III. Prozesadore morfologiko bat euskara estandarrerako.

Morfologiarako eredu konputazionalen barruan, egoera finituko morfologian sakondu ondoren bi mailatako morfologiaren ezaugarriez aritu gara aurreko kapituluan. Eredu horren euskararen gaineko aplikazioa da hirugarren kapitulu honen xede nagusia.

Horretaz aurreko kapituluan zer edo zer seinalaturik geratu bada ere, bi mailatako eredua aukeratzea justifikatzen da lehen pasartean, horrez gain eredua euskararen gainean aplikatzean egin diren hautapenak zehazten direlarik. Ondoren euskararen morfologiaren deskribapen labur bat egiten da bibliografi aipamenak erantsiz, horretan sakontzen laguntza emateko asmoz.

Lexikoa eta erregelak dira aipatutako ereduaren atal nagusiak eta horiexek azaltzen dira euskararen morfologiaren deskribapen zehatza eginez. Halako proiektu aplikatu batean informazioa edozein modutan gorde eta eguneratu ezin denez gero, sortua izan den datu-basearen berri ematen da. Horrekin batera lexikoan erabili den karaktere-multzoa, morfotaktika osatzen duten elementuak, azpilexikoak eta jarraitze-klaseak hain zuzen ere, eta lexiko honen dimentsioa zehazten dira ideia orokorra emanez. Beste aldetik, aldaketa morfofonologikoak nola gauzatzen diren deskribatzen duten erregelak ere aurkezten dira.

Ezagumendu linguistikoaren deskribapena egin eta gero egindako programaren egitura eta zenbait zehaztasun azaltzen dira, eta programa hori erabiliz eraginkortasunari, memoria-hartzeari, estaldurari eta gainsorrerari buruz lortutako neurriak eta ondorioak ere aurki daitezke. Azken urtetan aurreko kapituluan aipaturiko lexiko-itzultzaileen sorrerak eskainiko abantailez baliatzen gara euskararen inplementazioan ere, eta Xerox-ek utzitako tresnen erabilpenak zer-nolako hobekuntzak dakartzan ere azaltzen da.

Azkenik, informazio morfologikoaren trataerak dakarren arazoaz mintzatzen da, hau da, euskara hizkuntza eranskaria den aldetik morfema anitz biltzean azaltzen diren fenomenoak, elipsian eta deklinabide-kasu anitzetan sakonduko dugularik.

Kapitulu honen gida eta erreferentzia gisa taldekidea den M. Urkiaren tesia (1995) da gomendagarria.

III.1 Ereduaren egokitasuna eta jarritako mugak.

I.1 pasartean aipatutako irizpide orokorrei jarraituz, proiektu honen hasieran egin beharreko balizko prozesadore morfologikoari honako diseinu-baldintzak jarri genizkion:

- Estaldura handiko prozesadorea izatea, beraz, euskararen morfologia deskribatzeko gai zen mekanismoa aukeratu behar zen.
- Analisi zein sintesirako balio izatea, oinarrizko tresna izatean bere gainean tresna gehiago eta ahaltsuagoak eraiki ahal izateko.
- Ezagumendu linguistikoa eta programa erabat banandua edukitzea, aldaketak eta eguneratzeak errazteko asmoz.
- Eraginkortasunaren aldetik produktu erabilgarriak bideratzea, proiektu aplikatua zen aldetik.
- Gainsorrera ekiditea. Proiektuaren helburu nagusietako bat sorrera zehatza bideratzea bazen ere, gainsorrera ekiditeak garrantzi are handiagoa hartzen zuen bere lehen aplikaziorako, zuzentzaile ortografikorako hain zuzen ere.
- Alomorfoen erabilpena ekiditea ahal den neurrian, deskribapenaren eta mantenuaren ikuspuntutik sistema aldrebesten baitu.

Bibliografia aztertzean eraginkortasun-arrazoiak zirela eta egoera finituko ereduetan sakontzea deliberatu genuen —aurreko kapituluan nabarmena da eredu horien alde egiten den apustua— eta zenbait aproba eta maketa egin eta gero (Arregi & Urkia, 89) bi mailatako morfologia aukeratu genuen espezifikazio-baldintzak betetzeaz gain —kontutan hartu behar da hasiera batean suomierarako diseinatua izan zela eta euskararen flexio-sistema suomierarenarekin alderatu dela— bi ezaugarri hauek, oso garrantzizkoak bihurtu direnak, gaineratzen zituelako:

- Morfofonologia deskribatzeko eredu dotorea, morfotaktika eta morfofonologiaren arteko bereizte erabatekoa bideratzen duena, eta programa eta ezagumendu linguistikoaren arteko banaketa azken muturreraino ziurtatzen duena.
- Hizkuntza askotarako, ingurukoak barne, eredu izatea honek sistema eleanitzen eraikuntzan eta hizkuntzen arteko elkarlanean bultzada handia ematen diola gure sistemari.

Eredua aukeratu ondoren kapitulu honetan azaltzen den gauzatze konkretura pasatzean, hasieratik aurrean geneuzkan muga hauek kontuan hartu genituen:

- Euskaraz aurretik ez zegoen morfologiari buruzko lan sistematikorik, beraz lan horri ekin behar zitzaion (Urkia, 95).

- Flexio-morfologia nahiko aztertuta egonda eta erregularra izanda sakontasun osoz gauzatzeko aukera zegoen bitartean, eratorpen-morfologian eta elkarketan aukeratu egin behar zen: alde erregularrena bakarrik aztertu edo gainsorreraren arazoan erori. Erabakia lehen aukeraren ildotik joan zen. Horrela, eratorpenean generalizazioa onartzen duten kasuez gain termino lexikalizatuak bakarrik onartzen dira, eta elkarketan berdin, *izen-izen* ereduari jarraitzen dizkiotenak salbu.

- Testu-hitza da prozesadore morfologikoaren tratamendu-unitatea, beraz, hitz anitzeko terminoen trataera lan honetatik kanpo geldituko da oraintxe, helburu horrekin lanean jarraitu arren. Hala ere, hitzaren identifikazioa ez da berehalakoa, horretarako *token*-ezagutzailea edo iragazlea izeneko modulua erabili ohi baita.

- Aditz laguntzailearen zein trinkoaren banaketa morfologikoa ez da burutzen bi arrazoirengatik: alde batetik nahikoa konplexua eta ez-erregularra delako, aldaketa morfofonologiko anitz eta morfotaktikaren aldetiko urruneko menpekotasuna aurkeztuz; eta bestetik horrek eraginkortasunean duen eraginarengatik. Gaur egun, lexiko-itzultzaileen bidetik, ez legoke aditz laguntzailearen deskonposaketa egiteko arazorik eta zabaldutako ikerlerrotzat jotzen dugu.

III.2 Euskararen morfologia labURTUA.

Euskara hizkuntza eranskaria da, hau da, hitzen eraketa funtzio desberdinei, sintaktikoak barne, dagozkien osagaiez burutzen da. Horrela, izenen eta adjektiboen kasuan adibidez, determinazioari, numeroari eta deklinabide-kasuari dagozkien hizkiak hartzen dira ordena horretan eta elkarren artean independente lemaren ondoren. Eratorpena eta elkarketa aski emankor dira eta hitz-eraketan dezente erabiltzen dira.

Kasu askotako deklinabide-sisteman datza flexio-morfologiaren ezaugarri garrantzitsuenetako bat, inguruko hizkuntzetatik bereizten duena. Determinazioari, numeroari eta deklinabide-kasuari dagozkien osagaiak izen-sintagmako azken elementuan baino ez dira agertzen; hizkuntza erromantzeetan ez bezala, berauetan elementu guztietan itsasten baitira. Azken elementu hori izena izateaz gain adjektiboa edo determinatzailea ere izan daiteke. Adibidez “etxe zaharrean” izen-sintagman honako osagaiak aurki daitezke:

etxe: izena
zahar: adjektiboa
r eta *e*: epentesiaren ondorioak
a: singularreko determinatzailea
n: inesiboa

Latinaren bost deklinabide-paradigma ezagunetatik urrun, euskarak deklinabide-paradigma bakarra du, deklinabide-taula bakar bat baitago sarrera deklinagarri guztientzat.

Beste hizkuntzetako preposizioen funtzioa euskaraz atzizkien bidez burutzen denez gero, forma flexionatuak sortzeko ahalmena izugarria da. Adibidez, izen-sarrera batetik abiatuz 135 forma flexionatu lor daitezke gutxienez. Horietako 77 determinazioa, numeroa eta deklinabide-kasu konbinatuz lortutako forma ez-emankorrak diren bitartean, gainontzeko beste 58ak forma emankorrak dira bi genitiboetako batez bukatutako forma sinple edo deklinatuak baitira. Genitiboen atzetik teorikoki hasierako emankortasun-ahalmen guztia dago, genitiboaren atzetik flexiorik agertzean elipsi bat dago eta. Elipsi bat baino gehiago posible izanik, atzizki-hartzea errekurtsiboa izan liteke, maila teorikoan behintzat, eta ondorioz, emankortasun-ahalmena infinitua litzateke. Izan ere, elipsi bat baino gehiago agertzea ohizkoa ez bada ere oso arraro ez diren forma batzuek bi elipsi edo gehiago dute. Aurrekoaren ondorioz eta bi elipsi kontuan hartuz izen bati dagozkion forma flexionatuak honako hauek lirateke: 77 + 58 (77 + 58 (77 + 58)) = 458683 (Agirre *et al.*, 92). Izen bakoitzeko horiek baino gehiago ezagutzeko eta sortzeko gai izan behar du euskararako prozesadore morfologiko batek.

Azter ditzagun *seme* izenaren forma flexionatu batzuk

semea: seme+a (nominatibo mugatu singularra)
semeari: seme+ari (datibo mugatu singularra)
semearen: seme+aren (genitibo mugatu singularra)
semearena: seme+aren+asemearen (etxe[1])a
(genit. mugatu sing. + nomin mugatu sing.)
semearenera: seme+aren+(e)ra semearen (etxe)ra
(genit. mugatu sing. + alatibo mugatu sing.)
semearenekoak: seme+aren+(e)ko+ak semearen (etxe)ko (arazo)ak
(gen. mug. sing.+gen. mug. sing.+nom. mug. plur.)

Aipatutako emankortasuna antzekoa da elementu deklinagarri gehienetan baina adjektiboaren kasuan are handiagoa da, gradu-flexioa dela eta lau aldiz handiagoa baita.

[1] Elipsia bat dago, bera, aurreko elementu bati (etxea edo beste edozein) egiten zaio erreferentzia.

Konbinazio-aukera hauek errealitatean gertatzen direla egiazta daiteke corpusetan oinarriturik; eta horrela bi genitiboko hitzen bat agertzea oso-oso arraroa dela ondorioztatzen den bitartean, sei morfemaren metaketa arrunt samarra dela ikus daiteke: *amorratuenetakoak* (amorra+tu+en+eta+ko+ak), *argienetarikoa* (argi+en+eta+rik+ko+a), *egitekoetarako* (egin+te+ko+eta+ra+ko), etab.

Beste aldetik generoaren araberako flexioa ez dago euskarazko deklinabide-sisteman; beraz, maskulino eta femeninoa bereizteko hizkirik ez dago. Izan ere, aditz jokatuetan generoaren marka ager daiteke batzuetan, adibidez forma alokutiboetan solaskidearekiko konfidantzaren arabera.

Aditz-forma jokatuak aditz laguntzaileek eta aditz trinkoek osatzen dituzte. Aditz-flexioa aberatsa da euskaraz, askotan pertsona anitzeko hizkiak agertzen direla bakoitza ergatiboari, nominatiboari eta datiboari egoki dakiekeena. Hala ere flexioa aditz zahar erabilienetan baino ez dela erabiltzen hartu behar da kontutan.

III.3 Lexikoa.

Egin dugun bi mailatako morfologiaren egokitzapena aztertu baino lehen eta zenbait iturri aipatzeko probetxatuz, aipa dezagun lehen sistema osatzeko irizpideak.

Euskararen flexioa burutzeko Euskaltzaindiak (1985) proposatutako taulan oinarritu gara eta gure sistemara egokitu dugu; hau da, taula hori hartu eta lexiko-kategoria bakoitzari egokitzen zaizkion kasuak multzoka eratu ditugu.

Eratorpenean generaliza daitezkeen zenbait aurrizki eta atzizki landuta daude, baina gainontzeko hitz eratorriak hiztegi-sarrera bezala daude. Honetaz sakontzeko interesgarria da Adurizek eta Aldeazabalek egindako txostena (1995). Hitz-elkarketan ere, ohizkoena eta sistematizagarriena landu da momentuz, Euskaltzaindiaren LEF Batzordeak markatutako irizpideen arabera (Euskaltzaindia, 92).

Aditzari dagokionez, aditz laguntzailearen zein trinkoaren formak oso-osorik sartu dira, beti ere Euskaltzaindiak (1973, 1985) erabakiak. Forma neutroak, markatu gabeak nahiz hitanozkoak ezagutzen dira. Aditz faktitiboa ere sistematikoki landuta dago (1992ko Euskaltzaindiaren gomendioa eta 94ko erabakia).

Gramatikaren atalean Euskaltzaindia izan bada arau-iturri bakarra, bestela gertatzen da lexikoa lantzen hasi orduko. Puntu batzuetan emanak ditu kasuan kasuko gomendio eta erabakiak: *H* letra, *-a* berezkoa, zenbakien osaera eta idazkera, etab. Horiek jarraitu ditugu lexikoa osatzean, nahiz eta zenbakien kasuan oraingoz bi aukerak mantentzen ditugun

(*hogeita bost* eta *hogeitabost* onartuz). Beste hainbeste gertatu da pertsona- eta leku-izenekin, bai eta maileguen idazkeran ere.

Oinarrizko lexikoa osatzeko, hau da, edozein lexikotan maizenik agertzen diren lemen zerrenda, gaurko beste iturrietara jo behar izan dugu: Ibon Sarasolaren *Hauta-Lanerako Euskal Hiztegia*, UZEIko Euskalterm datu-bankua eta EEBS datu-base lexikografikoa, Xabier Kintana eta besteren *Hiztegia 2000* (1984), J.M. Etxebarriaren *Maiztasun- eta Prestasun-Hiztegia* (1987), etab. Euskaltzaindiaren irizpideekin bat ez zetozenean, sarrerak "egokitu" egin dira; eta, Euskaltzaindiak erabaki ez dituenetan, Ibon Sarasolaren hiztegia izan da irizpide-iturri.

Oinarrizko hiztegia osatu nahirik, UZEIko EEBStik hainbat esapide, lokuzio eta forma konplexu hartu da. Siglak eta laburtzapenak ere UZEIren (1988) irizpideen arabera landu dira. Hiztegi arruntetik abiatuz, terminologiaraino iritsi behar izan da zenbaitetan. Ezinbestekoa izan da Euskalterm (Urkia & Sagarna, 91) horrelakoetan.

Izen propioen zerrenda osatzeko (izen propioak hiztegi arruntetan ez badatoz ere), bi iturritara jo da: lehena Euskaltzaindiak proposatutako euskal pertsona- eta leku-izenen zerrenda (1979, 1983) izan da, baina munduko leku-izenen zerrendatua eskuratzeko Elhuyar-era (1990) jo da.

Iturri guzti hauetatik edanda, ondoan ikusiko den bezala, tamaina handiko lexikoa osatu dugu.

III.3.1 EDBL: Euskararako datu-base lexikala.

Eskala errealeko proiektu aplikatu bati ekitean datu linguistikoen egituraketa eta mantenua planifikatu egin behar aurretik. Prozesadore morfologikoaren muina den lexikoa datu-base batean antolatzea da hausnarketa horren ondorio berehalakoa. Nahiz eta bi mailatako formalismoaren bidez morfologia egiteko sortu, EDBLk (Agirre *et al*., 94) euskararen tratamendu automatikorako datu-base lexiko orokor bat izan nahi du eta horrexegatik morfologian erabiltzen ez diren informazioak ere metatzen dira bertan.

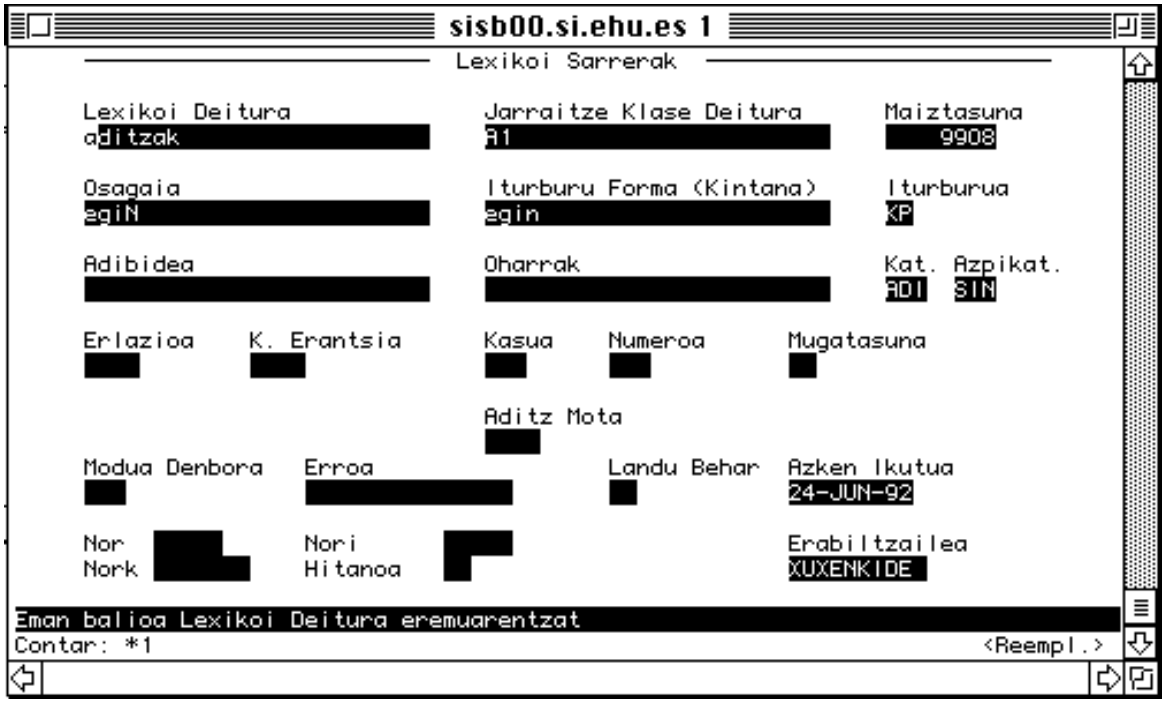
Hasiera batean VAXeko RDB softwarea erabili izan bazen ere, ondoren SUNeko ORACLEra eramana izan zen, gaur egun ORACLEren bidez kudeatzen delarik. Beraz, eredu erlazionalari jarraitzen dio, baina etorkizunerako, objektuei zuzendutako diseinu berri baten gainean ari gara lanean, hartzen ari den konplexutasunari ondo erantzun ahal izateko (Agirre *et al*., 94) (Agirre *et al*., 95).

Eredu eralazionaleko diseinu berri horri jarraitzeko lan-lerroa izanik, gaur egun darabilgun datu-basea azalduko da ondoren. Bertan nagusiki bederatzi taula daude definiturik:

1) **Karaktere arruntak**: lexiko-mailan onartzen diren karaktereak.

2) **Markak**: morfofonemak eta diakritikoak diren karaktereak.

3) **Azpilexikoak**: lexikoa osatzen duten azpilexiko guztiak zerrendatzeko eta bakoitzari bere ezaugarriak —hasierakoa izatea, lemaren parte izatea, irekitasuna, orokortasuna eta estandartasuna— definitzeko.



III.1 irudia.- EDBLren taula nagusia eguneratzeko pantaila.

4) **Morfemak**: euskararen morfema guztiak metatzeko taula. Taula hau funtsezkoa denez, bere eremuak zerrendatuko ditugu. III.1 irudian ikus daiteke taula honi dagokion eguneratze-pantaila.
   - lexikoi-deitura: azpilexikoa, zeine barruan dagoen morfema.
   - lexiko-mailako morfema, karaktere arruntez, morfofonemez eta hautapen-markez osaturik
   - dagokion jarraitze-klasea, ondoan itsats dakizkiekeen morfemak ondo definitzeko
   - erabilpena azaltzen duen adibide bat
   - kategoria eta azpikategoria sintaktikoak
   - kasua, numeroa eta mugatasuna, atzizkietan erabilia
   - erlazioa

- erroa, modua/denbora, nor, nori eta nork pertsonak eta hitanoaren marka aditz jokatuetarako
- kategoria erantsia, zenbait eratorpen-atzizkitan agertuko dena
- aditz-mota aditz-erroetarako
- morfemaren iturburua, hau da,nondik hartu izan den
- oharrak
- Kintana hiztegiaren forma
- agerpen-maiztasuna (Sarasolaren arabera)
- eguneratze-data
- zalantzazkoak
- erabiltzailea

Ikus daitekeenez eremu batzuk lemei soilik egokitzen zaizkie, beste batzuk aditz trinko eta laguntzaileei eta beste batzuk hizkiei. Horrexegatik diseinu berrian hiru azpitaulatan banatzea aurrikusi dugu.

5) **Jarraitze-klaseak**: morfemei dagozkien jarraitze-klaseak definitzen dira taula honetan. Bere identifikadoreaz gain zehazten diren osagaiak azpilexikoak edota definituriko jarraitze-klaseak dira.

6) **Jarraitze-klase hedatuak**: morfotaktikari dagokion urruneko menpekotasunak hala eskatzen duenean. Zehazten diren osagaiak jarraitze-klase arruntak dira, zuhaitz eran edo debekuen bidez konbinaturik.

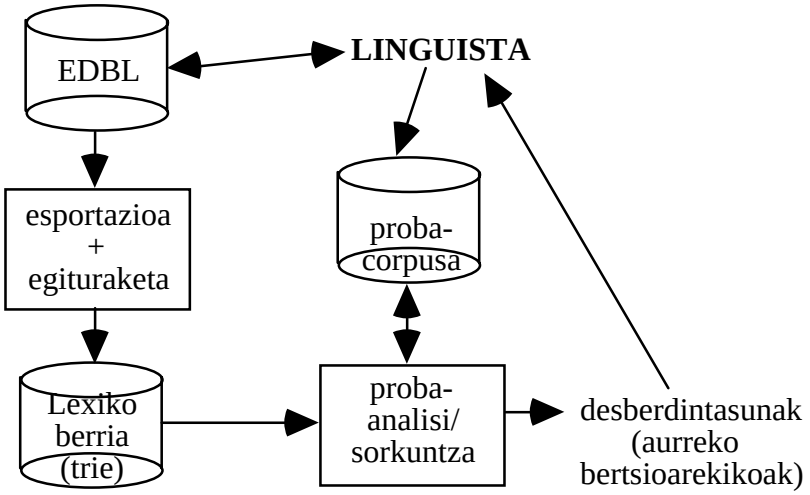
7) **Aldaeren azpilexikoak**: aldaeren trataerarako erabiltzen dira taula hau eta ondoko biak, eta hurrengo kapituluan azalduko dira.

8) **Aldaera morfemak**.

9) **Aldaeren jarraitze-klaseak**.

Informazioa linguistek datu-basean eguneratu eta mantentzen duten arren, prozesadore morfologikoak lexikoa *trie* egituraz behar duenez gero, esportazioa burutzen da bertsio berri bakoitzerako. Esportazioarekin batera proba-corpus batzuk prestaturik daude bertsio berria eta zaharraren arteko desberdintasunak lortu ahal izateko (ikus III.2 irudia). Desberdintasunak aztertuz linguistek errorerik detektatzen badute, datu-base zuzenduko dute prozesua berrabiatuz.

Datu-basearen aberasketa erabat eskuz egin zen hasiera batean, baina hiztegiak eta hitz-zerrendak lortuz joan garen heinean modu semiautomatiko bat ezarri da.



III.2 irudia.- EDBLren mantenua, esportazioa eta proba.

III.3.2 Lexikoko alfabetoa: morfofonemak eta hautapen-markak.

Lexikoan zehazten diren lexema eta hizkiak osatzeko ohizko diren karaktereez aparte morfofonemak eta hautapen-markak ere erabiltzen dira, erregela morfofonologikoetan eragin berezia lortzeko asmoz. Muga oso argia ez bada ere, morfofonema kasuaren arabera gauzatzen den karaktere bat den bitartean, hautapen-markak inoiz ez dira karaktere bihurtzen azalean, dagokien funtzioa zenbait erregelaren aplikazioa kontrolatzea baino ez baita.

Karaktere arruntei buruz bi ohar besterik ez:

- euskararen kasuan alfabeto arruntekoak ñ-a barne, elkarketarako marratxoa eta laburduretarako puntua dira karaktere hauek, beste karaktererik ez baitago onarturik euskara estandarrean.
- baliabideen ekonomia dela eta, trie egituran eta erregeletan minuskulak eta maiuskulak kontutan hartu beharrean, lehenak bakarrik erabiltzen dira. Letra maiuskula bat behartu behar denean, leku-izenetan adib., letra maiuskularen ordez izarra (*) eta dagokien minuskula adierazten da.

Morfofonemen kasuan ikus ditzagun zeintzuk izan diren euskararen deskribapen morfologikorako erabili ditugunak:

R esanahi bikoitza du: batetik *r* gogorra lemaren bukaeran eta bestetik *r* epentetikoa zenbait atzizkiren hasieran. Adib. *zakuR* eta *Rik* (partitibo mugagabea). Beste aukerak baziren, bi karaktere desberdin aukeratzea edo lehen kasuan *zakurR* adieraztea. Aukera horren aurrean alfabetoa eta lexikoa

minimizatzeko irizpideari jarraitu zaio.Adib. *zakuR+a:zakurra* eta *kale+Rik:kalerik.*

Q *e* epentetikorik hartzen ez duen bukaerako *r*-a. Adib. *haQ+Ek:hark.*

~ hiru eta lau zenbakiek gordetzen duten *r* zaharra. Adib. *hiru~+ak:hiruak* eta *hiru~+ak:hirurak*

E *e* epentetikoa. Deklinabide-atzizki askoren hasieran agertzen da. Adib. *zuhaitz+Eko:zuhaitzeko.*

N Bukaerako *n*-a galtzen duten aditz-erroetan jarria. Adib. *egiN+ten:egiten.*

M Bukaerako *n*-a galtzen duten atzizkiak. Adib. *lagun+areM+kiM+ko:lagunarekiko.*

**** Bukaerako *n*-aren galera aukeran duten atzizkiak. Adib. *e* (genitibo plurala); *norbait+e\+gatik:norbaitengatik* eta *norbait+e\+gatik:norbaitegatik*

A *a* organiko arrunta. Atzizkiekin lotzean batzuetan galtzen dena. Adib. *amA+a:ama.*

hitz-elkarketan gal daitekeen salbuespeneko *a* organiko. Adib. *kultur#_ekintzA: kultur_ekintza.*

@ aditz defektiboetan *e* bihur daitekeen bukaerako *a*. Adib. *ater@+a:aterea.*

& Leku-izenetan batzuetan galtzen den bukaerako *a* artikulua. Adib. **azpeiti&+Eko: Azpeitiko.*

^ hikako formak eta bukaeraren ondoan *a* har dezaketen batzuk. Adib. *dun^+En: dunan.*

Azkenik, euskararen deskribapen morfologikorako erabili ditugun **hautapen-markak** hauexek dira:

% *l, m, n, s, x, z,* eta *R*-z bukatutako leku-izenen marka, zenbait bihurketatan eragina duena. Adib. **usurbil%+Eko:Usurbilgo* eta **usurbil%+Eko:Usurbileko.*

: deklinabidea bokal bezala egiten duen sigla. Adib. **h*b:+Ek:HBk.*

/ deklinabidea kontsonante zein bokal bezala egiten duen sigla. Adib. **m*i*t/+Eko: MITeko* eta **m*i*t/+Eko: MITko.*

**$** Epentesia kontsonantez bukatutakoak bezala egiten duenbokalez bukatutako aditz jokatua. Adib. *du$+Ela:duela.*

! *garren* morfema markatzeko erabilia. Erregelak sinplifikatzearren zehazten da. Adib. *bi+garren!+Eko:bigarreneko* eta *bi+garren!+Eko:bigarrengo*.

+ morfemen arteko lotura adierazteko. Aurrizkien bukaeran eta atzizkien hasieran jarri ohi da baina gure inplementazioan programak jartzen du automatikoki.

III.3.3 Morfotaktika.

Azpilexikoen eta jarraitze-klaseen bidez definitzen da morfotaktika ohizko bi mailatako morfologiaren ereduan. Aurreko kapituluan esan dugun legez, eredu horri jarraitu diogu aldaketa batekin: morfemen arteko urruneko menpekotasuna deskribatzeko gai diren jarraitze-klase hedatuen erabilera.

Emankortasuna morfotaktikaren funtzioan dagoenez kapitulu honetako bigarren pasartean deskribatu den genitiboaren errekurtsibitatea jarraitze-klase eta lexikoen konbinaketaz ere gauzatuko da; genitibo bakoitzaren jarraitze-klasearen barruan berari dagokion azpilexikoa ere agertuko da, definitzen den grafoaren barruan bide zirkularrak hedatuz. Hau dela eta, elipsiari muga bat jarri gabe ezinezkoa izango da zehaztea zenbat forma ezagut edo sor dezakeen prozesadore morfologiko honek —erantzuna infinitua bailitzateke—; errekurtsibitate-fenomeno hau ez duten hizkuntzetarako aipatu ohi da zein den muga hau.

III.3.3.1 Azpilexikoak.

Esan bezala lexikoa azpilexikoetan banatzen da morfotaktikaren arrazoiak direla eta. Bi morfema azpilexiko berean egon daitezke baldin eta morfotaktika-ezaugarri berberak badituzte aurreko morfemekin, hau da, morfema berberei itsats dakizkiekeenean. Hala ere, eta argitasunari lehentasuna emanez, eraginkortasunarengatik azpilexiko berean egon zitezkeen morfemak banandu egin dira kategoriaren arabera.

Azpilexikoei bost ezaugarri egokitzen zaizkie: hasierakoa izatea, lemaren parte izatea, irekitasuna, orokortasuna eta estandartasuna. Hasierako azpilexikoetan dauden morfemak baino ezin dira jorratu analisiari zein sintesiari ekiteko. Lemaren parte diren morfemak izango dira analisiaren emaitzen artean agertuko den lema osatuko dutenak.

Gainontzeko hiru ezaugarrien baliagarritasuna hurrengo kapituluan azalduko da zehatz-mehatz, baina, modu laburrean bada ere, berauen sarrera egingo dugu. Irekiak direnak elementu gehiagoz osa daitezke —erabiltzailearen lexikoak erabiltzean— eta edozein karaktere-katez ordezkatuak izateko gai dira —lexikorik gabeko lematizazioa egitean. Orokortasunak azpilexiko irekietako morfemekin konbinatzeko gaitasuna adierazten du. Estandartasunaren ezaugarria ez dutenak ez dira kontutan hartzen

prozedura estandarrean baina bai aldaerak kontutan hartzen direnean. Ezaugarri hauen baliagarritasunaz sakontzeko hurrengo kapitulua kontsulta daiteke bertan hitz tekniko zein ez-estandarren prozesuaz aritzen baita.

Guztira 154 azpilexiko bereizi dira eta kopuru handi honen arrazoia bikoitza da: morfotaktika konplexu samarra izatea batetik eta alomorfoak ebitatzearren hizkiei dagozkien azpilexikoetan dispertsio handia gertatzea bestetik. III.3 irudian azpilexiko garrantzitsuenak eta dagozkien neurriak azaltzen dira.

40.000 baino sarrera gehiago daude kategoria nagusietan III.3 irudian ikus daitekeenez, baina hizkiak eta forma ez-estandarrak kontatzen baditugu 60.000tik gertu gaude. Etengabeko aberasketaren bidez laster 70.000 sarreratara iristea espero dugu.

Atzizkiak oso sakabanatuta daude Koskenniemiren filosofia jarraitu baitugu: alomorfoak erabili beharrean azpilexiko txiki anitz definitzea, hau eraginkortasunaren aldetik desegokia izan arren. Dena den, abiadura handitzearren eta prozedura automatiko batez, atzizki erabilien kasuan alomorfoak dituzten azpilexiko handixeago bananduetara jo dugu, eraginkortasunari buruzko hausnarketaren barruan azalduko den bidetik (ikus III.5.2 pasartea).

AZPILEXIKOA	NEURRIA
adberbioak	1.714
aditz laguntzailea eta trinkoa	7.387
aditz-erroak	4.324
adjektiboak	6.250
izenak	23.078
izenlagunak	308
gainontzeko lemak	1.957
siglak	314

III.3 irudia.- Azpilexiko garrantzitsuenak eta dagokien neurria.

Aldaketa morfofonologiko gutxi batzuk morfotaktikaren bidez konpondu dira eta halako kasuetan bakarrik erabili izan dira alomorfoak.

III.3.3.2 Jarraitze-klaseak.

Morfotaktikaren urruneko menpekotasuna ebazteko jarraitze-klase hedatuak erabiltzea proposatu badugu ere, menpekotasun hau oso fenomeno arraroa da euskaraz, eta ondoko hiru kasuetan bakarrik aurkitu dugu, beti menpekotasun murriztatzailea delarik (ikus §II.3.5):

- Aditz laguntzailearen eta trinkoaren eraketan, pertsonari dagozkion hizkien artean aurreko kapituluan azaldutakoaren ildotik. Dena den arazo hau saihestu dugu zeren, lehen esan den bezala, aditz hauek oso-osorik gorde dira eta ez morfemetan banaturik.

- Aditz jokatuarekin konbinatzen diren atzizki eta aurrizkien artean. Baldintzazko *ba* eta indarrezko *bait* aurrizkiak atzizki batzuekin ezin dira konbinatu. Horrela, ezinezkoa da *bait+dut+Ela*. BABALD eta BAIT ohizko jarraitze-klaseak murrizten dituzten @BABALD eta @BAIT jarraitze-klase hedatuak definitzea izan da hartu dugun irtenbidea.

- Marratxoaren ondokoa. Izen-izen elkarketa dela eta, izenek eta aditz nominalizatuek (*te* edo *tze* morfemaren bidez) marratxoa har dezakete atzean, marratxoaren atzean beste izen edo aditz nominalizatua egon daitekeelarik. Hala ere, hau behin bakarrik gerta daiteke, beraz, marratxoaren jarraitze-klasean izenetarako eta aditzetarako eman behar da aukera baina bi murriztapenekin: ondoren ezin da beste marratxorik agertu batetik, eta bestetik, aditzaren kasuan nominalizazioa beharrezko da. Beraz, ezin dira *kale+-+gizon+-+otso* edo *kale+-+egiN*, baina bai aldiz, *kale+-+gizon* edo *kale+-+egiN+te*. Arazo honi aurre egiteko erabili da @ELK jarraitze-klase hedatua.

Aurrekoa kontuan hartuz honako jarraitze-klase ez-konbentzional hauek gelditzen dira finkaturik:

@BABALD	(BABALD (I0))[1]
@BAIT	(BAIT (I0))
@ELK	(IZENAK (I1), ADITZAK (TETZE (I1)), I0)[2]

Gainontzeko jarraitze-klaseak konbentzionalak dira eta morfema bakoitzaren ondoren ondo-ondoko morfema-multzoa zehaztera murrizten dira. Ehun eta hogeita hamar

[1] @ sinboloa konbentzioz esleitzen zaie jarraitze-klase ez-konbentzionalei. I0-k jarraitze-klase hutsa adierazten du eta eraketa hor bukatzen dela adierazten du.

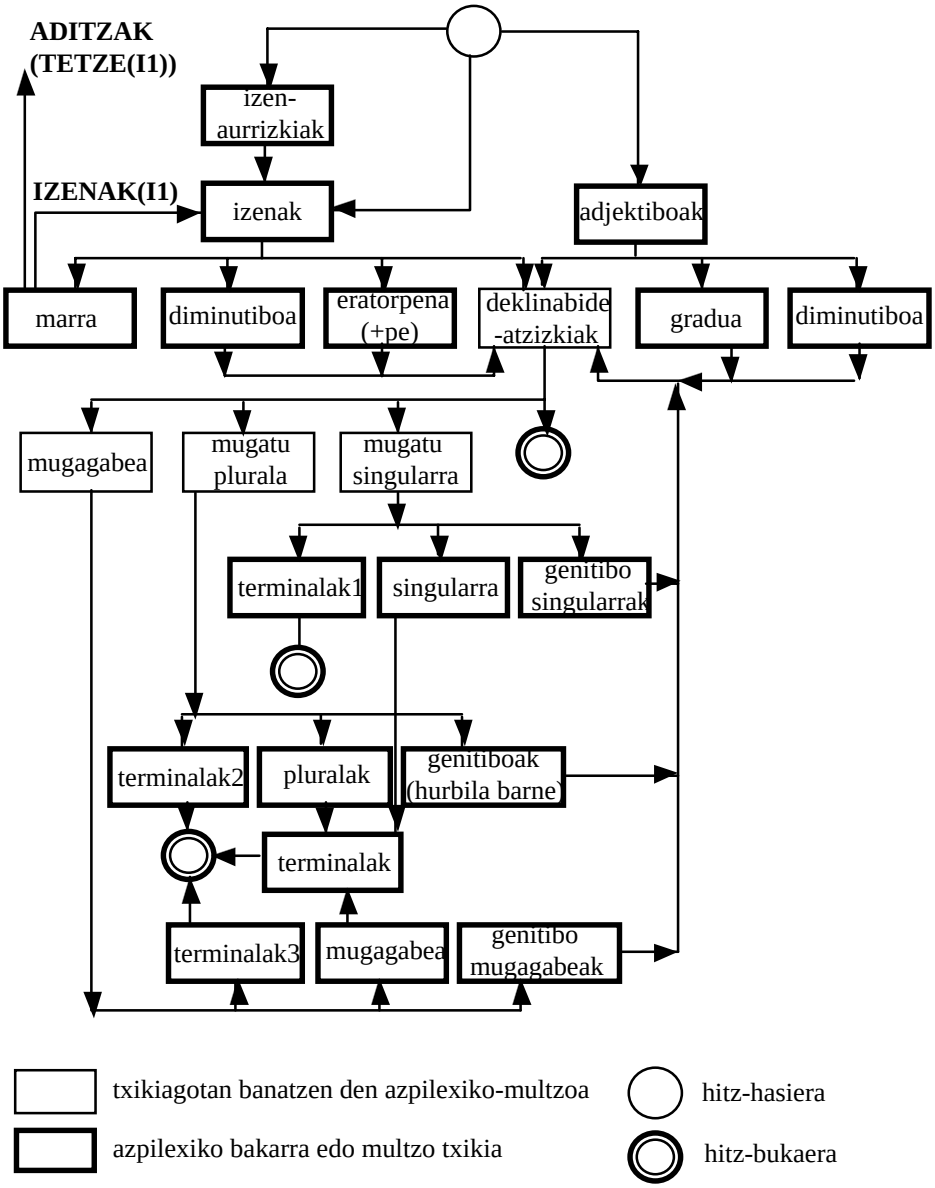
[2] I1-en barruan deklinabide-atzizkiak daude baina ez gidoia. Hitz bat gidoiaz buka daitekeelako agertzen da I0 izenekin eta aditzekin batera lehen maila.

jarraitze-klase desberdin bereizten dira, eta kopuru altua azpilexikoetan gertatzen den aipaturiko sakabanatzeak justifikatzen du.

Ondoren lema emankor ohizkoenen jarraitze-klaseak aztertzen dira; beti ere kasu erregularrei dagozkien jarraitze-klaseak aztertzen direla kontutan hartuz.

III.3.3.3 Izenaren eta adjektiboaren morfotaktika.

Euskarazko izenek eta adjektiboek flexio-morfologia bera dute salbuespen batekin: graduatzailea. Gainera, gure proiektuaren barruan eratorpena bere parte erregularrenean eta elkarketa izen-izen kasuan bakarrik landu denez, izenen eta adjektiboen morfotaktika hurbil samarra da III.4 irudian ikus daitekeenez.



III.4 irudia.- Izenaren eta adjektiboaren morfotaktikaren eskema.

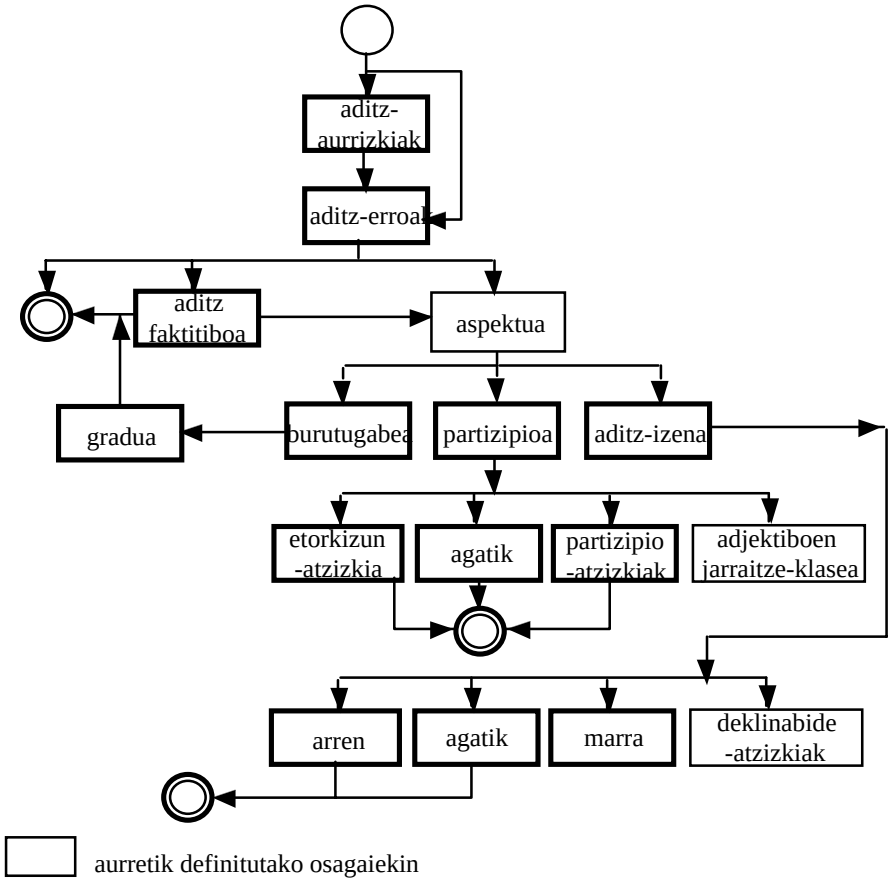
Izen zein adjektiboen atzean atzizki-multzo emankorrena onartzen da: deklinabideari dagokiona. Aurrizkiei, eratorpenari eta elkarketari dagokienean, berriz, desberdinak dira izenak emankorragoak izanik. Adjektiboak, graduatzaileak direla medio, deklinabidearen aukerak biderkatzen dituzte, graduatzaileen ondoren deklinabide osoa etor baitaiteke.

III.4 irudiari jarraituz, ikus daiteke nola banandu diren deklinabide-atzizkiak; batetik *terminalak* izenekoak numeroa/mugatasunarekin (*mugagabea*, *singularra* eta *pluralak*) independenteak direnak, eta bestetik kasuarekin batera numeroa/determinazioa adierazten duten *terminalak1*, *terminalak2* eta *terminalak3*.

Azpimarratzekoa da sinplifikazio txiki bat egin dugula banatuta zeuden zenbait azpilexiko bilduz eskema oso barreiaturik gera ez zedin. Eskeman jarraitze-klase hedatu

bat ikus daiteke (marraren ondoan agertzen dena[1]), morfema-multzo batzuk toki desberdinetatik zintzilik (*deklinabide-atzizkiak* izenetatik eta adjektiboetatik, *terminalak* zenbait atzizki mugagabetatik, mugatu singularretatik eta mugatu pluraletatik[2]) eta bide errekurtsibo edo zirkularra genitiboen bidez.

III.3.3.4 Aditz-erroaren morfotaktika.



III.5 irudia.- Aditz-erro erregularrenen morfotaktikaren eskema.

Aditz-erroen morfotaktika bi bide nagusitan bil daiteke, aditzarena batetik eta izenarena edo adjektiboarena bestetik, zeren *aditz-izen*ari dagokien hizkien bidez nominalizazioa gertatzen baita eta *partizipioa* adjektibo gisa flexionatu baitaiteke.

Aditz-erroetarako jarraitze-klase asko dago zeren aditz motaren arabera morfotaktika desberdina dagokio. Gainera erregelen bidez konpon zitezkeen aldaketa batzuk, *te*/*tze*

---

1 I1 jarraitze-klaseak deklinabide-atzizki guztiak biltzen ditu.

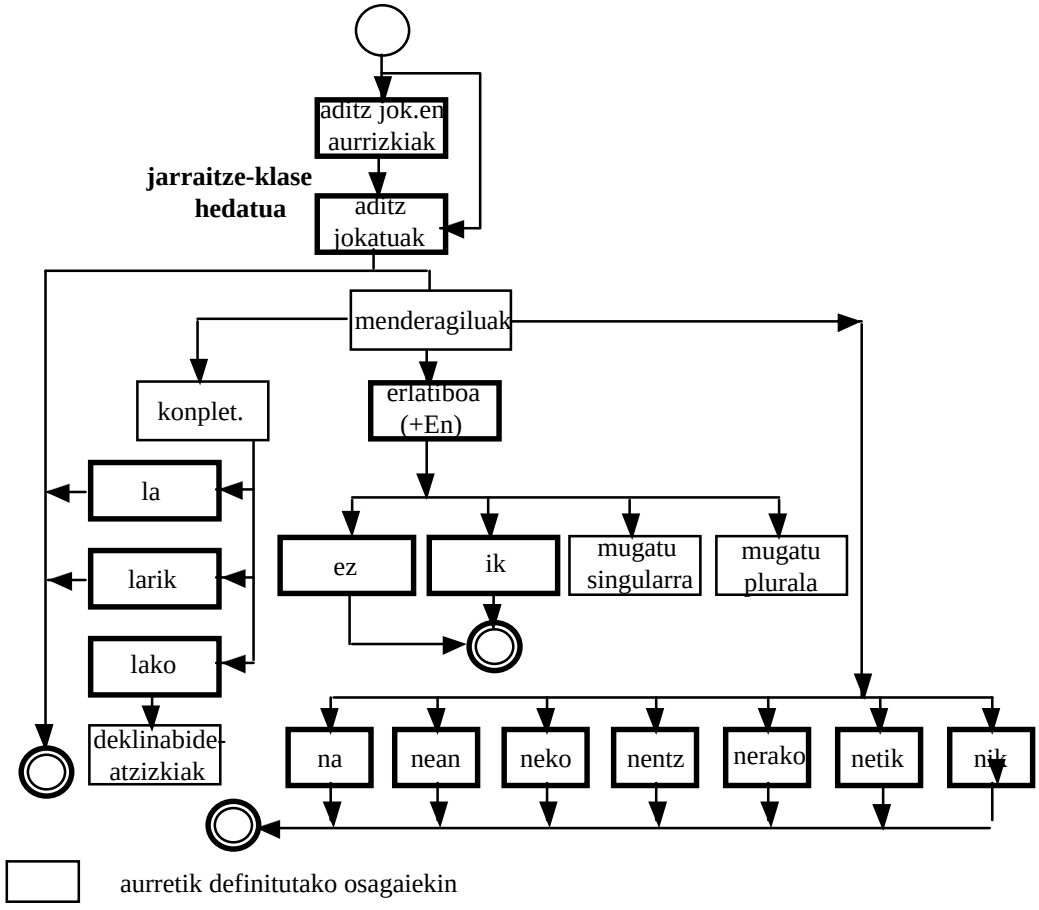
2 *terminalak* deitu dugun azpilexikoen multzoa deklinabide-atzizkiak numero-determinazio hizkiekin (0-ta-eta-ota) konbinatzen dira. *terminalak1*, *terminalak2* eta *terminalak3* izenekin deitutakoetan numero-determinazio eta kasua atzizki bakarrean daude.

ten/*tzen* adibidez, morfotaktikaren bidez konpondu dira Koskenniemik proposatutakoari jarraituz. III.5 irudian infinitiboa *tu* egiten duten aditzen morfotaktikari dagokion eskema ikus daiteke.

Irudian ikus daitekeenez *araz*, *tu* eta *tze* hizkien bidez oso emankorrak izan daitezke aditz-erroak nominalizazio eta adjektibaziora eramaten dute eta. Jarraipena definitu gabe duten osagaiak markaturik daude eta III.4 irudian definituriko jarraitze-klaseei dagozkie.

III.3.3.5 Aditz jokatuaren morfotaktika.

Aditz jokatua ere, laguntzailea zein trinkoa, oso emankorra izan daiteke, batez ere erlatiboaren atzizkiari esker. III.6 irudian eskema nagusia azter daiteke.



III.6 irudia.- Aditz jokatuaren morfotaktikaren eskema.

Aurrizkien artean aipatutako *bait* eta *ba* daudenez hauei dagozkien jarraitze-klaseak zehaztu den jarraitze-klase hedatua izango da.

III.4 Erregelak.

Aurreko kapituluan aipatu den bezala aldaketa morfofonologikoak itzultzaile bihurtzen diren bi mailatako erregelen bidez adierazten dira. Euskaran gertatzen diren aldaketak nahiko sinpleak dira, morfemen arteko loturen inguruan ematen dira eta ez dago hizkuntza batzuen bokal-armonia bezalako urruneko ondoriorik.

A eranskinean banan-banan azaltzen badira ere, ondoren euskararen aldaketa morfofonologikoak gobernatzen dituzten erregela batzuk azalduko dira. Erregelak idazteko erabiltzen den sintaxia *lexc* (Karttunen 93) programan erabilitakoa da bi arrazoirengatik: ondo definitutako sintaxia delako batetik, eta erregela-itzultzaile konpiladore honen bidez espezifikazioa eta exekuzioaren arteko bateragarritasuna lortzen delako bestetik[1]. Sintaxia espresio erregularretan dago oinarriturik, beraz, espresio erregularretan erabiltzen diren eragileak karaktere, morfofonema edo diakritiko gisa erabiltzeko ihes-karaktere bat ipini behar zaie aurretik —*%* karakterea hain zuzen ere[2], ikus aurreko kapituluaren II.3.2.3 pasartea—. Beste aldetik *!* sinboloak lerroko gainontzekoa ohar gisa interpretarazten du; eta adibideak azaltzeko erabiliko dugu. *#* sinboloak hitz-muga adierazten du, ezkerreko testuinguruan hitzaren hasiera eta eskuinekoan bukaera.

Erregelak sailkatzeko orduan morfofonologikoak eta ortografikoak bereizi ditugu, morfofonologikoen artean morfologikoak eta fonologikoak bereiztea batere argia ez da eta. Erregela definitzerakoan zehazten den kodeak —FONOL, MORFOL, MORFONOL— aldaketa gertatzeko arrazoia hurbiltzen du, askotan sailkatzeko zailtasunak badaude ere. Izan ere honetaz sakontzeko Urkiaren lana (1995) da gomendagarria.

III.4.1 Aurredefinizioak

Erregelak aztertu baino lehen erregeletan azaltzen diren karaktere-multzoak zehaztuko ditugu. Hona hemen multzo horiek:

A) *Diacritics* izenarekin definitzen diren sinboloak ez dira gauzatzen azaleko mailan eta ez dute eraginik erregelen testuingurua egiaztatzerakoan aipatzen ez badira, beraz hautapen-marka gehienak multzo honetan sartuko dira[3].

---

1 Tresna hau lortu baino lehen eskuzko konpilazioa egin dugu eta bateragarritasun-arazo txiki batzuk detektatu badira ere, erregelak bere sakontasunean ulertzeko baliagarria izan zaigu.

2 Akatsak ekiditearren alfabeto-karaktereak ez diren guztietan ihes-karakterea erabiliko da.

3 Batzuk ez dira sartzen erregelen testuinguruan eraginik izan dezaten.

B) Multzoak, *Sets*, ezaugarri morfofonologiko amankomunak dituzten karaktereek edota sinboloek osatzen dituzte.. Bereizi ditugun multzoak honakoak dira:
   - *Bokal*: bokal papera jokatzen duten karaktere guztiak.
   - *Bokalhutsa*: bost bokalak.
   - *BokIreki*: bokal irekiak.
   - *BokItxi*: bokal itxiak.
   - *Konts*: kontsonante papera jokatzen duten karaktere guztiak.
   - *Txis*: kontsonante txistukariak.
   - *AlboSud* kontsonante albokari eta sudurkariak.
   - *LehGor*: kontsonante leherkari gorrak.
   - *LehOzen*: kontsonante leherkari ozenak.

C) Errepikatzen diren espresio erregularrak modu esanguratsuagoan idazteko defini daitezke *Definitions* atalean. Honako definizioak erabili dira:
   - *Afrik*: kontsonante afrikatuak: *tz*, *ts* eta *tx*.
   - *MorfBuk*: morfema-bukaera adierazteko.
   - *Hasiera*: hitz-hasiera maiuskula kontuan hartu gabe.
   - *MM*: morfema-muga elipsia kontuan hartuz.
   - *LekKas*: kontsonantez hasitako lekuzko kasuak.
   - *KonpErl*: menderagailu konpletibo eta erlatibozkoak.
   - *Rez*: r epentetikoaren erregelan kontuan ez hartzeko sinboloak.

III.4.2 Erregela morfofonologikoak.

Euskararako definitu ditugun hogeita bat erregela morfofonologikotik bi aukera ditugu azalpen honetarako —guztiak azaltzen dira A eranskinean— k-ren ozenketarena eta t-ren galerarena. Erregela hauek tarteko konplexutasuna dutez, beraz, hizkuntza baterako erregela kopurua hogeitik gora bada erregelen idazketa hasiera batean pentsa zitekeena baino korapilatsuagoa da.

k-ren ozenketa

Sudurkariz edo albokariz bukatutako leku-izenekin eta n-z bukatutako morfemekin edo *garren!*-ekin konbinatzen den atzizkiaren hasieran gauza daiteke lexikoko k azaleko g-n. Aukeran edo behartua izatea atzizkiaren arabera izango da, zeren atzizkiak e epentetikoa badu aldaketa aukeran izan bailiteke —e epentetikoaren erregelaren arabera—.

Deskribapena (MORFONOL):

```
k:g <=> [ AlboSud %%: | :n | %!: ] MM (E:0) _ o ;
                    ! *usurbil%+Eko:*usurbilgo
                    ! *usurbil%+Eko:*usurbileko
                    ! egiN+ko:egingo
                    ! hemen+ko:hemengo
                    ! bi+garren!+Eko:bigarrengo
                    ! bi+garren!+Eko:bigarreneko
```

t-ren galera

Ondoko kasu hauetan galtzen da t karakterea:

- leherkari gor, albokari, sudurkari edo h-z hasten den morfema baten aurrean.
- Kontsonante afrikatuaren parte denean, leherkari gorrekiko zenbait konbinaziotan.

Guztiz fonologikotzat har daiteke eta espezifikazio zehatza ondoan dator.

Deskribapena (FONOL):

```
t:0 <=> _ MM [ :LehGor | AlboSud | h ] ;
          _ Txis %%:0 MM E:0 LehGor ;
          _ Txis MM t ;
          n _ Txis MM k ;
                    ! bait+gara:baikara
                    ! bait+naiz:bainaiz
                    ! *zarautz%+Eko:*zarauzko
                    ! utz+te:uzte
                    ! jantz+te:janzte
                    ! etxe+rantz+ko:etxeranzko
```

III.4.3 Erregela ortografikoak

Batere eragin edo arrazoi fonologikorik ez duten erregelak ortografikoak deitu ditugu. Dauden lauetako ordezkari gisa h-ren galerarena aurkezten da ondoren.

h-ren desagerpena

Aditz-erroa *beR* aurrizkiarekin lotzean —r gogorrarekin bukatutako beste aurrizkietara zabal daiteke— erroa h-z hasten bada, h hori desagertu egiten da.

Deskribapena:

```
h:0 <=> R: MM _ ;
                    ! beR+hasi:berrasi
```

III.5 Programa eta emaitzak.

Deskripzio linguistikoa aztertu eta gero, ikus ditzagun programaren inplementazioaren nondik-norakoak eta lortutako emaitzak.

III.5.1 Inplementazioa.

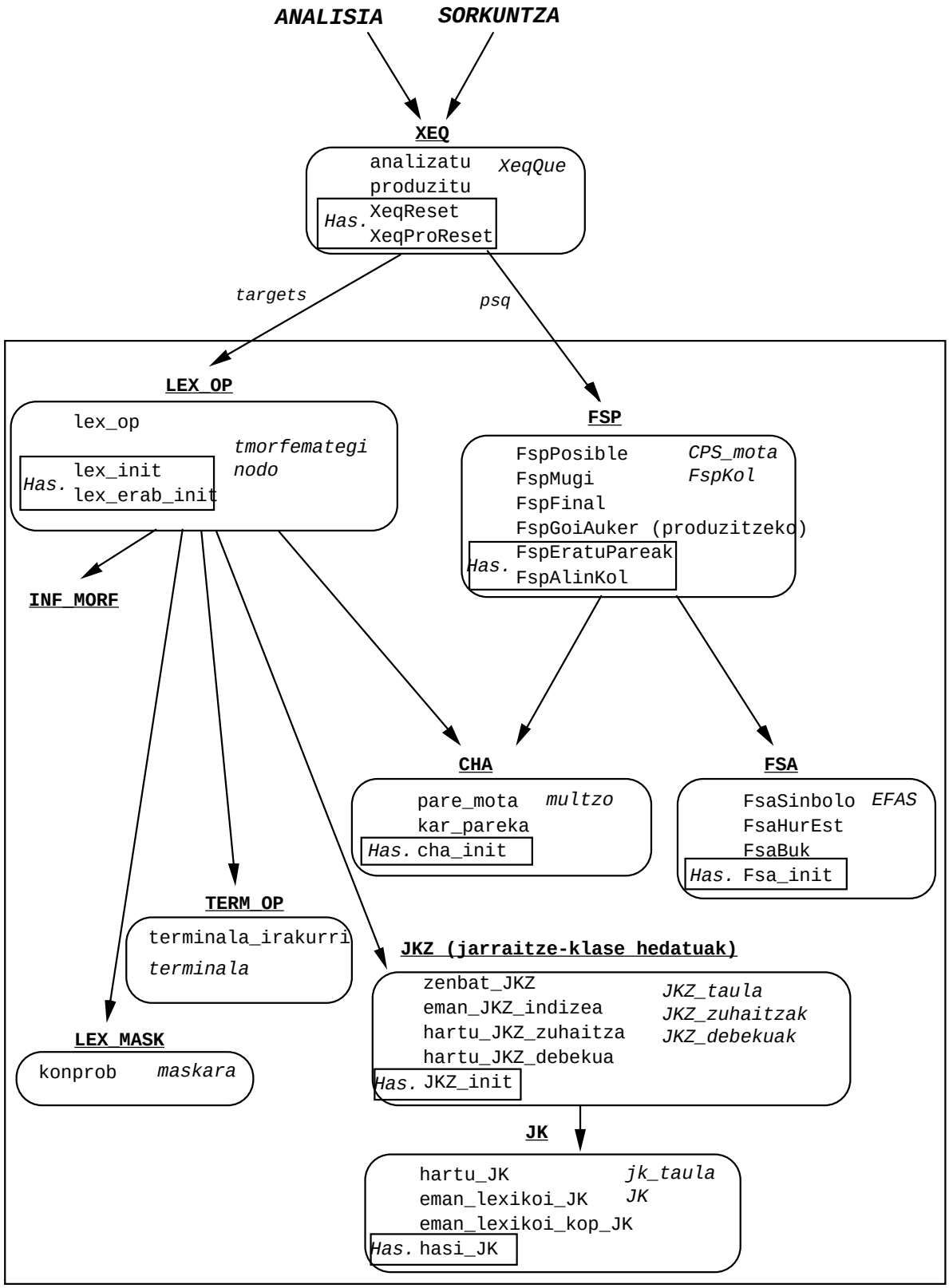
Proiektuan hasi ginenean bi mailatako morfologiari aurre egiten zion erabilpen libreko programarik ez zegoenez, geure inplementazioari ekin genion. Ondoren, PC-KIMMO (Antworth, 90) izeneko softwarea eskuragarri izan genuen, baina gure inplementazioarekin jarraitu genuen honako arrazoi hauengatik:

- Ez zuen ekarpen handirik egiten guk egindako programarekin alderatuz; gehien behar genuen erregelen konpiladorea ez zuen eta.
- Bi mailatako formalismoari egin nahi genizkion aldaketak, geure diseinuaren gainean geneuzkan pentsatuta eta hortik jarraitu genuen.
- Merkaturatze-asmoari begira bide egokiagoa zen gure inplementazioarekin jarraitzea.

Programa analisia zein sorkuntzarako baliagarria da, baina sorkuntzaren kasuan arazo bat gainditu behar izan dugu. Euskara hizkuntza eranskaria denez, eta genitiboen atzean berriro deklinabide osoa erants daitekeenez, lema batetik abiatuta sortzen diren formak infinituak dira, teorian behintzat. Honen aurrean, sorkuntza egiterakoan sorkuntza-ahalmenari muga bat jartzen dion parametro bat gaineratu behar izan da, morfema-kopuru maximoa edo genitibo-kopuru maximoa zehazten duena.

III.5.1.1 Programa

Programaren nondik-norakoak zehatz-mehatz azaldu zituen Koskenniemik bere tesiaren laugarren kapituluan (Koskenniemi, 83). Horretan oinarrituta, aldaketa txiki batzuk gora-behera eta proposatutako jarraitze-klase hedatuen mekanismoaren eransketarekin, III.7 irudian zehazten den eskemak irudikatzen duen inplementazioa egin genuen C programazio-lengoaia erabiliz.



III.7 irudia.- Bi mailatako morfologiaren gure inplementazioaren eskema.

Bertan nagusiki hiru modulu bereizten dira: backtracking-ilara kontrolatzen duen XEQ, lexikoa kudeatzen duen LEX_OP eta erregelen itzultzaileei dagokien FSP. Azter ditzagun banan-banan bakoitzaren zeregina, funtzio nagusiak eta datu-motak azaltzearren.

- XEQ moduluan *XeqQue* ilara definitzen da, bertan backtracking-aukerak meta daitezen. Lexikoa atzitzen aukera berriak sortzen dira, eta erregela-sistemak onartzen dituenak metatzen dira ilaran karakterez karaktere aztertzen jarraitzeko. Analisi zein sorkuntzarako datu-mota bera erabili arren, lehen kasuan azaleko karaktereek aukerak mugatzen dituzten bitartean, sorkuntzan lexikoa eta bertan adierazten den morfotaktika da aukerak mugatzeko bide bakarra.

- LEX_OP moduluan *trie* egiturari jarraitzen dion lexikoaren atzipena gauzatzen da. Bertatik funtzio-multzo laguntzaileak atzitzen dira ondoko funtzioetarako: karaktere-bikoteak eta multzoak kontrolatzeko, adabegi batetik zintzilik dauden arku edo lexiko-karaktereak jakiteko, morfema baten bukaera detektatzeko, morfemari dagokion informazio morfologikoa eta ondorengo azpilexikoak lortzeko. Modulu honen osagarri gisa, datu-base lexikotik *trie* egitura osatzen duen *LEXIKOI* izeneko modulua dago.

- FSP moduluak bi mailatako erregelak gauzatzen dituzten egoera finituko itzultzaileen kudeaketa burutzen du. Bertako funtzio garrantzitsuenak hauexek dira: karaktere-bikote bat posible denentz esatea, itzultzaileen mugimendua gauzatzea bikote baten eraginez, eta bukaerako egoeraz informatzea.

III.5.1.2 *Token*-ezagutzailea edo iragazlea.

Esan den bezala, zuriune batez bereiziko hitz-elkarketa, lokuzioak eta orokorrean hitz anitzeko terminoak ez dira analizatzen oraingoz; hala ere, beren tratamendua bideratzeko datu-base bat ari gara osatzen. Beraz, analisirako tratamendu-unitatea hitza da, baina formatoa kontuan hartzen badugu hitza mugatzea ez da hain prozesu erraza.

Ezaguna denez *token*-ezagutzaile[1] baten zeregina analizatzeko unitateak, hitz-zatiak, identifikatzea da. Edozein testurekin aritzeko diseinaturiko analizatzaile baterako ezinbesteko tresna dugu hau, bere eginkizunen artean ondoko elementu hauen identifikazioa eta tratamendua duelarik:

- zenbakiak, arruntak edo erromatarrak, dagokien deklinabidearekin.
- laburdurak eta siglak dagokien deklinabidearekin.
- lerro-bukaeran hitza banatzen duen marratxoa (*hyphenation*).

[1] *token* eta testu-hitza sinonimotzat hartzen da lan honetan zehar.

- zuriuneak eta puntuazio zeinuak, hitzen arteko bereizgarriak direlako.
- gainontzeko markak eta karaktere bereziak.
- maiuskulaz idatzitako hasierako letrak, zatiak eta izenburuak.
- corpusetan agertu ohi diren testuaren identifikazioak —urtea, testu-mota, idazlea, etab. zehazten duena—, orri-zenbakiak, beste hizkuntzen aipamenak etab.

Eman lezakeena baina lan neketsuagoa da euskararako halako ezagutzailea egitea, elementu batzuek —marra edo puntua adibidez— funtzio anitza dutelako eta beste hizkuntzetan formatoaren bidez oso erraz identifikatzen diren osagai batzuk, euskaraz deklina daitezkeenez, hain identifikaerrazak ez direlako.

Konplexutasun honen aurrean eta beste ezagutzaile batzuen bidetik, automata bat da identifikazioaz arduratzen den tresna. Lortzen den automata konplexu samarra da.

III.5.2 Analizatzailearen emaitzak eta estaldura-tasa.

Atal honetan analizatzailearen ezaugarri garrantzitsuenak aztertuko ditugu. III.8 irudian "Eta gauza aundirik ekartzerik ez zuen izan" esaldia analizatzean lortutako emaitza ikus daiteke. Bertan ikus daitekeenez, analisiaren emaitza zerrenda paranterizatu bezala ematen da, hitz bakoitzeko analisi-aukera desberdinekin —anal1, anal2,... identifikadoreaz bereiziak—, eta lerro bakoitzean morfema baten informazioa zehaztuz. Hitz baterako analisirik aurkitzen ez bada analisirik gabe agertuko da, ondoko kapituluan zehaztuko diren prozeduren zain. Adibidean *aundirik* hitza analizatu gabe agertzen da forma ez-estandarra da eta.

C eranskinean testu-zati luze samar baten adibide osoagoa azaltzen da, bertan datorren kapituluan azaltzen diren tratamenduak —forma ez-estandarren ezaguera eta analisia lema lexikoan egon gabe— buruturik daudelarik. Ondoko kapituluko IV.4 atalean deskribatzen den tratemendua burutua izan da emaitzen gainean.

Emaitzaren formatoa ikusita, pasa gaitezen estaldurari buruzko zenbait datu ematera. Datuak lehen kapituluan aipatutako corpusen gainean hartu dira, eta III.9 irudian azter daitezke. Corpus bakoitzeko bi neurri ematen dira, bat hitz guztiak kontuan hartuz (corpus) eta bestea hitz desberdinak bakarrik kontuan hartuz (zerrenda). Espero zitekeen bezala, zerrendetan tasa okerragoa da, analizatzen ez diren hitzak, hitz arruntak ez direnez, gutxitan errepikatzen baitira.

```
((forma "*eta")
  ((anal 1)
   ((lema "etA")((KAT JNT))))
)
((forma "gauza")
  ((anal 1)
   ((lema "gauza")((KAT ADI))))
  ((anal 2)
   ((lema "gauzA")((KAT IZE))))
  ((anal 3)
   ((lema "gauzA")((KAT IZE)))
   ((morf "a")((KAT DEK)(KAS NOM)(NUM S)(MUG M))))
)
((forma "aundirik")
)
((forma "ekartzerik")
  ((anal 1)
   ((lema "ekaR")((KAT ADI)))
   ((morf "tzerik")((KAT ERL)(ERL KONP))))
  ((anal 2)
   ((lema "ekaR")((KAT ADI)))
   ((lema "tze")((KAT ASP)(KER IZE)))
   ((morf "Rik")((KAT DEK)(KAS PAR)(MUG MG))))
)
((forma "ez")
  ((anal 1)
   ((lema "ez")((KAT ADB))))
  ((anal 2)
   ((lema "ez")((KAT IZE))))
)
((forma "zuen")
  ((anal 1)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun))))
  ((anal 2)
   ((lema "zu")((KAT IOR)))
   ((morf "eM")((KAT DEK)(KAS GEN)(NUM P)(MUG M))))
  ((anal 3)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun)))
   ((morf "En")((KAT ERL)(ERL ERLT))))
  ((anal 4)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun)))
   ((morf "En")((KAT ERL)(ERL ZHG))))
)
((forma "izan")
  ((anal 1)
   ((lema "izaN")((KAT ADI))))
  ((anal 2)
   ((lema "izaN")((KAT ADI)))
   ((morf "0")((KAT ASP)(ADM PART))))
)
```

III.8 irudia.- Esaldi baten analisia.

Corpus ezberdinetako emitzen arteko desberdintasuna testu-motak eragiten du; horrela, Argiako testu-zatietan atzerriko leku- edo pertsona-izen askoren agerpenak —kazetaritzan maiz gertatzen den fenomenoa— baldintzatzen du emaitza.

III.9 irudian ikusten denez, analizatzailearen estaldura-tasa orokorra %90etik gorakoa da corpus guztietan. Corpus handiko zerrendari dagokion emaitza txar hori, %70a, agertzen da bi arrazoiengatik: esan den bezala corpus horretan eredu estandarrari jarraitzen ez dioten testu, testu tekniko eta akats asko daudelako batetik, eta bestetik, oso

gutxitan agertzen diren hitzek —asko analizatu gabeak— eta askotan agertzen direnak berdin baloratzen direlako zerrenden gainean kalkulatzean. Batez-bestekoa %92 inguruan dago corpusetan, beste hizkuntzetarako analizatzaileetan ematen diren datuekin alderatuz baxua izanik, %95etik gora izan ohi baitira beti.

Testuak	**hitzak**	**analizatu gabe**	**tasa(%)**
1a.-Argia aldizkaria (corpus)	4.864	379	**92,2**
1b.-Argia aldizkaria (zerrenda)	2.607	307	**88,2**
2a.-Filosofiari buruzko artik.(C)	2.343	95	**95,9**
2b.-Filosofiari buruzko artik.(Z)	1.429	85	**94,1**
3a.-EEBSko azken urteak (C)	23.364	1.795	**92,3**
3b.-EEBSko azken urteak (Z)	9.313	1.312	**85,9**
4a.-EEBS estandarra (C)	396.840	36.172	**90,9**
4b.-EEBS estandarra (Z)	67.816	20.920	**70,0**

III.9 irudia.- Estaldura-tasari buruzko datuak.

Tasa baxu hauen arrazoiak, hauexek dira:

A) Euskara ez-estandarraren erabilera. Batasunaren historia labur, aldakor eta bukatugabean euskara estandarra ondo definitu gabe dago eta definiturik dagoena ez dago nahikoa hedatua. Gainera euskalkien aberastasunaren eraginez idazle batzuek, nahita ala nahi gabe, erabilpen dialektala egiten dute. Ondorioz, euskara estandartzat hartzen ez diren hitzak maiztasun handikoak dira; adib. *bait*, *haundi* edo *batzu*. Honen aurrean datorren kapituluan jorratzen den aldaeren tratamendua proposatzen dugu.

B) Lexikoan agertzen ez diren lemak. Hauen artean bereizketa egin behar dugu, lau iturri nagusi daudelako.
   - Lehenengoz, erdaren eraginez egiten diren mailegu desegokiak edota lexikoan jasogabeak. Hauen konponketa zail samarra da, baina corpusetan maiztasun-muga batetik aurrera agertu ahala lexikoan sartzeko asmoa dugu.
   - Bigarrenez, lexikoan agertzen ez diren lemak, gehienak leku- zein pertsona-izenak edo lexiko berezituei dagozkienak. Hauek konpontzeko bi bide proposatzen dira, zenbait leku- zein pertsona-izen lexikoan sartu behar diren bitartean, besteentzat lexiko berezituak proposatzen dira (ikus 4. kapitulua).

- Hirugarrenez eratorpen eta elkarketa “berriak” dauzkagu. Eratorpena irregularra denez eta euskararena ondo aztertu gabe dagoenez, egin dugun aukeraren arabera eratorpen zeharo erregularrak bakarrik sartu dira morfema gisa, gainontzekoetan eratorpen lexikalizatuak lema gisa sartu direlarik lexikoan.
- Azkenik, euskararako analizatzaile batek ezagutu ezin dituen beste hizkuntzetako hitzak.

Kontzeptua	**1b-n**	**2b-n**	**bietan**
Ezagutu gabeko hitzak (guztira).	307 (%100)	85 (%100)	392 (%100)
A.-Erabilpen ez-estandarra	101 (%32,9)	28 (%32,9)	129 (%32,9)
B1.-Erdararen eragina	31 (%10,1)	2 (%2,4)	33 (%8,4)
B2.-Lexikoan ez egotea	68 (%22,1)	16 (%18,8)	84 (%21,4)
B3.-Eratorpen/elkarketa “berria”	33 (%10,7)	13 (%15,3)	46 (%11,7)
B4.-Hitz arrotzak	39 (%12,7)	14 (%16,5)	53 (%13,5)
C.-Akatsak	30 (%9,8)	10 (%11,8)	40 (%10,2)
D.-Bestelakoak	5 (%1,6)	2 (%2,4)	7 (%1,8)

III.10 irudia.- Ez-estaltzearen arrazoiak ebaluatzen.

Hauez gain hizkuntzari dagozkion zenbait “eragozpen” daude. Euskararen flexio aberatsa dela eta, erro baten faltak forma ezezagun anitz eragiten dezake. Gainera juntagailurik ez egotean, corpusen kasuan ez dago juntagailuen maiztasun handien eraginaz baliatu.

Datorren kapituluan proposatuko diren hobekuntza batzuk burutuz emaitzak hobetzen dira, eta %95etik gorakoak izaten dira.

III.10 irudian bi testu-zatiren gainean egindako azterketaren emaitzak azaltzen dira, zehaztutako arrazoiei pisu bat egokitzearren. Aukeratutako testuak hitz-zerrendak dira, 1b eta 2b kodearekin identifikatu ditugun Argiako zatiena eta filosofi testuarena hain zuzen. Datu hauek hartu ditugu kontuan analizatzailea sendotzeko teknikak diseinatzerakoan, datorren kapituluan ikusiko den legez.

III.5.3 Gainsorreraren arazoaz.

Hasieratik proiektuaren helburuetako bat gainsorrera ekiditea izan da. Honen arrazoi berehalakoa egiaztatzaile/zuzentzaile bat eraikitzeko erabili behar zela bazen ere, ez da arrazoi bakarra izan, zeren gainsorrera ez duen sorkuntza morfologikoa oso elementu garrantzitsua da etorkizuneko erabilpenetarako.

Diseinu-erabaki honek azpimarratu beharreko ondoko bi ondorioak izan ditu:

- Aurreko atalean B3 kodearekin identifikatu dugun kasuetan analisirik ez lortzea. Eratorpena eta elkarketa lantzeko beste aukera genuen, generalizazioarena hain zuzen; ondorioz, estaldura-tasa handiagoa lortuko genukeen. Generalizazio hau egiteko bide errazetik —halako atzizkiak izen guztiekin, halakoak aditz-erroekin etab.— alde egin dugu, erabateko gainsorrera sortzen baitu alde batetik, eta atzizki hauek biltzean sortzen diren aldaketak konplexuak eta aztertu gabeak direlako bestetik. Arazo honen aurrean eratorpenaren azterketa sakon bat ari gara egiten (Aduriz & Aldeazabal, 95).
- Flexio-morfologiaren aldetik, morfotaktika konplexu samarra bihurtu da gainsorrera gerta ez zedin, salbuespenak kontuan hartu direlarik. Hitz batzuk defektiboak dira deklinabidearen aldetik —*batzu* adibidez—, aditz-erro batzuekin arazoak daude —itxi adib.— etab.

Beraz, sorkuntza legezko mugen barruan mantentzearren deskribapenaren, konplexutasuna handitu egin da, eta estaldura-tasa jaitsi.

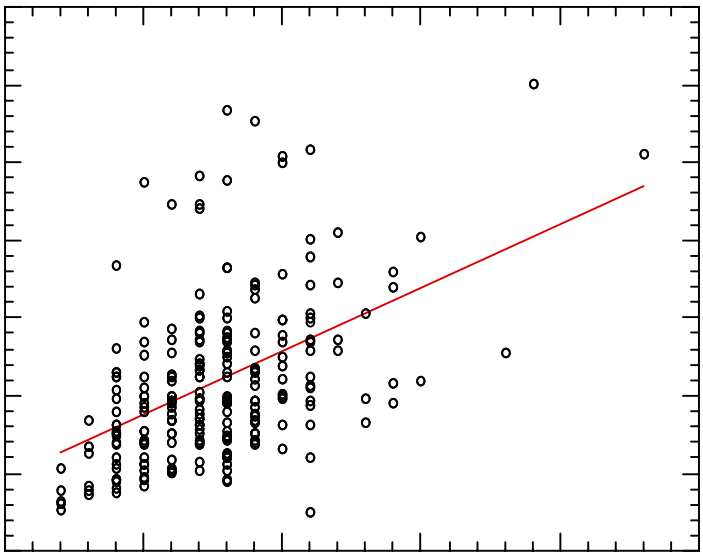
III.5.4 Eraginkortasunari buruzko zenbait datu eta gogoeta.

Estaldura-tasa aztertu ondoren abiaduraren eta leku-hartzearen neurriak emango dira atal honetan, beren azalpenarekin batera. Lexikoak bi Mega inguru hartzen du, eta laburtzeko teknikak erabiliz dezente jaits daiteke.

Abiaduraren aldetik, hona hemen filosofia izeneko corpusekin lortutako emaitza Sun-Sparc IPX baterako : 0,5 s/hitza, edo gauza bera dena 2 hitz/s. PC-KIMMO erabiliz antzeko emaitzak lortzen dira, eta antzekoak aipatzen ditu Oflazer-ek (1.994) turkierarako. Kontuan hartu behar da abiadura hau kalkulatu dela hitz guztien analisia burutzen denean eta gainera analisi posible guztiak sortuz. Hau azkar daiteke bi bidetik: batetik, hitz errepikatuak berriro ez analizatuz —horrekin abiadura ia bikoiztu egin daiteke, hitz bakoitza batez-beste ia bi aldiz agertzen baita corpusetan—, eta bestetik, maiztasun handieneko hitzen analisia buffer batean gordez —honela, eta gure corpusetan

egindako kalkuluetan oinarriturik, analisien erdia aurrez daiteke bufferrean 600 hitzen informazioa gordez, eta %80a 8000 hitzenarekin—.

Izan ere, hasiera batean espero zitekeena baino motelagoa gertatzen da analisia, eta honen zioa bilatzen saiatu gara. Koskenniemik eta Churchek konplexutasunaren inguruko bere artikuluan, analisi-urrats kopuru batzuk zehazten dituzte suomierarako. Beraiek ematen dituzten zenbakiak eta euskararako kalkulatu ditugunak oso bestelakoak dira. III.11 irudian gure sistemarako lortutako emaitzak, batez-beste, beraiek ematen dituztenak baino hamar bat aldiz gehiago baitira. Izan ere, beraien datuetan bezala, analisi-urrats kopurua luzeraren funtzio lineal batez hurbil daiteke.



III.11 irudia.- Analisi-urratsei buruz gure sistemaren gainean egindako estatistikak.

Aurreko kapituluaren II.4 atalean ematen da formalismoaren konplexutasunaren berri, konplexutasun hori handitzen duten faktoreak zehaztuz. Hortik abiaturik, bila daiteke alde horren zergatia. Arrazoia bikoitza da, batetik hizkuntzari dagozkion erregela diferenteek eta hitz bakoitzeko morfema-kopuruak eragina dute dudarik gabe; baina bestetik gure proiektuan egindako aukera baten[1] —ahalik eta alomorfo gutxien sartzearena— eragina ere bada, zeren lexikoa aztertuta atzizkiak oso sakabanatuta baitaude, osagai oso gutxiko azpilexiko askotan barreiaturik. Azken honen ondorioz analisi-aukera asko sortzen dira

1 Aukera hau bi mailatako morfologiak bultzatzen du, baina eztabaidagarria da nonraino aplikatu behar den.

jarraitze-klase bakoitzeko —gutxienez bat azpilexiko bakoitzeko—, horietako gehienak alperrik jorratuko direlarik.

Honen aurrean, eta II.4.3.1 paragrafoan azaldutakoaren arabera, lexiko-fusioa izango litzateke konponbidea; baina gure inplementazioaren gainean fusioa tratatzeko aldaketak egin eta gero neurriak hartu genituen eta denboraren aldetiko emaitzak antzekoak ziren, lexikoaren memori hartzea %10ean laburtu arren. Honen arrazoia hauxe da: jarraitze-klaseari dagozkion azpilexiko anitz korritu beharrean, azpilexiko bakar bat korritzen da, baina morfotaktikari buruzko informazio falta dela eta[1], morfema gehiago aztertzen da, eta gehienak alferrik aztertu ere.

Fusioaren emaitza kaskarra ikusita, eta jarraitze-klase batzuei dagozkien azpilexiko kopuru handiari erreparatuz, beste hobekuntza bati ekin genion, gehien erabiltzen diren jarraitze-klaseetako bakoitzari dagozkion azpilexiko guztiak azpilexiko bakar batean bilduz. Honen ondorioz, alomorfo anitz sortzen dira —ez espezifikazioan, baina bai benetan jorratzen den lexikoan—, memori hartzea %5ean igoz, baina denbora eta analisi-urratsak %15ean jaisten dira. Hobekuntza handiagoa lor daiteke bide honetatik jarraituz, morfotaktikari dagozkion urratsak optimizatzeko programa bat eginez, baina lortuko den hobekuntzaren muga nahikoa gertu dago.

III.6 Erabateko hobekuntza: lexiko-itzultzaileak.

Aurreko kapituluko II.4.3.2 atalean lexiko-itzultzaileen (Karttunen, 94) sarrera egiten da. Guk ideia hau jorratzen bideratzen duten tresnak, Xerox-eko *twolc* (Karttunen & Beesley, 92) eta *lexc*(Karttunen, 93) eskuratu eta ebaluatu ditugu. Pasarte honetan lexiko-itzultzaileen ezaugarriak eta beraien bidez egindako inplementazioa azaltzen dira.

III.6.1 Lexiko-itzultzaileen ezaugarriak.

Aipatutako II.4.3.2 atalean esandakoa laburtuz, hauek dira lexiko-itzultzaileen ekarpenak:

- Lexiko eta erregelei dagozkien egoera finituko itzultzaileak (FSTak) bakar bakar batean integratzean, eta optimizazio-teknika sofistikatuak erabiltzean, iraultzailea da abiaduraren aldetik, mila bat aldiz azkarragoa gertatuz.
- Erregelak itzultzaile bihurtzeko konpiladorea.

[1] Azpilexikoak fusionatzean morfotaktikari buruzko informazioa *trie* egituraren hostoetan baino ezin da egon.

- Morfemen desitxuratzea eragiten duten diakritikoen erabilpena bazter daiteke, horien ordez ezaugarri morfologikoak erabil daitezkeelako. Horrela lexikoa nahiz erregelak argiagoak dira.Horrez gain, lexikoan forma kanonikoa adieraz daiteke, ohizko erregelekin arituko den ohizko lexiko-mailarekin batera.

- Erregela paraleloen multzo desberdinak konposa daitezke, azkenean denen konposaketa itzultzaile bakar batean konpilatuz, tarteko adierazpideek zekartzaten arazoak ekidinez. Honek bi abantaila eskaintzen du: deskribapen ahalmen handiagoa batetik, eta deskribapena erraztea maila desberdinetan banatzeko aukeraz.

Morfotaktikaren aldetik berriz, lexiko-itzultzaileek ez dakarte aurrerapausorik, urruneko menpekotasunak adierazteko bide egokirik gabe jarraituz orain arte behinik behin. Urruneko menpekotasunak adierazteko orduan, beraz, II.3.4.3 atalean aipatutako bi mekanismo zakarrak baino ez dira gelditzen: erregela artifizial samarren bat erabiltzea (ikus §III.6.2), edo azpilexiko batzuen bikoizketa.

III.6.2 Euskararako aplikazioa.

Gure sistematik lexiko-itzultzaileetara pasatzeko ondoko urratsak eman dira:

- Lehen urrats batean, sistema ahalik eta aldaketa gutxienekin igaro sistema batetik bestera, horretarako formato-aldaketez gain egin behar genuen sakoneko lan bakarra urruneko menpekotasunak ebaztea zela. Honekin sistema bera lortzen da, baina askoz ere eraginkorragoa. Aipatutako urruneko menpekotasun horiek ebazteko erregela artifizial batzuk idatzi ziren.

- Lexiko-itzultzaileek eskaintzen dituzten ahalmen berriez baliatzea. Gure sisteman erabiltzen diren morfofonemak eta hautapen-markak baztertzea da helburu nagusia, horretarako erregelak aldatu behar direla. Era berean, erregela morfofonologikoak eta urruneko menpekotasunak ebaztekoak banatu dira bi maila desberdinetan, eta zenbait morfemari egokitu zaie forma kanonikoa.

III.6.2.1 Urruneko menpekotasunak ebazteko erregelak.

Kapitulu honetako III.3.3.2 pasartean aztertu dugu euskarazko urruneko menpekotasuna ebazteko modua guk proposaturiko jarraitze-klase hedatuen mekanismoaren bidez. Lexiko-itzultzaileekin halakorik erabiltzerik ez dagoenez, bi mailatako erregelen bidez ebatziko dugu arazo hau, mekanismo hori horretarako diseinaturik ez egon arren.

Euskarazko urruneko menpekotasunaren kasuetan aukerak murrizten direnez gero, laugarren motako erregelak erabiliko dira, debeku-ezarpenak hain zuzen (Ikus §II.3.2).

Lehen bi kasuak, *bait* eta *ba*-ren morfotaktikari dagokiona hain zuzen, erregela bakar batez ebatz daitezke, hasierako b karakterea debekatuz baldintzazko *ba* eta *bait* morfemen ondoren morfema bat baino gehiago baldin badager.

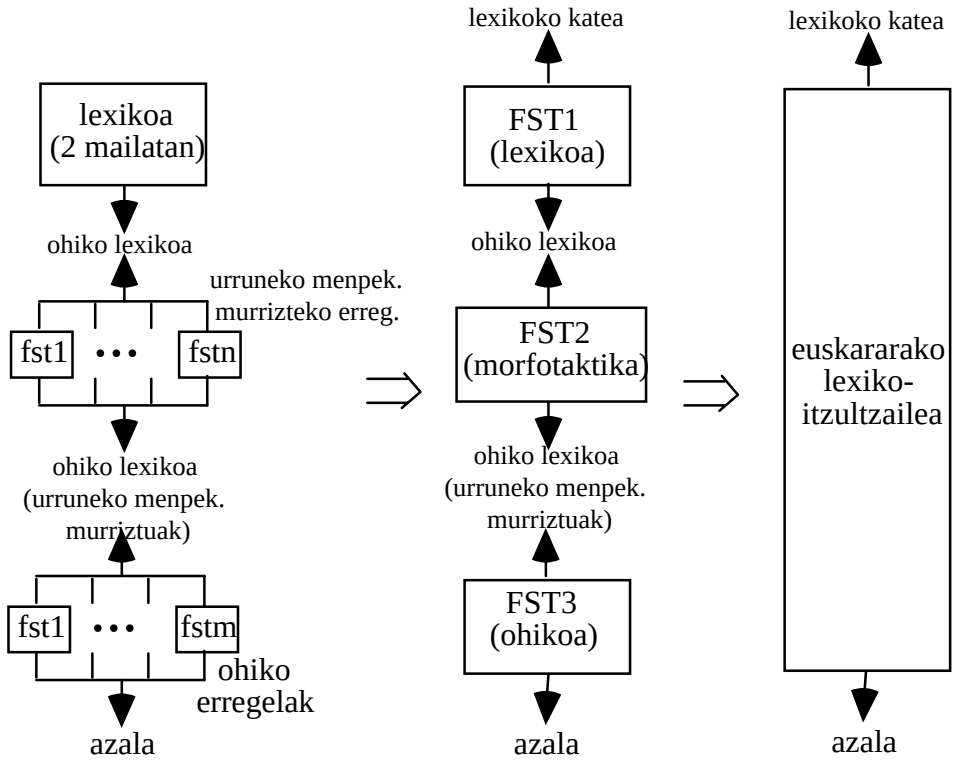
```
b:b /<= #: _ [ a {Baldin} | a i t ] MM [=:=]* MM ;
```

Azpimarratzekoa da baldintzazko *ba* morfeman marka edo ezaugarri morfologiko bat —azken hau da adibidean agertzen dena: *{Baldin}*— behar dela, beste *ba* morfematik bereiztearren.

Marratxoaren kasuan jarritako murriztapenean, bi murriztapen datoz, bi marren agerpena debekatzea batetik, eta bestetik ondoren aditza agertzen bada, aditz hori nominalizatua izan dadila *te* edo *tze* morfemaz.

```
%-:%- /<= _ MM [=:=]* %- ;
          _ MM [=:=]* {ADI} MM \[ t (z) e MorfBuk ] ;
```

Erregela hauek ohizko erregela morfofonologikoekin batera jar badaitezke ere, lexiko-itzultzaileek aukera ematen digute beste maila batean jartzeko, horrela morfotaktikari eta morfofonologiari dagozkien erregelak nahastu gabe utziz (ikus III.12 irudia).



III.12 irudia.- Euskararako lexiko-itzultzaile baten eraikuntza.

Esan bezala, lexikoan bi maila desberdin adieraz daiteke, bat emaitza gisa lortzeko eta bestea ohizko erregelekin aritzeko. Horren bidez lortzen da forma kanonikoa bultzatzea ohizko erregelak aldatu gabe. Azala eta forma kanonikoaren artean distantzia handi samarra eta ez-erregularra zenean analisiaren emaitza ez zen forma kanonikoa, bere aldaera baizik. Horrela hirugarren pertsonako *hau* erakuslearen flexio batzuen emaitza *hon* lema ez-kanonikoaz lortzen zen. Lexiko-itzultzaileetan posiblea da *hau:hon* bikotea adieraztea lexikoan; ondorioz analisiaren emaitza *hau*-z lortuko da, baina ohizko erregelak *hon*-en gainean aplikatzen dira.

III.6.2.2 Ohiko diakritikoen eta erregelen berrikuntza.

III.12 irudiko lehen bi mailak, lexikokoa eta morfotaktikakoa, aztertu ondoren; ikus dezagun zer egin daitekeen ohizko erregeletan —aipaturiko irudian hirugarren mailan daudenak— lexikoko diakritikoak desagertarazteko.

Bigarren kapituluan azaldu den bezala, Koskenniemik proposatu zituen morfofonemak eta hautapen-markak erregelen aplikazioa murrizteko. Hala ere, lexiko-itzultzaileetan informazio morfologikoa adierazpen lexikoarekin batera doa —eta ez bereizirik bi mailatako eredu klasikoan bezala—; beraz, posible da informazio hau erabiltzea, ezaugarri morfologikoak hain zuzen, erregelak baldintzatzeko. Horretarako, diakritiko guztiak aztertu behar dira, eta ahal den kasuetan existitzen den ezaugarri morfologiko baten bidez ordezkatuko da, ezin denean —morfofonemetan gertatu ohi dena— bere ordez ezaugarri berri bat sortuz.

Esan bezala, lexikoan bi maila desberdin adieraz daiteke, bat emaitza gisa lortzeko eta bestea ohizko erregelekin aritzeko. Horren bidez lortzen da forma kanonikoa bultzatzea ohizko erregelak aldatu gabe. Azala eta forma kanonikoaren artean distantzia handi samarra eta ez-erregularra zenean analisiaren emaitza ez zen forma kanonikoa, bere aldaera baizik. Hona hemen aipaturiko diakritikoak (ikus §III.3.2) eta dagozkien ezaugarri morfologikoak:

R esanahi bikoitza du: lemetan *r* gogorra, ezaugarri berri batez ordezka daitekeena: {Rgogor}; eta bestetik *r* epentetikoa, beste ezaugarri berri batez ere ordezka daitekeena: {Repent}. Adib. *zakur{IZE}{Rgogor}* eta *ik{DEK}{Repent}*.

Q Ezaugarri berria ere beharko luke *e*-ren epentesiari begira: {EpBerez}. Adib. *har{IOR}{EpBerez}*.

~ {Rzahar} ezaugarri berria. Adib. *hiru{ZNB}{Rzahar}*

E {Eepent} .ezaugarri berria. Adib. *ko{ZNB}{Eepent}*.

N {Ngal} ezaugarri berria. Adib. *egin{ADI}{Ngal}*.

M Aurreko ezugarri bera, kategoriaren arabera bereiz baitaiteke.. Adib. *aren{DEK}{Ngal}*.

\ {Nauk} ezaugarri berria. Adib. *en{DEK}{Nauk}*.

A {Aorg} ezaugarri berria. Adib. *ama{IZE}{Aorg}*

Aurreko ezaugarriaz gain {Asalbu} berria. Adib. *kultura{IZE}{Aorg}{Asalbu}*.

@ {Ebihur} ezaugarri berria. Adib. *atera{ADI}{Ebihur}*.

& Leku-izen batzuetan galtzen den bukaerako a artikulua. Adib. *Azpeitia{LIB}{Aartik}*.

^ Aditz jokatuetako informazio morfologikoa erabiliz ordezka daiteke.

% Lexu-izeneko ezaugarriarekin nahikoa. Adib. **usurbil{LIB}*.

: {DekBerez} ezaugarria izen bereziaren ezaugarriarekin batera. Adib. **h*b{IZB}{DekBerez}*.

/ lehengoaren aldaera {DekBerez2}. Adib. **m*i*t{IZB}{DekBerez2}*.

$ Aurreko {DekBerez} erabil daiteke aditz flexionatuarenarekin konbinatuz. Adib. *du{ADL}{DekBerez}*.

! *garren* morfemarekin egin daiteke erregelak zerbait zailduz.

+ morfema muga bere horretan mantentzen da.

Aldaketa hauekin erregelak aldatu behar dira baina ez asko, ulergarritasuna eta irakurgarritasuna irabaziz. Ondoren azaltzen dira n-ren galera gobernatzen duten erregelak bi formatoetan.

Deskribapena ohizko moduan:

```
N:0 <=> _ MM [ t e | k i MorfBuk ] ;
Cx:0 <=> _ MM k: ;
          where Cx in (M %\) ;
%\:0 => _ MM g a t i k ;
n:0 <=> _ MM E: KonpErl ;
n:0 => _ t:0 Txis %+: t ;
```

Deskribapen berria:

```
n:0 <=> _ {ADI} {Ngal} MM [ t e | k i MorfBuk ] ;
          _ {DEK} [{Ngal}|{Nauk}] MM k: ;
          _ {ADL} MM (0:) KonpErl ;
n:0 => _ {Nauk} MM g a t i k ;
          _ t:0 Txis {ADI} %+: t ;
```

Lexiko-itzultzaileen bidez, beraz, abiaduraren aldetik lortzen den aurrerapen ikaragarriez gain bestelako abantailak ere badaude, haien artean erregelen irakurgarritasuna ere azpimarra daitekeela.

III.7 Morfosintaxia.

Analisi morfologikoaren emaitzak ez dira zuzenean erabilgarriak beste aplikazioetarako, eta euskararen konplexutasun morfologikoa kontuan hartuz askoz ere gutxiago.

Sintaxian, etiketatzaile/lematizatzaileetan edota beste aplikazioetan erabiltzeko, emaitza morfologikoa tratatu begin behar da, emaitza trinkoagoa eta esanguratsuagoa izan dadin. Tratamendu honi morfosintaktikoa deituko diogu, eta euskararen kasuan kontuan hartzeko ezaugarri garrantzitsuenak hauek dira:

- Kasu anitzak. Izen eta adjektiboen kasuetan batez ere, atzizki desberdinak meta daitezke lema baten ondoren, kasuari buruz, eta honi dagokion numero eta determinazioari buruz, informazio anitz lortzen delarik. Morfologiaren ikuspuntutik informazio hau guztia interesgarri izan badaiteke ere, ondorengo tratamendurako batzuetan ez da esanguratsua eta tratamendua zailtzen du. Horri erantzuteko metatutako informazioaren prozesaketari ekin behar zaio. Gehienetan azken kasuari dagokion informazioa da esanguratsuena eta emaitza gisa lortu behar dena.

- Elipsia. Aurretik aipatutakoaz gain, kasu batek, genitiboaren ondoren beste kasu bat agertzeak, izen-elipsia adierazten du gehienetan; hau da, hitza horretan dagoen lema aparte beste izen bat erreferentziatzen da. Adib. *alabarena* analizatzen denean bi izen ari dira erreferentziatzen, batetik *alaba*, noski, eta bestetik berari dagokion zerbait. Kasu honetan bien informazioa mantentzea izan daiteke interesgarriena.

- Kategoria-eratorpena eta elkarketa. Eratorpen-atzizki batzuek eta zenbait elkarketak kategoria-aldaketa dakarte. Kasu honetan hitz osoaren kategoria

eratorria mantentzea bultzatu arren, zenbait aplikaziotarako, lematizatzailea adib., jatorrizko kategoria jakitea inportantea da.

Oraingo sisteman oso tratamendu sinplea egiten da irteera morfologikoa tratatzeko, UNIXeko *awk* tresnaren bidez egindako bi iragazle eskainiz:

1) Lema eta kategoria baino ez du ematen lehenengoak, baina kategoria eratorria kontutan hartuz.
   III.8 irudian azaltzen den analisia iragazle honetatik pasarazi ondoren lortzen den emaitza ondokoa da:

```
("*eta" (etA/JNT))
("gauza" (gauza/ADI)(gauzA/IZE))
("aundirik" )
("ekartzerik" (ekaR/ADI)(ekaR+tze/IZE))
("ez" (ez/ADB)(ez/IZE))
("zuen" (*edun/ADL)(zu/IOR))
("izan" (izaN/ADI))
```

   Ikus daitekeenez kategoria mailako anbiguetatea baina ez da isladatzen.

2) Bigarren iragazleak kategoria, azpikategoria eta testu-hitz batean metatutako kasu guztien arteko azkena eskaintzen ditu, kategoria eratorriaren eta elipsiaren kasuan informazioa bikoizten duela.

EUSLEM proiektuari begira (ikus §I.7) tratamendu hau hobetzeko diseinua egin da, Ritchie-ren taldeak proposatutako bidetik (Ritchie *et al.*, 92), baina horretarako lan teorikoa ari gara bukatzen.

Beste aldetik hitz anitzeko terminoen tratamendua eta anbiguetatea daukagu baina aipaturiko EUSLEM proiektuaren barruan irekitako ikerlerro bezala utzi da.