adin guztietakoak, 38 herri desberdinetakoak. 2-3 egunetan egindako elkarrizketak, egoera guztietan.

[34] SpeechDat: http://www.icp.inpg.fr/SpeechDat/home.html

[35] Idatzia besterik jasotzen ez dutela dirudien arren, egileek "mainly written Czech" aipatzen dute. Beraz, ahozkoa ere, neurri batean behintzat biltzen dutela pentsa daiteke, baina ez dute argitzen.

[36] Ahozkorik ez da sartu, baina ahoz erabiltzeko idatzi dena bai (hitzaldiak, legebiltzarreko aktak, etab.)

b) testuinguru berezikoa:
   i. hezkuntza, informatiboa: eskolak, mintegiak...
   ii. negozioak: saltzen, bilerak...
   iii. publikoak: sermoiak, hitzaldi politikoak, udaleko bilkurak...
   iv. bestelakoak: kirol-komentarioak, lagunarteko tertuliak, irratira egiten diren deiak...

Euskararen kasuan ere, ahozkoa lantzeak badu bere garrantzia eta, gainera, hainbat material eskuragai egon daiteke, hainbat lantalde aspaldi ari baita ahotsarekin lanean. Gehiago ere izango dira, beharbada, baina hona guk ezagutzen ditugunak:

a) Euskaltzaindiak Atlasa osatzeko egiten dituen elkarrizketak: 145 herritan egindako 2.800 galdera[37] (nahiz, zenbaitetan, bideratuegiak izan daitezkeen).

b) Ahotsaren tratamenduan diharduen AHOLAB taldea.

c) BIZKAIFONek[38] bildutako hainbat material.

d) Irrati-telebistetako sail desberdinak. Bereziki aipatzekoa ERESOINKAk bildutako material ugaria.

e) Legebiltzarreko bilerak, Foru Aldundietakoak, udaletxetakoak...

f) "Basque Spoken Corpus": 42 euskal hiztun, John Askek bildutako ahozko corpusa[39].

Hainbat material, b), c) eta f) sailekoak bai behintzat, tratatuak daude eta corpusean sartzea lan erraza izan liteke. Besteak modu errazean biltzeko adituekin batera landu beharko litzateke —ahotsaren tratamenduan dihardutenekin, hain zuzen—. Horiek horrela, corpusaren % 10 gehienez —estandarra kopuru horretan baitago, kasu egiten zaionean— ahozkoari eskain dakioke, idatziarekin paraleloan landu baitaiteke.

9. Egituratua

Corpusa, erabilgarri, hitz-bilduma hutsa baino gehiago denean da. Hizkuntza Naturalaren Prozesamenduan (HNP) diharduten ikertzaileek corpus etiketatuak, landuak baliatzen dituzte, horietatik lortzen baitute informaziorik osoena.

Corpusa egituratzea "korronpitzea" (John Sinclair, 1998) da neurri batean, markatu egiten baita eta, batez ere, interpretatu (kodetzean, markatzean). Hizkuntzalariak detektatzen dituen metahizkuntzari buruzko informazioak eta bestelako datu inplizituak, esplizitu jarri behar dira erabiltzaileak automatikoki eskuratu ahal izateko (eta, ondorioz, emaitza okerrik ez baliatzeko[40]). Aldi berean, gainera, edozein unetan testua hasierako formatura bueltatzeko moduan eduki behar da.

TEItik[41] sortutako kodetze-sistema da estandartzat hartu ohi dena, eta, bereziki, dokumentuak markatzeko honek baliatzen duen SGML[42] (XML), lau abantaila

[37] Datuok Beñat Oihartzabal Euskaltzaindiko Iker Sailburuak emanak dira.

[38] http://bizkaifon.ehu.es/

[39] Corpus hau ELDAn eskura daiteke, http://www.elda.fr/cata/speech/S0123.html helbidean.

[40] Adibidez, autore batek forma bat markatua (lodia, etzana edo dena delakoa) erabil dezake, hain zuzen ere gaitzesteko. Horri buruzko oharrik jartzen ez bada (metahizkuntza dela, esaterako), forma horren erabileretan autore hori ere azalduko da, forma berea balitz bezala, alegia, normal erabiliko balu bezala, kontrakoa denean. Erabiltzaileak testuinguru zabala irakurri beharko luke aldiro halakoak detektatzeko. Beraz, testua kodetzean horren berri ematen bada, erabiltzaileak informazio egokia jasoko du eta ez du horren erabilera okerrik egingo (ez, behintzat, corpusetik jasotako okerrik).

[41] TEI: Text Encoding Initiative.

[42] SGML: Standard Generalized Markup Language. SGMLren berrikuntza XML (eXtended Markup Language) da, orokorragoa eta malguagoa.

ikusgarri dituena: argitasuna, sinpletasuna, formalki zurruna/zehatza eta nazioarteko estandar gisa onartua. Eta, honen barruan, CES[43] (XCES) arduratzen corpusen kodetzeaz.

SGML/CESen kodetzean, informazio metatestuala (bibliografikoa) "header" atalean bilduko da eta testua bera "text" atalean, informazioa banatuaz. Testua markatzeko orduan, aukera ugari daude gainera: etiketa edo "tagging" gisa ezagutzen direnak (kategoria + informazio morfosintaktikoa: POS[44]), lematizazioa, analisi sintaktikoa (parsing), semantika, diskurtsoa, transkripzio fonetikoa, prosodia eta arazoei zuzendutako etiketatzea, besteak beste.

Guk, atal honetan, hiru irizpide hartuko ditugu abiapuntu gisa: corpusa kodetua izatea, eta honen barruan, bereziki, etiketatua (*tagged*) eta lematizatua izatea, hain zuzen, CESek proposatzen duena gure eginez.

1) Kodetua

CESek oinarrizko kodetze bat proposatzen du corpusak estandartzat hartu ahal izateko: errepresentazio deskriptiboari dagokionez —informazio estrukturala eta tipografikoa markatuz— eta arkitektura orokorrari dagokionez —datu-baseak baliatuz—. Bestalde, markatze linguistikorako zehaztasunak proposatzen ditu, aipatua dugunez. Gu, atal honetan, errepresentazio deskriptibora mugatuko gara.

Orain artekoak aipatuz, PAROLE TEI-SGMLn kodetua dagoela esan beharra dago. Proiektu horretan finkatu ziren, gainera, irizpide nagusiak, bai eta hizkuntza bakoitzak bereak zituen ezaugarri nagusiak batera adierazteko proposamenak ere.

XX. mendeko euskararen corpus estatistikoa kodetua da, TEI-SGMLn oinarritua, CREA gaztelaniaren erreferentzia-corpuseko irizpideetan oinarritua. Markatzea bi fasetan egin zen: marka tipografikoak aurrena, eta marka lexikografikoak hurrena. Horrela, corpus osoa modu estandarrean kodetua dagoela esan dezakegu.

Beste hizkuntzetan ere hori da jokabiderik hedatuena, nahiz ez duten guztiek egiten:

Corpusa	***Kodetua***
BNC	TEI-SGML
CTILC	sistema propioa[45]
CREA	TEI-SGML / CES
CORGA	Hasi gabe, baina XMLrantz[46]
CNC	Daturik ez
HNK	XML / CES
HNC-gr	TEI / CES
HNC-hu	SGML / CES
FIDA	Daturik ez
CNG	SGML
CORIS	Daturik ez
CRPC	Daturik ez
ANC	*TEI-SGML*

Hasiera batean SGML zen hedatuena, baina gerora XMLra egin dute jauzi corpus berriak kodetzen hasi direnak eta, noski, CES baliatzen ere bai, berriena[47] baita.

---

43 CES: Corpus Encoding Standards. CESen azken bertsioa XCES (Corpus Encoding Standards for XML) da.

44 Part of Speech (POS) annotation gisa ezagutzen dena.

45 Baina, egileek diotenez, PAROLEra pasa zen zatia SGMLratzeko ez zuten arazo larririk izan, sistema propioa bai, baina estandarretatik nahiko gertu omen dago-eta.

46 Erdara bakarrik markatzen dute oraingoz ("foreign" marka). Guztia XMLn jartzeko asmoa dute, bi DTD prest dituzte, baina etorkizuneko proiektu gisa planteatzen dute, ez berehalako lan bezala.

47 XCES, oraingoz, proba-fasean dago.

XXI. mendeko euskararen erreferentzia-corpusa erabilgarria, elkartrukerako egokia eta orotarikoa nahi badugu, estandarrok ezinbestean baliatu beharko dugu. Beraz, proposamena CESen irizpideak —oinarrizkoak behintzat— jarraitzea izango da. Oinarrizkoak, tamaina handiko corpusa ari garelako proposatzen eta, beraz, modurik automatikoenean egiten saiatu behar delako. Baina informazio lexikografikoa ezin da automatikoki landu eta, ondorioz, muga batzuk ezarri beharko dira, edo guztia eskuz begiratuko duen lantalde handi bat osatu.

2) Etiketatua

Corpus etiketatua diogunean etiketa morfosintaktikoak berekin dituela esan nahi da, alegia, "tagging" gisa ezagutzen direnekin.

Orain arteko datuekin jarraituz, PAROLEko 14 hizkuntzetako corpusak morfosintaktikoki etiketatuak daude. *XX. mendeko euskararen corpus estatistikoa*, bestalde, etiketatu gabe dago. Baina, nola jokatzen dute beste hizkuntzetako corpusek?

Corpusa	*Etiketatua*
BNC	Etiketa morfosintaktikoak
CTILC	Etiketa morfosintaktikoak
CREA	Etiketa morfosintaktikoak (ez osoa)
CORGA	Hasi gabe[48]
CNC	Daturik ez
HNK	Etiketa morfosintaktikoak (hasten)
HNC-gr	Etiketa morfosintaktikoak
HNC-hu	Etiketa morfosintaktikoak
FIDA	Daturik ez
CNG	Etiketa morfosintaktikoak
CORIS	Daturik ez
CRPC	Daturik ez
ANC	*Etiketa morfosintaktikoak*

Datuon argitan ikus dezakegu gehientsuenak etiketatuak —edo etiketatze bidean— daudela. Hain zuzen ere, horrek egiten baitu corpusa benetan erabilgarri eta orotariko.

Euskararen erreferentzia-corpusa bideratzeko proposamena ere bide honetatik plantea dezakegu: corpus etiketatua beharko genuke, minimoki etiketatua behintzat. Baina, nola ezarri etiketa morfosintaktiko horiek? Dagoeneko esku artean duguna baliatuz, EUSLEMek[49] eskaintzen duen informazioa erabiliz —eta, noski, berrikusiz—, betiere zuzenketak egin beharko baitira. Lexikografikoaren orrazketa-lana ezinbestekoa da kalitatezko corpusa lortu nahi bada behintzat.

3) Lematizatua

Corpusa lematizatzea testu-hitz bakoitzari lema estandar bat esleitzea da, alegia, erabilera errazteko helduleku bat ezartzea.

XX. mendeko euskararen corpus estatistikoa lematizatua dago, lana bi fasetan egina bada ere: 1900-1990 urteak tresna propio[50] bat erabilita lematizatu ziren, giza laguntza baliatuz. 1991-1999 urteak, aldiz, EUSLEM erabiliz lematizatu ziren, automatikoki, hemen ere guztia eskuz berrikusi bazen ere. Corpus estatistikoak eskaintzen duen lematizazioa xehea da, alegia: sarrerak, azpisarrerak (hitz elkartuak, aditz konposatuak,...), hitz anitzeko unitate lexikalak (lexiak, lokuzioak,

[48] Pentsatua bai, etiketatu eta desanbiguatu egingo dute corpusa, ahal den neurrian, automatikoki. Baina hasi gabe dagoen proiektua da hau ere, kodetzea bezala.

[49] EUSLEM (Euskararako Lematizatzaile Automatikoa), UZEIk eta IXA taldeak elkarlanean garatua. Hain zuzen, *XX. mendeko euskararen corpus estatistikoa* lematizatzeko erabilia, 1991-1999 urteetakoa.

[50] Baionako HIZKIA etxeak prestatutako *Rterm* baliatuz, nahiz eskuz berrikusi behar zen lematizatutako guztia, banaka.

esapideak,...) eta gramatikako osaerak biltzen ditu, besteak beste. Gainera, partez desanbiguatua da kategoriak eta adierak argituz, baina Hiztegi Batuko Lantaldearentzat hitzei buruzko txostenak prestatzen diren neurrian bakarrik, ez modu sistematikoan. Eta, informazio hau ez dago erabiltzaileen esku, UZEIko prestalaneko taldeak eguneroko lanean darabilen datu-basean besterik ez dago.

Beste hizkuntzetako corpusetan datuok aurkitu ditugu:

Corpusa	***Lematizatua***
BNC	daturik ez
CTILC	lematizatua + desanbiguatua
CREA	lematizatua (hasten)
CORGA	hasi gabe[51]
CNC	daturik ez
HNK	lematizatua (hasten)
HNC-gr	lematizatua
HNC-hu	lematizatua
FIDA	daturik ez
CNG	daturik ez
CORIS	daturik ez
CRPC	daturik ez
ANC	*lematizatua*

Corpus guztiak ez daude lematizatuak. Izan ere, lan astuna da eta, askotan, hizkuntza-tipologiaren araberako "beharra" egon daiteke. Alegia, batzuetan etiketa morfosintaktikoek helarazten duten informazioa aski izan daiteke.

Horiek horrela, eta euskararen tipologia kontuan hartuta —bai eta aurreko corpusen erabileran dugun esperientzia ere—, corpus lematizatua beharko genukeela uste dugu, zalantzarik gabe gainera. Are gehiago, lema estandarraz gain, aldaeraren lema izatea ere laguntza handia litzateke. Bilaketa-sistema erraz eta azkarra izatea oinarrizkoa da erabiltzailearentzat lagungarri izateko. Gainera, lemak, aldaeren lemak eta testu-hitzak konbinatzeko aukerak izanda, lexikoaren aldetik gutxienez aukerak ugariak lirateke (gaur *XX. mendeko euskararen corpus estatistikoa*rekin gertatzen denaren antzera, nahiz aldaeren lemaren beharra ikusi den eta, UZEIko barruko erabileran behintzat, partez landua dagoen).

Horretarako EUSLEM baliatzea oinarrizko dela deritzogu, gutxienez ondoko informazioa eskuratu ahal izateko: lema, kategoria[52] + azpikategoria eta markatze morfologikoa —are morfosintaktikoa—, alegia, etiketatua.

Orain arteko emaitzei begiratzen badiegu, EUSLEMen, testuen estaldura % 100ekoa da —hiztegirik gabeko lematizazioa egiteko ere prestatua dago, lexiko-itzultzaileen teknologia erabiliz, hau da, itzultzaile orokorra baliatuz— eta doitasuna, euskara estandarrari aplikatuta behintzat, ia % 99koa da. Horren adibide, *XX. mendeko euskararen corpus estatistikoa*ren azken bederatzi urteetako lematizazioa. Hori bai, UZEIko lantaldeak eskuz berrikusi, zuzendu eta desanbiguatu ditu EUSLEMetik etorritako lema guztiak, erabateko kalitatea bermatu ahal izateko, hitz anitzeko unitateei dagokienez batez ere.

Honaino corpusaren osaerari dagozkion irizpide orokorrak. Datuok finduz eta zehaztuz joan beharra egongo da, euskararen beraren egoera kontuan hartuta. Baina, gorago esan dugunez, aholkulari-batzorde bat eratu beharko da irizpideak zehazteko, gerora lantaldea materiala biltzen eta corpusa bera eratzen hasteko.

---

[51] 100.000 hitzeko azpicorpus bat desanbiguatzen ari dira, gero beste guztiari aplikatu ahal izateko.

[52] EUSLEMeko kategoria-sistema nolabait moldatu edo egokitu beharko litzateke beharbada, izendapen estandarretara ekarri ere bai, baina hori lan handirik emango ez lukeen aldaketa litzateke, nahiko automatizatua izan liteke.

Gogora dezagun puntu honen hasieran esan duguna, alegia, erreferentzia-corpusak ***euskara modernoaren erakusgarri ahalik eta zabalena*** izan behar duela. Aurreko orrietan proposatuarekin helburura hurbiltzen garelakoan gaude. Ez dezagun ahantz, baina, corpusa ez dela berez helburu, hizkuntzaren oinarrizko tresna baizik: alegia, ondoko lanen abiapuntu.

3.2. Baliabideak

Aurreko ataletan corpusari buruz aritu gara, nolako corpusa beharko genukeen aipatu dugu. Halere, ondoko lerroetan existitzen diren baliabideei buruzko proposamenak egingo ditugu, batzuk zeharka aipatu ditugun arren, horiek bateratu beharra dagoela uste baitugu. Honen oinarrian elkarlana dagoela uste dugu, eta hori erakusten saiatuko gara.

3.2.1. Giza baliabideak

Gaur ezagutza-maila handia dagoela uste dugu, hainbat esparrutan gainera, eta horiek guztiak elkartu beharko lirateke: lexikografoak, gramatikariak, arlo desberdinetako adituak, informatikariak eta hizkuntzalari konputazionalak, besteak beste.

UZEIko Lexikografia Sailak 1986tik dihardu corpusgintzan eta, beraz, badu eskarmentua. *XX. mendeko euskararen corpus estatistikoa*ren osaeran fase guztiak landu ditu (unibertsoa osatu, lagina aukeratu, eskuratu, testuak kodetu, lematizatu eta ustiatu batetik, baina baita datu-base egokiak aztertu eta aukeratu, aplikazioak egin eta Interneten kontsultagai jarri ere).

Orotariko Euskal Hiztegiaren corpusa eratzen ere aritua da Euskaltzaindia, bai eta corpusa bera lantzen eta ustiatzen. Eta Euskaltzaindiko beste batzorde eta lantaldeak ere bai: Gramatika Batzordea eta Hiztegi Batuko Lantaldea behintzat ohituak daude corpusak baliatzen haien lanetarako. Ber gauza esan daiteke unibertsitateetako ikertzaileei buruz, horiek baitira, hiztegigile eta terminologoekin batera, corpusaren orain arteko erabiltzaile nagusiak.

Bestalde, EHUko IXA taldeak eta UZEIk elkarlanean sortu dute EUSLEM lematizatzaile automatikoa eta, IXA taldea, beraz, ohitua dago bere ikerketa-lanetarako corpusetan oinarritzen. Ezinbesteko du, gainera.

Ahotsaren tratamenduan diharduen hainbat aditu ere bada gure artean, eta horien ezagutza eta esperientzia kontuan hartu beharko da, batez ere ahozko corpusa haiek landu beharko bailukete: Euskaltzaindia, AHOLAB, BIZKAIFON eta halakoak (gorago aipatu ditugunak, alegia).

Bada, beraz, corpusen inguruan aritu den eta ari den jendea, bai sortzaile gisa, bai erabiltzaile gisa. Aditu —eta ez hain aditu— bakoitzak ditu bere beharrak eta horiek definitu beharko lirateke, corpusa lantzen hastean helburuetan gehitu ahal izateko. Alegia, hasi aurretik jakin behar da zeintzuk izango diren erabiltzaile potentzialen beharrak (eskoletako irakasleengandik hasi eta ikertzaile espezializatuenganaino, alor guztiak ukituz).

3.2.2. Baliabide materialak

Corpusaren lehengaia testuak eta ahozko osagaiak dira, hau da, euskarazko materiala. Gaur egun erraz eskura daiteke informazioa.

a) Idatzia:

* Argitaletxeek kaleratzen duten gehiena —guztia ez bada— euskarri elektronikoan dute gaur, eta neke handirik gabe eskura daiteke (egile-eskubideak eta halakoak alde batera utzita, noski). Bestalde, EIMA programan laguntza jasotzen dutenen kopiak gordetzen dira eta horiek lortzea ere planteatu beharko litzateke. Are gehiago, Euskal Idazleen Elkartea, EIZIE eta gisa horretako erakundeek ere parte hartu beharko lukete maila honetan. Eta ez corpusa osatuko duten obrak helarazteko bakarrik, baita irizpideak ezartzeko orduan ere.

* Egunkariak eta aldizkariak kontuan hartu beharko lirateke, bai eta herri-aginteetako aldizkari ofizialak eta dokumentu publikoak (edo publikoaren eskura jar daitezkeenak).

* Azkenik, ahozkorako prestatutako idatzia kontuan hartzekoa da. Hemen ezin ahantz daiteke Eresoinkak urtetan bildu duen material guztia, ez eta irrati eta telebistetako materiala ere. Horiekin batera, idatziz prestatutako hitzaldiak eskuratzea egoki litzateke, publiko jar daitezkeen neurrian behintzat.

b) Elektronikoari dagokionez, euskararako bilatzaileak prestatu eta erabiltzen dituztenak izan behar dira gogoan, bai eta zerrendak sortu, mantendu eta erabiltzen dituztenak ere, handik eskura baitaiteke hainbat material.

c) Ahozkoa: gorago aipatu dugun arren, hainbat taldek dihardu ahotsaren tratamenduaren inguruan lanean eta material ugari dute jasoa eta, batez ere, landua. Egina dagoena, beraz, baliatu beharko litzateke. Horrekin batera, irrati eta telebistetako emanaldiak jaso beharko lirateke, hitzaldiak, sermoiak, kaleko elkarrizketak... corpusaren diseinuan aukeratutako guztiak, hain zuzen.

3.2.3. Baliabide teknikoak (tresneria)

Egun tresna egokiak daude testu-masa handi egituratuak ustiatzeko, baita gure artean ere. Euskararen tratamendu automatikorako baliabide tekniko berrerabilgarriak, malguak eta elkartrukerako egokiak ditugu eskura: estandarrak. Urteak dira hainbat aditu gai honen inguruan lanean ari dela eta ezagutza hori zein garatutako tresneria baliatu beharko litzateke: alegia, teknologia berriak erabili eta berrerabili beharko lirateke.

UZEIk ORACLE datu-basearekin lanean urteak daramatza, 1992.etik hain zuzen, eta, urteotako eskarmentuen ondoren, corpusen tratamendurako egokia dela erakutsi du, batez ere azkenaldiko bilaketa dokumentaletarako hobekuntzei esker. Corpus handi batek tresna eraginkorra behar du, eta informazio egituratu asko erraz kontsultatzeko moduan gainera. Euskalgintzan diharduten talde askok oraintxe egin dute ORACLE baliatzeko apustua, egokia dela ohartu baitira. Oraindik ere konkordantziak eskuratzearekin nahiko dutenak ageri dira, baina hori corpusak eskaini behar duen emaitzetako bat besterik ez da, eta ez landuenetakoa gainera. Hori baino gauza orokorragoa, osoagoa eta baliagarriagoa behar dela erakutsi du urteotako informatikarien eta lexikografoen arteko elkarlanak.

Hala ere, datu-baseaz gain, horretan trebatuak diren informatikariak behar dira, aplikazio lexikografikoetan adituak direnak. Hain zuzen, haiek bideratuko baitituzte corpusak kudeatzeko programak, hau da, corpusak eguneratu, kontsultatu eta ustiatzeko programak. Beraz, datu-baseak beharrezkoak dira, baina programak edo aplikazioak eta horretan arituak eta adituak ere ezinbesteko dira.

Baliabide teknikoei dagokienez, oinarrizko tresnez gain, corpusa osatzen lagunduko duten bestelako aplikazioak behar dira. Esaterako, EUSLEM lematizatzailea martxan da aspaldi (*XX. mendeko euskararen corpus estatistikoa*ren azken urteak lematizatzeko erabili da batetik, baina corpusa bera ere baliatu da lematizatzailea aberasteko eta hobekuntzak egiteko). Lematizatzailearen atzean daude, gainera, Hizkuntza Naturalaren Prozesamenduan egin diren bestelako lanak, bateratu eta

berrerabili beharko liratekeenak (analizatzaile morfologikoa, desanbiguatzailea, datu-base lexikala, aldaeren tratamendua), guztiek lematizazioa hobetzen lagunduko baitute.

3.2.4. Baliabide ekonomikoak

Aurreko puntuetan ezagutza eta tresnak badaudela ikusi dugu. Euskara ez dago, bada, beste hizkuntzengandik hain urrun: ezagutza badu, tresnak ere bai, estandarrak ezagutzen eta erabiltzen ditu. Baliabide ekonomikoak eskuratzeko ezinbestekoak dira Herri-Aginteak, baina baita enpresa pribatuak ere, gerora corpusa balia dezaketenak. Argi dago aplikazio askotarikoa izango dela eta, beraz, ez dela euskararen mundura bakarrik mugatu behar: "hizkuntzen industriari" dagokio eta, ondorioz, industriak ere parte hartu behar luke (bai enpresak, bai administrazioak).

Etorkizuneko lan askoren oinarri izango da, eta inbertsio gisa ulertu behar da zalantzarik gabe.

3.2.5. Egile-eskubideak eta bestelako lege-arazoak

Baliabideekin batera sartu dugu puntu hau, badelako horretan ere esperientzia. Corpuseko osagaiak ezin dira besterik gabe hartu eta datu-basean sartu, baimenak eta hitzarmenak behar dira. Euskaltzaindiak, *Lamiategi* proiektuan, egina du bere bidetxoa eta baliagarri izango da, noski. Baina legelariak ere beharko dira kontu hauek argitzeko, nahiz badiren nazioarteko irizpideak, *Oxford Text Archive* taldekoek landutakoak bereziki. Ikus, horrez gain, McEnery (2002).

Aurreko puntu hauek guztiek ondorio bakarrera garamatzate: *elkarlana* beharrezkoa da etorkizuneko erreferentzia-corpusa sortu nahi badugu. Eta, honek, goian aipatutako agenteen parte-hartzea, gutxienez, eskatzen du. Hala ikusten dugu guk:

1. Irizpideak ezarriko dituztenak
   - Corpusaren diseinuan: landu beharreko osagaiak eta estandarrak; arkitektura
   - Materiala aukeratzeko irizpideak zehazten
   - Lege-arazoak argitzen
2. Materiala helaraziko dutenak
3. Corpusa landuko dutenak
4. Dirua jarriko dutenak: administrazioa, enpresak

... eta erabiltzaileen iritzia kontuan hartu beharko da, haientzat egingo baita.

4. EMAITZAK ETA ONDORIOAK

Corpusa erabilia eta erabilgarri den neurrian bakarrik izango da baliozkoa. Ondorioz, ustiapen ahalik eta zabalena behar du: erabiltzaile-talde zabala izan behar da gogoan. Gure proposamenean erabilera batzuk besterik ez ditugu zehaztu, baina informazioa ondo eratua, antolatua eta landua badago, programazio-lana izango da behar zehatzei erantzun ahal izateko beharko dugun gauza bakarra.

Txostenaren hasieran aipatu dugunez, erabiltzaile-mota zabala izango da corpusa baliatuko duena: lexikografian hiztegigileek baliatuko dute gehienbat, hitzen erabilerak dokumentatzeko, adibide egokiak bilatu, adierak landu, besteak beste. Gramatikan ere, morfologia eta sintaxiko jokaerak, urruneko mendekotasunak, aditzen subkategorizazioa... Ezin ahantz daitezke kolokazioak ere.

Baina, erabiltzaile arruntek askotan nahikoa izango dute konkordantziak ikustearekin, adibidez. Edo maiztasunak, ikuspegi orokorrak izateko, esaterako. Ahotsaren tratamenduak hainbat datu berri ekarriko ditu, zalantzarik gabe.

Baina ez dira hauek corpusetik eskura daitezkeen datu guztiak. Beharrek markatuko dute zer eta nola ustiatuko den.

Txosten honen hasieran aipatu ditugun erabilera-aukeretara bueltatuz, guztiei erantzuteko gai izan beharko du corpusak: adierak, kategoria sintaktikoak, klase semantikoak, egitura argumentala, hitzen arteko kookurrentziak, unitateen agerpen-maiztasuna, unitateak bere testuinguruan, analisi probabilistikoa, erlazio lexikoak, erabilera-adibideak, murriztapen selektiboak, hitz elkartuak, lexiak, lokuzioak, etab.

Azkenean, etorkizuneko euskararen corpusa, *corpus nazionala*, beharko dugu, guztiona eta guztiontzat, orotarikoa. Eta, horrekin lotuta, *corpus publikoa* eta, noski, publikoaren esku egongo dena.

5. ERREFERENTZIAK

Euskaltzaindia (1996): II. Jagon Jardunaldiak (Tolosa), in *Euskera*, 1996-3, 41. liburukia (separata)

Euskaltzaindia (1986): *Euskera*, 1986, XXXI, 130. or.

Eusko Jaurlaritza / Hizkuntza Politikarako Sailordetza (1999): *Euskara Biziberritzeko Plan Nagusia (EBPN),* Gasteiz.

Hernáez, I., Navas, E., Sánchez, J.M., Madariaga, I., Gaminde, I. eta Zalbide, X. (2002): "BIZKAIFON: A sound archive of dialectal varieties of spoken Basque", in LREC 2002, Las Palmas de Gran Canaria.

López de Ipiña, K., Ezeiza, N. eta Bordel, G. (2002): "Automatic Morphological Segmentation for Continuous Speech Recognition of Basque", in LREC 2002, Las Palmas de Gran Canaria.

MacMullen, John (2002): "Requirements Definition and Design Criteria for Test Corpora in Information Science", in http://www.ils.unc.edu/~macmw/inls110/corpora-paper.pdf.

McEnery, Tony & Wilson, Andrew (2001): *Corpus Linguistics,* in: http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm.

McEnery, Tony (2002): "Ethical and legal issues in corpus construction", in: LREC 2002, Las Palmas de Gran Canaria.

Urkia, M. (2001): "Euskararen erreferentzia-corpusaren beharraz", in *Euskaltzaindiaren Nazioarteko XV. Biltzarra*, Bilbo (argitaragabea).