Testu-corpusak: ezaugarriak, eraketa eta tresnak

IXA taldea, Elhuyar Fundazioa

Sarrera: corpusak

Zer eta zertarako

Errealitateari buruzko hipotesia egin nahi duen zientzia orok errealitateak ematen dizkion datuetan oinarritu beharko lituzke haren ondorioak. Horrelako datuei natura-zientzietan datu enpiriko deitu ohi zaie. Giza zientziek ere datu enpirikoetara jo dezakete deskribatzen duten errealitateari buruzko hipotesiak ziurtatzeko. Baina hizkuntzalaritzaren kasuan, zeintzuk lirateke erreferentetzat hartu beharreko datu enpirikoak? Berez, hizkuntza da hizkuntzalariak deskribatu nahi duena, eta, hori egiteko, enuntziatu linguistikoetara jo beharko du, hau da, hizkuntza-jarduera errealaren emaitza diren esakuneetara bai ahozko hizketara, bai eta idatzizko testuetara ere. Bestela esanda, hizkuntzalaritzak ere beren teoriei eusten dieten erreferentzia-elementuak (datu enpirikoak) behar ditu, hizkuntzaren joera nagusiak agertzen dituztenak, eta corpusak dira, hain zuzen ere, erreferentzia hori sistematizatzen duten testu-multzoak (edo hizketa-bildumak). Hizkuntzalaritza enpirikoa deritzona mende hasieran indarrean zebilen korronteetako bat izan zen.

Hasierako hizkuntzalaritza sortzailearen [Chomsky, 1957] oinarrian, ordea, hizkuntza batek izan ditzakeen enuntziatuak ezin zenbatuzkoak direlako baieztapena zegoen, eta, haren jarraitzaileen ustez, ez dago hizkuntzaren mekanismoak osorik azalduko dituzten datu egokiak bil litzakeen testu-multzo (corpus) finiturik. Deskribatu behar den objektuaren adibidea bere hizkuntza hitz egiteko gaitasuna duen hiztun ideal batengan bilatu behar litzatekeela diote. Korronte horren ondorioz, hizkuntzalaritzaren ikuspegi enpirista batetik ikuspegi arrazionalista batera igaro zen. Orientazio berri horrek berarekin zekarren kritika hau da: corpusek ez dute balio hizkuntza deskribatzeko.

Aurrerago, eta batez ere hizkuntzalaritza aplikatua egiten hasi zenetik, corpusak, hizkuntzaren ikuspegi osoa ematera baino, beste helburu batera bideratu izan dira Corpusen helburu berria hizkuntzalaritzaren inguruko ikerkuntzaren oinarri izango den lagin adierazgarri bat izatea litzateke, bertan baitaude datu objektiboak. Corpusa ezingo da hizkuntza osoarekin parekatu, ezaugarri egokiak edo ez hain egokiak izango dituen datu-multzoa baino ez da izango. Helburu berri honen harira, hizkuntzalaritza enpirikoaren eta corpusen gaineko interesa handitu egin da berriro, hizkuntzalaritza arrazionalista eta enpirikoaren arteko eztabaidak bere horretan dirauen arren. Dena den, corpusen erabilgarritasuna ez da egun zalantzan jartzen.

Oraindik corpusa zehazki zer den guztiz definitu gabe ere, esan genezake corpusa hizkuntzari buruzko datu-bilduma dela. Corpusaren definizio zabal bat egitera, esan daiteke edozein testu edo testu-bilduma har litekeela corpustzat. Hala ere, gaur egun definizio zehatzagoa erabili ohi da: corpusa hizkuntza-erakusgarri 'errealen' multzo 'handi' bat da, irizpide batzuen arabera bildua eta formatu elektronikoan biltegiratua. Askok beste zerbait ere erantsiko liokete horri: corpusak, erabilgarria eta eraginkorra izango bada, informazio linguistikoaz hornitua behar du izan. Adibidez, lexikoaren morfologian interesatuta dagoen hizkuntzalari batentzat, corpusa hizkuntza bateko hitz eratorrien multzoa izan daiteke; edo, sintaxian lan egiten duen hizkuntzalari batentzat, hizkuntzaren sintagma-multzo zabala.

Baina hori guztia hala izanik ere, *corpus* hitza modu zorrotzago batean erabiltzen dela esan daiteke, eta oso lotua dago *hizkuntzaren teknologia* deritzon arloarekin. Adibidez, gaur egungo itzulpen

automatikoko sistema gehienak nola edo hala itzulpenen corpusetan daude oinarrituta.

Corpus-motak

Corpusen artean hainbat mota bereiz daitezke:

- hizketa/testua: ohikoenak testu-corpusak dira, baina, grabazioen eta komunikabideen digitalizazioa dela eta, gero eta hedatuagoak daude hizketakoak. Hala ere, artikulu honetan testu-corpusen gaiari helduko diogu

- orokorrak/espezializatuak: corpus berezia edo espezializatua hizkuntzaren erabilera-eremu espezifiko bateko edo hizkuntza-aldaera jakin bateko testuak biltzen dituen corpus-mota da, eremu edo aldaera horretako ezaugarriak aztertzeko asmoz eratua.

- elebakarrak/elebidunak/eleanitzak: hizkuntza bakarreko testu-bildumak diren corpus elebakarrak ohikoenak badira ere, gero eta corpus eleaniztun gehiago egiten ari dira, batez ere itzulpenari begira. Itzulpen-memorien datu-baseak eta *corpus paraleloak* dira horien artean erabilienak. Corpus horietan testu bera hizkuntza batean baino gehiagotan ematen da, (*bitestuak* ere esaten zaie). Corpus parekatuak corpus paraleloak dira, baina markatzen dute zein perpaus edo osagai den zeinen itzulpen. Hizkuntza desberdinetan antzeko ezaugarriak (gaia, urtea, mota...) dituzten testuak bilduz *corpus konparagarriak* deitzen diren corpusak osatzen dira, baina corpus konparagarri horien kasuan testuak ez dira beste hizkuntzetako testuen itzulpenak. Adibidez, urte bereko hainbat hizkuntzatako egunkariak bilduz corpus konparagarri bat lortzen da, baina ez corpus paraleloa.

- gordinak/etiketatuak: corpus gordinetan, testua argitaratu zen bezala dago, ez du informazio gehigarririk. Corpus etiketatuetan, testuko hitzei edo hizkuntza-unitateei buruzko informazioa eransten da, eskuarki etiketen bidez. Corpus gordinak sinpleagoak dira eratzen, baina aplikazioei begira mugatuagoak, etiketatuetan dagoen informazio gehigarriak (dokumentala, lexikala, morfologikoa, sintaktikoa, semantikoa) askoz interesgarriago bihurtzen baititu ikerkuntzari zein aplikazio informatikoei begira. Informazio gehigarrien artean, lema edo erroa funtsezkoa da bilaketak modu egokian egin ahal izateko, are gehiago euskara bezalako flexio handiko hizkuntzetan. Idazle batzuek *hizkuntza-corpusak* esaten diete linguistikoki etiketatutako corpusei.

- sinkronikoak/diakronikoak: garai desberdinetako testuak bilduz lortzen dira corpus diakronikoak, eta haien interes nagusia hizkuntzaren bilakaera aztertzea da. Sinkronikoen helburua egungo egoera aztertzea da.

- erreferentzia-corpusak: corpusaren helburua hizkuntzaren erabilera-eremu guztietarako baliagarria edo 'adierazgarria' izatea denean, 'erreferentzia-corpusa' edo 'orotariko corpusa' dela esan ohi da (Leech, 2002). Batzuetan *ereduzko* kontzeptuarekin nahasten da, baina desberdinak dira; erreferentziazkoak nola edo hala *maiztasuna* kontuan hartzen duelako, eta ereduzkoak, berriz, irizpide jakin batzuen arabera 'gomendatua' den eredua islatzen du.

- orekatuak: argitaratutako testuen artean metodo estatistikoak erabiliz adierazgarritasuna eta aniztasuna bilatzen duten corpusei esaten zaie. Erreferentziazkoak orekatuak izan ohi dira, baina horrez gain handi samarrak. Oreka hainbat irizpideren arabera bila daiteke: eremua, genero, dialektoak, garai historikoak, etab.

- nazionala: erreferentzia-corpusen sinonimotzat har daiteke, baina neurriarengatik, hedadurarengatik eta etiketatze-lan sakonarengatik hizkuntza baten erreferentzia nagusi eta ezinbesteko bihurtu (nahi) den corpusari deitu ohi zaio corpus nazionala.

- monitorea: etengabe eguneratzen den corpusa, hizkuntzaren egoera monitorizatzeko asmoz

eratua.

Artikulu honetan, corpus etiketatuei buruz ariko gara gehienbat, bestelakoak liburutegi digitalen eremuan kokatzen baitira corpusen eremuan baino gehiago.

Erabilpenak

Hizkuntzaren azterketan corpusek duten garrantzia ukaezina da gaur egun. Horietatik lortzen diren datu enpirikoen bitartez hizkuntzalariek adierazpen objektiboak egin ditzakete, eta ez adierazpen subjektiboetara mugatu, edo ez norberak hizkuntzari buruz izan dezakeen pertzepzio batera mugatu. Datu enpiriko horien bitartez, hizkuntzaren aldaerak aztertzeko aukera dugu, hala nola, dialektoak, edo lehenagoko hizkuntza-aroak, horrelakoak ezin baititugu gure sena erabiliz aztertu, ez modu arrazional batean, behintzat.

Kontuan izan behar da, zenbait hizkuntzalarik *corpus* hitza edozein testu-bilduma adierazteko erabiltzen duten arren, artikulu honetan erabiltzen denean, kontu handiz hautatutako testu-bildumatzat hartu behar dela, hizkuntza edo aldaera baten ahalik eta erakusgarri doiena izan nahi duen zerbait alegia.

Hona hemen corpusen aplikazio nagusia:

- Hizkuntzalarien artean: lexikografia eta terminologia izan ohi dira aplikazio-eremu aipatuenak; adibidez, hitzen erabilera aztertzeko eta haztatzeko erabil daitezke corpusak; lexiko orokorrerako zein espezializatu baterako erabakiak hartu aurretik ezinbestekoa baita erabilera ezagutzea. Corpusak elebidunak edo eleaniztunak badira, hiztegi edo glosario eleaniztunak sortzeko oinarrizko langai aproposa izan daitezke. Horrez gain, corpusek dialektologia, fraseologia, estilistika eta antzeko diziplinetan ematen dituzten datuek berebiziko garrantzia dute.

- Zalantzak kontsultatzeko tresna interesgarriak dira.

- Hizkuntza-ingeniaritzan, berriz, ezinbesteko langai dira hainbat tresna egiteko eta ebaluatzeko: ortografia-zuzentzaileetatik itzulpen automatikoraino, aplikazio guztietan erabiltzen dira corpusak, programen datuetarako zein emaitzen kalitatea neurtzeko. Are funtsezkoagoak dira gaur egun hainbeste erabiltzen diren metodo estatistikoak eta ikasketa automatikoa aplikatu ahal izateko.

Estandarrak: XML etiketatzea, TEI, ...

Corpusak etiketatzea da, beraz, arlo honetako egiteko garrantzitsu bat. Nola etiketatu corpusak? Zein informaziorekin?

Corpusak kodetzeko eta etiketatzeko proposatu diren ereduen eta formatuen artean, TEI eredua eta XML teknologia dira estandarizazioaren bidean arrakastatsuenak. TEI (*Text Encoding Initiative*) eta bereziki TEI P4 gidalerroen multzoa, nazioarteko estandar bat da, testu elektronikoak kodetzeko eta trukatzeko orientabideak proposatzen dituena (Arriola *et al*., 1997). Erabiltzaile askoren premiak betetzera datoz TEI P4 gidalerroak: zientzia eta giza arloko ikertzaile, argitaratzaile, bibliotekari, eta, oro har, dokumentuen bilaketa eta biltegiratzearekin zerikusia duten guztienak. Erantzun bat ematen dio, orobat, hizkuntzaren teknologiaren arloko jendeari, orotariko testu-corpus eta lexikoak biltzeari eta metatzeari emanak baitaude azken aldi honetan, hizkuntzaren ulerkuntzaren, sorkuntzaren eta itzulpenaren ikerkuntzan behar-beharrezkoak baitira horrelakoak. Kontuan hartzekoak dira, orobat, EAGLES (CES, XCES) eta ISO TC37 SC4 lantaldeen ekimenak eta proposamenak ere, corpusen kodeketa estandarizatzeko bidean elkarpen garrantzitsuak baitira.

TEI gidalerrootan jasotzen diren etiketatze posible guztien artean, honako hauek azpimarratuko

ditugu:

- egiturazko informazioa: normalean, dokumentuan bertan esplizituki dagoen informazioa izaten da, baina digitalizatzean informazio hori etiketatzen ez bada, edo galdu egiten da edo testuarekin nahastu. Besteak beste, honako informazio hauek etiketatu ohi dira: informazio dokumentala (egilea, data, generoa, hizkuntza edota dialektoa, etab.); dokumentuen egitura (titulu eta azpitituluak, paragrafo, esaldi, zerrenda, aipamenak etab.); formatu-ezaugarriak (letraren tamaina eta estiloa, komatxoak, irudien kokapena etab.)

- hizkuntza-informazioa: berez inplizitua da, baina corpusak ustiatzeko garrantzi handia du informazio hori esplizituki etiketatzeak. Konplexutasunaren arabera, honako informazio hau esplizita daiteke etiketen bidez: informazio lexikala (hitz bakoitzaren lema edo erroa, kategoria gramatikala, informazio morfologikoa, hitz anitzeko unitateak, datak, zenbakiak, erakundeak, pertsona edo toki-izenak, lokuzioak...); sintaktikoa (esaldi bakoitzaren egitura sintaktikoa); semantikoa (lemaren adiera hiztegi edo ontologia baten arabera), eta pragmatikoa (anaforaren edo korreferentziaren ebazpena, adib.)

Bistan dena, hori guztia etiketatzea lan handia da, eta nahiz eta modu erdiautomatikoan egiten den, programa informatikoen laguntzaz, oso garestia izaten da corpus etiketatuak eratzea, hizkuntza-informazioa sartzen bada batez ere. Hori dela eta, corpus handietan hizkuntza-informazio gutxi sartzen da, gehienetan *POS tag* delakoa (kategoria gramatikala) eta lemarekin osatutako informazio lexikala besterik ez. Informazio sintaktiko, semantiko zein pragmatikoa corpus espezializatu txikiagoetan etiketatu ohi dira.

Corpusen eraketa

Corpusa nolanahi bildutako testu-multzo hutsa izango ez bada, corpusgintza gidatuko eta egituratuko duen eredu bat da beharrezkoa. Corpusgintzan lau urrats nagusi bereizi ohi dira:

- Diseinua: corpusaren helburuak eta ezaugarriak zein izango diren, testuak zein irizpideren arabera corpuseratuko diren, testuak zein mailatan prozesatuko eta etiketatuko diren…

- Corpus gordina eratzea: corpusean sartuko diren testuak eskuratzea eta corpuserako hautatu den formatura bihurtzea

- Etiketatzea: corpusa osatzen duten testuei buruzko informazioa (metadatuak), egitura, formatu-ezaugarriak, informazio linguistikoa (lema, kategoria…)

- Corpusak analizatzeko eta ustiatzeko tresnak: corpusaren kontsulta diseinatzea eta inplementatzea

Corpus nazionalak

Garestiak izan arren, corpus etiketatuak oso baliabide ahaltsuak dira, eta ikerkuntzarako azpiegitura-inbertsiotzat hartu behar dira. Horren erakusgarri dira mundu osoan hizkuntza ahaltsu eta ez hain ahaltsuetarako eratu diren erreferentzia-corpus nazionalak, hizkuntza jakin baterako lantegi aurreratua izan nahi dutenak.

Corpus nazionalak corpus etiketatuak dira, eta, lehen esan den bezala, neurriarengatik, hedadurarengatik eta etiketatze-lan sakonarengatik hizkuntza baten erreferentzia nagusi izateko helburua dute. Beraz, horrelako corpusetan neurria garrantzitsua da, handiak izan behar dute, bestela neurri estatistikoak ez baitira adierazgarriak izango, eta hainbat fenomeno linguistiko ez dira behin ere agertuko. Ezaugarri horien guztien ondorioz, oso proiektu estrategikoak eta garestiak dira, eta, ezinbestez, administrazio publikoak lagunduak izan behar dute.

Mundu-mailako erreferentzia nagusia *British National Corpus* (BNC, www.natcorp.ox.ac.uk) delakoa dugu, oso zabalduta dagoena. 4000 testu-zati ditu, eta etiketatutako 100 milioi hitz biltzen ditu. TEIren gidalerroei jarraitzen die eta, corpus nazional gehienetan bezala, egiturazko informazioa eta informazio lexikala du etiketatua. Oxford unibertsitateak kudeatzen eta banatzen du, eta *Xaira* izeneko tresna ere eskaintzen du.

AEBetako ingelesaren ordezkari gisa, *American National Corpus* (ANC, americannationalcorpus.org) izeneko bilduma dugu, 22 milioi hitz etiketatu biltzen dituena.

Gertuago ditugu frantsesaren eta espainolaren erreferentzia-corpusak. Frantsesezkoak FRANTEXT izena du (www.frantext.fr/categ.htm), eta 150 milioi hitzek osatzen dute. *Institut National de la Langue Française*-k garatua da eta, harpidetzaren bidez banatzen da. Gaztelaniarako bi bilduma handi dago, biak ere *Real Academia Española*-k kudeatuak, eta Espainiako zein Ameriketako datuak jasotzen dituztenak: CREA izenekoa (corpus.rae.es/creanet.html), gaur egungo hizkuntza jasotzen duena (1975etik gaur arte), eta CORDE izenekoa (corpus.rae.es/cordenet.html), diakronikoa, eta hizkuntzaren bilakaera aztertzeko balio duena. Bakoitzak 100 milioi hitz etiketatu baino gehiago ditu, eta biek kontsulta publikoa eskaintzen dute.

Hiztun-kopuruan euskaratik hurbilago dauden hainbat hizkuntzatan ere corpus nazionalak garatu dira, hala nola eslovakiera, txekiera, gaelikoa, galiziera, katalana, etab.

Txekieraren kasuan, adibidez, oso corpus nazional zabala dute, alde sinkronikoan 100 milioi hitz dituena (SYN2000, ucnk.ff.cuni.cz). Gaelikoarena, berriz, 15 milioi baino ez du (www.ite.ie/pos.htm).

Galizian CORGA corpusa (*Corpus de Referencia do Galego Actual* corpus.cirp.es/corgaxml) garatu da 13,3 milioi hitz etiketaturekin eta Katalunian CTILC izenekoa (*Corpus Textual Informatitat de la Llengua Catalana* pdl.iec.es/entrada/paraules.asp) 52 miliorekin.

Corpusa	**Hitz-kopurua**	**Hizkuntza**
SYN2000	100 milioi.	Txekiera
British National Corpus	100 milioi.	Ingelesa
FRANTEXT	150 milioi.	Frantsesa
CRAE	130 milioi.	Gaztelania
CORDE	136 milioi.	Gaztelania
ANC *CTILC* *CORGA*	22 milioi. 52 milioi 13 milioi	Ingelesa (AEB) Katalana Galegoa

Gai honetan sakontzeko *Corpus Survey* (Corpus 2005) gunea oso interesgarria da, beste hainbat hizkuntzatako corpus nazionalen berri ematen duena (poloniera, hungariera, errusiera, greziera, eslovakiera, txinera, kroaziera, daniera, nederlandera, norvegiera eta abar)

Bestalde, hainbat erakunde daude corpusak eta beste hizkuntza-baliabideak zentralizatzeko eta banatzeko. Europan, ELDA (www.elda.org) da erreferentziazko erakundea arlo horretan, eta *American Linguistic Data Consortium* erakundea (www.ldc.upenn.edu) AEBetan.

Lehen esan den bezala, hizkuntza-informazio sakona esplizituki etiketatuta daukaten bestelako corpusak ere sortu dira, txikiagoak neurriz, baina interesgarriagoak zenbait aplikaziotarako. Honako egitasmo hauek aipa daitezke mota honetako corpusen artean: informazio sintaktikoa jasotzen duten *Penn Treebank* eta *Prague Dependency Treebank*, eta informazio semantikoa jasotzen duten *SemCor* eta *PropBank*. Corpus eleaniztunak ere badaude (MULTEXT edo PAROLE izenekoak esaterako), eta horietako batzuk itzulpengintzan erabiltzeko diseinatuta dauden corpus paraleloak dira (*Babel* txinera-ingelesa corpusa, adibidez).

Euskarazko corpusak

Euskararen kasuan, hainbat corpus eratu dira. Corpus gordinen artean azpimarratzekoa da Susak (www.susa-literatura.com) egindako bidea. Bertan aurki daitezke *Klasikoen Gordailua,* literatura klasikoa jasoz, eta *Ibinagabeitia Proiektua* prentsa historikoaren bilduma.

Horrez gain, azpimarratzekoa da Euskaltzaindiaren *Orotariko Euskal Hiztegia*rekin lotutako testu-corpusa. Egileek diotenez:

> *Hiztegi honen xedea garai eta leku guztietako euskal hitzen ondarea deskribatzea da; ahalik eta osoena izan nahi du. Antzinateko inskripzioetatik hasi, Erdi Aroko testuetatik jarraitu eta inprentaren sorreratik XVIII. mendearen erdiraino argitara emandako guztia hartzen du. Hortik aurrera argitaratutakoaren bolumena gure egunotara hurbildu ahala handituz doanez, corpusa ez zitekeen izan orohartzailea, baina zalantzarik gabe esan daiteke 1970. urtea arteko euskara idatziaren oso erreferentzia osatua dela OEHren corpusa. 6.000.000 hitzek osatutako testuak ditu gutxi gorabehera orotara.*

Zoritxarrez, corpus hori ez dago ez lematizatua ez etiketatua, eta ikerkuntzarako banatu den CD batean dago eskuragarri.

Corpus etiketatuen alorrean, berriz, hiru corpus aipatu behar dira: *XX. mendeko euskararen corpus estatistikoa*, *Ereduzko Prosa* izenekoa eta *Zientzia eta Teknologiaren corpusa*. Hirurak daude sarean kontsultagai, eta bilaketa-sistema aurreratua dute. Lehena orokorra, orekatua eta diakronikoa da; bigarrena, berriz, gaur egungo literatura eta prentsa biltzen du, oreka bilatu gabe. Zientzia eta Teknologiaren corpusa da berriena, espezializatua da eta orekatua izateko diseinatuta dago, eta TEI estandarreari jarraituz etiketatuta dago.

XX. mendeko corpusa

Euskaltzaindiaren *XX. mendeko euskararen corpus estatistikoa* (Urkia, 2002), UZEIk kudeatu du, eta erreferentziazkotzat har daitekeen bakarra da gaur egun. Hasiera batean, proiektuak EEBS-*Egungo Euskararen Bilketa-lan Sistematikoa* zuen izena. Egileen hitzetan, hauxe da helburua:

> *... Erabili izan den eta erabiltzen den euskararen lekuko eta erakusgarri izatea du egiteko nagusi eta ia bakarra, eta ez ereduzko hizkuntza proposatzea. ...*
>
> *... euskararen ikerle ororentzat baliagarri izango diren hizkuntza-datutegiak kontsultagai jartzea da gure asmoa.*

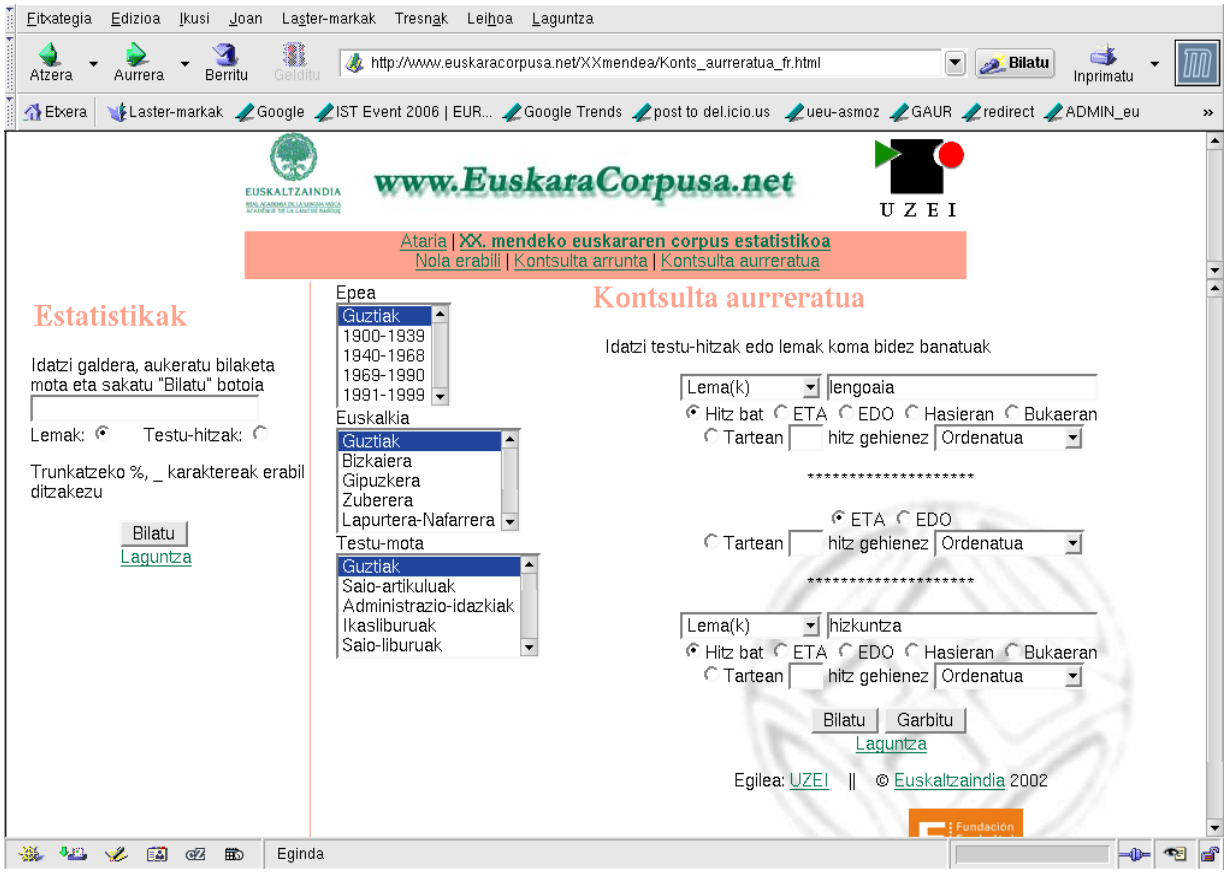
XX. mendeko testuak biltzen dira, mende horretako euskara idatziaren erakusgarritzat. 4.650.000 hitz guztira, 6.351 idazlanetatik hartuak. Lematizatuta dagoenez, lema-kopurua ere ematen dute: 101.585 (hitz anitzeko zenbait hizkuntza-unitate ere lema dira, hala nola marraz lotutako hitz-elkarteak, aditz-elkarteak, eraikuntza sintagmatiko batzuk…). Etiketatzean, XMLren aurrekaria izan zen SGML estandarra erabili da.

Corpus orekatua izan dadin hainbat irizpide hartu dira kontuan corpusa eratzerakoan:

- garaia: lau garai bereizi dira: 1900-1939, 1940-1968, 1969-1990, 1991-1999
- euskalkia: euskalki nagusiak eta euskara batua bereizi dira
- testu-mota: saio-artikuluak, administrazio-idazkiak, ikasliburuak, saio-liburuak, literatura-prosa, poesia, antzerkia, bertsoak, ikerketa-lanak, haur- eta gazte-literatura, ahozkoak, liturgia, egunkariak eta aldizkariak

Informazio gehiago eta corpusa kontsultatzeko intrfazea: www.euskaracorpusa.net

Hurrengo bi irudietan, kontsulta aurreratuaren eta emaitzen adibide bana jaso dugu:


Fitxategia Edizioa Ikusi Joan Laster-markak Tresnak Leihoa Laguntza
Atzera Aurrera Berritu Gelditu
http://www.euskaracorpusa.net/XXmendea/Konts_aurreratua_fr.html
Bilatu Inprimatu
Etxera Laster-markak Google IST Event 2006 | EUR... Google Trends post to del.icio.us ueu-asmoz GAUR redirect ADMIN_eu
EUSKALTZAINDIA
www.EuskaraCorpusa.net
UZEI
Ataria | XX. mendeko euskararen corpus estatistikoa
Nola erabili | Kontsulta arrunta | Kontsulta aurreratua
Estatistikak
Idatzi galdera, aukeratu bilaketa mota eta sakatu "Bilatu" botoia
Lemak: Testu-hitzak:
Trunkatzeko %, _ karaktereak erabil ditzakezu
Bilatu
Laguntza
Epea
Guztiak
1900-1939
1940-1968
1969-1990
1991-1999
Euskalkia
Guztiak
Bizkaiera
Gipuzkera
Zuberera
Lapurtera-Nafarrera
Testu-mota
Guztiak
Saio-artikuluak
Administrazio-idazkiak
Ikasliburuak
Saio-liburuak
Kontsulta aurreratua
Idatzi testu-hitzak edo lemak koma bidez banatuak
Lema(k) lengoaia
Hitz bat ETA EDO Hasieran Bukaeran
Tartean hitz gehienez Ordenatua
****
ETA EDO
Tartean hitz gehienez Ordenatua
****
Lema(k) hizkuntza
Hitz bat ETA EDO Hasieran Bukaeran
Tartean hitz gehienez Ordenatua
Bilatu Garbitu
Laguntza
Egilea: UZEI || © Euskaltzaindia 2002
Fundación
Eginda


Fitxategia Edizioa Ikusi Joan Laster-markak Tresnak Leihoa Laguntza
Atzera Aurrera Berritu Gelditu
http://www.euskaracorpusa.net/XXmendea/Konts_aurreratua_fr.html
Bilatu Inprimatu
Etxera Laster-markak Google IST Event 2006 | EUR... Google Trends post to del.icio.us ueu-asmoz GAUR redirect ADMIN_eu
EUSKALTZAINDIA
www.EuskaraCorpusa.net
UZEI
Ataria | XX. mendeko euskararen corpus estatistikoa
Nola erabili | Kontsulta arrunta | Kontsulta aurreratua
Estatistikak
Idatzi galdera, aukeratu bilaketa mota eta sakatu "Bilatu" botoia
Lemak: Testu-hitzak:
Trunkatzeko %, _ karaktereak erabil ditzakezu
Bilatu
Laguntza
Lema(k): lengoaia
<ETA>
Lema(k): hizkuntza
Atzera
1. 1991-1999 Euskara Batua Saio-liburuak I. Mendiguren 0067
Orduan lengoaia berean behartsuen hizkuntza eta aberatsen hizkuntza aurkitzen dira, plebeioen hizkuntza eta nobleen hizkuntza, hizkuntza kultua eta hizkuntza arrunta.
2. 1991-1999 Euskara Batua Saio-liburuak ZuzenbHizkera 0020
Aipaturiko dimentsio komunikatiborik gabe, lengoaiaren izanak ez du inolako esangurarik, ezta lengoaiaren gauzatze praktikoa besterik ez den hizkuntzen izateak ere.
3. 1991-1999 Euskara Batua Literatur prosa M. Hoyos 0056
Batik bat beren hizkuntza horrek, marruaren antzeko lengoaia bitxi bezain ulertezin horrek sortu du nigan itsasoetako ugaztun hauenganako miresmena eta zaletasuna (ez ote zuten gauza bera pentsatuko euskaldunak konkistatzera etorri zirenek?).
Aurkituak: 3 testu
Atzera
Eginda

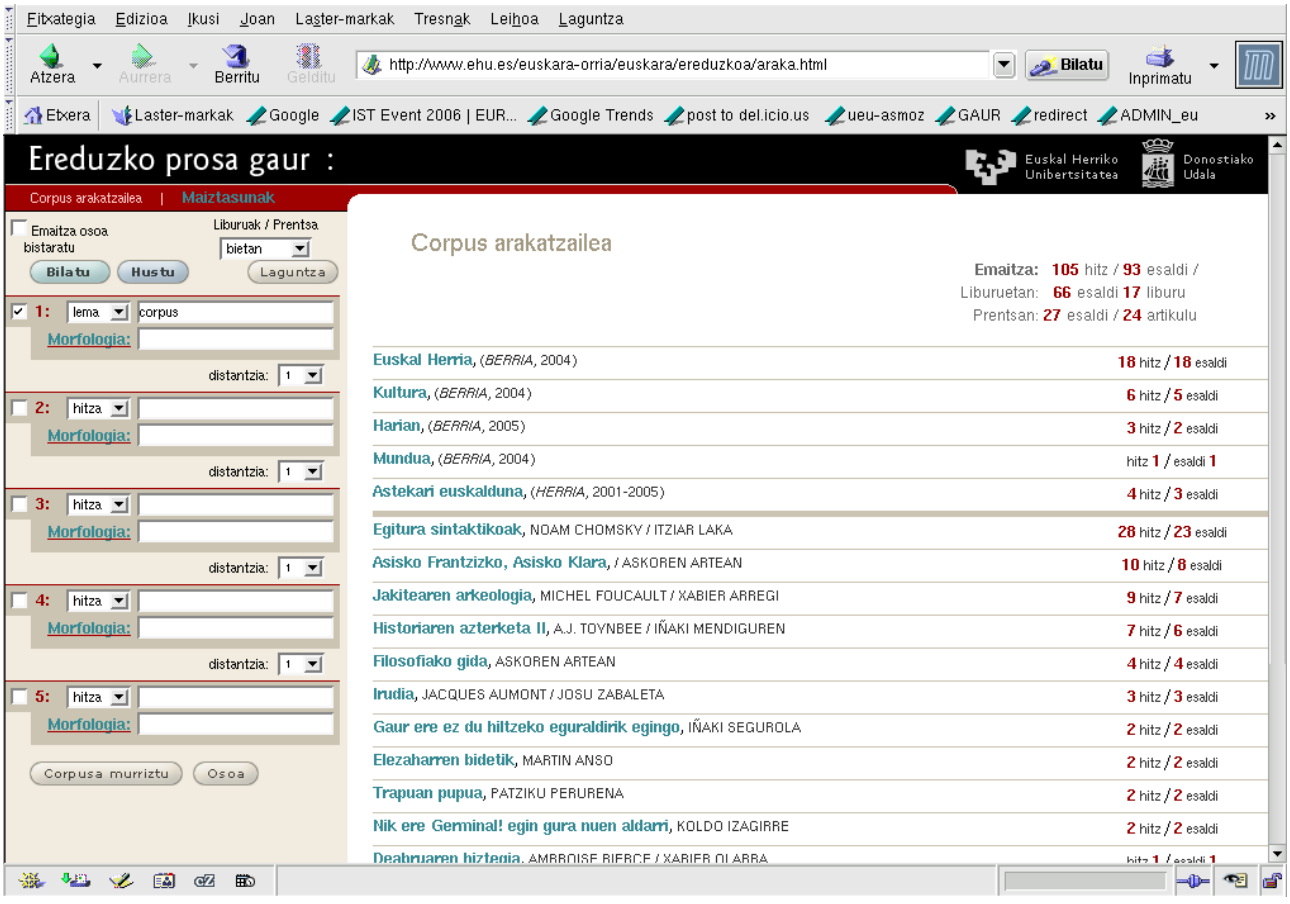
Ereduzko Prosa Gaur

Euskal Herriko Unibertsitateak eta Donostiako Udalak lankidetzan garatutako corpus honen helburua hauxe da, egileen arabera:

> *Ez dira gutxi, ezta unibertsitate munduan ere, artikuluak-eta taxutzean zalantzak eta duda-mudak dituzten idazleak, hitz forma egokienak, esapide aukerakoenak, joskera zuzenenak hautatzeko orduan. Egoera horri aurre egiteko EHUko Euskara Zerbitzuak honako lanabes hau eskaintzen dizu. Bertan, gaur egungo hainbat euskal idazle ereduzkoren azken urteotako testuak bildu ditugu, horiekin corpus aski zabal bat eratuz. Corpus horri etekinik beteena ateratzeko aztergailu ahaltsu eta erabilerraz bat erantsi diogu. Horiek horrela, lanabes aski egokia duzu hau, gaurko euskal autore eredugarriak zure duda-mudei eman dizkien irtenbideak ezagutzeko.*

235 literatura-liburutatik eskuratutako 10,2 milioi hitzek eta prentsatik (*Berria* egunkaria eta *Herria* aldizkaria) eskuratutako 9,6 milioi hitzek osatzen dute corpus hau (www.ehu.es/euskara-orria/euskara/ereduzkoa). Azken urteetako euskara jasotzen da (2000-2006), orekatu gabea da, baina bolumen aldetik oso interesgarria da corpusa. Etiketatze aldetik lema eta kategoria daude jasota corpusean, baina modu erabat automatikoan; beraz, emaitza ez da beti nahi bezain zehatza.

Hona hemen kontsulta-interfazearen adibide bat:



ZT corpusa

Zientzia eta Teknologiaren corpusa (Alegria et al., 2005) corpus berezi edo espezializatua da, eta euskaraz 1990-2002 bitartean argitaratu diren zientzia eta teknologiaren alorreko obren bilduma adierazgarria izatea du helburutzat. Bi data horiek bat datoz, hurrenez hurren, Euskaltzaindiaren araugintza berriaren hasierarekin, eta proiektuaren hasierarekin berarekin.

Elhuyar Fundazioaren eta EHUko IXA taldearen artean garatu da corpus hau, Hizking21 proiektu

zabalaren barruan. Proiektuaren helburu nagusiak honako hauek dira:

- Zientzia eta Teknologiaren alorrean euskaraz idatzi denaren lagin adierazgarria biltzea. Horrekin lantegi interesgarri bat lortu da ikerketa linguistikoari begira zein hizkuntza-ingeniaritzari begira
- Gainerako euskarazko corpusekin osagarria izatea, balizko corpus nazional bati begira
- Corpusgintzan aritzeko tresna-multzo ahaltsu, estandar eta integratua garatzea

Corpusa bi ataletan antolatuta dago. Batetik, adierazgarria izateko asmotan diseinatu den eskuz landutako gunea; eta bestetik, eskuragarritasunaren arabera corpuseratu diren obrez edo obra-zatiez osatutako atal irekia. Hain zuzen ere, eskuz landutako gunean ez dira obra osoak sartzen, obren lagin etenak baizik. Horrek berekin dakar gune horren obra baten pasarte ez hautatuak (lagin eten horien artekoak), eskura izanez gero, corpusaren atal irekian sar daitezkeela (laginerako hautatu ez diren baina eskura dauden obrekin batera).

Gune orekatuan zein obra sartu behar den eta obra bakoitzetik zein testu-masa eta zein pasarte sartuko diren ere irizpide jakin batzuen arabera erabaki da. Horretarako, lehenik 1990-2002 bitarteko zientzia eta teknologiaren alorreko obren inbentarioa egin da. Hurrena, adierazgarritasuna edo 'oreka' bermatuko duen lagintze-eredu estatistikoa landu da, obren *eremuan* eta *generoan* oinarrituta. Jakintza-alorrak eremuaren arabera sailkatu dira, eta testu-mota erregistroaren arabera:

- Eremuak: zientzia zehatzak, materiaren eta energiaren zientziak, lurraren zientziak, biziaren zientziak, teknologia, orokorra eta bestelakoak.
- Generoak: oinarrizko hezkuntzako materiala, goi-mailako liburua, artikulu espezializatua, dibulgazio-artikulua, dibulgazio-liburua eta administrazio publikoko dokumentua.

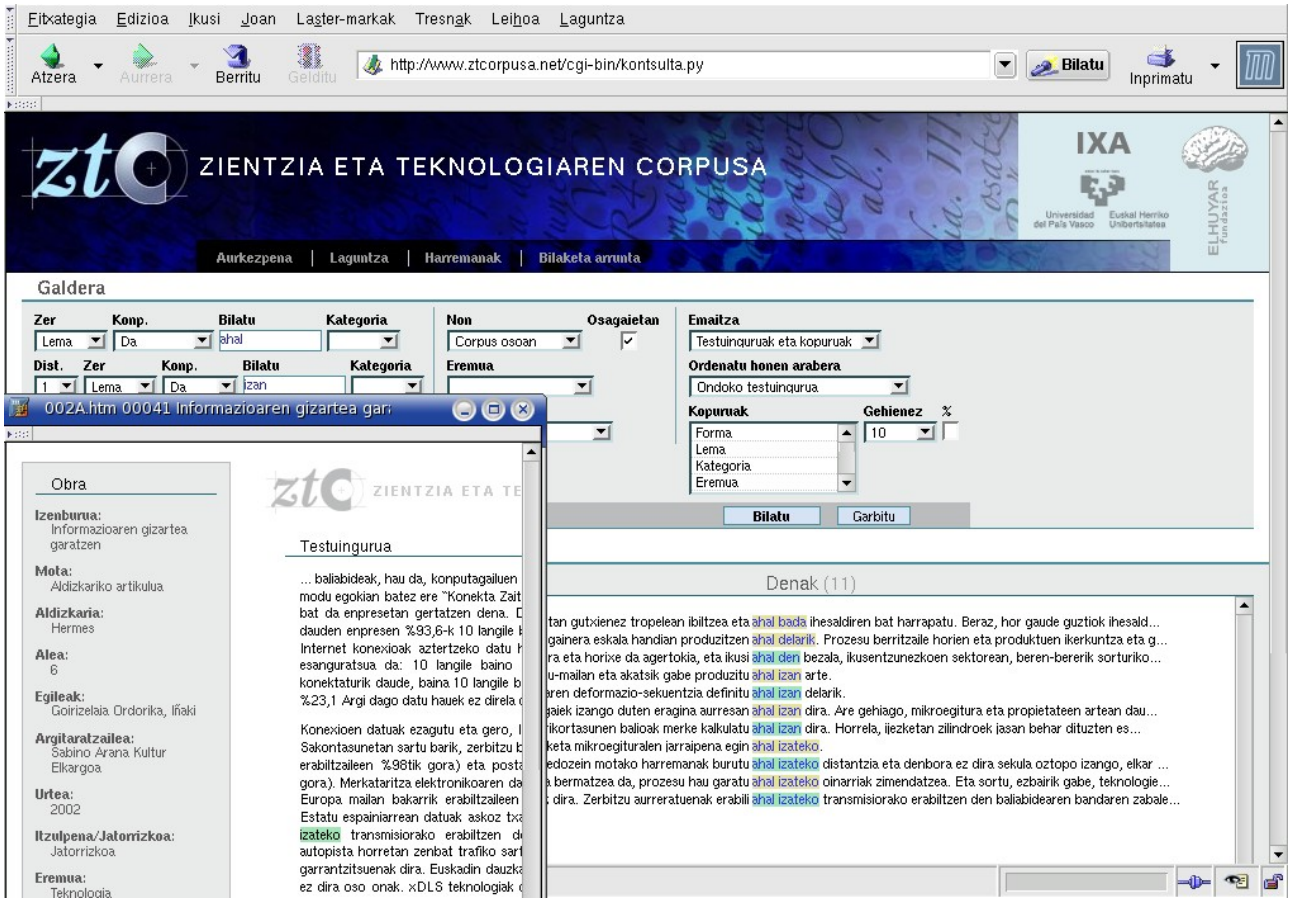
Adierazgarritasuna bermatzeko, kalkulatu da gune orekatuaren tamainak 5 milioi hitzekoa behar lukeela izan. Hala ere, corpusaren lehen bertsioan (*ZT corpusa 1.0*), gune orekatuan 1,6 milioi hitz

daude. Guztira, atal irekia barne, bertsio horrek 8 milioi hitz ditu.

Corpusa etiketatzean, gune orekatuko laginak automatikoki prozesatu dira lehenik, eta gero eskuz landu dira, etiketatze-lana, egiturazkoa zein linguistikoa, aberasteko, zuzentzeko eta desanbiguatzeko. Gune orekatukoak ez diren testu-zatiak, berriz, automatikoki baizik ez dira prozesatu, baina prozesamendu hori gune orekatuko lanak amaitutakoan egin da, sistemak eskuz landutakotik 'ikasi' duena aplika dezan, etiketatze automatiko hobea lortzearren.

XMLn etiketatuta dago, TEI gidalerroei jarraituz. Aurrekoak ez bezala, ikerkuntzarako eskuragarri dago, eta 2007tik aurrera, *ELDAren baliabideen artean egongo da, ustiapen komertzialerako eskuragarri, lizentzia bidez.*

ZT corpusaren lehen bertsioa Interneten ere kontsulta daiteke, www.ztcorpusa.net helbidean. Bere ezaugarriak, corpusa osatzen dituzten idazlanak eta antzekoak bertan kontsulta daitezke. Bertatik eskuratu dira irudiak:



Euskarazko corpus nazionalerantz?

XX. mendeko corpusa da euskarazko corpusa izan zitekeenetik gertuen dagoena, baina neurriarengatik eta, batez ere, eskuragarritasun-ezagatik, ez da nahikoa. Hainbat eragilek egin dute lehenago geure ere egiten dugun eskaera: euskarazko corpus nazionala garatzea.
Corpus horrek neurrian eta eskuragarritasunean aipatutako corpus nazionalekin parekagarria izan beharko luke; beraz, 50 milioi hitzetik gora jaso beharko luke, eta ikerkuntzarako doan banatu beharko litzateke, kontsultagai egoteaz gain noski. Arlo honetan aritzen garenon arteko lankidetzan egingo balitz, merkeago eta arrakastatsuago litzatekeela uste dugu. Gainera, etorkizunari begira, monitore izan beharko luke, etengabe eguneratzen den corpusa alegia.
Eusko Jaurlaritza prestatzen ari den teknologia-planaren barruan aztergai dagoela jakinda, espero dezagun laster abian ikusiko dugula halako egitasmo bat.

Corpusak kudeatzeko eta etiketatzeko tresnak

Corpusak kudeatzeko tresnen adibide gisa, Zientzia eta Teknologiaren corpusa eratzeko erabili ditugunak azalduko ditugu.

Corpusaren kudeaketa eta egiturazko etiketatzea: Corpusgile

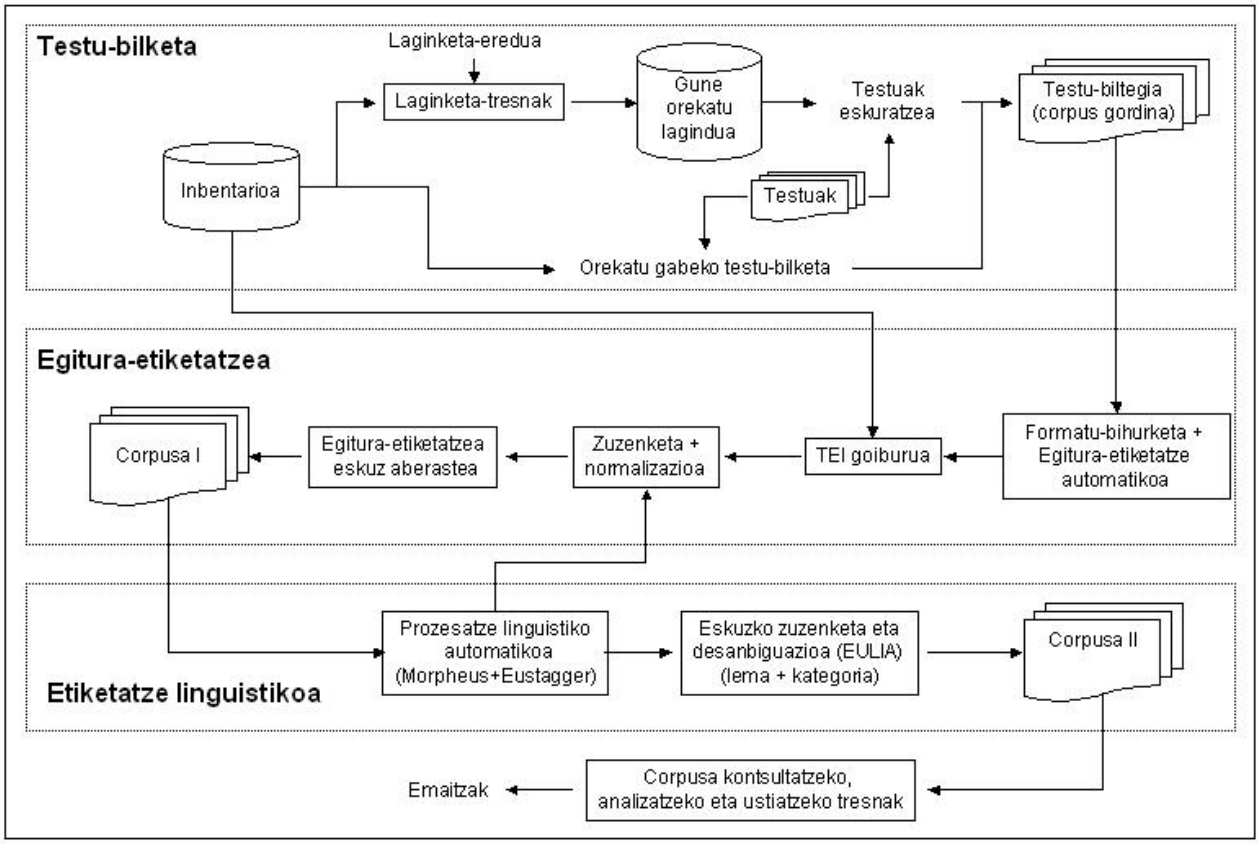
Corpusgintza-ereduko urratsak modu sistematiko eta egituratuan egiteko, corpus-metodologia bat landu behar da, eta, hori inplementatzeko, corpusgintza-tresna bat. Lehendik zeuden tresnak eta proiektu honetarako garatuak integratuz, *Corpusgile* aplikazioa sortu dugu. Corpus gordina eratzea eta etiketatze-lanak dira kudeatu behar dituen prozesu giltzarriak. Batetik, IXA taldeak euskara automatikoki prozesatzeko garatutako tresna batzuk (*Eustagger*, *Eulia*) moldatu eta areago garatu ditugu, eta, horrekin batera, corpusgintza bera kudeatzeko eta, oro har, corpus-lanak egiteko beharrezkoak diren tresnak ere sortu behar izan ditugu. Kontuan hartu behar da merkaturatu diren corpusgintza-tresna urriek ez dutela euskararen prozesamendu automatikorako beharrezkoak diren tresnak eta baliabideak integratzen, eta ez direla egokiak euskarazko testu-corpusak eratzeko. Halaber, *Corpusgile*-ren bidez corpusgintzaren etorkizuneko helburua den erreferentzia-corpus orokorra egiteko baliagarria izango den metodologia adostua eta kontrastatua eskaini nahi izan da.

Corpusgile hiru moduluz osatua da:

- TB: testu-bilketaren modulua (corpus gordina biltzeko modulua)
- EE: egitura-etiketatzea egiteko modulua
- EL: etiketatze linguistikoa egiteko modulua

Lehen bi moduluak Elhuyar Fundazioak garatu ditu; hirugarrena, berriz, IXA taldearen garapena izan da.

Diagrama honetan bildu ditugu urrats horien eta horien barneko prozesu nagusiak:



Formatu elektronikoan jasotzen dugunean, jatorrizko dokumentuaren formatu hauek onartu ditugu: *html*, *xml*, *doc*, *rtf*, *txt*, *pdf* eta *qk*. Horietako formatu batzuek arazoak sortzen dituzte formatu-bihurketa automatikoa egiteko. Bestetik, formatua bihurtzean jatorrizko formatu-ezaugarri batzuk gordetzea eta automatikoki prozesatzea interesatzen zaigu. Adibidez, egitura etiketatzean ikusiko dugu letra-estiloa (etzana, lodia…) atxikitzea interesgarria dela; beste hainbeste testuaren egiturari buruzko informazioa ematen duten estiloez (esaterako, *MS-Word*-en erabiltzen diren 'atalburua', 'buletdun zerrenda', eta abar).

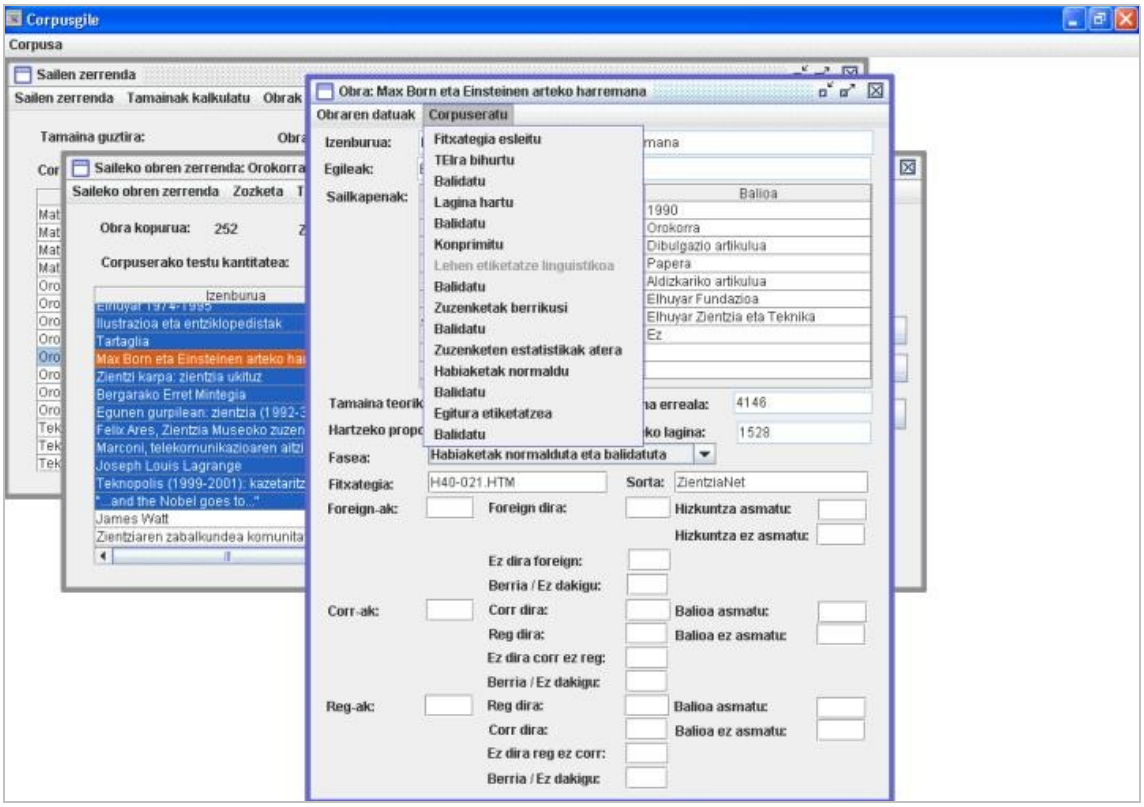
Corpusak kodetzeko eta etiketatzeko proposatu diren ereduen eta formatuen artean, TEI eredua eta XML teknologia hautatu ditugu. TEIk aukera ugari eskaintzen ditu testuak etiketatzeko.

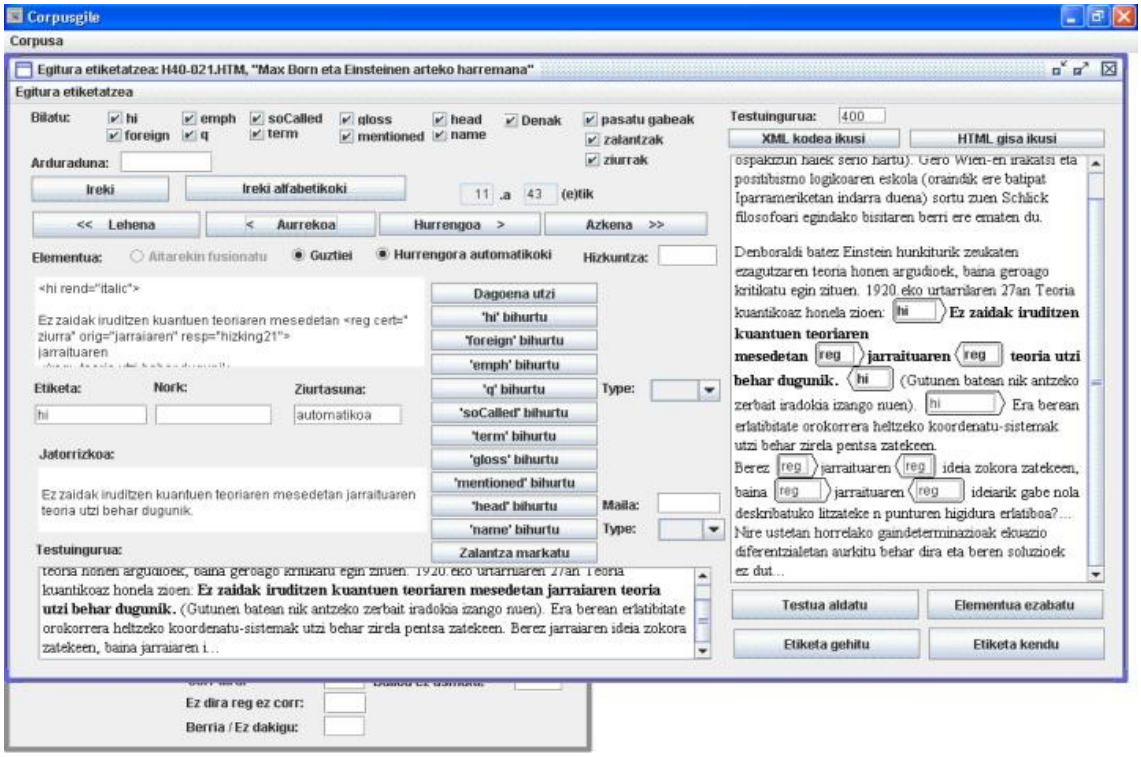
ZT corpusean, testuen egitura (atalburuak, atalak, azpiatalak, paragrafoak, zerrendak, taulak...) eta formatu-ezaugarri zenbait markatzea erabaki dugu. Egitura-elementuak hauek dira: `<text>`, `<body>`, `<div>`, `<head>`, `<p>`, `<table>`, `<row>`, `<list>` eta `<item>`. Testuaren joskeraren barnean irudi bat edo corpuseratuko ez den bestelako elementuren bat dagoenean (formulak, ekuazioak...), `<gap>` elementu hutsaren bidez adierazten dugu gune horretan zerbait 'falta' dela.

Etiketatze linguistikoen emaitzak hobetze aldera, zuzenketak eta aldaera ez-estandarrak etiketatzeko lana ere egiten dugu urrats honetan. Horretarako, `<corr>` eta `<reg>` elementuak erabiltzen dira. Etiketatzaileak `<corr>` edo `<reg>` proposamenak automatikoki markatzen ditu testuan, eta gero horiek denak eskuz aztertzen dira, balioesteko edo behar diren aldaketak egiteko.
Bestetik, testuaren ezaugarri tipografiko linguistikoki esanguratsuak (nabarmentzeak) automatikoki jasotzen dira, <hi> elementuaren bidez: letra-estiloak (lodia, etzana, azpimarratua...), komatxoak (bikoitzak, bakunak...)...; gune orekatuan, nabarmentzeak desanbiguatu egiten dira, hau da, nabarmentzeei balioa edo funtzioa esleitzen zaie `(<foreign>, <emph>, <soCalled>, <q>, <term>, <mentioned>, <name>...)`,

Azkenik, corpuseko obra bakoitzaren metadatuak obraren goiburuan (`<teiHeader>` elementuan) bildu ditugu (ISBN zenbakia, izenburua, egilea, argitaratze-urtea, argitaletxea, eremua, generoa...). Metadatu horiek inbentarioaren datu-basetik zuzenean ekartzen dira goiburura.

Ondoko irudian *Corpusgile* tresnaren interfaze nagusiak azaltzen dira, testu-bilketarena eta egitura-etiketatzearena hurrenez hurren.





Etiketatze linguistikoa: Eulia

Corpusa baliabide linguistikoa izango bada, ezinbestekoa da linguistikoki prozesatzea eta etiketatzea, alegia, corpuseko hitzak informazio linguistikoz aberastea.

Lematizazioan aipatutako desanbiguatze hori automatikoa da (ez da %100 zuzena, beraz), eta horren emaitza izango da corpuseko atal irekian geratuko dena. Gune orekatuan, ordea, eskuz berrikusiko dira emaitzak, eta, prozesua burututakoan, gune hori anbiguotasunik gabe eta erabat

zuzen lematizatua geratuko da.

Prozesu hauetan guztietan erabiltzen den informazio lexikala EDBL datu-base lexikaletik dator [Aldezabal et al., 2001]. EDBL lexiko-biltegi iraunkorra da, eta aparteko prozesu baten bitartez gobernatzen da. Emaitzen doitasuna handitzeko asmoz, EDBLko lexikoari erabiltzailearen lexiko partikular bat gehitu dakioke. Horretarako aurreprozesatze bat egiten da, arazoak ematen dituzten hitzak detektatzeko eta proposatzen diren lemen maiztasunaren arabera sailkatzeko. Zerrendaren buruan geratu diren lemak eskuz aztertu eta, egokitzat hartzen direnean, erabiltzailearen lexikoan barneratzen dira.

Testuak linguistikoki etiketatzeko, bi hurbilpen nagusi jarraitu ohi dira historikoki. Batean, informazio linguistikoa jatorrizko corpusean txertatzen da, hitzekin batera, orain arte ikusi ditugun etiketak bezala (`<text>`, `<body>`, `<hi>` etab.) erabiliz. Bestean, berriz, informazio linguistikoa hitzak dauden dokumentu nagusietatik at gordetzen da, horretarako berariaz sortutako dokumentuetan, alegia. Hitzak dagokien informazio linguistikoarekin lotzeko, bestalde, estekak erabiltzen dira. Azken hurbilpen horri anotazio banatua (*stand-off annotation* edo *markup*) esaten zaio, eta horixe erabili da gurean corpusa linguistikoki etiketatzeko.

Labur esanda, honako eragiketa hauek egin dira testuon gainean, *Eulia*-ren bitartez:

- Tokenizazioa: testua token edo analisi-unitatetan bereizi, puntuazio-ikur, maiuskula-minuskula, ezaugarri ortotipografiko eta abarren tratamendua eginez.

- Segmentazio morfologikoa: tokenak morfematan zatikatu, eta morfema bakoitzari dagozkion ezaugarriak esleitu. Prozesu honetan azaltzen da, estreinakoz, anbiguotasunaren arazoa, hitz-forma bat morfologikoki modu desberdinetan segmentatu ahal izango baita, eta, ondorioz, interpretazio bat baino gehiago izango dugu (kontuan hartu behar da segmentazioa testuingurua aintzat hartu gabe egiten dela, automatikoki).

- Analisi morfosintaktikoa: segmentazioaren emaitzatik abiatuz, hitz-formari dagokion lema osatu behar da (eratorpenaren kasuan, adib., oinarriari dagokiona aurrizki-atzizki lexikalekin elkartuz), eta forma osoari dagokion informazioa “goratzen” da morfema osagaien informaziotik (kasua, numero-mugatasunak, adib.).

- Hitz anitzeko unitateen tratamendua: hitz-forma solteen analisitik haratago, lexikalki unitatetzat har daitezkeen adierazpenak eta bestelako batzuk (entitate-izenak, data- eta zenbaki-adierazpenak, eta abar) ezagutzen dira fase honetan.

- Lematizazioa: prozesu honetan, bi aspektu bereizi behar dira: (1) geroko analisi-urratsetan –sintaxian, batik bat– pertinentea litzatekeen informazioa bereiztea: lema, kategoria-azpikategoriak, hitz-formaren kasua, numeroa eta mugatasuna, pertsona(k) adizkietan, funtzio sintaktikoa, erlazioa eta abar; (2) desanbiguazioa, hots, hitz-formari egoki dagokion interpretazioa zuzentzat markatzea (okerrak baztertuz), testuinguruari erreparatuz. Desanbiguazioa hizkuntza-ezagutzan oinarritua da (murriztapen-gramatika bat baliatuko da horretarako), alde batetik, eta estatistikoa, bestetik (ikaste automatikoko teknikak erabiliz, aldez aurretik eskuz desanbiguatutako corpus batean oinarrituz).

- Etiketatze linguistikoaren amaieran, corpuseko hitz orok zenbait informazio linguistiko izango du erantsita, hala nola: hitzaren lema eta kategoria lexikala (% 100 zuzen, eskuz desanbiguatutako atalean, eta automatikoki esleitutakoa, gainerakoan); hitzak duen kasua eta betetzen duen funtzio sintaktikoa (automatikoki esleituak); hitz anitzeko unitateen kasuan, unitate hauen egitura ere esplizitu errepresentatuko da, ezagutu direnen kasuan, jakina.

Etiketatze linguistiko automatikoa egindakoan, emaitzak eskuz lantzeko aukera dago. Lan hori corpusaren gune orekatua osatzen duten testuetan egiten dugu *Corpusgile*-ren EL moduluaren

bidez.

Bibliografia

Aldezabal I., Ansa O., Arrieta B., Artola Zubillaga X., Ezeiza A., Hernández G., Lersundi M. 2001. "EDBL: a General Lexical Basis for the Automatic Processing of Basque". *IRCS Workshop on linguistic databases*. Philadelphia (USA).

Alegria, I., Areta, N., Artola, X., Díaz De Ilarraza, A., Ezeiza, N., Gurrutxaga, A., Leturia, I., Saiz, R., Sologaistoa, A., Soroa, A. & Valverde, A. 2005. "Zientzia eta teknologiaren corpusa." In *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa.* Bilbo.

Alegria, I., Areta, N., Artola, X., Díaz De Ilarraza, A., Ezeiza, N., Gurrutxaga, A., Leturia, I., Saiz, R., Sologaistoa, A., Soroa, A. & Valverde, A. 2006. "Structure, Annotation and Tools in the Basque ZT Corpus." In *LREC 2006*. Genoa [http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1141404023/publikoak/pdf, 06-10-23an irakurria].

Arriola, J., Artola, X., Gojenola, K. & Soroa, A. 1997. "TEI: testu-kodeketarako gidalerroak." In *Ekaia*: *Euskal Herriko Unibertsitateko Zientzi eta Teknologi aldizkaria*, 7. zenbakia. Udazkena.[http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911707/publikoak/97EKAIA.ps ; 06-02-10ean irakurria]

Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Sologaistoa, A. & Soroa, A. 2004. "EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora." In *LREC 2004. Workshop on XML-based richly annotated corpora.* [http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1088448358/publikoak/04LREC_EULIA.pdf; 06-02-10ean irakurria]

Corpus Survey. 2005. [http://bowland-files.lancs.ac.uk/corplang/cbls/corpora.asp 06-09-25ean irakurria]

Euskara Corpusa. 2002. [http://www.euskaracorpusa.net/ 06-09-25ean kontsultatua]

Euskarazko Prosa Gaur. 2005 [http://www.ehu.es/euskara-orria/euskara/ereduzkoa/ 06-09-25ean kontsultatua]

Leech, G. 2002. "The Importance of Reference Corpora." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI. [http://www.uzei.org/corpusajardunaldia/06_gleech.pdf; 06-02-10ean irakurria]

Oihartzabal, B.. 2002. "Euskaltzaindiaren Corpusez" In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI [http://www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos60A.ppt; 06-02-10ean irakurria]

Urkia, M. 2002. "XX. mendeko euskara-corpusa." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI [http://www.uzei.org/corpusajardunaldia/03_murkia.pdf; 06-02-10ean irakurria]

Text Encoding Initiative. *The XML version of the TEI Guidelines*. [http://www.tei-c.org/P4X/; 06-02-10ean irakurria]