

EUSKARAREN SINTAXI
KONPUTAZIONALERANTZ: oinarrizko
baliabideak eta beren aplikazioa aditzen
azpikategorizazio-informazioaren
erauzketan eta errorearen tratamenduan

Koldo Gojenola Gallettebeitia



Udako Euskal Unibertsitatea

© Koldo Gojenola Gallettebeitia
© Udako Euskal Unibertsitatea

ISBN: 84-8438-063-7

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

URL: http://www.inguma.org/tesiak/Gojenola_Gallettebeitia_2000.pdf

Nola aipatu argitalpen hau:

Gojenola Gallettebeitia, Koldo. *Euskararen sintaxi konputazionalerantz: oinarritzko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta errorearen tratamenduan* [linean]. [Bilbo: Udako Euskal Unibertsitatea], 2004. http://www.inguma.org/tesiak/Gojenola_Gallettebeitia_2000.pdf [kontsulta: Urtea/hilea/eguna]

Edizio honen ontzailea:



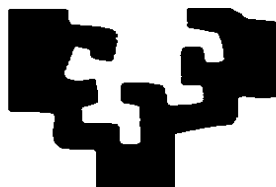
EUSKAL KOMUNITATE ZIENTIFIKOAREN DATU-BASEA

<http://www.inguma.org>
inguma@ueu.org

OHARRA: Galarazita dago dokumentu honen kopia egitea, osoa nahiz zatikakoa, edozein modutara delarik ere, Copyright-jabearen baimenik gabe. Dokumentu honen erabilera bakarra Jabego Intelektualaren legeak 31. eta 32. artikuluetan jasota dakarrena izango da.

LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA

eman ta zabal zazu



INFORMATIKA FAKULTATEA

EUSKARAREN SINTAXI KONPUTAZIONALERANTZ

Oinarrizko baliabideak eta beren aplikazioa aditzen
azpikategorizazio-informazioaren erauzketan eta errorearen
tratamenduan

Koldo Gojenola Gallettebeitiak

Informatikan Doktore titulua eskuratzeko aurkezturiko

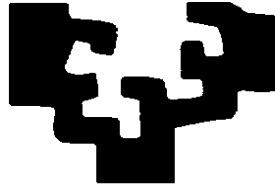
TESI-TXOSTENA

Donostia, 2000ko apirila.



LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA

eman ta zabal zazu



INFORMATIKA FAKULTATEA

EUSKARAREN SINTAXI KONPUTAZIONALERANTZ

Oinarrizko baliabideak eta beren aplikazioa aditzen
azpikategorizazio-informazioaren erauzketan eta errore
tratamenduan

**Koldo Gojenola Galletebeitiak Kepa Sarasola
Gabiolaren** zuzendaritzapean egindako tesiaren txostena,
Euskal Herriko Unibertsitatean Informatikan Doktore
titulua eskuratzeko aurkeztua.

Donostia, 2000ko apirila.

Eskerrak emanaz

Hasteko, Kepari, tesi honetan eta beste gauza askotan eman didan laguntza guztiagatik

IXA talde osoari, batez ere tesi honetan laguntza zuzena eman didatenei: Itziar, Eneko, Izaskun, Iñaki, X. Arregi, Jose Mari, X. Artola, Arantza, Nerea, Montse, Maite

Fakultateko lagun guztiei

Lekeitioko jendeari: familia eta lagunei

Encarri

```

kat perpausa
sint info kat as
    sint azpikat abs sint nag kom per 3
        num s
        kas abs
        mug m
        nor hu
            kat isk
            sar presotxipiroia
            erg sint nag kom per 3
                num s
                kas erg
                mug mg
                nrk hu
                    kat isk
                    sar huraxek
                    adizlagunak Singerren+bibotean
                    Anteroren+txamarrotean
                    gunalex beste mdl egi_ziu
                    mdn a1
                    err eduki
                    nag kat adt
                    lema dauka
                    aditz_mota nornork
twolevel
anteroentxamarrote0ansingerenbibote0anhuraxekeidaukapresotxipiroia
katea Anteroren+txamarrotean+Singerren+bibotean
+harexek+ei+dauka+preso+txipiroia

```

*Antzinako kanta herrikoia*ren ikuspegi berria

AURKIBIDEA

I PROIEKTUAREN AURKEZPEN OROKORRA	1
I.1 SARRERA GISA	1
I.2 TESIAREN EDUKI NAGUSIAK	3
I.3 TESIAREN ESKEMA ETA ARGITALPENAK	9

LEHEN PARTEA: ANALIZATZAILEAK

II ANALIZATZAILE MORFOSINTAKTIKOA	12
II.1 OINARRIZKO TRESNAK	12
II.1.1 Euskararen datu-base lexikala	12
II.1.2 Segmentatzaile morfologikoa	15
II.2 ANALISI MORFOSINTAKTIKOAREN DISEINUA	15
II.2.1 Analisi morfosintaktikorako hurbilpenak	15
II.2.2 Euskararen analizatzaile morfosintaktikoaren ezaugarriak	18
II.3 OINARRIZKO ERABAKIAK	21
II.4 GRAMATIKA MORFOSINTAKTIKOA	23
II.4.1 Erregelen adibide bat	23
II.4.2 Gramatikaren ikuspegi orokorra	26
II.4.3 Analisisien adibide bat	30
II.5 INPLEMENTAZIOA	31
II.6 LABURPENA ETA ONDORENGO PAUSOAK	32
III ANALIZATZAILE SINTAKTIKOA	36
III.1 SARRERA	36
III.1.1 Sintaxiaren tratamendu automatikoa	37
III.1.1.1 Ezagutza linguistikoa oinarritutako sintaxia	37

III.1.1.1.1 Testuingururik gabeko gramatiketan oinarritutako sistemak	41
III.1.1.1.2 Egoera finituko mekanismoetan oinarritutako sistemak	44
III.1.1.2 Teknika probabilitistikoetan oinarritutako sintaxia	48
III.1.1.3 Teknika linguistiko eta probabilitistikoen konbinazioak.....	49
<i>III.1.2 Euskararen sintaxia</i>	51
<i>III.1.3 Gure aukera</i>	55
III.2 BATERAKUNTZAN OINARRITUTAKO EUSKARAREN ANALIZATZAILEA	60
<i>III.2.1 Baterakuntza-formalismoen azterketa</i>	60
III.2.1.1 Zenbait formalismo sintaktikoren azterketa konparatiboa eta aplikazioa.....	60
III.2.1.2 Guneak zuzendutako egitura sintagmatikoen gramatika eta euskararako aplikazioa	62
III.2.1.3 Ondorioak	64
<i>III.2.2 Baterakuntzan oinarritutako analizatzaile sintaktikoa</i>	65
III.2.2.1 Oinarritzko erabakiak	66
III.2.2.2 Gramatikaren deskribapena	68
III.2.2.3 Gramatikaren aberasketa: postposizioak	76
III.2.2.4 Morfosintaxia eta sintaxiaren arteko muga	77
III.2.2.5 Inplementazioa	79
III.2.2.6 Ondorioak	81
III.3 EUSKARAREN EGOERA FINITUKO SINTAXI-TRESNAK	82
<i>III.3.1 Euskararen murriztapen-gramatika</i>	82
III.3.1.1 Murriztapen-gramatika aplikatzeko prozedura.....	84
III.3.1.2 Euskararen aplikazioaren emaitzak	86
III.3.1.3 Murriztapen-gramatikaren balorazioa	88
<i>III.3.2 Egoera finituko sintaxia (XFST)</i>	91
III.3.2.1 Sarrera	91
III.3.2.2 Eragiketa nagusiak	91
III.3.2.3 Euskararen egoera finituko analizatzaile sintaktikoa	94
III.4 FORMALISMOEN KONPARAZIOA ETA INTEGRAZIOA	95
<i>III.4.1 Formalismoen konparazioa</i>	95
<i>III.4.2 Formalismoen integraziorako ereduak</i>	98
III.4.2.1 Tresnen aplikazio sekuentziala.....	99
III.4.2.1.1 Murriztapen-gramatikatik baterakuntzan oinarritutako analizatzaile sintaktikora	101
III.4.2.1.2 Baterakuntzan oinarritutako analizatzaile sintaktikotik egoera finituko sintaxira	102
III.4.2.2 Formalismoen konbinaziorako beste aukera batzuk.....	105
<i>III.4.3 Irteeraren TEI deskribapena</i>	106
III.5 ONDORIOAK	109

BIGARREN PARTEA: APLIKAZIOAK

IV AZPIKATEGORIZAZIO-INFORMAZIOAREN ERAUZKETA	113
IV.1 BESTE LAN BATZUEN AZTERKETA.....	114
IV.2 ADITZEN AZPIKATEGORIZAZIO-INFORMAZIOAREN ERAUZKETARAKO TRESNA	117
IV.2.1 <i>Problemaren zehaztapena</i>	117
IV.2.1.1 Aditzen azpikategorizazioaren eremu teorikoa	118
IV.2.1.2 Lortu nahi den emaitzaren definizioa.....	119
IV.2.1.3 Batakuntzan oinarritutako analizatzailearen irteera	121
IV.2.2 <i>Tresnaren diseinua eta garapena</i>	123
IV.2.2.1 Inplementazioari buruzko datuak.....	128
IV.2.3 <i>Lehen emaitzak</i>	129
IV.3 ONDORENGO URRATSAK.....	133
V EZAGUMENDU SINTAKTIKOAREN ERABILERA ERROREEN DETEKZIOAN ETA	
ZUZENKETAN	131
V.1 SARRERA	131
V.1.1 <i>Errore motak</i>	132
V.1.2 <i>Erroreen detekziorako zenbait sistemaren azterketa</i>	135
V.2 ERROREEN DETEKZIOAN ETA ZUZENKETAN EGINDAKO ESPERIMENTUAK	138
V.2.1 <i>Euskarazko testuetako erroreen sailkapena</i>	138
V.2.2 <i>Murritzapen sintaktikoen erlaxazioa</i>	141
V.2.2.1 Metodoaren azalpen laburra	141
V.2.2.2 Egindako esperimenduak	143
V.2.2.3 Ondorioak	149
V.2.3 <i>Errore-patroien bidezko detekzioa</i>	150
V.2.3.1 Sarrera	150
V.2.3.2 Corpusetan oinarritutako patroien bidezko erroreen detekzioa	152
V.2.3.3 Ondorioak	156
V.2.4 <i>Errore ortografikoen zuzenketa</i>	157
V.2.4.1 Sarrera	157
V.2.4.2 Errore ortografikoen zuzenketa automatikoa	159
V.2.4.2.1 Erabilitako teknikak.....	159
V.2.4.2.2 Esperimenduak	161
V.2.4.3 Ondorioak	166
V.3 ERROREEN DETEKZIO ETA ZUZENKETARI BURUZKO LANEN ONDORIOAK ETA HURRENGO	
PAUSOAK	167
VI BESTE APLIKAZIOAK	173
VI.1 EUSLEM.....	173
VI.2 IKASLEEN TESTUEN EGITURA SINTAKTIKO OROKORRAK AZTERTZEKO TRESNA.....	176

VI.3 ONDORIOAK.....	178
VII TESIAREN ONDORIO NAGUSIAK ETA ETORKIZUNERAKO IKERLERROAK	179
VII.1 LORTUTAKO EMAITZAK.....	179
VII.2 ZABALDUTAKO IKERLERROAK ETA PERSPEKTIBAK	181
<i>VII.2.1 Tratamendu morfosintaktikoaren jarraipena</i>	<i>182</i>
<i>VII.2.2 Tratamendu sintaktikoaren jarraipena.....</i>	<i>182</i>
<i>VII.2.3 Erroreen tratamendurako lanen jarraipena</i>	<i>183</i>
BIBLIOGRAFIA	185

SARRERA ETA AURKEZPEN OROKORRA

I Proiektuaren aurkezpen orokorra

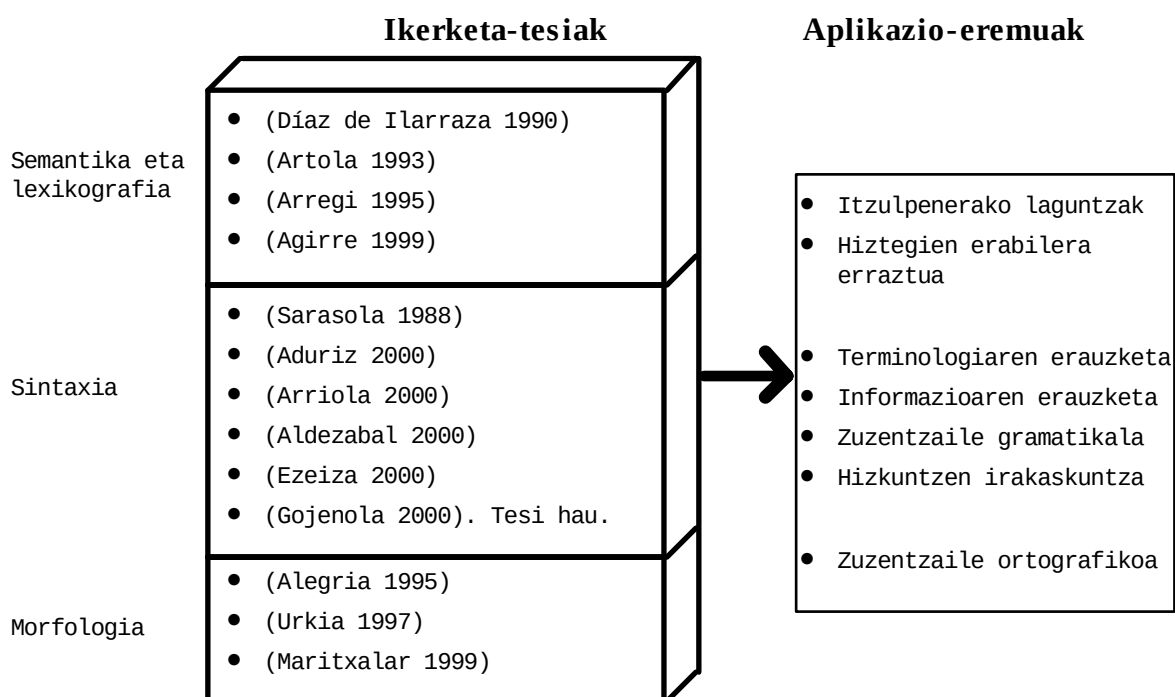
I.1 Sarrera gisa

Tesi hau Lengoaia Naturalaren Prozesamenduaren (LNP) barruan kokatzen da, eta bere helburua euskararen sintaxi konputazionalaren ikerketa zabaltzea da. Horretarako, beste hizkuntzen tratamendurako landu diren formalismo eta tresna linguistiko eta informatikoak aztertu dira, euskararen tratamendurako bide egokienak bilatzeko eta bukaeran euskararen baliabide linguistikoak eta tresna erabilgarriak inplementatzeko.

Hasierako helburu hori —*euskararen sintaxi konputazionalaren ikerketa zabaltzea*— hobeto zehaztea beharrezkoa da, sintaxi terminoak hizkuntzaren eremu oso zabala adierazten baitu. Horregatik, ohar hauek eman behar ditugu:

- Euskararen sintaxiaren tratamenduaren *lehen urratsak* direla esan behar dugu. Sintaxia beste hizkuntzetarako ere oraindik ikerketarako arlo emankorra da eta euskarari maila horretara igotzeko bide luzea geratzen zaio. LNPan hizkuntza zabalduenak (ingeleza, alemana, ...) bakarrik landu dira orain arte, eta horregatik euskara bezalako hizkuntzak ere aztertu beharra dago, bai fenomeno berrien tratamenduari ekitea eskatuko duelako, bai aztertutako hizkuntz multzoa handitzen joango den heinean LNP osoaren aurrerakada ekarriko duelako. Lehen urrats hauek euskararen sintaxi konputazionalari buruz egin diren edo egiten ari diren beste lan batzuen testuinguruan ikusi behar dira, ez baita hau euskararen sintaxiaren inguruan martxan dagoen azterketa bakarra (Abaitua 1988, Aduriz 2000, Aldezabal 2000, Arriola 2000, Ezeiza 2000).

- Sintaxi *konputazionala* aipatu dugu, hizkuntzaren deskribapen teorikoaz gain lan honen beste helburu bat euskara prozesatzeko tresna erabilgarriak eta aplikazioak garatzea delako. Lengoaiaren prozesamenduan egin diren lan askoren ideia hizkuntzaren deskribapena egiteko formalismoen inplementazioa izan da, indarra hizkuntzaren teoriaren garapenean jarritz, linguistika teorikotik hurbilago eta lengoiaia analizatzeko tresnen eraikuntza bigarren mailan utziz. Beste lan batzuek, aldiz, garrantzi handiagoa eman diote hizkuntzaren ingeniartzari, hau da, eguneroko lengoiaia (testuak, ahotsa) tratatzeko tresnak egiteari, teoria linguistikoak kontuan hartu gabe. Gure kasuan, bi alde hauek konbinatu nahi izan ditugu, euskarak eguneroko bizitzan erabiliko diren tresnak behar dituelako, baina jakinda, hala ere, deskribapen linguistiko egokirik gabe tresnen epe luzerako apustua galdu egingo litzatekeela.



I.1 irudia. IXA taldeko ikerketa-lanak.

Tesi hau ondo ulertzeko garrantzizkoa da lana IXA ikertaldearen barruan kokatzea. IXA taldeak hamar urte baino gehiago darama euskararen tratamendu informatikoan lanean. Taldeak zenbait aplikazio eta tresna sortu ditu, horien artean aipagarrienak euskararen datu-base lexikala (EDBL) eta XUXEN zuzentzaile ortografikoa. I.1 irudiak taldearen ikerketalerroen ikuspegia eman nahi du, horretarako egindako eta bidean dauden tesi-lanak aipatuz, tesi horiek taldeak jarraitzen dituen pausoen isla direlako.

Irudian agertzen ez den arren, oinarri lexikala ere badugu. Aplikazio guztiek erabiliko dute lexikoia eta horregatik ezinbestekoa izan zaigu aplikazioekiko independentea eta oinarri

linguistiko sendokoa izatea. Lexikoiaren gainean agertzen den lehen tratamendua hitzaren analisisa da. Euskararen morfologiari buruzko deskribapena eta inplementazioa egin dira Alegria (1995) eta Urkia-ren (1997) tesietan. Horien emaitzak analizatzaile morfologikoa eta bera oinarria duen zuzentzaile ortografikoa izan dira. Maritxalar-ek (1999) bigarren hizkuntzaren irakaskuntzarako sistema baten diseinua egin zuen alde morfologikoa landuz. Sintaxiaren eremu zabala dela eta, lerro desberdinak ireki dira bere tratamenduan, horien artean problemen enuntziatuen analisirako sistema (Sarasola 1988, gazteleraz idatzitako enuntziatuak analizatuz), euskararen murriztapen-gramatika (Aduriz 2000), hiztegiko definizioen analisi sintaktikoa murriztapen-gramatikaren bidez, azpikategorizazioari buruzko informazioaren erauzketan (Arriola 2000), aditzen azpikategorizazioaren azterketa (Aldezabal 2000), corpusen ustiaketarako tresnen garapena (Ezeiza 2000) eta tesi honetan azalduko duguna. Semantika eta lexikografiaren tratamenduarekin ere hasia da taldea, Díaz de Ilarraza (1990), Artola (1993), Arregi (1995) eta Agirre-ren (1999) tesiekin eta ondoren garatzen ari diren beste proiektuekin.

IXA taldearen jardunak argi uzten du zein garrantzitsu den hizkuntzalari eta informatikarien arteko koordinazioa. Horregatik, diziplina arteko koordinazioa behar izan da taldeko eginkizun guztietan eta, noski, horrek isla izan du tesi honetan ere, sintaxia lantzeko hizkuntzalari eta informatikarien lankidetzara aberatsa eta emankorra izan dugulako.

Tesi hau bere testuinguruan kokatu eta gero, § I.2n aurkeztuko dira landu ditugun aspektu nagusiak, analisi morfologiko eta sintaktikoa egiten duen sistemaren arkitektura eta bere aplikazio-erabilpenak. Kapitulu honetako azken atalean (§ I.3) tesi-txostenaren eskema azalduko da.

I.2 Tesiaren eduki nagusiak

Sintaxia era konputazionalen tratatzeko bidean, hauek izan dira lan honetan jorratu ditugun aspektuak:

- a) Morfosintaxiaren tratamendua. Alegria eta Urkia-ren lanetan morfologiaren lehen urratsa eman da: hitz-forma eta bere lexikoiko osagaien (lema eta morfemak) arteko korrespondentzia ebaztea, sorkuntzan zein analisisan. Horrela, hitz-forma bat emanda bera osatzen duten morfemak atera daitezke, bakoitza berari dagokion informazio morfosintaktikoarekin (kategoria, kasua, numeroa, mugatasuna, ...). Dena dela, prozesu hori hitzaren analisiaren lehen pausoa besterik ez da, morfema horiek eta

beraiek dakarten informazioa era askotan konbina daitezkeelako hitz osoa sortzeko. Hitz osoaren informazioa ez da lortzen beti bere osagaien informazioaren metaketa hutsez. Prozesu honen deskribapenari *hitzaren morfosintaxia* deitzen zaio. Termino horrek hitzaren barruko egitura sintaktiko konplexua adierazten baitu, hizkuntza eranskarietan hitzaren barruan morfologia eta sintaxia agertzen direlako. Tesi honen II. kapitulu

b) Sintaxiaren tratamendua. Hitzen analisi morfosintaktikoa egin eta gero, hurrengo pausoa sintaxia da. Euskara bezalako hizkuntzetan hitz-mailan gertatzen diren zenbait fenomeno beste hizkuntza batzuetan sintaktikotzat hartzen dira eta, beraz, esan liteke sintaxiaren tratamendua dagoeneko hasita dagoela analizatzaile morfosintaktikoaren bidez. Horregatik sintaxia morfosintaxiaren jarraipen gisa ikusi dugu. Bide horretan, formalismo edo hurbilpen bakar

- Murritzapen-gramatikaren formalismoa. Formalismo hau anbiguotasuna ebazteko landu zen bereziki, LNParen arazo nagusietako bat delako. Horregatik, euskararen hitz-mailako anbiguotasun-tasa altua kentzeko desanbiguazio-gramatika landu da (§ III.3.1en). Aduriz (2000) eta Arriola-ren (2000) tesi-lanak harantzago joan dira sintaxiaren eremuan murritzapen-gramatikaren eskutik, desanbiguazioaz gain osagai sintaktikoen bereizketan eta funtzio sintaktikoen esleipena egitera ere heldu direlako, baina tesi honetan ez dugu landu ildo horretatik lortzen den emaitza, eta desanbiguazio hutsarekin geratu gara.
- Baterakuntzan oinarritutako formalismoak. Hauek erabiltzeko arrazoi nagusien artean Shieber-ek (1986) erazagutzailetasuna eta ahalmen deskriptiboa aipatzen ditu. Arrazoi horiengatik aukeratu zuen Abaitua-k (1988) multzo honetako LFG formalismoa euskararen zenbait fenomenoren deskribapen formala egiteko. Arrazoi berak kontuan hartuta, baterakuntza sintaxian gertatzen diren fenomeno konplexuak deskribatzeko egokia dela ikusi dugu, eta horregatik euskararen baterakuntzan oinarritutako gramatikaren garapena PATR formalismoan egin dugu. Abaitua-ren lanarekin alderatuz gero, gure lanaren diferentzia nagusiak lexikoi handia erabiltzea (EDBL) eta testu erreale

direla eta, gure gramatikak ez ditu barruan hartuko testu errealetako esaldi guztiak, aukeratu ditugun oinarritzko hainbat egitura sintaktiko baino. Corpusetako maiztasunaren arabera eta EDBLn gaur egun dagoen informazioarekin tratagarriak diren osagaien arabera aukeratu dira ezagutuko diren egitura sintaktikoak. Gramatika partziala izango da, beraz, baina estaldura zabalekoa era berean, perpausetako osagai nagusiak harrapatzen dituelako eta oso lexiko zabala darabilako. Bere deskribapena § III.2n egingo da.

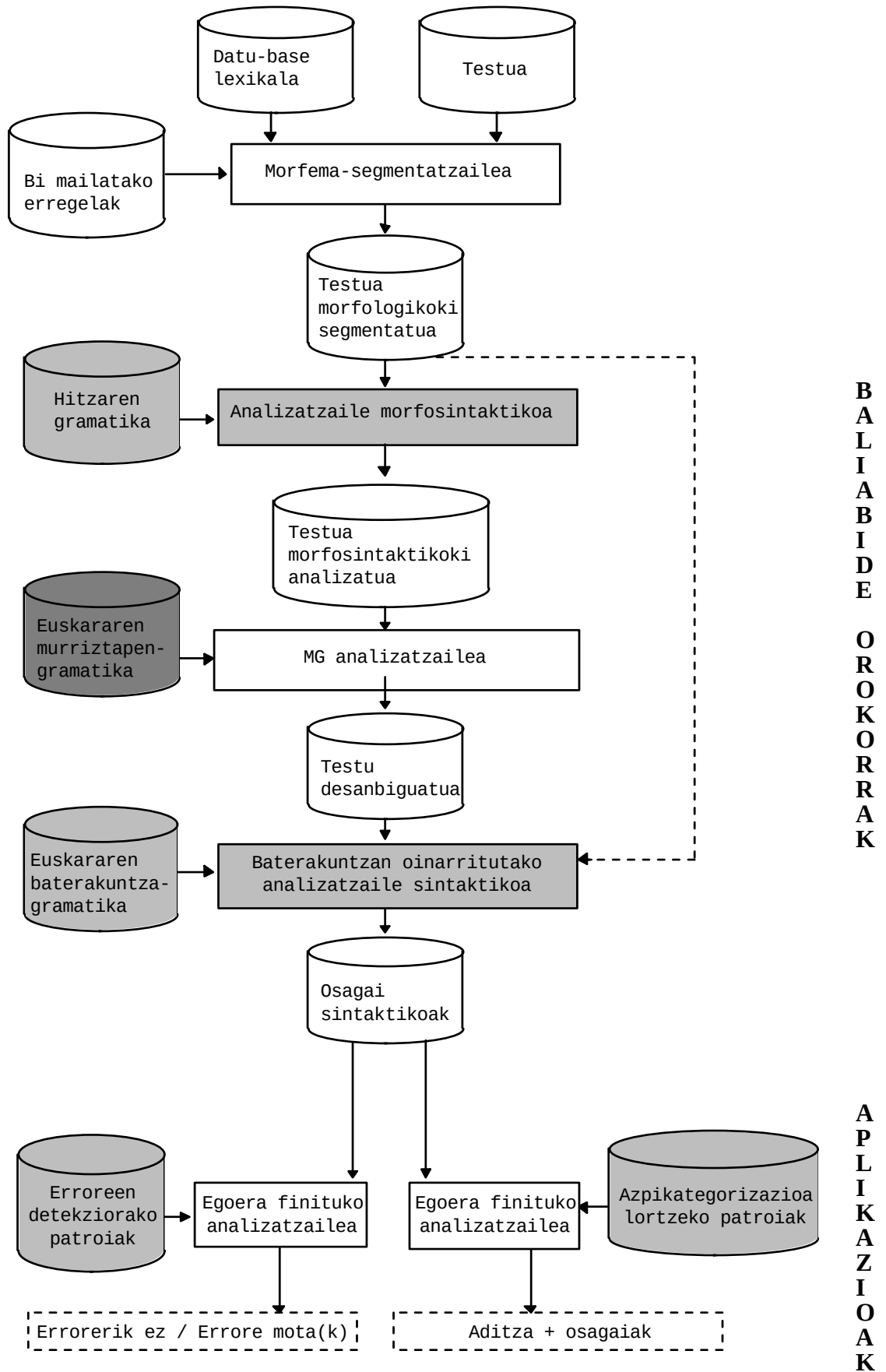
- Egoera finituko sintaxia. Murritzapen-gramatika eta baterakuntzan oinarritutako formalismoen mugak kontuan hartuta, adierazpen erregularren bidezko eredua ere aukeratu dugu. Eredua honetan, sintaxi osoa landu gabe patrietan oinarritutako prozesuak deskriba daitezke, beste formalismoekin lortu diren emaitzen gainean lan egiteko. Desanbiguazioa eta osagai sintaktiko konplexu berrien sorkuntza burutu daiteke hirugarren formalismo honekin (§ III.3.2).

c) Aplikazioen garapena. Tesi honen helburua euskararen sintaxiaren tratamendu automatikorako oinarritzko baliabideen garapena egitea da, berauek erabiliz aplikazioak lortzeko. Horregatik tesiaren bigarren fasea lortutako tresnak zenbait aplikaziotan probatzea izan da:

- Azpikategorizazio-informazioaren erauzketa. Landu diren baliabide sintaktikoak zenbait aplikaziotarako erabilgarriak diren arren, mugatuak dira neurri handi batean EDBLko informazioan hutsuneak daudelako. Horien artean garrantzitsuenetako bat aditzen azpikategorizazioari buruzko informazio-eza da. Arazo hau dela eta, tresna sintaktikoen lehen aplikazioa corpusetatik informazio hori era automatikoan ateratzen saiatzea izan da, horrela baliabide lexikal eta sintaktikoak aberastu ahal izateko gero. Lan hau IV. kapituluaz azalduko da.
- Ezagutza sintaktikoaren erabilpena erroreen tratamenduan. Morfologiaren tratamenduak zuzenketa ortografikoaren bidea ireki zuen bezala, antzera gertatu da tresna sintaktikoen garapenarekin. Gure kasuan, erroreen tratamenduaren hiru arlotan probatu dugu sintaxiaren ekarpena. Lehenengoan, baterakuntza-gramatikaren murritzapen sintaktikoen erlaxazioak aukera emango du komunztadura-erroreen detekzioan. Bigarren, egoera finituko patri sintaktikoen bidezko erroreen detekzioa probatu dugu. Azkenik, zuzentzaile ortografikoen proposamenen arteko hautaketa automatikoa egiteko

zenbait ezagutza-iturriren ekarpena aztertu dugu. Horiei guztiei buruzko esperimentuak V. kapituluaz azalduko dira.

- VI. kapituluaz beste aplikazio batzuk azalduko dira. Alde batetik, analizatzaile morfosintaktikoa lematizatzaile/etiketatzaile batean erabiliko da. Azken aplikazioa bigarren hizkuntzako ikasleen testuen azterketan kokatu dugu.



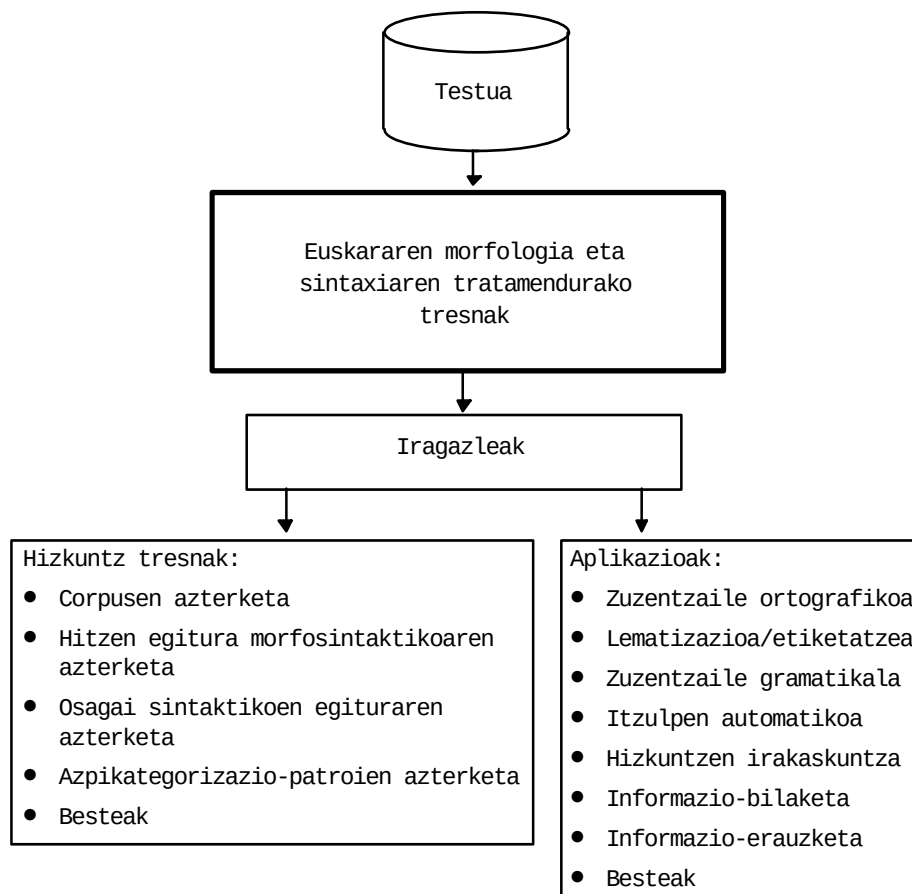
I.2 irudia. Euskararen analisi morfologiko eta sintaktikorako sistemaren arkitektura.

I.2 irudian euskararen analisi morfologiko eta sintaktikorako arkitektura aurkezten da. Bertan tesi honetan egindako lanak kolore grisez agertzen dira. Hasiera batean, EDBL eta segmentatzaile morfologikoaren bidez testuetako hitzen segmentazioa lortzen da, eta hitz bakoitza bere lema eta morfemetan zatitzen da. Segmentazio hau bi prozesuren abiapuntua izango da: tratamendu morfosintaktikoa eta tratamendu sintaktikoa, biek morfema oinarri gisa hartzen dutelako. Tratamendu morfosintaktikoak hitzen analisi osoa lortzen du, eta hau izango da murriztapen-gramatikaren bidezko desanbiguazio-prozesuaren sarrera, emaitzan hitzak ia guztiz desanbiguatuta uzteko (hau da, ia interpretazio bakarra hitzeko). Murriztapen-gramatikan egin dugun lana bere aplikazioaren alde informatikoan izan da, lan gramatikal nagusia beste tesi batzuetan (Aduriz 2000, Arriola 2000) egin delako. Horregatik irudian murriztapen-gramatika gris ilunez koloreztatu da, garapena lankidetzan egin dugula adierazteko. Morfemetatik abiatuko da baterakuntzan oinarritutako analizatzaile sintaktikoa. Horretarako desanbiguazioaren emaitzaz balia daiteke, aukera dezente baztertuz. Analizatzaile sintaktikoak ez ditu esaldi guztien analisi osoak lortuko, baina bai analisi partzialak, esaldiko osagai nagusiak barne egongo direla. Emaitza honetan osagai sintaktikoak daude, baina tresna bat behar da osagai horietatik aplikazio bakoitzak behar dituenak aukeratzeko, eta horretarako erabili da egoera finituko analizatzailea. Irudian orain arte landu diren bi aplikazio nagusiak azaltzen dira. Batetik, adierazpen erregularren bidezko gramatika bat definitu da osagai sintaktiko guztietatik aditzak beren inguruko elementuekin ateratzeko, modu horretan aditzen azpikategorizazio-patroiei buruzko informazioa lortzeko. Bestetik, gramatika baten bidez errore sintaktikoen patroien testuinguruak aztertu dira, errorearen agerpenak detektatzeko.

Sistema osoaren konplexutasuna dela eta, erabaki estrategiko bat hartu da arkitektura osoaren diseinuan: azpisistemen arteko interfazea markaketa-lengoaia formal batera egokitzearena. SGML lengoaia aukeratu da horretarako, Text Encoding Initiative (TEI) taldeak sortutako gidalerroak jarraituz. Era horretan tresna linguistikoen arteko komunikazio-arazoak ekidin nahi dira, markaketa-lengoaia formalizatu baten bidezko kodeketak formatuen arazo asko gainditzeko balio izango duelako.

I.2 irudian morfologia eta sintaxiaren prozesamendurako sistema konplexuaren barruko guztiak erakutsi dira. I.3 irudian sistema osoak kutxa beltz gisa ikusita eskain dezakeen informazioa erakutsi da. Bertan sintaxi eta morfologiarako tresnen erabilpenak azaldu nahi dira. Erabilpen horiek bi multzo handitan sailkatu ditugu. Lehenengo, tresnak baliagarriak izango dira hizkuntzalarien ikerketarako: corpusen azterketak egiteko, maiztasunak

ateratzeko, edo hainbat osagai sintaktiko eta morfosintaktiko erakusteko. Bigarren, tresna horiek LNPko aplikazioen oinarri izango dira.



I.3 irudia. Euskararen analisi morfologiko eta sintaktikorako tresnen erabilera.

I.3 Tesiaren eskema eta argitalpenak

Txosten hau bi zati nagusitan banatu dugu. Sarrera honen ondoren lehenengo partea dator (II. eta III. kapituluak), analisi sintaktikorako egindako oinarritzko baliabideen deskribapenarekin. Bigarren partean (IV., V. eta VI. kapituluak) tresna horiek erabiliz landutako aplikazioak azalduko dira. Bukatzeko, VII. kapituluan lanaren ondorioak eta etorkizunean aurreikusten diren lerroak emango dira.

Kapituluak banan-banan aztertuko ditugu orain. II. kapituluak euskararen morfosintaxiaren tratamendurako gramatika eta analizatzailearen garapenaren berri emango du. Morfosintaxiaren jarraipena sintaxia denez, III. kapituluan syntaxirako landu ditugun hurbilketen deskribapena emango da: euskararen murriztapen-gramatika, baterakuntzan oinarritutako gramatika eta egoera finituko patroien bidezko analisi sintaktikoa. Honekin oinarritzko baliabideen parteari emango zaio bukaera.

IV. kapituluak syntaxirako tresnen aplikazio bat garatuko du: corpusetarik aditzak eta eurekin doazen osagai sintaktikoen erabilera-adibideen erauzketa. Horrela azpikategorizazio-patroien bilketa automatikoa egiteko bidea irekitzen dugu.

V. kapituluak tresna sintaktikoen beste aplikazio bat aztertuko du: ezagutza sintaktikoaren erabilera errore ortografiko eta sintaktikoen tratamenduan. Erroreen tratamendurako mundu zabaletik hiru eremu aukeratu ditugu syntaxiaren ekarpena esperimendatzeko: alde batetik, baterakuntza-gramatikaren erabilera gramatikaren erregelen kontra egindako erroreak harrapatzeko (komunztadurak, adibidez); beste batetik testuinguru erroredunak deskribatzen dituzten errore-patroien bidezko detekzioa; eta bukaeran, zuzentzaile ortografikoek sortutako proposamenen arteko diskriminazioa, sintaktikoaz gain beste ezagutza mota batzuk ere erabiliz.

VI. kapitulan oinarritzko tresnen beste bi aplikazio azalduko dira: analizatzaile morfosintaktikoaren integrazioa EUSLEM lematizatzaile/etiketatzailean, eta tresna sintaktikoen erabilera bigarren hizkuntzaren irakaskuntzarako sistema batean, ikasleek erabilitako egitura sintaktikoen maiztasunak ateratzeko.

VII. kapituluak tesi honetatik ateratako ondorio nagusiak eta etorkizunean lan egiteko ikerlerroen aurkezpena emango ditu.

Sarrera-kapitulu honi bukaera ematen dioten I.1 eta I.2 taulek tesi honekin lotutako argitalpenen berri ematen dute, argitalpen bakoitza dagokion kapituluarekin lotuz. Argitalpenak IXA taldeko orritik jaso daitezke (<http://ixa.si.ehu.es/dokument/artiksail.html>).

Egileak	Argitalpena
Gojenola eta Sarasola 1994	Aplicaciones de la relajación gradual de restricciones para la detección y corrección de errores sintácticos.
Aduriz <i>et al.</i> 1995	Different Issues in the Design of a Lemmatizer/Tagger for Basque.
Aduriz <i>et al.</i> 1997	Morphosyntactic Disambiguation for Basque based on the Constraint Grammar Formalism.
Agirre <i>et al.</i> 1998ab	Towards a Single Proposal in Spelling Correction.
Aldezabal <i>et al.</i> 1998	Subcategorización verbal vasca: propuesta inicial y herramienta de validación.
Gojenola 1998	Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG) eta euskararako aplikazioa.
Aduriz <i>et al.</i> 1999	MORFEUS: Euskararako analizatzaile morfosintaktikoa.
Aldezabal <i>et al.</i> 1999	Combining Chart-Parsing and Finite State Parsing.
Aduriz <i>et al.</i> 2000a	A Word-Grammar Based Morphological Analyzer for Agglutinative Languages.
Aduriz <i>et al.</i> 2000b	A Word-Level Morphosyntactic Analyzer For Basque.
Aldezabal <i>et al.</i> 2000	A Bootstrapping Approach to Parser Development.
Gojenola eta Oronoz 2000	Corpus-Based Syntactic Error Detection Using Syntactic Patterns.

I.1 taula. Tesiarekin lotutako argitalpenak.

Kapitulua	Argitalpena
II. Analizatzaile morfosintaktikoa	Aduriz <i>et al.</i> 1999 Aduriz <i>et al.</i> 2000ab
III. Analizatzaile sintaktikoa	Aduriz <i>et al.</i> 1995, 1997 Gojenola 1998 Aldezabal <i>et al.</i> 1998, 1999, 2000
IV. Azpikategorizazio-informazioaren erauzketa	Aldezabal <i>et al.</i> 1998, 1999, 2000
V. Erroreen tratamendua	Gojenola eta Sarasola 1994 Agirre <i>et al.</i> 1998ab Gojenola eta Oronoz 2000
VI. Beste aplikazioak	Aduriz <i>et al.</i> 1995

I.2 taula. Kapitulu bakoitzarekin lotutako argitalpenak.

LEHEN PARTEA: ANALIZATZAILEAK

II Analizatzaile morfosintaktikoa

Euskararen sintaxiaren azterketa hasteko orduan ikusten dugu hizkuntza eranskaria izanik beste hizkuntzetan esaldi- edo sintagma-mailan gertatzen diren fenomenoak hitz-mailan agertzen direla. Hitz bat osatzen duten lema eta morfemek informazio morfosintaktikoa dakarte, eta bakoitzak bere ekarpena izango du hitzaren osaketan. Honengatik sintaxiaren lehen pausoak emateko hitzaren analisi morfosintaktikoa aukeratu dugu: morfemetatik abiatuta hitz osoaren analisisa lortzeko mekanismoen deskribapena eta inplementazioa.

Kapitulua hasteko, § II.1en tesi honetako lanen oinarri diren tresnak azalduko dira: euskararen datu-base lexikala eta segmentatzaile morfologikoa. § II.2n analizatzaile morfosintaktikoaren diseinuan kontuan izan behar diren faktoreak aztertuko dira, hurrengoan (§ II.3) gramatika morfosintaktikoa egiteko erabaki linguistikoekin jarraitzeko. § II.4 eta § II.5en lortutako gramatika osoaren azalpena emango da, inplementazioaren xehetasunekin batera. Bukatzeko (§ II.6), ondorio nagusiak eta etorkizunerako lerroak emango dira.

II.1 Oinarrizko tresnak

II.1.1 Euskararen datu-base lexikala

Aplikazio askotarako oinarri lexikala dugu Euskararen Datu-Base Lexikala (EDBL, Agirre *et al.* 1994, Aduriz *et al.* 1998a, Aldezabal *et al.* 1999b). Hiru atal nagusi ditu: hiztegi-sarrerak (hiztegi konbentzional batean aurkitzen direnak bezalakoxeak), aditz-formak eta morfema

ez-independenteak bakoitza bere informazio morfologikoarekin. Datu-base hau helburu askotarako pentsatua dagoen datu-biltegi erraldoi malgua eta irekia da.

Egun duen informazioa aldatu, osatu eta gehitu egingo da dudarik gabe, gaurko eta biharko beharrei erantzuteko, eta horretarako aurreko ezaugarriak ezinbestekoak dira. Hortaz, tresna aldagarria eta moldagarria da, eta eguneratuz joango da:

- Batetik, egunero datu-berriak sartuz (eta daudenak eguneratuz) aberasten ari gara (mantentzea). Adibidez, Euskaltzaindiaren erabakiak datu-basean adierazten dira. Semantika lantzen denean ere eremu berriak beharko dira (homografo-zenbakiak adierazten du nolabaiteko hurbilpena), bai eta azpikategorizazioa sistematikoki aztertzean. Hauetako hainbat aukera aurreikusita dago hala ere.
- Bestetik, diseinu-mailako egokitzapena ere egin behar da, informazio mota asko egoki integratu behar direlako.

Une honetan dagoen informazioari lotuko gatzazkio oraingoz, hori baita analizatzaile morfosintaktiko eta sintaktikoak erabiliko dutena. Informazio guztia hiru taula nagusitan banatua dago, baina kategoria bakoitzeko taula bat definituz, eta hauetako bakoitzak sarrerei dagokien informazioa errepresentatzen du. Hauek dira hiru taula nagusiak: hiztegi-sarrerak, adizkiak eta bestelako morfemak. Ikus ditzagun banaka eta hauetako bakoitzaren barruan definitutako taula desberdinak, esan bezala, kategoriaka eta erabiliko ditugun eremuekin¹:

- a) Hiztegi-sarrerak: kategoria nagusitzat definituta dauzkagun guztiek eta lagungarri batzuek osatzen dute lehen taula hau. Bertan, sarreraz gain, kategoriaren berri ematen da eta, osaera zehaztuz, hitz eratorria, elkartua edo hitz anitzeko unitate lexikala den aipatzen da.

Hauek dira definitutako kategoria nagusi sinpleak eta bakoitzaren eremuak (gure analisirako erabiliko ditugunak bakarrik, kategoria-sistema osoaren deskribapena (Aldezabal *et al.* 1999b) txostenean azter daiteke):

- Izena: kategoria honek azpikategoria, biziduna, zenbakarria, neurgarria, plurala eta mugatasun lexikala bezalako eremuak definituta dauzka. Gerora osatu beharko da gizaki, konkretu/abstraktu eta beste hainbat tasunekin ere.
- Adjektiboa: azpikategoriaren berri ematen du.

¹ Hemen aipatzen diren eremuez gain badira beste batzuk ere (Homografo-Id, Iturburua, Iturburuko_Forma, BIM-forma, Klausula-Muga, etab.), baina ez dugu hemen horien berri emango; izan ere, informazio morfologikoa/morfosintaktikoa besterik ez baitzaigu interesatzen analizatzaileen sarrerarako.

- Aditza: oinarri-forma, azpikategoria eta laguntzaile-mota.
- Adberbioa: azpikategoria eta adberbio-mota.
- Izenordaina: azpikategoria, pertsona, numeroa eta muga(gabe)tasun lexikala.
- Determinatzailea: azpikategoria, numeroa/mugatasuna, hurbiltasun-maila eta muga(gabe)tasun lexikala.
- Loturazkoak: azpikategoria.

Kategoria lagungarri gisa sartu ditugunen artean honakoak ditugu, osaera konplexukoak, alegia:

- Laburtzapenak/siglak: adierazia eta adierazlearen kategoria dute.
- Hitz eratorriak: oinarria, aurrizkia eta atzizkia(k)².
- Hitz elkartuak: mugakizuna, mugatzailea eta elkarketa-mota.
- Hitz anitzeko unitate lexikalak.

b) Adizkiak: aditz laguntzaile guztiak eta aditz trinko nagusienak listatuta biltegitatu dira informazio honekin: sarrera, kategoria, oinarri-forma, modua/denbora, nork, nori, nor eta hitanoa.

c) Bestelako morfemak: morfema ez-askeak dira hauek, hiztegi-sarrera izan ez daitezkeenak, hain zuzen ere. Taula honen barruko kategoriak honela sailkatu dira:

- Deklinabide-morfema: kasua, numeroa³, mugatasuna eta funtzio sintaktikoa(k).
- Erlazio-morfema: erlazioa eta funtzio sintaktikoa(k).
- Atzizki lexikala: oinarriaren kategoria eta eratorriaren kategoria.
- Aurrizki lexikala: oinarriaren kategoria eta eratorriaren kategoria⁴.
- Aspektu-morfemak: aspektua.
- Elipsia: elipsia adierazteko erabiltzen dugun zeinua.

² Atzizki bat baino gehiago izan baitaiteke.

³ Deklinabide-kasuak mugatasun- eta numero-informazioekin bilduta daude eta horregatik behar dira eremu horiek, berez deklinabideari ez badagozkie ere.

⁴ Nahiz badakigun, printzipioz, aurrizki lexikalek ez dutela kategori aldaketarik eragiten, atzizkien kasuan bezala bi eremu horiek errespetatzen dira.

- Elkarketa-marra: elkarketa adierazteko erabiltzen dugun zeinua.

Informazio hau guztiau baliatu ahal izango dugu, beraz, testu-hitz baten segmentazioa egiterakoan eta, ondoren, analisiaren azken emaitza eskaintzerakoan.

II.1.2 Segmentatzaile morfologikoa

Segmentatzaile morfologikoak testu-hitz baten analisia egiterakoan osagaien segmentazioa ematen du; hau da, lema eta morfemak banatu eta bakoitzari dagokion informazio morfologikoa eranstean zaie, horretarako EDBLko informazioa erabiliz. Alegria (1995) eta Urkia-ren (1997) tesi-lanetan garapen horren berri ematen da, ikuspuntu konputazionaletik lehenengoan eta linguistikotik bigarreanean. Tresna hau izango da tesi honen abiapuntua, eta horregatik helburua segmentatzaile morfologikotik analizatzaile morfosintaktiko eta sintaktikorako bidea eraikitzea da. II.1 adibidea erakusten du *mendiak* hitzaren segmentazio morfologikoaren emaitza: bi interpretazio daude, ergatibo singularra eta absolutibo plurala. Hitz osoaren analisia nahiz analisi sintaktikoa ateratzeko, lema eta morfema horiek konbinatu egin beharko dira.

```
((forma "mendiak")
  ((anal 1)
    ((lema "mendi") ((SAR mendi) (KAT IZE) (AZP ARR)))
    ((morf "ak") ((SAR ak) (KAT DEK) (KAS ABS) (NUM P) (MUG M) (FS1 @OBJ)
                  (FS2 @SUBJ) (FS3 @ATRIB))))
  ((anal 2)
    ((lema "mendi") ((SAR mendi) (KAT IZE) (AZP ARR)))
    ((morf "ak") ((SAR ak) (KAT DEK) (KAS ERG) (NUM S) (MUG M) (FS1 @SUBJ))))))
```

II.1 adibidea. *mendiak* hitzaren segmentazio morfologikoa.

II.2 Analisi morfosintaktikoaren diseinua

II.2.1 Analisi morfosintaktikorako hurbilpenak

Hitzen morfologia konputazionalari buruz egindako lanari erreparatuz gero, (Ritchie *et al.* 1992, Sproat 1992) izango dira orain arte egindako lanik sakon eta orokorrenak. Liburu

horietatik lehenengoan tratamendu morfologiko osorako proposamen bat egiten dute. Egile hauek hiru aspektu nagusi bereizten dituzte hitz-forma baten analisirako:

- a) Morfofonologia (honi *morfografemika* ere deitzen zaio). Honetan bi puntu nagusi tratatuko dira: hitz-forma bere morfemetan zatitzea (segmentazioa), eta morfemak lotzean gertatzen diren aldaketa ortografiko eta fonologikoen deskribapena. Adibidez, *zakur* eta *-a* morfemak lotzean *zakurra* ateratzen da, hau da, lehen morfemaren azken karakterea bikoiztu egiten da. Prozesu honek karakterea dauka oinarritzko unitatetzat, nahiz eta badagoen karaktere horiei eranstean zaien beste informazio motak ere erabiltzea.
- b) Morfotaktika. Termino honekin morfemen arteko kateaketa edo segida posibleak eta beraien ordenaren murriztapena adierazi nahi da. Adibidez, *-tze* atzizkia aditz bati lot dakioke (*ekartze*), baina ez adjektibo bati.
- c) Morfemen informazioaren konposizioa. Morfemak lotzeaz gain, oraindik beste galdera bat geratzen da erantzun gabe: aldaketa morfofonologiko horiek aplikatuz eta ordena horretan doazen morfemek zer osatzen dute? Beste era batera esanda: osatu den hitza zer da? Puntu honetan morfemen informazioa (numeroa, kasua, ...) nola konposatzen den adierazi beharko da, unitate osoarena (hitza) lortu arte. Adibidez, *handitasuna* hitza hiru morfemez osatua dago: *handi* (adjektiboa), *-tasun* (eratorpen-atzizkia, adjektibotik izena sortzeko) eta *-a* (kasu absolutiboa, mugatu singularra). Horiek konbinatuz lortzen den hitza izen bat izango da, kasu absolutiboan eta singularrean.

Morfofonologia definitzeko aukera desberdinak erabili dira: programaren bidezko definizioa, erregela berezien definizioa (eragiketa nagusiak karaktere baten aldaketa, ezabaketa edo kenketa izanda) (Allen *et al.* 1987), morfema-aukera guztien definizioak lexikoian sartzea (hau da, ez dago aldaketa morfofonologikorik) (Tzoukerman eta Liberman 1990), edo erregela sekuentzialen konposizioa (Kaplan eta Kay 1981). Dena dela, azken bi hamarkadetako pausorik garrantzitsuena Koskeniemi-ren (1983) lanean eman da, bertan bi mailatako (azaleko maila eta maila lexikala) morfologia definituz. Bi mailen arteko korrespondentziak egoera finituko transduktoreen bidez ematen dira. Formalismo honek bide egoki eta eraginkorra ematen du aldaketa morfofonologikoen deskribapenerako, eta lengoaia-multzo zabal bati aplikatu dakioke.

Morfotaktikaren definiziorako bi bide izan dira nagusi:

- Lehenengo batean, *egoera finituko morfotaktikan*, morfemen kateaketa posibleak honela deskribatzen dira: morfemak azpilexikoietan banatzen dira, eta azpilexikoen arteko jarraitze-klaseak (kateaketa posibleak) definitzen dira. Bide hau Koskenniemi-k berak deskribatu zuen (1983). Dena dela, hurbilpen honek arazoak dauzka morfema batek bere ondoren jarraian ez datozen beste morfemen kateaketa baldintzatzen badu (fenomeno honi urruneko mendekotasuna deitzen zaio, adibidez, *ez-onartze* hartzen bada, *ez-* aurrizkiaren ondoren izena beharko litzateke, baina *onar* aditza dator). Problema hori ebazteko, soluzio desberdinak egon dira, Alegriaren tesian (1995) aurkeztutako jarraitze-klase hedatuena bezalakoa.
- *Baterakuntzan oinarritutako morfotaktikan*, berriz, hitzaren barruko testuingururik gabeko gramatika baten gainean definitzen da informazio hori. Bide hau (Trost 1990, Ritchie *et al.* 1992, Antworth 1994) lanetan jarraitu da. Baterakuntzaren erabileraren aldeko arrazoi nagusia erazagutzailetasuna da, urruneko mendekotasunen arazoa konpontzeko adibidez, eta desabantailen aldetik baterakuntzaren inplementaziorako behar den abiadura-galera aipatu behar dugu.

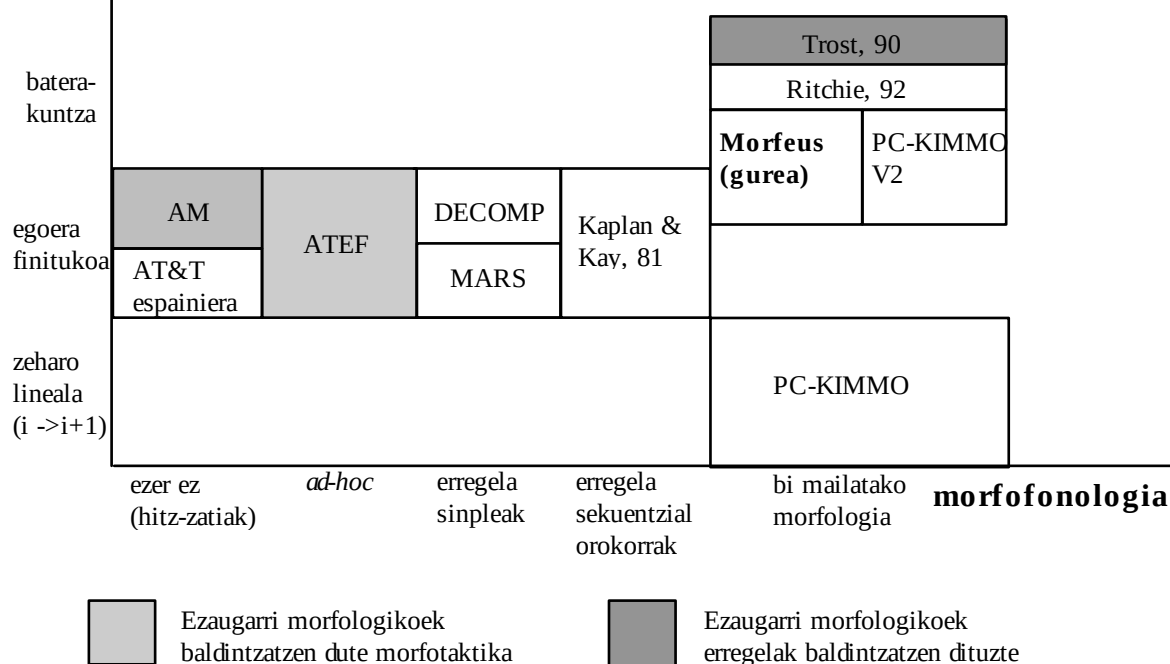
Morfemen informazioaren konposizioa ez da beharrezkoa edo garrantzitsua zenbait sistematan, askotan ez delako morfologiatik harantzago joan nahi izaten, adibidez, zuzenketa ortografikorako edo hitz isolatuen ezagumendurako. Baina ondoren sintaxia edo semantika aplikatu nahi badira, beharrezkoa izaten da formalismo bat erabiltzea hitz osoaren informazioa metatzeko morfemen informaziotik abiatuta. Hau hizkuntza batzuetan eredu sinpleen bidez lortu ahal da baina beste batzuetan, euskara kasu, agertutako fenomenoek konposizio-erregelen definizioa eskatzen dute. Pauso hau sistema gehienetan baterakuntzan oinarritutako gramatiken bidez egin da (Trost 1990, Ritchie *et al.* 1992, Antworth 1994, Carulla eta Oosterhoff 1996, Badia *et al.* 1996, Prószyński eta Kis 1999).

Hiru aspektu horiek era desberdinetan moldatu dira orain arte garatutako sistemetan. (Alegria 1995) tesiko II.1 irudiak analisi morfologikorako garatutako zenbait sistemaren ikuspegi orokorra emango digu. Gure helburua ez da izango egindako sistema guztien sailkapen orokorra egitea, sistema aipagarrienen arkitektura eta diseinuko erabaki nagusiak zehaztea baizik. Irudi horretan euskararako egin dugun azken inplementazioaren aldaketen berri ere ematen da.

Euskararen kasuan, lehenago esan dugu morfofonologia bi mailatako formalismoaren bidez inplementatu dela. Morfotaktikaren zati bat ere egoera finituko ereduaren definitu da, jarraitze-klaseen mekanismoaren bidez, jarraitze-klase hedatuen zabalkuntzarekin. Bi hauek konbinatuz segmentatzailea dugu, eta geratzen den azken aspektua, segmentatzaile

morfologikotik analizatzaile morfosintaktikorako bidea eraikitzea hain zuzen ere, horixe da tesi honetako II. kapitulu honen helburua.

morfotaktika (+ informazioaren konposizioa)



II.1 irudia. Prozesadore morfologikoen sailkapena.

II.2.2 Euskararen analizatzaile morfosintaktikoaren ezaugarriak

Ikusi dugu orain arte euskararen morfologiarako garatu diren tresnak erabiliz, testu-hitz baten analisisa egiterakoan osagaien segmentazioa besterik ez dela ematen, hau da, osagai guztiak banatu eta bakoitzari dagokion informazio morfologikoa (morfosintaktikoa, askotan) erantsi zaiela. Horrez gain, gaur egun informazio sintaktikoa ere eskaintzen digu gure EDBL datu-base lexikalak, formaren deskribapenez gain funtzio sintaktikoaren berri ere badugulako. Informazio hori guztia gehitzea interesgarria da hizkuntzaren azterketaren ikuspegitik, baina formaren analisisa egin eta irteera eskaintzerakoan arazo batekin baino gehiagorekin egin dugu topo. Alegria-ren (1995) tesiak aipatzen du analisi morfosintaktikoaren beharra sintaxian, etiketatze/lematizatzaileetan eta beste aplikazioetan, fenomeno hauek tratatzeko:

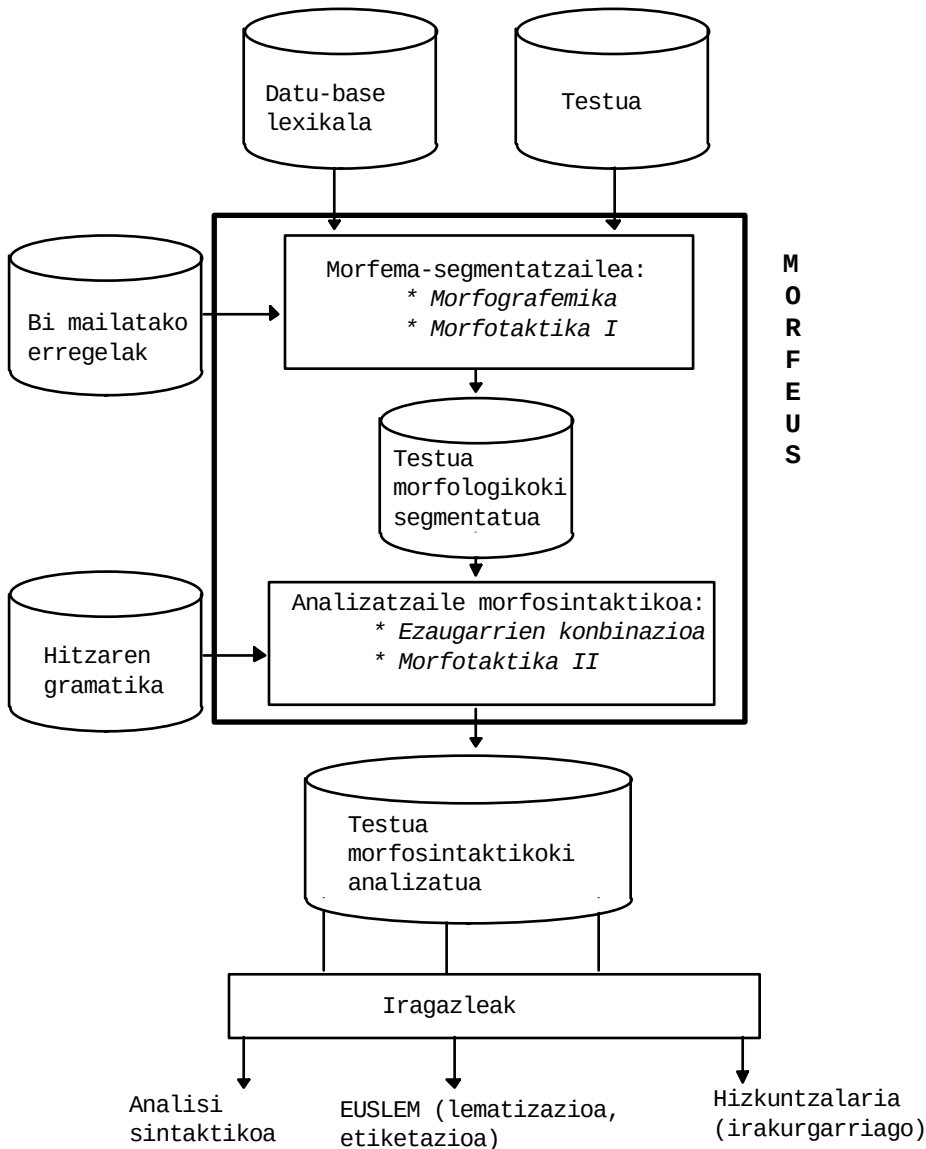
- Kasu, numero eta mugatasunaren informazio mota anitzak. Izen, adjektibo eta zenbait kategoria eratorrirekin atzizki desberdinak meta daitezke lema baten ondoren, kasu, numero eta mugatasunari buruzko informazio mota anitzekin. Askotan azken atzizkiaren informazioa da hitz osoari dagokiona, baina beste aukera batzuen tratamendua ere zehaztu beharko da.
- Elipsia. Genitiboaren ondoren beste kasu bat agertzen denean, izen-elipsia sortzen da kasu askotan. Hitz osoaren analisisian lema guztien informazioa mantentzea (lema eliptikoa barne) izango da interesgarriena.
- Kategoria-eratorpena eta elkarketa. Kasu hauetan lema baten kategoriaren aldaketa eta bi lemen bildura daukagu. Zenbait aplikaziotarako bai osagai diren lemen informazioa bai sortutakoarena jakitea inportantea da.

Lehen esan bezala, gure kasuan morfotaktikaren zati bat segmentazio morfologikoarekin batera landu da bi mailatako formalismoan txertatuz. Analisi morfosintaktikoa esaten dugunean morfemen informazioaren konbinazioaren emaitzari buruz ari gara, hau da, morfofonologia, morfotaktika eta morfemen konposizioa hartzen dituen prozesu osoaren emaitza. Lan hori ingeleserako sistema batean egin bada, problema latzagoa izango da hizkuntza eranskari batentzat egin behar denean. Gure asmoa, beraz, testu-hitzaren azken analisi morfosintaktikoa ematea da, beti ere emaitza osoaren berri emanez, barruan dagoen informaziorik galdu gabe; alegia, formaren barruko informazioa antolatu eta *analisi* eskaini, osagai-morfemen segida gaindituz.

Euskal morfosintaxiaren tratamendu automatikoaren ahalegin honetan, irteera sendo eta osoa lortu nahi dugu. Kategoria-sistema aukeratu zenean, hitzaren informazioaren etiketatze-maila desberdinak (Aduriz *et al.* 1995) zehaztu ziren aplikazio bakoitzaren helburuaren arabera irteera bat ala beste erabiltzeko. Eta, informazio morfosintaktikoaren irteera lantzeko, *analisi morfologikoaren emaitza osoa* egokiena zela ikusi zen (ikus § VI.1en EUSLEM lematizatzaile/etiketatzailearen azalpena), hain zuzen ere analisiaren funtzio anitzeko irteera dugulako helburu eta, printzipioz, ez dugulako inongo murriztapenetara lotuta egon behar.

II.2 irudiko eskemak argi dezake analizatzaile morfosintaktikoak eskainiko duen egoera eta, horrekin batera, analizatzaile morfosintaktiko orokorra eta erabilgarria izateko gure eginkizuna nola bideratu den.

erabileretarako ustia daitekeena, bestalde, iragazleak ezarrita.



II.2 irudia. Analizatzaile morfologikoa.

Testu segmentatu horrek tratamendu morfosintaktikoa ere izango du; beraz, testua morfologikoki segmentatua beharrean morfosintaktikoki analizatua izango dugu bere

⁵ Tresna hauei buruzko informazioa <http://ixa.si.ehu.es> helbidean aurki daiteke.

osotasunean. Hemen ere, aplikazioen arabera iragazleak erabiliko ditugu, hau da, ez dugu informazio mota bera beharko analisi sintaktikorako edo EUSLEMerako, adibidez.

II.3 Oinarrizko erabakiak

Analizatzaile morfosintaktikoa diseinatzeko hartu ditugun erabakiak hiru eremu nagusitan bana daitezke:

- 1) Linguistikoa. Hizkuntzaren teoriari berari buruzko erabakiak hartu ditugu, analisia egiterakoan “goratu”⁶ behar den informazioa erabakitzeko. Analisi morfosintaktikoari ekin aurretik, lan linguistiko handia egin zen horretarako beharrezkoa zen informazioa EDBLn agertzeko, eta bere konbinazioan aplikatzeko printzipio nagusiak erabakitzeko. Hauek (Urkia 1997, Aduriz *et al.* 1999, 2000ab) lanetan daude deskribatuta, eta honela labur daitezke:
 - Mugatasun lexikala. Kasu batzuetan, erroek bere baitan izango dute mugatasuna (pertsona-izen eta leku-izen bereziak, determinatzaileak eta izenordainak), eta beste batzuek (izen arruntak edo adjektiboak, esate baterako) atzizkien bidez lortuko dute mugatasun/mugagabetasuna. Erregela morfosintaktikoen ardura izango da kasu guztietan informazio egokia goratzea hitzaren analisia emateko.
 - Hitzaren barruko elipsia. Elipsia, euskaraz hitzaren baitan gertatzen den fenomeno, kontuan hartzekoa da, tratamendu berezitua eskatzen baitu. Esate baterako, *umearena* forma nola analizatu behar da? Elipsia jabego zein lekuzko genitiboaren ostean agertuko da (baina kontuz, horrelako guztiek ez dute beti elipsia adierazten; ‘*aingeru guardakoa*’ bezalakoetan, adibidez, ez da elipsirik egongo). Fenomeno honen deskribapen zabala (Urkia 1997) tesian dator.
 - Mugatasun, numero eta kasu arrunten metaketa. Elipsirik ez dagoenean eta kasu bat baino gehiago agertzen duten hitz-formetan, arau orokorra azken morfemaren tasunak hartzea izango da. Adibidez, *gizonarentzako* forma hartzen bada, segmentatzaileak honako interpretazio bat eskainiko digu:

⁶ Hau da, maila lexikaetik maila morfosintaktikora pasa.

((lema "gizon") ((SAR gizon) (KAT IZE) (AZP ARR)))

((morf "areM") ((SAR aren) (KAT DEK) (KAS GEN) (NUM S) (MUG M)))

((morf "tzat") ((SAR tzat) (KAT DEK) (KAS PRO) (MUG MG)))

((morf "ko") ((SAR ko) (KAT DEK) (KAS GEL)))

Adibide horretan, kasu eta mugatasunaren azken balioak (GEL eta MG) hartuko lirateke hitz osoarentzat.

- Lexikoien antolamendua. Arrazoi desberdinengatik, kasu-marka, numeroa eta mugatasuna morfema desberdinetan egon daitezke sakabanatuta ('-eta + -tik' kasuan, adibidez, lehen morfemak numeroa eta mugatasunaren informazioa dauka, eta bigarrenak kasuarena), edo beste batzuetan, berriz, bilduta agertuko zaizkigu (-arekin). Erregela morfosintaktikoen lana izango da, beraz, hauen tratamendu egokia egitea.

- 2) Hitzaren gramatikarako formalismoa. Baterakuntzan oinarritutako formalismoak (Shieber 1986) izan dira nagusi honen tratamendurako, baterakuntza-mekanismoa erazagutzaila izateagatik eta bere ahalmen deskriptiboari esker. Euskararen kasuan, bere hitz-mailako fenomenoek aberastasuna eta tratatu beharreko egituren konplexutasuna aintzat hartuta, are eta beharrezkoagoa izango da baterakuntza erabiltzea, egoera finituko mekanismoen bidezko soluziorik ez dagoelako.

Lehen aipatu dugu baterakuntzan oinarritutako zenbait formalismo garatu direla han eta hemen, horien artean PATR, LFG (euskararen sintaxiaren formalizaziorako erabilia Abaitua (1988) eta Zubizarreta-ren (1992) lanetan), GPSG (Gazdar *et al.* 1985) eta HPSG (Pollard eta Sag 1994) aipa ditzakegu. Hitzaren gramatika tratatzeko GPSG estiloko formalismoa erabili da (Ritchie *et al.* 1992) ingeleserako lanean, Trost-en (1990) lanak HPSG erabiltzen du alemanerako eta Antworth-en (1994) PC-KIMMO V2 ingeleserako sisteman, berriz, PATR erabiltzen da.

Gure kasuan azken formalismo hau, sinpleena, aukeratu dugu, PC-KIMMOren bigarren bertsioan egiten den moduan. Honek gauzak egiteko malgutasuna emango du, baina horren ondorioz erregelak erredundanteagoak izatea ekarriko du kasu batzuetan. Aipatutako beste formalismo konplexuagoetan (GPSG, LFG edo HPSG), printzipio orokorren bidezko generalizazioak adierazi ahal dira, eta horrekin erregela-kopuruak gutxitu eta sinplifikatu egiten dira. Euskararen formalizazioan oraingoz printzipio orokorren eta ezaugarri-multzoen lehen azterketa besterik ez dugu egin, ondoren azalduko den moduan. Bide hau urratzeari oso interesgarria deritzogu, baina honek lan linguistiko sakona eskatuko duelakoan gaude.

- 3) Estrategikoa. Erabaki horiek formalizatzeko *Text Encoding Initiative* (TEI) formatua erabili da (Ide eta Veronis 1995, Ide 1998, Arriola *et al.* 1997, Ide eta Greenstein 1999, Artola *et al.* 2000). Beraz, segmentatzailearen zein analizatzailearen emaitza *Standard Generalized Markup Language* (SGML) bidez eskainiko da, eta hau IXA taldeko informazioaren kudeaketarako lengoaia komuna izango da. Esan bezala, analizatzaile morfosintaktiko erabilgarria nahi dugunez, irekia izan behar du, eta ez du inongo formatu zehatzetara edo helburu bakanetara lotuta egon behar. Horregatik aukeratu da bide hau, aplikazio bakoitzak (EUSLEM, murriztapen-gramatika, analisi sintaktikoa, ...) beharko duen informazio partikularra/berezia iragazkien bidez emango delako. Ondorioz, analisi osoa izango da hitzaren gramatikak emango duena, barruko osagai guztiak gordez eta horren azaleratzea norberaren beharren arabera gauzatuz.

II.4 Gramatika morfosintaktikoa

Puntu honetan gramatikaren ikuspegi orokorra eman nahi dugu. Hasteko, erregeletan agertzen den notazioa zehaztuko da, ondoren erregela baten adibidea aztertzeke (§ II.4.1). Hurrengo azpipuntuan (§ II.4.2) gramatikaren azalpen orokorra egingo da eta, bukatzeke, hitz-forma baten analisisia eta bera lortzeke aplikatutako erregelak emango dira (§ II.4.3).

II.4.1 Erregelen adibide bat

II.2 adibidean agertzen den erregela (r_eratorpena izena eman zaio), atzizki lexikalak (eratorpena sortzen dutenak) tratatzeko erabiliko da. Horrela ‘etxe + -txo’, ‘erabil + -garri’ edo ‘egin + -araz’ bezalakoak tratatuko dira. Kasu horietan kategoria-aldaketa izango dugu, eta kategoria berriaren informazioa atzizkiak emango du.

```

% eratorri_kat -> oinarri_kat + erat_atzizkia
% ad.: lau+garren, lagun+garri, ikusgarri+tasun, iraki+te (te atzizki lexikala izanik)
% egin+araz
% kat_apreziatibo -> kat + apreziatibo_atzizkia
% ad.: etxe+txo, zuri+txo, apurtutxo, dagoen+txo, berandu+txo

r_eratorpena,
  X0 ----> X1, X2
          X2/ezaug/kat          <=> atz,          % X2ren kategoria
atzizkia da
          X0/ezaug/kat          <=> X2/ezaug/ker,    % X0ren (hitz osoaren)
kategoria
                                     % X2ren 'ker'
ezaugarriak
                                     % adieraziko du
          X0/ezaug/azp          <=> X2/ezaug/aer,
          X0/oina               <=> X1/oina,
          X0/ezaug/oin          <=> X1,
          X0/ezaug/oin/oina/twol <=> X1/oina/twol,
          X0/ezaug/oin/oina/sarrera <=> X1/oina/sarrera,
          X0/ezaug/atzl         <=> X2,
          X0/ezaug/erat         <=> plus,
          X0/forma              <=> X1/forma,
          X0/morf_lista         <=> gehitu_morf_lista(X1/morf_lista,
                                                         X2/morf_lista)).

```

II.2 adibidea. Erregela morfosintaktiko bat.

Erregela horretan ikusten dira formalismo honetan deskribapenak egiteko erabiltzen diren atalak:

- Hasieran, erregelaren izena (*r_eratorpena* adibidean).
- Ondoren, erregelan hiru osagai daudela esaten da: X0 (gurasoa) eta bi ume (X1 ezkerrekoa eta X2 eskuinekoa). Erregela bitarra dugu, beraz.
- Hurrengo lerroetan, baterakuntza-ekuazioak azalduko dira, erregelaren osagaiek betetzen dituzten murriztapenak adieraziz.

Hauek dira erregelen ekuazioetan erabili diren eragileak eta beren esanahiak:

- % zeinuaren ondoren datorrena, lerroaren amaieraraino, oharra izango da.
- “<=>” eragileak baterakuntza adierazten du, adib.: X0/kas <=> X2/kas

Horrela, X0 gurasoaren eta X2 eskuineko umearen kasuaren balioek bat etorri beharko dute. Kasuaren balioa atomikoa denez, horrek kasuak berdinak izatea eskatuko du, baina balio egituratuak direnean, baterakuntzaren bidez bi egiturak berdindu egingo dira.

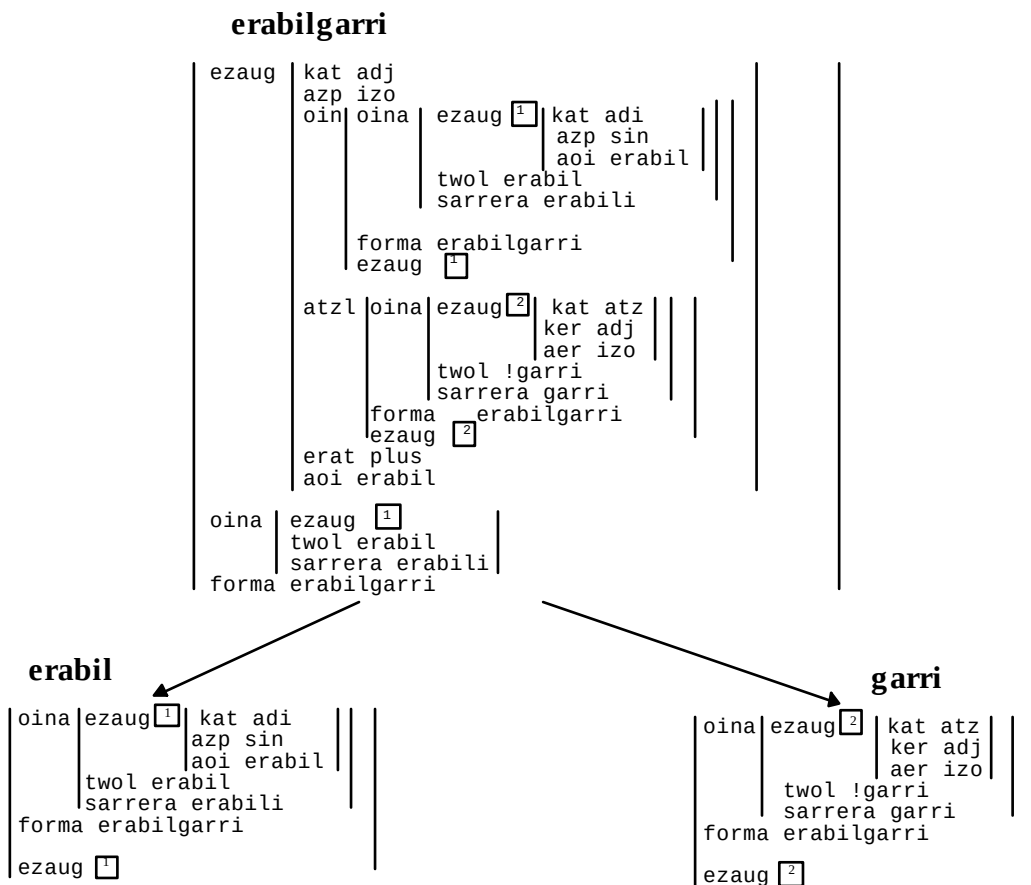
- “ez” eragileak ezaugarriaren balioa (atomikoa) bigarren listako elementua ez izatea ziurtatzen du. Adib.: X1/kas ez [gen, gel] murriztapena bete egingo da X1 egiturako kasuaren balioa genitiboa (jabegokoa zein lekuzkoa) ez bada.
- “badago” eragileak ezaugarriaren balioa (atomikoa) bigarren listako elementua izatea ziurtatzen du, adib.: X1/kas badago [gen, gel].

Adibide horretan kasua “gen” edo “gel” ez bada, orduan ekuazioa ez da beteko. Ondorioz, kasua definitu gabe balego ere, ekuazioa ez litzateke beteko.

- “\$” eragileak kateaketa adierazten du, adib.: $X0/sar \Leftrightarrow \$(X1/sar, X2/sar)$
- “edo/eta” eragileek baldintzetatik bat/denak betetzea eskatzen dute. Hau erregela-kopurua gutxitzeko definitu da, adib.:

*edo([X1/nor badago [hu],
X1/nrk badago [hu],
X1/nri badago [hu]])*

Murritzapen hori bete egingo da X1 ezkerreko umearen “nor”, “nri” edo “nrk” ezaugarrien balioa “hu” (singularreko hirugarren pertsona) bada.



II.3 irudia. II.2 adibideko erregelaren aplikazioa⁷.

⁷ Irudian indizeen bidez bi ezaugarri-egiturak balio bera dutela adierazten da.

II.3 irudian erregelaren aplikazio-adibide bat dugu. Hitz baten informazioan bi mota bereizten dira: bere ezaugarri morfologikoak (“ezaug” ezaugarrian), lema eta morfemetatik hartzen direnak, eta hitzaren oinarrian dagoen lemaaren informazioa (“oina” ezaugarriaren azpian). Eratorpenean, atzizkiak hitz eratorriaren kategoria eta azpikategoria adierazten ditu, “ker” eta “aer” ezaugarrien bitartez, horiek izango baitira hitz berriaren kategoria/azpikategoria. Horregatik adibidean *-garri* atzizkiak izenondoko adjektiboa (*adj izo*) sortzen duela adierazten du. Hitz eratorrien ezaugarrien artean eratorpenaren oina den lema eta atzizkiaren berri ematen da, “oin” eta “atzl” (atzizki-lista) izeneko ezaugarrien bidez. II.3 adibideak erregelaren aplikazio-adibide sinplifikatuak erakusten ditu.

Adib.: <i>erabilgarri</i>			
erabil + garri	=>	erabilgarri	
kat:adi	kat:atz	kat: adj	
azp:sin	ker:adj	azp: izo	
	aer:izo	erat:plus	
		oin: erabil	
		atzl:garri	
Adib.: <i>eginaraz</i>			
egin + araz	=>	eginaraz	
kat:adi	kat:atz	kat: adi	
azp:sin	ker:adi	azp: fak	
	aer:fak	erat:plus	
		oin: egin	
		atzl:arazi	
Adib.: <i>gizontxo</i>			
gizon + txo	=>	gizontxo	
kat:ize	kat:atz	kat: ize	
azp:arr	ker:ize	azp: arr	
	aer:arr	erat:plus	
		oin: gizon	
		atzl:txo	

II.3 adibidea. Eratorpenaren erregelaren aplikazio-adibideak.

II.4.2 Gramatikaren ikuspegi orokorra

Guztira hogeita lau erregela definitu ditugu. Diseinu-erabaki aipagarria erregela guztiak bitarrak direla izan da (hau da, bi umeko erregelak, ume bakarreko erregela baten salbuespenarekin). Horrela erregela bakoitzean fenomeno bakarra eta ahalik eta independenteena definituko da. Aukera honen justifikazioa Karttunen (1986) eta Uszkoreit-en (1986) lanetan aurki daiteke, eta gure kasuan oso lagungarria gertatu zaigu, hitzaren informazioa morfemen pausoz-pausoko konposizio modularren bidez adierazteko oso baliagarria izan baita. Erregelak era honetan daude banatuta:

- Hamaika erregela flexio-morfemen bilketak adierazteko, eta hauek kategoria nagusiekin konbinatzeko. Orokorrean, kategoria nagusiek hitzaren kategoria eta

azpikategoria ematen dute eta flexio-morfemek numeroa, kasua eta mugatasuna. Hauek dira erregelek tratatzen dituzten fenomenoak:

- Erroa (izena, adjektiboa, ...) gehi kasu-marka (numeroa eta mugatasuna barne).

Adibidez: ‘gizon + -a’, ‘gizon + -arentzat’, ‘mendi + -etara’,
‘mendi + -etaraino’, ‘gizon + -arengana’, ‘gorri + -etaraino’,
‘makurtu + -a’, ‘eman + -arekin’, ‘zenbait + -engana’, ‘ni + -ri’,
‘zu + -rekin’, ‘-0(elipsia) + -a’

- Aurrekoaren berdina baina hitz berezien eta mugatasun lexikala duten erroak lotzeko. Honetarako hiru erregela behar izan dira.

Adibideak: ‘guraize + -ak’, baina ez ‘*guraize + -a’, ‘Gabon + -ak’,
‘Kepa + -ri’

- -ko eta -ren genitibozko atzizkien hiru kasu berezi tratatzeko erregelak. Batzuetan, lekuzko genitiboaren -ko atzizkia kasu-marka duen osagai bati gehitzen zaio (‘mendiarentzat + -ko’, ‘mendiarekin + -ko’, ‘mendira + -ko’, ‘gizonarengana + -ko’, ‘mendiraino + -ko’). Adberbioekin, berriz, erro arruntekin ez bezala, -ko atzizkia numeroa-mugatasunik gabe lotzen da (‘gaur + -ko’). Azken aukera bizidunen adlatiboa eta antzeko kasuak sortzeko jabego-genitiboa da (‘gizonaren + -gana’).

- Azkenik, lau erregela daude kasu-marka, numeroa eta mugatasunak osatzen duten zenbait atzizki-multzo lortzeko.

Adibideak: ‘-0 + -ko’ (mendiko), ‘-0 + -tik’(menditik), ‘-eta + -ko’ (mendietako), ‘-a + -gana’ (gizonagana), ‘-on + -ekin’ (guztioneekin)

- Adjektibo, partizipio, determinatzaile eta adberbioen gradua tratatzeko erregela bat.

Adibideak: ‘handi + -ago’, ‘handi + -en’, ‘makurtu + -ago’, ‘maiz + -ago’,
‘honekin + -txe’

- Elipsiaren tratamendurako erregela bat. Erregela honek hitz-forma elipsoidunei syntaxitik oso hurbil dagoen egitura egokitzen die, hitzeko osagai desberdinen informazioa gordez. Adibidez, *zaldiarenekoa* analizatzeko orduan, gramatikako beste erregelek honako hiru osagai sortuko dituzte (non X ikurrak izen eliptikoaren tokia betetzen duen):

‘zaldiaren + X-eko + X-a’

Elipsiaren erregelak hiru osagai horiek bilduko ditu bi aplikazioen bidez, hiru osagaien segida adieraziko duen egitura bat sortuz.

- Zazpi erregela aditzen morfema-kateaketak deskribatzeko. Hauen artean menderagailuak eta aspektu-morfemak daude.

Alde batetik erregela bat dago aditzaren forma desberdinak emateko (partizipioa, aditz-izena, aditzoina, etorkizuneko forma, eta ez bukaerakoa).

Adibidez: ‘*etor + i*’, ‘*etor + -tze*’, ‘*etor + -0*’, ‘*etorri + -ko*’, ‘*etor + -tzen*’

Bestalde, aditzen bidez mendeko perpausak sortzeko, baldintzazko aurrizkia, *ba*-partikula eta *bait*- aurrizkia lantzeko sei erregela ditugu. Adibidez: ‘*eman + -da*’, ‘*eman + -ik*’, ‘*ikusi + -takoan*’, ‘*ikusi + -agatik*’, ‘*eman + -tea*’, ‘*eman + -teko*’, ‘*eman + -ten*’, ‘*eman + -tean*’, ‘*du + -en*’, ‘*du + -enean*’, ‘*du + -elako*’, ‘*du + -elarik*’, ‘*ba + -lego*’, ‘*ba + -dago*’ (partikula), ‘*bait + -du*’

- Bi erregela eratorpenaren tratamendurako, bata aurrizkientzat eta bestea atzizkientzat. Askotan eratorpenak kategoria-aldaketa dakar berarekin. Hau ezaugarri berezi bat erabiliz adieraziko da.
- Hitz-elkarketarentzako bi erregela. Momentuz izena-izena elkarketa modu arruntena aukeratu dugu tratamendurako. Bi erregela egotearen arrazoi nagusia lehen aipatutako erregela bitarren erabakia da, elkarketan hiru osagai daudelako (bi izenak eta marra). Horregatik erregela batek marra eta bigarren izena lotuko ditu, eta bigarren erregelak guztiaren analisisa emango du.

Erregelen bidez hainbeste tasun morfosintaktiko tratatu behar direnez, problema bat gertatzen da erregela guztietan adierazi behar denean tasun horien guztien jokaera, hau da, bakoitzeko ekuazio bat jarri beharko bagenu. Horregatik, ezaugarri guztien azterketa egin genuen, bakoitzaren portaera aztertu ondoren printzipio orokorrik asma zitekeen aztertzeke, (Ritchie *et al.* 1992) lanean definitutako *WordHead* edo *WordDaughter* izeneko printzipioen ildotik. Azterketatik arau nagusi bat atera genuen II.1 taulan eskuineko zutabean agertzen diren ezaugarri guztientzat:

Erregela bitarretan, erregelaren ekuazioak aplikatu ondoren, gurasoaren ezaugarrien balioak gurasoarengan ez badaude orduan beren balioak eskuineko umetik hartuko dira, eta ez badaude definituta orduan ezkerreko umetik.

Araua betetzen ez duten ezaugarriak	Araua betetzen duten ezaugarriak
kat, azp, ker, aer, oin, atzl, forma	mug, num, kas, biz, zenb, neur, fs_lista, per, plu, adm, asm, erl, grm, mdn, nor, nri, nrk, err, mdl, aoi, rare, erat, mugkz, mugtz, elkarketa, elk

II.1 taula. Ezaugarrien goratzerako portaera desberdinak.

Ikusten denez, ezaugarri gehienak atzizkietatik goratzen dira (numeroa, mugatasuna eta kasua dira aipagarrienak), baina badaude horren salbuespenak. Adibidez, numeroa normalean atzizkitik jasotzen da, baina kasu batzuetan lexikoitik dator ('*ni* + *-k*'), eta orduan lehen osagaitik goratuko da. Beste kasu batzuetan, ezaugarri baten balioa ez dator ez atzizkitik ezta lematik ere, baizik eta erregelak sortua da, eta ez dago lexikoian. Hori tratatzeko erregelaren lehen atala erabiltzen da: "ezaugarriak gurasoarengan ez badaude ...", eta gurasoarengan egongo dira ekuazio baten bidez sortuak izan badira. Horrela, onartu egiten da balio lehenetsien erabilera, erregelaren ekuazioen bidez beste balio bat emanez gainditu egin daitekeena. Printzipio orokor honen aplikazioak erregelen idazketa erraztu du, gramatika laburrago egiteko. Edozein kasutan, printzipio horren barruan irizpide linguistiko zein praktikoak (EDBLko lema eta morfemen banaketatik datozenak) daude, ezaugarrien multzo zabala hartzen baita, eta gu ez gara saiaturko justifikazio linguistiko bat egiten.

Beste alde batetik, lema bati fenomeno morfosintaktiko bat baino gehiago aplikatzen zaionean, zentzurik gabeko hipotesiak baztertzearren fenomenoak aztertzeko morfemen morfotaktika zehaztu behar izan dugu, horrela lotura-ordena jakin bat ezarri: lema, eta ondoren eratorpen-aurrizkiak, eratorpen-atzizkiak, elkarketa eta flexioa, ordena horretan. Ordena hau erregelen murriztapen-ekuazioen bidez finkatzen da. Horrela, 'berridazketa-egilea' bezalako forma bat agertuz gero, modu honetan aplikatuko lirateke erregela morfosintaktikoak (parentesiek ordena markatzen dute):

$((((ber- + idatzi) + -keta) + '-' + (egin + -le)) + -a$

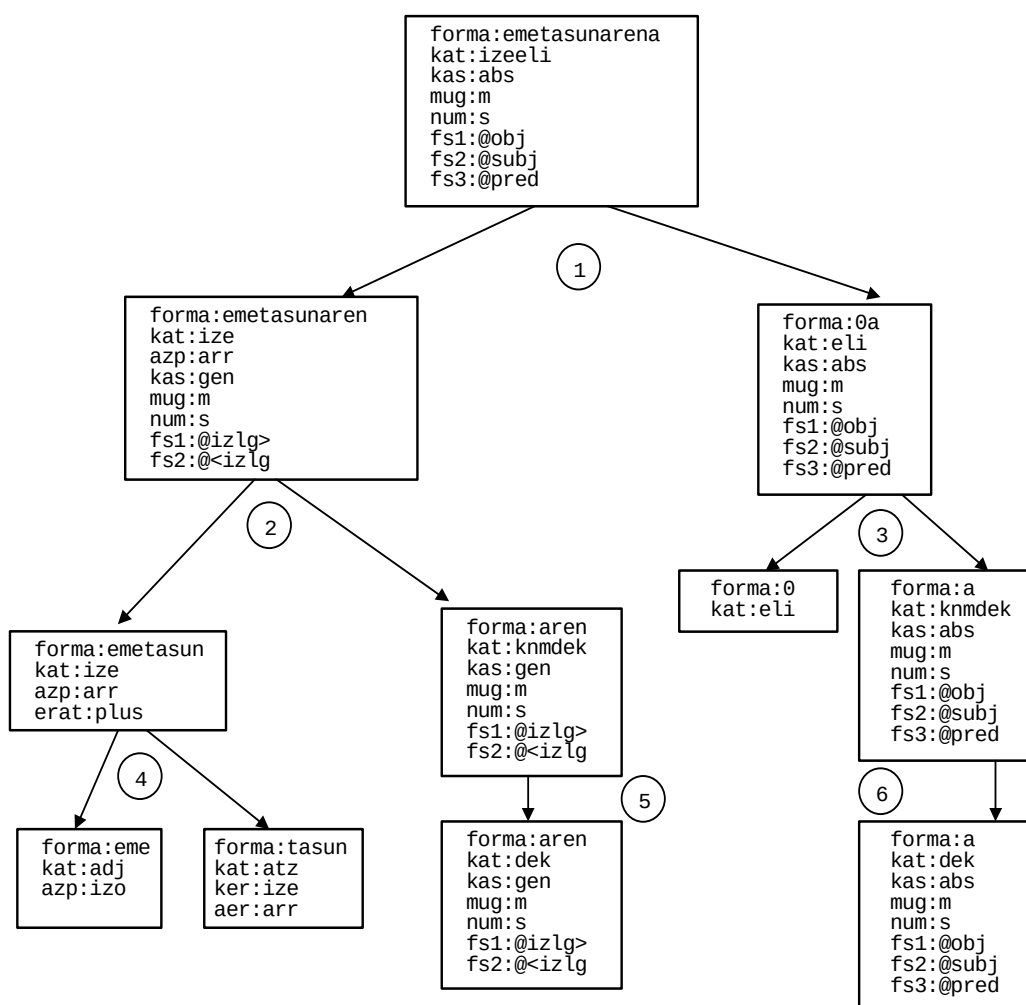
Azken emaitzan, EDBL datu-base lexikal aberatsa oinarri izanik eta morfosintaxiaren tratamendu zabal eta sendo hau emanda, lortu dugun analizatzaileak estaldura osoa du testu errealei aplikatuz gero. Horregatik, hitz-forma estandarrak zein aldaera dialektalak, hitz ezezagunak edo errore ortografikoak tratatzeko gai izango da azken analizatzaile morfologikoa.

Gramatika morfosintaktiko osoaren diseinua eta erregelak (Aduriz *et al.* 1999) barne-txostenean azaltzen dira sakon eta luzeago.

II.4.3 Analisien adibide bat

Ikus dezagun emetasunarena hitzaren analisia nola egingo den. Hitz horretan eratorpena, elipsia eta flexioa agertzen zaizkigu. II.4 irudian analisi-zuhaitza dugu.

4 zenbakiarekin adierazitako erregela eratorpen-atzizkiena da, eme eta -tasun lotzeko. Beste pauso batean flexio-morfemak tratatzeko erregelak aplikatzen dira (5 eta 6 zenbakiarekin markatuak). 2 eta 3 zenbakiko erregelen aplikazioak flexio-morfemak lemarekin lotzeko erregelarenak dira, kasu batean emetasun gehi -aren lotzeko, eta bestean osagai eliptikoa (-0 lema bidez adierazita) -a morfemarekin. Bukaeran, osagai eliptikoak lotzeko erregela aplikatzen da 1 zenbakiarekin markatutako posizioan.



II.4 irudia. Analisi baten adibidea.

II.5 Inplementazioa

Analizatzaile morfosintaktikorako PATR formalismoaren inplementazio bat erabili dugu (Douglas eta Dale 1992). Inplementazio horren aldeko motiborik garrantzitsuenak sinpletasuna eta malgutasuna izan dira, soluzio desberdinak esperimentatzeko aukera eman duelako. Geroago ikusiko denez, sistema hau sintaxiaren tratamendurako erabili dugun bera da, horrela morfologiatik syntaxiranzko bidea jauzirik gabekoa izateko (hau III. kapituluaren arrazoituko dugu).

Sistemaren eraginkortasunari buruz esan behar dugu hasierako gure helburua morfosintaxiaren deskribapen formalizatua egitea zela, abiadura bigarren maila batean utziz, baina beti sistema erabilgarria lortzeko. Egindako exekuzio-probetan, 10.832 hitzeko corpus batean analisi-denborak neurtu dira, Sun SPARC Ultra batean, II.2 taulan ikus daitezkeenak.

Deskribapena	Maiztegi a (formak + diskoa)	Tratatze ko hitzak	Hitz analizatu ak	Hitzak / segundoko (hitzaren gramatika hutsa)	Hitzak / segundoko (morfologia osoa)
Hitz-forma guztiak analizatuz	0	10.832	10.832	15,13	13,5
Hitz errepikaturik gabe, maiztegirik gabe	0	3.692	3.692	44	40
Hitz errepikaturik gabe, maiztegia erabiliz	10.000 (15 M)	3.692	1.483	111	95
	55.000 (75 M)	3.692	533	308	270

II.2 taula. Analizatzaile morfosintaktikoaren exekuzio-denbora.

Taula horretan bost lerro dauzkagu, kasu desberdinetan analisi-abiadura ematen dela:

- Lehen kasuan testuko hitz guztiak analizatu dira, banan-banan. Hau analizatzailearen kasu txarrena izango da, inplementazio-modurik sinpleena eginda. Honek 15 hitz/segundoko lortzen ditu hitzaren gramatikarako, eta 13,5 analisi morfologiko osoa egiten bada.

- Testuetan hitzak errepikatuta agertzen direla jakinda, bigarren neurketan testu horretatik hitz-forma desberdin bakoitza behin bakarrik analizatu da, lan bikoitza ekiditeko. Ikusten denez, abiadura hiru aldiz altuagoa da.
- Bigarren kasuan egin dugun bezala, pentsa daiteke gehien agertuko diren hitzak aldeztetik analizatuta egon daitezkeela, *maiztegia* deituko dugun datu-egitura batean. Horrela, testu berri bat hartzeko momentuan bakarrik testu horretako hitz berriak analizatuko dira. Soluzio honek datu-egitura baten erabilera eskatzen duenez, exekuzio-denborari egituraren kudeaketa ere gehitu egin behar zaio (maiztegiko hitz-formak eta euren analisiak ere gorde beharko dira). Soluzio honen proba bat egiteko 300.000 hitz-formako corpusa aukeratu genuen maiztegia ateratzeko (maiztegitako corpusak eta probarakoak ez dute ebakidura komunik, azken hau testu berria kontsidera dezagun). Bertan 55.500 hitz-forma desberdin aurkitu genituen. Emaizak maiztegiaren tamaina desberdinak probatuz lortu ditugu⁸.

Emaiz horietatik ondorioztatzen dugu lortutako analizatzailearen inplementazioa guztiz bideragarria dela taldearen beharrezanaren ikuspuntutik, eta gainera, lehen esan bezala, eraginkortasuna lehen mailako helburutzat hartu gabe. Hau guztia kontuan izanda, analizatzailea denbora errealean aplikatzeko moduan dago oraingo egoeran. Ondorengo atalean aipatuko dira (§ II.5) eraginkortasun hori hobetzeko ideia batzuk, jakinda alde horretatik oraindik irabazpen handiak lortzea badagoela.

II.6 Laburpena eta ondorengo pausoak

Laburbiltzeko, hauek dira sortu dugun analizatzaile morfosintaktikoaren ezaugarri nagusiak (Aduriz *et al.* 2000ab):

- Morfofonologia bi mailatako ereduaren bidez ebatzi da.
- Morfotaktikaren tratamendua banatuta dago: alde batetik egoera finituko ereduari, jarraitze-klaseen mekanismoaren bidez, eta bestetik hitzaren gramatikaren bidez.
- Hitz osoaren informazioa lortzeko, baterakuntzan oinarritutako hitzaren gramatika erabili dugu.

⁸ Emaizetan ez da agertzen maiztegiaren bilaketa egiteko denbora, bilaketa dikotomiko sinple bat nahikoa baitzen analisiaren aldean denbora askoz txikiagoa lortzeko.

Ikusten denez, morfotaktikaren lana, morfema/segmentuen arteko lotura posibleen deskribapena, bi ataletan dago banatuta: jarraitze-klaseen bidez eta hitzaren gramatika erabiliz. Banaketa era horretan egiteko hainbat arrazoi ditugu:

- Osagarritasuna. Nahiz eta bi faseetan osagaien ordena (morfotaktika) zehaztu, horrek ez du esan nahi bietan gauza bera errepikatzen denik. Bakoitzaren ekarpena zehaztekotan, egoera finituko morfotaktikak morfemen segmentazioa egiten du, morfemen arteko ordena finkatuz. Hitzaren gramatikaren morfotaktikak, berriz, sekuentzia horretako elementuei egitura hierarkikoa esleitzen die. Hau da, segmentatzeko orduan hitza morfemen segida gisa adierazten da, baina oraindik erabakitzeke dago morfema horiek nola konbinatuko diren (*'mendi + -aren + -zat + -ko + -a'*). Adibidez, hitz-elkarketa eta eratorpen-atzizkia agertzen badira (*kale-garbitzaile*), bi interpretazio eman litezke: *'tzaile'* atzizkia elkarketako bigarren osagaiari lotuz (*garbi*), edo elkarketa osoari. Momentuz anbiguotasun hau hitzaren gramatikan ebazten da, bertan eratorpen-atzizkiak beti osagai lexikal sinpleekin (*'garbi + -tzaile'* adibide horretan) lotzen baitira.

Dena dela, morfemen sekuentzia lineala hitzaren gramatikan ere modu inplizituan dago kodetuta (adibidez, izena eta kasu-atzikiak lotzen dituen erregelak beraien arteko ordena ere zehaztuko du). Zentzu honetan, morfotaktikaren zati bat bi aldiz kodetuta dagoela esan dezakegu.

- Eraginkortasuna. Aurretik esan den bezala, hitz osoaren analisisia emateko ezinbestekoa da hitzaren gramatika. Beraz, pentsa daiteke, (Ritchie *et al.* 1992) lanean egiten den modura, morfotaktika osoa gramatika horretan deskribatzea. Honek ez du inolako eragozpenik eta gainera, lan horretan aipatzen den bezala, erazagutzailetasuna gehituko lioke morfotaktikaren deskribapenari, jarraitze-klaseek eta azpilexikoiek erazagutzailetasuna ilundu egiten baitute. Baina morfosintaxi osoa hitzaren gramatikan ez egiteko arrazoi bat ere badago: morfotaktikaren egoera finituko inplementazioa eraginkorragoa da, eta horrela egiteak lan asko kenduko dio hitzaren gramatikaren bidezko analizatzaileari, analisirako aukera dezente baztertuz. Ingelesaren tratamenduan (Ritchie *et al.*, 1992) segmentazioak morfotaktika minimoa du, eta horregatik aukera asko sortzen dira (adibidez, atzizkia den morfema bat hitzaren hasieran kokatzen duen interpretazioa), gero hitzaren gramatikak baztertu egingo dituenak. Euskararen kasuan aukera-kopurua izugarri handia izango litzateke, eta tratamendua guztiz motelduko luke. Horregatik, morfotaktika segmentatzailean sartuz hitzaren gramatikak tratatuko dituen interpretazioak asko murrizten dira, eta abiaduraren igoera handia lortzen da.

- Modulartasuna. Analisi morfosintaktikoa bi modulu nagusitan banatu dugu eta beraien arteko interfazea argi definituta dago: hitza era sekuentzialean tratatuko da, lehenengo segmentazioa eta gero egituraren osaketa eginez.
- Praktikotasuna. Horrela lehendik garatutako tresna erabili da, zegoen bezala eta aldaketarik gabe. Dena dela, arrazoi hau ez litzateke pisuzkoa izan beharko aurreko hiru justifikazioak ez baleude, baina beraiekin bat etortzen denez, lana erraztea ekarri du.

Bukatzeko, esan behar dugu euskarazko hitzaren egitura analizatzeko sistema bat deskribatu dugula II. kapitulu honetan. Baina baieztapen horrek zuhurki adierazten du lan honen emaitza. Esan nahi baita, hitzaren egitura ez dela izan, ohitura den bezala, goi-mailako hizkera linguistiko batez egindako deskribapen hutsa, baizik eta formalismo baten beharretara moldatu eta inplementatu ere egin dela. Beraz, gramatika deskriptiboaz baino areago, gramatika baten inplementazio osoaz hitz egitea egokiago litzateke lan honen berri ematean.

Zuhurra da, gainera, hasierako baieztapena, hitzaren egitura esateak lan honen konplexutasuna erdi-ezkutuan uzten duelako. Hitza, inguruko erdaretan behinik behin, analisi-gai sinpletzat hartu ohi da. Hitzaren gramatika, horretara, fenomeno linguistiko benetan konplexuen tratamendurako abiapuntutzat hartzen da. Euskarazko hitzaren gramatikak, ordea, fenomeno linguistiko konplexuak ere tratatzen ditu. Esan daiteke, beraz, balio duela izen-sintagman eta adizlagunetan gertatzen diren fenomenoak, direnik eta konplexuenak, deskribatzeko. Horretan datza, besteak beste, lan honen garrantzia. Horregatik diogu, hasierako esaldia bere benetako mailara ekarriz, euskarazko hitzaren azterketatik abiatuta edonolako izen-sintagmaren (adizlagunaren) gramatika bat deskribatu eta inplementatu dela lan honetan. Hau hobeto ikusiko dugu III. kapituluan azalduko den gramatika sintaktikoaz hitz egiteko momentuan, morfosintaxiarekin egindako lana neurri handi batean syntaxian zuzenean integratu ahal dela azalduko baitugu.

Atal honetan deskribatutako sistema, nahiz eta hitz-mailako tratamendu morfosintaktiko osoa eta sendoa eman, ez da ikusi behar problemaren bukaeratzat, baizik eta oraindik sakonago jorratzeko dagoen bide bat gisa. Hauek dira egindako lanaren jarraipenerako aurreikusten ditugun bideak:

- Lehen esan dugun bezala, baterakuntzan oinarritutako analizatzailearen inplementazioa hobe daiteke. Egindako sistemaren helburu nagusia analizatzaile bat garatzea izan da, eta beraz indar nagusiak ez dira joan espazio edo abiadura bereziki azkartzera. Hala ere, egindako sistema erabilgarria da corpusen azterketarako.

Teknika bereziak erabiliz, eta agian inplementazio-lengoaia aldatuz, uste dugu oraindik etekin handiak lortuko direla.

Azkartzeko beste bide bat ezaugarri-egitura motatuen erabilerarena da, (Carpenter eta Penn 1993) lanean egiten den bezala. Mota emateak, nahiz eta konplexutasunaren igoera dagoela eman, konpilazio-prozesu baten ondoren analizatzaile azkarragoak ematen ditu. Honen moduko beste zenbait teknika (Kiefer *et al.* 1999) lanean aipatzen dira, guztira magnitude-ordena baten igoera lortzen dutela abiaduran (25etik 40ra aldiz azkarrago).

- Beste alde batetik, baterakuntza egoera finituko ereduaren bidez konpilatzeko proposamen desberdinak ari dira garatzen. Adibidez, (Beesley 1998a) lanean morfotaktikari buruzko murriztapenak baterakuntza formalismo sinplifikatu batean kodetzen dira, eta ondoren egoera finituko eredura konpilatzen dira, horrela abiadura azkartuz. Dena dela, azken automatarekin espazio-arazoak daude eta frogatzeko dago ea posiblea ote den ideia horren formalizazioa baterakuntza-murriztapen guztiak horrela tratatzeko. (Zajac 1998) artikuluan baterakuntza eta egoera finituko transduktoreak elkartu egiten dira analizatzaile hibrido bat sortuz.
- Hitzaren gramatikaren trinkotasuna eta orokortasuna hobetzeko, ezaugarrien portaera definitzeko erregela linguistiko orokorren azterketa interesgarria da, gramatika minimizatzeko eta aldaketak errazteko. Hau, (Ritchie *et al.* 1992) lanean egiten den antzera, gune- eta oin-ezaugarrien printzipioen antzekoak definituz lor daiteke. Mota honetako erregela baten definizioa jadanik egin dugula, honekin jarraitzeko asmoa dugu. Edozein kasutan, honek azterketa linguistiko sakona beharko du.
- Gramatikaren mantentzea. Analizatzaile morfosintaktikoa talde baten erabilera desberdinetarako baliagarria izateko, informazio-behar berrien tratamendura ere egokitu beharko da. Horretarako gramatika etengabeko aldaketa prozesuan egongo da, morfologia, sintaxia edo semantikaren tratamenduek eskatuko dutelako.

III Analizatzaile sintaktikoa

III.1 Sarrera

II. kapituluaren euskararen analizatzaile morfosintaktikoa aztertu dugu. Bertan ikusi dugu beste hizkuntza batzuetan esaldi-mailan gertatzen diren fenomenoak euskara bezalako hizkuntza eranskarietan hitz-mailan ere gertatzen direla (elipsiaren eta kasuen metaketaren modukoak). Hitzaren barruko konplexutasun hori onartuta ere, argi utzi nahi dugu oraindik badirela bi maila horien artean beste desberdintasun garrantzitsuak:

- Osotasuna. Esan dugunez, analizatzaile morfosintaktikoa sendoa da, eta edozein hitz-forma emanda, bere *interpretazio posible guztiak*⁹ lortuko ditu. Hau guztia hitzaren barruko fenomeno morfosintaktikoak guztiz deskribatuta daudelako da. Hitz-forma bat emanda gai gara zuzena ala okerra den bereizteko, eta bere egitura lortzeko. Esaldi-mailan, berriz, formalizazio teorikoa ez da osoa izan; horrela da gehien ikertu diren hizkuntzetan eta, are gehiago, euskara bezalakoetan. Linguistika teorikotik testu errealean analisira pasatzeko momentuan egoera zailagoa da, testu horietan linguistikoki deskribatu gabeko hainbeste egitura mota agertuko baitira (adibidez, berrogei hitzeko esaldi baten edo data-espresioen analisi sintaktikoa emateko orduan). Zenbait egilek (Sampson 1987) ohizko gramatiken baliagarritasuna zalantzan jartzera iritsi dira. Puntu horretara joan gabe, hizkuntza baten sintaxi osoaren deskribapena oraindik burutu gabeko lana da, eta sintaxiaren tratamenduak urte askotan bukatu gabeko ikerlerroa izaten jarraituko du.
- Anbiguotasuna. Hitz-mailan mugitzearen beste ondorio bat anbiguotasunaren problema saihestea izan da. Hitzaren muga baino harantzago joan gabe, ez da problema handia izango hitz bakoitzak bi edo hiru analisi posible izatea, baina helburua esaldi oso baten interpretazioa lortzea denean, hau arazo konplexua bihur daiteke. Lehen esan dugu sintaxiaren deskribapen osorik ez dagoela momentuz, baina deskribapen partziala emanda ere, edozein esaldi arruntentzat aukera asko sortuko dira, gehienak zentzugabeak testuingurua aztertuz gero. Beraz, esaldi-mailako tratamenduaren bigarren ikergai nagusia anbiguotasuna ezabatzearena

⁹ Egia esanda, ia guztiak, hitz ezezagunak eta aldaera dialektalak daudelako baina, lehen esan denez, hauek kontuan hartuz ere, emaitza %100etik oso hurbil dago.

izango da. Horren zailtasunaren neurri bat emateko, esan dezagun euskaraz hitz-mailako anbiguitasuna 2.6 interpretazio/hitzekoa dela, eta honi gramatika sintaktiko batek lortutako egiturak gehitzen bazaizkio, esaldi normal batek milioika aukera izango dituela. Helburu ideala esaldi bakoitzeko analisi bakarra lortzea izango da baina, ikusiko dugunez, analisi on hori zer den definitzea zaila izango da orokorrean, eta horregatik aplikazioaren arabera egin beharko da.

III. kapitulu honetan problema honi aurre egiteko gure proposamena deskribatuko dugu. Hasteko, sintaxiaren tratamenduari buruzko hurbilpen desberdinen azterketa egingo da (§ III.1.1). Ondoren, euskararen sintaxiaren ezaugarri nagusiak era laburrean aurkeztu ondoren (§ III.1.2), guk landutako bideak aurkeztuko dira (§ III.2 eta § III.3). Bukatzeko, bide horien konbinazio eta integrazioarako moduak aztertuko dira (§ III.4).

III.1.1 Sintaxiaren tratamendu automatikoa

Lehen esan den modura, sintaxiaren tratamendurako hurbilpenek bi arazo gainditu beharko dituzte: esaldi osoaren interpretazio posible guztiak identifikatzea eta beraien artean bat aukeratzea. Hauek biak modu bitara ebatz daitezke: ezagutza linguistikoa eskuz kodetuz edo automatikoki lortutako ezagutza erabili. Hiru aukera nagusi bereizten dira sintaxiaren prozesamenduan: eskuz kodetutako ezagutza linguistikoan oinarritutako hurbilpenak, estatistikan edo metodo automatikoetan oinarritutako lanak, eta bien konbinazioak. Ondoren alde bakoitzaren azalpena egingo dugu, abantailak eta eragozpenak aipatuz.

III.1.1.1 Ezagutza linguistikoan oinarritutako sintaxia

Ezagutza linguistikoa kodetzeko *gramatikak* erabili ohi dira. Hauek lehen aipatutako problema biak (analisiak lortu eta zuzena aukeratu) ebazteko erregelen edo prozeduren deskribapenak dira, gehienetan pertsona batek eginda. Gramatika bat idazteko orduan, irizpide desberdinak har daitezke, horietako batzuk elkarren kontrakoak eta besteak, aldiz, osagarriak direnak. Garatutako zenbait sistema aipatu aurretik, irizpide horiek definitu egingo ditugu:

- a) Hizkuntzaren teoria versus lengoia naturalaren prozesamendua (LNP). Shieber-ek (1986) dioenez, sistema batzuk teoria linguistikoak modelatzeko garatu dira, eta ez dituzte LNPrako sistemen helburu berdinak (Jensen 1988, Tomita 1988). Horrek ez du esan nahi analisi linguistikoak ez direla erabilgarriak LNPn, baizik eta formalismoen eta aplikazioen arteko lotura bat definitu behar dela (ikus III.1 taula). Egile horren ustez, LNPko sistemek gehienbat formaltasuna, erazagutzailetasuna eta

egokitasun linguistikoa mantendu behar dute. Antzeko gauza dio (Black *et al.* 1993:63) liburuak:

“... constructing a viable grammar for use in a computer analyzer of English is a very different process from constructing a “general linguistic theory”, or, to some extent, even a reference grammar of English for human use. It turns out that ‘Government-Binding Theory’, ‘Lexical-Functional Theory’, the theory of ‘Generalized Phrase Structure Grammar’, and their congeners, have little direct help to offer the practical grammarian, although they might indicate to some workers certain broad lines to approach. The large descriptive grammars that exist, ..., are vastly more data-driven than modern theorizing, and therefore more helpful for our purposes.”

	Hizkuntzaren teoriako formalismo sintaktikoak	Aplikazio errealetarako gramatika konputazionalak
Datuak	datu gutxi, “laborategikoak”	datu asko
Esaldien jatorria (orokortasuna)	esaldi asmatuak	esaldi errealak
Esaldien zuzentasuna	esaldi gramatikalak	esaldi ez-gramatikalak ere analizatu behar dira
Desanbiguazioa, puntuazioa, ...	ez dira tratatzen	tratatu behar dira

III.1 taula. Teoria sintaktiko eta implementazioen ezaugarriak.

Analisi sakonago baten ondorioz, gero aztertuko dugun bezala, hizkuntzaren teorien arabera aplikatutako sistemek bi aspektu nagusi izan beharko lituzkete kontuan. Batetik, beharrezkoa da lexikalizazioa¹⁰ (batez ere, azpikategorizazioaren tratamendua), osagai sintaktikoen egitura lexikoian sartuz, LFG edo HPSG (Sells 1985) moduko formalismoen antzera. Bestetik, informazio partzialeko egiturak (Shieber 1986) oso lagungarriak dira, hiztegiko elementuak nahiz osagai sintaktikoen informazioa ezaugarri-balio bikoteen bidez adierazteko. Ikusiko denez, dimentsio honetako bi alderdiak, teoria eta aplikagarritasuna, ez dira elkarren kontrakoak, eta bien konbinazioa posiblea eta komenigarria da.

- b) Testuingururik gabeko gramatikak (TGG) versus egoera finituko mekanismoak. Gramatikak idazteko orduan, notaziorik erabiliena testuingururik gabeko gramatikena izan da (Aho *et al.* 1986), esaldien egitura hierarkiko eta errekurtsiboak

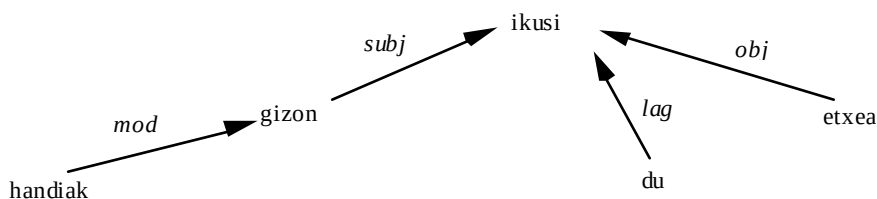
¹⁰ Lexikalizazioa hitza erabiltzen dugunean lehen syntaxian kodetzen ohi zen informazioa lexikoan gehitzeko joera adierazi nahi dugu. Adibidez, HPSG formalismoak ez du ia gramatikarik definitzen, analisi sintaktikorako behar den informazio gehiena lexikoan dagoelako.

definitzeko egokiak baitira. Egoera finituko mekanismoak (automatak eta transduktoreak), nahiz eta aspalditik ezagunak eta erabiliak izan informatikako alor askotan, TGGen itzalean geratu dira, ahalmen deskriptibo mugatuaren aitzakiarekin. Dena dela, azken urteotan lortu diren emaitza matematiko eta algoritmikoei esker (Roche eta Schabes 1997), lengoaia naturalaren prozesamendu aplikatua eta ikerketan teknologia ahalmentsu eta eraginkorra bihurtzen ari dira, eredu linguistikoetan oinarritutako patroik edo txantiloiak definitzeko erraztasuna emanez.

- c) TGG sinpleak versus baterakuntza oinarritutakoak. TGG sinple batean erregelek osagai atomikoak besterik ez dituzte deskribatzen. Honen ondorioz egitura sintaktiko sinpleenak (adibidez, izen-sintagma bat komunztadurak kontuan hartuz) zehazki definitzeko dozenaka erregela beharko lirateke. Hori ez gertatzeko, gramatikaren osagaiei informazioa gehi dakieke ezaugarri-egituren bitartez (Shieber 1986), horrela gramatikaren trinkotasuna eta sinpletasuna bultzatuz. Informazio linguistiko hori erabiltzeko baterakuntza izaten da eragiketarik inportanteena. Irabazpen honen kontra baterakuntza-ekuazioen kalkuluaren denbora-kostua dugu, eraginkortasuna moteldu egiten baita.
- d) Esaldiaren analisi osoa versus analisi partzialak. 1980ko hamarkadan, uste nagusia LNPko edozein lan, desanbiguazio morfologikoa kasu, esaldiaren ulerkuntzaren zati gisa ikustea zen, hau da, analisi sintaktikoa, analisi semantikoa, berbaldiaren analisisa eta munduaren ezagutza beharrezkoak zirela pentsatzen zen. Baina testu errealean erabilera zabaldu zen heinean, uste horren zati bat ezeztatuko zuten desanbiguazio morfologikorako tresna desberdinak garatu ziren, ulerkuntza osoa lortu gabe emaitza erabilgarriak lortzen dituztenak. Etiketatzearen bidean, *analizatzaile partziala* (Abney 1997, Basili *et al.* 1998) terminoak teknika desberdinen multzoa definitzen du. Analizatzaile mota hauek analisi sintaktiko tradizionalaren informazioaren zati bat, ez guztia, lortzen dute. Teknika hauek fidagarritasuna eta sendotasuna dute helburu, sakontasuna eta osotasuna neurri batean galduz. Beste aldetik badira sistemak, teoria linguistiko baten azterketarako pentsatuak, *analisi osoa* bakarrik lortzea helburu dutenak, fenomeno linguistiko interesgarriak aztertzeko pentsatuta daudelako, eta ez testu errealeko esaldiak.
- e) Ikuspegi eraikitzailea versus ikuspegi murriztailea. Analisi sintaktikoari ekiteko, lehenago esan dugu alde batetik egitura posibleak aztertu behar zirela (alde hau eraikitzailea kontsidera dezakegu), eta beste batetik aukera guztietatik bakarra aukeratu (alde murriztailea). Bi ikuspegi horien antolaketaren arabera, analizatzaile desberdinak garatu dira. Mutur batean, formalismo murriztaile hutsek lexikoiko informazioa erabiliz esaldia analizatzeko aukera posible guztiak sortzen dituzte, eta gramatikariaren lana onartezinak baztertzea da. Beste muturrean formalismo

eraikitzaile hutsa edukiko genuke, hitzetatik abiatuta, beraien konbinazioaz bakarrik esaldiaren azken egitura eraikitzen saiatzen dena, era deterministan alferrikako aukerak sortu barik (Hermjakob eta Mooney 1996). Ondoren ikusiko dugunez, gramatika/analizatzaile askotan bien arteko konbinazioa egingo da, egitura sintaktiko guztiak ez daudelako hasieratik sortuak, eta gainera azken analisiaren parteak ez direnak ere sortu egiten direlako.

- f) Osagai-egitura (*constituency-based*) versus mendekotasun-egitura (*dependency-based*). Osagai-egituraren bidezko analisisian emaitza lortutako osagaiak eta beren kategoriak definituz ematen da (izen-sintagmak, esaldiak, ...). Mendekotasun-egituraren kasuan (Järvinen eta Tapanainen 1998), berriz, osagaien arteko erlazioak deskribatzen dira (ikus III.1 irudia). Bi mutur hauen artean tarteko bideak ere erabiltzen dira. Adibidez, (Basili *et al.* 2000) artikuluan mendekotasun-egitura erabiltzen da esaldiaren oinarritzko osagaiak konbinatzeko (izen-sintagmak, preposizio-sintagmak eta aditza), baina horien barruan osagai-egitura lortzen da, mendekotasuna hitzetaraino eraman gabe.



III.1 irudia. 'gizon handiak ikusi du etxea' esaldiaren mendekotasun-egitura.

Ezaugarri horiek guztiak gure hasierako azterketa honetarako interesgarrienak direlako aukeratu ditugu, baina badaude sintaxiaren tratamendurako beste aspektu aipagarriak, horien artean analisirako estrategia (goitik beherakoa, behetik gorakoa edo mistoa) edo domeinu jakinetako ezagutza semantikoaren erabilera.

III.2 taulan sistema batzuk daude, bakoitza bere ezaugarri nagusien bidez deskribatuta. Irudiaren ondorengo lerroetan zenbait sistemaren ezaugarri nagusiak aztertzen saiatuko gara, aldeztetik jakinda analisi sintaktikorako sistemen zerrenda ikaragarri luze izan litekeela eta guk edo garrantzitsuenak edo gure lanetan eragin handiena izan dutenak bakarrik aipatuko ditugula. Deskribapena errazteko, lehen aipatutako b) dimentsioaren arabera sailkatu ditugu sistemak, hau da, testuingururik gabeko gramatiken edo egoera finituko mekanismoen erabilera kontuan hartuz.

		LFG (XRCE)	ANLT gramm ar	PLNLP	MG	XFST
a	Hizkuntzaren teoria	X	X			
	Lengoaia naturalaren prozesamendua	x	X	X	X	X
b	Testuingururik gabeko gramatikak	X	X	X		
	Egoera finituko mekanismoak				X	X
c	TGG sinpleak					
	Baterakuntzan oinarritutako TGGak	X	X	?		
d	Esaldiaren analisi osoa	X	X	X	x	x
	Esaldiaren analisi partzialak				X	X
e	Ikuspegi eraikitzailea	X	X	X		X
	Ikuspegi murriztailea	X	X	X	X	X
f	Osagai-egitura	X	X	X		X
	Mendekotasun-egitura			X	X	X

III.2 taula. Sintaxiaren tratamendurako zenbait sistemaren ezaugarriak.

III.1.1.1.1 Testuingururik gabeko gramatiketan oinarritutako sistemak

Alde batetik, hizkuntzaren teorian oinarritutako sistema desberdinak ditugu, gehienek eragiketa nagusi gisa baterakuntza erabiltzen dutela. Lehen esan denez, hauetako askoren helburu nagusia lengoaiaren teoria garatzea izan da, eta lengoaia analizatzeko tresnen eraikuntza bigarren mailan utzi da. Dena dela, kasu batzuetan teoria eta aplikagarritasuna lotzeko ahalegin berezia egin da, sistema eraginkorrak lortzeko asmoarekin.

Lexical Functional Grammar (LFG; Bresnan 1982) teoriaren analizatzaile desberdinak garatu dira. Teoriaren oinarria baterakuntza da, eta formalismo eraikitzailea da gehienbat, nahiz eta analisi posibleen ugalketaren ondorioz alde murriztailea ere gehitu behar izan duten. Inplementatutako sistemen artean *Xerox Research Centre Europe* (XRCE) taldeko lana (Maxwell eta Kaplan 1996, Brun 1998, Frank *et al.* 1998, Kuhn 1998, Butt *et al.* 1999) izan daiteke aipagarriena. Sistema honetan era paraleloan estaldura zabaleko LFG gramatikak ari dira garatzen ingeles, frantses eta alemanerako. Teoria eta praktika lotzeko lehen printzipiotzat analisisen motibazio linguistikoa dute, baina tratatzen diren egitura linguistikoak

testu-corpusen maiztasunen arabera erabakitzen dira, eta eraginkortasunaren galera ekar dezaketen fenomenoentzat soluzio bereziak (pragmatikoak beraien esanetan) landu izan dira. Horrekin batera analizatzailaren probarako corpusak prestatu dira eta garapenerako ingurune batean integratu.

Alvey Natural Language Tools (ANLT) ingeleserako analizatzaila eta gramatika dugu (Carroll 1993, Grover *et al.* 1993). Sistema honetan estaldura zabaleko gramatika sintaktiko eta semantikoa landu da, *Generalized Phrase Structure Grammar* (GPSG; Gazdar *et al.* 1985) baterakuntza-formalismoan oinarrituta. Ingelesaren egitura linguistikoen multzo zabala deskribatzen du, eta corpusen tratamendurako landu da bereziki, alde teorikoa eta praktikoa elkartuz. Inplementaziorako, behetik gorako LALR algoritmoa erabili da (Tomita 1986, Briscoe eta Carroll 1993) eraginkortasuna hobetzeko.

Head-Driven Phrase Structure Grammar (HPSG; Pollard eta Sag 1994, Borsley eta Przepiórkowski 1999) azken urteotan indar handia hartu duen lengoiaren teoria da, beste teoria askotatik (GPSG eta LFG barne) ideia desberdinak hartu dituen. Formalismoa baterakuntzan oinarrituta dago, teoria honetan lexikoari testuingururik gabeko gramatikari baino garrantzi handiagoa eman zaio, erregelak oso eskematikoak baitira, eta zenbait sistema inplementatu dira lengoia desberdinetarako (Uszkoreit *et al.* 1994, Kiefer eta Krieger 2000). Sistema horien asmo nagusia lengoiaren teoria garatzea izan da gehienbat.

PLNLP sistema (*Programming Language for Natural Language Processing*, Jensen 1988, Jensen *et al.* 1993) ez dago lengoiaren teoria jakin batean oinarriturik, asmoa sistema aplikatua eta malgua egitea baitzen. Egile batzuk hizkuntzalariak zirenez, arreta berezia jarri zen alde linguistikoan, teorietatik probetxagarriena hartzeko eta zailtasunak ematen zituzten aspektuak saihestu ziren. Horregatik testuingururik gabeko gramatika bat erabiltzen da oinarrian, baterakuntzaren antzeko ekuazioak adierazteko programazio-lengoia berezi batekin lotuta. Sistema, zenbait aplikaziotan probatu zen, horien artean gramatika-zuzentzailea eta hiztegi-definizioen analisisetan.

Tree-Adjoining Grammar formalismoa (TAG, Joshi 1985) ez dago ohizko TGG batean oinarrituta, zuhaitz-egitura partzialetan baizik. Zuhaitz horiek konbinatuz esaldien analisiak lor daitezke. Formalismo hau TGGak baino ahaltsuagoa da. Honen aldaera nagusia Lexicalized TAG da (LTAG, Schabes eta Joshi 1991, Doran *et al.* 1994, XTAG 1995), zuhaitzak lexikoiko sarrerekin lotuz, horrela testuinguru lexikalak lortzeko.

FIDDITCH (Hindle 1989) izan liteke analisi partziala egiteko sistemen arteko ‘zaharrenetarikoa’ eta baita arrakastatsuenetarikoa ere. Bere helburu nagusia mugarik gabeko testuak aztertzea zen, informazio mota desberdinak ateratzeko. Analizatzaileak esaldi baten osagai nagusiak ezagutzen ditu: sintagmen arteko mugak, edo subjektua eta predikatua. Arrakastaren arrazoiaren artean, batetik, anbiguotasuna ebazteko sistema aipa daiteke, ebatzi ezinezko anbiguotasunak ekiditen direlako (adibidez, preposizio-sintagmak ez dira lotzen ez baldin badago hori egiteko informazio nahikorik), eta bestetik, erregaletan oinarritutako determinismoa: sistemak ez du eraikitzen azken analisira ez daraman hipotesirik, honela abiadura azkartuz.

TACAT (Atserias *et al.* 1998) corpusak morfologikoki eta sintaktikoki etiketatzenko analizatzaile partziala da: hiru lengoaia desberdinetarako (gaztelania, ingelesa eta katalanerako) gramatikak definitu dituzte. Gramatika TGG sinple batean oinarrituta dago, eta asmo nagusia etiketatutako corpus horiek aplikazio desberdinetarako erabiltzea da.

Link Grammar Parser (Sleator eta Temperley 1993, Grinberg *et al.* 1995) ingeleseko analizatzaile sintaktiko bat da, sintaxiaren teoria berezi batean oinarrituta dagoena. Teoria horretan esaldi baten analisia hitz-bikoteen arteko estekadura-kate bat da, eta gramatikaren pisu guztia lexikoian kodetutako hitzen arteko konbinazio posibleen estekadurak dira. Alde horretatik esan liteke gramatika ‘gutxi’ duela formalismo honek, baina hala ere TGGen atal honetan sartu dugu. Honela lortzen den analisia hitzen arteko mendekotasun-egitura izango da. Analizatzailea sendoa da, lexikoian duelako pisua, eta horrela analisi partzialak, hitz ezezagunak eta antzekoak erraz trata daitezkeelako.

Aipatutako sistemez gain, beste asko daude, adibidez, *Core Language Engine* (Alshawarwa eta Moore 1992), *Government and Binding* (GB, Berwick *et al.* 1991), PC-PATR¹¹ (Antworth 1994), *Categorial Grammar* (Uszkoreit 1986), *Word Grammar* (Hudson 1990), *Slot Grammar* (McCord 1990), *Augmented Transition Networks* (ATN, Bates 78), edo (Giguët eta Vergne 1997), (Ciravegna eta Lavelli 1997). Hauen eta aurrekoen arteko konparaketa eginda ondorio nagusitzat hau har dezakegu: bai teoria linguistikoak bai sistema aplikatuak sintaxiaren pisu nagusia lexikoian jartzen ari direla, bertan informazio konplexua edo espezifikoa kodetuz, horrela erregela sintaktikoen orokortasun eta sinpletasuna handitzen dela.

¹¹ <http://www.sil.org/pcpatr>

§ III.1.2n sakonago aztertuko dira baterakuntzan oinarritutako analizatzaile sintaktikoak, beti ere euskarari aplikatzeko bideragarritasunaren ikuspuntutik.

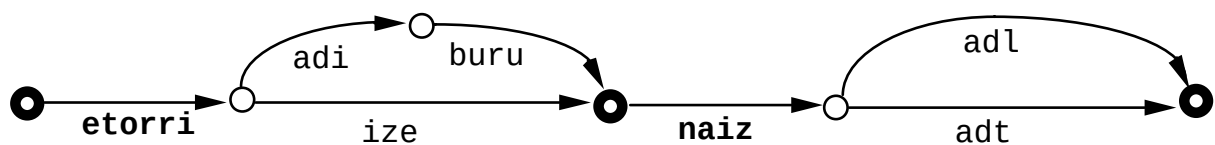
III.1.1.1.2 Egoera finituko mekanismoetan oinarritutako sistemak

Egoera finituko mekanismoen artean egoera finituko automatak eta transduktoreak ditugu (Roche eta Schabes 1997). Mekanismo hauek oso erabiliak izan dira informatikako alor desberdinetan, baina orain dela gutxira arte ez dira egokitzen hartu linguistika konputazionalerako eginkizun nagusietan: hiztegien kodeketan, testuen prozesamenduan eta ahotsaren prozesamenduan. Hala ere, azken urteotan egindako lanen ondorioz, esan daiteke teknika hauek etorkizun oparoa izan dezaketela lehenago TGGen bidez egiten ziren eginkizun askotan, beraien inplementazio eraginkorreki esker.

Egoera finituko automatak eta transduktoreak dira hurbilpen honen funtsa, batzuk lengoaiak eta besteak lengoaien arteko erlazioak definitzeko. Horien zehaztapena egiteko adierazpen erregularren lengoaia erabiliko da. Hauek dira beraien ezaugarri aipagarrienak (Roche eta Schabes 1997):

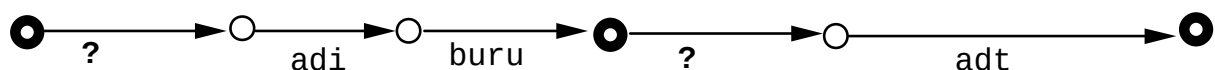
- Homogenotasuna. Tratamendu sintaktikoaren elementu guztiak egoera finituko automata/transduktoreen bidez egiten dute: hiztegia, gramatika, sarrerako esaldia eta analisiaren emaitza. Homogenotasun honek sinpletasunaren abantaila izango du orain arte garatu diren sistema askoren aurrean, hauetan normalean bakoitza modu desberdin batean kodetzen baita: sarrerako esaldia testu gisa, morfologia bi mailatako formalismoan eta sintaxia LFG edo HPSG moduko formalismoan. Gainera adierazpen erregularrak, automatak eta egoera finituko kalkulua notazio ezagunak dira eta finkatuak daude aspalditik hizkuntzalaritzan zein informatikan.
- Malgutasuna eta modularutasuna. Egoera finituko sistemetan gramatikak era modularrean osa daitezke, automaten propietate matematikoei esker. Horrela, desanbiguaziorako murriztapenak bildura erabiliz multzoka daitezke, automaten eragiketen itxidura-propietatei esker. Horren antzera, egoera finituko lengoaiak bata bestearen barruan egotearen problema (erregela berri bat asmatzen denean, adibidez) erantzun daiteke, automaten erabakigarritasunari (*decidability*) dagozkion emaitzak daudelako (hau ezinezkoa da testuingururik gabeko gramatikekin).
- Automaten propietate algoritmikoei esker (determinista bihurtzeko eta minimizatzeko garrantzitsuenak), gramatikak era trinkoan gorde daitezke, eta modu eraginkorrean aplikatu ere.

Puntu honekin hasteko, mekanismo hauen ahalmena adibide batzuen bidez azalduko dugu. III.2 irudian egoera finituko automata bat dugu, esaldi sinple bat adierazteko. Bertan bi hitz agertzen dira, bakoitza bi interpretaziorekin. Adibidez, lehen hitzak (*etorri*) bi interpretazio dauzka, bakoitza bere informazio morfologikoaren bidez adierazita. Lehenengo interpretazioak aditza eta burutua adierazten du, eta bigarren interpretazioak izena. Bigarren hitzak (*naiz*) aditz laguntzailea (adl) eta aditz trinkoaren (adt) interpretazioa ditu. Guztira lau bide desberdin daude automata hori zeharkatzeko, hau da, lau interpretazio desberdin dauzka anbigua den esaldi horrek. Esaldiaren informazioa horrela emanda, modu laburrean azalduko ditugu egoera finituko sistemen eragiketa nagusiak, hauek lehen aipatutako e) irizpidearen arabera banatuta: ikuspegi murriztailea eta eraikitzailea.



III.2 irudia. Esaldi sinple baten automataren adibidea¹².

Ikuspegi murriztailean oinarritutako formalismoetan, hasieran interpretazio posibleen informazioa esaldiaren automatan kodetzen da, eta gramatikariaren lana interpretazio horietatik bakarra aukeratzea izango da, aukerak baztertuz. Honetarako bi eratako murriztapenak erabiltzen dira gehienbat: ezinezkoak diren testuinguruak debekatzen dituztenak edo testuinguru zuzenak aukeratzen dituztenak. Bi murriztapen mota hauek osagarriak direnez, askotan gramatikariaren gustuaren arabera egongo da bata ala bestea erabiltzea. Debekuen kasua aukeratuz gero, III.3 irudiko testuingurua ez-onargarritzat kontsideratzeko automata defini dezakegu¹³.

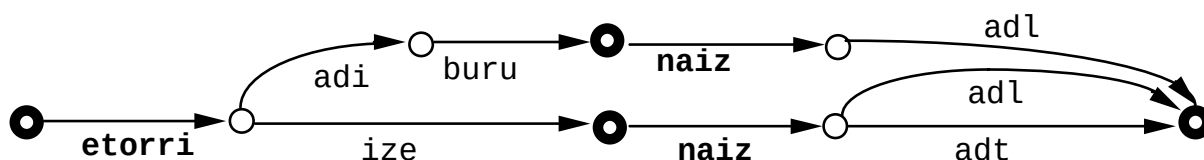


III.3 irudia. Testuinguru ezinezkoa debekatzeko automata.

¹² Sinplifikatzearen, automatan biribil lodia aukeratu da hitz-formen arteko muga adierazteko, eta biribil normala hitz-forma eta informazio morfosintaktikoa bereizteko. Benetako automata batean hori hitz- edo morfema-muga adierazteko ezaugarri berezi baten bidez egingo litzateke. Ezaugarri morfosintaktikoentzat EDBLko izenak erabili dira.

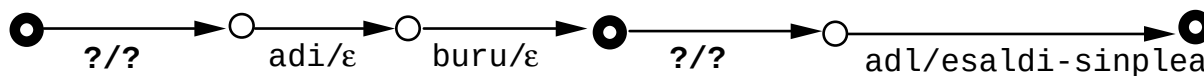
¹³ Adibidean automaten notazioko galdera-marka erabili da, edozein elementu adierazteko. Kasu horietan, horrek formekiko independentzia emango dio debekuari. Adibide hauek ilustrazio moduan jarri dira, eta ez dugu esango testu errealean baliagarriak izango direnik.

Irudi horretako automatak debekatu egingo luke¹⁴ aditz burutua agertzea aditz trinko baten aurrean, horrek (esaldi sinpleetan behintzat) bi aditz desberdin egotea ekarriko lukeelako (**ekarri dago**, adibidez). Esaldiaren eta debekuaren automaten ebakidura atereaz gero, III.4 irudiko automata izango genuke, hiru interpretazio dituena. Behar adina debeku-kopurua definituz gero, azkenean esaldi bakoitzeko analisi bakarra lortzea izango da helburua.



III.4 irudia. Esaldiaren automata murriztapenaren aplikazioaren ondoren.

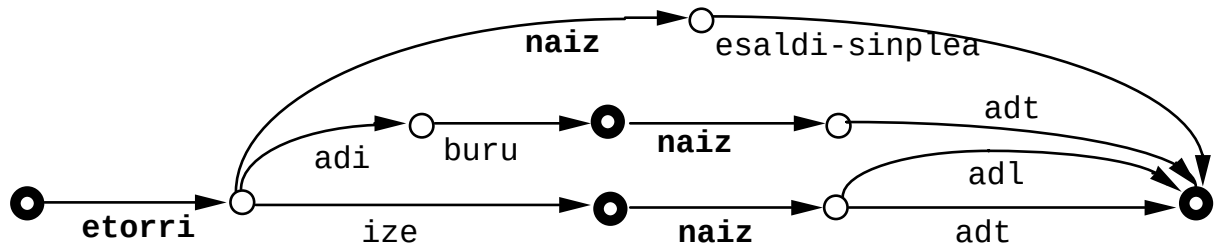
Ikuspegi eraikitzailean oinarritutako formalismoetan, berriz, esaldi baten hasierako informazioa hartuta emaitza gisa lortu nahi diren egiturak eraiki edo sortu egingo dira. Hori egiteko transduktoreak (*transducer*) izeneko automata bereziak erabiliko dira, III.5 irudikoa kasu.



III.5 irudia. Esaldi sinpleak lortzeko transduktorea.

Ikusten denez, transduktore horretako arku bakoitzean bi ikur jarri dira. Lehenengoak sarrerako lengoaia definitzen du eta bigarrenak irteerako lengoaia (goiko eta beheko lengoaia ere deitzen dira). Automata horren lana sarrerako lengoiaren instantziak irteerako lengoiakoekin ordeztzea izango da. ϵ ikurrak kate hutsa adierazten du, hau da, beraren bitartez nahi ez diren ikurrak irteeratik ezabatzen dira. III.5 adibideko transduktoreak aditz burutua eta ondoren aditz laguntzailea aurkitzean esaldi sinple baten arkua gehituko du. III.2 irudiko automata sarreratzat hartuta, III.6 irudian ikusten den emaitza lortuko litzateke. Modu horretan nahi diren egiturak sor daitezke.

¹⁴ Egia esanda, III.3 irudiko debeku hori adierazteko testuinguru hori ez den edozein kate deskribatzen duen automata (hau da, automataren osagarria) eman beharko genuke, baina automata hori konplexuagoa denez bere kontrakoa jarri dugu. Lengoaia erregular baten osagarria beste lengoaia erregular bat denez, ez da inolako arazorik egongo.



III.6 irudia. Transduktorearen aplikazioa.

Murritzapen-gramatika (MG) (Voutilainen eta Tapanainen 1993, Karlsson *et al.* 1994, Voutilainen 1994ab) formalismoa ikuspuntu murritzaitetik definitu da, eta azken urteotan azaleko sintaxia eta desanbiguazioa lortzeko egindako sistemetatik arrakastatsuenetarikoa bihurtu da, oso ezaugarri desberdinetako lengoaietara aplikatu delako. Esaldi bat emanda, lehen pausoa hitz-forma bakoitzari etiketa morfosintaktiko posible guztiak gehitzea da. Ondoren, murritzapen-erregelen multzoa aplikatzearen helburua hitz-forma bakoitzak interpretazio bakarra eta zuzena izatea da. Murritzapen-erregela horiek lehen aurkeztu ditugun debekuen modukoak edo interpretazio zuzenak aukeratzekoak (beraz, beste aukera batzuk baztertuz) izango dira. Eraitzan, beraz, hitz bakoitza morfosintaktikoki desanbiguatuta egongo da. Formalismo honen azalpen luzeagoa eta bere euskararen aplikazioa geroago (§ III.3.1) emango dugu.

Sintaxiaren bidetik jarraituz, *Finite State Syntax* izeneko formalismoa (Koskenniemi *et al.* 1992, Voutilainen 1994b) landu da, MG aberastuz, alde batetik formalismoaren zabaltzearekin, eta bestetik errepresentazio gramatikaren aberasketarekin. Formalismo honek, MGk bezala, hasieran informazio guztia dauka eskura, hitz bakoitzari etiketa morfosintaktikoen bidezko analisi posibleak jarritz, anbiguotasun-tasa oso handiarekin, eta murritzapen linguistikoen bidez aukera horiek ezabatzen joan beharko dira. Bi formalismo hauen abiapuntua lexikoia denez, gramatikariaren lan handienetarikoa deskribapen lexikal egokiak asmatzearena izango da.

Azken honen ildo beretik jarraituz, Xerox-eko ikerketa taldeak adierazpen erregularren¹⁵ bidezko tresna linguistikoak (*XFST* edo *Xerox Finite State Tool* izenekoa) garatu ditu (Karttunen *et al.* 1997, Aït-Mokhtar eta Chanod 1997, Chanod eta Tapanainen 1996ab). Adierazpen erregularrak analisi morfologikorako erabiltzen ziren orain dela gutxira arte, baina dagoeneko beste lan batzuetara aplikatuak izan dira, tokenizaziotik hasita azaleko analisi

sintaktikora arte. Tresna hauetan lehenago azaldutako bi ikuspegiak erabiltzea dago: eraikitzailea eta murriztailea. Eredu honi ‘formalismoa’ deitzea gehiegi izan liteke, ez baita funtsezko elementu teoriko berririk landu, automaten eta adierazpen erregularren teoriatik aparte, baina garapen matematikoa eta algoritmoen inplementaziorako ekarpen berriak egin dira, eta horrela aplikagarritasuna erraztu da. Lortutako tresnek malgutasuna, adierazpen erregularren sinpletasuna eta teoria matematiko baten sendotasuna dituzte ezaugarri nagusi aipagarriak. Tresna hauei buruz luzeago arituko gara euskararen gainean egindako aplikazioaren kapituluetan.

Oflazer-en (1999a) lanak Xerox-eko taldearen bidea jarraitzen du turkierarako egindako analizatzaille batentzat. Egoera finituko transduktoreen bidez mendekotasun-egitura bat lortuko du esaldi batetik abiatuta. Analizatzaille hau lehen fase batean dago, baina guretzat interesgarria izango da turkiera, euskara bezala, hizkuntza eranskaria delako.

INTEX sistema (Gross 1997, Silberztein 1997) antzeko ideietan oinarritzen da, baina enfasia lexikoiaren deskribapenean jarrita. Hasteko, oinarrizko forma lexikalen eta hitz konposatuen estaldura zabaleko hiztegiak landu dira. Horrez gain, *gramatika lokalak* deitutako egiturak daude, beraien bitartez esaldien zati berezi batzuk definitzeko: termino teknikoak, esaldi lexikalizatuak, eta lengoaia tekniko berezietako esaldiak. Berdin egin da datak, iraupenak, maiztasunak eta antzekoak definitzeko. Bukatzeko, *Lexicon-Grammar* izeneko informazioa du sistemak, bertan aditz bakoitzaren azpikategorizazio-ereduak gordetzeko, bai argumentuen ezaugarri sintaktikoak, bai argumentu lexikalizatuak. Sistema, egoera finituko automatetan oinarrituta dago, eta lexikoiaren sistematizazio handiak testu errealei aplikatzeko gaitasuna ematen dio. Sistema horren antzekoa dugu FASTUS (Hobbs *et al.* 1997).

III.1.1.2 Teknika probabilistikoetan oinarritutako sintaxia

Hurbilpen probabilistikoa (Black *et al.* 1993) indar handikoa bihurtu da azken hamarkadan, aurreko lanetan gramatikariek egiten zituzten atazak automatikoki egiteko. Sistema hauen ezaugarri nagusien artean hauek ditugu:

¹⁵ Automaten teoriaren emaitza jakina da egoera finituko automatak eta gramatika erregularrek lengoaia erregularren multzoa definitzen dutela, hau da, baliokideak dira adierazpen-ahalmenaren ikuspuntutik.

- Corpus etiketatuen beharra. Analizatzaile mota hauetan lan gehiena corpus etiketatuetatik (corpus hauei *treebank* edo *parse bank* deitzen zaie) ateratako probabilitateen bidez egiten da, hau da, gramatikak garatzeko orduan eskuzko lan gramatikal minimoa egiten da, ezagumendu linguistikoa corpusean agertzen diren elementuetatik (eta beren maiztasunetatik) ateratzen baita. Corpus horiek ingeleserako landu dira gehienbat (*Brown Corpus*-ak kategoriak etiketatuta ditu, edo *Penn Treebank*-a (Marcus eta Santorini 1991) sintaktikoki etiketatuta dago). Adibidez, *tagger* edo etiketatzaileetan probabilitate lexikalak kategoria edo hitzen bigrama edo trigramen bidez ateratzen dira (Church 1988, Garside *et al.* 1987, Brill 1995, Charniak 1993), gero testu berrietan aplikatzeko. Esan behar da corpus bat etiketatzea lan luzea eta zaila dela, informazio gramatikala automatikoki ateratzen denean ere. Gainera, gertaera linguistikoaren deskribapen zabala izateko, corpusak tamaina handia izan behar du. Adibidez, kategoria sintaktikoen trigrametan oinarritutako etiketatzaile batek, 10 kategoria ezberdin edukiz gero, 1.000 trigrama posible izango lituzke, eta milioi bat hitzeko corpus batean, fenomeno askoren agerpen-maiztasuna txikiegia gerta liteke. Etiketatzaileekin arazo hau gertatzen bada, are gehiago sintaxiaren tratamenduan, lengoaiaren eredu askoz aberatsagoa landu nahi delako.
- Azaleko sintaxia. Sistema probabilistiko gehienetan azaleko analisisa egiten da, etiketatzaileetan adibidez (hitz bakoitzaren kategoria sintaktikoa igarri behar da). Nahiz eta egitura sintaktiko osoak lortzeko zenbait lan egin (Atwell 1987, Bod 1993), oraindik frogatzeke dago estatistika hutsean oinarritzen den analizatzaileen bideragarritasuna.
- Muga gaindiezinak. Sistema hauek orain arte gaindiezinak izan diren mugak dauzkate. Adibidez, etiketatzaile estatistikoetan %95-97 inguruko neurriak (Voutilainen 1994a, Brill eta Wu 1998) agertu dira zenbait lengoaiatarako. Nahiz eta neurri hori ona izan lehenagoko etiketatzaileekin konparatuz gero, horrek problema bat suposatuko du edozein analizatzailerentzat, zenbaki hori muga maximotzat onartuko bagenu esaldi askotan errore bat egotea suposatuko lukeelako. Etiketatzaileak une honetan metodo probabilistikoekin lortu diren tresna hoberenak izanda, ezin izango dira espero emaitza hain onak sintaxi osoaren tratamendu probabilistikoan.

III.1.1.3 Teknika linguistiko eta probabilistikoen konbinazioak

Estatistika hutsa erabiltzeak arazoak izan ditu testuinguru mugatuetan gertatzen ez diren fenomenoak tratatzeko. Adibidez, trigrametan oinarritutako sistema batean zailtasuna dago

hiru hitz baino gehiago hartzen duten gertaera linguistikoak aztertzeko orduan, aditza eta osagarri nagusien artekoak kasu. Ez, ordea, inguru hurbileko erlazioak, izena, adjektibo eta determinatzaileen artekoak bezalakoak. Gainera, ikuspuntu estatistiko hutsean oinarritutako gramatikekin lortutako analisiek beste arazo bat dute, emaitza horiek linguistikoki interpretatzea ez baita erraza, eta horrek zailtasun handiak jar diezazkioke ondorengo prozesuei, interpretazio semantikoa kasu. Linguistek idatzitako gramatiketan, aldiz, maila altuko gertaera linguistikoak deskribatu dira gehienbat, sintagmak zein esaldi osoak konbinatzeko, baina arreta gutxiago eskaini zaio esaldi errealean agertzen den zenbait fenomenori, egitura jakin baten maiztasuna kasu. Horregatik metodo probabilistikoak eta ezagutza linguistikoa lotzeko saioak egin dira, bakoitzaren abantailak biltzeko asmoz.

Adibidez, (Black *et al.* 1993) lanean hizkuntzalariek egindako gramatika bat erabiltzen da, baina erregelen aplikazioa sintaktikoki etiketatutako corpus batetik ateratako probabilitateen bidez erabakitzen da. Hasiera batean analizatzaileak aukera posible guztiak proposatzen ditu, nahiz eta batzuk probabilitate gutxiak izan, ondoren probabilitate handienekoa aukeratzeko. Horrela, analizatzailearen lana erregela horietatik abiatuta corpuseko probabilitateetatik hurbilago dagoen analisia ateratzea da, eta emaitza zuzena emango da analizatutako esaldiaren egitura bat badator corpusean dauden erregelen aplikazioen maiztasunekin. Antzeko saioak egin dira Briscoe eta Carroll-en (1993) ANLT sisteman, hasiera batean garatu zen gramatikari eredu probabilistikoa gehitu baitzitzaion, esaldi bakoitzeko milaka analisi sortzen zirelako. LR behetik gorako algoritmoaren (Tomita 1986) bertsio probabilistikoa erabiliz, analisi horietatik corpusetik hurbilen dagoena aukeratzeko da. Bi adibide hauek erakusten dute sintaxiaren eredu eraikitzaile (baterakuntzan oinarritutako testuingururik gabeko gramatika) eta murriztaileen (corpusen probabilitateen bidez probabilitate handieneko analisia aukeratzeko) beharra.

Bod eta Kaplan-ek (1998) azaltzen duten sisteman antzeko bilketa egitea proposatu da. Abiapuntua LFG gramatika bat eta corpus batetik ateratako egitura sintaktikoak ditugu. Helburua corpus horretan agertzen diren egituren maiztasunekin hurbilago dagoen analisia bueltatzea izango da. Aurrekoekin duen desberdintasun nagusia da honetan egituren maiztasunak eta besteetan erregela sintaktikoen aplikazioen maiztasunak neurtzen direla.

(Collins *et al.* 1999) lanean analizatzaile sintaktiko estatistikoa proposatzen dute txekierarako, Collins-en (1997) ingeleserako egindakoa oinarritzat hartuz. Lengoia hau malgukaria da eta baita hitz-ordena nahiko librekoa ere, euskararekin zenbait ezaugarriekin

bat etorritz. Sistemak, informazioa ateratzeko eskuz markatutako corpora erabiltzen du (*Prague Dependency Treebank*, (Hajic 1998)), eta ekarpen interesgarriena ingeleserako parametro estatistikoak txekierari aplikatzean dago.

Aurreko sistemak eskuz etiketatutako corpus batean oinarrituta daudenez, aurretik etiketatze-lan hori egitea eskatzen dute. Egitura sintaktiko interesgarriak ateratzeko, corpus hori handia izan behar da, fenomeno askoren agerpenak ikusteko mega-hitza askotako testuak beharko direla, eta honek muga bat jartzen die euskara bezalako hizkuntzei, ezinezkoa baita oraingo egoeran etiketatze hori gauzatzea. Muga hori gainditzeko, (Carroll eta Rooth 1998) lanean etiketatu gabeko corpusak aztertzeke sistema bat deskribatzen dute. Bertan, oinarritzko gramatika bat erabiliz, corpus baten azterketarako gramatika probabilistikoa lortzen dute, EM (*Expectation-Maximization*) algoritmoa eta gramatika probabilistikoko lexikalizatuak (PLCFG, *Probabilistic Lexicalized Context-Free Grammar*) erabiliz. Gainera, hasierako gramatika lexikalizatu egiten dute, hau da, hasierako erregela sintaktikoei dagozkien hitzen probabilitateak gehitzen dizkiete, horrela corpus desberdinetarako gramatika egokitzeko aukera emanez, ikaste-prozesu baten ondoren. Bide hau interes handikoa da gure kasuan, etiketatze-lana ekiditen duelako. Dena dela, oraindik metodo honen lehen esperimenduak egiten ari dira, emaitza onekin, eta ikusi beharko da gramatika zabal eta lexikoi handien erabilerarekin mantentzen diren, une honetan dagoen konputazio-baliabideen arazoa (denbora eta espazioa, eredu probabilistikoko konplexuen ondorioz) gaindituz.

III.1.2 Euskararen sintaxia

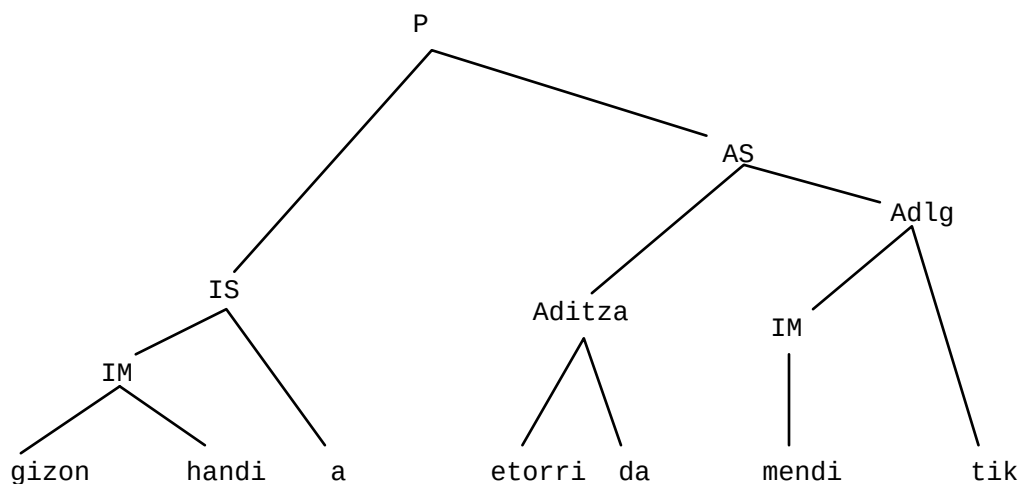
Puntu honetan euskararen sintaxiaren ezaugarri nagusiak azalduko ditugu, beti bi gauza gogoan izanda:

- Sintaxian interesa badugu, analizatzaile sintaktiko bat eraikitzeke izango da, eta analizatzailea testu errealei aplikatzea dugu helburu; hau da, tresna praktikoen inplementazioa teoria linguistikoen egokitzapenarekin elkartu nahi dugu. Beraz, ez gara luzatuko linguistikoki egitura sintaktiko interesgarriak baina aplikaziorako balio gutxikoak direnekin; hau epe luzerako helburua izango dela ahaztu gabe. Zehatzago esanda, tesi honetan azalduko den lanari *euskararen analizatzaile sintaktikoaren oinarria edo abiapuntua* dei diezaiokegu. Oinarri hori erabiliz aberasketa-prozesua egiteko asmoa dugu. Ingeleseztako *bootstrapping* terminoa erabiltzen da prozesu honen definiziorako: sistema baten lehen bertsio batek sistema bera hobetzeko balio du, horrela behar beste aldiz errepikatuz. Ideia hau

informatikako zenbait arlotan ere erabili da, hala nola, konpiladoreen diseinuan edo ezagutza lexikalaren aberasketan, adibidez.

- Analizatzaile (edo sintaxi) horren oinarrian EDBL datu-base lexikala eta segmentatzaile morfosintaktikoa izango ditugu; hau da, estaldura handiko sistema eta ez, oraindik lan batzuetan ikusten den modura, 100 edo 500 hitzeko jostailuzko hiztegia duena. Honek sendotasuna eta koherentzia izango du alde batetik, baina bestetik dagoen informazioarekin moldatu behar izatea ekarriko du, sintaxirako garrantzitsuak diren hainbat informazio ez daudelako, aditzen azpikategorizazioa kasu.

Sintaxiaren azalpen laburra egiteko (Goenaga 1980, Euskaltzaindia 1994, Zubiri eta Zubiri 1995) lanetan oinarritu gara alde linguistikotik, baina lan konputazionalagoak ere (Abaitua 1988, Abaitua *et al.* 1992) kontuan hartu ditugu. Has gaitezen III.7 irudian ikus daitekeen adibidetik. Esaldi horren zuhaitza egiteko Goenaga-ren notazioa jarraitzen saiatu gara¹⁶.

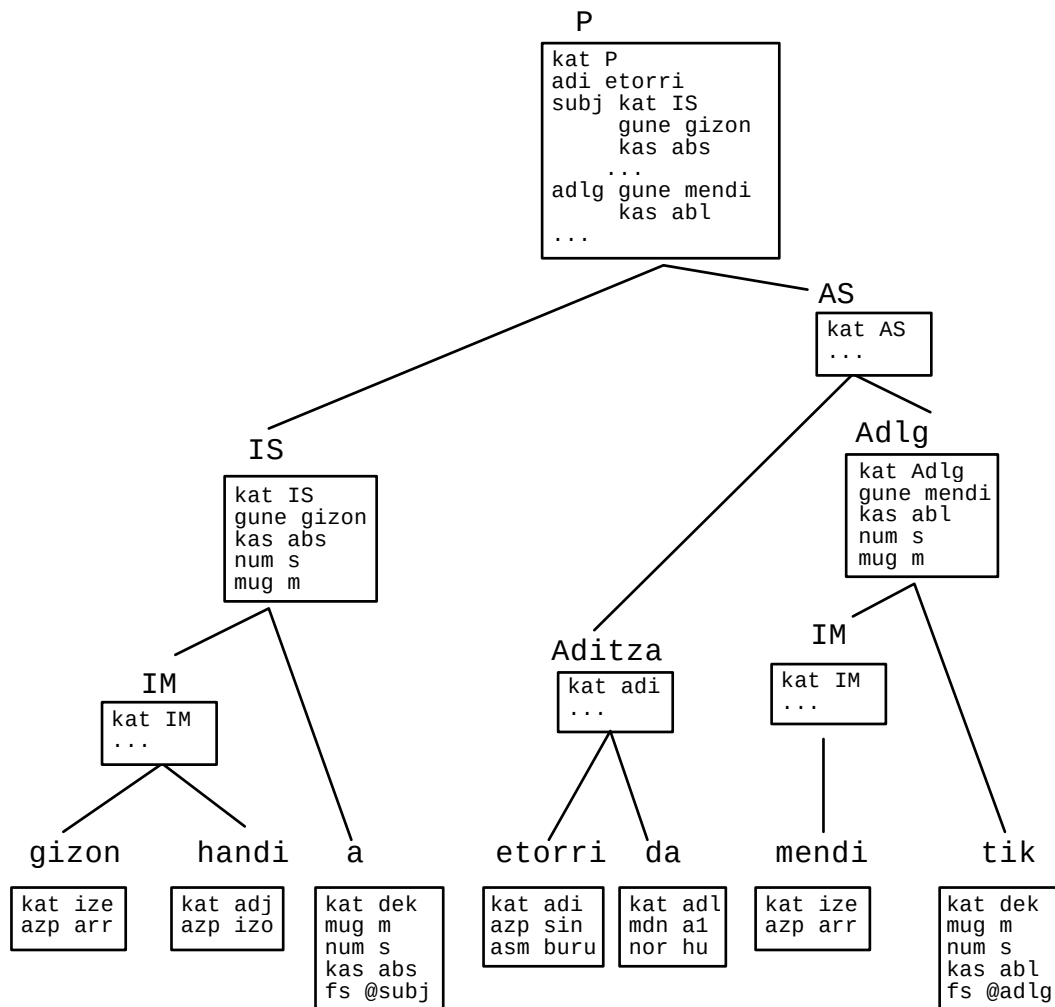


III.7 irudia. 'gizon handia etorri da menditik' esaldiaren analisi sintaktikoa.

Baina zuhaitza horrela azalduta, kategoria sintaktikoen etiketa hutsekin, galdu egingo genuke zuhaitz horren sorreran erabili den informazio asko, zuhaitzaren kategoria morfologiko eta sintaktikoen azpian informazio aberatsa dagoelako. Horrek ezkutatu egingo digu analisiaren konplexutasuna. Adibide bezala, egitura sintaktiko hori hartuta, EDBLko informazioa gehituko diegu zuhaitzaren osagai lexikalei, eta bertatik abiatuta, beste osagai sintaktikoetan egon litekeen informazioa ere saiatuko gara asmatzen (hau adibide bat da, gure

¹⁶ Notazio horretako ikurren esanahiak hauek dira: P perpausa, IS izen-sintagma, IM izen-multzoa, AS aditz-sintagma eta Adlg adizlaguna.

deskribapen sintaktikoa ondorengo puntuetan azalduko dugulako). III.8 irudian dugu zuhaitz osatua.



III.8 irudia. 'gizon handia etorri da menditik' esaldiaren analisi sintaktiko osatua.

Adibide horretatik atera ditzakegu euskararen sintaxiaren tratamenduan garrantzitsuak izango diren aspektuak:

- Morfema izan da analisiaren oinarritzko unitatea deskribapen gehienetan, Goenaga eta Abaitua-rena kasu, eta berdin euskararen antzeko hizkuntzentzat ere (Selkirk 1982, Prószyński 1996). Honek esan nahi du, bai morfologia bai sintaxia, egitura sintagmatiko beraren osagaiak kontsidera daitezkeela, beraien arteko muga argirik gabe. Adibidez, 'gizon + handi + -a' sintagman, -a morfema ez da lotzen adjektiboarekin, baizik eta izen-multzoarekin; horrela deskribapen gramatikala orokorrago eta sinpleagoa eginez. Hori perpaus-mailan ere gertatzen da (adibidez, 'gizona etorri dela' esaldiaren analisisetan -ela atzizkia aurreko esaldi osoari lotzen

zaio, eta ez *da* laguntzaileari). Guk bide honi jarraitu diogu, hurrengo puntuetan ikusiko den bezala, baina, beste aldetik, hitza unitate gisa tratatzeak ere bere abantailak izango ditu, eta horregatik honen ekarpena aztertu dugu euskararen murriztapen-gramatikaren garapenean.

- Informazio aberatsa dago lexikoian. Adibidean ikusten denez, osagai lexikal bakoitzak (eta osagai horiek oinarritzat hartuz sortutako osagai sintaktikoak ere), informazio desberdina dauka: kasua, mugatasuna, funtzio sintaktikoa, ... Informazio hori guztia konbinatzea izango da gramatika sintaktikoaren eginkizun nagusia.
- Aditzaren azpikategorizazioa oraindik landu gabe dago. Aditza elementurik funtsezkoena izango da sintaxiaren deskribapenean, teoria sintaktikoetan zein sistema aplikatuetan. Aditzaren informazioan azpikategorizazioarena da konplexuena, aditz bakoitza zein motatako osagaiekin konbinatzen den zehazten duena. Euskararen kasuan, nahiz eta aditz laguntzaileak informazio asko eman (subjektu, objektu zuzena eta zeharkako objektuaren kasua, numeroa, eta pertsona), oraindik EDBLn aditz nagusi bakoitzaren informazio propiorik ez dago.
- Aditzaren komunztadura subjektu, objektu zuzena eta objektu ez zuzenarekin. Euskaraz komunztadura dago aditza eta ergatibo, absolutibo eta datibo kasuko osagaien artean. Komunztadura lotuta dago azpikategorizazioaren informazioarekin. Horrela, *zekarzkigun* formak III.1 adibideko informazioa du (LFG notazioan idatzita).

kat	aditza		
pred	ekarri<subj obj obj2>		
denb	iragana		
subj	kasua	erg	
	per	3	
	num	sg	
obj	kasua	abs	
	per	3	
	num	pl	
obj2	kasua	dat	
	per	1	
	num	pl	

III.1 adibidea. *zekarzkigun* aditzaren informazioa.

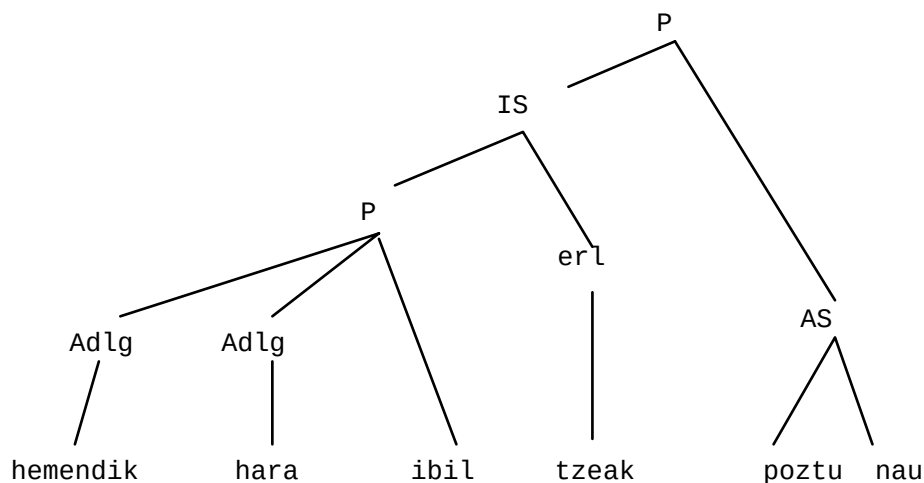
- Perpaus-mailako osagai sintagmatikoen ordena librea. Jakina da euskaraz perpausaren osagai nagusien ordena libre samarra dela. Hau da, subjektua, objektua, adizlaguna eta aditza emanda, beraien permutazio guztiak (hogeita lau) dira posible:

Txakurrak	egunkaria	ahoa	zekarren.
<i>subj</i>	<i>obj</i>	<i>adlg</i>	<i>aditza</i>

Esan behar da ere malgutasun hori perpaus-mailan bakarrik ematen dela, beste osagaietan (izen-sintagma edo mendeko perpausak adibidez) askoz mugatuago dagoelako. Euskara gunea eskuineko aldean duen hizkuntzat hartzen da, eta horregatik perpausen ordena naturalena aditza bukaeran joatea da (baldintza hau beharrezkoa da erlatibozko esaldietan eta mendeko batzuetan). Beste ordenak aldaketa pragmatikoen bidez azal daitezke.

Aurrekoak euskararen puntu aipagarrienak izanda, sintaxiaren tratamendurako ondoko aspektuak ere izan beharko dira gogoan:

- Eratorpena eta hitz-elkarketa. Hauek hitzaren barrukoak direnez, syntaxian ez diegu aldaketarik egingo morfosintaxirako emandako soluzioei.
- Perpaus elkartuak eta mendekoak: koordinazioa, mendekotasuna eta aditzen nominalizazioak. Azken hauek modu egokian tratatzeko, azpikategorizazioari buruzko informazio egokia beharrezkoa izango litzateke, III.9 irudian ikusten den bezala.



III.9 irudia. Nominalizazio baten adibidea.

Adibide horretan ikusten da adizlagunak (*hemendik* eta *hara*) perpaus berean biltzeko bai *ibili* bai *poztu* aditzen azpikategorizazioa edukitzea beharrezkoa izango dela, bestela ekidinezinak diren anbiguotasunak ebatzi ahal izateko.

III.1.3 Gure aukera

III.1.1en sintaxiaren tratamendurako aukera desberdinak aztertu ondoren, hemen guk euskararen tratamenduan hartutako bideak azalduko ditugu, eta ondorengo puntuetan

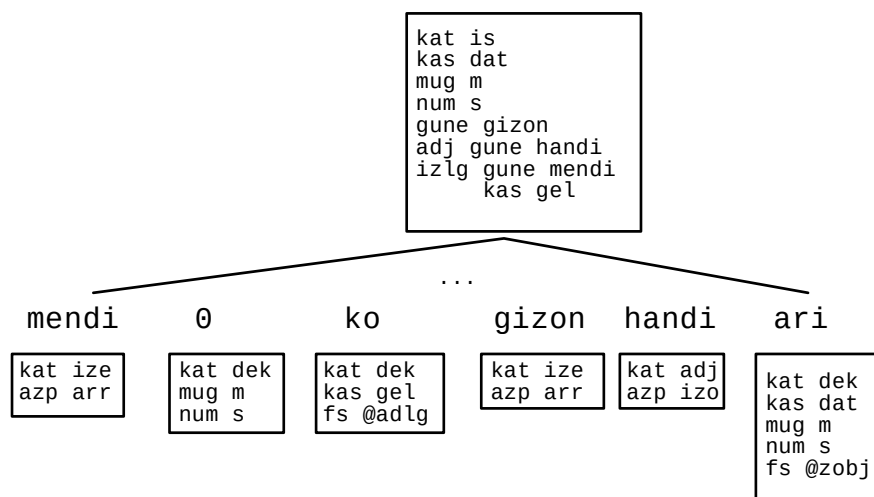
horietako bakoitzaren gainean egin dugun lana zehaztuko dugu. Hasteko, hurbilpen bakoitzaren alde onak eta arazoak bereizten saiatuko gara:

- a) Teknika probabilistikoetan oinarritutako metodoak. Bestelako hurbilpenak ere egin diren arren, teknika hauetatik gehienak alde aurretik etiketatutako corpusetan oinarritzen dira (Black *et al.* 1993), eta honek baldintzatu egiten du euskararako aplikagarritasuna. Ingeleserako, lehen esan dugunez, baliabide ugari daude, Brown Corpus edo Penn Treebank estilokoak, baina beste hizkuntzetarako aukera askoz murriztagoa da, azken aldian turkiera, txekiera edo alemana bezalako hizkuntzentzako saioak egiten ari diren arren (Skut *et al.* 1997, Hajic eta Hladká 1998, Oflazer *et al.* 1999b). Edozein kasutan, corpus hauek lortzea lan handia da. Euskararen lehen saioak morfologikoki etiketatutako 20.000 hitz inguruko eta sintaktikoki etiketatutako 10.000 hitzeko testuen gainean egin dira (Ezeiza *et al.* 1998), baina hoge mila hitz horiek gutxi badira desanbiguaziorako, zer esanik ez sintaxirako. Voutilainen (1997) lanean esaten da 200.000 hitzeko corpus bat sintaktikoki etiketatzeko, pertsona batek urtebete beharko lukeela. Bestalde, ezagutza linguistikoaren erabilerak, lan handikoa izanda ere, emaitza hobeak eman ditu estatistikoekin konparatuz gero (etiketatzean, adibidez, murriztapen-gramatikaren formalismoaren aurkezpenean esan den bezala). Horregatik, tesi honetan ez dugu estatistikaren bidea landuko, nahiz eta ondorengo lanetarako pauso hauek interesgarri ikusi.
- b) Ezagutza linguistikoan oinarritutako hurbilpenen artean, egoera finituko hurbilpenak hedatzen ari dira linguistika konputazionalaren munduan, eta honela labur ditzakegu beren abantaila nagusiak:
 - Erregela lokalen erabilera. Ez da gramatika osoa definitu behar sistema erabilgarria izateko. Honela, analizatzaile partzialak edo desanbiguaziorako erregelak (testuinguru hurbila aztertzen dutenak) defini daitezke.
 - Lexikalizazioa. Aurrekoarekin lotuta, sistema hauek bat datoz sintaxian nagusitzen ari den lexikalizazio-joerarekin, hitz edo adiera konkretuekin lotutako informazioaren erabilera ahalbidetuz.
 - Sintaxiaren alderdi eraikitzailea zein murriztailearen erabilpena. Lehengo adibideetan ikusi dugu formalismo hauetan bi aukerak edo beren konbinazioa erabiltzea badagoela. TGGen kasuan, adibidez, bakarrik alde eraikitzailea landu da, eta aukerak baztertzeko beste mekanismo bat gehitu behar izaten da.
 - Eraginkortasuna. Erregelak automata edo transduktoreetan konpilatu ondoren abiadura handiko sistemak lor daitezke, TGGen lortutakoen aldean.

Sistema hauekin gertatzen diren arazoak ere aipatuko ditugu:

- Adierazpen erregularrak konpilatzeko momentuan, konplexutasun linguistikoa igotzen doan heinean problemak daude lortutako automaten tamainekin (Tapanainen 1997, Beesley 1998a). Kasu horietan, automata bakar batean konpilatu beharrean, automata desberdinak sortu beharko dira, era sekuentzialean konposatzeko. Dena dela, oraindik lan hori hizkuntzalariak egin behar du, sortutako automaten tamaina aztertuz.
- Erlazio sintaktiko konplexuak adierazteko zailtasuna. Lehenago, morfosintaxiaren kasuan (§ II.6), ikusi ditugu ezaugarri-egitura konplexuak kudeatzeko egoera finituko metodoek dauzkaten zailtasunak. Hau areagotu egingo da sintaxia tratatzeko momentuan. Honen azalpenerako adibide bat dugu III.10 irudian. Bertan sintagma arrunt baten morfemak ditugu beheko partean, eta goian analizatzaile sintaktiko batetik espero genezakeen irteera bat. Ikusten denez, morfema horien guztien ezaugarrien kudeaketa lan konplexua da, eta horren moduko sintagma sinple batean gertatzen diren aukera guztien adierazpen erregularren bidezko tratamendua ikusteko dago oraindik (Beesley 1998a). Adibidez, deklinabide-morfema baten kasuaren goratzea ekuazio bakar batekin adieraz daiteke baterakuntza-formalismo batean, baina automata baten bidez egiteko, kasu guztien zerrendatzea egin beharko litzateke¹⁷. Kasuaz aparteko hainbeste informazio mota tratatu behar direnez, momentuz hau egitea ezinezkoa da.

¹⁷ Honek ez du esan nahi zerrendatze hori beti gramatikariak egin behar duenik, hau da, kasu batzuetan konpilazio-prozesu baten bidez izan liteke. Arazoa konbinazio askotan sortutako aukera-kopuru handiegia da.



III.10 irudia. 'mendiko gizon handiari' sintagmaren analisi sintaktikoa.

- c) Testuingururik gabeko gramatikek, berriz, azken arazo honetarako soluzioak dauzkate. Baterakuntzaren bidez, bai ekuazioak bai printzipio linguistiko orokorrak erabiliz, modu erazagutzailean deskribatuko dira fenomeno horiek. Gainera, baterakuntza eta TGGak deskribapen linguistikoak egiteko lengoia modura erabili ohi dituzte hizkuntzalariek, eta horregatik bi munduak, informatika eta hizkuntzalaritza, lotzeko erraztasuna emango du bide honek.

Eraginkortasunaren aldetik, baterakuntza eta TGGak ez dira egoera finituaren bidez landutako soluzioak bezain azkarrak. Beste alde batetik, lexikalizaziorako joerak erregelen (edo patroi lexikalen) biderketa dakar, eta horrek eraginkortasunaren arazoa areagotuko du (erregela sintaktiko orokorrak hitzen arabera biderkatu egiten direlako, edo hiztegiko sarrerak hitzaren erabilera ezberdinen arabera gehitu egiten direlako).

Euskararen sintaxia lantzeko, beraz, testuingururik gabeko gramatikak eta egoera finituko mekanismoak, biak, erabil ditzakegu. Lehen esan den bezala, konpromiso bat dago TGGen ahalmen deskriptiboaren eta egoera finituko eraginkortasunaren artean. Bestalde, ez dugu uste bide horiek kontrajarriak direnik, eta horregatik beraien konbinazio edo integrazioa bideragarriak eta interesgarriak direla pentsatzen dugunez, bien ekarpena aztertzen saiatu gara. Ondorengo ataletan (§ III.2, § III.3) euskararen tratamendu sintaktikorako testuingururik gabeko gramatikak (baterakuntzarekin hornituta) eta egoera finituko sintaxia aztertuko ditugu, bukaeran (§ III.4) beraien arteko konbinazio batzuen bideragarritasuna eta ekarpenak arrazoituz. Geroago azalduko dugun bezala, bi aukeren artean banaketa egiteko arrazoi nagusienetako bat azpikategorizazio informazioa ez edukitzeak emango du, informazio hori gabe etekin gutxikoa delako testuingururik gabeko gramatika bat lantzea esaldi-mailan,

ondorio nabarmenena anbigutasun ebatziezina izango baita. III.2 adibidean ikusten den bezala, kasu gehienetan baliorik gabekoa izango da bi aditzen arteko osagaiak aditzei esleitzea, horrek analisiak ugaltzea baitakar, eta gainera azpikategorizazio informazioari gabe ebatziezinak dira. Adibide horretan, lau modu daude esaldi hori ulertzeko, bi osagaiak bi aditzekin era desberdinetan lotuz. Dena dela, kasu batzuetan, aditz jokatuarekin gertatzen den bezala, lor daitezke esaldi-mailako egitura ez anbiguoak, komunztadurari buruzko informazioari esker. Hau guztia aintzat hartuta, TGGak erabili dira esaldietako osagai nagusiak definitzeko, EDBLk ematen duen informazioaren mugara iritsi arte.



III.2 adibidea. Bi aditzen arteko osagaiak aditzei esleitzeko (gutxienez¹⁸) azpikategorizazioari buruzko informazioa beharko litzateke.

Lehenago esan dugunez, tesi honetako lana euskararen sintaxiaren *lehen hurbilpena* da, eta horregatik, tresna desberdinen esperimentazioa beharrezkoa ikusten dugu sintaxiaren tratamendu sakonagoei ekin baino lehenago. Halere, morfosintaxiaren tratamendutik ateratako ondorioetako bat baterakuntzaren beharra izan da, informazio konplexuaren erlazioak eta metaketak adierazteko. Hauek TGG baten bidez tratatzeko egokiak ikusten ditugu, fenomeno sintaktiko orokorre dagokien heinean behintzat, baina badira beste fenomeno batzuk (sortutako analisi anitzen arteko aukeraketa edo hitz-mailako informazio lexikalizatua, azpikategorizazioa kasu, behar dutenak, adibidez) TGGekin edo ezinezkoak edo inplementaziorako arazoak eman ditzaketenak, egoera finituko mekanismoen aplikaziorako bideak irekitzen dituztenak. Hau da, analisi morfosintaktikoan egin dugun bezala (Ritchie 1992, Aduriz *et al.* 1999), syntaxirako ere bi formalismo motak integragarriak eta lagungarriak izan daitezkeen aztertuko dugu.

¹⁸ Gutxienez diogu azpikategorizazio-informazioarekin ere ez dagoelako argi guztiz ebatz litezkeen. Adibidean *Pirineotako bazter honetan* (inesiboa) eta *erakunde bat* (absolutiboa) bi aditzekin lotzea posible litzateke. Dena dela, informazio sintaktikoa edo maiztasunek lagun dezakete bi osagaiak *sortu* aditzarekin lotzen.

III.2 Baterakuntzan oinarritutako euskararen analizatzailea

Testuingururik gabeko gramatiken erabilera aztertzeke, baterakuntza-mekanismoarekin lotutako formalismoak landuko ditugu; berauek erabiltzeko arrazoen artean erazagutzailetasuna eta ahalmen deskriptiboa ditugu. Lehen puntu batean (§ III.2.1) baterakuntza-formalismoen azterketa orokorra egin ondoren, § III.2.2n guk inplementatutako gramatika aurkeztuko da.

III.2.1 Baterakuntza-formalismoen azterketa

Hemen formalismo batzuk aztertuko ditugu euskararen sintaxiaren ikuspuntutik. Lehenengo (§ III.2.1.1) formalismo desberdinen arteko konparazioa azalduko dugu, eta gero HPSG formalismoaren euskararen gramatika baten diseinu eta inplementaziotik ateratako ondorioak emango ditugu (§ III.2.1.2).

III.2.1.1 Zenbait formalismo sintaktikoren azterketa konparatiboa eta aplikazioa

Hemen kontatuko duguna (Abaitua *et al.* 1992) lanean luzeago azaldukoaren laburpena izango da. Lan honetan sintaxiaren deskripziorako formalismoetatik interesgarriak jo direnak konparatu genituen euskararen tratamendurako egokitasuna aztertzeke. GPSG eta LFG formalismoak eta beren inplementazioak (ANLT, Carroll 1993; GFU-LAB, Ruiz *et al.* 1990) arreta bereziz aztertu ziren. Hauek aukeratzeko arrazoi desberdinak daude: formalismoak oso zabaldua daude ingurune akademikoan, proposamenak daude hizkuntza desberdinetarako, eta beraien inplementaziorako tresnak eskuragarri daude.

Lehen balorazio hau ondoko ezaugarriei erreparatuta egin zen:

- 1) perpaus-mailako osagai sintagmatikoen ordena librea,
- 2) aditzaren azpikategorizazioa eta
- 3) aditzaren komunztadura subjektu, objektu zuzena eta bigarren objektuarekin.

Ordena librearen ikuspuntutik, problema nagusia da oinarritzko konfigurazioa (hierarkikoa) sortzen duten oinarritzko erregelak (1.a eta 1.b adibideak, subjektua eta

objektuaren agerpenak adierazteko) edo ordena librea onartzen duten erregela “lauak” (2 adibidea) aukeratzea.

- (1.a) Perpausa --> Izen-sintagma Aditz-sintagma
 (1.b) Aditz-sintagma --> Izen-sintagma Aditza
 (2) Perpausa --> X* Aditza X*

1.a eta 1.b moduko erregelak beharrezkoak dira zenbait teoriatan, printzipio orokorrak erregela horietan oinarritzen direlako; baina orduan arazo nagusia gero ordenazio konplexuak asmatu behar izatea da, eta gainera osagaien arteko zenbait ordena posiblek aditz-sintagmaren egitura hausten dutela:

- (3) Ahoan zekarren txakurrak egunkaria.
 adlg *subj* *obj*

Honek guztiak azalerazten ditu GPSGren zailtasunak euskara bezalako hizkuntzekin, teoriaren oinarritzko erabakiek 1.a eta 1.b moduko erregelen alde egiten baitute, eta teoriaren aldaketa sakonak ekarriko lituzkeelako egokitzapena egiteak. ANLT inplementazioan problema modu errazagoan konpon daiteke, baina arazo ugariarekin. Bestalde, ordena libreko hizkuntzen azterketak LFG teoriaren formulazioan eragina izan du, eta horregatik kodeketa ez-konfigurazionalari buruz hitz egiten da, 2 adibidean agertzen den moduan. Antzeko gauza egin da HPSG teorian, GPSGren ondorengoan (HPSGri buruz luzeago egingo dugu berba hurrengo puntuan).

Azpikategorizazioarekin antzeko problema gertatzen da: GPSG teorian, osagaien ordena murrizten dituzten erregelek baldintzatu egiten dute beren tratamendua. Lehen esan den bezala, teoria trinkoa eta koherentea da, baina ingeleserako ez diren hizkuntzentzat arazoak ematen ditu. Komunztadurarekin ere AGR (*agreement*) izeneko ezaugarri bat definitu da, ingelesaren komunztadura murriztuan oinarrituta. Euskaraz, aldiz, hiru osagaik dute komunztadura, eta gainera, hau lotuta dago azpikategorizazioarekin, nahiz eta teorian bi fenomenoak independentetzat jo. LFGk ez du arazorik bi fenomeno horiek tratatzeko.

Bukatzeko, GPSG-ANLT sistemarekin egindako analisiaren ondorioa da hiru fenomenoak (ordena librea, azpikategorizazioa eta komunztadura) tratatu ahal izateko GPSG teorian orokortzat hartzen diren printzipio batzuk birplanteatu egin behar direla. Azterketa honen beste ondorio bat izango da euskararen moduko hizkuntzentzako analisiak, egitura lauak eta ordena librekoa izan behar duela. GFU-LAB (LFG) sistemak ez du problemarik hiru fenomenoak deskribatzeko. Hau normaltzat har genezake, teoria euskara bezalako hizkuntzentzat pentsatua izan baita. Gainera, testuingururik gabeko gramatika eta

baterakuntza-ekuazioen erabilerak erraztasun eta malgutasunaren abantailak dauzka sistemaren erabiltzailearentzat.

III.2.1.2 Guneak zuzendutako egitura sintagmatikoen gramatika eta euskararako aplikazioa

Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG; Pollard eta Sag 1994) lengoaia naturalen sintaxia eta semantika deskribatzeko formalismoa da. HPSG baterakuntzan oinarritutako formalismoen artean nagusietakoa da momentu honetan. Bere erazagutzaitasuna eta koherentzia (gramatika osoa maila bakar batean deskribatu ahal izatea) izan daitezke arrakasta horren zergati nagusiak, eta hau izan da euskararekin proba bat egiteko saioaren arrazoia. Horretarako, inplementazio bat aukeratu genuen (Popowich eta Vogel 1991). Hemen (Gojenola 1998) barne-txostenean luzeago azaltzen denaren laburpena emango dugu.

HPSG teoria GPSG eta LFGren ondorengoa izanda, bietako aspektuak integratuz, pentsa liteke erraztasunak egongo direla lehen aipatutako fenomenoak deskribatzeko. Hainbeste lekutatik hartutako ideien bilduma honen ezaugarriak aipagarrienak hauek izan daitezke: baterakuntzan oinarrituta; semantika gramatikaren oinarritzko elementua da, fonologia, sintaxia eta berbaldiarekin integratuta; printzipio orokorrak edo unibertsalak definitzen dira, eta hauekin batera lengoaia konkretu baten printzipio lokalak; eta lexikalizazio-maila oso altua (erregela sintaktiko eskematikoak eta hiztegi oso konplexua).

HPSGn erregela sintaktikoak eskema sintaktiko orokorrak dira. Euskararentzat ondokoa izan liteke perpaus sinple baten erregela¹⁹:

[subcat <>] --> H [LEX+], C*

Erregela horrek, azpikategorizazioaren eta ordenaren printzipio orokorrekin batera, erregela tradizionalen multzo handia deskribatzen du, gunea eta bere osagarrien konbinazio ezberdinak tratatzen dituelako. Adibidez: “VP -> V NP” “VP -> V NP PP” (aditza (gunea) bere osagarriekin), edo “SN -> Det N” (izena (gunea) eta determinatzaileak (osagarriak)).

Euskararen gramatika sinple baten garapena egin zen ondoko fenomenoak tratatzeko:

¹⁹ ‘subcat <>’ espezifikazioak gune lexikalak azpikategorizatutako argumentu guztiak bete direla esan nahi du. H gunea (*head*), C osagaia (*complement*) eta LEX+ osagai lexikala adierazteko erabiltzen dira.

- Izen-sintagmak eta adizlagunak zenbatzaile, adjektibo eta kasuarekin. Adibideak: *ni, guk, guri, 'etxe polit -a', 'zenbait neska polit -i', 'bi gizon -ak', 'gizon bi -ek'*²⁰.
- Aditz-sintagma (perpaus sinpleak), aditz iragangaitz eta iragankorrekin. Adibidez: *'mutil bat dator', 'ni -k katu txiki -a daukat'*.

Gramatika hori egiteko, HPSGren hasierako ingeleserako formulaziotik zenbait gauza aldatu behar izan ziren:

- Ezaugarri sintaktikoak. Ingeleserako proposatutako ezaugarrien artean badira batzuk euskararen prozesamenduan (edo behintzat euskararen azpimultzo honetarako) interesgarriak kontsideratu ez direnak, eta beste batzuk, aldiz, existitu ez eta definitu behar izan direnak.
- Gunea-osagarriak-adjuntuak erlazioa. Kasua daukan atzizkia hartu da, XBAR teoria jarraituz, izen-sintagma eta adizlagunaren gunetzat, eta izena izen-multzoan. Erlazio hauek finkatzea lan konplexua da hartutako euskararen azpimultzoarentzat, zer esanik ez gramatika orokorra egin nahiko balitz.
- Erregela gramatikalak eta lehentasun linealeko murriztapenak (ordena finkatzeko). Hauek ere egokitu behar izan dira euskara tratatzeko, bere ordena libreko ezaugarriak kontuan hartuta.

HPSG formalismoaren azterketan eta euskararen oinarritzko gramatika baten prestakuntzan egindako lan horren ondorioz, aspektu hauek azpimarra daitezke:

- HPSG teoria, nahiz eta edozein lengoaia tratatzeko formalismo orokorra izan, ingeleserako landu da gehienbat eta, beraz, euskarari aplikatzeko lan linguistiko handia egin behar da, batez ere lexikoiaren definizioan. Hau LFG teoriaren kasuarekin konpara daiteke, LFG hasieratik lengoaia desberdinetarako pentsatuta zegoelako eta, beraz, euskarari aplikatzea sinpleagoa izan zen.
- HPSG teoriaren alde interesgarrienak bere erazagutzailetasuna (GPSG teoria baino askoz modu sinpleagoan definitua, semantika ondo definituarekin) eta bere homogenotasuna dira (teoria osoa maila bakarraren bidez definituta, ezaugarri-egituren bidez).
- Nahiz eta teoria oso modularra izan, informazio guztia ahalik eta modu sinpleenean definitu nahi izateak problemak ekartzen ditu batzuetan. Adibidez, gertaera

²⁰ Hemen ere, lehen azaldu dugun bezala, morfema hartu zen analisirako unitatetzat. Azalpen sakonagoa (Gojenola, 98) txostenean dago.

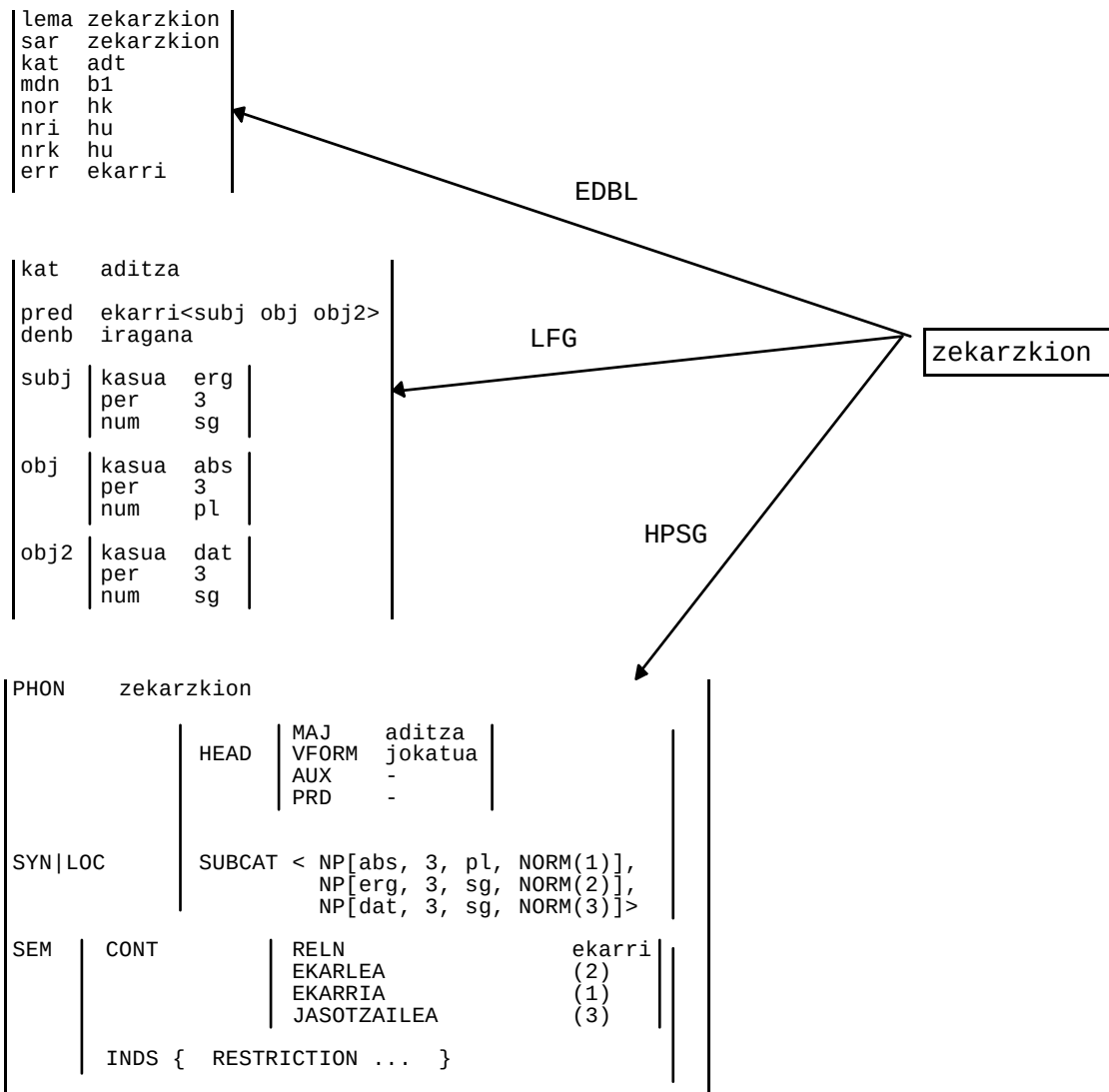
linguistiko bat tratatzeko erregela gramatikala edo printzipio orokorra aldatzen bada, orduan honek gramatika osoan du eragina, eta batzuetan interferentzia nahiko konplexuak izan daitezke.

- Baterakuntzan oinarritutako teoria guztiak bezala, oraindik zailtasunak daude sistema hauek benetako aplikazioetan erabiltzeko (adibidez, testu-tratamenduan). Uszkoreit-ek (1991) aipatzen duenez, eraginkortasuna hobetzeko mekanismoak gehitu behar zaizkie aplikagarritasun hitza erabili nahi bada (Briscoe eta Carroll 1993, Bod eta Kaplan 1998, Kiefer *et al.* 1999, Kiefer eta Krieger 2000).

III.2.1.3 Ondorioak

Baterakuntza-formalismo nagusien azterketaren ondoren atera dezakegun emaitzarik garrantzitsuen bat dator sintaxiaren tratamenduaren azken joerekin, lexikalizazioa proposatzen baitute sintaxiaren oinarritzat. Lexikoian oso informazio aberatsa gordetzen da, azpikategorizazioarena kasu, eta gure kasuan honek arazoak emango dizkigu sintaxiaren lanari ekiteko. Argi dezagun baieztapen hori III.11 irudiko adibidearen bidez. Bertan aditz baten informazioa deskribatu da hiru modutan: oraingo EDBLko informazioa, LFGkoa eta HPSGkoa.

EDBLn sintaxiaren tratamenduan beharrezkoa den hainbat informazio ez dagoenez, ezinezkoa izango litzaiguke LFG edo HPSG formalismoak zuzenean aplikatzea, hiztegia murriztu beharko genukeelako edo lexikoia osatzeko lan ikaragarria egin beharko litzatekeelako. Gainera, lan horrek formalizazio linguistiko handia dakar berarekin. Horregatik baterakuntza-formalismo sinpleenatarikoa aukeratu dugu, PATR, ondorengo lanetarako HPSG edo LFGren inplementazioak baztertu gabe. Heldu beharrezkoa izango da lexikoen konplexutasuna EDBLnosatzen denean. Horretarako gure analizatzaile sintaktikoaren lehen pausoa informazio hori corpusetatik ateratzea izango da (ikus IV. kapitula).

III.11 irudia. *zekarzkion* formaren informazio lexikala.

III.2.2 Baterakuntzan oinarritutako analizatzaile sintaktikoa

PATR (Shieber 1986) formalismoa aukeratu dugu sintaxiari ekiteko. Horren arrazoi nagusiak hauek dira:

- Formalismo malgu baten beharra dugu, HPSG edo LFG moduko teoria linguistiko konplexuekin konparatuz gero. Ez dugu ukatuko teoria linguistikoen beharra, baina berriro diogu teoria eta analizatzaileen helburu eta mekanismoek ez dutela berdinak izan behar (ikus III.1 taula).
- Azpikategorizazioaren informazio-eza. Oraingo datu-base lexikalean esan dugu azpikategorizazioarekin lotutako informazio-zati bat besterik ez dugula, komunztadura zuzena duten osagaiena (nor, nori eta nork), aditz jokatuetatik

datorrena. Informazio hori ez da nahiko komunztadurarik ez duten osagaiak tratatzeko eta, gainera, ezin da erabili kasu askotan (aditz jokatugabeak adibidez). Horregatik, gure formalizazioa ezin da alde honetatik ausartegia izan, bestela eskura ez dugun informazioaren gabeziak arazoak emango bailituzke (anbiguotasunaren igoera, adibidez).

- Formalismoaren sinpletasuna abantaila bat da analizatzailea euskararen tratamendurako lehen pausoa delako, horrela analisi partzialak ateratzeko aukera emango duelako. Lehenago komentatu dugu sintaxiaren tratamendu oso bat egitea oraindik bukatu gabeko lan bat dela, eta horregatik lan honek abiapuntu bat izateko asmoa du, ondorengo formalizazio (agian) konplexuagoetan berrerabili ahal izango dena.

Jarraian datozen puntuetan tratamendu sintaktikorako gure oinarritzko erabakiak azalduko ditugu hasteko (§ III.2.2.1), gero egindako gramatikaren deskribapenarekin segitzeko (§ III.2.2.2). Ondoren, § III.2.2.3n gramatikan postposizioen tratamendua gehitzean egin den aberasketa azalduko da. Gramatika morfosintaktiko eta sintaktikoaren arteko alde komunak eta ezberdintasunak aipatuko ditugu hurrengo puntuan (§ III.2.2.4), inplementazioari buruzko datuekin (§ III.2.2.5) eta laburpen batekin bukatzeko (§ III.2.2.6).

III.2.2.1 Oinarritzko erabakiak

Puntu honetan gramatikaren garapenean zehar hartutako erabaki nagusien berri emango da. Erabaki horietatik batzuk diseinukoak dira eta beste batzuk, berriz, historikoak, denbora eta esperimentazioaren ondoren hartu direnak:

- Gramatika partziala. Termino honekin analisi sintaktiko tradizionaleko informazioaren zati bat, ez guztia, ateratzeko sistemak izendatzen dira. Teknika hauek fidagarritasuna eta sendotasuna dute helburu testu orokorren aniztasunaren aurrean, baina hau osotasunaren kaltetan, errore-tasa txikia onartuz. Ideia nagusia, beraz, analisi osoak lortzeko zailtasunak ezagututa egitura sintaktiko txikiak ateratzen saiatzea da, fidagarritasunez atera daitezkeenak, askotan testuinguru lokalak erabiliz. Egitura sintaktiko hauei *chunk*²¹ deitu zaie testu askotan. Osagai sinple hauen ezagutzak lagun dezake esaldi-mailako fenomenoaren azterketaren konplexutasuna jaisten. Euskararen kasuan, testu orokorretarako gramatika

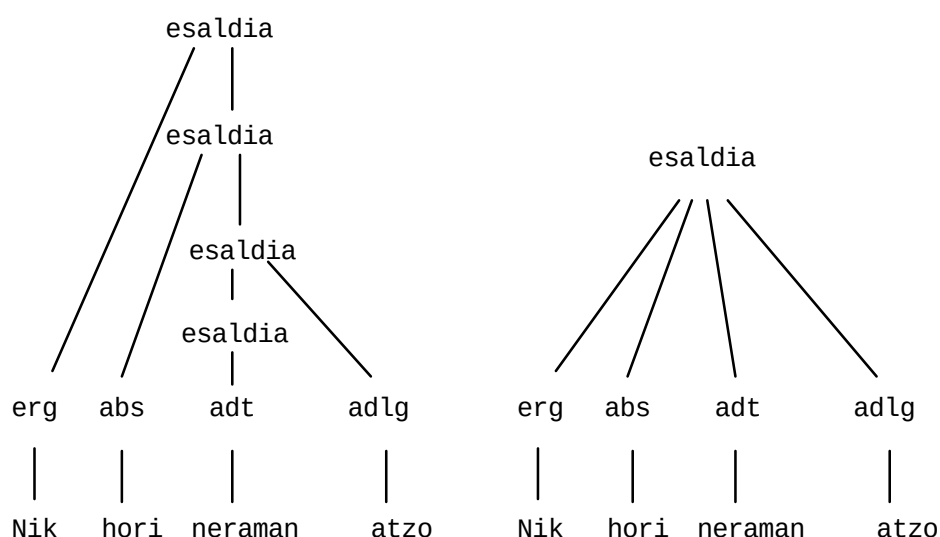
²¹ Termino honen definizioa era desberdinetan egiten da zenbait lanetan. Abney-k (1997) honakoa ematen du: “nonrecursive kernels of all ‘major’ phrases, regardless of category”.

sintaktiko konputazional baten lehen urratsak emateko bide hau aukeratu dugu, irizpide nagusia fenomeno orokorrak deskribatzea dela. Horregatik, sintaxiaren alde emankorrenak landu ditugu, corpusen tratamenduaren ikuspuntutik. Adibidez, izen-sintagma zein adizlagunen²² egitura erregela berekin definitu ahal da, eta gainera erregela horien bidez testuetako zati handien analisia lor daiteke, maiztasun handikoak direlako. Bestalde, esaldiaren analisi osoak azpikategorizazioaren informazioaren beharra du, eta etekin gutxikoa izango da honen gainean lan egitea informazio hori eskura ez badago (ikus III.2 adibidea). Honek esaldiaren *azaleko* tratamendua egitera bideratu gaitu, hau da, tratagarriak diren osagai sintaktikoak aztertzea (ondoren ikusiko denez, izen-sintagmak, adizlagunak, mendeko perpausak eta esaldi sinpleak). Baina honek ez du esan nahi gramatikan analisi osoa baztertu egin denik, hasierako analizatzailea era iteratiboan erabil baitaiteke azpikategorizazioaren informazioa lortzeko, *lexical bootstrapping* izeneko teknika (Abney 1997, Aldezabal *et al.* 2000) erabiliz. Gramatikan landutako egitura sintaktikoak § III.2.2.2n azalduko dira.

- Erregela bitarrak erabili dira gramatikaren osagai nagusiak definitzeko orduan, era horretan sistema linguistikoaren modulartasuna indartuz. Honen alde argumentuak daude alde teorikotik (gramatika sortzaile-transformatzaileetan bezala) zein praktikotik. Jensen-ek (1987ab) honako abantailak aipatzen ditu modu horretako erregelak erabiltzeko:
 - Ordena librea tratatzeko egokiak. Aurreko puntuan (§ III.2.1.1) ikusi denez, erregela hierarkikoak definituz gero, problemak daude osagaien posizio-aldaketak adierazteko: edo erregela-kopurua biderkatu egiten da, konbinazio posibleak emateko, edo beste mekanismo batzuk sartu behar dira, konplexutasuna handituz (adibidez, sistema batzuetan ordenaren portaera gidatzeko erregelak edo printzipioak definitu dira). Erregela bitarrak erabiltzearen ondorio nagusiak sinplifikazioa eta malgutasuna izango dira.
 - Aukerazko elementuak adierazteko erraztasuna. Ordena librearekin bezala, aukerazko elementuak (adjuntuak) toki askotan ager litezke.
 - Inplementaziorako erraztasuna. Puntu hau ez da inondik ere erabakigarria izan behar, beste arrazoi handiagorik gabe, baina geroago ikusiko dugunez, lagungarria izan zen analizatzailea inplementatzeko garaian.

²² Termino hauek era honetan definitu ditugu: izen-sintagmak kasu gramatikaldun (absolutibo, ergatibo eta datibo) sintagmak izango dira, XBAR teorian NP direnak, eta adizlagunak gainontzekoak izango dira (XBARen PP direnak).

- Analisi osoa askotan lortuko ez denez, behetik gorako estrategiarekin (*bottom-up*) aplikatuko dira erregela gramatikalak, horrela ahalik eta egitura gehienak ateratzeko analizatu behar den esalditik.
- Egitura lexikal *sinplea*. EDBLtik abiatuko gara, eta bertako informazioa erabiliko dugu. Esan dugunez, EDBL iturri lexikal aberatsa da, milaka sarrera dituelako, baina aditzen azpikategorizazioari buruz dagoena ez da nahikoa.
- Egitura sintaktiko *lauak* erabiliko dira analisi-zuhaitzaren ordean, III.12 irudian ikusten den bezala. Horrela, erregelen aplikazioa (erregela bitarrak erabilia zuhaitz altuak aterako dira), lortutako informaziotik bereiztuko da. Hau Jensen-en (1987ab) lanean edo LFGn egiten denaren antzekoa da²³.



III.12 irudia. Esaldi baten analisi-zuhaitza (ezkerrean) eta egitura sintaktiko lau (eskuinean).

III.2.2.2 Gramatikaren deskribapena

Hemen euskararen PATR gramatikak tratatzen dituen egitura sintaktikoak aztertuko ditugu. Osagai sintaktikoak errepresentatzeko definitu ditugun egiturak arrazoitu ondoren, gramatikak lantzen dituen fenomenoak eta tratatzen ez dituenak ere aztertuko ditugu. Bukaeran, adibide batzuk eta gramatikaren garapena jarraitzeko bideak azalduko dira.

Informazio interesgarri gehiena saiatu gara sartzen analisi sintaktikoaren emaitza adieraziko duten ezaugarri-egituretan, beti oinarri lexikalaren mugak kontuan hartuz, eta aplikazio jakinetan pentsatu gabe. Horrela tresna ahalik eta orokorrena egin nahi izan dugu,

jakinda ere aplikazio konkretuen beharrak murriztagoak izango direla eta informazio gehiago tratatzeak kostu konputazional handiagoa izango duela. Euskararen elementu sintaktiko nagusienak aipatzekotan, bi emango ditugu:

- Izen-sintagma eta adizlagunak. Bi osagai mota hauek egitura oso antzekoa dute, III.3 taulan sinplifika dezakeguna. Bertan zenbait adibideren bidez ikus daiteke biek aukerazko izenlagunak, adjektiboak eta determinatzaileak dituztela, bukaeran kasu-marka jarritz (absolutibo, ergatibo edo datibo izen-sintagmentzat, eta beste kasuak adizlagunentzat).

Aukerazko izenlaguna(k) / determinatzailea	izena	Aukerazko adjektiboa(k)	Aukerazko determinatzailea(k)	Kasu-marka
	etxe			a
zenbait	etxe	polit		0
gure	etxe	polit	bat	i
mendiaren puntako	etxe	polit	bi	rekin
ikusi ditudan	etxe	polit	bi	ak

III.3 taula. Izen-sintagma eta adizlagunen adibideak.

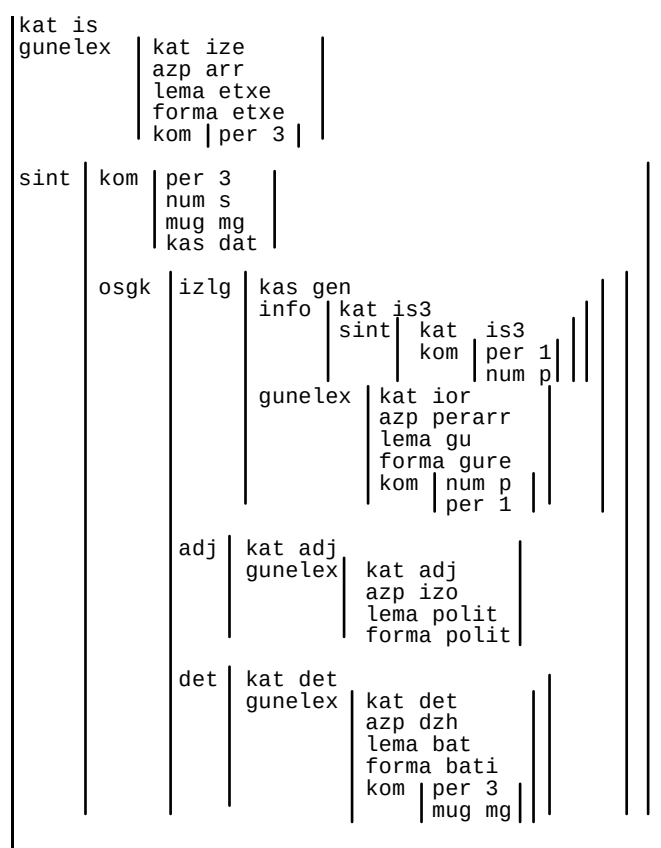
III.13 irudian agertzen da horiek adierazteko zehaztu dugun egitura sintaktikoa. Bertan osagai sintaktikoen ezaugarriak “sint” izeneko ezaugarriaren azpian jarri ditugu, lexikoitik datozenetatik bereizteko. Hauek “gunelex” ezaugarrian egongo dira, egitura sintaktiko osoaren “gunea”-ren²⁴ ezaugarriak gordetzeko. Ezaugarri sinpleak aparte (kategoria, azpikategoria, ...), “kom” ezaugarrian komunztadurarekin lotuta dauden kasua, numeroa, mugatasuna eta pertsona sartu ditugu. “osgk”-ren azpian gunearekin harremana duten osagaien berri emango da, horien artean izenlagunak, adjektiboak eta determinatzaileak daudela.

Egitura horiek osatzeko, egiaztatu egingo dira sintaktikoki okerrak diren egiturak, determinatzaile eta kasu-marken artekoak esaterako (**bi etxea*’,

²³ Irudian egitura sintaktikoa zuhaitz moduan irudikatu arren, analizatzaileak ezaugarri-egiturak erabiliko ditu.

²⁴ Gunetzat hartu ditugun osagaien justifikazioa praktikoa izan da gehienbat, eta alde batera utzi ditugu eztabaida teorikoen ekarpenak (adibidez, XBAR teorian izen-sintagma eta adizlagunaren gunetzat kasu-marka kontsideratu ohi da, baina guk izena hartu dugu).

‘*zenbait gizonak’, edo ‘*bi etxe hori’, adibidez). Erlatibozko izenlagunen kasuan, zalantzak izan genituen mendeko perpausaren aditzarekiko komunztadurak tratatzerakoan. Kasu batzuetan, komunztadura argia dagoela ematen du (‘*etorri diren gizona’), baina beste batzuetan ez da gertatzen mota horretako komunztadurarik (‘egon naizen tokietan’). Horregatik, ez ditugu orokorrean komunztadura horiek kontsideratu.



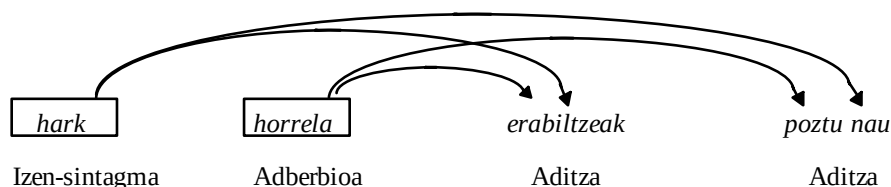
III.13 irudia. ‘*qure etxe polit bati*’ izen-sintagmaren egitura sintaktikoa²⁵.

- Perpausa eta bere eratorriak. Perpaus single baten osagai nagusia aditza dugu, eta bera gune moduan hartuta osagaiak eransten zaizkio, horien artean izen-sintagmak, adizlagunak, adberbioak eta mendeko perpausak.

Mendeko perpausak perpaus arrunt (murritzapen batzuekin) bati menderagailu bat txertatuz lortzen direnez, perpausaren definizio errekurtsiboa dugu. III.14 irudian esaldi baten analisiaren adibidea dugu. Egitura osoaren elementu nagusia *jakin* aditza da, esaldi nagusikoa. Berak bi osagai ditu lotuta, *zuk*

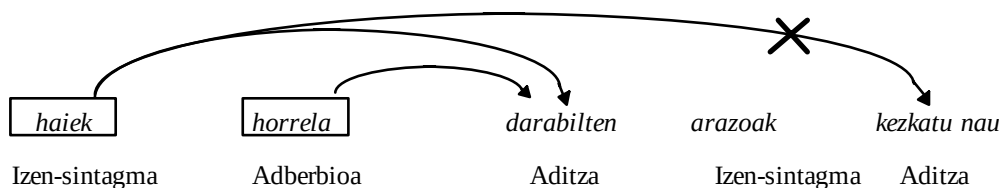
²⁵ Hauek dira irudian erabilitako ezaugarrien esanahiak: *is3* izen-multzoa, *ior* izenordea, *dzg* determinatzaile zehaztugabea, eta *izo* izenondoko adjektiboa.

analisiarekin bukatu dugu. Honen arrazoa lehen aipatutako azpikategorizazioaren informazio-eza da, oraingo gramatikaren egoeran esaldiaren analisia ematen saiatzeak anbiguotasun ebatziezina sortuko lukeelako, III.3 adibidean ikusten den modura.



III.3 adibidea. Nominalizazio baten adibidea. *hark* eta *horrela* bi aditzei lot dakizkieke, anbiguotasuna sortuz.

Antzekoa gerta liteke aditz jokatuarekin ere komunztadurarik ez duten elementuekin, baina gehienetan murriztu egin daiteke (ikus III.4 adibidea). Kasu hauetan, beraz, batzuetan anbiguotasuna sortuko da, adizlagunen moduko adjuntuak lotzeko orduan era bat baino gehiago dagoenean.



III.4 adibidea. Aditz jokatuaren komunztadurek lagundu egiten dute anbiguotasuna gutxitzen: *haiek* eta *horrela* osagaiak *darabilten* aditzarekin doaz, eta ez *kezkatu*-rekin

Irizpide praktikoa erabili da, beraz, analizatuko diren fenomenoak aukeratzeko orduan, eta arau nagusia ebatziezinak diren anbiguotasunak ekiditea izan da.

Hurrengo lerroetan gramatikaren erregelen laburpen txikia egiten saiatuko gara, tratatzen diren fenomenoak argitzeko asmoz. Guztira, laurogeita hamazortzi erregela ditu une honetan gramatika sintaktikoak, era honetan banatuta:

- Eratorpena eta hitz-elkarketa maila lexikalean gertatzen diren fenomenoak izanda, ez dago aldaketarik morfosintaxian dauzkagun erregelatan, eta horregatik eurentzako bost erregelak mantendu egin dira gramatika sintaktikoan. Beste bi erregela hitz anitzeko unitate berezi batzuk tratatzeko definitu dira.
- Izen-sintagma eta adizlagunak tratatzeko, honako elementuen segidak onartuko dira:

aukerazko izenlaguna(k) / adjektiboak +

aukerazko determinatzailea +

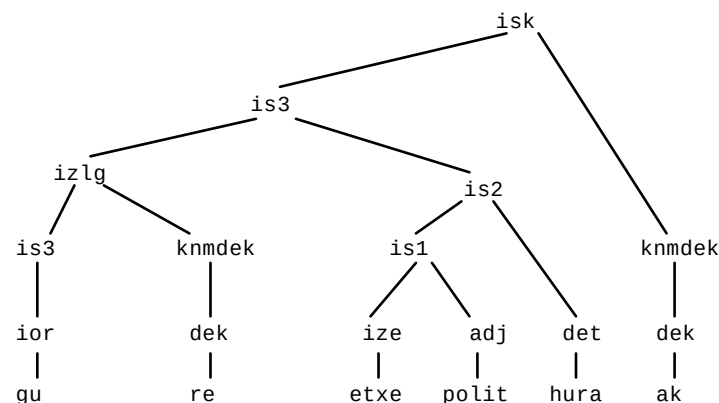
izena / adjektiboa / adberbioa / izenordaina / determinatzailea²⁶ +

aukerazko adjektiboa(k) +

aukerazko determinatzailea(k) +

kasu-marka

Hauek tratatzeko, berrogei erregela definitu dira. Izen-sintagmei eta adizlagunei “isk” izeneko kategoria orokorra esleitu zaie (‘izen-sintagma + kasu-marka’ lotuz sortzen direla adierazteko). Kategoria orokor horren erabilerak izen-sintagmak eta adizlagunak, biak, erregela-multzo berarekin deskribatzea ahalbidetuko du. Kategoria hori sortzeko tarteko kategoriak erabili dira, sintagma osoa sortzeko bidean hartzen diren elementuak kontuan hartuta: lehenengo adjektiboak lotuko dira, ondoren determinatzaileak eta izenlagunak gehitzeko. Bukatzeko, kasu-markak sintagmaren osaketa emango du. Izenordainen eta determinatzaile batzuen berezitasunak ere landu dira (adibidez, izenordainek ez dute adjektiborik edo izenlagunik onartzen). Behar den kasuetan komunztadurak egiaztatu egiten dira, ‘**bi etxea*’ edo ‘**zenbait gizonak*’ bezalakoak ez sortzeko. III.15 irudiak izen-sintagma arrunt baten adibidea du, egindako analisi-zuhaitzaren bidez adierazita (lortutako egitura lehen azaldu den III.13 irudiaren motakoa izango da).



III.15 irudia. ‘*gure etxe polit hark*’ izen-sintagmaren analisi-zuhaitza.

- Aditz-multzoa. Hamazazpi erregela erabiliz aditzarekin lotutako zenbait fenomeno landu ditugu, horien artean hauek:

²⁶ Letra beltzez izen-sintagma edo adizlagunaren gunea idatzi da (osagai horietatik bat agertzea beharrezkoa da).

- Aditz mota eta aspektua adierazteko morfemen kateaketa. Adibidez: ‘*etor + -0*’ (aditz-oinaren marka), ‘*etor + -i*’ (partizipioaren marka), ‘*etor + -i + -0*’ (burutuaren marka), ‘*etor + -tze*’ (aspektu ez-burutuaren marka), ‘*etorri + -ko*’, ‘*etor + -tzen*’.
- Aditz nagusia eta laguntzailearen arteko lotura. Bi hauek biltzen direnean (berdin aditz trinkoa dagoenean), esaldia osatzen da, ondoren beste osagai sintaktikoak erantsiko zaizkiola. Adibidez: ‘*etorri + dira*’, ‘*ez dira + etorri*’, ‘*ager + daiteke*’.
- Partikulak: *ez*, *ba*, *omen*, *ote* modukoak. Honela ‘*ba- + dator*’, ‘*ez- + omen + dute*’ eta antzekoak tratatzeko.
- Baldintzazko aurrizkia: ‘*ba- + letor*’.
- Esaldia. Hogeita bi erregela definitu dira esaldi-mailako egitura sintaktikoentzat. Erregela horietan aditza eta izen-sintagmen arteko tratamendua dugu, dagokion komuntaduraren egiaztatzearekin. Adizlagunak eta mendeko perpaus mota batzuk ere lotzen dira. Azken hauetan anbiguotasuna sortuko da, azpikategorizazioaren informaziorik ez baitugu.

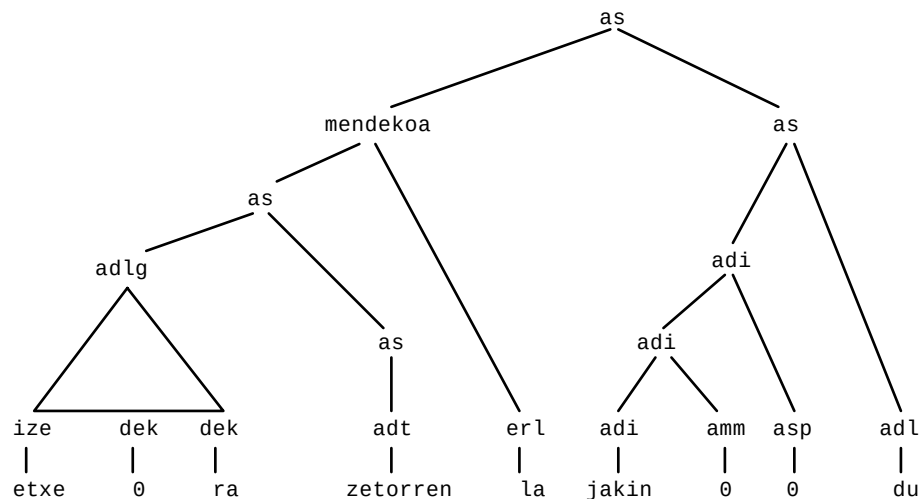
Multzo honetan erregela-kopuru altua izan arren, berriro azpimarratuko dugu maila honetako fenomenoek tratamendua hasi besterik ez dugula egin, hemendik aurrera bide luzea zabaldu zaigula.

- Mendeko jokatuak. Hauentzako erregela bat nahiko izan da, konpletiboak, zehar-galderak, denborazkoak, moduzkoak eta kausazkoak ateratzeko. Erregelak perpausa dagokion atzizkiarekin lotuko du.

Beste erregela berezi bat dago *bait-* kausazko aurrizkiaren kasua tratatzeko.

- Mendeko jokatugabeak. Hauentzat erregela bat nahiko izan da, aditza atzizkiarekin lotzeko (konpletiboak, moduzkoak, denborazkoak, kontzesiboak eta helburuzkoak). Mendeko hauek ez ditugu lotuko inguruko osagai posibleekin, hau da, ez da unitate sintaktiko bat sortuko ‘*palan jokatzea*’ moduko perpausarekin, eta bi osagai solte bezala azalduko dira. Adibidez: ‘*joan + -tea*’, ‘*joan + -teko*’, ‘*egite + -an*’, ‘*egin + -agatik*’.
- Bestelako egitura batzuk tratatzeko azken erregela-multzo bat dago, adjektiboaren graduatzailea (*-ago/-egi*) lotzeko, edo moduzko mendekoak (‘*eman + da*’).

III.16 irudian perpaus baten analisisa dago, fenomeno desberdinen tratamenduekin, adizlaguna eta mendeko perpausa esaldi nagusiari zelan lotzen zaien argitzeko.



III.16 irudia. ‘etxera zetorrela jakin du’ perpausaren analisi-zuhaitza²⁷.

Analisi sintaktiko osoaren konplexutasunaren aurrean, analisi partzialaren eraikuntzari ekin diogu, lengoaiaren oinarritzko egitura sintaktikoak landuz. Oraindik bidea dago landu gabe, kasu batzuetan denbora faltagatik, eta beste batzuetan horrela erabaki delako, zenbait tratamenduk momentu honetan ebatziezinak diren egoeretara eramango baikintuzkete. Aztertu ez diren fenomenoek zerrenda honakoa da:

- Azpikategorizazioa.
- Puntuazioa. Nunberg-ek (1990) eta Briscoe-k (1994) diotenez, puntuazio-zeinuek testu idatzien azpisistema linguistikoa osatzen dute, bere arau eta egitura propioekin. Guk ez dugu alde hau landu. Horregatik, gure analisisetan puntuazio-ikurrak esaldiko azpiesaldiak ezagutzeko baino ez ditugu erabiltzen.
- Osagai ez jarraiak. Oraingo gramatikak bere barruan hartzen dituen fenomenoetan ondoz ondoko osagaiak lotzen dira beti, eta ez gara saiatu ebazten ordena “normaletik” ateratzen diren zenbait ordena-aldaketa. Hauetatik kasu batzuk aipatuko ditugu:
 - Perpaus arruntetan aditz laguntzailea eta nagusia elkarren ondoan egotea eskatzen da. Horrela, ‘*ni etorri naiz*’ edo ‘*zuk ez omen duzu egin*’ modukoek unitate sintaktiko bat osatuko dute, eta ‘*zuk ez duzu hori egin*’ bezalakoak, aldiz, osagai desberdinen segida gisa emango ditugu, *hori* izen-sintagma aditzen artean tartekatuta doalako.

²⁷ Irudian analisi bat baino ez da agertzen, izan ere esaldi hori anbigua da, *etxera* adizlaguna bi perpausari lotzea badagoelako.

- Mendeko perpaus jokatuak aditzean bukatzen direla kontsideratu dugu, hori ematen delako egitura “normaltzat”. Hau dela eta, solte utziko da elementu bat aditzaren ondoren agertuz gero. Adibidez, *‘nik etorriko zela gizona pentsatu nuen’* adibidean *gizona* izen-sintagma *‘nik etorriko zela’* mendekotik kanpo utziko dugu.
- Mendekoak. Hauen artean sinpleenak, alegia, menderagailu-atzizkia morfema bakar batez osatuta daudenak, landu ditugu. Multzo zabala geratzen da tratatzeke, atzizki-bikote batez antolatzen diren kasuak adibidez. Hauen artean ditugu, besteak beste, kontzesiboak (*‘etorri + arren’*, *‘etorri da + -n arren’*), helburuzkoak (*‘egite + -ko asmotan’*), kausazkoak (*‘etorri + -z gero’*, *‘etorri da + -la eta’*), eta moduzkoak (*‘egin du + -en bezala’*).
- Perpausak lotzeko mekanismoak ez dira landu. Juntagailuentzako kasu sinpleena (izen gehi izen) tratatzeko bi erregela definitu dira, baina beste motak bazterrean utzi dira. Lokailuen aldetik ez da tratamendurik egin, oinarritzkoago ikusten genituen beste fenomenoak tratatu direlako.

III.2.2.3 Gramatikaren aberasketa: postposizioak

Oinarritzko gramatika hori egin ondoren, aplikazioak lantzen hasi ginenean, hala nola, azpikategorizazio-informazioaren erauzketa eta errore sintaktikoen tratamendua (tesiare IV. eta V. kapituluetan azalduak hurrenez hurren), konturatu ginen, izen-sintagma eta adizlagunak baino egitura konplexuago duten bestelako sintagma batzuen ezagupena beharrezkoa zela, Euskaltzaindiaren gramatikan postposizio izendatzen dituztenak osatutako sintagmak, hain zuzen ere. Hauek, hitz banatuak izan arren, deklinabide-atzizkien antzera jokatzen dute eta haiek bezala, aditzekiko lotura estua azaltzen dute (adibidez, maiztasun handiarekin agertzen dira *‘-ren alde’* eta *‘-ren kontra’* erabiliz osatutako postposizioak *agertu* aditzaren inguruan, seguruenik azpikategorizazio-erlazio bat erakutsiz). Bigarren tratamendurako ere, errore-iturri interesgarria zela ikusi genuen, esaterako *‘*mendiaren zehar’*, *‘mendian zehar’* erabili orde.

Honengatik, oinarritzko gramatikan egin dugun lehen aberasketa postposizioen tratamendua izan da. Euskarazko postposizioen multzoa zabala denez, arruntenak eta maizen agertzen direnak tratatu ditugu eta ez besterik, Euskaltzaindia-ren (1985) lanean oinarrituz. Postposizioen tratamendua gramatikan sartzeko, atzizki konposatu modura osatzea erabaki genuen (*‘etxe + -tik kanpo’*, *‘gaur + -tik aurrera’*), horrela lan nagusia atzizki horiek biltzea izango delako, ondoren kasu-markak lotzeko erregela orokorrak aplikatuz. Tratatu ditugun postposizioak III.4 taulan daude.

Postposizioak	Adibideak
-tik/-etatik + at/barna/gora/kanpo/zehar/landa	<i>etxetik at</i>
-n + barrena/gora/zehar	<i>mendian barrena</i>
-i + buruz/begira/esker	<i>zenbait etxeri buruz</i>
-ak + inguru/arte	<i>bostak arte</i>
-a/-ak + inguru/alde + -n/-ra	<i>bostak inguruan</i>
-a/-ak + arte/alde + -ko	<i>bostak aldera</i> <i>bostak arteko</i>
0 + (alde/antz/arte/atze/aurre/azpi/barne/gain/goi/inguru/ondo/oste/pe/pare/barru) + -n/-ra/-tik	<i>mendi inguruan</i> <i>mendi aldera</i> <i>mendi aldetik</i>
0 + bide + -z	<i>froga bidez</i>
0 + barna/bila/eske/gisa/gora/legez/truk	<i>etxe bila</i> <i>musu truk</i>
-en + bila/eske/gisa/zale/esku/jabe/ordez/zain/alde/gain/pare/arabera/kontra/bizkar/mende/truke	<i>diruaren bila</i>
-tik/-n + (atze/aurre/behe/landa) + -ra	<i>menditik behera</i> <i>gaurtik aurrera</i> <i>maldan behera</i>
-en + (alde/antz/arte/atze/aurre/azpi/barne/gain/goi/inguru/ondo/oste/pe/pare/mende/truke/buru/leku) + -n/-ra/-tik	<i>mendiaren inguruan</i> <i>mendiaren aldera</i> <i>mendiaren aldetik</i>
-en + (alde/arte/atze/aurre/azpi/barne/gain/goi/inguru/ondo/oste/pe/pare/arabera/aurka/kontra) + -ko	<i>etxearen aldeko</i>
-i + esker + -ak	<i>jaungoikoari eskerrak</i>
-tik + (gora/kanpo/landa) + -ko	<i>menditik gorako</i>
0 + (alde/arte/atze/aurre/azpi/barne/gain/goi/inguru/ondo/oste/pe/pare/barru) + -ko	<i>mendi aldeko</i>
-z + bestalde/kanpo/landa/gain	<i>legez kanpo</i>

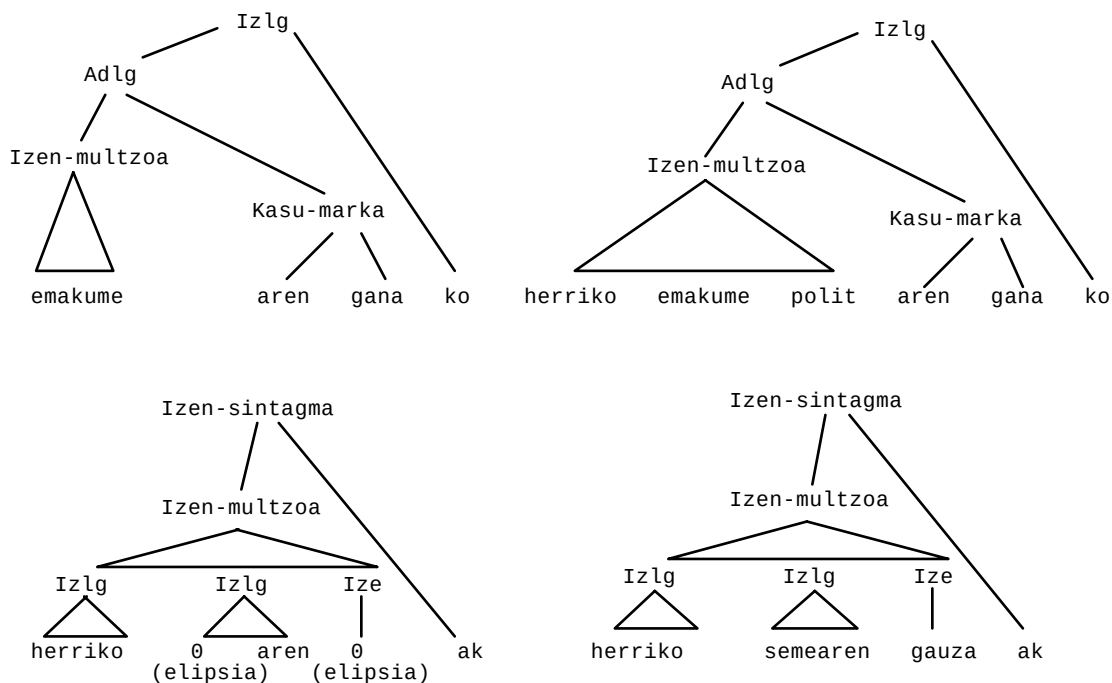
III.4 taula. Tratututako postposizioak eta adibideak.

III.2.2.4 Morfosintaxia eta sintaxiaren arteko muga

Gramatika sintaktikoaren azterketa egin ondoren, ikus daiteke morfosintaxia eta sintaxiaren arteko mugak ez direla argiak, bata bestearen hedapentzat kontsidera daitezkeelako, ebakidura-zati komuna dutelako. Hizkuntza eranskarien tratamendu sintaktikorako saio asko egin ez diren arren, egindako lan batzuetan (Prószyński 1996, Prószyński eta Kis 1999) biak

jauzirik gabeko sekuentzia gisa tratatzea proposatu da. Proposamen hau zentzuzkoa ikusten dugu, bere aldeko arrazoi desberdinak daudelako:

- Morfema oinarriko unitatea bada morfologian eta sintaxian (Goenaga 1980, Abaitua 1988), orduan hitzaren bereizketak bere pisua galtzen du. Hitza testu idatzietan bereizten den unitatea da, baina sintaxiaren ikuspuntutik ez du estatus berezi hori derrigorrez mantendu behar.
- Hitz-mailan gertatzen diren fenomeno askok, sintagma mailakoak diren unetik, sintaxi mailan ere beren paraleloak dituzte, kasu-markaren eransketan edo elipsiaren osaketan bezala. Hau III.17 irudiko adibideetan ikus daiteke hobeto.



III.17 irudia. Hitz-maila (ezkerrean) eta sintagma-mailen (eskuinean) arteko erlazioa.

Dena dela, lehen ere esan dugu badaudela bereziki sintaxiko fenomenoak direnak, hala nola, azpikategorizazioa, ordena librea edo komunztadurak. Gainera, desberdintasunak daude gramatikak kudeatu behar dituen egitura sintaktikoen aldetik:

- Morfosintaxian hitza denez analisiaren muga, kategoria guztiak lexikoitik sortuak dira, hau da, *gizonarengana* formaren analisi morfosintaktikoak bere kategoria *izena* dela adieraziko du.

- Sintaxian, berriz, kategoria sintaktiko berrien erabilera lagungarria dela ikusi da (Goenaga 1980), izen-sintagma, izenlagun, adizlagun edo perpausen modukoak. Lortutako egiturak, ondorioz, antzekoak baina desberdinak izango dira.

Horregatik, nahiz eta zati komuna izan sintaxia eta morfosintaxiaren artean, ezin izan ditugu erregelak automatikoki batetik bestera kopiatu, eta egokitzapen-pauso bat behar izan dugu batetik bestera pasatzeko. Gramatika sintaktikoa hitz-mailan, morfosintaxiaren emaitzaren gainean, definitzeko proba ere egin dugu, eta egiaztatu dugu morfemetan oinarritutakoa baino konplexuagoa, luzeagoa eta ulertzeko zailagoa geratzen dela. Honek bien aplikazio sekuentzialaren bidea uztera eramán gaitu.

Bestalde, pentsa liteke tratamendu morfosintaktikoa desagertu egin daitekeela, dena sintaxiaren barruan sartuz gero, baina hitz-mailako gramatika beharrezkoa zaigu zenbait aplikaziotarako:

- Lematizatzailea. Prozesu honek beharrezkoa du hitz bakoitzeko lema ateratzea, eta horregatik tratamendu morfosintaktikoa beharko da.
- Desanbiguazio morfosintaktikoa. Orain arte garatutako desanbiguaziorako metodo gehienak hitzetan oinarritzen dira, baita euskararako aplikatu direnak ere, horien artean murriztapen-gramatika eta desanbiguazio estokastikoa (Ezeiza *et al.* 1998). Honengatik, ezinbestekoa zen tratamendu hau burutzea.
- Finite State Syntax (Koskenniemi *et al.* 1992, Voutilainen 1994b) moduko lanetan, hitza hartzen da sintaxiaren tratamenduaren oinarritzat. Euskararen kasuan, Aduriz-en (2000) eta Arriola-ren (2000) lanek bide hau landu dute eta etorkizunean aztertuko den ildoak izango da.

Laburpen moduan, esan dezakegu morfosintaxi eta sintaxiak alde komunak izan arren, bakoitza bere aldetik tratatzea erabaki dugula, horrela aukera bakoitzaren onurak eta desabantailak aztertzeko, etorkizunerako egingo diren tratamendu bateratuaren oreka aurkitzeko.

III.2.2.5 Implementazioa

PATRren inplementazio desberdinak daude, horien artean erabilera librekoak, PC-PATR esaterako (Antworth 1994), baina gure kasuan Douglas eta Dale-rena (1992) erabili dugu. Lehen esan bezala, inplementazio horren aldeko motiborik garrantzitsuenetakoak sinpletasuna eta malgutasuna izan dira, soluzio desberdinak esperimentatzeko aukera emanaz. Horiek izan

dira gure lehentasunak beste batzuen gainean, eta horregatik eraginkortasuna bigarren mailan utzi dugu oraingoz, probak egin eta gero hau lantzeko denbora gehiago beharko delako. Dena dela, § III.4.2.1.1en azaltzen den III.12 taulako emaitzak ikusita, esan daiteke une honetan sistema erabilgarria dela testu errealean gaineko aplikazioetarako.

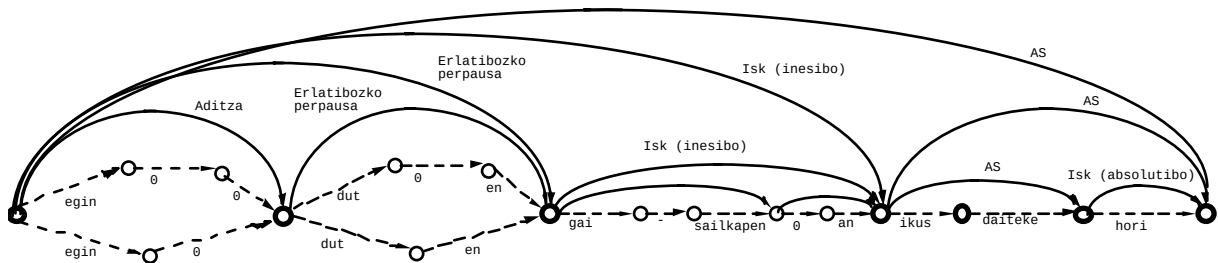
Analizatzailea implementatzeko, ezagututako osagaien taulan (ingelesezko *chart* terminoa erabiltzen da sarritan) oinarritutako Cocke-Kasami-Younger (CKY; Hellwig 1998) algoritmoaren egokitzapena egin behar izan dugu, algoritmoa hitz-formetan oinarrituta baitago eta morfemak unitateak direnean ezin delako zuzenean aplikatu. CKY metodoa behetik gorako eran atera ahal diren elementu guztiak lortzen saiatzen da eta, beraz, egokia da analisi sintaktiko partziala egiteko. Algoritmoak unitate guztiak aztertzen ditu, ezkerretik eskuinera, eta unitate bakoitzeko begiratu egiten du ea erregela baten eskuinaldeko azken osagaia izan daitekeen. Hori gertatzen bada, orduan erregelaren eskuinaldeko lehenengo elementuaren papera beteko duen osagaia bilatzen du *chart*-ean. Horrelakorik aurkituz gero, osagai berria sortzen du *chart*-ean. Bilaketak azkar egiteko, algoritmoan *chart*-eko osagaiak hasiera eta bukaerako posizioen arabera indexatzen dira. Euskararen kasuan, morfema analisirako unitatea dugula, aldaketa bat egin behar izan dugu: hitz-forma baten analisia tratatzeko orduan, bere morfema guztien sekuentzia aztertu beharko da, horrela hasierako algoritmoaren portaera berdina izateko. III.18 irudian algoritmo aldatuaren deskribapena dugu.

```
for H in 1.. Hitz-kopurua do
  for I in 1.. H-garren hitzaren interpretazio-kopurua do
    Aukerak <- H-garren hitzaren I-garren interpretazioko morfema-zerrenda
    while Aukerak ez hutsa do
      Unekoa <- Aukerak listako lehen elementua
      Unekoa chart-ean sartu
      Aukerak <- Aukerak ken lehen elementua
      for "Ezkerrekoa -> Elem1 Unekoa" moduko erregela bakoitzeko do
        for chart-eko Elem1 posible bakoitzeko (Unekoa-ren aurrean bukatzen
bada) do
          if erregela aplikagarria bada
            then Ezkerrekoa sartu Aukerak listan
          end for
        end for
      end while
    end for
  end for
```

III.18 irudia. CKY algoritmoaren egokitzapena morfema analisi-unitate gisa erabiltzeko.

Emaitza *chart* bat izango da non, esaldiaren hasierako morfema guztiez gain, aurkitutako osagai sintaktikoak egongo diren. Analisi behetik gorakoa izanda, ez dago arazorik esaldi

osoaren analisia lortzen ez bada, aurkitutako zatien berri gorde egiten delako. *Chart* horretan, geroago ikusiko dugun bezala, aukera dezente egongo da esaldi osoaren interpretazioa emateko, eta egin beharko den lanetatik bat desanbiguazioa edota interesgarriak diren osagaien bereizketa izango da. III.19 irudian esaldi baten emaitzaren adibidea dugu.



III.19 irudia. *Chart*-aren egoera ‘egin behar dudan gai-sailkapenean ikus daiteke hori’ esaldiaren analisiaren ondoren²⁸.

III.2.2.6 Ondorioak

Euskararen PATR gramatikaren aurkezpena bukatzeko, esan behar dugu euskararen estaldura zabaleko baterakuntza-gramatika diseinatu eta inplementatu dugula. Ondoko puntuak ditugu aipagarrienak:

- Testu errealean analisirako sistema garatu da. Nahiz eta esaldi batzuen analisia ez lortu, sistemak esaldiaren barruko osagaien analisiak emango ditu, horrela analizatzaile partziala lortuz. Lortzen diren osagaiak (izen-sintagmak, adizlagunak, esaldi sinpleak eta mendekoak) aukeratzeko justifikazio nagusia linguistikoa da: elementu horiek lor daitezke egun datu-base lexikalean daukagunarekin. Maila hori baino harantzago joateko, momentuz eskuragarri ez dugun informazioa (aditzen azpikategorizazioa gehienbat) beharko genuke.
- Lexikoi *osoa* baina *sinplea*. Alde batetik, erabilitako oinarri lexikalak EDBLren estaldura handia eta informazio aberatsa du, baina beste batetik, sintaxi osoaren tratamendurako beharrezkoa den informazioa falta du.
- Teoria linguistikorik gabe. Gramatika-formalismo teorikoen zenbait ideia harturik diseinatu da, baina beti ere helburu nagusia —testu errealean tratamendua—

²⁸ Irudian zirkulu lodiak erabili dira hitzen mugak adierazteko, eta zirkulu finak morfemen mugentzako. Marra jarraiek unitate sintaktikoak adierazten dituzte eta besteek, berriz, osagai lexikoak.

baldintzatu gabe. Alde honetatik, uste dugu Abaitua-k (1988) hasitako ildotik jarraitzea badagoela, formalizazio linguistikoan sakonduz.

- Gramatikan izen-sintagmak, adizlagunak, perpaus sinpleak eta mendekoak tratatu dira. Nahiz eta hauek euskararen sintaxiaren gune oinarritzkoena izan, oraindik bide luzea geratzen da sintaxiaren tratamendu osorantz. Horregatik, hau *chunker* izeneko sistemetan (Abney 1997, Basili *et al.* 1998) deskribatu den tarteko maila gisa ikus liteke, hitz-maila baino gorago baina esaldi osoaren analisisa iritsi gabe dagoen bitarteko pauso bezala.
- Sintaxia morfosintaxiaren jarraipena. Beste hizkuntza eranskariekin egin den bezala, bata bestearen hedapena direla ikusi dugu, analisi sintaktikoan morfema oinarritzko unitatetzat hartuz. Gainera, hitz-mailako morfosintaxia ere definitu dugunez, aukera biak lantzeko parada izan dugu, bien konparaketa eta integrazioa ahalbidetuz.

III.3 Euskararen egoera finituko sintaxi-tresnak

Puntu honetan egoera finituko mekanismoetan (automatak eta transduktoreak) oinarritutako bi formalismo aztertuko ditugu. Lehenengo euskararen murriztapen-gramatika eta ondoren XFST formalismoaren gainean egindako lana azalduko dira.

III.3.1 Euskararen murriztapen-gramatika

Baterakuntza-gramatikekin (§ III.2) desberdintasun dezente dauzkan formalismo honekin ere landu dugu tresna linguistiko bat. Helburu nagusia desanbiguazio edo etiketatze morfosintaktikoa (*part-of-speech tagging* deitutakoa zenbait lanetan) egitea izango da, hau da, hitz-forma bakoitzaren interpretazio guztietatik testuinguruaren arabera interpretazio bakarra aukeratzea. Egia esanda, termino horrek euskararen kasuan ez du ondo deskribatzen egin beharreko lana, izan ere ingelesaren moduko hizkuntzetan, hitz-forma bakoitzaren interpretazio bakarra lortzea bere kategoria sintaktikoa asmatzea da gehienbat, baina euskararen morfosintaxi konplexuaren ondorioz, sintaxiaren mailako gertaeren analisisa ere egin beharko da, problema are gehiago konplikatur.

Lehen esan bezala, orain dela hamarkada bat desanbiguazioaren problema analisi sintaktiko, semantiko eta pragmatikoaren ondoren bakarrik ebatz zitekeela zen uste nagusia, baina zenbait lanen ondorioz iritzi hori aldatu da, eta frogatu da emaitza onak lortu ahal direla tratamendu guztia egin barik. Gaur egun bi metodo nagusi daude problema honi ekiteko:

batek, eskuz markatutako corpusetatik ateratako lengoaiaren eredu estatistikoak erabiltzen ditu (Garside *et al.* 1987, Church 1988), besteak, aldiz, eskuz idatzitako erregela linguistikoak. Azken honek besteen aldean hedapen gutxiago duen arren, azken urteotan eredu estatistikoekin baino emaitza hobeak atera ditu. Historikoki eredu linguistikoekin zenbait saio desberdin egin dira urteetan zehar, baina English Constraint Grammar (ENGCG; Karlsson *et al.* 1995, Tapanainen 1996, Samuelsson eta Voutilainen 1997) da eskola honetako nagusia, eta horregatik gure taldea euskararen murriztapen-gramatikaren garapenean sartu da. Formalismo hau lengoaiarekiko independentea eta sendoa den tresna lortzeko diseinatu zen, testu orokorrak desanbiguatu eta analizatzeko. Bere deskribapenak esaldi errealetatik hurbil daude eta analisi sintaktikoaren problema nagusiari, anbiguotasunari, ekiteko pentsatuta daude.

```
"<Herriarenak>"
"herri" IZE ARR DEK GEN NUMS MUGM DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"herri" IZE ARR DEK GEN NUMS MUGM DEK ERG NUMS MUGM @SUBJ
"herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ERG NUMS MUGM @SUBJ
"<ziren>"
"izan" ADL B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"izan" ADL B1 NOR NR_HK ERL MEN MOS @+JADNAG_MP @+JADLAG_MP
"izan" ADL B1 NOR NR_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"izan" ADL B1 NOR NR_HK
"izan" ADT B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"izan" ADT B1 NOR NR_HK ERL MEN MOS @+JADNAG_MP @+JADLAG_MP
"izan" ADT B1 NOR NR_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"izan" ADT B1 NOR NR_HK
"zira" IZE ARR RARE+ DEK GEN NUMP MUGM @IZLG> @<IZLG DEK ABS MG @OBJ @SUBJ
"<mendiak>"
"mendi" IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"mendi" IZE ARR DEK ERG NUMS MUGM @SUBJ
"<ataldu>"
"ataldu" ADI SIN AMM PART ASP BURU NOTDEK @-JADNAG
"ataldu" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ @PRED
"ataldu" ADI SIN AMM PART NOTDEK @-JADNAG
"<eta>"
"eta" LOT JNT @PJ AORG
"eta" LOT MEN KAUS AORG @+JADNAG_MP @+JADLAG_MP
"<saldu>"
"saldu" ADI SIN AMM PART ASP BURU NOTDEK
"saldu" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ @PRED
"saldu" ADI SIN AMM PART NOTDEK
"<egin>"
"egin" ADI SIN AMM ADOIN NOTDEK
"egin" ADI SIN AMM PART ASP BURU NOTDEK
"egin" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ @PRED
"egin" ADI SIN AMM PART NOTDEK
"egin" ADT MDNC NOR_NORK NR_HU NK_HI RARE+
"egin" IZE ARR DEK ABS MG @OBJ @SUBJ @PRED
"egin" IZE ARR ZERO
"<zituzten>"
"*edun" ADL B1 NR_HK NK_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"*edun" ADL B1 NR_HK NK_HK ERL MEN MOS @+JADNAG_MP @+JADLAG_MP
"*edun" ADL B1 NR_HK NK_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"*edun" ADL B1 NR_HK NK_HK
"*edun" ADT B1 NR_HK NK_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"*edun" ADT B1 NR_HK NK_HK ERL MEN MOS @+JADNAG_MP @+JADLAG_MP
"*edun" ADT B1 NR_HK NK_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"*edun" ADT B1 NR_HK NK_HK
"<$.>"
PUNT_PUNT
```

III.5 adibidea. Murriztapen-gramatikaren sarrera ‘Herriarenak ziren mendiak ataldu eta saldu egin zituzten’ esaldiaren analisisian.

III.3.1.1 Murriztapen-gramatika aplikatzeko prozedura

Formalismoa ondorengo pausoetan oinarritzen da:

1. Testua analisirako unitateetan zatitzea (*tokenization*), hitzak, puntuazio-ikurrak eta antzekoetan. Formalismo hau hitza unitate modura erabiltzeko pentsatuta dagoenez ez genuen izan, testuingururik gabeko gramatiketan bezala, morfema unitate gisa hartzeko aukerarik.
2. Analisi morfologikoa (morfosintaktikoa euskararen kasuan). Honetan, hitz-forma bakoitzari bere interpretazio posibleak emango zaizkio, ezaugarri morfosintaktikoen zerrenda baten bidez. III.5 adibidean esaldi baten oraingo analisia dugu, formalismoak eskatzen duen formatuan idatzita.
3. Anbiguotasunaren ebazpena. III.5 adibidean argi geratzen da anbiguotasunaren problemaren tamaina, euskarazko hitz-forma bakoitzak informazio mota aberatsa baitauka. Murriztapen-gramatikak, beraz, sintaxiaren tratamendurako ikuspegi guztiz murriztailea du, hasieran aukera posible guztiekin hasten delako prozesua, eta helburua interpretazioak kentzeko erregelen bidez aukera zuzenarekin geratzea izango delako. Bukaeran, hitz bakoitzaren interpretazio zuzen bakarra eduki beharko genuke, baina errore-tasa jaistearren anbiguotasun-tasa ez-hutsa ere onartuko da, ezagumendu linguistikoa nahiko ez bada soberazko interpretazioak baztertzeko.

III.6 adibidean erregela bi agertzen dira, formalismoak desanbiguaziorako erabiltzen dituen forma biak erakusteko. Lehenengoa interpretazio zuzena aukeratzen saiatzen da, “adi” daukan irakurketa(k) hartuz hitz horrek izen ez-deklinatuaren interpretazioa badauka (“notdek”), ondoren postposizioa badago eta aurretik aditz trinkoa. Bigarren erregela motak aditzaren irakurketa ezabatuko du aurretik izenlaguna badago (izena eta lekuzko genitiboa: “ize” + “gel”) eta ondoren aditz-partizipio deklinatua.

<pre>SELECT (ADI) IF (0 NOTDEK) (1 ADPOSAG) (-1 ADT); # Adibidea: Ez dut ESAN beharrik REMOVE (ADI) IF (-1 IZE + GEL) (1 PART + DEK); # Adibidea: lau egunetako HIL kirastua ...</pre>
--

III.6 adibidea. Euskararen murriztapen-gramatikaren bi erregela.

Anbiguotasun mota desberdinei erreparatuz gero, hiru mota bereiz daitezke:

- a) Anbiguotasun kategoriala, izena/aditza (*harri*) edo aditza/adjektiboa/adberbioa (*erraz*) bezalakoa. III.5 taulan ikusten denez, etiketatzea oinarrizko hemeretzi kategorietara zuzentzen bada, batez besteko 1,55 interpretazio daude hitz bakoitzeko, hitz anbiguo zein ez-anbiguoak kontatuz.

	Interpretazioak hitzeke	Hitz guztien portzentajea %	Hitz-forma anbiguoak %
Forma estandarrak	1,44	93%	33,38%
Aldaerak	1,36	2%	34,44%
Hitz ezezagunak	3,83	5%	99,57%
Guztira	1,55	100%	36,54%

III.5 taula. Katgoria nagusiekiko anbiguotasuna.

- b) Anbiguotasun morfosintaktikoa. Flexioa eta ezaugarri morfologikoen bidez sortzen dena, absolutibo/ergatibo modukoak (*gizonak*), edo III.5 adibidean *saldu* formak “adi” kategoriatik sortzen dituen hiru analisiak. III.6 taulak 10.000 hitz-formako testu batetik ateratako datuak dauzka. Bertan erakusten da hitz bakoitzeko 2,65 analisi daudela batez beste, hitz ezezagunen kasuan 7,05 interpretaziotara iritsiz. Guztira, hitzen %64a baino gehiago dira anbiguoak. Honek desanbiguaziorako lan handia ekarriko du.

	Interpretazioak hitzeke	Hitz guztien portzentajea %	Hitz-forma anbiguoak %
Forma estandarrak	2,43	93%	62,7%
Aldaerak	2,61	2%	84,44%
Hitz ezezagunak	7,05	5%	99,57%
Guztira	2,65	100%	64,88%

III.6 taula. Analisi morfosintaktikoaren anbiguotasun orokorra.

III.5 adibidean ikusten da hizkuntza eranskarietan morfologia eta sintaxia oso lotuta daudela, askotan anbiguotasuna ebazteak funtzio sintaktikoa ebaztea baitakar (absolutibo/ergatibo anbiguotasuna *Herriarenak* moduko hitzetan tratatzean objektu/subjektu aurkaritza ari gara erabakitzen). Gu ahalik eta arinen saiatu gara anbiguotasun mota horiek tratatzen, erabakiak ondorengo tratamenduetarako utzi gabe.

Esan behar da ere anbiguotasun morfosintaktikoaren ebazpenak askotan aditzen azpikategorizazioen moduko elementuen formalizazioa dakarrela berarekin. Problema honi ekiteko, aditz jakin batzuen osagarrien ereduak islatzen duten multzo batzuk definitu dira, oraindik tratamendu orokor batetik urrun egon arren.

- c) Anbiguotasun sintaktikoa. Dena dela, kasu batzuetan anbiguotasuna sintaxiarekin bakarrik dago lotuta. Adibidez, *emateko* formak izenlaguna edo adizlagunaren funtzioa egin dezake (“@izlg” eta “@adlg” funtzio sintaktikoei dagozkienak). Anbiguotasun mota hauek tratatzen ari gara, baina oraindik ez dugu lortu besteekin bezain emaitza onak, eta horregatik ez ditugu aipatuko ondorengo emaitzetan.

Hasieran anbiguotasun mota desberdinen azterketa egin zen. Anbiguotasun kategorialak (datu-basetik datozenak) eta morfosintaktikoak bereiztu ziren. Horren ondoren 28.000 hitzeko corpus desanbiguatua lortu zen, gramatika ebaluatzeko. Corpus hori bi zatitan banatu zen: bata garapenerako eta bestea probarako (azken hau ezin du gramatikariak aztertu garapen-prozesuan, bestela gramatika corpus jakin horren analisira desbidera daitekeelako). Erregelak idatzi eta gero probatu egin ziren behin eta berriro, emaitzak onargarriak izan arte prozesu osoaren birfinketak eginez. Lerro hauek idazteko unean (1999ko abenduan) desanbiguazio-gramatika ia bukatua zegoen. 661 erregela daude, horietatik 593 hitz bat edo biko testuinguruetara mugatzen dira, eta beste 73k testuinguru mugatu gabeak erabiltzen dituzte. 298 erregeletan hitz jakinak tratatzen dira, eta 91 daude desanbiguazio sintaktikorako. Desanbiguazio-erregela orokorren lan handi bat bukatutzat eman daiteke, eta orain hitz konkretuen erregela lexikalizatuagoen idazketa geratzen da.

III.3.1.2 Euskararen aplikazioaren emaitzak

III.7 taulak desanbiguazio morfosintaktikoaren emaitzak aurkezten ditu (Aduriz *et al.* 1997). Interpretazio-kopurua ia erdira jaisten da, analisi zuzenen %97,86a baino gehiago mantenduz. Egokitzen jotzen dugu egindako desanbiguazioa, are gehiago kontuan hartuta lana oraindik jarraitu behar dela, eta gainera hasierako anbiguotasun-tasa altua dela beste lan batzuekin konparatuz gero (Voutilainen 1995). Bukaeran hitzen laurden bat anbigua da oraindik. Emaitzak txarragoak dira hitz ezezagunen kasuan, eta honek eragina izango du horien inguruko hitzetan ere.

Emaitza horiek aztertzeko, argigarriagoa izan daiteke esaldi-mailan lortzen den hobekuntza neurtzea. Hamar hitzeko esaldia hartzen badugu, orduan sarreran $2,65^{10} = 16.861$ interpretazio posible egon litezke, baina desanbiguazioa egin eta gero $1,62^{10} = 123$ interpretazio ditugu. Adibidez, baterakuntzan oinarritutako analizatzaile sintaktikoari lan handia ken diezaioke desanbiguazio-pauso batek.

	Hitzeko interpretazio- kopurua	Hitz-forma anbiguoak %	Interpretazio zuzeneko hitzak %
Sarrera	2,65	64,88%	100%
Irteera	1,62	25,85%	97,86%

III.7 taula. Desanbiguazio morfosintaktikoaren emaitzak.

Hemeretzi kategoria nagusiak aztertuz gero (III.8 taula), hitz bakoitzeko 1,09 interpretazio lortzen dira, lehengo testu berak erabiliz. Gogoratu behar da sarrerako anbiguitasun-tasa txikiagoa zela, hitzeko 1,55 analisirekin. Horrekin batera, interpretazio zuzenen kopurua ere igo egin da, %99,12ra iritsiz.

	Hitzeko interpretazio- kopurua	Hitz-forma anbiguoak %	Interpretazio zuzeneko hitzak %
Sarrera	1,55	36,54%	100%
Irteera	1,09	7,57%	99,12%

III.8 taula. Desanbiguazio kategorialaren emaitzak.

III.7 adibidean III.5 adibidearen emaitza ikusten da. Anbiguitasuna ia erdira jaitsi da, baina kasu batzuetan ez da guztiz ebatzi, errorea egiteko arriskua dagoenean nahiago izaten delako anbiguitasun-tasa txikia baina ez hutsa izatea. Letra etzanaz markatu ditugu interpretazio zuzenak, eta adibide horretan ikusten denez euskararen murriztapen-gramatikak ez du hutsik egin. Anbiguitasun kategorialari erreparatuz gero, hitz-forma guztiek daukate kategoria zuzena. Dena dela, oraindik interpretazio-kopuru garrantzitsua ebatzteke dago, eta hori gure hurrengo lanetarako lehentasun bat izango da.

```

"<Herriarenak>"
"herri" IZE ARR DEK GEN NUMS MUGM DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"herri" IZE ARR DEK GEN NUMS MUGM DEK ERG NUMS MUGM @SUBJ
"herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ERG NUMS MUGM @SUBJ
"<ziren>"
"izan" ADT B1 NOR NR_HK ERL MEN NOTERLT @+JADNAG_MP @+JADLAG_MP
"izan" ADT B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"izan" ADT B1 NOR NR_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"izan" ADT B1 NOR NR_HK AN2 @+JADNAG
"<mendiak>"
"mendi" IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"mendi" IZE ARR DEK ERG NUMS MUGM @SUBJ
"<ataldu>"
"ataldu" ADI SIN AMM PART ASP BURU NOTDEK
"ataldu" ADI SIN AMM PART NOTDEK
"<eta>"
"eta" LOT JNT @PJ AORG
"<saldu>"
"saldu" ADI SIN AMM PART ASP BURU NOTDEK
"saldu" ADI SIN AMM PART NOTDEK
"<egin>"
"egin" ADI SIN AMM PART ASP BURU NOTDEK
"<zituzten>"
"*edun" ADL B1 NR_HK NK_HK ERL MEN ERLT @+JADNAG_IZLG> @+JADLAG_IZLG>
"*edun" ADL B1 NR_HK NK_HK ERL MEN MOS @+JADNAG_MP @+JADLAG_MP
"*edun" ADL B1 NR_HK NK_HK ERL MEN ZHG @+JADNAG_MP @+JADLAG_MP
"*edun" ADL B1 NR_HK NK_HK
"<$.>"
PUNT_PUNT

```

III.7 adibidea. Murritzapen-gramatikaren emaitza III.5 adibideko esaldian.

III.3.1.3 Murritzapen-gramatikaren balorazioa

Desanbiguazioaz aparte, murritzapen-gramatikaren beste aplikazio bat analizatzaile sintaktiko bati lana kentzea da, tratatu beharreko aukera-kopurua dezente txikituz. Horregatik guk PATR gramatikaren aplikazioaren aurretik erabiliko dugu euskararen murritzapen-gramatika. Abiaduraren irabazia neurtzeko, bi eratara probatu dugu analizatzailea, eta III.12 taulan (§ III.4.2.1.1) horren emaitzak ikusten dira.

Esan behar da ere aukeren baztertze horrek bere alde negatiboa izango duela, murritzapen-gramatikak aukera zuzen bat kentzen duenean analizatzaile sintaktikoari ez zaiolako paradarik emango analisi ona ateratzeko. Hau neurtzeko, desanbiguazio morfosintaktiko osoaren emaitza ikusi beharko genuke. III.7 taulan horren estimazioa %2,16koa bada, orduan esaldien batez besteko luzera 15-20 hitzekoa izanda, gutxi gorabehera bost esalditik batean errore bat izango genuke. Dena dela, errore hauen maiztasuna eta beren eragina aztertzeak ikerketa sakonagoa egitea mereziko luke. Gainera, momentuz gure analizatzaile sintaktikoak analisi partzialak egiten dituenenez, errorearen eragina hori baino txikiagoa izango da, batzuetan ez baita esaldi osoaren analisisa lortuko, eta orduan errore-tasaren neurketa ez genuke esaldika egin behar, lortu nahi den egitura bakoitzeko baizik.

Bukatzeko, esan behar dugu murriztapen-gramatikaren formalismoa egokia ikusten dugula euskara bezalako hizkuntzen tratamendurako, osagaien ordena librea eta morfologia aberatsa dutenak, anbiguotasun-arazo zail baten aurrean emaitza onak ematen dituelako. Euskararen deskribapenaren azalpen osoa (Aduriz 2000, Arriola 2000) tesietan egiten da. Dena den, badaude zenbait arazo, lan gehiago egitea mereziko dutenak:

- Voutilainen-ek (1994) dioenez, zailtasunak daude perpausen arteko mugak bereizteko eta hori, nahiz eta desanbiguazio kategoriala ebazteko arazo handia ez izan, garrantzitsuagoa da anbiguotasun morfosintaktikoa kentzeko.
- Anbiguotasuna hitz-mailan adierazten denez, askotan ezin dira baztertu esaldi-mailan ezinezkoak diren aukerak. III.8 adibidean (adibidea desanbiguazioaren ondorengo emaitza batena da), *bere* eta *bizitzak* bi eta hiru analisi dauzkate, hau da, sei interpretazio guztira, baina horietatik zenbait ezinezkoak dira. Adibidez, *bere* determinatzailea plurala bada (bigarren interpretazioa) *bizitza*-ren bigarren interpretazioarekin (singularra) ezin da bat etorri. Murriztapen hau ezin da zuzenean aplikatu, *bizitza*-ren interpretazio singularra bat datorrelako *bere* formaren lehenengo irakurketarekin. Ondorioz, murriztapen-gramatika interpretazio bat baztertzeko ziur egon behar dugu hori ezinezkoa dela inguruko hitzen interpretazio guztientzat. Hau guztia murriztapen-gramatika hitzetara mugatuta dagoelako gertatzen da. Ondoren ikusiko dugu egoera finituko syntaxian arazo hau desagertu egingo dela, analisirako unitatea esaldia izanik zentzurik gabeko interpretazioak bazter daitezkeelako.

```

...
"<bere>"
"bera" DET ERKIND NUMS DEK GEN AORG @IZLG>
"berak" DET ERKIND NUMP DEK GEN @IZLG>
"<bizitza>"
"bizitza" IZE ARR DEK ABS MG AORG @OBJ @SUBJ
"bizitza" IZE ARR DEK ABS NUMS MUGM AORG @OBJ @SUBJ
"bizitza" IZE ARR
"<pertsonala>"
...

```

III.8 adibidea. Murriztapen-gramatikaren zailtasuna anbiguotasun mota batzuk ebazteko.

- Beste arazo bat dakar oinarria hitza izateak. Adibidez, izen-sintagmak eta adizlagunak ezagutzeko orduan, hau zeharkako moduan egin beharko da, balizko izen-sintagma eta adizlagun horien hitz desberdinen bidez. Honek erregelen ziurtasunean eta irakurgarritasunean izango du ondorioa, mota horretako elementu bat aipatu behar den bakoitzean bere aukera posibleen bidez egin beharko baita.

Honekin lotutako beste problema bat hitzaren barruko morfologia konplexua da (ikus II. kapitulua). Hau ekiditeko, § VI.1en tratamendu morfosintaktikoa eta

murritzapen-gramatikaren integrazioa aurkeztuko da, euskararen EUSLEM lematizatzaile/etiketatzailearen barruan.

- Oraindik ebatzi ez den anbigutasun-kasu ugari daude, batez ere alde morfosintaktikoan. Hauetatik batzuk ahalik eta azkarren tratatuko dira, mendeko perpausekin lotutakoak bezala, baina beste batzuk informazio semantikoa edo pragmatikoa beharko lukete (*‘eskola garaia’*).

Arazo hauek direla eta, sintaxiaren tratamenduarekin jarraitzeko murritzapen-gramatika baino harantzago doazen formalismoak definitu dira, lehen aipatutako Finite State Syntax modukoak (Koskenniemi *et al.* 1992, Voutilainen 1994b). Horrez gain, murritzapen-gramatika, sintaxiari ekin aurretik aplikatzen den desanbiguatzaile morfologikoa edo *tagger* gisa ere ikus daiteke. Guk bide hau probatu dugu eta horregatik, geroago (§ III.4) azalduko dugun bezala, baterakuntzan oinarritutako sintaxiaren aurreprozesuan aplikatuko dugu euskararen murritzapen-gramatika, desanbiguazio morfologikoa zehaztasun handiz egiteko, eta era horretan analizatzailearen lana errazteko.

Beste alde batetik, euskararen murritzapen-gramatikaren garapenak lagundu egin du desanbiguazio morfologikoaren tratamendurako ezagutza linguistikoan eta estatistikoan oinarritutako hurbilpenen arteko eztabaidan. Gure emaitzek Samuelsson eta Voutilainen-en (1997) baieztapenak indartzen dituzte, alde batetik frogatzen delako erregela linguistikoen bidezko corpusen desanbiguazioan zero inguruko errore-tasa lor daitekeela, eta bestetik argi geratzen delako murritzapen-gramatikaren arrakastak ez duela zerikusirik etiketatze-sistema sinplearen erabilerarekin, euskararako erabilitakoa oso konplexua delako beste lan batzuekin konparatuz gero.

Euskararen desanbiguazio morfosintaktiko osoaren bidean, estatistikan oinarritutako desanbiguazioa ere landu da (Ezeiza 1997). Honen emaitzak, era automatikoan lortzen badira ere, ez dira murritzapen-gramatikaren bidez lortutakoak bezain onak izan, lehen azaldu dugun testuinguru mugatuen (bi edo hiru hitz) arazoarengatik gehienbat. Dena dela, hitz bakoitzeko aukera bakarra atera daiteke, murritzapen-gramatikak batzuetan anbigutasun-tasa txiki bat uzten duen bitartean. Horregatik, bien konbinazioan (Ezeiza *et al.* 1998) lehenengo murritzapen-gramatika aplikatzen da, ondoren desanbiguazio estokastikoari ekiteko. Modu honetan desanbiguazio morfosintaktiko osoan %3ko errorrea lor daiteke, anbigutasuna eta doitasunaren elkarren kontrako oreka lortuz: errorerik ez bada egin nahi, orduan anbigutasun-tasa txikia onartu behar da eta desanbiguazio osoa nahi bada, berriz, errore-tasak gora egingo du. Honi buruz (Ezeiza 2000) tesian aurki daiteke informazio gehiago.

III.3.2 Egoera finituko sintaxia (XFST)

III.3.2.1 Sarrera

§ III.1.1.1.2n egoera finituko sareen oinarriak aipatu ditugu. Murritzapen-gramatika hauen adibidetzat ikusita ere, esan behar dugu formalismo mugatua dela (ikus § III.3.1.3n egindako kritikak). Beraz, puntu honetan hurbilpen horren orokortzera joko dugu, murritzapen-gramatikak egiten duen aukera-baztertzearaz aparte, egoera finituko mekanismoek beste eragiketa-multzo aberatsagoa baitakarte beraiekin (Karttunen *et al.* 1997, Aït-Mokhtar eta Chanod 1997, Chanod eta Tapanainen 1996ab, Grefenstette 1996, Gala 1999), oinarritzko formalismoan funtsezko aldaketarik egin gabe.

Gure kasuan, *XFST* edo *Xerox finite state tool* izeneko tresna erabili dugu euskararen aplikaziorako (Karttunen *et al.* 1997). Adierazpen erregularrak oinarri dituen sistema honek egoera finituko kalkuluaren pean era anitzeko eragiketak egiteko aukera emango du. Erregela lokalen bidezko sistema osoak (Silberztein 1997) egoera finituko kalkulurik gabe egin diren arren, kalkulu honen ahalmen espresiboari esker *XFST* tresnan erregela-multzoak mantentzeko eta aldatzeko erraztasunak emango dira, lengoaiaren tratamendurako aplikazioen garapena azkartuz. Hurrengo puntuan (§ III.3.2.2) *XFST* tresnaren egoera finituko kalkuluaren eragiketa nagusien azalpen laburra egingo dugu.

III.3.2.2 Eragiketa nagusiak

Egoera finituko eragiketa garrantzitsuenen azalpen laburra egingo dugu hemen, ondorengo puntuetan erabili egingo baititugu. Azalpen luzeagoa (Karttunen *et al.* 1997²⁹) artikuluan ematen da.

Hasteko, adierazpen erregular batek multzo bat definitzen du. Multzo hauek bi motakoak izan daitezke:

- Lengoaia erregularra, kateez osatutako multzoa da. Automata erregularren bidez kode daiteke.

²⁹ Eragiketa berriagoak aztertzeko, <http://www.rxc.xerox.com/research/mlttfst/home.html> helbidean daude azken eguneraketak.

- Erlazio erregularra, kate-bikoteez osatutako multzoa da. Transduktoreen bidez kode daiteke. Adibidez, $a:b$ adierazpenak a eta b kateen bikotea adierazten du. Erlazio batean bi lengoia daudenez, lehenengoari goikoa eta bigarrenari behekoa deitzen zaie.

Adierazpen konplexuak idazteko honako ikurrak erabiltzen dira:

- Makoak adierazpenak biltzeko erabiltzen dira. Adibidez, $[A]$ adierazpenak A lengoia adierazten du. $[]$ adierazpenak kate hutsa adierazten du.
- Ikur baten esanahi berezia alda daiteke komatxo bikoitzen artean jarritz. Adibidez, $"["$ adierazpenak ezkerreko makoa deskribatzen du.
- $?$ ikurrak edozein sinbolo adierazten du.

Lengoia eta erlazio erregularrak adierazpen erregularren bidez adieraz daitezke.

Adierazpen erregularren gaineko zenbait eragiketa definitu dira:

- Bildura: $A | B$.
- Kateaketa: $A B$.
- Errepikapena: A^+, A^* .
- Aukerazko elementua: (A) .
- Barruan egotea: $\$A$ ($'?^* A ?^*'$ adierazpenaren baliokidea).

Aurreko eragileak lengoia zein erlazio erregularrekin erabil daitezke, emaitza beti hasierako osagaien arabera izango delako. Ondorengo eragileak, aldiz, lengoia erregularrei bakarrik aplikatu dakizkieke, emaitzatzat lengoia erregularra emanez:

- Erlazio osagarria (ukapena): $\sim A$.
- Ebakidura: $A \& B$.
- Kenketa: $A - B$.

III.9 taulan adierazpen erregularren adibideak azaltzen dira, bakoitzak deskribatzen duen lengoia/erlazioaren azalpen batekin.

Adierazpen erregularren definizioa	Deskribatutako lengoia/erlazioa
<pre>define HH "@HH"; define HB "@HB";</pre>	<i>HH</i> (hitzaren hasiera) eta <i>HB</i> (hitzaren bukaera) izeneko lengoaiak definitu dira, eta bakoitzak kate bakarra (<i>@HH</i> eta <i>@HB</i> sinboloez osatua) du
<pre>define Muga [HH HB];</pre>	<i>Muga</i> izeneko lengoaiak bi kate ditu (<i>@HH</i> eta <i>@HB</i>).
<pre>define EzMuga ~\$[Muga];</pre>	<i>Ezmuga</i> lengoia infinituak bere barruan <i>@HH</i> eta <i>@HB</i> sinboloak ez dituzten kate guztiak dauzka.
<pre>define IskErg [OsagaiSintaktikoa & \$["+kat" "+isk"] & \$["+kas" "+erg"]];</pre>	<i>IskErg</i> lengoaiak <i>OsagaiSintaktikoa</i> multzoko kateak dauzka, kategoria izen-sintagma (+kat +isk) eta kasua ergatiboa (+kas +erg) dutenak.
<pre>define EsaldiAdibidea [Isk* Aditza Isk*];</pre>	<i>EsaldiAdibidea</i> multzoak izen-sintagmak (zero edo gehiago), ondoren <i>Aditza</i> , eta bukaeran berriro izen-sintagmak dauzkaten lengoaiak/erlazioak adieraziko ditu.

III.9 taula. Adierazpen erregularren adibideak.

Oinarrizko eragile horiek erabiliz, eragile berriak defini daitezke:

- Murriztapena: $A \Rightarrow B _ C$. A bakarrik onartuko da ezkerretik B eta eskuinetik C duenean. Mota honetako adierazpenak debekuak adierazteko erabiltzen dira, desanbiguazioan kasu (adibidez, murriztapen-gramatikan).
- Ordezpena: $A \rightarrow B \parallel L _ R$. Honek erlazio bat sortzen du, bi kate lotzeko. L A R testuingurua duen katea L B R katearekin erlazionatu egiten da, hau da, A Brekin “ordeztu” egiten du. Erlazioa identitatea izango da testuinguru hori ez bada agertzen hasierako katean. A multzoko kate bat era askotan aurki daitekeenean (adibidez, X+ moduko definizioa badu), bakoitzeko ordezen bat egingo da.
- Parekatze luzeenaren ordezena: $A @\rightarrow B \parallel L _ R$. Lehengoaren antzekoa, baina Aren aukera asko daudenean, eta batek barruan beste bat daukanean beti hartuko du luzeena.

- Konposizioa: [A .o. B]. Bi erlazio hartuta, erlazio berri bat emango du, Aren goiko lengoaia Bren beheko lengoaiarekin erlazionatuz. Adibidez, [[a | b]:c .o. c:d] konposizioaren emaitza [[a | b]:d] erlazioa da, hau da, a:d eta b:d bikoteen multzoa.
- Konposizio biguna edo malgua (*lenient composition*; Karttunen 1998): [A .O. B]. Konposizio arruntak duen arazo bat konpontzeko erabiltzen da eragile hau. Desanbiguazioa helburu nagusia denean, murriztapenak konposatu egiten dira hasierako kate anbiguoarekin, irakurketak gutxitzeko asmoz (EsaldiAnbigua .o. Murriztapen1 .o. Murriztapen2 .o. ...). Baina kasuren batean murriztapen batek irakurketa guztiak ezabatuko balitu, orduan emaitza kate hutsa izango litzateke, eta hori ezin da onartu, normalean emaitza bat nahi izaten delako. Eragile honek egoera hori ekidingo du. [Esaldia .O. Murriztapena] kasuan murriztapenaren aplikazioaren emaitza kate hutsa ez bada, hori izango da emaitzaren beheko lengoaia, konposizio arruntean bezala, baina emaitza kate hutsa bada (horrek esan nahi du murriztapenak interpretazio guztiak ezabatu dituela) orduan beheko lengoaia Esaldia-ren beheko lengoaia izango da. Lortzen den efektua hasierako esaldia “berreskuratzea” izango da. Konposizio-modu honen bidez murriztapenak aplika daitezke, gogorrenak (dena baztertzen dutenak) ez direla kontuan hartuko ziurtatzen baita.

Eragile berri hauek oinarritzko eragileen bidez definituta daude, eta ondorioz dena egiten da egoera finituko kalkuluaren pean³⁰, horrela eragileak era konplexuetan konbinatu eta konposatzeko malgutasuna lortzen dela. Erabiltzailearentzat askoz erosoagoa da goi-mailako eragiketa hauek erabiltzea eragiketa sinpleak baino, askoz adierazkortasun handiagoa dutelako. Nolabait bi eragiketa-maila hauen arteko diferentzia goi-mailako eta behe-mailako programazio-lengoaien artekoaren antzekoa da.

III.3.2.3 Euskararen egoera finituko analizatzaile sintaktikoa

Euskararen tratamendu automatikoaren azterketan, egoera finituko sintaxia aplikatzeko bide bat baino gehiago egon daiteke. Alde batetik, egoera finituko analizatzaileak sintaxi osoa deskribatzeko erabiltzen ari dira (Chanod eta Tapanainen 1996b, Grefenstette 1996, Gala 1999). Beste alde batetik, XFSTren bidez murriztapen-gramatikaren zein baterakuntzan oinarritutako analizatzaile sintaktikoaren ondorengo prozesuak deskriba daitezke, hau da, aurretik aplikatutako tresna linguistikoen ondoren erabiltzea ere badago. Adibidez, beste

³⁰ Adibidez, $A \Rightarrow B _ C$ murriztapena era honetan definitzen da: $[\sim[[\sim?^* B] A ?^*] | [?^* A \sim[C ?^*]]]$

tresnek egindako akatsen zuzenketak, geratzen den anbiguotasunaren ezabaketa edo sintaxiaren jarraipena egiteko egitura berrien osaketa; horrela, ikuspegi murriztailea eta eraikitzailea inolako arazorik gabe konbinatzeko aukera lortuko da. Beraz, osagai linguistiko desberdinen konbinazio eta integraziorako eredu desberdinak probatzeko tresna lagungarria izango da. § IV.4en XFST eta garatutako beste tresnak (murriztapen-gramatika eta baterakuntzan oinarritutako analizatzaile sintaktikoa) elkartu eta integratzeko moduak aztertuko dira. Atal horretan integrazio-modu desberdinen artean aukeratu duguna aurkeztuko da: egoera finituko sintaxia murriztapen-gramatikaren eta baterakuntzan oinarritutako sintaxiaren ondoren erabiltzea, modu sekuentzialean. Beraz, egoera finituko sintaxiaren gure erabilpena ez da izan sintaxia aztertzeko tresna autonomoarena, beste lan linguistikoen ondorengoa baizik. Horregatik, egoera finituko analizatzaile sintaktiko orokor bati buruz baino, aplikazio jakin baterako erabilpenak azalduko ditugu aplikazio bakoitza aipatzeko orduan, IV., V. eta VI. kapituluetan. Egoera finituko osagai linguistikoaren ebaluazioa ere, arrazoi berdinagatik, ez dugu hemen egingo, eta aplikazio bakoitzaren emaitzen gainean emango dugu.

III.4 Formalismoen konparazioa eta integrazioa

Aurreko ataletan azaldu dugunez, hiru eredu desberdin landu ditugu euskararen sintaxiaren tratamendurako: murriztapen-gramatika, baterakuntzan oinarritutako gramatika eta egoera finituko sareetan oinarritutako formalismoak. Bakoitza erabiltzearen aldeko eta kontrako arrazoiak daude. Inplementatu diren sistema gehienetan horietako bakarra landu da. Puntu honetan aztertu egingo ditugu bakoitzaren alde onak nahiz desabantailak (§ III.4.1), eta ondoren formalismo hauen konbinazioa egingarria dela eta gainera konbinazio hori eredu bakoitzaren abantailak ateratzeko moduan egin daitekeela frogatzen saiatuko gara (§ III.4.2).

III.4.1 Formalismoen konparazioa

Jatorriz eta helburuz oso desberdinak diren formalismoak konparatzeko, analisi sintaktikoaren zenbait problemari ekiteko gaitasuna aztertuz egingo dugu (ikus III.10 taula). Problema horiek analisi sintaktikoaren arazo orokorrak direnez, formalismo bakoitzak bere erara tratatuko ditu; batzuk modu errazean, formalismoaren diseinuan problema horiek kontuan hartu zirelako; beste batzuk, ordea, diseinu-fasean arazo horri erreparatu ez ziotelako edo, moldaketen bidez

tratatu behar izan ditu. Horregatik esan behar dugu formalismoak ez direla egokiak edo desegokiak orokorrean, aplikazio konkretuen arabera baizik.

	MG	Baterakuntzan oinarritutako gramatika	XFST
Analisia egiteko unitatea	hitza	morfema/hitza/osagai sintaktikoa	morfema/hitza/osagai sintaktikoa
Egitura sintaktiko konplexuak erabiltzeko ahalmena	-	+	-
Osagai sintaktikoen arteko komunztadura	+	++	+
Desanbiguazio-murritzapenak adierazteko ahalmena	+	-	+
Patroi sintaktikoak adierazteko ahalmena	+	-	++
Eraginkortasuna	+	+	+
Esaldiaren analisi sintaktiko osoa	-	gramatika osoa	+

III.10 taula. Sintaxiaren tratamendurako garatutako tresnen ezaugarriak.

Aipa ditzagun modu laburrean ezaugarriekiko baliagarritasuna:

- Analisia egiteko oinarritzko unitatea. MGk hitza soilik onartzen du analisia egiteko, eta honek problemak sortzen ditu kasu batzuetan, hitza baino harantzagoko osagaiak tratatu behar direnean, hitz anitzeko unitateak (hitz anitzeko unitatea anbigua denean gehienbat, hitz solteak edo unitate osoa hartzeko aukera dagoenean) edo osagai sintaktikoak adibidez. Baterakuntza-gramatikan eta egoera finituko gramatikan ez dago inolako arazorik edozein informazio mota erabiltzeko, ahalmen deskriptibo handiagoa dutelako.
- Egitura sintaktiko konplexuak erabiltzeko ahalmena. Baterakuntzak egitura hierarkiko konplexuak deskribatzeko ahalmena du, eta hauek beharrezkoak dira tratamendu-mota askotan (morfosintaxia kasu). MG eta egoera finituko gramatiketan, berriz, deskribapen sekuentzialak erabili behar dira. Ikuspuntu linguistikotik askotan egitura hierarkikoak behar izaten direnez, azken bi sistema mota hauetan moldaketak edo sinplifikazioak egin behar izaten dira.
- Osagai sintaktikoen arteko komunztaduren egiaztapena. Honekin formalismoek komunztadura tratatzeko duten erraztasuna neurtu nahi dugu. Gure ustez, kasu

honetan baterakuntza da ahaltsuena, ekuazio sinple baten bidez egiaztapen konplexuak egin daitezkeelako baterakuntzari esker. Adibidez, aditza eta osagai baten arteko komunztadura 'X1/kom $\Leftrightarrow$ X2/obj/kom' moduko ekuazio bat erabiliz defini daiteke, eta horrek kom ezaugarriaren barruko informazio guztia bat etortzea egiaztatuko du, horien artean kasua eta numeroa (kom/kas eta kom/num), balio posible guztientzat. MG eta egoera finituko formalismoetan, ordea, bi osagaien arteko kasuaren adostasuna ziurtatzeko balio posible guztien zerrendatzea egin behar da (biak absolutibo izan edo biak ergatibo edo ...).

Ondorioz esan behar dugu hiru formalismoekin adieraz daitezkeela mota horretako egiaztapenak, baina baterakuntzaren ahalmen espresiboari esker era trinko eta irakurgarriagoan egin daitekeela baterakuntzan oinarritutako formalismoetan. Aspektu hau garrantzitsua da tratamendu sintaktiko orokorrerako eta are gehiago gure kasuan tratatu nahi ditugun komunztadura-erroreerako.

- Desanbiguazio-murritzapenak adierazteko ahalmena. MG eta egoera finituko sintaxia bereziki landu dira patroien bidezko prozesuak deskribatzeko, eta horien artean desanbiguazioa da garrantzitsuenetariko bat. Horregatik, bi formalismo hauek egokiak dira desanbiguaziorako. Testuingururik gabeko gramatika eta baterakuntzan oinarritutakoak, aldiz, mekanismo sortzaileak dira, osagai sintaktikoak noiz zilegi diren adierazteko, aukera anitzen sorkuntza ekiditeko modurik gabe. Horregatik, gramatikei beste mekanismo batzuk gehitu zaizkie, gramatika estatistikoak erabiliz (Briscoe eta Carroll 1993, Collins 1997) edo analisia determinista bihurtzeko prozedurak gehituz (Marcus 1980, Hermjakob eta Mooney 1996).
- Patroi sintaktikoak adierazteko ahalmena. Patroiak definitzea ezinbestekoa da edozein aplikaziotarako. Testuingururik gabeko gramatikekin gertaera linguistiko orokorrak deskriba daitezke, baina maiz gertatzen diren osagai sintaktiko askok (hitz anitzeko unitateak edo kokakidetzak, adibidez) ez dituzte arau orokor horiek betetzen. MG eta egoera finituko sintaxia oso egokiak dira horretarako. Gure kasuan lehentasun handikoa da aspektu hau, buruan dauzkagun bi aplikaziotan, azpikategorizazio-informazioaren erauzketan eta errorearen tratamenduan, patroien erabilpena beharrezkoa ikusten dugulako (ikus IV. eta V. kapituluak).
- Eraginkortasuna. Une honetan murritzapen-gramatikaren inplementazioa da azkarrena. Baterakuntzaren tratamendua konplexua denez, berau erabiltzen duten inplementazioak motelagoak omen dira. XFSTren egoera finituko inplementazioa, berriz, azkarra omen da. Gure esperimentuetatik atera ditugun neurriak emango ditugu ondoren, baina badakigu tresnen bidezko konparazioa egiteko problema beraren gainean egin beharko litzatekeela.

Gure esperimentuetan MG segundoko mila hitzetik gora analizatzeko gai da. Baterakuntza-analizatzailearekin egindako probetan hamabost hitz segundoko lortu ditugu analizatzailearen oraingo egoeran, baina lehenago esan dugu oraindik abiadura azkartzeko aukera anitz dagoela, magnitude-ordena batekoak edo gehiago, (Kiefer *et al.* 1999) lanean aipatzen den antzera. Egoera finituko analizatzailearen kasuan, une honetan motelena da, segundoko lau hitz tratatzea lortu dugula. Dena dela, hori gure inplementazioaren ezaugarritzat hartu behar dugu, momentuz gramatikaren deskribapenean jarri dugulako arreta nagusiki, eta etorkizunerako espero dugu irabazpen handia lor daitekeela. Edozein kasutan ere, aspektu hau oraindik esperimentatzeko gaia da, eta gogoan izan beharko dira agertzen diren arazoak (Tapanainen 1997, Beesley 1998a).

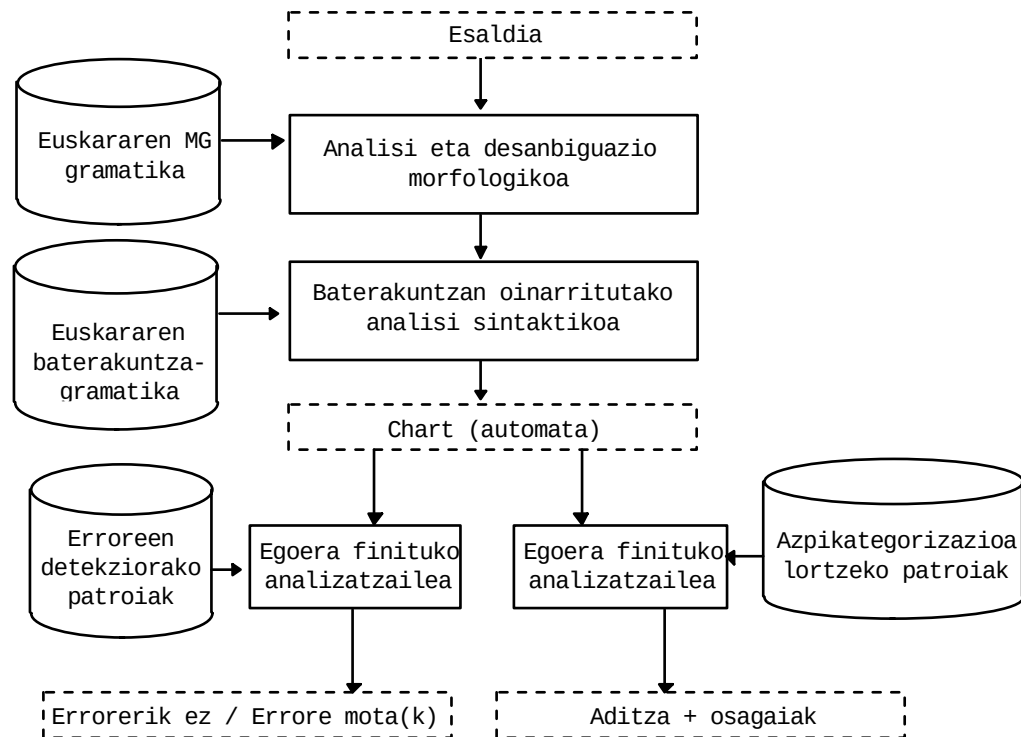
- Esaldiaren analisi sintaktiko osoa (*chunk*-en egituraketa). Lehenago esan dugu testu errealeko esaldiak analizatzeko hartu dugun hurbilpena osagai partzialen analisisa egitea dela. Honen abantaila nagusiak bi dira: osagai partzialak erabilgarriak direla zenbait aplikaziotarako eta gainera analisi partzial horiek analisi osoa egiteko abiapuntua izan daitezkeela, modulartasuna lortuz. Osagai partzialen lehen pauso hau izen-sintagma, adizlagun, aditz-multzo eta mendekoen mailan definitu dugu. Puntu honetan analisi partzial horretatik aurrera jarraitzeko bideragarritasuna ebaluatu nahi dugu, ondorengo urratsetan esaldi osoen analisiari ekiteko. MGren aldetik muga bat ikusten dugu, hitz-mailan definituta dagoelako batetik eta bestetik formalismoa desanbiguaziorako bideratuta dagoelako hein handi batean. Horregatik, formalismoaren asmatzaileak egoera finituko formalismora pasa ziren analisi osoa egiteko, *Functional Dependency Grammar* izeneko formalismora (Järvinen eta Tapanainen 1998). Baterakuntza-gramatiken ikuspuntutik badago jarraitzea gramatika osorantz (Briscoe eta Carroll 1993), baina arazoak egongo dira estalduran (testu errealeko esaldi guztiak analizatzeko zailtasuna) zein anbiguotasunaren gorakadan (normalean mekanismo estatistikoak gehitu behar izan dira, horrek lotuta dakarren corpus etiketatuen beharrarekin).

III.4.2 Formalismoen integraziorako ereduak

III.10 taulan ikusi ditugu formalismo bakoitzaren alde sendoenak. Formalismoak konbinatuz tresna ezberdinen abantailen gehitzea lor daiteke, eta puntu honen xedea zenbait konbinazioen azterketa egitea da. Hasteko, hiru tresnen aplikazio sekuentziala aztertuko dugu, hori baita gure aplikazioak lantzeko erabili duguna (§ III.4.2.1). Ondoren, beste aukera batzuk aurkeztuko dira (§ III.4.2.2).

III.4.2.1 Tresnen aplikazio sekuentziala

Tresnak konbinatu eta erabiltzeko aukera asko egonda ere, gure kasuan tratamendu sintaktikoari ekiteko sistema oso baten diseinuan ari garenez, proposamen desberdinen artetik aukera bat egin dugu, egoera finituko tratamendua baterakuntzan oinarritutako analizatzailearen ondoren egiteko (ikus III.20 irudia), eta hau izango da atal honetan aurkeztuko duguna.



III.20 irudia. Analisi sintaktikoaren tratamendurako antolaketa sekuentziala.

Antolaketa honek zenbait abantaila izango ditu (ikus III.11 taula):

- Murritzapen-gramatika aplikatuko den lehen tresna izango da. Honen arrazoi nagusia anbiguotasun morfosintaktikoa ezabatzea da, horrela ondorengo tresnei lan handia kentzeko.
- Baterakuntza-gramatikaren bidez sortutako egitura sintaktikoak erabili ahal izango dira egoera finituko patroien bidez. Lehen esan dugunez, hizkuntza eranskarietan informazio aberatsa dago morfema- zein hitz-mailan, eta horregatik arazoa da hitz soiletan oinarrituta egitura sintaktiko osoei buruzko tratamenduak adieraztea. Adibidez, izen-sintagma bat aipatzeko orduan, zaila da bera osatzen duten hitzen segida posible guztiak zerrendatzea, murritzapen-gramatikan egin behar den moduan. Egitura sintaktikoak sartuz gero, formalismoak eskura izango ditu morfemak, hitzak eta egitura sintaktikoak, kasu bakoitzean egokiena aukeratzeko.

Osagai sintaktikoen maila gehitzean, abstrakzio-maila igotzen ari gara, tresnaren ahalmena handituz.

	MG + PATR + XFST
Analisia egiteko unitatea	morfema/hitza/ osagai sintaktikoa
Egitura sintaktiko konplexuak erabiltzeko ahalmena	+
Osagai sintaktikoen arteko komunztadura	+
Desanbiguazioa adierazteko ahalmena	+
Patroi sintaktikoak adierazteko ahalmena	+
Eraginkortasuna	+?
Esaldi analisi sintaktiko osoa	+

III.11 taula. Sintaxiaren tratamendurako gure hurbilpenaren ezaugarriak.

- Desanbiguazioa. Baterakuntzan oinarritutako analizatzailearen emaitza aurkeztu dugunean, aipatu dugu esaldi bakoitzeko hainbeste analisi posible aterako direla, eta honen ondorioa anbiguotasun morfosintaktiko zein sintaktikoa izango da. Anbiguotasun hori ebazteko tresna egokia izango da egoera finituko sintaxia, debekuak, iragazkiak edo murriztapenak adierazteko adierazpen erregularren bidez. Lehen aipatu ditugun zenbait sistematan, testuingururik gabeko gramatiken gainean eredu probabilistikoa gehitzen da, analisi anitzak ebazteko. Gure kasuan, analisi-aukerak baztertzeko murriztapenak erabiliko ditugu. Aukera honetan desanbiguazio morfologikoan gertatu den dikotomiaren paralelotasuna ikus dezakegu: desanbiguazio morfologikoan etiketatzaile estatistikoen eta linguistikoen (murriztapen-gramatika nagusi) banaketa egin da, azken hauek emaitza hobeak lortu dituztela; desanbiguazio sintaktikorako eredu probabilistikoak erabili dira gehienbat, baina gure kasuan ezagutza linguistikoan oinarritutako desanbiguazio sintaktikoa gehitzen ari gara.
- Aplikazioaren araberako moldagarritasuna. Aplikazioaren arabera, informazio mota desberdinak atera nahi izango dira testuetatik, ondoren ikusiko diren kasuetan bezala. Adibidez, aditzen azpikategorizazioari buruzko informazioa ateratzeko (ikus IV. kapitula), izen-sintagmak, adizlagunak eta mendeko esaldiak interesatuko zaizkigu, baina errore sintaktikoen detekziorako (ikus V. kapitula), esaldi baten egitura orokorra ateratzeko (ikus VI. kapitula) edo terminoen erauzketarako, aldiz, beste osagai batzuk atera nahiko dira. Adierazpen eta erlazio erregularren bidezko

patroiak definituz, helburu horiek lortu ahal izango dira, sistema osoaren malgutasuna eta modulartasuna indartuz. Horrela, baterakuntzan oinarritutako sintaxiak oinarritzko osagai sintaktikoak lortuko ditu, eta egoera finituko iragazkiek informazio horretatik aplikazio bakoitzeko interesgarria dena aukeratuko dute. III.20 irudian landu ditugun bi aplikazio nagusiak bakarrik agertzen diren arren, sistemari beste aplikazioak gehitzea badago, beharrezko diren egoera finituko patroiak definituz. Irudian aplikazio bakoitzeko egoera finituko gramatikak banatuta agertzen dira, baina esan behar da biek neurri handi batean erregela berak erabiltzen dituztela: gertaera linguistiko orokorrak deskribatzeko erregelak berrerabilgarriak izango dira, eta aplikazio bakoitzeko erregela partikularrak definitu beharko dira (errore bat detektatzekoak, adibidez)

Konbinazio sekuentzialaren abantailak lortu ahal izateko, formalismoen arteko sarrera/irteeren moldaketarako lana egin beharko da, bakoitzak analisirako datuen sarrera eta irteera modu ezberdinean dituelako definituta: MGk ezaugarri morfologikoen zerrenda, PATRk ezaugarri-egitura baten moduan eta XFSTk adierazpen erregular edo automata baten moduan. Hurrengo bi puntuetan hiru analizatzaileen arteko komunikazioa definituko da, murriztapen-gramatikaren irteeratik baterakuntzan oinarritutako analizatzailearen sarrera emateko (§ III.4.2.1.1), eta baterakuntza-analizatzailearen irteeratik XFST formatura pasatzeko (§ III.4.2.1.2). Bide batez, aplikazio sekuentzialaren beste abantaila batzuk azalduko dira.

III.4.2.1.1 Murriztapen-gramatikatik baterakuntzan oinarritutako analizatzaile sintaktikora

Murriztapen-gramatikak analisi morfosintaktikoaren ondorengo informazioa datutzat hartuta, desanbiguazio morfologikoa egikaritzen du, interpretazio-kopurua gutxituz eta kasurik hoberenean hitz-forma bakoitzeko interpretazio bana utziz. Bere irteera, baterakuntzan oinarritutako analizatzailearen sarrera izango da. Lotura hau egitearen ondorio aipagarrienak hauexek dira:

- Anbiguitasun-tasa jaisten den heinean, lana kenduko zaio baterakuntzan oinarritutako analizatzaileari, horrela analisi-denbora gutxituz. Honen estimazioak III.12 taulan egin dira. Taulak erakusten du desanbiguazioa eginez gero (MG + estatistikoa) baterakuntzan oinarritutako analizatzailearen abiadura hirukoiztu egiten dela.

- Irakurketak kentzearen beste ondorioa batzuetan ez asmatzea da, hau da, desanbiguazio okerrak egitea. Desanbiguazio morfosintaktikoan, lehen esan den bezala, asmatze-tasa %97,86 da, hau da, gutxi gorabehera hitzen %2an errorea gertatzen da. Corpus batean esaldien batez besteko hitz-kopurua 25 bada, horrek esan nahi du ia esaldi erdietan erroreren bat gertatuko dela. Edonola ere, erroreen larritasuna eta garrantzia neurtzeko aplikazio bakoitzarekin lotutako ebaluazio sakonagoa egin beharko litzateke. § IV.2.3n aditzen azpikategorizazioari buruzko informazioaren erauzketan aurkitutako akatsak aztertuko dira, eta ondoren desanbiguazioan hobetu beharreko aspektuak zehaztuko dira.
- MGren irteera hitz bakoitzaren interpretazioen etiketa morfosintaktikoen sekuentzia da (adibidez: IZE ABS S). Horiek baterakuntza-gramatikaren sarrera izateko ezaugarri-balio bikoteetara pasa beharko dira (KAT IZE, KAS ABS, NUM S). Prozesu honek ez du inolako zailtasunik eta horregatik ez diogu toki gehiago eskainiko.

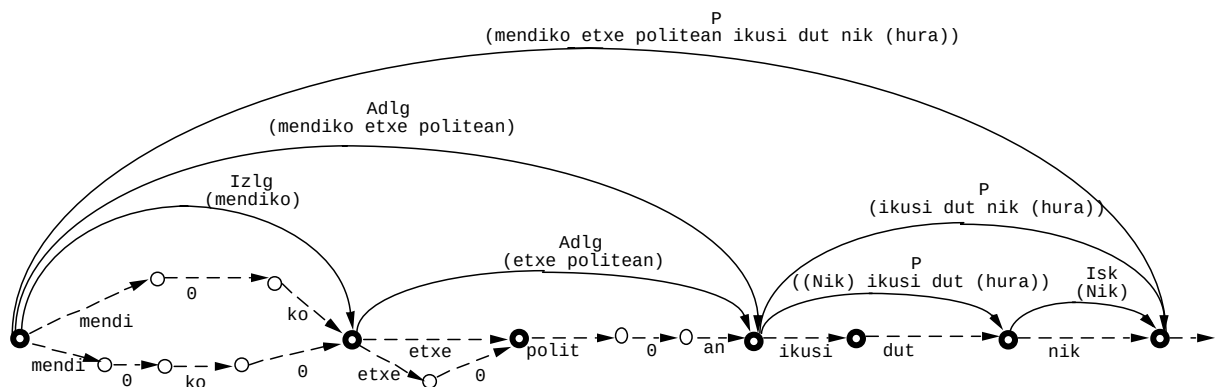
	Hitz-kopurua	Esaldi-kopurua	Denbora	Hitzak / segundoko
Murritzapen-gramatika aplikatu barik	22.272	989	4.391 s.	5,07
Murritzapen-gramatika aplikatuta	22.272	989	1.849 s.	12,05
Murritzapen-gramatika eta desanbiguazio estokastikoa aplikatuta	22.272	989	1.483 s.	15,01

III.12 taula. MG eta baterakuntzan oinarritutako analizatzaile sintaktikoen exekuzio-denbora.

III.4.2.1.2 Baterakuntzan oinarritutako analizatzaile sintaktikotik egoera finituko syntaxira

III.21 irudian agertzen da *chart*-a, baterakuntzan oinarritutako analizatzailearen emaitza. Lehenago esan bezala, hemen dagoen informazioa erabiliko da ondorengo aplikazioetan. Ikusten denez, maila linguistiko ezberdinak agertzen dira irudi horretan, datu-base lexikaleko osagaietatik hasita, tarteko osagai sintaktikoetatik pasatuz, maila altuagoko osagai sintaktikoetara iritsi arte (esaldia, izen-sintagmak, adizlagunak edo mendeko perpausak). Pentsa liteke, ezagumendu sintaktikoa erabiliko dituzten tresnek goi-mailako osagaiak beharko dituztenez, besteak bazter litezkeela, baina hori ez egiteko bi arrazoi aurkitu ditugu:

- Aplikazio bakoitzak osagai mota ezberdinak hartuko ditu. Horrela, errore sintaktikoak detektatzeko morfema bati buruzko informazioa interesgarria izan liteke, bestela errore horrek ez duelako onartzen zilegi den analisi sintaktiko bat.
- Nahiz eta analizatzailea sendoa izan, corpusen gainean erabiliz gero beti agertzen dira gramatikaren bidez tratatzen ez diren fenomenoak: akats ortografiko eta sintaktikoak, hitz ezezagunak, hitz anitzeko osagaiak edo gramatikan sartu gabeko osagai sintaktikoak. Gainera, hauen estaldura gero eta zabalagoa izango dela jakinda ere, beti egongo dira gramatikaz kanpoko elementuak. Beraz, analizatutako esaldietan ondo analizatutako osagaiak guztiz analizatu ez diren osagaien zatiekin tartekatuko dira. Horregatik *chart* osoa dagoen bezala mantentzea izan da hartu dugun erabakia, eta informazio aberats horretatik aplikazio bakoitzak berak behar duena aukeratu du, debekuak edo iragazkiak adieraziz.



III.21 irudia. Baterakuntzan oinarritutako analizatzaile sintaktikoaren irteeraren adibidea.

Osagai guztien informazioa mantentzearen erabakia hartu eta gero, beste arazo bat agertzen zaigu, baterakuntzan oinarritutako analizatzailearen irteera ezaugarri-egitura delako, eta egoera finituko analizatzailearen sarrera automata bat delako. Ezaugarri-egituren bidez informazio egituratua kodetzen da, eta gainera ez dago arazorik informazio berria gehitzeko baterakuntzaren bidez. Horrez gain, ezaugarrien arteko ordenak ez du garrantzirik. Automatetan, aldiz, ordena zehaztu behar da, eta informazio berria gehitzea ezinezkoa da, transduktoreen bidez ez bada, eta kasu honetan era nahiko mugatua. Arrazoi hauengatik, itzulpen-prozesua egin beharko da, ezaugarri-egituretatik (egituratuak) automata sekuentzial eta lauetara pasatzeko. Osagai sintaktikoen ezaugarri-egituretan, lehen azaldu dugunez, informazio aberatsa kodetu da, horrela aplikazio bakoitzean behar dena aukeratu ahal izateko. Momentuz orain arte aztertu ditugun aplikazioetan behar den informazioa bakarrik aterako dugu, jakinda beste informazio berria sartzeko ere arazorik gabe egin ahal izango dela.

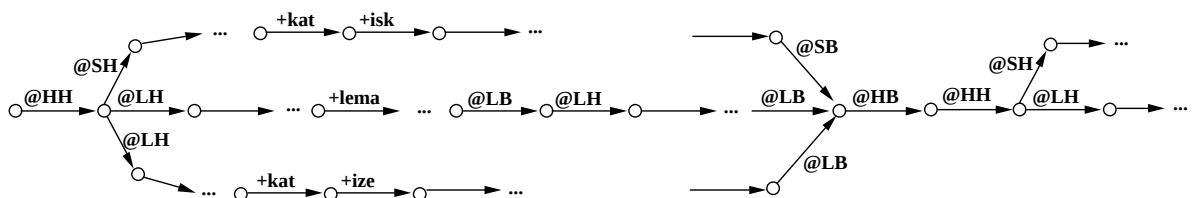


egoera finituko automata batera (behean).

Ikurrak	Esanahia
@HH @HB	hitzen arteko mugak (hitzaren hasiera eta bukaera)
@LH @LB	morfemen arteko mugak (morfemaren hasiera eta bukaera)
@SH @SB	osagai sintaktikoen arteko mugak (osagai sintaktikoaren hasiera eta bukaera)

III.13 taula. Osagai lexikal/sintaktikoak bereizteko ikurren esanahia.

Ezaugarri-egitura bakoitza nola kodetuko den erabaki ondoren, osagai sintaktiko eta lexisalak bereizteko beste ikur bereziak definitu ditugu, horrela lortzen den informazioa tratagarriagoa egiteko. III.13 taulan agertzen denez, hitzaren mugak, sintagmen mugak eta morfemen mugak bereiztu egin dira, une batean jakiteko zer moduko osagaia aztertzen ari garen. Era horretan, analisisen automatak III.23 irudian agertzen diren modukoak izango dira.



III.23 irudia. Egoera finituko automaten forma.

Automata hori zeharkatuz, esaldiaren irakurketa desberdinak aterako dira. Esaldi bakoitzeko irakurketa asko egongo denez, lehen lan bat desanbiguazioa eta behar den informazioaren aukeraketa izango da. Markaketarako ikurrak erabili direnez, irteera III.9 adibidean agertzen den bezalakoa izango da. Irakurtzeko zaila egiten denez, bere gainean iragazkiak aplikatuko dira informazioa formatu irakurgarriagoan emateko.

@HH@SH+atea+gure+etxe+polit+hari+sar+gureetxepolithurai+lema+etxe+kat+isk+azp+arr+kas+dat +mug+m+num+s+gunekat+ize@SB@HB@HH@LH+atea+.+sar+.+lema+.+kat+punt_punt +gunekat+punt_punt@LB@HB

III.9 adibidea. Egoera finituko automataren irakurketa bat ('gure etxe polit hari.').

III.4.2.2 Formalismoen konbinaziorako beste aukera batzuk

Gure aplikazioetan erabili dugun hiru formalismoen konbinazio sekuentziala azaldu ondoren, hemen labur azalduko ditugu beste proposamen batzuk, interesgarriak ikusten ditugulako edo beste sistema batzuetan erabili direlako:

- MG eta egoera finituko sintaxiaren konposizio sekuentziala. MGk desanbiguazioa baino harantzago joateko dauzkan arazoak ekiditeko, bere aplikazioaren ondoren egoera finituko sintaxia proposatu da zenbait lanetan (Koskenniemi *et al.* 1992, Järvinen eta Tapanainen 1998). Bide honetatik MGk desanbiguazioa eta oinarritzko unitate sintaktikoak ezagut ditzake, ondoren sintaxi osoari ekiteko.
- MG eta baterakuntzan oinarritutako sintaxia. MGren bidez egindako aplikazio bat izen-sintagma, adizlagun eta aditz-kateen markaketa izan da (Arriola 2000), baina osagai horien barruko elementuen konbinazioa ez dago deskribatuta (hau da, izen-sintagma batean determinatzaileak, adjektiboak eta hitz-elkarketak egon litezke, eta gunea zein den erabakitzea ez da sinplea izaten). Horregatik, ideia interesgarri bat MG sintagma horien mugak markatzeko erabiltzea da eta baterakuntza-gramatikak osagai horien analisi sintaktiko sakona egitea, ezaugarri-egitura batean sintagma horren elementu nagusien egitura hierarkikoa lortzeko. Modu horretan baterakuntzaren abiadura hobe liteke, bakarrik MGk aurreprozesatutako egiturak aztertuko direlako, eta optimizazioak aplikatu daitezkeelako (*top-down prediction* edo *determinization*).
- Aurreko azalpenetan tresnen konbinazio sekuentzialak aztertu ditugu, baina badaude atera berri diren ideiak analizatzaile sintaktiko ezberdinen irteerak konbinatzeko (Brill *et al.* 2000), analizatzaile-mota bakoitzaren ekarpena ahalik eta modu onenean erabiltzeko (adibidez, botoen teknikaren bidez). Metodo hauetan oraindik asko ikertu behar da, eta erabilgarriagoak izango dira ingelesa bezalako hizkuntzentzat, tresna ugari landuta daudelako, baina euskararen tratamendurako egin diren tresnekin probatzea ere interesgarria izango da.

III.4.3 Irteeraren TEI deskribapena

Aplikazio desberdinek tresna linguistikoen emaitzak erabiltzeko nahitaezkoa da emaitza horiek definizio formal bat izatea, horrela kodeketa-modu partikularren arazoa gaindituz. Ideia honekin, morfosintaxiaren kasuan egin dugun bezala, TEI gidalerroen arabera sintaxiaren emaitza adierazteko definizio formalizatuak landu ditugu (Artola *et al.* 2000). Era horretan, zenbait aplikazio desberdinek analizatzaile sintaktikoaren emaitzak jaso ahal izango dituzte, kasu bakoitzeko informazio-iragazleak erabiliz.

Gure sisteman konbinatutako hiru analizatzaile mota direla eta, aukera desberdinak daude emaitza ateratzeko. Alde batetik, saioak egin dira murriztapen-gramatikaren emaitzen gaineko unitate sintaktikoak ateratzeko hiztegi-definizioen analisisan (Arriola *et al.* 1999, Arriola

2000). Beste alde batetik, PATR analizatzailearen irteera atera liteke, baina esan dugu honek anbiguitasun handia izango duela, eta zailtasunak egon litezke informazio hori tratatzeko. Ideia interesgarriagoa iruditzen zaigu sistema konbinatuaren azken emaitzatzat zenbait informazio mota eskaintzea, aplikazioaren arabera gehiago zehaztu beharko direnak. Adibidez, izen-sintagma, adizlagunak, aditz-multzoak eta mendeko perpausen berri eman daiteke, XFST tresnaren bidez beharrezkoak diren iragazleak definitzen badira. Honetarako TEI estandarren araberrako definizioak prestatu ditugu, III.14 taulan agertzen diren informazio motak kodetzeko. III.10 adibidean esaldi baten osagaien adibide bat ikus daiteke.

Ezaugarriak	Deskribapena
kategoria, azpikategoria	kategoria sintaktikoak (izen-sintagma, adizlaguna, ...)
kasua, numeroa, mugatasuna	izen-sintagma eta adizlagunekin
oina	osagai sintaktikoaren oinarritzko elementu lexikala eta bere ezaugarriak
ergatibo, absolutibo, datibo	aditz batek azpikategorizatutako osagaiei buruzko informazioa
adizlagunak	aditz baten adjuntuen informazioa
adjektiboak	izen-sintagma edo adizlagunen izenondoak eta izenlagunak

III.14 taula. Analizatzaile sintaktikoaren emaitzen informazioa.

Bestalde, TEI estandarren definizioak tresna desberdinen arteko informazioa deskribatzeko erabiltzea oso interesgarria ikusten dugu, eta etorkizunerako pauso garrantzitsutzat jotzen dugu. Une honetan hasiak gara hauen definizio formalen prestaketan. Adibidez, baterakuntzan oinarritutako analizatzailearen TEIren bidezko deskribapen formala erabiliz, testuen gaineko unitate sintaktikoak aztertzeko tresna grafikoaren garapenean proiektu bat hasi dugu. Osagai sintaktikoak markatzeaz gain, tresnak osagai sintaktiko eta morfosintaktikoen barruko barne-egitura edo zuhaitza atera ahal izango du.

```
<text>
<body>
<p>
```

```
<fs type="osagai-sintaktikoa">
  <f name="ezaugarriak">
    <fs type="ezaugarri-lista">
      <f name="kat"><sym value="isk"></f>
      <f name="mug"><sym value="mg"></f>
      <f name="num"><sym value="s"></f>
      <f name="per"><sym value="1"></f>
      <f name="kas"><sym value="erg"></f>
      <f name="FSL" org="list">
        <sym value="@subj">
      </f>
    </fs>
  </f>
  <f name="oina">
    <fs type="lema">
      <f name="twol"><str>ni</str></f>
      <f name="sarrera">
        <fs type="gako">
          <f name="Sarrera"><str>ni</str></f>
        </fs>
      </f>
      <f name="ezaugarriak">
        <fs type="ezaugarri-lista">
          <f name="kat"><sym value="ior"></f>
          <f name="azp"><sym value="perarr"></f>
          <f name="per"><sym value="1"></f>
          <f name="plu"><sym value="minus"></f>
        </fs>
      </f>
    </fs>
  </f>
</fs>
```

**n
i
k**

```
<fs type="osagai-sintaktikoa">
  <f name="ezaugarriak">
    <fs type="ezaugarri-lista">
      <f name="kat"><sym value="isk"></f>
      <f name="mug"><sym value="m"></f>
      <f name="num"><sym value="s"></f>
      <f name="per"><sym value="3"></f>
      <f name="kas"><sym value="abs"></f>
      <f name="FSL" org="list">
        <sym value="@subj">
        <sym value="@obj">
      </f>
    </fs>
  </f>
  <f name="oina">
    <fs type="lema">
      <f name="twol"><str>etxe</str></f>
      <f name="sarrera">
        <fs type="gako">
          <f name="Sarrera"><str>etxe</str></f>
        </fs>
      </f>
      <f name="ezaugarriak">
        <fs type="ezaugarri-lista">
          <f name="kat"><sym value="ize"></f>
          <f name="azp"><sym value="arr"></f>
        </fs>
      </f>
    </fs>
  </f>
</fs>
```

**e
t
x
e
a**

```
...
</p>
</body>
</text>
```

III.10 adibidea. Esaldi baten osagai sintaktikoak TEI definizioak erabiliz (*nik* eta *etxea* izen-sintagmak agertzen dira).

III.5 Ondorioak

Tesi honen hirugarren kapituluan euskararen sintaxi konputazionalaren tratamenduan egindako lana aurkeztu dugu. Sintaxiaren mundua zabala da eta horregatik hasieratik lan honen mugak zehazten saiatu gara. Hauek izan dira gure hasierako helburuak:

- Sintaxirako tresna erabilgarriak lortzea. Alde teorikotik eta esaldi konplexuen egitura sakona lortu baino, gure asmoa testu errealeko esaldien tratamendua izan da. Horrek ez du esan nahi teoria linguistikoa albo batera utzi eta inplementazio-kontu hutsekin aritu garela, lan hau guztia hizkuntzalariekin batera egin dugu eta.
§ III.1.1.1en aipatu dugunez, bi dimentsiotan (lengoiaren teoria versus lengoia naturalaren prozesamendua, esaldien analisi osoa versus analisi partzialak) dauden elkarren kontrako muturren arteko aukera egin behar izaten da. Abaitua-k (1988) sakonean egindako LFG gramatikaren ezaugarri bat lexikoi txiki eta oso konplexua erabiltzea da, oraingo baliabide lexikaletatik urrun, aplikaziorako eremu oso mugatua emanez. Guk testuetako oinarritzko unitate sintaktikoen deskribapenari ekin diogu, jakinda hori gramatika oso baten lehen pausoa besterik ez dela, baina era berean aplikazioei bidea irekiz. Bestalde, egindako lan gramatikaren oinarri linguistikoa sendoa izan da, epe luzerako asmoa dugulako, eta horregatik aztertutako fenomenoaren tratamendua ahalik eta era sakon eta zorrotzenez deskribatzen saiatu garelako.
- Lehendik sortutako baliabide linguistikoen berrerabilpena. Aurrekoarekin lotuta dagoen arren, aspektu hau puntu berezi batean aipatzeak merezi duela uste dugu, taldean egindako lanaren adierazgarri delako. Horregatik esan behar dugu lan hau lengoia naturalaren prozesamendurako IXA taldearen barruan kokatu behar dela. Talde honek euskararen tratamendurako oinarri sendoak sortu ditu, horien artean lan honetan erabiliko diren Euskararen Datu-Base Lexikala eta euskararen bi mailatako morfologiaren bidezko segmentatzaile morfologikoa. Esan behar dugu ere momentuz ez daudela eskuragarri beste baliabide garrantzitsu batzuk, sintaktikoki etiketatutako corpusen modukoak. Arrazoi honengatik nagusiki ez diogu heldu analisi sintaktiko estatistikoari.

Horiek kontuan hartuta, ondokoak ditugu syntaxian egindako lanaren ekarpen aipagarrienak:

- Euskararen murriztapen-gramatikaren aplikazioa eta euskararen desanbiguazio morfosintaktikorako tresna. Tresna erabilgarria lortu da, une honetan zenbait aplikaziotan erabiltzen ari dena. Gainera, lan honek MG formalismoaren egokitasuna frogatu du euskararen moduko hizkuntza eranskarietan, lortutako

emaitzak ingeleserako ateratakoekin konparagarriak direla erakutsi delako. Egindako lana alde linguistikotik sakonean aztertzen da Arriola eta Aduriz-en tesietan (2000).

- Euskararen baterakuntzan oinarritutako gramatika partziala idatzi da, eta beretzako analizatzaile sintaktiko sendoa inplementatu da. Gramatika estaldura ertainekoa da, eta esaldien oinarritzko fenomenoak tratatzen ditu, horien artean izen-sintagmak, adizlagunak, esaldi sinpleak eta mendeko perpausak. Testu errealetako esaldi osoen analisiak askotan lortzen ez diren arren, ateratako osagai sintaktikoak baliagarriak dira aplikazio-multzo handi baterako. Indar handiena gramatikaren garapenean jarri den arren, analizatzailea gai da testu-sorta handiak modu eraginkorrean aztertzeko, nahiz eta oraindik abiadura azkartzeko bide zabala egon.
- Euskararen egoera finituko sintaxiaren bidezko tratamendua ere landu dugu, hurbilpen honen ekarpena frogatuz. Lortutako gramatikaren bidez anbiguotasuna landu daiteke (MGren zenbait arazo gaindituz), goi-mailako osagaien konbinazioa lortzeko aukerarekin batera.
- Tresna isolatuez aparte, tresnen konbinaziorako ereduak aztertu ditugu, formalismo bakoitzaren abantailak biltzeko asmotan. Hain zuzen ere, hiru tresnen aplikazio sekuentziala inplementatu eta probatu dugu. Erabilpen sekuentzialaren bidezko aplikazioak landu ditugu aditzen azpikategorizazio-informazioaren erauzketan eta errore sintaktikoen tratamenduan. Bietan emaitza onak lortu ditugu (ikus IV., V. eta VI. kapituluak).

Ekarpen horiek guztiak kontuan hartuz, esan behar dugu pauso hauek sintaxi osoaren tratamendurako lehenengoak izan direla, eta etorkizunean egiteko lan ugari aurreikusten dugula:

- *Bootstrapping* lexikala eta sintaktikoa. Oinarritzko baliabide sintaktikoak lortu ondoren aurrera eraman daitekeen pauso bat oinarritzko tresna horiek informazio lexikal eta sintaktikoen azterketarako erabiltzea da. Lan hori eskuz, automatikoki edo era mistoan egin daiteke. Adibidez, patroï sintaktikoen maiztasunak atera litezke, edo aditzen azpikategorizazio-patroiak eta hitzen arteko erlazio semantikoak (jakinda askotan patroï sintaktikoetan oinarritzen direla, aditza eta osagarrien artean bezala). Aberasketa hori, beraz, modu iteratiboan garatu daiteke, pauso bakoitzean aurrekoan sortutako lexikoi eta analizatzaileak erabiliz, gero eta doitasun eta fidagarritasun altuagoaz. Era horretan, sintaxiaren tratamenduan gailentzen ari den lexikalizaziorako joera integratu ahal izango da (Satta 2000). Honi buruzko lehen urratsak IV. kapituluan azalduko dira.

- Informazio lexikal osatuagoak lortzen diren heinean, orain arteko sintaxiaren zabalpena egin ahal izango da. Azpikategorizazioaren integrazioa funtsezko pausoa izango da, esaldi nagusi eta mendekoen arteko erlazioen deskribapenarekin batera. Bi bide nagusi ikusten ditugu honetarako:
 - Baterakuntzan oinarritutako teoria linguistikoen (HPSG edo LFG) euskararen gramatikaren diseinua eta inplementazioa tratatu ahal izango da. Hori interesgarria da, ez ikuspuntu linguistikotik bakarrik, azken aldiaren formalismo horien testu errealean tratamendurako arazoak gainditzen ari direlako, eta gero eta bideragarriago delako estaldura zabaleko gramatiken garapena (Kiefer *et al.* 1999, Butt *et al.* 1999, Kiefer eta Krieger 2000).
 - Teoria linguistiko formal horiek jarraitu gabe, gehiago testu errealean analisira jotzea. Hori egiteko eredu gehienak mendekotasun-egituren bidezkoak dira. Maiz aipatu da mendekotasun-egiturak egokiagoak direla osagaien ordena libreko lengoaien tratamendurako (Skut *et al.* 1997, Järvinen eta Tapanainen 1998, Oflazer *et al.* 1999ab, Basili *et al.* 2000). Abantailen artean azpikategorizazioaren tratamendu sinpleagoa eta sintaxia eta semantikaren egituren arteko integrazio errazagoa aipatu dira. Dena dela, aspektu honen gainean oraindik ikerketa sakona egin beharko da.
- Sintaktikoki etiketatutako corpusen edo *treebank*-en garapena oso interesgarria ikusten dugu, informazioa ateratzeko zein tresnen ebaluaziorako oso lagungarria izango baita. Ezin da ahaztu, dena dela, corpus horiek lortzeko beharrezkoa den lan handia, gurea bezalako talde txiki batean ondo neurtu behar dena. Gure kasuan, etiketatze hori hedatzen ari den mendekotasun-erlazioen bidetik ikusten dugu (Carroll *et al.* 1999).
- Albo-ondorioa kontsidera daitekeen arren, interesgarria izango da analisi sintaktikoak desanbiguazio morfologikoan duen eraginaren azterketa. Baterakuntzan oinarritutako analizatzaileak ez du sekula aukerarik baztertzen, baina egoera finituko gramatikan desanbiguaziorako murriztapenak eta debekuak adierazi dira, bukaeran esaldi bakoitzeko aukera bakarra uzten saiatuz. Frogatu gabe dagoen arren, gure hipotesia da osagai sintaktiko osoak erabili ahal izateak desanbiguazioaren tratamenduan ere onurak izango dituela.

BIGARREN PARTEA: APLIKAZIOAK

IV Azpikategorizazio-informazioaren erauzketa

Aurreko kapituluan, euskararen tratamendu sintaktikoari buruz hitz egin dugunean, aditzen azpikategorizazioari buruzko informazio-eza syntaxian eta semantikan lanean jarraitzeko lehen oztopoa dela arrazoitu dugu; bai azterketa teorikoak egiteko momentuan bai lengoia naturalaren aplikazioak ateratzeko orduan. Gainera, garatu ditugun tresnak direla eta, egoera egokian gaude gure baliabide linguistikoak modu automatikoan aberasteko, analizatzaile/desanbiguatzaile morfosintaktikoa eta azaleko analizatzaile sintaktikoak erabilgarri daudelako.

Modu honetan baliabide linguistikoak era iteratiboan garatu ahal izango dira, *bootstrapping* deitu ohi den teknika erabiliz: hasierako baliabideen bidez lor daitekeen informazioak baliabide berak hobetzeko balio dezake, modu iteratiboan aberasketa-prozesu bat definituz. Honen adibidetzat (Kuhn *et al.* 1998, Briscoe eta Carroll 1997) aipa daitezke, beste lan batzuen artean.

Gure kasuan, bi ikuspuntutatik dugu interesa tresna horien erabileran. Aurrenik, aditzaren azpikategorizazioaren azterketa teorikorako datu-base baten sorkuntzan (Aldezabal 2000), tresna informatiko eta linguistikoak erabil daitezke corpusetatik ezagutza ateratzeko, alde teorikotik lortutako informazioa egiaztatzeko edo aukera berriak proposatzeko. Bigarren, era horretan euskararen EDBL datu-base lexikalaren informazioa osatzen joan nahi dugu, aditzei syntaxia eta semantikarekin lotutako informazioa gehituz, informazio hori ezinbestekoa baita analizatzaile sintaktiko/semantikoen emaitzak hobetzeko. Kapitulua honetako edukia

(Aldezabal *et al.* 1998, Aldezabal *et al.* 1999c, Aldezabal *et al.* 2000) artikuluetan argitaratu da.

Ondorengo ataletan, beste hizkuntzetan aditzen azpikategorizazioari buruzko informazioa lortzeko egin diren sistema eta proiektu nagusiak aztertu (§ IV.1) eta gero, euskararekin egin dugun lanaren berri emango da (§ IV.2), hasieran problemaren definizioarekin (§ IV.2.1) eta ondoren inplementatutako tresna azaltzeko eta lehen emaitzak aurkezteko (§ IV.2.2, IV.2.3).

IV.1 Beste lan batzuen azterketa

Lexikoi konputazionalak osatzeko ahaleginak aspalditik egin dira. Hasiera batean informazio hori hiztegien bertsio elektronikoetatik (*Machine Readable Dictionaries*, MRD) edo hizkuntzalarien eskuzko lanari esker lortzen bazen ere, lanaren konplexutasun eta zabaltasunak kasu askotan ezagumendu hori era inplizitu edo esplizituan kodetuta duten testu-corpusak erabiltzea bultzatu du. Hauek dira gai honi buruzko proiektu interesgarrienetako batzuk:

- Eskuzko metodoak. *Complex Syntax* proiektuak (Grishman *et al.* 1994, Macleod *et al.* 1998) ingelesaren ezaugarri sintaktikoak kodetzeko estaldura ertaina/zabaleko lexikoi konputazionala (38.000 lema) lortzea du helburu. Kodetzeko informazioan azpikategorizazioarenari eman diote arreta berezia. Erabilitako metodo nagusia eskuzko lana da, hizkuntzalarietara egina, honen aldeko arrazoi nagusiak metodo automatikoen bereizketa finetarako murriztapenak eta maiztasun gutxiko osagaiak tratatzeko zailtasunak aipatzen direla. Eskuzko informazio hori hiztegi arruntekin zein corpusekin egiaztatu egiten da, honekin batera errore arruntenak identifikatzeko arau batzuk ezartzen direla. Erroreen artean adjuntu eta osagarrien nahasketak aipatzen dira (adjuntua den elementu bat osagarritzat, edo osagarria adjuntutzat eman), baita osagarri/adjuntuen ezaugarriak kentzea edo gehitzea ere.

Gross-ek (1997) frantseserako sistema baten deskribapena egiten du.

Lexicon-grammar izeneko egituran 12.000 aditz inguru daude sartuta, bakoitza bere argumentu-egiturekin. Egilearen ustez lengoaiaren deskribapena hitzen arteko mendekotasun lokalen bidez egin behar da, erregela sintaktiko orokorrekin baino, era honetan milaka erlazio lokal kodetuz. Lan hori eskuz eta corpusetan bilketa eta sailkapen sistematikoen bidez egitea proposatzen du, beste zientzia batzuetan (Biologia edo Geologia adibidez) egiten den bezala.

Gehienbat eskuz egindako bi proiektu hauen abantaila nagusia hizkuntzalarien ezagumendu guztia erabili ahal izatea da, baina arazorik handiena horrek eskatzen duen lan ikaragarria dugu, testuetan dagoen informazioa zabala eta ia bukaezina delako, eta kodeketa horren bideragarritasuna zalantzan jar daitekeelako. Horrez gain, corpus handiagoak erabiltzen diren heinean, handitu egiten da hizkuntzalariak errorea egiteko probabilitatea, edo antzeko fenomenoak modu desberdinean tratatzeko arriskua.

- Metodo automatikoak. Aspektu horiek kontsiderazioan hartuta, eskuzko lanean oinarritutako metodoen zenbait akats aipatu dira (Briscoe eta Carroll 1997, Carroll *et al.* 1998a). Lehenengo, zehaztasuna aipatzen dute, hiztegi zabal eta osoak lortzeko pertsonen erroreak (informazioa kentzea zein gehitzea) detektatzeko zailak baitira. Gainera, kostu handia du hedapenak egiteko orduan, bai neologismoak sartzeko bai azpilengoaia jakinekin lotutako informazio mota berriak gehitzeko. Honi generoen/azpilengoaiei arteko aldaketak azpikategorizazioan islatzen direla gehituz gero, eskuzko metodoen murriztapenak argi geratzen dira, azpilengoaia bakoitzeko

azpikategorizazio-aldaketak kodetzeko hizkuntzalariak erabiltzea ezinezkoa baita kasu gehienetan. Horrengatik, egile hauek informazio hori automatikoki ateratzeko metodoak aztertu dituzte. Beren sisteman sei osagai erabiltzen dituzte lan horretarako: etiketatzailea/desanbiguatzailea, lematizatzailea, analizatzaile sintaktiko probabilitistikoa, azpikategorizazio-patroien hipotesiak ateratzeko tresna, patroiak multzoetan sailkatzeko modulua, eta patroiak ebaluatzeko iragazlea. Asmoa aditzei 160 azpikategorizazio mota eta bakoitzaren maiztasuna esleitzea da, corpusean agertutako informazioaren arabera. Lortutako azpikategorizazio-ereduak analizatzaile sintaktikoan integratuz gero emaitzak hobetzen direla frogatzen dute. Hasierako helburua handinahikoa izanda (azpikategorizazio moten deskribapena fina da beste lanekin konparatuta) eta emaitza onak lortu dituztela kontuan hartuta, bide hau, informazioa automatikoki ateratzeko sistemena, etorkizun handikoa dela ikusten dugu.

Lan honen aitzindaria da Brent-ek (1994) egindakoa, bertan corpusetatik sintaxi lexikala ikasteko prozeduren azterketa egiten baita. Aurrekoarekin desberdintasun handiena ezagutza lexikal eta sintaktiko minimoa erabiltzeko apustua egitea da. Hori egiteko aditzak bukaeren arabera detektatzen ditu (lexikoirik gabe), eta balizko argumentuak ahalik eta anbiguotasun gutxienekoen artean aukeratzen ditu (izenordeak dira izen-sintagmen adierazle nagusiak (*I, me, he, him, ...*), eta mendeko perpausak *that the* moduko sekuentzien bidez detektatzen dira). Honek dakarren murrizketa corpus erraldoien adibide-kopuru handiaren bidez ekiditeko asmoa du

egileak. Beste alde batetik, esperimentua sei azpikategorizazio-eredutara mugatzen du, eredu gehiagotara orokortzeko metodoa argi ez dagoela, darabilen ezagutza sintaktiko ia 'hutsaren' ondorioz.

Carroll eta Rooth-en (1998) lanean antzeko asmoa dute, baina kasu honetan azpikategorizazio-ereduak ateratzeko eredu probabilistiko sofistikatua aplikatzen dute gramatika baten gainean, ikasketa-teknika automatikoen bidez. Gainera, patroiez aparte, aditza eta osagarrien arteko agerkidetzak neurriak ere ateratzen saiatzen dira, gramatika orokor baten gainean lexikalizazio-prozesua ahalbidetuz. Estatistikan oinarritutako beste metodo askorekin duen desberdintasun nagusia linguistikoki motibatua den gramatika baten gainean lan egitea da. Metodo honen beste abantaila bat etiketatu gabeko corpusen gainean lan egitea da. Nahiz eta emaitzak onak izan, oraindik metodoa garestia da, bai kostu konputazional handia duelako, bai corpus handien beharra duelako (50 M edo handiagoak).

- Metodo erdiautomatikoak. Sistema hauek aurreko bien tarteko bidea jorratzen dute. (Kuhn *et al.* 1998) artikuluan alemanerako estaldura zabaleko LFG gramatika baten erabilpena aztertzen da, corpusetik ezagutza lexikala ateratzeko. Helburu nagusia aditzen azpikategorizazio-patroien adibideak automatikoki ateratzea da, ondoren eskuzko tratamendua egiteko. Hori dela eta, alde automatikoaren doitasuna ahalik eta altuena izatearen garrantzia aipatzen dute, eskuzko lana gutxitzeko. Honetarako zalantzazko edo anbiguoak diren kasuak baztertu egiten dira, nahiz eta hori estalduraren aurka joan, corpus handien erabilerak arazo horren garrantzia minimizatuko omen duelako. Artikulu horretan corpusen galdeketa-sistema (hitzen analisisen gaineko adierazpen erregularren bidez, gramatkarik gabe) eta LFG gramatikaren arteko konparaketa azaltzen da, emaitza aipagarritzat gramatikaren nagusitasuna ateratzeko, doitasun zein estalduran. Ebaluazioa egiteko, aurretik aukeratutako hiru azpikategorizazio-ereduren adibideen bilaketan probatzen dute sistema. Gure kasuan, geroago azalduko dugunez, ez dugu mugatuko eredu-kopurua, aditz baten agerpen guztiak ateratzen saiatuko garelako.

Oesterle eta Maier-Meyer-ek (1998) lanak alemanezko izen- eta preposizio-sintagmak markatuta dituen corpus etiketatu (*treebank*) baten garapena deskribatzen du. Alemana hizkuntz malgukaria izanda, baterakuntzan oinarritutako formalismoa erabiltzen dute analisiak egiteko, komunztadura mota konplexuak egiaztatu behar baitira. Euskararen kasuan gertatzen den antzera, analisirako ez dago azpikategorizazioari buruzko informaziorik edo lehenago markatutako corpusik. Osagai bakoitzeko bere mugak (osagaiaren hasiera eta bukaera), kasua, pertsona, gunea eta zuhaitz sintaktikoa gordetzen dira. Lortutako emaitzaren aplikazioak

azpikategorizazioaren informazioa ateratzea edo informazioaren erauzketa izango dira.

Lapata-ren (1999) artikuluan egitura-alternantziak aztertzeko esperimentu baten emaitzak azaltzen dira. Bertan azaleko sintaxia erabiliz aditzen alternantzien adibideak ateratzen dira corpusetatik, bukaeran alternantzia bakoitzaren maiztasuna neurtzeko.

IV.2 Aditzen azpikategorizazio-informazioaren erauzketarako tresna

IV.2.1 Problemaren zehaztapena

Aurreko atalean ikusienez, sistema batzuetan hasierako baliabideak hutsetik hurbil dauden bitartean (Brent-en sistemak ez dauka ez gramatika ezta ia lexikoirik ere; Comlex proiektuan corpusetan bilaketa egiteko tresnak erabiltzen dira, syntaxirik gabe), besteetan tresna ahaltsuak erabili izan dira (Briscoe eta Carroll-ek ANLT gramatika eta hiztegi zabalak erabiltzen dituzte, aurretik Comlex eta ANLT sistemen azpikategorizazioari buruzko informazioarekin eta sintaktikoki etiketatutako corpusekin batera). Gure kasuan, tarteko bidean gaudela esan dezakegu, izan ere lexikoi aberatsetik abiatzen gara (EDBL), eta berarekin desanbiguaziorako tresna sendoa eta azaleko sintaxiaren analizatzaileak ere badauzkagu. Alde batetik, baterakuntzan oinarritutako analizatzaile sintaktiko partziala daukagu, eta bestetik bere gainean desanbiguaziorako eta patroiz linguistiko konplexuak idazteko XFST tresna. Ondorioz, oinarri sendoak dauzkagu lanari ekiteko, nahiz eta ingelesaren moduko hizkuntza baten baliabideetatik oraindik urrun egon, alde kuantitatibo zein kualitatiboan.

Ondoko puntuetan euskal aditzaren azpikategorizazio-erauzketari informatika-aplikazio gisa dagozkion sarrera eta irteera zehaztuko ditugu. Aplikazio hau burutzeko XFST tresna bakarrik landu behar dugu, beste tresnak lehenago eginda eta deskribatuta baitaude. Beraz, XFST tresnak lortu behar dituen emaitzak eta jasoko dituen datuak dira zehaztu behar direnak. § IV.2.1.1en azpikategorizazioaren azterketan kontuan hartu behar diren aspektu teorikoak komentatuko dira. Hurrengoan (§ IV.2.1.2) lortu nahi den emaitza zehaztuko da. Pauso hau garrantzi handikoa da, ebaluazio egokia egiteko aukera honen menpe egongo delako. § IV.2.1.3n baterakuntzan oinarritutako analizatzaile sintaktikoaren irteera deskribatu egingo da, hau izango baita egoera finituko adierazpenen bidezko patroiek hartuko duten

sarrera. Lehenago esan denez, irteera horretan analizatzaile morfologikoaren eta sintaktikoaren informazioa egongo da, bi motak izango direlako interesgarriak.

IV.2.1.1 Aditzen azpikategorizazioaren eremu teorikoa

Azpikategorizazioaren azterketa aditz bakoitzari esaldi zuzena lortzeko behar dituen ezaugarriak lotzean datza. Ezaugarri horien zehaztapenak zailtasun teoriko ugari sortu ditu teoria linguistikoa garatzen joan den heinean (Butt eta Geuder 1998), azpikategorizazioaren kontzeptuaren hausnarketa denborarekin aldatzen joan delako, eta hasiera bateko osagai lexikalen egituratze sintagmatikoetatik argumentuen erlazio eta rol tematikoetara pasa delako.

Euskara aztertzerakoan, osagai sintaktikoen ordena ez-finkoa eta kasuen aberastasuna kontuan hartuz, garrantzia argumentuen proiektzio diren kasuen azterketak dauka.

Egitura-alternantzieta, adibidez, kasuak dira, eta ez kategoria sintagmatikoak, aldatzen direnak. *ahaztu* aditzarekin, adibidez, bi eratarata ager daiteke IS-IS-Aditza egitura ('*gizonari galdera ahaztu zaio*' eta '*gizonak galdera ahaztu du*').

Bi ezaugarri horiei euskararen aditz laguntzailearen portaera gehitu behar zaie. Laguntzaileak aditzaren zenbait argumenturi buruzko informazioa darama, eta honen ondorio bat argumentu horiek (subjektua, objektua edo zeharkako objektua) aukerazkoak izatea da. Osagai hauek aukeran egoteak zailtasuna gehituko dio aditzen azterketari, batzuetan ezin izango delako bereiztu osagai bat ez dagoen ala modu inplizituan agertzen den. Aditzarekin komunztadura ez duten beste osagai batzuk, berriz, beharrezkotzat jo daitezke zenbait aditzentzat (kasu adlatiboa *joan* aditzarekin). Honela, beste hizkuntza batzuentzat aipatua izan den argumentua eta adjuntuen bereizketaren arazoa agertzen da.

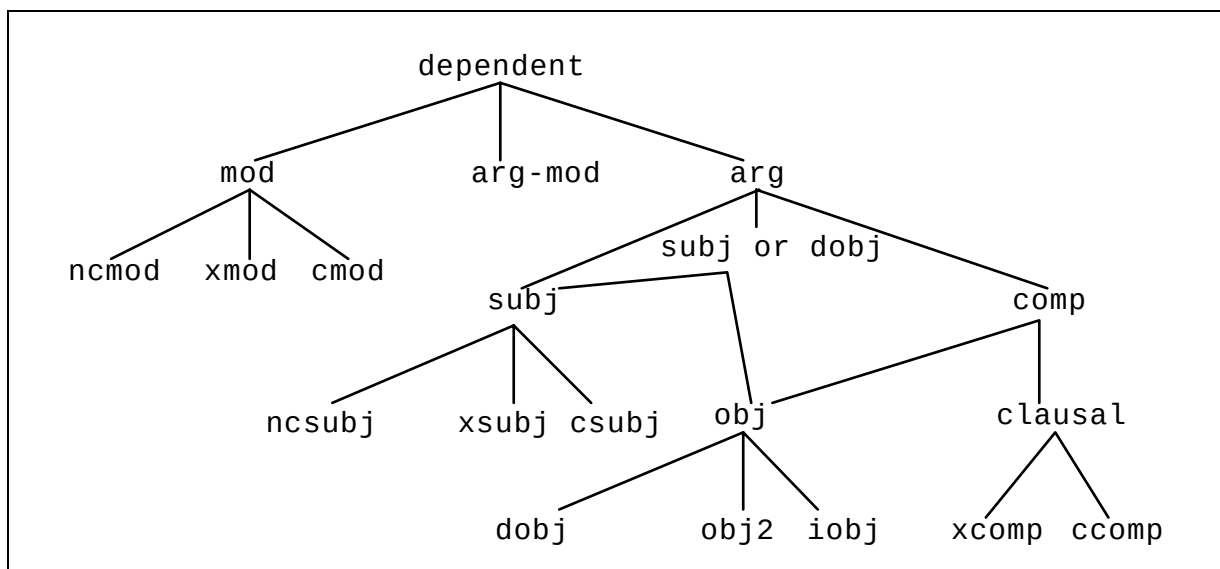
Lan honen beste aspektu azpimarragarria aditzak testuinguruan aztertzea da, helburua ez delako bakarrik aditz bakoitzaren forma kanonikoa (lema eta argumentuak) lortzea, baizik eta aditz bakoitza esaldietan agertzen den egitura sintaktikoak aztertzea, egitura bakoitzaren maiztasunarekin batera. Gainera, egile batzuen ustez (Levin 1993) aditz baten portaera sintaktikoa semantikak baldintzatuta dago eta, ondorioz, patroi sintaktiko berdinak jarraitzen dituzten aditzak multzo semantiko berean daude. Horrela, sintaxiaren azterketak semantikari buruzko azterketa egiteko bidea ireki dezake.

Hori guztiaren ondorioz, azpikategorizazioaren ikerketa honen helburua aditz bakoitzeko beharrezkoak diren osagaiak definitzea izango da, bere alternantzien kasu eta rol tematikoak

zehaztuz. Gainera, hurrengo pausoetan informazio hori aberasten saiatuko gara, kokakidetzak edo beste osagai lexikalen azterketa eginez.

IV.2.1.2 Lortu nahi den emaitzaren definizioa

Analizatzailearen emaitza definitzeko (Carroll *et al.* 1998b, 1999) lanetan oinarritu gara. Bertan analizatzaile sintaktikoen emaitzak eta sintaktikoki etiketatutako corpusetan erabilitako ebaluazio-eskemak aztertzen dituzte, bakoitzaren alde onak eta ahulak identifikatzeko, eta ondorioz erlazio gramatikalen hierarkia batean oinarritutako kodeketa-eskema proposatzen dute (ikus IV.1 irudia). Erlazio gramatikalak hierarkia batean kokatuz gero, erlazio batzuk beste batzuen espezializazioak izango dira, eta horrela aukera eman daiteke osagai sintaktiko baten erlazioa lortzeko orduan bera baino orokorragoa den erlazioa emateko, informazioa nahiko ez bada erlazio espezifikoen aukeratzeko.



IV.1 irudia. Ingeleserako definitutako erlazio gramatikalen hierarkia.

Paul intends to leave IBM

```

subj(intend, Paul)
xcomp(to, intend, leave)
subj(leave, Paul)
dojb(leave, IBM)
    
```

IV.1 adibidea. Ingeleseko esaldi baten erlazio gramatikalak.

Erlazio horiek erabiliz, IV.1 adibidean agertzen den moduan kodetuko dira esaldiak. Dena dela, euskararen tratamendurako arazoak daude aurreko erlazio gramatikalak zuzenean erabiltzeko. Alde batetik, linguistikoki erabaki egin beharko lirateke euskararen erlazio gramatikalak eta beraien arteko mendekotasunak. Adibidez, ingelesez subjektu eta

objektuaren bereizketa aditzarekiko posizioaren arabera markatuta dago, baina euskararen ordena librearen eraginez zailagoa izango da bereizketa hori egitea. Adibidez, maiz gertatzen da *-ak* deklinabide-atzizkiaren absolutibo plurala edo ergatibo singularraren arteko anbiguotasuna. Gainera, argumentu eta adjuntuen arteko banaketa sintaktikoki egin ote daitekeen argitu gabeko kontua da oraindik euskararako zein beste hizkuntza batzuetarako, poloniera kasu (Przepiórkowski 1999), nahiz eta IV.1 irudiko proposamenak modu argian definitzea badagoela ematen duen. Beste alde batetik, erabaki linguistiko horiek datu linguistikoen azterketaren ondoren har daitezkeela uste dugu, hau da, aditz bakoitza eta bere inguruko osagaien adibideak atera eta gero.

Horregatik, hasiera batean helburu apalagoa izango dugu eta, funtzio sintaktikoen bidezko erlazio gramatikalak bilatu ordeztuz, aditz batekin doazen osagarrien kasua, lema eta numeroa saiaturako gara ateratzen izen-sintagma eta adizlagun bakoitzeko, eta mendekotasun mota mendeko esaldiekin, IV.2 adibidean agertzen den moduan. Informazio gehiago lortzen den heinean, azpikategorizazioa lantzeko bigarren sistema batean eredua aberastu ahal izango da, horretarako hizkuntza eranskarietarako egindako proposamenak (Oflazer eta Okan 1996) kontuan hartuz.

Sarrera:	<i>Eta lepo honetatik Narbajara bide garbia doa, negurako pagaditik Larraingoitiko hareharrizko harrobitik zehar igaroaz</i>			
Irteera:	<i>abl(lepo, s)</i>	<i>ala(Narbaja, mg)</i>	<i>abs(bide, s)</i>	<i>doa</i>
	<i>lepo honetatik</i>	<i>Narbajara</i>	<i>bide garbia</i>	<i>doa</i>

IV.2 adibidea. Esaldi batetik ateratako informazioa (joan aditzaren azterketan).

Era horretako irteera bat zehaztuta, eraikitako sistema ebaluatzeko erabiliko ditugun neurri nagusiak aditz bakoitzarekin lortutako estaldura (*recall*) eta doitasuna (*precision*) izango dira, era honetan definituak:

$$\text{estaldura} = \frac{\text{lortutako osagai zuzenak}}{\text{esaldiko osagai zuzenak}} \quad \text{doitasuna} = \frac{\text{lortutako osagai zuzenak}}{\text{lortutako osagai guztiak}}$$

Estaldurak zenbat informazio lortzen den neurtzen du, eta doitasunak lortutako informazioaren kalitatea. Adibidez, aditz baten agerpen batean lau argumentutik tresnak hiru lortuko balitu, orduan estaldura %75ekoa (3/4) izango litzateke, eta doitasuna %100ekoa. Hiru osagai horietatik aparte beste bi osagai oker lortuko balira (aditzaren argumentuak ez direnak), estaldura %75 eta doitasuna %60 (3/5) izango lirateke. Bi neurri hauek elkarren kontrakoak izaten dira normalean, estaldura zabaltzen (osagai gehiago lortzen) saiatzeko den neurrian osagai oker gehiago lortzen direlako (doitasuna jaisten da), eta doitasuna gehitu nahi

denean (osagai zuzenak bakarrik lortu nahi dira) osagai gutxiago aukeratzen direlako (estalduraren jaitsiera). Geroago azalduko dugunez, bi balio hauek esaldi bakoitzeko erlazio sintaktikoen gainean kalkulatu dira, ondoren esaldi guztien batez bestekoa eginez.

Osagaia	Adibidea	Azalpena
<i>abl(lepo, s)</i>	<i>lepo honetatik</i>	adizlaguna ablatibo kasuan, singularra
<i>abs(bide, s)</i>	<i>bide garbia</i>	izen-sintagma absolutibo kasuan, singularra
<i>erg(iturri, p)</i>	<i>Gobernuko iturriek</i>	izen-sintagma ergatibo kasuan, plurala
<i>mendekoa(kaus, lako)</i>	<i>epaiketa egingo delako</i>	mendeko esaldia, kausala, <i>-lako</i> atzizkiaren bidezkoa
<i>mendekoa(agortu, rik)</i>	<i>agorturik (agortuta, agortzen, ...)</i>	mendeko esaldia, <i>-rik</i> atzizkiaren bidez
<i>adb(pozik)</i>	<i>pozik</i>	adberbioa

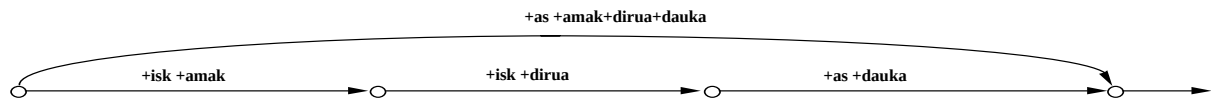
IV.1 taula. Euskararen tratamenduan aterako diren erlazio gramatikalen kodeketaren adibideak.

IV.1 taulan deskribatzen dira lehenengo hurbilpen honetan corpusetatik ateratzen saiatuko garen informazio motak. Gure sistemak lortuko dituen emaitzak aztertu ondoren, euskararen aditzen azpikategorizazioa lantzeko ondorengo pausoa funtzio gramatikalen zerrenda edo hierarkia zehaztea izango da. Gure iritziz, aurreikus dezakegu ingelesarentzat egindakoarekin ezberdintasunak izango dituela eta formalizazio linguistiko ez-sinpleak egin beharko direla, (Aldezabal 2000) tesian agertuko den modura.

IV.2.1.3 Baterakuntzan oinarritutako analizatzailearen irteera

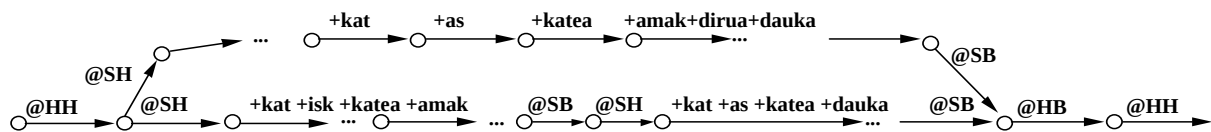
Emaitzaren definizioa zehaztu eta gero, modu laburrean aztertuko dugu XFST egoera finituko tresnaren sarreran dagoen informazioa, § III.3.2n azaldu dena. Baterakuntzan oinarritutako analizatzaileak izen-sintagmak, adizlagunak, esaldi sinpleak eta mendeko esaldiak lortzen ditu. Emaitza horiek ez dira zuzenean erabilgarriak oraingo aplikazio honetan, esaldi baten interpretazio ugari atera daitezkeelako, arazo hauengatik:

- Anbiguotasuna. Nahiz eta murriztapen-gramatikaren desanbiguazio-urratsa aplikatu, oraindik interpretazio bat baino gehiago geratzen da hitz bakoitzeko, eta gainera baterakuntzan oinarritutako gramatikak anbiguotasun sintaktikoak gehituko dizkio.



IV.2 irudia. Esaldi baten interpretazio-maila desberdinak (sinplifikatua).

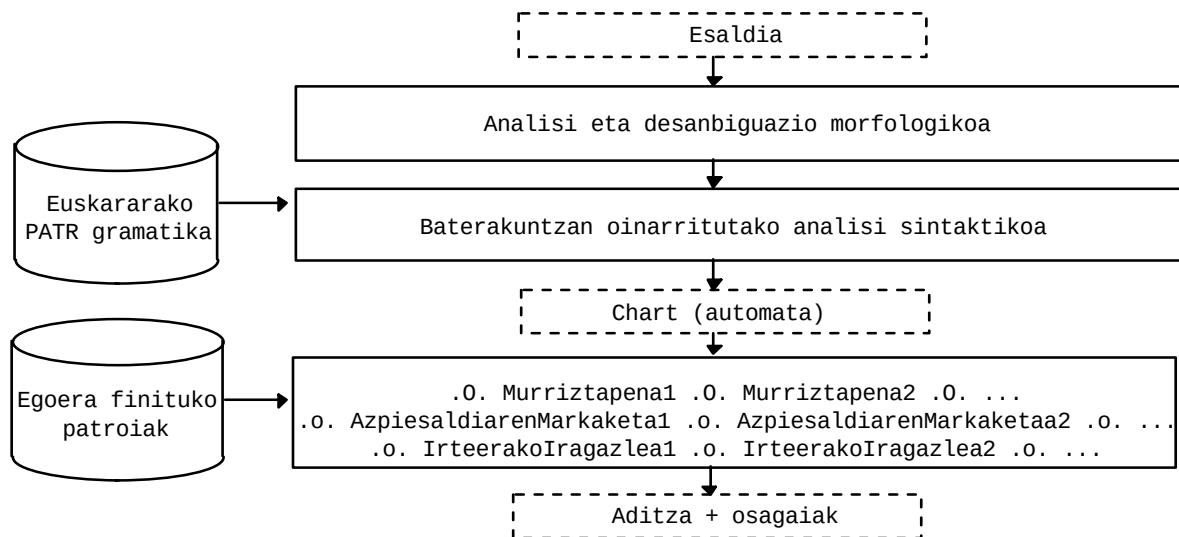
- Analisi-maila desberdinak. Maila guztietako osagaiak eta denak batera eskaintzen dira: hitz mailakoak, sintagma mailakoak eta perpaus mailakoak. Kontuan eduki behar da aplikazio batzuetarako izen-sintagmak lortu nahiko direla eta ez esaldi osoak. IV.2 irudian ikusten da nola '*amak dirua dauka*' esaldirako gutxienez bi analisi posible sortzen direla, lehenengoa (irudiaren goiko alde zeharkatuz lortzen dena) esaldi osoarena, eta beste bat bi izen-sintagma eta ondoren aditza hartuta (*amak*, *dirua* eta *dauka*). Nahiz eta kasu honetan benetako anbiguotasunik ez egon, osagai-maila desberdinak edukitzeak interpretazioen biderketa dakar berarekin, eta ondorioa da aplikazio bakoitzak behar dituen emaitzak iragazi beharko dituela. IV.2 irudiko automata era sinplifikatuan agertu da, benetako deskribapenean hitzak, morfemak eta osagai sintaktikoak bereizteko etiketak gehitzen direlako. IV.3 adibideak informazio osoa duen automata agertzen da.



IV.3 irudia. Esaldi baten interpretazio-maila desberdinak (informazio guztia).

- Gramatikaren estaldura mugatuaren eta beste fenomeno batzuen ondorioz (lehen esan bezala, hitz anitzeko unitateak, hitz ezezagunak, errore ortografiko eta sintaktikoak) ez dira esaldiaren osagai guztien analisiak lortuko, eta horregatik segmentatzaileak lortutako morfemak ere mantendu egingo dira irteeran.

Hiru faktore hauek bilduz gero, ondorioa da esaldi bakoitzeko milioika irakurketa ateratzen direla. Horietako asko kentzea erraza izango da, aplikazioaren araberrako informazioa iragazi beharko delako, baina hala ere lan zaila egin beharko da oraindik ebazteko konplexuak diren anbiguotasunak kentzeko. XFST tresna egokia da arazo horiei ekiteko.



IV.4 irudia. Aditzen azpikategorizazioari buruzko informazioa ateratzeko tresnaren diseinua.

IV.2.2 Tresnaren diseinua eta garapena

Aurreko puntuan aipatu diren arazoei aurre egiteko, mota ezberdinetako erregelak definitu behar izan dira. Ondoko multzoetan sailka ditzakegu erregelak (ikus IV.4 irudia):

- a) Desanbiguaziorako erregelak. Desanbiguazioa murriztapen-gramatikaren bidez egiten denaren antzekoa da, honako bi diferentzia nagusiak daudela (ikus § III.3.2):
 - Osagai sintaktiko osoak erabil daitezke, hitzaren muga estuetatik kanpo. Horrela, izen-sintagma edo mendeko esaldiei buruzko adierazpenak modu naturalean zuzenean erabil litezke, murriztapen-gramatikan zeharkako moduan egiten direnak.
 - Egoera finituko syntaxian analisirako unitatea aztertzen den esaldiaren interpretazio osoa denez, bazter daitezke zentzurik gabeko interpretazioak. Murriztapen-gramatikan, aldiz, anbiguotasuna hitz-mailan adierazten denez, askotan ezin dira baztertu esaldi-mailan ezinezkoak diren aukerak.

Desanbiguaziorako murriztapenak konposizio biguna edo malguaren eragilea (*lenient composition* .O.) erabiliz aplikatzen dira (ikus IV.4 irudia). Horrela ekidingo da murriztapen batek esaldi baten irakurketa guztiak ezabatzea, eta sistemaren sendotasuna lortzen da.

IV.3 adibideko erregelak absolutibo/ergatibo anbiguotasuna ebazteko egin dira. *IragazkiErg1* erregelak aditz jokatua hartzen du oinarritzat; *AditzaErgHaiek* motako osagaia (ergatiboan plurala duena, *dute* edo *dituzte* modukoa) aurkituz gero, orduan

kendu egiten du bere eskuineko izen-sintagma baten ergatibo singularraren interpretazioa (*‘ekarri dituzte gauzak’* moduko esaldian, *gauzak*-en ergatiboaren interpretazioa kenduko luke). Erregelak debekatu egiten du³¹ *“aditza-ergatibo-haiek edozer* izen-sintagma-ergatibo-singularra”* testuingurua. Ziurtasuna hobetzeko, erregela hori azpiesaldi baten barruan (aditz jokatu bakarrekoa) soilik aplikatuko da, beste esaldi bateko osagaia gaizki desanbiguatzeke arriskua ekidinez.

```
define AditzaErgHaiek [OsagaiLexSint & $["+nrk" "+hk"]];
define IragazkiErg1 ~$[AditzaErgHaiek
    ?*
    [OsagaiSintaktikoa & $[KAT "+isk"] &
    $["+kas" "+erg" "+mug" ? "+num" "+s"]
    ]
];
```

IV.3 adibidea. Anbiguitasuna ebazteko murriztapena.

```
define HartuLex0 ~$[ OsagaiLexikala ] ;
# osagai lexikalik ez duen katea
define HartuLex1 [$[ OsagaiLexikala ]]^1 & ~[ [$[ OsagaiLexikala ]]^2
];
# osagai lexikal bat du, baina ez bi

define HartuLex2 [$[ OsagaiLexikala ]]^2 & ~[ [$[ OsagaiLexikala ]]^3
];
# osagai lexikal 2 ditu, baina ez 3
...

define EsaldiaHartuLex0 [ Esaldia .0. HartuLex0];
# "lenient composition": bilatu lexikoko elementurik ez duena, eta
# ez badago elementu lexiko 1, 2, ... dutenak
define EsaldiaHartuLex01 [ EsaldiaHartuLex0 .0. HartuLex1];
define EsaldiaHartuLex02 [ EsaldiaHartuLex01 .0. HartuLex2 ];
define EsaldiaHartuLex03 [ EsaldiaHartuLex02 .0. HartuLex3 ];
...
```

IV.4 adibidea. Osagai sintaktiko luzeenak hartzeko murriztapenak.

- b) Osagai luzeenak hartzeko erregelak. Eraitza sintaktikoetan maiz gertatzen den beste anbiguitasun mota bat osagai sintaktikoak mota bereko osagaien barruan egotearena da. Adibidez, *‘teknika berriak’* izen-sintagma bakartzat har daiteke, edo bi daudela ere uler daiteke (*teknika* eta *berriak*). Berdin gertatuko da mendeko esaldiekin

(*‘gaur gizona etorri dela’* esaldian mendeko konpletibo ezberdinak atzematen ditu analizatzaileak: *‘etorri dela’*, *‘gizona etorri dela’*, *‘gaur gizona etorri dela’*). Hau ebazteko, gure heuristikoak osagai luzeenak aukeratuko ditu, kasu bakoitzaren salbuespenak kontuan hartuta. Heuristiko hori inplementatzeko IV.4 adibideko erregelak definitu ditugu. Bertan zero, bat, bi, ... osagai lexikal dituzten irakurketak hobesten dira, ordena horretan, horrela osagai lexikalen kopuru minimoa aukeratuko delako (hau da, osagai sintaktiko ahalik eta luzeenak hartuz; horregatik zero osagai

³¹ ‘~\$x’ adierazpenak, ‘~’ eta ‘\$’ bi eragileak konbinatuz, ‘x’ elementurik ez duen katea deskribatzen du.

lexikaleko kateek osagai sintaktikoak bakarrik dauzkate, osagai lexikal bateko kateek sintaktikoki ezagutu gabeko osagai bat dute, ...). Kasu honetan ere, *konposizio biguna* eragilea beharrezkoa da, bestela murriztapen horiek irakurketa guztiak kenduko lituzketelako.

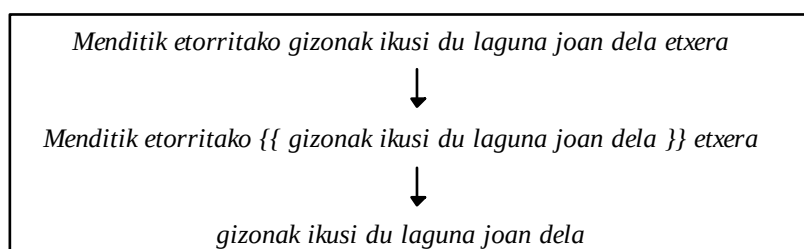
- c) Esaldiaren zatirik handienaren analisiak lortzeko erregelak. Antzeko problema agertzen da esaldi bateko osagai batzuentzat analisi sintaktikorik lortzen ez denean (gramatikaren hutsuneak, hitz ezezagunak, erroreak edo tratatu gabeko hitz anitzeko unitateak). Aurreko erregelak arazo hau saihesten ere laguntzen dute.

```
# Mendeko bat dagoenean eskuinaldean
# ikusi dut gizona etorri dela }} gaur (konpletiboa)
define MarkatuEskuinekoMendekoa31
    HelburuAditza HB HH]
    [[Isk & ~$["+gunekat" "+adi"]] HB HH]*
    [[OsagaiLexikala & $[KAT ["+lot" | "+adb"]]] HB
HH]*
    [OsagaiSintaktikoa & $"mendekoa"] HB HH
    -> ... "}}" ;

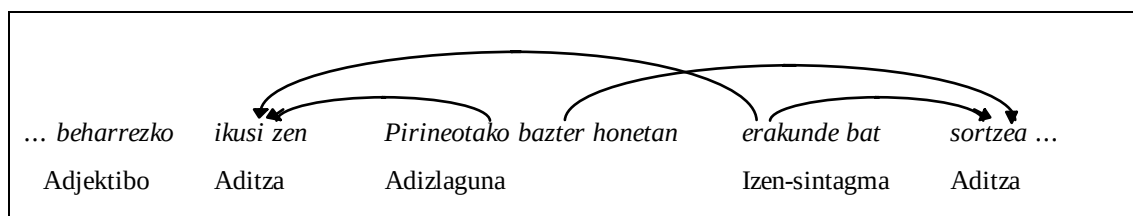
# Ezkerreko izenlagunak muga dira:
# etorri den {{ gizona nik ikusi dut
# etorritako {{ gizona nik ikusi dut
define MarkatuEzkerrekoErlatiboa [Isk HB HH]*
    [[OsagaiSintaktikoa|OsagaiLexikala] & $Aditza]
    @-> "{" "..."
    || [OsagaiSintaktikoa & $[KAT "+izlg"]
        & $["+gunekat" AdiAdt]
    ]
    HB HH
    _ ;
```

IV.5 adibidea. Azpiesaldiak murrizteko erregelak.

- d) Esaldi batean aditz bati dagokion azpiesaldia ateratzeko erregelak. Lehenago azaldu dugu aditz jakin bati dagozkion osagarriak lortu nahi direla. Anbiguotasuna asko jaisten da esaldi osoaren analisia hartu ordez aditzaren azpiesaldiarena bakarrik hartzen bada. Hau egin ahal izateko, erregela batzuek helburu den aditzaren testuingurua aztertuko dute, bere mugak markatuz. IV.5 adibidean ikusten dira lan hori egiten duten bi erregela. Lehenengoak (*MarkatuEskuinekoMendekoa31*) helburu den aditzaren eskuinaldean mendeko esaldi bat detektatzen duenean muga bat jartzen du (}}, mendeko horren ondorengo osagai sintaktikoak aditz horrekin ez doazela suposatuz. Helburu-aditza eta mendeko esaldiaren artean izen-sintagmak, adizlagunak, adberbioak edo loturazko osagaiak (lokailuak edo juntagailuak) onartuko dira, betiere aditz batez osatuak ez badira (adibidez, *etortzea*). Bigarren erregelak (*MarkatuEzkerrekoErlatiboa*) ezkerrean erlatiboazko perpaus bat (izenlaguna kategoriakoa) ikusten duenean muga bat jartzen du. Badakigu erregela hauengatik azpikategorizaziorako interesgarri diren osagai batzuk galduko ditugula (estalduraren galera), baina honela jokatuz anbiguotasuna asko jaisten da eta osagai okerrik hartzeko ziurtasun handia lortzen dugu (irabazia doitasunean).



IV.6 adibidea. Azpiesaldiak murrizteko erregelen aplikazioa.



IV.7 adibidea. Azpiesaldiak murrizteko zailtasuneko kasua.

- e) Iragazketa. Aplikazio bakoitzak bere informazio mota bereziak beharko ditu. Aditzen azpikategorizazioari buruzko informazioa lortzeko izen-sintagmak, adizlagunak eta mendeko esaldiak nahi dira, baina errore sintaktikoen detekziorako esaldi osoak tratatzea interesgarria izan liteke. Gainera, osagai jakin bati buruzko informazioa ere desberdina izan liteke aplikazioekiko. Adibidez, azpikategorizazioaren informazioa lortzeko kategoria sintaktikoa eta kasua dira informaziorik garrantzitsuenak, eta horregatik izena/adjektiboa moduko anbiguotasuna ekidin daiteke (*zuriekin* ize/adj soziation), kasu bietan izen-sintagma soziationa dugulako.

	Adierazpen/erlazio erregularren definizioa	Automata/transduktorearen tamaina
1	define IragazkiErgA IragazkiErg1 .o. IragazkiErg2 .o. IragazkiErg3 .o. IragazkiErg4 .o. IragazkiErg5 .o. IragazkiErg6; define IragazkiErgB IragazkiErg7 .o. ... IragazkiErg12;	14.697 egoera, 308.370 arku 1.585 egoera, 26.877 arku
2	define IragazkiErgA IragazkiErg1 .o. ... IragazkiErg7;	44.519 egoera, 978.586 arku
3	define IragazkiErgA IragazkiErg1 .o. ... IragazkiErg8;	104.769 egoera, 2.303.285 arku
4	define IragazkiErgA IragazkiErg1 .o. ... IragazkiErg9;	241.311 egoera, 5.546.341 arku
5	define IragazkiErgA IragazkiErg1 .o. ... IragazkiErg10;	395.105 egoera, 9.081.568 arku

IV.2 taula. Automata/transduktoreen tamainen igoera adierazpen/erlazioaren konplexutasunaren arabera.

Murritzapen eta iragazki guztiak adierazpen eta erlazio erregularren bidez definitu direnez, teorian posiblea izango litzateke lortzen diren automata eta transduktore guztiak egoera finituko sare bakar batean biltzea, horrek dakarren exekuzio-denboraren abiaduraren igoerarekin, baina hori ezin izan da egin, adierazpenak/erlazioak konplexuagoak diren neurrian sareen tamaina ere handitzen delako, eta memoriaren mugak gainditu egiten direlako. Adibidez, IV.2 taulan ergatiboaren desanbiguaziorako erregelak konpilatzeko zenbait aukera agertzen dira. Lehenago, ergatiboaren desanbiguaziorako hamabi murritzapen definitu dira (*IragazkiErg1*-etik *IragazkiErg12*-raino). 1 zenbakiko lerroan agertzen den aukeran murritzapenok bi erlazio erregularretan bildu dira (lehenengotik seigarreneneraino *IragazkiErgA* izeneko batean eta zazpigarrenetik hamabigarreneneraino *IragazkiErgB* izeneko bestean). Bigarren aukeran, murritzapen bat gehitzen zaio lehen *IragazkiErgA* erlazioari, eta ondorioz transduktorearen tamaina hirukoiztu egiten da. Hirugarren, laugarren eta bosgarren aukeran beste murritzapen bana gehitzen bada, orduan 400.000 inguru egoerako eta bederatzi milioi arku baino gehiagoko transduktorea sortzen da, sistemaren memoriaren mugetara hurbilduz. Honen ondorioz, kasu askotan egoera finituko sareak banatu egin behar izan dira, aplikazio sekuentziala eginez, honek lotuta dakarren abiadura-galerarekin, konpilazioa bideragarria izan dadin (Tapanainen 1997). Automata/transduktoreen arteko elkarrekintza oraindik ikertu behar den gaia da.

IV.2.2.1 Implementazioari buruzko datuak

Puntu honetan egoera finituko gramatikaren inguruko datuak zehaztuko ditugu, oraingo sistemaren azterketa orokor baterako interesgarriak direlako. Bi datu izango dira garrantzitsuenak: gramatikaren tamaina eta analizatzailearen abiadura. Lehenengoak honen moduko sistema baten garapenean egin beharreko lana neurtzeko balio dezake, eta bigarrenak lortutako tresnaren ahalmena datu-kopuru handiak tratatzeko.

Erregela-kopuruaren aldetik gramatikak 400 definizio erregular dauzka. Lehen ikusitako adibideetan hauetako batzuk agertu dira eta, ikusi denez, oso sinpleak direnetatik erlazio konplexuak adierazten dutenetara tarteko aukerak daude. Erregelen idazketaren aldetik esan behar da adierazpen erregularren lengoaia erazagutzailea izateak abantaila nabarmenak dauzkala, lengoaia hori tradizio handikoa eta oso erabilia baita bai hizkuntzalaritzan bai informatikan. Horrek argitasunean izango du eragina eta asko erraztuko ditu mantentzea eta dokumentazioa. Egoera finituko erregelak idaztearen konplexutasuna ez dator formalismoaren aldetik, baizik eta lanak behar duen ezagutza linguistikoaren partetik, hau da, kategoria lexikal zein sintaktikoen ezagutza, eta sintaxi orokorra edo corpusetako esaldien egituraren ezagutzatik. Esan behar da ere erregela hauek, funtsean aditzen azpikategorizazioaren azterketarako egin diren arren, erabilgarriak izango direla neurri handi batean beste aplikazioetarako, erregelek oinarritzko gertaera linguistikoen deskribapen orokorrak egiten dituzten heinean.

Denboraren aldetik, une honetan batez beste lau hitz prozesatzen dira segundoko. Abiadura handia ez izan arren, dauzkagun corpusak tratatzeko adinekoa da. Lehenago atera diren MG eta baterakuntzan oinarritutako analizatzailearen denbora gehituz gero (15 hitz segundoko, § III.4.2.1.1), ikusten dugu egoera finituko sintaxia dela urratsik motelena. Abiadura hori hobetzeko pausoak eman beharko dira, jakinda egoera finituko sare konplexuekin dauden arazoak gainditzea oraindik ikerketa-fasean dagoela. Lortutako automaten tamainak eragina du sistemaren abiaduran, era sekuentzialean aplikatzen den automata-kopuruarekin batera. Bestalde, definizio erregularrak egiteko moduak ere azken automata edo transduktorearen tamaina baldintzatzen du. Faktore hauek guztiak kontuan hartuta, sistemaren hobekuntzarako pausoak definitu beharko ditugu.

IV.2.3 Lehen emaitzak

Aurreko ataletan aurkeztutako egoera finituko gramatikaren garapenaren ondorioz, tresna guztiz inplementatua eta erabilgarria dugu. Puntu honetan tresnaren emaitzen ebaluazioa egingo dugu. Ebaluazioan lortutako datuek tresnaren ahalmena definitzeko eta hobetu beharreko aspektuak argitzeko balioko dute.

Analizatzailea testu errealei aplikatzeko nahi dugunez, lehen lana corpusak lortzea izan da, eta horretarako bi iturri nagusi aurkitu ditugu. *Egungo Euskararen Bilketa Sistematikoa* (EEBS; Urkia eta Sagarna 1991) alde batetik eta *Euskaldunon Egunkaria*-tik lortutako testuak bestetik, izan dira probarako esaldiak aukeratzeko erabili ditugun corpusak.

Aipatu dugunez, analizatzaile sintaktikoen ebaluaziorako gehien erabilitako neurriak doitasuna eta estaldura dira. Hauek neurtzeko, bost aditzen (*agertu*, *atera*, *erabili*, *ikusi* eta *joan*) 100 esaldiko multzoak aukeratu ditugu, hau da, 100 esaldi aditz bakoitzeko, horietatik erdiak EEBS-tik eta besteak egunkarietatik, guztira 500 esaldirekin. Aditzak aukeratzeko bi arrazoi nagusi izan dira kontuan. Lehenengo beren maiztasuna, hasierako probetarako maiztasun handiko aditzak hartu baititugu, datuak errazago jasotzearen. Bigarren, aditz horiek mota ezberdinetakoak dira (iragankorrak, iragangaitzak, ...), eta horrek azpikategorizatutako osagarri moten azterketa zabala egiteko aukera emango du. Esaldiak lortzeko ez da beste moduko iragazkirik pasa, hau da, lortu diren lehen esaldiak izan dira, eta ez dira hartu gramatikaren edo lexikoaren arabera, horrela sistemaren portaera benetako testuekin zein den ateratzeko.

Esaldi bakoitzeko, eskuz markatu ziren aditzaren agerpena eta berarekin lotutako elementu azpikategorizatuak, ondoren analizatzailearen irteerarekin konparatzeko. 350 esaldi gramatikak (baterakuntzan oinarritutakoa zein egoera finitukoa) garatzeko erabili ziren, eta beste 150 esaldiak (30 aditz bakoitzeko) azken probarako utzi ziren.

Anbiguotasunaren eta gramatikaren estalduraren hutsuneen problemei beste batzuk gehitu behar zaizkio:

- Esaldien luzera. Esaldi bakoitzak helburu-aditzaren agerpen bat du, beste esaldi batzuekin batera (esaldi nagusi edo mendekoak). Esaldien batez besteko luzera 22 hitzekoa da. Tratatu nahi den aditzaren azpiesaldiaren mugak zehazki bereiztea lan zaila izango da.

- Hitz anitzeko unitate lexikalak. Analisi morfologikoan sartzeko planak ditugun arren, oraindik hauen tratamendua inplementatu gabe dago, eta ondorio nagusia erroreen gorakada (alarma faltsuak) izango da, egitura berezi batean antolatzen diren hitzak baterakuntzan oinarritutako gramatikaren erregelen arabera okerreko moduan interpretatuko direlako.
- Hitz ezezagunak, izen bereziak eta errore ortografikoak. Analizatzaile morfologikoak batzuk ezagutzen ditu, baina besteek arazoak sortzen dituzte, normalean hipotesi asko sortzen dituztelako eta, ondorioz, errore-tasak gorantz egiten duelako. Lexikoaren estaldura handitzen doan neurrian ondorengo faseetako emaitzetan eragina izango du.

	Garapenerako corpusa (350 esaldi)		Probarako corpusa (150 esaldi)	
	Doitasuna	Estaldura	Doitasuna	Estaldura
agertu	95%	69%	87%	62%
atera	91%	65%	92%	64%
erabili	92%	70%	86%	55%
ikusi	91%	76%	87%	78%
joan	93%	74%	83%	70%
Guztira	92%	71%	87%	66%

IV.3 taula. Ebaluazioaren emaitzak.

IV.3 taulak emaitzak aurkezten ditu, esaldi guztien batez bestekoak hartuta. Nahiz eta beti oreka bat egon doitasuna eta estalduraren artean, lehenengoa maximizatzen saiatu gara, kasu batzuetan estaldura jaitziz. Analizatzailearen garapenean 350 esaldien analisiak aztertzea bazegoenez, bigarren eta hirugarren zutabeetako zenbakiak analizatzailearen oraingo egoeraren maximotzat uler daitezke, %92ko doitasuna eta %71ko estaldurarekin. Aditz baten agerpenaren interpretazio bat baino gehiago zegoenean esaldi hori baztertu egin dugu, (estaldura jaisten da horrelakoekin). Geroago azalduko dugunez, emaitza hauek hobe daitezke lexikoa eta gramatika sintaktikoak gehiago landuz gero. Hori guztia kontuan hartuta, emaitzak ontzat ematen ditugu, estaldura %66 eta doitasuna %87ra iristen direlako 150 esaldien multzoarekin, espero zen maximotik hurbil, eta emaitza horiek ikusi gabeko esaldiekin sistemaren portaera ona dela erakusten dutelako.

Erroreen jatorria eskuz aztertu genuen (IV.4 taula), zenbait errore motaren artean multzokatuz (estaldura eta doitasunean arazoak sortzen zituzten 68 errore identifikatu genituen guztira):

- Hitz anitzeko unitateak (5), hitz ezezagunak, izen bereziak (9) eta errore ortografikoetatik sortutako erroreak. Beren tratamendua analisi morfologikoaren moduluari dagokio gehienbat, eta aberasketa lexikalarekin lotuta dago.
- Desanbiguazio okerrarengatik sortutako erroreak. Hauek bi mota nagusitan bana daitezke. Desanbiguatzaileak irakurketa okerra aukeratzen duenean alarma faltsua sortzen da, doitasuna txikituz (9 errore). Beste alde batetik, kasu batzuetan aukera bat baino gehiago geratzen da, horien artean zuzena dagoela. Bere efektu nagusia estalduraren jaitsiera izango da (5 errore).
- Baterakuntzan oinarritutako gramatikaren hutsuneekin lotutako erroreak (32 errore). Errore mota hauek ematen dituzte analisi partzialaren mugak. Errore hauetatik erdia baino gehiago gramatikaren atal batzuetan landu gabeko aspektuak dira, eta arazo handirik gabe konpon litezke gramatika zabalduz gero, baina beste batzuek aldaketa kualitatiboak beharko lituzkete, aditzaren azpikategorizazioari buruzko informazioa txertatzea kasu. Azkenik, hirugarren multzo bateko erroreak corpusetako egitura sintaktiko berezi eta arraroengatik sortu dira. Gure ustez edozein analizatzailearentzat, aurreko bi arazoak ebatzi arren, zailtasun handikoak izango dira azken multzo honetako egiturak ondo analizatzea.

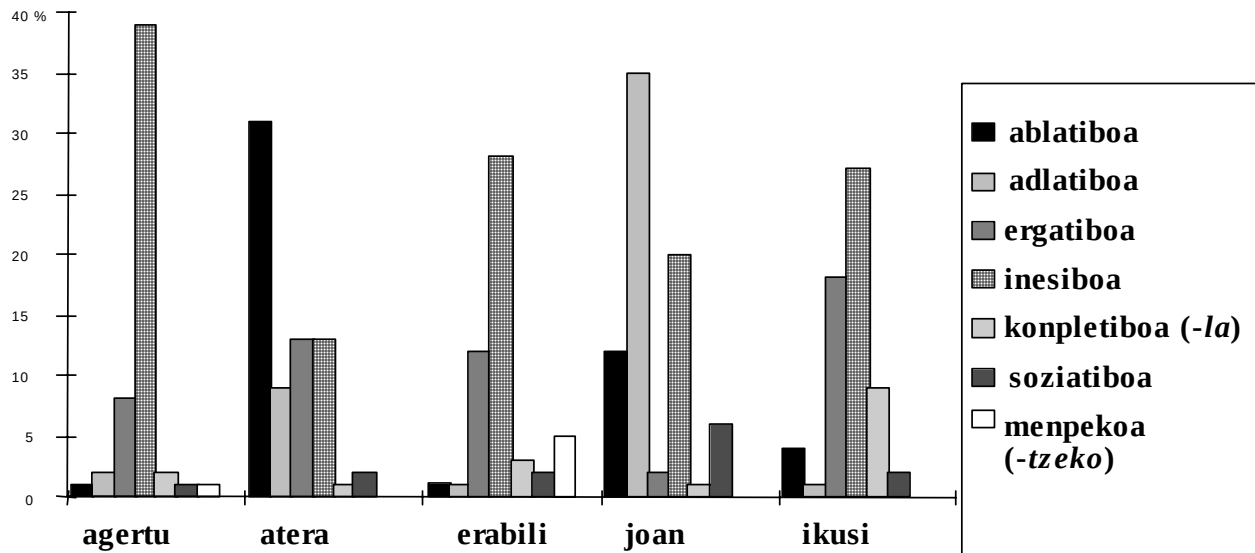
Errorea	Adibidea
Hitz anitzeko unitatea ezagutu ez denez, <i>aldez-instrumental</i> eta <i>aurretik-ablatibo</i> lortzen dira	ekimen batek izan ditzakeen emaitzak <i>aldez aurretik</i> ikustea
Hitz ezezaguna agertzen denez, ez da asmatzen ergatiboan doala	<i>PSOEk</i> begi onez ikusten du alderdi guztiak bilduko dituen gunea
Errore ortografikoaren ondorioz, ez da sintagma ezagutzen	berri <i>ori</i> bakarrik darabil
<i>duela</i> gaizki desanbiguatua, aditz trinkoaren interpretazioarekin	hori gertatzea ez <i>duela</i> posible ikusten
<i>aterako</i> gaizki desanbiguatua, izena	ez zen <i>aterako</i>
<i>ziren</i> aditzean erlatiboa/iragana anbiguotasuna ebatzi gabe dago	AABAk kezkatuta agertu <i>ziren</i> atzo
Sintaxiaren akatsa. ‘ <i>gauza bat bezala</i> ’ moduko egiturak ez daude sartuta	hau ez da betiko <i>gauza bat bezala</i> ikusten
Sintaxiaren akatsa. <i>erabiltzen</i> eta <i>erietxeetan</i> ez dira erlazonatzen, bien artean <i>hasi</i> dagoelako. Azpikategorizazioaren informazioa beharko da.	<i>erietxeetan</i> hasi zen erabiltzen

IV.4 taula. Azpikategorizazioaren erauzketan egindako errorearen adibideak.

Emaitzek erakusten dute errorearen erdia baino gehiago lexikoa, analizatzaile morfologikoa eta desanbiguzio morfologikoaren (36 errore guztira) hobekuntzekin konpon daitezkeela. Nahiz eta desanbiguzio morfologikoa zabaltzea eta aldatzea lan ez-sinplea izan, izen berezien, errore ortografikoen eta hitz anitzeko unitateen tratamenduan lan egiteak estaldura eta doitasunaren igoera nabarmena eragingo luke. Era berean, baterakuntzan oinarritutako gramatikaren zabalpena eginez gero, errore sintaktikoen erdiak ekidin daitezkeela aurreikusten dugu.

Emaitza horiek aztertu ondoren, ikusten dugu baliagarriak direla modu automatiko edo erdiautomatikoan informazio interesgarria ateratzeko. IV.5 irudian goian aipatutako bost aditzentzako emaitzen berri ematen da, aditz bakoitzeko berarekin agertzen diren osagarrien maiztasunak neurtu ondoren. Aztertutako osagarri-motak izen-sintagmak, adizlagunak eta mendeko esaldiak izan dira. Izen-sintagmen kasuan, kasu absolutiboa da maiztasun handiena duena diferentzia handiagatik. Arrazoi horrengatik, irudian ez da azaltzen beste osagarri-moten proportzioa hobeto ikusi ahal izateko. Inesiboa da aditz guztietan agertzen den kasu

bat, denbora-espazioari buruzko kokapena adierazteko erabileraren ondorioz. Ablatiboa eta adlatiboa *atera* eta *joan* aditzekin agertzen dira gehienbat. Emaizetan modu nabarian adierazten da mendekoen erabilera, helburuzkoa (*-tzeko*) *erabili*-rekin eta konpletiboa *ikusi*-ren kasuan.



IV.5 irudia. Bost aditzentzat lortutako osagaien maiztasunak.

IV.3 Ondorengo urratsak

Kapitulu honetan garatutako analizatzaile sintaktikoaren aplikazioa aurkeztu da, aditzen azpikategorizazioari buruzko informazioaren erauzketan. Lehenengo emaitzak aztertu ondoren, esan daiteke tresnak emaitza erabilgarriak ematen dituela, estaldura eta doitasuna kontuan hartuz. Egiaztatu da tresnak aditz desberdinen portaerak bereizteko gaitasuna ere baduela.

Bestalde, esan behar dugu oraindik aditzen azpikategorizazioaren tratamenduaren hasieran besterik ez gaudela, etorkizunean lantzeko zenbait pauso ikusten ditugulako:

- Tratamendu morfologikoaren hobekuntzak emaitzetan eragina izango du, hitz ezezagunen, izen berezien eta hitz anitzeko unitateen tratamenduan.
- Berdin gertatuko da oinarritzko baliabide sintaktikoekin. Alde batetik, murriztapen-gramatikaren desanbiguazioaren akatsak zuzentzea eta geratzen diren anbiguotasunen ezabaketarekin, eta bestetik baterakuntzan oinarritutako gramatikaren hedapenekin, batez ere esaldi-mailan, esaldiaren osagai nagusien arteko erlazioak aztertzea analizatzaile osoaren hobekuntzarako izango bailiteke.

- Orain arte bost aditz landu dira bereziki baina, tresnaren sendotasuna frogatuta dagoenez, hurrengo pausoa aditz-multzoa zabaltzea izango da, azken helburua datu-base lexikaleko aditz guztien azterketa egin arte. Une honetan 400 aditzen azterketa egiten hasi gara.
- Lehenago aipatu dugunez, aditzen agerpen-adibideak lortuta, oraindik azpikategorizazio-patroien definizioa geratzen da, hau da, agerpenak patroietan multzokatzea. Hemen, beste gauza batzuen artean, argumentua eta adjuntuak bereiztu beharko dira, edo patroï bakar baten egitura-alternantziak sailkatu. Aztertu beharko da prozesu hau zein puntutaraino egin daitekeen eskuz edo automatikoki. Aditz-multzoa zabaltzen doan heinean, eskuzko metodoen aplikazioa ezinezkoa bihurtuko da, eta eredu probabilistikoen erabilera (Carroll eta Rooth 1998) landu beharko da.
- Ikertu behar den lerro interesgarri bat aditzen agerpen-adibideak hartuta aditzak automatikoki sailkatzearena da. Hasierako hipotesia antzeko patroïak dituzten aditzak multzo semantiko berekoak direla izango da, eta hori corpusetako datuekin egiaztatu beharko da (Aldezabal 2000).
- Informazio hori guztia lortu eta gero, azpikategorizazioaren informazioa analizatzaile sintaktikoan integratzeko eredua aukeratu beharko da. Honetarako zenbait aukera daude (Basili *et al.* 1998, Voutilainen 1997), eta bakoitzaren ezaugarriak eztabaidatu egin beharko dira. Adibidez, azpikategorizazioa baterakuntzan oinarritutako analizatzailean sar daiteke, seguruenik PATR baino harantzago doazen LFG edo HPSG moduko formalismo batera egokituz. Beste aukera bat egoera finituko gramatikaren barruan kokatzea izango litzateke. Esan behar dugu pauso hau hartzeko azterketa sakona egin beharko dela.
- Informazio lexikala gehitzen denean, erlazio gramatikalen definizio aberatsagoak lortzea pauso interesgarria izango da, (Carroll et al. 1999) lanean agertzen den estiloan, orain erabiltzen ari garen kasu gramatikaletatik funtzio sintaktikoetara pasatzeko.

V Ezagumendu sintaktikoaren erabilera errorearen detekzioan eta zuzenketan

V.1 Sarrera

Lengoaia naturalaren prozesamendurako sistemek hasieratik tratatu behar izan duten arazo bat erroreena izan da, testu idatzietan zein ahotsaren tratamenduan mota askotako erroreak aurki baitaitezke. Testu idatzietan, nahiz eta konturik handiena jarri, beti gertatzen dira mota desberdinetakoak diren eta kausa desberdinak dituzten erroreak. Honen ondorioa testuak orrazten denbora handia pasatzea da eta, gainera, azkenean testua argitaratzen den momentua iritsi bezain pronto errorearen bilketaren fasea hasi behar izatea. Hau egia bada euskara neurri batean menperatzen dugunontzat, zer esanik ez euskara menperatzen ez duten edo ikasten ari direnontzat. Euskararen kasuan pentsa daiteke errorearen agerpena beste hizkuntza batzuetan baino maiztasun altuagoarekin gertatuko dela, bere egoera bereziagatik (hau da, hizkuntzaren normalizazio-prozesua, hiztun askoren ezagutza partziala edo inguruko erdaren eraginaren ondorioz). Honengatik, hain zuzen ere, euskararen morfologiaren tratamendu automatikoaren lehen aplikazioa zuzentzaile ortografikoa izan da (Aduriz et al. 1997, Aldezabal et al. 1999a). Errorearen tratamenduan informazio morfologiko hutsetik informazio sintaktikora pasatzeak asko zabaltzen du aplikaziorako eremua eta, horregatik, kapitulu honetan testu idatzietako errorearen tratamenduan ezagumendu sintaktikoaren ekarpena aztertzeke lehen saioak aurkeztuko ditugu. Asmo nagusia garatu diren tresnak erabiliz egin daitezkeen tratamenduen bideragarritasuna frogatzea izan da. Dena dela, badakigu, sintaxiaren tratamenduan aipatu dugun bezala, eremu hau neurri handi batean landugabea dela, eta etorkizunerako ikerlerroa izaten jarraituko duela zenbait urtetan.

Errorearen tratamenduan bi aspektu nagusi bereizten dira: *detekzioa* eta *zuzenketa* (Kukich 1992). Alde batetik errorearen detekzioak hitz batean edo esaldi batean akats bat gertatu den ala ez emango du. Zailagoa izaten da askotan zuzenketaren arazoa, errore bat dagoela jakinda ere proposamenen zerrenda batetik zuzena dena aukeratzea lan konplexua baita. Detekzioan nahiz zuzenketan, bereiztu behar dira hitz isolatuen prozesua eta testuingurua aztertzen duena. Hitz isolatuen kasuan, arrakastatsu suertatu dira hainbat zuzentzaile ortografiko komertzial. Beraien tratamendua hitzen zerrenda batean edo lexikoian oinarritzen denez mugak daukate,

kasu askotan beste motetako ezagumenduak, sintaxia edo semantika, behar izaten direlako detekzioa edo zuzenketa egiteko. Testuingurua kontuan hartuz gero, beste errore mota batzuetara zabal daiteke esparrua. Adibidez, hitz batean egindako erroreak beste hitz zuzena eman dezake (hauei *real-word errors* edo benetako hitzen erroreak deituko diegu), eta ezinezkoa izango da hitz isolatuen tratamendu hutsarekin detektatu. Berdin gertatuko da beste errore batzuetan (komunztadurakoak adibidez), hitzez hitz hartuta zuzenak direlako. Hauei errore sintaktikoak edo semantikoak deitzen zaie.

Kapitulua hasteko, erroreak sailkatzeko irizpideak deskribatuko dira (§ V.1.1), ondoren errorearen tratamenduaren munduan egin diren zenbait sistemaren azterketarekin jarraitzeko (§ V.1.2). § V.2n errorearen tratamenduaren inguruan egin ditugun lanak azalduko ditugu. Bertan errorearen detekzio eta zuzenketarako proposamen ezberdinak aurkeztuko dira, beren ebaluazioarekin batera. Egindako lanaren gehiena euskarari aplikatu zaion arren, ingelesezko errore ortografikoen zuzenketa automatikorako esperimentera ere aurkeztuko da. Kapituluak bukatuko da errorearen detekzio eta zuzenketari buruzko lanen ondorioekin eta geratzen zaizkigun eginkizunen deskribapenarekin.

V.1.1 Errore motak

Errorearen sailkapen bat egiteko momentuan lehen arazoa hori egiteko irizpidea aukeratzean datza, horren arabera sailkapen ezberdinak agertuko direlako (Douglas eta Dale 1991, Gómez Guinovart 1996, 1999). Hona hemen erabili diren irizpide batzuk:

- Irizpide deskriptiboa. Honek errorearen forma aztertzen du, errore ortografikoen tratamenduan egiten den antzera (Alegria 1995). Horrela aldaketa, sorrera, desagertzea eta trukearen ondorioz sortutako erroreak bereizten dira. Hitz-mailan eragiketa horiek karaktereen gainean egiten dira, eta beste mailetan (sintaxia, semantika), ezaugarri sintaktiko edo semantikoen gainean. Inplementazioaren ikuspuntutik, sailkapen hau egokiagoa da errore ortografikoen tratamenduan, hitz batetik karaktereak aldatuz sortzen diren proposamenak mugatuak direlako, eta esaldi bateko osagaien ezaugarri sintaktiko eta semantikoen aldaketek proposamen-kopuru ikaragarria sor ditzaketelako.
- Irizpide linguistikoa. Errorearen detekzio edo zuzenketarako behar den informazioaren arabera. Horrela, errore ortografikoak, sintaktikoak, semantikoak eta pragmatikoak bereiz daitezke. Mota honetako beste sailkapen sinpleagoa Kukich-ek (1992) egindakoa da: hitz isolatuen erroreak (ortografikoak) eta testuingurua behar

dutenak (testuinguru sintaktiko, semantiko edo pragmatikoa). Kasu honetan, zuzenketa egin behar duenaren ikuspuntutik sailkatzen dira erroreak, jatorri erabat ezberdinetakoak izan daitezkeen erroreak multzo berean sartuz, baldin eta zuzenketa behar den ezagumendua antzekoa bada. Adibidez, askotan teklatze-erroreek hitz zuzenak ematen dituzte (Peterson 1986), eta hauek errore sintaktiko edo semantikoetan sailkatuko dira, ezagumendu mota horiek beharko baitira tratamendu egokia egiteko, jatorrizko errore ortografikoak diren arren.

- Irizpide etiologikoa. Aurreko irizpidearen beste aldea kontuan hartuz gero, beste sailkapen mota bat erroreak jatorriaren arabera multzokatzearena da. Hemen enfasia ez da zuzenketarako tresnan jartzen, baizik eta errorea egiten duen pertsonaren gainean. Adibidez, pertsona batek *astakeritik* hitza idazten badu, arrazoia teklatze-errorea edo *a* organikoduna dela ez jakitearen ondorioa izan daiteke. Honek psikolinguistikarekin eta hizkuntzen irakaskuntzarekin ditu loturak (Díaz de Ilarraza *et al.* 1997, Maritxalar 1999), eta konplexutasuna gehitzen dio errorearen tratamenduari, detekzio-moduaz gain jatorriari buruzko informazioa erabili behar baita. Abantaila errorearen tratamenduaren egokitasunean dago, erabiltzailearen aldetik.

Gure kasuan, erroreetan dugun interesa egindako tresna informatiko eta linguistikoaren erabilera datza eta, beraz, bigarren motako errorearen sailkapenean, irizpide linguistikoan, oinarrituko gara. Gainera, tresnek morfologia eta sintaxia lantzen dutenez, ezagumendu mota horiek aplikatuz lor daitezkeen emaitzak azalduko ditugu, semantika³² edo pragmatika bigarren maila batean utziz. Azter ditzagun, beraz, errore mota ezberdinak:

- Errore lexikal edo ortografikoak. Lehen esan den bezala, multzo honetan lexikoi bat eta hitzen formaziorako erregelen bidez detekta daitezkeen erroreak sartuko ditugu, eta hitz zuzenak sortzen dituzten erroreak sintaktiko edo semantikoaren multzoan kokatuko ditugu. Errore lexikalen detekzioa morfologiarekin lotuta dago; bere tratamendua 80ko hamarkadan egiten hasi zen (Peterson 1980, 1986, Damerau eta Mays 1989), eta gaur egun emaitza onak lortzen dituzten zenbait sistema garatu dira (Ispell³³; Office 97rako tresnak³⁴; Aldezabal *et al.* 1999a). Arazo nagusietako bat hiztegiaren ez dauden baina zilegi diren hitzekin gertatzen da, hiztegiaren hutsunea

³² Errorearen tratamendurako semantikaren aplikazioa (Agirre, 1999) tesian azaltzen da.

³³ <http://fmg-www.cs.ucla.edu/geoff/ispell.html>

³⁴ <http://www.proofing.com> (*Proofing Tools for Microsoft Office 97*).

edo neologismoa dela eta. Tesi honetan sintaxiaren erabilera aztertu nahi dugunez, ez dugu errore hauen detekzioari buruzko gehiago esango.

Detekzioaren arazoa neurri handi batean ebatzita egon arren, zuzenketak beste problema batzuk dauzka. Hori argitzeko, ikus dezagun *problem* hitza sartuta zein diren ingeleserako zuzentzaile ortografiko baten proposamenak:

problem probe proem probed probes

Proposamen horien artean bereizteko beharrezkoa izango da testuinguruari buruzko informazioa, sintaxia (kategoria sintaktiko ezberdinetako proposamenak daude) edo semantika erabili beharko delako.

- Errore sintaktikoak. Esaldi bateko hitz guztiak lexikoaren bidez ondo analizatu eta gero, oraindik erroreak gerta daitezke esaldiak ez baditu betetzen gramatika baten arauak (gramatika bat zilegitzat eman daitezkeen egitura sintaktikoen deskribapen formala kontsideratuko dugu, gramatika nola adierazten den, eredu estatistikoak edo erregelak, zehaztu gabe). Definizio honekin ere errore ortografikoekin gertatzen zen arazo bera agertzen zaigu, zaila delako definitzea edozein testu ulertuko duen gramatika, eta horren ondorioz esaldi bat ezin denean ulertu, bereiztu egin beharko da errorea den edo gramatikatik at dagoen fenomeno den.

Errore sintaktikoen artean komunztadura-erroreak (deklinabide-atzizkia eta izen-multzoa, subjektua eta aditza), aditzen azpikategorizazio-eredu okerrean erabilera edo menpeko esaldiekin lotutako erroreak aipa ditzakegu. Nahiz eta oraindik emaitzak errore ortografikoentzat lortutakoak bezain onak ez izan, merkatuan errore sintaktikoen tratamendurako tresna ugari dago (Smith 1992, Rabinovitz 1993).

- Errore semantikoak. Sintaktikoki zuzenak diren esaldiek errore semantikoak izan ditzakete. Hauen artean bereizteko hautapen-murritzapenak erabili ohi dira lengoaia naturalaren prozesamenduan. Errore semantikoen detekzioaz gain, informazio semantikoa errore ortografiko edo sintaktikoetatik sortutako proposamenak diskriminatzeko ere erabil daiteke. Gure lanean sintaxiaren erabilera aztertzen ari garenez, informazio semantiko sinpleak bakarrik erabiliko ditugu eta horregatik ez diogu arreta handirik eskainiko errore mota honi.
- Diskurtsoaren erroreak edo errore pragmatikoak. Errore semantikoekin bereizketa egiteko esaten da pragmatikak hitzen esanahia testuinguruan hartzen duela. Dena dela, berbaldia eta pragmatikaren tratamendu konputazionala oraindik hasieran dagoenez ez dugu hauei buruz gehiago esango.

- Puntuazio-erroreak. Puntuazioak (Nunberg 1990) beste osagai linguistikoek bezala, bere arauak ditu, arau horien artean sintaxia, semantika eta berbaldia markatzeko erregelak daudela.

Errore mota bakoitzaren maiztasunen estimazio gutxi egin dira, batez ere hitz zuzenak sortzen dituzten erroreak ezin direlako automatikoki kalkulatu. Mitton-ek (1987), ikasleen 40.000 hitzeko testuak aztertu ondoren, errorearen %40a hitz zuzenak direla ondorioztatzen du. Honek sintaxia eta semantikaren erabilera eskatzen du, are gehiago hitz okerrearen proposamenen diskriminazioa ere kontuan hartzen bada, guztira errorearen erdiak baino gehiago tratatzeko sintaxia, semantika edo maila altuagoko ezagumendua erabiltzea beharrezkoa dela ondorioztatzen delako.

Hitz zuzena ematen duten erroreak neurtzeko saio bat egin zen Atwell eta Elliot-en (1987) lanean:

- Errorearen portzentaje handi bat (%52tik %4ra, testu motaren arabera) hitz okerrak dira
- Besteetatik gehienak (%48tik %28rako tartean) errore sintaktiko lokalak dira, hau da, hitz gutxi batzuen testuingurua aztertuz gero detektatu edo zuzendu daitezkeenak. Beste multzo bat errore sintaktiko globalena da (%16-%8 tartean), esaldi osoari buruzko ezagumendua beharko lituzketenak. Bukatzeko, errore semantikok aipatzen dituzte, %36-%10 tartean.

V.1.2 Errorearen detekzioarako zenbait sistemaren azterketa

Ezagumendu sintaktikoan oinarritutako errorearen tratamendua nagusiki 80ko hamarkadan hasi zen, gehienbat garai berean sintaxiaren formalizazioaren alorrean egindako aurrerapenei esker. Nahiz eta orduetik hona sistema asko diseinatu eta inplementatu diren (Douglas 1991, Kukich 1992, Stede 1992), puntu honetan sistema horietatik garrantzitsuenak edo eredu berriak ireki dituztenak aipatuko ditugu:

a) Analisi sintaktikoan oinarritutako sistemak.

Esperimentatu diren zenbait sistemak (Schank *et al.* 1980, Fass eta Wilks 1983, Carbonell eta Hayes 1983) ezagutza semantikoa erabiltzen du erroreari ekiteko, baina horrek eremu jakin batzuetara mugatzen ditu hurbilpen hauek, edozein testu tratatzeko behar den ezagumendu semantiko aberats hori eskuragarri ez dagoelako, eta horregatik metodoen orokortzea ezinezkoa da une honetan.

Ezagumendu gutxiago behar duten beste hurbilpen batzuek gramatika sintaktikoa dute oinarritzat (Heidorn *et al.* 1982, Weischedel eta Sondheimer 1983, Rodríguez 1991, Genthial *et al.* 1991, Jensen *et al.* 1993, Vosse 1992, 1994, Douglas eta Dale 1992, Menezo *et al.* 1996, Ramírez eta Sánchez 1996), *erlaxazioaren* teknika erabiliz: esaldi baten analisirik lortzen ez denean sistema huts egin duen erregelaren bilaketan ahalegintzen da, erregela horren murriztapen baten erlaxazioak analisi bat ematen duen egiaztatzeko. Honela komunztadura-erroreak, deklinazio okerrak edo izena-determinatzailearen arteko desadostasunak detektatu ahal dira. EPISTLE/CRITIQUE (Jensen *et al.* 1993) mota honetako sistema da. 300 erregela sintaktikorekin eta lexikoi zabal batekin (100.000 hitz) 100 errore gramatikal eta estilokoak detekta litezke. Tresna hau 2.254 esaldiko corpus batean probatu zen, eta emaitza nagusia errorearen zuzenketa (edo zuzenketarako laguntza, zuzenketa asmatzen ez denean) %82 eta %41 artekoa da, testu motaren arabera. Erlaxazioaren arazo bat testu errealei aplikatu ondoren ikusten da, oraindik ez delako garatu edozein esaldi (egunkari batean agertzen direnak kasu) analizatzeko gai den sistema, gramatiken mugak alde batetik eta testu errealeen lengoaiaren aberastasuna eta aldakortasuna bestetik (honi buruz, ikusi analisi sintaktikoari buruzko III. kapitulua). Horrela, esaldi batek analisirik ez badu, oso zaila da jakitea hori errore gramatikal baten edo gramatikaren hutsune baten ondorioa den. Bestalde, teknika honek aurkitzen duen beste oztopo bat erlaxazioak dakarren eraginkortasun-galera da, aukera-kopurua modu esponontzialean biderkatu egin daitekeelako.

Analisi sintaktikoan oinarria duten beste hurbilpen motak ere badaude. Sistema batzuek (Mellish 1989, Min eta Wilson 1998), *chart*-a erabiltzen saiatzen dira esaldi baten analisirik lortzen ez denean, bertan esaldiaren zatien analisiak daudelako, eta beraiekin zenbait konbinazio egin ondoren, batzuetan esaldi osoaren analisisia lor daitekeelako.

Beste ideia bat gramatika oso bat erabili gabe errorearen testuinguruak deskribatzen dituzten patroiak erabiltzea da. Era honetan, gramatika oso baten gainean baino, gramatikaren zati txikiak deskribatzen dituzten erregelekin, ziurtasun handikoak, lan egingo da. Ideia honen adibidea konpilazio-errore tipikoak inplementatzeko erabili da (In-Sig *et al.* 1993) eta murriztapen-gramatikan oinarria duten sistemak ere garatu dira (Lingsoft, <http://www.lingsoft.fi/>), nahiz eta bere funtzionamenduari buruzko publikaziorik ezagutzen ez dugun. Erroreen tratamendu mota hau sintaxiaren munduan gertatu den fenomenoaren isla dela uste dugu, gramatika oso batetik gramatikaren zati batzuen deskribapenera pasa delako, era horretan testu errealak modu eraginkorrean tratatzeko gaitasuna lortuz.

Antzeko beste tratamendu bat (Holan *et al.* 1997, Schneider eta McCoy 1998) gramatika bati errore jakin batzuei buruzko erregelak (*mal-rules* edo *error productions*) txertatzea da. Horrela bi ezagumendu mota, zuzena eta okerra, kodetzen dira gramatikan.

Bigarren hizkuntzaren irakaskuntzarako sistemetan lan ugari egin da, gehienetan aipatutako tekniken konbinazioak erabiliz, baina beti errorearen detekzio hutsa baino harantzago joanez, helburu nagusia irakaskuntza delako. Adibidez, Menzel eta Schröder-en (1999) sisteman errore-produkzioak eta erlaxazioa erabiltzen dituzte, azken hau gramatika sintaktiko baten gainean, baina baita semantika edo munduko ezagutzaren gainean aplikatuz. Lortutako sistema oso sendoa da, baina testuinguru jakin batzuetarako bakarrik erabil daiteke, edozein testuren aplikaziotik urrun.

- b) Metodo automatikoetan oinarritutako sistemak. Multzo honetan bi kategoria bereiztuko dugu: estatistikan oinarritutako sistemak eta ikasketa automatikoa erabiltzen dutenak. Biek komuna duten ezaugarri nagusia da eskuz edo automatikoki lortutako datu-multzo batetik ezagumendua ateratzen saiatzen direla.

Estatistikan oinarritutako sistemetan lengoaiaren modelizazio estatistikoa erabiltzen da errorearen detekzioa edo zuzenketa lantzeko, hitzen edo kategoria sintaktikoen bigrama/trigramak edo hitzen arteko agerkidetza-neurriak erabiliz. Atwell eta Elliott-en (1987) artikuluan 13.500 hitzeko corpus bateko erroreak detektatzen saiatzen dira, kategorien bigramen bidez, %62 detekzio-tasa eta %35eko doitasuna emaitza nagusia dela. Alarma faltsuen (doitasun baxua) kopuru altuegiak direnez emaitzak pbreak kontsideratu behar ditugu.

(Mays *et al.* 1991) lanak benetako hitzen errorearen (hitz zuzena sortzen dutenak) detekzio eta zuzenketaren problemari heldu nahi dio, 20.000 hitzeko lexikoiaren gaineko hitzen trigramen ereduak erabiliz. 100 esalditik 8.628 esaldi oker sortu zituzten ausaz, hitz batetik hurbil dauden aukerak sortuz. Errorearen detekzioa %76an egiten da ondo, %74ko zuzenketa-tasa lortuz. Lan honetako zenbait alde oraindik ikertu behar dira, esaldi multzo itxi batekin egin zelako, ez testu errealekin, eta horrek eragina du kostu handiko datu estatistikoak alde aurretik kalkulatu direlako.

Gale eta Church-ek (1990) hitz okerren zuzenketarako esperimentua deskribatzen dute, zuzenketarako proposamenak diskriminatzeke hitzen bigramen informazioa erabiliz. Era horretan ezagumendu sintaktiko eta semantiko primitibo baten erabilera egiten da. Emaitza testuingururik gabeko zuzenketarako beste hurbilpen batekin konparatzen da, %87tik %90erainoko igoera lortuz doitasunean. Emaitzek hobekuntza adierazten duten arren, oraindik esperimentatu beharko da bide hau.

Ikasketa automatikoaren arloan ere zenbait lan egin dira, ezagutza modu induktiboan ateratzeko. Golding eta Schabes-ek (1996) kategoria sintaktikoen trigramen eta testuinguruaren ezaugarriak konbinatzeko metodoa proposatzen dute benetako hitzen erroreak detektatu eta zuzentzeko. Bere esperimentuan agertzen den sistemak Microsoft Word-en zuzentzaile gramatikalak baino emaitza dezente hobeak lortzen ditu, baina estaldura hamazortzi errore-multzotara mugatzen da, bakoitzean okerra egiteko probabilitate altua duten antzeko bi edo hiru hitz daudela (adibidez: *weather*, *whether*). Mangu eta Brill-ek (1997) esperimentu bera egiten dute, ikasketa-metodo desberdina erabiliz. Antzeko emaitzak lortzen dituzte, baina beraien alde ikasitako erregela-kopuru txikia eta irakurgarritasuna aipatzen dituzte.

V.2 Erroreen detekzioan eta zuzenketan egindako esperimentuak

Errore motak eta erroreen tratamendurako sistemen aipamena egin eta gero, puntu honetan guk egindako lanak azalduko ditugu. Hasteko, euskarazko testuetan identifikatu ditugun erroreen sailkapena emango da (§ V.2.1). Ondoren, berorien detekzio eta zuzenketarako aukera ematen duten metodoen inplementazioa eta ebaluazioa azalduko dugu. Hiru esperimentu burutu ditugu. Hasieran, murriztapen sintaktikoen erlaxazioa aplikatuz lortutako emaitzak aipatuko dira (§ V.2.2), erroreen patroien detekzioaren hurbilpenaren lehen ebaluazioarekin jarraitzeko (§ V.2.3). § V.2.4en ingeleserako aplikatu den errore ortografikoen zuzenketarako sistema baten diseinua eta emaitzak aurkeztuko dira.

V.2.1 Euskarazko testuetako erroreen sailkapena

Ondorengo lerroetan euskarazko testuetan aurkitutako erroreen lehen sailkapen bat egingo dugu, hurrengo puntuetako aplikazioetan errore mota batzuk kokatzeko balioko duena. Sailkapen hau egiteko hainbat oinarri hartu ditugu: beste hizkuntzetarako egin diren saioak (Douglas eta Dale 1991, Gómez Guinovart 1996), eta euskara bereziki tratatu dutenak (Walz 1985, HABE 1985, Egunkaria 1992, Zubimendi eta Esnal 1992, Maritxalar 1999). Azken hauen artean badira argitaratu gabeko zenbait barne-dokumentu (euskararen irakaskuntzan diharduten instituzioetatik sortuak gehienbat). Erroreen deskribapen hauez gain, euskara ikasten ari diren ikasleen testu-sorta ere zuzenean aztertu ahal izan dugu (IRALE, AEK edo ILAZKI euskaltegiko ikasleen testuak), erroreen benetako agerpenen iturri aberatsa emanik. Ondorengo lerroetan azalduko diren erroreen adibide gehienak ikasleek egindakoak diren

arren, gure ustez errore horiek euskaldunek orokorrean egiten dituzten benetako errorearen adierazgarri dira, euskara-irakasleen iritzia jaso dugulako edo beste hizkuntzetan egindako erroreekin konparatu ditugulako.

Sailkapen honek ezagumendu sintaktikoa erabiliz errorearen detekzio eta zuzenketarako sistema baten definizioaren lehenengo pausoa emateko balioko du. Erroreak sailkatzearekin batera, errorearen kausa zein den aztertuko da, eta beren detekzioarako eta zuzenketarako inplementagarriak diren metodoak aztertuko dira:

- Errore ortografikoak. Hemen hitz solteetan egindako akatsak, hitz okerra ematen dutenak, kontsideratuko ditugu. Errore hauek hainbat arrazoiengatik sor daitezke, V.1 taulako adibideetan ikus daitezkeenez.

Adibide horietatik ikusten da errorearen jatorriak mota askotakoak izan daitezkeela, eta jatorri hori jakitea zuzenketarako proposamenak sortzeko edo sailkatzeko orduan erabil daitezkeela. Adibidez, errore fonologikoak tratatzeko, hitz okerretatik fonologikoki hurbil dauden proposamenak atera daitezke, eta hurbiltasun-neurri hori zuzenketa bat aukeratzeko erabili (Yannakoudakis eta Fawthrop 1983, Berkel eta DeSmedt 1988). Berdin gertatuko da teklatze-erroreekin, kasurik gehienetan karaktere baten distantzia egoten direlako hitz zuzenak (Pollock eta Zamora 1984), edo OCR (*Optical Character Recognition*) sistemek egindako erroreekin. Dena dela, informazio horiek lagungarriak izan arren, zuzenketa guztiz automatikoa egiteko doitasun handiaz sintaxiaren eta semantikaren erabilera ezinbestekoa izango da.

Errorea	Errorearen arrazoia
<i>araso, iztilua, naskagarri, herbestean, majikoen</i>	errore fonologikoak
<i>eskarrizketa</i>	teklatze-errorea
<i>Itziarri</i>	deklinabidearen aplikazio okerra
<i>bait da</i>	idazkera-arauen aldaketa
<i>inglesak, errelazio</i>	inguruko erdaren eragina
<i>dirazenean</i>	aldaera dialektala
<i>mexikatarrak, aurkeztzaileari</i>	eratorpen okerrak

V.1 taula. Errore ortografikoen adibideak.

- Errore sintaktikoak. Hauen jatorria mota askotakoa izan daiteke: errore ortografiko baten ondorioz hitz zuzen bat sortzen denean, komunztadura-erroreak,

deklinabide-erroreak, edo menpeko esaldien erabilera okerretik sortutako erroreak. V.2 taulan horien adibideak daude.

Adibide horietatik ikus daiteke sintaxiaren tratamendua behar duten errorearen aniztasuna errore ortografikoena baino askoz handiagoa dela. 1) adibideak erakusten du nola tekleatze-errore batek hitz zuzen bat ematen duen, eta hori detektatzeko ezagumendu sintaktikoa gutxienez edo seguruenik semantikoa ere beharko litzatekeela. Beste errore batzuk (2, 3, 4 eta 5) izen-sintagma baten edo testuinguru mugatu baten azterketarekin detektagarriak izan daitezkeela pentsa daiteke. Bukatzeko, badaude esaldi osoaren azterketa sintaktikoa beharko duten errore motak (6, 7, 8 eta 9).

Zuzenketarako proposamenak lortzeko, errore mota bakoitzak bere metodo propioen definizioa beharko duela uste dugu. Errore ortografikoekin bezala, proposamen horien artean zuzena automatikoki aukeratzeko informazio sintaktiko eta semantikoak sartu beharko dira.

	Errorea	Errorearen arrazoia
1)	<i>flotatzte dute</i>	tekleatze-errorea
2.a)	<i>gure herrialdetan lan egiteko</i>	mugagabearen erabilera okerra
2.b)	<i>itsasorentzat txarra da</i>	
2.c)	<i>edozein momentuan jasango du</i>	
2.d)	<i>zenbat gonbidatuak</i>	
2.e)	<i>hainbeste urtean</i>	
3.a)	<i>nahiz eta animalia basati izan</i>	deklinabidearen aplikazio okerra
3.b)	<i>Lurra osoa</i>	
4)	<i>Donostian, 1999ko urtarrilak 1ean</i>	dataren errorea (zenbakia ezin da deklinatu hila ergatiboan badoa)
5)	<i>ez ziren elkar bizi etxe kanpo</i>	komunztadura-errorea eta postposizio okerra
6.a)	<i>Hori eztiarekin zerikusirik ez dauka</i>	komunztadura-erroreak
6.b)	<i>nik ez naiz izan</i>	
6.c)	<i>hau ez du esan nahi hori</i>	
6.d)	<i>bueltak ematen nuen</i>	
6.e)	<i>Patxik kontatu zidan kuadrilako planak</i>	
6.f)	<i>biztanle guztiak goitizena dute</i>	
7.a)	<i>lan bat neretzat</i>	esaldiaren osagaien ordena
7.b)	<i>proposatzen dio plan bat</i>	
7.c)	<i>gauza batzuk gertatutakoak kontatuko dizut</i>	
8)	<i>prest dago guri zirkora eramateko</i>	kasuaren asignazio okerra edo aditzaren azpikategorizazioaren ezagutza-eza
9.a)	<i>lagunei molestatzen zebilen</i>	azpikategorizazioaren ezagutza-eza
9.b)	<i>Donostiara joatea behar nuela</i>	
9.c)	<i>nik ez dakit nola da Bretaña</i>	
9.d)	<i>alarma hautsi lortu zuten</i>	

V.2 taula. Errore sintaktikoen adibideak.

- Errore semantikoak eta pragmatikoak. Tesi-lan honetan sintaxiari eskainiko diogunez indar handiena, ez diegu semantikarekin lotura dituzten erroreari garrantzi handia eman. V.3 taulan zenbait adibide ditugu, sintaktikoki zuzenak diren esaldiak baina arazo semantikoak dituztenak (honi buruz ikusi Agirre *et al.* 1994, Agirre 1999). Taulako lehen esaldian, adibidez, komunztadura-errore bat agertzen da, interpretazio sintaktiko zuzena onartzen duena (*ekintza* subjektutzat hartuz gero).

Errorea	Errorearen arrazoia
<i>Edozein ekintzak ez duela onartuko</i>	komunztadura-errorea
<i>haizkorarekin ebaki dugu</i>	instrumentala (haizkoraz, irratis,
<i>irratitik zabaldu dute berria</i>	horretaz) erabili behar da
<i>horrena hitzegingo dut</i>	
<i>badaukazu ura?</i>	partitiboa erabili behar da
<i>beroa egiten du</i>	azpikategorizazioaren ezagutza-eza

V.3 taula. Errore semantikoen adibideak.

V.2.2 Murritzapen sintaktikoen erlaxazioa

V.2.2.1 Metodoaren azalpen laburra

Puntu honetan errore sintaktiko batzuei murritzapenen erlaxazioaren (*constraint relaxation*) metodoaren aplikazioaren emaitzak azalduko dira, (Gojenola eta Sarasola 1994) lanean eta ondorengo beste esperimentu batzuetan egin direnak. Metodo hori zenbait sistematan erabili izan da (Heidorn *et al.* 82, Douglas eta Dale 92, Tomabechi 93, Ramírez eta Sánchez 96, Mitjushin 1996), eta bere oinarria gramatika bat da. Analizatzailerik sintaktikoak, esaldi oso baten analisia lortzen ez duenean, esaldiaren zenbait osagai buruzko murritzapenak kentzen ditu (erroreen sorburua izan daitezkeenak, adibidez, komunztadura numeroan) eta, analisi bat lortzen bada, orduan errorearen kausa erlaxatutako murritzapenekin erlaziona daiteke. Era honetan, sistema gai izango da gramatikaren estalduratik kanpo dauden esaldiak ulertzeko ere.

Murritzapenen erlaxazioaren metodoaren aurkezpena egin dugunean, esaldi bat analizatzerik ez dagoenean bi arrazoiengatik izan daitekeela esan dugu: errore sintaktikoa dagoela edo esaldia zuzena dela baina gramatikaren estalduratik kanpo dagoela. Garatu dugun gramatika partziala denez, esan dugu testu errealetako esaldi asko ezin izango direla osorik analizatu (§ III.2). Horregatik, bi aukera horien arteko banaketa sinplifikatzeko, erlaxazioaren probak gramatikaren estalduraren barruan dauden esaldien gainean egingo ditugu, eta horrela esaldi batekin analisia ez lortzea errore baten agerpentzat ulertuko da. Sinplifikazio honen desabantaila handiena testu errealetako esaldiak saihestea da (edo testu errealetako esaldiak erabiltzen badira berauek gramatikaren barruan egon daitezzen moldaketak egin beharko dira), baina daukan abantaila nagusia da erlaxazioan oinarritutako metodoaren ebaluaziorako egokia dela, gramatikaren estalduraren arazoaren abstrakzioa egiten delako. Gainera, hau interesgarria izan daiteke beste testuinguru batzuetan, euskara irakasteko sistema batean

adibidez, gramatika sinplifikatu baten gaineko ariketak emateko. Era honetan, gramatika ideal bat bageneuka bezala joka dezakegu (kasu hoberena), erlaxazioaren ekarpena eta bideragarritasuna modu isolatuan neurtuz. Pauso hori eman eta gero, testu errealekin lan egiteko behar diren egokitzapenak diseina daitezke hurrengo fase batean. Horregatik, puntu honetan hartuko dugun hasierako hipotesia esaldi guztiak gramatikarekin analiza daitezkeela izango da, esperimentuak horren arabera eginez. Etorkizuneko lanen aipamenean (§ V.3), erlaxazioaren tratamendua testu errealeko esaldiekin nola lot daitekeen azalduko dugu.

Metodo horren euskararako aplikazioa baterakuntzan oinarritutako gramatikaren gainean egingo dugu. Baterakuntza-formalismoak egokiak dira murriztapenen erlaxazioaren aplikaziorako (Douglas eta Dale 1992), beraietan ezagumendu sintaktikoa murriztapenak deskribatzen dituzten ekuazioen bidez adierazten delako, eta ekuazio horietako batzuk kenduz inplementa daitekeelako. Kendutako ekuazio horiek emango dute akatsaren iturria, beraiek gabe esaldiaren analisia lortzea balego.

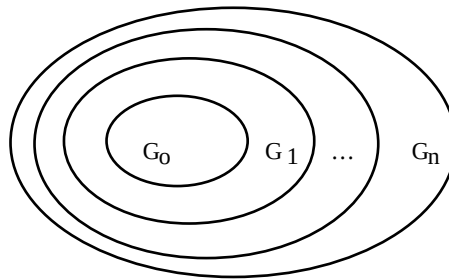
V.1 adibidean erregela bat azaltzen da, izen-sintagma eta aditzaren arteko numeroaren komunztadura eskatzen duena (4garren ekuazioa). Horrela, ez dira onartuko *'txakurra datoz'* bezalako esaldiak, baina laugarren ekuazioa askatuz gero lortuko da esaldi horren analisia. Ekuazioak numeroaren komunztadura egiaztatzen duenez, errorearen arrazoia numeroaren desadostasun batetik datorrela izan liteke zuzenketarako hipotesi bat. Hemen ikusten da erlaxazioaren beste abantaila bat, gramatika bera erabiltzen dela esaldi zuzenak eta okerrak ezagutzeko. Baterakuntzan oinarritutako gramatikak egokienak dira metodo honen aplikaziorako, murriztapenak modu erazagutzailean agertzen direlako, horrela murriztapen jakin bat askatzeko aukera emanez. PATR formalismoa, alde honetatik, LFG edo HPSG formalismoak baino errazago da metodo honekin erabiltzeko, formalismo konplexuagoetan murriztapen asko printzipio orokorren bidez adierazten direlako (adibidez, komunztaduraren printzipioa) eta zailagoa da esatea erregela jakin batean murriztapen bat (adibidez, bi osagaien arteko komunztadura numeroan) aska daitekeela.

X0	--->	X1, X2	
1)		X0/kat	<=> esaldia
2)		X1/kat	<=> izen-sintagma
3)		X2/kat	<=> aditz-sintagma
4)		X1/num	<=> X2/num

V.1 adibidea. Numeroaren komunztadura egiaztatzen duen erregela baten adibidea.

Murriztapen sintaktikoen erlaxazioa mota ezberdinetako lengoaietara aplikatu da, gaztelera edo ingelesa kasu (Heidorn 1982, Rodriguez 1993, Tomabechi 1993). Dena dela, hizkuntza eranskari edo malgukarietan (errusiera, suomiera edo euskara (Miller 1986)), metodoaren

aplikazioak analisirako aukera-kopuru trataezinak sortzen ditu. Horregatik, erlaxazioa modu gradualean aplikatu beharko da, aukera-kopurua tratagarria dela ziurtatzeko (Douglas eta Dale 1992). Era horretan, gramatika baten estaldura zabaldu egin daiteke, V.1 irudian agertzen den moduan. G_0 hasierako gramatika bada, zenbait pausotan G_1 , G_2 , eta G_3 gramatika zabalagoak lor daitezke, murriztapen sintaktikoak kenduz, eta horrela esaldi ez gramatikalak ulertzeko gaitasuna lortuko da.

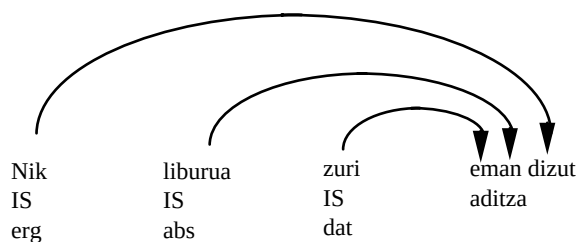


V.1 irudia. Murriztapen sintaktikoen erlaxazioaren bidezko gramatikaren zabalpena.

V.2.2.2 Egindako esperimentuak

Esperimentuetarako, 119 esaldi okerrekin osatutako multzoa aukeratu genuen, euskara ikasten ari diren ikasleek idatzitako testuak erabiliz. Testu hauek datu-iturri aberatsa osatzen dute, errore-kopuru altua baitute eta gainera egindako akatsak adierazgarriak dira jatorrizko euskaldunak egiten dituzten erroreekin alderatuz gero, euskara-irakasleen iritziz. Testuen corpusak euskararen morfologiaren irakaskuntzarako IDAZKIDE proiektuan lortu ziren (Díaz *et al.* 1997), eta 100 testu inguru ditu, bakoitza 150-200 hitzekin. Erlaxazioaren proba egiteko, aukeratutako erroreak maiztasun gehienekoak izan ziren, § V.2.1en aurkeztutako errore mota guztietatik:

- Aditza eta esaldiko osagaien arteko komunztadura kasu, numero eta pertsonan. Komunztadura mota hau ergatibo, absolutibo eta datibo kasuetan doazen izen-sintagmekin gertatzen da (ikus V.2 irudia). Osagaien ordena librea eta elipsia (osagai bat ez agertzea) gehituz gero, ondorioa da esaldi okerrekin kasuan errore-hipotesi kopuru handia aztertu beharko dela.



V.2 irudia. Esaldiko osagaien arteko komunztadura.

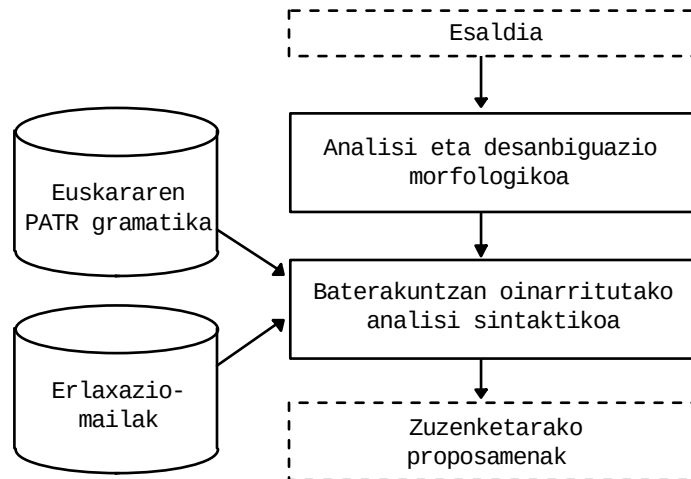
- Izen-sintagmaren barruko erroreak, hauen artean determinatzaile edota numeroaren gehitzea, kenketa edo ordezkapena daudela.
- Menpeko esaldiekin egindako erroreak, menpekoaren atzizkia edo markaren faltarengatik edo ordezkapenarengatik. Errore hauek komunztadura mota berezi gisa ere ikus daitezke, edo azpikategorizazio-murritzapenen akats moduan.

V.4 taulan errore horien adibideak agertzen dira, dagozkien zuzenketekin batera. Lehen esanenez, esaldiak gramatikaren estalduraren barruan daude, hau da, esaldi erroredunen zuzenketek analisi bat (edo gehiago) izango dute. Horrela egiteko lana gramatika zabaltzea izango da, murritzapen sintaktikoen ezabaketen bidez esaldi oker guztiak analizatu arte.

	Errorea	Zuzenketa
Komunztadura- erroreak	<i>bat-batean gizon bat hamabost zakurrekin agertu ziren .</i> <i>notizia horrek gauza onak ekarriko du .</i> <i>Zenbait ikerlari eman nion .</i> <i>Pezetaren debaluazioa alde onak ekarriko dizkigu .</i>	<i>bat-batean gizon bat hamabost zakurrekin agertu zen .</i> <i>notizia horrek gauza onak ekarriko ditu .</i> <i>Zenbait ikerlariri eman nion .</i> <i>Pezetaren debaluazioak alde onak ekarriko dizkigu .</i>
Izen-sintagmaren barruko erroreak	<i>Zenbait probintzietan hondartza dago .</i> <i>atezaina batekin egon zen .</i> <i>kotxeren historiak aldatu du .</i>	<i>Zenbait probintzietan hondartza dago.</i> <i>atezain batekin egon zen .</i> <i>kotxearen historiak aldatu du .</i>
Menpeko esaldien erroreak	<i>esanda dago inkestak datoz .</i> <i>gizon horrek esan du nik istilua izan dut .</i> <i>Nik eman dizut liburua polita da</i>	<i>esanda dago inkestak datozela .</i> <i>gizon horrek esan du nik istilua izan dudala .</i> <i>Nik eman dizudan liburua polita da</i>

V.4 taula. Tratatu diren errore sintaktikoen adibideak.

V.3 irudian era honetako errorearen tratamendurako egin dugun sistemaren arkitektura azaltzen da. Bertan gramatika eta erlazazio-mailei buruzko informazioa modu erazagutzailan kodetuko da. Analisi morfologikoaren ondoren, analizatzaile sintaktikoa esaldia analizatzen saiatuko da. Emaiztarik ez bada ateratzen, orduan lehen mailako murritzapenen askatzea onartuko da. Analisia lortzen bada, askatutako murritzapenek errorearen iturria markatuko dute. Analisisirik ez balego, orduan bigarren, hirugarren, ... mailak probatzen jarraituko du, analisi bat aurkitu arte edo maila guztiak aplikatu arte.



V.3 irudia. Erroreak tratatzeko sistemaren arkitektura.

```

% izenlaguna --> is3(mugagabe) + knmdek(gen/gel)
% adibideak: (zenbait gizon) + en
%            (gizon bi) + ren
%            (Peio) + ren
% gaizki:    *(zenbait gizon) + aren
X0 ----> X1, X2
1) X1/kat      <=> is3
2) X2/kat      <=> knmdek
3) X2/kom/kas  badago [gen, gel]
4) X0/kat      <=> izlg
5) X0/izlg_mota <=> en_ko
6) X0/info     <=> X1
7) X1/kom/mug  <=> X2/kom/mug
8) X1/kom/num  <=> X2/kom/num
9) edo[X1/osgk/det/gune/azp badago [dzh, dzg],
   X1/gune/azp badago [izb, lib]]
10) X2/kom/mug badago [mg]
...
Beharrezko murriztapenak: [1,2,3,4,5,6, 8, ...],
erlaxazio-posiblea([7,10], "izen-sintagmaren atzizkiak mugagabe izan behar
du")
erlaxazio-posiblea ([9], "izen-sintagma hau ez da mugagabea")
Maila: 1

% as --> is(abs) + as
% adibideak: gizonak + (etorri dira)
%            gozokiak + (nik eman dizkiot gizonari)
% gaizki:    *gozokia + (nik eman dizkiot gizonari)
X0 ----> X1, X2
1) X0/kat      <=> as
2) X1/kat      <=> isk
3) X2/kat      <=> as
4) X2/ordena   <=> ezkerra
5) X1/kom/kas  <=> X2/azpikat/abs/kom/kas
6) X1/kom/num  <=> X2/azpikat/abs/kom/num
7) X1/kom/per  <=> X2/azpikat/abs/kom/per
8) X1/kom/mug  <=> X2/azpikat/abs/kom/mug
...
Beharrezko murriztapenak: [1,2,3,4,8, ...],
erlaxazio-posiblea([5], "objektua-aditza komunztadura errorea (kasua)")
erlaxazio-posiblea([6], "objektua-aditza komunztadura errorea (numeroa)")
erlaxazio-posiblea([7], "objektua-aditza komunztadura errorea (pertsona)")
Maila: 1

```

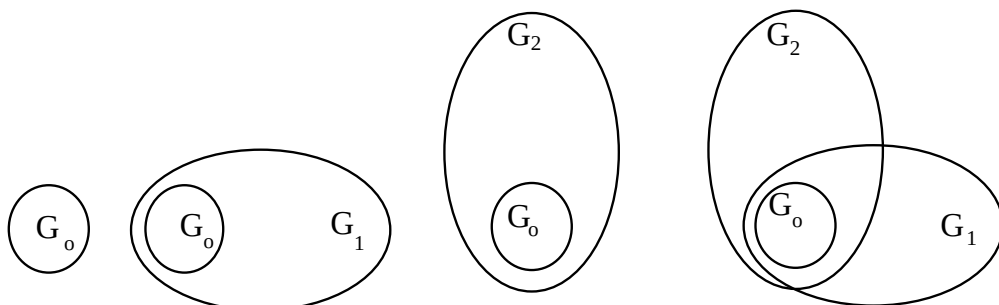
V.2 adibidea. Erlaxazioaren aplikazioa bi erregeletan.

V.2 adibidean bi erregela agertzen dira, bakoitzean erlaxatu daitezkeen murriztapenekin. Erregela bakoitzeko, beharrezko murriztapenak (erlaxatzen ez direnak) zehazten dira alde

batetik, eta bestetik errorea eman dezakeen murriztapena(k). Horietako bat kenduta analisi bat lortzen bada, dagokion mezuak errorearen kausa emango du. Adibidez, lehenengo erregelaren bi murriztapenek (7 eta 10) kasu-marka mugagabea dela egiaztatzen dute, izen-sintagma mugagabeak osatzeko. Akats arrunta mugatua jartzea izaten denez

(*zenbait etxe + ak'), murriztapen horiek kenduz gero izen-sintagma horiei analisi bat emango zaie. Antzeko gauza egiten da bigarren erregelarekin, aditz-sintagma bat absolutiboan doan izen-sintagma bat lotzeko erabiltzen dena. Seigarren ekuazioaren askatzeak, adibidez, numeroaren desadostasuna detektatuko du (*gizonak + ikusi dut').

Gramatika eta errorearen corpus osoak aztertu eta gero, 75 murriztapen aukeratu ziren, beraien ezabaketaren bidez errorearen analisia lortzen lagungarriak izan zitezkeenak. Lehen esperimentu batean metodoa modurik sinpleenean aplikatu zen, errorea detektatzen lagun zezaketen murriztapen guztiak aldi berean askatuz. Horrela, analizatzaileak bi pausotan funtzionatu zuen: lehen maila batean murriztapen guztiak bete behar ziren, esaldi zuzenak analizatuz; bigarren maila bat aplikatzen zen esaldi baten analisia lortzerik ez zegoenean, murriztapen guztiak aska zitezkeela. 63 esaldi okerrekin probatu ondoren, analizatzaileak 58ren analisia lortu zuen, metodoaren erabilgarritasuna frogatuz eta analizatzaileen sendotasuna lortzeko metodoaren baliagarritasuna ere bai. Geroago esango dugunez, ezagututako esaldi bakoitzeko zuzenketarako proposamen bat edo gehiago lortzen ziren. Dena dela, aukera honek eraginkortasun-arazoak zeuzkan, errore bat dagoenean 75 murriztapen guztiak batera askatzeak analisi-aukeren ugaltze ikaragarria eman baitezake, anbiguotasunaren betiko arazoari gehitzen zaiona. Horregatik bost esaldik ez zuten analisirik hartu, denbora eta espazioaren leherketa ekarri baitzuten. Esaldi okerrearen analisirako batez besteko denbora 29,41 segundo³⁵ izan ziren, zegozkien esaldi zuzenduen batez besteko 2,5 segundoekin konparatuz.



V.4 irudia. 4 erlaxazio-mailaren erabilera.

³⁵ Analizatzailearen inplementazioa ordutik hona hobetu denez, denborak oraingo analizatzailearenak baino txarragoak dira. Horregatik, emaitzak aztertzeko orduan denboren arteko erlazioa izan beharko da kontuan.

Eraginkortasun-aldaketa estimatzeko, beste proba bat egin zen hamar erlaxazio-maila definituz; horietako zortzitan errore-mota baten tratamendua egiten zen, eta beste bi mailetan esaldi zuzenak eta murriztapen guztien erlaxazioa deskribatzen ziren. V.4 irudiak lau erlaxazio-mailaren erabilera azaltzen du. G_0 hasierako gramatika da, eta G_1 eta G_2 gramatikek bi errore moten murriztapenen multzoak osatzen dituzte. Azken aukeran murriztapen guztiak aska daitezke. Zortzi erlaxazio-mailak errorearen maiztasunak gogoan izanda definitu ziren, ideia nagusia maiztasun handiagoko erroreak detektatzeko erlaxazioak lehenago probatzea zela (ikus V.5 taula).

Errorea	Errorearen adibidea
Mugagabea jartzea mugatua denean	<i>erleren laguna han dago</i>
Komunztadura objektuaren (absolutibo kasuan doana) numeroan	<i>etxeak joan da</i> <i>nik etxeak ikusi dut</i>
Komunztadura objektuaren kasuan	<i>nik gizonak ikusi ditut</i>
Komunztadura subjektuaren kasuan	<i>ni gizonak ikusi ditut</i>
Konpletiboetan menpeko atzizkia ez agertzea	<i>esan du etorri da</i>
Mugagabea ez jartzea behar denean	<i>zenbait erlearen laguna ikusi dut</i>
Komunztadura subjektuaren numeroan	<i>gizonak egin du</i>
Komunztadura bigarren objektuaren numeroan	<i>emaitzei dagokiona</i>

V.5 taula. Erlaxazio-mailak eta errorearen adibideak, maiztasunez (altuenetik txikienera) ordenatuta.

Esaldien batez besteko analisiaren denbora 27 segundokoa izan zen. Emaidza hauek besteekin alderatuta alde gutxi dutela eman arren, faktore bi eduki behar dira gogoan:

- Kasu honetan esaldi guztien analisisa lortzen da, eta analizatzailea ez da sartzen analisi-aukera esponentzialaren tratamenduan. Aurreko kasuan, murriztapen guztiak batera askatzean, kasu batzuetan ezin zen aurkitu errorearen kausa. Oraingo aukeran banan-banan aztertzen dira erroreak, eta kasu guztietan analisi bat ematen da. Beraz, denbora irabazteaz gain doitasunean ere irabazten da.
- Beste arazo bat analizatzailea optimizatu gabe egotea da, analisi-maila bakoitzean zerotik hasten delako, aurreko mailetan egindako lana berrerabili gabe. Adibidez, V.3 irudian G_0 -ren analisi partzialak maila guztietan erabil daitezke. Soluziorik hoberena pausoz-pausoko analizatzailea erabiltzea izango litzateke (Wiren 1993), maila berri baten analisiari ekiteko momentuan erlaxazio berriek alda ditzaketen osagaiak bakarrik erabiliz, lan berria minimizatuz. Gainera, zortzi erlaxazio-mailak ez dira optimizatu, banan-banan probatu direlako, eta eraginkorragoa izan daitekeelako maila batzuk biltzea, beren maiztasunen arabera. Aldaketa horiek

inplementatuz gero, espero dugu analisi-denborak ia antzekoak izatea esaldi oker eta zuzenentzat.

Emaitza horietatik zenbait ondorio atera ditugu. Batetik, erlaxazioaren metodoaren bideragarritasuna, esaldi okerrak ulertzeko ezagumendua gramatika bati lotzea ahalbidetzen duelako. Esperimentuetan esaldi guztientzako analisiak lortu ahal izan dira. Bestetik, erlaxazio hori modu eraginkorrean egiteko era gradualean egin behar dela, bestela denbora eta espazioarekin arazoak agertzen direlako, are gehiago errorearen multzoa zabaltzen doan heinean. Edozein kasutan, oraindik esperimentazio asko egin behar da arazo hau gainditzeko.

Laburbilduz, esango dugu aurreko puntuetan frogatu dugula erlaxazioak errore sintaktikoen tratamendua ahalbidetzen duela, esaldi okerrak analizatu ahal direlako murriztapen sintaktikoak erlaxatuz. Horrela, detekzioaren arazoa konpondu da gehienbat. Hori argitzeko, azter ditzagun 53 esaldi erroredunen gainean ateratako emaitzak (ikus V.6 taula). Bertan, esaldi bakoitzeko zuzenketa-proposamenen artean errore-kopuru minimoa daukatenak bakarrik aipatzen dira, kasu gehienetan horien artean egoten baita proposamen zuzena. Hasteko, 11 esaldik analisi bat hartzen dute, hau da, interpretazio sintaktiko zuzen bat daukate (adibidez, *‘gizona eman dit’*). Honek adierazten du errore sintaktikoen proportzio handi batek (%20aren inguruan kasu honetan) esaldi zuzenak ematen dituela, eta beraien konponketarako ezagumendu semantikoa, gutxienez, beharko litzatekeela. Bestalde, 14 esaldik zuzenketa-proposamen bakarra dute, baina 28k bi edo gehiago dauzkate. Beraz, proposamenak sailkatzeko irizpideak definitzea garrantzitsua da, Menzel-ek (1990) egin duen bezala, murriztapen gutxien hautsi dituzten interpretazioak hautatuz edo heuristikoen arabera hobetsiz.

Analisi zuzena	Zuzenketa bat	Zuzenketa bi	Zuzenketa bi baino gehiago
11	14	17	11

V.6 taula. Analizatutako esaldi okerren banaketa zuzenketa-proposamenen arabera.

Proposamenak zuzentzeko, ezagutza sintaktiko eta semantikoa beharko da. V.7 taulan bi esaldi eta sistemak proposatutako zuzenketak agertzen dira. Adibidez, lehenengo esaldiaren zuzenketan bazter liteke *harreman sozialak* subjektutzat hartzen duen interpretazioa (jakinez gero *daukate*-ren subjektua normalean biziduna dela). Bigarren esaldian, kontua da aukeratzea zein diren *dauka*-ren subjektua eta objektua, *haitzuloa* edo *garrantzia*. Dena dela, zuzenketaren aspektu honek azterketa sakona beharko du, erabakitzeke zeintzuk diren bereizketa hori egiteko behar diren ezagumendu sintaktiko edo semantikoak.

Errorea	Zuzenketarako proposamenak
<i>harreman sozialak daukate</i>	subjektua-aditza komunztadura errorea (kasua): <i>harreman sozialek daukate</i>
	objektua-aditza komunztadura errorea (numeroa): <i>harreman sozialak daukate</i> edo <i>harreman soziala daukate</i>
	objektua-aditza komunztadura errorea (kasua): <i>harreman soziala daukate</i> (ergatibo -> absolutibo)
<i>erreportaia honetan Ondarruko haitzuloa sekulako garrantzia dauka.</i>	subjektua-aditza komunztadura errorea (kasua): <i>erreportaia honetan Ondarruko haitzuloa sekulako garrantziak dauka.</i>
	subjektua-aditza komunztadura errorea (kasua): <i>erreportaia honetan Ondarruko haitzuloak sekulako garrantzia dauka.</i>

V.7 taula. Esaldi erroredunak eta zuzenketarako proposamenak.

V.2.2.3 Ondorioak

Aurreko puntuan (§ V.2.2) errore sintaktikoen detekzio eta zuzenketarako erlaxazioaren metodoa euskarari aplikatu zaio. Euskarak ezaugarri bereziak dauzka errorearen tratamendurako (komunztadura aberatsa adibidez), ingelesa bezalako hizkuntzekin konparatuta. Baterakuntzan oinarritutako gramatikaren gainean inplementatu da sistema, bere erazagutzailetasunari esker murriztapenak adierazteko ekuazioen askapen erraza onartzen duelako. Erlaxazioak errore-multzo zabal baten analisia onartzen duela frogatu dugu, eta eraginkortasuna hobetzeko murriztapenen askapen gradualak beharrezkoa dela. Emaitez metodoaren bideragarritasuna frogatu duten arren, badira oraindik egin beharreko lanak:

- Errore mota gehiago aztertu behar dira, eta dagozkien erlaxazio-mailak definitu. Oreka lortu behar da maila gehiago definitzeak dakartzan irabazien eta galeren artean. Mailakatzeko analisia posible egiten du, baina maila altuetako erroreak harrapatzeko kostu handiagoa izango du, aurreko maila guztiak probatu behar direlako. N errore posible definitu badira, batez beste $N/2$ aldiz analizatu beharko da esaldia errore sinple bat harrapatzeko (erroreen maiztasunak kontuan hartzen ez badira behintzat), esaldian bi errore badira $(N*(N-1))/2$ aldiz errore biak harrapatzeko, Adibidez, aditza eta izen-sintagmen arteko komunztadura egiaztatzeko, pertsona, numero eta kasuaren murriztapenak maila berean askatzeak errore-hipotesien kopuru altua aztertu behar izatea dakar, gainera gehienak baztertu egingo dira gero, errore asko dituztelako. Kasu honetan ideia hobea izan daiteke murriztapen horiek maila ezberdinetan askatzea. Maizen gertatzen diren komunztadura-erroreak, ergatiboaren kasu-marka ez jartzea adibidez, lehenengo mailatan jarri beharko dira.

- Errore bat detektatu eta gero, zuzenketa aukeratzeko lana geratzen da, askotan proposamen bat baino gehiago agertuko delako. Une honetan, proposamen guztiak aurkezten zaizkio erabiltzaileari. Horretarako desanbiguaziorako teknikak garatu beharko dira. Bide interesgarri bat errore ortografikoen zuzenketa automatikorako egin diren lanak dira (Golding eta Schabes 1996, Mangu eta Brill 1997, Agirre *et al.* 1998). Zuzenketa aukeratzeko orduan kontrako justifikaziok eman dira, lan batzuetan errorearen testuinguru hurbilaren azterketa gomendatzen delako baina beste batzuetan, aldiz, esaldiaren analisi orokorra egitea komenigarria dela aipatzen da. Zuzenketa bat egiteko, gehienetan hitz bat aldatu beharko da sortzaile morfologikoa erabiliz, baina beste kasu batzuetan osagai sintaktiko konplexuak aldatu behar dira, sorkuntza sintaktikoaren bidez. Ikusten denez, puntu honek ikertzeko aukera ugari ematen ditu.
- Landu ez dugun beste ideia interesgarria sistema ingurune ezberdinetara egokitzea da (adibidez, ikasleen testuen zuzenketa, bigarren hizkuntzen irakaskuntza, edo egunkari bateko testuak), maila-kopurua eta errore-motak testuaren ezaugarriei moldatuz.

V.2.3 Errore-patroien bidezko detekzioa

V.2.3.1 Sarrera

Aurreko puntuan ikusi dugu gramatika baten gainean murriztapen sintaktikoen askapenaren bidez errore batzuk, komunztadurak adibidez, detekta daitezkeela. Badaude beste errore batzuk, zenbait arrazoiengatik metodo horrekin trataezinak direnak. Egindako baterakuntza oinarritutako gramatikak euskararen osagai nagusiak definitzen ditu: izen-sintagmak, adizlagunak, menpeko esaldiak eta esaldi sinpleak. Horien bidez lengoaiaren osagairik arruntenak edo maizen gertatzen direnak estaltzen dira. Ondorioz, gramatika hori erabiliz erlaxazioaren bidez detektatu ditugun errore gehienek esaldiaren analisi osoa beharko dute, hau da, lehenago egindako errore sintaktikoen bereizketan (§ V.1.1) aipatutako errore sintaktiko globalak dira. Bestalde, gramatikaren estaldura osoa ez denez, ezin dira gramatikaz kanpoko testuinguruetako erroreak tratatu, erlaxazioa esaldi zuzenen gramatika baten gainean oinarritzen delako. Adibidez, daten deskribapen sintaktikoa egin gabe dugunez, ezin dira daten barruko erroreak metodo horrekin tratatu. Antzeko gauza gertatzen da testuinguru lokalak aztertuz detekta daitekeen errore-multzoarekin, erroreak hitz edo sintagma txikien arteko korrespondentzien artekoak direlako, normalean gramatikan sartu gabe. Honen adibide

bat *elkar* hitzarekin egindako erroreak dira (adibidez, ‘**elkar bizi ziren*’). Fenomeno horiek guztiak gramatika barruan sartzeko, lan handia egin beharko litzateke.

Aspektu hauek kontuan hartuta, puntu honetan errore lokalen patroien bidezko tratamendua (Douglas 1992, Ramírez *et al.* 1997) aztertuko dugu (ikus adibide batzuk V.8 taulan). Hurbilpen hau da merkatuko zuzentzaile gramatikal gehienek erabiltzen dutena. Gainera, gure kasuan ideia hori lortutako analisi partzialen egituraren (*chart*) gainean aplikatu ahal izango dugu, analizatzaileak emandako osagaiak berrerabiliz (Mellish 1989, Min eta Wilson 1998). Hurbilpen honek gramatika oso baten beharraren arazoa saihesten du, erregela lokalen bidez testuinguru batzuk identifikatzen direlako ziurtasun handiz, testu errealak aztertzeko sendotasuna lortuz. Azpitik analizatzaile partziala edukitzeak asko lagunduko du, egitura sintaktiko osoak (izen-sintagmak, menpekoak) erabili ahal izango baitira errorearen testuinguruen deskribapenean.

Errorea	Errorearen kausa
<i>Donostia, 1995eko urtarrilak 15ean</i>	hila ergatiboan badoa egunak deklinatu gabe joan behar du
<i>Lagunak elkar bizi dira.</i>	<i>elkarrekin bizi</i> idatzi behar da
<i>Joseba mailuarekin baliatu zen</i>	<i>mailuaz</i> instrumentala erabiliko da <i>baliatu</i> aditzarekin
<i>Etxe guzti horiek egin ditut.</i>	<i>horiek guztiak</i> idatzi behar da

V.8 taula. Patroien bidez tratagarriak diren errorearen adibideak.

Erlaxazioaren metodoaren desabantaila bat testu errealak tratatzeko zailtasuna zenez, atal honetan egingo dugun lana testu errealetako errore lokalen aplikaziora zuzenduko dugu, horrela etorkizunerako proposatuko dugun sistema konbinatu baten lehen pausoak emanez.

Oraingo testu-prozesadore gehienek gramatika-zuzentzaile komertzial bat duten arren, lan gutxi argitaratu da emaitzen ebaluazioaz, arrazoi ezberdinengatik:

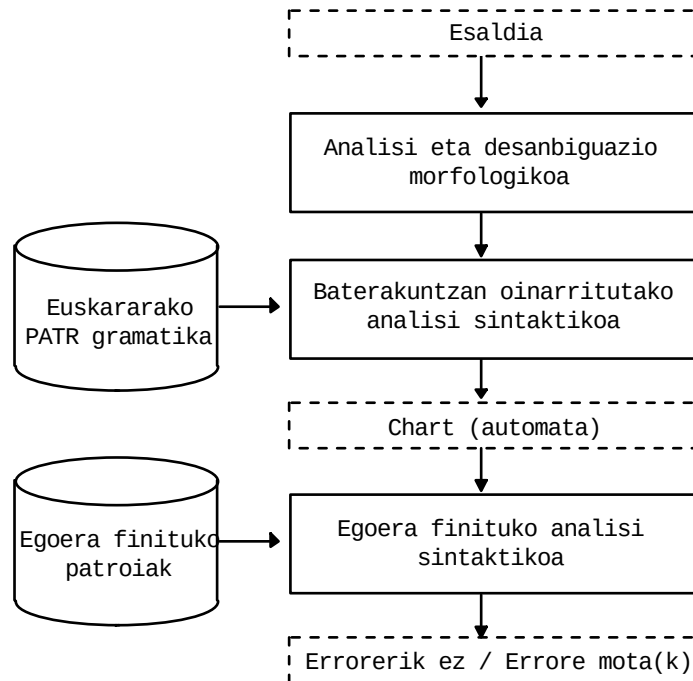
- Gramatiken estaldura partziala. Arazo hau esaldi zuzenekin gertatzen bada, zer esanik ez sintaktikoki okerrak diren esaldiekin, multzo ia infinitua osatzen dutelako. Beste alde batetik, sistema sendoek (analizatzaile estatistikoak, murriztapen-gramatika) askotan ezin dituzte bereiztu esaldi zuzenak eta okerrak. Maizen gertatzen diren errore-motak gramatikan sartu arren, oraindik errorearen kopuru osoaren portzentajea txikia izango litzateke. Beraz, esaldi baten analisisirik lortzen ez denean, askotan ezin da jakin errore bat edo gramatikaren hutsunea den.

- Alarma faltsuak. Erroreei buruz argitaratu diren lan gehienetan, errore sintaktikoak dagozkien esaldi zuzenen bidez definitu dira. Testu errealean erabilerak arazo desberdina dakar: esaldi zuzen bat okertzat ematea, errorearen testuinguruarekin antza duelako, nahiz eta askotan esaldi mota horrek erroredunekin inolako erlaziorik ez izan. Gertaera honek problemaren esparrua zabaltzen du: erroreak tratatzeko ez dira bakarrik esaldi okerren adibideak eta dagozkien zuzenketak aztertu behar, baizik eta corpus handi bateko esaldi guztiak.
- Corpus handien erabilera. Errore sintaktiko bakoitza maiztasun txikiarekin gertatzen da eta, ondorioz, probarako corpus handiak behar dira. Corpus erraldoi horiek eskuragarri egongo balira ere, errorearen agerpenak ezagutzea lan handia da, ez dagoelako sintaktikoki etiketatutako errorearen corpusik ebaluazio edo probarako. Horregatik, errorearen benetako adibideak aurkitzeko, milaka testu aztertu eta markatu beharko dira eskuz eta automatikoki.

Arrazoi hauengatik, puntu honetako helburua corpusetan oinarritutako errorearen detekzioaren bideragarritasuna aztertzea da, zuzenketa-tasa altua lortzeko alarma faltsuen kopuru baxua mantentzen den bitartean. Honelako tresnak beharrezkoak dira euskara bezalako hizkuntzentzat, estandarizazio-arazoak direla eta, beste arazoaren artean. Ondoren azalduko den esperimentua (Gojenola eta Oronoz 2000) artikuluan argitaratu dira.

V.2.3.2 Corpusetan oinarritutako patroien bidezko errorearen detekzioa

Errorearen tratamenduari ekiteko, azpikategorizazioari buruzko informazioarekin (IV. kapitulua) egin den modura, lehenengoz baterakuntzan oinarritutako analizatzailea aplikatuko da (hasieran analisi eta desanbiguazio morfologikoa eginda), eta ondoren egoera finituko patroiak definituko dira lortutako *chart*-aren gainean (V.5 irudia). Horrela, egoera finituko errorearen patroietan osagai sintaktiko konplexuak aipatu ahal izango dira, detekzio-ahalmena handituz, osagai lexikalen gaineko patroiak erabiltzen dituzten sistemekin konparatuz gero.



V.5 irudia. Errore-patroien bidezko detekzioarako sistemaren arkitektura.

Sistemaren proba egiteko, euskarazko daten adierazpenak aukeratu genituen, bi arrazoi nagusirengatik:

- Erraza zen probarako esaldiak aurkitzea, beste errore mota batzuekin konparatuz. Nahiz eta lan hau gehienbat eskuz egin behar, daten adierazpenek ezaugarri berezi batzuk dituzte (hilen izenak, urteen zenbakiak), adibideen bilaketa-prozesuan lagungarri direnak.
- Aplikazio-eremua zabala da, hau da, daten adierazpenek fenomeno aberatsak dituzte morfologian eta sintaxian, errore mota desberdinak egiteko era ematen dituztenak. Errore hauek errore sintaktiko lokalen multzokoen adierazgarritzat ikus ditzakegu. V.9 taulak datak adierazteko formatu erabilienak erakusten ditu (Euskaltzaindia). Ikusten denez, hasieran aukerazko tokia doa, absolutibo edo inesiboan, ondoren urtea genitiboan, eta bukatzeko hila eta eguna doaz. Hila genitiboan badoa, orduan ondorengo eguna deklinatuta joan behar da, eta hila ergatiboan badoa (*martxoak*), egunaren zenbakia deklinatu gabe agertuko da. Horregatik, datetan agertzen diren elementu gehienak deklinatuta doaz, bakoitza bere numero eta kasuaren morfemekin. Gainera, konbinazio batzuk bakarrik dira onargarriak, eta hori errorean iturburua da, errore gehienak zuzentzaile ortografikoentzat detektaezinak direlako, hitzez hitz hartuta zuzenak baitira.

Durango, 1999ko martxoaren 7a
Durango, 1999ko martxoaren 7an
Durango, 1999ko martxoaren 7tik ...
1999ko martxoaren 7an
Durango, 1999ko martxoak 7

V.9 taula. Datak adierazteko formatu zuzenak.

Ikasleen 267 testu bildu genituen (errore maiztasun handia zuten), egunkaririk eta aldizkariarik lortuak, guztira 500.000 hitz baino gehiago. Bertatik 658 esaldi hartu genituen (ikus V.10 taula), barruan data zuzenak, okerrak eta daten antzeko egiturak zeuzkaten esaldiak (azken hauek alarma faltsuak detektatzeko). Esaldiak multzo bitan banatu ziren, batean errorearen tratamendurako sistemaren garapenean erabiliko zirenak gordez, eta bigarrean azken probarakoak utziz (hau da, azken proban aztertu gabeko esaldiak prozesatzeko).

	Garapenerako corpusa		Probarako corpusa	
Esaldi-kopurua	411		247	
Data zuzenak	65		39	
Daten antzeko esaldiak	255		171	
Data okerrak	91		37	
Data okerrak errore batekin	43	% 47	6	% 16
Data okerrak errore birekin	42	% 46	27	% 73
Data okerrak hiru errorekin	6	% 7	4	% 11

V.10 taula. Probarako esaldiak.

Garapenerako corpusean aurkitutako errore guztiak aztertu eta gero, maiztasun handieneko erroreak aukeratu genituen (ikus V.11 taula). Errore mota bakoitza detektatzeko patroia bat edo gehiago definitu ziren. Dena dela, laster ikusi genuen hau ez zela aukera onena, askotan bi edo hiru errore agertzen direlako adierazpen berean. Fenomeno honek patroien gainean erlaxazio gradualaren antzeko hurbilpena eskatzen du, gogoan izateko errore asko batera ager daitezkeela. Errore bakoitza modu independentean tratatu ordez, patroiak egiteko orduan adierazpen zuzena eta bere bertsio okerrak izan behar ziren gogoan. Bestalde, data zuzena kontsidera daitekeen ereduaren erlaxazioak alarma faltsuen kopurua gehitzeko arriskua dauka. Esan behar da ere *erlaxazio* deitzen dugun horrek aurreko atalean (§ V.2.2) egin dugunaren antza duela, kasu bietan esaldi zuzen baten esparrua zabaldu egiten delako, aldaera okerrak onartuz, baina ezberdintasun nagusia da orain patroia bakoitzeko erlaxazio hori eskuz aplikatu

behar dela, eta lehengo ataleko inplementazioan, aldiz, behin askatzen diren ekuazioak zehaztu eta gero, automatikoki egiten zela. Erroreen arteko elkarrekintza modu esponentzialean hasten denez errore-kopuruarekin batera, gure errore-patroietan corpusean aurkitutako konbinazioak bakarrik landu genituen.

	Errorea	Errorearen kausa
1	<i>Donostian, 1995-eko martxoaren 22an</i>	Urtea zenbakiz idazten denean, ezin da marra jarri
2	<i>Donostian, 1995eko Martxoaren 22an</i>	Hila ezin da maiuskulaz jarri
3	<i>Donostian 1995eko martxoaren 22an</i>	Tokia idazten bada, ondoren koma jarri behar da
4	<i>Donostian, 1995eko martxoaren 22</i>	Hila genitiboan badoa eguna deklinatua egongo da
5	<i>Donostia, 1995eko urtarrilak 15ean</i>	Hila absolutiboan badoa eguna deklinatu gabe joan behar da
6	<i>Donostia, 1995eko martxoan 15ean</i>	Hila absolutibo edo genitiboan agertu behar da
	<i>karrera bukatu nuenean 1997ko Ekainaren 30</i>	2, 3, eta 4 erroreen konbinazioa

V.11 taula. Tratatu diren erroreak.

Bosgarren errore mota detektatzeko patroiak V.3 adibidean agertzen dira. Hasteko, errorearen testuingurua definitzen da (`ErroreMartxoak22an`), hau da, hila ergatibo edo absolutiboan (azken hau desanbiguazioaren balizko akatsak saihesteko) zenbaki deklinatu baten aurretik. Ondoren, transduktore batek (`MarkatuMartxoak22an`), errorearen hasieran eta bukaeran etiketa bi (`+ERRHAS5` eta `+ERRBUK5`) jarriko ditu. Erregelaren aplikazioa gehiago murrizteko, ezkerreko eta eskuineko testuinguruak zehaztuko dira, bi mailatako morfologiaren antzeko notazioa erabiliz, horrela alarma faltsuen agerpena gutxitzeko, baina aldi berean beste errore mota batzuk onartuz. Adibidean hori osagai ezberdinen bidez egiten da: `Urtea`-k urte zuzena eta okerrak onartuko ditu, `HilaAbsErg`-ek hila maiuskulaz idaztea ere onartuko du, eta `AukerazkoKoma`-k tokiaren atzetik koma ez jartzeko aukera emango du. Horregatik errore bi edo hiru dauden testuinguruetan ere bosgarren errorea detektatuko da (koma faltan eta eguna deklinatua: *Donostian 1999ko martxoak 13an*, edo hila maiuskulaz idazten bada).

```
define HilaAbsErg Hila & [\$["+kas" "+erg"] | \$["+kas" "+abs"]];
define ErroreMartxoak22an HilaAbsErg HitzMuga ZenbakiIzeDek; # martxoak 22an
define UrteaZuzena OsagaiSintaktikoa &
    \$["+kat" "+izlg"] &
    \$["+lema" ZenbakiIze];
define Urtea [UrteaZuzena | ErroreaMarraUrtean];
define MarkatuMartxoak22an [ErroreMartxoak22an] @-> %+ERRHAS5 ... %+ERRBUK5
    || AukerazkoTokia AukerazkoKoma Urtea HitzMuga
    - ;
```

V.3 adibidea. Bosgarren errorea (1999ko martxoak 12an) detektatzeko erregelak.

Errore-patroiak definitzeko orduan, arreta berdina jarri behar da data zuzen eta okerretan. Ondorioz, 60 patroik morfosintaktiko inguru definitu behar izan dira (bakoitza automata edo transduktore batez konpilatuta) sei errore motei dagozkien patroien definiziorako. Hauek murriztapen lokal txikienetatik hasita (45 automatek 100 egoera baino gutxiago daukate) patroik konplexuenetaraino doaz (transduktore bat 10.000 egoera eta 475.000 arkurekin).

V.12 taulak lortutako emaitzak erakusten ditu. Garapenerako corpusa egoera finituko erregelen prestaketan aztertu ahal zenez, bigarren eta hirugarren zutabeko emaitzak analizatzailearen (oraingo egoeran) kasurik hoberenaren adierazgarri kontsidera daitezke, %100eko doitasuna (alarma faltsurik ez) eta %91ren estaldurarekin. Aurretik ikusi gabeko 247 esaldien corpusaren gainean %84ko estaldura lortzen da. Doitasuna begiratzen badugu, 5 alarma faltsu daude, hau da, data zuzenak edo datak ez diren egiturak baina data okertzat hartzen direnak. Alarma faltsuak probarako corpuseko 247 esaldiekin zatitzen badira, alarma faltsuen tasaren estimazioa %2,02koa izango litzateke.

	Garapenerako corpusa		Probarako corpusa	
	0			
Esaldi-kopurua	411		247	
Detektatu ez diren data okerrak	7	9%	6	16%
Detektatutako data okerrak	84	91%	31	84%
Alarma faltsuak	0		5	

V.12 taula. Ebaluazioaren emaitzak.

V.13 taulak alarma faltsuak aurkezten ditu. Ikusten denez, horietako bi datak ez direnak baina datatzat hartu diren adierazpenak dira; beste bi egindako daten gramatikan kontuan hartu ez diren egiturak dira; eta azkena lexikoiaren hutsune baten ondorioz dator (*Primakov* pertsona-izen bereziaren interpretazioa eman ordez, hitz ezezaguntzat hartzen da). Emaitza onak diren arren, oraindik corpus handiagoak beharko genituzke, doitasuna hobetzeko.

Adibidea	Alarma faltsuaren kausa
<i>atxiloketa 1998ko urtarriletik irailaren 16ra ...</i>	Datatzat hartutako data osoa ez den egitura.
<i>Donostian 1960ko Urtarrilaren jaioa</i>	Okerra den esaldia datatzat hartu da.
<i>etorriko da 1997ko irailaren 26ko 1:15etan</i>	Ordua (1:15) hilaren eguntzat ulertu da.
<i>atzotik 1999ko abenduaren 31 arte</i>	Gramatikak ez du <i>arte</i> partikula tratatzen, eta horregatik data zuzena okertzat ematen da.
<i>Primakovek 1998ko irailaren 11n hartu zuen ...</i>	<i>Primakov</i> hitz ezezaguna lokatibotzat interpretatzen da.

V.13 taula. Sistemak emandako alarma faltsuak.

V.2.3.3 Ondorioak

Puntu honetan analizatzaile sintaktikoa errore-patroien bidezko detekzioan aplikatu dugu. Hauek dira egindako lanaren ezaugarri aipagarrienak:

- Corpusetan oinarritutakoa. Sistema erabilgarria egiteko, esaldi zuzen zein okerre benetako adibideen gainean probatu behar da. Hau begi-bistakoa dela eman arren, ez da orain arte errekeen detekzioan egin den lan gehien ezaugarria izan. Hau egiteko corpus handiak behar dira, neurri handi batean eskuzko markaketa-lana eginez.
- Errekeen detekzioarako metodo erabilienak konbinatu dira (erreke-patroiak, *chart*-aren berrerabilera eta erlaxazioaren aldaera bat) emaitza onak lortuz. Horrela, lengoaiari buruzko ezagumendu positiboa (gramatikaren erregeletan adierazitakoa: zer dagoen ondo) eta negatiboa konbinatzen dira (erreke-patroien debekuak: zer dagoen txarto (Heaton eta Turton 1987)), lengoaien irakaskuntzan egiten den antzera. Beste alde batetik, erlaxazioaren eskuzko idazketak lan linguistiko handia suposatzen du.
- Baterakuntzan oinarritutako eta egoera finituko analizatzaile sintaktikoen erabilera sekuentziala egokia da errekeen patroien tratamendurako. Baterakuntzan oinarritutako analizatzaileak osagai sintaktiko nagusiak aurkitzen ditu, eta horien gainean patroia konplexuak defini daitezke errekeak harrapatzeko.

Hauek aurreikusten ditugu izango direla sistema hau zabaltzeko hurrengo pausoak:

- Erreke mota gehiagoren azterketa. V.7 taulan, datez aparte, beste erreke sintaktiko lokal batzuk agertzen dira, metodo honen bidez tratagarriak izan daitezkeenak. Gure sistemak, errekeen tratamenduan ari diren sistema guztiak bezala, eskalatze-arazoa dauka, erreke-kopuruen gehikuntzak erreke-patroiak kodetzeko eskuzko lana eskatzen baitu. Erreke mota gehiago aztertzen direnean, erregelen arteko

elkarrekintza kontu handiz aztertu beharko den arren, espero dugu errore ezberdinetako erregelak independenteak izango direla neurri handi batean. Beste puntu interesgarri bat baliabide linguistikoaren berrerabilpena da: daten erroreak tratatzeko orduan, zenbait patroik arau linguistiko orokorrak definitzen dituzte, eta beste batzuk datentzako bereziak dira.

Horrez gain, patroien bidezko tratamendua lagungarria izan daiteke komuntadurak eta antzeko erroreak tratatzeko ere. § V.2.2n erlaxazioan oinarritutako metodoa aztertu dugu, bere arazo handienetako bat testu errealekara aplikatzeko zailtasuna zela. Bi metodoen konbinaziorako hipotesia aztertzen ari gara: patroien bidez azpiesaldi edo sintagmen testuinguruak aukeratuko dira, beren barruan erlaxazioa aplikatuz (Oliva 1997). Horrela, testu errealean tratamendua egingarria izan daiteke.

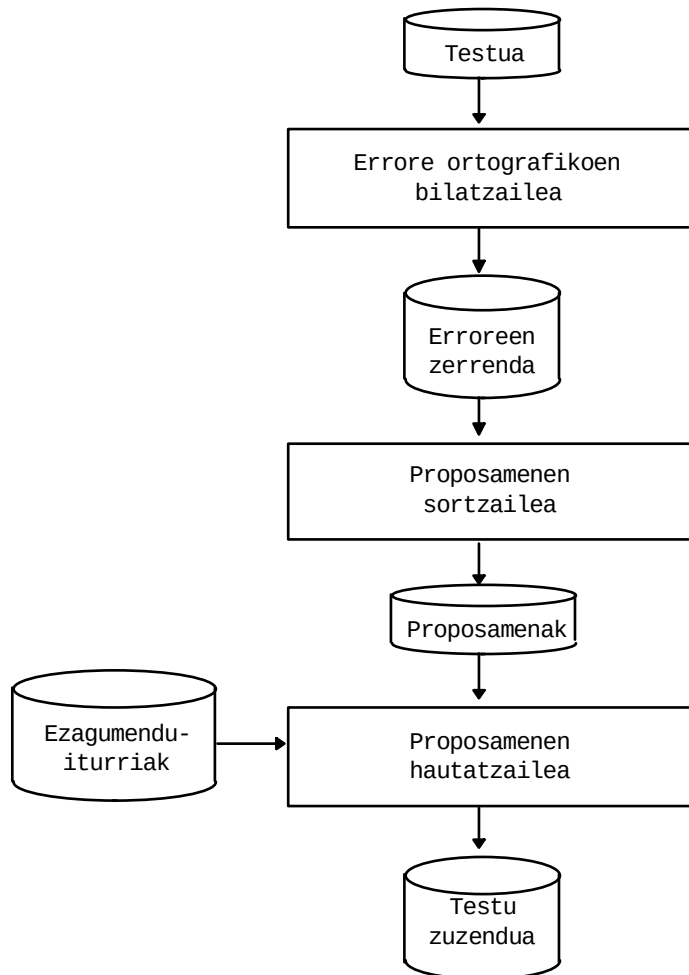
- Erroreak detektatzeko erregelen ikasketa automatikoa. Behin erroreak markatuta dauzkan corpora edukita, gure kasuan eskuz landu ditugu detekzio edo zuzenketarako erregelak. Horregatik, ikasketarako metodoen azterketa oso interesgarri ikusten dugu (Golding eta Roth 1996, Mangu eta Brill 1997), sistema garatzeko lan bakarra errorearen adibideak markatzea izango litzatekeelako.
- Corpus handien bilketa. Errorearen benetako agerpenak izateko, corpus handiak beharko dira etorkizunean. Internet-en gorakada dela eta, lan hau erraz daiteke, baina gogoratu behar dugu corpus horiek lortu ondoren oraindik errorearen eskuzko markaketa egongo dela.

V.2.4 Errore ortografikoen zuzenketa

V.2.4.1 Sarrera

§ V.2.2n eta § V.2.3n errore sintaktikoen detektziorako bi hurbilpen ezberdin aztertu ditugu. Lehenago, errore ortografikoak aipatu ditugunean esan dugu detektatu ondoren zuzenketa geratzen dela, proposamen bat baino gehiago egoten delako askotan, eta horien artean zuzena dena aukeratzeko ezagumendu sintaktiko eta semantikoa behar dela. Errore ortografikoekin gertatzen den bezala, metodoak definitu behar dira zuzentzaile sintaktiko baten zuzenketa-proposamenak baztertzeko, eta horregatik puntu honetan lan horri ekingo diogu. Eredu zabal hau murriztearren, gure kasuan zuzentzaile ortografikoen proposamenak diskriminatzea mugatuko dugu lan hau, eta horren arrazoi nagusia esperimentaziorako datuak lortzeko erraztasuna da, errorearen agerpenak ateratzeko nahikoa izango delako eskuragarri dagoen zuzentzaile bat testu bati pasatzea. Gainera, proposamenen diskriminazioa egiteko

ezagumendu-iturri ezberdinen ahalmena neurtu nahi dugu (ikus V.6 irudia). Ingelesa izan da linguistika konputazionalen gehien landu den hizkuntza, eta horregatik tresna ugari daude eskuragarri bere tratamendurako. Arrazoi honengatik, ingelesa aukeratuko dugu esperimentaziorako hizkuntzatzat, euskararentzat baino tresna gehiago garatu direlako eta probarako eskuragarri daudelako.



V.6 irudia. Erroreen zuzenketarako proposatutako sistemaren eskema.

Erroreen zuzenketa automatikoaren problema oraindik ikerlerroa da, erabilgarri dauden teknikak mugatuak daudelako estaldura eta doitasunean. Zuzentzaile ortografiko gehienek, erroreak detektatzeaz gain zuzenketarako kandidatuak diren hitzen multzoa aurkezten diote erabiltzaileari. Dena dela, kasu askotan beharrezkoa da zuzenketa guztiz automatikoa egitea (OCR programetan, edo denbora errealean lan egin behar duten ahotsaren tratamendurako edo itzulpen automatikorako sistemetan).

Testuinguruaren arabera errore ortografikoen zuzenketa, azken finean, proposamen bakoitza ordezkatzuz lortzen diren sententzien arteko aukeraketa egitea da (Mays *et al.* 1991).

Arazoa horretarako behar diren ezagumendu motak (lexikoa, fonologikoa, sintaktikoa, semantikoa edo estatistikoa) adierazi, erabili eta konbinatzea da.

Esperimentuak egiteko, ingeleserako Ispell zuzentzaile ortografikoa aukeratu dugu. Bere ezaugarri nagusien artean estaldura zabala, fidagarritasuna eta malgutasuna aipatuko ditugu. Proposamenak aukeratzeko, murritzapen-gramatikaren formalismoa aukeratu dugu ezagutza sintaktikoaren ordezkari izateko, bere desanbiguaziorako ahalmenagatik (ikus § III.3.1). Dentsitate kontzeptuala (Agirre eta Rigau 1996) izenen bereizketa semantikoa egiteko erabiliko da, izenen arteko distantzia neurtuko baitu Wordnet (Miller 1990) erabiliz. Testuinguruarekiko lotura semantikoa corpusetatik ere atera da, kokakidetza eta agerkidetzen ezaugarri estatistikoaren bidez (Yarowsky 1994). Azkenik, hitzen maiztasunak ere tratatuko dira, dokumentukoak zein corpus orokorretakoak, sistema konbinatuko ezagumendu-iturriak osatzeko. Lan honen beste helburu bat lengoaiaren eredu desberdinen azterketa egitea izan da, errorearen gain desanbiguaziorako beste aplikazioetan tratatu ahal izateko. Horregatik, nahita ez genuen errore-eredurik egin (teklatze-erroreak, errore fonologikoak, OCR erroreak, ...), jakinda horiek gehituz gero zuzenketa-tasa hobetu ahal izango dela. Sistemaren ebaluaziorako bi testu-sorta erabili dira: Brown corpusetik (Francis eta Kucera 1967) programa baten bidez sortutako errore artifizialak eta Bank of English³⁶ corpusetik ateratako benetako erroreak. Hemen aurkeztuko ditugun esperimentuak (Agirre *et al.* 1998ab) lanetan argitaratuta daude.

V.2.4.2 Errore ortografikoen zuzenketa automatikoa

Ingelesko errore ortografikoen zuzenketa automatikoan egin dugun lana aurkezteko, lehenengo erabilitako tekniken deskribapen laburra emango dugu, ondoren egindako esperimentuak eta emaitzak azaltzeko.

V.2.4.2.1 Erabilitako teknikak

Erabilitako tekniken azalpen laburra emateko, hau esango dugu:

- Murritzapen-gramatika (MG). Formalismoa bereziki prestatuta dago anbiguitasuna ebazteko, eta horregatik egokia da proposamenen diskriminaziorako. Gure aplikazio honetan, ezagutzen ez den hitz bakoitzeko Ispell-ek sortutako proposamenen analisi morfologikoak (ENGTWOL; Karlsson 1995) jarri genituen. V.4 adibidean *bos* hitz

³⁶ http://titania.cobuild.collins.co.uk/boe_info.html

okerraren *bop* zuzenketa kentzen du, testuinguru horretan aditzaren interpretazioa ez delako zuzena.

Ispell-en eta MGren estaldura zabaleko lexikoiak independenteak direnez, batean eta bestean oker edo ezezaguntzat emandako hitzak desberdinak ziren. Arazo hori saihesteko, hitz bat lexikoi bietako batean zegoenean zuzentzat eman genuen.

<pre>"<our>" "our" PRON PL ... "<bos>" ; ERRORE ORTOGRAFIKOA "boss" N S "boys" N P "bop" V S "Bose" <Proper> "<are>" ...</pre>	<pre>"<our>" "our" PRON PL ... "<bos>" ; ERRORE ORTOGRAFIKOA "boss" N S "boys" N P "bop" V S "Bose" <Proper> "<are>" ...</pre>
--	--

V.4 adibidea. Murritzapen-gramatikaren aplikazioak aditzaren interpretazioa kentzen du.

- Dentsitate kontzeptuala (DK). MGk askotan ezin du bereizketarik egin kategoria bereko proposamen bat baino gehiago dagoenean, ezagutza sintaktikoa erabiltzen duelako. Horregatik Wordnet-en oinarritutako hitzen adiera-desanbiguatzailea erabili genuen (Agirre eta Rigau 1996). Gure kasuan, desanbiguatzaileak testuingurutik semantikoki hurbilen dagoen proposamena aukeratu beharko luke. Wordnet-en orduko egoera zela eta, neurri hau proposamen guztiak izenak direnean bakarrik aplikatu ahal izan da.
- Maiztasunei buruzko estatistikak (DM eta BM). Hitz-formen maiztasunak errorea agertu den dokumentutik hartu ziren (Dokumentuko maiztasuna, DM) edo Brown corpuseko dokumentuetatik (Brown-eko maiztasuna, BM).
- Testuinguruaren estatistikak (TS). Yarowsky-k (1994) emandako ideiak jarraituz, proposamenek inguruko hitzekin duten hurbiltasuna neurtzeko metodoak erabili genituen. Testuinguruari buruzko ezaugarriak Brown corpusetik lortu ziren (esperimentu honen probarako testuak kenduta). Hartutako ezaugarriak hauek ziren: hitz-formen bigramak, hitz-formen trigramak eta ± 20 hitzeko leihoko hitzak.
- Beste heuristikokoak (H1 eta H2). Maiuskulaz hasitako proposamenak baztertu egin ziren hitz okerra minuskulaz hasi eta minuskulaz hasitako beste proposamenak zeudenean. V.4 adibidean *bos* hitzaren laugarren irakurketa (*Bose*) ken liteke, hasierako hitzarekin duen distantzia bikoia delako (H1 heuristikoa). Heuristiko hau

oso fidagarria izan da³⁷, eta horregatik ondoren azalduko diren esperimentu guztietan aplikatuko da.

Lehen emaitzak atera eta gero, 4 karaktere baino gutxiagoko erroreek (*si, teh*) proposamen gehiegi (askotan 30 proposamen baino gehiago) ematen zutela ikusi genuen, desanbiguatze oso zailak direnak. Hitz hauek gure metodoaren errore-iturri nagusietako bat zirenez, emaitzak hitz hauek sartuz eta sartu gabe lortu genituen (H2 heuristikoa).

- Oinarritzko tekniken konbinazioa botoak erabiliz. Teknika guztien konbinazioak izan genituen gogoan, adibidez: MG+BM, BM+DM, MG+DM+TS, ... Botoetan oinarritutako metodoak boto gehien jasotzen duen proposamena hobetsiko du, botoaren balioa zenbaki osokoa izanik. Metodo honen abantaila handia sinpletasuna da, eta beste teknika sofistikatuagoek (pisuen zenbaki zatiki onenak lortzeko optimizazio-algoritmoak erabiliz) bezain emaitza onak lortzen ditu.

Teknika bakoitzaren botoaren pisua alda daiteke, adibidez, MGri bi boto eta BMri boto bat emanez (konbinazio hau MG2+BM1 izendatuko dugu). Honek esan nahi du BMren kandidatua(k) bakarrik hautatuko dela MGk beste aukerarik ez badu hartzen. Pisuen konbinazio ezberdinak probatu ziren.

Testu-sorta batentzako teknika eta pisuen konbinazio onenak alda daitezkeenez, errorearen corpusa bitan zatitu genuen, lehenengo erdian konbinazio guztiekin saiaturaz, eta konbinazio onenen proba bigarren erdian eginez (ikus geroago proba-kasuen antolaketa).

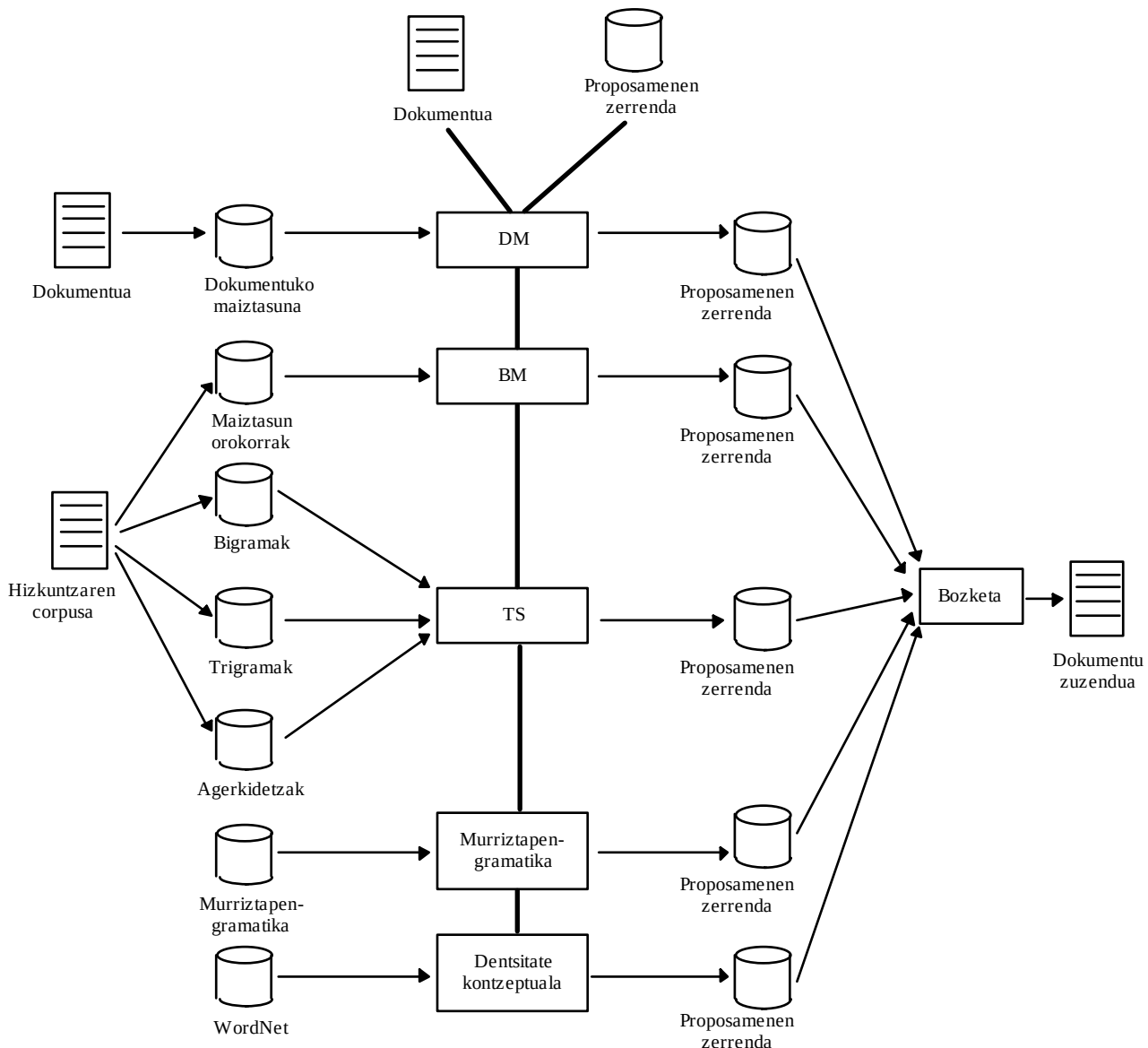
V.2.4.2.2 Esperimentuak

Ezagumendu mota bakoitza erabiliz, proposamen bat aukeratzeko zuzentzaile bana osatu genuen, eta era ezberdinetan konbinatu genituen (ikus V.6 irudia, Agirre-ren tesitik (1999) hartuta). Lehenbiziko fase batean aukera guztiak ebaluatu eta onenak hartu ziren ausaz sortutako errorearen corpusaren zati batean. Bukatzeko, konbinazio onenak benetako erroreetako testuetan probatu ziren.

Lehenago esan bezala, bi corpus erabili ziren esperimenturako. Lehenengoa erroreak ausaz modu sistematikoan sortuz atera zen Brown corpusaren zati batetik, eta bigarrena benetako erroreak zeukan corpus batetik. Lehenengo zatia egokia zen esperimentaziorako, erroreak

³⁷ Ispell-en estaldura handiko hiztegian izen propioen zerrenda luzea dagoenez, askotan minuskulaz hasitako erroreentzat izen propioen interpretazioak ematen dira, gehienetan zuzenak ez direnak.

automatikoki sortuak zirenez egiaztapena zuzenean egin zitekeelako. Bigarrenak, aldiz, erroreen zuzenketaren eskuzko markaketa eskatzen zuen, baina beste alde batetik egoera erreala islatzen zuen (Brown corpusean, bera sortzeko modua dela eta, ez dago ia errore ortografikorik, zuzendua izan baita).



V.6 irudia. Proposamenaren hautapenerako ezagutza-iturriak eta konbinazioa egiteko sistema.

Modu artifizialean sortutako erroreen corpusak (corpus artifiziala deituko diogu laburtzearen) honako ezaugarriak dauzka: SemCor-etik (Brown corpusaren azpimultzoa) lagin bat atera zen, ausaz 150 paragrafo hartuz. Honek 5050 esaldiko eta 12.659 token-eko hasierako corpusa ematen du. Errore ortografikoen simulazioa egiteko antispell izeneko programa bat exekutatu zen, Damerau-ren erregelak aplikatuz 20 hitzeko errore bat sortzeko (hitz ezezagunak ukitu gabe). Antispell 8 aldiz exekutatu zen hasierako corpus horren

gainean, testu bera baina errore ezberdinak dituzten 8 corpus sortuz. Esaldi batean errore bat baino gehiago egon liteke, eta paragrafo batzuetan ez zegoen errorerik.

Benetako errorearen corpusa (hemendik aurrera corpus erreala deituko diogu) Bank of English Corpus-eko aldizkariaren testuez osatua dago, lehenago zuzenketa ortografikorik egin gabea. Honelako corpusak lortzeak zailtasuna du, egun zuzentzaile ortografikoen erabilera zabalduta baita. Beste arazo bat errorearen markaketan eta zuzenketaren aukeran egin behar den lan handia da, ebaluazio automatikoa egin ahal izateko gero. Lehen aipatu dugu errore artifizialen corpusa bi azpimultzotan banatu zela. Lehenbizikoa, 1, 2, 3 eta 4 multzoez osatua, metodoa fintzeko (botoak eta konbinazio onenak ateratzeko) erabili zen. Bigarren azpimultzoa eta testu erreala azken ebaluazioa egiteko erabili ziren.

Corpus biak Ispell-etik pasatu ziren, hitz ezezagun bakoitzeko bere zuzenketarako proposamenak txertatuz. V.13 taulan ikusten da errore ortografikoak ausaz sortuz gero %23,5a benetako hitzak direla, eta lan honetatik kanpo geratuko dira. Testu errealean antzeko kontaketa egin ez genuen arren, gaineko azterketa eginda antzeko portzentajea gertatzea espero dezakegu.

	1. erdia	2. erdia	erreala
hitza	47584	47584	39733
erroreak	1772	1811	- ³⁸
ez-hitza diren erroreak	1354	1403	369
Ispell-en proposamenak	7242	8083	1257
proposamen bat baino gehiagoko erroreak	810	852	158
hitz luzeen erroreak (H2)	968	980	331
hitz luzeentzako proposamenak (H2)	2245	2313	807
hitz luzeak (H2) proposamen askorekin	430	425	124

V.13 taula. Esperimenturako datuak corpus artifizialean (2 zati) eta errealean.

Benetako errorearen corpusarekin jarraitutako prozedura hau izan zen: Ispell aplikatu ondoren ez zen zuzenketarik eman 150 hitzentzat (gehienak izen propioak eta erdarakadak), eta 300 inguru zeuden jarraian doazen bi hitz lotuz edo atzizkien lotura bereziak aplikatuz lortuak (Ispell-ek ondo ezagutu zituen gehienetan). Horiek kenduta, 369 hitz-forma oker geratu ziren. Proposamenak aztertu ondoren ikusi genuen proposamen zuzena Ispell-ek emandakoen artean zegoela ia kasu guztietan.

Hitz okarren proposamenen hautapena aztertuz, ia erdiek proposamen bakarra dute. Proposamen bat baino gehiagokoek emango dute egin behar den lanaren neurria. Adibidez,

corpus errealean 158 hitz-forma daude 1.046 proposamenekin, hitzeko 6,62 aukerako batez bestekoa emanez. Hiru letra edo gutxiagoko hitzak kontatzen ez badira, orduan 807 proposamen daude, hau da, 4,84 aukera hitzeko.

Hiru balio kontsideratu ditugu emaitzak ebaluatzerakoan:

- Estaldura: erroreetariko zenbatetan teknikak (edo tekniken konbinazioak) erantzuna ematen duen.
- Doitasuna: erroreetariko zenbatetan uzten duen proposamen zuzena hautatutako proposamenen artean.
- Errore bakoitzeko utzitako batez besteko proposamen-kopurua. Hau da, askotan doitasuna eta estaldura elkarren kontrakoak direnez, batzuetan teknika batek proposamen bat baino gehiago aukeratzen du, horrela doitasun handiagoa lortuz.

Konbinazio hoberenak aurkitzeko, garapenerako corpusean (errore artifizialen lehen erdia) lortutako emaitzak aztertu ziren. Konparazioa errazago egiteko, proposamenaren ausazko aukeraketaren emaitza ere atera zen. H1 heuristikoa kasu guztietan aplikatu zen, eta esperimentuak H2 erabiliz eta erabili gabe egin ziren. Bigarren fase batean, konbinazio hoberenak corpus artifizialaren bigarren erdiaren gainean ebaluatu ziren, emaitzak lehenengo zatia lortutakoen parekoak zirela, portzentajeak pixka bat jaitsi arren. Proba honekin konbinazio hoberenak errore ezberdinetan mantentzen direla baieztatzen zen.

³⁸ Hitz ezezagunak aztertu genituenez, ez ziren kontatu benetako hitzak ematen dituzten erroreak.

	estaldura %	doitasuna %	# proposamenak
Oinarrizko teknikak			
ausazko aukeraketa	100.00	69.92	1.00
ausaz+H2	89.70	75.47	1.00
MG	99.19	84.15	1.61
MG+H2	89.43	90.30	1.57
DM	70.19	93.05	1.02
DM+H2	61.52	97.80	1.00
BM	98.37	80.99	1.00
BM+H2	88.08	85.54	1.00
TS	97.02	89.10	1.02
TS+H2	85.64	91.50	1.01
Konbinazioak			
MG1+DM2	100.00	87.26	1.42
MG1+DM2+H2	89.70	90.94	1.43
MG1+DM1+BM1	100.00	80.76	1.02
MG1+DM1+BM1+H2	89.70	84.89	1.02
MG1+DM1+TS1	100.00	90.80	1.24
MG1+DM1+TS1+H2	89.70	93.10	1.20
MG1+DM1+TS2	100.00	89.70	1.04
MG1+DM1+TS2+H2	89.70	91.80	1.03

V.14 taula. Konbinazio hoberenak (testu errealeko erroreak).

Azken pausoa konbinazio hoberenak benetako errore ortografikoen corpusean ebaluatu genituen. Arreta proposamen bat baino gehiagoko kasuetan jarritz gero, hobeto ikus daiteke teknika bakoitzaren ekarpena. V.14 taulak emaitza orokorrak azaltzen ditu, eta V.15n proposamen anitzeko hitzenak (emaitza hoberenak grisez agertzen dira). Azken honetan, hitzeko 6,62 proposamen daude kasu orokorrean (corpus artifizialean baino 2 gutxiago), eta 4,84 H2 heuristikoa aplikatuz (corpus artifizialean baino 1 gehiago). Hurrengo lerroetan emaitzen balorazioa emango dugu, testu errealekin hasi eta ondoren corpus artifizialarekin konparatuz.

	estaldura %	doitasuna %	# proposamenak
Oinarrizko teknikak			
ausazko aukeraketa	100.00	29.75	1.00
ausaz+H2	76.54	34.52	1.00
MG	98.10	62.58	2.45
MG+H2	75.93	73.98	2.52
DM	30.38	62.50	1.13
DM+H2	12.35	75.00	1.05
BM	96.20	54.61	1.00
BM+H2	72.84	60.17	1.00
TS	93.21	74.16	1.05
TS+H2	67.28	75.36	1.03
Konbinazioak			
MG1+DM2	100.00	70.25	1.99
MG1+DM2+H2	76.24	75.81	2.15
MG1+DM1+BM1	100.00	55.06	1.04
MG1+DM1+BM1+H2	76.54	59.68	1.05
MG1+DM1+TS1	100.00	78.51	1.56
MG1+DM1+TS1+H2	76.54	81.58	1.53
MG1+DM1+TS2	100.00	75.94	1.09
MG1+DM1+TS2+H2	76.54	78.11	1.08

V.15 taula. Proposamen bat baino gehiagoko errorearen emaitzak (testu errealeko erroreak).

Teknika bakoitza modu independentean aplikatuz³⁹:

- Teknika guztiak dira ausazko aukeraketa baino hobeak.
- TSk dauka doitasun altuena (%74), %93ko estaldurarekin.
- BMk doitasun txikiagoa (%54) baina estaldura zabala du (%96).
- DMk doitasun txikiagoa (%62) eta estaldura txikiagoa (%30) ditu.
- MGk %62ko doitasuna eta ia %100eko estaldura du, baina proposamen asko mantentzen ditu (2,45).

Teknikak konbinatzerakoan, emaitzak hobetu egiten dira:

- MG+DM+TS konbinazioak emaitza onenak ematen ditu estalduran (%100etik hurbil) eta doitasunean (%78a V.15 taulan).
- TSk pisu bikoitza hartzen badu, MG1+DM1+TS2 konbinazioan, doitasuna gutxi jaisten da (%76), baina proposamen-kopurua onargarriagoa da (1,56tik 1,09ra).
- %70eko doitasuna lortzen da MG1+DM2 probatuz. Ikusten denez, MGk DM teknikaren estaldura igotzen du, proposamen-kopuruaren igoerarekin batera (1,9).

³⁹ Besterik ez bada esaten, aipamenak testu errealei buruz dira (VI taula).

DMren estaldura handiagoa izango balitz, proposamen-kopurua neurri berean jaitea espero genezake, corpus artifizialen parera iritsiz (1,28).

- MG1+DM1+BM1 konbinazioak antzeko estaldura ematen du hitz bakoitzeko ia interpretazio bakarra utziz, baina doitasuna %55era jaisten da.
- Estaldura osoa beharrezkoa ez bada, H2 heuristikoaren erabilerak doitasuna %3a igotzen du konbinazio guztietarako.

Argigarriagoa da corpus artifizialeko eta V.15 taularen arteko konparazioa. Azken taula honetan teknika guztien performantzia jaisten da. MG, TS eta BM 15, 10 eta 20 puntu jaisten dira, hurrenez hurren. DMk doitasunean 20 puntu eta estalduran 50 puntuko jaitiera du. Azken honen performantziaren okertzea espero zitekeen, benetako errorearen corpuseko dokumentuen luzera 50 hitzekoa delako batez bestekoan. Hau corpusaren ezaugarri berezitat jo behar dugu, ez baikenuen testu luzeagoak lortzerik izan, eta neurri handiagoko dokumentuak izatera, testu artifizialen antzeko estaldura espero genezake.

Konbinazio onenak mantendu egiten dira testu errealetarako. Doitasun altuena MG+DM+TS konbinazioarentzat da (H2rekin eta H2 gabe). Proposamen-kopurua handiagoa da testu errealetan (1,56 eta 1,12). Honen kausa bat DM eta TS errore guztiak ez estaltzea izan liteke, eta ondorioz MGren proposamen asko ez dira ukitzen.

Testu errealean performantziaren jaitiera azaltzeko faktore ezberdinak aipatuko ditugu. Lehenengo, dokumentuen luzera txikiagoak eragin handia izan du DMn. Bigarren, errorearen sorrera ere ezberdina da: errore ortografikoak sortzeko erabili dugun algoritmoak (antispell) gehiago hartzen ditu maiz agertzen diren hitzak, askotan hitz laburrak. Eredu hau benetako erroreekin zein puntutaraino bat datorren aztertu beharko dugu.

V.14 taularen emaitzak hartuz, zuzenketa automatikorako sistema baten irteera zein den aurreikusi dezakegu. MG1+DM1+TS2 konbinazioarekin testu bateko errore guztientzako 25etik 24 aldiz proposamen bakarra lor daiteke %90eko doitasunarekin (1,04 proposamen batez besteko) edo bestela doitasun altuagoa lor daiteke (%93), %90eko estaldura eta 1,20 proposamenekin (MG+DM+TS+H2).

V.2.4.3 Ondorioak

Egindako lana laburtzeko, esan behar dugu errore ortografikoen zuzenketarako teknika ezberdinen azterketa egin dugula, helburua errore bakoitzeko proposamen bakarra ematea izan

dela. Problemaren aspektu zailenetakoa emaitzen proba egitea izan da, datuak urriak izaten direlako. Horregatik, modu artifizialean sortutako errorearen corpusa erabili dugu, azken probarako benetako errorearen beste corpusarekin batera.

Aurretik errorearen zuzenketa automatikoan egindako lanekin konparatuz gero (Yarowsky 1994, Golding eta Schabes 1996), esan behar dugu sistema horiek lehenago aurreikusitako hitzen nahasketan oinarritzen zirela (azentuak edo antzeko hitzak). Gure lanean, berriz, sistemak aukera bat hartu behar zuen edozein errorerentzat, eta ezinezkoa zen aldeztu aurretik proposamenen diskriminaziorako aukerak biltzea. Gure kasuan, gainera, errore guztiak zuzendu nahi dira, nahiz eta kasu batentzako datuak urriak izan, eta lan horietan agerpen ugari zeukaten hitzak hartzen ziren.

Emaitzak aztertu eta gero, esan dezakegu hobetu egiten direla testuinguru gehiago hartzen den heinean. Hitz-formen maiztasuna irizpide gordina baina lagungarria da zuzenketa aukeratzeko. Doitasuna igo egiten da testuinguru hurbilagoak hartu ahala, dokumentuaren maiztasunak, murriztapen-gramatika edo testuinguruko ezaugarrien modukoak.

Benetako errorearen corpusetik ondorio batzuk atera ditzakegu. Lehenengo, zuzenketa Ispell-en proposamenen artean dagoela ia %100ean, honek errore guztiak zuzendu daitezkeela emanez. Bigarren, uneko sistematik espero daitekeen irteera errore ortografikoen 25etik 24tan proposamen bakarra lortzea dela, estaldura osoarekin eta %90eko doitasunarekin, edo bestela %93ko doitasuna eta %90eko estaldura, batez besteko 1,20 proposamen utziz.

Bi teknikak, Brown corpuseko maiztasuna eta Dentsitate Kontzeptualak, ez dute emaitza erabilgarririk eman. DK bakarrik errorearen proportzio txiki bati aplikatzen zaio, honek Wordnet-en estaldurarekin erlazio zuzena du, eta horregatik ezin dugu beste ondorioz atera.

Etorkizunerako emaitzak hobetzea espero dugu. Lehenengo, errore errealeen corpusak dokumentu oso laburrak zeuzkan, eta honek DMren performantzia jaisteak ekarri du. Testu luzeagoak lortzera, emaitza hobeak aterako genituzke. Bigarren, esan behar da Ameriketako ingelesaren testuinguruko ezaugarriak Erresuma Batuko ingeleseko testuak zuzentzeko erabili zirela. Lengoaien arteko korrespondentzia-arazoak kenduz gero, emaitzen hobekuntza lor daiteke. Bukatzeko, erabilitako teknikak hobetzea ere badago. Errore ortografiko arrunten eredurik ez genuen erabili (hauek errorearen sorrerarekin lotuta egongo dira: OCR, teklatze-erroreak, ...). Testuinguruko ezaugarriak buruz esan behar dugu hitz-formen ezaugarriak bakarrik erabili ditugula, baina kategoria edo lema erabiltzeak hobekuntzak ekar

ditzake. Bestalde, tekniken konbinaziorako beste metodoak proba daitezke, nahiz eta hemen ez dugun uste aldaketa handia gertatuko denik.

V.3 Erroreen detekzio eta zuzenketari buruzko lanen ondorioak eta hurrengo pausoak

Kapitulu honetan sintaxiaren erabileraren azterketa egin dugu errore ortografiko eta sintaktikoen detekzio eta zuzenketan. Aplikazio hori lehenago garatutako tresna morfologiko eta sintaktikoetan oinarritu da, eta tresna berriak osatzen eta azkartzen diren heinean errorearen tratamendurako ahalmenak ere gorantz egingo du.

Egindako lanaren ekarpen nagusiak hauek dira:

- Euskaraz egindako errore sintaktikoen sailkapena egin da. Euskararen ezaugarriak direla eta, berezitasunak dauzka gehien landu diren ingelesa edo espainiera bezalako hizkuntzekin konparatuz. Sailkapenarekin batera, errorearen lehen azterketa egin da, tratamendu automatikorrako bideragarritasunaren arabera.
- Murritzapen sintaktikoen erlaxazioa esperimentatu da errore sintaktikoen detekzioarako. Teknika hau egokia da euskaraz maiz gertatzen diren erroreak tratatzeko, komuntadurak kasu. Teknikaren ekarpena isolatzeko, probak gramatika sintaktikoaren estalduran zeuden esaldien gainean egin dira. Esperimentuek frogatu dute teknika honek errore gehienak detektatzen dituela, ia guztietan proposamen zuzena emanez (beste batzuen artean). Tratatutako errore motak zabaltzen diren heinean, murritzapenen askatze gradualak beharrezkoa dela frogatu dugu.
- Gramatikaren estalduratik kanpo dauden eta fenomeno sintaktiko lokaletan egindako erroreak tratatzeko, errore-patroietan oinarritutako detekzioa probatu dugu. Esperimentua corpusetan aurkitutako errorearen gainean egin da, horrela errorearen detekzioa eta alarma faltsuen arazoa elkarrekin tratatuz. Horretarako baterakuntzan oinarritutako analizatzailerik sintaktikoa eta egoera finituko patroiak erabili dira. Datu adierazpenak aukeratu dira sistema probatzeko, errorearen iturri aberatsa direlako eta beste errorearen adierazgarritzat har daitezkeelako. Emaizak onak lortu dira, doitasunean zein estalduran.
- Azkenik, zuzenketa automatikoa ere landu da, errore ortografikoen proposamen bakarra lortzeko asmoz. Horretarako, testu errealeko errorearen gainean egin dira probak, ezagumendu mota desberdinen (sintaktikoa eta semantikoa) ekarpena

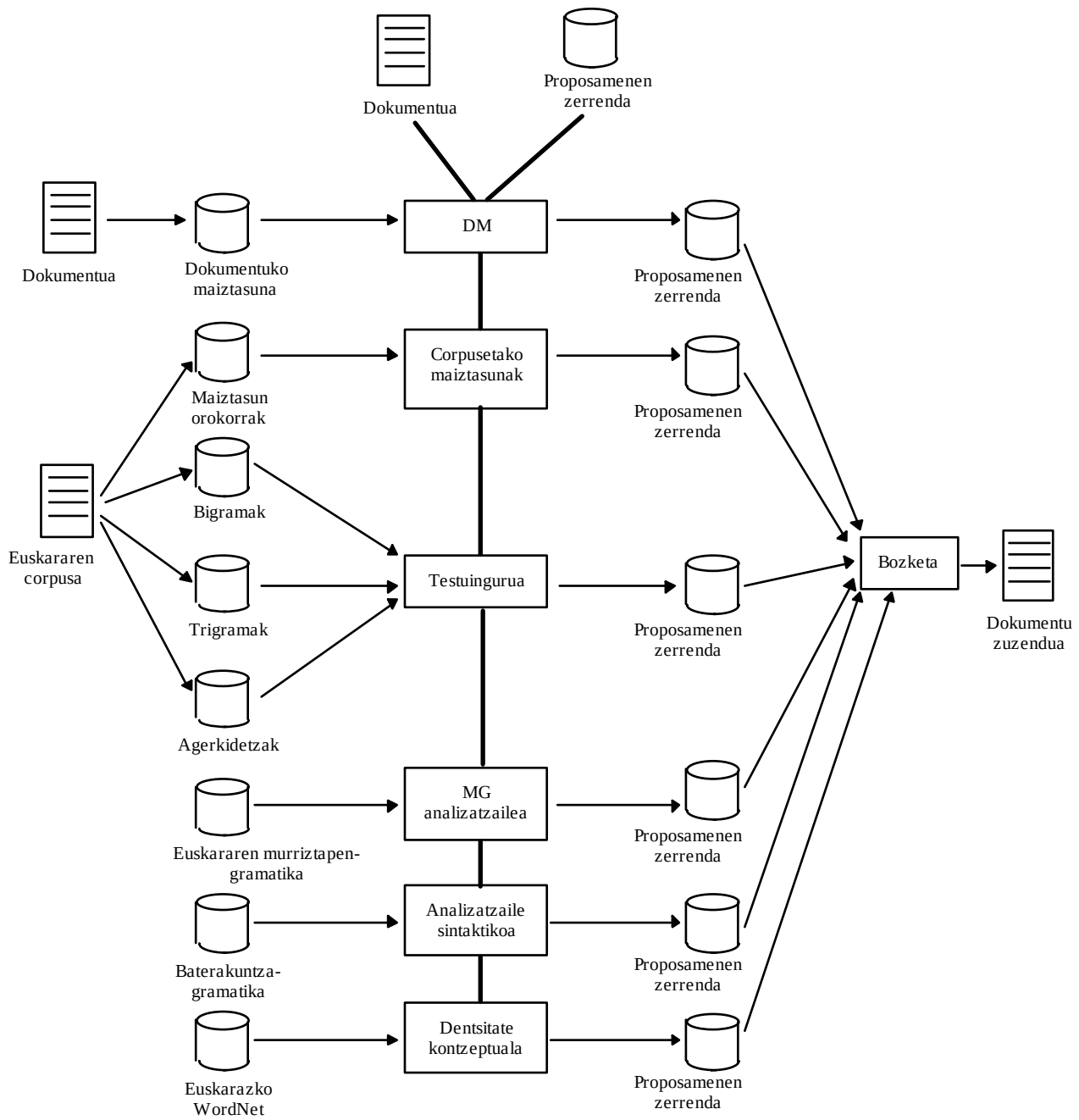
konbinatuz. Lortutako sistemak estaldura eta doitasun onak lortu ditu, kasu gehienetan proposamen bakarra utziz.

Bukatzeko, esan behar dugu etorkizunerako ikerlerroa izaten jarraituko duela errorearen tratamenduak, testu errealei aplikatuz modu eraginkorrean zuzenketa-tasa altuak eta alarma faltsuen kopuru baxua lortzen duten sistemak egiteko. Arlo honetan egin beharreko lanen artean hauek aurreikusten ditugu (Kukich 1992, Oliva 1997):

- Errore-patroien bidezko tratamenduaren zabalpena. Datekin egin den moduan, testuinguru sintaktiko lokaleko azterketarekin detekta daitezkeen errore gehiago aztertu eta tratatzeko asmoa dugu. Horien artean postposizioak edo azpikategorizazioarekin lotutako erroreak aipa daitezke (ikus V.8 taula).
- Analisi sintaktiko sendoa lortzen den neurrian, murriztapen sintaktikoen erlaxazioan oinarritzen diren sistemen inplementazioari ekin ahal izango zaio. Euskararen kasuan, erlaxazioa zuzenean edozein esaldiri aplikatzea kostu handikoa denez (exekuzio-denboraren aldetik, eta gainera etekin gutxikoa esaldi bakoitzean errore mota posible guztien agerpena frogatu nahi bada), gure ustez patroietan oinarritutako sintaxia eta erlaxazioa konbinatu beharko dira: patroiek detektatu beharko dituzte errorea eman dezaketen esaldien zatiak, eta ondoren baterakuntzan oinarritutako analizatzailea aplikatu dakioke zati horri. Horrela, bakarrik testuinguru horretan posible diren errore batzuk bilatuko dira.
- Linguistika konputazionalaren garapen-prozesuan gertatu den bezala, corpusen erabilera errorearen tratamenduan ere zabaldu egin beharko da. Ez da soluzio erraza daukan problema, corpusen lorpenak daukan problemez gain, errorearen tratamendurako corpusak lortzeko corpus berezituak (hau da, erroreak dauzkaten corpusak) behar dituelako. Gainera, corpus horiek lortu eta gero, ebaluaziorako erabilgarri izateko eskuzko lan handia egin beharko da. Aurrekoarekin lotuta, esan behar dugu errorearen detekzioarako sistemen mugarik handiena eskala zabaltzeko (*scale up*) zailtasuna dela, erroreak banan-banan edo errore-multzoka tratatu behar direlako, eta errore-kopurua infinitutzat kontsidera daitekeelako. Horregatik, testu errearen tratamendurako errorearen adibideak lortzea eskuz egin beharko da ia beti, gainditzeko zaila den muga jarritz (salbuespentzat hitz jakinen nahasketa (*weather/whether*), errore fonologikoak edo daten moduko erroreak har daitezke, hauetan lanaren zati bat automatikoki egin daitekeelako baina, hala ere, eskuzko lan handia eskatu dute).
- Erroreak detektatzeko erregelen ikasketa automatikoa. Behin erroreak markatuta dauzkan corpusa edukita, gure kasuan eskuz landu ditugu detekzio edo

zuzenketarako erregelak (gramatikaren gaineko erlaxazioak markatuz edo erroreak harrapatzeko patroiak definituz). Bestalde, errore ortografikoen zuzenketarako, automatikoki lortutako informazioak erabili dira (maiztasunak, testuinguruari buruzko erregelak). Horregatik, ikasketarako metodoen azterketa oso interesgarri ikusten dugu (Golding eta Roth 1996, Mangu eta Brill 1997), horrela sistema garatzeko lan bakarra errorearen adibideak markatzea izango litzatekeelako.

- Sintaxiaren prozesuan garatutako tresnak erabili ditugunez, ez ditugu asko landu errorearen tratamendu estatistikoak (hitz-formen bigrama/trigramak ezik). Gramatika sintaktikoa hedatzeko, euskararen neurri estatistikoen aplikazioaren erabilera (Ezeiza 2000) aztertzea interesgarri ikusten dugu, eta honek errorearen tratamenduaren hobekuntza ekarriko duela uste dugu. Antzera gertatuko da azpikategorizazioari buruzko informazioa lortzen den neurrian, errore asko azpikategorizazioaren erabilera okerrarekin lotuta daudelako.
- Detekzioaz gain, zuzenketak ere gainditu beharreko arazoak planteatzen ditu. Errore ortografikoen zuzenketak hitz baten aldaketa eskatzen du, eta ez dauka lan asko egin beharrik, baina errore sintaktiko baten zuzenketa egiteko, askotan errorea eta bere inguruko beste osagaiak aldatu beharko dira (komunztadura-errore bat dagoenean, adibidez, aditza edo izen-sintagmaren ezaugarriak alda litezke), beti ere esaldi osoaren zuzentasuna mantenduz. Honek sorkuntza morfologikoa eta kasu batzuetan sintaktikoa egitea eskatuko du.
- Ingeleserako errore ortografikoen zuzenketari ekin diogu emaitza onekin, eta horregatik oso interesgarri ikusten dugun urratzeko bidea euskarari aplikatzea da. Hizkuntza eranskarien tratamenduak arazo berriak dakartza, zuzenketarako proposamenak askotan deklinatuta joango direlako, eta proposamenen aukeraketarako joko aberatsagoak ager daitezkeela sumatzen dugulako. Adibidez *gizonk* hitzak bi proposamen izango ditu (*gizonak* eta *gizonek*), bakoitzak bi interpretaziorekin. Horien artean erabakitzeak sintaxiaren erabilera beharko du. Zuzenketarako ezagumendu-iturrien artean euskararen murriztapen-gramatika, baterakuntzan oinarritutako gramatika, maiztasunak eta agerkidetzak aipa ditzakegu. V.7 irudian modu horretako sistema baten arkitektura aurkezten da, ingeleserako egindakoan oinarrituta, baina euskararako bereziki landu diren baliabideak aipatuz. Baterakuntza-gramatika erabil liteke proposamenen aukeraketarako, hipotesi nagusitzat sintagma luzeenak ematen dituzten proposamenak hartuko direla. Egoera finituko patroiak ere erabil litezke errore jakin batzuetan proposamenak diskriminatzen.



V.7 irudia. Proposamenaren hautapenerako ezagutza-iturriak eta konbinazioa egiteko sistema.

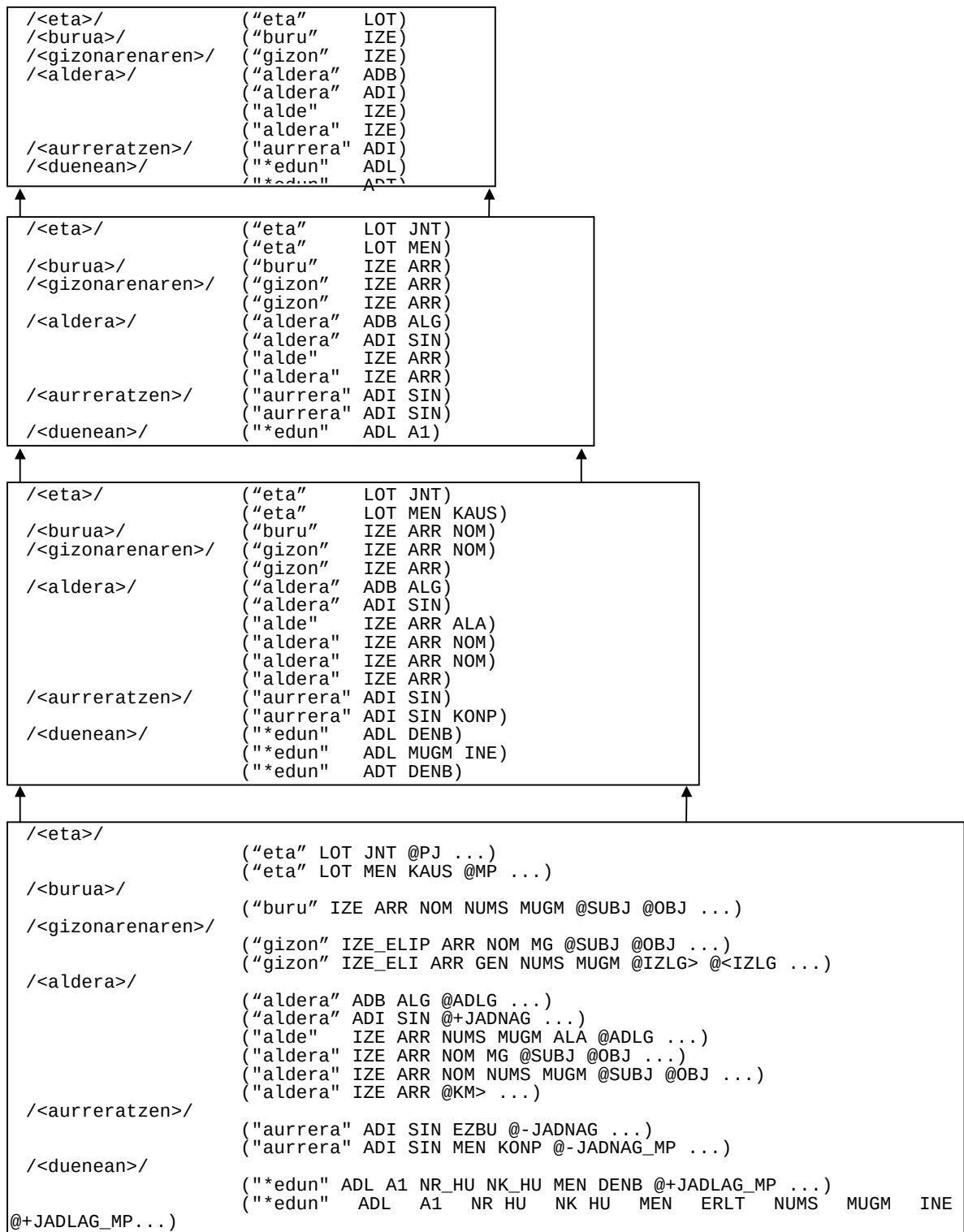
VI Beste aplikazioak

Aplikazioen atala bukatzeko, morfosintaxia eta sintaxian egindako lanaren ondorioz garatu diren bi tresna azalduko ditugu. Batetik, analizatzaile morfosintaktikoa euskararen lematizatzaile/etiketatzaile (EUSLEM izenekoa) baten osagaia izango da (§ VI.1). Bestetik, azaleko sintaxiaren tratamenduaren bidez ikasleen testuetako egitura sintaktikoen patroien maiztasunak neurtzeko tresna bat inplementatu da (§ VI.2)

VI.1 EUSLEM

Corpus bateko testu-hitz bakoitzari bere lema eta kategoria esleitzeko erabiliko den tresna informatikoa da EUSLEM izeneko lematizatzaile/etiketatzailea (Aduriz *et al.* 1996ab). Funtsezko tresna da corpus-analisirako, indexazio automatikorako edo analisi sintaktiko eta semantikorako. Testu errealei aplikatzeko tresna denez, egindako zenbait sistema konbinatuko ditu. Hona hemen bere osagai funtsezkoenak:

- a) Euskararen datu-base lexikala (EDBL). Lehen esan bezala, EUSLEMerako ez ezik beste aplikazio askotarako oinarri lexikala dugu EDBLn.
- b) Etiketatzeari begira, aurretik definituriko etiketa-sistema erabili da. Sistema mailakatua da, kategoria soila ematetik analisi morfologiko osoa (deskonposaketa, morfema bakoitzaren informazioa,...) emateraino heda daitekeena.
- c) Analizatzaile Morfologikoa (MORFEUS). Euskaraz, hizkuntza eranskaria eta morfologikoki konplexua izanik, ezinbestekoa da analisi morfologikoaren berri emango duen analizatzaile morfologiko bat.
- d) Analizatzaile morfologikoaren emaitzak hitz bakoitzaren analisi posible guztiak ematen ditu. Horren gainean aplikatzen da murriztapen-gramatika. Honek, informazio linguistikoan oinarrituz, desanbiguazio morfologikoa burutzen du. Horrez gainera, estatistikan oinarrituriko metodo bat ere aplikatuko da, aurrekoak desanbiguatu gabe uzten dituen kasu bakanetan (Ezeiza *et al.* 1998).
- e) Bestalde, EUSLEMek Hitz Anitzeko Unitate Lexikalak (HAUL) ere tratatzen ditu (Aduriz *et al.* 1996c). Unitate lexikal hauen tratamenduak zailtasun gehigarriak ditu. Horregatik, HAULak zehatz deskribatzeko adierazpide formal bat garatu da.



VI.1 irudia. EUSLEMen irteerarako informazio-mailak⁴⁰.

⁴⁰ Irudiaren beheko partean dagoen laugarren mailan, analizatzailer morfologikoaren emaitza (II.2 kapituluaz azaldutakoa) aterako da eta, beraz, irudia sinplifikazioa da. Horregatik jarri dira hiru puntu analisi bakoitzaren bukaeran, informazio gehiago dagoela adierazteko.

Beraz, tesi honetako lanak ekarpenak izango ditu EUSLEMen hiru atal garrantzitsutan:

a) Alde batetik, EUSLEMen irteera analisi morfosintaktikoaren emaitza izango da.

§ II.2n esan da emaitza hori osoa izango dela, hitzaren barruko interesgarri den informazio guztia kodetuz. Informazio-aberastasun honek malgutasuna emango dio EUSLEMi, eta hortik informazioa aplikazio-eremu askotarako iragazi ahal izango da. EUSLEMen diseinuan erabaki da lau informazio-maila egongo direla (ikus VI.1 irudia), kategoria soila ematetik informazio morfosintaktiko osorainoko tartea hartuz. Hau da definitutako mailen azalpen laburra:

- Lehen mailan, hemeretzi kategoria nagusiak (izena, aditza, ...) sartu dira. Hau lematizazio arrunta egiteko oinarritzko etiketa-sistema da.
- Bigarren mailan kategoria bakoitza azpikategorien etiketekin birfintzen da. Adibidez, aditzak sinple eta konposatuen artean desberdinduko dira.
- Beste informazio morfosintaktikoak gehituko zaizkio hirugarren mailari, kasua edo numeroaren modukoak (adibidez, kategoria, azpikategoria eta kasua hartuta, 318 etiketa konposatu desberdin aurkitu dira probarako testu batean). Maila honetan erabiltzaileak zehatz dezake behar duen informazioa.
- Azken mailan analizatzailer morfosintaktikoaren informazio guztia emango da, funtzio sintaktikoak barne. Maila honetako etiketa konposatuen kopurua asko igotzen da (2943 etiketa lehen aipatutako testuan).

```
"<Herriarenak>"
"herri" IZE ARR DEK GEN NUMS MUGM DEK ABS NUMP MUGM @OBJ @PRED
"herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ABS NUMP MUGM @OBJ @PRED
"<ziren>"
"izan" ADT B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG>
"<mendiak>"
"mendi" IZE ARR DEK ABS NUMP MUGM @OBJ @PRED
"<ataldu>"
"ataldu" ADI SIN AMM PART ASP BURU DU NOTDEK @-JADNAG
"<eta>"
"eta" LOT JNT AORG @PJ
"<saldu>"
"saldu" ADI SIN AMM PART ASP BURU NOTDEK @-JADNAG
"saldu" ADI SIN AMM PART NOTDEK @-JADNAG
"<egin>"
"egin" ADI SIN AMM PART ASP BURU NOTDEK @-JADNAG
"<zituzten>"
"*edun" ADL B1 NOR_NORK NR_HK NK_HK @+JADLAG
"<$.>"
PUNT_PUNT
```

VI.1 adibidea. Murriztapen-gramatikaren oraingo sarrera.

b) Beste alde batetik, analizatzailer morfosintaktikoa EUSLEMen

barne-funtzionamendurako ere lagungarria izango da, murriztapen-gramatikak hartzen duen sarrera morfosintaxiaren irteera izango delako. Une honetan murriztapen-gramatikak VI.1 adibideko formatuan hartzen ditu testuak.

Ikusten denez, oraingo deskribapen morfologikoa segmentatzailearen irteera da, eta horregatik kasu batzuetan bi balio daude numero, kasu eta mugatasunerako, hau da, lehen aipatutako problemak (ikus § II.2) agertzen dira. Arazo hori gainditzeko, murriztapen-gramatika kasu horiek tratatzeko erregela bereziak landu dira, erregelen definizioa konplikatuz eta orokortasuna galduz. Analisi morfosintaktikoak egoera honen konponketa ekarriko du, gramatikarien lana erraztuz, eta azken gramatikaren ulergarritasuna eta mantentzea hobetuko du.

- c) Desanbiguazioa euskararen murriztapen-gramatika erabiliz. Honen garapena § II.3n azaldu da. Taldean metodo estokastikoen desanbiguazioa ere aztertu da, baina honen emaitzak ez dira izan ezagutza linguistikoan oinarritutakoaren metodoarenak bezain onak. Horregatik desanbiguaziorako moduluan lehenengo murriztapen-gramatika aplikatuko da, eta ondoren berak utzitako anbiguotasunak ebazteko tratamendu estatistikoa, hitz bakoitzeko interpretazio bakarra geratzeko.

Puntu honi bukaera emateko, esan dezakegu egindako hitzaren gramatika zein murriztapen-gramatika funtsezko tresnak izango direla euskararen tratamendurako. Hitzaren gramatika hitz-mailan oinarritzen diren moduluetan erabiliko da, bai barne-funtzionamendurako bai kanporako emaitzak aurkezteko orduan. Murriztapen-gramatika euskara bezalako hizkuntzetan desanbiguaziorako ezinbesteko tresna izango da. Horregatik, tresna horiek funtsezkoak dira lematizatzaile/etiketatzaile batean.

VI.2 Ikasleen testuen egitura sintaktiko orokorrak aztertzeko tresna

Egindako analizatzaile sintaktikoa esaldi bateko elementu sintaktiko nagusiak analizatzeko gai denez, baliagarria izan daiteke testuen zenbait ezaugarriren azterketarako. Puntu honetan aipatuko duguna bigarren hizkuntza baten ikaskuntza-prozesua aztertzeko sistema batean erabili da. Sistema horretan euskara ikasten ari diren ikasleen testuak aztertu nahi dira, Maritxalar-en tesian (1999). Tesi horretan morfologiaren inguruan egin da lan gehiena, baina hitz barruko errorearen eta akatsen deskribapenaz aparte, maila desberdinetako ikasleen testuen egitura sintaktikoen maiztasunak atera nahi izan dira, hipotesi nagusia maila ezberdinetako ikasleek egitura sintaktiko ezberdinak erabiltzen dituztela zela. Hori egiaztatzeko, tresna linguistiko eta informatikoa garatu zen, lehen aipatutako oinarritzko analizatzaileak abiapuntutzat hartuz.

Adibidez, hasierako esaldia hau bada:

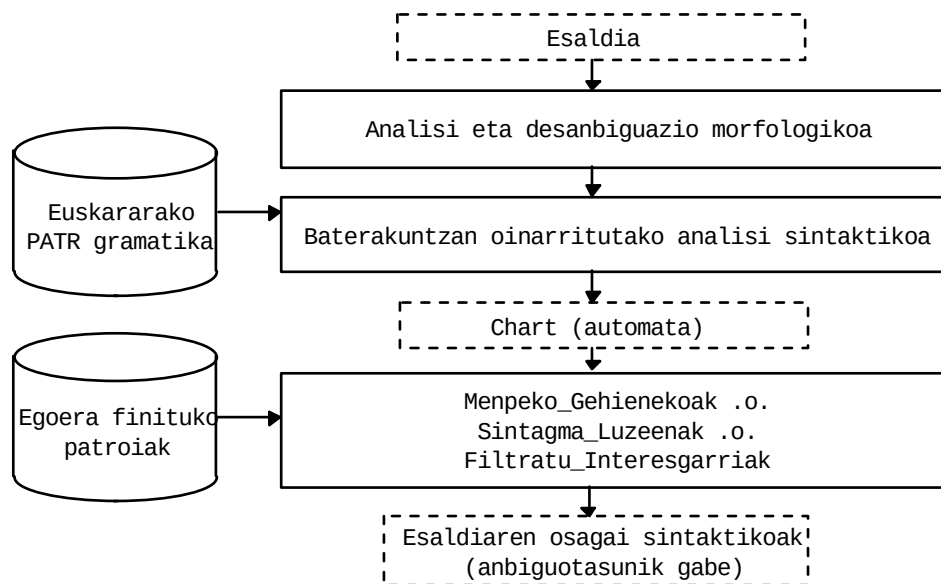
Horregatik, susmatzekoa da honekin batera, garraioak ere igo egingo direla, bidaiatzea gero eta garestiagoa izan dadila zeren eta dibisak erosteko pezeta gehiago ordaindu beharko da erebai.

Emitzen esaldi horretan zenbat menpeko dauden (menpeko bakoitzaren mota markatuz), zenbat esaldi arrunt, eta beraiek lotzeko nolako juntagailuak dauden jakin nahi da. VI.2 irudian deskribatzen da jarraitutako prozedura. Lehen pauso batean esaldiaren osagai sintaktikoak lortzen dira baterakuntzan oinarritutako gramatikaren bidez. Ondoren egoera finituko gramatika aplikatuko zaio testuari emaitza ateratzeko.

- Lehenengo menpeko gehien dauzkaten interpretazioak hartuko dira. Pauso honek baztertu egingo ditu interpretazio anbiguo batzuk, adibidez ‘*jatea eta edatea*’ esaldian, *-tea* bukaerako bi elementuek *-tea*-menpekoa/izena-abs (eratorpenaren bidez) anbiguotasuna dute, eta horregatik lau interpretazio posible ditugu:

<i>jatea</i>	<i>eta</i>	<i>edatea</i>
menpekoa		menpekoa
izena-abs		izena-abs

Horietatik, biak menpekotzat hartzen dituen interpretazioa hobesten da.



VI.2 irudia. Esaldi baten osagai sintaktikoak ateratzeko tresna.

- Bigarren pausoa sintagma luzeenak hartzen saiatuko da, hau da, *politak* eta ‘*etxe politak*’ bi izen-sintagmen artean bigarrena hartuko dugu. Honek aukera asko baztertuko ditu.

- Hirugarren pauso batean esaldiaren interpretazio guztietatik informazio interesgarria iragazten da, honela interpretazio bakarra utziz kasu guztietan. Adibidez, esaldi batek mota honetako lau interpretazio izan ditzake
esaldi sinplea + juntagailua + esaldi sinplea. Interesatzen dena patroio orokorra denez, honek aukera asko baztertuko ditu.

Aurreko esalditik lortutako emaitza VI.1 taulan agertzen den modukoa izango da. Informazio hori atera ondoren maila desberdinetako ikasleen esaldien konplexutasuna aztertu da, ondorio nagusiak maila baxuagoko ikasleek testu bera idazteko maila altuagokoek baino esaldi gutxiago erabiltzen dituztela, eta gainera esaldi horiek lotzeko mekanismo sintaktikoak pobregokoak direla, gehienetan koordinazioa erabiliz. Maila altukoek, aldiz, esaldi luzeagoak eta menpekoen erabilera aberatsagoa egiten dute. Adibidez, maila baxuko ikasleen esaldien laurdena (%25,25) *esaldi sinplea + eta + esaldi sinplea* patroikoak dira, eta maila altukoetan esaldi horiek %5,13a besterik ez du suposatzen. Beste alde batetik, menpekorik gabeko esaldien portzentajea handiagoa da behe mailako ikasleen testuetan (%32,35 versus %69,7).

Esaldia	Emaitza
<i>Horregatik,</i>	
<i>susmatzekoa</i>	Menpekoa(tzeko)
<i>da</i>	Esaldi sinplea
<i>honekin batera,</i>	
<i>garraioak ere igo egingo direla,</i>	Menpekoa(ela)
<i>bidaiatzea</i>	Menpekoa(tzea)
<i>gero eta garestigoa izan dadila</i>	Menpekoa(ela)
<i>zeren eta dibisak erosteko</i>	Menpekoa(tzeko)
<i>pezeta gehiago ordaindu beharko da erebai</i>	Esaldi sinplea

VI.1 taula. Testuen azterketarako tresnak lortutako emaitzaren adibidea.

VI.3 Ondorioak

Kapitulu labur honetan ikusi dugu egindako tresnak, analizatzaile morfosintaktikoa zein analizatzaile sintaktikoak, erabilgarri direla aplikazio-multzo zabal baterako. Analisi morfosintaktikoa lematizatzaile/etiketatzaile batentzako oinarritzko baliabidea da, lema eta morfemen informazioetatik abiatuta hitz osoaren informazioa lortzeko. Analizatzaile sintaktiko partziala tresna egokia da ikasleen testuetako egitura sintaktikoen erabilera aztertzeko, baita lengoaien irakaskuntzarekin lotutako beste hainbat aplikaziotarako ere.

AURRERA BEGIRAKOAK ETA ONDORIOAK

VII Tesiaren ondorio nagusiak eta etorkizunerako ikerlerroak

VII.1 Lortutako emaitzak

Tesi honetan euskararen sintaxi konputazionalaren azterketa egin da, testu errealean tratamendurako baliabide linguistiko orokorrak garatu dira eta zenbait aplikazio inplementatu dira. Sintaxia oraingoz eremu zabalegia dela kontuan hartuta, lan honetan bere tratamenduaren aspektu batzuk baino ez dira landu, zenbait irizpideren arabera aukeratuta. Lana ikerketa-talde baten barruan kokatzen denez, sintaxiaren deskribapenerako beste ikerketa-lanen testuinguruan kontsideratu behar da.

Hasteko, hitzaren barruko informazioa osatzeko hitzaren gramatika morfosintaktiko osoa garatu eta inplementatu da (II. kapitulua). Lehenago eginda zegoen hitzaren segmentazioaren bi mailatako deskribapenari gramatika hau gehituz euskararen morfologiaren tratamendu konputazional osoa biribildu egin da, corpusetako edozein hitzen analisisa egiteko tresna lortuz. Euskararen morfologia konplexua dela eta, analizatzaile morfologiko oso hau sintaxiaren tratamenduaren hasiera da, eta hitza edo osagai sintaktiko handiagoak unitatetzat hartzen dituen edozein prozesamendu motatan erabili beharreko tresna izango da. Beste alde batetik, analizatzaile morfologiko osoaren garapenak erakutsi digu formalismo ezberdinen konbinazioak abantailak dituela, formalismo bakoitzaren alde onenak har daitezkeelako.

Morfosintaxiaren kasuan bi mailatako formalismoa eta baterakuntza-gramatikak integratu dira osagarriak diren moduan, bukaeran sistema osoaren ahalmen deskriptiboa handituz eta testu errealeak tratatzeko eraginkortasuna lortuz (Aduriz *et al.* 2000ab).

Morfosintaxiaren tratamenduaren ondoren sintaxiaren mundu zabalagoan sartu gara III. kapituluan. Hasteko, euskararen estaldura ertaineko baterakuntza-gramatika garatu eta dagokion analizatzailea inplementatu da. Gramatika partziala da, baina corpusetan agertzen diren osagai sintaktiko nagusiak deskribatzen ditu (horien artean izen-sintagmak, adizlagunak, esaldi sinpleak eta menpeko esaldiak), eta horrela esaldien analisi sakon eta osorako lehen pausoa eman da. Analisi osoari ez ekiteko bi arrazoi aipatu dira. Batetik, EDBLren erabilerak sendotasuna eta estaldura lexikal zabala eskaintzen duen arren, oraingoz informazio sintaktiko garrantzitsuen gabezia ere kontuan hartu behar izan dugula (aditzen azpikategorizazioari buruzkoa, adibidez). Bestetik, sintaxiaren eremu zabalak tesi honetan tratatzeko osagaiak mugatzera eraman gaituela. Une honetan, analizatzailea gai da corpus handiak modu eraginkorrean tratatzeko eta corpus horietatik osagai sintaktiko nagusiak ateratzeko.

Egoera finituko sintaxiaren bidetik, euskararen murriztapen-gramatikaren garapenean lan egin dugu, testuetako hitzen desanbiguazioa lortzeko. Gramatikaren garapena lankidetzan egin da (Aduriz 2000, Arriola 2000). Gure ekarpen nagusia tratamendu informatikoan izan da. Beste hizkuntzetan lortutako emaitzekin konparagarria den desanbiguazio-tasa lortu dugu, euskara bezalako hizkuntzetarako formalismoaren egokitasuna frogatuz. Tresna erabilgarria lortu da, une honetan zenbait aplikaziotan erabiltzen ari dena. Tresnaren erabilera nagusia testuetako anbiguotasunaren jaitsieran egin dugu, ondorengo prozesuetarako lana aurreratuz.

Murriztapen-gramatikaren mugak kontuan hartuta, egoera finituko beste formalismo orokorrago baten egokitasuna aztertu dugu. Formalismo honek adierazpen erregularren bidezko patroietan oinarritutako sintaxia garatzea ahalbidetzen du, patroiekin anbiguotasuna kentzeko murriztapenak, informazioaren iragazleak edo osagai sintaktiko berrien sorkuntza adierazteko. Egin ditugun zenbait lanetan formalismo honen baliagarritasuna frogatu dugu.

Sintaxiaren tratamendurako hiru hurbilpen horien azterketak aukera eman digu bakoitzaren alde onak eta arazoak esperimintatzeko. Horregatik hirurak konbinatzen dituen sistema integratua garatu dugu, bakoitzaren alde positiboak biltzeko asmoz. Hain zuzen ere, hiru tresnen aplikazio sekuentziala probatu eta inplementatu dugu. Murriztapen-gramatika desanbiguazio morfosintaktikorako erabiliko da lehen pauso batean, emaitza desanbiguatuaren gainean baterakuntzan oinarritutako gramatikaren bidez osagai sintaktiko

posibleak eraikitze. Osagai sintaktikoen anbigutasuna eta aplikazioen behar desberdinak tratatzeko egoera finituko tresnaren bidezko patroiak defini daitezke, bukaeran corpusen azterketarako tresna malgua eta sendoa lortuz.

Baliabide sintaktiko orokor horien garapena egin ondoren, beraien gaineko aplikazioak landu dira. IV. kapitulan ezagumendu lexikal eta sintaktikoaren aberasketan egindako esperimentuak azaldu dira, corpusetatik aditzaren azpikategorizazio-informazioaren erauzketan. Modu horretan, hasierako baliabide sintaktikoak osatu egin ahal izango dira, era iteratiboan: oinarritzko gramatikak erabiliz informazio linguistikoa atera daiteke gramatika horiek aberasteko, ondoren aberasketa-prozesua behin eta berriro errepikatzeko. Esperimentu horietan azaldu dugunez, emaitzak erabilgarriak dira, estaldura eta doitasun altuekin, eta beraien gainean aditzei buruzko informazioa eskuz edo modu automatikoen bidez atera ahal izango dugu. Beste alde batetik, esperimentu hauek oinarritzko baliabide sintaktikoen baliagarritasuna frogatzeko ere balio izan dute.

Garatu diren tresna sintaktikoen erabilera errore ortografiko eta sintaktikoen detekzio eta zuzenketan ere esperimentatu dugu V. kapitulan. Lehenengo, euskarazko testuetako errore sintaktikoen sailkapena egin da, beste hizkuntzetan gertatzen direnekin konparatuz. Bigarren, baterakuntzan oinarritutako gramatikaren gaineko murriztapen sintaktikoen erlaxazio gradualak probatu da komunztadurak eta antzeko erroreak detektatzeko, metodoaren bideragarritasuna frogatuz. Ondoren, errore-patroietan oinarritutako detekzioa probatu dugu, corpusetan aurkitutako daten adierazpenen errorearen gainean, baterakuntzan oinarritutako analizatzaile sintaktikoa eta egoera finituko patroiak konbinatuz. Bukatzeko, zuzenketa landu da, errore ortografikoen proposamen bakarria modu automatikoki lortzeko, eta zenbait ezagumendu motaren (sintaktikoa eta semantikoa) ekarpena neurtu da.

Erroreei buruzko kapitulu hau laburtzekotan, esan behar dugu frogatu dugula ezagutza-iturri ezberdinen ekarpenak gehituz gero LNParren edozein aplikazioaren emaitzatan islatuko dela. Konbinazioak proposamen ezberdinen indarrak biltzeko ahalmena duela ikusi dugu.

Azkenik, VI. kapituluak beste bi aplikazio erakusten ditu, baliabide sintaktikoen malgutasun eta erabilpenaren adibidea emateko. Analisi morfosintaktikoa lematizatzaile/etiketatzaile batentzako oinarritzko baliabidea da, lema eta morfemen informazioetatik abiatuta hitz osoaren informazioa lortzeko. Bestalde, analizatzaile sintaktiko partziala tresna egokia da ikasleen testuetako egitura sintaktikoen erabilera aztertzeko.

VII.2 Zabaldutako ikerlerroak eta perspektibak

Tesi honetako ekarpen nagusiak aztertu eta gero, esan behar dugu sintaxiaren tratamenduaren azterketak bukaera argi eta zehatz batera eramane baino, ikertu gabeko bide berri zabal eta interesgarriak ireki dizkigula. Ondorengo lerroetan, tesiaren lana jarraitzeko aurreikusten ditugun ikerlerro nagusiak emango ditugu.

VII.2.1 Tratamendu morfosintaktikoaren jarraipena

Nahiz eta morfologiaren tratamendua ia bukatutzat eman, oraindik lan egiteko eremuak ikusten ditugu ondoko aspektuetan:

- Gramatikaren formalizazio linguistikoa. PATR formalismoa erabili dugu deskribapen morfosintaktikorako, horren alde malgutasuna aipatzen genuela beste formalismo konplexuagoak ez erabiltzeko. Baina honek daukan beste ondorio bat erregelen idazketan egin beharreko lana da, printzipio orokorrik definitu gabe ezaugarri guztien tratamendua baterakuntza-ekuazioekin egin behar delako. Printzipioen erabilerak gramatikaren trinkotasuna eta sinpletasuna ekarriko duenez, hasi dugun bide hau jarraitzea komenigarri ikusten dugu. Beste alde batetik, ezaugarri-egitura motadunak erabiltzeak gramatikaren koherentzia eta ziurtasuna gehituko du.
- Eraginkortasunaren hobetzea. Inplementatutako sistemaren abiadura nahikoa da orain taldearen beharretarako, baina gero eta corpus handiagoen erabilerak etorkizunerako eskakizunak gogortu egingo ditu. Lehenago esan dugunez, egin daitezkeen hobekuntza dezente daude, abiadura magnitude-ordena bat baino gehiago igotzeko. Bestalde, egoera finituko soluzio berrien ekarpena ere kontuan izan beharko da, une honetan ezinezkoak diren tratamenduetarako proposamenak egiten badira.
- Datu-base lexikaleko informazioak aldatzen edo gehitzen direnean gramatika morfosintaktikoa aldatu edo zabaldu egin beharko da. Adibidez, hitz anitzeko unitateen tratamendua lexikoan sartzeko asmoa dagoenez, horientzako erregelak definituko dira, edo azpikategorizazioa bezalako informazioak lexikoan sartzen diren heinean, berari dagozkion tratamendu morfosintaktikoen definizioak ere gehitu egin beharko dira.

VII.2.2 Tratamendu sintaktikoaren jarraipena

Sintaxiaren tratamendurako esperimentatzeko aukera anitz ikusten dugun arren, hurrengo puntuetan lehentasun handienekoak aipatuko ditugu:

- Aberasketa lexikal eta sintaktikoa. Lortutako tresnen bidez ikusi dugu corpusen azterketa egin daitekeela, doitasun eta estaldura altuekin. Horregatik, corpusetan kodetuta dagoen ezagumenduaren erauzketarako bidea irekita dago. Hau frogatu egin dugu azpikategorizazio-informazioa ateratzeko prozesuan, baina esan dugunez esperimentu horiek lehen pausoak besterik ez dira, eta egin beharreko lan hauek definitu ditugu:
 - ♦ Aditz eta osagarri posibleen adibideak lortu eta gero, azpikategorizazio-patroien definizioa egin daiteke, metodo estatistiko edo erdiautomatikoak erabiliz. Tresna automatikoak direla eta, patroi horiei agerpen-maiztasunak gehi dakizkieke, analizatzaile estatistikoen bidea jorratuz.
 - ♦ Ideia beretik abiatuta, egitura sintaktikoen maiztasunak ere neur daitezke, horrela gramatikako erregelen aplikazioa corpusetako maiztasunen arabera baldintzatzeko.
 - ♦ Sintaxitik harantzago joanez, osagai lexikalen arteko erlazioak ere azter litezke semantikaren mundua ukituz, kokakidetzak edo egitura sintaktiko batzuen bidez definitutako erlazio semantikoak ateratzeko (Berland eta Charniak 1999).
- Baterakuntzan oinarritutako gramatikaren zabalpena. Informazio lexikala aberasten joango denez, egingarria izango da euskararen LFG edo HPSG moduko formalismo ahaltsuagoen bidezko deskribapena. Lan hau moldaketa sinple bat baino askoz aldaketa handiagoa izango da, formalizazio linguistiko zehatza egitea eskatuko duelako, bai une honetan gramatikan tratatzen diren fenomenoena bai gramatikatik kanpo utzi diren guztiena (horien artean esaldi nagusi eta menpekoen arteko erlazioak).
- Mendekotasun-egituren bidezko analisia esaldi osoen tratamendurako. Hau aurreko puntuaren alternatiba izan liteke, baterakuntza-gramatika dagoen bezala utzi eta sortutako osagai sintaktikoen konbinazioak aztertuz, esaldi osoen analisiak zehazteko. Era honetan zenbait lanetan (Järvinen eta Tapanainen 1998, Oflazer *et al.* 1999ab, Basili *et al.* 2000) hasitako pausoak jarraituko genituzke.
- Sintaktikoki etiketatutako corpusen garapena. Pauso hau beharrezkoa da ebaluazioa edo estatistikan oinarritutako analisiak egiteko. Corpus hauek lortzeko egin behar

den lan handia kontuan hartuz gero, aztertzeko bide posible bat garatu diren tresna automatikoen laguntza erabiltzea izan daiteke.

- TEI estandarren definizioak tresna desberdinen arteko eta kanpoko komunikaziorako (Artola *et al.* 2000). Tresna linguistikoen emaitzak aplikazio desberdinek erabiltzeko nahitaezkoa da emaitza horiek definizio formal bat izatea, horrela kodeketa-modu partikularren arazoa gainditzeko. Honengatik baliabide linguistikoen sarrera/irteerak TEI gidalerroak jarraituz kodetzeko lanean ari gara une honetan.

VII.2.3 Erroreen tratamendurako lanen jarraipena

Erroreen tratamenduan egin ditugun lanak jarraitzeko honako bide hauek ditugu gogoan:

- Murritzapen sintaktikoen erlaxazioa testu errealen tratamendurako egokitu. Hau bideragarria izateko, patroietan oinarritutako sintaxia eta erlaxazioa konbinatu beharko dira: patroiek detektatu beharko dituzte errorea eman dezaketen esaldien zatiak, eta ondoren baterakuntzan oinarritutako analizatzailea aplikatu dakioke zati horri. Era horretan esaldi bakoitzeko errore posible guztien hipotesiak aztertzea ekidin ahal izango da, eta testuinguruaren arabera posible diren erroreak bakarrik aztertuko dira.
- Errore-patroien bidezko tratamenduaren zabalpena. Datekin egin den moduan, testuinguru sintaktiko lokaleko azterketarekin detekta daitezkeen errore gehiago aztertu eta tratatzeko asmoa dugu. Horien artean postposizioak edo azpikategorizazioarekin lotutako erroreak aipa daitezke. Azken hauen tratamendua azpikategorizazio-informazioaren erauzketan lortutako emaitzekin lotuta egongo da.

Esan behar da ere patroien bidezko erroreen tratamendua modu estuan lotuta dagoela erroreak markatuta dituzten corpusen garapenarekin. Une honetan hau da arazo handienetako bat euren tratamendurako.

Corpus horiek lortzen diren neurrian errore-patroi horientzako ikasketa-metodo automatikoen edo erroreen tratamendu estatistikoaren ekarpena probatu ahal izango da.

- Euskarazko testuetako errore ortografikoen zuzenketa automatikoa. Ingelesarekin egindako esperimentuaren ondoren, euskararekin probak egitea izango da hurrengo urratsa. Ingeleserako baliabide ugari dagoela aipatu dugu zuzenketa automatikoko proba egiteko hizkuntza aukeratu dugunean, baina euskararen baliabideen aberasketak aurrera jarraitzen badu, euskararen gaineko esperimentu berriak egiteko

materiala izango da, horien artean MG eta baterakuntzan oinarritutako gramatikak edo baliabide semantikoak (Agirre 1999) daudela.

- Tresna sintaktikoen integrazioa hizkuntzaren irakaskuntzarako sistema batean. Morfologiarekin gertatu den antzera (Maritxalar 1999), tresna sintaktikoak erabiliak izan daitezke euskararen irakaskuntzarako sistemetan. Aplikazio horietarako aukera ugari ikusten dugu: ikaslearen gaitasun linguistikoa neurtzeko, errorearen detekzio edo zuzenketa- edo ikasketa-prozesuen laguntza-moduan.

BIBLIOGRAFIA

Konferentzia nagusien erreferentziak egiteko laburdurak erabiliko dira:

- ACL: Association for Computational Linguistics.
- ANLP: Conference on Applied Natural Language Processing.
- COLING: International Conference on Computational Linguistics.
- EACL: European Association for Computational Linguistics.
- IWPT: International Workshop on Parsing Technologies.
- NAACL: North American Chapter of the Association for Computational Linguistics.
- SEPLN: Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural.

Abaitua, J. 1988. *Complex predicates in Basque: from lexical forms to functional structures*. Doktoretza-tesia, University of Manchester.

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X., Artola X., Arriola J.M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R. 1992. *Estudio comparativo de diferentes formalismos sintacticos para su aplicacion al euskara*. Barne-txostena, UPV/EHU/LSI.

Abney S. P. 1991. *Parsing by chunks*. R. C. Berwick, S. P. Abney, and C. Tenny, editoreak, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, Dordrecht.

Abney S. P. 1997. *Part-of-Speech Tagging and Partial Parsing*. S. Young eta G. Bloothoof, editoreak, Corpus-Based Methods in Language and Speech Processing, Kluwer, Dordrecht.

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. 1995. *Different Issues in the Design of a lemmatizer/Tagger for Basque*. From Texts to Tags: Issues in Multilingual Language Analysis. Association for Computational Linguistics SIGDAT Workshop, Dublin.

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. 1996a. *EUSLEM: A lemmatiser/tagger for Basque*. Euralex'96.

- Aduriz I., Aldezabal I., Alegria I., Ezeiza N., Urizar R. 1996b. *Del analizador morfológico al etiquetador/lematizador: Unidades léxicas complejas y desambiguación*. SEPLN'96, Sevilla.
- Aduriz I., Aldezabal I., Alegria I., Ezeiza N., Urizar R. 1996c. *MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque*. COMPLEX.
- Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997. *A spelling corrector for Basque based on morphology*. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford.
- Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. 1997. *Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism*. Conference on Recent Advances in Natural Language Processing, Bulgaria.
- Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., Insausti J. M. 1998a. *EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada.
- Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., ArtolaX., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. 1998b. *A Framework for the Automatic Processing of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada.
- Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., ArtolaX., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 1999. *MORFEUS: Euskararako analizatzaile morfosintaktikoa*. Barne-txostena, UPV/EHU/LSI/TR 1-99.
- Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 2000a. *A Word-Grammar Based Morphological Analyzer for Agglutinative Languages*. COLING'2000, Saarbrücken.
- Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 2000b. *A Word-Level Morphosyntactic Analyzer for Basque*. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000), Atenas.
- Aduriz I. 2000. *Morfologiatik Sintaxira Murriztapen Gramatika baliatuz*. Tesi-txostena, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.
- Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M. 1992. *XUXEN: A spelling Checker/Corrector for Basque Based on Two-Level Morphology*. ANLP'92, Trento.
- Agirre E., Arregi X., Arriola J. M., Artola X., Insausti J. M. 1994. *Euskararen Datu-Base Lexikala (EDBL)*. Barne-txostena UPV/EHU/LSI/TR 8-94.

- Agirre, E., Arregi, X., Artola, X., Díaz, A., Sarasola, K. 1994. *Lexical-semantic information and the automatic correction of spelling errors*. Proceedings of the Workshop on Semantics and Pragmatics of Natural Language: Logical and Computational Aspects, Sara, France.
- Agirre E., Rigau G. 1996. *Word sense disambiguation using conceptual density*. COLING'96, Copenhagen.
- Agirre E., Gojenola K., Sarasola K., Voutilainen A. 1998a. *Towards a Single Proposal in Spelling Correction*. COLING-ACL'98, Montreal.
- Agirre E., Gojenola K., Sarasola K., Voutilainen A. 1998b. *Towards a Single Proposal in Spelling Correction*. UPV-EHU / LSI / TR 8-98.
- Agirre E. 1999. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala*. Doktoretza-tesia, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Aho SA., Sethi R., Ullman J. 1985. *Compilers: Principles, Techniques and Tools*. Addison-Wesley.
- Aït-Mokhtar S., Chanod J-P. 1997. *Incremental Finite-State Parsing*. ANLP'97.
- Aït-Mokhtar S., Chanod J-P. 1997. *Subject and Object Dependency Extraction Using Finite-State Transducers*. ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications, Madrid.
- Aldezabal I., Alegria I., Artola X., Díaz de Illaraza A., Ezeiza N., Gojenola K., Aduriz I., Urkia M. 1994. *EUSLEM: Un lematizador/etiquetador de textos en euskara*. SEPLN'94, Córdoba.
- Aldezabal I., Goenaga P., Gojenola K., Sarasola K. 1998. *Subcategorización verbal vasca: propuesta inicial y herramienta de validación*. SEPLN'98, Alicante.
- Aldezabal I., Alegria I., Ansa O., Arriola J.M., Ezeiza N. 1999a. *Designing spelling correctors for inflected languages using lexical transducers*. Proceedings of EACL'99, 265-266. Bergen, Norway.
- Aldezabal I., Ansa O., Artola X., Ezeiza A., Gojenola K., Insausti J.M., Lersundi M. 1999b. *Euskararen Datu-Base Lexikala (EDBL): eskema berriaren proposamena*. Barne-txostena, UPV/EHU/LSI/TR 9-99.
- Aldezabal I., Gojenola K., Oronoz M. 1999c. *Combining Chart-Parsing and Finite State Parsing*. Proceedings of the Student Session of the European Summer School in Logic, Language and Computation (ESSLLI'99), Utrecht.
- Aldezabal I., Gojenola K., Sarasola K. 2000. *A Bootstrapping Approach to Parser Development*. IWPT'2000, Trento.
- Aldezabal I. 2000. *Euskal aditzaren azpikategorizazioa. Azterketa sistematiko-automatikoa*. Tesi-txostena, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

- Alegria I. 1995. *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktoretza-tesia, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Alegria I., Artola X., Sarasola K. 1995 *Improving a robust morphological analyzer using lexical transducers*. Conference on Recent Advances in Natural Language Processing, Bulgaria.
- Alegria I., Artola X., Sarasola K., Urkia M. 1996a. *Automatic morphological analysis of Basque*. Literary and Linguistic Computing. 11 (4). Oxford University.
- Alegria I., Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Urkia M. 1996b. *A Corpus-Based Morphological Disambiguation Tool for Basque*. SEPLN'96, Sevilla.
- Allen J., Hunnicut M., Klatt D. 1987. *From text to speech: the MITalk System*. Cambridge University Press.
- Allen J. 1995. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company.
- Alshawi H. (editorea) 1992. *The Core Language Engine*. Cambridge, MA: MIT Press.
- Alshawi H., Moore R. C. 1992. *Introduction to the CLE*. The Core Language Engine, Alshawi, H. editorea, MIT Press.
- Antworth E. L. 1994. *Morphological Parsing with a Unification-based Word Grammar*. North Texas Natural Language Processing Workshop, Texas.
- Arregi X. 1995. *ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza*. Tesi-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Arriola J.M., Soroa A. 1996. *Lexical Information Extraction for Basque*. Student Conference in Computational Linguistics, Canada.
- Arriola J. M., Artola X., Gojenola K., Soroa A. 1997. *TEI: testu-kodeketarako gidalerroak*. Ekaia, Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria 7.
- Arriola J.M., Artola X., Maritxalar M., Soroa A. 1999. *A Methodology for the Analysis of Verb Usage Examples in a Context of Lexical Knowledge Acquisition from Dictionary Entries*. Workshop on Linguistically Interpreted Corpora, EACL'99, Bergen.
- Arriola J.M. 2000. *Hauta-Lanerako Euskal Hiztegi-ko informazio lexikalaren erauzketa erdi-automatikoa eta bere integrazioa sistema konputazional batean*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.
- Artola X. 1993. *Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena/Conception et construction d'un systeme intelligent d'aide dictionnaire (SIAD)*. Tesi-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A. 2000. *Integration of NLP Tools using SGML-tagged texts*. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000), Atenas.
- Atserias J., Carmona J., Castellón I., Cervell S., Civit M., Màrquez L., Martí M.A., Padró L., Placer R., Rodríguez H., Taulé M., Turmo J. 1998. *Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text*. First International Conference on Language Resources and Evaluation (LREC'98), Granada.
- Atwell E. 1987. *Constituent-likelihood grammar*. The Computational Analysis of English, Garside R., Leech G., Sampson G. liburuan.
- Atwell E., Elliott S. 1987. *Dealing with Ill-Formed English Text*. In The Computational Analysis of English: a Corpus-Based Approach, Ed. Longman.
- Badia T., Egea A., Tuells A. 1996. *SEGMORF: un formalismo para analizadores morfológicos de dos niveles*. SEPLN'96, Sevilla.
- Badia T. 1997. *Especificaciones lingüísticas para gramáticas en estructuras de rasgos tipificadas*. Philologia Hispalensis, Vol. XI, fascículo 2, Facultad de Filología, Universidad de Sevilla.
- Basili, R., Pazienza M.T., Zanzotto F.M. 1998. *Efficient Parsing for Information Extraction*. Proceedings of the 13th European Conference on Artificial Intelligence, John Wiley & Sons Ltd.
- Basili, R., Pazienza M.T., Zanzotto F.M. 2000. *Customizable Modular Lexicalized Parsing*. IWPT'2000, Trento.
- Bates, M. 1978. *The theory and practice of ATN grammars*. L. Bolc (ed.) 'Natural Language Communication with Computers' liburuan, Springer Verlag, Berlin.
- Bear J. 1986. *A Morphological Recognizer with Syntactic and Phonological Rules*. COLING'86.
- Bear J. 1988. *Morphology with Two-Level Rules and Negative Rule Features*. COLING'88.
- Beesley K. 1998a. *Constraining Separate Morphotactic Dependencies in Finite-State Grammars*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.
- Beesley K. 1998b. *Arabic Morphological Analysis on the Internet*. Proceedings of the International Conference on Multi-Lingual Computing (Arabic & English), Cambridge.
- Berland M., Charniak E. 1999. *Finding Parts in Very Large Corpora*. ACL'99, Maryland.
- Berwick R., Abney S.P., Tenny C. 1991. *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht: Kluwer Academic.
- Bod R. 1993. *Using an annotated corpus as a stochastic grammar*. EACL'93, Utrecht.

- Bod R., Kaplan R. 1998. *A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis*. COLING-ACL'98, Montreal.
- Borsley R., Przepiórkowski A. 1999. *Slavic in Head-Driven Phrase Structure Grammar*. CSLI publications, Stanford.
- Brent M. 1993. *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax*. Computational Linguistics, vol. 19(2).
- Bresnan J., (editorea) 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, Massachusetts.
- Brill E. 1995. *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging*. Computational Linguistics.
- Brill E., Florian R., Henderson J.C., Mangu L. 1998. *Beyond N-Grams: Can Linguistics Sophistication Improve Language Modeling?*. COLING-ACL'98, Montreal.
- Brill E., Wu J. 1998. *Classifier Combination for Improved Lexical Disambiguation*. COLING-ACL'98, Montreal.
- Brill E., Henderson J., Ngai G. 2000. *Automatic Grammar Induction: Combining, Reducing and Doing Nothing*. IWPT'2000, Trento.
- Briscoe T., Carroll J. 1993. *Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars*. Computational Linguistics, vol. 19(1).
- Briscoe T. 1994. *Parsing (with) Punctuation etc*. Barne-txostena MLTT-002, Xerox Research Center.
- Briscoe T., Carroll J. 1997. *Automatic Extraction of Subcategorization from Corpora*. ANLP'97, Washington.
- Brun C. 1998. *Terminology Finite-State Preprocessing for Computational LFG*. COLING-ACL'98, Montreal.
- Butt M., Fortmann C., Rohrer. C. 1996. *Syntactic Analyses for Parallel Grammars: Auxiliaries and Genitive NPs*. COLING'96, Copenhagen.
- Butt M., Geuder W. 1998. *The projection of arguments - lexical and compositional factors*. CSLI publications, Stanford.
- Butt M., King T.H., Nino M.E., Segond F. 1999 *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Lecture Notes, CSLI Publications.
- Carbonell J.G., Hayes P.J. 1983. *Recovery strategies for parsing extragrammatical input*. American Journal of Computational Linguistics 9.
- Carpenter B., Penn G. 1993. *ALE User's Guide Version 1.0*. Laboratory for Computational Linguistics Technical Report, Carnegie Mellon University, Pittsburgh.

- Carroll J. 1993. *Practical unification-based parsing of natural language*. Computer Laboratory, Cambridge University, UK, PhD. thesis, Technical Report 314.
- Carroll G., Rooth M. 1998. *Valence Induction with a Head-Lexicalized PCFG*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Granada.
- Carroll J., Minen G., Briscoe T. 1998a. *Can Subcategorisation Probabilities Help a Statistical Parser?* Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora, Montreal.
- Carroll J, Briscoe T., Sanfilippo A. 1998b. *Parser Evaluation: a Survey and a New Proposal*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.
- Carroll J, Minnen G., Briscoe T. 1999. *Corpus Annotation for Parser Evaluation*. Proceedings of Workshop on Linguistically Interpreted Corpora, EACL'99, Bergen.
- Carulla M., Oosterhoff A. 1996. *El Tratamiento de la Morfología Flexiva del Castellano mediante reglas de dos niveles en una gramática de unificación*. SEPLN'96, Sevilla.
- Chanod J.P., Tapanainen P. 1995. *Tagging French - Comparing a Statistical and a Constraint-Based Method*. EACL'95, Dublin.
- Chanod J.P., Tapanainen P. 1996a. *A Non-deterministic Tokeniser for Finite-State Parsing*. ECAI '96 workshop on Extended finite state models of language. Budapest.
- Chanod J.P., Tapanainen P. 1996b. *A Robust Finite-State Grammar for French*. ESSLLI'96 Workshop on Robust Parsing, Prague.
- Charniak E. 1993. *Statistical Language Learning*. Cambridge, MA: MIT Press.
- Chomsky N. 1981. *Lectures on Government and Binding*. Dordrecht.
- Church K. 1988. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*. ANLP'88.
- Ciravegna F., Lavelli A. 1997. *Controlling Bottom-Up Chart Parsers through Text Chunking*. IWPT'97, Boston.
- Collins M. 1997. *Three New Probabilistic Models for Statistical Parsing*. ACL'97, Madrid.
- Collins M., Hajic J., Ramshaw L., Tillmann C. 1999. *A Statistical Parser for Czech*. ACL'99, Maryland.
- Damerau F.J., Mays E. 1989. *An examination of undetected typing errors*. Information Processing and Management 25, 6.
- Díaz de Ilarraza A., Maritxalar M., Oronoz M. 1997. *An implemented interlanguage model for learners of Basque*. Language Teaching and Language Technology, Swets and Zeitlinger Publisher.

- Díaz de Ilarraza A. 1990. *Diseño de un Módulo de Interacción Tutor-alumno para un sistema inteligente de enseñanza de la programación*. Tesi-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Doran C., Egedi D., Hockey B. A., Srinivas B., Zaidel M. 1994. *XTAG system - A Wide Coverage Grammar for English*. COLING'94, Kyoto.
- Douglas, S. 1991. *A Review and Bibliography of Approaches to Syntactic Error Correction and Robust Parsing*. The Editor's Assistant Project No. IED4/1/1679, Deliverable 4.3.1.
- Douglas, S., Dale R. 1991. *Towards a Taxonomy of Errors in Technical Texts*. The Editor's Assistant Project No. IED4/1/1679, Deliverable 3.4.
- Douglas, S. 1992. *Customising grammar and style checker rules*. *Intelligent Tutoring Media*, Vol. 3
- Douglas, S., Dale R. 1992. *Towards Robust PATR*. COLING'92, Nantes.
- Egunkaria 1992. *Estilo liburua*. Egunkaria bilduma.
- Euskaltzaindia 1985. *Euskal Gramatika Lehen Urratsak-I*, Iruñea.
- Euskaltzaindia 1994. *Euskal gramatika laburra: perpaus bakuna*. Euskaltzaindia gramatika batzordea.
- Euskaltzaindia. *Euskaltzaindiaren arauak. Datak nola adierazi*. 37. unitatea, Bilbo.
- Ezeiza N. 1997. *EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa*. Tesina-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Ezeiza N., Alegria I., Arriola J.M., Urizar R., Aduriz I. 1998. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98, Montreal.
- Ezeiza N. 2000. *Corpusak ustiatzeko tresna linguistikoak/ Herramientas lingüísticas para la explotación de corpus*. Tesi-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Fass D., Wilks Y. 1983. *Preference semantics, ill-formedness, and metaphor*. *American Journal of Computational Linguistics* 9.
- Francis S., Kucera H. 1967. *Computing Analysis of Present-Day American English*. Brown University Press.
- Frank A., King T.H., Kuhn J., Maxwell. J. 1998. *Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars*. In *Proceedings of the LFG98 Conference*, Brisbane, Australia. CSLI Online Publications.
- Gala N. 1999. *Using the Incremental Finite-State Architecture to create a Spanish Shallow Parser*. SEPLN'99, Lleida.

- Gale W.A., Church K.W. 1990. *Estimation procedures for language context: Poor estimates are worse than none*. Proceedings of Compstat-90, Springer-Verlag, New York.
- Garside R., Leech G., Sampson G. 1987. *The Computational Analysis of English*. Longman.
- Gazdar G., Klein E., Pullum G., Sag I. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.
- Genthial D. 1991. *Contribution à la construction d'un système robuste d'analyse du français*. Thèse de l'université Joseph Fourier, Grenoble I.
- Genthial D., Courtin J., Menezo J. 1994. *Towards a More User-Friendly Correction*. COLING'94, Kyoto.
- Giguet E., Vergne J. 1997. *From Part of Speech Tagging to Memory-based Deep Syntactic Analysis*. IWPT'97, Boston.
- Goenaga P. 1980. *Gramatika bideetan*. Erein
- Gojenola K., Sarasola K. 1994. *Aplicaciones de la relajación gradual de restricciones para la detección y corrección de errores sintácticos*. SEPLN'94, Córdoba.
- Gojenola K. 1998. *Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG) eta euskararako aplikazioa*. Barne-txostena, UPV/EHU/LSI/TR5-98.
- Gojenola K., Oronoz M. 2000. *Corpus-Based Syntactic Error Detection Using Syntactic Patterns*. NAACL/ANLP 2000 Student Research Workshop, Seattle.
- Golding A. 1995. *A Bayesian hybrid method for context-sensitive spelling correction*. Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA.
- Golding A., Roth. D. 1996. *A Winnow-based Approach to Spelling Correction*. Proceedings of the 13th International Conference on Machine Learning, ICML'96.
- Golding A., Schabes. Y. 1996. *Combining trigram-based and feature-based methods for context-sensitive spelling correction*. ACL'96, Santa Cruz, CA.
- Gómez Guinovart J. 1996. *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*. Tese de doutoramento, Departamento de Filoloxía Española, Teoría da Literatura e Lingüística Xeral, Universidade de Santiago de Compostela.
- Gómez Guinovart J. 1999. *La escritura asistida por ordenador. Problemas de sintaxis y de estilo*. Servicio de publicacións, Universidade de Vigo.
- Grefenstette G. 1996. *Light Parsing as Finite-State Filtering*. ECAI'96 workshop on Extended finite state models of language. Budapest.
- Grimshaw J. 1990. *Argument Structure*. MIT Press: Cambridge.
- Grinberg D., Lafferty J., Sleator D. 1995. *A robust parsing algorithm for link grammars*. IWPT'95, Prague.

- Grishman R., Macleod C., Meyers A. 1994. *Complex Syntax: Building a Computational Lexicon*. COLING'94, Japan.
- Gross M. 1997. *The Construction of Local Grammars*. Finite-State Language Processing liburuan, MIT Press.
- Grover, C., Carroll J., Briscoe E. 1993. *The Alvey Natural Language Tools grammar (4th release)*. Computer Laboratory, Cambridge University, UK, Technical Report 284.
- Gunji T. 1987. *Japanese Phrase Structure Grammar*. Dordrecht: Reidel.
- HABE 1985. *Akatsen bilduma analitikoa*. Zutabe Aldizkaria
- Hajic J., Hladká B. 1998. *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset*. COLING-ACL'98, Montreal.
- Heaton, J. B., Turton, N. D. 1987. *Longman Dictionary of Common Errors*. Longman.
- Heidorn G. E., Jensen K., Miller L. A., Byrd R. J., Chodorow M. S. 1982. *The EPISTLE text-critiquing system*. IBM Systems Journal, Vol. 21, No. 3.
- Hellwig P. 1998. *Natural Language Parsers: a "course in cooking"*. Tutorial Notes, COLING-ACL'98, Montreal.
- Hermjakob U., Mooney R.J. 1996. *Learning Parse Decisions From Examples With Rich Context*. ACL'96.
- Hindle D. 1989. *Acquiring disambiguation rules from texts*. ACL'89.
- Hobbs J.R., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. 1997. *FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text*. Finite-State Language Processing liburuan, MIT Press.
- Holan T., Kubon V., Plátek M. 1997. *A Prototype of a Grammar Checker for Czech*. ANLP'97, Washington.
- Hudson R. 1990. *English Word Grammar*. Oxford: Basil Blackwell.
- Ide N., Veronis J. K. 1995. *Text-Encoding Initiative, Background and Context*. Kluwer Academic Publishers.
- Ide N. 1998. *Encoding Linguistic Corpora*. Sixth Workshop on Very Large Corpora, COLING-ACL'98, Montreal.
- Ide N., Greenstein D., (Editors-in-Chief) 1999. *Tenth Anniversary of the Text Encoding Initiative*. Computers and the Humanities, Special Double Issue, Volume 33, Nos. 1-2.
- In-Sig Y., Kwang-Moo C., Taisook H. 1993 *Syntactic error repair using repair patterns*. Information Processing and Management 47.

- Järvinen T., Tapanainen P. 1998. *Towards an Implementable Dependency Grammar*. Proceedings of the Workshop on Processing of Dependency-Based Grammars, COLING-ACL'98, Montreal.
- Jensen K. 1987a. *Binary rules and non-binary trees*. Mathematics of Language liburuan, A. Manaster-Ramer editorea, Amsterdam: John Benjamins.
- Jensen K. 1987b. *Issues in Parsing*. Computer Science Research Report 13380 12/23/87, IBM Thomas J. Watson Research Center.
- Jensen K. 1988. *Why Computational Grammarians Can Be Skeptical About Existing Linguistic Theories*. COLING'98, Budapest.
- Jensen K., Heidorn G., Richardson S. 1993. *Natural Language Processing: the PLNLP Approach*. Kluwer Academic Publishers, Boston.
- Joshi A. 1985. *Tree-adjointing grammars: How much context sensitivity is required to provide reasonable structural descriptions*. D. R. Dowty, L. Karttunen, A. Zwicky editoreak, Natural Language Parsing, New York: Cambridge University Press.
- Kaplan R. Kay M. 1981. *Phonological rules an finite-state transducers*. Annual Meeting of the Linguistic Society of America, New York.
- Karlsson F., Voutilainen A., Heikkilä J., Anttila A. 1995. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.
- Karttunen L. 1986. *Radical Lexicalism*. Center for the Study of Language and Information (CSLI), 86-68, Stanford.
- Karttunen L. 1995. *The Replace Operator*. ACL'95, Boston.
- Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997. *Regular Expressions For Language Engineering*. Natural Language Engineering.
- Karttunen L. 1998. *The Proper Treatment of Optimality in Computational Phonology*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.
- Kempe A., Karttunen L. 1996. *Parallel Replacement in Finite-State Calculus*. COLING'96, Copenhagen.
- Kempe A. 1997. *Finite State Transducers Approximating Hidden Markov Models*. ACL'97, Madrid.
- Kiefer B., Krieger H., Carroll J., Malouf R. 1999. *A Bag of Useful Techniques for Efficient and Robust Parsing*. ACL'99, Maryland.
- Kiefer B., Krieger H. 2000. *A Context-free Approximation of Head-Driven Phrase Structure Grammar*. IWPT'2000, Trento.
- Koskenniemi K. 1983. *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. Ph D. thesis, University of Helsinki.

- Koskenniemi K., Tapanainen P., Voutilainen A. 1992. *Compiling and using finite-state syntactic rules*. COLING'92, Nantes.
- Kuhn J. 1998. *Towards data-intensive testing of a broad-coverage LFG grammar*. In Proceedings of KONVENS 98, Bonn. Peter Lang.
- Kuhn J., Eckle-Kohler J., Rohrer. C. 1998. *Lexicon Acquisition with and for Symbolic NLP-Systems -- a Bootstrapping Approach*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.
- Kukich K. 1992. *Techniques for automatically correcting words in text*. In ACM Computing Surveys, Vol. 24, N. 4, December, pp. 377-439.
- La Serna N., Díaz A., Rodríguez H. 1997. *Parsers Optimization for Wide-Coverage Unification-Based Grammars using the Restriction Technique*. IWPT'97, Boston.
- Lapata M. 1999. *Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations*. ACL'99, Maryland.
- Levin B. 1993. *English verb classes and alternations*. The University of Chicago Press.
- Macleod C., Grishman R., Meyers A., Barrett L., Reeves R. 1998. *NOMLEX: A Lexicon of Nominalizations*. Euralex'98.
- Mangu L., Brill E. 1997. *Automatic Rule Acquisition for Spelling Correction*. Proceedings of the 14th International Conference on Machine Learning, ICML'97.
- Marcus M. 1980. *A Theory of Syntactic Recognition for Natural Language*. The MIT Press.
- Marcus M., Santorini B. 1991. *Building very large natural language corpora: the Penn Treebank*. CIS report, University of Pennsylvania.
- Maritxalar M. 1999. *Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea*. Doktoretza-tesia, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Maxwell J.T., Kaplan R. 1996. *An efficient parser for LFG*. Proceedings of LFG'96, Grenoble.
- Mays E., Damerau F.J. Mercer R.L. 1991 *Context based spelling correction*. Information Processing and Management 27, 5.
- McCord M. 1990. *A System for Simpler Construction of Practical Natural Language Grammars*. Natural Language and Logic, Lecture Notes in Artificial Intelligence, Springer Verlag.
- Mellish C. 1989. *Some Chart-Based Techniques for Parsing Ill-Formed Input*. ACL'89.
- Menezo J., Genthial D., Courtin J. 1996 *Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs*. NLP + IA 96, Moncton, N. B., Canada.

- Menzel W. 1988. *Error Diagnosing and Selection in a Training System for Second Language Learning*. COLING'88.
- Menzel W., Schröder I. 1999. *Error Diagnosis for Language Learning Systems*. ReCALL, special edition, May.
- Miller L. A. 1986. *Computers for Composition: a stage model approach to helping*. Visible Language XX 2.
- Miller G. 1990. *Five papers on WordNet*. Special Issue of the International Journal of Lexicography, Vol. 3, N. 4.
- Min K., Wilson W. 1998. *Integrated Control of Chart Items for Error Repair*. COLING-ACL'98, Montreal.
- Mitjushin L. 1996. *An Agreement Corrector for Russian*. COLING'96, Copenhagen.
- Mitton R. 1987. *Spelling checkers, spelling correctors, and the misspelling of poor spellers*. Information Processing and Management 23, 5.
- Nunberg G. 1990. *The Linguistics of Punctuation*. CSLI Lecture Notes, 4 zenbakia, Stanford.
- Oepen S., Callmeier U. 2000. *Measure for Measure: Parser Cross.fertilization - Towards Increased Component Comparability and Exchange*. IWPT'2000, Trento.
- Oesterle J., Maier-Meyer P. 1998. *The GNoP (German Noun Phrase) Treebank*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.
- Oflazer K., Okan Y. 1996. *A Constraint-based Case-frame Lexicon*. COLING'96 Copenhagen.
- Oflazer K. 1999a. *Dependency Parsing with an Extended Finite State Approach*. ACL'99, Maryland.
- Oflazer K., Zeynep D., Tür H., Tür G. 1999b. *Design for a Turkish Treebank*. Proceedings of Workshop on Linguistically Interpreted Corpora, at EACL'99, Bergen.
- Oliva K. 1997. *Techniques for Accelerating a Grammar-Checker*. ANLP'97, Washington.
- Peterson J.L. 1980. *Computer programs for detecting and correcting spelling errors*. Communications of the ACM 23, 12.
- Peterson J.L. 1986. *A note on undetected typing errors*. Communications of the ACM 29, 7.
- Pollard C., Sag I. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. CSLI Lecture Notes no. 13, The University of Chicago Press.
- Pollard C., Sag I. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.
- Pollock J.J., Zamora A. 1984. *Automatic spelling correction in scientific and scholarly text*. Communications of the ACM 27, 4.

- Popowich F., Vogel C. 1991. *A logic-based implementation of HPSG*. Natural language understanding and logic programming III, C. Brown and G. Koch (editoreak), Elsevier Science Publishers.
- Prószeýy G. 1994. *Industrial Applications of Unification Morphology*. ANLP'94, Stutgart.
- Prószeýy G. 1996. *Morphological Analyzer as Syntactic Parser*. COLING'96, Copenhagen.
- Prószeýy G. 1998. *An Intelligent Multi-Dictionary Environment*. COLING-ACL'98, Montreal.
- Prószeýy G., Kis B. 1999. *A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages*. ACL'99, Maryland.
- Przepiórkowski A. 1999. *On Complements and Adjuncts in Polish*. Slavic in Head-Driven Phrase Structure Grammar liburuan, CSLI publications, Stanford.
- Rabinovitz R. 1993. *Better writing through electricity*. PC Magazine, Maizatza.
- Ramírez F., Sánchez-León F. 1996. *GramCheck: A Grammar and Style Checker*. COLING'96, Copenhagen.
- Ramírez F., Sánchez-León F., Declerck T. 1997. *Corrección gramatical y Preprocesamiento*. SEPLN'97, Madrid.
- Ritchie G., Pullman S. G., Black A. W., Russel G. J. 1987. *Computational Framework for Lexical Description*. Computational Linguistics, Vol. 13.
- Ritchie G., Russel G. J., Black A., W., Pullman S. G. 1992. *Computational Morphology: Practical Mechanisms for the English Lexicon*. The MIT Press.
- Roche R., Schabes Y. 1997. *Finite-State Language Processing*. MIT Press.
- Rodríguez, C. 1991. *CORRECTOR: un sistema de verificación sintáctica y estilística de textos*. SEPLN'91.
- Ruiz J.C., Zubizarreta J.R., Abaitua J. 1990. *Un compilador de LFG y su aplicación al euskara*. SEPLN'90, Donostia.
- Sampson G. 1987. *Evidence against the 'Grammatical/Ungrammatical' Distinction*. Corpus Linguistics and Beyond, W. Meijs editorea, Rodopi, Amsterdam.
- Samuelsson C., Voutilainen A. 1997. *Comparing a Linguistic and a Stochastic Tagger*. ACL-EACL'97, Madrid.
- Sarasola K. 1988. *Caprate: Un Sistema de Interpretación de Problemas en Lenguaje Natural / Caprate: Lengoia Naturalez idatzitako problemen interpretaziorako sistema*. Tesi-txostena, Lengoia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.
- Satta G. 2000 *Parsing Techniques for Lexicalized Context-free Grammars*. IWPT'2000, Trento.

- Schabes Y., Joshi A.K. 1991. *Parsing with Lexicalized Tree Adjoining Grammar*. Current Issues in Parsing Technology liburuan, Kluwer.
- Schabes Y., Waters R.C. 1993. *Stochastic Lexicalized Context-Free Grammar*. IWPT'93, Tilburg-Durbuy.
- Schank R., Lebowitz M., Birnbaum L 1980. *An integrated understander*. American Journal of Computational Linguistics 6.
- Schiller A. 1996. *Multilingual Finite-State Noun Phrase Extraction*. ECAI'96 Workshop on Extended Finite State Models of Language, Budapest.
- Schneider D., McCoy. K.F. 1998. *Recognizing Syntactic Errors in the Writing of Second Language Learners*. COLING-ACL'98, Montreal.
- Selkirk E. 1982. *The syntax of words*. MIT Press, Cambridge.
- Sells P. 1985. *Lectures on Contemporary Syntactic Theories*. CSLI, Stanford.
- Shieber S.M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, 4 zenbakia, Stanford.
- Shieber S.M. 1988. *Separating Linguistic Analyses from Linguistic Theories*. Natural Language Parsing and Linguistic Theories, U. Reyle eta C. Rohrer editoreak.
- Silberztein M. 1997. *The Lexical Analysis of Natural Languages*. Finite-State Language Processing liburuan, MIT Press.
- Skut W, Krenn B., Brants T., Uszkoreit H. 1997. *An Annotation Scheme for Free Word Order Languages*. ANLP'97, Washington.
- Skut W, Brants T., Krenn B., Uszkoreit H. 1998. *A Linguistically Interpreted Corpus of German Newspaper Text*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.
- Sleator D, Temperley D. 1993. *Parsing English with a Link Grammar*. IWPT'93, Tilburg-Durbuy.
- Smith J. 1992. *Mark your words with grammar-checking software*. PC/Computing, October.
- Sparck Jones K., Galliers J.R. 1996. *Evaluating Natural Language Processing Systems*. Lecture Notes in Artificial Intelligence, Springer.
- Sproat R. 1992. *Morphology and Computation*. The MIT Press.
- Stede M. 1992. *The Search for Robustness in Natural Language Understanding*. Artificial Intelligence Review 6.
- Tapanainen P., Voutilainen A. 1994. *Tagging Accurately-Don't guess if you know*. ANLP'94.
- Tapanainen P. 1996. *The Constraint Grammar Parser CG-2*. University of Helsinki. Publications n° 27.

- Tapanainen P. 1997. *Applying a Finite-State Intersection Grammar*. Finite-State Language Processing liburuan, MIT Press.
- ten Hacken P., Bopp S. 1998. *Separable Verbs in a Reusable Morphological Dictionary for German*. COLING-ACL'98, Montreal.
- Tomabechi H. 1993. *A soft graph unification method for robust parsing*. IWPT'93, Tilburg-Durbuy.
- Tomita M. 1986. *Efficient Parsing for Natural Language*. Boston: Kluwer Academic Publishers.
- Tomita M. 1988. *'Linguistic' Sentences and 'Real' Sentences*. COLING'88, Budapest.
- Trask, L. 1983. *Euskal izen sintagmaren egituraz*. 'Iker-2: Piarres Lafiteri Omenaldia' liburuan, Euskaltzaindia, Bilbo.
- Trost H. 1990. *The application of two-level morphology to non-concatenative German morphology*. COLING'90, Helsinki.
- Trost H., Matiassek J. 1994. *Morphology with a Null-Interface*. COLING'94, Japonia.
- Turcato D., Nicholson D., Heift T., Toole J., Tsiplakou S. 2000. *A Parsing Methodology for Error Detection*. IWPT'2000, Trento.
- Tzoukerman E., Liberman M. 1990. *A Finite-State Morphological Analyzer for Spanish*. COLING'90, Helsinki.
- Urkia M., Sagarna A. 1991. *Terminología y Lexicografía Asistida por Ordenador. La experiencia de UZEI*. SEPLN'91, Donostia.
- Urkia M. 1997. *Euskal morfologiaren analisi automatikorantz*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.
- Uszkoreit H. 1986. *Categorial Unification Grammars*. COLING'86, Bonn.
- Uszkoreit H. 1991. *Strategies for adding control information to declarative grammars*. ACL'91, Berkeley.
- Uszkoreit H., Backofen R., Busemann S., Diagne A.K., Hinkelman E.A., Kasper W., Kiefer B., Krieger H., Netter K., Neumann G., Oepen S., Spackman S. 1994. *DISCO - An HPSG-based NLP system and its application for appointment scheduling*. COLING'94, Kyoto.
- Van Berkel B., DeSmedt K. 1988. *Triphone analysis: A combined method for the correction of ortographical and typographical errors*. ANLP'88, Austin.
- Vosse T. 1992. *Detecting and correcting morpho-syntactic errors in real texts*. ANLP'92, Trento.

- Vosse T. 1994. *The Word Connection: Grammar-based Spelling Error Correction in Dutch*. PhD Thesis, Unit for Experimental and Theoretical Psychology, University of Leiden, Holland.
- Voutilainen, A., Tapanainen P. 1993. *Ambiguity resolution in a reductionistic parser*. EACL'93, Utrecht.
- Voutilainen, A. 1994a. *Three studies of grammar-based surface parsing of unrestricted English text*. Ph.D. thesis. University of Helsinki. Publications n° 24.
- Voutilainen, A. 1994b. *Designing a parsing grammar*. University of Helsinki. Publications n° 22.
- Voutilainen A, Järvinen T. 1995a. *Specifying a shallow grammatical representation for grammatical purposes*. EACL'95, Dublin.
- Voutilainen A. 1995b. *A syntax-based part-of-speech analyser*. EACL'95, Dublin.
- Voutilainen A. 1997. *Designing a Parsing Grammar*. Finite-State Language Processing liburuan, MIT Press.
- Weischedel R.M., Sondheimer N.K. 1983 *Meta-rules as a Basis for Processing Ill-Formed Input*. American Journal of Computational Linguistics, 9.
- Wiren, M. 1993. *Fully incremental parsing*. IWPT'93, Tilburg-Durbuy.
- XTAG Group. 1995. *A Lexicalized Tree Adjoining Grammar for English*. Technical Report IRCS 95-03, University of Pennsylvania.
- Yannakoudakis E.J., Fawthrop D. 1983. *The rules of spelling errors*. Information Processing and Management 19.
- Yarowsky D. 1994. *A comparison of corpus-based techniques for restoring accents in Spanish and French text*. Proceedings of the Second Workshop on Very Large Corpora, Kyoto.
- Zajac R. 1998. *Feature Structures, Unification and Finite-State Transducers*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.
- Zubimendi J.R., Esnal P. 1993. *Idazkera-liburua*. Eusko Jaurlaritzaren Argitalpen-Zerbitzu Nagusia.
- Zubiri I., Zubiri E. 1995. *Euskal gramatika osoa*. Didaktiker.
- Zubizarreta J.R. 1992. *Un modelo funcional de diálogo para diálogos orientados por la tarea*. Doktoretza-tesia, Euskal Herriko Unibertsitatea.
- Zubizarreta J.R., Jones C. 1994. *Modeling Dialogue by Functional Subcategorization*. COLING'94, Kyoto.