Zientzia eta Teknologiaren corpusa

Elhuyar Fundazioa
N. Areta, A. Gurrutxaga, I. Leturia, R. Saiz

Ixa taldea–Euskal Herriko Unibertsitatea
I. Alegria, X. Artola, A. Diaz de Ilarraza, N. Ezeiza, A. Sologaistoa, A. Soroa, A. Valverde

Zientzia eta Teknologiaren corpusa *IXA taldearen eta Elhuyar Fundazioaren lankidetza-proiektua da, eta Hizking21 ikerketa estrategikoko proiektuaren barnean ari da egiten. Artikulu honetan azalduko ditugu proiektu honen helburu eta ezaugarri nagusiak, erabili dugun metodologia, garatu ditugun baliabideak eta tresnak, eta aurkezteko moduan dauden lehen emaitzak.*

1. Corpus berezia edo espezializatua

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 hizkuntz erakusgarri 'errealen' multzo 'handi' bat da, irizpide batzuen arabera bildua eta forma-

tu elektronikoan biltegiratua.[1] Askok beste zerbait ere erantsiko liokete horri: corpusak, erabilgarria eta eraginkorra izango bada, informazio linguistikoz hornitua behar du izan.

Corpusa hizkuntza aztertzeko baliabide bat da, hainbat alorretan erabiltzen dena: lexikografian, sintaxian, diskurtsoaren analisian... Horrez gain, hizkuntz teknologiak corpusez eta corpusetan bildutako datuez baliatzen dira, hainbat helburutarako.

Corpus-motak bereizteko lehen ezaugarria direlako 'irizpideak' dira. Irizpideok corpusaren helburuarekin daude zuzen-zuzenean loturik. Esaterako, corpusaren helburua hizkuntzaren erabilera-eremu guztietarako baliagarria edo 'adierazgarria' izatea denean, 'erreferentzia-corpusa' edo 'orotariko corpusa' dela esan ohi da (Sinclair, 10; Leech, 1); aldiz, erabilera-eremu berezi bateko hizkuntza aztertu nahi denean, 'corpus berezia' edo 'espezializatua'.

1.1. Definizioa

Corpus berezia edo espezializatua hizkuntzaren erabilera-eremu espezifiko bateko edo hizkuntz aldaera jakin bateko testuak biltzen dituen corpus-mota da, eremu edo aldaera horretako ezaugarriak aztertzeko asmoz eratua (Sinclair, 10). Corpus berezia erabilera-eremu edo aldaera horren adierazgarria izan daiteke, horretarako berariaz diseinatua eta orekatua baldin bada. Corpus bereziaren adierazgarritasuna eremu edo aldaera horretara dago mugatuta, eta ezin da erabili hizkuntza orokorraren edo arruntaren ezaugarriak aztertzeko eta ondorioztatzeko. Corpus berezi edo espezializatuen adibide batzuk:

---

[1] Honela definitzen du B. Oihartzabalek: "Hizkuntza baten deskribatzeko eta ikertzeko baliatzen den hizkuntza-datu bilduma, baliabide elektronikoak erabiltzen eta eskaintzen dituena."

- *Corpus Científico-Técnico del español* (CCT). <http://www.rae.es/frame.html?URL=/rae/gestores/gespub000008.nsf/(voAnexos)/archC302921639D9C69CC1256ADB00378334/$FILE/CCT.htm&Cabecera= | Departamento de Lingüística Computacional%20|%20 Ingeniería Lingüística>
- IULA-Corpus textual especialitzat plurilingüe. (Dret, Economia, Medi Ambient, Medicina i Informàtica). <http://www.iula.upf.es/ corpus/corpus.htm>
- *Medicor: A corpus of contemporary American medical texts* (1998) Minna Vihla <http://citeseer.nj.nec.com/vihla98medicor.html>
- S*cientific Corpus of Modern French* (La Recherche magazine) (Béatrice Daille and Geoffrey Williams; Université de Nantes). <http://www.elda.fr/fr/cata/text/W0025.html>

1.2. Corpus berezien baliagarritasun eta interesa

Corpus berezien bidez, erabilera-eremu espezifiko baten edo aldaera jakin baten hizkuntz ezaugarriak hobeto aztertzeko aukera dago. Horrekin batera, espezialitate-arloetako hizkuntz erabileraren eta erabilera arrunt edo orokorraren arteko aldeak ere azter daitezke.

Aztergaiak hizkuntzaren aztertze-eremu askotakoak izan daitezke: lexikoa, morfosintaxia, semantika, pragmatika, diskurtsoa, estilistika, testugintza... Hona hemen batzuk:

- Terminologia, lexiko espezializatua: terminologia-azterketak, terminoen erauzketa erdiautomatikoa, termino-aldaeren azterketa eta tratamendua, neologismoen detekzioa
- Fraseologia-unitate espezializatuen azterketa
- Diskurtso espezializatuaren azterketa (hitz-ordena, gramatika-egiturak, joskera, estiloa, testu-egitura...)

- Kontzeptu-mailako informazioaren erauzketa, ontologiak eratzeko teknikak
- Testu-sailkapen automatikoa

Aztertze-eremu horiek hainbat aplikazio-eremutan izan daitezke baliagarri:

- Terminologiaren normalizazioa
- Hiztegigintza espezializatua (hiztegi terminologikoak eta teknikoak)
- Hiztegi orokorretan sartzekoak diren termino espezializatuen hautaketa
- Hitz-adieren desanbiguazioa
- Informazioaren berreskurapen eta erauzketa
- Xede berezietarako hizkuntz irakaskuntza (curriculuma, syllabus-ak)

Aipatutakoak corpus espezializatu elebakarrari dagozkio; corpus espezializatu paraleloen kasuan, hizkuntzen arteko baliokidetza-erlazioen azterketa egin ahal izatea da balio erantsia, eta horren aplikazio-eremu nagusiak: hiztegigintza eta datu-base lexikalak (hiztegi elebidun edo eleaniztunak, terminologia-lanak, HAU-Lak...), itzulpengintza (itzulpen automatikoa, itzulpengintzako laguntza-tresnak, itzulpen-memorien ustiaketa...), ezagutza-base lexiko-semantiko eleaniztunak (WordNet), IE-IR (dokumentu-indexazioa...).

Goian esan dugu hizkuntz erabilera espezializatuaren ezaugarriak 'hobeto' aztertzeko eta ezagutzeko aukera eskaintzen dutela corpus espezializatuek. Zergatik esan dugu hori? Bistan dena, aztertze-helburu jakin batekin diseinatu eta eratu den corpus espezializatuan, helburu horrekiko errelebanteak diren fenomenoei gertuagotik begi-

ratzeko aukera dago, aztertu nahi dugun hizkuntz erabileraren laginetа ebidentzia-dentsitate handiagoa egoteko aukera dagoelako. Corpusa gure interesekoa den alorreko hizkuntz erabileraren adierazgarri izatea da horretarako ezinbesteko baldintza. Baldintza hori betetzen bada, hizkuntz erabilera edo aldaera horretaz eskura ditzakegun datuak doiagoak eta aberatsagoak dira corpus orokor batetik eskura ditzakegunak baino. Horrexetan dago corpus espezializatuen baliagarritasunaren gakoa.

2. Corpusgintza-eredua

Corpusgintzan lau urrats nagusi bereizten ditugu:

- Diseinua
- Corpus gordina eratzea:
  - Inbentarioa
  - Laginketa
  - Testu-bilketa
- Etiketatzea
  - Egitura-etiketatzea
  - Etiketatze linguistikoa
- Corpusak analizatzeko eta ustiatzeko tresnak

Eredu horretako urratsak modu sistematiko eta egituratuan egiteko, corpus-metodologia bat landu behar da, eta, hori inplementatu ahal izateko, corpusgintza-tresna bat. Tresna hori *Corpusgile* da, eta testu-bilketa eta etiketatze-lana dira kudeatu behar dituen prozesu giltzarriak. Bestetik, garatzen ari garen analisi- eta ustiatze-tresnak honako hauek dira:

- Corpusak analizatzeko tresna generikoak
- *Erauzterm* (termino-erauzketa)
- Corpus paraleloak ustiatzeko tresnak

Hurrengo lerroetan, lehen urratsari buruzko ohar orokor batzuk egingo ditugu.

3. Diseinua

Corpus-diseinuan, corpusa osatuko duten testuak biltzeko erabiliko diren irizpideak zehaztu behar dira. Corpus-proiektu batean, diseinu-fasea urrats giltzarria da, arretaz eta xehe landu beharrekoa, corpusa baliabide eraginkorra izango bada. Corpusa, hemen ulertzen dugun moduan bederen, ez da nolanahi bildutako testu-multzo handia (Bach et al., 4). Corpusak hizkuntz datu adierazgarrietan, 'ebidentzietan', oinarritutako deskribapenak egiteko, ondorioak ateratzeko edo hizkuntz tresnak eratzeko balio duen baliabide eraginkorra eta ahaltsua behar du izan.

Horretarako, honakoak zehaztu beharra dago:

- Helburua: corpusa zeren adierazgarri izatea nahi den, zertarako nahi den, zein eduki-mota (testu-mota) izango dituen
- Adierazgarritasuna: corpusean biltzen diren testuen ezaugarriak eta datuak aztertuz helburuak ezarritako aztergairako ondorio fidagarriak atera ahal izatea

Corpusaren adierazgarritasunaren auzia oso gai eztabaidatua da corpus-diseinuan. Garrantzi handikoa da, eta horren baitan dago corpusaren analisiaren bidez hizkuntz erabilerari buruz aterako diren ondorioak fidagarriak eta baliagarriak izango diren.

Corpusa populazio oso baten lagin bat da (edo lagin-multzo bat, zehatzago esanda), eta, ikuspegi estatistikotik, lagin hori adierazgarria

izango da baldin eta hura aztertuz ateratzen ditugun ondorioek populazio osorako balio badute. Populazio hori helburuak determinatzen du, hau da, helburuak corpusa 'zeren' adierazgarri izatea nahi dugun mugatzen du. Diseinuaren xedea da helburu horretarako adierazgarria izango den lagin-bilketa bat egitea lortzea.

Esan gabe doa, ez da erraza lagina adierazgarria den ziurtasun osoz jakitea, populazio osoa aztertu behar genukeelako horretarako, eta laginaren funtzioa, hain zuzen ere, hori ez egitea delako. Beraz, adierazgarritasunaren arazoa konplexua da, ez da ebazten erraza, eta, oraingoz behinik behin, ez dago adierazgarritasuna ebaluatzeko prozedura objektibo bideragarririk. Nolanahi ere, gaur egun hizkuntzaz dugun jakintzak corpusaren adierazgarritasuna hein bateraino behintzat bermatzeko aukera ematen duten gutxieneko irizpide batzuk ematen dizkigu. Irizpide horiek nolabaiteko gida-lerroak dira.

Horretarako aztertu beharreko gaiak:

- Kantitatea (corpusaren tamaina): hizkuntz erakusgarriak biltzen dituzten baliabideen adierazgarritasunean eragin handienekoa duen faktorea da tamaina. Gehienetan, tamaina hitz-kopuruarekin lotzen da
- Oreka: corpus-diseinuan oreka lortzeak esan nahi du corpusean biltzen diren laginen tipologia, banaketa eta maiztasuna bat datozela errepresentatu nahi den populazioaren era bereko ezaugarriekin (edo horietatik gertu daudela), edo doituta dagoela corpusaren helbururako beharrezkoa den hizkuntz fenomenoen jasotze-proportzioarekin. Corpusa orekatua ez badago, adierazgarria ez izateko arriskua handia da. Landu beharrekoak:
  - Laginketa-teknika (proportzionala, geruzatua...)

- Lagin-tipologia: corpusean errepresentatu nahi den unibertsoaren edo populazioan ezaugarri jakin batzuen arabera bereiz daitezkeen motak
- Laginaren kalitatea: erreferentzia-testu eredugarriak vs lagin errealak (ausazkoak)
- Laginaren tamaina; testu osoak vs testu-zatiak
- Jatorrizko testuak/itzulpenak

• Corpus irekia (*monitor corpus*) vs corpus itxia; denbora-bitartea

Adierazgarritasunean eragina duten alderdi batzuk:

• Aztergaitzat hartzen den alor espezifikoaren (populazioaren) hedadura eta dibertsitatea, bai jakintza-alorraren barneko adarkopuruari, bai testu-tipologiaren dibertsitateari dagokienean

• Aztergaitzat hartzen den alor espezifikoaren 'dinamikotasuna' edo aldakortasuna; adibidez, teknologia berrien alorreko corpusa eratu nahi bada, corpusa etengabe elikatu behar da testu berriz, adierazgarria izaten segituko badu

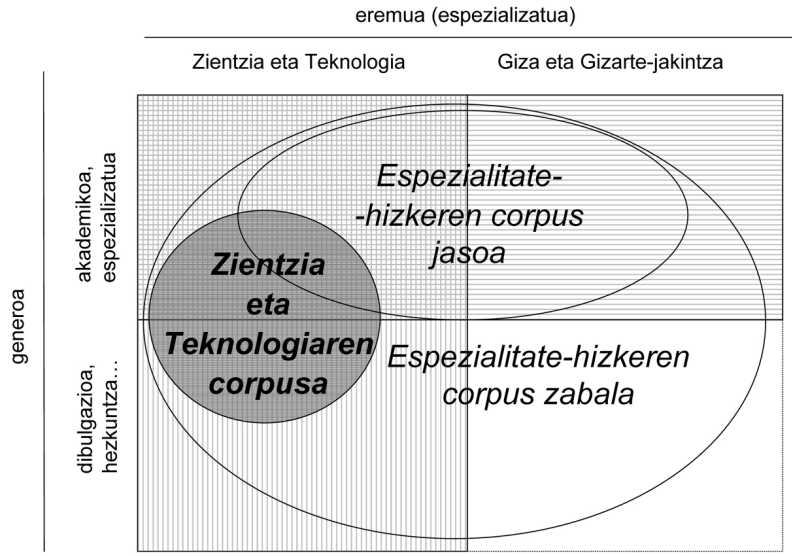
• Corpusean sartuko diren laginen hizkuntz kalitatea: corpusean hizkuntz fenomeno errealak osorik errepresentatu nahi diren ala hizkuntz kalitateko irizpide batzuen araberako 'hautaketa' egingo den; lehenean, hizkuntz produkzioaren populazio osoa da lagingai; bigarrenean, eredugarritzat hartzen diren testuak edo kalitate-baldintzak betetzen dituztenak soilik

Hurrengo ataletan, Zientzia eta Teknologiaren corpus bereziaren diseinu-ezaugarriak azalduko ditugu.

3.1. Helburua

Esan bezala, corpusa zertarako eratu nahi den zehaztu beharra dago lehenik eta behin. Helburuak diseinuan bertan du eragina

(Biber, 1993, 246; Kennedy, 70). Corpus berezia izaki, aurrena 'berezitasun' horren zerizana zehaztu beharra dago. Hurrengo irudian, bi parametroren arabera antolatu dira aztertu ditugun corpus-aukerak: a) eremua eta b) generoa.



1. irudia. ZT corpusaren esparrua: eremua eta generoa

- Eremuari dagokionez, gure helburua murritza da: Zientzia eta Teknologiaren alorreko testuak corpuseratzea
- Generoari dagokionez, zabal jokatu dugu. Zenbait testu-mota eta komunikazio-erregistro hartu ditugu kontuan; horiek guztiak bi mota nagusitara bil daitezke:
  - Adituen arteko komunikazioa (artikulu teknikoak, tesiak…)
  - Adituen eta ez-adituen arteko komunikazioa (testu-liburuak, eskuliburuak, dibulgazioa, erreportajeak…)

Bestetik:

- Gure asmoa erabilera-esparru zabala izan dezakeen corpus espezializatua egitea da; hau da, gure asmoa ez da hizkera espezializatuaren alderdi jakin batean zentratzea, ahalik eta erabilera gehien izan ditzakeen baliabidea lortzea baizik. Hori kontuan hartu beharko dugu diseinuan zehaztu behar diren ezaugarriak erabakitzean
- Baliabide linguistikoa eratu nahi dugu, ez dokumentala. Horren ondorioz, ezinbestekoa da corpusa linguistikoki prozesatzea eta etiketatzea (ikusiko dugu zein izango den gure konpromisoa); bestetik, ez da beharrezkoa jatorrizko dokumentuaren formatu- eta maketa-ezaugarri guztiak etiketatzea, ezta elementu grafikoak ere, baina bai oinarrizko testu-egitura eta linguistikoki interesgarriak izan daitezkeen formatu-ezaugarri batzuk (letra-tipoaren aldaketa, letra-estiloa, komatxoak…) eta horien balio desanbiguatuak (aipuak, nabarmentze hutsak, atzerri-hitzak, metahizkuntza, terminoak…). Bestetik, baliabide berrerabilgarria eta eramangarria izan behar luke, eta, horretarako, testuak kodetzeko eredu estandarren araberakoa
- Gure ikuspegia ez da eredu-emaile izatea. Proiektu honen helburua ez da zientzia eta teknologiaren alorreko 'ereduzko corpusa' eratzea. Aitzitik, inoiz 'eredutzat' har litekeen ikuspegi edo baliabide bat moldatu ahal izateko lehen urrastzat jotzen dugu gure proiektua.

  Hizkuntza aztertzeko corpus-ikuspegiaren alderdi enpirista hutsaren arabera, corpusean hizkuntzaren errealitatea islatzea da helburua. Horren arabera, laginketa diseinatzerakoan ez lirateke kontuan hartu behar testuen kalitatean oinarritutako irizpide edo iragazkiak. Hein handian, irizpide horiek aski aprioristikoak izaten dira, hizkuntzaren ikuspegi jakin batzuen arabera mol-

datuak. Horiek ezartzera, hau da, irizpide horien araberako baldintzak betetzen dituzten testuak soilik bilduz gero, esku hartzen ari gara, nolabait, aztertu nahi ditugun hizkuntz fenomenoen datuetan, hau da, aurkitu nahi duguna jartzen ari gara aztergaian, edo, bestela esanda, aztertzeko objektua aldez aurretik distortsionatzen ari gara, aurreiritziak edo hipotesiak egiazta daitezen (*petitio principii*). Espezialitate-alorreko hizkuntza kalitatearen ebazpena nekez egin liteke datuak begien bistatik ezkutatuz.

Corpuseko datuak aztertuz, hizkuntzaren aztertzaileek (hizkuntzalariek, euskara-teknikariek, irakasleek…) ondorioak atera ditzakete eta proposamenak egin ere bai, dagokion alorreko adituek hizkuntza-ereduari buruzko argibideak edo 'gidalerroak' izan ditzaten, eta erakunde arau-emaileek ere espezialitate-alorreko ebazpenak eman ahal izan ditzaten.

- Euskarria (*medium*): testu idatziak eta argitaratuak soilik corpuseratuko dira, corpusaren aldi honetan behinik behin; jakitun gara horrek kanpoan uzten duela ahozko komunikazioaren alderdia. Hala ere, gaur egun proiektu honen ahalmenetik at daude zientzia eta teknologiaren alorreko ahozko komunikazioaren laginak corpuseratu ahal izateko beharrezkoak diren tresnak eta baliabideak
- Itzulpenak: jatorriz euskarazkoak diren obrez gain, euskarara itzulitakoak ere cospuseratzea erabaki dugu

Horrenbestez, *Zientzia eta Teknologiaren corpusa* honelakoa izatea nahi dugu:

Zientzia eta Teknologiaren alorreko testu-bilduma egituratua, alorreko testu-produkzioaren eta -izaeraren adierazgarri izateko asmoz eratua, eta egitura aldetik eta linguistikoki etiketatua, gaur egungo estandarren arabera.

3.2. Adierazgarritasuna

3.2.1. Adierazgarritasuna eta corpusaren egitura

ZT corpusak gutxieneko adierazgarritasun-baldintza batzuk betetzea erabaki dugu. Horretarako, 'oreka' da kontzeptu giltzarria. Oreka bermatu ahal izateko:

- Populazioaren inbentarioa egin behar da
- Inbentario horretatik abiaturik, populazioaren adierazgarri den corpusa eratu behar da; horretarako, laginketa-eredu jakin baten bidez, testu-lagin multzo bat hautatu behar da

Dena den, alderdi hauek hartu ditugu kontuan adierazgarritasunak corpusean izan behar lukeen dimentsioa finkatzerakoan:

- Gaur egun testu-kantitate handia aski erraz eskura daiteke; aukera horretaz baliatzea komeni da, baina corpus orekatu eta adierazgarria ezin da horrela diseinatu (edo ezin da aurrez bermatu hori lortuko denik)
- Corpus-gune orekatu bat izanez gero, modu 'oportunistan'[2] bildutako testu-multzo handia baliabide osagarria izan daiteke; esaterako, corpusak konparatzeko eta gunearen diseinuan hartutako neurrien eraginkortasuna ebaluatzeko; edo LNPan behar diren datu-kopuru handiak (agerraldiak) lortu ahal izateko

[2] Laginketa testuen eskuragarritasun hutsaren arabera egiten denean, 'oportunista' dela esan ohi da. Beñat Oihartzabalek *bil-ahala-bil* jokabidea aipatzen du. Corpus oportunistak doan edo kostu txikian bildutako testu elektronikoez osatuak dira. Testuak bildu ahala, corpusaren egileak hutsuneak edo desorekak nabaritzen ditu, eta horiek konpontzen saiatu behar du. Laginketa oportunistak tamaina handiko corpusak, gehienetan corpus orekatu gabeak, kostu txikian eratzeko aukera ematen du. Gure ustez, ez da baztertu beharreko bidea, baina corpusaren osagai izango diren testuak ezin dira eskuragarritasun hutsaren arabera aukeratu. Horretan soilik oinarrituta, oso zaila da corpusa errepresentatu nahi den hizkuntz unibertsoaren adierazgarria izango dela bermatzea.

- Corpus-gune orekatua ongi desanbiguatua edukitzea komeni da, baina, beharbada, corpus-atal oportunistan ez litzateke ezinbestekoa izango, aski izan liteke prozedura automatikoz prozesatzea, eskulanik egin gabe
- Aurrerantzean ere, corpus-atal oportunistako testuak corpus-gunera pasatzeko prozedurak azter litezke

Ildo horretatik aurrera egiteko, corpusak bi atal izango lituzke:

- *A atala*: orekatua, inbentario eta laginketaren bitartez diseinatua; automatikoki eta eskuz etiketatua eta desanbiguatua
- *B atala*: masa handia, modu digitalean lortua, ahal den handiena (bilketa 'oportunista'); automatikoki etiketatua eta desanbiguatua

Corpusa analizatzean eta ustiatzean, A, B zein A+B corpusak erabil litezke, beharren eta interesen arabera.[3] Horiengatik guztiengatik, *Corpus-gune orekatua + bilketa-lan oportunista* izenda genezakeen corpus-estrategia onartu da.

3.2.1.1. Corpus-gune orekatua

3.2.1.1.1. Tamaina (n)

Hizkuntz erakusgarriak biltzen dituzten baliabideen adierazgarritasunean eragin handienekoa duen faktorea da tamaina. Gehienetan, tamaina hitz-kopuruarekin lotzen da, besterik gabe, baina honakoak ere kontuan hartzekoak dira: testu-tipologiaren araberako mota (kategoria) bakoitzeko testu-kopurua, testu bakoitzetik hartzen den lagin-kopurua eta lagin bakoitzaren hitz-kopurua (Kennedy, 66; Grönqvist et al.).

---

[3] Fisikoki, corpus bakarra da; gune orekatuko testu-zatiak etiketaturik daude (Ik. 5.1.2.1 atala).

Corpus-diseinuan hutsik egin gabe aipatzen da corpusaren tamainaren 'default value' delakoa 'handia' dela (Sinclair, 6). Arrazoia da lagin 'asko' jaso behar direla hizkuntzaren dibertsitatea eta aldakortasuna jasotzeko eta emaitza estatistiko esanguratsuak lortu ahal izateko. 'Handi' esateak ez du gauza handirik adierazten ordea. Gainera, corpusen tamainan erabakigarri izan diren faktoreak, adierazgarritasuna lortzeko irizpide objektiboak baino gehiago, corpusak eratzeko unean-unean izan diren baliabide edota teknologien menpekoak izan dira gehienetan. Egia da ez dela lan erraza irizpideok objektiboki zehaztea. Gaur egun, teknologiak testuak bildu eta corpusak eratzeko eskaintzen dituen baliabideak izugarri garatu dira, eta tamaina gero eta arazo txikiagoa da. Nabarmendu egin da, horren ondorioz, tamaina jakin batetik aurrera corpusa handitzearen errentagarritasuna gutxituz doala (gero eta hizkuntz fenomeno berri gutxiago sartzen dira corpusean). Konpromiso bat bilatu behar da, beraz, tamaina handitzeko egiten den ahaleginaren eta lortzen den etekinaren artean. Horrez gain, aipatu beharrekoa da tamaina handiagoa izateak, besterik gabe, ez duela esan nahi corpusa adierazgarriagoa denik (Bowker et al., 45). Aditu batzuek diotenez, corpusak behar bezain handia eta ahal bezain txikia izan behar luke (McMullen, 14) (horretarako gakoa 'oreka' litzateke, inondik ere).

Zenbat hitzeko corpusa behar da alor jakin bateko hizkuntz erabileraren testu-lagin adierazgarria eta ondorio fidagarriak ateratzeko modukoa lortzeko? Lehen begiratu batean, zentzuzkoa da pentsatzea erreferentzia-corpus orokorretan baino hitz-kopuru txikiagoa behar dela gutxieneko adierazgarritasuna bermatzeko. Dimentsio apalagoa behar dela, alegia.

Gaur egun helburu orokorretarako diseinatzen diren corpusen tamaina estandar moduko bat 100 milioi hitzekoa dela onar liteke. Euskarazko corpusetan, askoz ere tamaina txikiagoetara baino ez gara heldu (*OEH-ko testu-corpusa* ≅ 5,5 milioi testu-hitz; *XX. men-*

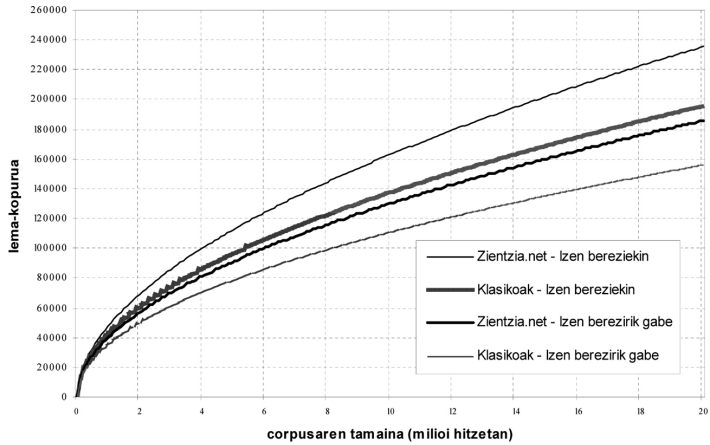
deko Euskararen Corpus Estatistikoa ≅ 4,5; *Ereduzko prosaren corpusa* ≅ 7).

ZT corpusaren tamaina erabakitzeko datu nolabait 'objektiboagoetan' oinarritu ahal izateko, corpusaren tamainaren eta lema-kopuruaren arteko erlazioa landu dugu. Horretarako, Yang et al.en lanean oinarritu gara. Horien arabera, lema-kopurua/corpus tamaina erlazioari gehien hurbiltzen zaion funtzioa $y = f(x) = \alpha x^{\beta}$ da (x = corpusaren tamaina, y = lema-kopurua), eta errore karratu minimoen metodoa erabiltzen dute datu errealei gehien hurbiltzen zaizkien α eta β parametroak kalkulatzeko. Metodo hori euskarazko corpusetara aplikatzeko, azterketa bat egin dugu, bi corpus hauek hartuta: *Pentsamendu Unibertsalaren Klasikoak* bilduma (6.500.000 hitz) eta Zientzia.net-eko *Elhuyar Zientzia eta Teknika* aldizkaria (2.400.000 hitz). Lehenengoa osatzeko testuak Internetetik lortu ditugu (www.klasikoak.com). Uste dugu eratzen ari garen ZT corpusaren ezaugarriak eta bi corpus horien ezaugarriak ez direla oso desberdinak izango *tamaina/lema-kopurua* alderdiari dagokionez: Zientzia.net.eko testuak espezializatuak dira eremua aldetik, nahiz eta erregistroa dibulgazio-mailakoa den; *Pentsamendu Unibertsalaren Klasikoak* bilduma, berriz, erregistro jasokoa da, eremu aldetik zientzia eta teknologiakoa ez den arren (liburu gutxi dira alor horretakoak, zoritxarrez)

Horietako bakoitzean azpicorpusak osatu ditugu, bakoitza aurrekoa baino 100.000 hitz gehiagokoa, eta hainbat ezaugarriren bilakaera aztertu dugu: lema-kopurua, testu-hitz kopurua, izen-kopurua, adjektibo-kopurua, aditz-kopurua... Testua lematizatzeko, IXA taldearen EUSLEM etiketatzailea erabili dugu. Lema-kopuruaren kalkulua egiteko, EUSLEM lexikorik gabeko lematizazioan lortzen duen etekinaren estimazioa jakin behar da; lagin bat azterturik, asmatzetasaren balioa kalkulatu da, eta hori erabili da lema-kopuruak ateratzeko.

Corpusa	**Analisi-mota**	α	β
Pentsamendu Unibertsalaren Klasikoak	Izen bereziak kontuan harturik	56,81	0,4472
	Izen bereziak kontuan hartu gabe	55,25	0,4487
Zientzia.net	Izen bereziak kontuan harturik	44,17	0,4741
	Izen bereziak kontuan hartu gabe	42,31	0,4763

Horien irudikapena:



2. irudia. Corpusaren tamainaren eta lema-kopuruaren arteko erlazioa

Datu horiek ikusita, eta proiektuari eslei lekizkiokeen baliabideak kontuan harturik, *corpus-gune orekatua 5 milioi hitzekoa* izatea erabaki dugu. Izen bereziak aparte utzita, 80.000-90.000 bitarteko lema-kopurua lortzea espero daiteke.

3.2.1.1.2. Laginketa-eredua

Laginketa-sistema geruzatua erabiltzea erabaki da. Laginketa geruzatuan, populazioa zenbait multzo edo 'geruzatan' banatuta dago.

Corpusean sartuko diren testu-laginak ausaz hautatzen dira, geruza bakoitzaren barnean betiere. Horretara, geruza bakoitzak corpusean halako proportzioa izango duela bermatzen da. Proportzio horiek geruzek populazioan duten proportzio berberak izan daitezke, edo horiek ez bezalakoak. Izan ere, populazioaren izaeraren arabera, gerta daiteke geruza batzuetako testu-produkzioa kuantitatiboki txikia izatea, baina linguistikoki interesgarria. Laginketa proportzional batean, horrelakoak oso ezkutuan gera daitezke. Geruzen arteko 'oreka' handiagoa nahi bada, geruza jakin bakoitzak corpusean izango duen proportzioa aldatzeko aukera dago, beraz.

Horiek horrela, hauek dira gure laginketa-ereduaren ezaugarriak:

- Geruzak: Geruzak edo 'sailak' eratzeko, parametro batzuk erabil daitezke, eta parametro horien balioen konbinazioak dira geruzak. Guk bi parametro erabili ditugu: a) eremua; b) generoa. Hona hemen horien balioak:

  – Eremua

    - Zientzia zehatzak (Matematika eta Logika)
    - Materiaren eta energiaren zientziak (Fisika eta Kimika)
    - Lurraren zientziak (Geologia, Ozeanografia, Geografia...)
    - Biziaren zientziak (Biologia, Medikuntza, Ingurumena...)
    - Teknologia (Teknologia Mekanikoa, Teknologia Elektrikoa/Elektronikoa, Telekomunikazioak, Informatika, Aeronautika...)
    - Bestelakoak (Ekonomia, Arte-teknologiak, Antropologia...)[4]

[4] 'Bestelako gaiak' eremuan, zientzia eta teknologiaren alorrean sartu ohi ez diren baina mugakotzat jo litezkeen zenbait alorretako testuak sartu ditugu. Ez da batere samurra horrelakoetan erabaki argi eta zalantzagabea hartzea, eta irizpideak zehaztea ere zaila da.

- Orokorra

– Generoa

- Oinarrizko hezkuntzako materiala
- Goi-mailako liburua (espezialistentzako liburua + goi-mailako hezkuntzako liburua)
- Artikulu espezializatua
- Dibulgazio-artikulua
- Dibulgazio-liburua
- Administrazio publikoko dokumentua

• Geruza bakoitzaren tamaina, hasiera batean behintzat, geruzak populazioan duen proportzioaren araberakoa izatea:

$$n_i = N_i \frac{n}{N}$$

Horretarako, N eta N_i balioak jakin beharra dago. Horien kalkuluak gutxi gorabeherakoa behar du izan, halabeharrez. Inbentarioak obra bakoitzaren tamainari (hitz-kopuruari) buruz ematen dizkigun datuak hauek dira: a) orrialde-kopurua; b) orrialdearen neurriak. Azterketa bat egin dugu bi parametro horien eta hitz-kopuruaren arteko erlazioa zehaztu nahian.[5] Hala ere, azterlan horren emaitzetatik ez da ondorio argirik atera, eta batazbestekoarekiko desbideratzea handia da. Nolanahi ere den, orrialdeko batez besteko hitz-kopurua (175) aski gertu dago UZEIk kalkulatu zuenetik (180), eta hori da erabili dugun

[5] Elhuyar liburutegitik 30 liburu aukeratu ditugu, zientzia eta teknikari buruzkoak eta tamaina eta genero askotakoak.

balioa inbentarioan obra baten hitz-kopuruaren estimazioa egiteko.

Lehen esan dugunez, inbentarioa egindakoan n_i/N_i balioa aldatzea komenigarria den azter daiteke; hurrengo atalean azalduko ditugu horrekikoak.

- Geruza bakoitzetik hautatu beharreko obra-kopurua. Geruzako obra-kopurua bada, eta geruzatik hartzen den kopurua bada, honakoa bete behar da:

$$\frac{m_i}{o_i} \cdot \frac{n_{i(j)}}{N_{i(j)}} = \frac{n_i}{N_i}$$

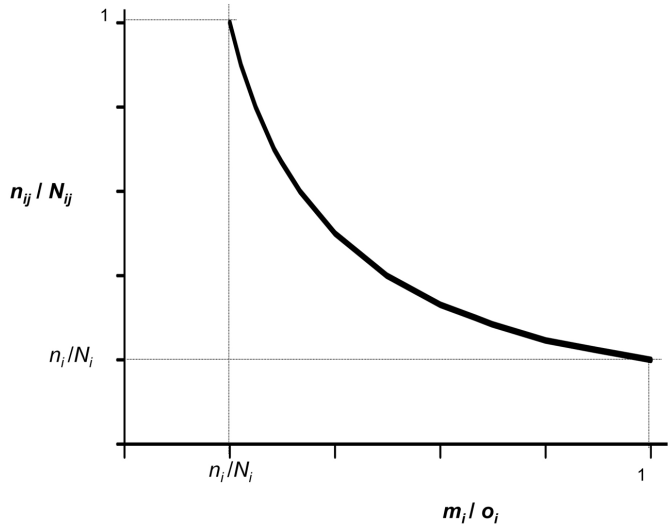
$\frac{n_{i(j)}}{N_{i(j)}}$: geruzako obra batetik hartuko den proportzioa

$N_{i(j)}$ aukeratutako obra bakoitzaren tamaina da, eta aukeratutako obra bakoitzetik hartzen den testu-kantitatea. Mutturreko aukerak hauek dira:

– Geruza bakoitzari dagokion hitz-kopurua obra osoak corpuseratuz betetzea:

$$\frac{m_i}{o_i} = \frac{n_i}{N_i}$$

– Geruza bakoitzari dagokion hitz-kopurua obra guztietatik lagin bana hartuz betetzea: $\frac{m_i}{o_i} = 1$; obra bakoitzetik $\frac{n_i}{N_i}$ proportzioa corpuseratuko litzateke



3. irudia. Lagindutako obra-ehunekoaren (m_i/o_i) eta obratik lagindutako proportzioaren (n_{ij}/N_{ij}) arteko erlazioa

Bi mutur horien arteko puntu bat interesatzen zaigu, hau da, obretako batzuk hartu eta horietako bakoitzetik zati bat. Bi aldagaiak berdintzen diren puntua hobetsi dugu:

$$\frac{m_i}{o_i} = \frac{n_{i(j)}}{N_{i(j)}}$$

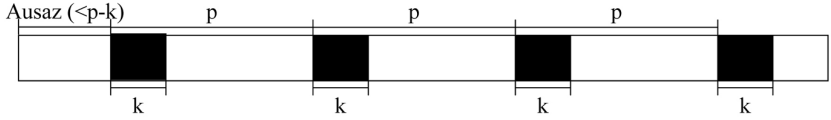
Beraz:

$$m_i = o_i \sqrt{\frac{n_i}{N_i}}$$

- Geruza bakoitzeko obren zozketa: denek probabilitate bera izatea
- Obra bakoitzetik hartuko den testu-masa: obraren tamainaren araberakoa izatea

$$n_{i(j)} = N_{i(j)} \frac{n}{N} \bullet \frac{N_i}{N_{i(t)}}$$

- Obra bakoitzetik hatu beharreko testu-masa jarraitua ez izatea hobetsi dugu (Badia et al., 2), halako karaktere-kopuruko tartean behin (*p*) hartutako *k* karaktereko laginak izatea baizik. *k*-ren balioa 300 hitz izatea erabaki dugu (orrialde estandarraren inguruko balioa)



4. irudia. Obra baten lagin etenak antolatzeko prozedura

- Lagin-tamaina minimoa: obra baten tamaina lagin jarraituaren tamaina baino txikiagoa denean, laginketan ez sartzea erabaki da

3.2.2. Denbora-bitartea

Lehenik eta behin, corpusa itxia ala irekia (*monitor corpus*) den erabaki behar da, baita testuen sortze- edo argitaratze-daten balio-tartea zehaztu ere (corpus irekia bada, hasiera-data soilik).

- *Corpus itxia/irekia*. Corpusgintza irekia bultzatu behar litzatekeela uste dugu. Hala ere, HIZKING21en barnean egitekoak diren corpusak, une honetan bederen, corpus itxiak izango direla pentsatzera behartuta gaude
- Denbora-bitartea atalkatzeko modua dela eta, *XX. mendeko euskararen corpus estatistikoa*-n erabiltzen diren epeekin bat etortzea komeni da (Urkia, 6). Corpus horretan, 1969-1990 da bitarteetako bat (euskara batuaren sorrera eta araugintza berriaren hasiera, hurrenez hurren). Hala ere, gure kasuan bederen, interesga-

rria da dokumentuak formatu elektronikoan jasotzeko aukera izatea, eta 1984 jo da edizio elektronikoaren hasieratzat. Baina urte horrek testu elektronikoen eskuragarritasunaren muga markatzen du, eta ez, beharbada, hizkuntzaren bilakaeran esanguratsua izan daitekeen muga bat

Horiek horrela:

- *Hasiera-data*. Inbentarioa 1990etik aurrera landu da, eta aurreko produkzioa geroko urrats batean landuko da
- *Amaiera-data*. Corpusgintza-aldi honen amaiera-data 2002ren bukaera da

Beraz, ZT corpusaren lehen bertsio honetan, 1990-2002 bitartean argitaratutako testuak dira corpusean sartzeko hautagai.

4. Corpus gordina eratzea

4.1. Inbentarioa

Diseinu-ezaugarriak finkatuta, gune orekatua eratzeko lehen eginkizuna inbentarioa egitea da. Horretarako, Euskal Herrian egin diren lanak eta ISBN datu-basearen CD-ROMa hartu dira abiapuntutzat:

- Joan Mari Torrealdairen *XX. mendeko euskal liburuen katalogoa* (http://www.jakingunea.com/grafikoak/katalogoa.htm)
- *Inguma - Euskal Komunitate Zientifikoaren Datu-Basea* (http://www.inguma.org/berria/index.cfm)
- *Agencia Española del ISBN* (http://www.mcu.es/bases/spa/isbn/ISBN.html#A01)

Inguma-ren jabe den UEUrekin hitzarmena sinatu da, baliabide hori HIZKING21eko corpusgintzan erabili ahal izateko.

1990-2002 bitarteko Zientzia eta Teknologia alorreko obrak sartu dira inbentarioaren datu-basean. Horretarako, aipatu iturrietako datuen

bidezko hautatze-prozedura automatikoak erabili dira lehen bilketa-lana egiteko. Jatorrizko datuak iragazteko, SHU sailkapeneko balioak erabili dira.[6] Batzuetan, iturburuan ez dago horri buruzko informaziorik. Adibidez, Jakinen bibliografian 'Hezkuntza' atal generikoan bildu dira zientzia eta teknologiaren alorreko obrak. Beraz, atal hori osorik landu behar izan dugu, ZT alorrekoak ez direnak baztertzeko.

Bilketa hori datu-base bakarrera ekarri da, eta, ondoren, eskuz berrikusi dira emaitzak, bat ez datozen datuak bateratzeko, osatzeko, eta estaldura eta doitasuna hoberentzeko.

Prozesu hori antolatzeko irizpide batzuk:

- ISBN zenbaki bereko argitalpenak: azken edizioa hartu da
- Liburu edo aldizkari bakarrean argitaratutako artikuluak: bakoitza obratzat hartu da

Inbentarioa egin ondoren, obrak 'Laginketa-eredua' atalean zehaztu ditugun eremu- eta genero-balioen arabera sailkatu dira.

Emaitzak:

- Obra-kopurua: 9.481
- Hitz-kopuruaren aurreikuspena: 86.360.275
- Eremuaren araberako banaketa

---

[6] Zientzia eta Teknologiaren alorrak hartzen dituzten 5. eta 6. sailez gain, beste zenbait sail ere hartu ditugu kontuan: 004 (Computer science and technology. Computing); 007 [Activity and organizing. Information. Communication and control theory generally (cybernetics)]; 311 (Statistics as a science. Statistical theory); 33 (Economics. Economic science); 77 (Photography and similar processes); 911 (General geography. Systematic geography. Physical, human, theoretical geography etc.). Azkenik, beste zenbait sailetako kodeak dituzten obrak banan-banan aztertu dira: 159.9 (Psychology); 72 (Architecture); 76 (Graphic arts. Graphics); 66 (Applied graphic techniques. Commercial graphics); 7 (Photography and similar processes); 78 (Music); 91 (Cinema. Films)

Eremua	Obrak	%	Eremua	Hitzak	%
Biziaren zientziak	4.010	42,30	Biziaren zientziak	28.478.966	32,98
Teknologia	2.210	23,31	Zientzia zehatzak	17.191.685	19,91
Materiaren/energiaren zientziak	876	9,24	Teknologia	15.890.121	18,40
Zientzia zehatzak	864	9,11	Bestelako gaiak	10.570.116	12,24
Bestelako gaiak	641	6,76	Materiaren/energiaren zientziak	6.047.952	7,00
Orokorra	454	4,79	Lurraren zientziak	4.773.169	5,53
Lurraren zientziak	426	4,49	Orokorra	3.408.266	3,95

• Generoaren araberako banaketa

Generoa	Obrak	%	Generoa	Hitzak	%
Dibulgazio-artikulua	5.136	54,17	Oinarrizko hezkuntza	39.090.780	45,26
Oinarrizko hezkuntza	1.751	18,47	Goi-mailako liburua	22.226.400	25,74
Artikulu espezializatua	1.051	11,09	Dibulgazio-liburua	12.151.080	14,07
Dibulgazio-liburua	524	5,53	Administrazio publikoa	7.967.160	9,23
Administrazio publikoa	473	4,99	Dibulgazio-artikulua	2.876.140	3,33
Goi-mailako liburua	546	5,76	Artikulu espezializatua	2.048.715	2,37

Nabarmena da 'Dibulgazio-artikulua' generoko obra-kopurua handia. Kontuan hartu behar da, dena den, artikulu horietako asko (2.915, obra guztien % 30,75) lagin-tamaina minimoa baino laburragoak direla (<300 hitz). Bistan dena, dagokien hitz-kopurua txikia da (382.724, hitz, % 0,44).

4.2. Laginketa

Inbentarioaren datuak ikusita, honako hauek erabili ditugu gogoan laginketarako geruzen proportzioak erabakitzerakoan:

• Geruzen arteko alde handiak nabari dira; eremu eta genero batzuetako produkzioa oso handia da beste batzuen aldean. Zenbaitetan, alde horiek eremu- edo genero-banaketaren zentzuzko

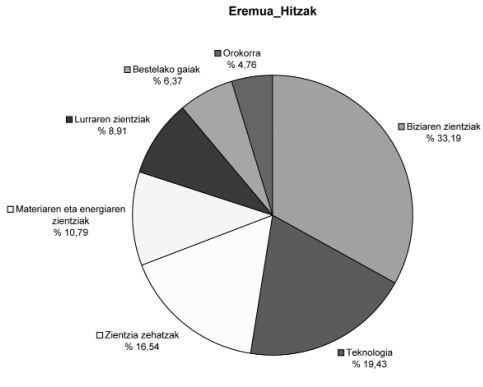
ondorio dira. Esaterako, 'Biziaren zientziak' oso eremu zabala da (Biologia, Zoologia, Botanika, Ekologia, Ingurumena, Medikuntza, Psikiatria…). Nolanahi ere den, proportzio horiek hein batean doitzeak aukera emango luke corpusean biltzen diren edukien arteko 'oreka' handiagoa izateko

- Laginketarako geruzen proportzioek ez lukete gehiegi urrundu behar inbentarioko proportzioetatik; doitze-lanak ez lituzke desitxuratu behar ZT alorreko benetako testu-produkzioaren ezaugarriak

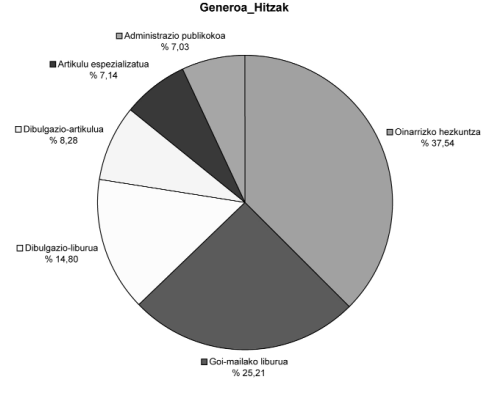
Horiek horrela, proportzioak zertxobait doitzea erabaki da, baina inbentarioak iradokitzen duen irudi orokorra gehiegi aldatu gabe: 'Hitzak' parametroan, 'Zientzia zehatzak', 'Biziaren zientziak' eta 'Bestelako gaiak' eremuen pisu erlatiboa gutxitu egin da; orobat 'Oinarrizko hezkuntza' generoarena.

Bestetik, kontuan hartu behar da 'Dibulgazio-artikulua' generoko obra asko ez direla lagin-tamaina minimora iristen (300 hitz), eta baztertu egin dira laginketatik.

Horren guztiaren emaitza:



5. irudia. Eremuaren araberako lagintze-proportzioak (hitzetan)



6. irudia. Generoaren araberako lagintze-proportzioak (hitzetan)

Bi parametro horiek konbinatuz, 42 geruza edo 'sail' sortzen dira. Lehen azaldu dugun laginketa-eredu estatistikoa aplikatu da. Sail bakoitzean honako prozesua egin da: saileko hitz-kopurua (zenbatetsia) da abiapuntua; hortik, cospuseratzekoa den obra-kopurua kalkulatu, eta obrak ausaz hautatu dira (zotz eginez); hurrena, obra bakoitzetik hartu beharreko hitz-kopurua kalkulatu da; azkenik, obra bakoitzetik hartu behar diren lagin etenak zehaztu dira.

Sail bakoitzeko obren artean zotz egin ondoren, guztira 2.019 obra sartu dira gune orekatuan.

4.3. Testu-bilketa eta formatu-bihurketa

Testuak biltzeko hiru bide aipatu ohi dira: a) testua formatu elektronikoan jasotzea; b) testua eskaneatzea; eta c) testua eskuz idaztea ordenagailuan. Esan gabe doa, a) bidea da erosoena eta fidagarriena, testuak argitaratu bezala jasotzen direlako. Testuak formatu horretan jaso ahal izateko, argitaratzaileengana jo dugu. Horretarako, corpusaren helburua, erabilera eta testuak corpuseratzeko baldintzak zehazten dituen hitzarmena sinatzen da hornitzaileekin. Zenbaitetan ordea, ezin izan da testua formatu elektronikoan eskuratu, eta eskaner bidez digitalizatu behar izan dugu. Azken bidea (testuak idaztea) ez dugu erabili behar izan.

Formatu elektronikoan jasotzen dugunean, jatorrizko dokumentuaren formatu hauek onartu ditugu: .html, .xml, .doc, .rtf, .txt, .pdf, .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, Word-en erabiltzen diren 'atalburua', 'buletdun zerrenda', eta abar).

5. Etiketatzea

Corpusak kodetzeko eta etiketatzeko proposatu diren ereduen eta formatuen artean, TEI eredua eta XML teknologia hautatu ditugu. Horretara, gure etiketatze-eredua koherentea da TEI P4ren orientabideekin, orokorrekin zein hizkuntz corpusetarako emandako orientabide bereziekin (23. atala; http://www.tei-c.org/P4X/CC.html).

5.1. Egitura-etiketatzea

Egitura-etiketatzearen helburua bikoitza da:

- Testu-egituraren ezaugarriak jasotzea; horretara, atalburuak, atalak, azpiatalak, paragrafoak, zerrendak, taulak, oin oharrak, irudi-oinak, etab. etiketa daitezke
- Testuaren ezaugarri tipografiko linguistikoki esanguratsuak jasotzea, baita bereizte horren balioa edo funtzioa ere

Jatorrizko testu-laginetatik testu-egiturari eta testu ezaugarriei buruzko zein informazio etiketatuko den erabaki da. Erabaki hori hartzeko kontuan hartu beharreko alderdiak:

- Etiketatze-lana automatizatzeko dagoen aukera; jokaera malgua aurreikusten dugu, hau da, egitura-ezaugarri batzuk automatikoki etiketatzea osa zaila eta kostu handikoa izan daiteke batzuetan, lana eskuz egin beharra dagoelako; horrelakoetan, azaleko egitura etiketatzea egingo da
- Corpusaren helburua: bideratu nahi den corpusgintzak, hala aitortu da HIZKING21en, baliabide linguistikoak eratzea du helburu; beraz, testuaren jatorrizko itxura berreskuratu ahal izatea ez da helburutzat jo. Horrek berekin dakar linguistikoki esanguratsua ez den egitura- edo tipografia-ezaugarriak ez direla ezinbestean etiketatu behar

5.1.1. Corpusaren oinarrizko egitura

TEI ereduan, corpusak *composite text* ("testu konposatua" edo "testu anizkoitza") erakotzat hartzen dira. Elementu nagusia `<teiCorpus.2>` da; corpusa bera deskribatzen duen goiburuaz (`<teiHeader type='corpus'>`) eta corpusa osatzen duten dokumentuez osatua da. Dokumentu horietako bakoitza `<TEI.2>` elementuan biltzen da. Hori, berriz, goiburuaz (`<teiHeader type='text'>`) eta dokumentuaren testua bera biltzen duen `<text>` elementuaz osatua da.

Honelakoa izan liteke corpusaren egitura:

```
<teiCorpus.2>
   <teiHeader type='corpus'>
      <!-- TEI header for corpus-level information -->
   </teiHeader>
   <TEI.2 id='T1'>
      <teiHeader type='text'> <!-- ... --> </teiHeader>
      <text> <!-- ... --> </text>
   </TEI.2>
   <TEI.2 id='T2'>
      <teiHeader type='text'> <!-- ... --> </teiHeader>
      <text> <!-- ... --> </text>
   </TEI.2>
   <!-- ... etc. -->
</teiCorpus.2>
```

Edozein XML dokumenturen egitura deskribatzeko, DTD (Document Type Definition) izeneko dokumentuak erabiltzen dira (Arriola eta al, 6). Horietan, elementuak, elementu bakoitzaren barnean izan daitezkeen elementuak eta atributuak, datu-motak, eta abar. deskribatzen dira formatu jakin batean. DTDen bidez, programa batzuk erabil daitezke gure XML dokumentua zuzen eratuta dagoen automatikoki egiaztatzeko.

Gure corpusaren baliozkotasuna egiaztatzeko ere DTD bat erabiltzen dugu. TEI ereduak badu bere DTDa, baina TEI mota askotako dokumentuak kodetzeko erabiltzen denez, dituen elementuen kopurua eta DTDaren tamaina oso handiak dira. Horregatik, TEI-koek *Pizza Chef* izeneko tresna paratu dute (http://www.tei-c.org/pizza.html), TEItik guk nahi ditugun elementuak hautatuta DTDa sortzen duena. Tresna horren bidez, gure DTDa sortu dugu, TEItik erabili behar ditugun elementuak soilik dituena. Elementu horiek jarraian deskribatuko ditugu.

5.1.1.1. TEI GOIBURUA (<TEIHEADER>)

TEI goiburuan corpus osoaren edo dokumentu baten informazio deskriptiboa edo deklaratiboa adierazten da. Bestela esanda, corpustestuen deskribapen formala egiten duen informazio erreferentziala biltzen du goiburuak.

Corpusaren goiburuan, corpus osoari dagokion informazioa biltzen da: helburuak, diseinu-ezaugarri eta -irizpideak, sailkapenmotak (esaterako, eremuaren eta generoaren sailkapenak), kodetze-irizpideak, eta abar.

Bestetik, corpuseratu den obra bakoitzaren datuak `<teiHeader type='text'>` elementuan biltzen dira (ISBN zenbakia, izenburua, egilea, argitaratze-urtea, argitaletxea, eremua, generoa...). Metadatu horiek inbentarioaren DBtik zuzenean ekartzen dira goiburura.

5.1.2. Etiketatze-eskema

TEIk aukera ugari eskaintzen ditu testuak etiketatzeko. ZT corpusean, testuen egitura eta formatu-ezaugarri zenbait markatzea erabaki dugu. Horiez gain, analisi linguistikoaren emaitzak hobetze aldera, zuzenketak eta aldaera ez-estandarrak etiketatzeko aukera ere baliatzen da.

5.1.2.1. EGITURA-ELEMENTUAK

- `<text>`: obra bat edo obra baten laginak hartzen ditu bere baitan
- `<body>`: obra baten gorputza edo testua bera[7]
- `<div>`: testuaren atal bakoitza hartzen du; `maila` atributuaren bidez, `<div>`-en arteko habiatzea adierazten da[8]
- `<head>`: atalburua
- `<p>`: paragrafoa
- `<table>`, `<row>`, `<cell>`: taula, errenkada, gelaxka
- `<list>`, `<item>`: zerrenda, zerrenda-elementua
- `<note>`: oin-oharra

Bestetik, TEIren DTDari atributu bat erantsi diogu: `orekatua`. Horren bidez, corpus-gune orekatuan sartzen diren laginak markatzen dira. Horretara, obra baten testu osoa corpuseratu denean, gune orekatuko laginak bereiz etiketaturik daude, eta gune orekatuko laginak soilik edukiko lituzkeen azpicorpusa eratzea erraza da, beraz.

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.

---

[7] `<body>` elementuaren aurretik eta ondoren antola daitezkeen `<front>` eta `<back>` elementuak ez ditugu erabili; elementu horietan, azala, aurkibideak, eskaintzak, bibliografia, aurkibide analitikoa, eta abar antolatzen dira. Elementu bereizietan etiketatzeko lana eskuz egin behar izaten da, eta, gainera, batzuek ez dute interes linguistiko berezirik (bibliografiak, adibidez). Horregatik, corpuseratu direnak `<body>` elementuaren barnean antolatu dira.

[8] `<div>` elementua automatikoki etiketa daitekeenean baino ez da gauzatu; jatorrizko hainbat dokumentutan, testuak ez dakar egituratze-informaziorik, eta horrelakoetan ez da `<div>` elementua erabili.

5.1.2.2. NABARMENTZEA ETA AIPUAK

Letra-estiloaz (letra lodia, etzana, azpimarratua...), letra-tipoa aldatuz edo komatxoen bidez nabarmentzen diren zatiak <hi> elementuaren bidez etiketatzen dira testua jatorrizko formatutik TEIra bihurtzen dugunean. Nabarmentze tipografiko mota rend atributuaz markatzen da:

```
McDonnell-en NOTAR sisteman (<hi rend="italic">No Tail
Rotor</hi> edo isats-errotorerik gabea hitzen laburdura
da), bihurdura-momentua...
```

<hi> elementua hitz baten barnean gertatzen denean, <seg> elementuaren bidez markatu da hitz osoa. Hori garrantzitsua da etiketatze linguistikoa egiten denean, hitz osoa token bakartzat prozesatu ahal izateko.

Hurrengo urrats batzean, <hi> elementuak eskuz aztertzen dira, eta honako balio hauetakoren batez ordezkatzen ditugu:

- <foreign>: testuko hizkuntzakoa ez den hitza edo pasartea
- <emph>: enfasi linguistiko edo erretorikoa
- <distinct>: linguistikoki berezia den hitz edo pasartea
- <q>: aipua; elkarrizketak ere elementu honen bidez etiketatzen dira; type atributuaren bidez bereizten dira galderak (type="answer") eta erantzunak (type="answer")
- <soCalled>: idazleak adiera berezia ematen dion (edo eman ohi zaion) hitz edo pasartea[9]

---

[9] TEIn honela definitzen da <soCalled> elementua: "Contains a word or phrase for which the author or narrator indicates a disclaiming of responsibility, for example by the use of scare quotes or italics. Common examples include the 'scare' quotes often found in newspaper headlines and advertising copy, where the effect is to cast doubts on the veracity of an assertion. (...) The same element should be used to mark a variety of special ironic usages."

- <term>: terminoa
- <gloss>: terminoaren azalpena edo definizioa
- <mentioned>: metahizkuntza
- <name>: izen bereziak (atributuak: pertsona, lekua, erakundea, objektua, artelana, produktua...)

Batzuetan, TEIk aurreikusi bezala, <hi> elementua bere horretan utzi da, aurreko funtzioetako bat esleitzerik izan ez dugunean. Elementu horietako batzuetan, lang atributua (hizkuntza) zehaztu da: <q>, <term>, <socalled>, <mentioned>, <name>.[10]

Aurreko adibidea honela agertuko da eskuz desanbiguatu ondoren:

```
McDonnell-en NOTAR sisteman (<term cert="ziurra"
lang="en" rend="italic" resp="hizking21">No Tail
Rotor</term> edo isats-errotorerik gabea hitzen laburdura
da), bihurdura-momentua...
```

5.1.2.3. Zuzenketak, aldaketak

TEI ereduak akats tipografikoak eta testu-hitzen aldaera estandarrak markatzeko aukera ematen du. Horrelakoak etiketatzea interesgarria da etiketatze linguistikoa errazteko eta eraginkorragoa izateko.

Bi eratara marka daitezke: a) testu-hitza aldatu gabe, forma zuzendua edo aldaera estandarra atributuan markatzea; b) testuan forma zuzendua edo estandarra jartzea, eta jatorrizko testu-hitza, atributuan. Bigarren aukera hobetsi dugu, etiketatze linguistikorako erosoagoa delako.

---

[10] Beraz, <foreign> elementua testuko hizkuntzakoa ez den eta beste elementu horietako bat esleitzerik ez dagoen hitz edo pasartea markatzeko soilik erabiltzen dugu.

- Akats tipografikoak: `<corr>` elementua (jatorrizko testu-hitza: `sic` atributuaren balioa)
- Aldaera estandarrak: `<reg>` elementua (jatorrizko testu-hitza: `orig` atributuaren balioa)

Adibidez:

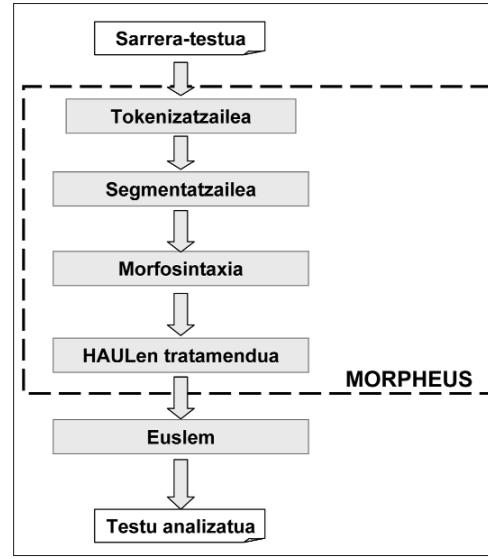
```
<corr cert="ziurra" resp="hizking21" sic="baztuk">bat-
zuk</corr>

<reg cert="ziurra" resp="hizking21" orig="zientzila-
ri">zientzialari</reg>
```

Prozesu hau nola egiten den *Corpusgile*-ren EE moduluaren atalean azalduko dugu.

5.2. Etiketatze linguistikoa

Corpusa baliabide linguistikoa izango bada, ezinbestekoa da linguistikoki prozesatzea eta etiketatzea, alegia, corpuseko hitzak informazio linguistikoaz aberastea. Hitzen informazio linguistikoa lortzeko, IXA taldearen hainbat tresna linguistiko erabili dira.



7. irudia. Etiketatze linguistikoaren oinarrizko eskema

Sarrera-testutik (gure kasuan, egitura-etiketatzearen irteeratik) abiatuz, *tokenizatzaileak* testuan ezagutu dituen tokenen zerrenda ematen du emaitzatzat. *Morpheus*-en baitako *segmentatzaileak* lexiko orokorra sarreratzat hartzen du, eta testuaren segmentazio morfologikoa egiten du, eta token bakoitza osagai dituen morfemetan zatitzen du. Horren ondoren, tratamendu morfosintaktikoa egiten da, hitzmailako interpretazio morfologikoak eratuz, eta, azkenik, HAULak (Hitz Anitzeko Unitate Lexikalak) tratatzen dira. Bukatzeko, EUSLEM lematizatzaileak aurreko emaitzak fintzen ditu, testuinguruaren arabera okerrak diren interpretazioak baztertuz.

Informazio lexikala EDBL datu-base lexikaletik dator. EDBL lexiko-biltegi iraunkorra da, eta aparteko prozesu baten bitartez gobernatzen da. Emaitzen doitasuna handitzeko asmoz, EDBLko lexikoari erabiltzailearen lexiko partikular bat gehitu diogu. Hiztegi horretan, hizkuntza arruntean erabiltzen ez diren termino zientifiko-teknikoak gehitu dira; beraz, termino horiek lematizatzen/etiketatzen direnean sistemak ez du beste aukerarik aztertuko. Hiztegi hori osatzeko, bi irizpide jarraitu dira: Elhuyarren hiztegigintzako datu-basea eta ZT corpusa bera. Lehena EDBLrekin erkatu da, hor ez dauden lemak erabiltzailearen lexikoan sartzeko edo, batzuetan, terminoak orokor samarrak zirenean, EDBL bera aberasteko. Corpusaren erabilerari dagokionean, aurreprozesaketa bat egin da lematizatzeko arazoak ematen dituzten hitzak detektatu eta maiztasunaren arabera sailkatzeko. Zerrendaren buruan geratu diren terminoak eskuz aztertu eta, egokitzat hartu denean, erabiltzailearen lexikoan barneratu dira. Eginkizun hau burutzeko, programa bat garatu da, geroago aztertuko den Corpusgile tresnaren parte dena. Bi lan horiek etiketatze linguistikoaren beraren aurretik egin dira, egitura-etiketatzearekin batera.

Prozesu honen amaieran, corpusean dagoen hitz orok zenbait informazio linguistiko izango du erantsita, hala nola:

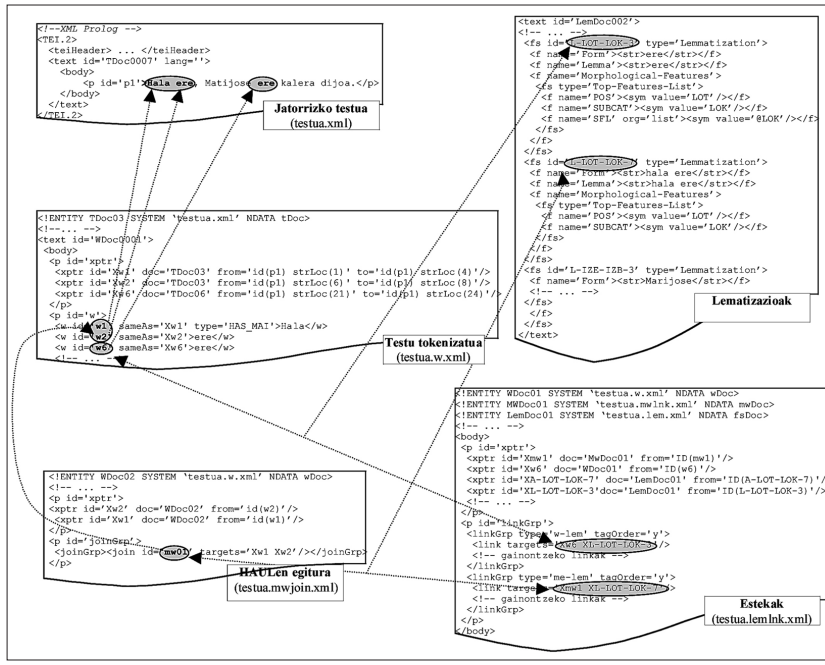
- Hitzaren lema
- Hitzaren kategoria lexikala
- Hitzak duen kasua
- Hitzak betetzen duen funtzio sintaktikoa

Testuak linguistikoki etiketatzeko, bi hurbilpen nagusi jarraitu ohi da, historikoki. Batean, informazio linguistikoa jatorrizko corpusean txertatzen da, hitzekin batera, orain arte ikusi ditugun etiketak bezala (<text>, <body>, <hi> etab.). 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. Informazio linguistikoaren konplexutasuna kontuan harturik, hurbilpen honek abantaila hauek eskaintzen dizkigu, besteak beste:

- Informazio linguistikoa hainbat mailatan edo geruzatan antola daiteke, eta geruza bakoitza independentea izan daiteke besteekiko. Geruza batean aldaketak egin behar badira, aldaketek eragin txikia izango dute gainerako geruzetan
- Informazio teilakatua adierazteko aukera ematen du, eta, ondorioz, analisi linguistiko anbiguoak adierazteko
- Etiketatze-sistemaren hedagarritasuna errazten du, corpusaren gainean informazio linguistiko osagarria txerta baitaiteke, dagoen informazioaren gainean oinarriturik

Azken hurbilpen horri *standoff markup* esaten zaio (anotazio edo etiketatze 'banatua'), eta horixe erabili da corpusa linguistikoki etiketatzeko.

Irudi horretan, corpusaren etiketatze-sistema ageri da (Aldezabal et al., 65). Bost dokumentu daude, eta horien bidez gauzatzen da etiketatze linguistikoa. Goi-ezkerraldean, linguistikoki analizatu nahi



8. irudia. Etiketatze linguistikoa. Anotazio banatuaren 'amarauna'

den dokumentua dago, hots, corpusa edo, gure kasuan, egitura-etiketatzearen emaitza. Horren azpian dagoen dokumentuak (*Testu tokenizatua* izenekoa) tokenizazioaren emaitza adierazten du, hau da, dokumentu honetan zehazten da jatorrizko testuen tokenak zein diren. Horretarako, token bakoitzak jatorrizko dokumentuan duen posizioa adierazten da, estekak erabiliz. Adibidez, dokumentu horretan honako lerro hau dago:

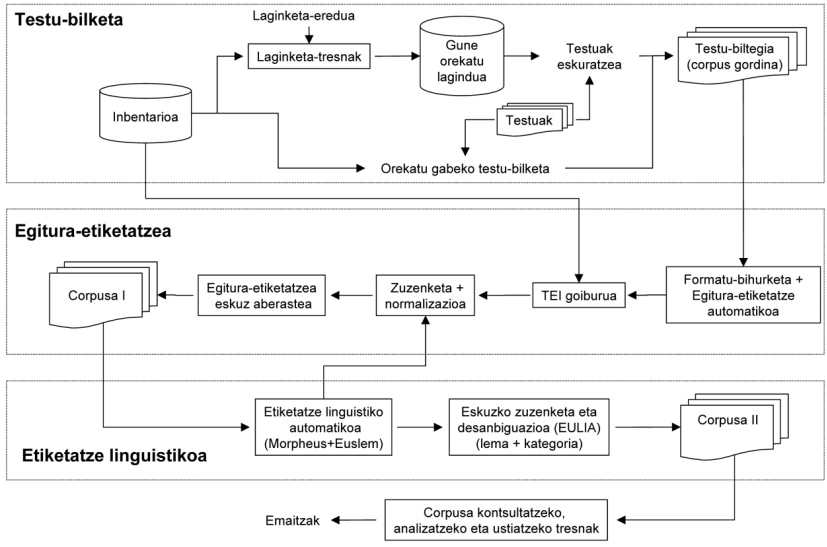
```
   <xptr id='Xw1' doc='TDoc03' from='id(p1) strLoc(1)'
to='id(p1) strLoc(4)'/>
```

Horrek adierazten du 'Xw1' tokena 'p1' identifikatzailea duen elementuaren barruan dagoela, eta horren lehenengo posiziotik laugarrenera doan testu-zatia dagokiola (*Hala* hitza). Erabilitako estekak,

berriro ere, bat datoz TEI P4ko gidalerroekin, eta, adibide xume honekin ikusi dugunez, ahalmen handia eskaintzen dute kanpoko dokumentuen zatiak identifikatzeko. Tokenizatzailearen dokumentuak soilik izango ditu esteka zuzenak jatorrizko dokumentuarekin, eta gainerako dokumentuek, analisi-katearen tokenizatzailearen ondorengo urratsak izanik, tokenizatzailearen dokumentuko elementuak izango dituzte iturri nagusitzat.

HAULen egitura izeneko dokumentuak testuan identifikatu diren hitz anitzeko unitate lexikalak adierazten ditu (adibidez, *Hala ere* unitatea). *Lematizazioak* dokumentuan, berriz, testuaren analisiak gordetzen dira, eta, besteak beste, hitzaren lema, kategoria lexikala, kasua eta funtzio sintaktikoak gordetzen dira bertan. Azkenik, *Estekak* dokumentuak jatorrizko testuko tokenak beren analisiekin lotzen ditu.

Markaketarako erabili dugun ereduak hitzen analisi anbiguoak adierazteko aukera ematen du; nolanahi ere, ZT corpusaren gune



9. irudia. Corpusgintzaren diagrama

orekatuan, prozesamendu automatikoa egin ondoren gelditzen diren analisi anbiguoen lema eta kategoria eskuz desanbiguatzeko konpromisoa hartu dugu (horrek esan nahi du gune orekatuan lema eta kategoria % 100 desanbiguatuta egongo direla).

Orain arte, corpusgintzaren hiru urratsak azaldu ditugu: corpus gordina eratzea edo testu-bilketa, egitura-etiketatzea eta etiketatze linguistikoa. 9. irudian ageri den diagraman bildu ditugu urrats horien eta horien barneko prozesu nagusiak.

6. Corpusa egiteko tresna: *Corpusgile*

Tresna horren helburuak:

- Corpusgintza modu sistematikoan antolatzeko metodologia eta teknologia eskaintzea
- Corpusgintzan arituko diren lantaldeek lan-eredu eta metodologia bera erabiltzea une oro
- Corpusgintzaren etorkizuneko helburua den erreferentzia-corpus orokorra egiteko baliagarria izango den metodologia adostua eta kontrastatua eskaintzea
- Corpusgintza nazioarteko gaur egungo estandarren araberakoa izatea

Corpusgile tresna eratzeko arrazoiak:

- Hizkuntz Teknologien alorrean, corpus-beharra handia da, eta oso garrantzitsua da egiten diren corpusak berrerabilgarriak izatea
- Corpusgintza prozesu konplexua da, baliabide eta tresna askoren integrazioa eskatzen du, eta prozesua osatzen duten urratsen gaineko kontrola eta horien arteko informazio- eta dokumentu-fluxuaren kontrola behar-beharrezkoa da

- Corpusgintza diru-ezartze handiak eskatzen dituen prozesua izaki, kostuak minimizatzeko prozedurak eskaintzea interes handiko ideia da
- Merkaturatu diren corpusgintza-tresna urriek ez dute euskararen prozesamendu automatikorako beharrezkoak diren tresnak eta baliabideak integratzen, eta ez dira egokiak euskarazko testu-corpusak eratzeko

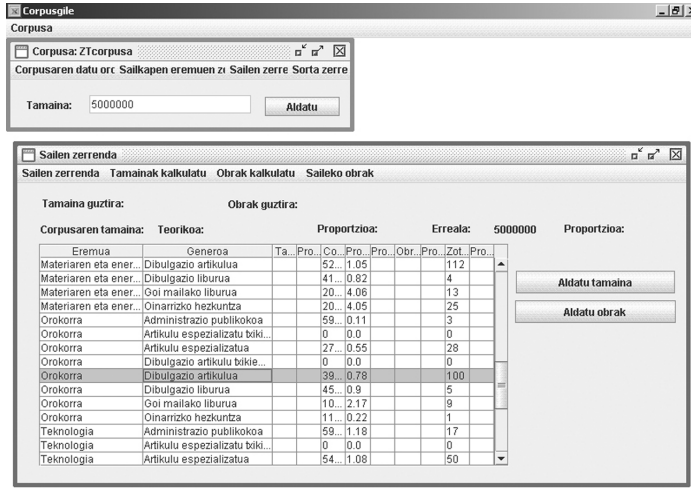
6.1. TB modulua

Urrats honen helburua corpusean jasoko diren testuak hautatzeko, jasotzeko eta biltegiratzeko sistema diseinatzea eta inplementatzea da. Modulu horren helburuak:

- Prozesuen gaineko kontrola bermatzea
- Prozesuak albait automatizatzea

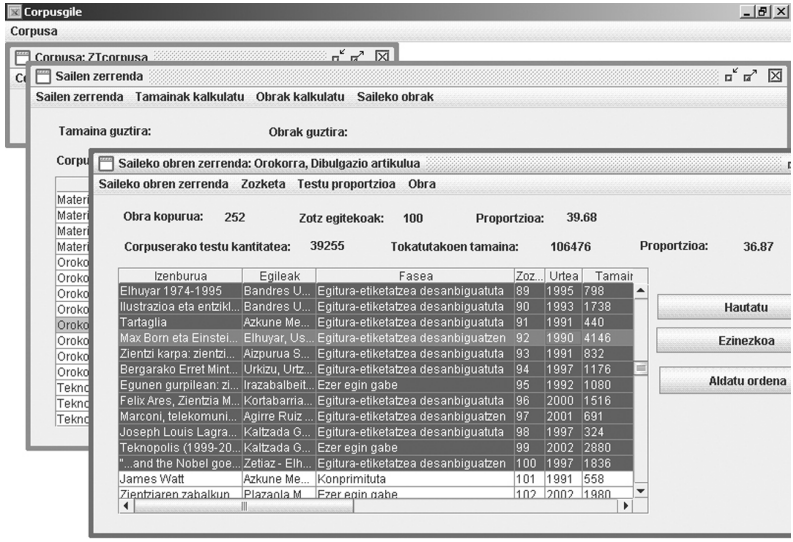
Horretarako, honakoak egin ditugu:

- *Corpusean sartzeko hautagai diren obren inbentarioaren datu-basea*



10. irudia. Inbentarioa sailka (geruzak)

- *Corpusean sartuko diren testu-laginak hautatzeko laginketa tresnak*. Corpus-diseinuan erabakitzen den laginketa-eredua modu automatikoan gauzatzeko tresna bat diseinatu eta inplementatu da. Tresna horren bidez, inbentarioan zehaztutako populaziotik lagin batzuk hautatzen dira, laginketa-ereduan ezarritako irizpideak automatikoki erabiliz



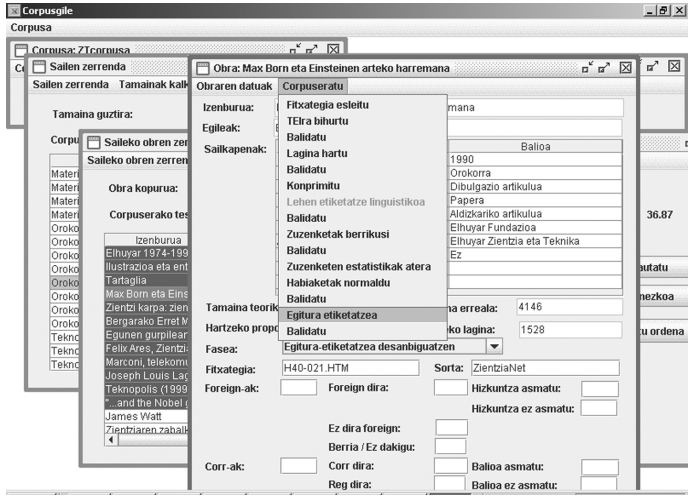
11. irudia. Sail baten laginketaren emaitza
(corpuseratzeko obrak urdinez nabarmenduta daude)

- *Corpusean sartuko diren testu-laginen biltegi egituratua*. Laginketan hautatutako testu-laginak (testuak edo testu-zatiak), horien jabe edo hornitzaileengandik jasotakoan, biltegi egituratu batean gordetzen dira. Biltegi hori diseinatu eta inplementatzean, honakoak hartu dira kontuan:
  - Testu-lagina eskuratzeko bidea (dokumentu elektronikoa, OCRz digitalizatu eta ezagututako dokumentu inprimatua, tekleatutako dokumentua)

– Dokumentu elektronikoen kasuan, onartuko diren jatorrizko formatuak (.txt, .doc., .rtf, .html, .xml, .pdf, .qk...)
– Corpusean testu-lagin bakoitzari buruz etiketatuko den informazioa (metadatuak: izenburua, egilea, urtea, etab.); informazio hori zein izango den diseinuan erabakitzen da, eta inbentariotik eta laginketa-eredutik sortzen da. Hurrengo urrats batean, testu-laginaren `<teiHeader>`-en kodetzen da, egitura-etiketatzearen aurretik edo ondoren

6.2. EE modulua

Modulu honetan, bilketa-moduluaren irteeratik jatorrizko formatuan datorren testu-lagina egitura-etiketatzerako erabakitzen den ereduaren araberako formatura bihurtzen da. Etiketatze-lan horretarako tresna automatikoak edo erdiautomatikoak inplementatu dira. Modulu honen irteera egitura-etiketatzea formatu jakin batean gauzaturik duen dokumentua edo dokumentu-multzoa da, eta etiketatze linguistikoa gauzatzen den moduluaren sarrera izateko prest dago.



12. irudia. Egitura-etiketatzearen moduluaren menu nagusia

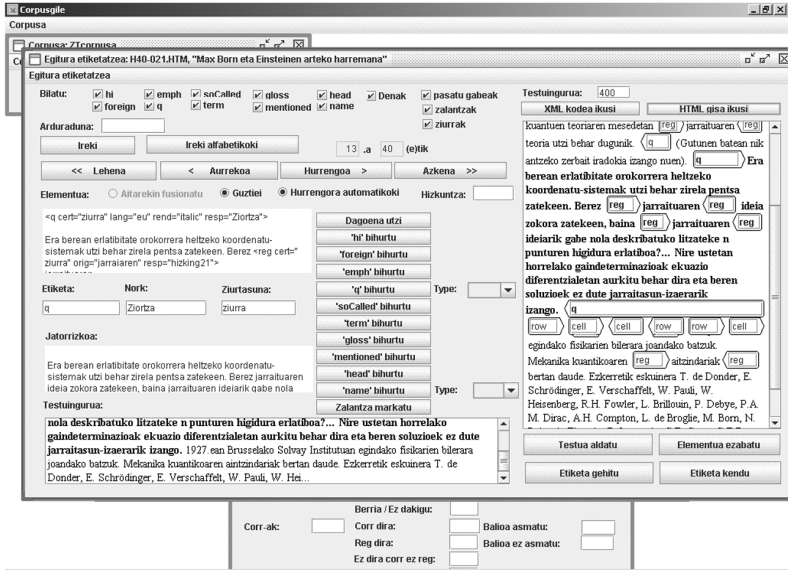
Horretarako guztirako, honako urratsak egiten dira:

- *Jatorrizko formatuak TEI_XML bihurtzeko parserrak*. Urrats honetan, egitura-etiketatzea bera gauzatzen da. Prozesuaren inputa bilketa-moduluaren irteeratik jatorrizko formatuan datorren testu-lagina da. Outputa, berriz, automatikoki etiketatutako eta TEI_XML formatura bihurtutako dokumentua. Hori egin ahal izateko, jatorrizko formatu bakoitzerako bihurtze- eta etiketatze-programak (parserrak) egin dira
- *Gune orekatuko testu-laginak automatikoki markatzea*
- *Erabiltzailearen hiztegia elikatzea*. Etiketatze linguistikoaren etekina handitzeko, EUSLEMek darabilen hiztegian ez dauden, eta, beraz, ezagutzen ez dituen testu-hitzen lemak 'Erabiltzailearen hiztegian' sartzeko aukera dago. Kontuan hartu behar da zientzia eta teknologiaren alorreko testuetan ugari direla lema espezializatuak, eta espero izatekoa da horietako hainbat hiztegi orokorretan ez egotea. EUSLEMi testu-lagin gordinak prozesarazi eta analisi zuzenik ematen ez duten testu-hitzak itzultzeko eskatzen zaio. Horrelako testu-formak maiztasunaren arabera ordenatu, eta erabiltzaileari bistaratzen zaizkio, egoki deritzonean landu eta erabiltzailearen hiztegira esportatzeko
- *Akats tipografikoak eta estandarizazioa:* `<corr>` eta `<reg>` elementuak etiketatu ahal izateko, benetako etiketatze linguistikoaren aurreko prozesatze linguistiko berariazkoa egiten dugu. EUSLEM etiketatzaileak automatikoki markatzen ditu ezagutzen ez dituen hitzak, eta bi proposamen-mota egiten ditu:
  - Testu-hitza aldaera ez-estandartzat identifikatzen duenean, aldaera estandarrari dagokion testu-hitza txertatzen du testuan, `<reg>` elementuaren barnean (jatorrizko testu-hitza `orig` atributuaren balioa da)

– Testu-hitza ezagutzen ez duenean, EUSLEMek `<corr>` elementuan sartzen ditu zuzentzat dauzkan proposamenak; hitza ez ezagutzeko arrazoia akats tipografikoa izaten da batzuetan; beste batzuetan, berriz, hitza EUSLEMen hiztegian (EDBLn) ez egotea

Gero, horiek denak eskuz aztertzen dira, balioesteko edo behar diren aldaketak egiteko. Horretarako interfazean, lema berriak sar ditzake 'Erabilzailearen hiztegia'-n. Eskuz egiten den orrazketa horren emaitzak berrerabili egiten dira, sistemak 'ikas' dezan eta asmatze-tasa handiagoa lortzeko

• *Egitura-etiketatzea lantzea:* urrats honetan, `<hi>` elementuak desanbiguatu egiten dira (**5.1.2.2 Nabarmentzea eta aipuak** ataleko elementuak erabiliz), elkarrizketetako galdera-erantzunak markatu egiten dira, `<head>` elementuen maila erabakitzen da, eta abar



13. irudia. Egitura-etiketatzea lantzeko interfazea

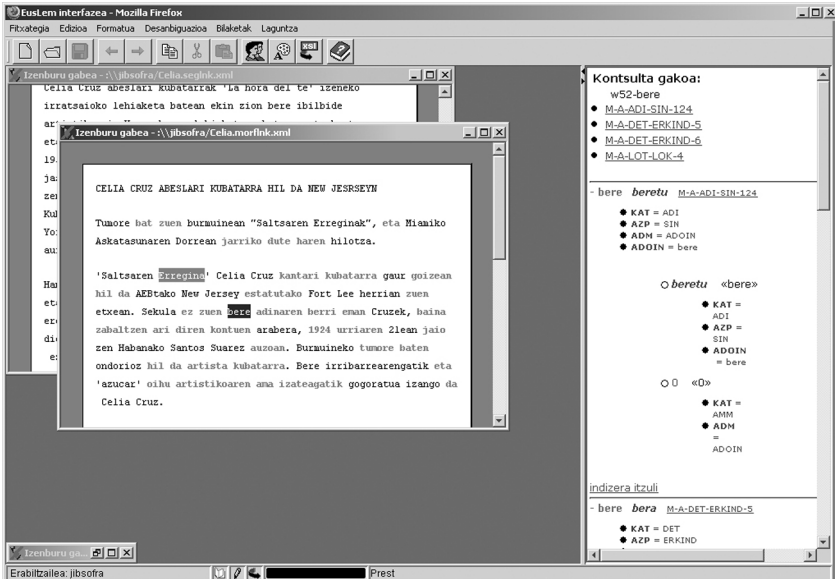
- *Testuen egitura-etiketatzea balioestea*: aurreko urratsetako bakoitzaren ondoren, xml dokumentuak automatikoki ebaluatzen dira DTDa erabiliz, eta sortzen diren erroreak zuzendu egiten dira, dokumentuak hurrengo urratsera iragan aurretik baliozkoak izan daitezen

6.3. EL modulua (EULIA)

EL modulua (EULIA) corpusaren gainean etiketatutako informazio linguistikoa gainbegiratzeko, orrazteko eta desanbiguatzeko ingurunea dugu, eta giza erabiltzaileari zuzenduta dago. EULIA lagun, linguistek zein etiketatzaileek aurreko urratsetan sortutako informazio linguistiko guztia aztertzeko aukera dute, eta, nahiz izanez gero, informazioa gehitu, aldatu edo zuzendu.

Besteak beste, honako aukera hauek eskaintzen ditu EULIAk:

- *Analisiak ikustatzea eta nabigazioa*: erabiltzaileek aurreanalizatutako corpusaren analisiak ikustatu eta bertan nabigatzeko aukera dute. Auzi honetan, EULIAk ingurune intuitiboa eskaintzen du
- *Eskuzko zuzenketa eta desanbiguazioa*: erabiltzaileak analisi baten gainean lan egiteko aukera du, alegia, analisi linguistikoen gainean zuzenketak eta eskuzko desanbiguazioa egin dezakete. Aldaketak edo zuzenketak egiterakoan, sistemak bermatzen du analisi berriak TEIko gidalerroen araberakoak direla
- *Oinarrizko bilaketak eta bilaketa aurreratuak*: erabiltzaileak desanbiguaziolanean eta analisien azterketan lagunduko dioten bilaketak egin ditzake. Bi bilaketa mota aurreikusten dira, bilaketa aurredefinituak, sinpleak eta parametrorik gabekoak, eta bilaketa aurreratuak, aukera zabalagoa ematen dutenak. Azken mota horretako bilaketetan, analisien ezaugarrien araberako parametroak jasotzen dira eta baldintza hauek betetzen dituzten hitzak bilatzen dira



14. irudia. EULIAren lan-interfazea

Irudi horretan, EULIAren interfaze grafikoa ageri da. Irudiaren ezkerraldean *Testu-leihoa* dugu, eskuinekoa *Analisi-leihoa* da, eta beheko aldean dagoena, *Egoera-kutxa*. Ikus ditzagun hirurak banan-banan.

6.3.1. Testu-leihoa

Testu-leihoan, sarrera-testua (gure kasuan, egitura-etiketatzearen emaitza), tokenizazioaren emaitza eta HAULen fitxategia prozesatzearen ondorioz sortutako testu-egitura bistaratzen da. Leiho honetan bi motatako osagaiak nahasten dira:

- *Tokenizatzaileak ezagutu dituen zatiak*: linguistikoki interesgarriak diren testu-zatiak dira, hau da, tokenak. Tokenek interes linguistikoa dute, eta hauei lotuta analisi linguistikoak egon daitezke

- *Tokenizaziotik at gelditu diren zatiak*: multzo hau hutsuneek, lerro-jauziek, eta abarrek osatzen dute. Mota horretako osagaiek ez dute analisi linguistikorik

Itxurari dagokionez, ezin dira osagai horiek ezberdindu. EULIA-ren helburuetako bat jatorrizko testua idatzita dagoen modu berean erakustea da; beraz, tokenizaziotik at gelditu diren zatiak, linguistikoki interesgarriak izan ez arren, erakutsi egin behar dira.

Testu-leihoan token baten gainean klik egiten dugunean, horrekin erlazionatzen diren tokenen araberako ekintzak abiaraz daitezke:

- *Klikatutako tokena ez da HAUL baten parte:* kasu honetan, klikatutako tokena tokenizazio-fitxategian azaldutako offseten arabera markatzen da, eta bere analisi guztiak Analisi-leihoan erakusten dira
- *Klikatutako tokena HAUL baten edo gehiagoren parte da:* tokena eta dagokion HAUL bakoitza markatzen dira, eta Analisi-leihoan tokenaren analisiak eta markatutako HAULenak erakusten dira

Testu-leihoko hitzei erreparatzen badiegu, bertan zenbait hitzek marka bereziak dituztela ikusiko dugu; horiek interfazearen komandoak erabiliz egindako bilaketa edo markatze bereziak dira. Bilaketa baten emaitzak Testu-leihoko tokenak formatu berezietan markatuz erakusten dira. Adibidez, irudian token bat letra lodiz agertzeak adierazten du analisi anbiguoak dituela; itzalduta dagoen tokena (*bere*) erabiltzaileak klik eginez aukeratutako hitza da; azkenik, laukiaren barnean dagoen tokena (*Erregina*) erabiltzaileak arrazoi jakin batengatik marka batez nabarmendu duen tokena dugu.

Marka horiez gain, interfazeak beste marka mota batzuk egiteko aukera ere ematen du. Marka guztien itxura erabiltzaile bakoitzarentzat pertsonaliza daiteke.

6.3.2. Analisi-leihoa

Leiho honetan, Testu-leihoan markatutako tokenekin erlazionatutako analisiak erakusten dira. Analisia erakusteko, zenbait estilo-orri definitu dira, erakutsi beharreko analisi-mota eta ikusi nahi den informazioaren espezializazio-maila kontuan izanik. Horri esker, XML fitxategiak EULIAren azpian gelditzen dira, eta erabiltzaileak modu gardenean ikus dezake informazio linguistikoa.

Estilo-orriak horrela erabilita, leiho honek izan ditzakeen funtzionalitateak irekita gelditzen dira. Hemen erakusten den informazioa eta erabiltzaileekin duen harremana estilo-orri baten bidez defini daiteke. Erabilpen berezietarako, estilo-orri konplexuak defini daitezke, eta Analisi-leihoan komandoak edo bilaketa berriak egiteko aukerak gehitu daitezke. Hau tresna indartsua da, eta, unean tratatzen den informazioaren arabera, interes gehien dituzten ekintzak eskain ditzakegu, modu adimentsuan.

Testu-leihoko dokumentu bakoitzeko, Analisi-leiho bat dago; horretara, aktibo dagoen dokumentuaren arabera, analisi bat edo beste erakutsiko dugu.

6.3.3. Egoera-kutxa

Osagai hau sinplea da; erabiltzailearen inguruko informazioa (izena eta baimenak) eta interfazearen egoera orokorra azaltzen ditu.

7. **Ondorioak**

Zientzia eta Teknologiaren corpusaren bidez, baliabide egoki eta ahaltsu bat eskaini nahi dugu espezialitate-alor horietan erabili den hizkuntza aztertzeko. Euskara ez da hasiberria alor horietako testugintzan. 30 urte baino gehiago iragan dira zientzia eta teknologiako lehen testuak argitaratzen hasi zirenetik. Handik hona egin den bide-

aren zati bat baino ez du bere baitan hartzen ZT corpusak, 1990-2002 bitartekoa alegia, baina gure iritzia da aski datu-bilketa egokia izan daitekeela, batez ere kontuan hartzen badugu aurreko urteetako hizkuntza erabileraren 'heldutasuna' urte-bitarte horretan erdietsi bide duela euskarak, eta horretan eragin handia izan dutela Euskaltzaindiaren araugintza berriak eta Hiztegi Batuak.

Bestetik, corpusa eratzeko metodologia zehaztu eta ezarri dugu, corpusgintzan behar diren tresnak eta baliabideak moldatu edo garatu ditugu (TB eta EE moduluak, EULIA), eta prozesu osoa bere baitan hartzen duen aplikazio batean, *Corpusgile*-n, integratu.

Hizkuntza orok bezala, euskarak ere corpusak behar ditu; hizkuntzalariek, terminologoek, hizkuntz teknologien ikertzaileek, hizkuntzaren estandarizazioaren ardura dutenek, hainbatek behar ditu corpusak, gaur egun hizkuntza aztertzeko ezinbesteko baliabide direlako. Baina corpusak berak ez ezik, horiek eratzeko teknologia ere behar dugu, corpusgintza-prozesua behar bezala bideratu eta kudeatzeko, eta hain handiak izaten diren kostuak gutxitzeko. Horiek biak dira, baliabide eta tresna bana, hain beharrean gauden alor honetara egin nahi ditugun ekarriak.

8. Esker ona

Andoni Sagarna eta Yosu Yurramendiri, haien jakintza sakonaz eta zabalaz aberastu dutelako proiektua, eta bidegurutze zein ataka gaiztoetan bide-erakusle izan ditugulako.

Lan hau Eusko Jaurlaritzako Industria Sailaren Etortek programaren diru-laguntza jaso duen Hizking21 proiektuaren baitan egin da.

Bibliografia

ALDEZABAL, I., ALEGRIA I., ANSA O., ARREGI X., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., GOJENOLA K., HERNÁNDEZ G., MAYOR A., ORONOZ M. & SOROA A. 2002. *Hizkuntza prozesatzeko tresnen integrazioa, SGML erabiliz.* Barne-txostena. UPV/EHU/LSI/TR 2-2002.

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.

BACH, C., SAURI, R., VIVALDI, J. & CABRÉ, M.T. 1997. *El corpus de l'IULA: descripció*. Bartzelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. [on line] [kontsulta: 05-01-22] <ftp://ftp.iula.upf.es/pub/publicacions/97inf017.pdf>

BADIA, T., CABRÉ, T., PUJOL, M., TUELLS, A., VIVALDI, J. & DE YZAGUIRRE, Ll. 1998. "IULA's LSP Multilingual Corpus: compilation and processing", In *ELRA conference*, Granada, 1998. [on line] [kontsulta: 05-01-22] <ftp://ftp.iula.upf.es/pub/publicacions/granact.pdf>

BIBER, D., CONRAD, S. & REPPEN, R. 2000. *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BIBER, D. 1993. "Representativeness in Corpus Design." In *Literary & Linguistic Computing* 8. 243-257. orr.

BOWKER, L & PEARSON, J. 2002. *Working with Specialized Language. A prtactical guide to using corpora.* New York: Routledge

GRÖNQVIST, L & HELGADÓTTIR, S. 2002. *Literature review of representativeness of linguistic resources.* GSLT course on Linguistic Resources.

KENNEDY, G. 1998. *An introduction to Corpus Linguistics*. Londres: Longman. Studies in Language and Linguistics.

LEECH, G. 2002. "The Importance of Reference Corpora." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI. [on line] [kontsulta: 05-01-22] <http://www.uzei.org/corpusajardunaldia/06_gleech.pdf>

MacMULLEN, W. J. 2002. *Requirements Definition and Design Criteria for Test Corpora in Information Science*. SILS Technical Report 2003-03 School of Information and Library Science University of North Carolina at Chapel Hill [on line] [kontsulta: 05-01-22] < http://ils.unc.edu/ils/research/reports/TR-2003-03.pdf >

OIHARTZABAL, B. 2002. "Euskaltzaindiaren corpusez." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI [on line] [kontsulta: 05-01-22] <http://www.uzei.org/corpusajardunaldia/07_boihartzabal.ppt>

SINCLAIR, J, 1996. *Preliminary Recommendations on Corpus Typology*. EAGLES. [on line] [kontsulta: 05-01-22] <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>

URKIA, M. 2002. "XX. mendeko euskara-corpusa." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI [on line] [kontsulta: 05-01-22] <http://www.uzei.org/corpusajardunaldia/03_murkia.pdf>

Text Encoding Initiative. *The XML version of the TEI Guidelines*. [on line] [kontsulta: 05-01-22] <http://www.tei-c.org/P4X/>

VIVALDI, J., DE YZAGUIRRE, Ll., SOLÉ, X. & CABRÉ, M.T. 1996. *Marcatge Estructural i Morfosintàctic del Corpus Tècnic amb l'estàndar SGML*. Bartzelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. Serie Informes, 1. [on line] [kontsulta: 05-01-22] <ftp://ftp.iula.upf.es/pub/publicacions/96inf001.pdf>

YANG, D.H., CANTOS, P. & SONG, M. 2000. "An Algorithm for Predicting the Relationship between Lemmas and Corpus Size." In *ETRI Journal*, 22/2: 20-31. [on line] [kontsulta: 05-01-22] <http://etrij.etri.re.kr/Cyber/servlet/GetFile?fileid=SPF-1042453354988>