EUSKALTZAINDIAREN CORPUSEZ

B. Oyharçabal,
***Iker** sailburua*

Euskaltzaindiarentzat garrantzi handia dute testuek, hauen ikerketan oinarritzen baitira bereziki haren lanak. Aski da gure erakundearen iraganeko gorabeherei behatzea, berehala horretaz konturatzeko. Nork dudan jar dezake testuen euskarri sendoa izan gabe biziki zail dela deus funtsezkorik egitea hizkuntzaren finkatzeko? Testu oinarri horren garrantzia are handiagoa egin zaigu azken hamarkadan, teknologia berriei esker, duela guti ezin asmatuzkoak eta ezin amestuzkoak ziren ikerbideak eskuen artean izan baititzakete orain hizkuntzalariek, eta ondorioz lan egiteko erak, erran nahi baitu hizkuntza datu askoren azterkatzeko baldintzak, osoki aldatu baitira.

Euskaltzaindia, beharrik, hasia da azken urteetan gisa horretako baliabideez baliatzen, eta orok dakigu, adibide baten emateko, OEHak ez zuela orain duen itxura izanen euskal literatura zaharreko testu asko eta asko era digitalizatu batean jaso izan ez balira, eta EEBS ere egina izan ez balitz.

Haatik, teknologia berriek mugatzen duten eremu hau gelditu gabe aldatuz doa, eta laster zaharkiturik geratzen dira sail horretan egiten diren aitzinamenduak. Pentsa zitekeen, beraz, denbora joan arau, horrelako zerbait gertatuko zela Euskaltzaindiak abiarazi programa zenbaitekin, eta, hain zuzen, hala gertatu da: franko laster zaharkiturik gelditu dira. Horrela, bada, OEH bukatzera doalarik eta EEBS ere amaitutzat jo dezakegularik, ordu da, berant ere gabiltzala erran dezakegu, Euskaltzaindiak kontu honetaz gogoeta sakonak egin ditzan eta ondoko urteei buruz ildo berriak urra ditzan, deus ikustekorik ez baitute gaur egungo teknologia baldintzek eta duela hogei edo hamar urtekoek, ikergai izan daitekeen testu-multzoaren tamainari dagokionaz bereziki.

Areago dena, Euskaltzaindia ez dabil bakarrik langintza horretan. Euskalaritzako partaide franko hasi dira hizkuntzaren gaineko teknologia horretan, baita honek ekarri ekoizpenetan ere, lan egiten, bai unibertsitatean, bai hartarik kanpoan ere, eta guztiz beharrezkoa da gure erakundeak argiki erran dezan zer dioen horretaz, zer har dezakeen bere gain, zer ez, eta nola joka daitekeen euskarak gau eta bihar ahal bezainbat probetxu atera ahal dezan teknologia berriek corpusen eratzeko ematen dituzten ahalmen berezi horietarik. Euskaltzaindiak Autonomia Erkidegoko herri agintariekin izenpetu duen

aurtengo protokoloan, hitz eman die corpusak zirela-eta zerbait proposamendu aitzinatuko zuela eta eskatu die horretaz zerbait erabakitzerakoan haren hitza bereziki kontuan hartzea. Badakigu proposamenak eta galdeak plazaratzen direla, gero eta gehiago, eta deliberoak gaurtik biharrera beti ezin gibela daitezkeela. Tenorea dugu, beraz, gure erantzunkizunei buru egiteko, eta corpusak direla-eta erran dezagun argiki zer den Euskaltzaindiaren hitza.

Txosten hau ideia horretan egina da: gogoeta egitera laguntzeko eta Euskaltzaindiari proposamen zenbaiten egiteko. Hiru puntutan aurkeztuko dut:

- oinarrizko termino batzuen argitzea;
- Euskaltzaindiak orain arte eratu dituen corpusen deskribatzea;
- Euskaltzaindiak geroari buruz izan ditzakeen jokabideen aurkeztea, eta proposamenen egitea.

1. ZENBAIT OINARRIZKO KONTZEPTU

1.1. Corpus kontzeptuaz aitzin-ohar labur bat

Lehenbiziko hizkuntza corpusak aspaldikoak ditugu, franko goizik, filologoek eta filosofoek, hizkuntzen arteko konparantzen egin ahal izateko, corpus konparagarriak moldatu izan baizituzten, iturri bereko testua mintzaira askotan emanaraziz. Horrela *gure aita* otoitza anitz hizkuntzatara itzularazia izan zen 16. mendean, eta geroago Haur prodigoaren parabolarekin gauza bera egina izan zen, adibidez, Coquebert de Montbreten inkestetan, hemeretzigarren mendearen hastapenean. Dakigun bezala, Bonapartek ere baliatu zituen gisa horretako corpus konparagarriak euskal dialektologiaren oinarrien finkatzeko. Corpus labur-laburrak ziren horiek, eta oso erabilera mugatua zuten, konparaketa lanen laguntzeko bereziki baliatzen baitziren. Gogoan izan halere, haietarik batzuk, hala nola Bourciez-ek Iparraldean egin bilduma edo Sacazek egina, ez direla oraino argitaratuak izan.

Hogeigarren mendean, ordea, hizkuntzalaritzako ikerbideek sinkroniari bereziki lehentasuna eman ziotelarik, corpusek beste manera batean hartu zuten garrantzi berezia, hizkuntzalari estrukturalisten arabera hizkuntza estudioen oinarri enpirikoa haietan (eta, teorian behintzat, haietan bakarrik) finkatu behar baitzen ikerketa. Horrela egina izan zen, adibidez, euskal mintzamolde bat sinkronian ikertzen zuen lehenbiziko doktoretzako tesia (G. N'Diayerena), 60ko hamarkadaren hondarrean.

Azken mende laurdenean ere, are funskiago aldatu dira hizkuntza ikerketen oinarriak eta ikerbideak, eta gaur egun hizkuntza corpusez hitz egiten

delarik corpus kontzeptua ez ohi da gehiago lehen egiten zen maneran aditzen eta erabiltzen, beste zerbaiten adierazteko baizik.

Hizkuntza corpusen oinarrizko definizioa, halere, ez da aldatu: hizkuntza baten deskribatzeko eta ikertzeko baliatzen den hizkuntza-datu bilduma da corpusa, gaur lehen bezala. Ordea, zabalegi da mugatze hau, egun hizkuntza corpusen hitz egiten delarik, hitzak hartzen duen adiera behar bezala jasotzeko. Izan ere, **baliabide elektronikoak erabiltzen eta eskaintzen dituen hizkuntza-datu bildumak** izendatzeko bereziki erabiltzen da *corpus* hitza, orain corpus hizkuntzalaritzaren testuinguruan erabiltzen denean, eta horrela ulertuko dugu hemendik aitzineko lerroetan ere.

1.2. Corpus mota desberdinak

Corpus hitza horrela ulerturik ere, askotan arras eduki desberdineko datu bildumak izenda ditzake. Corpus horiek nolako datuak biltzen dituzten, eta zer helburutan eratzen diren.

Helburuak biga izan daitezke:

– testuen beren hartan biltzea eta eskaintzea (ingelesez, *artxibo* hitza erabiltzen da askotan mota horretako corpusen izendatzeko, eta argiago izateko guk ere hala eginen dugu txosten honetan);

– testu azterketa automatikoen bideratzea, tamaina handi-handiko testuen ikertu ahal izateko (gaur egun ehun milioi hitz baino gehiago dituzte ingeles, frantses edo gaztelerazko corpus nagusiek).

Edukiak ere mota desberdinetakoak izan daitezke, garranzizkoenak hauek direlarik:

– literaturako obrak (idatzizkoak, ahozkoak, ikus-entzunezkoak);

– testu idatziak (orotarikoak, edo sail ala gai batzuetan murriztuak);

– ahozko testu grabatuak (orotarikoak, edo sailka antolatuak: elkarrizketak, hitzaldiak, berriak);

– hizkuntza datu base bereziak: hiztegiak, lexikoi bereziak, onomastika zerrendak, galdakizunen erantzunak, ...

Irizpide hauek baliaturik lau corpus mota nagusi bereizten ahal dira:

Testu artxiboak: Testu osoak, berezko interesa dutenak, biltzen dituzte testu artxiboek. Idatzizkoak edo grabaziozkoak izan daitezke; (adibidez literaturako obrak, errefrau bildumak, bertsu txapelketen grabaketak, ipuin kontaketak, etab.).

Testu artxiboetan testuak bere hartan eta bere osoan du balioa, eta beraz artxibaketa haren formato linguistikoa aldatu gabe, edo izatekotan filologia klasikoaren irizpideen arabera egokiturik, eskaintzen da. Bistan dena, testu artxibo horretaz baliaturik beste bertsio bat egin daiteke, testuaren era linguistikoa testu corpus gisa baliatzeko gisan moldaturik, baina orduan beste corpus bat eratzen da. Hori gertatu da (partez behintzat) OEHaren testu corpusarekin: testuen grafia aldatua izan da, hitzen bilaketa automatikoen bideratzeko ezinbestekoa baitzen grafia kontuan batasun zerbait izatea (osoa ez izanik ere). Ondorioa, ordea, garbia da, OEHko testuak ezin balia daitezke gure literatura zaharraren testu artxibo gisa. Erran gabe doa, haatik, euskara batuan idatzi testuen kasuan distantzia hori kentzen dela (nahiz ez osoki ezabatzen ere, adibidez, testu baten bertsio desberdinak, edo berdin erdarazko itzulpenak, atxiki baitaitezke testu artxibo batean, horrek ez baitu interes bera testu corpus batean).

Testu corpusak: Helburu linguistiko batekin osatzen diren bildumak dira testu corpusak, grafia bateratua duten testu gordinak biltzen diztuztenak.

Kontzeptu hau ez da ondokoarekin (hizkuntza corpusak) nahasi behar, nahiz hori egiten den askotan, zeren hizkuntza corpus guztiek testu corpus bat baitute oinarrian. Diferentzia halere argia da: testu corpusak azterkatu ez diren corpusak dira, eta hizkuntza corpusak, berriz, testu corpus linguistikoki azterkatuak (lematizatuak gutienenetik). Diferentzia horrek egiten du denboran, diruan, eta tamainan baldintza desberdinak baitituzte bi corpus mota horiek.

Azterkatuak ez izanik ere, testu corpusak ez dira beti aldatugabeak, grafia bateratua izan behar baitute. Izan ere, nahitaezkoa da testu bat horrelako corpus batean sarrarazi aitzin zer grafia mota darabilen ikustea, eta beharrez, corpusean baliagarri izateko gisan haren grafia aldatzea. Testu aztertugabeak izanik, ez pentsa informazio guti dakartela horrelako corpusek. Hizkuntzaren beraren izaeragatik (euskaraz hizka idazten da eta funtzio morfologia ere franko erregularra da), gauza asko eta asko atera daitezke testu corpus hutsetarik, haien eratzea franko merke eta aise delarik: OEHaren testu corpusa, adibidez, testu corpus hutsa da.

Hizkuntza corpusak: Hizkuntza corpusak testu aztertuak (lemak, etiketa morfosintaktikoak, lokuzioak, …) eskaintzen dituzten corpusak dira, besteak beste azterketa estatistikoak bideratzeko egoki direnak. Hizkuntza corpusaren testu oinarria helburu baten arabera antolatua eta orekatua delarik, erreferentzia corpus gisa balia daiteke. Testu idatziak eta testu grabatuak izan ditzakete oinarrian.

Hizkuntza corpusak testu corpus landuak –erran nahi baitu linguistikoki azterkatuak– dira. Hizkuntza corpusak ez dira denak itxura be-

rekoak, eta testu corpus batetik hizkuntza corpus desberdinak sor daitezke. Hiru parametro bereziki kontuan hartzeko dira (hiruen artean lotura badelarik noski):

- azterketaren barnatasun maila;
- azteketaren automatizazio maila;
- testu oinarriaren nolakotasuna (eta tamaina).

Azterketaren barnatasuna: testuari egiten zaion azterketa aitzinatua izan daiteke edo ez. Adibidez, corpus bat lematizatua izan daiteke, baina denbora berean lexemen kategoriak eta homonimia kontuak batere argitu gabe utz ditzake, edo alderantziz gisa horretako informazioa eman dezake. Gaur egun, halere, hizkuntza corpus bateko hitz guztiak morfosintaktikoki etiketatuak izatea eskatzen ohi da, askotan horretarako hitz edo morfemaren ingurumena ere kontuan hartzea beharrezkoa delarik.

Azterketaren automatizazio maila: testu corpusa aski handia denean, nahitaezkoa da azterkatzaile automatizatu baten erabiltzea hizkuntza corpusaren eratzeko. Azterketa automatikoa beti hobekitzen den tresna batekin egiten da: gauzak behar bezala eginez gero, denbora joan arau, testu azterkatzaileek gero eta testu luzeagoak, gero eta lasterrago, gero eta huts gutiago eginez, gero eta sakonkiago, azterkatzen dituzte. Halere, nolako hizkuntza corpusa eratu nahi den, eta batzuetan datu mota batzuen kasuan bederen, azterkatzaile automatikoaren ondotik eskuz egiaztatzea komeni da hutsik gabeko corpusa sortu ahal izateko. Erran gabe doa, azterketa automatizatua hobekitzen den arau, baztertzen direla horrelako prozedurak, ezen denbora gehiago eskatzen dute eta kostu handiagoa ere badute gisa horretako corpusek. Ondorioz, ezaugarri gutiago kontuan harturik, tamaina tikiagoko testu oinarriekin lan egitera behartzen dute.

Testu oinarriaren nolakotasuna (eta tamaina): hizkuntza corpusen testu-oinarria helburuen arabera modu desberdinetan antola daiteke: garaiak, dialektoak, testu motak, gaiak, testuen zabalkundea, testuen kalitatea, testuen arrakasta, edo bestelako irizpideak gogoan edukiz. Askotan orori begia atxikitzen zaie, nork zer egin nahi duen. Helburu baten arabera ordezkaritasun maila gora duten testu oinarrien bidez, ***erreferentzia corpusak*** deitzen dira. Kontuz! Erreferentzia corpusen zer-egina ez dagokio egokitasunari baina erabilera hutsari, hots, ez da hartan agertzen nola erran behar den (ikuspegi normatibo edo eredu emaile batekin), baina nola eta zer maiztasunekin agertzen diren delako corpusean hizkuntza datuak. Ereduzkotasuna, izatekotz, **corpusak duen testu oinarritik, erran nahi baitu haren**

eratzeko baliatu diren irizpideetarik, sor daiteke, ez beste ezertarik.

Hizkuntza datu-base bereziak: Hizkuntza atalen arabera egituratuak (hiztegia, morfologia, joskera).

Hizkuntza datu-base bereziek hizkuntza atalen araberako oinarria dute. Ez dituzte, beraz, aztertu gabeko testu jarraikiak eskaintzen. Adibidez, hiztegi elektronikoak halakoak dira, baita adiztegiak, edo onomastikako datu baseak ere.

Corpus hitza darabilgunean, guztiz garrantzizkoa da zertaz hitz egiten dugun ongi jakitea, nahiz azpimarratzekoa den, bestalde, agerian eman dugun sailkapeneko osagaiek ez dutela elkar kanporatzen, datu bilduma bera modu batean baino gehiagotan balia baitaiteke. Eman dezagun OEHaren literatura testuen corpusa testu artxiboa izan daiteke gisa horretan antolatuz gero; baina testu corpus gisa ere balia daiteke, testuaren grafiazko bateraketa segurtatu ondoan, bai eta hizkuntza corpus gisa ere, testu corpus hori lematizatzen bada, eta aztertzaile automatiko batetik iraganarazten bada.

2. EUSKALTZAINDIAK ORAIN ARTE ERATU DITUEN CORPUSAK

Lehen errana dugun bezala, Euskaltzaindiak badu jadanik bide puska bat eginik corpusgintzan, eta oker legoke orain arte deus egin izan ez dela uste lukeena. Eginak izan direnak, gainera, guztiz garrantzi handikoak izan dira Euskaltzaindiaren lanetan, eta ezin hobeki erakustera eman dute zein baliosak diren gisa horretako baliabideak.

Kontuan hartzekoa da, bestalde, corpus horiek izan dezaketen euskarri mota: corpus mota batzuk on-line dira, beste batzuk CD-ROMetan, batzuk bi euskarrietan izan daitezkeelarik.

2.1. Testu artxiboak

Euskaltzaindiak orain arte ez du testu artxiborik eratzeko politikarik izan, eta ez da harritzekoa gibelamendu zerbait baitu alor honetan. Euskal literatura urria izanik pentsa zitekeen sail honetan egin zitezkeela aurreramendurik handienak, baina ez da horrela gertatu, dudarik batere gabe edizio kritikoetan euskaldunek dugun atzerapenaren ondorioa baita egoera hori. On-line diren edizio elektronikoetan mugatzen bagara, gaur egun, euskalgintzan aurkitzen ditugun mota horretako materialak jende batzuen iniziatibari esker sortuak izan dira bereziki. Euskaltzaindiak, aldiz, gauza gutxi eskaintzen du, bi lan bakarrik aipa baitaitezke, alde batetik, Frai Bartolomeren obra osoaren

edizio kritikoa, eta atlas proiektuari dagokion euskal mintzamoldeen antologia, bestetik. Bestelako euskarrietan diren artxiboak kontuan hartuz gero, *Euskera* aldizkariaren CD-ROMa aipa daiteke, bai eta Euskararen Herri Hizkeren Atlaseko soinu-artxiboa ere, nahiz hau bibliotekan baizik entzun ez daitekeen. Deus gutxi bestela testu artxiboei dagokienik. Gure literatura obra nagusiak ere ez dira eskaintzen: ez Etxeparerena, ez Axularrena, ez *Peru Abarka* edo *Buruxkak*.

2.2. Testu corpusak

Euskaltzaindiak testu corpusak eratu ditu, nahiz, bere batzordeetan erabiltzeko sortuak izanik, ez diren behar eta merezi bezainbat zabaldu ahal izan. OEHaren testu-corpusarekin moldatua izan den CD-ROMa dugu hemen bereziki aipagai. 300 bat dokumentu biltzen dituen testu corpusa da, hizkako, hitz zatikako eta hitz multzokako bilaketen egiteko aukera eskaintzen duena. Oso tresna zerbitzu egilea da erabiltzen dakienarentzat. Halere, akatsak ere baditu, bereziki zenbakitu zen testua ez baitzen zuzendua izan. Grafiaren bateratzeari dagokion aldaketak bazterrean utzirik ere, hastapeneko akatsekin gelditu da, beraz, CD-ROMean dagoen testua. Baina dudarik batere gabe testu corpus honen mugarik handiena zabalkundeari dagokio, arrazoi juridikoengatik ez baita on-line jartzen ahal (duen formatoan bederen) eta Euskaltzaindiaren lanetan ari direnen artean, edo ikerketarako behar berezia adierazi duten ikertzaileen artean baizik ez da zabaldu.

Beste testu-corpus baten asmoa agertu izan da azken urteetan, *Lamiategi* deitua. Ideia zen OEHaren esperientziaz eta ingelesezko *Oxford Text Archive* delakoaren ereduaz baliatuz biltze prozedura bat, idazleekin, argitaletxeekin eta itzultzaileekin adostua, sortzea, egungo euskararen corpusaren eratzen hasteko (eta CD-ROMen bidez zabaltzeko). Zorigaitzez orain arte ez da obratu ahal izan asmo hau, eta denbora anitz galdu da. Ikustekoa da orain, Euskal idazleen elkarteko eta Euskal itzultzaileen elkarteko buruekin eta legelariekin adostua izan zena modu berean obratzen ahalko den.

2.3. Hizkuntza corpusak

Euskaltzaindiaren hizkuntza corpus bakarra XX mendeko euskararen corpus estatistikoa da (ex-EEBS). UZEIk Euskaltzaindiarentzat moldatu duen corpus honen ezaugarriak ezagunak dira: testu zati laburrez osatua da, zatiak modu aleatorioan testu moten araberako proportzio batzuk errespetatuz aukeratuak izanik. Erreferentzia corpusen ereduari darraio beraz corpus hau.

Eskaintzen dituen hizkuntza datuak lexikografiari dagozkio ororen gainetik, ez baita harritzekoa hiztegi batuaren lanen aitzinatzeko eratua izan baitzen. Hortaz, lematizazioari dagozkionez kanpo ez da aurkitzen hartan beste hizkuntza daturik, ez lemen kategoriarik, ez datu morfosintaktikorik.

Corpus honen beste ezaugarria da haren fidagarritasun maila, guztiz altua dena. Fidagarritasun hori, azterketa automatikoaren ondotik, datuak berriz 'eskuz' azterketaturik lortu da.

Euskaltzaindiaren WEB gunean on line kontsulta daiteke euskararen corpus hau. Ez, ordea, CD-ROMetan.

2.4. Hizkuntza datu-base bereziak

Ororen buru gaur egun hizkuntza datu-base bereziak ditu gehienik eskaintzen Eukaltzaindiak: arauei dagozkien corpus laburrak (hiztegi batua, onomastika zerrendak, jagonet datu-basea, ...) Ikus daitekeenaz, Euskaltzaindiak euskararen normagintzan sortzen dituen ekoizpenak aurkitzen dira bereziki hor gaur egun.

Modu horretako datu-baseek garrantzi handia dute, Euskaltzaindiaren lanen ezagutarazteko bereziki: ikerketen ondorio gisa agertzen dira, beraz, ez ikerketen egiteko tresna gisa. Nola gaineratiko hizkuntza corpusek ez bezalako tratamendua eskatzen baitute, ez dira gehiago aipatuko horrelako datu baseak txosten honetan.

3. HIZKUNTZA-CORPUSEI BURUZKO JOKABIDEA

Egoeraren irudia marraztatu ondoan, zer proposamen mota egin daiteke Euskaltzaindiak hizkuntza corpusak direla-eta segitu behar lukeen politikaz?

Hiru partetan zatitzen da proposamena:

– Euskal testuen biltegia;

– Euskaltzaindiaren hizkuntza corpusa;

– Euskararen corpus orokorra.

3.1. Euskal testuen biltegia

Hizkuntza corpusen eratzeko lehenbizikorik testuak bildu beharrak dira eta formato bateratu batean eman haien kanpoko ezaugarri nagusiak zehaztuz (tamaina, garaia, testu mota, egilea, ...). Euskara bezalako hizkuntza baten ka-

suan, lan horren egiteko mugak ez datoz tamaina handiko biltegiak eratzeko ezintasunetik, ezen milioika eta milioika hitzetako corpusak baliatzen dira gaur egun, eta badakigu ondoko urteetan are handiago eginen direla biltegi elektronikoen ahalak. Buruhausteak, horrelako biltegi baten antolatzeko orduan agertzen dira: nola joka euskaraz egiten diren testuak behar den forman biltegi batera igorriak izan daitezen eta corpus batean baliatuak izateko baldintza juridikoak aldez aurretik gaindituak izan daitezen?

Ezen erran gabe doa testu horiek erosi behar balira edo, egileek urririk utzirik ere, ekoizleen laguntzarik gabe jaso behar balira (banaka berriz idatzi edo eskanerrizatu behar balira), ezinezkoa litzatekeela horrelako biltegi zabal orokorrik moldatzea. Haste-hastetik muga aski hertsiak jarri behar litzaizkioke. Garestiegi litzateke bestela, bai dirutan, baita denboran ere. Hitz batean, euskal testuen sortzaileen eta argitaratzaileen onespen eta lankidetzarik gabe ezin era daiteke horrelako biltegi zabalik, kontuan izanik ez dela hau hemendik zenbait urte bukatuko den proiektu bat, baina bururatze jakinik ezin izan dezakeen bat (piska bat biblioteken moduan).

Hemen guztiz garrantzizkoa da Euskaltzaindiaren izaera berezia: erakunde guti dira Euskal Herrian, euskalgintzan diharduten jende gehienen partetik onespen maila bera duenik horrelako lan bati fidagarritasun maila gutieneko batekin lotzeko. Horregatik, hain zuzen, iduri zait eni, Euskaltzaindiak behar lukeela bere gain hartu horrelako biltegi baten antolatzeko erantzunkizuna. Are gehiago, eginbidea duela erran nezake. Ideia horretan asmatu zen duela bizpahiru urte Lamiategi proiektua, eta orduan ikusi ahal izan zen oso harrera ona egin ziotela literatura munduko partaide nagusiek: idazleen elkarteak, itzultzaileen elkarteak, argitaletxeen elkarteak.

Duda egin dezake batek gure erakundeak horretarako ahalik ba ote duen egiazki, eta horrelako erantzunkizuna bere gain har dezakeen orain ondoko urteetako. Galdea arrazoizkoa da, baina beste aukera handirik ere ez da, ez baita aise ikusten nor besterik izan daitekeen: EHU, menturaz, nahiz ez dirudien, ahal handiagoak izanarren, toki anitzez hobean den hura proiektuari orokortasun eta iraunkortasun bera emateko.

Kontuan izan horrelako biltegi zabal komun bat eratzen ez bada euskalgintzako partaideak zein bere aldetik ariko direla beren biltegien moldatzen, hizkuntza corpusen egin ahal izateko. Hasiak ere omen dira eskuin eta ezker bide horri lotzen, eta horretarako diru laguntzen eskatzen.

Testu corpusaren ezaugarriak.

Horrelako testu corpus baten ezaugarriak zein izan litezke? Hiru puntutan bil daitezke ezaugarri nagusi horiek:

- Euskalgintzako partaideen lankidetzari eta borondate onari esker sortua izanik ez luke helburu eta izaera komertzialik izan behar;

- Corpus publikoa izan behar litzateke, erran nahi baitu corpus eratzaile guztiek, erabilera baldintzak betetzen dituzten ber, hartara jotzeko ahala behar luketela beren corpus oinarriak berek nahi bezala eratzeko;
- Corpus irekia izan behar litzateke, irekitasuna bi alderditarik ulertuz: irekia edozein euskal testu, baldintza teknikoak beteak diren ber, hartzeko; eta irekia denboran ere, urteak joan arau haziz joateko.

Testu biltegi honen eratzeko Euskaltzaindiaren barnean ere egitura berezia sortu behar litzateke (testu biltegiaren batzorde behin-behinekoa), hastapenean, bereziki, egitasmo honen abiarazteko, baita *lamiategi* proiektua ere obratzeko. Ahozko corpusen arazoaz ere gogoeta egin behar litzateke eta alde horretarik komeni litzateke dialektologia batzordeko buruak (edo ordezkari batek) hartan parte hartzea. Batzorde aski arina behar litzateke, Iker sailburua, Dialektologia batzordeko burua edo ordezkaria, Azkue Bibliotekako idazkaria eta informatikaria bil litzakeena. Gauzak finkatu ondoan, urrunago so eginez zer egitura mota sortu behar litzatekeen eta zer etekin mota sor ahal litekeen honetarik, gero ikustekoa litzateke, baita kanpoko erakundeekin izan daitezkeen egiturek nolakoak izan behar luketen ere; ikus horretaz ondoko puntua ere).

3.2. Euskaltzaindiaren hizkuntza corpusa

Gorago aipatu *euskal testuen biltegia*, izenak dioen bezala, biltegia da. Biltegi irekia, gainera. Hari zuzenean probetxu atera daiteke, adibidez, hizkako datuak zuzenean eskain litzaketen azpi-corpusak sortuz; beharrez, CD-RO-Metan emanik. Ez da haatik, hizkuntza corpusa, eta ez lematizaziorik, ez hitzen etiketatzerik ez da agertuko hango testuetan. Horrelako hizkuntza corpusen eratzeko beste lan bat egitekoa da, informatikari dagokiona ororen gainetik, eta nahitaez Euskaltzainditik kanpo eginaraztekoa dena; hots, beste corpus bat sortu beharra da: hizkuntza corpusa.

Euskaltzaindiak bere lanetako, eta bereziki hiztegi batuko lanetarako hizkuntza corpus berezien eratzea beharrezkotzat jotzen badugu, argi da hizkuntza corpus horien testu oinarria Euskaltzaindiak emanikako irizpideen arabera finkatzekoa dela. ***Euskaltzaindiaren hizkuntza corpusa*** sor liteke orduan, aipatu irizpideen arabera moldatua, eta berak bere lanen egiteko eraikia.

Guztiz garrantzizkoa da ulertzea euskal testuen biltegia eta Euskaltzaindiaren hizkuntza corpusa guztiz desberdinak direla, nahiz bigarrenak lehenbizikoaren parte bat oinarrian duen.

Hona diferentziarik nabarmenenak bien artean:

<u>Euskal testuen biltegia</u>:

- formazko baldintzak bete dituen edozein euskal testu onartzen du;

- irekia da;
- ez da lematizatua, ez etiketatua;
- ez daiteke estatistikarik egin (hitz zenbaketak bakarrik);
- berehala irakur daiteke nahi den aukera eginik testuen artean;
- formazko bateraketaz kanpo ez du lanik eskatzen (ahozko testuez kanpo);
- Euskaltzaindiak antola eta egin dezake.

Euskaltzaindiaren hizkuntza corpusa:

- testu oinarria eraikia da testu moten, garaien, euskalkien, etabarren arabera;
- irekitasuna baldintzatua da;
- lematizatua da;
- etiketatua da;
- azterkatzaile automatiko batek azterkatu ondoan eratzen da;
- garbiketa lana beharrezkoa izan daiteke (automatizatzearen ondorioen arabera);
- estatistikak egin daitezke;
- Euskaltzaindiak ezin egin dezake, baina bere beharren arabera muga dezake.

Funtsean beraz, OEHko testu corpusa eta EEBSen artean dauden aldeak aurkitzen ditugu hemen ere, areagoturik. Garbi da, bi corpus mota horiek ez direla elkarren aurkakoak, baina elkarren osagarri, eta beraz Euskaltzaindiak biei ateak ireki behar dizkiela.

3.3. Euskaltzaindiaren hizkuntza corpusa eta euskararen corpus orokorra

Euskaltzaindiak hizkuntza-corpus baten beharra badu. Beharrezkoa da jakin dezan bere normak ematen dituenean zer den egoera.

Ordea, hizkuntza corpusen eragina Euskaltzaindiak egiten dituen lanetarik kanpo ere agertzen da, eta geroan ere gero eta gehiago agertuko da. Duela hamar urte Euskaltzaindiaz kanpo inor guti zen corpus kontu horietan zebilenik. Gauzak, zorionez, aldatu dira eta euskalaritzako partaideetan gero eta gehiago dira horretan dabiltzanak. Burugabekeria litzateke, euskalgintzako

partaide desberdinak, zein bere aldetik eta zein bere testuekin, hasten balira, beren hizkuntza corpusaren sortzeko azterkatzaile automatikoen eratzen, eta lematizatze eta etiketatze programen moldatzen. Nekez gainera horrelakorik egin daiteke, kostu handi-handia baitu holako lan batek, eta baitakigu diru-iturriak aski murritzak direla. Aterabide bakarra da, beraz, hizkuntza corpus nagusi baten eratzea euskalaritzako partaide guzien zerbitzuan jar litekeena.

Hizkuntza corpus horren testu oinarria Euskaltzaindiak bere lanetarako behar dituena baino zabalagoa litzateke nahitaez, eta ez lituzke testu oinarriaren hautatzerakoan oreka arazo berak. Argi da, adibidez, testuen azterketa automatikoa lantzen dutenentzat (adibidez, azterketa automatikorako tresnak egiten dituztenentzat), corpusen tamainak garrantzi handia duela, haiek Euskaltzaindiarentzat baliorik ez duten datu asko eta asko landu beharrak baitituzte. Emaitzen garbitasunari dagokionez ere, askotan behintzat, haiek ez dituzte Euskaltzaindiak izan ditzakeen kezka berak. Ez luke zentzurik Euskaltzaindiak bere irizpideak holako lanetan dabiltzainei inposatzen balizkie. Eta alderantziz ere berdin. Malgutasunik gabe, nekez sortuko da horrelako proiektu zabal batean hain beharrezkoa den lankidetza.

Nork nola antola horrelako corpusa? Ez da erraz erantzutea. Argi da Euskaltzaindiak ezin bere gain har dezakeela haren obratzea, parterik handienean informatikako lana baita hor egitekoa dena, guztiz puntakoa gainera, eta Euskaltzaindiak ez baitu horretan gaitasunik. Bestalde, euskalgintzako beste partaideen beharrak eta eskaera bereziak (informatikariena, hizkuntzalariena, hiztegileena, euskararen teknikariena, ...) kontuan hartzera behartua litzateke, aski lan duelarik, bizkitartean, eskuen artean dituen egitekoei buru egiteko.

Badirudi euskalgintzako partaide desberdinen beharrak bete ditzakeen hizkuntza corpus zabal honen eratzeko *consortium* gisako zerbait sortu behar litzatekeela. Lankidetza eremu zabal hori sortzen balitz, haren erdigunean egon behar luke Euskaltzaindiak, besteak beste, testuen biltegiaren antolatzaile gisa testu ekarle handia litzatekeelako eta, bere corpusa ere hortik aterako lukeelako.

4. EUSKALTZAINDIAREN TESTU ARTXIBO ELEKTRONIKOA

Aurreko atalean corpusak, testu corpusa eta hizkuntza corpusak, aipatu ditugu, haiek baitute antolamenduaren aldetik buruhauste gehienik ekartzen eta haien baitute geroari buruz presarik handiena.

Horiek horrela Euskaltzaindiaren ohiko lanak ere ezin ahantz ditzakegu, eta haietan sartzen da euskal testuen altxorraren eratzea. Ideia ez da berria, aspaldi hasia baitu Euskaltzaindiak bere argitalpenetan *Euskararen lekukoak* deitu saila. Hasi bai, baina nekez segitu, liburu guti atera baitugu azken urteetan. Badakigu zergatik: langile eta diru eskasez, azken arrazoi honek argitalpen horiek sortzen duten beste buruhauste batera garamatzalarik: salbuespen bat edo

beste kendurik, liburu horiek ez dira batere saltzen, ez zabaltzen, eta finantziamendu berezia aurkitzen ez bazaie, ezin argitara eman daitezke.

Artxibo elektronikoek horrelako testuen argitaratzeko baldintzak osoki aldarazten dituzte, eta uste dut Euskaltzaindiak berriz ikusi behar duela euskal testu klasikoen argitalepenen kontua, bai edizio lanaren egiteko moduari dagokionez, bai argiraratzeko moduari dagokionez.

Partez, gaur egun dugun egoera antolamenduaren ondorioa ere da: azken urteetan argitalpen batzordearen gain utzia izan da sail horretako ardura, hots, praktikan, zuzendaritzaren gain. Eta esperientziak erakustera ematen digu zuzendaritza ez dela horretaz behar bezala arduratzeko baldintzetan. Beraz, kontu hau egiazki kudeatua izan gabe eramaten da, autoreen urtemugen eta jendeen gogoaren arabera, eta editatzeko prozedura jakinik finkatu gabe.

Bizkitartean hondar urteetan euskalaritzak aitzinamendu handiak egin ditu sail horretan eta Euskararen lekukoen sailean atera diren azken argitalpenetan ikusi da hori. Horra zergatik iduritzen zaidan Euskaltzaindiak beste bide bat ireki beharko lukeela euskal testuen artxibo elektronikoa sortzeko.

Nire ustez, behin behineko batzorde teknikoa sortu behar litzateke euskal testuen artxibo elektronikoa nola antola daitekeen aztertzeko. Batzorde honetan, Literatura batzordeko burua (edo ordezkaria), Hiztegi lantaldeko burua (edo ordezkaria), eta gaur egun lan horietan adituak diren filologo batzuk, Euskaltzaindiaren gisa horretako argitalpenetan lanean azken urteetan ari izan direnak bil litezke, ... Batzorde honen egitekoa Euskaltzaindiari proposamen baten aurkeztea litzateke, hartan:

– artxibo elektronikoaren edukia eta egitura mugatuz;

– editatzeko prozedura bat finkatuz;

– argitalpen programa bat aurrikusiz;

– antolamendurako egitura bat proposatuz.