ZIENTZIA ETA TEKNOLOGIAREN CORPUSA DISEINUA ETA METODOLOGIA

N. Areta, A. Gurrutxaga, I. Leturia, Z. Polin, R. Saiz

Elhuyar Fundazioa

I. Alegria, X. Artola, A. Diaz de Ilarraza, N. Ezeiza, K. Fernández, A. Sologaistoa, A. Soroa, A. Valverde

Ixa taldea. Euskal Herriko Unibertsitatea

1 Sarrera

Corpusak azken urteotan hizkuntza-baliabide gisa hartu duen garrantzia inork gutxik uka lezake gaur egun. Corpusa hizkuntza aztertzeko ezinbesteko baliabidea da, hainbat alorretan erabiltzen dena: lexikografian, sintaxian, semantikan, diskurtsoaren analisian... Adibidez, gaur egun, mundu zabalean egiten diren lexikografia-lanetan hutsik egin gabe aipatzen den hitza *corpus* da. Dela hiztegigintzaren osagarri, dela hiztegia bera egiteko lehengai eta abiapuntu, corpusa hiztegigilearen tresna eta euskarri ezinbestekotzat jotzen da gero eta maizago, hainbesteraino, non corpusean oinarritua edo, gutxienez, corpusaren laguntzaz taxutua ez den hiztegiari nekez aitortzen baitzaio zehaztasuna, zorroztasuna, fidagarritasuna eta, oro har, kalitatea. Esan gabe doa, corpusak ez dira lexikografian soilik erabiltzen, diziplina oso bat da corpus--hizkuntzalaritza. Corpusek datu linguistikoak jasotzen dituzte eta baliozko lanabesak dira hizkuntzaren erabilera erreala aztertu nahi bada. Urte askoan, horrelako azterketa enpirikoak egitea hizkuntzalaritzaren korronte nagusitik kanpo egon bada ere, azken urteotan gero eta tresna estimatuagoak dira, ez noski gramatika sortzailearen alternatiba edo aurkari gisa, beste ikuspegi baten eta ebidentzien ekarle gisa baizik. Gainera, corpusetan bildutako datu-kopuru handien bidez, hizkuntza-teknologien alorreko behar eta eginkizun batzuei beste era batera erantzuteko modua egoten da (prozedura estatistikoetan oinarritutako desanbiguazio-teknikak, itzulpen-memoriak, etab.).

Berez, edozein testu-bilduma har liteke corpustzat; hala ere, gaur egun baldintza batzuk ezarri ohi dira testu-bilduma bat 'corpustzat' jotzeko: hizkuntza-erakusgarri 'errealen' multzo 'handia' izatea, irizpide batzuen arabera bildua, formatu elektronikoan

biltegiratua eta informazio linguistikoz hornitua (Bach *et al.* 1997: 4). Baldintza horien guztien helburua, azken buruan, corpusa hizkuntza-baliabide eraginkorra izatea da, hau da, corpusetik datu linguistiko asko, aberatsak eta esanguratsuak lortzeko aukera izatea.

2 Corpus berezia

Badira hogeita hamar urte baino gehiago euskara zientzia- eta teknologia-gaietan erabiltzen hasi zela. Geroztik egindako lanaz eta, bereziki, handik hona argitaratu diren lanez, iritzi desberdinak daudela esan daiteke; mutur banatan, honako hauek: batzuentzat aintzat hartzekoa dena, 'tradizio berria' dena, baztergarritzat edo ez ikusia egiteko modukotzat dute beste batzuek. Ez gaude ados azken urteotako testu-produkzio horren deskalifikazio orokorrarekin eta baztertzeko joerarekin, ezta ezinbestean segitu beharreko eredua finkatu delako ustearekin ere; hau da, uste dugu jarrerak 'kritikoa' izan behar duela. Gure iritzia da testu-produkzio hori aztergaitzat hartu behar litzatekeela, eta, horretarako, corpusa behar dela.

Euskaraz azken urteetan eratu diren corpus lematizatuak 'orotarikoak' dira (*XX. mendeko euskararen corpus estatistikoa;* Urkia 2002: 6), edo, *Ereduzko Prosa*-ren *zein Ibinagabeitia Proiektua*-ren kasuan, literatura edota prentsa jasotzen dituzte. Lematizatu gabeko corpusak ere badaude (*OEHko Testu-corpusa*, *Klasikoen Gordailua*...).

Zientzia eta Teknologian erabiltzen den euskara aztertzeko, alor horietako testuak biltzen dituen corpusa erabiltzea litzateke zentzuzkoena. Hau da, behar berezi horri erantzuteko, egokiena baliabide 'berezi' bat eratzea delakoan gaude, hartarako berariazkoa hain zuzen ere. Asmo horri 'corpus berezia' dagokio. 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 (Sinclair 1996: 10). Corpusaren helburua hizkuntzaren erabilera-eremu guztietarako baliagarria edo 'adierazgarria' izatea denean, 'erreferentzia-corpusa' edo 'orotariko corpusa' dela esan ohi da (Sinclair 2002: 10; Leech 2002: 1).

Corpus berezien bidez, erabilera-eremu espezifiko baten edo aldaera jakin baten hizkuntza-ezaugarriak hobeto aztertzeko aukera dago. Horrekin batera, espezialitate-arloetako hizkuntza-erabileraren eta erabilera arrunt edo orokorraren arteko aldeak ere azter daitezke. Aztergaiak hizkuntzaren aztertze-eremu askotakoak izan daitezke:

lexikoa, terminologia, fraseologia, morfosintaxia, semantika, pragmatika, diskurtsoa, estilistika, testugintza... Aztertze-eremu horiek hainbat aplikazio-eremutan izan daitezke baliagarri: terminologiaren normalizazioa, hiztegigintza espezializatua (hiztegi terminologikoak eta teknikoak), hiztegi orokorretan sartzekoak diren termino espezializatuen hautaketa, terminoen erauzketa erdiautomatikoa, kontzeptu-mailako informazioaren erauzketa, ontologiak eratzeko teknikak, testu-sailkapen automatikoa, hitz-adieren desanbiguazioa, informazioaren berreskurapen eta erauzketa; xede berezietarako hizkuntza-irakaskuntza (curriculuma, syllabus-ak)…

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. Beraz, gure ikuspegia ez da eredu-emaile izatea, ez ditugu corpuserako obrak 'kalitate-irizpide' baten arabera bahetuko. 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.

3 Corpusgintza-eredua

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 eta nola prozesatuko eta etiketatuko diren…
- Corpus gordina eratzea: corpuseratzekoak 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

Hurrengo ataletan, eredu horren araberako corpusgintza-prozesua azalduko dugu.

4 Diseinua: ZT Corpusaren ezaugarriak

Zientzia eta Teknologiaren corpusean, euskaraz 1990-2002 bitartean argitaratu diren zientzia eta teknologiaren alorreko obrak jaso nahi ditugu. Bi data bat datoz, hurrenez hurren, Euskaltzaindiaren araugintza berriaren hasierarekin, eta proiektu honen hasierarekin berarekin. Corpusa bi ataletan antolatuta dago. Batetik, adierazgarria izateko asmotan diseinatu den gune orekatua; bestetik, eskuragarritasunaren arabera corpuseratzen diren obrez edo obra-zatiez osatutako atal irekia. Hain zuzen ere, gune orekatuan ez dira obra osoak sartzen, obren lagin etenak baizik. Horrek berekin dakar gune orekaturako aukeratu den obra baten pasarte ez hautatuak (lagin eten horien artekoak), eskura izanez gero, corpusaren atal irekian sar daitezkeela (gune orekaturako 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 erabaki egin behar da. Horretarako, lehenik 1990-2002 bitarteko zientzia eta teknologiaren alorreko obren inbentarioa egin dugu. Hurrena, adierazgarritasuna edo 'oreka' bermatuko duen lagintze-eredu estatistikoa landu dugu. Eredu horren lehen oinarria da laginketa geruzatua izatea, eta geruzak sortzeko erabili ditugun parametroak 'Eremua' eta 'Generoa' dira. Jakintza-arloak 'Eremua' parametroaren arabera sailkatu ditugu, eta testu-motak 'Erregistroa' parametroaren arabera:

- 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...)[1]

[1] '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 edo 'sail' bakoitzean eremu-genero konbinazio bakoitzeko obrak daude, eta laginketaren ausazkotasuna geruza horietako bakoitzean bermatzen da. Horrela jokatuz, ziurta dezakegu mota guztietako obrak ordezkaturik egongo direla gune orekatuan. Bigarren oinarria da geruza bakoitzaren tamaina, hasiera batean behintzat, geruzak populazioan duen proportzioaren araberakoa izatea; inbentarioa amaitutakoan, zenbait doikuntza txiki egin dira, geruza edo sail batzuen proportzio handiak txikiagotzearren. Landu den lagintze-eredu estatistikoan, honako hauek ere automatikoki zehazten dira: a) geruza bakoitzetik zenbat obra hartu behar diren; b) obra bakoitzetik zenbat hitz hartu behar diren (obraren tamainaren arabera); c) obra bakoitzetik lagin etenak hartzea (automatikoki egiten da XML dokumentuan). Lagin-tamaina minimoa 300 hitz da.

Gune orekatuan zenbat hitz sartu behar liratekeen kalkulatzerakoan, kontuan hartu ditugu, batetik, inbentarioko datuak aztertuz zenbatetsi den hitz-kopurua (98 milioi hitz), eta, bestetik, euskarazko bi corpus txikiren forma/lema erlazioaren azterketa eta estrapolazioak aurreikusarazi digun corpus-tamainaren eta lema-kopuruaren arteko erlazioa (erreferentzia). Horiek horrela, 5 milioi hitzeko gune orekatua diseinatu da.

Gune orekatuan biltzen diren laginak automatikoki prozesatu ez ezik, eskuz ere berrikusten dira, corpusgintzaren urrats bakoitzean egiten diren lanak zuzentzeko edo desanbiguatzeko. Atal irekia, berriz, automatikoki baino ez da prozesatzen. Dena den, etiketatze linguistikoan, atal irekiko testua eskuz halako masa handi bat landu ondoren prozesatzen da, eskuz egindako lanetik 'ikas' dezan, eta asmatze-tasa handiagoa izan dadin.

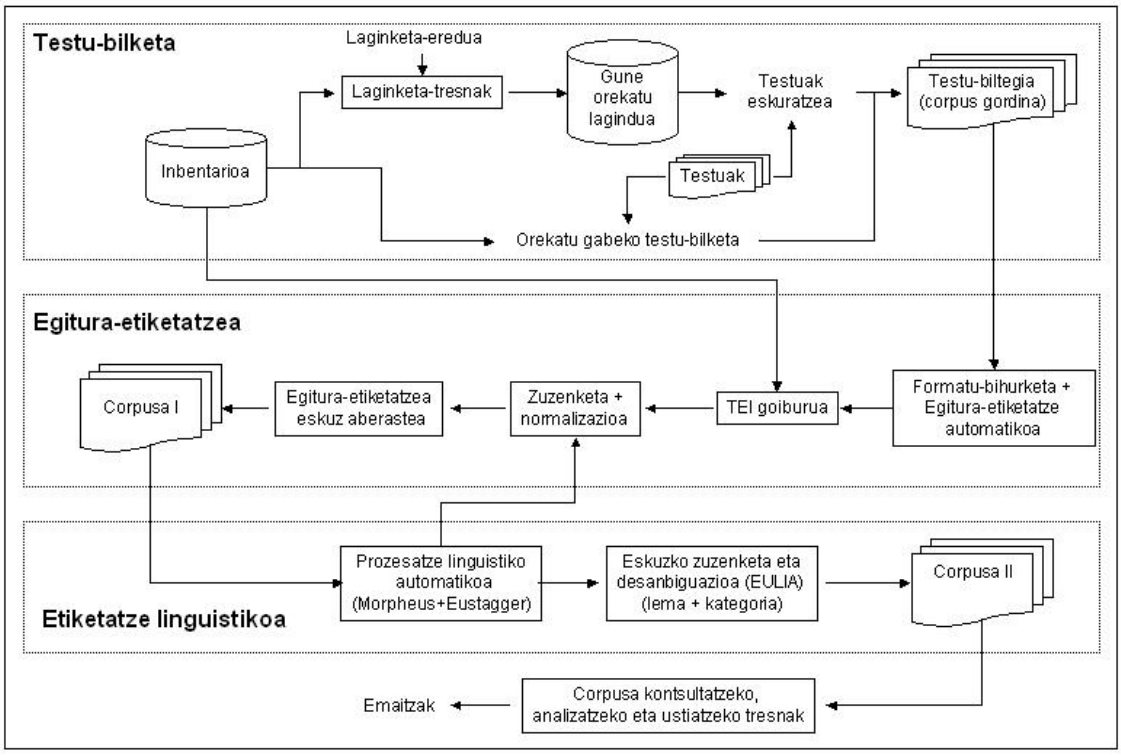
5 Cospusgintza-lana

Corpusgintza-ereduko urratsak modu sistematiko eta egituratuan egiteko, corpus-metodologia bat landu behar da, eta, hori inplementatzeko, corpusgintza-tresna bat. Lehendik garatutako 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 garatu 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, CORPUSGILEren bidez corpusgintzaren etorkizuneko helburua den erreferentzia-corpus orokorra egiteko baliagarria izango den metodologia adostua eta kontrastatua eskaini nahi izan dugu.

CORPUSGILE hiru moduluz osatua da:

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

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

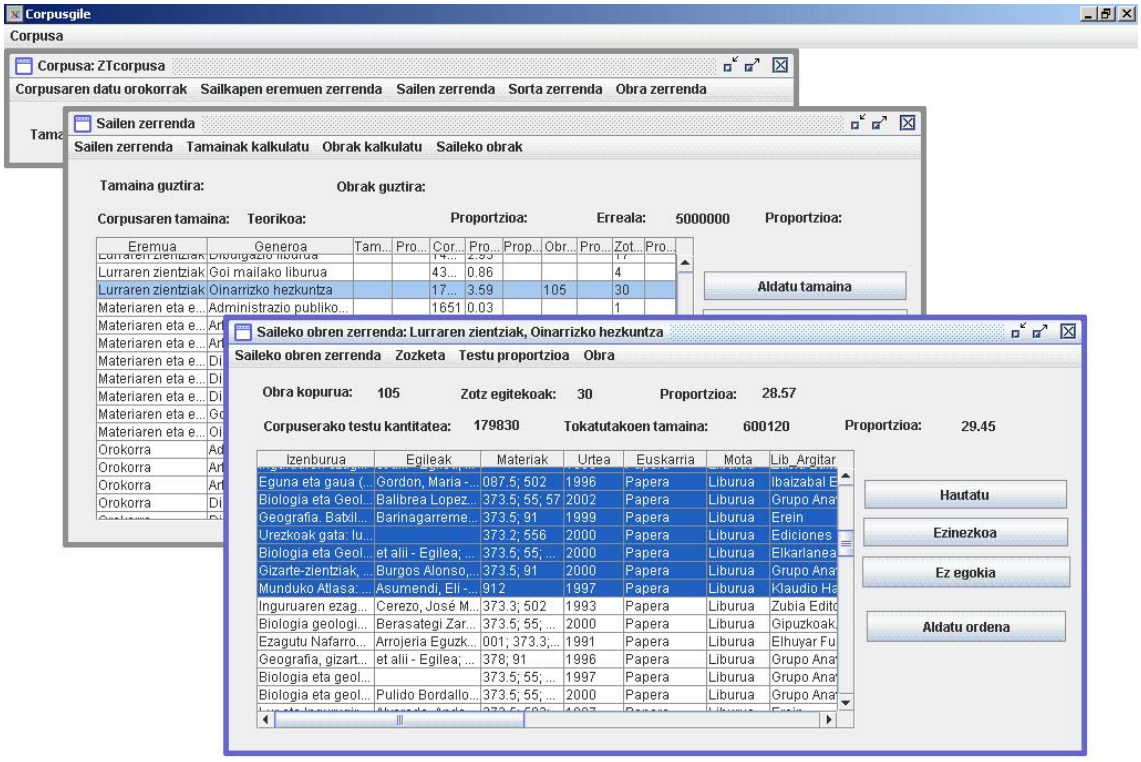


1. irudia. Corpusgintzaren diagrama

5.1 Corpus gordina

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 formatu horretan jaso ahal izateko, argitaratzaileengana jo dugu. Horretarako, corpusaren helburua, erabilera eta testuak corpuseratzeko baldintzak zehazten dituen hitzarmena sinatzea proposatu zaie hornitzaileei. Zenbaitetan ordea, ezin izan da testua formatu elektronikoan eskuratu, eta eskaner bidez digitalizatu behar izan dugu.

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).



2. irudia. TB modulua: sail baten laginketaren emaitza (corpuseratzeko obrak nabarmenduta daude)

5.2 Etiketatzea

Corpusak kodetzeko eta etiketatzeko proposatu diren ereduen eta formatuen artean, TEI eredua eta XML teknologia hautatu ditugu. TEI (Text Encoding Initiative) nazioarteko estandar bat da, testu elektronikoak kodetzeko eta trukatzeko orientabideak proposatzen dituena (Arriola *et al*. 1997: 6). Gure etiketatze-eredua koherentea da TEI P4ren orientabideekin, orokorrekin zein hizkuntza-corpusetarako emandako orientabide bereziekin (23. atala; http://www.tei-c.org/P4X/CC.html).

5.2.1 Egitura-etiketatzea

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>`, `<item>...` Testuaren joskeraren barnean irudi bat edo corpuseratuko ez den bestelako elementuren bat

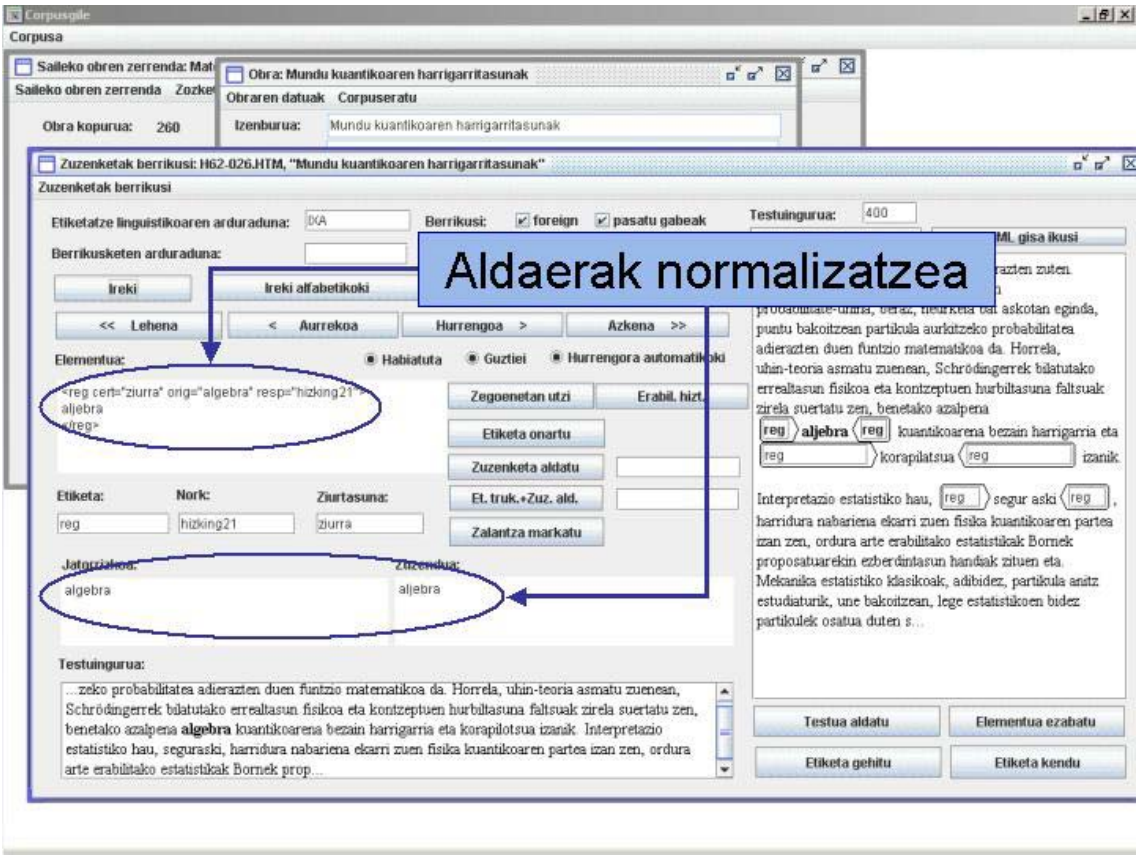
dagoenean (formulak, ekuazioak...), <gap> elementu hutsaren bidez adierazten dugu gune horretan zerbait 'falta' dela.

Etiketatze linguistikoen emaitzak hobetze aldera, zuzenketak eta aldaera ez-estandarrak etiketatzeko lana ere egiten dugu urrats honetan. Horretarako, <corr> eta <reg> elementuak erabiltzen dira. Adibidez:

```
<corr cert="ziurra" resp="hizking21" sic="baztuk">batzuk</corr>
```

```
<reg cert="ziurra" resp="hizking21" orig="zientzilari">zientzialari</reg>
```

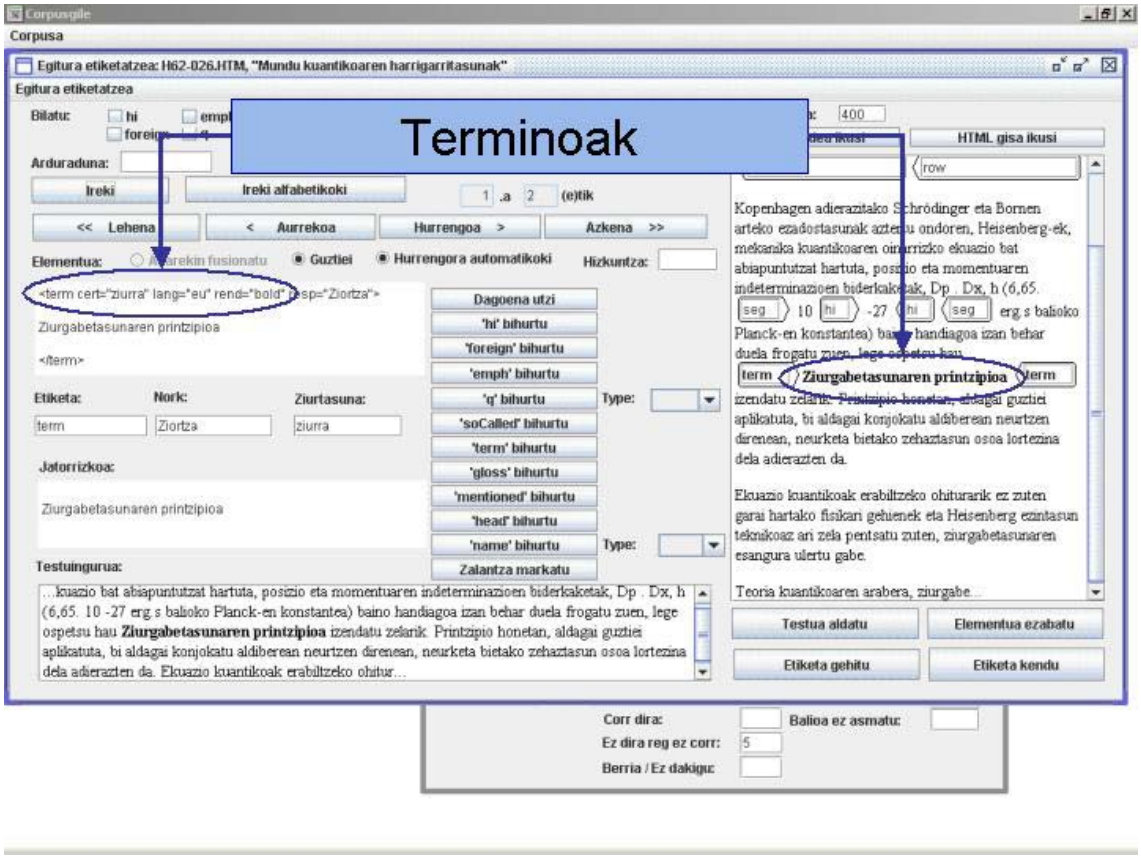
EUSTAGGER etiketatzaileak <corr> edo <reg> proposamenak automatikoki markatzen ditu testuan, eta gero horiek denak eskuz aztertzen dira, balioesteko edo behar diren aldaketak egiteko (eskuz landutako corpus-atalean, noski).



3. irudia. EE modulua: aldaera ez-estandarren normalizazioa

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>...),: Bestetik, TEIren DTDan, orekatua atributua erantsi diogu <p> elementuari. Horren bidez, corpus-gune orekatuan sartzen diren laginen paragrafoak markatzen dira.

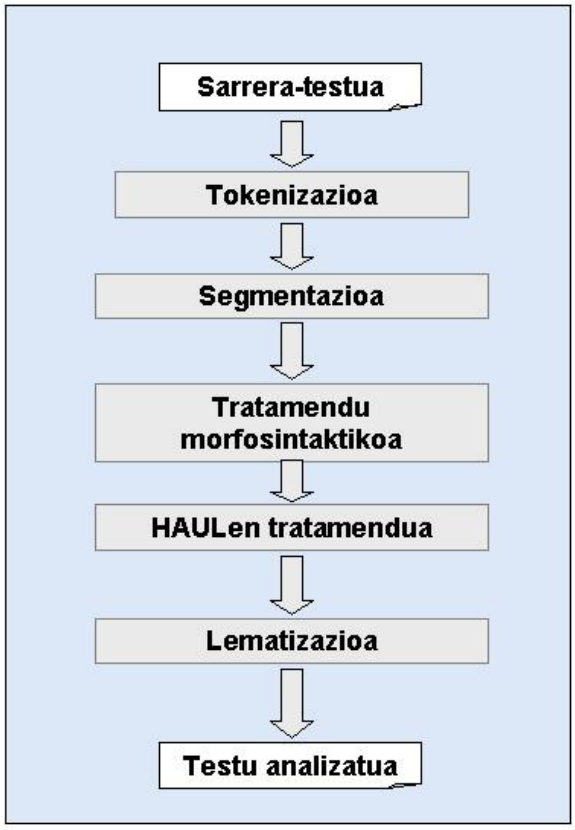


4. irudia. EE modulua: nabarmentzeen lanketa

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

5.3 Etiketatze linguistikoa

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



5. irudia. Prozesatze linguistikoaren oinarrizko eskema

Irudian (5. irudia) ikus daiteke testuek jasan duten prozesatze linguistikoaren eskema. Labur esanda, honako eragiketa hauek egin dira testuon gainean:

- 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 izango da (murriztapen-gramatika bat baliatuko da horretarako), alde batetik, eta estatistikoa, bestetik (ikaste automatikoko teknikak erabiliz, aldez aurretik eskuz desanbiguatutako corpus batean oinarrituz).

Lematizazioan aipatutako desanbiguatze hori automatikoa da (ez % 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. Eskuzko berrikuste hori EULIA izeneko tresnaz baliatuz egingo da.

Prozesu hauetan guztietan erabiltzen den 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 (hots, EDBLn ez dauden) termino zientifiko-teknikoak gehitu dira; beraz, termino horiek lematizatzen/etiketatzen direnean sistemak ez du beste aukerarik aztertuko, ez da saiatuko lexikorik gabeko lematizazioa egiten, alegia. Hiztegi hori osatzeko, bi irizpide jarraitu dira: Elhuyarren hiztegigintzako datu-basea (ElhDB) 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, aurreprozesatze bat egin da lematizatzeko arazoak ematen dituzten hitzak detektatu eta EUSTAGGERek proposatzen duen lemaren maiztasunaren arabera sailkatzeko. Zerrendaren buruan geratu diren lemak eskuz aztertu eta, egokitzat hartu direnean, erabiltzailearen lexikoan barneratu dira. Bi lan horiek etiketatze linguistikoaren beraren aurretik egiten dira, egitura-etiketatzearekin batera. Bigarren eginkizunerako, gainera, programa eta erabiltzaile-interfaze berezia garatu dira (CORPUSGILEren EE moduluan integratu da, zuzenketak eta aldaera ez-estandarrak etiketatzeko egitekoen aurretik).

Prozesu honen 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 (EDBLn ziurtzat jotzen diren eta testuan etenik gabe agertu ohi direnak).

Testuak linguistikoki etiketatzeko –anotatzeko–, 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 (Aldezabal *et al.*, 2002).

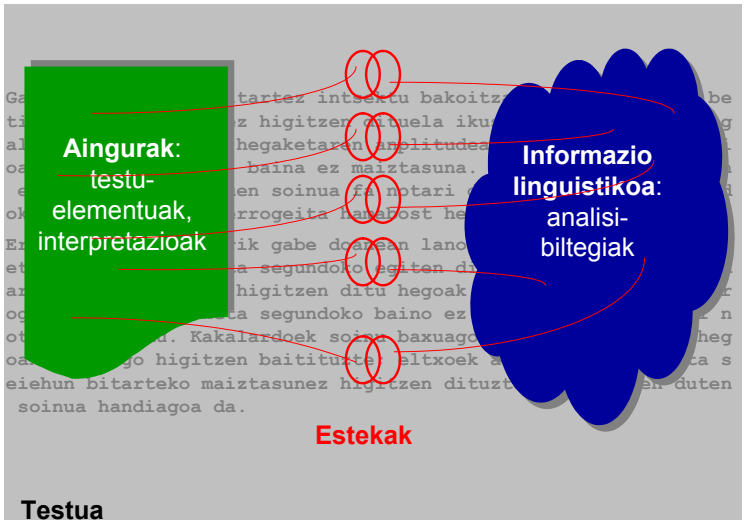
Informazio linguistikoaren konplexutasuna kontuan harturik, hurbilpen honek abantaila hauek eskaintzen dizkigu, besteak beste:

- Informazio teilakatua adierazteko aukera ematen du, eta, ondorioz, analisi linguistiko anbiguoak adierazteko.

- 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. Beraz, anotazioaren hedagarritasuna errazten du, corpusaren gainean informazio linguistiko osagarria txerta baitaiteke, dagoen informazioaren gainean oinarriturik.
- HAULen osaera errepresentatzeko modu egokia eskaintzen du, baita hitz anitzeko unitate horiek testuan etenda gertatzen direnean ere.
- Desanbiguatze-egoera (eskuzkoa, automatikoa) zein den adierazteko era egokia eskaintzen du.

Diagraman (6. irudia), anotazio banatuaren oinarrian dagoen eskema irudikatu da. Funtsean, hiru elementuk hartzen dute parte guk amarauna esaten diogun anotazio-arkitektura honetan:

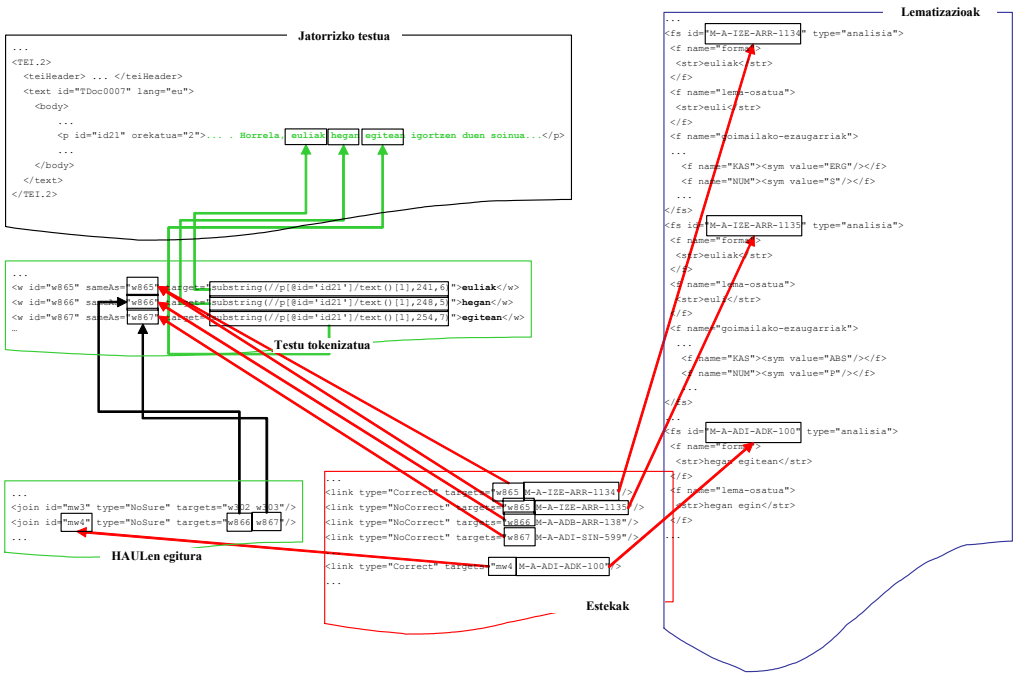
- Aingurak: testu-elementuak edo, oro har, aurreko prozesuek sortutako interpretazioak izan daitezke anotazioen jomuga edo helduleku.
- Informazio linguistikoa: prozesu linguistikoek eraikitako analisi-biltegiak, ezaugarri-egiturak (*feature structures*) erabiliz errepresentatuak.
- Estekak: aurreko biak lotzen dituzten elementuak, hau da, aingura bat (testu-hitz edo -zati bat, esate baterako) dagokion informazio linguistikoarekin (lematizazioaren emaitza, adibidez) estekatzen duen elementua.



6. irudia. Anotazio banatua, eskematikoki.

Hiru elementu horiek, praktikan, elkarrekin lotutako hainbat XML dokumenturen bitartez gauzatzen dira, testu batek jasan dituen prozesu linguistikoen emaitzak (anotazioak) biltzen dituztenak. Beheko irudian (7. irudia) adibide konkretu bat ikus daiteke: *Horrela, euliak hegan egitean igortzen duen soinua...* esaldiaren lematizazioaren ondoren izango genukeen anotazio-amarauna dago bertan irudikatua, eskematikoki. Kasu honetan, bost dokumentuk osatzen dute amarauna: jatorrizko testua (egitura-etiketatzearen emaitza), tokenizazioaren emaitza (*Testu tokenizatua*, irudian), lematizazioen bilduma (*Lematizazioak*), HAULen egitura errepresentatzen duena eta esteken dokumentua. Ikus daitekeenez, aingurak testu tokenizatuan eta HAULen egitura errepresentatzen duen dokumentuan aurki daitezke. HAULen egitura adierazteko, tokenen dokumentuko unitateen erakusleak erabiltzen dira, eta horrela errepresentatzen da, adibidez, *hegan* eta *egitean* tokenak lematizazio-unitate beraren osagai direla. Estekei erreparatuz gero, berriz, aise ohartuko gara interpretazio-anbiguotasuna nola errepresentatzen den (*euliak* formak bi lematizazio posible ditu: ergatibo singularra eta absolutibo plurala, eta, hortaz, bi estekak dute helduleku token horretan), eta baita desanbiguatze-egoera adierazten duen `type` atributuaren funtzioaz ere (`Correct` balioak adierazten du, desanbiguazioaren ondoren, interpretazio zuzena zein den). Azkenik, lematizazioen bilduma dugu informazio linguistikoaren atalean, non, ezaugarri-egitura batek errepresentatzen baitu hitz-forma desberdin bakoitzaren lematizazio-informazioa:

forma bera, lema osatua eta goi-mailako zein morfemaz morfemako informazio morfologikoa (kasua, funtzio sintaktikoa eta abar).



7. irudia. Etiketatze linguistikoa. Anotazio banatua: dokumentu-amarauna.

Etiketatze linguistiko automatikoa egindakoan, emaitzak eskuz lantzeko aukera dago. Lan hori corpusaren gune orekatua osatzen duten testuetan egiten dugu. Lan hori CORPUSGILEren EL moduluan egiten da, eta hurrengo atalean xeheago azalduko dugu.

5.3.1 EL modulua

EL modulua corpusaren gainean etiketatutako informazio linguistikoa gainbegiratzeko, orrazteko eta desanbiguatzeko ingurunea dugu, eta giza erabiltzaileari zuzenduta dago. Modulu honen osagai nagusia EULIA izeneko tresna bat da, eta berorri esker linguistek zein etiketatzaileek aurreko urratsetan sortutako informazio linguistiko guztia aztertzeko aukera dute, eta, nahi izanez gero, informazioa gehitu, aldatu edo/eta zuzentzekoa (Artola et a. 2004:.

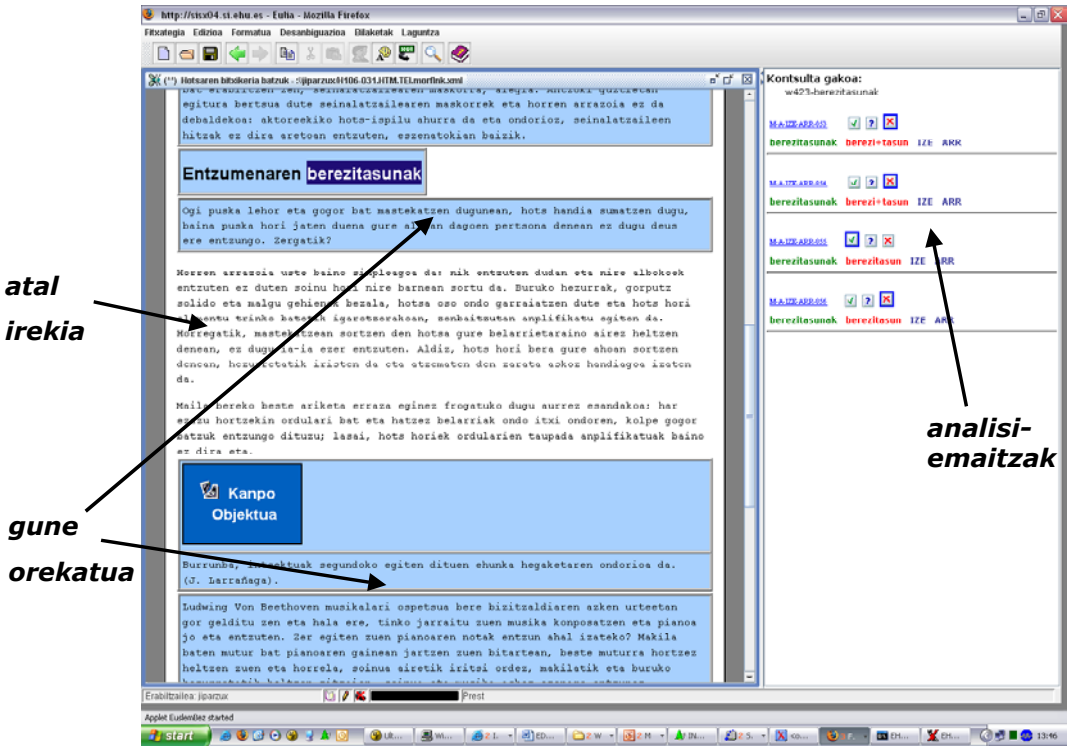
EULIAren helburuak honako hauek dira:

- Analisi-tresnak modu integratu batez koordinatu, eta tresna horien emaitzak kudeatzeko laguntza eskaini.
- Hizkuntzalariari anotazioen adierazpidea eta konplexutasuna ezkutatu, bere lana modu atsegin batean egin ahal izan dezan.

EULIA honako prozesu hauetan erabiltzen da:

- Analisi-prozesu automatikoak, oro har: testu bat hautatu, eta berorren tokenizazioa, segmentazioa, analisi morfosintaktikoa, lematizazioa eta abar abiarazi eta egikaritu.
- Aldez aurretik prozesatutako testu baten analisi-emaitzak ikuskatu, eta berorien gaineko bilaketa oinarrizko nahiz konplexuak egin.
- Eskuzko desanbiguazioa: interpretazio bat baino gehiago daudenean zuzena markatu, akatsak zuzendu, interpretazio zuzenik ez dagoenean sortu, eta abar. Sistemak bermatzen du zuzenketak edo/eta analisi berriak TEI gidalerroen arabera egiten direla, hau da, sortzen diren ezaugarri-egiturak dagokien motaren arabera eratuta sortzen direla.

EULIA proiektuan erabiltzeko egokitu da, eta CORPUSGILEn integratuta dago.



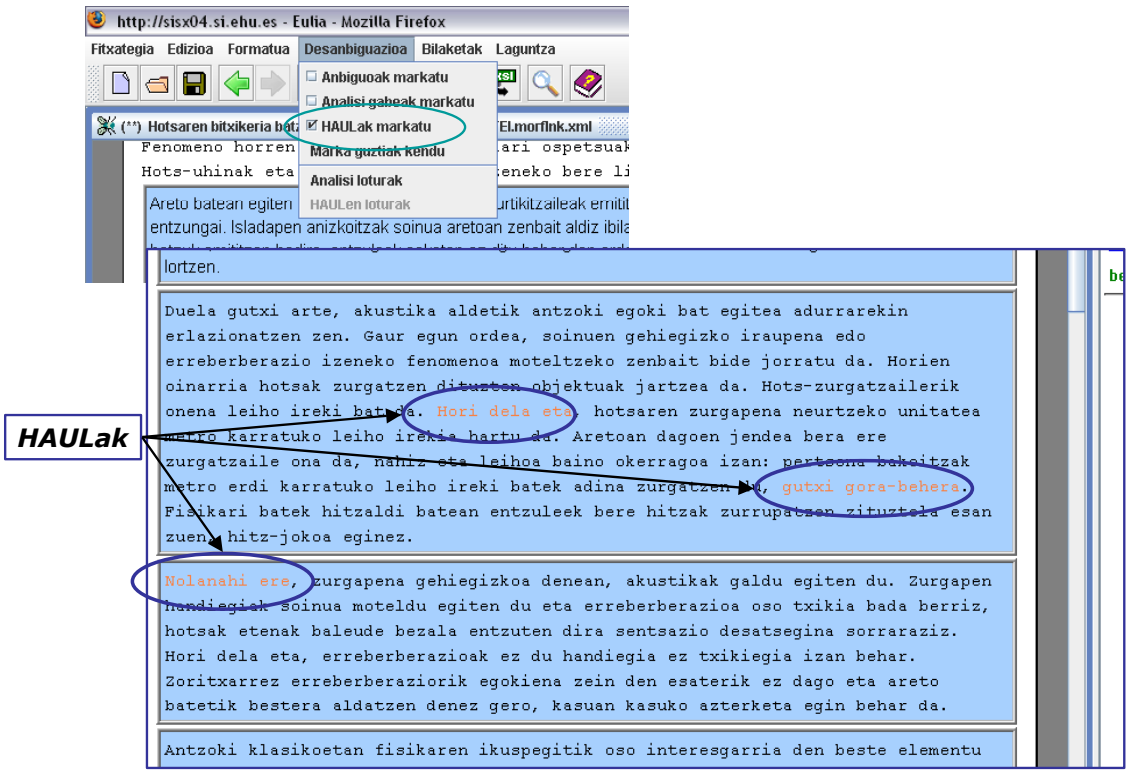
8. irudia. EULIAren lan-interfazea.

Irudian (8. irudia), EULIAren interfaze grafikoa ageri da. Irudiaren ezkerraldean *Testu-leihoa* dugu eta eskuinekoan *Analisi-leihoa*. Beheko aldean ohiko *Egoera-barra* ere ikus daiteke. Ikus ditzagun, bada, bi leiho nagusiak.

5.3.1.1 Testu-leihoa

Testu-leihoan, sarrera-testua (gure kasuan, egitura-etiketatzearen emaitza), tokenizazioaren emaitza eta HAULen fitxategia prozesatzearen ondorioz sortutako testu-egitura bistaratzen da. Irudian ikus daitekeenez, corpusaren gune orekatukoak diren paragrafoak nabarmenduta ageri dira eta atal irekikoaz zuriz. Leiho honetan bi motatako osagaiak nahasten dira:

- *Tokenizatzaileak ezagutu dituen zatiak*: linguistikoki interesgarriak diren testu-zatiak dira, hau da, tokenak. Tokenei 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.



9. irudia. HAULak markaturik, EULIAren interfazeko testu-leihoan.

Itxurari dagokionez, ezin dira osagai horiek bereizi. EULIAren 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.

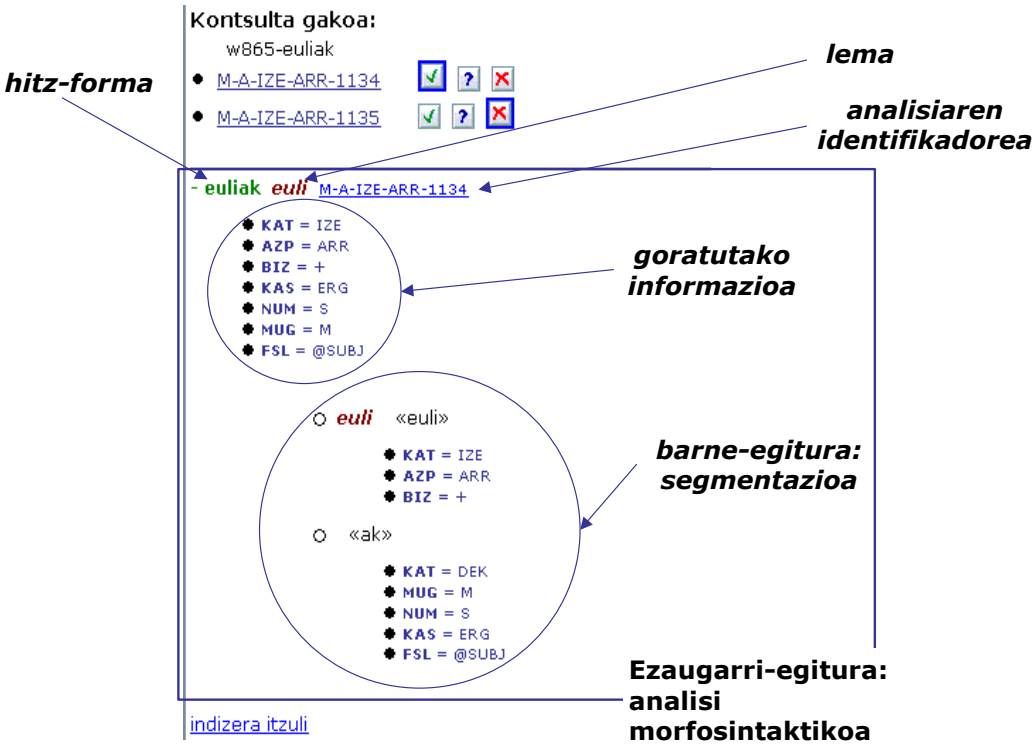
Testu-leihoan token baten gainean klik egiten dugunean, horrekin erlazionatzen diren tokenen araberako ekintzak abiaraz daitezke. Hona hemen suerta daitezkeen kasuak:

- *Klikatutako tokena ez da HAUL baten parte:* kasu honetan, klikatutako tokena tokenizazio-fitxategian azaldutako testu-erreferentzien 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 (erabiltzaileak hala nahi badu, noski: ikus 9. irudia), eta analisi-leihoan tokenaren analisiak eta markatutako HAULenak erakusten dira.

Interfazeak aukera ematen du, bada, testu-leihoko hitzen gainean klik egin eta dagokien informazioa ikuskatzeko. Horretaz gain, markak erabiltzen dira hitzak bereizteko: analisi anbiguoak dituztenak modu berezi batez bistaratzen dira, erabiltzaileak hautatutakoa(k) beste modu batez, eta abar. Marka hauen guztien itxura pertsonaliza daiteke erabiltzaile bakoitzarentzat.

5.3.2 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 xehetasun-maila kontuan izanik. Horri esker, amarauna osatzen duten XML dokumentuak ezkutuan gelditzen dira, eta erabiltzaileak modu gardenean ikus eta erabil dezake informazio linguistikoa. Irudiaren goiko aldean (10. irudia) ikus daitezke *euliak* hitz-formaren bi lematizazio posibleak zerrenda batean, non lehena zuzentzat markaturik ageri den (desanbiguazio automatikoaren ondorioz edo hizkuntzalariak analisi hori eskuz hautatu duelako). Beheko aldean, berriz, lematizazio horren xehetasunak ikus daitezke: informazio goratua eta morfemaz morfemako informazio xehatua.



10. irudia. EULIAren interfazeko analisi-leihoa (xehetasuna).

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 duten ekintzak eskain daitezke, modu adimentsuan.

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

6 Ondorioak

Hizkuntza orok bezala, euskarak ere corpusak behar ditu; hizkuntzalariek, terminologoek, hizkuntza-teknologien ikertzaileek, hizkuntzaren estandarizazioaren ardura dutenek, hainbatek behar ditu corpusak, gaur egun hizkuntza aztertzeko ezinbesteko baliabide direlako. Zientzia eta Teknologiaren corpusaren bidez, baliabide

egoki eta ahaltsu bat eskaini nahi dugu espezialitate-alor horietan erabili den hizkuntza aztertzeko.

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. Bestetik, corpusa eratzeko metodologia zehaztu eta ezarri dugu, corpusgintzan behar diren tresnak eta baliabideak moldatu edo garatu ditugu, eta prozesu osoa bere baitan hartzen duen aplikazio batean, CORPUSGILEn, integratu.

Horiek biak dira, baliabide eta tresna bana, hain beharrean gauden alor honetara egin nahi ditugun ekarriak.

Bibliografia

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.

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.

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]

Bach, C., Saurí, R., Vivaldi, J. & Cabré, M.T. 1997. *El corpus de l'IULA: descripció*. Bartzelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística

Aplicada. [ftp://ftp.iula.upf.es/pub/publicacions/97inf017.pdf; 06-02-10ean irakurria]

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]

Sinclair, J. 1996. *Preliminary Recommendations on Corpus Typology*. EAGLES. [http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html; 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]