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.

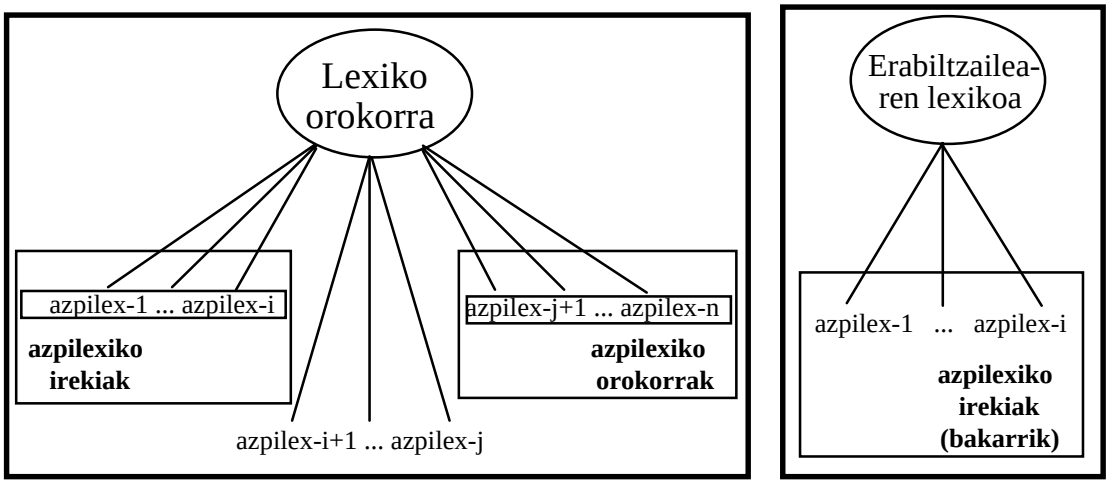
erabiltzailearen lexikoetan. Adibidez, aditz laguntzailea ondo mugaturik dago euskaraz, eta ez du erabiltzailearen lexikoan agertzeko arrazoirik.

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.

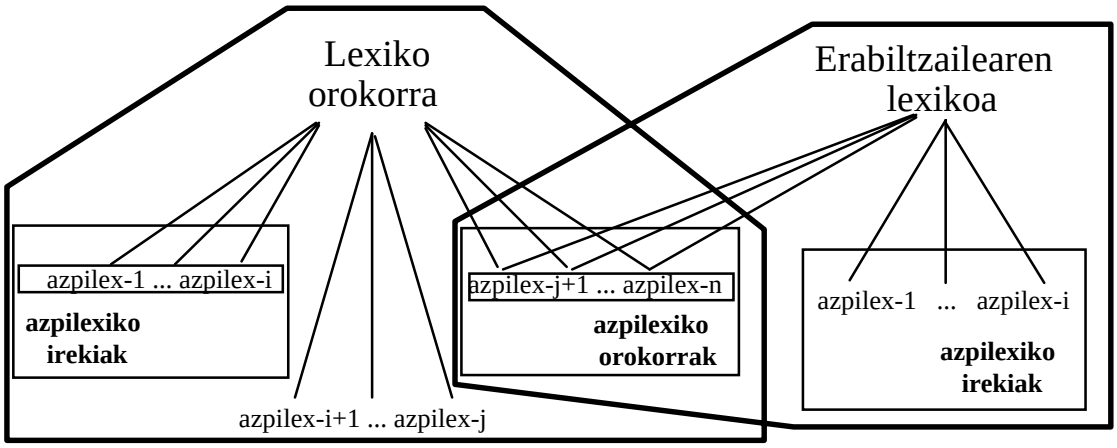
Azaldutako bi ezaugarri horiek erabiltzailearen lexikoak kudeatzeko erabiltzeaz gain, “lexikorik gabeko analisia” egiteko ere dira baliagarriak.

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

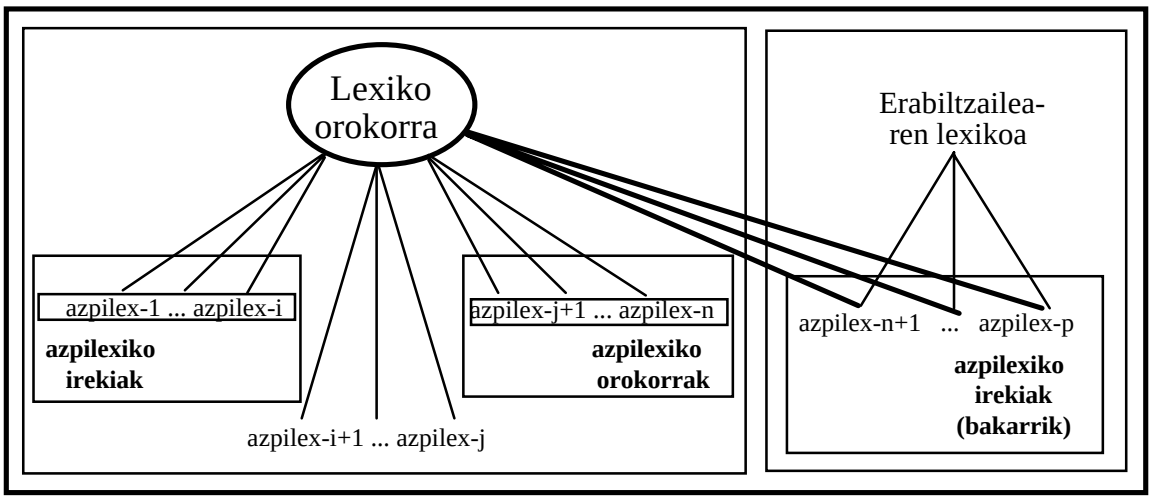
IV.1 irudia.- Lexiko orokorra eta erabiltzailearen lexikoa osatzeko modua.

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.

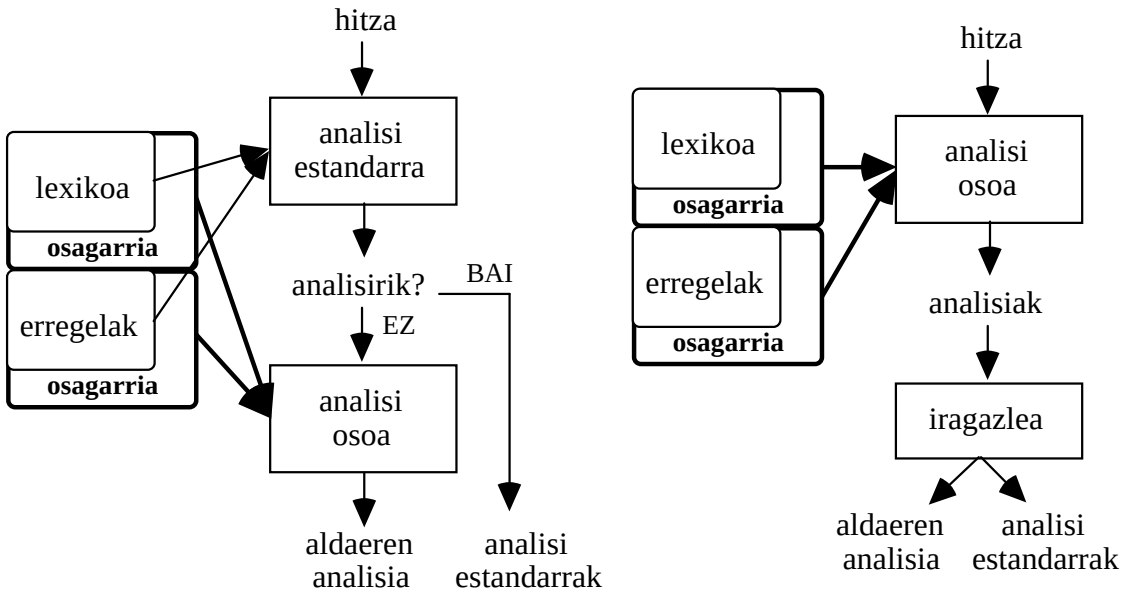
Ikus dezagun forma ez-estandarren tipifikazioa eta mota bakoitzeko analisia ebazteko modua. Hiru multzotan banatuko ditugu aurkitutako aldaerak: morfemen aldaera, morfotaktikaren aldaera, morfofonologiaren aldaera.

- **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.

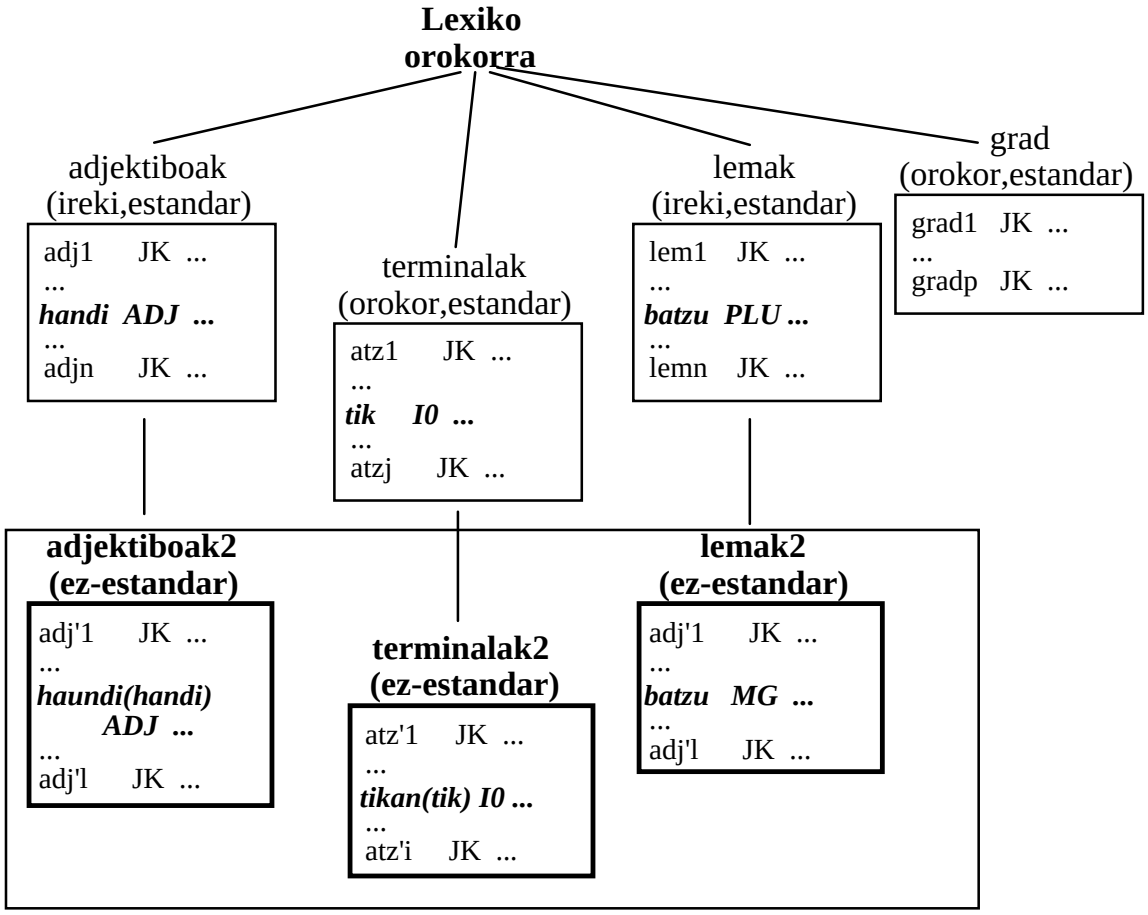
Aurrean aipatu den bezala aldaeren tratamendua bideratzeko bi mailatako formalismoaren oinarria diren lexikoari eta erregela-multzoari gehigarri bana eransten zaie: azpilexikoak eta erregelak, hurrenez hurren.

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.

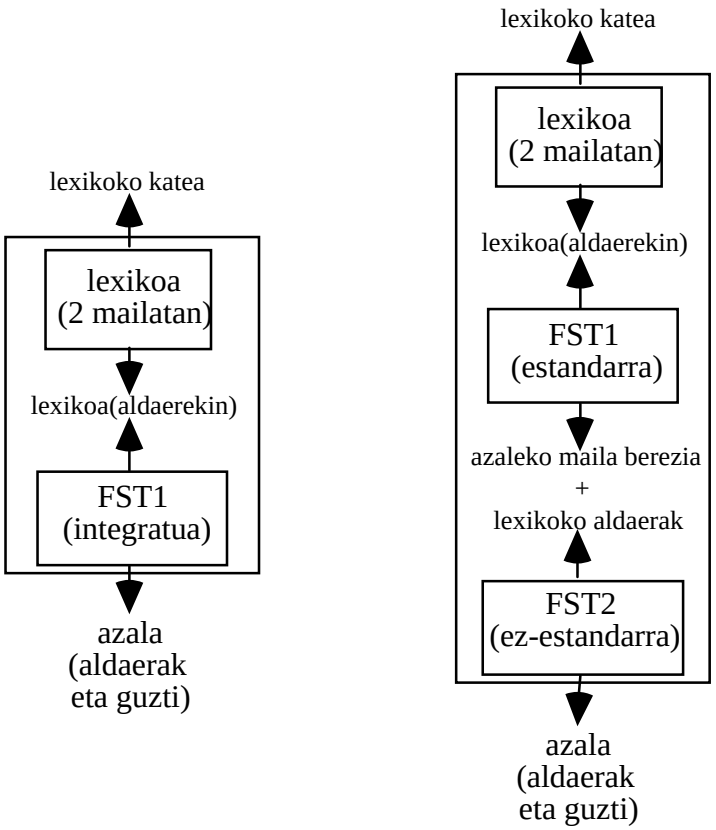
Lexikoan adierazten diren aldaeren aldetik arazorik ez dago, ezta forma estandarra lortzeko ere, lexikoan adieraz daitekeen tarteko adierazpidearen bidez lor baitaiteke. Horrela *haundi* eta *tikan* forma ez estandarren sarrerak hauek izango lirateke:

```
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.

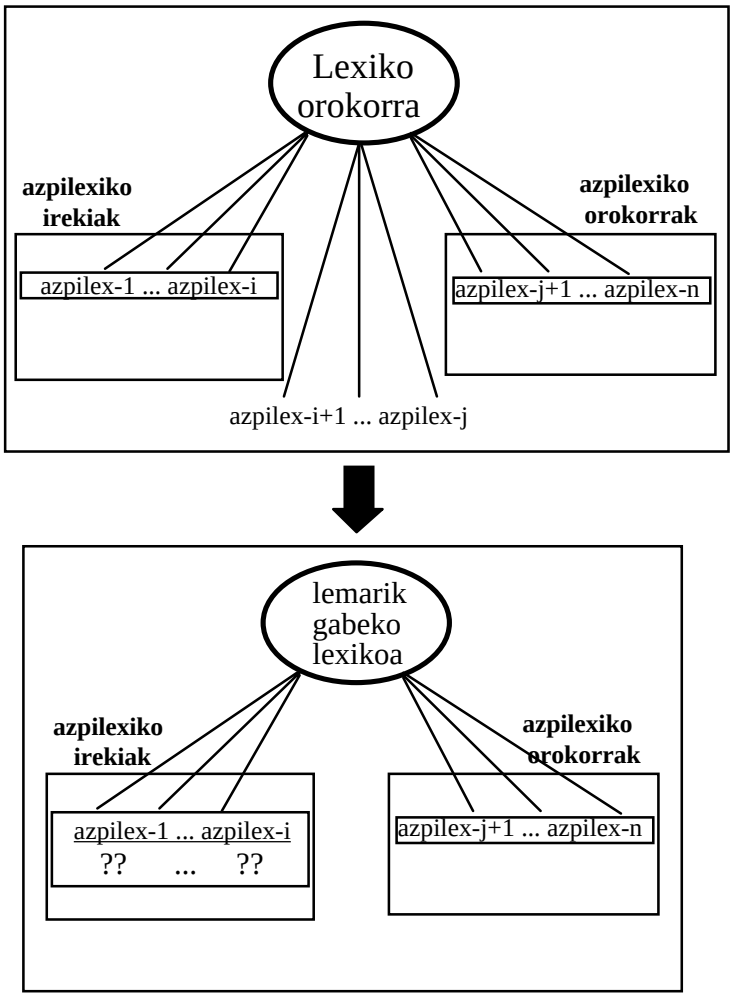
IV.3.1. Gakoa: bi mailatako erregela bereziak.

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:

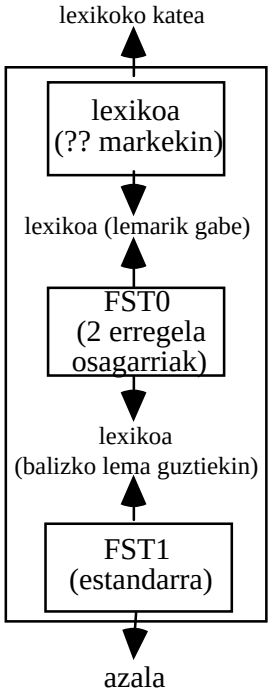
```
%?:0 => [ Hasiera | MM ] _ 0: ;
        0: MorfBuk ;
0:Cx => %?: [0: ]* _ [0: ]* %?: ;
            where Cx in (Kons Bokal) ;
```

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:

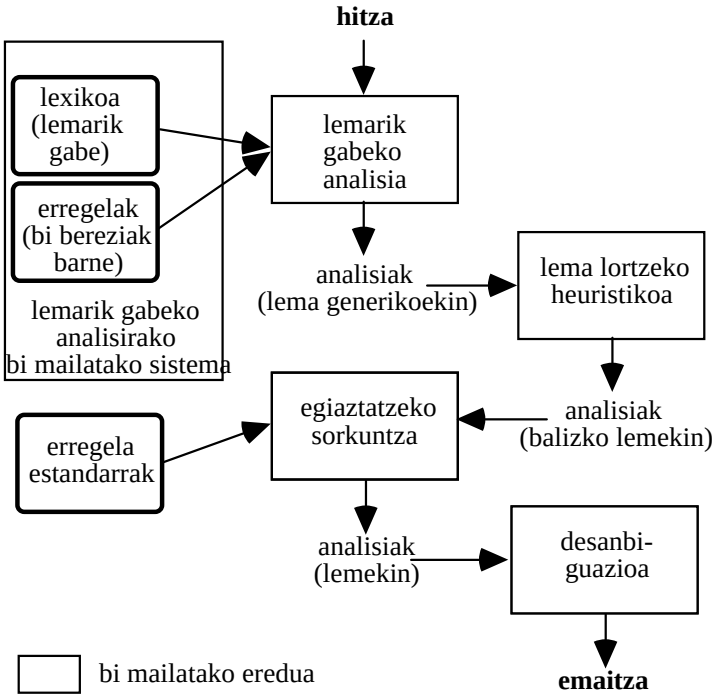
- lortzen diren analisietan agertzen den lema lema generikoa da, baina emaitza garden baterako benetako lema lortu behar da.
- analisi zilegiarekin batera beste analisi asko lortzen dira. Aukera guzti horien artean aukeratzea da desanbiguazio lokalaren helburua.

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.