EUSKAL MORFOLOGIAREN TRATAMENDU AUTOMATIKORAKO TRESNAK

Iñaki Alegria Loinaz

Udako Euskal Unibertsitatea

ISBN: 84-8438-062-9

Argitaratzailea: UEU. Erribera 14, 1.D, 48005 Bilbo. http://www.ueu.org

URL: http://www.inguma.org/tesiak/Alegria_Loinaz_1995.pdf

LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA

EUSKAL MORFOLOGIAREN TRATAMENDU AUTOMATIKORAKO TRESNAK

Euskararako prozesadore morfologiko sendo baten diseinua eta eraikuntza. Oinarri horrekin osatutako zuzentzaile ortografikoa.

Iñaki Alegria Loinazek

Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 1995eko apirila.

LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA

EUSKAL MORFOLOGIAREN TRATAMENDU AUTOMATIKORAKO TRESNAK

Euskararako prozesadore morfologiko sendo baten diseinua eta eraikuntza. Oinarri horrekin osatutako zuzentzaile ortografikoa.

Iñaki Alegriak Xabier Artolaren eta Kepa Sarasolaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 1995eko apirila.

Euskal Unibertsitatearen alde diharduen orori

Bereziki Joserrari eta Koldori, garai zail hauetan

“Maite ditut
maite
gure bazterrak
lanbroak
izkutatzen dizkidanean
zer izkutatzen duen
ez didanean ikusten
uzten
orduan hasten bainaiz
izkutukoa ...”

(J. A. Artze)

“Hegazti errariak
pausatu dira
leihoan
argia eta itzala
bereizten diren lekuan
argia eta itzala
leihoan
pausatu dira
hegazti errariak”

(J. Sarrionaindia)

eskerrak ematen edo zorrak kitatzen

- Xuxenkide guztioi, guztion lana baita hau, adiskideok. Kepa, Xabier, Arantza, Xabier, Eneko, Montse, Nerea, Miriam, Itziar, Jose Mari, Izaskun, Koldo, Jon Mikel, Aitor, Alexander, ... norberaren zein taldearen laguntzarik gabe tesi hau ez litzateke existituko. Euskaraz egindako ikerketa aplikatuaren aldeko apustu honek aurrera jarrai dezan.

- Konputagailuen Arkitektura eta Teknologia nere Saileko lagunoi eta bereziki Olatzi eta Agusi; tesi honen arkitekturan erru handia izan duzue eta.

- “Kaxkarin” sindikatukideoi, eta bereziki tesi honen alde apustua egin zuenari, “tesi berdin krisi” leloari kontraadibidea bilatzearren. Gurekin batera kaxkarin izateko nahikoa “curriculum” egin duzuen guztioi.

- Irakaslego kontratatuaren “mugida”ko lankide eta borrokakideoi; unibertsitateko jauntxoen burugogorkeriaren aurrean, aurrera jarraitzeko emandako laguntzarengatik.

- Fakultateko garai zahar-zailetako ausarti guztioi.

- Lauri Karttunen-i eta Ken Bessley-ri lexiko-itzultzaileekin emandako laguntza eskuzabalarengatik.

- *awk* programaren egileei, lan asko aurreratzen laguntzeagatik.

AURKIBIDEA

SARRERA ETA AURKEZPEN OROKORRA

I. Lanaren nondik norakoak eta aurkezpen orokorra.

I.1. Sarrera gisako aurkezpena.

Euskararen prozesaketa automatikoan lehen urrats bat izan nahi du aurkezten dugun *Euskal morfologiaren tratamendu automatikorako tresnak* izeneko lan honek.

Euskararen prozesaketa automatikoa bultzatu eta garatzeko epe luzerako egitasmo zabal batean kokatu behar da lan hau, horretarako hizkuntzalari eta informatikarien artean osaturiko talde bat elkarlanean aritzen garelarik.

Lengoaia Naturalaren Prozesaketak, bere gorabeherak eta guzti, garapen handia izan du azken hamarkadetan, baina garapen eta aplikazio gehienak ingeleserako egin dira. Bada hizkuntz sorta bat garapen honetaz baliatu dena hein txikiago batean izanda ere, batzuetan garapen berri hauetatik ideia eta eredu interesgarri orokorrak sortu direlarik. Azkenik, tratamendu informatikotik at gelditu diren hizkuntzak dauzkagu, normalean hiztunen kopuru txikiarengatik edota ezagutza ofizial ezarengatik merkatu-interesetatik kanpo daudelako.

Euskara azken multzo honetan kokaturik zegoela ikusirik —Abaituarena (1988) zen arlo honetan aipa daitekeen lan bakarra—, eta euskal gizarteak normalizazioaren bidean halako tresnak edukitzea ezinbesteko urratsa zelakoan, Donostiako Informatika Fakultateko Lengoaia eta Sistema Informatikoen Sailean Lengoaia Naturalaren

Prozesaketarako (LNP) talde bat sortzea erabaki genuen. Aipatutako arrazoietan oinarrituz egitasmo bat planteatu genuen ondoko helburu metodologiko hauekin:

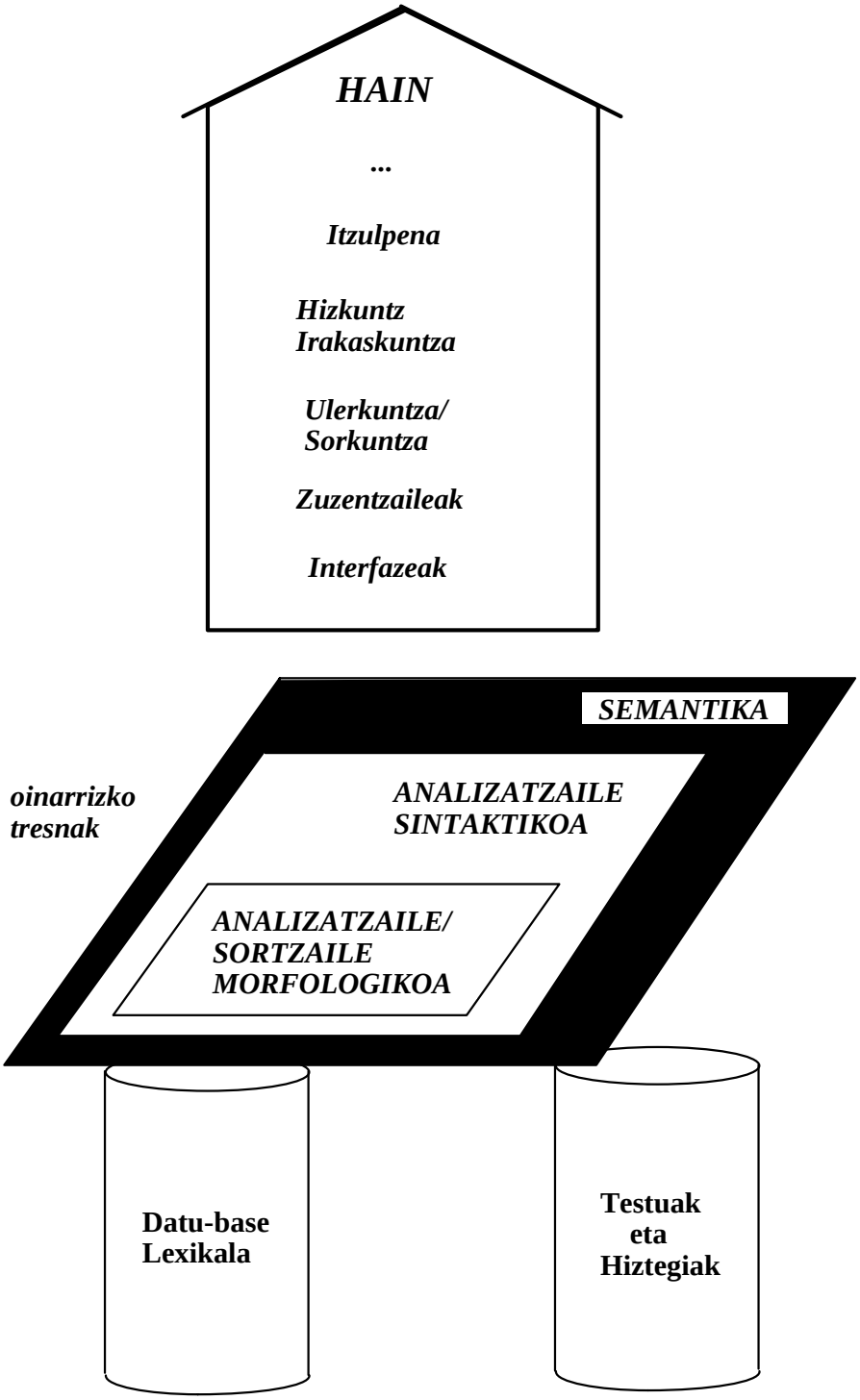
- **LNPren ikerkuntza-esparrua jorratzea**. Oinarrizko tresnetatik hasiz, oinarri sendoa osatzeko asmoz, etorkizunean helburu zabalagoetara heltzea da helburua.
- **Jakintza-arloen arteko elkarlana**. Hizkuntzaren alorra eta hizkuntza zehatz bat uztartzea teknologia informatikoaren eredu eta pentsamoldeekin, helburu bikoitza lortzeko asmoz: hizkuntzaren ezagumendua ustiatzea tresna automatikoak eraikitzeko batetik, eta teoria zein ekarpen linguistikoak egiaztatzea edota frogatzea informatikak eskaintzen dituen tresnak erabiliz. Uztartze honetan UZEI —Unibertsitate-Zerbitzuetarako Euskal Ikastetxea, terminologian eta lexikografian aritzen dena— izan da osagarri egokiena Informatika Fakultate batean sortutako talde honetarako.
- **Aplikazioa**. Garapen teorikoak baztertu gabe aplikazioa da gure lanaren zio nagusia. Hala ere, eta beste kasuetan gertatu den legez, hizkuntza berrien aplikazioan arazo berriak sortzen dira eta, hortik abiaturik, teoria eta ekarpen berriak ere.
- **Eskala erreala**. Erabakiak hartzerakoan maketen eta antzekoen erabilgarritasuna kontutan hartuz, arazo eta eskala errealeko aplikazioei erantzuten dieten sistemen eraikuntza da gure helburu nagusia.
- **Berrerabilgarritasuna**. Burutzen diren aplikazioak berrerabilgarriak izan daitezela zentzu bikoitzean: batetik, aplikazio horien gainean eta ondoko urratsetan aplikazio konplexuago eta osotuagoak eraiki ahal izatea, eta bestetik, irekiak izatea aplikazio hauek beste erabiltzaileen esku jarriz.
- **Corpus idatzietan oinarrituta** baina **arauak errespetatuz**, eta corpusekin egiaztatuta eta neurtuta. Hau da, Euskaltzaindiak eta beste batzuek sortutako arauak eta teoriak kontuan hartzen dira, baina sistemen baliagarritasuna hizkuntzaren erabilera errealarekin alderatuz neurtu behar da.

Aipatutako ezaugarri horiek egitasmo osorako pentsaturik badira ere, hemen aurkezten den lanari aplika dakizkioke ere banan banan. Aurkezten dugun lanean bi tresna diseinatu dira prozesadore morfologikoa eta zuzentzaile ortografikoa, eskala errealekoak eta berrerabilgarriak biak, arau eta ezagutza morfologikoan oinarrituak, baina corpusekin egiaztatuak.

Tresnen nondik-norakoak azaldu baino lehen, egitasmo orokorraren barruan duen kokapena eta euskararen ezaugarri garrantzitsuenak azalduko ditugu.

I.2. Hizkuntzaren prozesaketa automatikoaren oinarria eta aplikazioak. Proiektuaren helburuak.

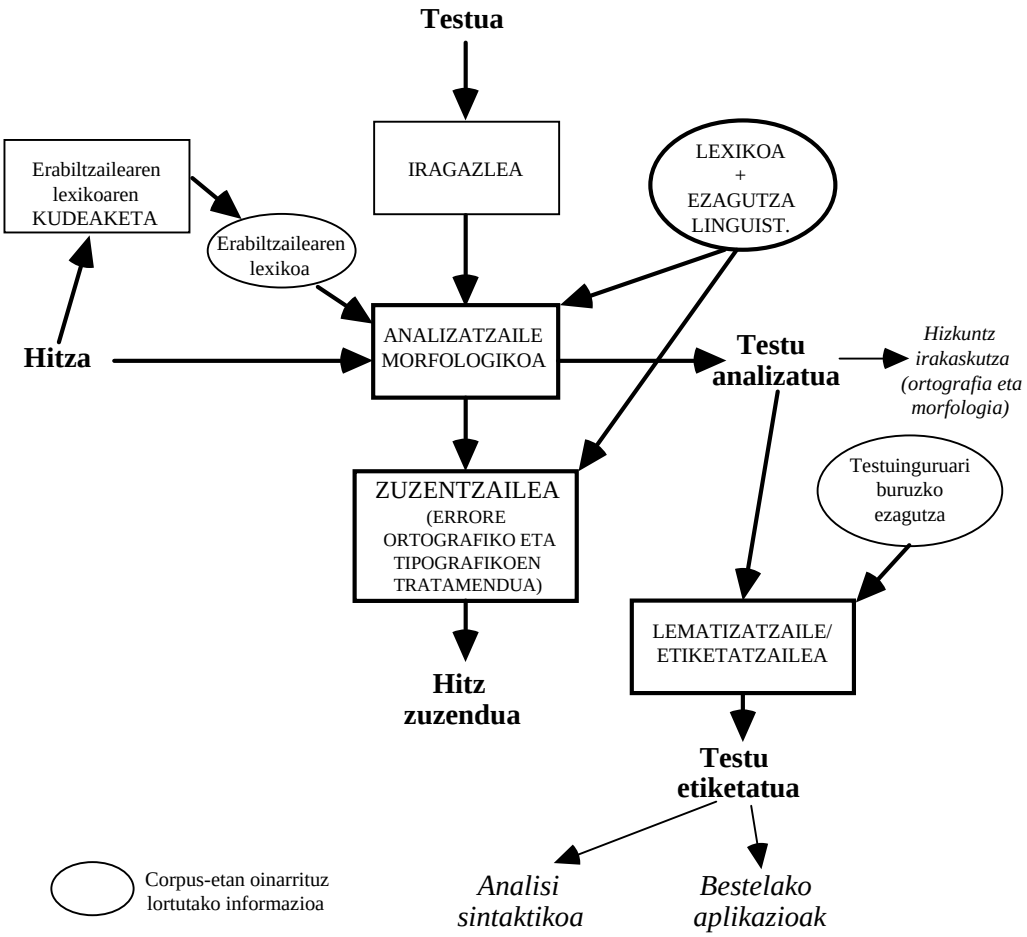
I.1 irudian gure ikertaldean euskara idatziaren prozesaketa automatikorako ezarri ditugun urratsak eta maila desberdinak irudikatzen dira, etxe baten eraikuntza simulatuz.



I.1 irudia.- LNPrako urratsak eta mailak etxe baten eraikuntza irudikatuz.

Prozesagarriak diren datu-base lexikala, testuak edo corpusak, eta hiztegiak dira etxeko zimenduak. Funtsezko informazio horiek ez baditugu, ezinezkoa izango da tresnak eraikitzen hastea, zeren tresna horiek errealitateari egokituak eta kalitatekoak izan daitezen, ezinbesteko oinarri eta erreferentzia baitira. Zimendu horiek bideratzen dute etxeko zorua eta egitura eraikitzea. Morfologia, sintaxia eta semantika dira garrantzi

handieneko lanbideak LNPrako sistemetan; zuzenean aplikazio komertzializagarriak ez izan arren, tresna komertzial gehienen funtsa baitira. Zimenduak, zorua eta egitura osoa buruturik dagoenean, etxeko paretak eta teilatua diren produktu erabilgarriak egitea lan gogorra da baina beti ziurra, oinarria ondo jarrita baitago. Produktuak aipatzean honako hauek azpimarra daitezke: zuzentzaileak, lematizatzaileak, lengoaia naturalaren bidezko interfazeak, testuen sorrera eta ulermena, ordenadorearen bidezko hizkuntz irakaskuntza, itzultzaile automatikoak edo semiautomatikoak, etab. Etxeko elementu guztiak kudeatzeko eta ustiatzeko ingurune bat ere aurrikusten da, HAIN —Hizkuntz Aplikazioetarako INgurunea— izenekoa.



I.2 irudia.- Aurkezten den lanaren eskema orokorra.

Dena den aurkeztu den egitasmo orokorrean esandakotik ez da ondorioztatu behar edozein produktu egiten hasi baino lehen zimenduek, zoruak eta egiturak erabat bukatuta egon behar dutenik, baina bai produktu bakoitzari dagokion oinarriari merezi duen garrantzia eman behar zaiola.

Lan honetan azaldutako egitasmoaren atal bat aurkezten da, morfologiaren inguruan dagoena hain zuzen ere. Morfologia lantzeko bidean EDBL izeneko datu-base lexikal bat

prestatu da, eta hizkuntzari buruzko ezagumendua lortzeko eta emaitzak ebaluatzeko testu-multzo bat lortu eta erabili ere. Eraikitako prozesadore morfologikoa erabiliz zuzentzaile ortografiko bat egin da, eta lematizatzaile/etiketatzaile bat proposatzen da. Prozesaketaren unitatea hitza denez, token-ezagutzailea edo iragazlea oso elementu garrantzitsua da.

Lanaren eskema orokorra I.2 irudian ikus daiteke.

I.3. Euskararen ezaugarriak modu laburrean.

Euskara da Europako hizkuntzen artean zaharrena, hizkuntza indoeuroparrak iritsi baino lehenago Europan zeudenen artean gelditu den bakarra baita. Teoria desberdinak badaude ere, bere jatorria zeharo egiaztatu gabe dago gaur egun.

Gaur egun hiztunak 650.000 inguru dira, Euskal Herriko populazioaren laurdena baino gutxixeago. Lurraldeei dagokienean, azken mendeetako murrizte-prozesua gelditzeko zorian dago Hego Euskal Herrian, baina ez Iparrean.

Ofizialtasuna du Araban, Bizkaian eta Gipuzkoan eta koofiziala da Nafarroan. Lapurdin Behe Nafarroan eta Zuberoan ez du ofizialtasun-aitorpenik.

Dialektoei dagokienean, oso hizkuntza aberatsa da zortzi euskalki bereizten dira eta. Aberastasun hori eta tradizio idatzi murritza dela eta, orain dela gutxi arte ez dira batasunerako urrats eraginkorrak eman. 1968an, zeuden kezkak eta saioak ikusita, Euskaltzaindiak bultzatu zuen euskara idatziaren batasuna, erabat arrakastatsua gertatu dena eta oraindik osatzen ari dena.

Morfologiaren aldetik honako ezaugarri hauek azpimarra daitezke:

- Oso flexio aberatsa, hamalau kasu desberdinekin, generorik gabe eta singularra eta pluralaz gain mugagabea ere bereiziz. Flexioa amankomuna da izen eta adjektiboetarako, oro har. Hizkuntza eranskaria da, askotan ezaugarri morfologiko bakoitzari morfema edo hizki bat dagokio eta. Horrela zenbait atzizkiren atzean atzizki gehiago metatu daitezke.
- Ergatiboa, kasu hau ez da hizkuntza indoeuroparretan agertzen.
- Aditza oso aberatsa da, aberastasun hau aditz laguntzailean eta trinkoan ere agertzen dela, eta forma bakar batean hiru pertsona-marketaraino irits daitekeela. Euskarak egiten duen genero-banaketa bakarra aditzean aurki daiteke, bigarren pertsona hurbilaren tratamenduan.

Egin dugun lanak ekarpen bat izan nahi du batasunaren bide horretan; helburu horrekin analizatzaile morfologiko estandarrak ez ditu ezagutuko forma estandartzat hartu ez diren hitz asko. Beraz, batasunaren aldeko apustu horrek oinarrizko analizatzailearen estalduran —ezagutzen eta analizatzen diren hitzen portzentaia— ondorio negatiboak ekarriko ditu. Hala ere analizatzaile estandarra baino harantzago doan analizatzaile hedatuaren bidez (ikus laugarren kapitulua), euskarazko hitz ez-estandar asko ezagutzeko aukera dago. Oinarri-lan honen fruitu gisa sortutako eta merkaturatutako ordenadore pertsonaletarako zuzentzaile ortografikoa batasunerako oso tresna baliagarria delakoan gaude.

Gure lanari ekiteko garaian izan ditugun oztopo nagusiak bi izan dira: batetik morfologiari buruzko lan sistematikoen falta, eta bestetik, batasunarekin loturiko gatazkak, irizpide finkoak ez zeudelako edo aldatuz joan direlako. Izan ere, azken urteetan euskararen inguruan egindako hainbat lan —gramatikak, hiztegiak, ikerketa-lanak, etab.— behar-beharrezko laguntza izan dira gure proiektuan.

I.4. Zimenduak: Euskararako Datu-Base Lexikala eta corpusak.

Proiektu hau bideratzeko, eta etorkizuneko aplikazioetarako datuak biltzeko funtsezko osagaiak dira bi hauek.

I.4.1. Euskararako Datu-Base Lexikala (EDBL).

Euskararako Datu-Base Lexikala (EDBL) funtsezko ezagumendu-oinarria da Lengoaia Naturaleko Prozesaketaren arlo askotan, eta bereziki morfologiaren alorrean. Eskala errealeko proiektu erreal bati ekitean pentsaezina da dimentsio errealeko informazioa testu arruntetan edo fitxategi konbentzionaletan biltegiratzea, eta datu-basea da dudarik gabe dagokion errepresentazio-sistema. EDBLk euskararen tratamendu automatikorako datu-base lexikal orokor bat izan nahi du, eta horrexegatik bertan mota guztietako informazioak biltzen dira, morfologikoak, sintaktikoak eta semantikoak.

Hasiera batean morfologiarekin lotutako proiekturako erabili denez, informazio morfosintaktikoari bultzada handia eman zaio, semantikari buruzko datuak gerorako utziz. Jakintza-arloen arteko talde-lana izatean, datu-basearen eguneratzea, zuzenketa eta mantenua linguisten zeregina izan den bitartean, informatikariena izan da datu-basearen eta intefazearen diseinua, esportaziorako prozeduren idazketa eta integritate- zein osotasun-egiaztapenerako murriztapenen definizioa.

Datu-basean informazio anitz metatzen da, baina inportanteena informazio lexikala dugu. Hirurogei mila sarreratik gertu daude erabat landuta, eta beste asko guztira osatu gabe. Sarrera bakoitzeari dagokion informazioa honako multzo hauetan bil daiteke:

- forma kanonikoa
- bi mailatako forma (morfologian erabiliko den ereduari egokitua)
- itsats dakizkiokeen morfemei buruzko informazioa
- erabilpenaren adibide bat
- kategoria, azpikategoria eta aditz-mota
- flexioari buruzko informazioa: kasua, zenbakia, mugatasuna, erlazioa, modu/denbora, pertsona
- kategoria erantsia
- iturburua, oharrak eta zalantzak
- maiztasuna (Sarasolaren maiztasun-hiztegiaren arabera, 1982)
- eguneratze-data eta berau egin duen hizkuntzalaria

Datu-basearen diseinuan eredu erlazionalari jarraitzen zaio, baina etorkizunerako, objektuei zuzendutako diseinu berri baten gainean ari gara lanean, gorde behar den informazioak duen konplexutasunari ondo erantzun ahal izateko, bertan lokuzioak, hitz anitzeko terminoak etab. biltegiratu nahi ditugu eta. Eredu berri horretan saihestuko da gaur egun datu-baseak bi mailatako morfologiarekin duen menpekotasuna, morfologia-formalismotik independentea bihurtuz. Datu-base honi buruz zehaztasun gehiago azaltzen dira hirugarren kapituluan.

I.4.2. Corpusak.

Testuek edo corpusek ematen dute benetan erabiltzen den hizkuntza idatziaren neurria. Gaur egun beren erabilpena areagotu egin da arrazoi horregatik, baita erregela-sistemen aurrean corpusetan oinarritutakoek duten eraginkortasuna eta sendotasunagatik ere. Leech-ek esaten duen bezala (Garside *et al.* 87: 3), corpusetan oinarritutako hurbilpenek eta erregeletan oinarrituek ezaugarri osagarriak dituzte, eta beraz osagarriak dira:

> ... The strength of the corpus-based approach is that, through probabilistic predictions, it is able to deal with any kind of English language text which is presented to it: it is eminently robust. Its weakness is that the very reliance on probability admits the possibility of error. The probabilistic system makes the best “guess” available to it, based on textual material that has been analysed in the past.
>
> This combination of strength and weakness is the exact opposite of the AI-based system which assumes (...) that 100% successful processing is possible, but which falls short of the ability to deal with uncensored, unrestricted text.
>
> We would argue that the two approaches are complementary: ...

Sistemen zehaztasuna neurtzeko erabilpenaz gain, beste zereginetan ere erabiltzen dira; sistemak azkartzea helburua duten maiztasun handieneko osagaien *buffer*-ak eta etiketatze-lanetan erabiltzen diren Bayes-en ereduak eta eredu markoviarrak dira corpusen erabilpen ezagunenak ezagumendu-iturri gisa. Izan ere, gaur egun corpus eleanitzak ere proposatzen dira haien hasierako eremutik kanpo ziruditen aplikazioetarako ere, haien artean itzulpenarena azpimarra daitekeela.

Corpusen artean ondoko sailkapen sinplea egin daiteke:

- **Orekatuak/ez-orekatuak**. Orekatuetan testu-moten artean halako oreka bat bilatzen da, testu-mota berezituei dagozkien ezaugarri partikularretatik aldenduz. Horretarako, iturburu desberdinetatik testu-zati txiki samar anitz, esanguratsuak eta aberasgarriak biltzen dira, teknika estatistikoak erabiliz. Corpus orekatuak ezinbestekoak dira ezagumendu-iturri gisa; besteek, ez-orekatuek hain zuzen ere, zehaztasuna neurtzeko bakarrik balio dute, helburu bereziturako sistemen eraikuntzan ez bada behinik behin.
- **Etiketatuak/etiketatu gabeak**. Gehienak etiketatu gabeak badira ere, ugaltzen ari da corpus etiketatuen eskaintza, ingeleserako behintzat. Etiketatuetan, aurreprozesaketa batez —eskuzkoa, semiautomatikoa edo automatikoa izan daitekeena— testuak zuen informazioaz gain beste datu batzuk gehitzen dira, zenbait erabilera erraztearren.

Testuak	**Ezaug.**	**hitzak**	**hitz kopurua**	**agerpen kopurua**[1]
1.- Argia aldizkaria (zatiak)	ez-orek.	4.864	2.607	1,86
2.- Filosofiari buruzko artikulua	ez-orek.	2.343	1.429	1,64
3.- EEBSko azken urteak	orekatua	23.364	9.313	2,51
4.- EEBS estandarra	orekatua	396.840	67.816	5,85

I.3 irudia.- Lanean zehar erabilitako zenbait testuren neurriak.

Gure kasuan, testu ez-orekatu batzuez gain, UZEIrekin izandako elkarlanari esker, EEBS proiektutik (Urkia & Sagarna, 91) banandutako corpus orekatu bat eskuratu dugu

[1] Forma bakoitzeko batez-beste agerpen kopurua.

euskarri prozesagarrian, honek lana izugarri erraztu digularik. I.3 irudian erabili ditugun corpus batzuen neurriak agertzen dira.

Corpus orekatu orokorrari “EEBS estandarra” deituko diogu lan honetan zehar, eta bere ustiapenerako zenbait arazo egon dira. Arazo garrantzitsuena hauxe izan da: euskara estandarrerako corpus orekatua zen helburua, baina EEBSn hogeigarren mendeko euskara idatziaren mota guztietako laginak daude; eta data, testu-mota eta euskalkiaren arabera sailkaturik egon arren, euskara batuaren garaiko testu neutroak —euskalkiaren aldetik— aukeratu arren, erabilpen ez-estandarrak agertzen dira maiz, euskara estandarraren arauak eta irizpideak aldatzen ari baitira. Honen ondorioz aukeratutako corpus orekatuan forma ez-estandar anitz agertzen dira, haien arteko batzuk maiztasun handiz. Euskara batua/estandarra bultzatzeko tresnak eraiki nahian, ezin izan diogu eman corpus orekatu honi beste hizkuntza normalizatuagoetan ematen zaion garrantzia; finkatzen ari diren irizpide batzuk corpusetan agertzen diren datuekin kontraesanean daudenean, irizpide horiei lehentasuna eman baitiegu.

Corpus orekatuan oinarriturik bi taula garrantzitsu lortu dira: maiztasun handieneko hitzena, eta maiztasun handieneko trigramena —hiru karaktereko multzo gainjarriak aurreko eta ondorengo zuriuneak kontuan harturik—.

I.5. Egiturazko tresna: prozesadore morfologiko automatikoa.

Prozesadore morfologiko baten eraikuntza eta beraren erabilpena beste tresnak diseinatzeko izan da lan honen muina. Konputagailuaren bidezko morfologiari ekin aurretik eredu desberdinak aztertu dira eta zenbait proba egin ere. Bide horretan, eta burutzapenaren lehen fase batean, bi “maketa” eraiki ziren bi formalismo desberdinen arabera: bi mailatako formalismoari (Koskenniemi, 83) jarraituz bat, eta ATEF sistema (GETA, 82) erabiliz bestea. Bibliografiatik ateratako ondorioak eta esperientzia praktikoetatik ateratakoak bat etorri ziren, eta bi mailatako morfologia izan zen aukeratu genuen eredu konputazionala.

Euskararako prozesadore morfologikoaren eraikuntza bi fasetan izan da burutua: euskara estandarrerako prozesadore morfologikoa batetik, eta aurreko prozesadore morfologikoak ezagutzen duen hitz-multzoa —*coverage* edo estaldura-tasa— handitzen duen “analizatzaile sendoa” bestetik.

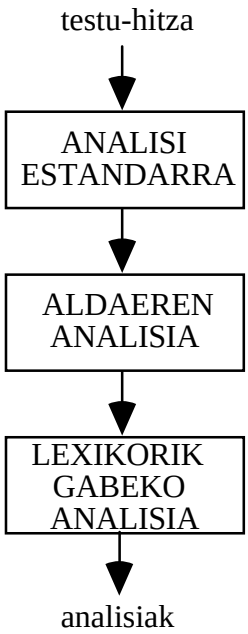
Bi faseetan erabili diren teknikak bi mailatako morfologian (Koskenniemi, 83) daude oinarrituta, eta horri esker sistema osoa homogenoa da irtenbide partikularretatik aldenduz. Hiru hobekuntza burutu dira bi mailatako formalismoaren inguruan: lehenengoz erabiltzaileen lexikoen erabilera bideratu da, bigarrenik bi mailatako paradigmaren

erabilpen “berri” bat egin da, aldaera deitu ditugun forma ez-estandarren tratamendurako; eta azkenik fonologiarako bakarrik erabilia zen “lexikorik gabeko analisia” testuen analisirako izan da hedatua.

Bi mailatako morfologiari jarraituz, euskara estandarraren morfologia deskribatzeko definitu dira oinarrizko bi osagaiak: morfemak eta haien arteko loturak zehazten dituen lexikoa batetik, eta morfemak biltzen direnean gertatzen diren aldaketa (morfo)fonologikoak deskribatzen dituzten erregelak.

Bi mailatako morfologiaren eredu klasikoa ez da nahikoa morfotaktikaren barruan kokatzen den gertakizun bat, urruneko menpekotasuna deitutakoa hain zuzen ere, modu egokian adierazteko. Euskararen morfologian urruneko menpekotasuneko kasu arrunt batzuk daude, eta horiek modu egokiagoan adierazteko proposamen bat egin dugu: jarraitze-klase hedatuak.

Dugun lexikoarekin, eta normalizatzen ari den hizkuntza batean hizkuntza estandarrera mugatzearen ondorioz, probatu diren testuetako %90 hitz inguru ezagutzen dira lehen hurbilpenaren bidez. Honen aurrean, eta emaitza hauek osatzearren, lehen aipatu diren hobekuntzak proposatzen dira prozesadore morfologikoa sendotzeko: erabiltzailearen lexikoen edo lexiko berezituen kudeaketa, forma ez-estandarrei dagozkien aldaeren tratamendua, eta analisia lema lexikoan egon gabe.



I.4 irudia.- Analisi morfologikoaren urrats desberdinak.

Lexiko orokorrean ez dauden lemak erabiltzailearen lexikoetan gorde daitezke; horretarako azpilexiko irekien eta itxien artean bereizketa egin delarik. Honen helburua

zera da: ezagututako hitzen kopurua handitzea, askotan termino teknikoak edo pertsona-zein leku-izenak ez baitaude jasoak lexiko orokorrean. Erabiltzaileak, horrela, aberats dezake lexikoa bere beharretara egokituz.

Aldaeren tratamendua funtsezkoa da hain batze-bide laburra duen hizkuntza baterako. Aldaerak, bi mailatako morfologiaz kudeatzen direnez, bi multzotan banatu ditugu: oso orokorrak direlako erregela morfofonologikoen bitartez adieraz daitezkeenak batetik, eta morfema zehatzei dagozkielako lexikoan adierazten direnak bestetik. Tratamendu honen bidez analizatzailearen estaldura-tasa hobetzeaz gain, forma ez-estandarrei dagozkien estandarrak lor daitezke, prozedura hau zuzenketan eta ordenadorez lagunduriko irakaskuntzan aplikazio zuzenekoa izanik.

Aurreko metodoez hitz bat analizatzerik ez dagoenean, analizatzaile morfologiko sendo bat lortzeko behinik behin, analisia lortzeko bideren bat bilatu behar da. Gure ebazpideak, bi mailatako formalismo barruan kokatzen denak, lemarik gabeko lexiko txiki bat erabiltzen du fonologiarako erabilitako metodo bati (Black et al., 91) jarraituz. Prozesu honi “lexikorik gabeko analisia” deitu diogu eta aipaturiko azpilexikoaz gain bi mailatako erregela berezi pare bat erabiltzen du.

Tratamendu-multzo horrekin aberasturiko analizatzaileak honako ezaugarriak ditu:

- Orokorra: euskara estandarraren forma gehienak analizatzeko eta sortzeko gai.
- Malgua: erabiltzailearen lexikoek eta aldaeren tratamenduak bideratzen dute ez-orokorrak edo ez-estandarrak diren formen ezagutza, prozesadore morfologikoari malgutasuna emanez.
- Sendoa: Lexikorik gabeko lematizazioari esker beste urratsetan ezagutzen ez ziren hitzen analisia bideratzen da, sistemari sendotasun handiagoa emanez.

Deskribatutako prozesadore morfologiko hau oinarria da eraiki dugun Xuxen izeneko egiaztatzaile-zuzentzaile ortografikorako, garatzen ari garen EUSLEM izeneko euskararako lematizatzaile/etiketatzaile orokorrerako eta etorkizun hurbilerako helburu dugun analizatzaile sintaktikorako.

I.6. Prozesaketa morfologikoa hobetzen: Lexiko-itzultzaileak.

Bi mailatako morfologiaren arrakasta izugarria izan da, eta gure proiektua aurrera joan den bitartean beste talde batzuk haren inguruan hobekuntzak burutzen joan dira.

Hobekuntza horien artean azpimarratzekoa da *lexiko-itzultzaile* izenarekin ezagutarazi dena, Xerox-en garatua izan dena. Oinarri teorikoa (Karttunen *et al*., 92) eta aplikazio

praktikoa (Karttunen, 94) azken bi urteetan eman dira aditzera eta ekarri dituzten hobekuntzak bi arlotan bana daitezke:

- Eraginkortasunaren aldetik, lexikoa eta erregelak automata bakar batean biltzean, automata horren optimizazioari esker lortzen da abiadura handitzea oso modu garrantzitsuan.
- Deskribapen-ahalmenaren aldetik, bi mailatako morfologiaren erregela paraleloen abantailak mantendu arren, erregela paraleloen multzo desberdinen arteko konposaketa sekuentziala bideratzen dute, deskribapen ahalmena handitu eta deskribapena bera erraztuz.

Aldaketa garrantzitsu honen aurrean, eta diseinatutako tresnak erabiltzeko eman diguten aukeraz baliatuz, tresna berri hauen aplikazioa eta baliagarritasuna ebaluatu dugu, baina ez bakarrik analisi estandarrerako, baizik eta bi mailatako morfologiari buruz guk egindako aldaketen eta proposamenen gainean ere lexiko-itzultzaileek joka dezaketen papera ikertu dugu. Horretarako gure inplementaziorako geneuzkan datuak egokitu ditugu eta beraiekin euskara bezalako hizkuntza eranskari bati dagokion sistema erreal baterako aplikazioa aztertu dugu. Tratamendu gehienetarako hobekuntzak besterik ekartzen ez badituzte ere, zenbait muga igarri ditugu beraiengan, hirugarren eta laugarren kapituluetan ikus daitekeenez.

I.7. Produktu komertziala: Xuxen zuzentzaile ortografikoa.

Euskararako zuzentzaile ortografiko bat burutzeko ideia taldearen helburu nagusien artean zegoen hasiera hasieratik. Arrazoi nagusiak honako hauek ziren:

- Inguruko beste hizkuntzetarako eskuragarri zen aipaturiko produktu hori, baina ez euskararako.
- Helburu nagusien artean aipatu diren *aplikazioa* eta *eskala erreala* irizpideekin bat zetorren bete-betean.
- Euskarak bizi duen batasun-prozesurako are garrantzi handiago du halako tresna batek. Zentzu berean, gertatzen diren idazketa-erroreetan batasuna erabat finkatu gabe egoteak problematika berria eta aberatsa dakar, ikergai erakargarria bihurtuz.

Euskararen ezaugarriek, flexio aberatsa eta eranskaria edukitzeak, zuzentzailea morfologian oinarritzera eraman gintuen ezinbestean, hitzak onartuz banaketa morfologiko posiblea baldin badute. Bibliografian agertzen ziren erreferentzia gehienak ez

dira oso baliagarriak, euskara bezalako hizkuntza eranskarietan zuzenketa-prozesua korapilatsuagoa da eta.

Aipatutako testuinguruan, zuzenketari ekin aurretik ondoko erizpide hauek geneuzkan buruan:

1) Testuingurua kontuan hartzen duen zuzentzailearena —bigarren belaunaldiko zuzentzaileak edo estilo-zuzentzaileak ere deituak— proiektu erakargarria izan arren, lehen urrats batean zuzentzaile konbentzional batera murriztu ginen, hartarako behar ziren beste oinarriak —analisi sintaktikoa etab.— egin gabe daudelako.

2) Batasunarekiko zalantzak sortutako erroreak —orokorrean gaitasun-erroreak edo aldaerak deituko ditugunak— ziren tratatzeko lehentasuna zutenak, gainontzekoen tratamendua —errore tipografiko deitutakoena— bigarren maila batean utziz.

Ondorioz, zuzenketa bi modulu osagarriren bidez burutzen da, gaitasun-erroreena batetik eta errore tipografikoena bestetik; eta lehen motako erroreak tratatzeko bi mailatako morfologian oinarritutako metodo berritzaile batera iritsi garen bitartean, errore tipografikoak tratatzeko prozedura klasiko bat erabiltzen dugu, azkartzeko bideetan zenbait ekarpen egin badira ere.

Zuzenketa-aplikazioetan ohizko diren beste moduluez gain, iragazlea adibidez, erabiltzailearen hiztegiarekin ekimen berezi bat egin da, hizkuntzalari ez den erabiltzaile bati informazio morfosintaktikoa eskatzeko modua sakonean aztertu da, informazio horrekin sistemak hitz berri baten flexio guztiak ezagutuko baititu.

I.8. Hurrengo urratsa: EUSLEM.

Aurkezten den lanean hitza da tratamendu-unitatea. Tesi hau mugatzeko orduan, testuingurua kontuan hartzen duen oro kanpoan utzi dugu, arlo horretan lanean ari bagara ere.

EUSLEM diseinatu den eta gauzatze-bidean dagoen lematizatzaile/etiketatzailea da, euskararako eta bi mailatako analisi morfologikoan oinarrituta. Diseinu-filosofia orain arte azaldutako bera da: eskala erreala, aplikagarritasuna, garatutako beste tresnen berrerabilpena garrantzitsuenak izanik. Horrez gain lematizatzaile/etiketatzailea aplikazio konkretutik independente diseinatu da, helburu desberdinekin erabili ahal izateko. Tresna honen oinarrizko osagaiak hauek dira:

- *Token*-ezagutzailea deitzen den aurreprozesadorea, hitzak, puntuazio-karaktereak, zenbakiak etab. identifikatzeko. Analisi morfologikorako egindakoa erabiliko da aldaketa gutxi batzuekin.
- Analizatzaile morfologikoa, hitzei dagozkien lema eta etiketa posibleak zehazteko. Egindakoa berrerabiliko da.
- Hitz ezezagunen etiketatzailea edo *guesser*-a, analizatzaile morfologikoak ezagutzen ez dituen hitzen lema eta etiketa hipotetikoak lortzeko. Egindako aldaeren analisia eta lexikorik gabeko analisia erabil daitezke helburu horrekin.
- Etiketen definizioa eta analisi morfologikoarekiko egokitzapena.
- Hitz anitzeko terminoen identifikazioa, horien tartean lokuzioak, hitz-elkarketa eta bestelako kasu asko sartzen direlarik.
- Testuinguruan oinarritutako desanbiguazio, metodo estokastiko, linguistiko edo bion konbinaketaren bidez egina.

Esan bezala, hitza baino harantzago doazen tratamenduak lan honen esparrutik kanpo gelditzen dira; beraz, aplikazio hau irekitako ikerlerro gisa aurkezten da.

I.9. Egindakoaren aplikazio berri posibleak.

Lan honetan aurkezten diren tresnek morfologia dute oinarritzat; hala ere prozesadore morfologiko batean oinarriturik egin daitezkeen aplikazioak askoz gehiago dira. Honako hauek dira garrantzitsuenak:

- Esan den bezala lematizazioa analisi morfologikoan oinarritzen da, eta lematizazioa funtsezko tresna da lan lexikografikoetan nahiz informazioa biltegiratzen eta berreskuratzen laguntzen duten sistemetan, dokumentuen datu-baseak adibidez.
- Hizkuntz aplikazio sakonagoetarako —sintaxia, itzulpen automatikoa, etab.— lehen urrats gisa.
- Hizketaren sintesia edo testu-sorkuntza lortzeko sorkuntza morfologikoa funtsezko osagarria da. Hizketaren kasuan, ahoskatzeko testua eduki arren, ahoskatzeko orduan informazio morfologikoa garrantzizkoa izan daiteke.
- Itzulpenerako laguntza-tresnak. Hiztegi elebidun baten laguntzaz iturburu-testu batetik abiatzen bagara, hiztegian adierak bilatzeko jatorrizko testuaren lema posibleak lortu behar dira hiztegian bilatu ahal izateko. Gaur egun gorantz doan

ikerlerroa den testu-parekatzean analisi morfologikoak ere funtsezko funtzioa eduki dezake.

- Beste aplikazioak. Zaila izango litzateke aplikazio guztiak banan-banan zerrendatzea. Hona hemen gure sistemarekin buruturiko bat: koherentzi egiaztatzea. Entziklopedia-hiztegi batean zenbait sarrera kendu behar ziren baina hauek kentzean testuak kontsistentzia gal zezakeen, definizioetan sarrera horiek edo berauen flexioak ager baitzitezkeen. Honen aurrean, eta lexiko-sarrerak arruntak ez zirenez, sarrera horiek eta flexio-hizkiek osatutako lexiko bat osatu genuen, eta testua analizatzean analisirik lortzen zuten hitzak baztertzekoak ziren, kendutako sarreren forma flexionatuak baitziren.

Zuzentzaile ortografikoarenak, berriz, hauexek lirateke testu-edizioaz gain:

- Ordenadorez Lagunduriko Irakaskuntzako sistemetan (OLI), morfologia eta ortografia irakatsi eta zuzentzeko.
- OCR dispositiboen bidez jasotako euskarazko testuen zuzenketa.
- Hizketaren analisiaren emaitza zuzentzea. Zuzentzailea egokitu beharko litzateke, testu idatzian eta hizketan agertzen diren hizkuntzen ezaugarriak desberdinak dira eta. Gainera hizketaren kasuan, OCRan bezala, zuzenketa automatikoa behar da.
- Pertsona-makina elkarrekintza aplikazio orotarako, pertsonak barneratzen duen informazioan akatsak egon daitezkeela suposatzen bada behintzat. Datu-baseen zein entziklopedien kontsulta-sistema lokalak zein sarearen bidezkoak, eta elbarritu edo behar bereziak dituzten pertsonekiko komunikazio-sistemak sartzen dira multzo honetan.

I.10. Txostenaren eskema.

Ondoan azaltzen den txostena lau partetan dago banaturik.

Lehendabizikoan euskararako prozesadore morfologikoaren diseinua azaltzen da hiru kapitulutan banatuta. II. kapituluan morfologiaren oinarri minimoak azaldu eta gero, morfologiaren tratamendurako eredu konputazionalen azterketa egiten da, egoera finituko ereduetan eta, batez ere, bi mailatako morfologian sakonduz. Urruneko menpekotasunak ebazteko gure ekarpena den “jarraitze-klase hedatuak” izeneko mekanismoa ere azaltzen da bertan. III. kapituluan bi mailatako morfologiaren aplikazioa den euskara estandarrerako prozesadore morfologikoaren diseinu eta gauzatzea azaltzen da, lexiko-

itzultzaileen bidez ere egiten dena. IV. kapituluan azkenik, aurreko analizatzaile morfologikoa estaldura handiko eta sendo bihurtzeko urratsak azaltzen dira, bertan bi mailatako paradigmari jarraitzen dioten erabiltzailearen lexikoen kudeaketa, aldaeren ezagutza eta lexikorik gabeko analisia adieraziz.

Bigarren parteak zuzenketa du oinarritzat eta bi kapitulutan dago banaturik. V. kapituluan zuzenketaren nondik-norakoak azaltzen dira flexio handiko hizkuntzak eta hizkuntza eranskariak zuzentzeko dauden arazoetan sakonduz. Era berean ez-jakiteak bultzatutako erroreak, gaitasun-erroreak deitutakoak, tratatzeko garrantzia ere azpimarratzen da. VI. kapituluan morfologian oinarritutako euskararako zuzentzaile ortografiko baten diseinu eta gauzatzea deskribatzen da, morfologiarako garatutako hainbat tresna berrerabiltzen direlarik.

Hirugarren partean berriz, egindako lanaren ondorioak eta etorkizuna ditugu hizpide. Bertan azalduko dira egindako lanaren alde azpimarragarrienak, EUSLEM izeneko lematizatzaile/etiketatzailean egindako lanak duen tokia, eta lan honen ondorioz aurkitutako ikergai interesgarrienak.

Azkenik, erabilitako bibliografia gaika azalduta, eta testuan zehar proposatutako eranskinak gehitzen dira.

LEHEN PARTEA: ANALISI MORFOLOGIKOA

II. Egoera finituko morfologiaren inguruan.

Morfologiarako eredu konputazional desberdinen aurkezpena egitea eta horien barruan bi mailatako morfologiarena kokatzea eta sakontzea da bigarren kapitulu honen xede nagusia.

Formalismo morfologikoak aztertu baino lehen berauek ezagutzen lagunduko diguten zenbait funtsezko ezaugarri aurkezten dira hasteko. Ezaugarri horietan oinarrituz sailkapen bat proposatzen da, klasifikazio honen barruan literaturan agertzen diren zenbait sistema kokatuz. Sistema hauetako batzuen nondik-norakoak azaldu eta gero guk erabiliko dugun bi mailatako formalismoa azaltzen da zehaztasun handiagoz.

Bi mailatako morfologia izeneko formalismoaren osagaiak banan-banan aztertuz, deskribapen-ahalmenaren aldetik bere aldeko eta aurkako irizpideak zehazten dira, eta ondorioz, bere puntu ahulenaren gainean, morfotaktikarenean hain zuzen, proposatutako hobekuntzak aztertzeaz gain gure proposamenaren berri ematen da. *Jarraitze-klase hedatuak* deitu dugun mekanismo honek, jarraitze-klase arruntei leporatutako arazo guztiak konpontzen ez baditu ere, euskararako aplikazio zuzena duen funtsezko bat bideratzen du, morfemen arteko urruneko menpekotasunarena hain zuzen.

Azkenik, bi mailatako morfologiaren ereduaren konputazio-konplexutasuna aztertzen da, honek sistema errealetan erabiltzeko bideragarritasuna adierazten baitu. Gaia nahikoa polemikoa izan da eta formalismo hau euskarari egokitu izanak honetaz erakutsi diguna ere azalduko da.

Kapitulu honetan gida gisa erabili dugun liburua, R. Sproat-en Morphology and Computation (1992), gai honetan gehiago sakontzeko oso baliagarria da.

II.1 Analisi morfologikoa: sarrera gisakoa.

Morfologia teorikoan sakontzeko batere asmorik gabe, ondoren azalduko diren eredu konputazionalak alderatu eta sailkatu ahal izateko beharrezkoa da morfologiaren kontzeptu orokorrak gainbegiratzea.

Aurretik aipatutako liburuan (Sproat, 92:17) egiten diren galderak izan daitezke kontzeptu horiek azaltzeko iturburua:

> "In particular, I shall discuss the following issues:
> What sort of things can morphology mark in different languages?
> How are words built up from smaller meaningful units-morphemes? ...
> What are the constraints on the order of morphemes within words?
> Do phonological rules complicate the problem of morphological analysis?"

Galdera hauei erantzutean konputazio-ereduei begirako morfologiaren kontzeptu garrantzitsuenak ondorioztatzen dira:

- Funtzioei begira hiru kontzeptu azaltzen dira nagusiki, *flexio-morfologia* eta *eratorpen-morfologia* eta *elkarketa*. Lehenengoa sintaxiak eskatua da, erregularra da normalean kategoriaren arabera, eta ez du funtzio sintaktikoa aldatzen. Eratorpena aldiz, sintaxiak ez du eragiten, ez da erregularra eta kategoria gramatikalaren aldaketa gerta daiteke. Elkarketa lema bat baino gehiago biltzean sortzen da eta, askotan hitzaren muga gainditzen duenez, bere tratamendua korapilatsuagoa da flexio edo eratorpenarena baino.

- Morfemen arteko loturari begira fenomeno desberdinak gerta daitezke, gure inguruko hizkuntzetako ohizko aurrizki eta atzizkiak erabiltzen dituen *kateatze sinple*tik, arabiera bezalako hizkuntzetako *erro-patroi* eredu konplexuraino. Konplexutasunaren aldetik tartean egongo liratekeen beste fenomenoak ere aipa daitezke: artizkien bidezko lotura, bikoizketa, etab.

- Loturak gertatzeko murriztapenak, askotan *morfotaktika* deritzana, inguruko morfemen funtzioa izatea da arruntena, nahiz eta hizkuntzaren arauera erregulartasun- eta hurbiltasun-gradu oso desberdinak aurkitu.

- Aldaketa fonologikoak batzuetan fonologiak zuzenean eraginda izan daitezke (suomieran adib.) eta beste batzuetan ortografiak (ingelesa adib.), horrexegatik *morfofonologikoak* deituko ditugu, bibliografian aurreko izenez gain

morfografemika ere agertu arren. Aldaketa hauen kopurua eta aldaketa gertatzeko baldintzak oso bestelakoak izan daitezke hizkuntza desberdinetan. Fenomeno honen aurrean, analisi morfologikoa egiterakoan, sistema batzuetan *alomorfoak* erabiltzen dira, hau da, morfema bera adierazteko forma bat baino gehiago erabiltzea. Aldaketa fonologiko konplexuaren adibide gisa, hizkuntza batzuetan gertatzen den bokal-armoniaren fenomenoa dugu, non puntu batean gertatzen den bokal baten aldaketak ondoko bokalen aldaketa ere eragin baitezake.

II.2 Morfologiaren eredu konputazionalak eta zenbait adibide.

Atal honetan deskribatuko dugu zein eredu erabili diren analisi zein sintesi morfologikoa ordenadorez burutu nahi izan denean. Gure helburua euskararen morfologiaren tratamendua izan denez, honen ezaugarri den morfemen kateatzeari aurre egiten dioten teknikak azaltzen dira, bestelako hizkuntzen fenomeno batzuk —hizkuntza semitikoenak adibidez— aipatzen badira ere.

II.2.1 Eredu konputazionalak: sailkapenerako irizpideak

Ingelesaren flexio-morfologia sinplearen[1] eraginaz ordenadorez egindako analisi/sintesi morfologikoari kasu handiegia ez zitzaion egiten (Winograd, 83). Programa eta ezagumendu linguistikoa nahasten zuten sistema primitiboak ziren ohizkoak orain dela urte gutxi arte. Azken urteetan aldiz, arlo honetan egindako lanak ugaritu egin dira honako arrazoiak direla medio: beste hizkuntzetarako sistema automatikoen garapena batetik, eta corpusetan oinarritutako analisirako eskaintzen duten abantaila bestetik.

Gaur egun prozesadore[2] morfologiko asko aurki daiteke bibliografian, bakoitza bere ikuspuntu eta ezaugarriekin. Beraien arteko konparaketa egin ahal izateko irizpide batzuk zehaztu behar dira aurretik. Irizpideak zehazterakoan aurreko atalean azaldutako galderetatik ere abiatuko gara honako irizpide hauek ondorioztatuz:

1) Formalismo edo ereduaren *deskribapen-ahalmena*, hau da, zein fenomeno adieraz edo analiza daitezke eredu hori erabiliz. Aztertuko ditugun adibideak, morfologikoki euskaratik urrutegi ez egotea nahi dugunez, ezaugarri honi

[1] Ingelesaren morfologia sinpletzat hartu bada ere, hau guztiz zalantzazkoa da; horrela Sproat-ek (1992:152-53) azpimarratzen du nola morfologia konplexuaren fama duten hizkuntzetan (suomiera edo turkiera, adib.) konplexutasuna luzeraren sinonimotzat hartu den, erregulartasuna eta aldaketen kasuistika kontutan hartu gabe.

[2] Analisi edota sintesi morfologikoa burutzen duen programari prozesadore morfologiko deituko diogu.

dagokionean antzekoak izango dira; flexioa, eratorpena zein hitzaren mailako elkarketa adierazteko gai izanda ere, morfemen arteko loturen konplexutasuna kateatze maila hutsean geldituko baita, erro-patroi bezalako eredu konplexuak kontuan hartu gabe. Era berean, irizpide honen barruan *analisi* eta *sorkuntza* burutzeko gaitasuna edo bietako bat bakarrik burutzekoa bereiziko dugu.

2) Morfologiari ekiteko modua. Teoria linguistikoak eraginda, eta hizkuntzaren egiturak zein sistema eraikitzeko konputazio-ikuspegiak ere, bi eredu bereizten dira:
   - lexikoan oinarritutakoak, erroa eta hizkiak[1] dira abiapuntua eta beraiek dira gainontzekoa gobernatzen dutenak.
   - paradigman oinarritutakoak, paradigma desberdinak dira sistemaren funtsa eta gainontzeko osagaiak paradigmen menpe daude (Calder, 89; Anick & Artemieff, 92). Horrela, lexikoaren osaketa egiterakoan paradigma da erabiltzen den irizpide nagusia.

   Sistema gehienak erro-hizkian oinarritzen dira, eta ondoren aztertuko ditugun adibideetan horrela suposatuko da besterik esaten ez bada behintzat.

3) Morfotaktika ebazteko modua. Aurretik aipatu den bezala morfemen arteko lotura posibleak zehazteko moduarekin dago lotuta. Morfemetan oinarritutako sistemetan bi prozesamendu-mota agertzen dira nagusiki: *egoera finituko morfotaktika* deituko duguna eta *baterakuntza-mekanismo*etan oinarritutakoak. Lehenengoetan morfemen arteko erlazioak grafo-eran ikus daitezke, korapiluneak morfemak eta arkuak onartutako kateatzeak izanik. Baterakuntza-mekanismoek sintaxian erabili ohi diren ezaugarrietan oinarritutako gramatikak aintzakotzat hartzen dituzte, eta ondorioz malguagoak dira, tratamendu morfologiko —edo morfosintaktikoa— errazten dute baina konplexuagoak dira konputazioaren ikuspuntutik. Horietan, eredu paradigmatikotik egindako hurbilketak dira askotan, objektuei zuzendutako ereduetan ohizko diren herentzia bezalako kontzeptuak erabiltzen direlarik (de Smedt, 84; Calder, 89; Anick & Artemieff, 92).

4) Aldaketa morfofonologikoak adierazteko modua. Bestelakoak badaude ere, bibliografian bi metodo gailentzen dira: orain dela urte batzuk ohizkoa zen

[1] Hizki terminoa aurrizki, atzizki eta artizkien multzotzat hartzen dugu lanean zehar.

programa bidezko metodo *ad-hoc*ak eta gaur egun oso arrakastatsu bihurtu den egoera finituko *itzultzaile*etan[1] oinarritutakoa.

5) Lexikoan gordetzen diren osagai-motak. Ohizkoa da morfemak gordetzea, sistema batzuetan erroak gordetzen ez badira ere; baina batzuetan gordetzen dena hitz-zatiak dira aldaketa morfofonologikoak adierazteko modurik ez dagoelako edo eraginkortasun-arrazoiengatik. Beste aukerak ere badira, silabekin lan egiten dutenak adibidez (Cahill, 90). Irizpide honen barruan kokatzen dira lexikoan batzuetan agertzen diren bi fenomeno:
   - alomorfoen erabilpena, hau da, morfema bera adierazteko lexikoan forma bat baino gehiago erabiltzea.
   - morfemen desitxuratzea, morfema bere forma ezagunean ez gordetzea hain zuzen; KIMMOn (Koskenniemi, 93) erabiltzen diren diakritikoak dira honen adibide.

Azken irizpidea eraginkortasunarena litzateke, dena den ez da aintzakotzat hartu irizpide-multzo honetan, bere formalizazioari garrantzia eman nahi izan diogulako eta batzuetan eraginkortasuna inplementazioaren araberakoa delako eta ez formalismoaren ezaugarri. Izan ere, beste atal batean sakonduko dugu honetaz bi mailatako morfologiaren eraginkortasuna eztabaidatzean.

Adibideak aztertzean aipaturiko irizpide horien arabera sailkatzen saiatuko gara; kasu batzuk, sailkapen ia-ia guztietan gertatzen den legez, alde batean edo bestean kokatzea oso korapilatsua bada ere.

Moreno Sandoval-ek (1991) ondoko sailkapena proposatzen du:

- hitza-paradigman oinarritutakoak
- “osagaiak eta prozesuak” motakoak (edo automatetan oinarritutakoak)
- “osagaiak eta kokapena” motakoak
- bi mailatakoa eta baterakuntza

Sailkapen hori lehenago aipatutako irizpideen arabera ere adieraz daiteke, eta gainera guk aurreko puntuetan proposatutakoa zabalagoa eta zehatzagoa delakoan gaude. Moreno Sandoval-ek sistemen arteko konparazioak egiteko beste hiru irizpide proposatzen du: aipatutako eraginkortasuna, egokitzapen formala eta gainsorkuntza.

Gainsorkuntzaren arazoa ereduaren arauera baino inplementazioaren menpe dago —hizkuntza-espezifikazioa egitean alearen tamaina da funtsezkoa— askotan, eredu batzuetan gainsorkuntza ekiditea oso zaila gerta badaiteke ere.

---

1 *Itzultzaileak*: etiketa gisa n-tuplak dituzten automatak edo arkuetan n-tuplak dituzten grafo zuzendu finitoak.

Egokitzapen formalaren inguruan berak proposatzen ditu lau eredu, formalizazio prozedurala eta erazagutzailea batetik eta inplementazio prozedurala eta erazagutzailea bestetik konbinatuz lortzen direnak hain zuzen. Sailkapen hau konparaketak egiteko erakargarria bada ere, inplementazioa formalismoari egokitzen zaion zerbait izaten da, edo izan beharko luke; eta guk nahiago izan dugu formalismoak sailkatzea eta ez inplementazioak. Azken finean, berak proposatutako inplementazio erazagutzailea formalismoak erabiltzen duen eredutik finkatuta datorren ondorioa besterik ez da. Baterakuntzan oinarritutako inplementazio erazagutzailearen alde sintaxiarekin eduki ohi duten homogenotasuna aipatzen da, baina azken urteotan indartuz joan den egoera finituko sintaxia (Karlsson *et al*., 92) erabiltzen bada, estatu finituko morfologia homogenoa da ere.

II.2.2 Adibideak

Ondoren bibliografiako zenbait prozesadore morfologiko aurkezten dugu. Egiten den aurkezpena ez da osoa, adibide adierazgarri batzuk besterik ez baitira azaltzen; hala ere sistema bakoitzarekin berarengandik gertu dauden beste batzuen bibliografia-erreferentzia ematen da. Aurpezpenean jarraitu dugun ordena kronologikoa izan da (ordezkaria hautatzeko garaian behintzat, nahiz eta aldamenean antzeko adibide berriagoak zehaztu), alde batetik kontzeptuen bilakaeraz konturatzeko, eta bestetik ezaugarrien arabera aurkeztea nahikoa konplexu suerta zatekeelako.

II.2.2.1 DECOMP

Analizatzaile hau, ondorengo bertsioak izan baditu ere, hirurogeiko hamarkadaren erdialdean garatu zen MITn, MITalk izeneko proiektuaren barruan (Allen *et al*., 87). Lehenengo analizatzaileetako bat da. Lexikoa edukitzeko tokiak zein sistemaren hedadura nahikoak bere garrantzia bazuten ere, analisia burutzeko izan zuten arrazoi nagusia ingelesez morfologia eta hizketaren artean dagoen lotura da.

DECOMPen funtsezko ezaugarriak honako hauek dira:

- Flexioa, eratorpena zein hitzaren mailako elkarketa hartzen ditu kontuan. Analisirako tresna da eta ez du sorkuntzarako aplikaziorik.
- Egoera finituko morfotaktika erabiltzen du, morfemen motetan oinarritua. Morfotaktika definitzeko erregela sinple batzuk erabiltzen dira.
- Aldaketa morfofonologikoak oso erregela sinpleen bidez deskribatzen ditu. Aldaketa hauek morfemen artean gertatzera mugatuta daude, eta oso aldaketa

sinpleak adieraz daitezke. Morfema baten azken letraren aldaketa, ezabaketa edo sorrera besterik ez da kontutan hartzen sistema honen erregeletan.

- 12.000 morfemak osatzen dute lexikoa, Brown corpuseko 50.000 hitzetatik abiaturik lortu zirenak. Morfema bakoitzari kode bat egokitzen zaio —morfema-mota definitzen duena—, bere gainean erregelak nola aplika daitezkeen definitzen duena. Horietako kode batek erregela bakoitzak eragiten duen aldaketa behartu, debekatu edo aukeran utz dezake.

Morfemetan banatzeko erabiltzen den algoritmoak eskuinetik ezkerrera tratatzen du hitza, errekurtsiboa da, eta anbiguitateak ekiditeko morfotaktikari dagozkion egoera-aldaketei pisu bat esleitzen die analisi-eredu batzuk beste batzuei gailentzearren eta analisia azkartzearren. Horrela *scarcity*-ren eratorpen gisako analisia "*scarce+ity* " lortuko da eta ez "*scar+cite+y*" elkarketa. Emaitzen aldetik, analisien %95a zilegia dela diote, baina badirudi neurri hori beste moduluen lana kontutan hartuz egiten dela.

Sistema hau aspaldikoa da baina urteetan zehar hobetu dute. Oso ezaugarri interesgarriak ditu: morfotaktikaren tratamendu dotorea, desanbiguazio-mekanismoa eta eraginkortasuna. Eragozpenak ere leporatu behar zaizkio: analisirako baino balio ez izatea —bere aplikaziorako nahikoa bada ere— eta aldaketa morfofonologikoen aldetiko ahalmen eskasa —ingelesaren tratamendurako honetan nahikoa izan arren—. Azken arrazoi hauek direla eta, ez da morfologiarako eredu orokorra eta ez du jarraitzaile asko izan.

Espainierarako MARS (Meya, 87) izeneko analizatzaile morfologikoak antz handia du DECOMP sistemarekin, ezaugarri guztiak, analisia egin ahala burututako desanbiguazioa barne, pareka baitaitezke: analisirako bakarrik balio izatea, egoera finituko morfotaktika, aldaketa morfofonologikoak oso erregela sinpleen bidez —nahiz eta arlo honetan DECOMPekin desberdintasunak izan—, eta lexikoan morfemei dagozkien erregelei buruzko informazioa ere gordetzea. Lexikoan alomorfoak erabiltzen dira beren aldaketari dagokion erregela morfofonologikoa orokorra ez denean. MARS (Morphological Analisis for Retrieval Support) datuak berreskuratzen laguntzeko sistema baten barruan erabiltzen da.

II.2.2.2 ATEF

ATEF itzulpen automatikorako ingurune baten barruan dagoen analizatzaile morfologikoa da. Ingurune hau GETA Grenobleko laborategian erabiltzen da (GETA, 82), eta 70.eko hamarkadaren bukaeran garatu izan da. Ingurune horren barruan harreman estua du ROBRA izeneko analizatzaile/sortzaile sintaktikoarekin. Hizkuntza askotarako erabilia izan da, frantsesa, alemanera, errusiera eta Asiako ekialdeko zenbait hizkuntzatarako

analizatzaileak eraiki baitira. Gure taldeak prototipo bat burutu du euskararako (Arregi & Urkia, 89).

ATEFen osagaiak honako hauek dira:

- Aldagaiak: analisi morfologikoaren emaitza den informazio morfologikoa jasotzen duten aldagai sinbolikoak.
- Hiztegiak: morfemak biltzen dituzten azpilexikoak. Gehienez zazpi dira, erregeletatik kudea daitezke, eta bertan honak informazio hauek azaltzen dira: hitz-zatia, hau da, aldatzen ez den morfemaren zatirik luzeena, dagokion formatoa eta unitate lexikoa —erro amankomuna duten hitz-zatiak biltzeko erabilia— eta gainerako informazio morfologikoa.
- Formatoak: hiztegiko unitate-multzo bati dagokion informazioa biltzen duen eredua. Ohizkoa da atzizki berdinak hartzen dituzten lexikoko unitateei formato bera egokitzea.
- Gramatika (erregelak): erregelen multzoa, hiztegietan aurkitutako hitz-zatiei dagozkien formatoen arabera aktibatzen direnak eta zenbait ekintza buru daitezen eragiten dutenak. Berauetan bestelako baldintzak zehatz daitezke, ekintza garrantzitsuenak ondokoak izanik: aldagaien gaineko eragiketak, hiztegien irekitzea edo ixtea, eta testu-aldaketa.

Programak etengabe bilatzen ditu hitz-zatiak hiztegietan, eta aurkitutakoei dagokien informazioa aldagaiei esleitzeaz gain, berauen formatoen arabera aplikatzen ditu erregelak.

Aipatutako osagaiekin morfotaktikaren zein tratamendu morfosintaktikoaren deskripzioa erraza eta malgua den bitartean, salbuespenak modu dotorean adieraztea bideratuz, aldaketa morfofonologikoen tratamendua kaxkarra da oso, horretarako morfotaktika helburua duten erregelak erabili behar baitira. Hori dela eta, aldaketa morfofonologiko sinpleak adierazteko ere, zenbait amarru eta zeharkako bide erabili behar dira beti.

Horrez gain, beste bi eragozpen ditu sistema honek:

- Programa ezagumendu linguistikotik independente bada ere gramatikaren idazketa ez da erazagutzailea, metalengoaia agintzaile batetik gertu dagoen zerbait baizik.
- Programa ez da eskuragarria eta bere zehaztasunak ez dira ezagunak, eta gainera garaiko IBM *mainframe-tean* baino ezin zen erabili.

Martí-k (1987) espainierarako proposatutako AM analizatzaileak, lematizatzaile baten parte denak, zenbait ezaugarri du amankomunean aurrekoarekin:

- Analisirako bakarrik balio du.
- Morfotaktika azpilexikoetan oinarritutako erregelen bidez burutzen da eta erregela hauek informazio morfologikoarekin lotutako ezaugarrien menpe jar daitezke. AM-n eratorpen-morfologia definitzeko erabiltzen da aukera hau.
- Lexikoan UD (hiztegi-unitate) izeneko hitz-zatiak gordetzen dira.
- Aldaketa morfofonologikoetarako ez dago mekanismorik.

II.2.2.3 KIMMO

Aurretik ikusitako ereduetan morfotaktikaren aldetik nahikoa ahaltsuak baziren ere aldaketa morfofonologikoetarako pobreak ziren. Horren zioa aplikazio-hizkuntzen ezaugarrietan aurki daiteke, zeren eta normalean oso flexio pobre eta aldaketa erregular gutxi duten ingelesa bezalako hizkuntzetarako egiten ziren prozesadore morfologikoak.

Koskenniemik (1983) bere tesian eredu berri bat proposatu zuen, bi mailatako morfologia deitutakoa, oso arrakastatsua gertatu dena bere ezaugarri garrantzitsuenari esker: analisi zein sintesirako aldaketa morfofonologikoak adierazteko formalismo ahaltsu, orokor eta eraginkorra[1] izatearena hain zuzen. Suomierarako gauzatu bazuen ere, berehala etorri zen KIMMO[2] izeneko ingeleserako bertsioa, Karttunen-ek (1983) eginda. Aldaketa morfofonologikoak adierazteko egoera finituko itzultzaileetan konpilatzen diren bi mailatako erregela paraleloak erabiltzen dira. Formalismo hau da euskararako oinarrizko tresnak diseinatzerakoan aukeratu duguna.

Hala ere KIMMO ez zen izan lehena aldaketa morfofonologikoak deskribatzeko erregela orokorrak diseinatzen. Beste batzuen artean, aldaketa morfofonologikoak adierazteko Kaplan-ek eta Kay-k (1981) automatatan konpilatzen ziren erregela sekuentzialak erabiltzea proposatu zuten —Koskenniemirengan eragin handia izan zuena—, baina tarteko egoerekin arazoak gertatzen ziren. Ondoren *keçi* izeneko prozesadore morfologikoa egiterakoan Hankamer-ek (1986) *sortu eta egiaztatu* filosofiarekin zebiltzan erregela sekuentzialak ere proposatu zituen.

---

1 Ezaugarri hauen gainean ñabardurak egingo dira geroago.

2 KIMMO izenarekin Koskenniemik proposatutako bi mailatako morfologian oinarritutako prozesadore morfologiko guztiak izendatuko ditugu.

Bi mailatako morfologiaren ezaugarriak zehatz-mehatz hurrengo atalean aztertuko ditugun arren, formalismo guztien artean sailkapen bat egiteko ezinbestekoa da ezaugarri horiek laburtzea:

- Analisi zein sintesirako baliagarria da. Kateatze mailako fenomenoak bakarrik deskriba daitezke, baina bi mailatako ideia n mailatara zabalduz beste fenomeno konplexuagoak, hizkuntza semitikoenak adibidez, ebatz daitezke (Kay, 87) (Beesley, 90) (Kiraz, 94).

- Azpilexikoetan oinarritutako morfotaktika. Morfema bakoitzari bere ondotik etor daitezkeen morfemen multzoa definitzen duen *jarraitze-klasea* egokitzen zaio. Hori dela eta, morfemen kateatze-sekuentzia bateko i-garren morfemak (i+1)-garrena baino ezin du baldintzatu, urruneko menpekotasuna deituriko fenomenoa deskribaezina bihurtuz. Beraz, mekanismoa oso sinplea da, baina batzuetan ez da nahikoa esanguratsu, eta horrexegatik proposatu dira aldaketak arlo honetan. Horrela, ondoko atalean aztertuko dugunez, Bear-ek (1986), Ritchie-ren taldeak (1987) Alvey sistemaren barruan, eta Trost-ek (1990) ezaugarri morfologikoetan oinarritutako baterakuntza-mekanismoak proposatzen dituzte honetarako, eta guk, aldiz, *jarraitze-klase hedatuak.*

- Aldaketa morfofonologikoena da aipatu den bezala gailentzen den aspektua, egoera finituko automaten ideia modu arrakastatsuz erabiltzen duelako xede honetarako.

- Lexikoan morfemak gordetzen dira eta, beharrezkoa ez bada ere, alomorfoak erabiltzea ez du baztertzen Koskenniemik. Morfema desitxuratu baina erregelen aplikazioa kontrolatzen duten *diakritikoak* (berak *hautapen-markak* deitzen dituenak) erabili ohi dira.

Formalismo honen arrakasta izugarria izan da. Cahill-ek (1989) horrela zioen:

> "The field of computational morphology was revolutionized by the work of Kimmo Koskenniemi, whose two-level model of morphonology has been used for the description of several languages, including English, French, Finnish and Japanese."

Literaturan gehiago aurkitzea erraza bada ere, hona hemen eredu honetaz ari diren erreferentzia garrantzitsu batzuk:

- Hobekuntzak: (Karttunen *et al*., 87), (Kay, 87), (Ritchie *et al*., 87), (Bear, 88), (Trost, 90), (Karttunen *et al*., 92), (Karttunen, 93).

- Inplementazioak (aldaketa handirik proposatu gabe): (Karttunen & Wittenburg, 83), (Karlsson, 92), (Clemenceau & Roche, 93), (Oflazer, 94), (Kim *et al*., 94), (Kiraz, 94).

- Banaketa libreko tresnak : PC-KIMMO (Antworth, 90), (Karp *et al*., 92).

Gaur egun gutxienez bi enpresa ari dira formalismo hau erabiltzen produktu komertzialak lortzeko: Xerox eta Lingsoft. Azken hau, 1994an alemanera analizatzen sistema onena hautatzeko Morpholympics izeneko lehiaketan, izan da irabazle.

II.2.2.4 Tzoukermann eta Liberman

Aurretik ikusitako azken bi aukerak biltzen dituen sistema baten berri ematen digute Tzoukermann-ek eta Liberman-ek (1990), analisi zein sintesirako erabil daitekeena. Sistema honetan linguistak egiten duen espezifikazioa eta programak tratatzen duena nahikoa desberdinak dira, tarteko konpilazio-prozesua dela eta. KIMMOren kasuan erregeletatik automatetara iragateko egiten den konpilazioa —eskuzkoa edo automatikoa— bazegoen, baina honek ez zituen aldatzen sistemaren ezaugarriak.

Espainierarako AT&T laborategietan egindako lan honetan, KIMMOk proposatzen zituen erregelen ideia hartzen da, baina eraginkortasunari begira lexikoarekin konpilatzen dira eta alomorfoei dagozkien hitz-zatiak sortzen dira automatikoki. Sortutako hitz-zatien gainean ez da aldaketa morfofonologikorik gertatuko eta, horrela, lexikoan egingo den bilaketa karakterez karaktere egin beharrean zatiz zati egin daiteke. Horretarako, konpiladorearen emaitza ez da izango morfema eta alomorfoari dagokien hitz-zatia bakarrik, baizik eta morfotaktika kontrolatzen duen grafoan dagokion hasiera- eta bukaera-egoera ere. Adibide gisa, bere artikuluan zehazten dituzten konpilatu ondorengo lexikoko osagai batzuk:

hasiera-egoera	**bukaera-egoera**	**hitz-zatia**	**morfema**	**informazio morfologikoa**
1	7	jug	jugar	aditza
1	8	jueg	jugar	aditza
1	9	juegu	jugar	aditza
1	10	jugu	jugar	aditza
1	300	buen	bueno	adjektiboa
1	500	mariscos[1]	mariscos	izena, mask., plur.
150	500	o		sing.,orain.,indik.,1.

[1] Izenetan generoari eta zenbakiari dagozkien flexioak ebazten ditu konpiladoreak.

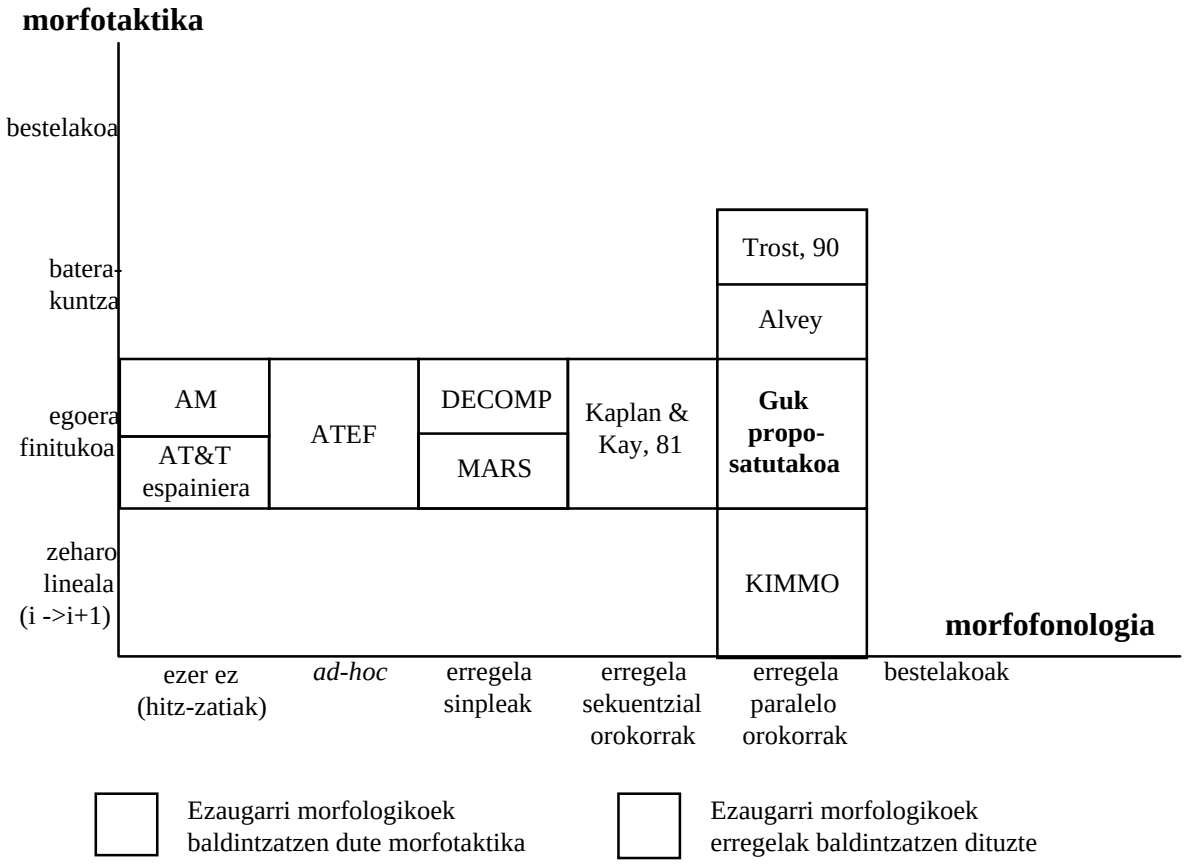
150 500 as sing.,orain.,indik.,2.

Horrela espezifikazioaren ikuspuntutik ezaugarriak KIMMO sistemarenak bezalakoak badira ere, programaren ikuspuntutik ATEF eta AM izenarekin aztertutako kasuetatik gertuago dago.

Beraiek oso interesgarri jotzen dute hau espainiera bezalako hizkuntz erromantzeetarako, eta alemanerarekin esperimentatzeko asmoa azaltzen dute. Bi mailatako morfologiaren barruan aztertuko dugun bezala, Karttunenek (1993, 1994) *lexc* izeneko konpiladorea erabiltzen duten lexiko-itzultzaileak proposatzen ditu, Koskenniemiren formalismoaren inplementazioa hobetu eta azkartzen dutenak. Karttunenen proposamen honetan grafoaren arkuetan morfemak edo hitz-zatiak egon beharrean, karaktere-bikoteak agertzen dira, bi mailatako formalismoaren ezaugarriak bere horretan mantenduz.

II.2.3 Sailkapen bat

Aurreko adibideak aztertu eta gero, II.1 irudian ikus daitekeen sailkapena ezar daiteke. Sailkapen honetan sartzeko sistemek honako ezaugarri hauek bete behar dute: morfemez osatutako lexikotan oinarriturik egotea eta kateatze-mailako fenomeno morfologikoak deskribatzea. Halako sistemetarako eskema orokorra da morfotaktika eta morfofonologia bereiztea dagoenean behintzat; eta kasu berriak aztertu ahala egunera liteke irudia. Beste idazle batzuek aipatzen duten bezala (Ritchie *et al*., 92), beste ereduekin konparaketa zehatzak egitea oso zaila gertatzen da, eta horrexegatik horiek iruditik kanpo gelditzen dira.



II.1 irudia.- Azaldutako prozesadore morfologikoen sailkapena morfotaktikaren eta morfofonologiaren tratamenduaren arabera

II.3 Bi mailatako morfologia.

1983an Koskenniemik bi mailatako morfologiaren eredu konputazionala definitu zuen, aurreko sailkapenean KIMMO izenarekin aipatu duguna. Egoera finituko morfologiari bultzada handia eman zion eredu honek harrera bikaina jaso du ondorengo urteetan, besteak beste, dituen ezaugarri hauengatik:

- Eredu orokorra da, kateatze-mailako morfologian behinik behin, eta beraz, gure inguruko edozein hizkuntzari aplika dakioke.
- Ezagutza linguistikoa eta algoritmoa bereizi egiten ditu eta, ondorioz, programa berak edozein hizkuntzatarako balio dezake.
- Baliagarria da hitzen analisi morfologikorako zein sorkuntzarako.
- Analizatu edo sortuko den hitzaren *azaleko maila* eta hiztegian (lexiko-sisteman) errepresentatzen den *lexikoko maila*—sakonekoa ere esaten zaio— argi eta garbi

bereizten ditu. Hau dela eta, ez dago aldaketa morfofonologikoengatik sortutako morfema baten forma desberdinak (alomorfoak) gorde beharrik.

- Fonologia sortzaileko berridazketa-erregelak erabili beharrean erregela paraleloak erabiltzen dituzte, kontzeptualki zein konputazionalki errazago bihurtuz.

- Aurreko ezaugarriak kontutan hartuz, esan daiteke sistemaren konputazio-konplexutasuna ez dela altua, eta ondorioz, makina txikietan sistema errealak ezartzea bideratzen duela.

Ondoko pasarteetan formalismo honen osagai garrantzitsuenak diren lexiko-sistema, morfotaktika ere definitzen duena, eta erregela morfofonologikoak sakonean azaldu ondoren, sistemari egindako kritikak zerrendatuko dira, eta bukatzeko, morfotaktikari dagokionean, proposamen desberdinak eta guk egindako ekarpena azalduko dira.

II.3.1 Lexiko-sistema.

Lexiko-sistemak morfema-multzoa eta morfotaktika definitzen ditu. Hiru elementu funtsezkok osatzen dute sistema lexikoa: lexiko-sarrerak, azpilexikoak eta jarraitze-klaseak.

Lexiko-sarrera bakoitzak hiru eremu ditu:

- Lexiko-adierazpena, alfabeto lexikoko karaktere-sekuentzia bat da. Karaktere hauek azaleko karaktereak, *morfofonemak* eta *hautapen-markak* izan daitezke. Erregelen bitartez karaktere arruntei zein morfofonemei azaleko beste karaktere bat edo hutsa egoki lekizkiekeen bitartean, hautapen-markei hutsa besterik ez zaio egokituko. Bibliografian hautapen-markei diakritiko deitzen zaie morfema desitxuratu egiten dutelako, baina beharrezkoak dira erregelen aplikazioa murrizteko.

- Dagokion *jarraitze-klasea*. Geroxeago azalduko den bezala morfemen arteko sekuentzia posibleak erregulatzen dituzte jarraitze-klaseek.

- Sarrerari dagokion *informazio morfologikoa*, analisiaren emaitza gisa agertuko den informazioa alegia.

Lexikoa morfemen artean egon daitezkeen kateamenduen arabera sailkaturik dago azpilexikoen bidez. **Azpilexikoek** aurreko morfemekiko kateatzeari dagokionean ezaugarri berdineko elementu lexikoak biltzeko balio dute. Hori dela eta, morfotaktikak behartzen du azpilexikoen eraketa, linguistikoki elementu artifizial samarrak gertatuz.

Horrela, deklinabide-atzizki guztiak, adibidez, ezin dira azpilexiko berean egon, batzuk lema-motaren arabera aukeran badira.

Azpilexiko guztien definizioek egitura bera dute: identifikadorea den izena, ezaugarriak eta sarrera-multzoa. Amankomuneko informazio morfologikoa duten morfema-multzoa markatzeko erabil daitezke ezaugarriak. Horrela hitz-hasieran egon daitezkeen morfemak ezaugarri batez ezagut daitezke.

Jarraitze-klasea azpilexiko multzo bat da, morfotaktikaren aldetik unitate bat dena, eta paradigma baten osagaiekin identifika daitekeena. Identifikadore batez ezagutzen da. Esan bezala jarraitze-klase bat egokitzen zaio lexikoan morfema bakoitzari, eta jarraitze-klasean biltzen diren osagaiak dira definitutako sarreraren ondoren ager daitezkeen morfema bakarrak. Beraz, jarraitze-klaseak hitz batean ager daitezkeen morfemen arteko konbinaketa posibleak definitzeko mekanismoaren oinarria dira.

Morfotaktika-sistema honen deskribapen-ahalmena, esan bezala, txikia da oso, eta honengatik, kasu batzuetan beharrezkoa izango ez litzatekeen zenbait deskripzio-bikoizketa gertatu ohi da, aipatutako morfemen arteko urruneko menpekotasunaren kasuan esaterako. Hori dela eta, aldaketak proposatu dira arlo honetan.

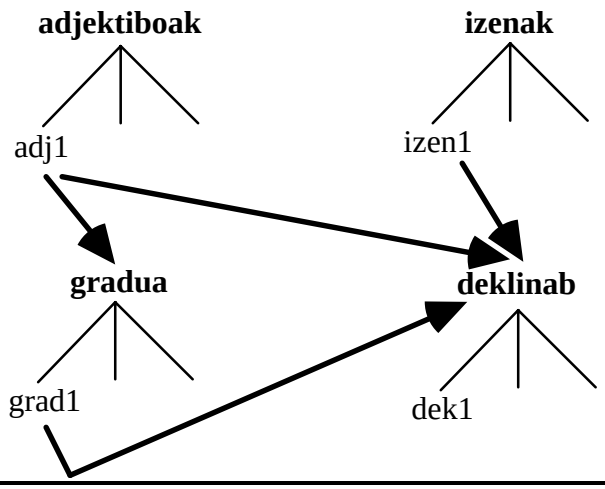
Jarraitze-klaseen eta azpilexikoen arteko bereizketa ez da funtsezkoa zeren lexiko-sistema beste modu honetaz ikus baitaiteke: estekatutako azpilexiko-multzoa. Ikuspegi honetatik morfotaktika definitzen duen grafoa ondo eratzeko elementuak besterik ez dira jarraitze-klaseak eta azpilexikoak.

Adibidez, eta sinplifikazio bat eginez, euskarazko adjektiboek eta izen arruntek deklinabide-atzizki berberak hartuko dituzte, baina lehenek baino ezin dute gradu-flexiorik hartu. II.2 irudian ikus daiteke lexiko sinplifikatu honen eraketa.

LEXIKOA **izenak**
ize1 izen_jk ...
ize2 izen_jk ...
...
LEXIKOA **adjektiboak**
adj1 adj_jk ...
adj2 adj_jk ...
...
LEXIKOA **gradua**
grad1 izen_jk ...
grad2 izen_jk ...
...
LEXIKOA **deklinab**
dek1 #[1] ...
dek2 # ...
...

JARRAITZE-KLASEA **izen_jk**
deklinab

JARRAITZE-KLASEA **adj_jk**
deklinab
gradua



II.2 irudia.- Lexiko-sistemaren erazagutzearen eta egituraren adibidea

Aldaketa morfofonologikoak burutzeko erregela-multzoa dagoenez, lexikoan **alomorfo**rik definitzko beharrik ez dago. Hala ere Koskenniemik ez ditu baztertzen bi kasuta berezitan:

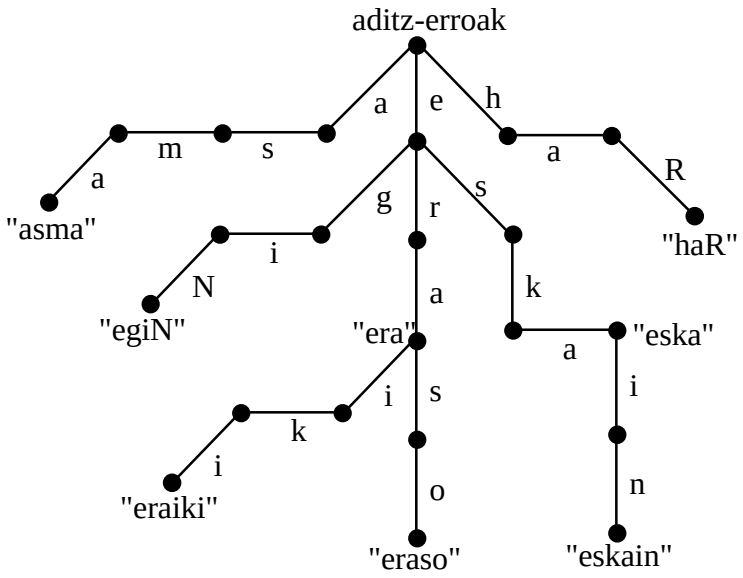
1 # sinboloak jarraitze-klase hutsa adierazten du.

- azala eta lexikoaren artean aldaketa handiak edota ez-erregularrak gertatzen direnean. Horien artean daude berak erabiltzen dituen *aukera-patroiak*[1] izeneko azpilexiko txikiak.

- jarraitze-klaseak eratzeko orduan azpilexiko txiki asko sortzea lexiko-sarrerak errepikatzea baino egokiago jotzen badu ere, gehiegizko dispertsioa ekiditeko horrelakorik egin daiteke.

Aurretik aipatutako morfotaktikaren deskribapen-ahalmen apalak eragindako lexiko-sarreren bikoizketa ere bada alomorfoen iturria.

Informazio-egituraketaren aldetik lexiko-sistema azpilexiko-zuhaitz bat da eta azpilexiko bakoitza errepresentatzeko hostoetan informazio morfologikoa gordetzen duen *trie*[2] egitura erabiltzen da. ***Trie*** **egitura** grafo bat da non arku bakoitzak lexikoko karaktere bat adierazten duen eta adabegi bakoitzak horretarainoko arkuek osatutako morfemari dagozkion jarraitze-klasea eta informazio morfologikoa gorde ditzakeen. Esan bezala, morfemari dagokion jarraitze-klasea zenbait azpilexikorekin lotura gauzatzen duten arku-multzo bat bezala ikus daiteke.



II.3 irudia.- Azpilexiko sinple baten *trie* egitura

[1] *aukera-patroiak*: azpilexiko txiki hauekin zera egiten da: aldaketa morfofonologiko ez-erregularretarako gertatzen diren aldaketak jarraitze-klase desberdinetan antolatzea, eta jarraitze-klase berbera behar zuten morfemei desberdinak esleitzea (morfofonologiaren zenbait arazo morfotaktika bihurtuz).

[2] *trie* terminoa re*trie*val hitzetik hartua da. Zuhaitz erako egitura bada ere ez da *tree*-rekin nahastu behar. Informazio gehiagorako ikus Knuth-en *The art of Computer Programming*, vol. 3, 481-489, 1973.

Adibidez *asma*, *egiN*[1], *era*, *eraiki*, *eraso*, *eska*, *eskain* eta *haR*[2] aditz-erroek osatutako azpilexiko baten egitura II.3 irudian ikus daiteke.

Karakterez karaktere atzitzeko eta informazioa modu trinkoan gordetzeko ahalmena aipa daitezke *trie* egituraren alde.

II.3.2 Bi mailatako erregelak.

II.3.2.1 Sarrera.

Aldaketa morfofonologikoak deskribatzeko Koskenniemik sortutako bi mailatako erregela-sistema izan zen zalantzarik gabe Koskenniemiren ekarpen handiena. Hori argi eta garbi gelditzen da bibliografian, eta idazle batzuek hori kontutan harturik eredu honen izena zalantzan jartzera iristen dira, bere ordez bi mailatako fonologia proposatuz. R. Sproat-en aipatutako liburuan (1992: 92-93) horrelako aipamena egiten da:

> "... the bulk of the machinery is designed to handle phonological rules, and in the original version of that system the actual *morphology* done is particularly interesting. Indeed, the term *two-level morphology*, used by Koskenniemi to describe the framework, is really a misnomer: the "two-level" part of the model describes the method for implementing phonological rules and has nothing to do with morphology per se."

Bi mailatako erregelek errepresentazio lexikoaren eta azalekoaren artean parekatzea kontrolatzen dute. Erregelak egoera finituko itzultzaile (FST)[3] paralelo bihurtzen dira eta karaktere-bikoteak onartuko dira baldin automata guztietan onartzen badira. Bi errepresentazioen artean, lexikokoa eta azalekoaren artean hain zuzen, ez dago tarteko egoerarik, eta hauxe da fonologia sortzailearekiko diferentzia nagusia. Beraz, hitzen analisia azaleko formari dagozkion errepresentazio lexiko onargarriak aurkitzean datza. Alderantziz gertatzen da sorkuntzan, errepresentazio lexiko ezagunetik abiatu eta berari dagozkion azaleko errepresentazioak bilatzen baitira.

Aipatu den bezala Koskenniemik proposatutakoaren (morfo)fonologia eredu honen aurretik Kaplan-ek eta Kay-k (1981) berridazketa-erregeletan oinarritutako beste eredu bat proposatu zuten, erregela sekuentzialena hain zuzen ere. Beren ereduan ere, erregelak

---

1 N karaktereak —morfofonemak, Koskenniemiren terminologiaz— gal daitekeen n karakterea adierazten du.

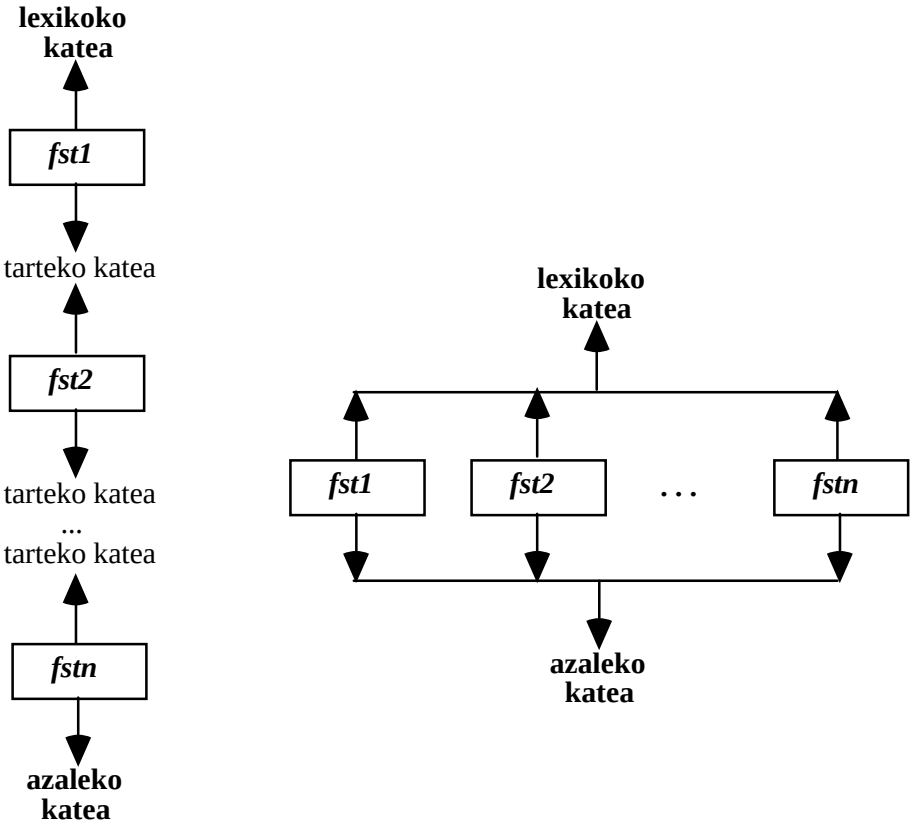
2 R karaktereak r gogorra markatzen du.

3 Egoera finituko itzultzaile (finite state transducer - FST) eta egoera finitutako automata (finite state automaton - FSA) baten artean dagoen desberdintasuna zera da: FSAren alfabetoko osagaiak sinbolo sinpleak diren bitartean FSTarenekoak bikoteak dira. Izena itzultzailea izan arren errazago ulertzen da bikoteak ezagutzeko edo ez ezagutzeko automata bezala.

egoera finituko itzultzaileetan konpilatzen ziren[1], eta ondorioz analisi zein sorkuntzarako balio du.

Aurreko eredu horiek, Kaplan eta Kay-rena, Koskenniemirenarekin konparatuz — eratorpen-fonologia eta erazagutze-fonologia izenekin bereizten ditu Karttunenek (1992)—, honako desberdintasunak zeuzkaten:

- Kaplan eta Kay-ren ereduak ikuspegi sortzailea du, beraz aldaketa morfofonologikoen arazoa urrats sinpleen bidez adieraz daiteke, eredu teoriko sinplea izanik.
- Tarteko egoerak sortzen ditu, eta beraz sekuentziak garrantzi handia du. Ondorioz testuinguruaren espezifikazioa konplexua bihur daiteke praktikan, ordenaren araberako aurreko erregelen emaitzak kontutan hartu behar direlako.
- Analisi zein sorkuntzarako baliagarria bada ere, sorkuntzaren kasuan prozesua ez-determinista gerta daiteke.



II.4 irudia.- Egoera finituko itzultzaile ordenatu eta paraleloen arteko konparaketa

[1] Ideia hau Jhonson-en (1972) ekarpena da.

Hala ere, Kaplan-en arabera (Kaplan, 88) (Kaplan & Kay, 94) bi ereduetan karaktere-kateen arteko erlazio erregularrak adierazten dira, erlazio hau itxia izanik konposaketarekiko baina ez ebaketarekiko. Izan ere bi mailatako morfologiaren kasuan ebaketa ere itxia da. Honetan oinarriturik Karttunen-ek bi ereduak erabiltzen dituzten lexiko-itzultzaileak proposatzen ditu, kapitulu honen amaieran azalduko direnak.

II.4 irudian konputazioaren aldetik bi ereduen artean dagoen desberdintasuna azaltzen da. Azpimarratu behar da bi sistemetan erregelak egoera finituko itzultzaile bihurtu arren, hauek bi mailatako morfologian itzultzaile baino bikote-kontrolatzaileak direla.

Bi ereduen ezaugarriez eta formalizazioaz sakontzeko oso egokiak dira Karttunen-en (1991) eta Kaplan & Kay-ren (1994) artikuluak.

II.3.2.2 Osagaiak

Esan bezala bi mailatako morfologiaren ezaugarririk garrantzitsuena lexiko-maila eta azalekoa parekatzen duten erregela-multzoan datza. Erregela hauen sintaxia zehaztu baino lehen defini dezagun beraiekin lotuta dauden zenbait kontzeptu:

- *bi mailatako konfigurazioa*: lexikoko eta azaleko karaktereak banan-banan sinkronizatzen dituen karaktere-kate parea. Adibidez:

  ```
  e g i N + t e n      (lexikoa)
  e g i 0 0 t e n      (azala)
  ```

  Konbentzioz pareetan beti lexiko/azala ordena mantenduko da. Zeroak karaktere hutsa adierazten du —beste lanetan ε sinboloarekin adierazten dena—.

- *azaleko alfabetoa*: analizatzeko formetan ager daitezkeen karaktere guztiak.

- *lexiko-alfabetoa*: lexikoan ager daitezkeen karaktere guztiak. Bertan azaleko karaktereez gain morfofonemak eta hautapen-markak daude.

- *karaktere-bikote konkretua*: lexikoko eta azaleko karaktere banaz osatutako bikotea. Øak ager daiteke bi mailetan, azalean morfofonema zein hautapen-markekin parekatzeko eta orokorrean azaleko karaktereren baten desagerpena edo sorrera adierazteko.

- *karaktere-multzoa*: amankomuneko ezaugarriak dituzten karaktereez osatutako multzoa, identifikadore batez ezagutzen dena. Horrela V bokalen multzoa izaten da eta C kontsonanteena. Konbentzioz `=` karaktereak alfabetoko edozein karaktere edo Ø adierazten du.

- *karaktere-bikote abstraktua*: lexiko-mailan eta azalekoan karaktere konkretu bat eduki beharrean karaktere-multzoa duen bikotea. Maila batean karaktere konkretua eta bestean multzoa dutenei bikote erdiabstraktua deritze.

II.3.2.3 Erregelen formatua

Erregelen hasierako sintaxiak aldaketa batzuk izan ditu (Koskenniemi, 85; Ritchie *et al.*, 92, Karttunen, 93) deskribapen-ahalmena irabazi eta konpiladoreak inplementatu ahal izateko, eta automatetarako itzulpena eskuz egin behar ez izatea ahalbideratzen duena. Aipatutako azkena erabiliko da hemen, konpiladore hori erabil dezakegu eta.

Erregelen formatua eta osagaiak honako hauek dira:

formatua `cp op lc _ rc`

Korrespondentzia (cp), karaktere-bikote bat da. Karaktere hauek konkretu nahiz abstraktuak izan daitezke, azken hauek erregelen generalizazioa ahalbidetzen dutelarik. Gehienetan bikote konkretuak dira.

Eragilea (op), testuinguruaren eta korrespondentzian adierazitako bikotearen artean zer-nolako erlazioa dagoen finkatzen duena. Lau eratakoa izan daiteke: *testuinguru-murriztapena* (=>), *azalekoaren derrigortzea* (‹=), aurreko biak batera (<=>) edo *debeku-ezarpena* (/‹=). Azaleko derrigortzearena berridazketa-erregeletan erabiltzen denaren baliokidea izango litzateke. Azken mota, debekuarena hain zuzen, Bear-ek (1986) proposatu zuen salbuespenen tratamendua errazago adierazi ahal izateko.

Bakoitzaren esanahia honako hau da:

op	**adibidea**	**interpretazioa**
=>	l:a => lc_rc	*l* lexikoko karakterea azalean *a* bihurtzen da **baldin** testuingurua lc_rc bada.
<=	l:a <= lc_rc	lc_rc testuinguruan *l* **beti** gauzatzen da *a* bezala.
<=>	l:a <=> lc_rc	*l* karakterea *a* bezala gauzatzen da lc_rc testuinguruan eta ez beste inoiz.
/<=	l:a /<= lc_rc	l **ez** da **inoiz** a bezala gauzatzen lc_rc testuinguruan.

Gehien erabiltzen dena eragile konposatua da, baina aldaketa baten gauzatzea aukeran denean, testuinguruaren murriztapena ere erabili ohi da.

Testuingurua (lc_rc), korrespondentzia gertatzen deneko kasuak mugatzen dituena, aurreko eta ondorengo karaktereen arabera. _ karaktereak ezkerreko edo aurreko testuingurua (lc) eta eskuineko edo ondoko testuingurua (rc) banatzen ditu, bietako bat hutsa izan badaiteke ere.

Ezkerreko zein eskuineko testuinguruetan karaktere-bikoteen segidak adierazten dira, eta bikotearen bi osagaiak berdinak direnean karaktere bakarra zehaztea nahikoa da. Bi puntuak eta karaktere bakar bat zehazten bada orduan karaktere horrekin era daitezkeen bikote posible guztiak erreferentziatzen dira. # sinboloak hitzaren muga adierazten du, beraz, ezkerreko testuinguruan hitz-hasiera eta eskuinekoan bukaera adieraziko du.

Testuinguruko bikoteak zenbait eragileren bidez konbina daitezke. Aukera hau urtetan zehar zabaldu da eta espresio erregularretan erabiltzen diren zenbait sinbolo hartu izan dira. Makoak **[ ]** espresioak biltzeko erabiltzen diren bitartean, parentesiek **( )** aukeran dagoena biltzeko balio dute. Hona hemen testuinguruko eragile batzuk (eragile hauek diakritiko gisa erabiltzen badira % ihes-karakterea jarriko zaie aurretik):

eragilea	**adibidea**	**interpretazioa**
kateaketa:	k:g o	k:g bikotea o:o bikoteaz jarraituta
bilketa: |	k:g | k:k	lexikoko k azaleko g edo k-rekin
errepikapena : ***** edo +	[%+:]+ [V:V]*	lexikoko + karak. behin edo gehiagotan bokala 0, 1 edo n aldiz
osagarria: \	\V	lexikoan bokala ez duen edozein bikote
kenketa: -	C-h	edozein kontsonante h izan ezik
ahaztea: /	V+ / h	bokalak h-ak kontutan hartu gabe

Posiblea da erregela bakoitzean testuinguru bat baino gehiago jartzea (**;** karakterea erabiliko da testuinguru bakoitzaren bukaeran).

Testuinguruan bikoteak zehaztea badago ere, askotan bietako bat zehaztea nahikoa da eta gainera honek erregela orokorrago bihurtzen du, gehiegi zehaztearen akatsari aurre eginez. Adibidez, testuinguru batean lexikoko k zehazteko `k:k` bikotea edo `k` zehaztea gehiegizkoa izan daiteke `k:g` bikotea ere zilegia delako; beraz, kasu horretan `k:` zehaztea da egokiena. Aldaketa fonologikoak gobernatzen dituzten erregeletako testuinguruan azaleko karaktereak izango dira nagusi, aldaketa morfologikoei dagokienetan aldiz, lexikoko karaktereak.

Adibideak hirugarren kapituluan azalduko dira. Izan ere oso komenigarria da Anworth-en liburuaren (1990) seigarren kapitulua kontsultatzea fenomeno morfologiko desberdinei dagozkien erregelen adibide zehatzak baitaude bertan.

II.3.2.4 Erregelatik automatara

Erregelatik egoera finituko automatara iragan ahal izateko konpiladoreak egon badaude baina hala ere interesgarria da eskuz nola egiten den jakitea, horrela erregela-mota desberdinen esanahia hobeto ulertuko baitugu.

l:i bikote bat bada, a:b ezkerreko testuingurua eta c:d eskuinekoa ikus ditzagun lau erregela motak eta dagozkien egoera finituko itzultzaileak. Kontutan hartu ondoko konbentzioak: bikoteetan lehen karakterea lexikokoa da eta bigarrena azalekoa, automataren 0 egoerak ezinezko bidea adierazten du, egoeraren ondoko bi puntu sinboloak bukaera-egoera eta puntu bakarrak ez-bukaerakoa. Karaktere bakarreko testuingurua suposatu dugu luzeagoa denean orokortzea oso erraza baita[1].

- testuinguru-murriztapena: **l:i => a:b _ c:d**

  gertatzeko testuingurua bete behar denez, ezkerreko testuingurua egiaztatu arte l:i bikotea ez da onartzen, eta ezkerreko testuingurua egiaztatu ondoren, eskuineko testuingurua egiaztatu arte gainontzeko bikoteak debekatzen dira eta egoera ez-bukaerakoa zehaztuko da.

```
    a l c =
    b i d =
 1: 2 0 1 1
 2: 2 3 1 1
 3. 0 0 1 0
```

- azalekoaren derrigortzea: **l:i <= a:b _ c:d**

  ezkerreko testuingurua egiaztatuz joaten da eta lexikoko l karakterea duen l:i ez den beste edozein bikote debekatzen da eskuineko testuingurua egiaztatzen denean.

```
    a l l c =
    b i = d =
 1: 2 1 1 1 1
 2: 2 1 3 1 1
 3: 2 1 1 0 1
```

[1] Ezkerreko zein eskuineko testuinguru luzekoa azpiautomata bat bezala ikus daiteke eta.

- aurreko bion konposaketa: **l:i <=> a:b _ c:d**

```
    a l l c =
    b i = d =
1: 2 0 1 1 1
2: 2 3 4 1 1
3. 0 0 0 1 0
4: 2 0 1 0 1
```

- debeku-ezarpena: **l:i /<= a:b _ c:d**

  testuingurua egiaztatuz joaten da eta l:i bikotea debekatzen da testuinguru horretan.

```
    a l c =
    b i d =
1: 2 1 1 1
2: 2 3 1 1
3: 2 1 0 1
```

Honetaz gehiago sakontzeko Antworth-en (1990) PC-KIMMO liburuaren hirugarren kapitulua kontsulta daiteke.

Eskuzko konpilazioa lagungarria da ondo ulertzeko erregelen sintaxia, baina, oso zaila ez bada ere, erregela-kopurua eta testuinguruen konplexutasuna handitu ahala lana neketsu bihurtzen da eta akatsak aurkitu eta zuzentzeko oso zorriketa nekagarria burutu behar da. Gainera eskuzko konpilazioaren ondoren gerta liteke erregelen esanahia eta automatena bat ez etortzea. Hau guztia ekiditeko zenbait konpiladore sortu dira, haien artean Koskenniemik (Koskenniemi, 85), Ritchie-ren taldeak (Ritchie *et al.*, 92) eta Xerox-eko Karttunen eta Beesley-ek proposatutako (1992)*Twolc* . Gainera erregelen arteko gatazkak detekta daitezke eta kasu batzuetan automatikoki konpondu ere. Gure proiektuan PC-KIMMOn bezala eskuzko konpilazioa egin badugu ere, gaur egun Xeroxeko Twolc tresna dugu erabilgarri.

Pena merezi du, hala ere, halako konpiladore baten nondik-norakoak zehazteak. *Computational Morphology* (Ritchie *et al.*, 92) liburuaren 7.3 kapituluan honetaz aipatzen dena oso baliagarria da zeregin honetan. Konpiladoreak burutu behar dituen urratsak honako hauek lirateke:

- Multzoak hedatu eta aldagaiak onartzen badira hauek instantziatu.
- Eragilearen motaren arabera erregela bakoitza automata bihurtu. Ezkerreko zein eskuineko testuinguruak bi automata bihurtu ondoren, erregela-mota kontutan hartuz ondokoa egiten da:

a) testuinguru-murriztapena (=>) bada korrespondentzian zehaztutako bikotearen bidez lotzen dira aipatutako automatak,
b) azalekoaren derrigortzea (<=) bada korrespondentziarena ez den baina lexikoko karaktere bera duen edozein bikoteren bidez lotzen dira, automata berria baztertze-automata gisa birdefinituz.
c) biak batera (<=>) direnean eskuineko testuinguruaren automata bikoiztu ondoren aipatutako loturak gauzatzen dira.
d) debeku-ezarpenaren (/‹=) kasuan derrigortzerenean egindakoa errepikatzen da baina lotura korrespondentzian zehaztutako bikotearekin eginez.

- Sortutako automatak ez-determinista direnez determinista bihurtu.
- Automatak bildu eta minimizatu.

Aurreko guztia eginkizun sinpletzat har daiteke kontutan hartzen ez bada helburua ez dela lengoaia bat ezagutzen duen automata bat egitea, momentuero berrabiatu behar den bat baizik. Gainera automaten artean gatazkak sor daitezke eta konpiladoreak, honetaz konturatzeaz gain, ebatz ditzake kasu askotan (Karttunen & Beesley, 92: 22-25).

II.3.3 Programa eta exekuzio-eredua.

Koskenniemik bere tesiaren laugarren kapituluan programaren zehaztasun guztiak ematen ditu. Zehaztasun handiegitan sartu gabe, berak proposatutakotik oso gertu dagoen gure inplementaziokoari buruz hitz egitean sakontasun handiagoz azalduko baita ondoko kapituluan, azpimarra ditzagun puntu garrantzitsuenak:

- Lexikoa eta erregela-sistema modulu independenteetatik kontrolatzen dira.
- Analisian, ilararen aukera bakoitzeko, azalaren arabera lexikoko karaktere parekagarriak bilatuz doa, ondoren bikotearen bideragarritasuna testuinguruan egiaztatzen da automatak mugituz, eta bideragarriak direnean analisi-ilaran kokatzen dira. Lexikoan morfema baten bukaera aurkitzean jarraitze-klaseari dagozkion azpilexiko guztion aukerak ilararatzen dira. Beraz, *backtracking*-ean oinarritutako prozedura da. II.5 irudian algoritmoa aurkezten da.
- Sorkuntzan —hasierako Koskenniemiren proposamenean lexikoko maila osoa duen sarrera suposatzen da— lexikoa ez da erabiltzen, beraz, automaten arabera sortzen dira azaleko aukera desberdinak, ilaran kokatzen direnak eta ondoren bazter daitezkeenak. Kasu honetan morfotaktika ez da egiaztatzen.

- Morfema (edo morfema-segida) batetik abiatuta sor daitezkeen forma flexionatu zein eratorri guztiak lortzeko, analisiaren algoritmoan aldaketa txiki pare bat egin behar dira: azalak gobernatu beharrean prozesua lexikoak gobernatzen du sarrera-morfema aurkitu arte, eta ondoren analisiaren prozedura jarraitzen da baina azaleko murriztapenik gabe.

```
algoritmoa analizatu
hasiera
Ilaratu_Hasierako_Lexikoak
bitartean Ilara_Ez_Hutsa
    hasiera
    egoera = IlarakoLehena
    baldin AzalaBukatuta eta BukaerakoMorfema eta AutomatakBukaeran
        orduan IrteeraAnalisia(egoera)
    baldin JarraitzekoEgoera orduan
        hasiera
        LexLortuUmeakEtaJarraitzeLexikoak
        egin UmeBakoitzeko
            hasiera
            baldin LexKaraktereEzabatzeaPosible
                orduan AutomatakMugituEtaIlaratu
            baldin LexKaraktereEtaAzalaPosible
                orduan AutomatakMugituEtaIlaratu
            amaia
        egin JarraitzeLexikoBakoitzeko
            Ilaratu
        baldin Azaleko_elipsia orduan Ilaratu
        amaia       (* aurreko egoera tratatua *)
    amaia           (* aukera guztiak *)
amaia       (* algoritmoa *)
```

II.5 irudia.- Analizatzeko sasialgoritmoa

Hobeto ulertu ahal izateko ikus dezagun *egingo* hitza analizatzen ari denean gertatzen den kasu bat. Azaleko *egi* ezagutu izan denean aukera desberdinak daude: lexikoan, besteen artean, *egiA*, *egiN* eta *eginbehaR* agertzen dira, Aren parekatze-karaktere posibleak A:a eta A:0 eta Nrenak N:n eta N:0 izanik lau bide irekitzen dira *egiN*-en kasuan aukera bikoitza baitago. *egiN:egin* aukera izango da arrakasta izango duen bakarra gainontzekoak antzuak dira eta: *egiA:egi* aukera morfotaktikak baztertuko du aurretik

bazterturik gelditzen ez bada erregela morfologikoen bidez, *egiN:egi* aukera *N:0* aldaketa gobernatzen duen erregelak eragozten du, eta *egin:egin* aukerak ondorengo karaktereetan egingo luke porrot lexikoan bikote onargarriak ez direlako aurkitzen.

II.3.4 Sistemaren gaineko kritikak eta proposamenak.

Bi mailatako morfologiaren gainean lan handia eta publikazio asko egin da 1983an egindako lehen proposamenetik. Garapen handi honen ondorioz bi mailatako formalismoen artean "kutsu" desberdinak bereizi arren, bere funtsa, morfofonologia deskribatzeko erregelak hain zuzen, bere horretan mantentzen da. Garapen horretan, aurretik aipatutako hobekuntzez aparte —erregela-mota berria (/<=), sintaxi esanguratsuagoa eta konpiladore automatikoa—, gertatutako kritikak eta proposamenak bilduko dira ondoko multzoetan:

- deskribapen-ahalmena
- diakritikoen beharra
- morfotaktika

II.3.4.1 Deskribapen-ahalmena.

Koskenniemiren proposamenaren gaineko kritika garrantzitsuenak morfotaktikari buruz izan badira ere, badaude morfofonologiaz —morfografemika, batzuen definizioan— aritzen diren aldaketa-proposamenak. Black-ek eta zenbait lankidek (Black *et al*., 87) ondokoa kritikatzen dute erregelen gainean: bikote bakarra egotea korrespondentzian eta erregela berriak idatzi ahala gatazkak eta nahasketak sortzea. Arazo hauek ebazteko beste erregela-eredu bat proposatzen da 1993an zehazten dena (Pullman and Hepple, 93). Eredu berri honen abantaila bakarra fenomeno morfologiko konplexuak adierazteko malgutasuna da, baina gure kasuan ez dugu egokitzat jo gutxitan agertu baitzaizkigu lehen artikuluan aipatutako arazoak.

Ritchie-ren (Ritchie *et al*., 92:181) taldeak ezaugarrietan oinarritutako erregelak erabiltzeko komenientziaz hitz egiten du baina ikerketa-lan handiagoaren beharra ikusten du horretarako. Aipaturiko lan hauetan oinarriturik Carter-ek (1995) *Core Language Engine* (Alshawi, 92) delakoarentzat egindako ekarpen berrian, beste berrikuntzez gain, erregelen formatoaren aldaketa proposatzen du, bertan ezaugarrien erabilera proposatuz. Hala ere erregeletako korrespondentzian karaktere bat baino gehiago adierazteko aukera eman arren, karaktere bakar batera murriztea gomendatzen du.

Beste aldetik, kateatze-morfologia ebazteko diseinatua izan zen hasiera batean bi mailatako morfologia. Hala ere formalismo honetan oinarrituta erantzuna eman nahi izan zaio kateatze-morfologiatik kanpo geratzen diren zenbait arazori: artizkiak, bikoiztea eta hizkuntza semitikoen *erro-patroi* motako fenomenoak. Euskararen kasuan aurkitzen ez badira ere oso labur aipatuko ditugu lerro honetan egindako lanak.

Lehen kasurako Antworth-ek (1990:156) tagalogeraz erabiltzen den *in* artizkiaren arazoa ebazteko —lehen kontsonantearen atzean kokatzen dena— ondokoa proposatzen du:

- marka (X) batez ezagutzen da artizki hau lexikoan, eta aurrizki bezala adierazten da morfotaktikaren aldetik.
- erregela baten bidez 0:i eta 0:n (in-en sorrera) bikoteak onartzen dira ezkerreko testuinguruan X marka badago.

Beste kasuetan halako erregela(k) asmatzea zailagoa da ez-naturalak baitira, eta gainera konputazioaren ikuspuntutik, erregela hauek konplexutasuna igotzen dute.

Bikoiztearen arazoan PC-KIMMOren egileak (157-159 orr.) antzeko zerbait proposatzen du tagalogeraz bikoizten den lehen kontsonante-bokal silabaren bikoizketaren kasuan. Bi morfofonemaren bidez adierazten da edozein bikoizte lexikoan eta erregela pare batez kontrolatzen da morfofonema hauen gauzatzea azaleko mailan. Hala eta guztiz ere adibide honetan sinplea dena beste kasuetan askoz konplexuagoa gerta daiteke. Adibidez, walpirieraren bikoizteetarako hamalau mila egoera inguruko automata bat aurrikusten du Sproat-ek.

Hizkuntza semitikoetarako ere zenbait proposamen egin dira bi mailatako morfologian oinarriturik. Inportanteenak izan dira Kay-k 1987an proposatutako 4 mailatako eredua eta 1990ean Beesley-k egindako “desbiderapena”, lexiko anitzen arteko bi mailatako prozesua aplikatzen duena. Kay-ren proposamenean lau maila edo zinta proposatzen dira: sarrerakoa edo azalekoa lehena, kontsonante-erroena bigarrena, patroiena hirugarrena eta bokal-morfemena laugarrena. Egoera finituko itzultzaileak bi zintatatik irakurri beharrean lautatik irakurtzen du, baina zintaren batean ez aurreratzea adieraz daiteke kode batez. Teoria honetan oinarriturik eta lehentxeago aipatutako Pullman-en erregela-eredu berria erabiliz Kiraz-ek (1994) inplementazio bat proposatzen du. Beesley-renean aldiz, ohiko 2 mailak erabiltzen dira baina bi lexiko bereizten dira erroena eta bokalekin konpilatzen den patroiena, biak *trie* egiturarekin. Patroienak agintzen du baina kontsonanteei dagokien hutsuneak aurkitzean lexiko-trukea gertatzen da, horrela bi lexikoak ireki eta ixten

direlarik[1] . Bokalizazioak sortzeko lexikoko bokalen desagerpena onartzen da erregelen bidez, horrela bokalizazio partzialak ere onartzen direla.

II.3.4.2 Hautapen-markak edo diakritikoak.

Lexikoko karaktere hauek erregelen aplikazioa kontrolatzeko erabiltzen dira. Bide eskas eta nahasgarritzat hartu izan den honek abantaila bat dakar: erregelen sintaxia oso sinplea da bere osagai bakarrak karaktere-bikoteak eta eragileak baitira.

Horren truke lexikoan agertzen diren osagai ez-naturalak dira karaktere hauek, hizkuntzalariaren lana narrasten dutenak eta datuen ulergarritasuna galarazi. Honen aurrean zenbait proposamen agertzen dira bibliografian: Bear-ena eta Trost-ena.

Bietan ideia nagusia bera da, lexikoko elementuek dituzten ezaugarri morfologikoen bidez erregelen aplikazioa baldintzatzea; baina lehen proposamenean (Bear, 1988) erregelak anulatzeko ezaugarri berezi bat eransten den bitartean, bigarrena ahalmentsuagoa da (Trost, 91), gainontzeko ezaugarri morfologikoek baldintza baitezakete erregelen aplikazioa.

Gure proiektuan markak mantendu ditugu bi arrazoi nagusirengatik: batetik morfotaktika egoera finituko mekanismoen bidez burutzen delako —kontuan hartu behar baita aurreko bietan morfotaktika burutzeko ezaugarri morfologikoetan oinarritutako baterakuntza-mekanismoak proposatzen direla—, eta bestetik eskuragarri dauden tresnetan (PC-KIMMO, Alvey, Twolc, etab.) halakorik ezin delako erabili.

II.3.4.3 Morfotaktika: jarraitze-klaseak vs. baterakuntza-mekanismoak.

Koskenniemik proposatutako formalismoari aldaketa morfofonologikoen trataeratik datorkio arrakasta eta, horregatik, hasierako izena hori izan gabe, idazle askok *bi mailatako fonologia*ren izena egokitu diote. Horren ondoan, proposatutako morfotaktika —hitzaren gramatika edo sintaxia beste batzuen terminologian— oso pobrea da, zeharo lineala baita, morfema bakoitzean ondoren etor daitezkeenen multzoa zehaztea izanik morfotaktika adierazteko aukera bakarra. Bide honi jarraitu diote zenbait inplementazio eta tresna: Karttunen-en inplementazioa (1983), PC-KIMMO (Antworth, 1989), Xerox-eko Lexc (Karttunen, 1993).

Morfotaktikaren aldetik aipatutako pobrezia horrek duen arazo nagusia *urruneko menpekotasun*ari dagokiona da, hau da, morfema baten agerpenak ondo-ondokoa ez den beste morfemen agerpena baldintzatzen dueneko kasua. Adibidez, ingelesez *en*, *joy* eta

---

[1] Gogoratu behar da ideia hau ATEF izeneko prozesatzailean agertzen zela.

able morfemak zilegiak dira eta hirurak jarraian joan badaitezke ere, azken biak bakarrik ezin dira lotu. Alegia, *en*-ek *able*-en agerpena baldintzatzen du, edo beste era batean esanda, *en* eta *able*-ren artean urruneko menpekotasuna dago.

Eredua aldatu gabe arazo honen ebazpena bihurria da. Bi modutan egin daiteke, erregela baten bidez edo morfotaktikaren bidez buru daiteke ondoko prozeduraretako bat aukeratuz:

- urruneko menpekotasun-erlazioa duten morfemak markatu, baldintzatzailea hautapen-marka batez (bukaeran) eta baldintzatua beste batez, eta erregela batek bigarrenaren gauzatzea kontrolatuko du ezkerreko testuinguruaren arabera.
- tartean egon daitezkeen morfemek osatzen duten azpilexikoa(k) bikoiztu, bat aurrizkia onartzeko eta bestea ez onartzeko, atzikien eraketarako bakoitzari jarraitze-klase desberdin bat emanez. Aztertutako adibidean *en* har dezaketen morfemen jarraitze-klaseei morfema baldintzatua (*able*) egokituko zaio. Bigarren irtenbide hau ez da gomendagarria kasu hoentan alomorfo asko sortu behar baita.

Beste proiektu batzuetan morfotaktikaren eredua zeharo aldatzea proposatu da, bi mailatako morfologiaren jatorrizkoa oso pobrea dela eta. Proposamen ezagunenak Bear-ek (1986), Trost-ek (1990, 1994), eta Alvey-n (Ritchie et al., 87; Ritchie et al., 92) azaldutakoak izanik. Hiruretan baterakuntza-mekanismoak proposatzen dira; Bear PATR formalismoan oinarritzen den bitartean, Ritchie-ren taldean GPSG[1] aukeratzen dute. Alvey-ren kasuan baterakuntzak morfotaktika burutzea baino helburu handiagoa du, eratorpenak sortutako kategoria aldaketa zein elkarketen eta informazio morfosintaktikoaren tratamendua bideratzen baitu[2] —aipatutako liburuaren (1992) hirugarren, laugarren eta bosgarren kapituluak oso gomendagarriak dira—. Honetaz arituko gara luzeago III.4 pasartean.

Trost-en kasuan azpimarratzekoa da alemaneraren umlaut izeneko fenomenoa ebazteko egindako lana. Carter-ek (1995) hauekin bat dator aipatu den lanean.

Baterakuntza-mekanismoen alde aipatzen diren abantailak hauek dira: potentzia, malgutasuna eta sintaxiarekin bat etortzea. Moreno Sandoval (1991) EUROTRA proiektuaren barruan, eta bi mailatako morfologiari jarraitu gabe, eredu hauen alde

---

1 Formalismo hauek ez ditugu azalduko sintaxiaren gaitzat hartu izan baitira eta gure aplikazioan ez dira erabili. Hala ere, honetaz sakontzeko honako liburu hau gomendatzen dugu: Shieber S.M. *An introduction to unification-based approaches to grammar*. CSLI Lecture Notes 4. Chicago U. Press. 1986.

2 Tratamendu honi *morfosintaktikoa* deituko diogu eta ez da harritzekoa kasu honetan halako tratamendua burutzen *hitzaren gramatika* terminoa erabiltzea eta ez morfotaktika.

agertzen da beste aldeko irizpide bat aipatuz, inplementazio erazagutzailea. Izan ere baterakuntzak aurkako irizpide bat du, eraginkortasunarena hain zuzen. Eraginkortasun-galera horrez gain aldaketa honekin morfofonologia eta morfotaktika prozesu banatu eta sekuentzial bihurtzen dira, eraginkortasunaren aldetik kaltegarria izan daitekeena, konputazio-konplexutasuna aztertzean azalduko dugun bezala.

II.3.5 Ekarpen bat: jarraitze-klase hedatuak.

II.3.5.1 Deskripzioa

Gure proiektuan morfotaktikari ekiteko garaian tarteko irtenbide bat hartu dugu honako filosofia honi jarraituz: hasierako eredua ahalik eta gutxien aldatuz urruneko menpekotasuna adierazi ahal izatea. Horretarako jarraitze-klase hedatuak proposatzen ditugu (Agirre, Alegria, *et al*. 92). Hitzaren gramatikak dotoreagoak eta ahaltsuagoak badira ere, proposatutako mekanismoaren alde arazoa ebazteko nahikoa izanik oso mekanismo sinplea dela argudia daiteke. Hala eta guztiz ere oso interesgarritzat jotzen dugu Alvey-n proposatutako bidea morfotaktikarengatik baino tratamendu sintaktikoarengatik.

Euskararen morfotaktika nahikoa sinple eta lineala izanda ere urruneko menpekotasuna izenaz deskribatu dugun fenomenoa agertzen da. Ondoko kapituluan honi berrekiteko tokia egongo bada ere, bi kasu azalduko dugu orain, proposatutako morfotaktika-sistemaren aplikazioa adibide hauekin erabiltzearren:

- Aditz jokatuaren hizkiak. Aditz hauei zenbait aurrizki eta atzizki lot dakieke. Aurrizkiak *ba* baldintzazko morfema, *ba* baieztapenekoa eta *bait* kausala dira. Atzizkien artean *la* konpletiboa eta *n* erlatiboa ditugu. Aditzak morfema bakar bat hartzen duenean ez dago arazorik baina aurrizki batzuek debekatzen dituzte beste atzizki batzuk. Horrela *ba* indarrezkoa *la*-rekin konbinatzea zilegia den bitartean, *ba* baldintzazkoa eta *bait* ezin dira harekin konbinatu.

- *Nor-nori-nork nor-nork* eta *nor-nori* motako aditz laguntzaile eta trinkoen eraketa. Aditz hauetan pertsona berari dagozkion morfema konbizazio batzuk ez dira zilegiak—singularreko zein pluraleko lehen eta bigarrenekoak—, eta erroa tartean egon daitekeenez urruneko menpekotasunaren kasuaren aurrean gaude. Horrela, *nor-nork* laguntzaileak orainaldian honako patroi honi jarraitzen dio: *nor-morfema+erroa+nork-morfema* baina *naut* —*na*(nor-1.perts-sing) + *u*(ukan

laguntz-orainaldia) + *t* (nork-1.perts-sing)— ezinezkoa da[1]. Gauza bera gertatzen da *hauk* (2.perts-sing / 2.perts-sing-mask), *haun* (2.perts-sing / 2.perts-sing-mask), *naugu* (1.perts-sing / 1.perts-sing-plu), *gaitut* (1.perts-plur /1.perts-sing) eta beste batzuekin.

Adibide hauek azaldu ondoren, jarraitze-klase hedatuak zertan dautzan aztertzeko momentua iritsi da. Izenak dioen bezala, jarraitze-klaseen hedapen bat da eta hedapen honetan deskribapen-ahalmen aldetik bi posibilitate gehiago eskaintzen dira: debekuak eta jarraitze-klaseen zuhaitzak.

Debekuak jarraitze-klaseari eransten zaizkio eta morfema honen atzetik agertu ezin duten —debekatuta daudela esango dugu— azpilexiko-multzoak zehazten dira beraietan. Debekuak bereizteko ken sinboloa erabiltzen da aurrizki gisa. Horrela lehen adibidean agertzen zen *bait* aurrizkiak lexikoan duen definizioa honako hau da:

bait (ADITZ_JOK - LA - N)[2]

Honekin adierazten dena zera da: aurrizkiaren ondoren ADITZ_JOK jarraitze-klase arruntari dagozkion azpilexikoetako morfemak etor daitezke baina atzerago etor litezkeen morfemak edo atzizkiak ezin dira LA edo N jarraitze-klaseei dagozkien lexikoetako morfemak izan. Beraz, debekuek jarraitze-klase arrunt batean urruneko murriztapenak ezartzen dituzte.

Jarraitze-klaseen zuhaitz batek zuhaitz-moduko “jarraitze-bideak” zehazten ditu parentesien bidezko espresio batean zehaztuta. Ondo-ondoko morfemari dagozkion azpilexikoak bakarrik zehaztu beharrean ondoko maila desberdinetakoak zehatz daitezke, maila bakoitza sekuentziako morfema bati dagokiolarik. Zuhaitza zerrenda bat bezala adierazten da non maila berri bat adierazteko parentesi bat irekitzen den eta maila berean jarraitze-klase arrunt bat baino gehiago bereizteko koma karakterea erabiltzen den.

Nor-nork egiturarako definizio batzuk hauek lirateke:

na (ERROA (NORK23))

ha (ERROA (NORK13))[3]

---

1 Egungo inplementazioan aditzaren formak oso-osorik daude lexikoan arrazoi praktikoak direla eta; hala ere aipatutakoa jasotzen duen eredu teorikoa landuta dago.

2 Multzo hau etiketa batez adierazten da beste edozein jarraitze-klase bezala.

3 Hau lortzeko pertsoneei dagozkien azpilexikoa beste txikiago batzuetan banatzera behartuta gaude.

Honekin adierazten dena argi da erroaren ondoren etor daitekeena pertsona batzuei dagokien nork kasua dela. Adibide honetan ondoko morfemen murriztapenerako erabili bada ere —konturatu honetarako debekuak erabil zitezkeela—, zuhaitz hauekin aukera dago ondoko morfemen esparrua zabaltzeko edo murrizteko. Horrela, ingelesezko aztertutako adibidearen kasuan, *en-joy-able* , irtenbide bat eskaintzen digu guk proposatutako hobekuntza honek. Honako hau egin beharko litzateke urruneko menpekotasunaren arazo hau ebazteko:

joy JK_ADITZ

en (EN_ADITZAK (JK_ADITZ, ABLE))

EN_ADITZAK jarraitze-klasean *joy* aditza duen azpilexikoa badago eta ABLE-n *able* atzizkia duena aipatutako arazoa konponduta dago.

II.3.5.2 Sintaxia

Jarraitze-klase hedatuen sintaxia honako hau litzateke:

```
<jarraitze-klase hedatua> ::=    <jarraitze-klasea>
                                 | <jarraitze-klaseen arbola>
                                 | <jarraitze-klase murriztua>
<jarraitze-klaseen arbola> ::=   “(“<jarraitze-klaseen zerrenda>
                                   [<jarraitze-klaseen arbola>]”)”
<jarraitze-klase murriztua> ::=  “(“<jarraitze-klasea>
                                   <jarraitze-klase ezeztatua>*”)”
<jarraitze-klaseen zerrenda> ::= <jarraitze-klasea>
                                   [“,”<jarraitze-klasea>]*
<jarraitze-klase ezeztatua> ::=  “-”<jarraitze-klasea>
```

II.3.5.3 Semantika

Semantikaren aldetik ondoko definizioak proposatzen ditugu:

Izan bedi W hitza, $W=m_1+m_2+...+m_n$, non m_i $(1 \leq i \leq n)$ morfemak diren.

- **Ohiko jarraitze-klaseekin** morfema bakoitzari jarraitze-klase bat egokitzen zaio (m_i -> jk_i) eta jarraitze-klase horrek lexiko multzo bat definitzen du ($lex(jk_i)$).

  Zera egiaztatu behar da:

$\forall i\ (1 \leq i < n : m_{i+1} \{ lex(jk_i))$
eta m_n bukaerako morfema da.

- **Jarraitze-klase hedatuekin** morfema bakoitzari jarraitze-klase hedatu bat egokitzen zaio (m_i -> jkh_i), eta jarraitze-klase hori arrunta (jk), arbola (jka) edo murriztua (jkm) izan daiteke. Izan bitez
  $jkm_i = jk_i$ - deb_{i1} - ... - deb_{ip}
  $jka_i = jk_i$ **(**azp_{i1} **(** ... **(**azp_{ip}**))**
  non azp_{ij} eta deb_{ij} jarraitze-klase arruntak diren.

  Zera egiaztatu behar da:
  $\forall i\ (1 \leq i < n : m_{i+1} \{ lex\_hed_i)$
  non
  $lex\_hed_i = (lex(aurre_i « jk_i)) - lex(deb_i)$
  a«b = baldin a•Ø a, bestela b
  $aurre_i = azp_{jk} : j+k=i$ ñ $j = max_j \mid j<i$
  $deb_i = :(deb_{jk}) : j<i$
  eta m_n bukaerako morfema da.

II.4 Bi mailatako ereduaren konputazio-konplexutasuna eta azkartzeko bideak.

Koskenniemik, bi mailatako morfologia enuntziatu zuenean, egokitu zion ezaugarrietako bat eraginkortasuna izan zen. Hasiera batean honetaz asko eztabaidatu ez bazen ere ondoren oso gai polemikoa izan da, eta horren froga ondoko lanak dira: (Barton, 86), (Barton *et al.*, 87), (Koskenniemi & Church, 88), (Sproat, 92: 3.5).

II.4.1 Eraginkortasunaren aldetiko arazoak.

Bi mailatako morfologian konplexutasun-iturri nagusia erregeletatik dator. Lexikoko zein azaleko karaktere bat beste mailako batekin baino gehiagorekin egokitu ahal izatean, analisian zein sorkuntzan bide bati baino gehiagori jarraitu beharko zaio zenbait momentutan, eta honen ondorioz *backtracking*-a ekidin ezinezkoa izanik (ikus kapitulu honetan emandako algoritmoa). Bide honetatik eta intuitiboki honako ondorio honetara heltzen gara: zenbat eta aldaketa gehiago onartu, eta zenbat eta testuinguru murriztugabeagoa izan aldaketa horietan, orduan eta eraginkortasun-galera handiago gertatuko da aztertzeko bideak gehiago dira eta. Gainera, lexikoko karaktereak desagertzea dagoenean, ezabapenak onartzen direnean alegia, bide-kopurua handitzen da

ikaragarri, posizio bakoitzean ezabatzen den edozein lexikoko karaktereren agerpena kontutan hartu behar baita.

Erregelak konplexutasun-iturri nagusia badira ere —ondoko pasartean beraiei dagokien konplexutasuna baino ez dugu aztertuko— lexikoa ere konplexutasun-iturri bada. Hona hemen lexikoarekin lotutako zenbait puntu eraginkortasunarekin zerikusia dutenak:

- Jarraitze-klase bati dagozkion lexiko anitzak. Honek zera esan nahi du: analisia egitean bide asko jorratu beharko dira, haien artean gehienak antzuak direla.
- Hasierako azpilexiko anitz egoteak aurretik aipatutako ondorio berbera dakar. Nahiz eta bitxia irudi ahal izan, hasierako lexiko anitz egotea arruntzat jo behar da, egoera finituko morfotaktika mantentzen bada gutxienez, aurrizkiek baldintza baititzakete ondoren doazen lemak.
- Analisi-prozesuan erregelek sortzen duten konplexutasuna lexikoak murritz dezake, posible diren aukeretako bat lexikoan ez dagoenean. Sorkuntzan aldiz ez dago halako murriztapenik, beraz bide guztiak jorratuko dira.

II.4.2 Konputazio-konplexutasuna zehaztuz.

Eraginkor izatearen hasierako uste hura honetan oinarritzen zen: egoera finituko makinak oso sinpleak eta azkarrak dira eta egoera- zein arku-kopuruak ez du eragin handia abiaduran, zenbait inplementaziotan frogatu ahal izan zen bezala.

Barton izan zen gai honi sakonean eta formalki ekin zion lehena eta eztabaidaren sortzailea. Bere argudiotan sakondu gabe —honetaz sakontzeko 1987an publikatutako liburua (Barton *et al.*, 87) kontsulta daiteke— bere arrazonamendua eta ondorioak ondoko puntu hauetan labur daitezke:

- Ereduaren konputazio-konplexutasuna kalkulatzeko konplexutasun ezaguneko problema baliokide bat bilatzen du, teknika honi *laburtzea* esaten zaio.
- Bi mailatako morfologia erabiliz egindako sorkuntza *Boolean satisfiability* (SAT hemendik aurrera) problemara laburgarria dela aurkitzen du.
- Ondorioz, sorkuntza NP-gogorra dela esan dezake.
- Antzeko bidetik ezagutza edo analisiaren konplexutasuna aztertzen du eta konplexutasun berekoa dela dio mugarik gabeko ezabapenak ez badira onartzen, zeren eta kasu horretan konplexutasuna handiagoa baitagokio, PSPACE-gogorra hain zuzen.

Ondorioz esan daiteke bi mailatako eredua ez du zertan eraginkorra izan behar, beraren bidez oso problema konplexuak kode baitaitezke. Barton urrutiagora doa eta desegokitzat jotzen du, ez baitu bereizten lengoaia naturalaren problema bat —morfologiarena— konplexuago diren besteetatik.

Aurrekoa ikusita, Koskenniemik eta Church-ek (1988) erantzuten diote praktikan oinarrituz eta honako ideia hauek azpimarratuz:

- Barton-en frogapena ondo dagoela baina laburtzeko aukeratutako problema dute desegokitzat.
- SAT motako problemetan denbora modu esponentzialean hazten den bitartean, hizkuntza desberdinetarako bi mailatako morfologiaren inplementazioetan lineala dela frogatutzat ematen dute.
- Aurkitutako kasu konplexuenean, urruneko murriztapenenean eta horren barruan bokalen armoniarenean, abiadura ez da esponentzialki hazten; gainera hau ez da ohiko fenomenoa.

Laburbilduz, beren ustez bi mailatako eredua egokia da, hizkuntza anitz desberdinetako morfologia modu eraginkorrean deskribatu da eta.

II.4.3 Proposatutako hobekuntzak.

Aurretik esandakoaz honako gomendio hauek luza daitezke bi mailatako prozesadore morfologiko eraginkorrak egiteko:

- Erregelen aldetik ahal den neurrian murriztapenak ezkerreko testuinguruan jartzen saiatzea aukerak lehenago murriz daitezen, eta ahalik eta ezabapen gutxien zehaztea, maiztasun handiko karaktereen ezabapena ekidituz.
- Lexikoaren eraginaz konplexutasuna haz ez dezan, lexiko bakarreko edo gutxiko jarraitze-klaseak hobestea. Izan ere azken irizpide honek alomorfoen erabilera bultzatzen du eta Koskenniemik aipatutako beste baten kontra doa, morfemak errepikatzea baino lexiko txiki anitzen erabilera bultzatzen zuen eta.

Gomendio taktiko hauez gain zenbait aldaketa proposatu dira mota honetako prozesadore morfologikoak azkartzearren; haien artean bi hauek azpimarratu daitezke[1]: lexiko anitzei aurre egiteko Bartonek proposatutako lexikoen fusioa eta Karttunen-ek proposatutako lexiko-itzultzaileak.

---

1 Kasu batzuk ezin dira azaldu arrazoi komertzialak direla eta dokumentaturik ez daudelako.

II.4.3.1 Lexikoen fusioa.

Oinarrizko ideia oso sinplea da, lexiko guztiak bakar batean biltzen dira —edo logikoen izan daitekeena, hirutan: aurrizkiak, erroak eta atzizkiak— horrela bilaketa lexiko bakar batez burutzen da, bide-kopurua murriztuz eta hiztegia trinkotuz —*trie* egitura horrela trinkoagoa baita—.

Arazo bat sortzen da ordea, morfotaktikarena. Lexikoaren egitura aldatzen ez bada behintzat, morfotaktikari buruzko informazioa —azpilexikoak eta jarraitze-klasea— zuhaitzaren hostoetan dago, beraz bide desegokiak aukera daitezke jorratzen jarraitzeko morfofonologiaren aldetiko murriztapenik ez dagoen bitartean, baita alperrik jorratu ordea, morfotaktikak bide horiek debekatzen dituenean. Hauxe bera gertatzen da morfofonologia eta morfotaktika prozedura independiente gisa inplementatzen direnean.

Beraz, abiaduraren ikuspuntutik fusioaren interesa zalantzazkoa da, alde batetik irabaz daitekeena bestetik gal daitekeelako. Espazio kontuan bere egokitasuna zalantzarik gabekoa da. Euskararekin guk egindako fusioko probetan, ondoko kapituluan zehaztuko diren datuetan oinarriturik, abiaduraren aldetik hobekuntza aipagarririk ez dela lortzen ondorioztatzen da. Bere momentuan luzatuko gara honetaz.

II.4.3.2 Lexiko-itzultzaileak.

Karttunen-en proposamena (Karttunen *et al.*, 92), (Karttunen, 93), (Karttunen, 94) harantzago doa. Berak proposatutako *lexiko-itzultzaile*ek konplexutasunaren aldetik dakarten iraultzaz gain, bestelako abantailak ere dauzkate formalismo morfologikoaren funtsaren ikuspuntutik: batetik lexikoko adierazpide arbitrarioak ekiditen dira forma kanonikoa bultzatuz eta lexiko mailan karaktere berezirik egon beharrean informazio morfologikoa egonda; bestetik tarteko adierazpideak eta erregela-multzo anitzak konbinatuz deskripzio morfologikoa erraz daiteke eta deskribapen-ahalmena handitu[1].

Horretarako forma flexionatuak forma kanonikoekin ezkontzen ditu —adibidez *better* eta *good*—, nahiz eta horretarako lexiko eta azalaren arteko distantzia handitu. Distantzi handitze honi aurre egiteko tarteko egoerak onartzen dira lexikoko eta azaleko mailen artean, Kaplan & Kay-ren ideia zaharrak berreskuratuz.

Eredu honen funtsa ideia hauetan datza eta II.6 irudian isladatzen da:

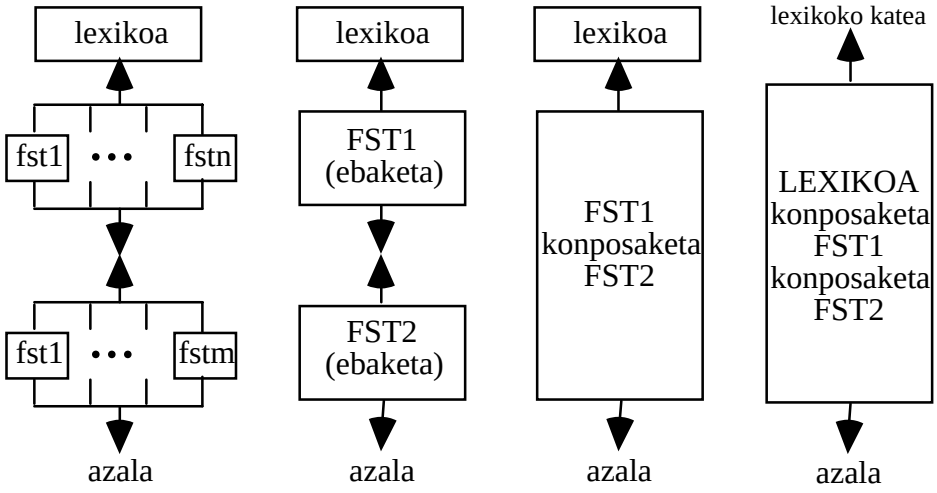
---

[1] Kaplan eta Kay-ren (1994) ideietan oinarriturik ere, Carter-ek (1995) *Core Language Engine* delakoproiekturako egindako lanean antzeko ideia proposatzen du, baina konpilazioan hizkiak sartzen baditu ere erroak kanpoan gelditzen dira sistema malguagoa izan dadin, horretarako eraginkortasuna galtzen duen arren.

- Normalean tarteko maila bat bereizten da, ohiko bi mailatako ereduan lexikokoa zena (adierazpide arbitrarioak, diakritikoak eta guzti). Lexikoko mailan informazio morfologikoa sartzen da baina ez hostoetan, arkuetan baizik.

- Ondoan dauden mailen arteko aldaketak bi mailatako morfologiari dagozkion egoera finituko itzultzaileen bidez gobernatzen dira, guztiak bakar batean konpila daitezkeenak —horretarako *twolc* konpiladorea dagoela.

- Tarteko egoeren arazoaren aurrean, sekuentzia ekartzen duena, itzultzaileen arteko konposaketa[1] proposatzen du, itzultzaile bakar bat lortuz —honetaz *lexc* konpiladorea arduratzen da.



II.6 irudia.- Itzultzaileen arteko ebaketa eta konposaketa lexiko-itzultzaile bat lortzeko (Karttunen *et al.*, 92)

- Lexikoko maila eta ondoko tarteko mailaren arteko bi mailatako erregelak idatzi beharko lirateke, itzultzaile paralelo eta bateragarriak sortzeko ohiko bidea baita. Erregela hauek asko eta oso lokalak liratekeenez, horren ordez *lexc* konpiladoreak bikoteak onartzen ditu —*be+Pres+Sg+P3+Verb : is* da horren adibide bat— bi maila horien arteko parekatzea definitzen dutenak —*lexc* konpiladorea arduratzen da bihurketa hauei dagozkien grafoen eraketaz.

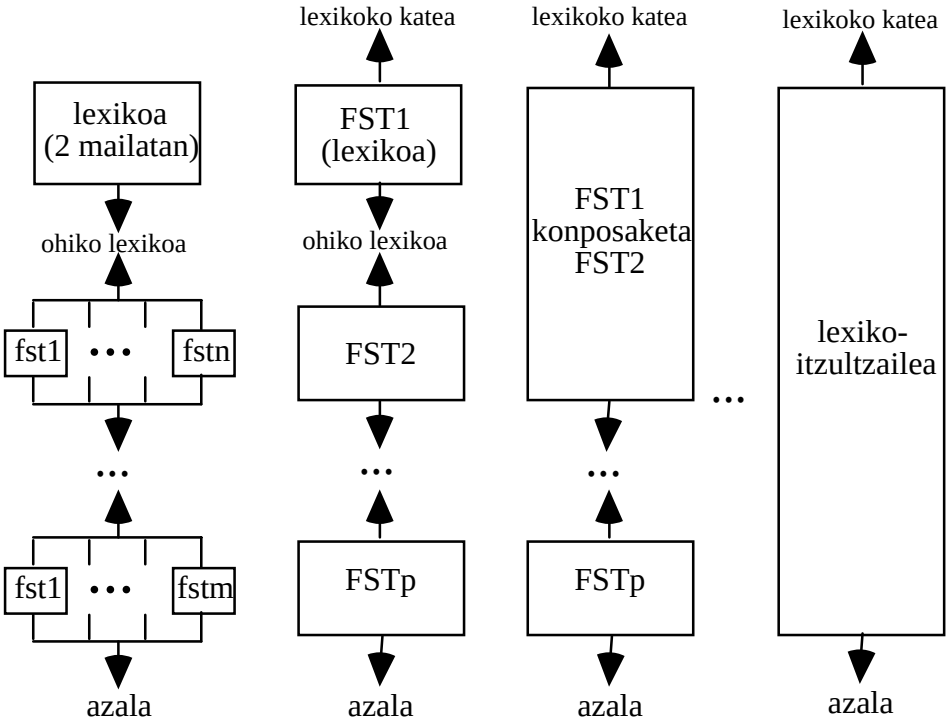
- Lexikoko mailaren eta aurretik zegoen itzultzailearen arteko konposaketa itzultzaile bakar batean sistema osoa edukiz burutzen da. Hau da funtsezkoena eraginkortasunari begira.

---

[1] Hau sinplifikazio txiki bat da. Errealitatean gertatzen dena ez da ondoko maila bakoitzeko ebakidura lehen eta ondoren konposaketa, baizik eta Karttunen-ek (1994) deitzen duen ebakitze-konposaketa (*intersecting composition*) prozesua.

Diseinu hau Kaplan-ek eta Kay-k (Kaplan, 88) (Kaplan & Kay, 94) egindako ekarpen teorikoan oinarritzen da, II.3.2.1 pasartean azaltzen dena.

Diseinu hori praktikan jartzean espero zutena baina askoz ere emaitza hobeak lortu zituzten. Erregelen ebaketa eta konposaketaren ondorioz lortzen den egoera-kopuruaren magnitude-ordena lau edo bost zifra hamartarrekoa da, teoriaz izan zitezkeen hamabi zifretakoetatik oso urrun. Beste aldetik, lexikoarekin konposaketa burutu ondoren egoera zein arku kopurua laburtu egiten da hasiera batean asko handitzea espero zitekeenean; nonbait lexikoak abstrazioa murrizten du eta.



II.7 irudia.- Lexiko-itzultzaile orokor bat lortzeko urratsak (Karttunen 94)-n oinarritua.

Lortutako emaitzak ikaragarriak dira, frantseserako honako datu hauek ematen dituztelarik: 50 K-egoera eta 100 K-arku inguru, denak Mega bat baino gutxiago hartzen duena[1] eta zenbait mila hitz/segundo abiadura-ordena analisian. PC-KIMMO baina ehundaka aldiz azkarrago.

Bukatzeko hausnarketa bat: lexiko-itzultzaile hauek bi mailatako morfologiaren hobekuntza dira edo eredu berri bat? Kapituluaren hasieran egiten genuen sailkapenera eta kasu-errebisiora itzuliz Tzoukermann-ek eta Liberman-ek proposatutakoarekin du zerbait

[1] Honetan kodetzeko teknikak ere badu bere garrantzia. Honetaz gehiago sakontzeko (Karttunen, 90) erreferentzia da lagungarri.

amankomunean: erregelak desagertu dira exekutatzen den prozesadore morfologikotik. Honen ondorioz, exekuzioaren ikuspuntutik eredu berri baten aurrean gaudela esan badaiteke ere, lengoaia baten morfologia deskribatzeko garaian bi mailatako eredutik oso gertu dago erregelak bi mailatakoak baitira.

Lexiko-itzultzailearen kasuan eredu berri baten aurrean gaudela argiagoa da, erregela-sistema desberdinen konposaketak eskaintzen duen deskribapen-ahalmena batetik, eta deskripziorako erraztasuna bestetik, bide berriak irekitzen baititu (ikus II.7 irudia). Beraz, bi mailatako eredu batetik askoz orokorrago eta ahalmentsuago den maila anitzeko beste batera pasa gara, aplikazio-esparruk asko zabaltzen delarik. Aplikazioez sakontzeko interesgarriak dira Chanod-ek (1994) egiten duen frantses-aditzaren deskribapena eta Kwon-ek eta Karttunen-ek (1994) egiten duten korearraren deskribapen inkrementala.

Hurrengo kapituan, euskararen gaineko aplikazioan hain zuzen ere, lexiko-itzultzaile hauen ahalmenaz eta dagozkien tresnen erabileraz sakontzeko aukera egongo da.

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 labur tua.

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+a	semearen (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.

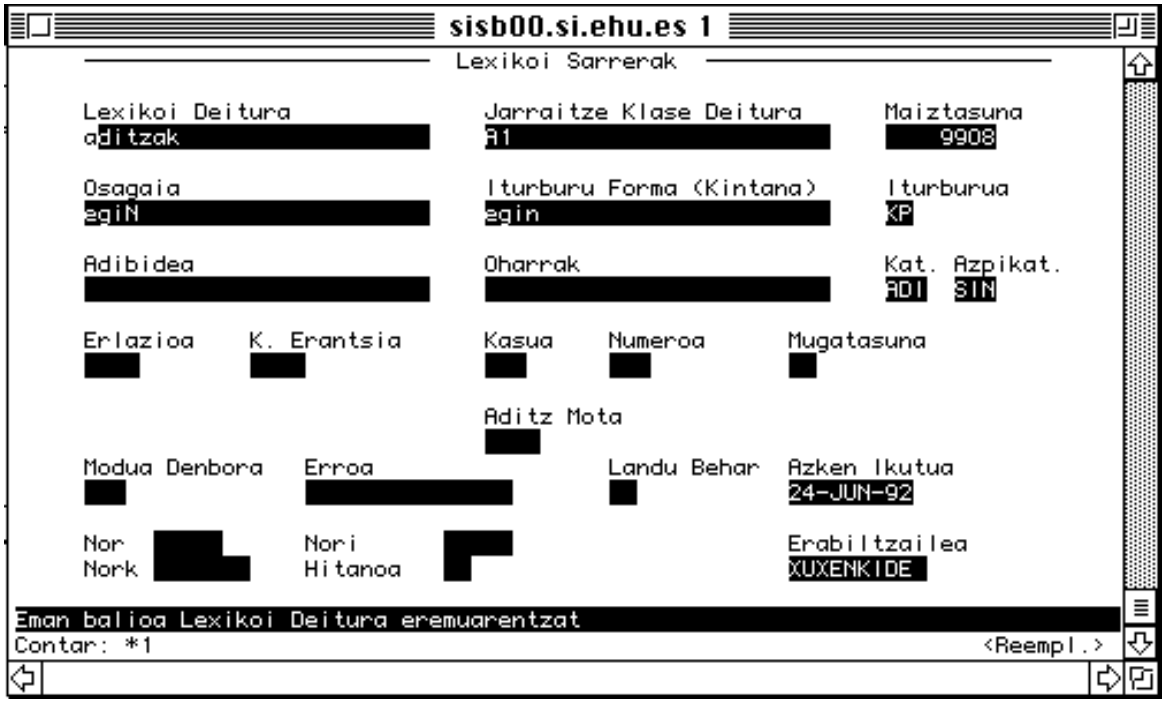
Hasiera batean VAXeko RDB softwarea erabili izan bazen ere, ondoren SUNeko ORACLEra eramana izan zen, gaur egun ORACLEren bidez kudeatzen delarik. Beraz, eredu erlazionalari jarraitzen dio, baina etorkizunerako, objektuei zuzendutako diseinu berri baten gainean ari gara lanean, hartzen ari den konplexutasunari 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.

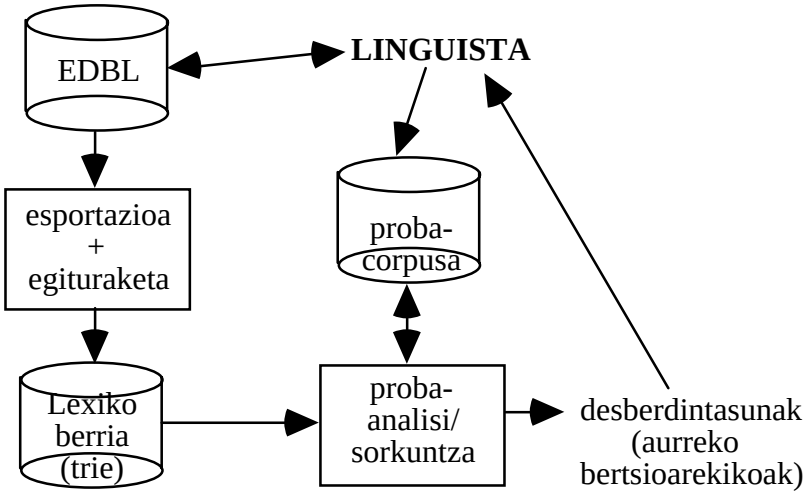
7) **Aldaeren azpilexikoak**: aldaeren trataerarako erabiltzen dira taula hau eta ondoko biak, eta hurrengo kapituluan azalduko dira.

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.

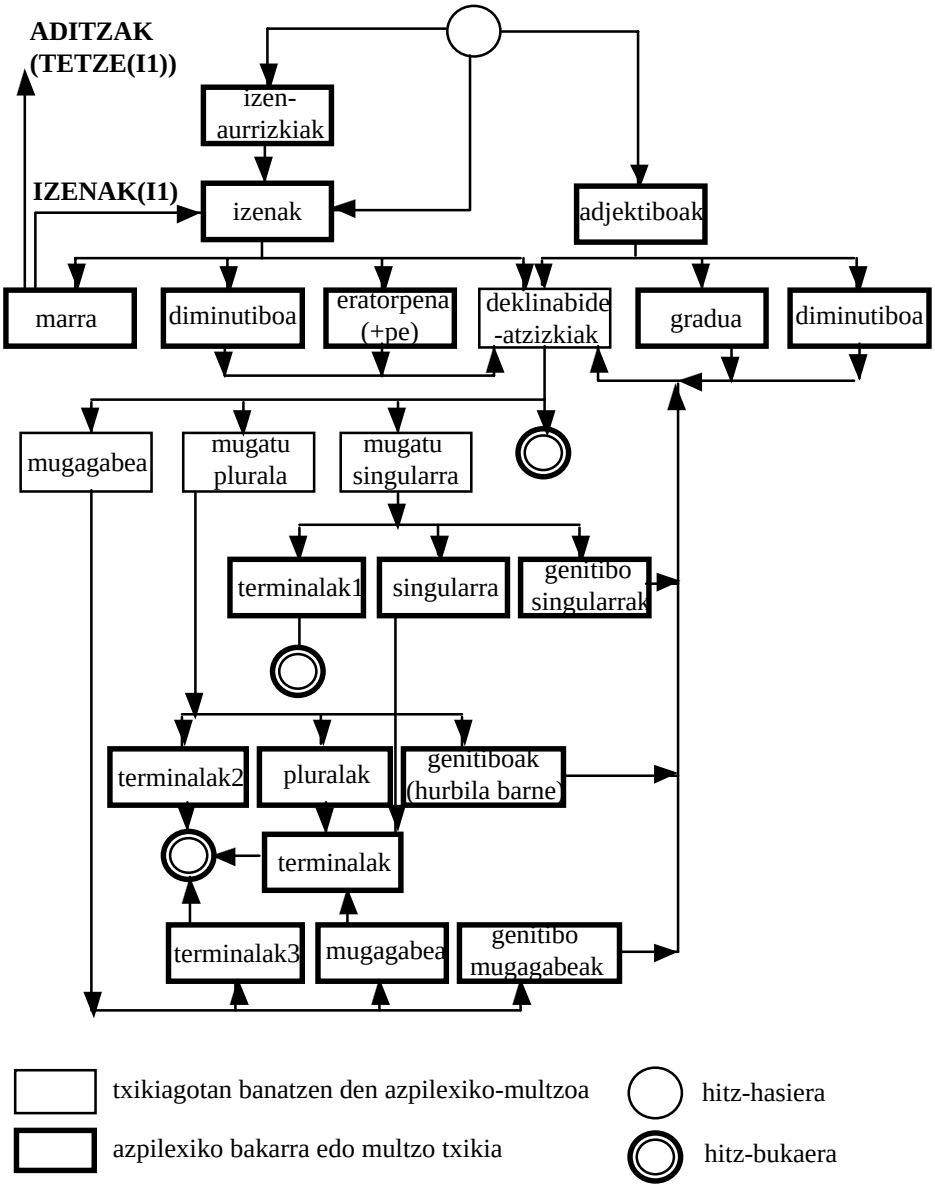
[2] I1-en barruan deklinabide-atzizkiak daude baina ez gidoia. Hitz bat gidoiaz buka daitekeelako agertzen da I0 izenekin eta aditzekin batera lehen maila.

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.

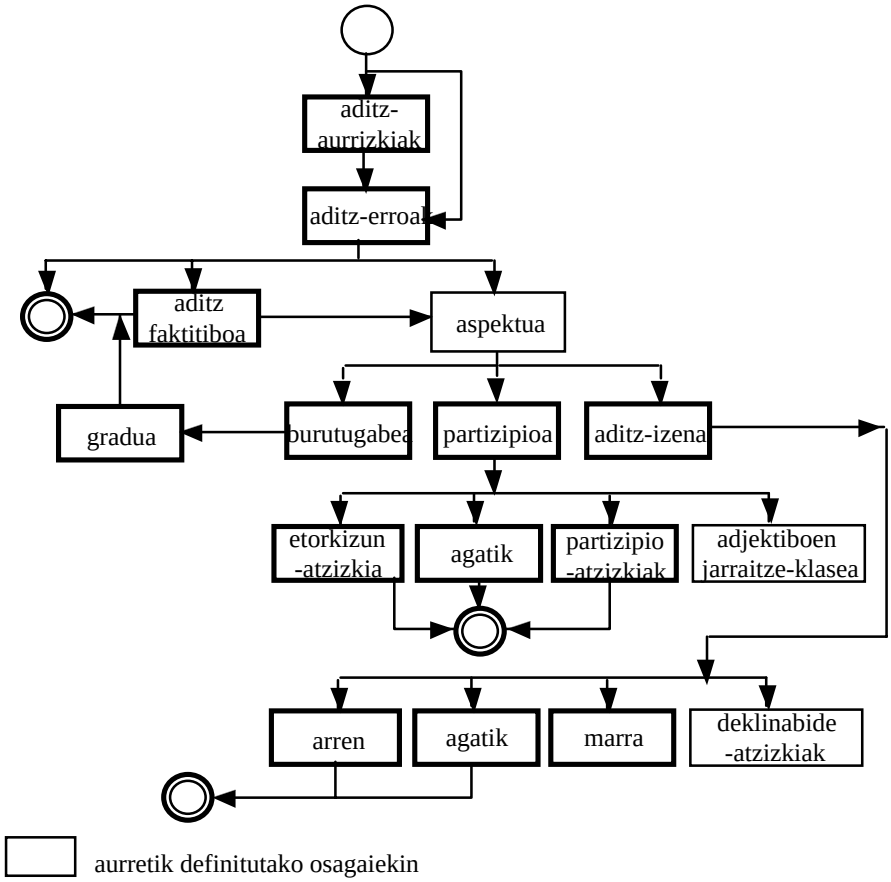
Izen zein adjektiboen atzean atzizki-multzo emankorrena onartzen da: deklinabideari dagokiona. Aurrizkiei, eratorpenari eta elkarketari dagokienean, berriz, desberdinak dira izenak emankorragoak izanik. Adjektiboak, graduatzaileak direla medio, deklinabidearen aukerak biderkatzen dituzte, graduatzaileen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

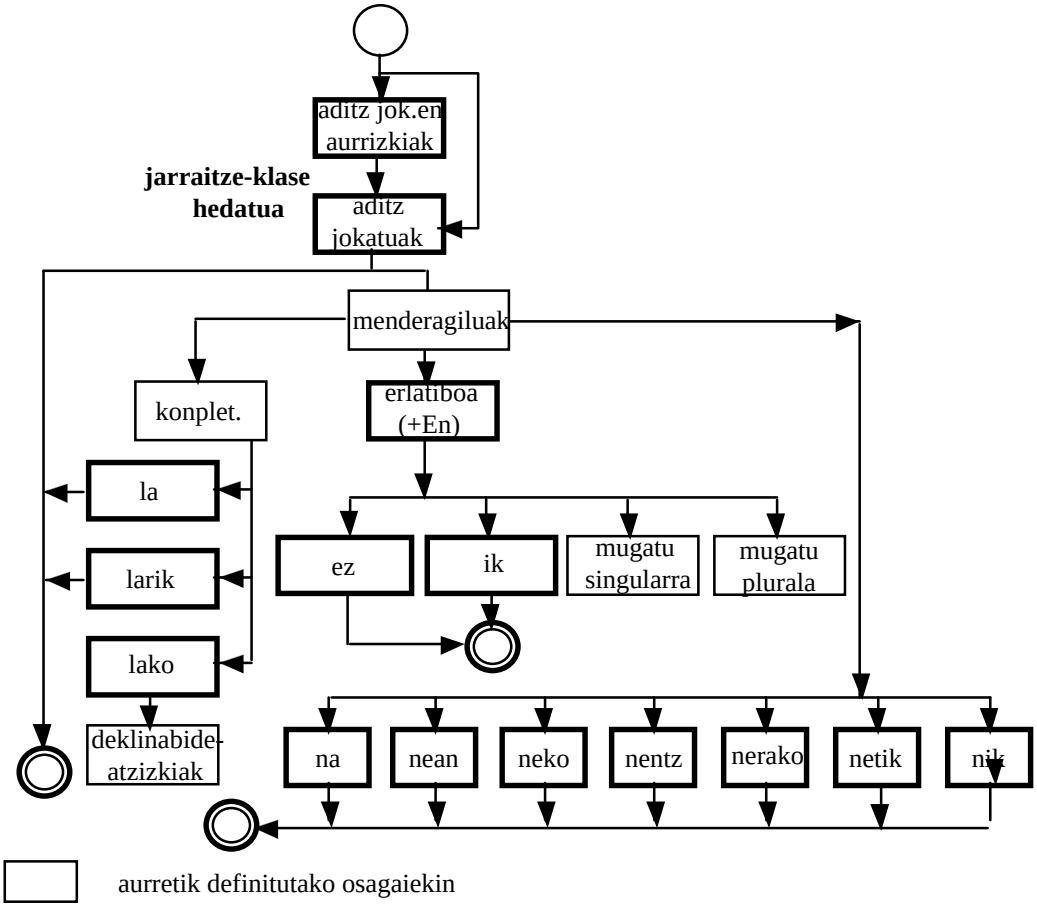
2 *terminalak* deitu dugun azpilexikoen multzoa deklinabide-atzizkiak numero-determinazio hizkiekin (0-ta-eta-ota) konbinatzen dira. *terminalak1*, *terminalak2* eta *terminalak3* izenekin deitutakoetan numero-determinazio eta kasua atzizki bakarrean daude.

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.

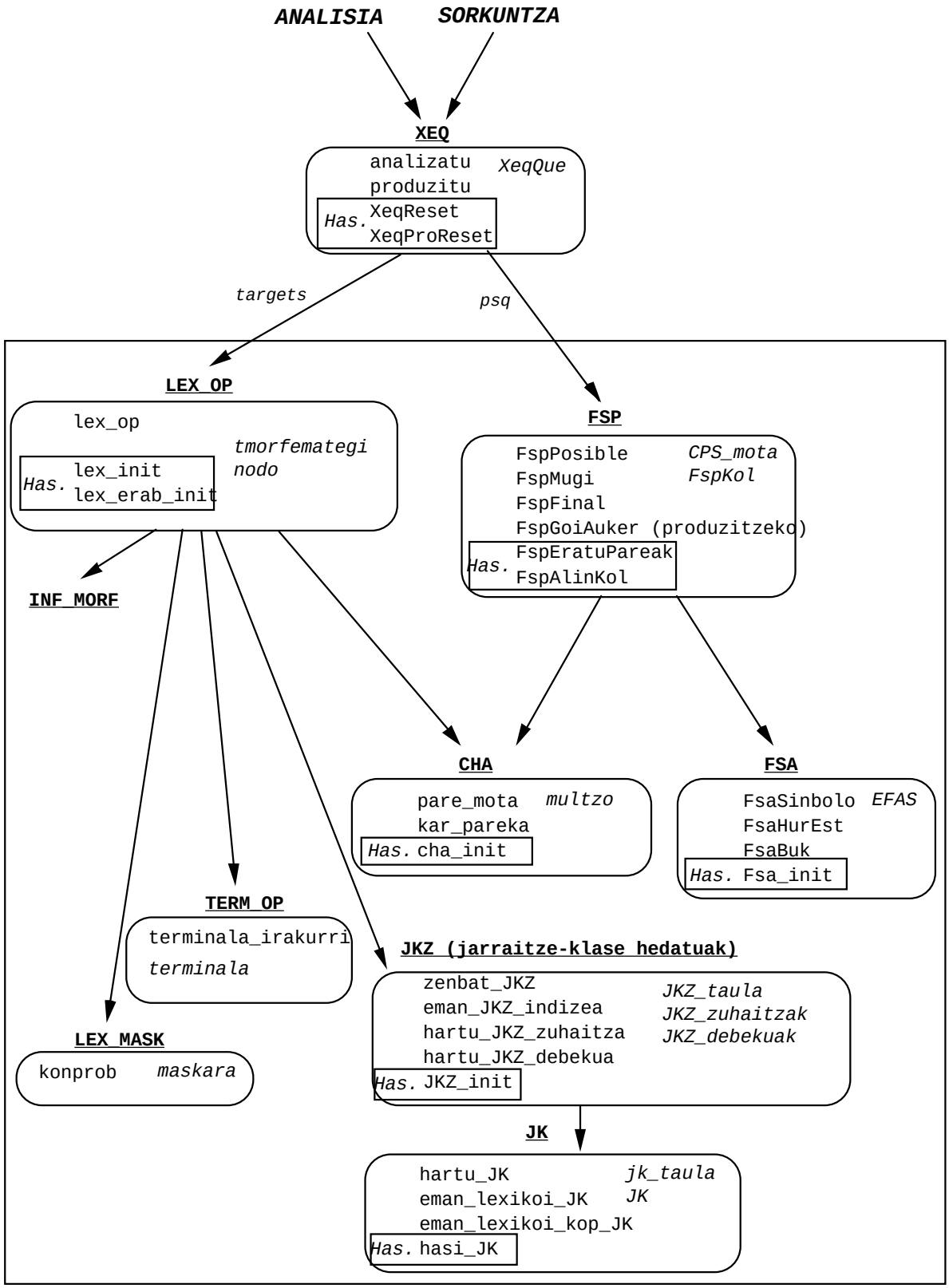
Proiektuan hasi ginenean bi mailatako morfologiari aurre egiten zion erabilpen libreko programarik ez zegoenez, geure inplementazioari ekin genion. Ondoren, PC-KIMMO (Antworth, 90) izeneko softwarea eskuragarri izan genuen, baina gure inplementazioarekin jarraitu genuen honako arrazoi hauengatik:

- 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.

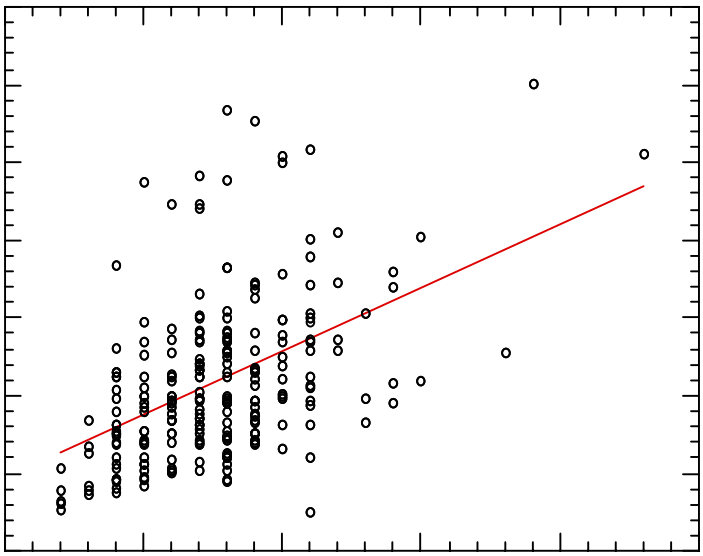
III.5.4 Eraginkortasunari buruzko zenbait datu eta gogoeta.

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.

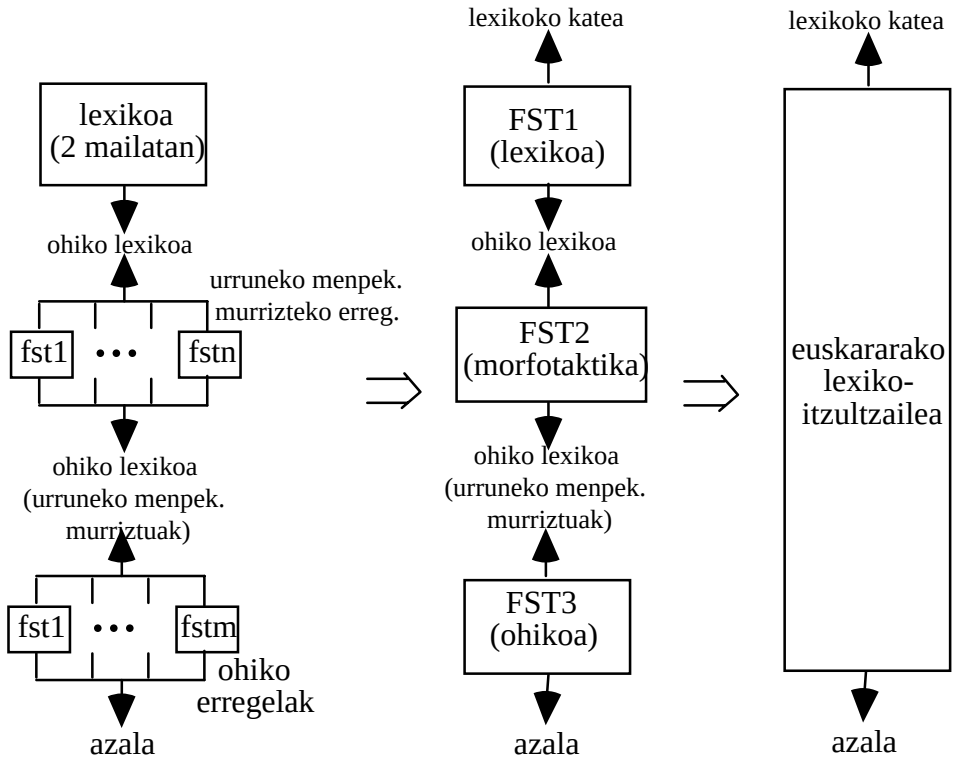
```
b:b /<= #: _ [ a {Baldin} | a i t ] MM [=:=]* MM ;
```

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.

IV. Analizatzaile sendoa osatzen.

Euskara estandarraren analisia aztertzean estaldurari buruzko emaitzak ez zirela behar direnak azaldu da (ikus §III.5.2), eskala errealeko sistema bat eraikitzeko asmoa badugu behintzat. Aipatutako emaitzak zuzentzeko urrats desberdinetan burutu den lana kapitulu honetan azaltzen da, eta ikusiko denez, tratamendu guztiak bi mailatako morfologian daude oinarriturik.

Aipaturiko emaitzetan oinarriturik, bi dira prozesadore morfologikoaren emaitzak mugatzen dituzten arrazoi nagusiak: lexikoan ez dauden lemak batetik, eta erabilpen ez-estandarrak bestetik.

Analizatzen ez diren hitzen erdien ingurua ez dira ezagutzen dagokion lema lexikoan ez dagoelako —ikus III.10 irudia—. Lexikoan ez egotearen arrazoiak desberdinak izan arren, eta, kasu batzuetan oraindik, lexiko orokorra aberasten lan gehiagoren beharra antzematen bada ere, askotan ezin da edo oso zaila gertatzen da lema guzti horiek lexiko orokorrean egotea, testuaren edota idazlearen erabilpen espezifikoak baitira askotan. Gainera, gure proiekturako oso inportantea izan da lexiko estandar bat definitzea; batetik hizkuntzak bizi duen batasun-prozesuari begira hau zehaztea funtsezkoa iruditu zaigulako, eta bestetik lexiko hori erabili delako euskararako dagoen egiaztatzaile/zuzentzaile ortografiko bakarrerako. Aurreko hori guztia kontuan hartuz, komenigarritzat jo dugu erabiltzaileari lexiko partikularrak eguneratzeko aukera ematea, aukera guzti-guztiak lexiko orokorrean sartu gabe.

Aipaturiko emaitza motz horien beste iturri nagusia aldaeren erabilera da, hau da, euskara estandartzat hartzen ez diren formen erabilera. Erabilpen dialektalak, forma estandarrei buruzko ezjakintasunak, zalantzek edo gertaturiko aldaketek edo erregelen aplikazio okerrak eragiten dute aldaeren agerpena. Horren aurrean, eta maiztasun handiko agerpenak daudela kontuan hartuz —ez analizatutako heren bat gutxi gorabehera, III.10 irudian agertutako datuen arabera— forma horiek analizatzeko bi mailatako morfologian oinarritutako mekanismoak burutu dira.

Azkenik, analizatzaile sendoa lortzeko, eta etiketatzaile/lematizatzaile bati begira, sistemak forma guztiak analizatzeko gai izan behar du, nahiz eta dagokion lema lexikoan orokorrean egon ez. Prozesu honi “lexikorik gabeko analisia” deituko diogu, lexikoko atal bat soilik, hizkiena hain zuzen, erabiltzen duelako.

IV.1 Erabiltzailearen lexikoa.

Erabiltzailearen lexikoa edo lexiko berezitua erabiltzeko arrazoiak kapituluaren hasieran aurkezten badira ere, arrazoi nagusia zera da: euskararako lexiko orokor oso bat eratzeko dauden zailtasunak, lexiko arrunta zeharo finkatu gabe, maileguak orokorrean, eta espainieratikoak bereziki, noraino iritsi behar diren eztabaidagai da, atzerriko leku-izenen idazkera ez dago erabakita, termino zientifiko-tekniko berriena ere ez, etab. Adibide batzuk aipatzearren, testuetan bilatuz gero ez da arraroa aurkitzea honelako arazoak: lema bera adierazteko grafia desberdin asko agertzea —adib. injinero, injineru, injeniero, injenieru, injinadore, ingeniari, inginari, ...—, edo euskal forma anitz agertzeaz gain mailegu desberdinen agerpena—ogibitarteko, ogitarteko, otarteko, sandwich, bokadilo.

Aipatutako arazoek, besteak beste, oso eragin negatiboa dute analizatzaile morfologikoaren estalduran; beraz, honen aurrean, eta lexiko orokorra ahalik eta gehien osatzeari uko egin gabe, erabiltzailearen lexiko berezituen kudeaketa bideratzea erabaki dugu; horrela lexiko orokorra lexiko estandarra bihur dadin, ahal den neurrian behintzat. Honekin bi abantaila lortzen dira: malgutasuna eta lexiko estandarra/ez-estandarra bereiztea, hau guztia analizatzailearen emaitzak hobetzeko aukera galdu gabe. Horren truke, erabiltzaileari lexikoa eguneratzeko ahalegina eskatzen zaio.

IV.1.1. Azpilexikoen ezaugarri garrantzitsuak.

Erabiltzailearen lexikoa kudeatzeko orduan badago funtsezko elementu bat: azpilexikoei egokitzen zaizkien ezaugarrien multzoa. III.3.3.1 atalean azaldu den bezala gure sistemaren azpilexiko bakoitzari lau ezaugarri esleitzen zaizkio: hasierakoa izatea, irekitasuna, orokortasuna eta estandartasuna. Lehenengoak eta laugarrenak erabiltzailearen lexikoen kudeaketarekin zerikusirik ez duten bitartean, morfotaktikarekin eta aldaeren tratamenduarekin loturik baitaude, irekitasunak eta orokortasunak oinarrizko funtzioa dute.

Azpilexiko bat **irekia** da, eta irekitasun ezaugarria esleitzen zaio, baldin eta dagozkion forma guztiak ez badira bukatzen bere osagaiekin, eta ondorioz, azpilexiko horri legokizkiokeen formak erabiltzailearen lexikoan ager badaitezke. Azpilexiko irekietako osagaiak beti dira lemak eta gure sisteman ondoko sei azpilexiko ireki hauek daude: izenak, adjektiboak, aditz-erroak, adberbioak, siglak eta bestelakoak[1]. Gainontzeko azpilexikoak itxiak dira, eta ondorioz beraiei dagokien morfemarik ezin da agertu

[1] Azken kategoria honetan kokatzen dira forma bereziak, eta flexiorik gabeko formatzat hartzen direnak.

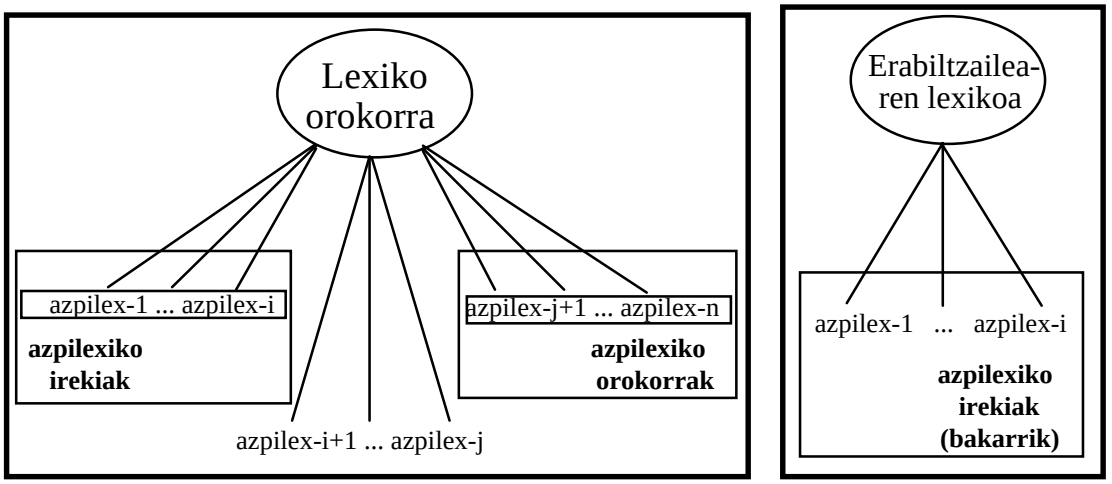
erabiltzailearen lexikoetan. Adibidez, aditz laguntzailea ondo mugaturik dago euskaraz, eta ez du erabiltzailearen lexikoan agertzeko arrazoirik.

Azpilexiko bat **orokorra** da morfotaktikaren arabera bere osagaiek azpilexiko irekietako sarrerekin konbina badaitezke. Dagozkien osagaiak beti dira hizkiak eta ez lemak. Lemak dituzten azpilexiko itxiek eta hauei bakarrik dagozkien hizkien azpilexikoek ez dute bi ezaugarrietako bat ere izango.

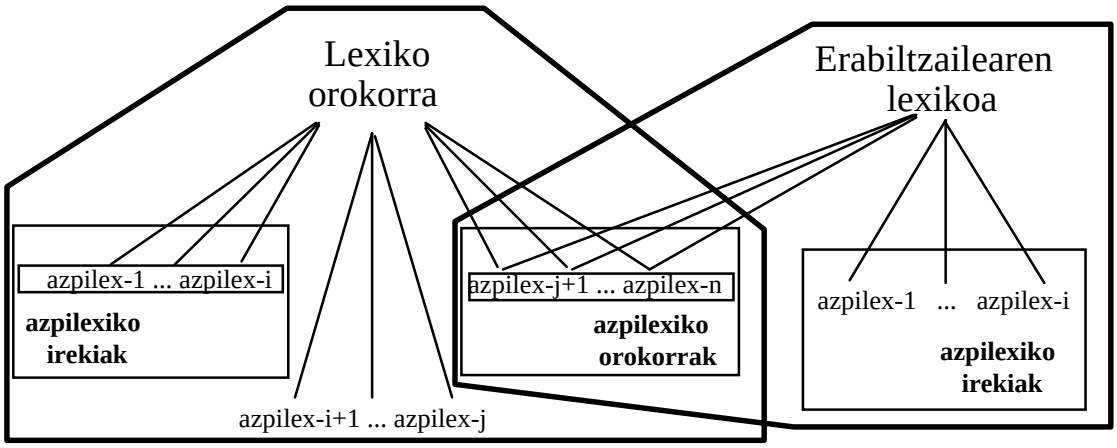
Azaldutako bi ezaugarri horiek erabiltzailearen lexikoak kudeatzeko erabiltzeaz gain, “lexikorik gabeko analisia” egiteko ere dira baliagarriak.

IV.1.2. Burutzapena.

Aurreko ezaugarri horien erabilera IV.1 irudian azaltzen da.



(A) LEXIKO OROKORRA ETA ERABILTZAILEARENA FITXATEGIETAN



(B) LEXIKO OROKORRAREN ETA ERABILTZAILEARENAREN KUDEAKETA

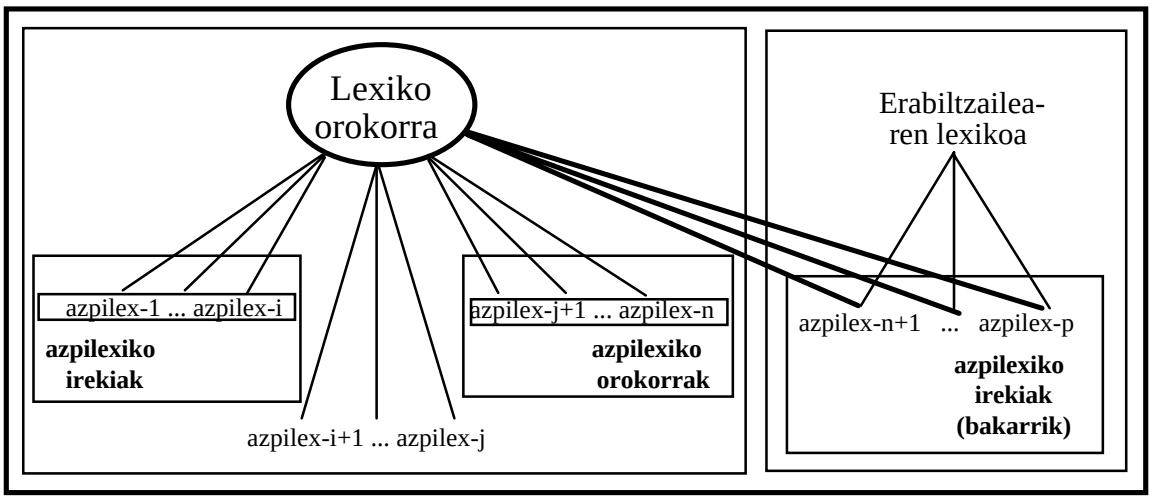
IV.1 irudia.- Lexiko orokorra eta erabiltzailearen lexikoa osatzeko modua.

Irudikatzen den ideia honetan datza: gordetzen den erabiltzailearen lexikoan lemak baino ez badaude ere, lema hauek lexiko orokorrean dauden azpilexiko orokorrekin konbinatuko dira, erabiltzailearen lexikoa erabiliz bertako lemen flexioa eta eratorpena ezagutzea posible izan dadin.

Beraz, erabiltzailearen lexikoa erabiliz analisia egiteko bi urrats burutuko dira: bat lexiko orokorraren bidez, eta ondoren bestea, aurrekoa arrakastatsua ez denean normalean[1], erabiltzailearen lexikoaren bidez.

[1] Hau parametriza daiteke, posible baita beti analisi posible guzti-guztiak lortzea.

Erabiltzailearen lexikoa maneiatzeko bazegoen beste aukera bat, azpilexiko irekiak bikoiztea eta analisi bakar batean burutzea analisia (ikus IV.2 irudia).



IV.2 irudia.- Erabiltzailearen lexikoa osatzeko beste modua.

Buruturiko aukerak bestearekin dituen aldeak hauek dira:

- Malguagoa da, erabiltzailearen lexiko berezitu anitz onartzen duelako, eta analisi estandar/berezitua banatzea modu naturalean bideratzen duelako.
- Morfotaktika ez da ukitu behar, beraz adierazpen morfologikoa zeharo gardena da berrikuntza honekiko. Beste aukerak ukitu txiki batzuk eskatzen ditu zenbait aurrizkiren jarraitze-klasean.
- Arazoak daude hitz-elkarketan lexiko orokorreko eta erabiltzailearenekoa lema bana elkartzen badira, analisia ez baita ezagutzen. Hau ez litzateke gertatuko beste eskemarekin.
- Eraginkortasunaren aldetik antzekotasun handia susmatzen da bi sistemen artean, batek paraleloan egiten duena bestea sekuentzian burutuko duelako.

IV.1.3. Eguneratzeko prozedura.

Aipatu den bezala, lexikoa eguneratzea da lexiko berezituen erabilerak dakarren eragozpen bakarra. Eguneratze hori ezin da automatikoa izan, eta erabiltzaileari eskatu behar zaizkio informazio desberdinak morfotaktika eta ezaugarri morfologikoei buruz.

Gure inplementazioan eskatzen diren informazioak honako hauek dira:

- **kategoria**: azpilexikoa identifikatzeko, beraz sei hauen artean aukeratu beharko du erabiltzaileak: izena, adjektiboa, aditz-erroa, adberbioa, sigla eta besterik.
- **azpikategoria**, izenaren kasuan: bereizi behar dira izen arruntak, leku-izenak eta pertsona-izenak, beren deklinabidea desberdina da eta.
- ***r* mota**: gogorra ala biguna *r*-z bukatutako lemetan, kasuaren arabera zenbait erregelaren aplikazioa aldatzen baita.

Informazio hau eskatuz lexikoa eguneratzen duen prozedurak osagai okerrak edo zaharkituak ezabatzeko aukera ere badu. Seigarren kapituluan azalduko denez, prozedura honetarako elkarrizketa erabilterraza eta atsegina diseinatu da zuzentzaile ortografikoari begira.

Informazio horiez gain beste informazio batzuk suposatu dira erabiltzaileari galdetu gabe. Batetik, aditz-erro berri guztien morfotaktika *tu* bukaera duen infinitiboaren paradigmaren ildotik suposatu da, gainontzekoak aditz zaharrei dagozkielakoan, eta hauek guztiak jaso ditugulakoan *itxiak* izanik. Beste aldetik, kontsonantez bukatutako siglen deklinabidean gerta daitezkeen epentesiak aldakorrak dira haien ahoskeraren arabera, baina erabiltzaileari galdetu beharrean —askotan ez dago hain argi zein den dagokion ahoskera— hautapen-marka berezi bat definitu da halako kasuetarako, / diakritikoa hain zuzen (ikus §III.3.2), beraren bidez bi ahoskerei dagokien deklinabidea onartzen delarik. Automatikoki ezartzen da marka hori siglaren azken letraren arabera.

Modulu honen erabilpena testu-zuzenketan izango bada ere, corpusen analisian ere aplika daiteke, analisia egin ahala eguneratze semiautomatikoa bidera baitaiteke, ezagutzen ez diren hitzen analisia lortzeko eta etorkizunerako analizatzaile sendoagoa lortzeko asmoz.

Jakintza-arlo desberdinetarako lexiko berezituen ekoizpena ere sartzen da gure proiektuen barruan.

Erabiltzailearen lexikoen gauzatzea gure bi mailatako sisteman integratu dugu arazorik gabe. **Lexiko-itzultzaile**en bidez bideratzeko garaian arazoak daude, lexiko-itzultzaileek aurrekonpilazioa eskatzen dutelako. Beraz, prozedura aldatuz, lexiko hauek aurreprozesu zein postprozesu baten bidez eguneratu beharko lirateke, ez baita oso egokia hitz bakoitza sartzean konpilazio berri bat burutzea. Honek arazoak dakartza lexiko-itzultzaileetan oinarriturik zuzentzaile ortografiko malgu bat diseinatu nahi dugunean. Gainera, lexiko-itzultzaileetan ezin dira azpilexiko mailako ezaugarri morfologikoak kudeatu, beraz, IV.1 eta IV.2 irudietako egiturak konplexuago izango lirateke.

IV.2 Forma ez-estandarren analisia.

Euskara estandartzat hartzen ez diren formen erabilera da prozesadore morfologiko estandarraren emaitzen mugatzaile nagusietako bat. III.10 irudian agertutako datuen arabera, analizatu gabe geratutako formetako heren bat -gutxi gorabehera erabilpen ez-estandarrei dagokie. Proportzio hau igo egiten da corpusa kontuan hartzen bada eta ez hitz-zerrenda, zeren erabilpen ez-estandar batzuek maiztasun handiko agerpenak dituztelako, *batzu*-ren flexio mugagabeak edo *haundi* lemaren agerpenak, adibidez.

Forma ez-estandar hauei *aldaerak* deituko diegu. Erabilpen dialektalak, forma estandarrei buruzko ezjakintasunak, zalantzek edo gertaturiko aldaketek edo erregelen aplikazio okerrak eragindako formak kokatzen dira multzo honetan. Hauetako forma batzuen erabilpen zabalaren arrazoia hauxe da: garai batean estandarrak izan zirela edo estandartzat hartu izana, euskararen batasunaren historia laburra izan arren iritzi batzuk aldatuz joan direlako eta zehaztu gabe zeuden irizpide batzuk zehaztu direlako. Gainera, aurreko urteetako testuak analizatzeko asmoa baldin dugu —beti ere batasunerako oinarrizko irizpideak betetzen dituztenak, deklinabidearena eta aditz laguntzailearena batez ere—, aldaeren tratamendu hau are ezinbestekoago bihurtzen da.

IV.2.1. Oinarria: bi mailatako mekanismo osagarria.

Aldaeren tratamendua bideratzeko ere bi mailatako morfologia izan da oinarria. Hitz batek analisi estandarrik ez badu analisi berri bati ekiten zaio, horretarako analisi estandarraren funtsezko osagai diren lexikoari eta erregelei beste bi mailatako sistema osagarri bat eransten zaio; sistema osagarri honen elementuak ere lexikoa eta bi mailatako erregelak dira.

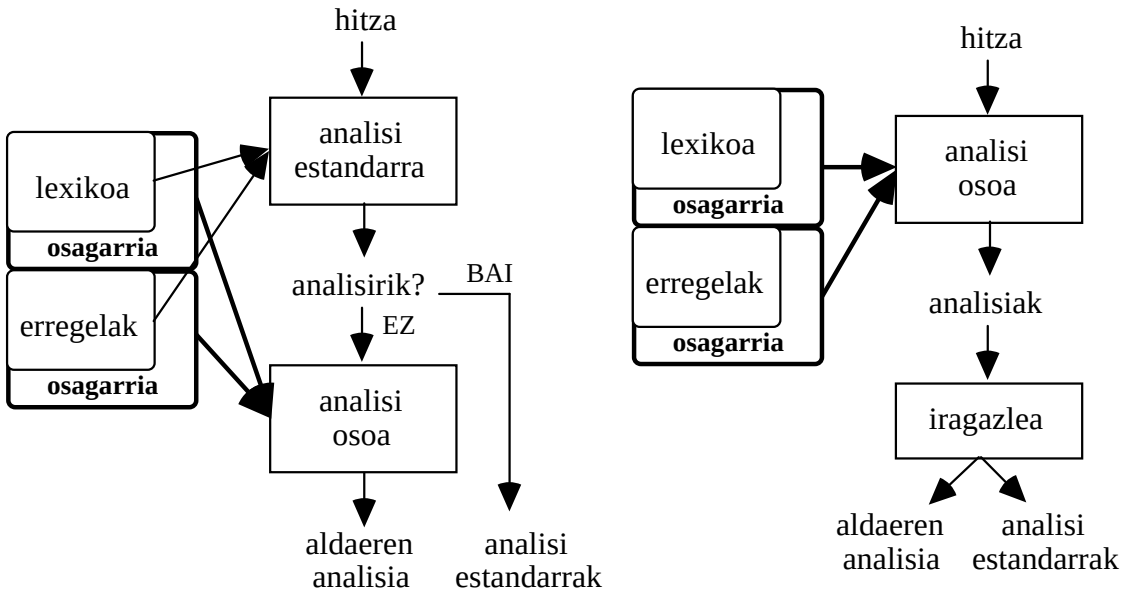
Ikus dezagun forma ez-estandarren tipifikazioa eta mota bakoitzeko analisia ebazteko modua. Hiru multzotan banatuko ditugu aurkitutako aldaerak: morfemen aldaera, morfotaktikaren aldaera, morfofonologiaren aldaera.

- **Morfemen aldaera**: morfema baten ordez, erroa edo hizkia izanda, beste bat erabiltzen denean, aldaera mota honetakoa da. *haundi* lema eta *tikan* atzizkia ditugu adibide gisa: Euskaltzaindiak *handi* hobesten du *haundi*-ren kaltetan, eta *tik* Gipuzkoako euskalkiko *tikan* -en ordez.

- **Morfotaktikaren aldaera**: Morfema baten edo multzo baten ondoren etor daitekeen morfema-multzoa aldatzen denean. Adibide gisa *bait* aurrizkia eta *batzu* bezalako moduko determinatzaileak —*nortzu*, *zeintzu*, etab.— ditugu. Aurrizkiaren kasuan aurreko arauaren arabera banandua idatzi behar zen, beraz

terminala izan behar zuen, eta arau berriaren arabera aditz jokatuari eransten zaio. Determinatzaileen kasuan, berriz, gaur egun beren flexio estandarra pluralari dagokiona den bitartean, duela zenbait denbora mugagabean deklinatzea ere onargarria zen.

- Erregela morfofonologikoak gaizki erabiltzetik edota berriak erabiltzetik datoz fonologia edo **morfofonologiaren aldaerak** deiturikoak. Erregularrak diren aldaera-multzoak biltzen dira hauetan. *s*/*z* eta *z*/*s* aldaketak edo *h*-ren erabilera okerra dugu hauen adibidea; bietan idazlearen ezjakintasunari lepora badakioke ere, lehenaren iturburua euskalkiaren eragina da, eta bigarrenarena hegoaldean ez ahoskatzearena.



(A) bi urratsetan **(B) urrats bakarrean**

IV.3 irudia.- Aldaeren analisia lortzeko prozedurak

Aldaera horien analisia bideratzeko bi mailatako formalismoari eutsi egin diogu, ebazpen partikularretatik alde eginez. Helburu horrekin, lehen bi motako aldaerak ezagutzeko lexiko osagarri bat diseinatu da, lexiko nagusiaren azpilexiko bakoitzari beste bat definitzeko aukera emanez, eta bertan morfemen aldaera edota jarraitze-klase berria zehaztuz.

Aldaera morfofonologikoak deitutakoak analizatzeko bi mailatako beste erregela-multzo bat definitu da, baina, erregela hauetako batzuk hasierakoekin kontraesanean egon daitezkeenez gero, arazo honi aurre egin behar zaio geroago ikusiko dugun bezala.

Prozeduraren aldetik, eta ondorengo aplikazioei begira (etiketatzailea, analisi sintaktikoa, etab.), analisia urrats desberdinetan egitea deliberatu dugu, hau da, edozein formatarako analisi estandarra burutzen da aurretik, eta arrakasta ez dagoenean baino ez da burutzen aldaerak kontuan hartzen dituen analisia (ikus IV.3 irudiko A aukera).

Beste aukera bat dago IV.3 irudian, B-ri dagokiona hain zuzen, analisi osoa burutzean eta emaitzen artean bereiztean datzana. Azken aukera hau baztertu egin dugu, ondoko bi arrazoiengatik:

- Analisi guztiak batera egitean analisi estandarren eta ez-estandarren artean bereiztea ez da berehalakoa: azpilexiko osagarrietatik hartzen dena markaturik egon daitekeen bitartean —bereizteko erraza izanik—, erregelen bidez bideratutako analisi ez-estandarrak bereiztea gatza da (ikus §IV.2.3 atala).
- Forma gehienek analisi estandarrik badutenez eta erregela-multzo osagarriak konputazio-konplexutasuna erruz igotzen duenez, azkarragoa izaten da lehen aukera bigarrena baino.

Hala ere forma guztien analisi osoa lortzeko aukera ere badago.

IV.2.2. Azpilexikoak eta erregela osagarriak.

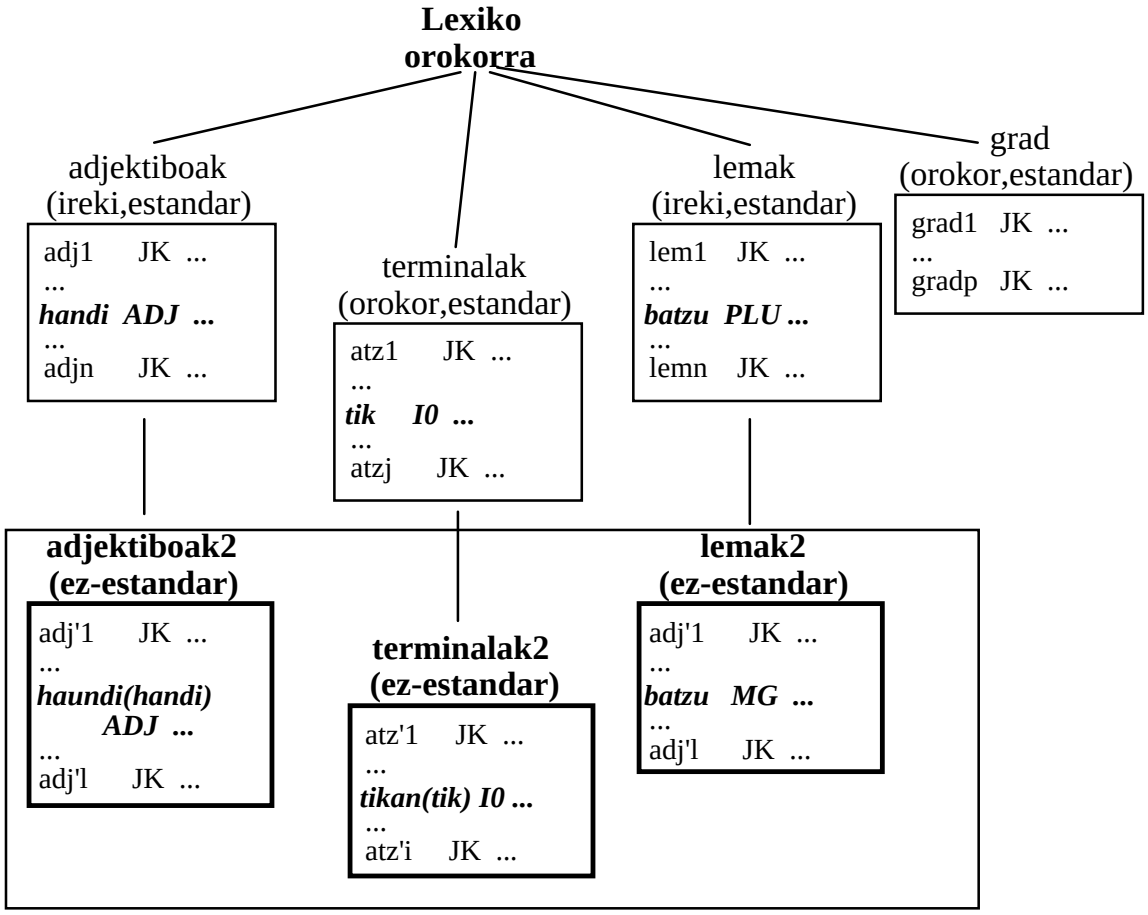
Aurrean aipatu den bezala aldaeren tratamendua bideratzeko bi mailatako formalismoaren oinarria diren lexikoari eta erregela-multzoari gehigarri bana eransten zaie: azpilexikoak eta erregelak, hurrenez hurren.

IV.2.2.1. Azpilexikoak.

Aurretik esan den bezala, lexiko orokorreko azpilexiko bakoitzeko posiblea da dagokion azpilexiko osagarria definitzea eta erabiltzea morfema zehatzen aldaera eta morfotaktikaren aldaeren tratamenduari aurre egiteko asmoz. Azpilexiko osagarri hauek ezaugarri berezi bat dute ez-estandartasunarena hain zuzen. Ezaugarri honen arabera erabakiko da azpilexikoak kontuan hartzen direnentz formak analizatzerakoan. Analisi estandarra egiten denean *estandar* ezaugarria duten azpilexikoak bakarrik hartuko dira kontuan; aldaerak ezagutzen dituen analisia egitean, aldiz, denak jorratuko dira.

Adibide gisa IV.4 irudia dugu. Bertan ikusten denaren arabera, eta sistema orokorraren sinplifikazio bat dela kontuan hartuz, lau azpilexikok osatzen dute lexiko orokorra: *adjektiboak, lemak, terminalak* eta *grad,* laurak estandarrak, jakina. Hiru azpilexiko ez-estandar osagarri dago: *adjektiboak2*, *lemak2* eta *terminalak2*, horietako bakoitza estandar bati dagokiolarik. Bakoitzean sarrera bat azpimarratu dugu, horrela *haundi* da *handi*-ren

aldaera eta *tikan tik*-ena, beraz jarraitze-klasea mantentzen dute, *ADJ* eta *I0*, eta morfema dagokion estandarrarekin loturik dago: *haundi(handi)* eta *tikan(tik)*. Lotura honek helburu bikoitza du: analisian forma estandarra lortzea, eta zuzenketan aldaeraren ordez forma estandarra lortzea. Beste kasua, *batzu*-rena, desberdina da morfotaktikaren aldaera da eta. Kasu hauetan lema bera agertzen da bi azpilexikoetan, estandarrean eta dagokion ez-estandarrean, baina bakoitzean jarraitze-klase desberdina. Lexiko ez-estandarrean agertzen den jarraitze-klasea berria izan daiteke baina normalean estandarretako bat izaten da —kontuan hartu behar da erabilera "tipikoa" dela eta, beraz, beste batekin nahastetik datorrela—.



IV.4 irudia.- Aldaeren tratamendurako azpilexiko osagarriak

Sistema osoa ibil dadin ondoko hau ez da ahaztu behar:

- Ezin da pentsatu aldaeren analisia egitea azpilexiko osagarriak soilik erabiliz, zeren hitzetan erabilera ez-estandarrak eta estandarrak konbinatzen baitira. Horrela *haunditik* edo *handitikan* analizatzeko lexiko osoa behar da, *haundi* eta *tikan* lexiko osagarrian dauden bitartean *handi* eta *tik* lexiko estandarrean daude.

- Morfotaktikari begira, analisi estandarrerako definitutako jarraitze-klaseetan zehazten diren azpilexiko bakaoitzari azpilexiko osagarria erantsi behar zaio berau existitzen bada. Hori modu automatikoan egin daiteke bestelako lan gehiagorik hartu gabe.

Proiektuaren definizio-kopuruen aldetik, 33 azpilexikori egokitu zaie bere azpilexiko ez-estandarra, guztien artean ia mila sarrera osatuz. Azpilexiko ez-estandar handienak dira izenena, 468 sarrerarekin, eta aditz-erroena, 180rekin.

IV.2.2.2. Erregelak.

Aipatutako azpilexikoekin batera aldaeraren bat duten hitzak ezagutzeko aldaketa morfofonologikoak ere kontuan hartu behar dira, eta horretarako bi mailatako erregela-multzo gehigarria definitu da.

Erregela guzti hauek *testuinguru-murriztapena* (=>) motakoak dira, zeren aipatzen diren aldaketak gerta daitezke baina ez dira behartu behar; kontuan hartu behar baita aldaketa estandarrak eta ez-estandarrak konbina daitezkeela hitz bakar batean, lexikoan gertatzen den legez.

Erregela hauek, osagarriak direla esan badugu ere, eragina dute jatorrizko multzoko batzuetan, eta hau gertatzen da bi sistemetan aldaketa bera deskribatzen denean eta jatorrizkoan *azalekoaren derrigortzea* (<=) azaltzen bada. Horrelako kasuetan jatorrizko erregela berrikusi egin behar da[1], bi erregelen artean dagoen kontraesana ebazteko asmoz, horretarako jatorrizkoaren derrigortzea lasaituko delarik. Beraz, aldaeren tratamendurako erregela-multzoa ez da jatorrizkoa gehi osagarria baizik eta berri bat, atal honetan egingo dugun azalpena erraztearren independentetzat hartuko ditugun arren.

Erregelak hiru ataletan banatu ditugu: fonologikoak (hein handi batean, behintzat), ortografikoak eta morfofonologikoak. Jatorrizkoekin erkatzen baditugu gailentzen diren aldeak bi dira: jatorri fonologikoa nagusitzen da batetik (erregela kopuruan baino aldaketa kopuruan batez ere), eta bestetik diakritikoak ez erabiltzea, aurrekoarekin lotuta dagoena.

B eranskinean guztien deskribapena azaltzen denez gero, ondoren hiru erregelaren deskribapena azaltzen da adibide gisa. Erregeletan erabiltzen diren alfabetoak, markak, multzoak eta espresioak III.4.1en azaldutako berberak dira.

---

[1] Lexiko-itzultzaileen kasuan saiatu izan gara bi sistemen arteko independentzia mantentzen (ikus §IV.2.5).

g/j aldaketa

Letra hauen ahoskatze desberdinak eta espainieraren eragina direla eta aldaera hau maiz gertatzen da.

g/j aldaketa

```
Cx:Cy => _ [ e | i ] ;
            where Cx in (g j)
                  Cy in (j g)
                matched;
                ! filologia:filolojia
                ! erlijio:erligio
```

h-ren erabilpen okerra.

h duten hitzak ez ezagutzetik dator aldaera hau.

h-ren sorrera eta galera

```
0:h => [ Hasiera | Bokal ] _ Bokal ;
                ! ziur:zihur
                ! esparru:hesparru
h:0 => [ Hasiera | Bokal ] _ Bokal ;
                ! mehe:mee
                ! hau:au
```

Maileguetako u/o bukaera.

Zenbait mailegutako bukaera o/u izan daiteke jatorriaren arabera, eta honen ondorioz akatsak egiten dira.

u/o aldaketa

```
u:o => Kons _ MorfBuk ;
                ! exenplu:exenplo
o:u => Kons _ [ MorfBuk | a ] ;
                ! alfabeto:alfabetu
                ! agoanta:aguanta
```

IV.2.3. Aldaera-motaren identifikazioa. Desanbiguazio lokala.

Aldaerak, beti ez bada ere, euskara estandar idatziaren ikuspuntutik erroreak dira, beraz, aldaeraren ezaguera, identifikazioa eta zuzenketa interesgarria izan daiteke aplikazio

batzuetarako, Ordenadorez Lagunduriko Hizkuntz Irakaskuntzan (OLI) esaterako (Maritxalar & Diaz de Illarraza, 94).

Beste aldetik, aldaeren tratamendua egiterakoan anbiguetatea sor daiteke, ez bakarrik forma estandar eta ez-estandarren artean, baizik eta analisi ez-estandar desberdinen artean. Azken kasu honetan komenigarria izan daiteke, lematizatzaile edo etiketatzaile bati begira adibidez, analisi desberdinen artean sailkapen bat eta desanbiguazioa burutzea, eta honi desanbiguazio lokala deitu diogu. Horretarako jakin behar da zenbat aldaera gertatu diren hitzaren barruan analisi bakoitzeko, eta aldaera hauen mota.

IV.2.3.1. Aldaera-mota eta kopurua.

Analisi batean agertzen diren aldaera-motak ezagutzeko beraiei buruzko informaziora jo beharko da. Informazioa bi tokitan dagoenez, lexikoan eta erregeletan, bi kasu hauek bereiziko ditugu.

Lexikoaren kasuan ez dago arazorik, lexikoan informazio morfologikorako aurrikusitako tokian aldaerari buruzko kode bat ezar daiteke aldaeren tipifikazioa egin eta gero, eta analisiaren emaitzaz informazio hori lortu. Hau izan da guk egin duguna. Horrela, *kaletikan* analizatzean ondoko analisia lortuko da:

```
((forma "kaletikan")
  ((anal ALDAERA1)
   ((lema "kale")((KAT IZE))))
   ((morf "0")((KAT DEK)(NUM S)(MUG M))))
   ((morf "Etik") (ald3 "Etikan")((KAT DEK)(KAS ABL)))))
)
```

Bertan ikus daitekeenez, *Etikan* atzizkia 3. motako aldaera gisa gorderik dago azpilexiko ez-estandarrean, berarekin ordezko morfema estandarra, *Etik*, lotzen delarik.

Erregelen kasuan irtenbidea askoz ere konplexuagoa da. Kontuan hartu behar da, bigarren kapituluan esan den bezala (ikus §II.3.2), bi mailatako formalismoaren arabera erregelak bikote-kontrolatzaileak direla eta, bikote bat onartzeko, erregela guztietan onartua izan behar da. Beraz, nola jakin erregela osagarriren bat arrakastatsu izan dela dagokion kasua kontuan hartzeko? Automatetako egoera batzuk markatzea izan da gure ebazpidea: egoera markatu horietara iristeko, aldaera bati dagokion erregelaren eskuineko testuingurua egiaztatutakoan iritsiko da[1].

Ondoko parrafoan erregela bati dagokion automata markatua ikus daiteke. Bertan ikus daitekeenez, guk nahiago izan dugu arkuak markatzea —zeinu negatiboaren bidez

[1] Ritchie-ren taldean antzeko zerbait proposatzen da ere: bere lanean erregelen testuinguru osoa betetzen deneko egoeretan *TERMINAL* motako egoera ezartzen da. Ikus (Ritchie *et al*. 92:151)

automatan— korapiluneak baino, horren arrazoia egoera kopuruen minimizazioa izan delarik. Beraz, eskuzko konpilazioak aukera eman digu hau burutzeko. Konpiladore bat egiterakoan erregelen sintaxian zerbait gehitu beharko litzateke adierazteko zein testuinguru markatu behar diren.

e-ren galera eta *r*/*err* aldaketarako erregela eta dagokion automata markatua.

```
e:0 => e MM _  n MorfBuk ;       !  MorfBuk: +: | #:
       Hasiera _ r:0 r Bokal ;   !  Hasiera: #: (*: )

        # * e e n r r Bokal + =
        = = e 0 n 0 r Bokal = =
    1: 6 1 2 0 1 1 1   1   1 1
    2: 1 1 2 0 1 1 1   1   3 1
    3: 1 1 2 4 1 1 1   1   3 1
    4. 1 1 2 0 5 1 1   1   1 1
    5.-1 1 2 0 1 1 1   1  -1 1
    6: 1 6 2 7 1 1 1   1   1 1
    7. 1 1 2 0 1 8 1   1   1 1
    8. 1 1 2 0 1 1 9   1   1 1
    9. 1 1 2 0 1 1 1  -1   1 1
```

Lexikoan eta automatetan gehitutako informazio hau erabiltzen duten aldaketa batzuk gehitu ditugu programan, aldaeren analisiarekin batera dagokion zera lortzen duena: aldaerei dagozkien morfemetan dagoen kodea eta aplikatutako erregela osagarriak. Horrela, forma estandarretik distantzia handian dauden *suaitxetikan* (*zuhaitzetik*-en aldaera) moduko formak ezagut daitezke eta ondoko analisia lortu:

```
((forma "suaitxetikan")
  ((anal ALDAERA1)
     ((lema "zuhaitz")(ald "suaitx")((KAT IZE))(er24,er18,er24))
   ((morf "0")((KAT DEK)(NUM S)(MUG M))))
   ((morf "Etik") (ald3 "Etikan")((KAT DEK)(KAS ABL)))))
)
```

Adibidean ikusten diren informazio azpimarragarrienak hauexek dira: *er24* eta *er18* dira bi erregelari dagozkien kodeak —txistukarien arteko aldaketa eta h-ren galerari dagozkienak— eta *ald3* da lexikoan jasotako aldaeraren kodea —euskalkiaren eragina adierazten duena—. Informazio hau beste aplikaziotarako erabilgarri izateaz gain, desanbiguazioari begira ere interesgarria gertatuko da.

IV.2.3.2. Desanbiguazio lokala

Aldaeren analisia lortuz gero haien artean ordena, eta bere kasuan batzuen bazterketa, burutzen duen desanbiguazio lokal izeneko prozesua buru daiteke ondorengo prozesaketei begira. Horren arrazoia aldaeren analisian gertatzen den anbiguetatea da.

Horrela *kaletikan* formak analisi bakar bat eman beharrean —aurretik sinplifikatzeko aipatu den bezala— bi ematen ditu: *kaletik* eta *kalatik* formei dagozkienak hain zuzen ere; baina desanbiguazio lokalean lehenengoari dagokiona hautatuko da, aldaera bakar batez eratzen baita (*tik*-en ordez *tikan*), besteari bi aldaera dagozkion bitartean (aurreko bera gehi *a* organikoaren erabilpen okerra).

Desanbiguazio-prozesurako aurreko atalean aipatutakoa erabiltzen da, ondoko irizpideak jarraituz:

- Analisi bakoitzeko aldaeren kopuruak, lexikokoena, erregeletakoena eta guztirakoa kalkulatzen dira.
- Guztira zenbateko txikiena dutenak baino ez dira mantentzen, eta haien artean erregeletako aldaera kopuru txikienekoak. Honen arrazoia zera da: lexikoko aldaerak konkretu eta zehatzagoak dira erregeletakoak baino, beraz probabilitate gehiago dago hauek gerta daitezen.

Irizpideak sakontzeko corpus baten gainean egindako desanbiguazioak aztertu beharko lirateke, eta eskuzko prozesu batez akatsak detektatu eta zuzendu. Hori eginez gero irizpide konplexuagoak ondorioztatzea posible izango litzateke, aldaera-mota batzuei besteei baino pisu gehiago emanez. Hitzen edo lemen maiztasuna kontuan hartzea ere inteesgarria litzateke.

Desanbiguazio-prozesu hau *awk* programarako idatzitako *script* batez dago idatzita.

IV.2.4. Integrazioa lexiko-itzultzaileetan.

Aurreko bi kapituluetan aztertutako lexiko-itzultzaileak ere erabili ditugu aldaeren tratamenduari aurre egiteko. Emaitzak ez dira analisi morfologikoarenak bezain erabatekoak, baina zenbait ondorio interesgarri atera daitezke.

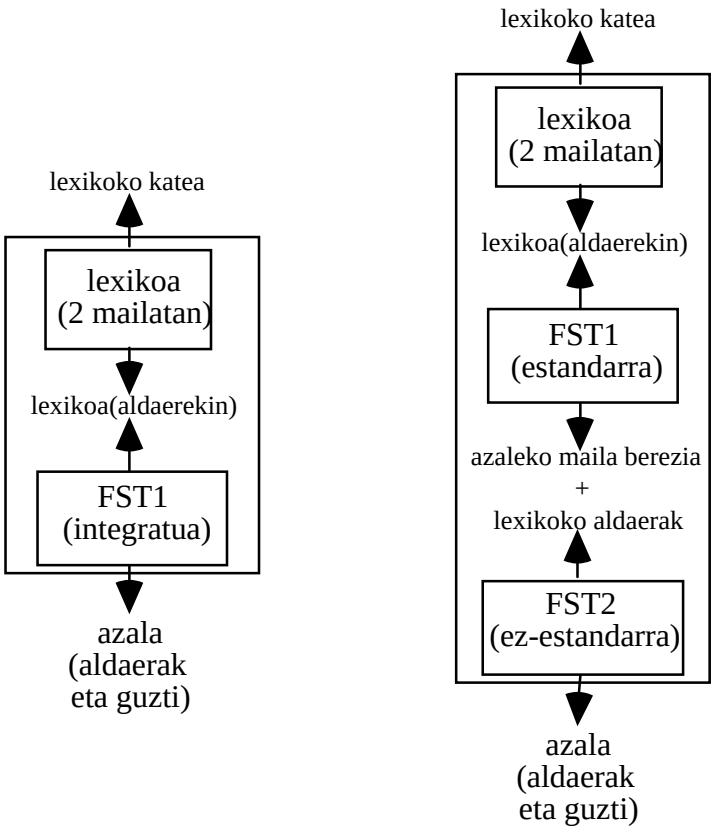
Lexikoan adierazten diren aldaeren aldetik arazorik ez dago, ezta forma estandarra lortzeko ere, lexikoan adieraz daitekeen tarteko adierazpidearen bidez lor baitaiteke. Horrela *haundi* eta *tikan* forma ez estandarren sarrerak hauek izango lirateke:

```
haundi ALDAERA/handi:haundi
tikan ALDAERA/tik:tikan
```

Hala ere guk egindako bi mailatako inplementaziotik desberdintasun bat badago: estandar/ez-estandar ezaugarria ezin da kudeatu azpilexiko mailan, baina lexikoko sarreretan *ALDAERA* ezaugarria jarriz bereiz daitezke lexikoko forma estandar eta ez-estandarrak.

Erregelen kasua, berriz, korapilatsuagoa da; eta, IV.5 irudian ikus daitekeenez, bi aukera aztertu dugu: erregela-sistema integratua eta banandua.

Jatorrizko erregelak eta erregela osagarriak batera daitezke erregela-sistema integratu batean, haien arteko gatazkak lehen aipatu den moduan ebatziz.



(A)
Erregela estandarrak eta ez-estandarrak integratuak

(B)
Erregela estandarrak eta ez-estandarrak maila desberdinetan

IV.5 irudia.- Aldaeren tratamendua lexiko-itzultzaileen bidez.

Izan ere, eta erregela-sistema anitz maila desberdinetan jartzeko lexiko-itzultzaileek ematen duten aukeraz baliatuz, saiatu gara ohiko erregela estandarrak ukitu gabe sistema osagarri oso bat eraikitzen. Ahalegin honetan arazo larri bat sortu da: aldaerak ezagutzeko erregela batzuek zenbait informazio morfologiko behar dute, morfema eta *a* organikoaren marka esaterako. Beraz, IV.5 irudian FST1[1] eta FST2ren artean dagoen adierazpidea ez da, hasiera batean pentsa zitekeen bezala, azaleko adierazpide hutsa —horrexegatik deitu dugu azaleko maila berezia. Honen ondorioz erregela-sistema estandarrean ukitu batzuk

[1] FST1 deitu dugun erregela-sistema bakarra edo bat baino gehiago izan daiteke (adibidez III.12 irudian daudenak: morfotaktikarako bat eta morfofonoligiarako beste bat)

egin behar dira, zenbait informazio morfologiko tarteko maila honetan manten dadin. Dena den, aldaketa horiek bi erregela-sistemak integratzeko egin behar direnak baino sinpleagoak dira.

Lexiko-itzultzaileak erabiliz, hala ere, eta aurreko bi aukeretako edozein hartuta, ezin da egiaztatu zenbait erregelaren arrakasta, horrek desanbiguazioari begira duen mugarekin.

IV.2.5. Emaitzak, konplexutasuna eta erabilpenak.

Aurretik azaldutako azpilexiko eta erregela osagarrien bidez, aldaerak analizatzeko eta sortzeko prozesadore morfologikoa osatu da. Hala ere, sorkuntzari begira esan behar da prozesadorea gainsortzailea dela; zeren eta, erregela osagarriak ahalik eta modu zehatzenean egiten saiatu arren, erregela hauek paraleloan aplikatzean aukera desberdin asko sortzen baita. Aipaturiko desanbiguazio-prozesuan aipatzen diren irizpideetan oinarriturik, aukera batzuk bazter litezke post-prozesu batean gainsorreraren arazo hau murriztearren.

Dena den ez da ahaztu behar aldaeren tratamenduaren helburu nagusia, eta ia bakarra, analisia dela, sorkuntza egitean modu estandarra interesatuko zaigu eta.

Kontzeptua	**1b-n**	**2b-n**	**bietan**
Ezagutu gabeko hitzak (guztira).	307	85	392
Erabilpen ez-estandarra	101 %100	28 %100	129 %100
Analizatutakoak	85 **%84,2**	22 **%78,6**	107 **%83**

IV.6 irudia.- Aldaeren analizatzailearen estaltze-tasa hitz-zerrendekin.

Aldaeren analizatzaileak duen estaldura-tasa aldaeren %90etik gorakoa da Corpusetan eta %80 inguruan hitz-zerrendetan; corpusetan gehien agertzen diren aldaerak jasota baitaude. Horrela, III.10 irudian azaltzen ziren datuetatik abiatuz, IV.6 irudian azaltzen diren emaitzak lortu dira.

Tasa hauek hobetzeko erregelak baino lexiko osagarria aberastu egin behar da, eta horretan jarraitzeko asmoa dugu.

Aldaeren analisia lortzeko, eta, batez ere erregela berriek eragiten duten aukeren biderkatze handiaren eraginez, analizatzeko denbora eta analisi bakoitzari dagokion analisi-urrats kopurua 2 edo 3 aldiz handiagoa da analisi estandarrekoa baino.

Erabateko abiadura-hobekuntza dakarten lexiko-itzultzaileen erabilera alde batera utziz, aldaeren analisia azkartzeko, gehien azaltzen diren aldaeren analisia *buffer* batean gorde daiteke eta, esan den bezala, aldaeren analisiari estandarrak emaitzarik ematen ez duenean soilik erabili.

Aldaeren tratamenduak bideratzen dituen aplikazioak bilduz hona hemen inportanteenak:

- analizatzailearen lana hobetzea, batez ere estaltze-tasa igoz, horren ondoren etor daitezkeen prozesuetarako abantaila izanik.
- zuzentzaile ortografikoari begira, forma ez-estandarren ordez forma estandarrak proposatzeko aukera ematen du; analisi ez-estandarren lexiko mailako emaitzak erabiliz sorkuntza estandarren bidez azaleko proposamen egokiak lor baitaitezke.
- ordenadorez lagunduriko irakaskuntzaren arloan euskaren morfologia eta ortografia lantzeko tresnak egin daitezke analisi estandarra eta ez-estandarra oinarritzat hartuz.
- dialektologia lantzeko tresna bezala erabiltzeko, eta beste garai bateko testuen azterketarako

IV.3 Lema lexikoan ez duten hitzen analisia.

Aurretik ikusitako hobekuntzak —erabiltzailearen lexikoarena eta aldaeren tratamendua— erabili arren, beti agertuko dira analisirik gabe gelditzen diren hitzak. Ondorengo aplikazioetarako, etiketatzaile/lematizatzaile edo analizatzaile sintaktikoa esaterako, funtsezkoa da analizatzailea sendoa izatea, hau da, edozein hitzetarako analisiren bat lor dezala. Bide horretan eta, III.5.2 atalean aipatu den bezala, kontuan hartuz ez analizatzearen arrazoia lexikoan lema ez egotea dela, bilatu dugu halako kasuetan analisirik sortzeko metodo bat, hau ere bi mailatako morfologian oinarriturik.

Nahiz eta lexikoko atal batzuk erabili “lexikorik gabeko analisia” deituko dugun prozesu hau, erabiltzailearen lexikoaren bidez konpon daitezkeen formen analisia beste modu batez ebazten da, bi metodoen arteko desberdintasunak hauek izanik:

- Erabiltzailearen hiztegian lemak sartzen badira, analisi zehatzagoak lortzen dira, baina horretarako eskuzko aurreprozesu bat behar da.
- Lexikorik gabeko analisiaren bidez eskuzko aberasketa ekiditen da, analizatzailea sendoa bihurtuz, baina horren truke anbiguetatea dezente altuagoa izango da.

Gure sisteman bi aukerak aurrikusi dira, eta lexikorik gabeko analisia bakarrik burutuko da aurreko analisi-saioak ezer lortzen ez dutenean. Gainera anbiguetatea jaisteko prozedura bat diseinatu da metodo honi dagokion alde negatiboena, ahal den neurrian behintzat, murrizteko.

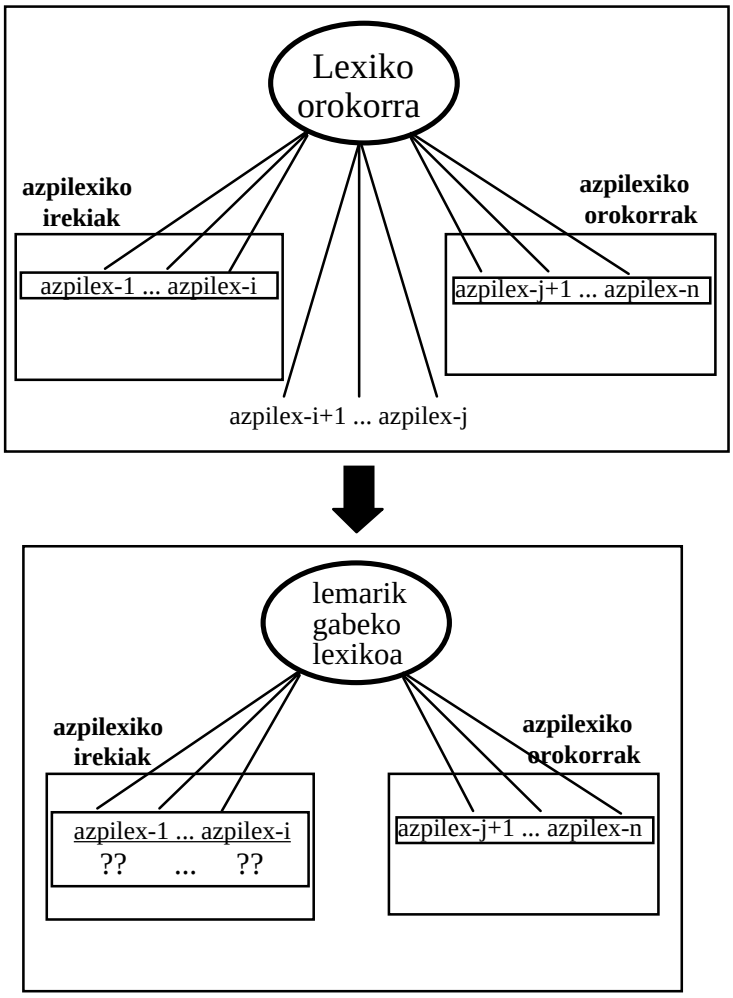
IV.3.1. Gakoa: bi mailatako erregela bereziak.

Lema lexikoan egon gabe, eta orokorrean lexikorik gabe, analisi morfologikorik lortu ahal izateko, azterturiko sistema gehienak atzizkien tratamenduan oinarritzen dira. Sistema batzuetan morfemak erabili beharrean bukaerako hitz-zatiak erabiltzen dira eredu probabilistikoan oinarriturik. Hau interesgarria izan daiteke, etiketa baino lortu nahi ez denean, baina ahalik eta analisi morfologiko osoena lortu nahi denean sistema hori ez da interesgarria unitate horiei informazio morfologikoa ez dagokielako, eta are interes gutxiagokoa euskara bezalako hizkuntza eranskarietarako.

Gai honen inguruan guk egindako aplikazioa fonologiarako egindako lan batean (Black *et al*., 91) oinarritzen da, horren gainean gure egokitzapena burutu dugularik. Lan horretan proposatzen zen muina izan da guk erabili duguna eta honetan datza:

- Analisia burutzeko lemak ez diren morfema-multzoak, aurrizki eta atzizkiak hain zuzen, bakarrik jartzen dira lexikoan. Gure kasuan orokortasunaren ezaugarria duten azpilexikoak izango dira morfema-multzo horiek.
- Lemen ordez, lema generiko batzuk kokatzen dira lexikoan, bakoitza interesatzen den informazioarekin, eta *??* bi karaktere[1] bereziren bidez ezagutzen direnak. Gure kasuan lema generiko hauek azpilexiko irekitan kokatuko dira, bat kategoria/azpikategoria bakoitzeko (lexikoaren aldetiko bihurketa IV.7 irudian ikus daiteke).
- Lexikoko bi karaktere berezi horien eta azaleko aukeren artean ezkontzea gobernatzeko bi erregela osagarri zehazten dira, karaktere berezien desagerpena kontrolatzeko bat, eta azaleko karaktere guztien sorrera bestea.

[1] Aipatutako erreferentzian ** karaktereak proposatzen ziren, baina guk horiek maiuskula-markatzat erabili ditugu.



IV.7 irudia.- Lexiko orokorretik lexikorik gabeko lexikora.

Morfotaktikaren informazioari dagokionez analisi orokor estandarrarena erabiltzen da funtsean, zenbait informazio soberan egon badaiteke ere desagerturiko lemei dagokielako. Morfofonologiaren aldetik, eta gehitutako bi erregelez aparte, analisi estandarrerako oinarrizko erregelak, edo behintzat gehienak, mantendu egin behar dira morfemen arteko loturetan eta hizkien barnean gertatzen diren aldaketak gobernatzen jarrai dezaten.

Zehaztasun gehiagotan sartu baino lehen azter ditzagun bi erregela berriak:

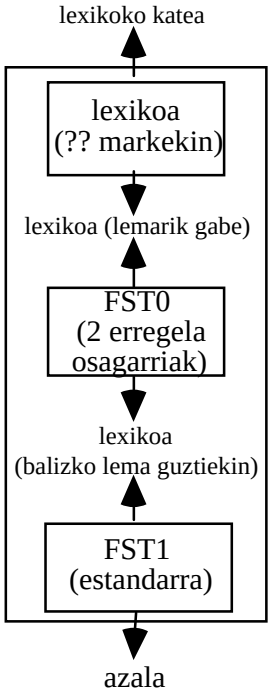
```
%?:0 => [ Hasiera | MM ] _ 0: ;
        0: MorfBuk ;
0:Cx => %?: [0: ]* _ [0: ]* %?: ;
            where Cx in (Kons Bokal) ;
```

Lehen erregelak marken desagerpena bideratzen du morfema baten hasieran eta bukaeran. Bigarrenak aldiz, bi marken artean azaleko karaktereen agerpena onartzen du lexikoan ezer agertzeko beharrik gabe. Erregela hauen testuinguruak konplexuagoak egin daitezke, horrela egiten zuten aipaturiko erreferentzian, sortzen diren azaleko karaktereen

konbinazioak hizkuntzaren konbinazio zilegiak direla egiazta dadin, bide batez anbiguetatea murriztuz; baina, horren truke, mailegaturiko hitz arrotzen analisi eragotz daiteke.

Erregela hauen agerpenak eragina du gainontzeko sistema orokorraren gainean, bi arrazoirengatik:

- Lemak desagertzean diakritikoak ez daude; beraz, gerta daiteke analisi morfologiko zilegia lortzeko aplikatu beharreko erregela batzuk ez aplikatzea hautapen-marka edo morfofonemaren faltarengatik. Honen aurrean lema generikoak errepika daitezke, bakoitzean lemetan agertu ohi diren diakritiko bat erantsiz (honek aipaturiko bi erregelak "ukitzera" eramango gaitu).
- Erregela estandar batzuetako testuinguruan lemei dagozkien lexiko-mailako karaktere arruntak zehazten dira, baina testuinguru hori ez da inoiz egiaztatuko, lexiko mailako karaktere guztiak, aurreko puntuan zehaztutakoak salbu, desagertu baitira, bi ikur bereziek ordezkatu dituztela eta. Arazo hau ebazteko erregelak banan banan aztertu dira eta kasu batzuetan ukituren bat egin da.



IV.8 irudia.- "Lexikorik gabeko analisia" lexiko-itzultzaile baten bidez.

Azken eragozpena ebatz daiteke askoz modu erraz eta dotoreagoan **lexiko-itzultzaileak** erabiliz. Beste behin erabiliko dugun erregela-sistema anitzen aukerari esker, lemaren gauzatzea kontrolatzen duten bi erregela berriak banandurik jar daitezke lexikotik hurbileneko mailan, eta ohiko erregelak ondoren, azalekiko bihurketak bidera

ditzaten (ikus IV.8 irudia). Honen bidez lortzen da ohiko erregelak bere horretan mantentzea, batere ukiturik gabe.

IV.3.2. Emaitza, lemaren bilaketa eta desanbiguazio lokala.

Lexikoan lema egon gabe burutzen den analisiaren emaitzak zilegiak dira kasu gehienetan, baina bi arazo azpimarratu behar dira:

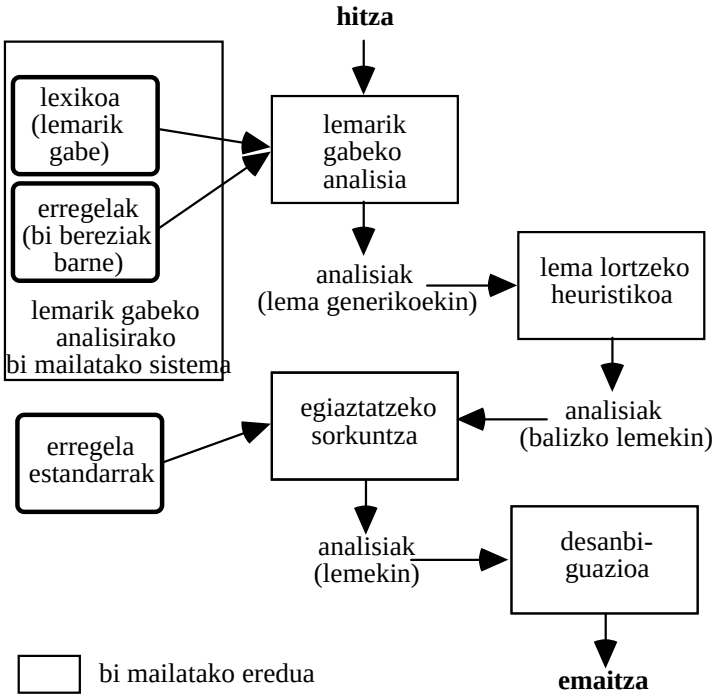
- lortzen diren analisietan agertzen den lema lema generikoa da, baina emaitza garden baterako benetako lema lortu behar da.
- analisi zilegiarekin batera beste analisi asko lortzen dira. Aukera guzti horien artean aukeratzea da desanbiguazio lokalaren helburua.

Zilegitasunaren aldetik salbuespen bat dago:

- hitzak erroreren bat duenean eta bere lema azpilexiko ez-ireki bati dagokionean. Hau zuzentzea gaitza da, dagoen irtenbide bakarra hauxe baita: lexiko estandarrean lema duten azpilexiko guztiekin irekiekin bezala jokatu. Honekin gutxitan gertatzen den arazo bat konpontzen da baina horren truke anbiguetatea asko igotzen da.

IV.3.2.1. Lemaren bilaketa

Bi mailatako formalismoari jarraitzen dion analisi osagarri honen emaitzan, lema zehatzaren ordez, lexikoan adierazi den lema generikoa agertuko da. Gainontzeko morfemen informazioa zehatza bada ere , emaitza hau ezin da zuzenean eskaini erabiltzaileari.



IV.9 irudia.- Lexikorik gabeko analisia lortzeko urratsak.

Lema generikotik benetako lemara pasatzeko prestatu dugun heuristikoak azaleko adierazpidea hartzen du erreferentzia nagusitzat, eta gerta daitezkeen aldaketak kontuan hartuz analisiari dagokion balizko lemak sortzen ditu. Balizko lema hauek gainontzeko morfemekin batera sorkuntza estandarretik iraganarazten dira, emandako azaleko forma lortzen ez dituztenak baztertuz (ikus IV.9 irudia).

IV.3.2.2. Desanbiguazio lokala

Lexikorik gabeko analisiak burutzean analisi asko lortzen da hitz bakoitzeko. Ez da arraroa forma batetik hogei analisi desberdin baino gehiago lortzea, eta hau ez da erabilgarria. Aipatutako artikuluan Black-ek eta bere lankideek hau aurrikusi zuten eta eragozpen hori konpontzeko desanbiguaziorako zenbait irizpide eman zuten, honako analisi hauek lehenetsiz: lema motzenak dituztenak —edo gauza bera dena, hizkien bidez zati luzeena ezagutzen dituztenak—, aplikatutako erregelen eta bereizitako hizkien probabilitatea.

Beraiek desanbiguatzeko zuten premia gure sistemarena baino handiagoa zen, zeren ahoskerarako aukera bakarra aukeratu behar baitzen. Gure kasuan aukera bat baino gehiago hauta daiteke, desanbiguatzeko gainontzeko lana testuingurua kontuan hartzen duten beste prozesuetarako utz baitaiteke.

Gure desanbiguazio lokalean jarraitu diren irizpideak hauexek izan dira:

- Kontrakoa erabakitzen ez den bitartean kategoria bakoitzeko gutxienez analisi bat lortuko da.
- Kategoria bereko analisien artean lema motzenak dituztenak aukeratuko dira, letra bateko aldea duten guztiak ere mantentzen direlarik.
- Puntu ondoren etorri gabe maiuskulaz hasten diren hitzetan, pertsona- eta leku-izena ez diren aukerak baztertzen dira.

Desanbiguazio-prozesu hau arintzeko, egokia iruditu zaigu eratorpen-atzizki ohizkoenak integratzea lexikorik gabeko lexikoan, horrela hobeto bideratuko baitira aurreko kapituluko III.10 irudian B3 kodearekin jasotzen diren eratorpen “berri”en analisiak —ez ezagututakoen artean %10etik gora direnak—. Halako eratorpen berrietan eratorpen-morfema ezagutzen bada lema motzago izango da, desanbiguazio-prozesua argituz.

IV.4 Analizatzaile sendoa. Emaitzak.

Kapitulu honetan aztertutakoarekin aurreko kapituluan azaltzen zen prozesadore morfologiko estandarra osatu egiten da, analizatzaile sendo eta orokor bat lortzeko asmoz.

```
/<zergatik>/
    ("zergatik"   ADB)
    ("zerga"   IZE + DEK NUMS MUGM + DEK ABL)
    ("zergati"   IZE + DEK ERG MG)
 /<ez>/
    ("ez"   ADB)
    ("ez"   IZE + DEK NOM MG)
    ("ez"   IZE)
 /<zuen>/
    ("*edun"   ADL B1 NOR3 NRK3 + ERL ERLT)
    ("*edun"   ADL B1 NOR3 NRK3 + ERL ZHG)
    ("*edun"   ADL B1 NOR3 NRK3)
    ("zu"   IOR + DEK GEN NUMP MUGM)
 /<gorputza>/
    ("gorputz"   IZE + DEK NOM NUMS MUGM)
 /<haundituta>/
    ("handi"  /haundi/  ADI + ASP PART + ERL MOD)
 /<.>/<PUNT_PUNT>/
 /<Baina>/<HAS_MAI>/
    ("baina"   ADI)
    ("baina"   IZE + DEK NOM MG)
    ("baina"   IZE + DEK NOM NUMS MUGM)
    ("baina"   IZE)
    ("baina"   JNT)
 /<,>/<PUNT_KOMA>/
 /<ur>/
    ("ur"   ADI)
    ("ur"   IZE + DEK NOM MG)
    ("ur"   IZE)
 /<guti>/
    ("gutxi"  /guti/  ADI)
    ("gutxi"  /guti/  ADJ + DEK NOM MG)
    ("gutxi"  /guti/  ADJ)
    ("gutxi"  /guti/  IOR + DEK NOM MG)
 /<irentsi>/
    ("irents"   ADI + ASP PART + DEK NOM MG)
    ("irents"   ADI + ASP PART)
 /<zuela>/
    ("*edun"   ADL B1 NOR3 NRK3 + ERL DENB)
    ("*edun"   ADL B1 NOR3 NRK3 + ERL KONP)
    ("*edun"   ADL B1 NOR3 NRK3 + ERL MOD)
 /<,>/<PUNT_KOMA>/
/<pentsatu>/
    ("pentsa"   ADI + ASP PART + DEK NOM MG)
    ("pentsa"   ADI + ASP PART)
 /<nuen>/
    ("*edun"   ADL B1 NOR3 NRK1 + ERL ERLT)
    ("*edun"   ADL B1 NOR3 NRK1 + ERL ZHG)
    ("*edun"   ADL B1 NOR3 NRK1)
 /<gero>/
    ("gero"   ADB)
    ("gero"   IZE + DEK NOM MG)
    ("gero"   IZE)
    ("gero"   JNT)
/<.>/<PUNT_PUNT>/
```

IV.10 irudia.- Testu-zati baten analisia hiru urratsak pasa eta gero.

Horrela, testuak analizatzeko orduan, hitzen analisi estandarrak lortzen diren bitartean —erabiltzailearen hiztegiak horretan lagun dezakeela— ez dira kontuan hartzen aldaera izateko aukerak, eta analisi estandarrean zein aldaeren analisian ezer lortzen ez denean bakarrik burutuko da lexikorik gabeko analisia.

Analisiaren emaitzak anbiguoak izan daitezke eta irekitako ikerlerrotzat dugu hitz bati dagozkion analisi guztien arteko desanbiguazioa, lan hau EUSLEM proiektuaren barruan garatzen ari garela (Aldeazabal *et al*., 94) (Aduriz *et al*., 95).

Testu bat hiru analisi-aukeretatik —estandarra, aldaerena eta lexikorik gabekoa— pasa eta gero lortzen den emaitza C eranskinean ikus daiteke. Hala ere IV.10 irudian zati txiki bat azaltzen da. Ematen den emaitza tratatua izan da, token-ezagutzailea lortutako informazioa erantsiz eta analisi-aukera bakoitza lerro bakar batean azalduz. Analisi bakoitzean lema eta aldaera ager daiteke, baina, analisi estandarretan lema bakarrik agertzen da, eta lexikorik gabeko analisietan lema hipotetikoa aldaera bezala agertzen da.

Kontzeptua	**A** (Argia)	**B** (Filosofia)	**A+B**
Hitzak (corpusa)	4.864	2.343	7.207
Hitz desberdinak (zerrenda)	2.607	1.429	4.036
Zerrendako hitzen artean ezezagunak	307	85	392
analizatzaile estandarrerako	%12	%6	%10
Aldaerak	101	28	129
Analizatutako aldaerak	85 (%84)	22 (%79)	107 (%83)
Erroreak	21	4	25
Zehaztasuna	**%99,2**	**%99,7**	**%99,4**

IV.11 irudia.- Analizatzaile morfologikoari buruzko estatistikak

Estaldura-tasari dagokionean %100 da ia lexikorik gabeko analisiari esker, baina gerta daiteke hitz batzuen analisia desegokia izatea; beraz, zehaztasun-tasari begiratu beharko zaio orain, hau da, analisi egokirik dutenen proportzioari. Corpus txikiekin egindako probetan zuzentasuna[1] %99tik gora dela egiaztatu da (ikus IV.11 irudia).

[1] Aldaeren analisian eta lexikorik gabeko analisian emaitza zehatzat jotzen da analisi zilegia agertzen baldin bada, beste analisi hipotetiko desegokiak egon arren. Desanbiguazio-prozesuaren lana izango da analisi egokia aukeratzea.

BIGARREN PARTEA: ZUZENKETA ORTOGRAFIKOA

V. Erroreen zuzenketa.

Euskararako zuzentzaile ortografiko bat burutzeko ideia taldearen helburu nagusien artean zegoen hasiera hasieratik. Horri ekiteko arrazoi nagusia, premia zen, ez baitzegoen halako produkturik euskararako, baina, baita ere, martxan dagoen batasun-prozesuak are interesgarriago bihurtzen zuelako halako tresna bat.

Berrerabilgarritasun eta zehaztasun irizpideak hobetsiz, horrek eraginkortasunaren aldetiko galera eragin bazezakeen ere, bi ideia nagusitu ziren laster: egiaztatzaile/zuzentzailea morfologian oinarritu behar zela batetik, eta martxan den batasun-prozesu irekiak eragiten dituen akatsei aurre egiteak lehentasuna zuela bestetik.

Izan ere, ideia horiek ondo finkatzeko eta arazoei aurre egiteko, gai honen inguruko kontzeptu eta ideia nagusiak eztabaidatzen eta argitzen joan ginen, bibliografia lagun genuela —azpimarratzekoa da Kukich-ek *ACM Computing Surveys*-erako (1992) idatzitako bilketa, gaur egun arlo honetan sartzeko oso gomendagarria dena.

Ideia horiek kapitulu honetan biltzen dira, euskararako egin dugun gauzatze konkretua hurrengorako utziz. Zuzenketaren arlo honetan aurrerapen eta proposamen asko dagoenez, lehenengoz mugatu dugu gure lanaren helburua, hitz isolatua aztergai hartuz, eta testuingurua kontuan hartzen duen zuzenketa ondorengo proiektuetarako utziz. Muga hori ezarri ondoren, hartutako esparruan kokatzen diren teknikak gainbegiratzen dira: hitzen ezagutza batetik, eta ez-onartuei dagozkien ordezkapen/proposamenei buruzkoak bestetik.

Zuzenketari buruz asko ikertu arren, helburu-hizkuntza eranskaria edo flexio-aukera handikoa denean, problematika konplexuagoa izanda ere, erreferentzia bakan batzuk baino ez dira aurkitzen.

Kapitulu hau lau ataletan banatu da: aplikazioak, hitzen egiaztapena, hitzen zuzenketa eta flexio handiko zein hizkuntza eranskarietan aurkezten diren arazo bereziak.

V.1. Aplikazioak, sailkapena eta irizpideak.

Testuen zuzenketa aplikazio desberdinetarako garatzen ari den ikerkuntza-arlo irekia da. Erabilpen ezagunenak testuen edizioaren eta karaktereen ezagutza optikoaren (OCR) esparruetan kokatzen badira ere, pertsona-ordenadore interfaze sistema orok, komando-lengoaia erabiltzen dutenak barne, hobetzen dira halako teknikak erabiliz. Horien artean honako hauek azpimarra daitezke: datu-baseen gaineko biltegitze/berreskuratze interfazeak, lengoaia naturalezko interfazeak, OLI sistemen interfazeak —eta baita beraien ezagutza-basea ere OLIren helburua idazketa denean—, testu-hizketa hizketa-testu bihurketak, eta elbarritu edo behar berezietako pertsonentzako komunikazio-sistemak.

Aplikazioaren arabera betebehar eta ezaugarri desberdinak egon arren, aplikaturiko estrategiaren arabera bi multzo handi bereizten dira:

- Elkarrekintzazko zuzenketa: erabiltzailearen esku utzi ohi da azken erabakia erroreak ordezteko orduan —gutxienez programak zalantza-tarte handia duenean—. Aplikazio tipikoa zuzentzaile ortografikoa da. Zuzenketarako proposamen bat baino gehiago lor daiteke emaitza gisa, eta hauen artean probabilitate-sailkapen bat egiten da. Proposamenik ez eskaintzea ez da oso egokia baina onar daiteke salbuespen gisa. Testuingurua kontuan hartzea ez da ezinbestekoa baina bai lagungarria, beste moduz detektaezinak diren akatsak aurki baitaitezke eta proposamenak zehatzago ordena baitaitezke. Testuingurua kontuan hartzen bada, zuzentzaile hauen zehaztasuna igotzeaz gain, sintaxia eta estiloa kontuan hartzen duten zuzentzaile aurreratuak egin daitezke.

- Zuzenketa automatikoa: zuzenketarako proposamen bakar bat lortu behar da, giza-laguntzarik ez dago eta. Testuingurua kontutan hartzea ezinbestekoa da emaitza dotoreak lortzeko. Denbora errealeko hizketa-ezagutza bezalako aplikazioetan behar da.

Ohizkoa da, Kukich-ek aipaturiko artikuluan (Kukich, 92) horrela egiten duen bezala, gai honi buruzko ikerkuntza hiru ataletan banatzea[1]:

- hitz ezezagunen detekzioa edo testu-egiaztatzea
- testuingururik gabeko hitz-zuzenketa
- testuinguruaren araberako hitz-zuzenketa

Lehen atalean ez dira sartzen Mitton-ek (1987) *real-word errors* deitutakoak —*benetako hitzaren erroreak* deituko ditugunak—, hitz ezaguna sortzen duten errore hauek detektatzeko testuingurua aztertu behar delako, hori dela eta hirugarren atalari dagozkion tekniken bidez tratatu behar direlarik.

Hirugarren atalari dagozkion teknikak txosten honek jasotzen duen proiektutik at daudenez, taldearentzat irekitako ikerlerro bat bada ere (Agirre *et al.*, 94) (Gojenola & Sarasola, 94), teknika-multzo horretaz ez gara ariko lan honetan zehar. Bibliografia gisa, Kukich-en aipatutako artikuluaren hirugarren kapitulua eta Vosse-rena (1992) aipa daitezke.

V.2.Egiaztatzea.

Zuzenketa bideratzeko ezinbesteko lehen urratsa hitzen arteko bereizketa edo egiaztatzea da, hau da, testuan zehar agertzen diren hitzen artean aukeratzea zeintzuk diren zilegiak edo ezagutuak eta zeintzuk ez.

Prozesu honetan egiaztatzailearen helburua ahalik eta zehaztasun handienaz lan egitea da. Zehaztasun-tasa jaisten duten bi akats-mota gertatu ohi dira egiaztatze-prozesuan (Peterson, 80):

- Erroretzat hartzen diren hitz zilegiak. Muga batetik behera mantentzen diren bitartean behintzat, elkarrekintzazko aplikazioetan oso kezkagarriak ez diren bitartean, erabiltzaileak ontzat emango baititu, zuzenketa automatikoan erabat ekiditera jo behar da, erroreen kopurua handitzen dute eta.
- Zilegitzat hartzen diren erroreak. Bi multzotan bereizten dira, hizkuntzan existitzen ez diren hitzak direnak batetik, eta, bestetik, *benetako hitzaren erroreak* deitu ditugunak, hau da, errorearen ondorioz testuinguru horretan ez baina hizkuntzan zilegia den hitza sortzen dutenak. Esan den bezala azken hauek lan honetatik kanpo utziko ditugu.

[1] Kukich-en terminologiaren arabera *nonword error detection*, *isolated-word error correction* eta *context-dependent word correction*.

Egiaztapena burutzeko metodoak sailkatzeko orduan irizpide desberdinak erabiltzen dira. Askotan metodo heterogenoak erabiltzen diren arren, bilketa honetan hiru multzotan banatu ditugu metodo hauek, erabiltzen dituzten datuen arabera: hitz-zerrenden bidezkoak, hitz-zatien bidezkoak eta morfologian oinarritutakoak.

Bakoitza bere aldetik aztertuko dugu, dagozkien abantailak eta eragozpenak azpimarratuz eta erreferentziaren bat aipatuz[1].

V.2.1 Hitz-zerrendetan oinarritutako metodoak

Sistema hauetan hitza onartu egiten da baldin eta hitz-zerrendan agertzen bada. Zehaztasun minimo bat lortzeko behinik behin, hitz-zerrenda oso luzea izaten da: hizkuntzaren arabera aldatzen da, baina 50.000 edo 100.000 hitzetik gorakoa izan ohi da. Datu-kopuru horiekin lan egiteko datuak gordetzeko eta atzitzeko bideak sakonean aztertu behar direlarik.

Gainera, halako hiztegi erraldoi bat eraikitzeko hitz askoko corpus zuzen bat (errorerik gabekoa) behar da iturburu gisa, miloi bat hitzetatik gorakoa gomendatzen da. Horrela egiten ez denean zehaztasunaren kalterako izaten da eta honako ondorio hauek izaten ditu: corpus txikien bidez hitz zilegi batzuk erroretzat hartuko dira, eta zuzendu gabeko corpusen bidez, berriz, errore batzuk ontzat hartzeko arrisku dago. Azken hau oso kaltegarria litzateke euskara bezala ondo batu gabeko hizkuntzetan.

Metodo horien ezaugarriak hauek izaten dira orokorrean:

- Flexio aberatsa duten hizkuntzetarako desegokia, gordetzeko zerrenda izugarri luzatzen delako, eta ondorioz “ikasteko” corpusak askoz handiagoa izan behar duelako. Gainera, sistema ez da malgua izango jakintza-arlo berri edo termino berrientzat, erro berri bakoitzeko bere deklinabide guztia sartu beharko litzatekeelako.
- Flexio txikiko hizkuntzetarako aurreko arazoak saihets badaitezke ere, beti iraungo du batek, koherentziarenak alegia. Erabiltzaileari oso ulergaitz egiten zaio erro baten flexio batzuk onartzen dituen bitartean beste batzuk errefusatzen direla ikustea, azken hauek probabilitate txikiagokoak izan arren.
- Garatzeko azkarrak eta merkeak.

[1] Kasu praktikoen adibide asko eman litezke, oso bibliografia oparoa dago eta. Hala ere, nahiago izan dugu ezaugarrien arabera gauzatze garrantzitsuenak azpimarratzea erreferentzia gehiegi ematea baino. Kukich-en aipaturiko artikuluan bibliografia osoa aurki daiteke.

- Ondoren aipatuko diren teknikak erabiliz, abiaduraren aldetik azkarrak izaten dira eta arazorik ez dute biltegi-tokiaren aldetik.

- Zehaztasuna, iturburu-corpusaren eta egiaztatzeko testuaren araberakoa izango da, baina aipatutako corpusa zuzena bazen ziurta daiteke ez dela ontzat emango hizkuntzan existitzen ez den hitzik. Dena dela testua erreferentzia-corpusarekin bat ez badator —jakintza-arlo bereziak, mailegu berriak, etab. direla eta— hitz zilegi batzuk ez dira ezagutuko.

Metodo honen bidez hitz-zerrenda edo hiztegia oso luze izan daitekeenez, hiztegia gordetzeko teknika desberdinak ikertu dira zehaztasunaren, toki-hartzearen eta atzipen-denboraren artean ahalik eta orekarik handiena lortzeko.

Teknika horietako batzuk aztertuko ditugu ondoren:

- Mailakako biltegitzea (Peterson, 80) memoria-hierarkien ideiari jarraituz hiru maila bereizten dira: lehena txiki samarra eta azkarra, maiztasun handieneko hitzak azkar atzitzeko —testuetako hitzen %50a hartzen duten hitzak kokatu ohi dira bertan—; bigarrena, dokumentuan bertan lehenago azaldu diren hitzak tarteko abiaduraz atzitzeko, eta hirugarrena, masa-biltegia, atzipen-abiadura motelenarekin. Hizkuntzaren arabera alda badaiteke ere, lor daiteke azken maila hori atzitu behar izatea %10ean baino gutxiagotan.

- *Hashing*-teknikak erabiltzea, taula batzuk eraikiz hiztegiaren atzipena azkartzeko. Gakoa kolisio gutxi sortzen duen *hash*-formula egokia aurkitzea da. Sistema batzuetan, Unix-eko *spell*-en (McIlroy, 82) adibidez, hitza gorde beharrean bere agerpena adierazten duen bit bat baino ez dute gordetzen. Horrek trinkotzen du hiztegia, baina, horren truke, hitz zilegiekin kolisioa sortzen duten akatsak onartzera darama.

- Bilaketa azkarra bideratzen duten zuhaitz-egiturak eta bigarren kapituluan aztertu den *trie egitura*.

Hitz-zerrendetan oinarritutako metodo hau izan da erabiliena, orain dela gutxi arte eta flexio-sistema sinplea duten hizkuntzetarako behintzat, zuzentzaile ortografikoen arloan; joera aldatzen ari da, ordea, eta morfologiaren bidezko tratamenduak —sarritan sinplifikatuak— ugalduz doaz.

V.2.2 Hitz-zatietan oinarritutako metodoak

Hitzak ezagutu gabe akatsak detektatu nahi direnean edo hitz-zerrenda luzeegiak gertatzen direnean aplikatzen dira aurrekoen aldaeratzat har daitezkeen teknika hauek.

Corpus batean oinarriturik ere, hitz-zati horiei buruzko informazioa lortu eta gordetzen da. Corpusak, normalean, ez du aurreko teknika-multzokoa bezain handia izan behar, baina zuzen-zuzena izatea horietan bezain komenigarri edo gehiago da.

Zatiak bi motakoak izan daitezke:

- *n-grama*k: *n* karaktereko luzera duten hitz-zatiak. *n* handitu ahala zehaztasun handiagoa lortzen da, baina toki-hartzea ere handiagoa da. n-gramen posizioa kontuan har daiteke zehaztasuna hobetzeko toki-hartze handiagoaren truke (Hull & Srihari, 82). Trigramak dira n-grama erabilienak, zehaztasun eta toki-hartzearen artean oreka onena lortzen delakoan (Zamora *et al*., 81).

- Luzera aldakorreko zatiak. Trataera eta burutzapen konplexuagoa eskatzen du, sasi-morfemak bilatzea baita helburua. Horretarako aipatutako corpusa konpilatu behar da, hitz-zatien artean loturak inferitzeko, adibidez egoera finituko automata bat sortuz (Aho, 90) (Meddeb, 94).

Hitz-zatietan oinarritutako metodo horien ezaugarriak hauek izaten dira orokorrean:

- Zehaztasunaren aldetik du arazo nagusia, bi motako akatsak egiten direlako: existitzen ez den hitzak zilegitzat hartu —hau da arazo nagusia— eta zilegi diren hitzak erroretzat. Hala ere, azken multzo honi dagokionez, corpusetan agertzen ez ziren forma zilegiak onartzeko gaitasuna du, hitz-zatien arabera zilegiak badira.

- Malgutasun falta, ia ezinezkoa baita sistema aberastea zehaztasuna hobetzeko.

- Garatzeko azkarrak eta merkeak, n-grametan oinarrituak behintzat.

- Azkarrak izaten dira eta, biltegi-tokiaren aldetik, oso trinkoak.

- Koherentziaren arazoa. Aurreko teknika-multzoan bezala, baina probabilitate txikiagoarekin, flexio batzuk onartzen diren bitartean beste batzuk errefusaturik izan daitezke.

n-grametan oinarrituriko sistemak OCR motako aplikazioetan erabili ohi dira, honako arrazoi hauengatik: sortzen diren erroreak nahikoa espezifikoak dira n-grama arraro samarrak sortuz, eta hizkuntzaren ohizko n-gramekin bat datozen hitz “berriak” onartzen direlako ezer eguneratzen ibili gabe.

V.2.3 Morfologian oinarritutako metodoak

Hitz bat zilegia den ala ez erabakitzea, hitzak deskonposaketa morfologiko zilegia duenentz aztertzea da; hau da, hitz bat ontzat ematen da deskonposaketa morfologikorik baldin badu. Beraz, analizatzaile morfologiko baten bidez bidera badaiteke ere, lan honetarako ez da analizatzaile osoa behar, deskonposaketa morfologiko posiblerik duenentz zehaztea nahikoa delako.

Metodo hauek hitz-zerrendetan oinarritutakoen alternatiba dira, hizkuntzaren flexioa aberatsa denean eta baita hitz-zerrenda laburtu nahi denean ere. Hajic-ek eta Droza-k (1990) sarreran diotena aldatuko dugu hona:

> " ... From different reasons, among which the speed of processing prevails, they are usually based on dictionaries of word forms instead of words. This approach is sufficient for languages with little inflection such as English, but fails for highly inflective languages such as Czech, Russian, Slovak or other Slavonic languages. ..."

Aurreko sistemen bilakaera "logikotzat" har daiteke teknika-multzo hau. Morfemak lirateke luzera aldakorreko hitz-zatiak, eta corpusetik lortzen den informazioa morfologiari buruzko ezagutza.

Morfologian oinarritutako metodoen ezaugarriak hauexek dira:

- Garatzeko garestiak dira denbora eta kostuaren aldetik. Horren truke berrerabilgarritasuna dugu, egindako lana helburu anitzekoa baita.
- Koherentziaren eta malgutasunaren aldetik sistema onenak dira. Erro berri bat behin sartuz gero bere forma flexionatu guztiak, eta kasu batzuetan forma eratorriak eta elkartuak ere, ezagutzen dira; ondorioz, sistemaren aberasketa erraztu egiten da.
- Biltegi-tokiaren aldetik aurreko bi teknika-multzoen artean dago. Abiaduraren aldetik, berriz, formalismo morfologikoaren menpe izanda ere, besteak baino motelago izan ohi da. Azkartzeko maiztasun handieneko hitzen zerrenda batekin konbinatu ohi da.
- Zehaztasunaren aldetik inoiz ez da ontzat emango hizkuntzan existitzen ez den hitzik, morfologiaren bidez gainsorkuntza onartzen ez bada behintzat. Ezagutzen ez diren hitz zilegien kopurua lexikoaren araberakoa izango da, baina, esan den bezala, aberasketa erraza eta koherentea bideratzen duenez, erabiltzailearen esku utz daiteke aberasketa hori. Hori dela eta, idazlearen estiloari eta jakintza-arloari ondo egokitutako erabiltzailearen lexiko baten bidez oso zehaztasun handia lor daiteke.

- Lortutako informazio morfologiko partziala interesgarria izan daiteke zuzenketa garaian. Inguruko hitzen informazio morfologikoak testuingurua kontuan hartzen duen zuzenketa bidera dezake.

Flexio aberatsa duten hizkuntzetarako ezinbestekotzat jo daitekeen bitartean, gainontzeko hizkuntzetarako gero eta erabiliagoak dira, elkarrekintzazko aplikazioetan batez ere: (Means, 88), (Hajic & Droza, 90), (Solack & Oflazer, 93), (Aduriz *et al*., 93), (Oflazer & Guzey, 94), (Vagelatos *et al*. 95).

V.3.Zuzenketa.

Ezagutzen ez diren hitzak zein diren jakin eta gero —aplikazioaren arabera hitz susmagarriak edo erroreak deitzen zaie— forma horien kudeaketa dator. Kudeaketa hori automatikoa izan daiteke, eta, orduan, zuzenketa automatikoa deitzen zaio, edo erabiltzailearen laguntzaren bidezkoa, kasu honetan proposamenen sorkuntza eta sailkapena izena egokiago delarik. Gure helburua bigarren kudeaketa-mota hori bada ere, bi multzoetan erabiltzen diren teknikak nagusiki amankomunak dira: proposamenak egiteko erroreen tipologia eta ezaugarriak aztertu behar direlako, eta, hautapen bakarra edo sailkapena egiteko, hurbilpen- edo antzekotasun-neurriak aztertu behar dira irizpideak finkatzeko.

V.3.1 Errore-motak eta ezaugarriak.

Erroreak zuzentzeko, edo dagozkien proposamenak lortzeko, erroreen ezaugarriak aztertzea interesgarria da oso. Erroreei buruz sailkapen desberdinak egin badaitezke ere, honako hiru multzotan sailkatzea proposatzen dugu argigarria delakoan:

- Oinarrizko erroreak: aplikazio guztietan agertzen dira mota honetako erroreak, jatorriak desberdinak izan arren. Askotan errore tipografikoak deitzen dira.
- Aplikazioarekin lotutako berezitasunak: testua lortzeko bidearekin lotura zuzena duten erroreen ezaugarriak kokatzen dira honetan.
- Hizkuntzaren ezaugarriekin lotutako erroreak, hizkuntzaren zenbait ezaugarri ez ezagutzeak, nahasteak edo aldaketa dialektalak eraginda, gizakiek modu kontzientean egindako erroreak dira beti. Aldaerak edo gaitasun-erroreak deituko ditugu.

V.3.1.1 Oinarrizko erroreen sailkapena.

Damerau-ren (1964) tipifikazioa hartzen da oinarritzat garatutako zuzenketa-aplikazio ia guztietan. Sailkapen honen arabera errore gehienak —testu-edizioan %80a da Damerauk ematen duen neurria— hauetako lau gertakizunetako bakar baten eraginez sortzen dira:

- Karaktere baten **aldaketa**. Karaktereetako bat ez da jatorrizkoa, bere ordez beste bat kokatu delako. Adib. *kame kale*-ren ordez
- Karaktere baten **sorrera**. Karaktere bat gehiago dago jatorrizko forman baino, bertako biren artean edo mutur batean txertatu da eta. Adib. *ssistema sistema*-ren ordez
- Karaktere baten **desagerpena**. Karaktereetako bat desagertu da, jatorrizko hitza karaktere batean laburtuz.Adib. *ed edo*-ren ordez
- Bi karaktere jarrairen arteko **trukea**. Bi karaktere jarrairen artean ordena aldatu egin da, hitzaren luzera mantenduz baina bi posizioetako karaktereak aldatuz. Adib. *bania baina*-ren ordez. Hauek ez dira orokorrak teklatua erabiltzen den aplikazioetan bakarrik agertzen baitira; beraz, gainontzeko aplikazioetan ez da kontuan hartuko.

Hitz batetik sor daitezkeen errore bakunak —hitz osoan aipatutako erroreetako bakar bat duten formak— kuantifikatu egin dira, eta *n* luzera duen hitz baterako hauexek dira kalkuluak (hizkuntzaren alfabetoko karaktere-kopurua *k* izanik): *n(k-1)* aldaketa, *(n-1)k* sorrera, *n* desagerpen eta *(n-1)* truke. Beraz, konbinazio kopurua *2nk*-tik oso gertu dago. Hala ere, hauetako konbinazio asko eta asko bazter daitezke, hasieratik metodo estatistikoak (n-gramen analisien bidez adibidez) erabiliz (Pollock & Zamora, 84).

Bakunak ez diren erroreak gertakizun hauen konbinazioaren bidez adieraz daitezke, eta **errore anitzeko akatsak** —*multi-error misspelling*— deitu ohi zaie. Hauen kopuruaz oso datu kontrajarriak daude: Pollock-ek eta Zamorak %6a aipatzen duten bitartean, Mitton-en (1987) ustez %31raino iristen dira. Badirudi datuen eta aplikazioaren arabera oso aldakor izan daitekeela.

Hitzen luzera eta erroreen arteko eraginaz zenbait zehaztasunen berri ematen da. Errore gehienak bakunak direnez, zera inferi daiteke: erroredun hitzaren eta jatorrizkoaren arteko luzera-diferentzia bat edo gutxiago dela. Hori dela eta, hitz-zerrendetan oinarritutako zuzentzaileak luzeraren arabera antolatzen dute hitz-zerrenda askotan. Hitz luzeetan motzetan baino errore gehiago egiten ote den ez dago argi, baina gaizki zuzendutako hitzak motzak izaten dira askotan.

Gertakizun bakunen bidez gerta daitezke aipatu behar diren bi fenomeno: benetako hitzaren erroreak eta hitz-mugaren gaineko erroreak.

Benetako hitzaren erroreak, aurretik esan den bezala gure lan-eremutik at daude, baina beren kuantifikazioa interesgarria da —ideia ematen baitigu testuingururik gabeko tratamenduen mugaz—. Honetaz ere neurriak ez datoz bat; horrela eta beti ingeleserako hartutako neurriez, Peterson-en (1986) ustez behe-muga %16a den bitartean Mitton-ek (1987) %40a aipatzen du. Kontuan hartzekoa da bi esperimentuen arteko desberdintasuna zuzentzaile arrunten erabileraren eragina izan daitekeela, hein batean behintzat. Alegia, zuzentzaile arrunten erabilerak errore-kopurua laburtzen du benetako hitzarenak kenduta, ondorioz azken hauen portzentaia igoz.

Aipatutako datuak ingeleserako datuak dira, eta beste hizkuntzetarako ez da halako daturik aurkitzen. Euskara bezalako hizkuntza eranskarietan portzentaia hori txikiagoa izatea espero daiteke hitzak luzeagoak izan ohi direlako eta zera baitago frogatuta: benetako hitzaren errore bat sortzeko probabilitatea txikiagoa dela hitz luzeetan, motzetan baino. Gainera, zuzentzaile ortografikoen erabilera murritzagoa dela eta bestelako erroreak ez dira gutxiagotzen.

Hitz-mugaren gaineko erroreak askotan ez dira kontuan hartzen, beren tratamendu konplexuagoa dela eta, *token* bat baino gehiago kontuan hartu behar baita. Hala ere, oinarrizko errore bezala ikus daiteke zuriunea[1], karaktere arruntzat jotzen baldin bada. Bi multzotan bana daitezke errore hauek: zuriunearen galerarengatik bi hitz bakar batean biltzekoak (*run-on words*), eta zuriuneren baten agerpenarengatik hitz bat bitan zatitzekoak (*split words*). Kukich-en ustez (1992), detektaturiko erroreen %15a mota honetakoak dira (%13 eta %2 hurrenez hurren). Tratamendu egokirik gabe hauei dagozkien zuzenketak edo proposamenak desegokiak izango dira. Mota honetako erroreek hitz ezagun bat noiz sortzen duen ere azterturik dauka Mitton-ek.

Errore hauen tratamenduari dagokionez, berriz, tratatzen dituztenen artean bi multzo bereiz daitezke: tratamendu berezitu partikularra ematen dutenak (Pollock &Zamora, 84) (Kernighan, 91) batetik, eta *lattice*[2] izeneko sare batez teilakatze-aukera guztiak aztertzen dituztenak (Carter, 92).

---

1 Zuriunea hitz-mugaren sinonimotzat hartuko dugu.

2 Vosse-k (1992) *lattice* egitura bera proposatzen du lokuzioen eta hitz anitzeko terminoen tratamendurako.

V.3.1.2 Aplikazioarekin lotutako erroreak.

Aurreko atalean azaldutako baieztapen edo neurri batzuk berraztertu egin behar dira aplikazioaren eta testu-iturriaren arabera, zeren desberdinak baitira pertsona batek teklak sakatzean sortzen dituen erroreak, OCR unitate batek sortzen dituenak edo mikrofono eta hizketa-testu sistema batean sortzen direnak. Kasuaren arabera, aurretik ikusitako zenbait errore maiztasun handiagoz edo gutxiagoz gertatuko dira, edo kasuistika bereziak sortuko dira. Horren aurrean teknika berriak edo teknika orokorren egokitzapenak izango dira gomendagarri. Ongien aztertutako aplikazioak OCR motakoak eta testu-edizioa direnez bi horietan zentratuko gara.

OCR bidezko irakurketetan honako ezaugarri hauek detektatu dira:

- Gertatzen diren erroreen ondorioz n-grama arraroak sortzen dira askotan, beste aplikazioetan baino gehiagotan. Horregatik erabiltzen da, besteak beste, n-grametan oinarritutako metodoa erroreen detekziorako.
- Erroreen iturburu nagusia karaktereen arteko antzekotasuna izaten denez, errore gehienak aldaketa baten bidez gertatzen direla suposa daiteke, aldaketen probabilitatea antzekotasunaren arabera defini daitekeela. Sistema batzuetan antzekotasun hori definitzeko irakurritako testuaren letra-mota hartzen da kontuan.
- Aurreko puntuan esandakoaren arabera, erroredun hitzetan jatorrizko luzera mantentzen dela esan badaiteke ere, aldaketa batzuek ondorioak dituzte luzeraren gainean. Horrela *ri =n* edo *m=iii* aldaketak sarri gertatzen dira. Kasu berezi horiek behintzat aztertu ohi dira, beti luzera mantentzen delako erregela gaindituz.
- Hitz-mugaren gaineko erroreei dagokionez, berriz, aipatutako bi motetakoen arteko banaketa alderantzizkoa da testu-edizioarekin konparatuz, hau da, hitz-zatiketa gehiago gertatzen da biren bilketa baino.

Testu-edizioari dagozkion ezaugarriak hauek dira:

- Teklatuan karaktereek duten posizioaren arabera karaktereen arteko distantzia fisikoa eta antzekotasun-neurria lot daitezke. Irizpide hau, interesgarria badirudi ere, sistema gutxitan erabiltzen da.
- Yannakoudakis-en (1983) iritziz errore gehiago gertatzen dira hitzaren azken karaktereetan hasierakoetan baino. Lehen karakterean errore gutxi egon ohi delakoan, zerrendetan oinarritutako zuzentzaile ortografiko askok hiztegia lehen

karakterearen arabera antolatzen dute; honek, kasu batzuetan, zuzenketa zilegia ez aurkitzera eramaten ditu.

V.3.1.3 Gaitasun-erroreak.

Idazlearen arabera izen desberdinak esleitzen zaizkie hizkuntzaren ezaugarriekin lotutako erroreei: *orthographical errors* (van Berkel & de Smedt, 88); *competence errors* (Veronis, 88), *cognitive errors* eta *phonetic errors* (Kukich, 92). Gaitasun-erroreak eta, morfologian (ikus IV kapitulua) ikusitakoarekin bat etorriz, aldaerak deituko ditugu.

Hizkuntzaren ezaugarriekin lotutako errore hauek bereziki tratatzea ez da ohizkoa, baina, egiten denean, emaitza onak lortzen dira. Errore fonetikoak izaten dira tratamendu berezia merezi ohi dutenak. Mitton-en arabera aztertutako corpusean aurkitzen diren erroreen artean %44ak homofonoekin du zerikusirik. Hala ere, bibliografian agertzen diren aplikazioak, leku-izenekin eta pertsona-izenekin lotuak dira. Mota honetako bi adibide ditugu: orri horiak Minitelaren bidez frantsesez kontsultatzeko Veronis-ek (1988) garatutako zuzentzailea, batetik, eta bestetik Van Berkel eta De Smedt-ek holanderazko pertsona-izenen gainean egiten dituzten neurriak, beren Triphone sistemarako. Argi dirudi halako eremuetan handiago dela fonologiaren eragina erroreetan.

Sistema batzuetan bestelako erroreekin batera tratatzen dira errore hauek, maiztasun handia duten erroreak eta dagozkien zuzenketak buffer batean gordez.

Normalean silaba/fonemen bidez eta ezagumendu linguistikoa erabiliz tratatzen dira, eta errore tipografikoen trataerarekin konbinatzen diren azpisistemak izan ohi dira. Honetaz V.4 atalean sakonduko badugu ere, zenbait sistemaren oinarriak zerrendatuko ditugu hemen:

- Aipatutako Triphone-n homofonoak bilatzeko fonemen araberako lexiko bat erabiltzen da.
- TWBn (Kese *et al*., 92) alemanerarako erabilitako zuzenketan lexiko berezi bat eratzen da errorea, testuingurua, dagokion zuzenketa eta arrazoiaren azalpena edukitzeko.
- Fonemen arteko baliokidetasun-taulen eraketa erabili da frantsesezko interfaze-sisteman (Veronis, 88), eta greziera modernorako zuzentzaile batean (Vagelatos *et al*., 95).

Gure proiektuan, eta hurrengo kapituluan sakonduko dena laburbilduz, laugarren kapituluan aipatu den aldaeren tratamendurako informazioa erabiltzen da hizkuntzaren ezaugarriekin lotutako erroreak zuzentzeko. Alegia, azpilexiko-multzo bat morfemetan eta

morfotaktikan gertatu ohi diren akats edo gaitasun-erroreetarako definitzen dira, eta bi mailatako morfologiaren araberako erregela-multzo bat aldaketa morfofonologiko erregularrak adierazteko (Aduriz *et al*., 93). Honekin bibliografian agertzen den gaitasun-erroreen tratamendurik osoena egiten da.

V.3.1.4 Tratamenduaren garrantzia errore-mota eta aplikazioaren arabera.

Aplikazioaren arabera errore-mota batzuk ager daitezke eta beste batzuk ez. Erroreen tratamendua erabakitzeko orduan, irizpide nagusia maiztasuna izan ohi da, baina honek ez du zertan beti horrela izan behar.

Elkarrekintzazko zuzenketan askoz inportanteagoa da gaitasun-erroreen tratamendua errore tipografikoena baino, hauek maiztasun handiagokoak izan badaitezke ere. Azken hauetan erabiltzaileak hitz egokiaren idazkera gehienetan dakien bitartean, aldaeretan askoz arazo gehiago dauka berak bakarrik zuzentzeko. Aldaeren tratamenduak are garrantzi handiagoa hartzen du OLIren arloko aplikazioetan, edo hizkuntzaren ezagumendua baldintzaturik dagoenean —euskararen kasuan aipaturiko batasun-prozesuarengatik dago baldintzatua, adibidez—.

Baieztapen honekin bat datoz ikerlari bat baino gehiago, ondoren ikus daitekeenez:

> "... In man-machine communication, the correction of competence errors is far more important than the correction of performance ones. ..." (Veronis, 88:708).

> "... Most of the correction methods currently in use in spelling checkers are biased toward the correction of typographical errors. We argue that this is not the right thing to do. Even if orthographical errors are not as frequent as typographical errors, they are not to be neglected for a number of good reasons. First, orthographical errors are *cognitive* errors, so they are more persistent than typographical errors: proof-reading by the author himself will often fail to lead to correction. Second, orthographical errors leave a worse impression on the reader than typographical errors. Third, the use of orthographical correction for standardization purposes (e.g. consistent use of either British or American spelling) is an important application appreciated by editors. ..." (van Berkel & de Smedt, 88:77).

Zaila egiten zaigu beste zerbait eranstea; bakarrik gehitzea ideia horiek buruan geneuzkala proiektuari ekin genionean. Estandarizazioari egiten zaion aipamenarekin lotuz, euskararen estandarizazio-prozesuan lagungarri den zuzentzaile bat diseinatzea izan da gure lanaren helburu nagusietako bat.

V.3.2 Antzekotasun-neurriak.

Aurreko ezaugarriak oinarritzat hartuz, zuzentzeko metodo/algoritmoak diseinatzen dira. Metodo hauetan, askotan, zuzenketa automatikoa edo errore tipografikoak zuzentzeko

egindakoetan batez ere, agertzen den arazo nagusia zera da: zenbait hitzen artean nola aukeratu erroredun hitzarekiko “antz” handiena duena. Honetarako neurri desberdinak erabiltzen dira ondoan ikusiko ditugunak dira erabilienak.

Gai hau oso zabala eta korapilatsua da, beraz, sarrera bat besterik ez da egingo. Honetaz sakontzeko Kukich-en artikuluko bigarren kapitulua gomendagarria da oso.

Edizio-distantzia

Damerau-ren tipifikazioaren arabera, bi formaren artean dauden bihurketa bakunen kopuru minimoa da distantzia hau. Horrela diren *baina* eta *bania* testu-hitzen artean dagoen edizio-distantzia batekoa da, bi karaktere jarrairen truke bakar batez (*in* - *ni*) batetik bestera igaro baitaiteke.

Neurri horren arabera, bateko distantzia minimoan hitz asko egon daitezkeenez, haien artean aukeratzeko bestelako irizpideak ere erabil daitezke: hurbilpen fonologikoa edo teklatuaren araberakoa testu-ediziorako, hurbilpen grafikoa OCR aplikazioetarako, maiztasun handieneko hitzak, etab.

Neurri horrek bi eragozpen aurkezten du: (1) edozein bi karaktere-kateren arteko edizio-distantzia kalkulatzea ez da berehalakoa; eta inportanteena (2), hitz bat ondo zuzentzeko hitz posible guztiekin alderatu behar da hitz akasduna, dagokion zuzenketa zehatzena lortzeko, eta hau oso garestia da konplexutasunaren aldetik.

Bigarren puntua izan da oso ikerlerro garrantzitsua, batez ere OCR aplikazioetan. Aldaera kopuru izugarria laburtzeko, besteak beste —programazio dinamikoa, luzeraren araberako bilaketa, etab.—, distantzia-neurri berriak proposatu dira. Horien artean inportanteenak diren honako bi hauek aipa daitezke: kodeen arteko distantzia eta n-gramen arteko distantzia.

Kodeen arteko distantzia

Hashing tekniketan oinarriturik hiztegiko forma guztiei kode bat esleitzen zaie; ondorioz, hiztegia kodeen arabera antolatzen da eta distantzia hitzen artean kalkulatu beharrean kodeen artean kalkulatzen da.

Normalean kontsonanteei, batez ere hasierakoei, balio handiagoa ematen zaie eta karaktere errepikatuei ez zaie jaramon handirik egiten. Pollock eta Zamora-k (1988) teknika hau erabiltzen dute SPEEDCOP sisteman, baina kode bakarraren ordez bikoitza, *skeleton key* eta *omission key*, erabiltzen dute, bilaketa zehatzago eta azkarrago burutzearren.

Multzo honetako emaitzak oso onak izan daitezke, baina horretarako kodeketa eta informazioaren antolaketa konplexu samarrak behar dira.

n-gramen arteko distantzia

Kodeak erabili beharrean n-gramak (trigramak normalean) erabili ohi dira karaktere-kateen arteko distantziak kalkulatzeko eta konplexutasuna txikitzeko. Hitzen arteko distantzia haien arteko amankomuneko trigramen araberakoa izango da. Erabiltzen diren trigrama-egiturak (edo orokorrean n-gramenak) bitarrak izaten dira —trigrama onartzen den ala ez esanez— eta trigramaren posizioa kontuan har daiteke edo ez. Hiztegia dagoenean, hitzen trigramen arabera indexatu ohi da bilaketa errazteko.

Honen adibidea ACUTE sistema (Angell *et al.*, 83) dugu. Trigrametan eta hitzen luzeran oinarritutako sistema honetan distantzia neurtzeko formula honako hau da:

$$d = c / \max(n, n')$$

non *c* amankomuneko trigrama kopurua den eta *n* eta *n'* hitzen luzerak.

Oso emaitza onak azaltzen dituzte, karaktere jarraien arteko trukearen kasuaren salbuespenaz.

Saio hauez gain, hitzak trigrama-bektoreen bidez adierazten dituzten teknika sofistikatuak ere proposatzen dira; horien gainean Hamming-en distantziak bezalako neurriak aplikatzen direlarik.

V.3.3 Zuzenketa-metodoak

Berriro azpimarratu behar da hitz isolatuak zuzentzeko teknikak beti oso mugatuak direla, ondo zuzentzeko testuingurua kontuan hartzea ezinbestekoa da eta. Honen adierazgarri pertsonekin egindako testak ditugu: zenbait laguni emandako testuingururik gabeko akatsen aurrean eskatzen zitzaien hiru edo lau hitzen artean aukera zezaten, batez-besteko asmatze-tasa %75ean ezarriz (Kukich, 92:411). Sistema automatiko onenetan antzeko zenbakiak lortzen dira.

Aurretik esandakotik ondoriozta daitekeenez hitz baten zuzenketa dirudiena baino eginkizun konplexuagoa da. Horren lekuko da PF-474 txipa (Yianilos, 83), helburu berezitu honetarako diseinatu dena.

Ondoren zuzenketa-metodo adierazgarrienak azalduko dira; oinarrizkoak aurretik eta konbinatuak ondoren. Azpimarratzekoa da zenbait sistemaren inguruan dagoen informazio-falta, gaiak duen interes komertziala dela eta.

V.3.3.1 Oinarrizko metodoak

Zuzenketara aplikatzen diren funtsezko metodoak azaltzen dira honako lau multzo hauetan bereiziak: (1) alderantzizko edizio-distantziaren bidezkoa; (2) hitz guztiekiko distantziaren bidezkoa; (3) erregelen bidezkoa eta (4) metodo estatistikoak.

Alderantzizko edizio-distantziaren bidezko metodoak.

Metodoaren funtsa hauxe da:

- Akasdun formatik abiatuta Damerau-ren legeak aplikatzen dira alderantziz. Horrela, eta bateko distantziara mugatuz, V.3.1.1 atalean aipatutako zenbakiak erabiliz *2nk* hipotesi sortzen dira, *k* alfabetoaren karaktere-kopurua eta *n* hitzaren luzera izanik.
- Hipotesiak egiaztatzen dira hizkuntzaren hitzak diren ala ez jakiteko, horretarako V.2 atalean aipatzen diren metodoak erabil daitezkeela.
- Ontzat hartutako hipotesiak sailkatzen dira; lehena aukeratzeko zuzenketa automatikoan edo lehenengo batzuk elkarrekintzazko zuzenketan.

Sistema askotan erabiltzen da metodo hau: (Peterson, 80), (Kernighan *et al*., 90), (Church & Gale, 91) (Vagelatos *et al*., 95). Gure proiektuan erabilitako metodoaren barruan ere, teknika hau errore tipografikoei aurre egiteko erabiltzen da.

Metodo honen ezaugarriak hauek dira: hiztegi osoa ez duten sistemetan aplikagarria, programatzeko eta memoria-hartzearen aldetik sinplea, baina bakun ez diren erroreentzat ez du proposamen egokirik eskaintzen. Azken eragozpen honen aurrean, metodo bera biko distantziarekin aplika daiteke, baina horren ondorioz, aukera-kopurua izugarri haziko litzateke eraginkortasunaren kalterako.

Hitz guztiekiko distantziaren bidezko metodoak.

Helburua zera da: errorea eman duen testu-hitza hizkuntzaren hitz posible guztiekin —edo askorekin, optimizazio-teknikak erabiltzen badira— alderatzen da, berarekiko distantzia txikiena duena aukeratuz zuzenketa automatikoan eta hurbilen dauden lehenak elkarrekintzazko zuzenketan.

Hau da metodo erabiliena eta gehien ikertu dena, testu-edizioan batez ere. Sistemen artean hauek daude: (De Heer, 82), (Angell, 83), (Pollock & Zamora, 84), (Hull & Srihari, 82), (Tanaka, 87). Aipatutako PF-474 txipa eragiketa hau arintzeko diseinatua da.

Metodo honen ezaugarriak hauek dira: emaitza onak lortzen dira, eta gainera, beti aurkitzen da zuzenketaren bat; baina horretarako hiztegi osoa biltegiraturik eduki behar

da. Aurreko atalean aipatu den bezala, distantzia- edo hurbiltasun-neurri desberdinak erabil daitezke, teknika horien artean zehaztasun, memoria-hartze eta eraginkortasun neurri desberdinak lortuz; hiru irizpideak batera optimizatzeko metodorik ez dago ordea.

Erregelen bidezko metodoak.

Ezagumendu linguistikoa erabiliz erregelak sortzen dira akasdun formatik dagokion forma zilegia lortzeko. Erregela hauek gehienetan morfofonologikoak izaten dira, baina aztertutako corpusetan gertatzen diren errore tipografikoetatik inferitutako erregelen bidezko sistemak ere sartzen dira multzo honetan.

Sailkapena zehaztearren bi multzotan bana daitezke metodo hauek:

- Zuzenean hitz zilegiak lortzen dituzten metodoak; hauetan, erregelak zeharo linguistikoak direnez, lortzen diren zuzenketak hizkuntzaren formak izanik. Hauek dira benetako “metodo linguistikoak”. Multzo honetan kokatzen da gure proiektuaren barruan Xuxen zuzentzaile ortografikorako egindako aldaeren tratamendua (ikus §VI.4 atala).

- Proposamen hipotetikoak lortzen dituztenak, ondoren hipotesi hauek egiaztatu behar direla. Multzo hau lehen multzoaren aldaketa bezala ere ikus daiteke, Damerau-ren erregelak aplikatu beharrean beste batzuk aplikatzen direlarik. Mota honetako metodoa dugu Yannakoudakis eta Fawthrop-ek (1983) proposatutakoa, non erregelak baino heuristikoak erabiltzen diren.

Ezaugarrien aldetik, berriz, metodo hauek oso emaitza onak ematen dituzte erroreak aurrikusitako parametroen barruan gertatzen badira, baina oso txarrak gainontzekoetan; eta horrexegatik osatu ohi dira beste metodo batzuekin sistema konbinatuak eginez.

Metodo estokastikoak.

Aurreko erroreetan oinarriturik, automatikoki inferitutako informazioa erabiliz zuzentzen dira akats berriak. Normalean, ikasketa-prozesu bat behar dute aurretik; prozesu horretan, eskuz prestatutako edo zuzendutako corpus batez, akatsak eta hitz zilegien artean erlazioak bilatzen dira. Ezagumendua inferitzeko teknika nagusiak hiru dira: taula estatistikoak, eredu markoviarrak, eta sare neuronalen eredua. Teknika hauek etiketatze-lanetan (*tagging*) erabiltzen diren berberak dira, eta lexikorik erabili gabe lan egiten duten zuzentzaileetan erabiltzen dira batez ere, egiaztapena hitz-zatietan oinarrituz.

OCR aplikazioak izan ohi dira teknika hauen helburua, OCR dispositiboek akatsak modu erregularrean egiten baitituzte, pertsona desberdinen akatsak askoz irregularragoak izanik —pertsona bakoitzeko ikasketa-prozesu berezia beharko litzateke—. Gainera, lexikorik edo hiztegirik gabe lan egiten den aplikazioetan —OCR eta hizketaren

tratamendua normalean— eta lehen bi teknika-multzoak ezin direla erabili kontuan hartuz, lortzen diren emaitzak aipagarriak dira.

Adibide gisa *correct* programa (Kernighan *et al*., 90) dugu. Corpusen gainean lortutako probabilitateetan oinarritzen dira bere oinarrizko sistema osatzeko —oinarria alderantzizko edizio-distantziaren bidez eraikitzen da— eta oso emaitza onak azaltzen dituzte. Damerau-ren lau errore motetako bakoitzerako "nahasketa-matrize" bat osatu dute datuen arabera, eta hitz akasduna ordezkatzeko gai direnen artean sailkapen bat egiten dute, hitzaren probabilitate absolutua eta akatsekiko desberdintasunaren probabilitatea biderkatuz. Metodo estokastiko honekin oso emaitza onak lortzen dira errore bakunetarako.

V.3.3.2 Metodo konbinatuak.

Testu-edizioan aplikatutako zuzenketarako metodo konbinatuak ari dira proposatzen azken urteotan, eta hauetan sakonduko dugu.

Berriro De Smedt eta Van Berkel-en hitzak aldatuko ditugu hona:

> "Of the method described in the previous chapter, no single method sufficiently covers the whole spectrum of errors. Because each method has its strengths and weaknesses, it is advantageous to combine two methods which supplement each other." (van Berkel & de Smedt, 88:80)

Lehentxeago aipatutako *correct* da hauetako bat, alderantzizko edizio-distantzia eta metodo estokastikoak konbinatzen dituena. Ematen duten asmatze-tasa %87a da, baina neurria ez da estandarra. Normalean lehen edo lehen hiru proposamenekin zenbatetan asmatzen den izaten bada neurria, beraiek testuingurua kontuan hartu gabe hiru pertsonek emandako epaien artean gutxienez birekin bat etortzea hartzen dute neurri-unitatetzat.

Ondoren beste bi metodo konbinatu azaltzen dira gainbegirada osoa lortzearren.

Triphone (van Berkel & de Smedt, 88)**.**

Entziklopedia bat kontsultatzeko diseinaturiko sistema honek bi metodo konbinatzen ditu: errore fonetikoak tratatzeko fonemen gaineko erregelak erabiltzen zituen Spell Therapist batetik, eta forma guztiekiko distantzian oinarritutako trigramen bidezko FUZZIE[1] (De Heer, 82) metodoa.

---

[1] Metodo hau egokiagoa da aipatutako ACUTE (Angell, 83) baino, azken honetan luzerak funtsezko papera duelako eta fonema batek karaktere kopuru aldakorra duelako.

Proposatzen duen irtenbidea hau da: distantzia kalkulatzeko trifonemen arteko distantzia erabiltzea trigramena erabili beharrean, ondorioz hiztegia trifonemen arabera antolatuz. Jarraitzen den algoritmoa honako hau da:

- bere fonemen arabera hitza trifonemetan banatzen da (banaketa-aukerak optimizatuz)
- trifonema bakoitzari dagokion maiztasuna lortzen da
- zenbait trifonema aukeratzen dira, maiztasun-muga batetik behera dauden hautapen-trifonemak deitutakoak, eta horien arabera antolaturiko fitxategian bilatzen da.
- bide honetatik aurkitutako hautagai guztiekin amankomuneko trifonemen arabera antz handiena dutenak aukeratzen dira.

Azaltzen dituzten emaitzak oso onak dira, %92 lehen proposamenean, baina eremua oso mugatua da pertsona-izenekin bakarrik probatzen delako.

Forma guztiekiko distantzia + morfofonologia (Veronis, 88)**.**

Minitel kontsulta-sistemarako garatutako sistema honetan, hitz isolatuak zuzentzeko metodoa —komunztadura-akatsak zuzentzeko beste metodo bat ere badago— ezaguna den beste batean (Durham *et al*., 83) oinarritzen da eta hiru multzotan banatzen da:

- Forma guztiekiko distantzia da jarraitzen den oinarrizko metodoa. Errore tipografiko bakar bat hartzen du kontuan.
- Fonologiaren menpe dagoen karaktere-multzoen artean antzekotasun-taula bitarra (antzekoak ala ez esaten baitu) erabiltzen du, antzeko multzoak berdintzat joz distantziak neurtzerakoan.
- Aurreko guztia erroekin egiten da, ondoren atzizkien tratamendu morfologiko ad-hoc sinple bat eginez.

V.4.Hizkuntza flexionatuen eta eranskarien zuzenketa.

Flexio handiko hizkuntzetan zein hizkuntza eranskarietan erroreen zuzenketa korapilatsuagoa da beste hizkuntzetan baino. Hala ere, eta ingelesaren flexio-sistema sinplea dela eta, hizkuntza hauen gaineko zuzenketa ez da sakonean aztertu.

Lexikoa erabiltzen duten aplikazioetara murriztuz —lexikorik gabekoetan, hitz-zatietan oinarritutako egiaztapenean, eta erregelen zein metodo estokastikoen bidezko zuzenketan ez baitago alde nabarmenik hizkuntza-motaren arabera—, honela labur daiteke hizkuntza hauetarako zuzenketa-mekanismoa:

- **Egiaztapena** analisi morfologikoaren bidez burutzen da, hitz-zerrenda oso bat desegokia eta ezinezkoa edo memoria-hartzearen aldetik garestiegia baita. Mota horretako hizkuntzetarako lortutako erreferentzia guztietan horrela egiten da. Izan ere, lexikoa aberasteko orduan, geroxeago aztertuko denez, informazio linguistikoa jaso behar da erabiltzailearengandik.

- **Zuzenketa** egiteko arazoak handiak dira, eta proposamen desberdin egin dira. Forma posible guztiekiko distantzia kalkulatzea ezinezkoa da —forma guztiak ez baitira inon gordetzen—, eta, ondorioz, morfologia eta zuzenketa-metodoak konbinatu behar dira.

V.4.1 Lexikoaren aberasketa.

Esan bezala, euskara bezalako flexio handiko hizkuntzetan errorea dagoenentz jakiteko testu-hitzen analisi morfologikoa egin ohi da. Horretarako sisteman, lexikoan normalean, morfologiari buruzko informazioa metatzen da. Lexikoa itxia denean ez dago arazorik, baina lexikoaren aberasketa erabiltzaile arruntaren esku uzten denean ondoko arazoak sortzen dira:

- sarrera berriei buruzko informazio morfologikoa beharrezkoa da berauen flexioa ondo era dadin.

- informazio linguistiko hori modu automatikoan sortzea ezinezkoa izaten denez, elkarrizketa-protokolo bat diseinatu eta erabili behar da erabiltzaileari informazio hori eskatzeko.

- informazio linguistikoa linguistikan aditua ez den erabiltzaile bati eskatzen zaionez gero, elkarrizketa sinpleaz baina ahalik eta informaziorik zehatzena lortu behar da, eta hau, askotan, ez da erraza.

Gai honetaz bibliografian dagoen informazioa hutsetik hurrena da. Gure zuzentzailearen barruan elkarrizketa-modulu hau diseinatu dugu helburu bikoitz horrekin: sinplea izatea baina zehaztasunik galdu gabe (ikus §VI.6 irudia).

V.4.2 Zuzenketa. Zenbait sistema.

Flexio aberatseko hizkuntzetan, V.3 atalean ikusitakoaren arabera, eta morfologian oinarrituriko sistema baterako aritzen garela suposatuz, akats baten gaineko zuzenketarako ondoko algoritmoa har daiteke oinarritzat:

- *alderantzizko edizio-distantzia*ren metodoa erabiliz, hitz-forma honetarako proposamen hipotetiko guztiak sortu.

- Hitz-forma hipotetiko guzti hauek benetako hitzak diren ala ez jakiteko berauen analisi morfologikoa burutu, arrakastatsuak aukeratuz.

- Benetako hitzen artean sailkapena egin.

*Forma hipotetikoen analisi*an oinarritutako algoritmo horren eragozpen nagusiak bi dira:

1) Oso motela gerta daiteke, batez ere analisi morfologikoa konputazio-konplexutasun handi samarrekoa denean. Azkartzeko metodoak —maiztasun handieneko hitzak, trigrama okerren bidezko bazterketa, etab.—bidera badaitezke ere, sakoneko arazoa izan daiteke.

2) Akatsa bakuna ez denean ez da proposamenik (edo proposamen egokirik) sortuko, alderantzizko edizio-distantziaren metodoa bateko edizio-distantziarekin lan egiten du eta. Biko distantziara heda liteke metodoa baina horrekin izugarri areagotuko litzateke lehen eragozpena.

Oinarrizko algoritmo hori izan da, funtsean, guk errore tipografikoen tratamendurako erabili duguna. Ikus dezagun bibliografian agertzen diren zuzenketa-tratamendu garrantzitsuenak morfologian oinarrituriko sistemetarako:

- **Tratamendurik ez**. Zenbait sistematan ez da zuzentzeko laguntzarik ematen: (Hajic & Droza, 90), (Solack & Oflazer, 93).

- **Forma hipotetikoen analisia**: Aurretik ikusitako tratamendua, alderantzizko edizio-distantziaren metodoan oinarritzen dena, edo horren deribazioren bat da hedatuena, errore tipografikoetarako behinik behin: (Means, 88), (Vagelatos *et al*. 95).

- **Erro-hizkiak**. Hitza ezagutzen ez den analisian erro eta hizkien banaketa burutzen da, eta bakoitzaren zuzenketari ekiten zaio V.3 atalean aipatutako metodoren batez. Bukaeran sorkuntza morfologikoaren bidez erroa eta hizki zilegien bidez lortzen dira zuzenketarako hitzak. Metodo honen eragozpena lehen banaketan datza, baina trukean hipotesien analisia saihesten da. Horren adibidetzat har daiteke gure zuzentzailean egiten den aldaeren zuzenketa. Veronis-ek (1988) frantseserako sistema batean ideia honi jarraitzen dio. Autore horrek burutzen duen tratamendu morfologikoa, hala ere, ez da osoa, atzizkien tratamendu partikular bat besterik ez baitu egiten; beraz, ez da egokia izango hizkuntza eranskarietarako.

Oflazer eta Guzey-k ondoren azalduko dugun tarteko tratamendua proposatzen dute, erro guztiekiko distantzian eta sorkuntza morfologikoan oinarritzen dena.

Hizkuntza eranskarietarako proposatutako metodoa (Oflazer & Guzey, 94)**.**

Turkiera hizkuntza eranskaria denez erroreen zuzenketa korapilatsua da. Aipatu den bezala, lehen zuzentzaile batean ez da zuzenketarako mekanismorik eskaintzen (Solack & Oflazer, 93). Azken urteetan, eta bi mailatako ereduan oinarritutako prozesadore morfologiko bat burutu ondoren (Oflazer, 93), zuzenketaren arazoari ekin diote.

Algoritmoaren funtsa bi urratsetan banatzen da:

1) Akasdun hitzaren erro posible guztiak lortzea erro-hiztegitik.

2) Lortutako erroetatik abiaturik akasdun formarekin antza duten formak sortzea.

Lehen urratsa burutzeko erroen azalak eta akasdun hitzaren hasierako azpikateak[1] alderatzen dira edizio-distantziaren metodoa erabiliz, distantzia honi muga bat jarriz. Bilaketa hau laburtzearren, bigramen araberako bektoreen indizeak eratzen dira lexikoa atzitzeko —bakarrik akasdun formarekin amankomuneko *k* bigramak dituzten erroak hartzen dira kontuan.

Bigarren urratserako bi hipotesi egiten du: balizko erroari dagokion eskuineko azpikatea —atzizki-multzoa osatuko lukeena— zuzena dela suposatuz batetik eta ez dela suposatuz bestetik.

Lehen kasuari "*on the left edge of the word*" izena eman diote eta tratamendu erraza du: erroari dagokion bukaerako azpikatea erantsi eta analizatu, analizatuz gero proposamen bat lortzen delarik. Erroa eta hasierako azpikatearen arteko distantzia mugan bada, hori da erro horretarako saio bakarra.

Bukaerako azpikatean akatsik egon daitekeela suposatzen bada, erroaren araberako sorkuntzari ekiten zaio, eta hasierako formarekiko distantzia osoa —erroari eta atzizkiei dagozkienak batuz— mugatik behera duten kasuak aukeratzen dira. Aztertzeko kasuak laburtzearren, "*Cut-Off Paths*" izeneko teknika (Du & Chang, 92) erabiltzen dute.

Erroaren eta atzizkiaren artean gerta daitezkeen karaktere-aldaketak eta galerak kontuan hartzeko, distantzia-mugaren gainean "ukituak" egiten ditu.

Azkenik, distantzia bera dutenen proposamenen artean sailkapena egiteko aztertutako datuen gainean egindako estatistikak darabiltzate.

Azaltzen dituzten emaitzak onak dira zehaztasunaren aldetik, baina ez hainbeste eraginkortasunaren aldetik. Adibidez, bateko distantzia-muga ezarriz —errore bakunak

[1] Aurrizkiak ez ditu kontuan hartzen.

bakarrik zuzenduko dira—, zehaztasun handia lortzeko aipaturiko *k* faktoreak hiru izan behar du eta, ondorioz, batez-besteko analisiak 31 dira, sorkuntzak 311 eta distantzia-eragiketak 2500. Biko distantziarekin, zehaztasuna mantentzeko, neurri horiek bost aldiz handiagoak dira.

Neurri hauekin eta datorren kapituluan ikusiko dugun gure sistemarekin konparatuz, zera ondoriozta daiteke: metodoaren sofistikazioak ez du ebazten *forma hipotetikoen analisi*an oinarritutako metodo klasikoaren eragozpen nagusia, eraginkortasunarena hain zuzen.

V.4.3 Ondorioak.

Zuzenketa-metodoen gainean egindako azterketa honetatik ondorio hauek lortzen dira:

- Bi metodo multzo bereizten dira: informazio lexikoa erabiltzen dutenak eta hitz-zatietan oinarritzen direnak. Lehenak zehatzagoak dira, baina ez dute malgutasun handirik, lema edo hitza lexikoan ez dagoenean hitz zilegi bat akastzat hartzen baita. Bigarrenak OCR motako aplikazioetan erabiltzen dira nagusiki, zuzenketa automatiko, azkarra eta malgua behar izaten delako aplikazio hauetan.

- Lexikoa erabiltzen duten metodoek eraginkortasunaren eta memoria-hartzearen artean oreka lortu behar dute, eta horretan hizkuntzaren ezaugarri morfologikoek zerikusia handia dute. Forma-zerrendak oso azkarrak dira baina memoria hartze handia dagokie, eta alderantziz gertatzen da morfemetan oinarritutako metodoetan. Mailatan eratutako sistemak interesgarriak dira.

- Testuingurua kontuan hartzen ez bada zuzenketa, automatikoa batez ere, ez da zehaztasun handikoa izango.

- Elkarrekintzazko zuzenketan errore tipografikoen zuzenketak interes txikiagoa du fonetikak edo ez-jakiteak eragindakoenak baino; haiek nola zuzendu jakin ohi den bitartean besteak ez.

- Hizkuntza eranskarietan hitzen egiaztapena morfologian oinarritu ohi da, eta zuzenketa konplexu samarra gertatzen da.

VI. Xuxen: bi mailatako morfologian oinarritutako zuzentzaile ortografikoa.

Euskararako zuzentzaile ortografiko bat burutzea zen gure helburua hasiera-hasieratik; euskara bizi den batasun-prozesurako oso garrantzi handikoa baita halako tresna bat.

Hasieratik argi zegoen zuzentzailea analisi morfologikoan oinarritu behar zela; ondorioz, eraikitzen ari ginen analizatzaile morfologikoa berrerabiltzea zen irtenbiderik, logikoena ez ezik, merkeena eta interesgarriena.

Euskararen batasuna erabat finkatu gabe egoteak problematika berria eta aberatsa dakar, ikergaia interesgarriago bihurtuz. Gainera, bibliografian agertzen ziren erreferentzia gehienak ez ziren oso baliagarriak izaten, euskara bezalako hizkuntza eranskarietan zuzenketa-prozesua korapilatsuagoa da eta.

Arazo hauen aurrean, eta egiaztatzea analisi morfologikoan oinarrituz burutzeaz gain, problema ebazteko sinpleago diren azpiproblematan banatzea izan zen erabili den taktika: errore tipografikoak alde batetik eta ez-jakiteak eragindako erroreak edo gaitasun-erroreak —analisi morfologikoaren aldaeren tratamendu bera erabiliko denez aldaerak ere deituko ditugunak— bestetik. Tratamenduaren diseinua egiterakoan eraginkortasuna eta zehaztasunaren arteko oreka abiapuntua izan da, erabilpen komertziala bilatzen ari ginen eta.

Analisi tipografikoen tratamendurako, aurreko kapituluan aztertutako alderantzizko edizio-distantziaren ohizko metodoa erabili da, flexio konplexuko hizkuntzetarako baliagarria dena. Metodo honi jarraituz, akasdun formaren gainean proposamen hipotetikoak eratu eta egiaztatzen dira, ondoren benetako hitzak direnak sailkatuz. Proposamen hipotetiko guztien analisia ekiditeko, egiaztatzea analisi morfologikoaren bidez egiten baita, hipotesiak murrizteko eta sailkatzeko zenbait metodo erabiltzen dira.

Gaitasun-erroreen zuzenketarako metodo berritzailea erabili dugu. Hitz bat erroretzat hartzen da analisi morfologiko estandarrik ez dagokionean. Horren aurrean, eta laugarren kapituluan azaldutako aldaeren tratamendu morfologikoa berrerabiliz, lexikoa eta erregela morfofonologikoak hedatzen dira aldaeren tratamendurako informazioarekin, eta analisi-saio berri bat burutzen da. Saio berri horretan analisirik lortzen bada, erroredun hitza gaitasun-erroretzat har daiteke eta baita sorkuntza morfologikoari esker dagokion forma estandarra sortu ere.

Bi tratamenduetan proposatzeko benetako hitzak sor daitezke. Proposamen hauek nola ordenatu behar diren erabakitzea ez da erraza, bibliografian dauden proposamenak oso heterogenoak izanik. Elkarrekintzazko testu-edizioan erabiliko den honetan funtsezkoena ez bada ere, estatistikan oinarritutako sailkapen bat erabaki da.

Proposamenen sorkuntza eta sailkapena egiten duten moduluez gain beste bi modulu garrantzitsu azpimarratu behar dira: iragazlea eta erabiltzailearen hiztegia eguneratzekoa. Biak morfologian erabilitako token-ezagutzailea eta erabiltzailearen lexikoak egokituz burutu dira.

Azkenik modulu guztiak integratzeko eta ingurune atsegina lortzeko elkarrekintza ere diseinatu da.

Inplementaturiko zuzenketa-sistemak zehaztasuna/eraginkortasuna oreka mantentzen duelakoan gaude. Izan ere, aukera berriak ari gara aztertzen bi bide nagusitatik: batetik, lexikorik gabeko analisia berrerabiliz erro-atzizki banaketan oinarritutako metodo bat diseinatzea errore tipografikoetarako; eta bestetik, eraginkortasuna hobetzearren, lexiko-itzultzaileen erabilera testu-zuzenketan.

VI.1. Sarrera.

Euskararako zuzentzaile ortografiko baten diseinuaren aurrean, aurreko kapituluan zehaztutako kontzeptuak gogoan hartuz, hauek izan ziren programaren funtzionalitateak mugatzen eta zehazten dituzten oinarrizko irizpideak:

- Eskala errealeko zuzentzaile erabilgarria burutzea, produktu komertzial baten oinarria izanik, euskararako horrelakorik ez baitzegoen. Gainera, indarrean den batasun-prozesuan horrelako tresna bat oso lagungarria da.
- Lehen belaunaldiko zuzentzaile ortografikoa zen helburua, hau da testuingurua kontuan ez duen zuzentzaile arrunta, testu-edizioan erabili ohi direnak bezalakoa. Irekitako ikerlerro bezala gelditu dira testuingurua kontuan hartuz zuzentzaile hau hobetzea eta estilo-zuzentzailea edo zuzentzaile gramatikala egitea.
- Morfologian hitz anitzeko terminoen tratamendua baztertzeko erabilitako arrazoiek hitz-mugaren gaineko erroreak lan honen eremutik kanpo uzteko ere balio dute, tratamendu partzial bat egin bada ere aldaeren kasuan.

- Egiaztatzea analisi morfologikoan oinarrituz egitea, beste metodoak, n-grametan oinarritutako metodo estatistikoak zein forma-zerrendan oinarritutakoak, baztertuz.

- Euskararen batasunerako arauak ondo ez ezagutzeak zein erabilpen dialektalak eragindako erroreak zuzentzea lehenestea gainontzeko erroreen aurrean, haiek direlakoan erabiltzaileei zuzentzeko buruhauste handiena ematen dietenak eta batasun-prozesuan lagungarrien gertatzen direnak[1].

- Euskararen egoera kontuan hartuz erabiltzailearen hiztegiak sortu eta eguneratzeko aukera ezinbestekoa da, lexiko orokorra zeharo finkatu gabe baitago, pertsona- zein leku-izenen eta termino teknikoen aldetik batez ere. Gainera, hiztegi hauetan termino bat sartuz gero bere flexio guztia ere ezagutu beharko du zuzentzaileak.

VI.2. Egiaztatzea.

Esan den bezala hitzen ezagutza tratamendu morfologikoaren bidez burutzen da; hau da, hitz bat onartzen da analisi edo deskonposaketa morfologikorik baldin badu, bestela erroretzat hartzen da.

Beste metodoak, n-grametan oinarritutako metodo estatistikoak, zein forma-zerrendan oinarritutakoak, baztertu egin ziren honako hiru arrazoi hauengatik:

1) Berrerabilgarritasuna. Irtenbide *ad-hoc*etik ihes egitea eta berrerabilgarritasuna bultzatzea, bide batez helburu orokorreko prozesadore morfologikoa burutzeko asmoa indartuz.

2) Ortogonaltasuna. Erabiltzaileari oso ulergaitza gertatzen zaio lemaren flexio batzuk onartzea eta beste batzuk ez. Hori gerta daiteke beste metodoekin baina ez ondo eratutako analizatzaile baten bidez.

3) Segurtasuna. Testu-edizioan funtsezkoa da, eta euskararen egoera kontuan hartuz are gehiago, ez ematea ontzat hizkuntzan existitzen ez diren hitzak. n-grametan oinarritutako metodoetan hau gertatu ohi da, eta horixe da metodo hauek baztertzeko arrazoi nagusia aplikazio-mota honetan.

[1] Lan honetan zehar euskara batuarekin edo hobetsitako lexikoarekin bat ez datozen testu-hitzei erroreak edo akatsak deituko diegu, batzuentzat termino hauek gogor samarrak izan arren.

Behin analisi morfologikoaren oinarria aintzat harturik, hirugarren kapituluan azaldu den bi mailatako morfologian oinarritutako analizatzaile morfologiko estandarra egokitu zen, honako ukitu hauek eginez:

- Informazio morfologikoaren bazterketa memoria-hartzea laburtzearren, aplikazio honetan ez baita beharrezkoa. Beraz analisi morfologikoa baino, burutzen dena segmentazio morfologikoa da.

- Hitz zilegien segmentazio morfologiko guztiak ez dira beharrezkoak, analisi zuzen bat duela jakitea nahikoa baita. Hori dela eta, hobekuntza hau burutu da: lehen segmentazioa posiblea lortu bezain laster prozesua eten eta hitza zilegitzat ematen da. Gainera, lehen segmentazioa lehenbailehen lortzeko backtracking-aren bidezko analisi-aukerak sakonean[1] jorratuko dira (ikus §II.5 irudia backtracking-prozedura gogoratzeko). Honen ondorioz hitz zilegien egiaztatze-prozesua azkarragoa izango da erroreena baino, haietan egiaztatzea lehen segmentazioan bukatzen den bitartean, bigarrenetan bide guztiak jorratu behar baitira segmentazio-aukerarik ez dagoela egiaztatu arte.

Dena den, eta aipatutako hobekuntza kontuan hartuz ere, hitz bakoitzaren segmentazio morfologikoak eraginkortasun-galera dakar beste egiaztatze-metodoekin alderatzen badugu. Galera hori ahalik eta gehien murriztearren Peterson-ek (1980) aipatzen zituen buffer-en ildotik egin dugu lan. Buffer horiek lehen kapituluan aipatutako corpusaren bidez osatu dira eta berauetan beti testu-hitzak daude, bilaketa bitarra da. Ondoko hiru mailatan banatzen dira:

- Maiztasun handieneko hitz zilegien bufferra. 8000 bat sarreraz osaturik, horrekin testuetako hitzen %75-80 bat ezagutuz.

- Maiztasun handieneko erroreen bufferra. Hitza buffer honetan aurkituz gero egiaztatzea eten egiten da, zuzenean hitza erroretzat hartuz. Buffer honetan hitz hauei dagozkien proposamenak ere gordetzen dira, zuzenketa azkartzeko asmoz. 800 bat hitzez osaturik dago, eta testuaren arabera estaldura desberdina bada ere, %5-10 tartean dago. Beste hizkuntzetan ez da ohizkoa halako bufferrik erabiltzea, baina, aipatutako batasun-prozesua dela eta, errore “tipikoen” kopuru handiak bultzatzen du buffer honen erabilera.

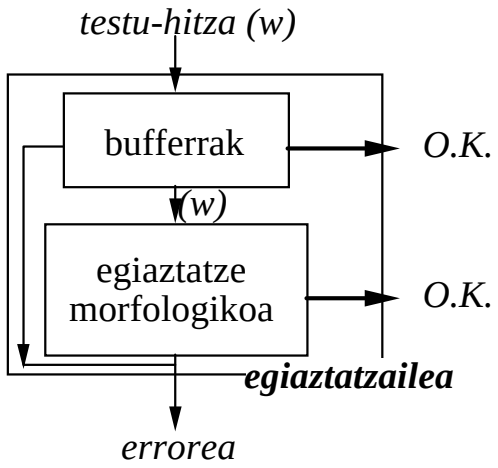
- Testuan aurretik agertutako hitz analizatuak, zilegiak diren ala ez zehaztuz. Hauek dokumentuko bufferra osatzen dute, eta dokumentu-motaren arabera emaitza desberdinak eman ditzake.

---

[1] Analisi morfologikoan sakonean edo zabalean aritzea ez da axola, analisi posible guztiak lortu behar direlako.

VI.1 irudian egiaztatzaile ortografikoaren eskema sinplifikatua ikus daiteke.

Hitz bat ezagutzen ez bada erroretzat hartzen da. Errore baten aurrean zuzentzaile ortografikoak bi bide jorratzen du: errore tipografikoei dagokiena, eta aldaerak edo gaitasun-erroreak deitu ditugun ezjakintasunak bultzatutako erroreei dagokiena. Bi bide hauek paraleloz jorra litezke ordenadorearen ezaugarriek horrela gomendatuko balute; ondoren proposamenak ordenatu egingo baitira.



VI.1 irudia.- Egiaztatzaile ortografikoaren eskema orokorra.

VI.3. Errore tipografikoen tratamendua.

Aurreko kapituluan aztertutako alderantzizko edizio-distantziaren metodoa erabili da zuzenketak edo, aplikazio honetarako egokiago den izenaz, zuzenketa-proposamenak sortzeko.

Metodo honi jarraituz (aipatutako metodoa §V.3.3.1 atalean azaldu zen), akasdun forma batetik abiatuta honako urrats hauek jarraitzen dira:

- Bateko edizio-distantzia duten proposamen hipotetiko guztiak sortu.
- Lortutako proposamen hipotetiko guztiak egiaztatu, VI.2 atalean azaldutako egiaztatzailea erabiliz. Arrakastaz egiaztatzen direnak dira benetako hitzak, gainontzekoak zuzenketa-proposamen gisa baztertuz.

Metodo hau sinplea da, baina bi eragozpen inportante du zehaztasun aldetik, emaitza onak eman baditzake ere:

1) Akatsa eta dagokion zuzenketaren artean edizio-distantzia bat baino gehiago denean, ez da zuzenketa egokirik eskainiko.

2) Forma hipotetiko guztiak egiaztatu behar izatea ez da eraginkorra, morfologian oinarritutako egiaztatze-prozesu bat burutzen denean batez ere.

Lehen eragozpenaren aurrean edizio-distantzia bira zabal liteke baina horrekin zuzenketak lortzeko denbora izugarri haziko litzateke, zuzenketa-prozesua ia ezinezkoa bihurtuz —gogoratu behazehaztasuna/eraginkortasuna oreka bermatu behar dela. Horren arrazoia zeran datza: proposamen hipotetikoen kopuru izugarria eta hauetako bakoitza morfologikoki egiaztatzeko beharra.

Hala ere, eta lehen eragozpen hori erlatibizatuz, bi irizpide hartu behar dira kontuan:

- Zuzentzaile ortografikoa biko edo edizio-distantzia handiagoko erroreak zuzentzeko gai izango dela, erroreek aldaeren ezaugarriak dituztenean.
- Aurretik esan dugun bezala, errore tipografikoen zuzenketa ez da diseinu-helburu nagusia.

Dena dela kapitulu honen bukaeran (ikusi §VI.8) eragozpen hauek berraztertzeari eta aurreko kapituluan hizkuntza eranskarietarako egindako proposamenarekin alderatzeari ekingo diogu.

VI.3.1. Azkartzeko bideak.

Aipatutako bigarren eragozpenaren aurrean, proposamen hipotetiko guztien egiaztatzearen beharraz hain zuzen, zenbait heuristiko erabili dugu proposamenen sorkuntza ahalik eta azkarren buru dadin. Bestela, aurreko kapituluan esan zen bezala, bateko edizio-distantzian egon daitezkeen forma hipotetikoak *2nk* inguru dira *n* hitzaren luzera eta *k* alfabetoaren karaktere-kopurua izanik (ikus §V.3.1.1), eta denak egiaztatu beharko lirateke akats bakoitzeko. Gainera, hauetako proposamen hipotetiko gehienak benetako hitzak ez direnez, haien egiaztatzea motelagoa izango da benetako hitzena baino.

Azkartzeko teknika horiek bi motakoak dira: batetik, zuzenean aplikatzen direnak, zehaztasunean eragin adierazgarririk ez dutelako; eta aukeran jartzen direnak bestetik, zehaztasunean eragina izanik zehaztasuna eta eraginkortasunaren artean hautatzen den orekaren arabera.

Lehen multzoan daude proposamenen bufferra eta trigramen bidezko proposamenen sorrera. Bigarrenean aldiz, morfemen selekzioa eta proposamenen zein analisien kopurua murriztea koka daitezke.

VI.2 irudian prozesu osoaren eskema azaltzen da.

Proposamenen bufferra.

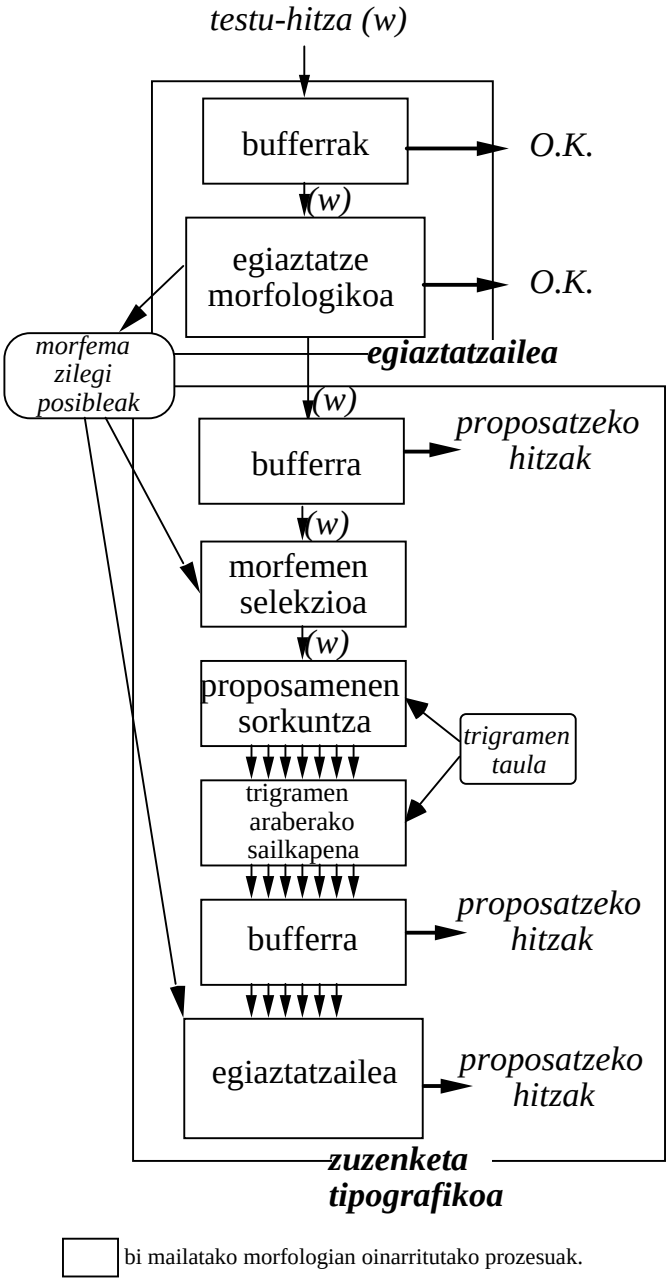
Egiaztatze-prozesua azaldu denean, esan den bezala maiztasun handieneko hitz zilegiak aparte, maiztasun handiko erroreak ere gordetzen dira, azken hauek dagozkien proposamenekin batera.

Azken buffer honen bidez maiztasun handieneko akatsak ezagutu eta zuzen daitezke urrats batean, hipotesien sorrera eta egiaztatzea saihestuz. Bufferrean gordetzen diren proposamenak errore tipografikoei zein aldaerei dagozkienak dira, eta aurreprozesu batean lortzen dira.

Prozedura azkartzeko helburuari jarraituz eta VI.2 irudian ikusten denez, proposamen hipotetikoak bilatzen dira hitz zilegien bufferrean, egiaztatzera joan baino lehen.

Trigramen bidezko proposamenen sorrera.

Aurreko kapituluan aipatu den bezala n-gramen erabilerak arrakasta handia izan du zuzenketaren alorrean, Trigramak dira n-grama erabilienak memori hartzearen eta esanguratasunaren artean oreka handiena eskaintzen dutelako. Xuxen zuzentzailerako trigramen taula bat osatu da, trigrama posible bakoitzari dagokion maiztasuna esleituz, corpusetatik lortutako datuen arabera.



VI.2 irudia.- Errore tipografikoen tratamendu eraginkorra.

Proposamenen sorkuntza azkartzeko trigramek tratamendu bikoitza bideratzen dute:

- Erroredun formaren trigramak aztertzen dira eta zilegi ez diren trigramak —taulan agertzen ez direnak hain zuzen— bakarrik hartzen dira kontuan Damerau-ren oinarrizko tipifikazioa aplikatzean. Trigrama guztiak zilegiak badira, aldiz, hitz osoaren gainean aplikatzen dira.

- Damerau-ren oinarrizko tipifikazioa aplikatutakoan trigrama ez-zilegirik duten proposamen hipotetikoak egiaztatu gabe baztertzen dira. Horrez gain,

proposamen hipotetikoak haien barneko trigramen pisuaren arabera sailkatzen dira egiaztatzeari begira.

Morfemen selekzioa.

Esan den bezala, aukeran den tratamendu hau hiru urratsetan burutzen da:

1) Hitzaren egiaztatze morfologikoa burutzen den bitartean, aurkitutako morfemak edo morfema-multzoak eta dagozkien lexikoko posizioa zein automaten egoerak gordetzen dira datu-egitura batean. Prozesua eskerretatik eskuinetara burutzen denez aurkitutako morfemak edo morfema-multzoak beti dira hasierakoak.

2) Hitza zilegia bada aurretik gorde dena baztertu egiten da baina ezagutzen ez bada proposamenen tratamendua metatutako morfemetatik abiatzen da. Horrela hasierako morfemak ontzat emango dira eta Damerau-ren araberako aldaketak ezagutu ez den partean baino ez dira aplikatuko. Oinarrizko proposamen kopurua *2nk* inguru izan beharrean, *2(n-l)k* inguru izango da, *l* ezagututako morfemaren luzera izanik.

3) Proposamen hipotetikoak egiaztatzeko analisi morfologikora jo behar baldin bada, gordetako egoeratik hasiko da analisia, eta ondorioz askoz azkarragoa izango da.

Aurkitutako morfemak anitzak badira, egoera bakoitzetik abiatzen da analisia; hipotesiak sortzeko, ordea, morfema edo morfema-multzo luzeena hartzen da erreferentziatzat.

Esan den bezala zehaztasunaren kaltean izan daiteke tratamendu hau, zeren akats baten bidez hitzaren ezkerreko partea morfema desberdin bat bilakatzen bada, morfema horretatik abiatuta ezinezkoa izango baita akatsa zuzentzea. Izan ere, §V.3.1.2 atalean azaldu den bezala, edizioan probabilitate handiagoz egon daiteke errorea hitzaren bukaeran hasieran baino (Yannakoudakis, 83).

Proposamenen kopurua murriztea.

Proposamenak sortzeko prozesua azkartzeko funtsezko parametroa analisi morfologikoen kopurua da, hauxe baita atalik konplexuena konputazioaren ikuspuntutik. Horren ondorioz hipotesi batzuk baztertu egingo dira egiaztatu gabe; baina litekeena da horietako batzuk benetako hitzak izatea, eta are gehiago hitzari zegokion zuzenketa. Beraz, ondorio kaltegarria ekar diezaioke zehaztasunari eraginkortasun hobetze honek, analisi-kopurua murriztea proposamen kopurua murrizteak ekar baitezake. Ondorio kaltegarri hori dela eta, tratamendu hau aukerazkoa eta parametrizagarria izango da.

Parametrizatze hori bi aldagairen arabera egin daiteke —argi egon arren haien artean erlazio zuzena dagoela—:

- proposamen kopuruari muga jartzea analisiak burutzen hasten denetik.
- analisi kopurua mugatzea proposamen kopuruari jaramonik egin gabe.

Bi irizpideak konbina daitezke eta horrela egin dugu gure produktu komertzialean (ikus §VI.6 atala).

Beste aldetik, eta VI.2 irudia aztertuz ikus daitekeenez, proposamen hipotetikoak banan-banan egiaztatu beharrean bi urratsetan burutzen da: lehenean hipotesi guztiak bilatzen dira hitz zilegien bufferrean; horrela ez da analisi morfologikorik egin behar, han aurkitutakoekin proposamen kopurua osatzen baldin bada.

VI.4. Gaitasun-erroreen zuzenketa.

Errore mota hau da diseinu irizpideen arabera zuzentzeko lehentasuna duena. V.3.1.4 atalean aipatzen zen bezala, errore hauek tratatzeko arrazoi nagusia zera da: beste motako erroreen zuzenketa nola egin erabiltzaileak jakin ohi duen bitartean, hauena normalean ez du jakiten. Hainbestetan aipaturiko euskararen egoera dela eta, hauen portzentaia beste hizkuntzetakoa baino handiagoa denez, are interesgarriago bihurtzen da prozesaketa hau.

Analisi estandarraren bidez ezagutu ez den hitz bat aldaeratzat hartzeko, laugarren kapituluko bigarren atalean azaldutako aldaeren analisi morfologikoaz ezagutu behar da. Han azaldutakoa puntu hauetan labur daiteke:

- Sistema estandarreko lexikoa eta erregela-sistema osatzen dira, azpilexiko multzo berri eta erregelen azpisistema berri bana erantsiz.

- Azpisistema osagarri horren bidez horrelako kasuak aurrikusten dira:
  1) Morfemen aldaera: morfema baten ordez beste bat erabiltzetik datozen akatsak. Jatorrizko morfemaren eta aldaerari dagokionaren arteko desberdintasuna diakritikoren bat bada, morfema konkretu hau lotzean egiten diren akatsak ere ezagutzen dira. Morfemaren aldaera eta dagokion zilegia lotzen dira lexikoan.
  2) Morfotaktikaren aldaera: morfema baten ondoren etor daitezkeenak aldatzetik datozen erroreak.
  3) Aldaera erregularrak: errore fonologiko, morfologiko eta ortografiko erregularrak bi mailatako erregela osagarrien bidez ezagutzen dira.

- Analisia burutzean azpisistema osagarria estandarrarekin batera ibiltzen da, hitzetan bi motako morfema zein aldaketa morfofonologikoak gerta daitezke eta.

Analisirik aurkitzen baldin bada, morfemen aldaerei dagozkien morfema estandarrak bilatu behar dira lexiko-loturaren bidez.

Orain arte ikusitakoa laugarren kapituluan sakonean azaltzen da, baina zuzentzaile ortografikoan gaitasun-erroreetarako aldaeren analisi morfologikoan egiten ez zen prozedura bat burutu behar da: aldaerari dagokion zuzenketa edo forma estandarra lortu behar da. Zuzenketa hau burutzeko sorkuntza morfologikoa erabiliko da: analisian lortutako morfema estandarrak lotzen dira erregela estandarrak erabiliz.

Hala ere **arazo** bat dago, morfotaktikaren aldaerak sortutako akatsak ezagutu arren ezin baitira zuzendu. Arrazoia sinplea da, aldaera honen bidez erabiltzaileak eraikitako hitzaren osagaiak zilegiak dira baina ez kateatze konkretu horretan. Lexikoa diseinatu den bezala behintzat, ezinezkoa da zuzentzea, aldaeraren deskribapenean dagoen jarraitze-klasea estandarrarekin loturik ez dagoenez gero, ez baitago jakiterik bi jarraitze-klase horietan dauden morfemen artean zer erlazio dagoen. Aldaera hauen deskribapen konplexuago baten bidez zuzentzeko aukera egon liteke, baina ez dirudi halakorik egiteak pena merezi duenik, aldaera mota hau bereziena dela kontuan hartzen badugu.

Beste irtenbidea hauxe izan daiteke: morfotaktikaren aldaera morfemen aldaera multzo bezala adieraztea, hau neketsua izan badaiteke ere.

Adibide gisa *batzu* morfemaren jarraitze-klasea dugu. IV.3 irudian azaltzen zen bezala, jarraitze klase estandarra *PLU* (plurala) da, baina aldaera gisa *MG* (mugagabea) erabili ohi da[1]. Honen zuzenketa burutzeko zeharkako aukera hau legoke:

- *PLU* azpilexiko bakarra denez, berau osatzen duten morfemak azpilexiko berri batean bikoiztu, alomorfoak sortuz. Horrez gain, *batzu*-ren jarraitze-klasea aldatu beharko litzateke.
- Azpilexiko hori bikoiztu osagarri ez-estandar batean, morfema bakoitzari dagokion *MG*-ko morfema zehaztuz morfemaren aldaera gisa.

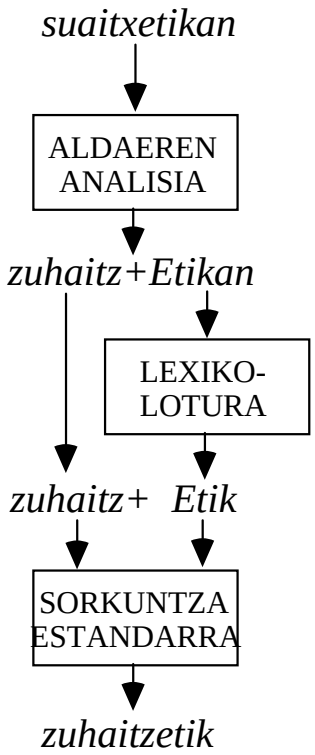
Lehen urratsa aurrez daiteke baina horrela izugarrizko gainsorrera bultzatzen da, *PLU* jarraitze-klasea erabiltzen duten lema guztietarako suposatzen ari baikara aldaera lokal bat.

Maiztasun handia duten mota honetako aldaeren zuzenketa *ad-hoc* edo partikularra bidera daiteke maiztasun handieneko hitzen bufferraren bidez. Morfotaktikaren aldaerei dagozkien formatarako, eta aipatu den aurreprozesaketaren bidez proposamen egokirik lortu ez bada, salbuespen gisa ukituak egin daitezke bufferrean.

---

1 Orain dela gutxi arte bien erabilera onartuta zegoen, baina orain pluralarena bakarrik onartzen da.

Gaitasun-erroreen zuzenketa bere osotasunean azaltzeko har dezagun *suaitxetikan* hitzaren **adibidea**. Forma honi dagokion zuzenketa-prozesua VI.3 irudian agertzen da.



VI.3 irudia.- *suaitxetikan* hitzaren zuzenketa aldaeren analisi eta sorkuntza morfologiko estandarraren bidez.

suaitxetikan zuhaitzetik forma estandarraren aldaera bezala ikus daiteke —adibidea muturrera eraman da, baina zuzentzeko aukerez jabetzeko oso egokia—. Azpimarratzekoa da zuzentzea badagoela edizio-distantzia bostekoa izan arren. Dagokion analisia IV.2.3.1 atalean azaldu zen, bertan zera ikus daitekeela:

1) *zuhaitz* lema lortzen da *suaitx* hitz-zatitik bi erregela osagarriri esker: arrakasta bi aldiz duen txistukarien arteko aldaketarena (z-s, z-x) eta h-ren galerarena.

2) *Etikan* atzizkia lortzen da azpilexiko osagarriei esker.

3) Aipaturiko azpilexikoetan *Etikan Etik* forma estandarrekin agertzen da lotuta.

Behin analisia burutu ondoren, eta zuzentzeko arazorik ez dagoela ikusita sorkuntzara pasatzen da, *zuhaitz+Etik* lexiko-karaktereetatik erregela estandarrei dagozkien itzultzaileen bidez *zuhaitzetik* lortuz.

Bibliografian aipatutako errore fonologikoen zuzenketarekin alderatuz gero, gaitasun-erroreak tratatzeko sistema orokorra, dotorea zein beste moduluekiko homogenoa bihurtzen da gure burutzapenean.

Gure sistemarako 30 aldaketa baino gehiago deskribatzen duten 18 erregela osagarri (ikus §IV.2.2.2 atala), eta ia mila morfema ez-estandar landu dira, oso azpisistema ahaltsua osatuz.

Azkenik aipatu behar da metodo honen bidez “hitz-mugaren gaineko errore” batzuk zuzen daitezkeela. Banaturik idazten diren zenbait forma, *hitz egin* adibidez, batera idaztea aldaera bezala jaso daiteke morfema ez estandarretan, *hitzegiN hitz_egiN* forma estandarrarekin lotuz; eta horrela, morfema horien flexioa zuzen daiteke.

VI.5. Sistemaren arkitektura eta ezaugarriak.

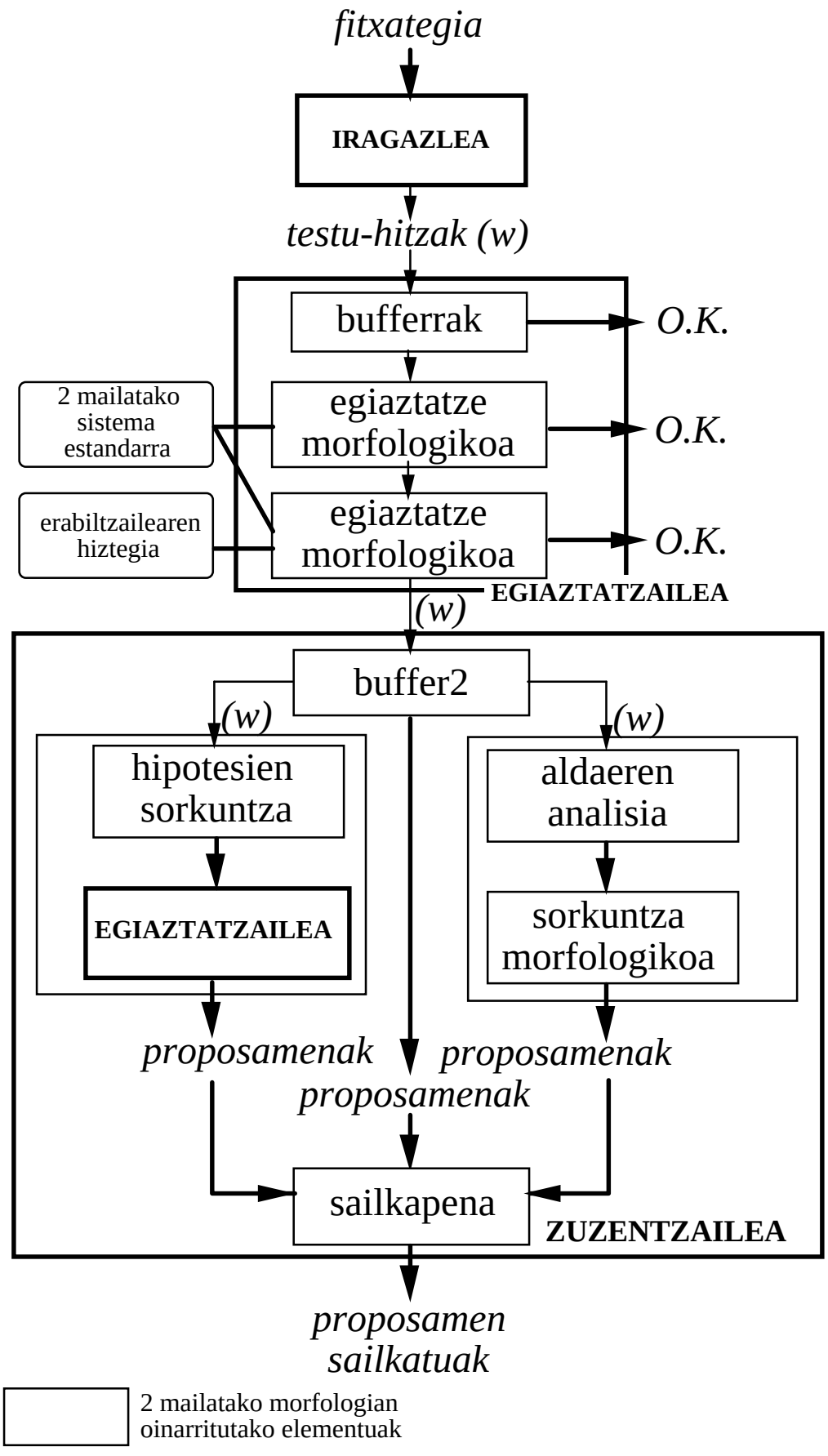
Aztertutako modulu egiaztatzailea eta zuzentzailea zuzentzaile ortografikoaren funtsa dira baina ez osagai bakarrak. Horiekin batera ondoko modulu hauek gehitu behar dira, kalitateko zuzentzaile bat burutu nahi bada:

- Proposamenak sailkatzeko modulua. Zuzentzaileak sortutako proposamenak ordenatzea da horren helburua. Zuzenketa automatikoa behar den sistemetan funtsezko modulua da.
- Erabiltzailearen hiztegiaren kudeatzailea. Egiaztatzailearekin lotuta dago, hiztegi hau kontuan hartu behar delako egiaztatzean, baina eguneratzeko aukera ere eskaini behar zaio erabiltzaileari.
- Iragazlea edo token-ezagutzailea. Fitxategi bat irakurriz hitzen eta testu-unitate berezien ezagutzaz arduratzen da, eta egiaztatu aurretik burutzen da.

Aipatutako osagai hauek banan-banan aztertuko dira ondoren, eta aurretik deskribatutako funtsezko moduluekin VI.4 irudian azaltzen den arkitektura osatzen dute.

VI.5.1. Proposamenen sailkapena.

Proposamenak sailkatzerakoan komenigarria litzateke testuingurua kontuan hartzea baina, esan den bezala, hau gure lanaren esparrutik kanpo dago. Testuingurua erabiltzen ez bada proposamenak sailkatzeko honako teknika hauek erabil daitezke (ikus §V.3.3.1):



VI.4 irudia.- Xuxen zuzentzaile ortografikoaren arkitektura.

- Edizio-distantzia edo aipatutako bestelako distantziaren neurriak. Azken hauek dira zuzenketa automatikoan erabiltzen direnak. Dena den edizio-distantzia erabiltzen bada metodoren batez bereizi beharko dira distantzia bera dutenak.

- Errore-corpusen gainean aplikatutako metodo estokastikoak, taula estatistikoak, eredu markoviarrak edo sare neuronalen bidezkoak erabiliz.

- Proposamenen maiztasuna: estatistikak, errorearen eta dagokion zuzenketaren arteko erlazioan oinarritu beharrean, hitz zilegien datu sinpleetan oinarri daitezke. Metodo hau aurrekoa baino askoz sinpleagoa da, baina errore-corpusik ez da behar.

Gure kasuan sailkapena proposamen hipotetikoen gainean egin zitekeen, baina kontuan hartu behar da kasu horretan aldaeren tratamenduaz sortutako proposamenak sailkapenetik at geldituko liratekeela. Beste aldetik errore-corpus fidagarririk ez dagoenez bigarren puntuko irizpideak ezin izan dira aplikatu, eta beraz, hirugarrenekoak aplikatu dira.

Proposamenak egiteko honako algoritmoari jarraitzen zaio, bai maiztasuneko handieneko akatsei dagozkien zuzenketak sortzean, bai zuzenketa-prozesuan zehar:

1) Bateko edizio-distantzian eta maiztasun handieneko hitz zilegien bufferrean dauden proposamenak. Hauek maiztasunaren arabera sailkatu beharko lirateke, baina *maiztegi* izeneko buffer horretan maiztasunaren balioa ez da jasotzen memoria-hartze arrazoiak direla eta. Horren ordez barneko trigramen pisuaren arabera sailkatzen dira.

2) Aurreko puntuan sartzen ez diren aldaeren tratamenduz lortutako proposamenak, hurrenez hurren edizio-distantziaren arabera eta barneko trigramen eraketaren arabera sailkatuak.

3) Gainontzeko proposamenak morfologikoki egiaztatu ahala, aurretik barneko trigramen eraketaren arabera sailkaturik baitaude.

Algoritmo honekin lortutako emaitzak VI.7 atalean azaltzen dira.

VI.5.2. Erabiltzailearen hiztegia.

Hizkuntza eranskarien zuzenketan dauden arazo berezien artean, hiztegiaren aberasketa aipatu da aurreko kapituluan (ikusi V.4.1 atala).

Sistema komertzial batzuetan egiten den hitz-zerrendaren bidezko erabiltzailearen hiztegia ez da, inola ere, egokia hizkuntza eranskarietarako. Hitzen ordez morfemak —gehienetan lemak— metatu eta erabili behar dira.

Xuxenerako IV.1 atalean azaldutako erabiltzailearen lexikorako burutzapena berrerabili da, horrekin helburu bikoitza lortuz:

- Azpilexiko **irekiak** definitzeko aukeraren bidez lema berriak erabiltzailearen hiztegietan gordeko dira. Lemari dagokion azpilexikoa, jarraitze-klasea eta bi mailatako morfologiaren araberako lexiko-maila (diakritikoak eta guzti beharrezkoak bada) lortu behar dira linguistikan aditua ez den erabiltzailearengandik. Horretarako IV.1.3 atalean zehazten den informazio morfosintaktikoa —kategoria eta, kasuaren arabera, azpikategoria eta ezaugarriren bat— eskatzen zaio erabiltzaileari interfaze atsegin eta sinple batez (ikus VI.6 irudia).
- Egiaztatze morfologikoa bi urrats edo gehiagotan burutzen da: lehenean lexiko orokorra kontsultatzen den bitartean, ondorengoetan erabiltzailearen hiztegiak kontsultatzen dira lexiko orokorreko azpilexiko orokorrak ere erabiliz, VI.4 irudian ikus daitekeenez. Aplikatzen diren erregela morfofonologiko estandarrak ez dira aldatzen urrats desberdinetan.

Horrela hiztegi desberdinak erabil daitezke jakintza-arloaren arabera eta “benetako hitzaren errore” batzuk saihestu egingo dira, hitz orokor bat, akatsen baten eraginez, beste jakintza-arloko hitz berezitu bat bihurtzen denean.

VI.5.3. Iragazlea edo token-ezagutzailea.

Hitzak eta beste testu-unitateak bereiztea da modulu honen helburua. Analizatzaile estandarrerako egindakoa berrerabili da zenbait aldaketa eginez. Bi automatatan oinarritzen da eta bere zeregina ondoko puntuetan bana daiteke:

- Zenbakiak, arruntak edo erromatarrak, bereiztea dagokien deklinabidearekin batera. Deklinabidea ondo dagoen ala ez ziurtatzeko egiaztatzera bidaltzen da.
- Laburdurak eta siglak identifikatzea dagokien deklinabidearekin. Egiaztatzera bidaltzen dira.
- Lerro-bukaeran hitza banatzen duen marratxoa (*hyphenation*) ezagutu eta kontuan ez hartzea. Marratxoaren tratamendu korapilatsua azpimarra daiteke: aipatutako funtzioaz gain elkarketarena, erdal hitzen atzizkiekiko lotura eta bereizgarri-funtzioa ere izan baititzake.
- Zuriuneak, puntuazio-zeinuak eta gainontzeko hitzen arteko bereizgarriak ezagutzea.

- Maiuskulaz osoki idatzitako hitzekin tratamendu berezia egitea, maiuskulaz ezagutzen ez badira minuskulaz ere egiaztatuko direlarik.
- Testuetan karaktere arraroak, bereizgarriak edo hizkuntzarenak ez direnak agertzen badira, horren berri ematea erabiltzaileari.

Zenbait informazio metatzen da zuzenketak eskaintzeko orduan kontuan har dadin. Horrela, hitza osoki edo hasieran maiuskula duen ala ez, zenbaki eta laburduren kasua, etab.

Zuzentzaile komertzial batzuetan aurreko eragiketa batzuk aukeran ematen zaizkie erabiltzaileei, horrela hauek maiuskulaz idatzitakoa, zenbakiak, siglak, laburdurak, karaktere arraroak etab. egiaztatu nahi dituzten ala zuzenean ontzat eman nahi dituzten hauta dezaten.

Proiektuan gure diseinu-filosofiarekin bat etorriz —zalantza kasuan nahiago izan dugu abisua ematea ontzat ematea baino, eta horrexegatik ekidin dugu gainsorrera— nahiago izan dugu dena egiaztatzea. Hala ere, etorkizunean halako parametrizazioak egitea litzateke egokiena.

VI.6. Produktu komertzialaren diseinua.

Aurreko ataletan azaldutako osagaiekin zuzenketarako prototipo parametrizagarria osatu genuen VI.4 irudian agertzen den arkitekturari jarraituz. Prototipo hori produktu komertzial bihurtzeko eman behar izan diren urrats garrantzitsuenak honako hauek izan dira:

- Aukerazko azkartze-mekanismoak erabakitzea, makinaren arabera zehaztasuna/ eraginkortasuna oreka mantenduz.
- Interfaze atsegina diseinatzea, erabiltzailearekiko hartu-emanak ahalik eta modu sinple bezain esanguratsuenean buru daitezen.
- Makina zein testu-editore zehatz bakoitzerako egokitzapena. Macintosh eta PC izan ziren aukeratutako oinarri-ordenadoreak eta Word eta WordPerfect testu-editoreak.

Lehen bi puntuak interesgarriak izan daitezkeelakoan zabaldu egingo ditugu ondoren.

VI.6.1. Zehaztasuna/eraginkortasuna oreka.

Zehaztasuna/eraginkortasuna oreka mantentzea da LNPko aplikazio guztien ardatzetako bat, eta produktu komertzial baten arrakastarako aldagai nagusietako bat.

Xuxen zuzentzaile ortografikoa merkaturatzeko orduan prototipoa erabiliz neurriak hartu genituen, ondoko ondorio kualitatiboak lortuz:

- Zehaztasun aldetik kalitate handiko egiaztatze/zuzenketa egiten zuela, bateko distantziatik gorako errore tipografikoen kasuaren salbuespenaz.
- Abiaduraren aldetik motela zela konputagailu pertsonaletan korritzeko beste produktu komertzialekin alderatuz gero, hizkuntza eranskarietarako emandako datuekin parekagarria bada ere.

Horren aurrean abiaduraren aldera orekatzea erabaki genuen aukerazko azkartze-metodo guztiak erabiliz, zehaztasunean ahalik eta eragin txikiena eraginez noski. Ondorioz, horrela moldatu genituen azkartze-metodoak:

- Hitzaren hasieran aurkitutako morfemei buruzko informazioa erabiltzea hipotesi kopurua laburtzearren eta hipotesien analisi morfologikoa azkartzearren.
- Proposamenen kopurua mugatzea analisi morfologikoen kopurua laburtzeko. Bide horretan gaitasun-erroreen tratamendutik eta bufferrean bilatzetik sor daitezkeen proposamen guztiak lortzen dira, baina hipotesien analisi morfologikoari aurreko bideetatik proposamenik sortu ez bada soilik ekiten zaio. Beraz, proposamen bat lortuz gero ez da analisi morfologiko gehiagorik egiten. Gainera, analisi kopuru mugatu batera iritsiz gero, proposamenik lortu ez bada ere, zuzenketa-prozesua eten egiten da, eta horrexegatik da funtsezkoa hipotesien sailkapena barneko trigramen arabera. Analisi kopuruaren muga hitzaren luzeraren menpe dago, hitza zenbat eta luzeago analisi morfologikoa hainbat eta motelago baita.

Bide honetatik lortzen da zuzenketarako abiadura onargarria konputagailu pertsonaletan.

VI.6.2. Erabiltzailearekiko interfazea.

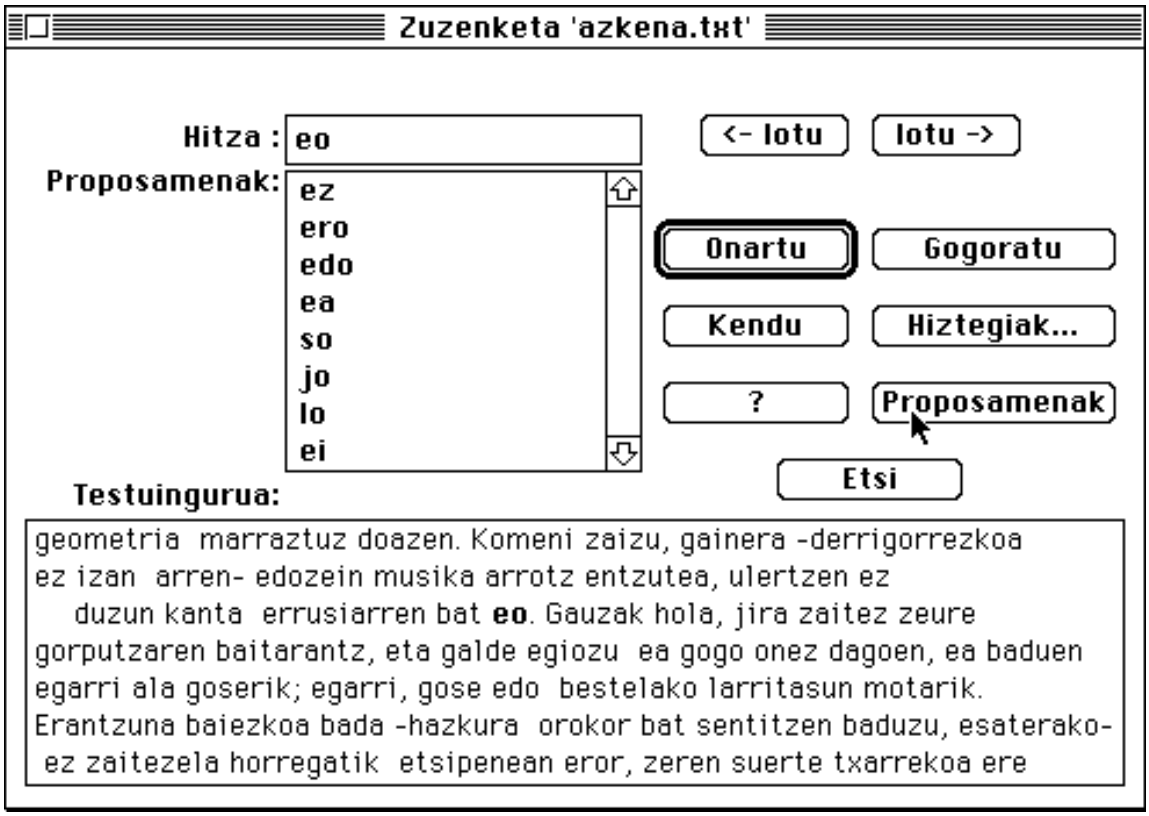
Zuzentzailea merkaturatzeko beste funtsezko ekimen bat interfazea egoki eta atsegina izaten da.[1]- GUIen garaian, leihoetan oinarritutako interfaze-sistema bat, objektuei

[1] GUI: Graphical User-Interface (Erabiltzaile-Interfaze Grafikoa)

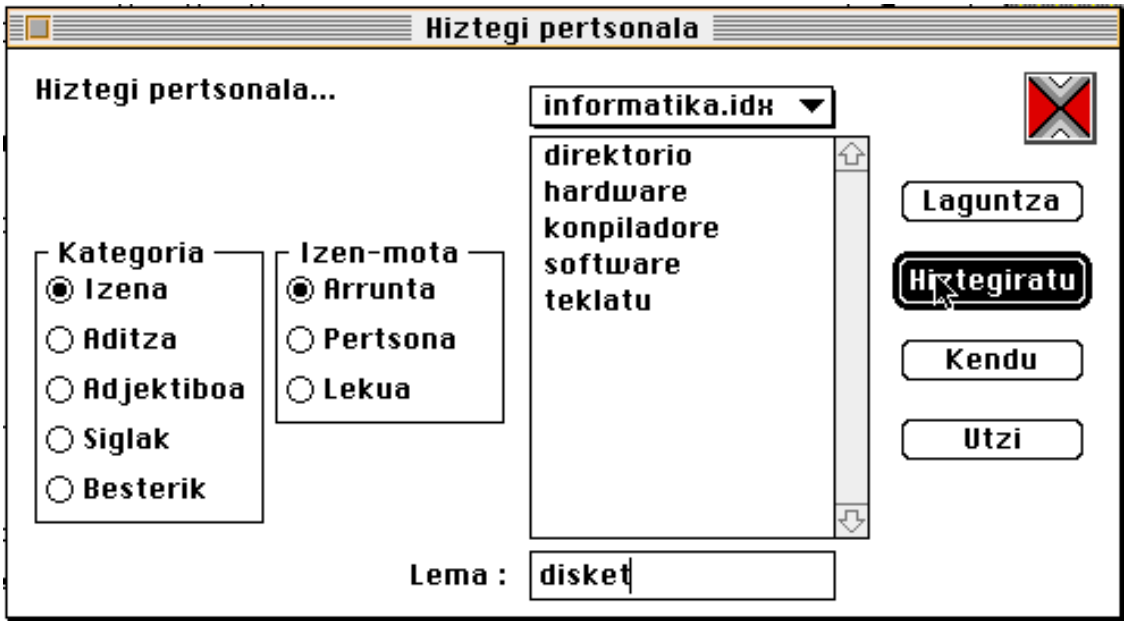
zuzendua eta atzean gertaerei zuzendutako programazio-eredua duena, ia-ia ezinbestekoa da produktua arrakastatsua gertatzeko.

Ildo honetatik garraigarritasuna ziurtatzen duen ingurune batean programatzea ezinbestekotzat jo daiteke zuzentzailea plataforma desberdinetarako eskaini nahi bada. XVT izan zen ingurune garraiagarria garatzeko aukeratu genuen software-paketea.

Interfazeari buruzko zehaztasun gehiago ematearren VI.5 eta VI.6 irudietan bi leiho nagusiak azaltzen dira: zuzenketa-prozesua kudeatzekoa eta erabiltzailearen hiztegia aberastekoa.



VI.5 irudia.- Xuxen zuzentzaile ortografikoaren leiho nagusia



VI.6 irudia.- Xuxen zuzentzailean erabiltzailearen lexikoa aberasteko leihoa.

VI.7. Zehaztasuna eta eraginkortasuna.

Aurreko ataletan deskribatutako egiaztatze- eta zuzenketa-prozesuen gainean hartutako datuak azaltzen dira pasarte honetan.

VI.7.1. Egiaztatzea.

Zehaztasunaren aldetik III.9 eta III.10 irudietan azaldutako emaitzak aipa daitezke hastapen gisa. Horren arabera testu-hitzetako %91-96a ezagutzen da, beren analisia lortzen da eta. Erabiltzailearen lexikoa erabiliko balitz portzentaia hori igo egingo litzateke, analizatzen ez diren hitzen erdia baino gehiago ez baitira erroreak; baina ez dira ezagutzen dagokien lema lexikoan ez dagoelako. Hala ere, kontuan hartu behar da aipatutako testuetan akats tipografiko gutxi dagoela, argitaraturiko testuak dira eta.

Egiaztatu gabeko testuak izaten dira zuzentzaile ortografikoen helburua eta horrelako testu batzuk aztertuz lortu dira VI.7 irudian azaltzen diren emaitzak. Bertan hiru iturburu desberdinetatik eskuratutako testuak agertzen dira, antzeko tamainakoak direnak:

- Ilazki euskaltegian euskara ikasten duten azken urratseko ikasleek egindako testuak[1], ikasle batek sakaturik —ikasleen testuak deitutakoak. Hauetan aldaera gehiago dago besteetan baino, aldaeretan gaitasun-erroreak ere sartzen baitira.

Testuak	**hitzak**	**ezagutu gabe**	**ondo[2] daudenak**	**ezagutut. aldaerak**	**errore tipograf.**
1.-Ikasleen testuak	3.959	326 **%8,2**	63 [3]%19,3	171 **%52,5**	92 **%28,2**
2.-Testu teknikoa	3.891	220 **%5,6**	32 %14,5	32 **%14,5**	156 **%71**
3.-Prentsako testuak	4.319	205 **%4,7**	42 %20,5	79 **%38,5**	84 **%41**
GUZTIRA	12.097	751 %6,3	137 %18,2	282 %37,6	332 %44,2

VI.7 irudia.- Egiaztatzeari buruz hartutako estatistikak.

- UZEIn sortu eta sakatutako testu tekniko zuzendu gabeak. Terminologia teknikoa kenduta testu estandarra da eta horregatik aldaera gutxi detektatzen dira. Terminologiaren arazoak ebazteko hiztegi berezitu bat aberastu da aurretik.

- *Egunkaria*tik lortutako prentsako testu zuzendu gabeak. Estandarrak dira hein batean, hala ere izen nagusien eragina saihesteko maiuskulaz hasitako izenak ez dira kontuan hartu, eta horregatik ezagutu gabeko hitzak gutxiago dira beste testu-zatietan baino.

Beraz, irudian azaltzen diren datuetatik atera daitezkeen ondorioak espero zitezkeenak dira.

Jakintza-arloarekin lotutako lexikoa hiztegi berezituetan antolatzeagatik, eta egindako deskribapen morfologikoari dagokion gainsorrera-ezarengatik, **benetako hitzaren erroreak** ekiditen dira ahal den neurrian. Horien zenbatekoa corpusetan oinarriturik kalkulatzea zaila da, automatikoki egitea ezinezkoa da eta, ondorioz lagin adierazgarria aztertzea oso neketsua izango litzateke. Horren ordez hurbilpen estatistikoa egin dugu.

---

1 Testu hauek M. Maritxalarrek bildu ditu bere ikerketarako OLIren esparruan (Maritxalar & Diaz de Illarraza, 93).

2 Lexikoan ez egoteagatik edo beste hizkuntzetako hitzak izateagatik ondo dauden baina ezagutu ez diren hitzak.

3 Ondoko hiru portzentaiak egiaztatu gabeko hitzen gainean kalkulatzen dira.

Luzera	Hitz Kopurua.	Benetako hitzak(%)
2	18	9,7
3	74	6,8
4	81	6,0
5	56	5,5
6	67	3,0
7	61	2,6
8	39	1,7
9	41	1,5
10	21	0,9
11	20	1,0
12	13	1.0
13	3	0.5
14	3	0.4
best.	3	0.0
GUZT.	500	4.0

VI.8 irudia.- Benetako hitzaren erroreen probalititatea (hurbilpena).

Hurbilpen horretan testu bateko 500 hitz[1] zilegi jarrai hartu dira, eta bateko edizio-distantzian dauden forma guztien artean benetako hitzen portzentaia kalkulatu da, %4 ingurukoa izanik errore hauen probabilitatea. VI.8 irudian luzeraren araberako datuak aurkezten dira.

Datu hauek minimotzat jo behar dira; hurbilpenean egin diren bi sinplifikazioen artean, erroreak beti bateko distantzian daudela eta bateko distantzian daudenek probabilitate bera dutela alegia. Bigarrenak batez ere eragin nabarmena eduki dezake emaitza hori txikiagotuz, frogatutzat har baitaiteke erroreak sortzean benetako hitzak idazteko joera handiago dagoela arrazoi sikolinguistikoak direla eta.

Abiaduraren aldetik analisi morfologikoarena 2 hitz segundoko da—III.5.4n esaten zen bezala Sun-Sparc IPX baterako—, eta egiaztatzearena 25-30 hitz segundoko izatera iristen da bufferren erabilerarengatik eta lehen analisiarekin analisi-prozesua bukarazteagatik. Izan ere, zuzenketa-fasean egiten diren egiaztatzeak, existitzen ez diren hitzei dagozkienez, analisi morfologiko osoen denboratik gertuago daude hitz arrunten egiaztatzeenetik baino.

[1] Neurri honekin lortzen diren emaitzak egonkorrak direla egiaztatu da.

VI.7.2. Zuzenketa.

Zuzenketaren zehaztasunari eta abiadurari buruzko datuak lortzea oso inportantea da bi arrazoirengatik: bi aldagai horien artean bilatu beharreko oreka-puntua bilatzeko eta bibliografian dauden beste sistemekin alderatzeko.

Egiaztatzeari buruzko estatistikak lortzeko erabilitako corpus bakoitzetik 100 errore hartu eta horren gainean VI.9 irudian azaltzen diren estatistikak lortu ditugu. Datu kopurua dela eta fidagarritasuna mugaturik egon arren, estatistiken emaitzak interesgarriak dira. Estatistika horietan erabilitako aldagaiak hauek dira:

A) Zutabeetan lau aukera agertzen dira: 1) Xuxen zuzentzaile komertzialean erabili dugun zuzenketa azkarra, proposamen hipotetikoen analisia aurkitutako morfemetatik abiatzen duena eta hauen kopurua mugatzen duena. 2) Aurreko berbera baina analisi kopurua murriztu gabe. 3) Proposamenen analisi aukera guztiak kontuan hartzen direnekoa baina analisi kopuru mugatuarekin. 4) Proposatutako metodoa batere murriztapenik gabe zuzentzen duena. Lehena azkarrena den bitartean, laugarrenean ziurtatzen da bateko edizio-distantzian dauden errore guztien zuzenketa agertuko dela[1] proposamenen artean. Bigarrena eta hirugarrena tarteko ebazpideak dira. Kontuan hartu behar da guztietan amankomunean dagoena: akats ohizkoenen bilaketa, proposamen hipotetiko guztien bilaketa maiztasun handieneko hitz zilegien bufferrean, eta gaitasun-erroreen tratamendua.

B) Corpus bakoitzeko eta aurretik aipatutako zuzenketa-metodo bakoitzeko lau neurri ematen dira: azkena hitz bakoitzari dagozkion proposamenak sortzeko batez-besteko denbora den bitartean, lehenengo hirurak zehaztasunarenak dira, erroreari dagokion zuzenketa zenbatetan proposatzen den *(n)*, zenbatetan proposatzen den lehen hiruren artean *(3)* eta zenbatetan lehena *(1)*.

[1] Bi edo edizio-distantzia handiagoko erroreen zuzenketa proposamen gisa agertuko da, baldin eta aldaera bezala ezagutzen bada.

Testuak		**Xuxen** (mugatua)	**Xuxen** (muga gabe)	**morfema guztiak** (mugatua)	**guztiak**
1.-Ikasleen testuak	(n)	%82	%87	%81	%89
	(3)	%81	%85	%80	%86
	(1)	%74	%76	%73	%75
	denb.	0,3 s	6,2 s	0,6 s	15 s
2.-Testu teknikoa	(n)	%63	%73	%64	%88
	(3)	%62	%72	%63	%86
	(1)	%49	%56	%50	%68
	denb.	0,4 s	2,7 s	0,5 s	12,5 s
3.-Prentsako testuak	(n)	%70	%80	%71	%89
	(3)	%68	%78	%69	%85
	(1)	%59	%64	%60	%71
	denb.	0,35 s	4,5 s	0,6 s	16,7 s
GUZTIRA	(n)	%72	%80	%72	%89
(300 akats)	(3)	%**70**	%**78**	%**71**	%**86**
	(1)	%61	%65	%61	%71
	denb.	**0,35** s	**4,5** s	**0,6** s	**14,7** s

VI.9 irudia.- Zuzenketaren zehaztasuna eta abiadurari buruz hartutako estatistikak.

Irudian azaltzen diren datuak aztertuz ondoko ondorioak aipa daitezke:

- Denborak: Kontuan hartu behar da emandako denborak batez-bestekoak direla, baina proposamen kopurua mugatzen duten bi metodoetan desbiderapen handia dagoenez gero, kasu batzuetan proposamenak sortzeko denbora luzeagoa izan daiteke.

- Zehaztasuna: Egiaztapen oso guztiekin hiru corpusekin emaitza antzekoak lortzen badira ere, gainontzekoetan askoz emaitza hobeak lortzen dira aldaera edo gaitasun-errore asko duten testuetan (ikasleenak) besteetan baino. Horrekin baieztatzen da aurretik esan dugun zerbait: aldaeren tratamendua osoagoa da errore tipografikoena baino. Egiaztapen guztiekin zehaztasunak muga bat du %90ean, eta hori bateko edizio-distantzian dauden ordezkagaiak bakarrik aztertzetik dator nagusiki —distantzia handiagoan dauden batzuk zuzentzen dira aldaeren tratamenduari esker—.

- Metodoen arteko desberdintasunak: Xuxen zuzentzaile ortografiko komertzialerako erabilitako metodoak (lehen zutabekoa) nahikoa oreka ona lortzen du zehaztasuna eta eraginkortasunaren artean, batez ere gaitasun-errore

asko dagoenean. Hala ere, eta egiaztatze morfologikoa azkartu ahala, bigarren eta laugarren zutabeei dagozkien metodoetara jo beharko da, hirugarrenekoan zehaztasuna apenas hobetzen ez baita.

Flexio handiko hizkuntzetan aplikatzen diren beste sistemetarako ematen diren neurriekin konparatuz gero, nahikoa emaitza onak lortzen direla esan daiteke, gaitasun-erroreen tratamendua horretan garrantzi handia izanik.

VI.8. Proposatutako hobekuntzak.

Egindako sistemaren gainean aztertzen ari garen hobekuntzak azaltzen dira kapituluaren azken atal honetan. Hobekuntza hauek, normala denez, bi bidetatik joan behar dute, abiadura eta zehaztasunaren aldetik, alegia.

Abiadura da zuzentzailearen alde ahulena beste hizkuntzetarako merkatuan dauden zuzentzaileekin alderatzen badugu. Hobekuntza horretarako funtsezkoa da analisi morfologikoaren denborak laburtzea, horretan aztertutako lexiko-itzultzaileek markatutako bidea jorratu behar delarik. Abiadura hobetzea zehaztasunaren onerako izango da, oraingo metodoarekin egiten diren mozketak, VI.6.1 atalean azaldu direnak, bazter daitezkeelako.

Zehaztasunaren aldetik zuzentzaile zehatza bada ere, hauek dira puntu ahulenak eta hobetu behar direnak:

- Bateko distantzia baino gehiago duten errore tipografikoen tratamendua. Dagoen zuzentzailearekin eginez gero, abiaduran oso eragin handia jasango genuke.
- Testuingurua kontuan hartzea, proposamenak sailkatzea, eta benetako hitzaren erroreen detekzioa hobetu.

Hobekuntzarako bide horiek bi urratsetan laburtuko ditugu: proposamen-sistema erro-hizkiaren bidez burutzea batetik, eta lexiko-itzultzaileak erabiltzea bestetik.

VI.8.1. Lexiko-itzultzaileen erabilera.

Lexiko-itzultzaileen ezaugarri nagusietako bat analisi morfologikorako eskaintzen duten abiadura dugu. Beraz, III.6 atalean azaldutako euskara estandarrerako prozesadore morfologikoa egiaztapen morfologikorako erabiliz, eta IV.2.4 atalean azaldutakoa aldaeren tratamenduari aplikatuz emandako denborak ehun bat aldiz azkar litezke, ondorioz zehaztasuna laburtzen duten mugak kentzeko aukera emanez, eta morfologikoki sinpleago diren beste hizkuntzen zuzentzaileekin parekatuz.

Erabiltzailearen hiztegiarekin dira arazo bakarrak lexiko-itzultzaileak zuzenketan aplikatzeko garaian. Honen integrazioa Carter-ek (1995) adierazitako bidetik etor liteke, lexiko-itzultzaileetan egiten den lexiko osoaren konpilazioaren ordez lema irekiak ez direnak soilik konpilatuz.

VI.8.2. Erro-hizkiaren bidezko proposamen-sistema.

Erro-hizkian oinarritutako metodoa honetan datza:

- Hitz baten erro-hizkien banaketa posibleak lortu. Hau da egitekorik zailena.
- Alde bakoitzaren zuzenketa posibleak sortu, horretarako aurreko kapituluan azaldutako mekanismoak erabil daitezkeela.
- Sorkuntza morfologikoaren bidez proposamenak eskuratu. Lan honetan morfologia nahiz morfotaktika eduki behar da kontuan, morfotaktikaren aldetik jatorrizko zatien kateatzea zilegia bada ere, zati horietatik sortutako zuzenketa-zatiak elkartezinak izan baitaitezke morfotaktikaren aldetik.

Lehen urratsari dagokionean zerbait egin da aurretik. Batetik, lema hipotetikoak gordetzen dira analisia egin ahala, VI.3.1 atalean aipatutako morfemen selekzioaren bidez. Bestetik, laugarren kapituluan azaldutako lexikorik gabeko analisiari esker lema ezezagun bati dagozkion atzizki-multzo posibleak lor daitezke.

Arazo nagusia ondokoa da: akatsek erroari eta hizkiren bati batera eragiten badiete —biko edizio-distantziaz edo mugako bi karaktere jarrairen arteko trukeaz— lehen urratsa egitea oso konplexu bihurtzen da.

Aurreko eragozpenen aurrean, hobekuntza hau irekitako ikerlerro gisa utzi dugu.

ONDORIOAK ETA AURRERA BEGIRAKOAK

VII. Ondorioak eta zabaldutako ikerlerroak.

VII.1. Ondorioak.

Ikerlan honen emaitza gisa bi oinarrizko tresna eraiki dira: euskararen morfologia ezagutzen duen prozesadore sendo eta estaldura handiko bat batetik, eta prozesadore horrek burutzen duen analisi eta sorkuntza morfologikoa erabiliz zuzentzaile ortografiko bat bestetik.

Tresna horiek euskararen prozesaketa automatikorako egitasmo orokor baten barruan kokatzen dira. Bide horretan, eta egitasmo honi oinarria emateko, EDBL izeneko Euskararako Datu-Base Lexikala egituratu eta osatzeaz gain, corpus multzo bat bildu da, horren gainean maiztasunaren araberako hitz- eta trigrama-zerrendak lortu direlarik.

Prozesadore morfologikoa Koskenniemik definitutako bi mailatako morfologian dago oinarriturik. Formalismo honen egokitasuna frogatuta gelditu da, euskal morfologiaren deskribapen dotore, eroso eta malgua bideratzen baitu; eskala errealeko definizioa erraztuz. Egokitasuna eta malgutasuna frogatu da berriro Euskaltzaindiak Leioako 1.994ko kongresuan arau berriak onartu dituenean, oso lan sinplea izan baita sistema egokitzea arau berri hauei.

Bi mailatako formalismoaren barruan, morfemen arteko urruneko menpekotasuna adierazi ahal izateko, morfotaktikaren funtsa diren jarraitze-klaseen deskribapen-

ahalmena handitzen duten “jarraitze-klase hedatuak” izeneko mekanismoa proposatzen da.

Prozesadore morfologikoa hedadura handikoa izan dadin lau modulutan banatzen da; euskara estandarrerako modulua, erabiltzailearen lexikoa edo hiztegi berezituaren kudeaketarakoa, erabilera ez-estandarrak edo aldaerak tratatzeko modulua eta lexikorik gabeko prozesaketa morfologikoa bideratzen duena. Hizkuntza estandarrerako tratamendurako bi mailatako morfologiaren erabilera ohizkoa bada ere, gainontzeko moduluetan erabiltzea proposamen berritzailea da. Eta berritzaileena dena zera da: prozedura guztiak bi mailatako morfologian oinarriturik egotea, sistema homogeno eta trinkoa osatuz.

Bi mailatako formalismoaren gure inplementazio burutzea lan neketsua baina interesgarria izan da; alde batetik formalismoan sakontzeko balio izan duelako, eta bestetik, beraren gainean aldaketak eta hobekuntzak esperimentatzeko aukera eman digulako. Kode hau merkaturatu da Xuxen zuzentzaile ortografikoaren barruan.

Bi mailatako morfologiak, 1983an sortu zenetik, bilakaera garrantzitsua izan du, eta hobekuntza aipagarrienetako bat lexiko-itzultzaileena da, eraginkortasuna eta deskribapen-ahalmena izanik metodo horren abantailarik garrantzitsuenak. Ikerlan honetan metodo hori ebaluatu egin dugu, morfologia banatu dugun lau moduluetan emaitza azpimarragarriak lortuz, erabiltzailearen hiztegian aplikatzeko arazoak aurkitu badira ere.

Xuxen izeneko zuzentzaile ortografikoa izan da prozesadore morfologikoan oinarriturik egin dugun lehen produktu komertziala. Aurretik aipatutako modulu gehienak berrerabiltzen dira zuzentzaile honetan, horrela bi mailatako morfologian oinarritutako zuzentzaile “linguistikoa” lortuz.

Zuzenketa ortografikoaren problematika aztertu ondoren eta euskararen ezaugarriak kontuan hartuz, gaitasun-erroreak tratatzeko beharra da funtsezko ondorioa. Errore horiek detektatzeko aldaeren analisi morfologikorako erabilitako mekanismo bera erabiltzen da eta analisi horretan lortzen den informazioa gorde egiten da, ondoren akatsari dagokion zuzenketa lortzeko, sorkuntza morfologiko estandarra erabiliz helburu horrekin.

Errore tipografikoen tratamendua arras konbentzionala denez —bibliografia-azterketan azaldutako zailtasunen aurrean bigarren mailako lehentasuna baitzuen gai honek gure sisteman— zehaztasuna/eraginkortasuna faktoreen arteko orekan zentratu da gure lana, proposamenak sortzeko azkartze-metodo desberdinak aztertuz.

Azpimarratu behar da egindako tresnen izaera: eskala errealeko produktu bukatuak baitira eta ez prototipoak edo maketak. Prozesadore morfologikoa euskararen gaineko

edozein aplikaziotarako oinarrizko tresna den bitartean, zuzentzaile ortografikoa salgai dago eta oso harrera ona izan du.

VII.2. Zabaldutako ikerlerroak eta perspektibak.

Lan honetan zabaldutako ikerlerroak anitz dira. Alde batetik daude lanaren alde ahulenak hobetzeko bideak eta lanean zehar hausnartutako etorkizuneko hobekuntza posibleak. Beste aldetik daude proiektu zabalago batean integratuta egotetik datozen ikerlerroak, egindako tresnak beste urratsetan oinarri-tresna gisa erabiliko baitira. Aldez aurretik esan behar da ez zaizkigula denak berdin interesatzen, eta zenbaitetan dagoeneko lanean hasiak garen bitartean, beste batzuk aipatu besterik ez ditugu egingo.

Aurkezteko orduan lau multzotan banatu ditugu etorkizuneko lan hauek: lehenengo bietan ikerlan honekin zuzenean lotutako bi gaiak hartzen dira mintzagai, prozesaketa morfologikoa eta zuzenketa hain zuzen; hirugarrenean, epe laburrean burutu nahi dugun tresna, EUSLEM lematizatzaile/etiketatzailea, aurkezten da; eta, azkenik, egindako tresnak oinarritzat hartuko dituzten bestelako aplikazioak aipatzen dira.

VII.2.1 Prozesaketa morfologikoa hobetzen.

Prozesaketa morfologikoan lan sakona egin den arren, alde ahulena eraginkortasunarena dugu. Aipatu den bezala ahulezia hau konpontzeko aurrekonpilazioa da bide nagusia, Karttunen-ek (1994) proposaturiko lexiko-itzultzaileen bidea edo Carter-ek (1995) proposatutakoa interesgarrienak izanik. Lexiko-itzultzaileak erabili ditugu eta eraginkortasunaren zein deskribapen-ahalmenaren aldetik oso emaitza onak eskaintzen dituzten arren ez dira nahiko malguak erabiltzailearen hiztegiak integratzeko. Carter-en ekarpena interesgarria da malgutasunaren aldetik, azpilexiko itxiak baino ez baititu aurrekonpilatzen baina arazoak ditu lemen konposaketan. Beste aldetik, sistema hauetan ezin da jorratutako erregela morfofonologikoei buruz informaziorik lortu, eta hau interesgarria da aldaeren tratamendurako zein OLIren arloko aplikazioetarako. Azkenik, lexiko-itzultzaileetan morfotaktikak Koskenniemiren hasierako definizioari jarraitzen dio, urruneko menpekotasuna adierazteko deskribapen-ahalmenik gabe jarraituz. Ezaugarri horiek guztiak integratuko lituzkeen eredu berri bat definitzea irekitako ikerlerro bat da.

Euskararako aplikazioari dagokionean berriz, bi lan nagusi gelditu dira formalki landu gabe: aditz laguntzailea eta trinkoa eta eratorpena. Honez gain, datu-basea eguneratzen jarraitzea eta sistema martxan dagoen hizkuntzaren batze-prozesuari egokitzen joatea da etengabe burutzen ari den lana.

Aditz laguntzailea eta trinkoa hitzez hitz landu da, eta honen arrazoia bikoitza da; batetik eraginkortasuna, morfemak oso motzak eta aldaketak ugariak baitira, eta bestetik barneko morfemen arteko urruneko menpekotasuna. Arazo hauek konpondu ondoren aditz laguntzailearen deskribapen “formala” egingarri bihurtzen da.

Eratorpena gai korapilatsua da, irregularra izanik ondo aztertu gabe dagoelako. Taldeko linguistak egiten ari diren lan teorikoaren emaitzaz, emankortasun handiko morfema sinpleetan oinarritutako eratorpena landuko da, orain arte landutako eratorpen lexikalizatua osatuz.

VII.2.2 Zuzenketa.

Zuzenketan egindako lana aldaerak deitu ditugun gaitasun-erroreen aldetik oso interesgarria den bitartean, errore tipografikoen trataera bi aldetatik hobe liteke: bateko edizio-distantzia baino handiago duten hitzen zuzenketari ekinez batetik, eta bestetik, testuingurua kontuan hartuz, proposamenen artean zuzenketa ondo aukeratzeko zein “benetako hitzaren erroreak” detektatzeko asmoz.

Lehen ekimenean oso abiapuntu interesgarria da Oflazer eta Guzey-rena (1994), emaitza onak lortzeko oraindik eragiketa asko burutu behar badira ere, honek eraginkortasunean duen ondorio kaltegarriarekin. Ekarpen hori hobe daitekeelakoan gaude, lexikorik gabeko analisiak horretan lagun dezakeelarik.

Proposamenen artean zuzena zein den erabakitzea oso zaila gertatzen da testuingurua kontuan hartzen ez bada. Hori egiaztatzeko eskuz egitea baino ez da egin behar, testuingurua ikusi gabe pertsonek ere zalantza handiak dituztelako hitzak zuzentzeko. Testuingurua kontuan hartu gabe hobekuntza batzuk oraindik egin badaitezke ere —adib. hitzen maiztasunak kontuan hartzea, teklatuaren arabera zein aztertutako erroreen arabera aldaketei pisuak esleitzea— mugatik nahikoa gertu gaude eta testuingurua kontuan hartzea ezinbestekoa da emaitza horiek nabarmen hobetzeko, batez ere OCR eta hizketaren prozesaketarako erabili nahi bada zuzenketarako tresna hau. Gainera, testuingurua kontuan hartzen bada, benetako hitzaren erroreak detektatu eta hitz-mugaren gaineko erroreak zuzendu daitezke, bigarren belaunaldiko zuzentzailea sortuz. Testuingurua kontuan hartzeko lau bide bereizten dira nagusiki: corpus-en gaineko estatistikak, sintaxia, semantika —hitzen arteko erlazioak lortuz—, eta soluzio partikularrak. Metodo hauek konbina daitezke baina lehen hiruetarako oinarrizko lanen garapena behar da aurretik: analizatzaile sintaktiko osoa edo partziala, esanahien bilketa eta egituraketa datu-basean eta haien arteko distantzia kalkulatzeko metodoa semantikarako, eta corpusak ustiatzeko tresnak. Oinarrizko tresna hauetan ari gara lanean epe erdian zuzenketan aplikatzeko asmoz.

VII.2.3 EUSLEM.

Garatzen ari garen euskararako oinarrizko lematizatzaile/etiketatzailea da EUSLEM. Hitzaren esparrua gainditzen duenez, lan honetatik kanpo utzi dugu baina aurreratu samarra dago, aurkeztutako analisi morfologikoan oinarriturik baitago. Bere diseinurako aztertutako bibliografia azaltzen da eranskinetan, hemen azalduko dena bertako ideietan oinarritzen da eta. Tresna hau oinarrizko lanabesa izango da beste aplikazioetarako, adibidez analisi sintaktikoa, corpus-en ustiapena, dokumentuen datu-baseen indexazioa, lexikografia etab.

Hasierako lan garrantzitsu bezain korapilatsua etiketen definizioa eta analisi morfologikoarekiko egokitzapena izan da. Maila desberdinetan eratutako etiketa-sistema bat diseinatu da, oraindik ebaluatu gabe dagoena. Euskaran gertatzen den elipsiaren arazoa ere aztertu dugu bere tratamendurako proposamen bat eskainiz (Aduriz *et al*., 95).

Diseinatutako tresna hau (Aldezabal *et al*., 94) honako elementuek osatzen dute funtsean:

- Hitzak, puntuazio-karaktereak, zenbakiak etab. identifikatzen dituen aurreprozesadorea. Analisi morfologikorako egindakoan zenbait aldaketa eginez lortzen da.

- Analizatzaile morfologikoa, hitzei dagozkien lema eta etiketa posibleak zehazteko. Analizatzaile estandarra erabiltzen da lan horretan, ondoren analisiaren arabera etiketak egokitzeko prozedura burutuz.

- Analizatzaile morfologikoak ezagutzen ez dituen hitzen lema eta etiketa hipotetikoak lortzen dituen hitz ezezagunen etiketatzailea edo *guesser*-a. Horretarako, egindako aldaeren analisia eta lexikorik gabeko analisia erabili ondoren, laugarren kapituluan deskribatutako desanbiguazio lokala eta aurreko puntuan aipatutako etiketen egokitzapena burutzen da.

- Hitz anitzeko terminoen identifikazioa, horien artean lokuzioak, hitz-elkarketa eta bestelako kasu asko sartzen direlarik. Flexio handiko hizkuntzetan tratamendu honek ere berezitasunak ditu. Momentu honetan hitz anitzeko terminoekin datu-basea ari da osatzen, terminoei lau ezaugarri egokituz: (1) segurua/anbiguoa, lehen kasuan hitz banatuen analisiak baztertu ahal izateko; (2) finkoa/deklinagarria, finkoetan terminoaren identifikazioa hitzen arabera egingo den bitartean deklinagarrietan lemak identifikatu behar dira; (3) jarraian/ez-jarraian, bigarren kasuan terminoaren osagaiak sakabanatuak egon daitezke eta; (4) ordenan/ez-ordenan, azken hauen identifikazioa korapilatsuago baita.

Ezaugarrietatik ondorioztatzen denez gero, hitz anitzeko terminoen identifikazio-prozesua konplexua da oso.

- Testuinguruan oinarritutako desanbiguazioa, interpretazio morfologiko anitz duten hitz/terminoei lema/etiketa bakarra esleitzeko asmoz. Horretarako metodo desberdinak erabil daitezke: estokastikoak (Garside *et al*., 87), (De Rose, 88), (Cutting *et al*., 92), (Elworthy, 93), linguistikoak (Karlsson *et al*., 92), (Voutilainen, 94) (Tapanainen, 94) edo bion konbinaketa direnak (Leech *et al*., 94). Metodo estokastikoen bidez emaitza hobeak lortzen zirela iritzi zabalduaren aurrean, bestelako iritziak ugaldu egin dira azken urteotan (Brill, 92), (Chanod & Tapanainen, 95). Lanean ari gara bi bideak jorratzeko, tresna linguistikoaren garapenean sintaxirako ere erabiltzen den Murriztapen-Gramatika erabiliz (Karlsson, 95).

Proiektu honen oinarrian EEBS —Egungo Euskararen Bilketa Sistematikoa— (Urkia & Sagarna, 91) dago, eta bertatik lortu ditugu EUSLEMen erabiltzen diren corpus-ak. Momentu honetan testu-zati bat ari gara desanbiguatzen eskuz, geroago ezagutza-iturri gisa zein emaitzen ebaluaziorako erabiliko dena.

VII.2.4 Beste aplikazioak.

Ikusi den bezala, bi mailatako eredua oso aplikagarria da fonologian ere; beraz, hizketaren tratamenduan aplikazio zuzena izan dezake. Hizketaren tratamenduaren inguruan ari den beste ikertalde batekin elkarlanean aztertzen ari gara gure lanaren integrazioa hizketaren sorkuntzarako sistema batean.

Beste aldetik hemen azaldutako tresnak oinarrizkoak dira beste askoren eraikuntzan (ikus §l.9 atala). Analisi sintaktikoa eta itzulpenerako tresna lagungarriak daude taldeko helburu hurbilen artean.