Corpusgintzaren garrantzia hizkuntzalaritzan eta euskararen egoera

Miriam Urkia
UZEI

Aurkezpen honen helburua ez da corpusgintzari buruzko sarrera teoriko bat egitea, literatura ugari dago horren inguruan. Corpusgintzaren beharra zalantzan jarri gabe, hau da, gaur egun ezinbestekoa dela onartuta, bere garrantzia azaltzen saiatuko naiz eta euskara zein puntutan dagoen erakustera ausartuko naiz. Corpus idatzi elebakarretara mugatuko naiz batez ere, paraleloak beste aurkezpen batean azalduko baitira. Ahozkoa ere axaletik bakarrik aipatuko dut. Horretarako corpusen eta corpusgintzaren ikuspegi azkar bat emango dut, azalekoa ezinbestean, denbora-mugagatik batetik, baina bereziki badakidalako ezagutzen ez dudan hainbat corpus eta corpusgintzarako tresna izango dela, txikiak, pribatuak eta ezagutzera eman ez direnak. Gaur egun euskara lan-tresna duten askok baliatzen dutelako corpusa, baina halaxe, lan-tresna gisa bakarrik, ez produktu edo gizarteratzeko baliabide gisa. Hala ere, erakusgarrienak, erabilienak, ezagunenak, orokorrenak.... horiek bilduko dituen argazkia egiten saiatuko naiz nondik gatozen eta non gauden kokatzeko eta, ahal den neurrian, guztion artean nondik jo asmatzen laguntzeko.

1. Corpusak eta corpusgintza

Sinonimo gisa erabiltzen ditugu askotan bi kontzeptuok, baina badago argitu beharreko zerbait. Corpusa, adiera modernoan, baliabidea da, zehatzago: hizkuntzaren atal baten **erakusgarri** gisa erabiltzen den **testu-multzo elektroniko egituratua**, erabilera errealak jasotzen dituena, betiere irizpide zehatz batzuen arabera. Hala definitu dute Sinclair-ek eta McEnery eta Wilsonek:

> *"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).
>
> "*A finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration*" (McEnery & Wilson, 2001).

Corpusgintza hori baino zabalagoa da, corpusaren osaeran parte hartzen duen oro hartzen baitu bere baitan: corpusak eratzeko eta ustiatzeko metodologia, baliabideak, tresnak eta corpusak berak. Horren inguruan arituko naiz ondoko orriotan.

Bada, baina, corpusgintza dela-eta argitu beharreko kontu pare bat.

Bat, corpusgintza azken urteotan askoren ahotan baitabil eta, maiz, hizkuntzalaritzaren adar berria ote den eztabaidatu baita, gramatika, semantika eta horien pareko. Horiek guztiek hizkuntzaren atal batzuek deskribatzen dituzte batez ere. Corpusgintza, berriz, adar guztietan balia daitekeen **metodologia** bat da, adar bakarrera mugatu gabe. Eta bi, horiek guztiak esanda ere, ez al da "moda" kontu huts bat, halako batean pasatuko dena? Badirudi orain aurreko gauza asko ukatzen dugula corpusek kontrakoa diotela argudiatuta. Eta egia da neurri batean, baina honek ez du esan nahi aurreko guztia okerra edo faltsua zenik, nahiz hainbat baieztapen egin izan diren historian zehar egia absolututzat hartu ditugunak, besteak beste eskura genituen datuekin egia zirelako,

baina datuekin osatu den neurrian bideratu dira argudio eta “egia” berriak ere. Lehen datuak teoriak indartzeko erabiltzen ziren bezala, gaur teoria erabileran oinarritzen da, datu enpirikoetatik abiatzen dira ikertzaileak teoriak eratzeko, eta corpusek eskaintzen dituzte datu enpiriko horiek. Hortik corpus handiak, ondo eratuak eta egituratuak izatearen garrantzia, datu horietatik sortuko baitira “egia” berriak.

2. Corpusgintzaren garrantzia: ikuspegi azkarra

2.1. Corpusgintzak lan egiteko modua aldatu du

Corpusek, beraz, lan egiteko modua aldatu dute: datu enpirikoetan oinarritzen garela esan dugu, baina ez da hori izan aldaketa bakarra. Hizkuntzaren teorizazioan ere eragina izan du lan-prozedura berri honek, corpusgintzak egituratzea, formalizatzea eta sistematizatzea behar baititu, eta honek tresna berriak garatzea ekarri du: etiketatzaileak, lematizatzaileak eta gramatika-formalismoak, besteak beste. Alegia, honek guztiak orain arteko lan deskriptibo asko berrikusi beharra ekarri du, hutsuneak agerian utzi baititu.
Hemendik abiatuta, corpusgintzaren garrantzia hiru galderekin erantzuten saiatuko gara.

1) *ZER jasotzen du corpus batek?*

Gorago aipatutako “egia” berriak lortzeko testu-masa handiak (edo horien transkripzioak, ahozkoen kasuan) behar dira, baina hizkuntzaren erakusgarri izateko antolatuak. Helburuen arabera, mota desberdinetako corpusak osatuko dira:

a) idatziak, ahozkoak edo bietarikoak
b) elebakarrak, elebidunak, eleaniztunak
c) diakronikoak / sinkronikoak
d) irekiak / itxiak
e) erreferentziazkoak, monitoreak, estatistikoak, paraleloak, konparatuak, espezialitatekoak, esperimentalak, literarioak, etab.

Bakoitzak bere egitekoa du, bere testuinguruan kokatu behar da, kasuan kasuko beharren arabera. Hala ere, edukiaren kalitateak, orekak eta egituratzeak baldintzatuko dute emaitza.

2) *NOLA jasotzen du informazioa corpus batek? Zeintzuk dira erabiltzen dituen baliabideak eta tresnak?*

a) Etiketatzaileak: biltzen den informazioa etiketatu egin behar da. Batetik, goiburukoek informazio bibliografikoa eta landutako sailkapena bilduko dute, dokumentua identifikatu eta ondoko bilaketak errazteko. Bestetik, testuko forma guztiak linguistikoki etiketatuko dira, edukia bera ustiatu ahal izateko.
b) analizatzaileak/lematizatzaileak: formak linguistikoki etiketatzeko erabiliko dira bi tresna hauek, bakoitzak, gutxienez, informazio morfosintaktikoa eskaini beharko baitu (lema, kategoria, azpikategoria, kasua, aditz-jokoa, etab.).
c) desanbiguatzaileak: informazio linguistikoa bere testuinguruan kokatu eta leku horretan dagokion analisi-aukera bakarra utziko da. Hiru modutara egin daiteke hau: ezagutza linguistikoa erabilita, eredu estatistikoak baliatuta, edo aurreko biak konbinatuko dituzten eredu hibridoetara jota.
d) ahotsaren transkripziorako tresnak: ahozko corpusetan, testuak transkribatu eta, aurreko informazioaz gain, fonemak ere zehaztuko dira, behar den bestelako informazioaz gain.

e) testu-parekatzaileak: corpus elebidunetan edo eleaniztunetan, aurrena parekatu egingo dira paragrafoak, esaldiak, eta ondoren aurreko guztia aplikatuko da hizkuntza bakoitzean.
f) Eta, guztiaren oinarrian, baliabideok: datu-base lexikalak, gramatika-formalismoak eta ezagutza-baseak landu beharko dira analizatzaileek-eta oinarri sendoa izan dezaten.

3) *ZERTARAKO behar da corpusa?*

Bi erabilera eta erabiltzaile mota nagusi ditu corpusak: hizkuntzaren erabilera aztertzeko (hizkuntzalariak, ikertzaileak, ikasle/irakasleak, itzultzaileak, hizkuntza aztergai duen edonork, azkenean) eta proba-banku gisa erabiltzeko (informatikariak batez ere). Hala izan da tradizionalki, nahiz azken urteotan hizkuntzalaritza konputazionala eta hizkuntza-teknologiak bateratu eta elkarlana nagusitu den.

a) Hizkuntzaren erabilera ikertzeko. Lehen hizkuntzaren deskribapenari begiratzen zitzaion batez ere, hitzaren mailan mugitzen zen, hitzaren egituran, ahalik eta osagairik txikienak (morfemak) deskribatu eta horien eta osagai handiagoen arteko konbinazioa zen axola zena. Gaur ez, gaur hori baino gehiago bilatzen da: erabilera. Eta corpusaren osaeran egituratzea findu bada (dialektoak, jakintza-arloak, garaiak, autoreak, erregistroak, etab.), murriztapenak erabili ahal izango dira beharren arabera fintzeko.

   Corpusek hizkuntzaren atal hauek ikertzeko lagundu izan dute, besteak beste: lexikoa (forma soilak zein konplexuak, morfemak, maiztasunak, neologismoak, adierak, kolokazioak, terminoen erauzketa...), gramatika (sintagmen egiturak, ordena, aditzen erregimena,...), itzulpengintza, dialektologia, semantika, ontologiak, pragmatika, ahotsaren analisia, hizkuntzen ikaskuntza, psikolinguistika, idatziaren *vs* ahozkoaren arteko bilakaera, sexuaren araberako erabilerak.

   Historikoki, baina, erabilera nagusia hiztegigintzara mugatu da: *BNC*[1] corpusa *Longman*, *Larousse* eta *Oxford* hiztegien abiapuntu izan da, *BoE*[2] *Collins* eta *Cobuild* hiztegiena, *FRANTEXT*[3] *Trésor de la Langue Française*ren oinarria da, *CREA*[4] *Diccionario de la Real Academia Española* eta *CTILC*[5] *Diccionari de la Llengua Catalana*, besteren artean. Eta gurean ere bai, ikus dezakegunez: *Orotariko Euskal Hiztegia* osatzen duten 16 liburukiak izen bereko corpusean oinarritu dira; corpus hau berau eta *XX. mendeko euskararen corpus estatistikoa*[6] dira *Hiztegi Batua*ren abiapuntu.

b) Tresnak probatzeko, proba-banku gisa. Testu-masa handiak behar dira corpusak proba-banku gisa erabiltzeko, formalismoak-eta datu askorekin probatu eta frogatu behar baitira, aldi berean horiek corpusak modu fidagarrian trata ditzaten.

2.2. Corpus esanguratsu batzuk

Corpusgintzaren historian eragin handiena izan duten corpus batzuk aipatuko ditugu azaletik, euskal corpusgintzaren historia errazago ulertzen lagunduko digutelakoan.

[1] British National Corpus: *http://www.natcorp.ox.ac.uk/*
[2] Bank of English: http://www.collins.co.uk/Corpus/CorpusSearch.aspxhttp://www.titania.bham.ac.uk/
[3] FRANTEXT: http://www.frantext.fr
[4] Corpus de Referencia del Español Actual: http://corpus.rae.es/creanet.html
[5] Corpus Textual Informatizat de la Llengua Catalana: http://ctilc.iec.cat/
[6] XX, mendeko euskararen corpus estatistikoa; www.euskaracorpusa.net/XXmendea

Ingelesa izan zen aitzindaria eta horietara mugatuko gara, mota desberdinetako erakusgarrienak aipatuz: lehen corpustzat hartzen dena, estatistikoa, *Brown* corpusa; erreferentzia-corpus nagusia, *British National Corpus*; corpus monitore bat, oreka bat mantenduz etengabe eguneratzen dena, *Bank of English*; eta corpus oportunista bat, aurkitzen duen guztia, orekari begiratu gabe, jasotzen duena, *Collins corpus*.

Lehen corpustzat Nelson Francisek eta Henry Kucerak osatutako *Brown corpusa*[7] (1964) hartu izan da, 2000 hitzeko 500 lagin-zatitan banatutako milioi bat hitzeko corpus estatistiko txikia, Ameriketako ingeles idatzia jasotzen zuena. Bere muga guztiekin ere, aldaketa nabarmena ekarri zuen hizkuntzalaritzaren lan egiteko moduan, eta eztabaidarik ere piztu zuen, garai horretakoak baitira Chomskyk corpusen kontra egindako adierazpenak.

Corpus honek berehala izan zuen segida, 1970-1978 urte bitartean Geoffrey Leech eta S. Johansson buru zirela, lagin-eredu bera baliatuta Britainia Handiko ingelesa jasoko zuen Lancaster-Oslo-Bergen (LOB) corpusa osatu baitzuten. Bi lurraldeetako ingelesa erkatzeko erabili ziren corpusak, besteak beste.

Baina hau abiapuntua besterik ez zen izan, berehala hasi baitziren beste corpus batzuk lantzen.

Historian jauzi bat eginez, gaur eredutzat hartzen den *British National Corpus* (BNC) aipatuko dugu, lehen erreferentzia-corpusa baita, hau da: hizkuntzaren erakusgarri orokorra eta orekatua. 100 milioi hitzek osatzen dute Britainia Handiko ingelesa bakarrik biltzen duen corpus hau, % 90 idatzia eta % 10 ahozkoa eskaintzen duena (eta hau berrikuntza handia izan zen bere garaian). Orekatua da osaeraren aldetik, kodetua dago informazioa erraz eskuratzeko eta berrerabiltzeko (TEIn[8] oinarritua eta XMLn kodetua). Corpus itxia da, 1975 inguruko edukiarekin hasi eta 1994koarekin bukatutzat eman baitzen. Gorago aipatu dugunez, batez ere hiztegigintzari begira sortu zen, baina egitura eta sailkapen osoak beste hainbat aplikazio bideratu dute, eta egun ere erreferentetzat hartzen da. Are gehiago, asko dira eredu honetan oinarritu diren, eta oraindik ere oinarritzen diren, corpusak, munduko hizkuntza askotan gainera.

Collins corpusak ez du edukiaren orekaren ardurarik: bil dezakeen guztia corpuseratzen du eta 2,5 bilioi hitz ditu dagoeneko, nahiz horietako 56 milioi bakarrik kontsulta daitezkeen sarean. Elebakarra da hau ere, ingelesa du helburu, baina mundu osokoa, eta idatzia (mundu guztiko webguneak, egunkariak, aldizkariak, liburuak) zein ahozkoa (irratia, telebista eta elkarrizketa arruntak) hartzen ditu. Irekia da, noski, hilero eguneratzen baita. *Collins* hiztegiak etengabe eguneratzeko erabiltzen da, hitz eta adiera berriak sortu ahala sartzen baitituzte.

Hain zuzen, aipatuko dugun azken corpusa, *Bank of English*, *Collins* corpusaren parte da, baina mugatuagoa. Corpus monitorea da, alegia, irekia, etengabe eguneratzen dena, baina oreka bat mantenduz. Corpus handia da, 650 milioi hitz jasotzen ditu, batez ere gaur egungo Britainia Handiko, Ameriketako Estatu Batuetako eta Australiako ingelesa eta, gehiena idatzia bada ere, ahozkoari leku egiten dio. *Collins Cobuild* hiztegien oinarria da corpus hau.

Azken urteotan beste korronte berri bat ari da indarra hartzen, *Web as Corpus*[9] gisa ezagutzen dena: Internet baliatzen dute corpus gisa.

[7] *http://icame.uib.no/brown/bcm.html*
[8] Text Encoding Initiative: http://www.tei-c.org/index.xml
[9] http://webascorpus.org

Gaur corpusak edonon aurki daitezke, ia hizkuntza guztiek dute berea, baina batez ere erreferentzia-corpusak dira lantzen direnak, orekatuak, hizkuntzaren azterketarako eredugarritzat hartzen direnak. Izan ere, hizkuntza batek bere "hiztegia" behar duen bezala behar du "corpusa" ere. Tamainari begiratuta, 100 milioi testu-hitzen bueltan dabiltza gehientsuenak, eta batez ere BNCren bidetik osatzen dira gainera. Gorakada 90. hamarkadan[10] etorri zen, neurri handi batean Europar Batasunak bultzatuta hizkuntza nazionalak biltzen zituen PAROLE[11] proiektua jarri zelako martxan, *EAGLES*[12] irizpideen arabera corpusak eratu eta irizpide bateratuen arabera kodetzea helburu zuena.

3. Euskal corpusgintza: egungo egoera

Aurreko atalean gainetik aurkeztutako corpusak ez ditut besterik gabe hautatu: erakusgarriak izateaz gain, badute loturaren bat euskal corpusgintzak izan duen bilakaerarekin.

3.1. Euskal corpusgintzaren abiapuntua: Euskaltzaindia

Euskal corpusgintzak badu bere tradizioa, eta Euskaltzaindiari zor dio neurri handi batean, bera izan baita aitzindari, eta berak eutsi baitio etorkizuneko corpusgintzaren aldeko apustuari. Euskarak eman dituen lehen bi corpusak Euskaltzaindiaren eskutik etorri zaizkigu, Hiztegigintza Planean aipatzen zituenak: *Orotariko Euskal Hiztegia*ren corpusa (euskararen tradizioa jasotzen duena) eta *EEBS corpusa* (*Egungo Euskararen Bilketa-lan Sistematikoa*), gerora *XX. mendeko euskararen corpus estatistikoa* izatera pasatu dena (euskara modernoa biltzen duena, UZEIk Euskaltzaindiaren enkarguz egina). Corpusen tradizio orokorrean bezala, *Euskaltzaindiaren Hiztegia*ren oinarri izateko sortu ziren bi corpusok, oso modu desberdinean osatu baziren ere.

Orotariko Euskal Hiztegiaren corpusa

1984an abiatu zen corpus historiko gisa (XVI. mendea – XX. mendearen erdia, osatuz joan dena), 310 obratako euskara idatzia bakarrik jasoz, eta guztira 5.600.000 hitz inguru bilduz. Testu gordina da, etiketatu gabea, sailkapen orokorra duena (epea, euskalkia, testu-mota zabala) eta lematizatu gabea (gerora, UZEIk, Hiztegi Batuko prestalanerako behar izan duen neurrian, % 10 inguru lematizatu badu ere). Hain zuzen, horregatik batzuetan zalantzan jarri izan da ea corpusa ala testu-bilduma hutsa den. Ez dago sarean, baina 2009an sareratu da corpus honekin osatu den *Orotariko Euskal Hiztegia*, Euskaltzaindiaren webgunean dagoena[13].

XX. mendeko euskararen corpus estatistikoa[14]

1986-2000 urte bitartean osatu zuen UZEIk Euskaltzaindiaren eskariz. Euskara idatzia biltzen du, baina baita ahozkoa, idatzi den neurrian. 1900-1999 epea jasotzen du, XX. mendeko euskal argitalpenen inbentario sailkatuan (garaia, euskalkia, testu-mota eta obraren tamaina) oinarrituta, erabileraren araberako lagin estatistikoa jasoz. Guztira

[10] Lehen aipatuez gain, CTILC (katalana, 52,3 milioi), CREA (espainiera, 160), CORGA (galegoa, 25), CNC (txekiera, 100), HNK (kroaziera, 101), HNC.gr (greziera, 47), HNC-hu (hungariera, 187), FIDA (esloveniera, 100), CORIS (italiera, 120), CRPC (portugesa, 334), ANC (Ameriketako ingelesa, 100). Kopuruok gora egin dute hizkuntza batzuetan, eta gerora sortu dira hizkuntza gehiagotarako corpusak ere.

[11] www.ilc.pi.cnr.it/parole/parole.html

[12] Expert Advisory Group on Language Engineering Standards: http://www.ilc.cnr.it/EAGLES96/browse.html

[13] http://www.euskaltzaindia.net/oeh

[14] http://www.euskaracorpusa.net/XXmendea/

SGMLn kodetutako ia 5.000.000 hitz biltzen ditu, erabat lematizatuak eta eskuz berrikusiak, 104.000 lema desberdin bilduz. Lema horiek, gainera, Euskaltzaindiaren arau berriekin eguneratzen dira etengabe. Garrantzi handia izan zuen corpus honek bere garaian, euskaraz modu "modernoan" osatutako lehenengoa izan zelako, eta gerora sortu ziren beste batzuen eredu ere izan zelako. Oraingoz, tamainaz txiki geratu bada ere, hau da XX. mendearen euskal erakusgarri orokor bakarra.

Gerora hainbat corpus sortu dira, espezialitatekoak gehienak, ondoren aipatuko ditugunak, baina Euskaltzaindiak beste urrats bat gehiago egin du corpusgintzaren ibilbidean:

Lexikoaren Behatokia

2008an jarri zuen abian corpus monitore hau. Lau urtean 50 milioi hitz bilduko ditu, baina etengabe eguneratuko da, eta erreferentzia-corpusaren oinarri izateko pentsatua dago. Horregatik eman zaio garrantzia katalogazioari, metadatuak xeheki jasotzeari, egituratzeari, berrerabilgarri izan dadin. 2000. urtetik honako euskara idatzia biltzen du *Lexikoaren Behatokia*k, batez ere komunikabideetako materiala lehen fase honetan, lexiko orokorra betiere, eta paperean zein elektronikoki argitaratua jasotzen du[15]. Guztia TEIren arabera kodetzen da eta automatikoki analizatzen da, morfosintaktikoki oraingoz. Gerora, lehen aplikazioa Hiztegi Batua osatzea eta eguneratzea izango denez, XXI. mendeko erabilerak baliatuz, beharren arabera berrikusi eta desanbiguatu da.

Berrikuntza handia ekarri du proiektu honek. Batetik, elkarlana bultzatu nahi izan du Euskaltzaindiak, eta corpusgintzako hiru erakunde bildu ditu berera: EHUko IXA taldea, Elhuyar Fundazioa eta UZEI. Bestetik, erreferentzia-corpusaren beharraz ohartuta, estandarren aldeko apustua egin du eta gerora berrerabiltzeko moduan antolatu du guztia, besteren artean eskura ditugun tresnak eta baliabideak bateratuz eta hobetuz. Laster izango da sarean guztion eskura.

Aurreko atalean ikusitakoaren bidetik, Euskaltzaindiak ere corpus estatistikoarekin abiatu zuen egungo corpusgintza, *Brown* corpusak izan zuen arrakasta gurean XX. mendekoak beteaz, baina gerora bidea erreferentzia-corpusa eta, honekin batera, corpus monitorea ezinbesteko direla ere ikusi du, *BNC* edo *BoE* izan diren bezala.

3.2. Beste euskal corpus batzuk

Euskaltzaindia aitzindari izan da corpusgintzan, baina ez da bakarra izan. Neurri batean beharrek sortuta, hainbat espezialitateko corpus bideratu dira gurean azken urteotan. Aipatu besterik ez ditut egingo, eta ez guztiak gainera. Besteak beste, jakin baitakit izango direla ezagutzen ez ditugunak, barruko erabilerarako sortuak eta inoren eskura ez daudenak. Ezagunenak aipatuko ditut bakarrik, eta corpus idatziez gain, bestelakoak ere aipatuko ditut, baina aipatu bakarrik, argazkia osatze aldera.

1. Corpus gisa egituratuak, idatziak, espezialitatekoak batez ere: EHUko Euskara Institutuaren eskutik sortuak dira hiru (*Ereduzko prosa, gaur* (EPG), *Ereduzko prosa dinamikoa* (EPD) eta *ZIO corpusa* (ZIO)), EHUko IXA taldearen eta Elhuyarren eskutik *Zientzia eta teknologiaren corpusa* (ZTC) eta erabiltzaileen eskura ez dagoen

[15] Lehen fase honetan *Berria* egunkaria, ETBko dokumental batzuk, eitb.com-eko berriak eta *Argia* aldizkaria katalogatu ditu.

Eusko Jaurlaritza - EJIEren eskutik landutako *Euskarazko corpus etiketatua eta segmentatua*[16].

Ereduzko prosa, gaur (EPG)[17]

EHUko Euskara Institutuak sortutako corpus honek 25,1 milioi hitz biltzen ditu bi multzotan banatuak: liburuak (2000-2006 bitarteko 287 liburu, literatura eta saioa batez ere, 13,1 milioi hitz, egileen ustez "ereduzkoak" diren autoreak biltzen dituztenak) eta prentsa (12 milioi hitz guztira, 2004-2006 bitarteko *Berria* egunkariko 10 milioi hitz eta 2001-2005 *Herria* astekariko 2 milioi hitz). Euskara idatzia eta ereduzkoa biltzen dute, kodetu gabea da eta *stemmer* bidez lematizatua dagoela dirudi, linguistikoki ez oso motibatua. *Egungo euskararen hiztegia* (EEH)[18] osatzeko erabili da, besteak beste.

Ereduzko prosa dinamikoa (EPD)[19]

EHUko Euskara Institutuaren beste corpus honek *EPG* corpusak ematen duen informazioa osatzea du helburu. Urtez urte aldatzen da, egileek diotenez unean unean erabiltzen den prosaren ispilu izan nahi baitu. Horregatik, berritze bakoitzak aurreko bost urteetan argitaratutako testu aukeratuak biltzen ditu. Oraingoz, 2004-2008 urte bitarteko 15 milioi testu-hitz ditu *EPG*k, erdia liburuetatik jasoa eta beste erdia prentsatik. Argitalpen bakoitzean urte zaharrena kendu eta berria sartzeko asmoa dute, betiere guztiak gordez, etorkizunean alderatze historikoak egin ahal izateko.

ZIO corpusa[20]

EHUko Euskara Institutuaren azken corpus honek egungo ezagutza zientifikoaren erakusgarri diren irakurgai hautatuak, "bikainenak" biltzen ditu, egileen esanetan. Euskarazko prosa zientifikoa biltzen du, espezialitateko itzulpenak direnak eta kalitateagatik erreferentzia moduan erabil daitezkeenak, egileen irizpideen arabera. Corpus honen lantze maila EPGrena da, kodetu gabea eta *stemmer* bidez lematizatua.

Zientzia eta teknologiaren corpusa[21]

Elhuyar Fundazioak eta EHUko IXA taldeak garatutako espezialitateko corpus honek 1990-2002 epean argitaratutako zientzia eta teknologiako eduki idatziak biltzen ditu. 8,5 milioi dituen corpus hau egungo estandarren arabera etiketatua da, sailkatua eta automatikoki lematizatua, 1,9 milioi eskuz zuzendu badituzte ere.

Euskarazko corpus etiketatua eta segmentatua

Eusko Jaurlaritza - EJIEren eskariz osatutako corpusa da, IKT inbentarioan jasotzen denez. Izenak dioen bezala, etiketatua eta segmentatua da eta aldizkari ofizialetako testuak biltzen ditu, onomastika eta toponimia kontuan izanda. Hala ere, aldizkari ofizialen bilduma, euskara-gaztelania corpus paraleloa etiketatzea eta segmentatzea aipatzen dute helburu gisa. Beraz, corpus elebakarra izan gabe, gerora elebidunen taldean aipatzekoa izan daiteke beharbada.

---

[16] Azken honen erreferentzia bakarra Eusko Jaurlaritzaren IKT inbentarioan aurkitu dut: http://www.euskara.euskadi.net/r59-19678x/eu/t06aInventarioWar/t06aBuscProyectoServlet?idioma=e&accion=PROYECTO_PORAREA&limpiar=s.

[17] http://www.ehu.es/euskara-orria/euskara/ereduzkoa/

[18] http://www.ehu.es/eeh/

[19] http://ehu.es/ehg/epd/

[20] Zientzia Irakurle Ororentzat: http://www.ehu.es/ehg/zio/

[21] http://www.ztcorpusa.net/cgi-bin/kontsulta.py

2. Testu-bilduma gisa jasoak: Susa literaturak sarean jarri ditu *Klasikoen gordailua* eta *Ibiñagabeitia proiektua*, egitura aldetik corpus gisa nekez onar daitezkeenak, baina testu-bilduma aparta eskaintzen dutenak. Horiekin batera, berriz ere Eusko Jaurlaritza - EJIEren eskariz bildutako *Euskarazko Testu Corpusa* aipatu behar da[22].

Klasikoen gordailua[23]

Susa literatura taldeak sortu du ia 500 liburu biltzen dituen 11,9 milioi hitzeko testu-bilduma, batez ere literaturaren alorrekoa, baina ez bakarra. Euskal literaturako lehen agerpenetatik 1936ko gerrak ezartzen duen mugara arte ekoitzi diren euskarazko literatur testuak biltzeko asmoa azaldu dute egileek. Epea, euskalkia eta generoaren arabera sailkatua dago. Egileen esanetan, testuak TEI eta RTF formatuetan jasota daude eta ez dago lematizatua[24].

Ibiñagabeitia proiektua[25]

Susa literatura taldearen bigarren testu-bilduma honek 451 aldizkaritako 10.967 artikulu jasotzen ditu, batez ere XX. mendeko bigarren erdikoak. Generoaren arabera dago sailkatua eta lematizatua dago, baina *stemmer*a baliatuta, ez linguistikoki motibatuta.

Euskarazko Testu Corpusa

Eusko Jaurlaritza - EJIEren eskariz osatutako 25 milioi hitzeko testu-bilduma hau testu adierazgarrien bilduma da, erabilera-eremu eta euskalki desberdinetatik jasoa, eta ahots-teknologietan erabiltzeko pentsatua, betiere ere IKT inbentarioko datuen arabera, ez baitago eskuratzerik.

3. Corpus elebidunak / eleaniztunak: hainbat corpus elebidun/eleaniztun daudela esan dezakegu, besteak beste itzulpengintzak gurean duen indarra ukaezina delako eta, ondorioz, laneko laguntza gisa ezinbestekoak direlako. Horietako asko ez dira corpus mailara iristen, baina laguntza baliagarria eskaintzen dute itzulpen-memorietan, adibidez. Aipagarrienak sartzearren, IVAPen *IDABA*, EHUren *Itzulpenen Kontsulta*, EIZIEren corpusa, Gipuzkoako Foru Aldundiaren itzulpenen datu-basea, Bizkaiko Foru Aldundiaren datu-base dokumentala, Deustuko Unibertsitateko DELi taldearen *LEGE-Bi*, UZEIren itzulpen-corpusa eta, bereziki, gaur aurkeztuko den *Consumer* corpusa izenda daitezke.

IDABA

IVAPek, Herri Arduralaritzaren Euskal Erakundeak, landu eta eskuratutako itzulpen-pare guztiak biltzen ditu *IDABA*k, metadatuekin, eta Eusko Jaurlaritza osoan dago kontsultagai. Administrazioko dokumentuak dira nagusi, baina legeak eta bestelakoak ere biltzen ditu. 1.100.000 itzulpen-unitate ditu. Oinarrian UZEIren *eLENA* aplikazioa dago.

EHUko itzulpenen kontsulta

EHUko Euskara Zerbitzuak unibertsitatean itzulitako liburuak, zientziari dagozkionak oraingoz, parekatuta biltzen dira, UZEIren *eLENA* aplikazioa erabiliz. Sailkatua dago, metadatuekin.

---

[22] Corpus honen erreferentzia, aurrekoa bezala, IKT inbentariotik eskuratua da.
[23] http://klasikoak.armiarma.com/
[24] Lematizazio automatikoa ezinezkoa da kasu honetan gainera, egungo euskararako (baturako eta bizkaierarako, oraingoz) bakarrik prestatu baita. Eskuzko lan ikaragarria eskatuko luke honek.
[25] http://andima.armiarma.com/

Gipuzkoako Foru Aldundiaren itzulpenen datu-basea

Gipuzkoako Foru Aldundian egiten diren itzulpen-pareak jasotzen ditu aplikazio honek, batez ere administrazioari dagozkionak, baina badira tartean legeak, dekretuak eta bestelakoak ere. Metadatuz hornitua dago, eta dokumentu motaren zein gaiaren araberako bilaketak onartzen ditu. Oraingoz barne-erabilerarako badago ere, Gipuzkoako udal guztietara zabaltzeko asmoa dute. UZEIren *eLENA* aplikazioa dago honen azpian.

Bizkaiko Foru Aldundiko datu-base dokumentala

Bizkaiko Foru Aldundiko administrazioko testu estandarizatuak biltzeko sistema gisa aurkezten dute, itzulpen-memoriekin bateragarri den datu-base dokumental edo dokumentu-biltegi hau, autoreek berek diotenez. Code&Syntax enpresak prestatua da.

EIZIEren corpus-hiztegia[26]

Webgunean corpusa aipatzen badute ere, corpus-hiztegia da erakusten eta definitzen dutena: gaztelania - euskara norabideko itzulpen-memoria gisa, administrazioko eta lege alorreko itzulpenetatik jasoak. Iturrien artean, Eusko Jaurlaritzako sail guztiak, foru aldundiak, Nafarroako Gobernua eta Osakidetza. Herri-administrazio erakunde nagusiek aldizkari ofizialetan eta gainerakoetan argitara emandako testu itzulien bilduma dela esan daiteke. 87.000 itzulpen-unitate dituzte, TMX formatuan.

UZEIren itzulpen-corpusa

UZEIk urtetan egindako itzulpen guztiak parekatuta biltzen ditu itzulpen-corpus honek. Eduki aldetik oso zabala da, bai gaiari begiratuta bai espezialitateari begiratuta, itzulpen oso espezializatuez gain arruntak ere jasotzen baititu. Eleaniztuna da, euskara, gaztelania, frantsesa, ingelesa eta, maila apalagoan, alemana eta katalana biltzen dituena. 48.286 dokumentu sailkatutatik eskuratutako 2.104.364 itzulpen-pare biltzen ditu, metadatuz hornituak, eta UZEIn garatutako eLENA aplikazioak kudeatzen du.

LEGE-Bi[27]

Deustuko Unibertsitateko DELi taldeak 1992-2001 urte bitartean bildu eta etiketatutako testu elebidunak biltzen ditu, euskara-gaztelania norabideko buletinak batez ere, baina atal eleaniztunak ere jasotzen ditu tarteka. Buletinok Euskal Herrikoak, Arabakoak, Bizkaikoak eta Gipuzkoakoak dira eta HTML, TMX, TXT eta XML gisa kodetuak daude. 82.549 itzulpen-unitate jasotzen ditu eta Vigoko Unibertsitateko CLUVIn[28] kontsulta daitezke.

Consumer corpusa

Eroskiren Consumer aldizkariaren urtetako lana biltzen du lau hizkuntza jasotzen dituen corpus honek: gaztelania, katalana, galegoa eta euskara. Gaur aurkeztuko da, beraz, aipatu besterik ez dezakegu egin hemen.

Corpus hauek elebakar gisa ere erabil daitezke. Aurreko ataleko edukia osatzeko ere balio dezakete, beraz.

4. Ahozko corpusak:

Euskaltzaindiaren *Euskararen Herri Hizkeren Atlasa*, EHUko Aholab taldearen *BIZKAIFON* eta *Basque FDB-1060 database*, Deustuko Unibertsitateko Fonetiker

[26] www.eizie.org
[27] http://www.deli.deusto.es/Resources/LEGE-Bi/
[28] Corpus Lingüístico da Universidade de Vigo: http://sli.uvigo.es/CLUVI/index.html#legebidun

taldearen *FonAtari* eta Jon Askeren *Basque Spoken Corpus* dira aipagarrienak, gehienak sarean eskura daitezkeenak (ELRA[29] katalogoaren bidez edo webguneetan).

Euskararen Herri Hizkeren Atlasa

Euskaltzaindiaren hizkuntza-atlasak 4000 ordutik gorako soinu-artxiboa jasotzen du eta, horrekin batera, ohiko euskal mintzamoldeen antologia eta EHHAren datu-basea. Oraingoz ez dago kontsultagai, baina eduki handia eta zabala da gerora ahozko corpus gisa erabiltzeko.

BIZKAIFON

EHUko Bilboko Ingeniaritza Goi Eskola Teknikoko Aholab taldeak prestatuak dira BIZKAIFONen bi bertsioak. Batak, Bizkaieraren Fonoteka gisa 2003. urtetik ELRAren webgunean salgai dagoenak (S0153 kodea), ahots-artxiboak eta hauei lotutako aldaera dialektalei buruzko informazioa eskaintzen du 21 orduko grabazioen bidez, bat-batekoa eta irakurria, transkripzio ortografikoa ondoan duela. Besteak, Bizkaiko Foru Aldundiarekin batera Bizkaieraren bideoteka[30] gisa sarean aurkezten dutenak, bideoak ere biltzen ditu.

Basque FDB-1060 database (SpeechDat-like)

EHUko Aholab taldeak 2003. urtean jarri zituen ELRA katalogoan (S0152 kodea) salgai 1.060 hiztunen telefono bidezko grabazioak (480 gizonezko eta 580 emakumezko). Ahoskeraren lexikoa eta transkripzio fonemikoa eskaintzen ditu.

FonAtari[31]

Deustuko Unibertsitateko Fonetiker taldearen corpus hau euskal fonetikaren atari gisa aurkezten da eta hizkuntzaren maila fonikoari buruzko informazioa eskaintzen du. Euskararen grabazio-bilduma aberatsa eskaintzen du, ahozko corpus desberdinetan banatua, egileen esanetan.

Basque Spoken Corpus

Jon Askek 2002. urtean ELRA katalogoan (S0123 kodea) eskuragarri jarri zituen 42 grabazio biltzen ditu ahozko corpus honek, dagozkien transkripzioekin. Adin, euskalki eta euskara-maila desberdinetako hiztunak hautatu zituen helburu jakin batekin: hitz-ordenaren azterketa.

ahotsak.com

Oraingoz corpus gisa landu gabe badago ere, Badihardugu elkartetik abiatu eta egun Euskal Herri osora zabaldu duen *ahotsak.com* egitasmoaren planetan badago Euskal Herriko hizkerak eta ahozko ondarearen corpusa osatzea. Une honetan[32] 108 herritako 1.690 hizlariren 15.954 pasarte dituzte, horietatik 3.501 transkribatuak, eta informazioa herrika eta gaika dago landua. Helburu gisa, 2012. urterako euskal herri guztietako lekukotasunak izatea da.

5. Trebatze-corpusak: tresna lagungarri gisa erabiltzen diren eta ezagutzen ditugun corpusak bi motatakoak dira: erabat etiketatuak eta eskuz zuzenduak, eredu gisa erabiltzekoak (EHUko IXA taldearen *EPEC* eta *UZEIren corpus etiketatua*), eta erroredun corpusa (IXA taldearen ERREUS).

[29] http://catalog.elra.info
[30] http://bizkaifon.ehu.es
[31] www.fonatari.org
[32] 2010-01-13ko datuak, www.ahotsak.com helbidean eskuratuak.

EPEC

EHUko IXA taldearen esanetan, “euskararen tratamendu automatikorako erreferentzia-corpusa” da EPEC. Euskara batu idatzia erabili dute corpusa etiketatzeko (*XX. mendeko euskararen corpus estatistikoa*ren zati bat + *Euskaldunon Egunkaria*), 300.000 hitz guztira. Linguistikoki eskuz etiketatua dago, maila desberdinetan (morfosintaxia, sintaxia eta semantika) eta IXAko tresnak hobetzeko darabilte, barruko erabilerarako, IKT inbentarioak jasotzen duenez.

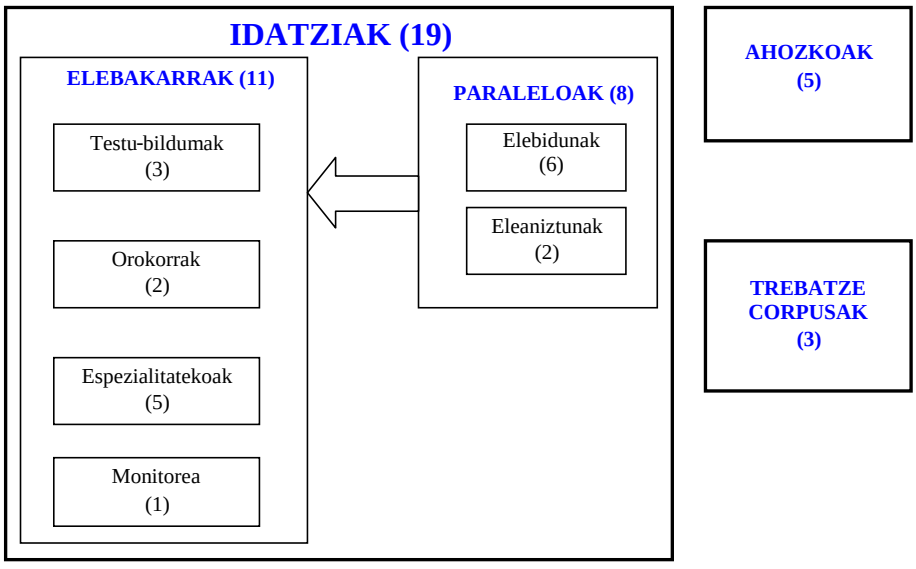
UZEIren corpus etiketatua 2008

UZEIk 2008. urtean argitaratutako 100.000 hitz (erdia lexiko orokorra, *Berria* egunkaria, eta beste erdia lexiko espezializatua (ekonomia, zuzenbidea eta medikuntza, originalak eta itzulpenak) morfosintaktikoki etiketatu eta eskuz desanbiguatu ditu etxe barruko tresnak probatu eta hobetzeko. Barruko erabilerarako da. Honekin batera, gaztelaniako beste 100.000 hitz ari da automatikoki etiketatzen eta eskuz berrikusten, corpus paraleloetan erabiltzen diren tresnak hobetzeko.

ERREUS

Euskara ikasten ari direnen testuen bilduma antolatua da IKT inbentarioan aipatzen denez. Ikasleen ezaugarrien arabera eta euskaltegietako programazioen arabera antolatua dago eta testuetako erroreen etiketatzea ere badu. Ez dago eskuragarri.

Honaino euskal corpus eta testu-bilduma esanguratsuenen argazkia, 27 guztira, ez-osoa ezinbestean.



Ahozkoak eta trebatze-corpusak hemen utziko ditugu, eta testu-corpusetara mugatuko gara. Baina, aurrera egin aurretik, ohar pare bat:

a) ahozko corpusak urri samarrak dira tamainaz, hauek eskuratzea, transkribatzea eta etiketatzea lan astuna baita. Kasuren batean ikus daiteke tratamendu automatizatuari eman diotela garrantzia eta, beraz, ondoko lanetarako aurrelanak egin dituztela, corpusaren edukiaren kaltetan. Oinarriak ezarrita daudela esan dezakegu.

b) trebatze-corpusak, izenak dioen bezala, ez dira erabilera corpusak *per se*, laguntzarako baliabideak baizik.

Bi corpus mota hauek, beraz, ondoko atalean ere sar zitezkeen, corpusgintzarako laguntza gisa, trebatze-corpusak bereziki.

Testu-corpusetara itzuliz, hauek dira aipatu eta sailkatu ditugun euskal corpus idatzi nagusiak:

Euskarazko corpus etiketatua eta segmentatua	espezialitatekoa	idatzia	Administrazioa
Ereduzko prosa, gaur (EPG)	espezialitatekoa	idatzia	Literatura, saioa, prentsa
Ereduzko prosa dinamikoa (EPD)	espezialitatekoa	idatzia	Literatura, saioa, prentsa
ZIO corpusa	espezialitatekoa	idatzia	Prosa zientifikoa
Zientzia eta teknologiaren corpusa	espezialitatekoa	idatzia	Zientzia eta teknologia
Lexikoaren Behatokia	monitorea	idatzia	Komunikabideak
Euskarazko Testu Corpusa	orokorra	idatzia	Orokorra
XX. mendeko euskararen corpus estatistikoa	orokorra	idatzia + ahozkorako idatzia	Orokorra
IDB	paraleloa	eleaniztuna	Orokorra
Consumer corpusa	paraleloa	eleaniztuna	Orokorra
EHUren Itzulpenen Kontsulta	paraleloa	elebiduna	Zientzia
LEGE-Bi	paraleloa	elebiduna	Administrazioa, zuzenbidea
IDABA	paraleloa	elebiduna	Administrazioa, zuzenbidea
EIZIEren corpus hiztegia	paraleloa	idatzia	Administrazioa, zuzenbidea
GFAren itzulpenen datu-basea	paraleloa	elebiduna	Administrazioa, zuzenbidea
BFAren datu-base dokumentala	paraleloa	elebiduna	Administrazioa
Klasikoen gordailua	testu-bilduma	idatzia	Literatura
Ibiñagabeitia proiektua	testu-bilduma	idatzia	Komunikabideak
OEH	testu-bilduma	idatzia	Orokorra

Taula honek bi gauza uzten ditu agerian: a) euskaraz badira corpusak, baina, Euskaltzaindiaren egitasmoak kenduta, partikularrak edo/eta espezialitatekoak dira eta, egun, ez daukagu corpus orokorrik; eta b) lan partikular horiek proiektuak bikoiztera eraman dute corpusgile ugari, eta hori nabarmena da corpus elebidunen kasuan, buletinak behin eta berriro aipatzen baitira egitasmoetan, adibidez.

Edukien sailkapenera etorriz, eta ezagutzen ditugun corpusei begirada azkar bat emanda, literatura, saioa, zientzia eta teknologia, komunikabideak, administrazioa eta zuzenbidea jasotzen dute corpusek, baina ez dira ageri filosofia, hizkuntzalaritza, ekonomia, gizarte-zientziak, kirola, erlijioa eta beste hainbat. Informazio hori guztia, txikia eta mugatua bazen ere, eskaintzen zuen *XX. mendeko euskararen corpus estatistikoa*k, orekatua gainera, baina oraingoz ez dugu XXI. mendeko euskararen erakusgarri izango den corpus bateratu orokor bat, egungo corpusak integra ditzakeena, azpicorpus gisa nahi bada. Egungo euskararen erreferente izango den *XXI. mendeko euskararen erreferentzia-corpusa* behar dugula uste dugu, Euskaltzaindiak berak behin baino gehiagotan planteatu izan duena, guztiona eta guztiontzat izango dena. Esan dugunez, eduki asko berrerabil daiteke, batzuk arazorik gabe gainera, erabat kodetuak eta etiketatuak baitaude, beste batzuk moldatu egin beharko lirateke. Eta egun argitaletxeek, komunikabideek, itzulpen-enpresek material guztia dute formatu elektronikoan. Are gehiago, ondoko orrietan ikusiko dugunez, edukiaz gain, baliabideak eta tresnak ere baditugu halako proiektu bati heltzeko, beste hizkuntzen parean jartzeko. Eta honek ez luke kenduko corpus monitoreak, oportunistak, *Web as Corpus* moduko proiektuekin jarraitzea, horiek ere behar baitira, baina guztien arteko loturak finkatuta.

3.3. Corpusgintzarako euskal baliabideak eta tresnak

Euskal corpusgintzak lan handia egin du hasieratik, eta gurea moduko hizkuntza eranskari bat automatikoki tratatzen saiatzea erronka handia bezain erakargarria izan da.

1. Baliabide lexikoak eta gramatikak: euskara automatikoki tratatzeko prestatu behar izan dira baliabide lexikoak (hiztegiak etiketatu, atzizkiak tratatu, informazio morfosintaktikoa gehitu), baina baita gramatikak sistematizatu ere, horretarako formalismo berriak ezagutu eta gure hizkuntzaren gramatika berriak sortuz.

Datu-base lexikalak

Hiru datu-base lexikal nagusi ezagutzen ditugu, eta horietako bi ELRA katalogoan eskura daitezke. IXA taldeak osatutako *EDBL*[33], 75.000 sarrera dituena (ELRA, L0057 kodea), eta UZEIren *euLEX* (ELRA, L0085 kodea), 120.000 sarrera biltzen dituena, lexiko orokorra eta terminologia orokorrena. Hirugarrena, *LBDBL*[34], Lexikoaren Behatokiaren baitan sortu da, Euskaltzaindiaren zuzendaritzapean, eta oinarrian EDBL dauka, baina Elhuyar Fundazioko lexiko orokorrarekin eta UZEIko terminologiarekin osatua, eta corpusak berak ematen duen lexiko berriarekin aberasten joateko asmoa du.

Datu-banku terminologikoak

UZEIk urtetan landutako terminologia biltzen du *UTH*k, UZEIren Terminologia Hiztegiak. 121.457 euskal termino etiketatu eta sailkatu dauzka, eta etengabe eguneratzen da. Erdal terminoak ere etiketatzen dihardu UZEIk.

Ezagutza-baseak

IXA taldeak lantzen duen *Euskal WordNet* hitzen esanahiak erlazio lexiko-semantikoen inguruan egituratzen dituen ezagutza-base lexikala da.

Baliabide gramatikalak

Euskal morfosintaxiaren deskribapen zehatza egin eta formalismo batzuen arabera sistematizatu da, euskara automatikoki tratatu ahal izateko. Guk ezagutzen ditugun gramatikak bi dira: IXA taldeak eta UZEIk deskribatu dituztenak. Ondoko atalean ikusiko ditugun tresna guztiak benetan linguistikoki motibatuak badira, izan beharko lukete gehiago ere, baina ez dugu informaziorik eskuratzerik izan.

Bestalde, UZEIk sortuak ditu esaldi mailako erregela linguistikoak ere, batez ere termino-erauzketan baliatzen dituenak.

2. Corpusgintzarako tresnak: hainbat motatako tresnak garatu dira euskal corpusgintzan.

Corpusen eratze-prozesurako tresnak: corpusgintzak landuko den edukia katalogatu eta eskuratzea eskatzen du, ondoren guztia etiketatu, linguistikoki tratatu eta ustiapenerako prestatzeko. Prozesu hori bideratzeko *Corpusgile* prestatu zuen Elhuyar Fundazioak, eta, maila apalagoan, *SASKI*, UZEIk.

Corpus paraleloak lerrokatzeko eta ustiatzeko tresnak: Ametzagaiñak sortu ditu *Ametra* proiektuaren baitan, Code&Syntax-ek *TuMaTxa* eta *SARE Bi*, Elhuyarrek ere *ElexBI* eta *AzerHitz* prestatu ditu, eta UZEIren *eLENA* aplikazioa darabilte UZEIko *IDB* itzulpen-corpusak, IVAPeko *IDABA*k, EHUko Itzulpenen Kontsultak eta Gipuzkoako Foru

[33] EDBL (Euskararen Datu Base Lexikala), IXA taldeak eta UZEIk osatua, gerora IXA taldeak bere gain hartu duena. http://ixa2.si.ehu.es/edbl/

[34] LBDBL (Lexikoaren Behatokiko Datu Base Lexikala).

Aldundiko itzulpenen datu-baseak, besteak beste. UZEIk *IMEBI*, itzulpen-memorientzako arakatzaile aurreratua ere prestatzen dihardu.

Informazio linguistikoa tratatzeko tresnak, corpusetarako eta beste hainbat aplikaziotarako dutenak, berriz, segidan ikus ditzakegu.

Analizatzaileak / lematizatzaileak: lau tresna aipatzen dira merkatuan, baina maila oso desberdinetakoak direla pentsa daiteke corpusek eskaintzen dituzten emaitzak ikusita. Guk dakigula, eta aurreko ataleko datu-baseekin eta gramatikekin ikusi ahal izan dugunez, bi dira euskararako analizatzaile/sortzaile erabat linguistikoki oinarrituak, corpusen lematizaziorako baliatzen direnak: IXA taldearen *Eustagger* eta UZEIren *tLEMA*. Ametzagaiñak lematizatzailea duela aipatzen du, baina *stemmer* mailan dabilela dirudi *EPG*, *EPD*, *ZIO* eta *Ibiñagabeitia* corpusen emaitzak ikusita, hau da, ezagutza linguistiko minimoa eta oinarrizko erregela batzuk darabiltzala ematen du, atzizkien zerrendak baliatuta. Azkenik, EmergiaTech enpresak bere lematizatzailea sortu du, baina ez dugu horri buruzko daturik eskuratzerik izan. Garai batean *Snowball* sortu zuen Code&Syntax-ek, baina badirudi hor geratu zela.

Corpusen tratamendurako, hala ere, gaur *sistema hibridoak* erabiltzen dira askotan, *ezagutza linguistikoa eredu estatistikoekin aberastuta*. Bide hori lantzen dute IXA taldeak eta UZEIk, eta hala aipatzen dute Ametzagaiñakoek ere, baina, aurrekoa kontuan izanda, ez dakigu zein mailatan aplikatzen duten hori.

Termino-erauzleak: terminologiaren tratamendurako tresnak Elhuyar Fundazioak eta UZEIk eskaintzen dituzte, *Erauzterm* eta *Itzulterm* lehenengoak, *UZEIren termino-erauzlea* bigarrenak.

Transkriptoreak: ahotsa testu bihurtzeko *AhoPhonTranskrip* garatu zuen Aholab taldeak. Grabazio batetik abiatuz, esaldi bat irakurri eta hitzetan banatzen du, dagokion forma emanez, eta horretarako lexikoak baliatuz.

Honekin batera *AhoTools* ere garatu dute, seinaleen analisirako softwarea (soinu-uhina, energia eta espektrograma ikus daitezke, besteak beste). Elementu hauek guztiak balio handikoak dira segmentazioa eta transkripzioak gauzatzeko.

Azkenik, gaur Internet corpus gisa erabiltzen dela jakinda, *Corpeus* garatu zuen Elhuyar Fundazioak Internet euskarazko corpus gisa kontsultatzeko eta ustiatzeko.

Izango dira gehiago, baina irudi azkar bat egiteko balio dezakete orain arte aipatutakoak, bai eta gogoeta pare bat sorrarazteko ere: a) baditugu baliabideak eta tresnak, merkatuari aurre egiteko ongi posizionatuak gaudela esan daiteke, baina b) zein bere aldetik ari gara sortzen, lanak bikoizten, corpusak bezala. Horrek erakusten du prestatuta gaudela erronka berriei erantzuteko, jende asko dela gai garai berrietara egokitzeko eta euskaraz lan egiteko gainera. Baina, behar al dira lau lematizatzaile euskara moduko erabilera-eremu urriko hizkuntza eranskari baterako? Zergatik sortu dira hainbeste? Erantzuna ere berehala datorkigu burura: beharra dagoelako, eta daudenak eskuratzerik ez dagoelako. Lematizatzaileak sortuta daude eta beude horretan, baina, aurrera begira, ez al genituzke indarrak batu beharko, lanak banatu beharko, eta bikoizten edo n-koizten ez ibili? Guztiok irabaziko genuke gainera, gauza berri gehiago garatzeko aukera izango genukeelako, elkarlana bideratuz. Corpusekin Euskaltzaindiak bidea markatu duela dirudien bezala, tresnekin eta baliabideekin ere zerbait egin beharko litzateke. Sortzear da *Hizkuntza Teknologien Clusterra*. Hortik bidera daitezke etorkizuneko lanak eta indarrak batzeak, beharbada.

Bukatzeko, aurreko guztia ikusita, euskal corpusgintza sendo dagoela esan dezakegu: baditugu baliabideak, baditugu tresnak, badugu ezagutza, eta baditugu corpusak. Erabilera-eremu urriko hizkuntza (eranskaria, gainera) izanda ere, leku onean gaude. Beste hizkuntza batzuetarako egin diren urratsak egin ditugu guk ere corpusen osaeran: historikoa, estatistikoa, monitorea, espezialitatekoak, ahotsa, trebatze-corpusak,... Eta lematizatzaileak, desanbiguatzaileak, corpus-eraketarako eta -ustiapenerako tresnak garatu dira. Euskal corpusgintza osasunez ondo badago ere, ezin dugu hutsune handi bat aipatu gabe utzi, argazkian falta dena: erreferentzia-corpusa.

3.4. Euskal corpusen erabilera

Baina, non eta zertarako erabiltzen ditugu corpusak? Izan, baditugu, baina zeinek jotzen du corpusetara? Orain artekoa ikusita, sortu diren corpusak begiratzen baditugu, beharrek eraginda sortu dira sortu direnak, eta interes konkretu batzuei begira osatu dira corpus konkretuak. Beraz, euskararen bilakaera diakronikoa jasotzeko bildu zen *OEH*, XX. mendeko euskararen berri jasotzeko *XX. mendeko euskararen corpus estatistikoa*, euskara "eredugarria" biltzeko *EPG* eta *ZIO*, zientzia eta teknologia aztertzeko *ZTC*. Kasu hauetan guztietan, baina, hiztegiek hartu dute garrantzia.

Euskaltzaindiaren Hiztegi Batuaren aztergaia osatzeko, adibidez, XX. mendeko corpusetik eskuratu ziren maiztasun handieneko lemen zerrendak eta maiztasunen araberako multzoetan oinarrituta joan da hiztegia osatzen, OEHk eskaintzen duen tradizioarekin batera. Aurrerantzean Lexikoaren Behatokiak eskainiko du informazio hori, erreferentzia-corpusik ezean, eta XXI. mendeko lexikoarekin aberastuko da oinarria dagoeneko bildua duen egungo Hiztegi Batua.

Baina lexikoaren atalean badira behar gehiago (XX. mendeko corpusa dela-eta jasotako eskaerak dira ondoren aipatzen direnak, erabilera errealak, jakinik askoz gehiago izango direla, dudarik gabe).

Lexikoa: hitz-elkarketaren eta eratorpenaren azterketak, Zuzenbideko hitz gakoen bilakaera, jakintza-alor konkretuetako terminoen erauzketa, neologismoen detekzioa, kolokazioak, lemen aldaerak, euskalkiak, onomastika (entitateen detekzioa eta tratamendua), adierak, lexiko-estatistika (maiztasun-zerrendak, Swadesh zerrenda), zenbakien idazkera, ordinalak.

Eskaerak Euskaltzaindiko batzordeetatik, hemengo zein kanpoko unibertsitateetatik, eta bestelako erakunde eta enpresetatik ere etorri dira, atzerrikoak hainbat kasutan. Ikerketarako zein produktuen garapenerako behar zuten corpusaren edukia.

Eta beste adar guztiak?

Gramatika: aditz-motak, aditzen erregimena, postposizioak, menderagailuak izan dira eskaera nagusiak, Euskaltzaindiko Gramatika Batzordetik eta unibertsitateetako ikertzaileen eskutik. XX. mendeko corpusarekin batera, OEH corpuseko datuek izan dute eskabiderik gehien gramatikaren azterketan.

Semantika: adieren tratamendurako zerrenda zehatzak (adieren neologiarako, adibidez), zentzumen-aditzak, ikertzaileengandik iritsi dira eskaerak.

Psikolinguistika: psikolinguistika estatistikoa (euskal hitzen zerrenda maiztasunduna, bigramen azterketa), e-Hitz proiektuaren oinarri izan dena.

Horiek dira UZEIn jaso diren eskaera formaletako batzuk, erabiltzaile arruntek webgunean zuzenean egiten dituzten kontsulta puntualagoak alde batera utzita.

Badira, hala ere, askoz ere erabilera gehiago, baina corpusaren tamainak, edukiak, sailkapenak eta etiketatze-mailak baldintzatuko dituzte erabilera batzuk ala besteak.

Psikolinguistikan, lehen aipatuez gain, afasia, erroreak, haur-hizkuntza aztertzeko balio handia dute corpusek.

Pragmatika eta diskurtsoaren analisia aztertzeko.

*Estilistika*n, azpicorpusa erabilita normalean, ez corpus osoa, autore konkretuei buruzko informazioa izaten baita aztertzen dena.

Estatistika: aipatu dugu lexikoaren atalean, baina maiztasunak adar guztietan aplikatzekoak dira, datuek eta kopuruek (edo kopuru-ezak) garrantzia hartzen baitute corpusgintzan.

Soziolinguistika: erregistroen eta estiloaren azterketa, baina ikuspegi soziologikotik.

Hizkuntzen irakaskuntza: hizkuntza-ereduak detektatzeko, baina baita adibidetegiak eskuratzeko, ezezagutzak eragindako akatsak detektatzeko, etab.

Ikerketa historikoak bideratzeko ere erabiltzen dira corpusak, hizkuntzaren garapena adar guztietan kontutan hartuta gainera.

Ahozko corpusetatik fonetika, erregistroak, sexuen araberako bereizketak, hitz-ordena elkarrizketan, elipsia, etenak, ezaugarri lexiko-fonetiko-morfologiko-sintaktiko bereziak eskuratzen dira, besteak beste.

Adar hauek guztiek ez dute corpus erraldoi bat eskatzen ezinbestean, batzuetan azpicorpus txiki bat besterik ez dute beharko. Adibidez, estilistika ikertu nahi duenak autore bakar bat hautatuko du seguru asko, edo dialektologian diharduenak euskalki bakar bat, edo euskalki bakar horretako garai zehatz bat, edo finago. Horrek corpusa ondo egituratua izatea eskatzen du, eta informazio ugariz eta aberatsez hornitua, gero erabiltzaileak komeni zaiona erraz eta zehatz hauta dezan, *zaratarik* gabe. Eta azkar. Corpusgileak egin behar du lana erabiltzaileak corpusaren emaitzak balia ditzan, garbitze-lanik hartu gabe.

Eta ez dezagun ahaztu *proba-banku* gisa ere behar direla corpusak, hori ere corpusen erabilera garrantzitsua da.

4. Ondorio gisa, aurrera begirakoak

Euskal corpusgintza ondo samar prestatua dagoela ikusi dugu: baditugu baliabideak, tresnak eta ezagutza. Baina etorkizun hurbilerako erronka batzuk aipatu nahi nituzke, zuen baimenarekin:

a) Erreferentzia-corpusa bideratzeko urrats sendoak egin beharko genituzke
b) Corpus monitorea osatzen joan behar dugu eta, ahal dela, sakabanatutako informazioa bateratzen saiatu, gerora azpicorpus gisa erabili ahal izateko
c) Baliabideen eta tresnen optimizazioa bideratu beharko litzateke, lanak bikoiztu gabe
d) Informazioa ireki eta guztien eskura jartzea komeni da, zeinek bere beharren arabera ustia dezan

Euskaltzaindiari eskatuko nioke corpusen ardura hori bere gain har dezala, bera baita corpus nagusiak orain arte bideratu dituena eta, lanak bateratuko badira, guztion babesa duena. Corpusgintzarako metodologia, baliabide eta tresnen kudeaketa, berriz,

Hizkuntza Teknologien Cluster berriak har lezake bere gain, berak izango baitu dagoenaren, horren egoeraren eta beharren ezagutza zabalena. Aukera ezin hobea iruditzen zait horri heltzeko, batez ere talde zabala biltzen badu bere baitan.

Utz iezadazue, bukatzeko, bi hitz esaten Euskaltzaindiari eta Eroskiri. Euskaltzaindiari eskerrak orain arte egindako lan mardulagatik, corpusgintzan hasieratik sinetsi eta aurrera egin duelako, eta eskerrak gaurko jardunaldi hau antolatu eta zuen artean egoteko eskaintza egin didalako. Eskerrik asko, bihotzez.

Eta eskerrak Eroskiri bere eskuzabaltasunagatik, *Consumer* corpusak corpusgintzari, kasu honetan euskal corpusgintzari, egiten dion ekarriagatik, Euskaltzaindiaren esku jartzeagatik. Eta zorionak 40 urteengatik.

Eskerrik asko guztioi.