

Lengoaia eta Sistema Informatikoak Saila



Informatika Fakultatea

**EUSKARAZKO ERRORE SINTAKTIKOAK
DETEKTATZEKO ETA ZUZENTZEKO
BALIABIDEEN GARAPENA: DATAK,
POSTPOSIZIO-LOKUZIOAK ETA KOMUNZTADURA**

Maite Oronoz Anchordoquik

Informatikan Doktore titulua eskuratzeko aurkezturiko
TESI-TXOSTENA

Donostia, 2008ko urria.

Lengoaia eta Sistema Informatikoak Saila



Informatika Fakultatea

EUSKARAZKO ERRORE SINTAKTIKOAK DETEKTATZEKO ETA ZUZENTZEKO BALIABIDEEN GARAPENA: DATAK, POSTPOSIZIO-LOKUZIOAK ETA KOMUNZTADURA

Maite Oronoz Anchordoquik Arantza Diaz de Ilarrazaren eta Koldo Gojenolaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2008ko urria.

Etsekoei, bereziki amari eta aitari

Eskerrak ematen:

Gezurra dirudien arren, tesi-lan hau bukatutzat jotzeko ordua iritsi da. Hori ez litzateke posible izango jende askoren laguntzarik gabe, eta horregatik, eskerrak eman nahi dizkiet hauei guztiei:

- IXA taldeko diren eta izan diren lagun guztiei, tesi hau hein handi batean zuena ere bada.
- Zuzendariei, Arantzari eta Koldori, bide luzea izan da, baina horrelako gidari finekin errazagoa da amaierara “sano” iristea.
- HIREKIN taldeari (Montse, Larraitz, Itziar, Arantza, Bertol, Ianire). Izena oso egokia da. NIREKIN sentitu zaituztet.
- Maritxalar & Maritxalar senar-emazteari. Zuekin partekatu nahiko nuke ume honen gurasotasuna.
- Nereari eta Kikeri, beti hor, beti laguntzeko prest. Kateak lotu izan nauenean, zuei giltzarrapoko giltza eskatzea besterik ez zegoen.
- Zuzenketa automatikoa ez, baina tesi-txostenean idatzitakoaren eskuzko zuzenketa fina egin duten guztiei: Arantza, Koldo, Nerea, X. Artola, Mikel Lersundi, Montse eta Larraitz.
- Erroreen detekzioa ahalbidetzeko hainbeste zuhaitz marraztu dituen Larraitz artistari, eta hainbeste errore markatu dituen Edurne Aldasorori. Larraitz, zu lasai, oraindik ere jarraituko dugu hainbeste gus-tatzen zaizkigun erroreen inguruko *eztabaida filosofikoekin*.
- Maxuxi, ni bezala zuhaitz artean erdi galduta ibili arren, beti topatu duelako uneren bat irtetzeko bidea erakusteko. Azkenean atera gara baso zerratutik!
- Marratxoak, mugagabeak, postposizio-lokuzioak... Nire duda linguis-tiko guuzti guztiak pazientzia handiz argitu dituzten hizkuntzalariei: Ruben, Maxux, Mikel, Larraitz, Klara, X. Artola. Hainbeste galdera “txoro” entzun beharra ere!
- Ana Arruarteri bere esku zabaltasunagatik. Hainbat ardura eta lan hartu ditu bere gain, niri tesirako denbora uztearren.
- Latex-en tripak erakutsi dizkidazuen horiei (Aitor Soroa, Oier, Roberto, Gorka, Eli...). Zuei esker honek itxura hobe du.

- A las “Euskericas” (Ainhoa, Sonia, Itziar eta Olatz). Oraingoan afaria nire kontu, baina, hurrengoan, eta laster, Olatzen kontu. Zuk emandako laguntza bueltatzeko asmoa dut, beraz, ez pentsa zurearen amaiera hain urrun dagoenik.
- Maiderri, azala edertzen hartu duzun lanagatik. Zu bai artista!
- Mikel Lersundiri, goizetako kafeak eta postposizioei buruz (postposizio zuzena) dakizuna nirekin konpartitu duzulako.
- Aitziberri eta Izaskun Aldezabali, aditzekin egin duzuen lan guztia nire eskura ipintzeagatik.
- Aitor Sologaistoa, eulien moduan anotazio-amaraunean trabatuta geratu naizenean, askatzen lagundu nauzulako.
- Joxe Mariri eta Itziarri anbiguotasunak ebazten laguntzeagatik.
- Kataluniako lagunei, Jordiri eta Monseri.
- Lagun-koadrilari (Ainhoa, Amaia, Nekane, Maider, Maite, Izaskun, Dolo, Meli eta Jaione). Zuen ondoan lan kontuak alde batera uzten saiatu banaiz ere, noizbehinka nire tesiko “txapak” kafe baten ondoan pazientzia handiz entzun dituzuelako.
- Etxekoei, horrenbeste ordu ebatsi dizkidan lana ulertu ez arren, beti nirekin egon zaretelako.
- Itsasne nire besoetako nexka pollitari. Zure bizitzako lehen bost hilabete hauetan, lana alde batera uzteko aitzaki ezin hobe izan zara.
- Zure izena hemen ikustea espero zenuen horri. Berdin eskertzen dizut egindakoa, baina une honetan buruak ez dit gehiagorako ematen!!

Denoi, eskerrik asko! Mersi!

LINGUAE PRIMITIAE

Hainbeste mendeetan kolpatua
eta aurrera,
han-hemen pozoindua zein desitxuratua,
aurrera,
badirudi zenbaitetan arrastaka,
ailegatzear,
egunera.

Ez luke izan behar
aukera,
ez genuke beharko bultza
denok batera,
ez legoke lotuta egon behar
geure lurraldera,
berezko bizitza merezi du
borroka orotik
kalera.

Horregatik guztiagatik
jalgi plazatik
bidera,
zabaldu hegoak,
egizu jauzi
airera,
aske izan erabat
urrundu mendien
gainera,
ahoan gozatu eta bizi
bihotzetik haratagoko
erabilera,
maitatuko zaitut
betidanik - betirako neure:

Euskera.

Belen Oronoz

“utopia zeruertzean dago; bi urratsez hurbildu natzaio eta berak beste bi atzera egin du. Hamar urrats egin ditut eta zeruertza hamar urrats urrunago daukat. Etengabe ibilita ere, ez dut inoiz harrapatuko. Zertarako balio du utopiak? Horretarako: ibiltzeko...”

Eduardo Galeano. Itzulpena.

Euskaraz egin daitezkeen errore guztiak detektatzea “*utopia*” iruditu arren, tesi-lan honek urrats batzuk egin nahi ditu bidean.

Gaien aurkibidea

MOTIBAZIOA ETA SARRERA	1
I Proiektuaren nondik norakoak	3
I.1 Sarrera	3
I.2 Lanaren kokapena	6
I.2.1 IXA taldea eta ortografia- eta gramatika-zuzentzai- leak	7
I.2.2 IXA taldea eta hizkuntzen irakaskuntza	9
I.3 Aurkezpen orokorra eta helburuak	10
I.4 Tesi-txostenaren eskema eta argitalpenak	13
ERROREEN DETEKZIO ETA ZUZENKETA AUTOMATIKOA	19
II Erroreak, corpus erroredunak, erroreen sailkapena eta ebaluazioa	21
II.1 Sarrera	21
II.2 Errorea	22
II.2.1 Errorearen definizioa	22
II.2.1.1 Desbideratzeak	22
II.2.1.2 Estilo desegokia	25
II.2.1.3 Hitz zuzenen erabilera nahasketak	25
II.2.2 Detekzioa, diagnosia eta zuzenketa	26
II.3 Corpus erroredunak	29
II.3.1 Irizpideak	30
II.3.2 Testu-bilketa	32
II.3.3 Etiketatzeta	34
II.4 Erroreen sailkapena	35
II.4.1 Irizpideak	35
II.4.2 Ezaugarriak	36
II.4.3 Aukeratutako errore-kategoriak	40

II.5	Ebaluazioa	41
II.5.1	Kalitate eskaerak ezartzea	42
II.5.2	Zehaztapena eta diseinua	43
II.5.3	Exekuzioa	45
II.6	Laburpena eta ondorioak	47
III	Erroreak detektatzeko eta zuzentzeko teknikak	49
III.1	Sarrera	49
III.2	Erroreak detektatzeko teknikak	49
III.2.1	Hizkuntza-ezagutzan oinarritutako teknikak (sin- bolikoak)	51
III.2.1.1	Ezagutza morfologikoa (hitz-maila)	51
III.2.1.2	Ezagutza sintaktikoa (esaldi-maila)	53
III.2.1.3	Ezagutza semantikoa	63
III.2.2	Corpusetan oinarritutako teknikak (enpirikoak)	64
III.2.2.1	Metodo estatistikoak	65
III.2.2.2	Ikasketa automatikoa	66
III.3	Erroreak zuzentzeko teknikak	68
III.3.1	Ortografia	68
III.3.2	Gramatika	70
III.4	Gramatika-zuzentzaileak	71
III.5	OLHIA sistemak	73
III.6	Laburpena eta ondorioak	75
IV	Abiapuntua. Hizkuntzaren analisia eta errepresentazioa	81
IV.1	Analisia: katea	81
IV.1.1	Deskribapena	84
IV.1.1.1	MORFEUS, analizatzaile morfosintaktikoa	84
IV.1.1.2	EUSTAGGER, lematizatzaile/etiketatzailea	88
IV.1.1.3	IXATI zatitzailea	89
IV.1.1.4	Mendekotasun-erlazioen ezartzailea	92
IV.1.2	Analisi-katearen egokitasuna informazio ez-grama- tikala lantzeko	96
IV.2	Errepresentazioa: XML bidezko anotazio-amarauna	99
	TESTUINGURU MUGATUKO ERROREEN DETEKZIOA ETA ZUZENKETA	105
V	Patroien erabilera datetako eta postposizio-lokuzioetako erroreak detektatzeko eta zuzentzeko	107

V.1	Sarrera	107
V.2	Transduktoreak eta datak	109
V.2.1	Erabilitako tresna: XFST	110
V.2.2	Datetan gertatzen diren errorearen sailkapena	112
V.2.3	Sistemaren ikuspegi orokorra	115
V.2.3.1	Detekzioa	118
V.2.3.2	Zuzenketa	119
V.2.4	Emaitzak	121
V.3	Murritzapen Gramatika eta postposizio-lokuzioak	124
V.3.1	Erabilitako tresna: CG2	125
V.3.2	Postposizio-lokuzioak	126
V.3.2.1	Definizioa eta sailkapena	126
V.3.2.2	Aukeraketa-prozedura	128
V.3.3	Detekzioa	133
V.3.3.1	Arte/artean	135
V.3.3.2	Aurre	145
V.3.3.3	Bitarte	149
V.3.3.4	Buruz	151
V.3.3.5	Zehar	155
V.3.3.6	Laburbilduz	156
V.3.4	Zuzenketa	157
V.3.5	Integrazioa XUXENG gramatika-zuzentzailean	159
V.4	Laburpena eta ondorioak	160

TESTUINGURU ZABALEKO ERROREEN DETEKZIOA

165

VI	<i>Saroi</i> : mendekotasun-zuhaitzetan informazioa kontsul- tatzeko tresna	167
VI.1	Sarrera	167
VI.2	Arkitektura nagusia	169
VI.2.1	Analisi sintaktikorako modulua	173
VI.2.1.1	Sarrerako dokumentuak	174
VI.2.1.2	Basoa	181
VI.2.1.3	Zuhaitza	193
VI.2.1.4	Irteerako dokumentuak	196
VI.2.2	Kontsulta-erregelen itzulpenerako modulua	200
VI.2.2.1	Kontsulta-lengoaia linguistikoak	200
VI.2.2.2	<i>Saroi</i> ren kontsulta-lengoaia	205
VI.2.2.3	Erregelak. Deskribapena eta adibidea	210

VI.2.2.4	Itzulpen-prozesua	213
VI.2.3	Kontsulta-erregelen aplikaziorako modulua	218
VI.2.4	Emaizten aurkezpenerako modulua	221
VI.2.4.1	Eskaeren analisisa	221
VI.2.4.2	Emaizten aurkezpenera	223
VI.3	Laburpena eta ondorioak	228
VII	<i>Saroiren erabilera komunztadura-erroreak detektatzeko eta hiztegiak aberasteko</i>	229
VII.1	Sarrera	229
VII.2	Komunztadura-erroreak	230
VII.2.1	Definizioa	230
VII.2.2	Aditzaren eta subjektuaren/objektuaren/zehar-objektuaren arteko komunztadura	233
VII.2.3	Komunztadura aldatzen duten fenomenoak	234
VII.2.4	Komunztadura lantzeko zailtasunak	236
VII.3	Komunztadura-erroreen sailkapena	239
VII.4	Informazio linguistikoaren aberasketa	242
VII.4.1	Sintagma-koordinazioa esplizitu egitea	242
VII.4.2	Aditzen aberasketa	244
VII.5	Gramatika	251
VII.6	Aplikazio-estrategia eta errore-erabakia	257
VII.7	Komunztadura-erroreen eta anbiguitasunaren arteko erlazioa aztertze metodologia. Emaizak	257
VII.7.1	Desanbiguaziorako aukera egokienaren bila	259
VII.7.2	Sintaxiko desanbiguazio-maila egokienaren bila, mendekotasun-gramatika erlaxatuz	263
VII.7.3	Aukeraketa eginda, azken probak	269
VII.7.3.1	Corpus erroreduna	269
VII.7.3.2	Corpus zuzena	277
VII.8	Hiztegi aberasketa	285
VII.8.1	Sarrera	285
VII.8.2	Aurreprozesua	286
VII.8.3	Corpuseko eta EH hiztegiko aditzen arteko inkoherentziak	288
VII.8.3.1	Esperimentua	288
VII.8.3.2	Emaizak	290
VII.8.4	EH hiztegian sarrerak gehitzen	291
VII.8.4.1	Esperimentua	291
VII.8.4.2	Emaizak	292

VII.9	Laburpena eta ondorioak	296
VII.9.1	Komunztadura-erroreen detekzioari dagokionez . .	296
VII.9.2	Aditzen aberasketari dagokionez	298
ONDORIOAK ETA ZABALDUTAKO IKERLERROAK		299
VIII	Ondorioak eta etorkizuneko lanak	301
VIII.1	Ondorio nagusiak	302
VIII.1.1	Garapena	302
VIII.1.2	Ebaluazioa	306
VIII.2	Ekarpenak	307
VIII.3	Etorkizuneko lanak	311
VIII.3.1	Erroreen tratamendua	311
VIII.3.2	<i>Saroi</i> ren moduluen hobekuntza	312
VIII.3.3	<i>Saroi</i> ren erabilerak	314
Bibliografia		315
ERANSKINAK		1
A	Gramatika-zuzentzaileak eta OLHIA sistemak	1
A.1	Gramatika-zuzentzaileak	1
A.2	OLHIA sistemak	19
B	<i>Saroi</i> . Klase-diagramak	25
C	<i>Saroi</i> . Erregelen itzulpena	33
C.1	Analisi lexikoa	33
C.2	Erregelen analisi sintaktikoa. TGG eta SZIE	35
C.2.1	Testuingururik gabeko gramatika (TGG)	35
C.2.2	Sintaxiak zuzendutako itzulpen-eskema (SZIE) . .	35
D	<i>Saroi</i> . <i>Document Type Definition</i> (DTDa)	49
E	<i>Saroi</i> . Komunztadura-erroreen detekzioa. Emaitzak.	53

Algoritmoen zerrenda

1	<i>Permutazioak</i> sortzeko algoritmoa.	188
2	<i>Basoa</i> sortzeko algoritmoa.	189

Irudien zerrenda

I.1	Tesi-lanaren kokapena zuzentzaileen ingurunean.	8
I.2	Tesi-lanaren kokapena hizkuntzen irakaskuntzarako ingurunean.	10
II.1	Euskarazko errorearen sailkapena. Katēgoria nagusiak.	39
II.2	Gramatika eratzeiko zikloa. Fluxu-diagrama.	46
II.3	Ebaluazioan erabilitako corpusak eta neurriak.	47
IV.1	Euskararako geruza anitzeko sintaxi partzialeko analizatzailea.	84
IV.2	MORFEUS, analizatzaile morfosintaktikoa.	85
IV.3	EUSTAGGER, lematizatzaile/etiketatzailea.	88
IV.4	IXATI zatitzailea (chunker).	90
IV.5	Mendekotasun-erlazioak adierazten dituzten etiketak.	93
IV.6	EDGK, mendekotasun-erlazioen ezartzailea.	94
IV.7	(9) adibideko esaldiaren mendekotasun-zuhaitza.	95
IV.8	*Zenbait gizonak sintagma erroredunaren analisisa.	99
IV.9	Dokumentu motak anotazio-amaraunean.	102
IV.10	Dokumentu anitzeko anotazio-amarauna (1/2).	103
IV.11	Dokumentu anitzeko anotazio-amarauna (2/2).	104
V.1	Sistemaren arkitektura nagusia.	115
V.2	Sistemaren arkitektura nagusia eta tarteko emaitzak.	117
V.3	Errore-patroia XFSTz definitzeko adierazpen erregularrak. . .	119
V.4	Datak XFSTz sortzeko adierazpen erregularrak.	120
V.5	<i>Buruzen</i> erabilera oker bat deskribatzen duen islapen-erregela baten lehen hurbilpena.	133
V.6	bidaia buruz egituraren analisi morfosintaktikoa, errore etike- tekin.	134
V.7	<i>Buruzen</i> erabilera oker bat deskribatzen duen islapen-erregela.	135
V.8	bidaia buruz egituraren analisi morfosintaktikoa, errore-etike- tekin.	158
VI.1	<i>Saroi</i> ren arkitektura nagusia.	170

VI.2	<i>Saroiaren eskema.</i>	172
VI.3	<i>*Zentral nuklearrak zakar erradiaktiboa eratzen dute</i> esaldiari dagokion mendekotasun-zuhaitza.	175
VI.4	XML anotazio-amaraunaren adibidea (1).	177
VI.5	XML anotazio-amaraunaren adibidea (2).	178
VI.6	Anbigutasuna duen mendekotasun-zuhaitza.	180
VI.7	Mendekotasun-zuhaitz anbiguotik eratorritako zuhaitz ez-anbiguoak.	181
VI.8	Basoa sortzeko algoritmoari jarraituz 1 konbinazioari dagokion mendekotasun-zuhaitza.	187
VI.9	Basoaren domeinu-eredua.	192
VI.10	Mendekotasun-zuhaitzen domeinu-eredua.	196
VI.11	Zuhaitz errepresentazio <i>implizitua</i>	197
VI.12	Zuhaitz-errepresentazio <i>esplizitua</i>	199
VI.13	Osagai-egituretan (1) eta mendekotasun-egituretan (2) oinarritutako zuhaitzak eta nabigazioa.	208
VI.14	<i>Saroiaren</i> kontsulta-lengoaian definitutako komunztadura-erroreen detekziorako erregela bat.	211
VI.15	Erroreen <i>detekziorako</i> erregelaren adibidea.	212
VI.16	<i>Saroiaren erregelak</i> lantzeko domeinu-eredua.	216
VI.17	Analisi-zuhaitza.	218
VI.18	<i>Estrategiak</i> kudeatzeko domeinu-eredua.	220
VI.19	<i>Saroiaren</i> kontsulta-interfazearen diseinua.	226
VII.1	Analisi partzialen adibidea.	238
VII.2	Sintagmen koordinazioa esplizitu egitea. Adibidea.	243
VII.3	Sintagmen koordinazioa esplizitu egitearen ondorioz, zenbait mendekotasun-erlazioen garraiatzea. Adibidea.	243
VII.4	Ezkerraldean dute aditzaren analisia. Eskuinaldean analisi bera <i>aberasketa</i> -prozesuaren ondoren.	245
VII.5	etorri aditza <i>aberasketa</i> -prozesuaren ondoren.	249
VII.6	Mendekotasun-zuhaitzak <i>aberasteko</i> domeinu-eredua.	250
VII.7	Komunztadura-erroreak detektatzeko “aditzen patroiak” erabiltzen dituen erregelaren adibidea.	252
VII.8	Komunztadura-erroreak detektatzeko “zehir-objektuaren sor-kuntza” kontuan hartzen duen erregelaren adibidea.	253
VII.9	Komunztadura-erroreak detektatzeko “perpausaren barruko komunztadura” aztertzen duen erregelaren adibidea.	254

VII.10	Subjektuari dagokionez, komunztatzen duten elementuen kokapen ohikoena zuhaitzetan (kokapena X ikurra erabilita adierazi dugu).	255
VII.11	Komunztatzen duten elementuen kokapen ohikoenaren aldeak (kokapena X ikurra erabilita adierazi dugu).	255
VII.12	Elipsia, aditzari lotuta bi subjektu.	280
VII.13	Corpuseko eta Euskal Hiztegiko aditzak.	287
VII.14	Corpuseko eta EH hiztegiko aditzen arteko inkoherentziak.	289
VII.15	NOR paradigmako aditz laguntzaileak dauzkaten aditzetan inkoherentziak topatzeko erregela.	289
VII.16	EH hiztegiaren sarrerak gehitzen.	292
VII.17	<i>Treebank</i> ean “ <i>ados etorri</i> ” lemadun aditza agertzen ote den aztertzen duen erregela.	293
B.1	Mendekotasun-zuhaitzen <i>basoa</i> (saroia) inplementatzeko erabilitako klase-diagrama.	27
B.2	Mendekotasun-zuhaitza inplementatzeko erabilitako klase-diagrama.	28
B.3	Aditz nagusien aberasketarako eta aditz laguntzaileen ezaugarriak esplizitu egiteko inplementazioan erabilitako klase-diagrama (1/2).	29
B.4	Aditz nagusien aberasketarako eta aditz laguntzaileen ezaugarriak esplizitu egiteko inplementazioan erabilitako klase-diagrama (2/2).	30
B.5	Erregelen itzulpeneko moduluaren inplementaziorako erabilitako klase-diagrama.	31
C.1	<i>Saroiren erregelak</i> lantzeko domeinu-eredua.	38
C.2	<i>Saroiren</i> itzultzailearen sarrerako erregela eta irteeran lortutako kodea.	47

Taulen zerrenda

I.1	Tesi-txostenaren antolaketa.	13
II.1	Euskaraz egiten diren erroreak maiztasunaren arabera sailkatuta.	41
III.1	Teknika sinbolikoen eta enpirikoen arteko desberdintasunak. .	51
III.2	Gramatika-zuzentzaileek erabiltzen dituzten teknikak (1/2). . .	72
III.3	Gramatika-zuzentzaileek erabiltzen dituzten teknikak (2/2). . .	73
III.4	OLHIA sistemek erabiltzen dituzten teknikak.	74
III.5	Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (1/3).	78
III.6	Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (2/3).	79
III.7	Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (3/3).	80
IV.1	EDGK-I sintaxi-analizatzailearen emaitza.	95
V.1	Datetan gertatzen diren errore mota ohikoenak (erroreak letra lodiz eta makoen artean markatu dira).	114
V.2	Garapenerako eta probetarako corpusak.	121
V.3	Ebaluazioko emaitzak.	122
V.4	Alarma faltsuak daten detekzioan.	123
V.5	Anbiguitasunaren eragina datetako erroreak ezagutzerakoan. .	124
V.6	Postposizio-lokuzioen osagaiak.	127
V.7	Postposizio-lokuzioen azterketan erabilitako corpus erroreduna.	129
V.8	Postposizio-lokuzioen azterketan erabilitako euskaltegietako cor- pus erroreduna ikasketa-mailaren arabera zehaztua.	129
V.9	Corpus zuzenean eta erroredunean agerpen kopuru handieneko 20 postposizio-lokuzioak.	130
V.10	Postposizio-lokuzioen erabilera okerrari buruzko informazio- -iturriak.	132
V.11	Dokumentatutako erroreak postposizio-lokuziotan.	132

V.12	Aukeratutako postposizio-lokuzioen erabilera zuzena.	136
V.13	<i>artean</i> OBa duen postposizioaren inguruko erroreak.	138
V.14	<i>arte</i> OBa duen postposizioaren inguruko erroreak.	140
V.15	Garapenerako eta probetarako corpus erroredunak.	141
V.16	Garapenerako eta probetarako corpus zuzenak.	141
V.17	<i>Arte</i> . Emaitzak.	143
V.18	<i>Arte</i> . Alarma faltsuak.	143
V.19	<i>Arte</i> . Anbiguitasunaren eragina emaitzetan.	145
V.20	<i>aurre</i> OBa duen postposizioaren inguruko erroreak.	146
V.21	<i>Aurre</i> . Emaitzak.	148
V.22	<i>Aurre</i> . Alarma faltsuak.	149
V.23	<i>Aurre</i> . Anbiguitasunaren eragina emaitzetan.	149
V.24	<i>bitarte</i> OBa duen postposizioaren inguruko erroreak.	150
V.25	<i>Bitarte</i> . Emaitzak.	151
V.26	<i>Bitarte</i> . Anbiguitasunaren eragina emaitzetan.	151
V.27	<i>buruz</i> OBa duen postposizioaren inguruko erroreak.	152
V.28	<i>Buruz</i> . Emaitzak.	153
V.29	<i>Buruz</i> . Alarma faltsuak.	153
V.30	<i>Buruz</i> . Anbiguitasunaren eragina emaitzetan.	154
V.31	<i>zehar</i> OBa duen postposizioaren inguruko erroreak.	155
V.32	<i>Zehar</i> . Emaitzak.	155
V.33	<i>Zehar</i> . Anbiguitasunaren eragina emaitzetan.	156
V.34	Emaitzen laburpena.	156
V.35	Doitasunaren emaitzak probarako corpusetan.	157
V.36	Postposizio-lokuzioen zuzenketarako eskemaren adibidea.	159
VI.1	Adibideko mendekotasun-zuhaitz anbiguoari dagozkion <i>dep.xml</i> eta <i>deplnk.xml</i> fitxategiak.	184
VI.2	Konbinazio-taula baten adibidea.	185
VI.3	<i>Saroi</i> ren kontsulta-lengoaiaz erantzun daitezkeen zenbait kontsulta.	207
VI.4	<i>Saroi</i> ren kontsulta-lengoaiaz erantzun ezin diren zenbait kontsulta.	208
VI.5	Mendekotasun-zuhaitzak nabigatzeko eta arakatzeko <i>Saroi</i> ren erregeletan erabil daitezkeen <i>eragileak</i>	209
VI.6	Analisi sintaktikoaren oinarri den TGGaren produkzio-erregelak.	215
VI.7	Kontsulta-erregelen itzulpenean erabilitako TGGaren eta domeinu-ereduaren arteko baliokidetasunak.	217
VI.8	Hizkuntzalariek <i>treebank</i> ean egin nahiko lituzketen kontsultak.	222
VI.9	Adibidea. Kontsulta bat modu grafikoan adieraztea.	227

VII.1	Komunztadura-erroreak testu-zatietan.	232
VII.2	Emendiozko juntadura eta numero aldaketa.	235
VII.3	EDGKren eta <i>MaltIx</i> aren ebaluazioa, komunztadurarekin lo- tura zuzena duten mendekotasun-erlazioei dagokienez.	239
VII.4	Aditzarekiko komunztadura-erroreen maiztasuna.	241
VII.5	<i>etorri</i> aditzaren <i>bssak</i>	246
VII.6	<i>etorri</i> aditzak corpusean paradigma bakoitzeko aditz laguntzai- leekin duen agerpen-proporzioa.	248
VII.7	Hiztegiko eta corpuseko etiketen arteko baliokidetasunak.	249
VII.8	EUSTAGGERen desanbiguaziorako aukerak.	258
VII.9	Komunztadura-erroreak eta anbiguotasuna.	262
VII.10	Laburpen-aula. Komunztadura-erroreak eta anbiguotasuna.	263
VII.11	Laburpen-aula (75 esaldi erroredun). Komunztadura-erroreak eta anbiguotasuna.	264
VII.12	Laburpen-aula (75 esaldi zuzen). Komunztadura-erroreak eta anbiguotasuna.	266
VII.13	Sintaxi-anbiguotasuna eta erlaxazioa. Emaizak.	268
VII.14	Corpus erroredunaren banaketa ebaluaziorako.	270
VII.15	Komunztadura-erroreen detekzioa <i>MaltIx</i> a eta EDGK erabilita.	271
VII.16	<i>MaltIx</i> ako (M) eta EDGK-ko (E) alarma faltsuak.	274
VII.17	<i>MaltIx</i> an (M) eta EDGKn (E) alarma faltsuak zein erregelare- kin sortzen diren.	275
VII.18	<i>Treebank</i> aren banaketa ebaluaziorako.	278
VII.19	Alarma faltsuak.	279
VII.20	EH hiztegia falta diren azpikategorizazio-etiketak	282
VII.21	EH hiztegia gehitzeko hautagai diren aditzak.	294
VII.22	EH hiztegia gehitzeko aditzen onartze-tasak.	295
VIII.1	Emaizak	307
C.1	<i>Saroiko</i> erregelen analisi lexikoan erabiltzen diren osagai lexi- koak.	34
C.2	Analisi sintaktikoaren oinarri den TGGaren definizio formala.	36
C.3	Analisi sintaktikoaren oinarri den TGGaren produkzio-erregelak.	37
D.1	Zuhaitz <i>inplizitu</i> aren DTD-zehaztapena.	52
D.2	Zuhaitz <i>esplizitu</i> aren DTD-zehaztapena.	52
E.1	Komunztadura-erroreen detekzioa. Emaizak.	59
E.2	Komunztadura-erroreak ez detektatzeko arrazoiak.	61
E.3	Komunztadura-erroreen detekzioa esaldi zuzenetan. Emaizak	66

E.4	Komunztadura-erroreen detekzioa. Emaidzak (2).	71
E.5	Sintaxi-anbiguotasuna eta erlaxazioa. Emaidzak.	71

MOTIBAZIOA ETA SARRERA

I. KAPITULUA

Proiektuaren nondik norakoak

I.1 Sarrera

Nire bizitzaren egunik zoriontsuena izango da, eguna nik oso ongi euskaraz ulertzen, idatzi, eta hitz egiten dudala. Betikoa euskara (...) zait baina denbora inoz ezan nuen. Lehen oso txikia nintzelako; gero asko ikasi behar nuen, gero ez dakit zer... Iaz, Unibersitatean erabaki nintzen ordua zela! eta euskaraz ikastera hasi nuen. Orain oso pozik nago! Oraindik ez dut hitz, idatzi...baina ongi nabil. Etorriko du.

AEK-ko 2. urratseko euskara-ikasle bat.

Irakurri berri dugun testuan euskara-ikaslearen asmoa eta igorri nahi duen mezua ulertzeko gai gara: komunikazioa ematen da. Adibidean, helburutzat komunikazioa duen idatzizko zeinu-sistema batek osatzen du *hizkuntza*. Zeinu horiek gramatika (edo arau) batzuen arabera antolatu ohi dira. Arauak jarraitzean egiten du, hain zuzen ere, testuak huts. Ez ditu euskararen arauak zuzen, edo egoki, erabiltzen. Testuko gramatikak, ortografiak eta ortotipografiak euskarazko arauak jarraituko balitu, testua ulergarriagoa litzateke, eta honen haria errazago jarraituko genuke. Hizkuntzaren araua kontuan hartuz, komunikazioa hobetu egiten da.

Hala ere, ikaste-prozesuan arautik aldentzea, hots, erroreak egitea, ezinbestekoa da. Jendeak ezin du hizkuntzarik ikasi lehenengo sistematikoki errorerik egiten ez badu (Dulay *et al.*, 1982). Erroreak garai batean gaizki ikusiak baziren ere, ikaste-prozesuaren parte kontsideratzen dira gaur egun. Hizkuntza zuzen ikasteko, ikasketa-prozesuaren uneren batean egindako erroreak detektatzea, eta beharrezkoa kontsideratzen denean zuzentzea, garrantzitsua da.

Euskararen kasura etorrita, arautik aldentzen den euskararen erabileran eragin zuzena duten faktoreak biltzen dira ondorengo ataletan:

1. **Euskararen estandarizazio-prozesua.** Euskal hizkuntzaz arduratzen den Euskaltzaindia¹ (1919) erakunde ofizialak eman ohi ditu hizkuntzaren normalkuntzarako arauak eta, hiztegi- eta morfosintaxi-mailako arau ugari zehaztu dituen arren, oraindik ez dira asko sintaxi-mailan definitutakoak. Arau akademikoen faltan, zenbait hizkuntza-egitura gaizki erabiliak ote diren edo ez erabakitzea zaila izaten da batzuetan. Ugartebururen arabera (2004) arau akademikoaz arautu gabe dauden egituren eta erabileren inguruan maiz hartu behar dira erabakiak. Erabaki horiek hartzeko orduan ezinbestekoa da ereduren bat gogoan hartzea eta erabilerari begiratzea. Tesi-lan honetan arau akademikoen falta sumatu dugunean, horixe egin behar izan dugu, egituraren erabilera ohikoenei begiratu.
2. **Euskalkiak.** Zuazoren azterketak (2003) dio Euskal Herrian sei euskalkitan egiten dela hitz. Aberastasun honen eraginez, adibidez, x eta h hizkien erabilerari buruzko auziak sortu ziren (Iparraldeko euskaltzainak, 1978) eta horiek ebazteko asmotan, euskara bateratzeko beharra nabaritu zuen Euskaltzaindiak. Baterakuntzarako bide horretan, Euskaltzaindiak euskara idatzian ortografia bakarra onartzea erabaki du, batuan nahiz euskalkietan, “izan ere, euskara hizkuntza bakarra baita, ez hizkuntza asko”². Alfabetatu gabeko euskaldun askok euskalkia barneratua izanik ere, arauak ezagutzen ez dituenek, euskara batuan egitura “ez-estandarrik” erabiltzeko joera du. Egitura horietako batzuk euskara batuan erabilera ez gomendatukoak kontsideratzen dira.
3. **Inguruko erdaren eragina.** Euskara, bi erdara indartsuz inguratuta bizi da: gaztelania hegoaldean eta frantsesa iparraldean. Horrek, euskaraz ari garela, euskararen arau eta kodeetan aritu beharrean, batzuetan gainerako kodeetan aritzeak eragiten du. Gartziaren arabera (2005) “erdararen sistemak ez dezan (hasi fonetikatik eta estiloraino, maila guztietan) euskararen sistema —itsuski gero eta *sistematikoagoki*, hain zuzen— ordeztu”.

Ezaugarri hauen guztien eraginez euskaraz gaizki erabiltzen diren egiturak detektatzea eta zuzentzea garrantzitsua iruditzen zaigu. Gainera, ataza betetzeko testuetan egingo dugun egitura okerrean azterketa enpirikoak

¹<http://www.euskaltzaindia.net>

²Euskaltzaindiaren 137. adierazpena: *Euskalkien erabileraz*.

arauen zehaztapenean lagun lezake, euskararen baterakuntzarako bidean gure ekarpena eginez.

Erroreak edo gaizki erabilitako egiturak detektatzea ataza oso garrantzitsua da hizkuntzaren tratamendu automatikoko ondorengo bi alorretan:

- **Ortografia- eta gramatika-zuzentzaileetan.** Zuzentzaile ortografikoak (*spell checker*) eta gramatika-zuzentzaileak (*grammar checker*) erabiltzen duten informazio linguistikoko kopuruagatik bereizten dira. Zuzentzaile ortografikoek bi zuriuneren arteko karaktere-kateen zuzentasuna egiaztatzen dute, unitate bakarreko hitzena, hain zuzen ere. Gramatika-zuzentzaileek, berriz, hitzaren testuinguruko informazioa erabiltzen dute, sintaxiaren inguruko informazioa. Esaldia izan ohi dute azterketarako unitatea. Hori horrela izanik, ohikoa da zuzentzaile ortografiko batek egokitzen hartutako karaktere-katea, gramatika-zuzentzaile batek oker moduan markatzea hitza testuinguru zehatz batean okerra izateagatik. Gramatika-zuzentzaile gehienek ez dituzte sintaxi-mailako erroreak soilik nabarmentzen. Horietaz gain, puntuazio-erroreak, letra larriekin eta zenbaki/daten formatuekin gertatzen diren erroreak ere, detektatzen dituzte. Zuzentzaile automatikoen testua erretean bila aztertzen dute, errore posiblea markatu egiten dute, eta azalpena eman ondoren, ordain zuzenen zerrenda proposatzen dute.
- **Ordenagailuz Lagundutako Hizkuntzen Irakaskuntzan³ (OLHI).** *Erroreen Analisiak* (EA) hizkuntzen eskuratzeari buruzko informazioa ematen du. Gainera, ikasleek egin ditzaketen erroreak ezagututa, ordenagailuz lagundutako hizkuntzen irakaskuntzarako aplikazio eraginkorragoak eta adimentsuagoak egin ditzakegu, eta ikasleen eta sistemen arteko *feedback* edo berrelikadura sendotu. Sistema horiek normalean mugatuak dira erabiltzen duten hizkuntzaren gramatikari dagokionez. Orokorrean, ez dute hizkuntza osoki lantzen, *oinarrizko* hizkuntza baizik. Oinarrizko hizkuntzan erroreak detektatu ahal izateko, maiz, analizatzaile edo *parser* bat erabili behar izaten da. Holland *et al.* (1993) egileen arabera “the use of parsers in language instruction is commonly referred to as intelligent CALL⁴ or ICALL”. Euskaraz OLHI Adimenduna edo OLHIA.

Gramatika-zuzentzaileetan erroreak automatikoki detektatzea eta zuzentzea ezinbestean egin beharreko lana da. Ordenagailuz Lagundutako Hizkun-

³**Irakaskuntza** terminoa modu honetan idatzita, *ikaskuntza* eta *irakaskuntza* terminoak bildu nahi ditugu.

⁴*CALL: Computer Assisted Language Learning.*

tzen Irakaskuntzarako sistemetan, ataza honek garrantzi handia izan dezake. Hori horrela izanik, tesi-lan honetan garatuko ditugun tresnak eta gramatikak etorkizunean sortuko diren euskararako gramatika-zuzentzaileetan eta hizkuntzen irakaskuntzarako sistema adimendunetan erabili ahal izango dira. Bi alorrak erabat aplikatuak dira, eta gure ustez, CALL alorrari ten Hacken-ek (2008) eskaintzen dizkion hitzok ongi egokitzen zaizkie:

«CALL ... is very much in line with modern trends in computational linguistics because it aims for the solution of real-life, practical problems rather than abstract, general problems such as understanding natural language »

Aipuaren arabera, alor horiek hizkuntzalaritza konputazionalaren egungo joerarekin lotura handia dute, bizitza errealeko arazo praktikoei irteera ematea baitute helburu, eta ez, lengoaia naturala ulertzearen moduko arazo orokor eta abstraktuak ebaztea.

Hurrengo atalean, IXA taldeko lanen artean kokatuko dugu tesi-proiektu honetako jomuga: *euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena*. Zehazkiago, taldearen barruan jorratzen diren ondorengo alorretan: ortografia- eta gramatika-zuzentzaileetan eta hizkuntzen irakaskuntzan. Tesi honetan egindako ekarpenek alor horiek aberastuko dituztelako ustean gaude.

1.2 Lanaren kokapena

Tesi-lan hau Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldean sortu da. Taldeak hogeitaz inguru daramatza hizkuntzaren tratamendu automatikoa (Lengoaia Naturalaren Prozesamendua —LNP— izenez ere ezagutua) egiten, eta urte horietan batez ere euskara landu dugu. Denbora tarte horretan hizkuntzalarien eta informatikarien elkarlanari esker, euskararako sortutako baliabideak eta tresnak⁵ ugariak izan dira. Hala nola, *Euskararen Datu-Base Lexikala* (EDBL) (Aldezabal *et al.*, 2001), MORFEUS analizatzaile morfologikoa (Aduriz *et al.*, 1998), hainbat analizatzaile sintaktiko (ikus informazio zehatzagoa IV. kapituluan), corpusak, hiztegi elektronikoak, sare semantikoa... Baliabide horiek oinarrian hartuta, errearen tratamendu automatikoan lanean gabiltza 90. hamarkadaz geroztik.

Erroreen tratamendua aplikazio zehatzen bat buruan edukita egin daiteke (zuzentzaile automatiko bat edo OLHIA sistema bat kasu) eta horien

⁵Informazio zabalagoa <http://ixa.si.ehu.es> web orrian.

berezitasunetara egokitu; edo modu independente eta orokorrean, eta ondoren, behar izanez gero, moldatu. Azken ikuspuntuarekin sortu da esku artean dugun tesi-lana. Ildo beretik zihoan Gojenolaren doktoretza-tesia (Gojenola, 2000), zeinetan euskararako testuingururik gabeko gramatikan baterakuntzarako murriztapenak erlaxatzen ziren erroreak detektatzeko. Eta, bide bera jarraitzen du garapenean dagoen Arrietaren doktoretza-tesiak (Alegria *et al.*, 2006). Azkenekoan erroreak detektatzeko asmoarekin corpusetan oinarritutako teknikak (ikasketa automatikoa, kasu honetan) erabiltzen dira. Horietaz gain, aipatutako bi alorrei egokitzen zaien lana aspaldi ari da egiten taldean, eta horixe da, ondorengo bi azpiataletan laburbilduko duguna.

I.2.1 IXA taldea eta ortografia- eta gramatika-zuzentzaileak

Euskararako *Xuxen* egiaztatzaile/zuzentzaile ortografikoa garatu ondoren (Agirre *et al.*, 1992), azken urteotan XUXENG gramatika-zuzentzailea euskaldun guztien eskura ipintzea du helburu IXA taldeak. Errore sintaktikoak era askotakoak dira eta ortografikoak baino konplexuagoak detektatzeko, beraz, analisi- eta inplementazio-lan sakona eskatzen dute. Hementxe egin dezake tesi-lan honek bere ekarpena. Sintaxi-erroreak detektatzeko eta zuzentzeko definituko ditugun erregelak, XUXENG zuzentzailean integratu ahalko dira.

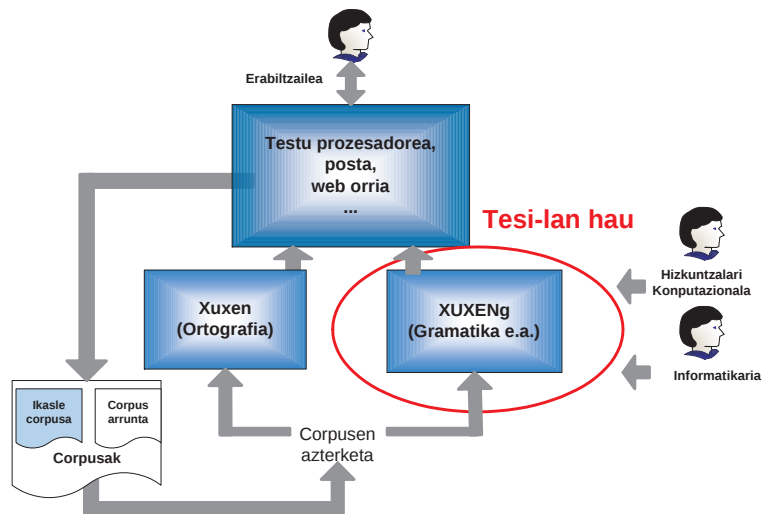
Tesi-lana ortografia- eta gramatika-zuzentzaileen ingurunean zein tokitan kokatuko litzatekeen adieraztea du helburu I.1 irudiak. Irudiko elementuak labur-labur azalduko ditugu ondorengo lerroetan.

*Xuxen*⁶ egiaztatzaile/zuzentzaile ortografikoa 1992 urtean sortu zen euskararako analizatzaile morfologikoaren garapena aurreratua zegoenean (Urkia, 1997; Alegria, 1995). Oinarrian Kukich-en (1992) ideia hau zuen: deskribapen morfologikoa egina dagoenean, erraza da ortografia-zuzentzaile bat garatzea. Analizatzaileak estandar moduan analizatzen dituen hitzak zuzenak lirateke, eta ezagutzen ez dituenak okerrak (ikus informazio zabalagoa III. kapituluan). Orduz geroztik, zuzentzailea datu berriei elikatua izan da eta testu-prozesadore desberdinetarako, web-erako eta OCR-rako⁷ bertsioak sortu dira. 2008 urtean, GNU/Linux munduko aplikazioetan erabilgarria den *Xuxen*en bertsio bat jarri da eskuragarri (Alegria *et al.*, 2008).

Morfologia edo morfosintaxia lantzen jarraitzen badugu ere, hainbat tesi-lanek egiaztatzen duten moduan, azken urteotan IXA ikerketa-taldean sintaxian lan handia egin da. Besteak beste honakoak egin dira: euskararen murriztapen-gramatikari buruzkoa (Aduriz, 2000), Murriztapen Gramatikaren bidez hiztegiko definizioen analisi sintaktikoa (Arriola, 2000), aditz-azpi-

⁶<http://www.xuxen.com> helbidean topa ditzakezue *Xuxen*en bertsio guztiak.

⁷*Optical character recognition* edo karaktere-ezagutze optikoa.



I.1 Irudia: Tesi-lanaren kokapena zuzentzaileen ingurunean.

kategorizazioaren azterketa (Aldezabal, 2004), euskararen sintaxia lantzeko oinarritzko baliabideen garapena (Gojenola, 2000), eta azkenik, (Aranzabe, 2008) dependentzia-gramatiken formalismoa jarraituz garatutako analizatzailer sintaktikoa. Sintaxi-lan horiek baliatuz, gramatika-zuzentzailea sortzeko bidean lehen urratsak egin ditugu.

XUXENg euskararako gramatika-zuzentzailearen lehen prototipoan (Otegi, 2003) informazio linguistikorik gabe detekta zitezkeen estilo- eta puntuazio-erroreak integratu ziren Microsoft Word-en (informazio zabalagoa V. kapituluaren). Hala nola, puntu eta komaren ondorengo espazioak, esaldi luzeegiak, letra larriaren erabilera puntuazio-marken ondoren... Bigarren prototipo batean (Otegi, 2006), informazio linguistikoa behar duten erroreen integrazioa egin zen. Hain zuzen ere, tesi-lan honetan sortu dugun postposizio-lokuzioetan gertatzen diren erroreak detektatzeko gramatikaren lehen bertsioa txertatu zen. Horretarako, hizkuntz informazioa lortzeko beharrezko tresnak integratzeko lana egin behar izan da. Tokenizatzailea, analizatzailer morfosintaktikoa, desanbiguaziorako modulua eta errore sintaktikoak detektatzeko gramatika liburutegi dinamiko batean jarri dira.

I.2.2 IXA taldea eta hizkuntzen irakaskuntza

John Nerbonne-k LNPa hizkuntzen irakaskuntzaren alorrean nola erabil daitekeen aztertu zuen (Nerbonne, 2003) lanean. Haren arabera, “a central technical issue in NLP applied to CALL has been that of recognizing, diagnosing and reacting to students’ errors”. Irakaskuntzaren alorrean, ordea, ezin da errekeen detekzioa eta zuzenketa egokitzapenik gabe erabili, besteak beste, ez baita komenigarria ikasleei errore guzti-guztiak adieraztea eta zuzentzea. Aztertzeko arazo pedagogikoa da honakoa: hizkuntzaren ikasketaren fase bakoitzean ikasleen zein errore detektatzea eta zuzentzea den gomendagarria aztertu behar da, baita erroreei buruzko informazioa zein modutan erakutsiko zaien ere. Azken bi gaien inguruko hausnarketa eta analisi sakona egin ondoren, tesi-lan honetan egingo dugun errekeen detekzioa eta zuzenketa lana IXA taldean garatutako OLHIA tresnetan integratu ahal izango da.

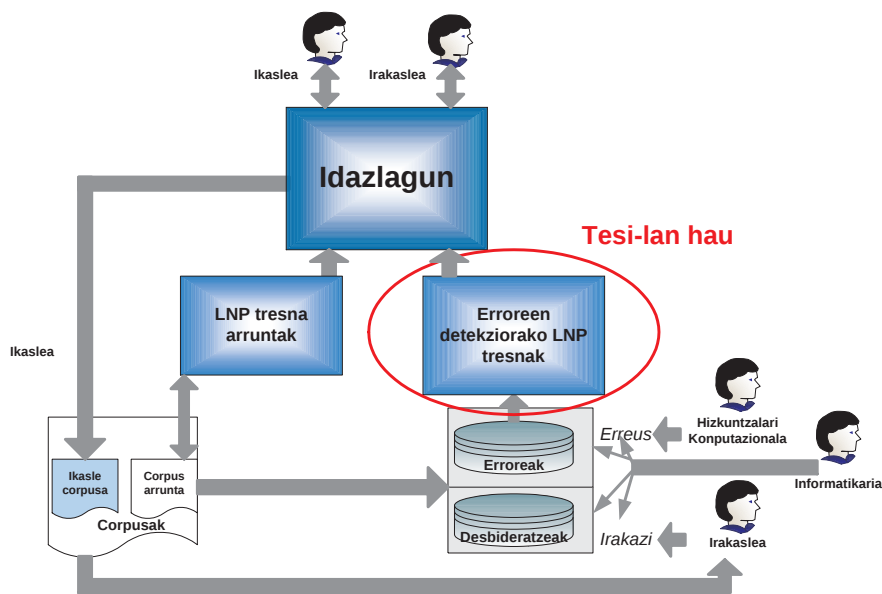
Gure ustez, tesi-lan hau ongi kokatzen da IXA taldeko irakaskuntzarako ingurunean. Jarraian deskribatuko dugun ingurune horretan tesiak duen toki I.2 irudian zehaztu dugu.

IXA taldean ordenagailu bidezko irakaskuntzaren alorra 90. hamarkadan hasi zen jorratzen Díaz de Ilarrazaren (1990) tesiarekin. Orduz geroztik, lan handia egin da hizkuntzen irakaskuntzaren eta ikasleen hizkuntzaren azterketaren inguruan. Horren lekuko dira argitalpen anitz eta Maritxalarren (1999) tesi-lana. Garapenean dauden bi tesi-proiektu, bata ordenagailu bidezko errekeen analisiaren ingurukoa (Aldabe *et al.*, 2005b), eta bestea, ariketen sorkuntza automatikoaren ingurukoa (Aldabe *et al.*, 2006), lan-eremu honetan kokatzen dira.

IXA taldean hizkuntzen irakaskuntzarako aplikazio eta sistemak garatu ditugunean, ikasle, irakasle eta hizkuntzalari konputazionalen arteko elkarrekintzari garrantzi handia eman diogu. Irudian (ikus I.2 irudia) Idazlagun⁸ web-aplikazioa ageri da. Idazlagunen zenbait tresna linguistiko integratzen dira; besteak beste, hiztegiak, aditzen kontsulta, hitz edo lema konkretuen erabilera-adibideen kontsulta, datak, zenbakien idazkera. . . Hauek guztiek idazketa-prozesuan ikaslea laguntzea dute helburu. Alor honetan egingo diren tresnek informazio-iturritzat ikasleen testuak izan ditzaten garrantzitsua denez, corpusak biltzen, errore tipikoen bilduma osatzen, eta errereak etiketatzen, lan handia egin dugu. Erroreei buruz bildutako informazio teknikoa ERROREAK⁹ izeneko datu-basean gordetzen dugu (Aduriz *et al.*, 2002b) eta informazio psikolinguistikoa DESBIDERATZEAK datu-ba-

⁸Erabilgarri <http://sisx04.si.ehu.es/idazlagun> helbidean

⁹ERROREAK datu-basea *Erreus* web-aplikazioaren bitartez da atzigarri. <http://ixa.si.ehu.es/Erreus>



I.2 Irudia: Tesi-lanaren kokapena hizkuntzen irakaskuntzarako ingurunean.

sean¹⁰. Irakaskuntzarako sistemetan erroreen detekzioarako baliabideak erabiltzeko beharrezkoa den egokitzapena egiteko, azken datu-basean biltegiratutako informazioa erabili ahal izango da.

I.3 Aurkezpen orokorra eta helburuak

Tesi-lan honen helburu nagusia *euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena* da, aztertu dugun moduan, euskararen erabilera egokirako erroreak automatikoki detektatzea eta zuzentzea garrantzitsua baita. Garatutako baliabideak, euskararako gramatika-zuzentzaileetan eta hizkuntzen irakaskuntzarako sistemetan erabiliko dira. Helburua gauzatzeko asmoz pauso hauek emango ditugu:

- Erdaretan erabilitako tekniken gainbegiraketa doitasun handia eskainiko duten tekniken bila.

¹⁰ DESBIDERATZEAK datu-basea *Irakazi* web-aplikazioaren bitartez atzi daiteke. <http://ixa.si.ehu.es/ikasleDB/menua>

Azterketa bibliografikoa egingo dugu erdaretan erabili izan diren eta egun erabiltzen diren teknikak ezagutzearen. Azterketaren helburua doitasun altua eskaintzen duen, eta IXA taldeko baliabideetara eta beharretara ongi egokitzen den teknika (edo teknika multzoa) aurkitzea da. Egindako gainbegiratua amaitutakoan, hartutako erabakia arrazoituko dugu.

- Ezagutza linguistikoaren azterketa eta erabilera.

Erroreen azterketa testuinguru erreal batean egingo dugu, eta ez laborategiko esaldiekin. Horrek estaldura handiko tresna sendoak eta ezagutza linguistiko aberatsa eskatzen ditu. Beraz, IXA taldeko analisi-katearen (Aduriz *et al.*, 2004) eta analisisen errepresentaziorako egituren erabilera (Artola *et al.*, 2000, 2004a) funtsezkoa da (ikus IV. kapitulua). Oinarritzko ezagutza linguistikoaz gain, aztertutako egitura bakoitzak zein hizkuntza-baliabide espezifiko behar duen ikertu beharko dugu kasuz kasu, eta tesian erabilitako tresnak haien erabilerarako egokitu.

- Hizkuntza-egituren eta hauetan erroreak detektatzeko eta zuzentzeko erabiliko ditugun tekniken/tresnen analisia. Egituren testuinguruaren araberako aukeraketa.

Garrantzitsua da aztergai izango ditugun hizkuntza-egitura sintaktikoen aukeraketa ongi egitea. Aukera corpus erroredunetan duten agerpen maiztasunaren arabera, eta haien interes linguistiko eta praktikoaren arabera, egingo dugu. Gaizki erabiltzen diren egiturak dokumentatuta, haien detektziorako eta zuzenketarako egokiena iruditu zaigun tresna hautatuko dugu. Sintaxi-mailako erroreak lantzea helburu izanik, bi kasu nagusi aurreikusi ditugu. Alde batetik, testuinguru edo leiho mugatuan gertatzen diren erroreak lantzeko, esaldiko ordena linealean jarraian dauden bospasei hitz aztertzeko gai diren tresnak beharko ditugu. Bestetik, errorea detektatzeko esaldiko ordenan urrun dauden elementuak aztertu behar direnean —testuinguru edo leiho zabalak— esaldien analisi-zuhaitzak arakatzeko iruditzen zaigu egokiena. Testuinguru lokaleko fenomeno linguistikoak arakatzeko tresnak sortuta daude eta eskuragarri ditugu. Analisi-zuhaitzetan egiturak bilatzeko tresna egokirik topatu ezean, garatu egin beharko dugu.

- Gramatiken garapena.

“Gramatiken garapena” izenburuarekin aukera inplizitu bat adierazi dugu: tesi-lanean erreoren detekzioarako hizkuntza-ezagutza oinarritutako teknikak erabiliko ditugu, hots, hizkuntzari buruzko ezagutza erregeletan edo beste adierazpideetan modu esplizituan kodetzen duten teknikak. Erregelak bilduko dituzten gramatikak idaztea izango da hurrengo urratsa erreoren detekzioarako bidean. Gramatiketan doitasun altua lortzeak lehentasuna izango du estalduraren aurrean, erabiltzaileak ezin baitu bere hizkuntza-ezagutza dudatan jarri.

- Gramatiken ebaluazioa corpus erroredun eta zuzenetan.

Testu errealekin lan egitea erabaki dugunez, eta gainera, gure emaitzen erabiltzaile posibleek euskarari buruz duten ezagutza dudatan ipin dezaketenaren arriskua dagoela jakinda, erreoreak detektatzea bezain garrantzitsua da, gure aburuz, zuzenak diren egiturak erroretzat ez jotzea. Tresnek sortutako alarma faltsuen kopurua ebaluatzeko corpus erroredunak erabiltzeaz gain, corpus zuzen handiak ere erabiliko ditugu. Jakin badakigu, ordea, erabaki honek ebaluazioko emaitzetan eragin negatiboa izan dezakeela.

- Erreoren detekzioaren eta anbiguotasunaren arteko erlazioaren azterketa.

Ukaezina da hizkuntzaren analisi automatikoan (hizkuntzan bertan bezala) anbiguotasunak duen garrantzia. Corpus erroredunen analisirako erabilitako zenbait tresnatan desanbiguazio-maila parametrizagarria denez, anbiguotasun morfosintaktikoak eta anbiguotasun sintaktikoak erreoren detekzioan duten eragina aztertuko dugu. Urratsez urrats eta metodologia bati jarraituz egingo dugu aipatutako analisisia.

- Azpiegitura-lanen garapena: corpusak eta ebaluazio-metodologia.

Hizkuntza-baliabideez gain, gure lanerako ezinbestekoa izango da corpusak biltzea, horietan topatutako erreoreak sailkatzea, eta lortutako emaitzen ebaluaziorako irizpideak ezartzea. Corpus erroredunak jasotzea lan neketsua izan ohi da.

Tesi-txostenean deskribatuko dugun lana aurkeztu dugu eta gure helburuak azaldu ditugu. Txosteneko edukiak antolatzeke moduaren berri emango dugu hurrengo atalean.

I.4 Tesi-txostenaren eskema eta argitalpenak

Tesi-txostena bost ataletan banatu dugu, bertako zortzi kapituluak haietan bilduaz, I.1 taulan agertzen den eskemari jarraituz (taularen ezkerrean atalak eta eskuinaldean kapituluak zehaztu ditugu). Multzokatze honen nondik norakoak eta haietan agertzen diren kapituluetakako bakoitzean landu ditugun gaiak labur-labur deskribatuko ditugu ondorengo lerroetan.

MOTIBAZIOA ETA SARRERA	I. Proiektuaren nondik norakoak
ERROREEN DETEKZIOA ETA ZUZENKETA AUTOMATIKOA	II. Erroreak, corpus erroredunak, erroreen sailkapena eta ebaluazioa III. Erroreak detektatzeko eta zuzentzeko teknikak IV. Abiapuntua. Hizkuntzaren analisia eta errepresentazioa
TESTUINGURU MUGATUKO ERROREEN DETEKZIOA ETA ZUZENKETA	V. Patroien erabilera datetako eta postposizio-lokuzioetako erroreak detektatzeko eta zuzentzeko
TESTUINGURU ZABALEKO ERROREEN DETEKZIOA	VI. <i>Saroi</i> : mendekotasun-zuhaitzetan informazioa kontsultatzeko tresna VII. <i>Saroi</i> ren erabilera komunztadura-erroreak detektatzeko eta hiztegiak aberasteko
ONDORIOAK ETA ZABALDUTAKO IKERLERROAK	VIII. Ondorioak eta etorkizuneko lanak
ERANSKINAK	A-(III. Teknikak) Gramatika-zuzentzaileak eta OLHIA sistemak B-(VI. <i>Saroi</i>) Klase-diagramak C-(VI. <i>Saroi</i>) Erregelen itzulpena D-(VI. <i>Saroi</i>) <i>Document Type Definition</i> (DTDa) E-(VII. Aplikazioak) Komunztadura-erroreen detekzioa. Emaizak

I.1 Taula: Tesi-txostenaren antolaketa.

Lehen atala esku artean dugun kapituluak osatzen du. Tesian egindako lanaren helburuak zehaztea, lana IXA taldean egiten denaren barruan kokatzea, eta txostenean topatuko dugun informazioaren berri ematea du helburu.

Bigarren atalean *erroreen detekzio eta zuzenketa automatikoarekin* zerikusia duten hiru kapitulu bilduko ditugu. II. kapitulan erroreen kontzeptua definituko dugu bere aldaera guztiekin batera, eta erroreen detekzioaren, diagnosiaren eta zuzenketaren arteko bereizketa egingo dugu. Corpus erroredunen inguruan arituko gara gero, beranduago hauetan detektatutako erroreen sailkapena aztertzeke; kapituluak amaitzeko, ebaluazio-prozesuaren

zehaztapena egingo dugu eta honen exekuzioa deskribatuko dugu. Erroreak detektatzeko eta zuzentzeko erabiltzen diren teknikai gainbegiratu emango diegu III. kapituluan. Aztertutako teknikak bi multzotan banatuko ditugu: hizkuntza-ezagutza oinarria duten tekniketan eta corpusetan oinarria dutenetan. Azterketa bibliografikotik ateratako ondorioen artean lan honetarako garrantzi handikoa da hau: hizkuntza-ezagutza oinarritutako teknikak erabiliko ditugu. Erabakia ahalik eta hobekien arrazoitzen sailatuko gara. Atal honetako azken kapituluan, IV. kapituluan, erroreen bila corpus erroredunak zein zuzenak automatikoki analizatzeko erabili dugun analisi-katea eta honen osagaiak azalduko ditugu. Informazio linguistikoaren trukerako XML bidezko errepresentazioak berebiziko garrantzia du tesi-lanean, sortutako gramatiken eta tresnen sarrera-dokumentuak horrela kodetuta baitaude. XML bidezko anotazio-amarauna deskribatuz amaituko dugu IV. kapitulua.

Hirugarren atalari *testuinguru mugatuko erroreen detekzioa eta zuzenketa* izena eman diegu eta berau kapitulu bakarrarekin, V. kapituluarekin, osatuta dago. Kapitulu horretan testuinguru sintaktiko lokalari erreparatuz detekta daitezkeen egitura sintaktikoetako erroreak (leiho mugatukoak) landuko ditugu. Azterketarako aukeratu ditugun fenomeno linguistikoak datak eta postposizio-lokuzioak dira. Egiturotan erroreak detektatzeko eta zuzentzeko hizkuntzaren tratamendu automatikoaren munduan ezagunak diren bi tresna erabiliko ditugu: Xerox Finite State Tool (XFST) (Karttunen *et al.*, 1997; Aït-Mokhtar eta Chanod, 1997) datetako erroreak lantzeko eta Murritzapen Gramatika (MG) (Karlsson *et al.*, 1995) postposizio-lokuzioen erabilera okerrak tratatzeko.

Laugarren atalean *testuinguru zabaleko erroreen detekzioa* landuko dugu. Testuinguru zabalekoak deitu ditugu aplikazio-eremua esaldi osoa duten egiturak, zeintzuetan landu beharreko osagaiak esaldian edozein ordenatan ager daitezkeen, eta edozein distantziatar. Modu horretakoa da landuko dugun fenomenoak: subjektu, objektu eta zehar-objektu funtzioa betetzen duten osagaien eta aditzaren arteko komunztadura. Esan dugunez, egitura horietarako esaldiaren analisi-zuhaitza lantzea da egokiena. Ataza honetarako tresna aproposik topatu ez dugunez, *Saroi* tresna sortu dugu. *Saroi* tresna orokorra da, erroreen tratamendua egiteko ez ezik, analisi-zuhaitzetako egituren gaineko kontsultak egiteko eta zuhaitz horietan egitura linguistikoak bilatzeko balio duena. *Saroi* tresna VI. kapituluan deskribatuko dugu, eta bertan, tresnako modulu bakoitzaren diseinua eta inplementazioa azalduko ditugu. Tresnari eman dizkiogun erabilerak VII. kapituluan deskribatuko ditugu: komunztadura-erroreen detekzioa eta hiztegien aberasketa aditzen azpikategorizazioari dagokionez.

Tesi-txostena amaitzeko, ateratako zenbait ondorio eta zabaldu ditugun

ikerkuntzarako zenbait bideren berri emango dugu VIII. kapituluan.

Ezin aipatu gabe utzi tesi-txostenarekin lotutako eranskinak. Eranskine-tako bakoitzean tesi-txosteneko kapituluren batean azaldutakoak zehatzago deskribatuko dira:

- A eranskinak [gramatika-zuzentzaileak eta OLHIA sistemak](#) du izena eta bertan III. kapituluan aztertutako sistemak zehatzago deskribatu-ko dira.
- B eranskinean, *Saroi* tresnaren (VI. kapitulua) [klase-diagramak](#) topa-tuko ditugu.
- C eranskinak *Saroi*ren kontsulta-lengoaia inplementatzeko definituta-ko itzultzailearen [analizatzaile lexikoa](#) eta [analizatzaile sintaktikoa](#) des-kribatuko ditu.
- D eranskinean XML bidezko anotazio-amarauna osatzen duten doku-mentuen sintaxia definituko dugu *Document Type Definition* (DTD) bat dela medio.
- E eranskinean komunztadura-erroreak detektatzerakoan lortutako emai-tzak zehatzago ditugu (VII. kapituluan emaitzak laburtuta erakutsiko ditugu).

Sarrera-kapitulu honi bukaera emateko, tesi-lanarekin lotutako eta egi-leen artean doktoregaia duten zenbait argitalpenen berri emango dugu, ho-rietako bakoitza dagokion kapituluarekin lotuz. Zuzen-zuzenean tesiari lo-tzen zaizkion artikulua aipatu ditugunean, erreferentzia osoa idatzi du-gu. Zerikusia dutenak zerrendatu ditugunean, berriz, izenburua eta aipame-na soilik idatzi ditugu, irakurleak hala nahi izanez gero, bibliografia oro-korrean bila ditzan. Argitalpenak IXA taldeko web orritik jaso daitezke (<http://ixa.si.ehu.es>).

- I. kapitulua. Proiektuaren nondik norakoak
 - (Maritxalar *et al.*, 1997). *From Psycholinguistic Modelling of Interlan-guage in Second Language Acquisition to a Computational Model*.
 - (Díaz de Ilarraza *et al.*, 1998a). *An Implemented Interlanguage Model for Learners of Basque*.
 - (Díaz de Ilarraza *et al.*, 1998b). *Integration of NLP Tools in an Inte-lligent Computer Assisted Language Learning Environment for Basque: IDAZKIDE*.

- (Díaz de Ilarraza *et al.*, 1999). *IDAZKIDE: an intelligent CALL environment for second language acquisition*.
- (Aduriz *et al.*, 2002b). *The design of a digital resource to store the knowledge of linguistic errors*.
- (Aldezabal *et al.*, 2003a). *Toward a punctuation checker for Basque*.
- (Díaz de Ilarraza *et al.*, 1997). *Reusability of NLP tools for detecting rules and contexts when modelling language learners' knowledge*.
- (Arrieta *et al.*, 2003). *A database system for storing second language learner corpora*.
- (Ansa *et al.*, 2004). *Integrating NLP Tools for Basque in Text Editors*.
- (Aldabe *et al.*, 2005d). *Learner and Error Corpora Based Computational Systems*.
- (Aldabe *et al.*, 2006). *The Use of NLP tools for Basque in a multiple user CALL environment and its feedback*.
- (Aldabe *et al.*, 2005c). *IRAKAZI: a web-based system to assess the learning process of Basque language learners*.
- II. kapitulua. Erroreak, corpus erroredunak, erroreen sailkapena eta ebaluazioa
 - (Aldabe *et al.*, 2005b). *Propuesta de una clasificación general y dinámica para la definición de errores*.
- IV. kapitulua. Abiapuntua. Hizkuntzaren analisisa eta errepresentazioa
 - (Aduriz *et al.*, 2002a). *Construcción de un corpus etiquetado sintácticamente para el euskera*.
 - (Aduriz *et al.*, 1998). *A framework for the automatic processing of Basque*.
 - (Aduriz *et al.*, 2003b). *Construction of a Basque Dependency Treebank*.
 - (Aduriz *et al.*, 2004). *A Cascaded Syntactic Analyser for Basque*.
 - (Aduriz *et al.*, 2006). *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing*.
- V. kapitulua. Patroien erabilera datetako eta postposizio-lokuzioetako erroreak detektatzeko eta zuzentzeko
 - (Aldezabal *et al.*, 1999). *Combining Chart-Parsing and Finite State Parsing*.
 - (Gojenola eta Oronoz, 2000). Koldo Gojenola eta Maite Oronoz. *Corpus-based syntactic error detection using syntactic patterns*. NAACL-ANLP00, Student Research Workshop, Seattle, USA, April 30, 2000.

- (Díaz de Ilarraza *et al.*, 2007b). Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, Maialen Otaegi eta Iñaki Alegria. *Syntactic error detection and correction in date expressions using finite-state transducers*. Proceedings of Finite-State Methods and Natural Language Processing (FSMNL07), Postdam, Germany, 14-16 of September 2007.
- (Díaz de Ilarraza *et al.*, 2008) *Detecting Erroneous Uses of Complex Postpositions in an Agglutinative Language*. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, 18-22 August 2008.
- VI. kapitulua. *Saroi*: mendekotasun-zuhaitzetan informazioa kontsultatzeko tresna.
 - (Díaz de Ilarraza *et al.*, 2004). *Abar-Hitz: An Annotation Tool for the Basque Dependency Treebank*.
 - (Díaz de Ilarraza *et al.*, 2005). Arantza Díaz de Ilarraza, Koldo Gojenola eta Maite Oronoz. *Design and Development of a System for the Detection of Agreement Errors in Basque*. Computational Linguistics and Intelligent Text Processing: 6th International Conference (CICLing2005), Mexico City, Mexico, February 13-19, 2005.
- VII. kapitulua. *Saroi*ren erabilera komunztadura-erroreak detektatzeko eta hiztegiak aberasteko
 - (Aldezabal *et al.*, 2003b). *Application of finite-state transducers to the acquisition of verb subcategorization information*.
 - (Díaz de Ilarraza *et al.*, 2007a). Arantza Díaz de Ilarraza, Koldo Gojenola eta Maite Oronoz. *Reusability of a corpus and a treebank to enrich verb subcategorisation in a dictionary*. Conference on Recent Advances in Natural Language Processing (RANLP07), Borovets, Bulgaria, September 2007.

ERROREEN DETEKZIO ETA ZUZENKETA AUTOMATIKOA

II. KAPITULUA

Erroreak, corpus erroredunak, errorearen sailkapena eta ebaluazioa

II.1 Sarrera

Hizkuntza —ikusmenezko, entzumenezko eta ukimenezko zeinuak alde batera utzita—, idatzizko zeinuekin eta hauek antolatzeke arauekin osatuta dagoela esan dugu. Arau linguistikoen bidez hizkuntzaren erabileran zer dagoen ongi eta zer dagoen gaizki adierazten zaigu. Gaur egunean, pentsatzeko era murriztaile hau ez da hizkuntzaren erabilera ikasten duten ikertzaile guztien gustuko. Ikertzaile batzuek hizkuntzalaritzak deskribatzailea izan behar duela diote, gizartean ematen den erabilera deskribatu egin behar omen da inolako balio-judiziorik egin gabe, eta ez “halabeharrez” adierazi zer den zuzena eta zer ez. Gure helburuetarako, ordea, okerra zuzenetik bereiztea ezinbestekoa da.

Errorea, arau linguistikoak oinarrian hartuta definituko dugu II.2 atalean eta ondoren, errorearen aldaerak deskribatuko ditugu. Corpus erroredunen inguruan arituko gara gero II.3 atalean. Corpusetatik bildutako erroreak sailkatzeko moduz jardungo dugu II.4 atalean. Atala amaitzeko, tesi-lan honetan lantzeko aukeratu ditugun errore motak zeintzuk diren azalduko dugu, eta egindako aukeraketaren arrazoiak emango ditugu. Kapitulu azken atalean, II.5 puntuan, errore horiek detektatzeko garatutako sistemen emaitzak ebaluatzeko metodologia eta neurriak azalduko ditugu. Has gaitezten, bada, errorea definitzen.

II.2 Errorea

II.2.1 Errorearen definizioa

Hizkuntzaren erabilera estandarraren paradigmetatik hizkuntza-erabiltzaile indibidualek egiten dituzten desbideratzeei *errore* deitzen diete batzuek (adibidez, www.wikipedia.com hiztegi entziklopedikoan¹). Vandeventer-en arabera (2003) erroreak *zuzena* zer den adierazten duen araua existitzen delako soilik detekta ditzakegu. Laburtuz, errorea honela definituko dugu: egitura zuzenetik, arautik, desbideratzen den egitura da.

Egin dugun definizioa orokorra da eta hizkuntzarekin lotutako edozein erroretarako, errore ortografiko, sintaktiko, semantiko edo pragmatikoetarako izan daiteke baliagarria. Oraingoz ideia orokorrarekin geratuko bagara ere, hurrengo kapituluan errorearen detekziorako erabilitako hizkuntza-informazioaren arabera egiten den bereizketa zehazteari ekingo diogu. Lan honetan *edizio mekanikoaren* ondorioz sortutako erroreak alde batera utzi eta errore *kognitibo*ei erreparatuko diegu.

Errorearen definizioak sinplea dirudien arren, kontzeptua azken urteotan analisi ugariaren helburu izan da eta ikerketa-lerro ezberdinak sortu dira bere inguruan. Hasiera batean *Erroreen Analisia* (EA) izeneko teoria oso kritikatu izan zen baliotasunik ikusten ez zitzaiolako. Hala eta guztiz ere, erroreen inguruan egindako ikerketek emaitza positiboak zituztela ikusirik, laster hasi zitzaion baliagarri irizten. Teoria hau indartzen hasi zen 1967an argitaratutako S. Pit Corder-en *The Significance of Learners' Errors* lana-ekin. Oinarri hura izanik eta teknologia berriek eskaintzen dituzten abantailak baliatuz, gaur egun *Ordenagailuz lagundutako Errore Analisia* (OEA) egiten da. Hori baliagarria da bai *erroreen tratamendu automatikoaren* alorrean, baita *hizkuntzen ikasketa eta irakaskuntzaren* alorrean ere.

Testu zati batean errorea dagoela esatea ez da lan erraza izaten, eta askotan eztabaidagarria izan daiteke kontu hau. Ondorengo lerroetan errore kontzeptuaren zehaztapenen inguruan jardungo dugu. II.2.1.1 atalean, desbideratzeak izango ditugu mintzagai. Estilo desegokia eta hitz zuzenen erabilera nahasketak aztertuko ditugu gero II.2.1.2 eta II.2.1.3 ataletan, hurrenez hurren.

II.2.1.1 Desbideratzeak

Bigarren hizkuntzen (H2) ikasketa eta irakaskuntzaren alorrean, ikertzaile batzuk errore terminoari ez diote egoki irizten eta *desbideratze* terminoa

¹Ingeleseko bertsioan, *error in oral and written language*.

erabiltzen dute. Hizkuntzen ikasketaren alorra tesi-lan honen oinarri ez izan arren, bertan egiten den lana zuzenean dago errearen azterketarekin lotuta. Hori dela eta, zenbait ideia emango ditugu alor honen inguruan eta zenbait termino azalduko ditugu. Hizkuntzen ezagutzarekin lotutako honako ideia emango dugu lehenik: Chomsky-ren arabera (1965), hiztunetan *hizkuntza-gaitasuna* edo hizkuntzaren ezagutza linguistiko inkontzientea, eta *hizkuntza-ariketa*, hizkuntzaren erabilera erreala, bereizten dira. Orokorrean, hizkuntza bat erabiltzen dugunean (performantzia), ez ditugu beti arauak jarraitzen nahiz eta hauek ezagutu (gaitasuna). Tesi-lan honetan hizkuntza aztertzeko hiztunen idatzizko ekoizpena erabili dugu, honenbestez hiztunen performantzia da neur dezakeguna, eta ez gaitasuna.

Kontzeptu batzuk argitu ondoren, desbideratzearen definizioarekin aurrera joko dugu. James-ek bere *Errors in Language Learning and Use* liburuan (James, 1998) honakoa zehazten du:

«When we say that a piece of learner language is “wrong”, we are merely using the label as shorthand to refer to a **discrepancy** between what this particular learner (or some typical population of learners) tends to say and what the collective entity of Native Speakers (or the ideal Native Speaker) tend to say.»

Aipuan, hizkuntza-ikasleen eta xede-hizkuntzako hiztun osoen hizkuntza-egituren arteko *desberdintasunak* zehazten dira. Hau da, errorea *arautik* aldentzean ematen dela esan dugu; ikasleei buruz hitz egiten dugunean, eza gutzen dituen arauetatik aldentzeaz gain (arau guztiak ez ditu ezagutzen), ikaslearen egiturak *hiztun osoaren ohiko erabilerako hizkuntza-egituretatik* desbideratzen dira.

Alexopoulou-k (2005) Corder-en hitzak (Corder, 1991) erabiltzen ditu desbideratzeak deskribatzeko:

«Corder (1971/1991) considera inadecuado caracterizar “erróneas” las oraciones idiosincrásicas de los aprendientes de una L2, ya que no implican “una ruptura premeditada de reglas” (...). El aprendiente produce oraciones idiosincrásicas precisamente porque no conoce las reglas que lo ayudarían a evitarlas; en otras palabras, su competencia lingüística en ese momento dado no le permite producir oraciones correctas según la norma de la L2.(...) Sería, pues, inapropiado caracterizar de agramaticales las oraciones idiosincrásicas. Como hace Corder, debemos

considerarlas *gramaticales*, ya que son coherentes con la lengua del aprendiente.»

Maritxalarrek (1999) bere tesi-lanean desbideratzea terminoa erabili zuen ikasketa-prozesuan hizkuntza-ikasleek erabiltzen dituzten berezko egiturak deskribatzeko. Deskribatutako desbideratzeetako batzuk ikasleentzat gramatikalak diren arren, hizkuntzen ikasketa/*irakaskuntza*tik kanpoko alorretan erroretzat jotzen dira. Egitura bat desbideratzea da honako kasuetan:

- Egitura linguistikoa zuzena ez denean (errorea beste alorretan).
- Egitura bera behin eta berriro erabiltzen denean egitura baliokide bat erabili beharrean.
- Egitura linguistiko bat saihestu egiten denean.

Irakasleak ikasleari zein egitura erakutsi dizkion dakienez, gai izaten da desbideratutzat jo ditugun egitura linguistikoak testuetan eskuz markatzeko. Irakasleak sistema automatikoei baino “errazago” igar dezake ikasleak zein egitura saihesten duen eta “errazago” marka dezake zein errepikatzen duen. Hori automatikoki egin ahal izateko ikasleei hizkuntza-maila bakoitzean erakusten zaizkien hizkuntza-egitura guzti-guztiak ezagutu beharko genituzke eta modu egokian kudeatu. Lan hori ez da tesi honen helburua eta, beraz, lehen motako desbideratzeak, errore kontsidera daitezkeenak, landu ditugu. Aipatutako azkenak dira testuetan marka “argiena” uzten dutenak. Adibidez, **Zenbait gizonak etorri dira* esaldi okerrean, *zenbait* determinatzailea mugagabea da eta *gizonak* mugatua. Bi hitzek ez dute mugatasunari dago-kionez komunztatzen eta hori detektatu ahal izateko, nahikoa da *zenbait* eta *-ak* bikoteari eta gainera aditzari² begiratzea. Desbideratze mota hau testuan atzeman daiteke, aztarna bertan utzi duelako. Egitura bat errepikatzen denean ere, adibidez *eta* juntagailua behin eta berriro erabiltzen dugunean, egitura horrek testuan isla du. Azken kasuan, berriz, egitura linguistiko bat saihestu bada, ez da testuan azalduko. Ondorioz, zailagoa da desbideratze mota hau markatzea.

Hizkuntzen *irakaskuntza*ren alorrean desbideratze terminoaren inguruan zehaztapen anitz egin dira. Beren lanetan termino honetaz hitz egin duten ikertzaileek (Norrish, 1983; Edge, 1989; Maritxalar, 1999), desbideratzeak *errakuntza*, *hutsegite*, *lapsus*, *saialdi* eta besteetan banatu dituzte.

²Aditzari begiratu behar zaio, *Zenbait gizonak ikustera etorri da* esaldian komunztadura-errererik ez markatzeko.

II.2.1.2 Estilo desegokia

Lan honetan estiloa landu ez dugun arren, errore-kontzeptuarekin duen harremanagatik beharrezkoa deritzogu kontzeptua kokatzea. *Estilo desegokiaren* adierazgarri izaten diren fenomenoak egitura linguistiko zuzenak dira. Ondorioz, ez dira erroreak, baina, hala ere, ez dira egokiak. Testuan edo hitz egiterakoan adierazi nahi izan dena adierazteko modu egokiago bat baino gehiago egon daiteke eta hobetsitako formak erabiltzea gomendatzen dute hizkuntzaren erabilera zuzena bermatzen dutenek. Ravin-en arabera (1993) estilo ahultasunak idazlearen, irakurlearen eta idazten ari garen dokumentu motaren arabera dira eta kalitate neurriak hartu behar dira kontuan horiek ebaluatzeko. Gainera, erroreetan ez bezala, zuzenketa ez da garbia izaten. Zentzu batean, aurreragoko lerroetan desbideratzeen definizioa egin dugunean *saihestutako egiturak* eta *behin eta berrir*o errepikatzen diren *egiturak* ere, *estilo desegokiaren* barruan sailka ditzakegu. Diferentzia, ordea, hizkuntza-erabiltzaileetan dago. Lehenago esan dugun moduan, desbideratzeetan hizkuntza-erabiltzaileak ikasleak izaten dira. Estilo desegokia berriz, edozein hitzunek erabil dezake, ikaslea izan edo ez.

Euskaltzaindiak, zenbait erakundek (Herri Arduralaritzaren Euskal Erakundea, I.V.A.P. (2005)), zein egunkarik (Egunkaria, 1995; Aranguren *et al.*, 2006), baita idazle batzuek ere (Alberdi eta Sarasola, 2001), gomendioak eman ohi dituzte estilo egokiaren erabilerarako.

II.2.1.3 Hitz zuzenen erabilera nahasketak

Ohiko ortografia eta gramatika-egiaztatzaile eta -zuzentzaileak ez dira normalean gai izaten hitz zuzenen erabilera nahasketez ohartzeko. Har dezagun **Atetik begiratu eta esan zegoela ikusi zuen* esaldia. Bertan, *esan* eta *esna* hitz-formak nahastu egin dira, baina bi hitzak zuzenak direnez, latza gertatzen da nahasketaren ondorioz sortutako errorea automatikoki detektatzea. Testuingurua behar da hitza egoki erabili den ala ez jakiteko. Arazoa zaila eta zabala da: batzuetan inguruan hitz konkretu bat topatzearen baitan dago, beste batzuetan inguruko hitzen kategoriak aztertu behar dira. Kukichen arabera (1992), “real-word errors” deitutakoak, hots, baliozko hitzetan gertatzen diren erroreak, errorearen % 25-40 izaten dira; beraz, honakoa garrantzi handiko gaia da. *Context-sensitive spelling correction* edo *testuingurua aztertuta ortografiaren zuzenketa* du izena arazo honetaz arduratzen den alorrak; hots, hitz zuzenen erabilera nahasketak zuzentzen dituen alorrak. III. kapituluaren berrir aipatuko dugu alor hau erroreak detektatzeko teknikak azaltzen ditugunean.

11.2.2 Detekzioa, diagnosia eta zuzenketa

Errorearen definizioan ikuspuntu eta ñabardura anitz topatu baditugu, ez da gutxiago gertatzen *detekzioari*, *diagnosiari* eta *zuzenketari* buruz hitz egiten dugunean.

Batzuetan, detekzioa eta diagnosia terminoak fenomeno bera izendatzeko erabiltzen dira. Terminoak argitzeko asmotan, beraien definizioak aztertzea joko dugu. *Dictionary of Contemporary English* hiztegian, *detekzioa*, *diagnosia* eta *zuzenketa* hitzak honela definitzen dituzte:

- “Detection: when something is found that is not easy to see, hear etc., or the process of looking for it”.
- “Diagnosis: the process of discovering exactly what is wrong with someone or something, by examining them closely”.
- “Correction: “a change made in something in order to make it right or better”.

Gure kasura ekarrita, testu bat aztertu ondoren, erroreak daudela esatea eta erroreak *kokatzean* datza erroreak detektatzea, zergatirik edo argibiderik eman gabe. Erroreak agerian jartzen dira inolako deskribapenik egin gabe. *Diagnosia* egiten denean, ordea, errorea deskribatu egiten da; gaizki dagoena aztertu egiten da zehatz-mehatz arazoa zein den adierazteko; kokatzea baino gehiago egiten da. *Zuzenketa* egitean egitura erroredunaren ordezkoko zuzenak ematea da helburua, gehienetan aldezturik erroreak diagnosia eginez, eta gutxi batzutan, diagnosirik egin gabe (ikus III. kapitulua). Ikus ditzagun definizioak adibideak erabiliz:

- (1) *Esaldia*:
*Zentral nuklearrak zakar erradiaktiboa eratzen dute.
- (2) *Detekzioa*:
*[Zentral nuklearrak] zakar erradiaktiboa eratzen [dute].
- (3) *Diagnosia*:
Esaldiko subjektuak, Zentral nuklearrak, eta aditz laguntzaileak, dute, ez dute numeroan edo/eta kasuan komunztatzen.
- (4) *Zuzenketa*:
a. Zentral nuklearrak zakar erradiaktiboa eratzen du.
b. Zentral nuklearek zakar erradiaktiboa eratzen dute.

Aipatzekoa da hemen erakusten diren ordain zuzenak aldaketa gutxien behar dituztenak direla (aldaketa bakarra, lehen zuzenketa aditz laguntzaileria aldatu da, eta bigarrenean, subjektuaren kasu-marka). Aldaketa posible bat baino gehiago eginda, zuzenketa gehiago sortuko lirateke (adib. [Zentral nuklearrak zakar erradiaktiboak eratzen ditu](#)). Foster eta Vogel-ek (2004a) corpus erroredun bat sortu dute, eta paraleloan, corpus erroreduneko erroreei dagozkien zuzenketekin osatutako corpus zuzena. Esaldi ez-gramatikalen % 20 haien esanahia aldatu gabe modu batean baino gehiagotan zuzen zitezkeela jakinarazi dute. Alegia, ohikoa da errore bati zuzenketa posible bat baino gehiago lotzea.

Tesi-lan honetan erroreen diagnosi automatikoa landuko dugu batez ere, horretarako lehenik erroreak detekzioa eginez. Zenbait kasutan zuzenketa ere egingo dugu.

Diagnosia OLHIan

Azpimarratzekoa da egin dugun diagnosiaren definizioa ez dela *Ordenagailuz Lagundutako Hizkuntzen Irakaskuntzaren (CALL-Computer Assisted Language Learning)* alorrean orokorrean erabiltzen dena. OLHIan, desbideratzearen diagnosia egiten denean, desbideratzea *zergatik* egiten den azaltzen da: desbideratzearen arrazoiak ematen dira. Lehenago ere aipatu dugun James-ek (1998), guk diagnosi izendatu duguna *deskribapen* deitzen du eta OLHIaren alorrean erabiltzen den diagnosia eta deskribapena honela bereizten ditu:

«The diagnosis question is therefore one that transcends description and invokes **ascription** or explanation, tracing errors to their causes, sometimes called error **etiology**.»

Diagnosia erroreak kausa edo jatorriarekin lotuta azaltzen da aipu honetan, erroreak zergatia bilatzen baita. Ildo honetan, Maritxalarrek (1999) bere tesi-lanean ikasleek desbideratzeak egiteko izan ditzaketen sakoneko arrazoiak sailkapen bat egin zuen. Tesi-lan honetan ez dugu arrazoiak oinarria duen diagnosirik egingo. Nahiz eta jakin badakigu, batzuetan oso zaila dela erroreak deskribapena eta errorea egiteko arrazoiaren arteko bereizketa garbia egitea. Esaterako, [*Sara bezala bataiatu zuten](#) esaldian *erdararen* kalko okerra (Gartzia, 2005) erabili da³. Hau da, euskaraz egitura bat gaztelarako eruditik jaso da. Kasu honetan errorea deskribatzerakoan arrazoiak ematen ari gara.

³Esaldi egokia [Sara izenez bataiatu zuten](#) litzateke.

Zuzenketa

Zuzenketa kontzeptua dagoeneko azaldu dugu eta III.2.1.1 atalean berriro ere landuko dugu zuzenketarako teknikak azaltzen ditugunean. Hala ere, zenbait ideia aipatzea garrantzitsua iruditzen zaigu. Egiaztatzaile eta zuzentzaile automatikoek ezezaguna eta, beraz, seguraski okerra den hitz bat markatzen dutenean, hitza zuzendu ahal izateko proposamen zerrenda eman ohi dute. Gehienetan, *ordain zuzenen zerrenda* azaltzen den lehen hitz-forma probabilitate handienarekin gerta daitekeena izaten da. Beste batzuetan ez da izaten inolako ordenarik proposamenen artean. Proposamenen arteko ordena egokia ipintzea ez da berehalako lana eta horretarako, fonologia, morfologia eta testuinguruan oinarritutako zenbait teknika erabiltzen dira (ikus III.3.1 atala).

Zuzenketaren inguruan kontuan hartu behar diren zenbait puntu aipatuko ditugu jarraian, horretarako adibideak erabiliz:

- Errore bat zuzentzeak beste errore baten sorkuntza eragin dezake:

(5) *Hango [jende guztiek] dakite lurrikara denean zer egin.

Esaldi honetan *jende guztiek zatiaren barnean komunztadura-errore bat topa dezakegu. Zehazkiago, numero singularra berezkoa duen jende hitzak ez du plurala duen guztiek hitzarekin komunztatzen. Errore hau modu bakarrean zuzen daiteke, guztiek hitza numero singularrean deklinatuz, eta ondorioz, jende guztiak sintagma sortuz.

Errorea zuzentzeak beste errore baten sorkuntza dakarkigu.

(6) *Hango jende [guztiak dakite] lurrikara denean zer egin.

Lehenago zuzena zen sintagma (guztiak dakite) orain okerra da (*guztiak dakite).

Ondorio moduan esan dezakegu, erroreak zuzentzean sortzen den egitura behin eta berriro analizatu behar dela errore gehiagoren bila. Ziklo bat sortu behar da zuzenketa bakoitzaren ondoren erroreen bilaketa berriro egiteko. Detekzio-zuzenketa zikloa ez da amaituko azken egitura guztiz zuzena den arte. Zikloko iterazio bakoitzean ahalik eta errore gehien detektatzea eta zuzentzea komeni da, ziklo eta analisi gutxiago behar izateko, eta ondorioz, zuzenketa azkarragoa egiteko.

- Automatikoki sortutako lehen zuzenketak, ez du zertan egokiena izan:

(7) *Pareta egin zuen langileak zuzendu behar izan dute

Adibideko komunztadura-errorea zuzentzeko giza-zuzentzaile baten joera, segur aski Pareta egin zuen langileak zuzendu behar izan du esaldia sortzea litzateke. Ordain zuzen hori, informazio pragmatikoa erabiliz eta idazleak esan nahiko lukeena suposatuz egingo luke giza-zuzentzaileak, detekzioan gehiegi pentsatu gabe. “Teorian” zuzenketa egokiena litzateke. Ordain hori lortzea, ordea, makinenzako ezinezkoa da, oraingoz, ez baitute testuinguruko informaziorik erabiltzen. Makinari, abiapuntutzat aditz laguntzailea (dute) emango genioke, eta honek subjektuarekin langileak ez duela komunztatzen detektatuko luke. Horrela, (7) adibideko esaldi erroredunaren zuzenketa moduan *Pareta egin zuen langileak zuzendu behar izan dute esaldi okerra proposatuko luke. Bigarren analisi batean (gogoratu aurreko puntuan aipatutako detekzio-zuzenketa zikloa), geratzen den errorea detektatzeari (*zuen langileak) eta zuzentzeari ekingo lioke, azkenean, Pareta egin zuten langileak zuzendu behar izan dute lortuz.

- Esaldi batean errore bat baino gehiago daudenean erroreak banan-banan, urratsez-urrats zuzendu behar ditugu, ala errore guztiak batera zuzenduko ditugu?

Errore guztiak aldi berean zuzenduta ikustea ohikoagoa egiten zaigu errore batzuk soilik zuzentzea baino. Hizkuntzen irakaskuntzaren alorrean, ordea, segur aski egokiago litzateke ikaslearen hizkuntza-ezagutza mailako kontzeptuekin lotutako erroreak soilik zuzentzea. Corдер-en aipuan irakurri dugun moduan, ikasleari ez litzaizkioke ezagutzen ez dituen hizkuntza-egiturak zuzendu behar. Gramatika-zuzentzaile batek, berriz, errore posible guztiak zuzendu beharko lituzke.

Erroreak zuzentzeko teknikak III. kapituluaren azalduko ditugu zehazkiago. Orain, corpusen inguruan arituko gara.

II.3 Corpus erroredunak

Tesi-lan honetan corpusen erabilera oso garrantzitsua da. Corpus erroredunak zein zuzenak, erredeen detektziorako gramatikak garatzen laguntzeko eta gramatika horiek ebaluatzeko erabili ditugu. Erabili ditugun corpusak deskribatzen hasi aurretik, corpus bat zer den definituko dugu. Corpusen inguruan definizio asko egin badira ere, honakoa ongi egokitzen da gure helburuetara:

«(...) corpora hizkuntzari buruzko datu-bilduma da. Corpusaren definizio zabal bat egitera, esan daiteke edozein testu edo testu-bilduma har litekeela corpustzat. Hala ere, gaur egun definizio zehatzagoa erabili ohi da: corpora hizkuntza-erakusgarri “errealen” multzo “handi” bat da, irizpide batzuen arabera bildua eta formatu elektronikoan biltegitua. Askok beste zerbait ere erantsiko liokete horri. Corpusak, erabilgarria eta eraginkorra izango bada, informazio linguistikoaz hornitua behar du izan. »

(IXA taldea eta Elhuyar Fundazioa, 2007)

Definizio horri jarraiki, puntu honetan corpus erroredunak biltzeko irizpideak zehaztuko ditugu lehenengo (ikus II.3.1 azpiatala); corpusak formatu elektronikoan nola bildu ditugun azalduko dugu gero (ikus II.3.2 azpiatala); eta azkenik, informazio linguistikoaz nola hornitu ditugun adieraziko dugu (ikus II.3.3 atala).

II.3.1 Irizpideak

Corpusak osatzeko irizpide zehatzak ematen hasi aurretik, gure corpora zein motakoa den adieraztea egokia iruditu zaigu. Aipatu berri dugun lanean corpus motak zehazten dira (IXA taldea eta Elhuyar Fundazioa, 2007) eta zehaztapen horren arabera, gureak honela deskriba ditzakegu: testu-corpusak dira, espezializatuak (erroreak dituzte), elebakarrak, etiketatuak (sintaktikoki analizatuak), sinkronikoak eta erreferentziazkoak⁴.

Ezaugarri orokor horietaz gain, erroreen tratamendurako espezifikoak diren beste zenbait ezaugarri zerrendatu ditugu jarraian. Hauek dira, gure ustez, corpus erroredunetan kontuan hartzekoak:

- **Ikasle corpusak *vs* corpus erroredun orokorrak.** Wagner *et al.* (2007) egileek ongi zehazten duten moduan:

«A corpus of ungrammatical sentences can take the form of a learner corpus (Granger, 1993; Emi *et al.*, 2004), i.e. a corpus of sentences produced by language learners, or it can take the form of a more general error corpus comprising

⁴Erroredun corpus bati azken etiketa hau ipintzea zalantzazkoa izan daiteke. Erreferentzia corpusaren definizioa honakoa da: corpusaren helburua hizkuntzaren erabilera-eremu guztietarako baliagarria edo “adierazgarria” izaten denean (Leech, 2002). Erroredun corpusak hizkuntzaren erabilera-eremu guztietarako direla baliagarriak ulertzen dugu. Kasu honetan, *erabilera ez-gramatikal* guztietarako.

sentences which are not necessarily produced in a language-learning context and which contain competence and performance errors produced by native and non-native speakers of the language (Becker *et al.*, 1999; Foster eta Vogel, 2004a; Foster, 2005)»

Gure lanerako, corpus erroreduna hizkuntza-ikasleena izatera mugatu gabe, corpus orokorrago bat erabili nahi izan dugu, zeinetan hiztun osoen eta hiztun ez-osoen erroreak bilduko diren. Erroreekin egindako lanaren azken erabilera esparrua bikoitza denez, gramatika-zuzentzai-leak eta hizkuntzen irakaskuntzarako sistemak, corpus orokorra erabiltzea iruditu zaigu egokiena.

- **Corpus errealak vs sortutako corpusak.** Erroreak dituzten corpus errealak biltzea lan neketsua da, ez baita soilik errorerik ote dagoen ala ez aztertu behar, honez gain, zein errore gertatu den etiketatu behar izaten da. Azken boladan, lan hori aurreztearren, esaldi gramatikalekin osatutako testuetatik abiatuz corpus erroredunak sortu dituzte, adibidez, hitzak gehituz, ezabatuz edo aldatuz (Wagner *et al.*, 2007; Bigert, 2004). Ingeles ez-gramatikalaz osatutako *treebank* bat ere sortu dute (Foster, 2007), helburu bikoitzarekin: batetik, analizatzaile sintaktikoen sendotasuna neurtzeko balio lezake, erroreak egon arren analizatzaileek analisia egiteko gai izan behar baitute; beste aldetik, esaldi erroredunak analizatzeko gai den gramatika bat ikasteko balio luke. Automatikoki sortutako corpus erroredunak era naturalean sortutakoen antzekoak izateko ahaleginak egin diren arren, gure ustez, “errealistak” ez izatearen arazoa izan dezakete. Gainera, erroreak sortzeko prozesuan, esaldi zuzenak sor daitezke eta automatikoki okerrak moduan etiketatu. Hori horrela izanik, corpus errealak biltzea erabaki dugu.
- **Probarako zerrendak vs corpusak.** Lortzeko lan handiagoa ematen duten arren, probarako kasu-zerrendak ebaluatu beharreko egituretara hobeto egokitzen omen dira, eta beraz, gure ustez, emaitza zehatzagoak ematen dituzte. Corpusek, ordea, tamaina egokikoak izanik, erreoren maiztasun erreala zehazten dute. Bi testu mota hauen abantailak biltzeko asmoz, lan honetan corpusak bildu egin ditugu, horrek ematen duen lanarekin, eta berauetatik abiatuta probarako zerrendak sortu ditugu. Horrela, berariazko fenomenoak zehatzago landu ditugu, testu osoan fenomeno horiek duten agerpen-maiztasuna galdu gabe.

- **Corpus erroredunak vs zuzenak.** Tesi-lanarekin aurrera egin ahala, corpus zuzenak (errorerik gabekoak) erabiltzearen garrantziaz jabetu gara. Erroreak detektatzea bezain garrantzitsua da, gure aburuz, zuzenak diren egiturak erroretzat ez jotzea. Hau da, alarma faltsuak ekidin behar ditugu, eta hori ongi egiten ari garen ezagutzeko, testu zuzenak behar ditugu. Lan honetan azalduko diren gramatika guztiak testu erroredunetan eta testu zuzenetan probatuko dira. Azkenak erroredunak baino errazago lortzen dira, eta garrantzitsuak izanik, ez zaizkio erroreen detekzioaren atazari zuzenean lotzen; beraz, horiek biltzeko modua ez dugu II.3.2 atalean zehaztuko. Atal hori testu erroredunak biltzeko prozesua azaltzeari eskainiko diogu.

Laburtuz, corpusei dagokienez, honako aukeraketa egin dugu: i) erabiliko ditugun corpus erroredunak orokorrak izango dira, ii) corpus horiek erreala izango dira, eta iii) corpus erroredunak zein zuzenak erabiliko ditugu. Corpus mota aukeratzerakoan (probarako zerrendak vs corpusak), bereizketa bat egin dugu landutako hizkuntza-fenomenoaren arabera:

- Testuinguru edo leiho mugatuko erroreak lantzen ditugunean (ikus V. kapitulua) probarako zerrendak erabiliko ditugu. Zerrenda hauek osatzeko, “gako-hitza” zehatz bat duten esaldiak jasoko ditugu. “Gako-hitza” duten esaldietan, landu nahi den egitura zuzen erabilita, gaizki erabilita (errorea) eta beste egituraren batean erabilita agertzen da. Adibidez, datak lantzerakoan, [martxo](#) gako-hitza duten esaldiak jasoko ditugu. Horrela, data zuzenak ([2008ko martxoaren 8a](#)), data okerrak ([2008-ko Martxoaren 8](#)) eta egin dugun definizioaren arabera (ikus V. kapitulua) datak ez diren adibideak ([Martxoan itzuli zen Ameriketatik](#)) bilduko ditugu. Zerrendak erabilita, erroreak modu zehatzagoan landuko ditugula uste dugu.
- Testuinguru zabalekoak deitu ditugun erroreak landu ditugunean, adibidez komuntadura (ikus VII. kapitulua), “gako-hitza” erabiltzeak ez du zentzurik. Esaldiak edo perpausak aztertu behar ditugunez, corpusak erabili ditugu. Corpus zuzena, esaldi erroredunekin soilik osatutako corpora, eta azkenik, esaldi zuzenak zein okerrak dituen corpora erabiliko ditugu.

II.3.2 Testu-bilketa

Gurea bezalako ataza zehatz baterako, erroreen tratamendu automatikorako, testu-bilketa lan neketsua da. Idazleek, beraien idazlanak denon eskura

ipini aurretik, normalean, gramatika aldetik zuzendu egiten dituzte. Modu erraz samarrean lor daitezkeen corpusak, Interneten atzigarri egon daitezkeen euskarazko egunkariak, aldizkariak, liburuak etab., ia beti gramatikalki zuzenak dira. Zenbaitzuk, hori horrela izanik, errorearen tratamenduaren alorrean Internet hizkuntza gramatikaren adibide moduan erabiltzen duten arren (Sjöbergh, 2006; Moré *et al.*, 2004), hori egiteak sor ditzakeen arazoak ikusten dituzte:

“There are many web sites with intentional examples of incorrect language use, and recognizing these can be hard”

(Sjöbergh, 2006)

Interneten topa daitezkeen testu erroredunak zuzenetatik bereiztea lan astuna izan daitekeenez, erroreak izateko probabilitate handia duten testuak biltzea erabaki dugu. Honako iturrietako testu erroredunak erabili ditugu errorearen detekzioarako eta zuzenketarako:

- *Euskaltegietakoa*. Euskaltegi hauetako testuak jaso dira, besteak beste: Nafarroako AEK-koak, Trintxerpeko AEK euskaltegiak, irakasleen trebakuntzarako Irale programak eta Donostiako Ilazki euskaltegiak. Jasotako testu-bilduma testuen zenbait ezaugarriaren arabera sailkatu da. Adibidez, i) irakasleak proposatutako ariketa motaren arabera (laburpena, proposatutako gai bati buruzko idazlana, gutuna...) edo ii) egilearen arabera. 1990 urtean hasi ginen testu hauek jasotzen eta gaur egun, oraindik, horretan dihardugu. Guztira 125.000 hitz inguruko corpusa osatu eta erabili dugu.

Euskaltegietan lantzen diren testuetan zailtasun-maila eta errore-tasa oso desberdinak direnez (ikasleen jakintza-mailaren arabera), testuok hiru mailatan banatu ditugu: *behe-mailan* garai bateko 1-5 urratsetako testuak bildu ditugu, *maila ertainean* 6-8 urratsetakoak, eta *goi-mailan* 9-12 urratsetakoak (Maritxalar, 1999). Mailak gutxi gorabehera bat etorriko lirateke HABEko “Helduen Euskalduntzearen Oinarrizko Kurrikuluko (HEOK)” lehen hiru mailekin (4 daude).

- *EuskaraZ zerrenda*. EuskaraZ zerrenda⁵ 1996an sortu zen eta euskarekin loturiko informazioa trukatzeko zuen helburu. Zerrenda horretatik 2000 eta 2001 urteetan idatzitako posta elektronikoko 3.049 mezu (542.866 hitz) jaso genituen. Corpus hau erraz atzigarria da, modu

⁵<http://www.sarean.com/artxiboak/000401.html>

elektronikoan lor baitaiteke, eta errore linguistikoak ditu. Desabantaila bat du, ordea, hizkuntz informalean idatzita dago eta batzuetan laburtzapenak eta bukatu gabeko hitzak topa ditzakegu bertan. Horrek analisia zailtzen du.

- *Euskara teknikoa*. Euskal Herriko Unibertsitatean *euskara teknikoa* izeneko irakasgaia irakasten duten irakasle batzuen eskutik jaso ditugu haien ikasleen lanak. Lan horiekin 19.549 hitzeko testu-bilduma osatu dugu.
- *Karrera bukaerako proiektuak*. Informatika Fakultateko ikasleek karrera amaitu ahal izateko, proiektu bat garatu eta honi buruzko txosten bat idatzi behar izaten dute. Euskaraz idatzitako hainbat txosten jaso ditugu erreoren bila. Txostenak zuzendariek zuzendu aurretiko bertsioak dira. Testu hauetan 305.796 hitz daude.

Euskaltegietako testuetan euskara-ikasleen erroreak topatuko baditugu ere, gainontzeko hiru testu multzoetan hiltun osoek egiten dituzten erroreak bildu dira. Hori horrela izanik, corpus erroredun orokor baten aurrean gaudela esango dugu.

Corpus erroredunaz gain, corpus zuzena ere erabili dugula esan dugu. Euskaldunon Egunkariko 2002 urteko albiste guztiak erabili ditugu testuinguru mugatuko erreoren alarma faltsuak probatzeko (8.207.919 hitz), eta euskarazko *treebanka* erabili dugu (Aduriz *et al.*, 2003b), testuinguru zabaleko erreoren alarma faltsuak probatzeko (50.000 token inguru).

V. eta VII. kapituluetan zehatz-mehatz azalduko dugu zein corpus, corpus horietako zein testu eta zenbat hitz erabili ditugun errore-erregelak garatzeko eta probatzeko.

II.3.3 Etiketatzeara

Corpusak etiketatzea lan garestia den arren, corpus etiketatuak baliabide ahaltsuak dira. Testu-bildumak modu ezberdinetan etiketa daitezke (IXA taldea eta Elhuyar Fundazioa, 2007): *egiturazko informazioa* etiketatuaz (informazio dokumentala, dokumentuen egitura eta formatu-ezaugarriak) edo *hizkuntza-informazioa* gehituaz (informazio lexikala, sintaktikoa, semantikoa eta pragmatikoa). Tesi-lan honetan ez dugu egiturazko informaziorik etiketatuko, baina hizkuntza-informazioa etiketatzea ezinbestekoa izango zaigu erroreak automatikoki detektatzeko eta zuzentzeko, eta hau guztia ebaluatzeko.

Testuinguru mugatuko erroreak detektatzeko, etiketatze morfosintaktikoa erabiliko dugu, zenbait kasutan informazio semantikoa erabiliko badugu ere (adibidez, postposizio-lokuzioen detekziorako). Testuinguru zabaleko errorearen detekziorako, komunztadura-erroreen detekziorako kasu, etiketatze morfosintaktikoaz gain, sintaxi-informazioa ere baliatuko dugu.

Honetaz gain, tresnen ebaluazioaren atazarako hizkuntzalari batek corpus erroredunetan erroreak etiketatu egin ditu, jarraian zehaztuko dugun sailkapenean zehaztutako errore-kategorien etiketak erabiliz.

II.4 Erroreen sailkapena

Erroreen sailkapen egoki eta zehatza egitea interesgarria da. Hizkuntzaren eta honen irakaskuntzaren aldetik interesgarria dela argi badago ere, alde praktikoari begira honako atazarako baliagarria izango da: sailkapenean definitutako errore-kategoriak erroreak detektatzeko erabiliko diren erregelak identifikatzeko, eta ondorioz tesi-lanean garatutako tresnen ebaluaziorako balioko dute. Gainera, erroreen sailkapen hau I. kapituluaren aurkeztutako erroreen biltegitratzeko bi datu-baseetan, ERROREAK eta DESBIDERATZEAK (Arrieta *et al.*, 2003), eta erroreak eskuz etiketatzeko *Etikerro* tresnan erabiltzen da. Kapitulu berean aipatutako XUXENG gramatika-zuzentzailean detektatzen diren errore guztien identifikaziorako etiketek ere, sailkapeneko kodeak erabiltzen dituzte.

II.4.1 Irizpideak

Erroreen sailkapena osatzeko, lehenengo, sailkapenerako irizpideak finkatu behar dira. Gojenolak bere tesi-lanean errore-sailkapenak osatzeko irizpide deskriptiboa, irizpide linguistikoa edo irizpide etiologikoa erabil daitezkeela dio (Gojenola (2000), 132. orrialdean). Irizpide berberak aipatzen ditu Alexopoulouk (2005):

- *Irizpide deskriptiboa*: Erroreen forma aztertzen du. Zehazkiago, instantzien azaleko egituraren zein aldaketa ematen den. Horrela, deskribatutako errore motaren arabera, hizkiaren, morfemaren, hitzaren, hizkuntza-egituraren... ordena aldaketa, sorrera (edo gehitzea), desagertzea (edo kentzea) eta trukearen ondorioz sortutako erroreak bereizten dira. Adibidez, **hemendikan* hitzeko errorea *-an* morfema gehitu dela esanez deskribatuko genuke; **jendea aldaketa maite duelako* esaldian *-k*, edo ergatiboaren kasu-marka (subjektua) desagertu dela esango genuke.

- *Irizpide linguistikoa*: Errorea zein maila linguistikotan kokatzen den zehazten du. Beste modu batera esanda, detekzio edo zuzenketarako behar den informazioaren arabera egiten da sailkapena. Horrela, errore ortografikoak, morfologikoak, sintaktikoak, semantikoak, pragmatikoak e.a. bereiz daitezke.
- *Irizpide etiologikoa*: Errorea sortzeko arrazoiaren, errorearen jatorriaren, arabera egiten da sailkapena. Enfasia errorea egiten duen pertsonaren gainean jartzen da. Psikolinguistikarekin eta hizkuntzen irakas-kuntzarekin ditu harremanak. Erroreen zergatia adierazteko irizpide etiologikoan azaltzen diren errore multzoetan, hizkuntzen arteko erroreak (interferentziak, itzulpen literala eta ihes egiteak) eta hizkuntzaren barruko erroreak (sistema sinpleago batera murriztea eta orokortzea) bereizten dira.

11.4.2 Ezaugarriak

Lehen bi irizpideak aukeratu ditugu guk euskararako erroreen sailkapena egiteko. Hona hemen sailkapen horren ezaugarriak:

- *Erroreen sailkapenaren bizkarrezurra irizpide linguistikoan oinarritu dugu eta hizkuntza-maila bakoitzaren barruan errore-kategoria zehatzenean irizpide deskriptiboa* erabili dugu erroreak deskribatzeko. Adibidez, errore *ortografikoetan* irizpide deskriptiboa erabili dugu barneko mailan (letrak kendu, gehitu, ordezkatu...). Bi ikuspuntuak bateratzeak sailkapena aberatsagoa egiten duela iruditzen zaigu. Alde batera utzi dugu irizpide etiologikoa. Ez dugu erroreak egiteko arrazoietan sartu nahi izan, psikolinguistikako alorreko lana kontsideratzen baitugu.
- *Hierarkikoa eta dinamikoa da*. Sailkapen irekia egiten saiatu gara, erre-kategoria berriak topatuz gero, erraz hedagarria izan dadin. Zuhaitz moduko hierarkia honetan, barne-adabegietan erroreen kategoriak eta azpikategoriak topatuko ditugu, eta hostoetan, errore-instantziak beraien adibideekin. Egun, 7 kategoria nagusi zehaztu ditugu eta 155 azpikategoria (horietako 129, hostoak). Kategoria/azpikategoria bakoitza kode batekin, kodearen definizioarekin eta gutxienez errore adibide batekin osatuta dago. Une honetan lau mailako sakonera du hierarkiak, baina errore kategorietako bakoitza zehatzago aztertu arte, ez dugu jakingo sailkapenak beharko duen sakontasuna zein den. Deskribapen osoa ikusi nahi izanez gero, jo (Aldabe *et al.*, 2005a) txostenera.

- *Errore-kategoriak ahalik eta gehien zehatzu* ditugu multzo orokorra-goak egin beharrean. Erroreak zehatz deskribatuta, horien diagnosis, detekzioa eta zuzenketa zehatzagoak izango dira.
- Sailkapenean agertzen diren *kategoria guztiak ez dagozkie erroreei*. Hau da, kategoria gehienek egitura ez-gramatikalak deskribatzen dituzten arren, egitura gramatikalek ere bere tokia dute sailkapenean. Estiloari erreferentzia egiten dion kategoriak, adibidez, egitura zuzenak deskribatzen ditu, baina gramatika-zuzentzaile edo OLHIA sistema batek estilo desegokia markatu beharko lukeela uste dugunez, sailkapenaren parte dira.

Euskaraz egiten diren erroreen sailkapena sendoa eta osatua dela esan dezakegu. Prozesu luze baten ondorioa da, eta bere garapenean probak egin ditugu zenbait euskara-aditurekin:

1. Hasierako bertsioa garatzeko oinarri moduan erabili genituen iturriak ugariak izan ziren: Maritxalarren (1999), Gojenolaren (2000) eta Guinovart-en (1996b) tesi-lanetan ikuspuntu zeharo desberdinekin zehaztutako sailkapenak; euskarazko erroreak lantzen dituen gramatika bat (Zubiri, 1994); beste hizkuntzetarako egindako lanak (Becker *et al.*, 1999) e.a. baliatu genituen. Gainera, euskara-ikasleen ikasketa prozesuko hizkuntza-ezagutza adierazteko egin genituen lanak (Díaz de Ilarraza *et al.*, 1997) eta irakasle izandako hizkuntzalarien esperientzia lagungarri izan genituen. Jasotako informazio hori guztia bildu ondoren, antolatu eta osatu egin genuen.
2. Sailkapena ebaluatzeko asmoarekin, inkesta bat pasa zitzairen 14 adituri erroreen sailkapena birfin eta osa zezaten. Adituei, aurkeztutako errore-kategoria bakoitzaren maiztasuna markatzeko eskatu zitzairen. Aztertzen zituzten testuetan errore-kategoria zehatz bat “askotan”, “batzuetan” edo “gutxitan” gertatzen ote zen adierazi behar zuten. Hiru multzotan banatu genituen inkesta jasotzaileak: egunkari zuzentzaileak eta editoreak, *euskara teknikoa* irakasgaiko irakasleak, eta eskalategietako irakasleak. Inkestak multzo berezitu hauetan banatuta, ahalik eta kasuistika zabalena jasotzen saiatu ginen. Era horretan gainera, etorkizunean bai zuzentzaileak baita OLHIA sistemak erabiltzaile motaren arabera egokitze bidea prestatu dugu. Adituek egindako proposamenekin erroreen sailkapena osatu eta hobetu genuen.
3. Azkenik, erroreen sailkapena hobeto zehazteko ariketa bat prestatu zen hamaika hizkuntzalari partaide zituelarik. 28 esaldi eman zitzaiz-

kien, gure iritziz errore bat edo gehiagorekin, erroreak detektatu, sailkatu eta zuzen zitzaten. Ondorio interesgarriak lortu genituen ariketa honetatik: i) hizkuntzalariak, oro har, nahiko konforme agertu dira sailkapenarekin; ii) zailtasun handienak errore morfosintaktikoak sailkatzerakoan aurkitu dituzte; iii) joera orokorra, errore bat baino gehiagoko egitura kategoriatan bakar batean sailkatzea da; iv) errore askotako esaldietan, joera esaldi osoa zuzentzearena da, eta ez zaio erroreetako bakoitzari bere kategoria propioa esleitzen; v) hizkuntzalarien % 45i errazagoa egin zaio errorea zuzentzea eta ondoren sailkatzea, % 27ri lehenengo sailkatzea eta gero zuzentzea eta gainontzekoei bi teknikak nahastu dituzte; vi) hizkuntzalarien % 63ri zehatzegia egin zaio sailkapena, baina gainontzekoei interesgarria iruditu zaie halako zehaztasuna (Aldabe *et al.*, 2005a, b).

Aditu batzuentzako egitura batean errorea dagoela jakinda ere, oro har ez da erraza izan errorea sailkapeneko kategoria batean kokatzea. Noiz edo noiz, egitura bat kategoria batean baino gehiagotan kokatzea posible da. Kasu horretan ez genuke errore bakarra, kategoria adina errore baizik. Adibidez, ***Bonboiak Aitor jan du** esaldian hiru errore daudela esan dezakegu:

- *Semantikoa*: **bonboia** bizigabea denez, ezin du ezer jan. Esaldiari dagozkion zuzenketak **Bonboia Aitorrek jan du** edo **Aitorrek bonboia jan du** lirateke.
- *Sintaktikoa*: **jan du** aditzaren subjektua **Aitor** dela erabakiz gero, ergatibo kasuan deklinatuta egon beharko luke (*aditza-subjektua* kategoriatan errorea). Objektua, **bonboia** izanez gero, honek absolutibo kasuan deklinatuta egon beharko luke (*aditza-objektua* kategoriatan). Bi errore horien zuzenketa eginez gero, besteen artean **Bonboia Aitorrek jan du** esaldi zuzena lortuko genuke⁶.

Era berean, **Bonboiak Aitor jan du** esaldia zuzena dela erabaki dezakegu, adibidez, erregistro poetikoan edo narratiboan erabili denean. Erregistroaren arabera, testu motaren arabera eta, noski, idazlearen ezagutza-mailaren arabera egitura bat *erroretzat* hartzea edo ez hartzea erabaki daiteke. Informazio hori guztia biltzeak eta kudeatzeak ekarri dezakeen lana handia denez, oraingoz ez dugu erroreekin halako bereizketarik egin.

Erroreetako bakoitzak bere zuzenketak izango ditu.

⁶Aditz laguntzailea finkatu dugu esaldia zuzentzeko. Zuzenketarako aditz laguntzailea aldatuz gero, **Bonboiak Aitorrek jan ditu** edo **Aitorrek bonboiak jan ditu**. . . zuzenketak lortuko genituzke.

1. Ortografikoak
2. Lexikoak
3. Morfologikoak, sintaktikoak eta morfosintaktikoak
 3. 1. Deklinabidea
 3. 2. Determinatzailea
 3. 3. Izenordainak
 3. 4. Adjektiboak eta adberbioak
 3. 5. Aposizioak
 3. 6. **Postposizio-lokuzioak**
 3. 7. Aditza
 3. 7. 1. Aditzaren denbora, aspektua edota mota nahastea
 3. 7. 2. **Aditz-paradigmen nahasketa**
 3. 8. **Komunztadura**
 3. 8. 1. **Sintagma barruko komunztadura eza**
 3. 8. 2. Komunztadura eza aposizioan
 3. 8. 3. **Komunztadura eza perpausean**
 3. 8. 3. 1. **Aditza - subjektua**
 3. 8. 3. 2. **Aditza - objektua**
 3. 8. 3. 3. **Aditza - zehar-objektua**
 3. 8. 3. 4. Aditza - predikatua
 3. 8. 4. Komunztadura eza esaldian (mendekoetan)
 3. 8. 5. Komunztadura eza testuinguruan
 3. 9. Mendeko esaldiak
 3. 9. 1. Konpletiboak
 3. 9. 2. Zehar galderak
 3. 9. 3. Helburuzkoak
 3. 9. 4. Kausazkoak
 3. 9. 5. Baldintzazkoak
 3. 9. 6. Erlatibozkoak
 3. 9. 7. Denborazkoak
 3. 9. 8. Kontzesiboak
 3. 9. 9. Moduzkoak
 3. 9. 10. Konparaziozkoak
 3. 9. 11. Nominalizazioa
 3. 10. Perpausen egitura
 3. 11. Juntagailuak eta lokailuak
 3. 12. Bestelakoak
4. Nozioak
 4. 1. Ordua
 4. 2. **Datak**
 4. 3. Bestelakoak
5. Semantikoak
6. Puntuazio-markak
7. Estilo kontuak

II.1 Irudia: Euskarazko erroreen sailkapena. Kategoria nagusiak.

II.4.3 Aukeratutako errore-kategoriak

Ikusi berri dugun sailkapeneko hiru errore-kategoria aukeratu eta landu ditugu tesi-lan honetan. Kategorია horiek *datak*, *postposizio-lokuzioak* eta *komunztadura* dira (II.1 irudian letra lodiz adierazita agertzen dira). Erroreen aukeraketa egiterakoan kontuan hartu ditugun ezaugarriak ugari izan dira (aukeraketa-arrazoiak errore bakoitzaren deskribapenean adieraziko dira dagozkien kapituluetan, V. eta VII. kapituluetan, hain zuzen ere), eta horien artean bat, erroreen maiztasuna izan da. Maizen egiten diren erroreak landu beharko liratekeela uste dugu. Datu hori, ordea, ez dugu inolako dokumentutan argitaratuta topatu, eta ez dugu kalkulatzeko aukerarik izan. Egokiena, maiztasun-informazioa euskaraz egiten diren errore guztiakin eskuz etiketatutako corpus batetik erauztea litzateke, baina corpus hori ez da existitzen. Corpus erroredun gordina bilduta daukagu eta zenbait errore etiketatzen hasiak gara (determinatzaile- eta komunztadura-erroreak, adibidez), baina errore guztiak etiketatuta dituen testurik ez dugu. Hori dela eta, maiztasunak arestian aipatutako inkestan jasotako erantzunen azterketatik jaso ditugu. Jakin badakigu, hau ez dela datuak jasotzeko modurik egokiena, subjektiboa baita hamalau iritziren ondorio izanik, eta gainera, sailkapenaren mende dago⁷. Dena den, oraingoz modu bakarra da, eta adituen esperientzia handiaz fio gara. Aipatutako hiru multzoetako adituen (*euskara teknika* irakasgaiko irakasleak, euskaltegiako irakasleak eta zuzentzaileak) ustez, orokorrean gehien egiten diren erroreak II.1 taulan bildu ditugu. Taula horretan bi zati bereiz ditzakegu: lehen zatian, adituen ustez *askotan* gertatzen diren erroreak bildu ditugu; bigarrenean, lehen zatian azaltzen ez diren arren, maiztasun handikoak direnak (*askotan*, A, eta *batzuetan*, B, aukeren batura altua dutenak).

Datak idazterakoan akatsak oso maiz egiten direla erakusten du II.1 taulak (lehen postuan agertzen da). Era berean, esaldiaren barruan zein sintagmaren barruan, elementuen arteko komunztadurak ez direla ongi egiten argi geratzen da. Postposizio-lokuzioetan gertatzen diren erroreak ez dira maiztasunen taula honetan agertzen, ez baitziren adituek jaso zuten inkestan agertzen. Hala ere, duten interes linguistikoagatik aukeratu ditugu.

Datak idazterakoan egiten diren erroreen sailkapen zehatza V. kapituluko V.2.2 atalean egingo dugu. Postposizio-lokuzioekin egiten direnak kapitulu bereko V.3.2 atalean izango ditugu aztergai. Azkenik, komunztadura-erroreak sakon aztertuko ditugu VII. kapituluko VII.3 atalean.

⁷Inkesta egiteko erroreen sailkapeneko lehen bertsioa erabili zen. Bertsio honetako errore-kategoria guztiak azken bertsioan agertzen dira, nahiz eta batzuetan izen ezberdina duten.

ASKOTAN ematen diren erroreak:	
Inkestak (14tik)	Errorea
11	Datak gaizki adieraztea
11	Hitzak oker ordenatzea
10	Kontagarria/kontaezinak nahastea
9	hobe/hobeto nahastea
8	Aditzen erregimena okerra
8	Esaldien barneko osagaien komunztadura okerra
7	Adberbioak gaizki erabiltzea
7	Moduzkoak nola/nolako nahastea
6	Baso bat ur/ur baso bat
5	Esaldi erreflexiboak gaizki
ASKOTAN + BATZUETAN maiztasunen batura:	
B (4) + A (3)	Zatia osatzen duten elementuen arteko komunztadura okerra
A (4) + B (2)	Esaldia osatzen duten perpausen arteko komunztadura okerra
B (4) + A (1)	Helburuzko perpausetan “joan” eta “etorri” aditzekin “-tzeko” erabiltzea “-tzera” beharrean

II.1 Taula: Euskaraz egiten diren erroreak maiztasunaren arabera sailkatuta.

II.5 Ebaluazioa

Kapitulu honetan, erroreen definizio bat eman ondoren, errore horiek izan ditzaketen corpusak biltzeko eta sailkatzeko modua erakutsi dugu. Beranduago, erroreen sailkapen bat eman dugu, eta azkenik, honako azpiatal honetan, erroreen detekziorako eta zuzenketaarako teknikak aplikatu ondoren, emaitzak ebaluatzeko moduari buruz arituko gara.

Ebaluazioak berebiziko garrantzia du sistema baten baliagarritasuna neurtzeko. Ebaluazioaren bidez, aldezturik ipinitako helburuak bete ote diren ezagutu ahal izango dugu, eta ondorioak atera. Bi sistema konparatu nahi direnean, ebaluaziorako metodo eta metrika berdinak erabiltzea ezinbestekoa da; beraz, horien estandarizazioan lanean ari dira azken urteotan. Starlander eta Popescu-Belis-en arabera (2002) “The methodology of evaluation in Natural Language Processing (NLP) has witnessed considerable standardization efforts in the past decade.” Lan berean, ebaluazio prozesu estandar bat aplikatzen diete frantseserako bi ortografia- eta gramatika-zuzentzaileari, erroreen ebaluaziorako espezifikazio bat sortuaz. Espezifikazio horretan aipatzen diren ebaluazio-prozesuko bost faseak jarraituko ditugu tesi-lan honetan. ISO-k (*International Organization for Standardization*) eta IEC-k (*International Electrotechnical Commission*) zehaztutako ISO/IEC 14598-5 (for Standardization/ International Electrotechnical Commission, 1998) txostenean ebaluazio-prozesua fase hauetan banatzen da:

- Kalitate eskaerak ezartzea
- Ebaluazioaren zehaztapena
- Ebaluazioaren diseinua
- Ebaluazioaren exekuzioa
- Ebaluazioaren ondorioak (eta idazketa)

Azkena ezik, gainontzeko faseak hurrenez hurren deskribatuko ditugu. Ebaluazioaren ondorioak, landutako errore bakoitzari dagokion kapituluaz azalduko ditugu.

11.5.1 Kalitate eskaerak ezartzea

Sarrerako kapituluaz aipatu genuen moduan, erroreak detektatzea eta zuzentzea oinarritzeko lana da *ortografia- eta gramatika-zuzentzaileetan* edota *Ordenagailuz Lagundutako Hizkuntzen Irakaskuntza Adimendunean (OLHIA)*. Bi aplikazio mota horietan erabiltzaileak berebiziko garrantzia du, eta hortaz, erabiltzaileen profila eta beharrak ezagutu behar dira tresnaren kalitatea neurtu nahi bada. Adibidez, *Microsoft* etxean haien gramatika-zuzentzaileak diseinatzeko eta ebaluatzeko, erabiltzaile taldeak aztertzen dituzte. Aztertutakoen artean haien lehentasunak (zuzentasunari edo ulergarritasunari ematen dioten garrantzia, adibidez), idazten dituzten dokumentu motak, eta dituzten beharrak daude (Helfrich eta Music, 2000; Riley *et al.*, 2004). Domeij *et al.* (2002) egileen arabera ebaluazioa egiterakoan, erabiltzailearen azterketa egiteaz gain, erroreen konplexutasuna eta aurkezpena neurtu behar dira: “(...) aspects such as user abilities and needs, variability of text genres and user groups, the complexity of error types and error presentations must also be taken into consideration”. Gure kasuan, erabiltzaileak, testu-prozesadore erabiltzaileak eta euskara-ikasleak dira. Hau horrela izanik, honako ezaugarriei emango diegu garrantzia:

- **Doitasuna eta alarma faltsuak.** Riley *et al.* (2004) egileen arabera, gramatika-zuzentzaileen erabiltzaileek sistemaren doitasun-akatsak ez dituzte ongi onartzen, eta alarma faltsuak gertatzen direnean, sistemaren balioari buruz duten iritzia okertu egiten da. Gainera, erabiltzaile askori ordenagailuak hutsik gabekoak direla iruditzen zaionez, alarma faltsuek zalantza ipin dezakete gramatikalki zuzena denari buruzko gure ezagutza. Hori ez da inola ere onargarria hizkuntzen irakaskuntzako sistemetan. Argi geratzen da, beraz, doitasun altua

lortzea eta, batez ere, alarma faltsuak ekiditea oso garrantzitsua dela errorearen detekzioan. Tesi-lan honetan bereziki landu ditugu alarma faltsuak.

- **Mezua.** Erabiltzaileak ongi ulertu behar du egin duen errorea. Errorea detektatu denean erabiltzaileari erakutsiko zaion mezuak argia eta ahalik eta zehatzena izan behar du. Hori lortzeko, erroreak kodetu ditugunean, kodean bertan gero mezua “sortzeko” erabiliko dugun informazioa gehitu dugu. Adibidez, KOM_NAH_SUB-KAS_N-NK kodetik mezu hau lortuko genuke:”KOMunztadura errorea, NAHasketa gertatu da SUBjektuaren KASu-markan; ez dator bat NOR-NORK paradigmako (N-NK) aditz laguntzailearekin”.

Erabiltzailea alde batera utzita, eta tresnen alderdi teknikoari begira, sistemen kalitatea ISO/IEC 9126 estandarrak (Engineering, 2000) neurtzen du honako ezaugarriak aztertuta: funtzionalitatea (*“functionality”*), fida-garritasuna (*“reliability”*), erabilgarritasuna (*“usability”*), eraginkortasuna (*“efficiency”*), mantentzeko erraztasuna (*“maintainability”*) eta garraiagarritasuna (*“portability”*). Gurera ekarrita, funtzionalitateari eman diogu garrantzi handiena eta hau honela definitu dugu: *gramatika-erroreak detektatzeko eta zuzentzeko gaitasuna*.

II.5.2 Zehaztapena eta diseinua

Ebaluazioa nola egin dugun azaltzeko, lehenik, erabili ditugun “neurriak” azalduko ditugu, geroago diseinua azalduko dugularik.

Ebaluaziorako neurriak

Gure sistemen *“gramatika-erroreak detektatzeko eta zuzentzeko gaitasuna”* neurtzeko, errorearen *detekzio* eta errorearen *zuzenketarako* neurriak bereizi beharko genituzke (Starlander eta Popescu-Belis, 2002; Guinovart, 1996a). Gure kasuan, zuzenketarako lehen urratsak egin ditugun arren, ez dugu ebaluaziorik egin; beraz, ez dugu zuzenketarako neurririk emango. Etorkizunean egin beharrekoen artean utzi dugu zuzenketaren ebaluazioa.

Detekzioari buruz ari garenean, ohikoa da *doitasuna* eta *estaldura* ematea⁸. *Doitasunak* sistemak detektatzen dituen errore guztien artean (ongi edo gaizki detektatuak) ongi detektatutakoak zenbat diren adierazten digu. *Estaldurak*, berriz, detektatu beharko liratekeen errore guztietatik, sistemak

⁸Emaizta-tauletan, doitasuna **D** hizkiarekin eta estaldura **E** hizkiarekin adierazi ditugu.

ongi detektatuak zenbat diren adierazten digu. Ondorengo ekuazioek kalkulatuko lituzkete neurri horiek:

$$doitasuna = \frac{\text{ongi detektatutako erroreak}}{\text{ongi detektatutako erroreak} + \text{alarma faltsuak}}$$

$$estaldura = \frac{\text{ongi detektatutako erroreak}}{\text{detektatu beharko liratekeen errore guztiak}}$$

Horien baliokideak dira (Starlander eta Popescu-Belis, 2002) lanean sistemek esaldi erroredunak detektatzeko duten gaitasuna ebaluatzeko ematen dituzten neurriak⁹:

$$doitasuna = \frac{TP}{TP + FP}$$

$$estaldura = \frac{TP}{TP + FN}$$

Doitasunak eta estaldurak Oren eta 1en arteko balioa hartzen duten arren, guk emaitzak ehunekoetan emango ditugu.

Guinovart-en arabera (1996a), gramatika egiaztatzen duen sistema batek erredeen detekzioarako mekanismo oso murritzailak dituenean, doitasun oso altua izan ohi du eta estaldura txikia. Mekanismo ez-murritzailak dituen sistema batek, ordea, estaldura handia izango luke eta doitasun eskasa. Aurreratu dugun moduan, erabiltzaileak doitasuna estalduraren aurretik jartzen du normalean, ez baititu alarma faltsuak ongi toleratzen. Hori dela eta, tesi-lan honetan lehentasuna emango diogu doitasun on bat lortzeari, estaldura ahalik eta balio altuenetan mantentzen saiatuz.

⁹Ekuazioetako laburtzapenen esanahia hau da:

- Egiazko positiboak (*true positives (TP)*): egitura okerrak sistemak “oker” moduan markatuta. Erroreak ongi detektatu dira.
- Egiazko negatiboak (*true negatives (TN)*): egitura zuzenak sistemak markatu gabeak (zuzenak zuzen markatzea izango litzateke). Kontzeptu hau sistemek esaldi zuzenak detektatzeko duten gaitasuna neurtzeko erabiltzen da.
- Gezurrezko positiboak (*false positives (FP)*): alarma faltsuak. Egitura zuzenak sistemak “oker” moduan markatuta.
- Gezurrezko negatiboak (*false negatives (FN)*): egitura okerrak sistemak “zuzen” moduan markatuta. Hau da, egitura okerrak markatu gabeak.

Erakutsitako neurriek *esaldi edo egitura okerrak detektatzeko gaitasuna* neurtzen dutela esan dugu. Era berean neur zitekeen *esaldi edo egitura zuzenak detektatzeko gaitasuna* baina, gure ustez, lehen neurria ulerterrazagoa da esku artean dugun ataza kontuan hartuta.

Riley *et al.* (2004) egile taldearen ustez, erabiltzaileari sistema batek *ehuneko* hainbateko doitasuna eta *ehuneko* bestelako estaldura duela esatea ez omen da oso esanguratsua, eta ez omen du sistemaren erabilgarritasunaren aztarnarik ematen. Haiek, errore motako doitasuna eta estaldura emateaz gain, *orrialdeko alarma faltsu kopurua* (“false flags per page” (FFPP)) neurria ematea proposatzen dute. Balio hori kalkulatzea erraza den arren, ez ematea erabaki dugu. Beste sistemekin konparaketak egin ahal izateko, ohikoagoak diren neurriak erabiltzea erabaki dugu.

Azkenik, gogora dezagun estalduran eta doitasunean eragin zuzena dutela aukeratutako corpusek (corpusen deskribapena II.3.1 atalean egin da). Ondorengo atalean, besteak beste, ebaluaziorako corpus horietan erabaki dugun banaketa azalduko dugu.

II.5.3 Exekuzioa

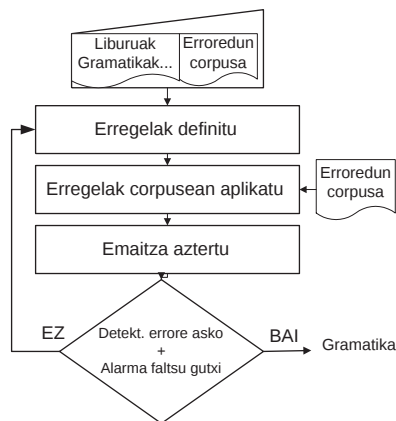
Aipatutako corpusetako bakoitza (erroreduna zein errorerik gabea) bi zatitan banatu dugu. Adibide bilduma bakoitzaren % 60 erregelen *garapenerako eta birfinketarako* erabili dugu, eta gainontzeko % 40 *probarako*.

Garapenerako corpusa behin eta berriro aztertu dugu azterketa ziklo bakoitzean bi urrats eginez: a) erregelak corpusean aplikatu ditugu, eta b) emaitzak aztertu ditugu detektatu ez diren errorean eta alarma faltsuen bila. Topatutako hutsuneak aztertu ondoren, erregelak birfindu egin ditugu hobetze aldera eta zikloarekin jarraitu dugu, emaitzak ahalik eta hoberenak zirela uste izan dugun arte. Lan hau gramatiken garatzaileak egin du.

Probarako corpusean erregelak behin soilik pasa ditugu, ez dugu inolako erregela birfinketarik egin, eta lan hau, noski, garatzailea ez den pertsona batek egin du.

Helburu desberdinekin aztertu ditugu corpus erroreduna eta zuzena. Lehenengoan, *errore-erregelen detekzio-maila* neurtzea izan dugu helburu, eta bigarrenean, berriz, errore-erregelek sor ditzaketen *alarma faltsuak* neurtzea. Helburu zehatzak horiek izan diren arren, erroreak eta alarma faltsuak bi corpusetan aztertu ditugu (zuzenean ere erroreak topatu ditugu). Honako neurriak kalkulatu ditugu¹⁰:

¹⁰Oharra: corpus zuzenak oso handiak direnez, ez dugu bertan dagoen errore kopurua zehaztu, ezta detektatu gabeko errorean kopurua ere.



II.2 Irudia: Gramatika eratze zikloa. Fluxu-diagrama.

- Adibide kopurua
- Erroreak: ongi detektatutako erroreak + errore detektatu gabeak
- Alarma faltsuak

Normalean, errazagoa izaten da adibide “zuzenen” corpora topatzea adibide “erroredunena” baino. Hori horrela izanik, corpus “zuzenak” (zuzen deitu ditugun corpus arruntetan erroreak ager daitezke) dezente handiagoak izaten dira. Erroreen eskuzko etiketatzea lan neketsua eta luzea denez, ez dugu erroreekin etiketatutako corpus arruntik lortzeko aukerarik izan. Corpusean dagoen errore kopurua ezagutu ezean, ezinezkoa da estalduraren neurria ematea. Beraz, aipatutako tarteko emaitzak erabilita, *doitasuna* eta *estaldura* neurtu ditugu corpus erroredunetan, eta *doitasuna* soilik corpus zuzenean.

Corpus zuzenean, alarma faltsuen kopurua adierazteaz gain, *alarma faltsuen tasa* ere kalkulatu dugu. Azkenak erlatibizatu egiten du gezurrezko positiboaren kopurua, kopuru hau adibide guztiak kontuan hartuta ematen baita. Hau da, corpusean 29 alarma faltsu gertatu direla esatea baino, 7769 adibideen artean % 0,37 kasuetan alarma faltsua gertatu dela esatea esanguratsuagoa da gure ustez. Atal honetan aipatutako guztiaren laburpen bat ikus dezakegu II.3 irudian.

Corpusak	Corpus erroreduna (Erroreak + alarma faltsuak)	Corpus arrunta (Alarma faltsuak + erroreak)
	Garapenerako % 60 Probarako % 40	Garapenerako % 60 Probarako % 40
Neurriak	Doitasuna + Estaldura	Doitasuna
Tarteko emaitzak	Elementu kopurua Errore detektatu gabeak Ongi detektatutako erroreak Alarma faltsuak	Elementu kopurua Ongi detektatutako erroreak Alarma faltsuak Alarma faltsuen tasa

II.3 Irudia: Ebaluazioan erabilitako corpusak eta neurriak.

II.6 Laburpena eta ondorioak

Erroreen inguruan erabiltzen diren termino anitz deskribatu eta azaldu ditugu kapitulu honetan. *Erroreak*, *desbideratzeak*, *estilo desagokia* eta *hitz zuzenen erabilera nahasketa* bereizi ditugu. Egitura horien guztien *detekzioa*, *diagnosia* eta *zuzenketa* egin daiteke hizkuntzaren tratamendu automatikoa baliatuz. Erroreak dira tesi-lan honetan landu ditugunak, errore sintaktikoak zehazkiago, eta hauetan erroreen detekzioa, diagnosia eta zuzenketa egingo ditugu.

Terminoak ezagutzen baditugu ere, lanean hasi arte ez gara jabetuko haiek bereizten dituen marraren fintasunaz. Errore eta estilo desagokia bereiztea maiz zaila da. Bestalde, estandarizazio falta dela eta, batzuek errore kontsideratzen duten egitura, besteek zuzentzat har dezakete. Adibidez, **gosea daukat** egitura Ilari Zubiriren gramatikaren arabera (Zubiri, 1994) okerra da, baina Euskaltzaindiako JAGONET zerbitzuko¹¹ erantzunen arabera, oraindik ez da ezer erabaki gai horren inguruan, eta *Orotariko Euskal Hiztegia*n badaude gosea izan/eduki modukoak.

Errorea edo estilo desagokia izan, gramatika-zuzentzaile batek edo OLHIA sistema batek markatu behar al ditu egitura horiek? Gure ustez, bai. Erroreak, desbideratzeak, estilo desagokia eta nahastutako hitzak, guztiak, markatu behar dira. Erabiltzaileari egiturak aurkezteko moduan egin behar dugu bereizketa. Erroreak edo hitz nahasketak topatzen ditugunean diagnosia egin ondoren ordain zuzenak ematea litzateke aproposena. Errore ez diren desbideratzeak eta estilo desagokia erabiltzen direnean, berriz, egitura egokiago bat eskainiko litzaioke erabiltzaileari, baina ez hasierakoa zuzendu.

¹¹<http://www.euskaltzaindia.net>

Egitura horiek corpus erroredunetan topatzen dira, baita batzuetan corpus arruntetan ere. Corpus erroredunak biltzeko irizpideak ezarri ditugu eta atazak aurrera eramateko prozedura ere azaldu dugu. Prozesua neketsua izanagatik, oraindik testu erroredunak biltzen, kodetzen eta etiketatzen dihardugu. Hala ere, tamaina handi samarreko testu multzoa bildu dugu besteak beste tesi-lan honetan erabiltzeko.

Corpusetan eskuz topatutako erroreak aztertu ditugu eta errorearen sailkapen bat osatu dugu. Sailkapenerako aurkeztutako kategoria-sisteman irizpide linguistikoa eta horren barruan irizpide deskriptiboa erabili ditugu. Kategoriak zehaztasunez deskribatuta egonik ere, zenbait egitura erroredun kategoria batean edo bestean kokatzea zaila suerta daitekeela ikusi dugu. Erroreen detekzio automatikoa egiten dugunean, arazo bera izango dugu, egitura bati hainbat errore-etiketa desberdin egokituko baitzaizkio.

Azkenik, errorearen tratamendu automatikoan lortutako emaitzak ebaluatzeko metodoa azaldu dugu. Estandarrei jarraituta prozesua zehaztu ondoren, ebaluazioa diseinatu dugu eta ebaluaziorako erabiliko ditugun neurriak deskribatu ditugu. Ebaluazioaren emaitzak, errore mota bakoitzari lotutako kapituluetan emango ditugu.

Prozesu honen guztiaren ondoren, hiru fenomeno linguistiko aukeratu ditugu lantzeko: *datak*, *postposizio-lokuzioak* eta *komunztadura* (zatietan zein perpausetan).

Ondorengo kapituluan, erroreak detektatzeko eta zuzentzeko tekniken eta metodoen inguruan arituko gara.

III. KAPITULUA

Erroreak detektatzeko eta zuzentzeko teknikak

III.1 Sarrera

Errore kontzeptuaren inguruan aritu ondoren, hauek detektatu eta zenbait kasutan zuzendu ahal izateko *hizkuntzaren tratamendu automatikoa* erabiltzen duten teknikak azterketari ekingo diogu.

Kapitulua hasteko, erroreak detektatzeko erabiltzen diren teknikari begiratu zabal bat emango diegu. Tekniken deskribapena *erroreak detektatzeko teknikak* izena duen atalean kokatu dugu (III.2 atala). Geroago, III.3 atalean, erroreak zuzentzeko teknikak inguruan arituko gara. Erroreen detekzioaren eta zuzenketaren aplikazio-esparru nagusiak gramatika-zuzentzaileak eta OLHIA sistemak direla esan dugu. Azterketa bibliografikoan bi alor horietan topatu ditugun sistemak zerrendatuko ditugu, III.4 eta III.5 ataletan, hurrenez hurren (sistema guztien laburpenak egin ditugu A eranskinean). Kapituluaren amaitzeko, bertan esandakoen laburbilpen bat egingo dugu III.6 atalean. Atal berean, erroreen detekzioarako tesi-lan honetan aukeratu ditugun teknikak berri emango dugu, aukeraketa arrazoituz.

III.2 Erroreak detektatzeko teknikak

Erroreak detektatzeko eta zuzentzeko erabiltzen diren teknikak ugariak dira, hizkuntzaren tratamendu automatikoan (lengoaia naturalaren prozesamenduan, LNP) erabiltzen diren ia teknika guztiak eginkizun zehatz honetarako erabil daitezkeela esan baitezakegu. Teknika guztiek zenbait eredu formaletan dute oinarria. Jurafsky eta Martin-en arabera (2000), eredu horiek

egoera-makinak, erregela-sistema formalak, analisi logikoa, probabilitatearen teoria eta ikasketa automatikoan oinarritutako erremintak dira.

Egoera-makinak eta *erregela-sistema formalak* fonologiako, morfologia-ko eta syntaxiko ezagutza lantzeko tresna nagusiak dira. Syntaxian duten erabilera III.2.1.2 atalean aztertuko dugu.

Semantika, pragmatika eta diskurtsoarekin lan egitean erabiltzen den eredu formala *analisi logikoa* da normalean, baina eremu hauek gure helburuetatik kanpo geratzen dira.

Probabilitatea da aitatu dugun azken eredu formala. Probabilitatearen erabilera nagusia anbigutasun-arazoak ebaztea da eta LNPko arazo gehienak “sarrera (*input* ingelesez) anbiguo batentzako N aukera emanda, probabilitate handienekoa aukeratu” moduan definitu daitezkeenez, oso erabilia da. *Ikasketa automatikoarekin* oso lotuta dago probabilitatea. Bi eredu hauek III.2.2 atalean aztertuko ditugu.

Tekniken eredu formalak azertu ondoren, oinarrian erabiltzen duten informazio mota irizpide hartuta, honako sailkapena egin dugu:

1. Hizkuntza-ezagutza oinarritutako teknikak.
2. Corpusetan oinarritutako teknikak.

Hizkuntza-ezagutza oinarritutako teknikak, teknika *sinboliko* izena ere jasotzen dute. Metodo sinbolikoak honela definitzen dira (Dale, 2000) lanean: “symbolic methods —where knowledge about language is explicitly encoded in rules or other forms of representation—”. Hau da, teknika sinbolikoetan hizkuntzari buruzko ezagutza erregeletan edo beste adierazpideetan kodetzen da modu esplizituan. Teknika sinbolikoak oso erabiliak izan dira azken 40 urteetan, baina mundu errealeko arazoei aplikatzen zaizkienean *hedagarritasunaren arazoa* dute.

Arazo honen aurrean, azken 10 edo 15 urteotan *corpusetan oinarritutako teknikak* edo teknika *enpirikoak* hazkunde izugarria izan dute. Horietan, datu kopuru oso handiak erabiltzen dira (LNPrek kasuan corpusak), eta horiekin batera, estatistikaren inguruko prozedurak. Ezagutza testuetatik automatikoki erauzi egiten da, erregeletan eskuz kodetu beharrean.

Gaur egun, giza-hizkuntzaren prozesamenduaren gakoa teknika sinboliko eta enpirikoen konbinazio egokian dagoela onartzen dute ikertzaile gehienek.

III.1 taulan, hizkuntza-ezagutza oinarritutako teknikak (sinbolikoak) eta corpusetan oinarritutako teknikak (enpirikoak) arteko desberdintasun nagusienak adierazten saiatu gara:

Metodo sinbolikoak erabiltzen dituzten teknikak hurrengo atalean aztertuko ditugu (III.2.1). Enpirikoak, berriz, III.2.2 atalerako utzi ditugu.

	Sinbolikoak	Enpirikoak
Erraz hedagarria	Ez	Bai
Hizkuntza-ezagutzarekiko dependente	Bai	Ez
Corpus handien beharra	Ez	Bai

III.1 Taula: Teknika sinbolikoen eta enpirikoen arteko desberdintasunak.

III.2.1 Hizkuntza-ezagutzan oinarritutako teknikak (sinbolikoak)

Hizkuntza-ezagutzan oinarritutako teknikan erregelak eskuz garatzen dira eta analisirako informazio linguistiko anitz behar izaten dute. Teknika hauek hiru azpimultzotan banatu ditugu behar duten informazio linguistikoaren arabera:

1. Ezagutza morfologikoa edo hitz-mailako informazioa behar duten teknikak.
2. Ezagutza sintaktikoa edo esaldi-mailako informazioa behar duten teknikak.
3. Esaldi-mailatik haratago dagoen ezagutza —semantika, pragmatika eta diskurtsoa— behar dituzten teknikak.

Multzo hauek III.2.1.1, III.2.1.2 eta III.2.1.3 ataletan aztertuko ditugu, hurrenez hurren. Garrantzirik handiena sintaxia oinarrian duen atalari eman diogu, hauxe baita gure lanean bereziki landu dugun atala.

III.2.1.1 Ezagutza morfologikoa (hitz-maila)

Hitza, sintagma eta goragoko mailetak egituren osagaia den unitatea da. Hitzen barne-egitura *morfologiak* lantzen du eta hitza lemekin eta zenbaitetan morfemekin osatuta egoten da. Batzuetan, morfemetan esaldi-mailako funtzioak topa ditzakegu, adibidez, morfema batek hitzak objektu edo subjektu funtzioa bete dezakeela adieraz dezake. Kasu horietan, hizkuntza-unitateak propietate morfologikoak eta sintaktikoak edukitzen ditu, eta hauen definiziorako *morfosintaxia* terminoa erabiltzen da. Euskara hizkuntzaren izaera eranskaria dela eta, morfosintaxia terminoa oso erabilia da.

Hitzeekin maila morfologikoan lan egiteko *adierazpen erregularrak* eta *automatak* erabiltzen dira normalean. Teknikaren aukeraketa, hizkuntzaren ezaugarriek erabakitzen dute maiz. Adibidez, ingelesaren antzeko hizkuntzetan errore ortografikoak lantzeko, hitz baten forma flexionatu guztiak

zerrenda batean gorde daitezke, eta zerrenda horrekin hizki-zuhaitz moduko egoera finituko automata bat eraiki. Horrela, hitz zuzenak ezagutuko lirateke eta hitz erroredunak baztertu. Suomiera, hungariera, turkiera eta euskara moduko flexio-maila handiko hizkuntzetan forma flexionatu kopuru izugarria sortzen da, eta arazoak egoten dira forma guztiak hiztegian gorde ahal izateko. Arazoa konpontzeko premiazkoa da hitzak morfologikoki behar bezala lantzen dituzten teknikak erabiltzea.

Ondorengo lerroetan, egiaztatzaile-zuzentzaile ortografikoetan erabiltzen diren tekniken errepasoa egingo dugu.

- Egiaztatzaile-zuzentzaile ortografiko batzuetan, egiaztatzaileak hiztegi bat gordetzen du sisteman erabilgarria den hizkuntza bakoitzerako. Hiztegi horietan lemak eta morfemak gordetzen dira hitzen sorkuntzarako zenbait erregelekin batera. Adibidez, ingeleserako hiztegian **twenty/H** bikotea gordeko genuke. Lehena hitzaren lema izango litza-teke, eta bigarrena, hitzaren flexioa adieraziko lukeen honako H izeneko erregela:

(8) H etiketa:

Y >-Y, IETH # As in twenty >twentieth
 [^ Y] >TH # As in hundred >hundredth

Zuzentzaileak **twentieth** topatzen duenean, erregela atzekoz aurrera erabiltzen du **iet** ezabatuz eta **y** gehituz. Gero hiztegian begiratzen du **twenty** zuzena ote den. Teknika hau *Ispell*¹ (*Unix* sistema eragileak, *emacs* editorea), *Aspell*² (*Opera*, *AbiWord*, *LyX* eta *gedit* tresnak), *Hunspell*³ (morfologia aberatsa eta hitzen konposaketa konplexua duten hizkuntzetarako, *OpenOffice*ko aplikazioetan) eta *MySpell* (*OpenOffice*, *Mozilla firefox* eta *Vim*) kode irekiko zuzentzaile ortografikoetan erabiltzen da.

- Aurreko paragrafoan aipatutako teknika flexio handiko hizkuntzetan erabiltzea oso zaila dela esan dugu dagoeneko. Erabili beharreko hiztegiak tamaina itzela izango luke, bertan lema, eta batez ere, morfema kopuru izugarria gorde beharko bailitzateke. Horren aurrean, erroreak onartzen dituzten egoera finituko eredu ezagutzaileak erabiltzen ditu Oflazer-ek (1996) turkierarako. Lemak eta morfemak automata batean

¹<http://fmg-www.cs.ucla.edu/geoff/isspell.html>

²<http://aspell.sourceforge.net>

³<http://hunspell.sourceforge.net>

kodeitzen ditu, eta hitz erroredunen analisi morfologikoa eta zuzenketa egiten ditu aldi berean. Erroreak onartzen dituen egoera finituko ezagutzaile honek, hitz zuzenetatik *urruntzen* diren karaktere-kateen ezagutza ahalbidetzen du. Karaktere-kate bat beste batean bihurtzeko zenbat edizio eragiketa (eransketa, ezabaketa, aldaketa, transposizioa) behar diren neurtzen duen *edizio-distantzia* izeneko kontzeptua erabiltzen du. Automatan hasierako egoeratik bukaerako egoeretakoren batera dauden bide guztiak aztertzen dira, eta bideren bateko etiketa guztiak kateatzen direnean lortzen dugun karaktere-kateak sarrerako kate okerretik edizio-distantzia zehatz batera egon behar du.

- Antzeko gauza egiten da turkieraren moduan eranskaria den euskararako. Hitzen analisirako Koskeniemi-k (1983) garatutako *bi mailatako (lexikoa, azalekoa) morfologian* oinarritutako analizatzailea erabiltzen da, zeinetan analisisa eta sorkuntza egiten diren. Hitzen analisisa egiteko, sarreran *azaleko* forma (adibidez, **etxean**) jasotzen da eta emaitzan dagokion adierazpen *lexiko* zuzena itzultzen da (adibidez, **etxe[IZE][ARR]+an[INE]**). Sorkuntzan, forma lexikala emanda, azaleko forma lortzen da. Azpilexikoen multzoa eta erroen eta afixuen sekuentzia posibleak erregulatzen dituzten *jarraitze-klaseak* osatzen dute sistema. Bi mailatako erregelek, transformazio morfofonologikoen eraginez sakoneko eta azaleko mailen artean sortzen diren diferentziak adierazten dituzte. Eta honetaz gain, eransketa eta ezabaketaren ondorioz sortutako erroreak deskribatzen dituzte. Analizatzaile honek hitzen bat ezagutzen ez duenean, hitzak erroreren bat duela suposatzen du, eta beharrezkoak diren proposamenak ematen ditu (ikus zuzenketa ortografikorako tekniken atala). *Xuxen* zuzentzaile ortografikoan (Agirre *et al.*, 1992) erabiltzen da.

III.2.1.2 Ezagutza sintaktikoa (esaldi-maila)

Hitza aztertu ondoren, unitate handiago bat aztertzerako joko dugu: esaldia. Sintaxiak hitzak sintagma izeneko egituretan antolatzen ditu eta egitura sintagmatiko horiek, era berean, esaldietan antolatzen ditu. Guk esaldiaren eta sintaxiaren arteko parekotasuna egin dugu, baina parekotasuna egitean kontuz ibili behar dugu, jakin badakigu eta, maiz hitzek esaldien arteko harremanak ezartzen dituztela; beraz, sintaxia esaldiaren mugatik haratago doala.

Esaldi bat sintaktikoki analizatzen denean, egituraren bat esleitzen zaio-lar esan dugu. Egitura horrek esaldiko osagai linguistikoak errepresentatzen

ditu, eta beren arteko harreman gramatikalak azalarazten ditu. Lan hori guztia modu mekanikoan gauzatzen duten algoritmo eta programek *parser* edota analizatzaile sintaktiko izena hartzen dute. Analizatzaileetako askotan hizkuntza-ezagutza *gramatiketan* kodetzen da erregelen bidez. Erregelek, *hizkuntza-egiturak* deskribatzen dituzte. Erroreen detekzioaren lan eremuan, hizkuntza-egiturak gramatikalak edo ez-gramatikalak izan daitezke.

Gramatika bat idazteko garaian, aukera ezberdinak egin daitezke (Gojenola, 2000). Erroreen detekzioaren ikuspuntua kontuan hartuta erabil daitezkeen hurbilpenen zerrendapena egingo dugu hurrengo lerroetan:

- **Hizkuntzaren teoria vs hizkuntzaren tratamendu automatikoa.** *Hizkuntzaren teoria* hizkuntza baten (edo hizkuntzen arteko) oro deskribatzeaz arduratzen den bitartean, hizkuntza modu automatikoan ulertzea eta sortzea du helburu *hizkuntzaren tratamendu automatikoak*. Azken honetarako ordenagailuak erabili beharrak, eragin handia izaten du sintaxiaren eta semantikaren teorizazioan, teoriak konputagarriak izan behar duten eragiketetara murriztu behar baitira. Hori egitean, LNP sistemek gehienbat formalitasuna, erazagutzaile izateko gaitasuna eta egokitasun linguistikoa mantendu behar dituzte (Shieber, 1986). Hizkuntzaren teoriako formalismoetan normalean datu gutxi eta laborategikoak erabiltzen dira, erabilitako esaldiak gramatikalak dira eta desanbiguazioa, puntuazioa eta antzeko fenomenoak ez dira tratatzen. Aplikazio errealetarako erabiltzen diren gramatiketan, berriz, datu askorekin eta esaldi errealekin lan egin behar da: esaldi gramatikalekin eta esaldi ez-gramatikalekin. Gure lana, hizkuntzalaritza aplikatuaren eta hizkuntzaren tratamendu automatikoaren eremuan kokatzen da.
- **Esaldiaren analisi osoa vs analisi partziala.** Testu errealekin lan egin beharrak, *analisi partzialaren* erabilera bultzatzen du. Edozein egitura, gramatikala zein ez-gramatikala, analizatzeko gai den sistema behar dugu, eta sistema horietan ezin izaten dira aurreikusi hizkuntza-egitura posible guztiak; beraz, ezin izaten da beti analisi osoa egin. Analisi partzialak, fidagarritasuna eta sendotasuna ditu helburu, sakontasuna eta osotasuna neurri batean galduaz. Beste aldetik, badira sistemak, teoria linguistiko baten azterketarako pentsatuak, *analisi osoa* bakarrik lortzea helburu dutenak, fenomeno linguistiko interesgarriak aztertzeko pentsatuta daudelako, eta ez testu errealetako esaldiak.
- **Gramatikak vs egoera finituko mekanismoak.** Lengoaia erregularrak, morfologia deskribatzeko nahikoak izan ohi diren bitartean, sintaxia deskribatzeko baliabide ahaltsuagoak behar izaten dira. Orokorrean,

zeregin horretarako, *testuingururik gabeko gramatikak (TGG)* erabili izan dira (Aho *et al.*, 1985), esaldien egitura hierarkiko eta errekursiboak definitzeko egokiak baitira. Bestalde, egoera finituko mekanismoak (automatak eta transduktoreak) oso erabiliak izan dira hizkuntzaren tratamendu automatikoan, eredu linguistikoetan oinarritutako patroiak definitzeko duten erraztasuna dela eta. Egin dugun erreten detekziorako sistemen analisian, bai *TGG*ak baita *egoera finituko mekanismoak* ere, testu erreal erroredunetan gertatzen diren egiturak deskribatzeko erabili direla ikusi dugu. Hori izango da erreten detekziorako teknikak eta sistemak sailkatzeko erabiliko dugun irizpidea.

- **Ikuspegi eraikitzailea vs murriztailea.** Informazio linguistikoarekin lan egiteko orduan, ikuspegi ezberdinak erabili izan dira. *Ikuspegi murriztailea* delakoan, lexikoiko informazioa erabiliz esaldi bat analizatzeko aukera posible guztiak sortzen dira, eta gramatikarekin edo erregelekin aukera horien artean onartezinak direnak baztertzen dira. *Ikuspegi eraikitzailean*, hitzetatik abiatuta konbinazio posibleak soilik deskribatuz zuhaitz-egitura eraikitzen saiatzen dira. Gramatikak eraikitzeko orduan, bi formalismoen konbinazioak egin ohi dira, egitura sintaktikoak ez daudelako hasieratik sortuta, eta gainera, azken analisien parte ez diren egiturak ere sortu egiten direlako. Egoera finituko mekanismoek ikuspegi murriztailerako joera, eta gramatikek oinarrian ikuspegi eraikitzaileko joera dutela, uste dugu. Beraz, bi ikuspegiak erabiltzen dira erreten detekziorako sistemetan.
- **Osagai-egitura vs mendekotasun-egitura.** Osagai-egituretan oinarritutako analisiaren ideia nagusia hau da: hitz multzoak *osagai* deituriko unitate bakar moduan kontsidera daitezke (Jurafsky eta Martin, 2000). Adibidez, izen-sintagma izeneko hitz multzoak unitate bakar moduan jokatzen du maiz. Mendekotasun-egituretan egitura sintaktikoak dependentzia izeneko erlazio bitar asimetrikoekin elkarren artean lotutako elementu lexikalekin osatuta daudela suposatzen da (Nivre, 2005).

Sintaxia lantzeko teknikak era askotarikoak dira, baita sintaxi-mailako erroreak detektatzeko teknikak ere. Guk honela sailkatu ditugu:

- Egoera finituko mekanismoak erabiltzen dituzten teknikak.
- Gramatikak oinarrian dituzten teknikak.
- Bestelakoak.

Egoera finituko mekanismoak

Lan honetan zehar, egoera finituko mekanismoei buruz aritzen garenean, patroi-erregelak aipatuko ditugu, behin eta berriro. Egia esateko, guk baliokidetzat hartuko ditugu, patroi-erregetan oinarritutako formalismo gehienek patroiak automata edo transduktore bihurtzen baitituzte.

Patroi-erregelak egitura edo ereduren bat definitzen duten erregelak dira. Gramatika-zuzentzaileetan, erregelak egitura ez-gramatikalak deskribatzeko erabiltzen dira. OLHIA sistemetan, berriz, batzuetan funtzio hori dute, baina maiz, sistemak espero duen erantzuna eta ikasleak eman diona konparatzeko erabiltzen dira.

Erregelen bidez deskribatzen ditugun patroiak dagoeneko morfosintaxi-mailan analizatuta dagoen testu baten kontra parekatzen ditugunean, patroi-parekatze teknika erabiltzen ari gara. Erregelek egitura oker bat deskribatzen dutenean, metodoa testuetan erroreak topatzeko erabil dezakegu. Hori horrela izanik, *erroreak gidatutako* teknika bat erabiltzen ari garela esaten da, erroreen detektziorako eta diagnosirako ezagutza negatiboa erabiltzen delarik. Patroi-parekatzearen teknikak abantaila anitz ditu; besteak beste: i) ez da esaldi osoa analizatu behar egitura mugaturen batean errorea dagoela salatzeke, ii) erregelak modu indibidualean erabil daitezke eta iii) erroreari buruz xehetasun handiko mezuak eskaini daitezke komentario lagungarriekin. Gainera, erregelak ulerterrazak izan ohi direnez, erabiltzaileek erregela multzoak erraz heda ditzakete errore-kasu sinpleak modu inkrementalean garatuz, kasu konplexuagoak definitu arte. *Patroi-erregelak* erregela-lengoaia ezberdinak erabiliz idatz daitezke, batzuk ezagunak diren formalismoetan (*Murritzapen Gramatika* eta *Xerox Finite State Tool* tresnak erabiliz), besteak norberak garatutakoetan.

Murritzapen Gramatika formalismoa (MG) (Karlsson *et al.*, 1995; Voutilainen eta Tapanainen, 1993; Tapanainen, 1996) ikuspuntu murritzaitetik definitu da, eta azken urteotan azaleko sintaxia eta desanbiguazioa lortzeko sistemarik arrakastatsuenetarikoa da. Desanbiguazioa egiten duela diogunean, forma baten interpretazioen artean, analisi zuzenak/egokiak aukertzen dituela esan nahi dugu.

Nahiz eta tresnaren helburu nagusia desanbiguazioa den, formalismoa testuari etiketak esleitzeko gai da islapen-erregelen bidez. Islapen-erregelak erabilita errore baten hasiera eta erreorearen bukaera deskribatzeko gai bagara, MGk errorea mugatzeko eta, beraz, detektatzeko aukera ematen digu.

Erroreen detektziorako maiz erabili da *Murritzapen Gramatika*. Katalanerako (Badia *et al.*, 2004), suedierarako (Birn, 2000; Arppe, 2000) eta norvegierarako (Johannessen *et al.*, 2002) sistemetan erabili dute formalis-

mo hau modu arrakastatsuan (ikus A eranskinean sistema hauen deskribapen osoagoa). MG erabiliz definitutako errore-patroiak analizatzaile morfologiko baten bidez analizatutako testuari aplikatzen zaizkio. Analizatzaileak informazio zuzena analizatzeko prestatuak egoten direnez, egokitzenak, murriztapenen “erlaxazioak” egin behar izaten dira testu ez-gramatikalak analizatu ahal izateko (Johannessen *et al.*, 2002; Badia *et al.*, 2004).

Xerox-eko ikerketa-taldeak sarreratzat adierazpen erregularrak jasotzen dituen eta horiek automata transduktore bihurtzen dituen *Xerox Finite State Tool (XFST)* tresna garatu du (Karttunen *et al.*, 1997; Aït-Mokhtar eta Chanod, 1997). Tresnak eragiketa multzo aberatsa du eta ezaugarri honek adierazpenerako indar handia ematen dio. Eragiketen artean lengoaien eta adierazpen erregularren arteko bildura, kateamendua, itxidura... aurki ditzakegu, baita desanbiguaziorako egokia den murriztapenaren eragiketa, ordezpenean, luzeenarekin parekatzea eta antzekoak ere. Aipatzekoak dira *konposizio* eta *konposizio biguna edo malgua* eragiketak, zeintzuen bidez lengoaiak eta horiek ezagutzen dituzten automatak konposa daitezkeen.

Hashemi *et al.* (2003) lanean erroreen detektziorako XFST erabiltzen duen sistema bat deskribatzen da (ikus deskribapen zehatzagoa A eranskinean). Lan horretan erabilitako teknika transduktoreen arteko kenketan oinarrituta dago. Egitura linguistiko zuzen eta okerrak deskribatzen dituen *gramatika zabal* bat konpilatuz lortutako transduktorearen, eta egitura zuzenak soilik deskribatzen dituen *gramatika estuaren* bidez lortutako transduktorearen arteko kenketa egiten dute. Kenketa egin ondoren lortutako automatek, egitura okerrak detektatzeko gaitasuna dute.

Aipatutako formalismo ezagunez gain, patroietan oinarritutako errore-erregelak idazteko modu gehiago ere definitu dira. Bere tesi-lanean, Naber-ek (2003) *Extensible Markup Language*-n (XML) eta *Python* programazio-lengoaian idatzitako erregelak erabili zituen erroreak deskribatzeko.

Patroi-erregelak definitzeko erregela-lengoaia egituratu propioa erabiltzen dute *Granska* gramatika-zuzentzaile hibridoan (Carlberger *et al.*, 2002; Domeij *et al.*, 1999). Sistema horretan, patroi-erregelak aplikatzeko estatistika erabiltzen dute. Aldez aurretik, kategoria bakoitzeko aplikatzeko probabilitate handieneko erregela kalkulatzeko da. Testuan hitz edo etiketa konkretu bat agertzen denean, estatistika horiei kasu eginez, dagokion erregela edo erregelak soilik aplikatzen dira, ez guztiak.

OLHIA sistemetan patroi-erregelak erroreen diagnosia egiteko eta aldi berean erabiltzailearen eta sistemaren arteko berrelikadura sendotzeko erabiltzen dira. Patroi-erregelen bidez erantzun moduan jasotako esaldi osoaren edo esaldi zatien zuzentasuna, esperotako erantzunarekin pareka daiteke. Erregeletan bai erroreak buruzko informazioa, baita berrelikadurarako

mezuren bat ere, idatz ditzakegu. Teknika honek emaitza egokiak ematen ditu domeinu oso mugatuetan eta ekintza sinpleetarako (Vandeventer, 2003). Aztertutako ia sistema guztietan ikaslearen erantzuna eta sistemaren arteko erantzunaren arteko konparaketa sintaxi-ezaugarriak kontuan izanik egiten den arren, konparaketa karakterez karaktere eta diagnosirako inolako informazio linguistikorik erabili gabe egiten duen sistemarik ere badago (*DRIP*; Güvenir (1992)).

Diagnosia eta berrelikaduraz gain, beste zereginik ere badute egoera finituko mekanismoek ikasleekin lan egiten denean. Patroi-erregelak ikasleen testuei aplikatzen zaizkienean, bereziki ikasleen erroreak detektatzera bideratutako errore-erregelak definitu behar izaten dira, baita ikasleentzako estatistika eta heuristikoko bereziak ere.

Testuingururik Gabeko Gramatikak (TGG)

Lehenago ere aipatu dugu testuingururik gabeko gramatiken egokitasuna esaldien egitura hierarkikoa definitzeko. TGG sinpleak eta *baterakuntzan* oinarritutakoak bereiz ditzakegu. TGG sinpleetan erregelekin osagai atomikoak deskribatzen diren bitartean, gainontzeko TGGetan osagaiei informazioa gehi diezaiekegu ezaugarri-egituren bitartez (gramatikaren sinpletasuna eta trinkotasuna bultzatzen da). Ezaugarri-murritzapenen arteko bateragarritasuna egiaztatzeari eta bateragarriak direnean informazio hori trinkotzeari, baterakuntza deitzen zaio. Baterakuntza oinarritzat duten formalismorik ezagunenak, *Lexical Functional Grammar (LFG)*, *Generalized Phrase Structure Grammar (GPSG)* eta *Head-Driven Phrase Structure Grammar (HPSG)* ditugu.

Oinarrian testuingururik gabeko gramatikak dituzten sintaxi-analizatzailerak asko erabili izan dira errorearen detektziorako, emaitza onargarriak lortuaz. Erroreen detektziora bideratuta ez dauden TGGak ez dira prestatuta egoten sintaxi-egitura ez-zuzenak analizatzeko, beraz, gaizki eratutako esaldi bat jasotzen dutenean, ez dira gai izaten analisi-zuhaitz bat eraikitzeke. Hori egin ahal izateko, moldaketaren bat behar izaten dute. Analisi-zuhaitza ez sortzeko arrazoia gramatikako murritzapenak ez betetzea denean, murritzapenak *erlaxatzearekin* nahikoa da. Beste batzuetan aldatzen dena gramatika-egitura bera denean, *errore-gramatikak* erabiltzen dira. Ondorengo puntuetan azalduko ditugu bi teknika hauek.

Baterakuntza + Erlaxazioa

Baterakuntza-formalismoak egokiak dira *murritzapenen erlaxazio* izeneko teknikaren aplikaziorako. Baterakuntzan, ezagutza sintaktikoa ezaugarri-balio egituren bidez adierazten dela esan dugu. Egitura horietan bete beharreko murritzapenak ekuazioen bidez deskribatzen dira. Esaldi baten analisi-zuhaitza lortzen ez denean, ekuazio horietako batzuk kendu egiten dira (erroreen sorburua izan daitezkeenak, adibidez, numeroa komuntaduran) analisisia lortu ahal izateko. Kendutako ekuazioek emango dute akatsaren iturria. Erroreen detekziorako eta diagnosirako *ezagutza positiboa* erabiltzen da. Patroi-erregetan gertatzen ez zen bezala, errore-mezu zehatzak ematea zaila da analizatzaileak esaldi zuzenen analisisa egiten baitu.

Erroreen detekziorako murritzapenenen erlaxazioa erabiltzen duten sistema ugari (ikus A eranskina) teknika hau beste batzuekin konbinatzen dute. Testu erroredunetan topa daitezkeen erroreak era askotakoak direnez, estrategia ezberdinak jarraitzen dira errore mota bakoitzaren detekziorako. Sistema batzuetan, *erroreei aurre hartzen* dieten erregela esplizitu lokalekin eta heuristikoeekin konbinatzen da erlaxazioa (Bustamante eta León, 1996; Teixeira Martins *et al.*, 1998; Ravin, 1993). Beste zenbaitetan, murritzapenen erlaxazioa errore-gramatika batekin konbinatzen dute (Vosse, 1992).

Erlaxazioaren teknika erabiltzean, sintaxi-analizatzaileak analisi posible ugari eman ohi ditu, esaldi bat analizatzeko erregela erlaxatu gabeak eta erlaxatuak erabiltzen baitira. Erlaxazio-konbinazioak murritzeko, (Gojenola eta Sarasola, 1994; Gojenola, 2000) lanetan ezaugarrien erlaxazio mailakatua proposatzen dute (ikus A eranskina).

Baterakuntza + Errore-gramatikak

Esaldi erroredunetan hizkuntza-egiturak aldatzen direnean, esaldiaren analisisi-zuhaitzak lortzea ez da posible izaten. Analisisi ahalbidetzeko, hizkuntza-egitura okerrak *errore-gramatiketan* deskribatzen dira. Errore-gramatika bat errore tipikoak adierazten dituzten esaldi egiturak deskribatzen dituen gramatika bat da. Hizkuntza-egitura zuzenak deskribatzen dituen gramatikarekin sarrerako testu erroredunaren analisisa egitea ezinezkoa denean, errore-gramatikari pasatzen zaio testua. Errore-gramatikako egituraren batekin analisisi posible bada, errore-mezu bat erakusten da. Beti, lehenengo egitura zuzenak lantzen dituzten erregelak aplikatuko dira, eta gero, egitura ez-zuzenak lantzen dituztenak. Kasu honetan ere, egoera finituko mekanismoetan gertatzen zen moduan, erroreei aurrea hartu behar zaie, hau da,

aldez aurretik ezagutu behar ditugu errore posibleak. Gainera, errore posible guztietarako errore-erregelak idazteko gai izan behar dugu.

Foster eta Vogel-en lanean (2004b) errore-gramatiken alde egiten da eta erlaxazioaren eta analisi partzialen bateraketa erabiltzen duten teknikari zenbait kritika egiten zaie. Hona hemen kritika horietako batzuk:

1. Erlaxazioaren teknika erabiltzean esaldi zuzen batean eta esaldi oker batean hitz kopuru bera egoten dela suposatzen da. Eta, beraz, hitzak esaldian ekiditen edo gehitzen direneko kasuak ez omen dira lantzen. Esaldi baten analisi osoa lortzen ez denean, lortutako analisi partzialak elkartuaz analisi bat lortzearen teknika, berriz, hitzen ekiditea eta sortzearen kasurako egokia omen da, baina ez hainbeste hitzen forma okerrak erabiltzen direnerako. Errore-gramatikak fenomeno horiek guztiak lantzeko gai dira.
2. Esaldi okerra analizatu ahal izateko erlaxazioa erabiltzean, murriztapen ugari erlaxatzeko beharra egon daiteke, eta horien ondorioz, zentzurik gabeko analisiak lor ditzakegu. Hori ez da errore-gramatikekin gertatzen, errore-erregelak modu kontrolatuan gehitzen baitira.
3. Errore-gramatikek gramatikala ez denaren eredu linguistiko bat ematen dute eta azpitik ezagutza dutenez, diagnosian informazio zehatza emateko gai dira.

OLHIA sistemetan testuingururik gabeko gramatikek duten paperari buruz ez dago ezer berezirik esateko. Hizkuntzen irakaskuntzarako zenbait sistemek erroreak detektatzeko eta zuzentzeko modulu bat izaten dute eta horretan erabiltzen dira TGGak.

Orain arte, Chomsky-ren hierarkiarekin zerikusia duen sailkapen bat erabili dugula esan dezakegu: teknikak egoera finituko mekanismoetan eta gramatiketan sailkatu ditugu. Irizpide horren orde irizpidetzat esaldiko egitura sintaktikoak errepresentatzeko modua hartzen badugu, erroreak lantzeko teknikak *osagai-egituretan* eta *mendekotasun-egituretan* sailka ditzakegu. Gure kasuan, interes berezia dugu mendekotasun-egituretan oinarritutako tekniketan, beraz, horien inguruan arituko gara hurrengo lerroetan.

Mendekotasun-egiturak oinarri dituzten teknikak

Mendekotasun-gramatikak sintaxia adierazteko modu bat izan dira tradizionalki. Tradizio horrek gramatika-teoria eta -formalismo multzo handi bat du azpian, eta horiek ideia bera (Nivre, 2005):

«The fundamental notion of *dependency* is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of a sentence.»

Mendekotasun-gramatiken inguruan egiten diren lanek maiz Tesnière-ren lanetik edaten dute (Tesnière, 1959). Mendekotasun-erlazioak, hitz-formen artean definitzen dira gehienetan, baina adabegitzat sintaxi-osagaiak, lemak eta morfemak (Mel'čuk, 1988) dituzten mendekotasun-zuhaitzak existitzen dira. Erlazio motei dagokionez, nahiz eta gehienetan sintaxi-mailakoak diren (subjektu, objektu...), zenbait kasutan semantika-mailakoak ere izaten dira (egile, jasale...). Mendekotasun-gramatiketan oinarria duten teoria ugari daude (ikus Nivre (2005)). Gure lanean, eta atal honetan, sintaxi-mailako erlazioak dituzten mendekotasun-zuhaitzen inguruan arituko gara. Ohi-ko mendekotasun-egituretaz gain, bereziki esteka-gramatikak⁴ (Sleator eta Temperley, 1993) aipatuko ditugu, erroreak detektatzeko saiakerak egin baitira sistema gramatikal formal honekin.

Esteka-gramatiketan hitzen arteko estekak ipintzeko bi parametro nagusirekin egiten da lan: noranzkoarekin eta distantziarekin. Hiztegi batean hitz bakoitzak eskuinera (“+”) edo ezkerrera (“-”) zein gramatika-kategoriako hitzekin egin dezakeen lotura (D-, O- e.a.⁵) definitzen da eta hitzen artean beharrezko baldintzak betetzen badira, hitzen arteko lotura sortzen da. Arku bat ezarri ahal izateko honako baldintzak bete behar dira: i) hitz bakoitzaren baldintza lokalak bete behar dira, ii) estekak ezin dira gurutzatu esaldia osatzean eta iii) hitzek grafoa osatu behar dute.

Esteka-gramatiken bidez esaldi erroredunak analizatu ahal izateko, detektatu nahi den errore motarentzako garrantzitsuak diren hitzen hiztegi-definizioa aldatu egin behar da (Brehony eta Ryan, 1994).

Mendekotasun-egiturak hitzen ordena librea duten eta beraz, sintaxi-egitura ez-jarraituak dituzten hizkuntzetarako egokiak dira (Mel'čuk, 1988). Erroreak detektatzeko mendekotasun-egiturak erabiltzen dituzten txekierarako (Holan *et al.*, 1997), errusierarako (Mitjushin, 1996) eta frantseserako (Courtin *et al.*, 1991) sistemak topatu ditugu.

Txekierarako egindako lanean zuhaitzetan gerta daitezkeen trinkotasun ezak metaerregeletan deskribatzen dira (Holan *et al.*, 1997). Esaldia analizatzerakoan metaerregelaren bat erabiltzen bada, eraikitako zuhaitzean ikur negatibo bat txertatzen da. Erregeletan murriztapenen erlaxazioa egiteko

⁴*Link grammar* ingelesez

⁵Hitz baten definizioan D- topatuz gero, ezkerrera D konektore bat (ustez determinatzailea) behar duela adierazten da. Adibidez, *cat* hitzak ezkerrera *the* edo *a...* behar du hauek sintagma batean lot daitezen.

aukera dago. Analisia faseka egiten da eta *ondo* *ondoko harreman sintaktikoak deskribatzen dituen ezaugarria* (projectivity⁶) eta zuhaitzeko ikur negatiboak kontuan hartzen dira, zuhaitzak errorerik ote duen edo ez erabakitzen duen modulura bidaltzeko.

Sintaktikoki ez-konektatuta egotearen ezaugarria erabiltzen dute errusieraz erroreak detektatzeko (Mitjushin, 1996). Konektatutako segmentuen kopurua 1 denean, egitura sintaktiko osoa lor daiteke analisisian. Bestela, hasierako esaldian aldaketa batzuk egin behar dira esaldi erroredunaren aldaeraren baten analisia lortu arte. Horrela, esaldi erroredunaren ordain zuzena —zuzenketa— lortzen dutela diote. Diagnosia egitea ez da beharrezkoa.

Frantseserako egindako lanean (Courtin *et al.*, 1991), mendekotasun-zuhaitzei baldintza-ekintza moduko erregelak aplikatzen dizkiete komunztadura-erroreak detektatu ahal izateko.

Bestelakoak

Atal honetan inolako multzotan sailkatu ezin ditugun teknikak azalduko ditugu motz motzean.

1. DeSmedt-ek (1995) *predikatuak zuzendurikoa* izeneko teknika proposatzen du. Teknika honetan lehenik aditza identifikatzen da, eta gero, esaldiaren gainontzeko elementuak aditzaren osagarri moduan identifikatzen saiatzen da. Adibidez, aditz trantsitibo bat edukita, objektua izan daitekeen sintagma baten bila hasiko da analizatzailea. Esaldiko elementu guztiei, aditzarekiko betetzen duten funtzioa esleitzen zaie. Elementuren bat funtziorik gabe geratzen bada, esaldia ez dela guztiz zuzena suposatzen da.
2. Vandeventerrek (2003) *zatieen berrinterpretazioaren* inguruan egin du lan. Ideia nagusia honakoa da: “sintaxi-analizatzaile bat analisi osoa lortzeko gai ez denean, analisi partzialak edo zatiak lortzen ditu. Zati hauek bilduz gero, esaldi osoa lortuko dugu. Beraz, zatiak aztertuiko ditugu, esaldi osoaren analisia ez lortzearen zergatia ezagutzeko.” Analizatzailearen estaldura nahikoa bada, analisi osoa ez lortzearen arrazoa errore bat izan daiteke.

⁶A dependency graph satisfies the constraint of projectivity with respect to a particular linear order of the nodes if, for every arc $h \rightarrow d$ and node w , w occurs between h and d in the linear order only if w is dominated by h (Nivre, 2005).

III.2.1.3 Ezagutza semantikoa

Erroreen sailkapenak aztertuz gero, *errore semantikoaren* kontzeptua ez dela gehiegi aipatzen ikusiko dugu. Ortografia- eta sintaxi-mailan zuzenak diren baina kokatuta dauden testuinguruan zentzurik ez duten hitz edo esaldiek *errore semantiko* bat sortzen dutela esan ohi da. Hori, ordea, II. kapituluaren aipatutako *hitz zuzenen erabilera nahasketen* definizioaren berdina da. Bi kontzeptu horien artean diferentzia bat egitekotan, errorea egitearen arrazoi posiblerik jo dezakegu. Hitz testuinguru okerrean kokatu bada esanahi zehatza ezagutzen ez delako, *errore semantikoa* dugula aurrean esan dezakegu (adib. **bilatu** vs **topatu**). Tipografia edo teklatze-arazoak direla eta antzeko hitza sortu bada, edo hitzak homofonoak izateagatik nahastu badi-ra, baina nahastutako hitzen esanahia ezaguna eta zeharo ezberdina bada, *hitzen nahasketa* sinple baten aurrean gaudela esango dugu, inolako kutsu semantikorik ez duena (adib. **hura** vs **ura**).

Errore semantiko bat egiten denean, esanahi zehatza ezagutzen ez dela esan dugu. Berrirori ere, Jamesen (1998) lanera jo behar dugu zehaztasun gehiagoren bila. Honek, zentzu-erlazioen nahasketak, i) esanahi orokorragoa duten hitzak zehatzagoen ordeztaraztea (hiperonimoa, hiponimoaren ordeztaraztea), ii) termino zehatzegia erabiltzea (hiponimoa, hiperonimoaren ordeztaraztea), iii) bi hiponimo anaien artean desegokiena erabiltzea edo iv) hurbileko sinonimoen artean okerra erabiltzea, errore semantiko mota desberdinak kontsideratzen ditu. Horietaz gain, kolokazio-erroreak ere aipatzen ditu. Hau da, esanahiagatik elkarren ondoan ager ez daitezkeen hitzak. Adibidez, **makila okertu** dela esatea egokia da, baina **eguraldia** ezin da **okertu**. Antzekoak dira lekuz alda ezin diren hitz bikoteak. Adibidez, **karakola eta barea** esango dugu eta ez ***barea eta karakola**.

Malapropismoa, eduki-hitz⁷ bat soinu aldetik antzekoa den, baina semantikoki testuinguruarekin guztiz bateraezina den beste batekin aldatzea da. Kontzeptu hau landu da Bolshakov-en lanean (2005) eta hauen detektatzeko eta zuzentzeko estatistika-teknika Interneten aplikatu da.

Idazleak gaizki erabilitako hitzaren esanahia ezagutzen ote duen ala ez jakitea oso zaila denez, hitz zuzenen erabilera nahasketak eta errore semantikoak detektatzeko eta zuzentzeko teknika berdina erabiltzen dira. Teknika horiek ikasketa automatikoari dagokion atalean (III.2.2.2 atalean) aipatuko ditugu.

⁷Errusieraz, izenak, adjektiboak eta adberbioak *eduki-hitzak* dira. Preposizio edo aditz laguntzaileak *hitz funtzionalak* dira; eta izenordainak eta antzekoak, *stop hitzak*.

III.2.2 Corpusetan oinarritutako teknikak (enpirikoak)

Azken 10 edo 15 urteotan aipatzekoa izan da hizkuntzaren tratamendu automatikoan teknika enpirikoen garapena. Teknika hauek datu kopuru izugarrien erabileran dute oinarria, datuak arakatzeko prozedura estatistikoak erabiltzen direlarik.

Hurbilpen enpirikoen ezaugarri nagusia corpusen erabilera da. Eskuar-tean darabilgun eremuan, corpora testuzkoa izango da eta kasu gehienetan nolabait *etiketatu*a egongo da. Teknika enpirikoetan ataza ezberdinak lantzeko beharrezko hizkuntza-ezagutza corpusean agertzen diren elementuetatik (eta beren maiztasunetatik) erauzten da. Horrela, analizatzaile sintaktiko bat garatzeko orduan, gramatika eskuz idatzi beharrean, beharrezko informazioa sintaxi-mailan etiketatutako testuetatik jasoko litzateke. Gertaera linguistikoen deskribapen zabala izateko, corpusak tamaina handia behar izaten du. Teknika enpirikoek erabiltzen duten etiketatze mota ohikoenetako bat, morfologia-mailako etiketatzea da. Hala ere, erabiltzaileak duen helburuaren arabera, corpusak sintaxi-mailan, hitzen adiera-mailan e.a. etiketatuak egon daitezke. Internetek darabilen informazio kopurua izugarria da eta LNPko zenbait alorretan Internet corpus moduan erabiltzeko joera gero eta handiagoa da. Nahiz eta zenbait corpus etiketatu existitzen diren (hala nola, Brown Corpora⁸, British National Corpora⁹ e.a.), horiek ez dituzte beti ikerlarien beharrak asetzen. Hori dela eta, gaur egun indar izugarria egiten da corpusak helburu zehatzekin eta hizkuntza ezberdinetarako eskuz etiketatzen.

Erroreen detekzioari dagokionez, corpusean oinarria duten teknikak ez dira gehiegi erabili. Oso zaila da teknika enpirikoak erabili ahal izateko ego-kia den testu erroredun kopuru handia biltzea. Gainera, testuan topa daitezkeen errore posible guztien etiketatzea ataza konplexua eta garestia da. Hala ere, zenbait abantaila badituzte teknika enpirikoek. Ez dira alde zuzenak ezagutu behar errore posibleak. Testuetan dute isla. Desabantaila moduan esan dezakegu, erroreak detektatu ondoren sistema hauetan normalean ez dela errore-mezurik ematen, horretarako erregela bereziak erabiltzen ez ba-dira, behintzat.

Zenbaitetan Interneteko corpus izugarria erabili da terminoren baten zuzentasuna neurtzeko (Moré *et al.*, 2004). Ideia sinplea da: Interneten oso maiz azaltzen diren terminoak eta testu-zatiak zuzenak dira. Gutxi azaltzen direnekin, ordea, duda izan dezakegu. Hurbilpen honek Interneten argitaratzen diren testuen zuzentasuna suposatzen du. Hipotesia, ordea, ez da

⁸<http://nora.hd.uib.no/whatis.html>

⁹<http://natcorp.ox.ac.uk>

erraz neurgarria. OLHIaren alorrean, berriz, azken urteotan iraultza izugarria gertatu da Interneten. Sarea komunikaziorako tresna boteretsua da eta bigarren hizkuntzaren irakaskuntzaren eremuan, irakasle eta ikasleen artean informazioa, ariketak, irakurgaiak e.a. konpartitzeko modu erraza eta azkarra da. Baliabideak kudeatzeko Moodle¹⁰ sistema da Internet erabiltzen duen sistema ezagunetako bat.

Teknika enpirikoak orokorrean *teknika estatistikoetan* eta *ikasketa automatikoan* sailkatzen dira. Batzuetan, ordea, teknika horien arteko banaketa ez da argia. Teknikak teilakatu egiten dira, bi diziplinetan datuen analisia egiten baita, eta maiz ikasketa automatikoa erabiltzen duten sistemak estatistikarekin konbinatuta erabiltzen baitira. Hurrengo lerroetan teknika hauen inguruko azalpentxo bat emango dugu:

III.2.2.1 Metodo estatistikoak

Esan dugun bezala, testu kopuru handia eskura duten hizkuntzetan estatistika erabili ahal da informazio linguistiko aberatsa testuetatik erauzteko. Zenbait gramatika-zuzentzailek metodo estatistikoak baliatzen dituzte errorearen detekzioarako atazak garatzeko. Normalean, maila morfologikoan etiketatutako corpus bat erabiltzen da gramatika-kategorien etiketen zerrenda bat eraikitzeke. Zenbait etiketa-sekuentzia (n-grama) oso arruntak izango dira (adibidez *izena determinatzailea*) eta beste zenbait oso gutxitan gertatuko dira (adibidez, *izena determinatzailea determinatzailea*). Corpusetan maiz gertatzen diren sekuentziak *zuzenak* kontsideratuko dira, eta arruntak ez direnak, *okerrak*. Metodo estatistikoekin honakoak hartu behar dira kontuan: i) Etiketa zehatzen agerpen ezaren arrazoia *datu urritasuna* izan daiteke, eta beraz, egitura ez-ohikoak errore kontsidera ditzakegu, *agerpen urrikoak* besterik ez direnean. ii) n-grametan hitzaren kategoria soilik kontuan hartuz gero, gramatika-kategoria berdineko baina oker erabilitako hitzetan gertatutako erroreak ez lirateke detektatuko. Metodo estatistikoekin alarma faltsu kopuru handia sortzen da.

Estatistika hutsezko zuzentzaile bat sortzea zaila da, eta normalean, hauek erregeletan oinarritutako teknikekin hedatzen dira. Erregeletan erroreak deskribatzen dira eta zuzenketarako proposamenak egiten dira.

Informazio zuzena oinarrian hartu beharrean, *errore-probabilitatea* neurten saiatu izan dira zenbait egile (Atwell, 1987): erroreetan zein gramatika-etiketa konbinazio agertzen den aztertu nahi izan dute. Hurbilpen hau baztertua izan da erroreekin etiketatutako corpus ezagatik.

¹⁰<http://moodle.org>

Estatistika hutsezko sistemei hizkuntza-ezagutza gehitu ondoren, erroreen detekzioaren doitasuna izugarri hobetzen dela, eta alarma faltsuen kopurua asko jaitsi dela, egiaztatu dute zenbait lanek (Bigert eta Knutsson, 2002; Sjöbergh, 2005).

Azken aldian, hizkuntzaren tratamendu automatikotik aldentzen diren teknikak erabiltzen hasiak dira erroreak detektatzeko. Adibidez, Kumar eta Nair-en lanean (2007) immunologia-sistema artifizialeko (*Artificial Immune System (AIS)* ingelesez) oinarriak erabiltzen dituzte gramatika-erroreak detektatzeko. Erroreak testuetako patogenoak bailiran tratatzen dituzte, eta antigorputz detektatzaileak eraikitzen dituzte horiek detektatzeko. Hizkuntza-informazioa kategoriatan gramatikalen bigrametan, trigrametan eta tetragrametan biltzen denez, estatistika oinarria duen lanen artean kokatu dugu honakoa, nahiz eta algoritmoak AIS eremukoak diren. Teknika honekin detekta daitezkeen zortzi errore mota topatu dituzte soilik, eta gramatika-liburu batetik ateratako esaldiak erabili dituzte probarako, hots, laborategiko esaldi motzak.

III.2.2.2 Ikasketa automatikoa

Ikasketa automatikoaren helburu nagusia, ordenagailuek *ikas* dezaten teknikak garatzea da. Zehazkiago, adibide moduan emandako informazio ez-egituratutik abiatuta, portaera zehatz batzuk garatu nahi dira. Hizkuntzaren tratamendu automatikoari dagokionez, informazio ez-egituratu hori corpusetan egoten da. Testu-bildumetatik datuak jaso eta testuan egon daitezkeen egiturak inferitzen dira datu-meatzaritza izenekoak eginez. Teknika hau alor desberdinetan erabil daiteke, hala nola, hitzen adiera-desanbiguazioan, terminoen erauzketan edo sailkapen-arazo moduan defini daitezkeen edozein alorretan. Ikasketa automatikoak estatistikan aipatu ditugun ezauzgarrien antzekoak ditu, datu urritasunaren eta alarma faltsuen arazoak, corpus etiketatuen beharra... Ikasketa automatikoa erroreen detekzioaren eta zuzenketaren, eta hitz zuzenen erabilera nahasketaren atzetan erabiltzen da (ikus A eranskina).

Corpus erroredunak biltzeko, batzuek eskuzko etiketatzea erabili duten bitartean (Izumi *et al.*, 2003), beste batzuek lan hau ekiditeko erroreak automatikoki sortzen dituzte, eta hori egin ahala, testuak erroreekin etiketatzen dituzte (Sjöbergh eta Knutsson, 2005). Hitz zuzenen erabilera nahasketaren atazarako ez da testu erroredunik behar, ongi samar idatzitako edozein testu erabil daiteke; beraz, ez da arazo handirik egoten corpusak lortzeko. Hori izan da ikasketa automatikoarekin azken ataza honetan lortutako arrakastaren arrazoi nagusia.

Erroreak detektatzeko lanetan erabilitako algoritmoen dagokionez, *trans-formazioan oinarritutako ikasketa*¹¹ (Sjöbergh eta Knutsson, 2005), *memoryan oinarritutako ikasketa*¹² (Örvar Kárasen, 2005) eta *entropia handieneko eredu*¹³ (Izumi *et al.*, 2003) erabiltzen dituzten lanak topatu ditugu. Ikasketarako, errore mota bakoitzeko ikasketa automatikoko modulu bat entrena daiteke, edo errore mota desberdinetarako modulu bakarra. Lehen hurbilpenak, zuzenketak proposatu ahal izatearen abantaila du.

Hitz zuzenen erabilera nahasketaren atazarako ikasketa automatikoa oso teknika egokia da, arazoa desanbiguazio-ataza moduan ikustea erraza baita. Era formalean, honela definitzen da arazoa (Carlson *et al.*, 2001) lanean: sarrerako esaldi bat eta *helburu hitz* bat emanda, jakin nahi da hitz hori noiz den zuzena edo noiz aldatu behar dugun. Hitzen arteko anbiguotasuna *nahasketa multzoekin* modelatzen da (adibidez, ingelesezko, (**among, between**), (**cite, sight, site**), (**I, me**)...). $C = w_1, \dots, w_n$ nahasketa multzo batean w_i hitz bakoitza anbigua da multzoko beste hitzekiko. Testuingurua kontuan izanik zein hitz den egokiena aukeratzea da egin beharrekoa. Arazo honen ingurukoak dira (Golding, 1995; Golding eta Roth, 1996, 1999; Carlson *et al.*, 2001) lanak. Horietan, nahiz eta emaitza onak lortzen diren (% 92-95eko zuzentasuna), lan eremua nahasketa multzo zehatz batzuetara mugatuta dago.

Datu kopuru eskergarekin aipatutako teknikak probatuz gero, emaitzek bat egiten dutela ikusi da. Hori dela eta, LNPko ikerlarion etorkizuneko lana etiketatutako corpusak sortzera bidera dadin proposatzen dute Banko eta Brill-ek (2001), ikasketa teknika berriak probatu eta konparatzera baino.

Ikasketa automatikoak hizkuntzen irakaskuntzan izan dezakeen erabilerrari dagokionez, Zock-ek (1996) zera dio:

«It is also worth mentioning that no cross fertilization has taken place between the ICALL community and people working in the Machine Learning paradigm». »

Irakaskuntzarako sistema gutxi topatu ditugu ikasketa automatikoa erabiltzen dutenak. Adibidez, Interneten oinarritutako Web-EasyMath algebrako tutoreak ikasketa automatikoko teknikak erabiltzen ditu ikasle-eredua sortzeko eta mantentzeko (Tsiriga eta Virvou, 2003). Ikasleen ezaugarriak inferitzeko ere erabili da ikasketa automatikoa. Adibidez, ikasle batek erantzun zuzena emateko zein probabilitate duen eta erantzuna sortzeak zenbat

¹¹Transformation-Based Learning.

¹²Memory-based Learning.

¹³Maximum Entropy (ME) model.

denbora hartuko dion kalkulatzeko (Beck eta Woolf, 2000). Topatutako sistemak hizkuntzak irakasteko ez diren arren, aipatutakoak OLHIA sistemetarako baliagarriak direla uste dugu.

III.3 Erroreak zuzentzeko teknikak

III.3.1 Ortografia

Ortografia-mailako erroreak detektatzeari buruz III.2.1.1 atalean aritu garenez, zuzenketari eskainiko diogu honako atala. Cucerzan eta Brill-en arabera (2004) helburu orokorreko zuzenketa ortografikoaren xedea hitz ezezagune-tan (hizkuntzaren lexikoan agertzen ez diren hitzetan) errore tipografikoak (txertatzeak, ezabatzeak, aldaketak eta transposizioak) konpontzea da. Ortografia-zuzentzaile gehienek, errore ortografikoa detektatu ondoren, lexiko-koak diren ordezeko hitz zuzenen zerrenda bat ematen dute zuzenketa posible moduan. Zerrenda osatzeko, lexikoko hitzen maiztasunak, teklatzerakoan egiten diren ohiko akatsak eta akats fonetiko eta kognitiboak hartzen dira kontuan.

Ordezko zuzenen zerrenda gehienak sailkatuak egoten dira ordezeko ego-kienetik hasita beheranzko ordenan. Zerrendak lortzeko erabiltzen diren tek-nika batzuk, honakoak dira:

Peterson-ek (1980) zuzenketa ortografikorako zenbait teknika aipatzen ditu, hala nola, testu bateko tokenak maiztasunaren arabera sailkatzea, bi-grama eta trigrama arraroak *erreore posible* moduan markatzea, hiztegi-tan begiratzea, grafematik fonemara eta berriro grafemara transkripzioa egitea eta hizkien (aurrizkien, artizkien, atzizkien) analisia. Hauetako teknika anitz oraindik erabiltzen dira zuzentzaile ortografikoetan.

Vosse-k (1992) bere zuzentzaile ortografikoan trigramak eta trifonemak¹⁴ erabiltzen ditu bozketa mekanismo batekin. Gainera, zuzenketa posibleen ar-tean batzuk iragazteko, sintaxia erabiltzen du. Hau da, esaldian hitz okerra bere ordain zuzenekin ordezkatu ondoren, sortutako esaldien analisia egi-ten du. Analisi zuzena sortzen duten proposamenak soilik erakusten zaizkio erabiltzaileari.

Lexiko- eta sintaxi-mailan eta frantsesez idatzitako testuetan erroreak detektatu eta zuzendu dituzte (Courtin *et al.*, 1991) lanean. Hiru teknika aipatzen dituzte erroreak zuzentzeko:

¹⁴Trifonema hiru fonemen sekuentzia errepresentatzen duen unitatea da. Testuinguruaren mendeko fonema.

- *Skeleton key* izeneko teknikan, ideia nagusia honakoa da: karaktere-kate bat hiztegiakoa den hitz batean bihur badaiteke oinarritzko edizio-eragiketak erabilia (txertaketa, ezabaketa, aldaketa, transposizioa), orduan, hitz hori karaktere-kate horrentzako zuzenketa posible bat da.
- Bigarren metodoan, gaizki osatuta dagoen hitzari dagozkion *baliokide fonetikoak* lortzen dira. Transdukzio-prozesu baten bidez, karaktere-kate bat oinarri hartuta, horri lotutako transkripzio fonetiko guztiak ematen dira. Adibidez, *chevale* hitz okerrerako “ch.é.v.a.l.”, “k.é.v.a.l.” ...transkripzioak lortuko lirатеke eta horietatik, idatzizko formak (*ke-vale*, *cheval*, *keval*, *chéval*, *chèval*). Azkenean, grafikoki zuzenak diren karaktere-kateak lortzen dira. Kasu honetan, *cheval*, zaldi, hitz zuzena lortzen dute egileek.
- Azkenik, analisi eta sorkuntza morfologikoa erabiltzen dute, zehazkiago egoera finituko transduktore bat, leman edo morfeman gertatzen diren errorearen ordeko zuzenak emateko.

Agirre *et al.* egileek (Agirre *et al.*, 1998a, b), zuzenketarako proposamenen zerrendako lehen posizioan aukerarik egokiena eskaintzeko lan bat egin zuten. Lana *IsPELL* erabilia, ingeleserako eta hitz zuzenetan gertatzen diren erroreetarako egin zen. Helburua lortzeko, ezagutza sintagmatikoa, paradigmatikoa eta estatistikoa konbinatu zituzten. Ezagutza sintagmatikoa Murritzapen Gramatikaren baliatuz adierazten da, ordain zuzenen artean gramatika-kategoria okerrekoak kendu egiten baitira. Horrela, kategoria berdinekoen artean ezin da desberdintasunik egin. Izena kategoria duten ordainen artean testuinguruarekiko semantikoki hurbilen dagoen hitza aukeratzeko erabiltzen den osagai paradigmatikoa, *dentsitate kontzeptuala* da (Agirre eta Rigau, 1996). Honek WordNet-eko izenen arteko *afinitate-distantzia* erabiltzen du. Azkenik, hitzen agerpenen-maiztasuna eta testuinguruaren informazioa erabiltzen da (ezagutza estatistikoa). Zenbait heuristikore erabili dira proposamenak ezabatzeko. Ez da erroreari buruzko inolako informaziorik erabili nahi izan. Teknika hauen guztien konbinazioak egin dira modu ezberdinetan, teknikei pisu ezberdinak emanez. Dokumentu bereko neurri estatistikoa eta MG oso erabilgarriak izan dira. Dentsitate kontzeptualak, ordea, ez du ekarpen handirik egin garai horretako tresnen garapen urria dela eta. % 90 inguruko doitasuna lortu da eta proposamen bakarra kasuen % 100ean.

Cucerzan eta Brillen arabera (2004), Interneten bilaketa bat egiten dugunean, erabiltzailearen *logean* hizkuntzari buruzko informazio inplizitu eta esplizitu kopuru handia gertatzen da. Egileek, hizkuntza-informazio hau eta

estatistika erabili dituzte bilaketa-galderak zuzentzeko. Milioika pertsonen egindako kontsulten *log*-ak aztertuta (bilatzaile batek jasotako bilaketa-galderen histogramak), hitz baten baliagarritasuna neurtzen dute. Horretarako, sarrerako galdera tokenizatu egiten dute. Ondoren, token bakoitzari oso era zehatz eta formalean definitu duten edizio-funtzio bat eta atalase funtzio batzuk aplikatzen dizkiote, aldaera posible batzuk lortuz. Aldaera horiekin osatutako hitz-zerrendak, bilaketa-galderen *logetan* bilatzen dira, bertako maiztasuna kontuan hartuta. Bilaketa-galdera maiztasunaren arabera ordenatutako aldaeren zerrendan goiko postuetan agertzen bada, ongi idatzita dagoela suposatzen dute. Bestela, zerrendaren behealdean badago, gaizki idatzia egongo da, baina dagokion zuzenketa zerrendaren hasieran egongo da. Esperimentu honetan, ordain zuzenen % 85,5eko zuzentasuna lortzen dute, % 88,4ko doitasuna eta % 85,4ko estaldura.

III.3.2 Gramatika

Azterketa bibliografikoan gainbegiratutako lan gehienetan gramatika-mailako zuzenketa errorea detektatu ondoren egiten dela ikusi dugu. Hala ere, detekziorik gabeko zuzenketa egiten duten lanak ere topatu ditugu. Azkeneko moduko zuzenketa egiten da adibidez, Mitjushin-en (1996) lanean. Esaldi erroredunari ezin zaionean egitura sintaktikorik egokitu, esaldia eta analisi-zuhaitza aldatu egiten dira, analisi bat lortzea posible den arte. Lortutako esaldi-aldaerak, abiapuntutzat hartutako esaldiaren zuzenketa posibletzat jotzen dira. Hori egin daiteke, gertatutako erroreak identifikatu gabe.

Hurbilpen hori baino ohikoagoa da, behin errorea detektatu dela, erroreen parte diren esaldiko osagaiak zuzentzea. Adibidez, komunztadura-erroreetan, esaldi osoa zuzendu beharrean, komunztatzen ez duten osagaien zuzenketara jotzen da. Osagai horien ordain zuzenak lortu ondoren, bi osagai berrien arteko *egokitasuna* aztertu beharko litzateke, behin eta berriro, errorea desagertzen den arte. Ordain zuzenak lortzeko erabili daitezkeen teknikak, hitz-mailako zuzenketak lortzeko erabiltzen diren teknika berberak dira. Aipatutakoez gain, honakoak ere topatu ditugu:

Bustamante eta Leónen lanean (1996) analisisia-transferentzia-sintesia ziklo bat erabiltzen da zuzenketa egiteko. Informazio okerra duten egitura linguistikoak *zuzen* moduan definitu den gaztelerara transferitzen dira itzulpen automatikoko tekniken antza duen teknika bat erabiliz.

Bender *et al.* (2004) lanean linguistikoki okerrak diren karaktere-ka-teak ongi eratutako adierazpen semantikoekin parekatzen dira. Adierazpen semantikoetatik abiatuta, erroreak buruzko diagnosia egin daiteke, baita erroreei dagozkien ordezkio zuzenak sortu ere.

III.4 Gramatika-zuzentzaileak

Gramatika-zuzentzaile baten garapenarekin hasi aurretik, bi gauza hartu behar izaten dira kontuan:

1. **Landu beharreko hizkuntzaren ezaugarri sintaktikoak zeintzuk diren.** Ezaugarri sintaktikoek zuzentzaileak detektatuko duen fenomeno multzoa zehaztuko dute. Adibidez, euskara moduko hizkuntza eranskari batean, non aditz laguntzaileen osaketa nahiko konplexua den, aditza eta osagai gramatikalen (subjektu, objektu, zehar-objektu) arteko komunztadura oso garrantzitsua izango da; eta suedierarako, adibidez, ez (Arppe, 2000).
2. **Zein formalismo linguistiko daukagun eskuragarri hizkuntzaren analisia eta geroago erroreen detekzioa eta zuzenketa egin ahal izateko.** Zerotik hasi nahi ez badugu behintzat, aurretik garatutako tresnak berrerabiltzeak garrantzi handia izango du. Hori, are eta garrantzitsuagoa izango da, enpresek merkaturatzen dituzten gramatika-zuzentzaileetan.

Bibliografia azterketan topatutako gramatika-zuzentzaileak zerrendatu ditugu eta sistema bakoitzaren deskribapen labur bat egin dugu A eranski-nean. Haiei gainbegiratu bat eman diegu III.2 eta III.3 tauletan, gramatika-zuzentzaileetako bakoitza bi parametroren arabera sailkatuz: i) erroreak detektatzeko erabiltzen duen teknika eta ii) helburu duen hizkuntza. Kapitulu amaierako laburpen-tauletan (ikus III.5, III.6 eta III.7 taulak) gramatika-zuzentzaile berei buruzko datu gehiago erakutsi ditugu. Irakurritako artikuluan informazioa agertzen bada, honakoak eman ditugu: i) erroreak detektatzeko erabilitako teknika, ii) landutako errore mota batzuk, iii) probak laborategiko esaldiekin egin ote dituzten¹⁵, eta azkenik, iv) lortutako estaldura eta doitasuna¹⁶. Azken tauletako datuak tentu handiz jaso behar dira, egileetako bakoitzak errore-sailkapen desberdina erabili baitu (fenomeno bera modu desberdinetan deskriba dezake) eta emaitzak kalkulatzeko modua ere, desberdina izan baitaiteke. Hala ere, zenbait ondorio atera daitezkeela uste dugu, eta horiek III.6 atalean idatzi ditugu.

¹⁵Testu errealak erabili dituzteneko kasuetan “-” ikurra erabili dugu.

¹⁶Atal honetan doitasuna eta estaldura agertzen ez direnean, honako laburduraren bat idatzi dugu: EZ-Artikuluan ez da inolako ebaluazio-emaitzarik eman. LG-Artikuluan ebaluazioari buruzko datu xehatuak eman dira, baina ez doitasuna eta estaldura. Z-*Accuracy* edo zuzentasuna adierazi da.

Hizkuntza	Gramatika-zuzentzaileak eta erabilitako teknikak							
	Sinb.						Enp.	
	Egoera finituak			Gramatikak			Estat.	IA
	MG	XFST	Best.	TGG		Mend.		
				+ Erlax.	+ Errore gram.			
Katalana (Badia <i>et al.</i> , 2004)	✓	-	-	-	-	-	-	-
Suediera <i>Grammatifx</i> (Arppe, 2000)(Birn, 2000)	✓	-	-	-	-	-	-	-
(Hashemi <i>et al.</i> , 2003)	-	✓	-	-	-	-	-	-
<i>Granska</i> (Carlberger <i>et al.</i> , 2002)	-	-	✓	-	-	-	-	-
(Domeij <i>et al.</i> , 1999)	-	-	-	✓	-	-	-	-
<i>Scarrie</i> (Paggio, 2000)	-	-	-	-	-	-	✓	-
(Bigert eta Knutsson, 2002)	-	-	-	-	-	-	✓	-
(Sjöbergh, 2005)	-	-	-	-	-	-	-	✓
(Sjöbergh eta Knutsson, 2005)	-	-	-	-	-	-	-	✓
Islandiera (Örvar Kárasón, 2005)	-	-	-	-	-	-	-	✓
Norvegiera (Johannessen <i>et al.</i> , 2002)	✓	-	-	-	-	-	-	-
Espainiera <i>GramCheck</i> (Bustamante eta León, 1996)	-	-	-	✓	-	-	-	-
Italiera <i>JDII</i> (Bolioli <i>et al.</i> , 1992)	-	-	-	✓	-	-	-	-
Brasilgo portugesa <i>ReGra</i> (Teixeira Martins <i>et al.</i> , 1998)	-	-	-	✓	-	-	-	-
Arabiera <i>Arabic GramCheck</i> (Shaaan, 2005)	-	-	-	✓	-	-	-	-
Euskara (Gojenola, 2000) (Gojenola eta Sarasola, 1994)	-	-	-	✓	-	-	-	-
Nederlandera (Vosse, 1992)	-	-	-	✓	-	-	-	-
Txekiera (Holan <i>et al.</i> , 1997)	-	-	-	-	-	✓	-	-
Errusiera (Mitjushin, 1996)	-	-	-	-	-	✓	-	-
Frantsesa (Courtin <i>et al.</i> , 1991)	-	-	-	-	-	✓	-	-

III.2 Taula: Gramatika-zuzentzaileek erabiltzen dituzten teknikak (1/2).

Hizkuntza	Gramatika-zuzentzaileak eta erabilitako teknikak						
	Sinb.					Enp.	
	Egoera finituak			Gramatikak		Estat.	IA
	MG	XFST	Best.	TGG + Erlax.	Mend. + Errore gram.		
Ingelesa							
(Naber, 2003)	-	-	✓	-	-	-	-
(Foster eta Vogel, 2004b)	-	-	-	-	✓	-	-
<i>AbiWord</i>	-	-	-	-	-	✓	-
(Brehony eta Ryan, 1994)	-	-	-	-	-	✓	-
(Atwell, 1987)	-	-	-	-	-	✓	-
(Chodorow eta Leacock, 2000)	-	-	-	-	-	✓	-
(Cucerzan eta Brill, 2004)	-	-	-	-	-	✓	-
(Kumar eta Nair, 2007)	-	-	-	-	-	✓	-
(Izumi <i>et al.</i> , 2003)	-	-	-	-	-	-	✓
Ingeleseko nahasketa multzoak							
(Golding, 1995)	-	-	-	-	-	✓	-
(Golding eta Roth, 1996)	-	-	-	-	-	-	✓
(Mangu eta Brill, 1997)	-	-	-	-	-	-	✓
(Golding eta Roth, 1999)	-	-	-	-	-	-	✓
(Carlson <i>et al.</i> , 2001)	-	-	-	-	-	-	✓

III.3 Taula: Gramatika-zuzentzaileek erabiltzen dituzten teknikak (2/2).

III.5 OLHIA sistemak

Zenbait kasutan gramatika-zuzentzaileak ordenagailuz lagundutako hizkuntzen irakaskuntzarako sistema adimendunen (OLHIA) parte dira. Kasu horietan sistemako modulu batek ikasleek egindako erroreak detektatzen ditu. Beste sistema batzuetan, ordea, ez da erreoreen inolako detekziorik egiten, edukien kudeaketari, irakaste-estrategiei eta, batez ere ikaslearen ezagutzari eta berarekin sistemek eduki dezaketen berrelikadurari, ematen baitzaio garrantzia. Hori dela eta, OLHIA sistemak deskribatzeko laburpen-taula bakarra sortu dugu (ikus III.4 taula): xede-hizkuntzaren eta erabilitako teknikaren arabera antolatutakoa. Taulako sistemen deskribapena A eranskinean egin dugu.

Hizkuntza	OLHIA sistemak eta erabilitako teknikak			
	Eg. finituak	Gramatikak		Best.
		TGG		
		+Erlaxazioa	+Errore gramatika	
Frantsesa <i>CLEF</i> <i>VINCI</i> (Levison <i>et al.</i> , 2000) (Vandeventer, 2003)	✓ ✓ -	- - ✓	- - -	- - Fonologia Zati-berrinterp.
Turkiera <i>DRIP</i> (Güvenir, 1992)	✓	-	-	-
Suediera <i>CrossCheck</i> (Kann, 2002) <i>Grim</i> (Bigert <i>et al.</i> , 2005)	✓ ✓	- -	- -	- Estatistika IA
Espainiera <i>MILT</i> (Weinberg <i>et al.</i> , 1995) (Holland <i>et al.</i> , 1993) <i>ALICE</i> (Evans eta Gates, 1991)	- - -	✓ - -	- ✓	- -
Alemaniera <i>BRIDGE</i> (Weinberg <i>et al.</i> , 1995) (Holland <i>et al.</i> , 1993) <i>MS</i> (Menzel eta Schröder, 1999) <i>Herr Komissar</i> (DeSmedt, 1995)	- - -	✓ ✓ -	- - -	- - Predikatuak zuzendutakoa
Arabiera <i>BRIDGE</i> (Weinberg <i>et al.</i> , 1995) (Holland <i>et al.</i> , 1993)	-	✓	-	-
Japoniera <i>ALICE-chan</i> (Levin eta Evans, 1995)	-	-	✓	-
Ingelesa <i>Grammar-debugger</i> (Si-Qing eta Luomai, 1990) <i>ICICLE</i> (Schneider eta McCoy, 1998) <i>Arboretum</i> (Bender <i>et al.</i> , 2004)	- - - -	✓ - -	- ✓	- -
	-	-	✓	-

III.4 Taula: OLHIA sistemek erabiltzen dituzten teknikak.

III.6 Laburpena eta ondorioak

Erroreak detektatzea, haien diagnosia egitea eta ordain zuzen posibleak ematea, ezinbestekoa da ortografia- eta gramatika-zuzentzaileetan, eta garrantzitsua, OLHIA sistemetan. Hizkuntzaren tratamendu automatikoaren alorrean erroreak detektatzeko eta zuzentzeko erabiltzen diren tekniken eta hauek erabiltzen dituzten sistemen errepasoa egin dugu kapitulu honetan.

Aipatutako teknikak, bi multzo nagusitan banatu ditugu: alde batetik hizkuntza-ezagutzan oinarritutako teknikak edo teknika sinbolikoak daude, zeintzuetan hizkuntzari buruzko ezagutza erregeletan edo beste adierazpideetan kodetzen den modu esplizituan (gogoratu III.2.1 atala). Beste alde batetik, corpusetan oinarritutako teknikak edo teknika enpirikoak daude (III.2.2 atala). Teknika sinbolikoak egoera finituko makinetan eta testuingururik gabeko gramatiketan sailkatu ditugu. Teknika enpirikoek informazio iturri nagusiena corpusetan duten prozedurak erabiltzen dituzte: estatistika eta ikasketa automatikoa.

Ortografia- eta gramatika-zuzentzaileetan sistemaren azkartasunean eragin handiegirik ez duen edozein teknika erabil daitekeen bitartean, aztertutako OLHIA sistema gehienetan patroi-erregelak eta baterakuntza formalismoak erabili direla ikusi dugu. Horren zergatia, ikasleari eskaini behar zaion berrelikaduran dago. Teknika horiek diagnosirako informazio oso zehatza emateko aukera dute, gainontzeko teknikak (estatistikak eta ikasketa automatikoak), ordea, ez dute normalean horrelako informaziorik eskaintzen.

Zenbait ondoriotara iristeko parada izan dugu azterketa bibliografikoa egin ondoren:

- **Corpusen tamainari dagokionez**, corpus gero eta handiagoak lortzea erraza denez, azken urteotako joera corpusetan oinarritutako teknikak —estatistika eta ikasketa automatikoa— erabiltzea da. Gizakiok egindako erregelak ez omen dira inoiz gai izango hizkuntzak duen konplexutasuna jasotzeko. Corpus handiak etiketatuz gero, ordea, testuetatik jasoko dugu behar dugun ezagutza. Corpusen etiketatzean izaten dute teknika enpirikoek zailtasun handiena. Lan hau egina duten hizkuntzetan, ingelesean adibidez, teknika erabilerrazak eta egokiak dira. Euskara modukoetan, berriz, informazio zuzena duen corpora biltzea zaila bada, interesatzen zaigun atazari dagokion erroredun corpora biltzea, are eta neketsuagoa da. Hori dela eta, tesi-lan honetan *hizkuntza-ezagutzan oinarritutako teknikak* erabiliko ditugu. Bestelako teknikak etorkizunerako, testu-bilduma handiagoak ditugun garaietarako, utziko ditugu.

- **Ezagutza linguistikoari dagokionez**, egindako azterketatik ondorioztatu dugu *corpusetan oinarria duten teknikek* lema eta batzuetan kategoria gramatikala besterik ez dutela erabiltzen. Teknika sinbolikoak oinarrian dituzten sistemek, berriz, informazio linguistiko kopuru handiagoa erabiltzen dute maiz. Zehatzagoak izateko, sintagmei buruzko informazioa (izen- eta aditz-kateak), eta zenbait kasutan unitate lexikoen arteko erlazioei buruzko informazioa (funtzio sintaktikoak eta beren arteko erlazioak, mendekotasun-erlazioak kasu) erabiltzen dituzte. IXA ikerkuntza-taldean, une honetan, hizkuntza-egagutza hori guztia lortzeko moduan gaude. Azterketatik ateratako beste ondorio bat honakoa da: egagutza linguistikoa erabiltzeak sistemen doitasuna izugarri hobetzen du. Erroreen detekzioa eta diagnosia ahalik eta zehatzenak izan daitezen nahi dugunez, eta informazio sintaktikoa eskura dugunez, *hizkuntza-egagutzan oinarritutako teknikak* erabiltzea erabaki dugu.
- **Detektatzen dituzten errore motei dagokionez**, erroreak hitz zuzenen nahasketarekin edo gabeziarekin zerikusia dutenean teknika enpirikoak oso egokiak direla ikus dezakegu III.7 taulan. Sintagma-mailako erroreak patroierregelekin (MG, XFST tresnekin, adibidez) lantzea ohikoa dela ondorioztatu dugu; esaldi-mailako erroreak detektatzeko, ordea, osagaien arteko lotura egiteko hizkuntza-egitura konplexuak, analisi-zuhaitzak, esaterako, behar direla uste dugu. Badirudi, gramatika-zuzentzaileek erabiltzen dituzten teknikak aztertuta, aukeratu ditugun erroreak tratatzeko egokienak *hizkuntza-egagutzan oinarritutako teknikak* direla. Hala ere, beste hainbat errore detektatzeko (erroreak baliozko hitzetan, komaren kokapen egokia. . .) ez dugu baztertzen *corpusetan oinarritutako tekniken* erabilera.
- **Emaitzei dagokionez**, aztertutako lan bibliografikoetan maiz ez da ebaluazioari buruzko daturik ematen (ikus III.5, III.6 eta III.7 taulak). Teknika enpirikoak deskribatzen dituzten lanetan zenbait emaitza ematen dira, baina ez, estaldura eta doitasuna; beraz, zaila da batzuetan gainerako sistemekin konparazioa egitea. Teknika sinbolikoak deskribatzen dituzten lanetan, berriz, neurri horiek erabiltzea ohikoa da. Gainbegiratutako sistemek testuinguru mugatuko erroreekin (sintagma barruko komunztadura, adibidez) era askotariko emaitzak lortu dituzte (ikus balio zehatzak tauletan). Emaitzarik onenak alde zurretik eskuz tratatutako laborategiko esaldi motzekin (batez beste 4 hitz) lortu dira (Shaan, 2005): % 96,6ko doitasuna eta % 95,5eko es-

taldura. Sintagma-mailako erroreak detektatzeko testu errealak erabili dituzten sistemen artean emaitzarik onenak MG erabiltzen dutenek lortu dituzte, % 70 inguruko doitasunarekin (Birn, 2000; Johannessen *et al.*, 2002). Testuinguru zabaleko erroreak landu dituzten sistemek (subjektu-aditza komunztadura, adibidez) orokorrean ez dute emaitzarik eman.

- **Diagnosirako informazioari dagokionez**, erroreak detektatzeko esaldirik analizatzen ez duten teknikak ez dute errorearen deskribapenik emateko aukerarik izaten. Eskuz sortutako erregelak erabiltzen dituztenek, ordea, errazagoa izaten dute errorearen deskribapena erakustea eta diagnosis egitea (tresnaren eta erabiltzailearen arteko berrelikadura). Tesi-lan honetan errorearen diagnosirako informazio egokia emateari garrantzitsua irizten diogu; beraz, *hizkuntza-ezagutzan oinarritutako teknikak* erabiliko ditugu.

Hizkuntza-ezagutzan oinarritutako teknikak erabiltzeko arrazoiak azaldu ondoren, hurrengo kapituluetan teknika horiek erabiliz egin ditugun saiakeren berri emango dugu. V. kapitulan egoera finituko mekanismoak erabili ditugu datetan eta postposizio-lokuzioetan gertatzen diren erroreak detektatzeko eta zuzentzeko. Bi fenomeno linguistiko hauek ondoz ondoko hitzetan, bospasei hitzeko leihoetan gertatzen dira; orokorrean, sintagma berean. Aipatu dugun moduan, XFST eta MG tresnak honelako fenomenoak lantzeko egokiak direnez, hauek erabiltzeko modua eta lortutako emaitzak azalduko ditugu. Komunztadura-erroreak, ohiko errore sintaktikoak dira edozein hizkuntzatan. Euskaraz, komunztaduraren maiztasuna handia denez, horiek landuko ditugu VII. kapitulan. Komunztadura-erroreak maiz auzokideak ez diren esaldiko bi elementuren artean gertatzen direnez, ezinbestekoa da esaldiaren egitura ezagutzea errore hauek detektatzeko. Mendekotasun-zuhaitzen gainean patroierregelen bidez erroreak detektatzeko gai den *Saroi* ize-neko tresna aurkeztuko dugu VI. kapitulan.

Gramatika-zuzentzaileak				
	Teknika	Landutako fenomenoak	Emaitzak	
			Laborategiko esaldiak?	Emaitzak
Katalana (Badia <i>et al.</i> , 2004)	MG	Komunztadura sintagman, subj-adi komunztadura...	-	EZ
Suediera <i>Grammatifix</i> (Arppe, 2000)(Birn, 2000)	MG	Komunztadura sintagman, koherentzia aditz katean... Errore sinpleak (egileen hitzetan)	-	D(% 70) E(% 85)
(Hashemi <i>et al.</i> , 2003)	XFST	Komunztadura sintagman, aditz-forma finituak...		D(% 45) E(% 92)
<i>Granska</i> (Carlberger <i>et al.</i> , 2002) (Domeij <i>et al.</i> , 1999)	Eg. finitu	Komunztadura-erroreak, hitz-konposatuak banaketa	-	D(% 53) E(% 52)
<i>ProbGranska</i> (Bigert eta Knutsson, 2002)	Estat.	Hitz zuzenen erabilera okerrak, gaizki zatitutako hitz-konposatuak, ortografia-erroreak...	-	LG
(Sjöbergh, 2005)	Estat.	Hitz zuzenen erabilera okerrak, komaren kokalekua...	-	LG
<i>Scarrie</i> (Paggio, 2000)	TGG+R	Hitz arrunt eta konposatuak, izen-sintagmetan eta aditz-kateetan komunztadura	-	D(% 20,6) E(% 58,1)
<i>SnalGranska</i> (Sjöbergh eta Knutsson, 2005)	IA	Gaizki zatitutako hitz-konposatuak, hitzen ordena okerra	-	LG
Islandiera (Örvar Kárasen, 2005)	IA	Gaizki zatitutako hitz-konposatuak	-	D(-) E(% 65)
Norvegiera (Johannessen <i>et al.</i> , 2002)	MG	Komunztadura sintagman, hitzen ordena...	-	D(% 70) E(-) AF(% 70)
Espainiera <i>GramCheck</i> (Bustamante eta León, 1996)	TGG+R	Sintagma barruko eta arteko erroreak: generoa eta numeroa, <i>dequeismoa</i> , <i>queismoa</i> ...	-	D(-) E(% 28,2)

III.5 Taula: Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (1/3).

Gramatika-zuzentzaileak				
	Teknika	Landutako fenomenoak	Emaitzak	
			Laborategiko esaldiak?	Emaitzak
Italiera <i>JDII</i> (Bolioli <i>et al.</i> , 1992)	TGG+R	Deskribapen teorikoa	-	EZ
Brasilgo portugesa <i>ReGra</i> (Teixeira Martins <i>et al.</i> , 1998)	TGG+R	Komunztadura zatian, aditz-kateko buruaren eta subjektuaren arteko komunztadura		D(% 31,21) E(% 45,6)
Arabiera <i>Arabic GramCheck</i> (Shaan, 2005)	TGG+R	Komunztadura sintagman	BAI	D(% 96,6) E(% 95,5)
Euskara (Gojenola, 2000)	TGG+R	-	-	EZ
Nederlandera (Vosse, 1992)	TGG+R	Errore tipografikoak, ortografikoak (gaizki banatutako hitz-konposatuak barne) eta morfosintaktikoak	-	D(% 47,8) E(-)
Txekiera (Holan <i>et al.</i> , 1997)	Mend.	-	-	EZ
Errusiera (Mitjushin, 1996)	Mend.	Diagnosirik ez, zuzenketa soilik egiten da. Hitzak forma ez-zuzenetan erabiltzea	-	EZ
Frantsesa (Courtin <i>et al.</i> , 1991)	Mend.	Komunztadura-erroreak	-	EZ
Ingelesa (Naber, 2003)	Eg. finitu	Sintagma-mailakoak	-	EZ
(Foster eta Vogel, 2004b)	TGG+E	-	50E	Z(% 84)
<i>AbiWord</i>	Link	-	-	EZ
(Brehony eta Ryan, 1994)	Link	-	-	EZ
(Atwell, 1987)	Estat.	-	-	EZ
(Chodorow eta Leacock, 2000)	Estat.	Gaizki erabilitako hitzak, hitz zuzenen erabilera nahasketak.	-	D(% 80) E(% 20)
(Kumar eta Nair, 2007)	Estat.	Komunztadura-errore batzuk, gaizki erabilitako artikulak, aditzaren denbora...	Esaldi motzak	EZ
(Izumi <i>et al.</i> , 2003)	IA	Hitza falta (ez-egite errorea), hitz zuzenen erabilera nahasketak (ordezkatzeko errorea)	-	D(% 50) E(% 30)
(Bernth, 1997)	ESG	Koordinazioa, ezezeko bi-koitzak, esaldi ez-osoak	Hizkun. kontrolatua	D(% 81) E(% 87)

III.6 Taula: Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (2/3).

Gramatika-zuzentzaileak				
	Teknika	Landutako fenomenoak	Emaitzak	
			Laborategiko esaldiak?	Emaitzak
Ingelesezko nahasketa multzoak (Golding, 1995)	Estat.	Hitz nahasketak	-	EZ
(Golding eta Roth, 1996)	IA	Hitz nahasketak	-	LG
(Mangu eta Brill, 1997)	IA	Hitz nahasketak	-	LG
(Golding eta Roth, 1999)	IA	Hitz nahasketak	-	LG
(Carlson <i>et al.</i> , 2001)	IA	Hitz nahasketak	-	LG

III.7 Taula: Gramatika-zuzentzaileak. Teknikak, fenomeno linguistikoak eta emaitzak (3/3).

IV. KAPITULUA

Abiapuntua. Hizkuntzaren analisia eta errepresentazioa

Erroreak detektatzeko, hauen diagnosia egiteko eta ahal denean erroreak zuzentzeko, *hizkuntza-ezagutzan oinarritutako teknikak* erabiltzea erabaki dugu aurreko kapituluan aipatutako arrazoiengatik. Teknika horiek erabiltzeko, ezinbestekoa da aztertu beharreko testuaren hizkuntza-analisia. Horretan dihardugu azken urteotan LNPrako IXA taldean. Kapitulu honetan IXA taldean euskararen analisi linguistiko konputazionalerako eta lortutako analisisien errepresentaziorako eta kudeaketarako erabiltzen diren baliabideak aurkeztuko ditugu. Bai analisi-katea baita informazio linguistikoa errepresentatzeko egiturak ere, erroreak detektatzeko eta zuzentzeko sistemetan erabiliko ditugu.

Hasteko, euskararako sintaxi-analizatzailea eta hau osatzen duten moduluak eta geruzak aurkeztuko ditugu IV.1 atalean. Atala amaitzeko, informazio ez-gramatikala lantzeko informazio gramatikalean oinarria duten analisi-tresnak erabiltzearen inguruko hausnarketa egingo dugu. Bukatzeko, IV.2 atalean analisirako katea osatzen duten tresnen sarrerako eta irteerako dokumentuen errepresentaziorako erabilitako eredua aurkeztuko dugu.

IV.1 Analisia: katea

Euskararen analisi linguistiko konputazionalerako, IXA taldean mendekotasun-egituretan oinarritutako sintaxi-analizatzaile sendo bat garatu da (Adu-riz *et al.*, 2004). Sintaxi-analizatzaileak analisia geruzaka egiten du geruzetako bakoitzean hizkuntza-ezagutza maila desberdina erabiliz. Analisi-geruzak

katean erabiltzen dira modu sekuentzian eta moduluetan bilduta. Moduluetakoz, mendekotasun-ezartzailea kasu, trukagarriak dira; hots, analisi-kateko azken geruzan erregeletan oinarritutako *Euskararako Dependentsia Gramatika Konputazionala* (EDGK) (Aranzabe, 2008) erabil dezakegu, edo *Maltparser* (Nivre *et al.*, 2007) mendekotasun-erlazio ezartzaile estatistikoaren euskararako egokitzapena (*Maltixa*) Bengoetxea eta Gojenola (2007). Moduluak aurrerago deskribatuko ditugu. Lehenago, analizatzailearen ezauzgarriak aipatzea egokiagoa iruditu zaigu ikuspegi orokorragoa izateko:

- **Sendoa.** Analisi-katea sendoa da Mokhtar *et al.* (2002) lanean ematen den definizioari jarraituz: “Robustness can be broadly defined as the ability for a language analyzer to provide useful analyses of real-world input text”. Sarreran edozein karaktere-kate duela, gure analizatzaileak beti ematen du nolabaiteko analisi erabilgarriren bat. Analisi erabilgarriak *zuzenak* direnak (batzuetan partzialki) eta aplikazio edo ataza automatikoren batean erabil daitezkeenak dira. Analizatzaile sendoek gutxienez analisi bat ematen dute mundu errealeko sarrerekin (ez zero), eta analisisien artean preferentziaren bat ezar dezakete.
- **Orokorrean, hizkuntza-ezagutzan oinarritutakoa.** Analisi-katean, orokorrean, hizkuntzari buruzko ezagutza erregeletan kodetzen da modu esplizituan, ez da testuetatik erauzten (hurbilpen sinbolikoa). Hala ere, geruza batzuetan, lematizatzailean desanbiguazioa egiten denean adibidez, hurbilpen sinbolikoa teknika estatistikoekin (ikuspegi enpirikoa) konbinatu egiten da edo teknika estatistiko hutsak erabiltzen dira (*Maltixa* analizatzaile sintaktikoan, adibidez).
- **Analisi partziala.** *Analisi partziala* terminoak deskribatzen duen kontzeptua izendatzeko *azaleko analisia* eta *analisi arina* terminoak ere erabili izan dira. Analisi partziala testuen gramatika-analisi ez-oso da. Sintaxi-analizatzaile partzialek sintaxi-egitura minimoak sortzen dituzte, zeintzuetan anbiguotasuna ez den guztiz ebatzi (Mokhtar *et al.*, 2002). Carroll-en arabera (2003), *azaleko sintaxian* sarrerako karaktere-katea teilakatzen ez diren unitate-sekuentzietan edo *zatieetan* banatzen omen da. Hitz-sekuentzia bakoitza kategoria sintaktiko batekin etiketatzen da, eta batzuetan, zatiaren *burua* adierazten da. Azken definizio honek deskribatutakoa baino analisi sakonagoa egiten du gure sintaxi-analizatzaileak, hitzen arteko erlazioak ezartzeko gai baita. Hala ere, esaldia osorik ulertzea eta analizatzera iritsi ez garenez, analizatzaile partzial moduan definitu dugu gurea. Analisi partzialak,

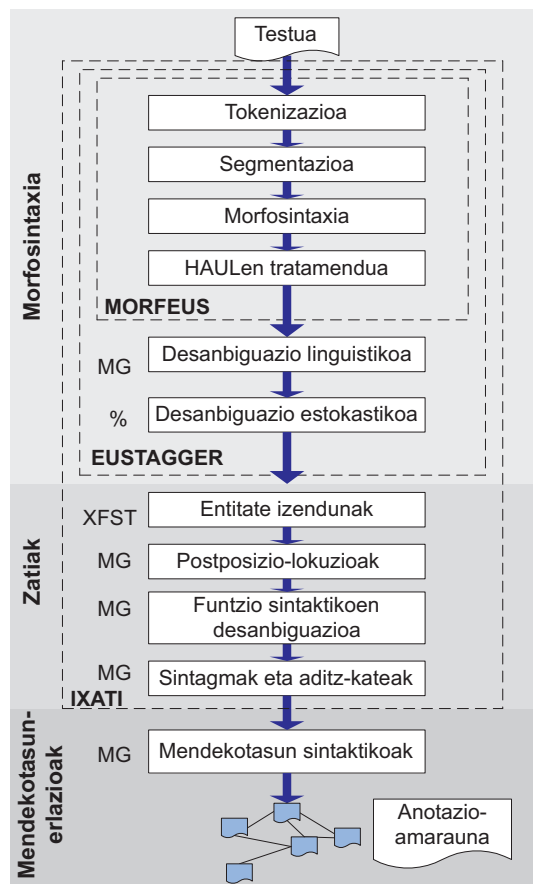
fidagarritasuna eta sendotasuna ditu helburu, sakontasuna eta osotasuna neurri batean galduaz. Testu errealak analizatu ahal izateko, analisi partziala egitea ezinbestekotzat jotzen dugu.

- **Mendekotasun-egiturak adierazpide.** Sintaxi-analizatzailearen emaitza, erlazio bitar asimetrikoekin elkarren artean lotutako elementu lexikalekin osatutako (Nivre, 2005) zuhaitzak dira. Beraz, mendekotasun-egituretan, eta ez osagai-egituretan, oinarritutako sistema bat dugu.
- **Orokorrean, egoera finituko mekanismoak.** Analisi-kateko geruza anitz Xerox Finite State Tool tresna (XFST) (Karttunen *et al.*, 1997; Aït-Mokhtar eta Chanod, 1997) eta Murriztapen Gramatika (MG) formalismoak (Karlsson *et al.*, 1995; Voutilainen eta Tapanainen, 1993; Tapanainen, 1996) erabiliz definitu dira. Formalismo horietan adierazpen erregularren pareko egin ditzakegun patroiak definitzen direla esan dugu III. kapituluan. Egoera finituko mekanismoak oinarrian dituzten sistemak oso *eraginkorrak* dira eta maiz erabiliak izan dira sintaxi-analisan. Esan beharra dago, testuingururik gabeko gramatika bat ere erabili dela hitzaren gramatika (morfosintaxia) definitzeko.

Analisi-kateko geruza bakoitzak aurreko geruzak eskaintzen dion informazioa erabiltzen du sarrera moduan, eta jasotako analisia informazio linguistiko berriarekin aberasten du. Hizkuntza-ezagutzari dagokionez egitura modular honek sintaxi-analisia geruza bakoitzean sakontasun ezberdinarekin egin dadin ahalbidetzen du. Gainera, modulartasunak hizkuntza-datuen mantenuan laguntzen du eta sistema erabilgarria eta era errazean moldagarria egiten du. Sintaxi-analisia urratsez-urrats egiten da eta erabiltzailearen esku geratzen da erabili nahi duen hizkuntza-ezagutza mailaren aukeraketa.

Geruzetako bakoitzean bereizketa argia egiten da gramatiken eta hauek aplikatuko dituzten programen artean. Gramatikak, MG eta XFST tresnek definitutako erregela-formatuetan kodetzen dira. Tresna horiek dira, hain zuzen ere, gramatika horiek aplikatzeko erabiliko direnak. Bi formalismo horiek ordena askeko elementuekin lan egiteko metodologia eta tresna egokiak eskaintzen dituztelako aukeratu dira.

Analisi-kateko moduluak eta haien analisi-geruzak IV.1 irudian ikus ditzakegu.



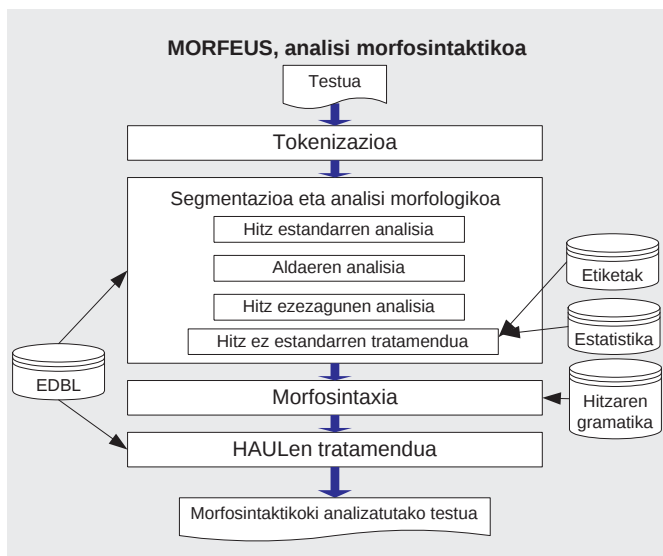
IV.1 Irudia: Euskararako geruza anitzeko sintaxi partzialeko analizatzailea.

IV.1.1 Deskribapena

IV.1.1.1 MORFEUS, analizatzaile morfosintaktikoa

Analisi-prozesua MORFEUS (Aduriz *et al.*, 1998) analizatzaile morfosintaktikoarekin hasten da. MORFEUSEk sarrerako testua jaso eta tokenetan banatu ondoren, horietako bakoitzarentzako lema eta morfema konbinazio posible guztiak ematen ditu, beraiei buruzko informazio morfologiko guztiarekin batera. Euskaraz, hizkuntza-unitateak propietate morfologikoak eta sintaktikoak edukitzen dituen *morfosintaxia* terminoa erabiltzen dugu morfologia terminoaren ordez (gogoratu III. kapitulua).

Ezeizak (2002) bere tesi-lanean informazio zabala ematen du MORFEUSEn inguruan. Analisi-kateko modulu hau osatzen duten geruzak (ikus IV.1.1.1 irudia) deskribatuko ditugu jarraian:



IV.2 Irudia: MORFEUS, analizatzaile morfosintaktikoa.

1. **Tokenizazioa.** Tokenizatzaileak testua unitate edo tokenetan, eta gero esaldietan, banatzen du. Hots, testuan honakoak identifikatzen ditu: hitzak, zenbakiak, arruntak eta erromatarrak, deklinatu gabeak eta deklinatuak; laburdurak eta siglak; zuriuneak eta puntuazio-markak¹. . . Tokenei beharrezkoa den informazio tipografikoa gehitzen zaie etiketen bidez (hasiera maiuskulaz, hitz osoa maiuskulaz; siglak eta zenbaki deklinatuak). Tokenak banatzeari dagokionez, analisirako tratamendu-unitatea hitza izanik, esan beharra dago formatua soilik kontuan hartzen bada, hitza zehaztea ez dela lan erraza, baina bai garrantzitsua, tokenizazioak ondorengo prozesuak elikatzen baititu. Tokenizatzaileak sarreran testu gordina eta XML formatuak onartzen ditu.
2. **Segmentazioa edo analisi morfoloikoa.** Segmentatzaileak, testu-hitz

¹Tokenizatzailearen deskribapen zehatzagoa ematen da Ezeizaren (2002) lanean, 28. orrialdean.

bakoitza esanahia duten unitate txikienetan, lema eta morfemetan, banatzen du. Osagai hauei buruz, morfotaktika (hots, hitz bakoitzaren ondoren ager daitezkeen morfema segida posibleak eta beraien ordenaren murriztapenak) eta informazio morfologikoa (hitzaren segmentu bakoitzari dagozkion analisi morfologikoak) ematen dira. Tokenizatzailetik jasotako testu-hitz bakoitzeko interpretazio posible guztiak identifikatzen ditu segmentatzaileak. Interpretazio bakoitzean izen eta adjektiboen kasuan hitz-formaren kategoria, azpikategoria, deklinabide-atzizkia, numeroa eta mugatasuna ezaugarriak ematen dira; modu/denbora eta aspektua, aditzen kasuan. Hizkuntza-informazio hau guztia *Euskararen Datu-Base Lexikaletik* (EDBL) jasotzen du. Datu-base aberats honek 80.000 sarrera ditu, eta bertan, hiztegi-sarrerak, adizkiak eta morfema ez-askeak² gordetzen dira bakoitza bere informazio morfologikoarekin.

Analizatzaile morfologikoa bi fasetan eraiki zen:

2. 1. Lehen fasean, *euskara estandarreko analizatzaile morfologikoa* egin zen % 95eko estaldura-tasa lortu zelarik.
2. 2. Hitzen % 100 ezagutu ahal izateko, bigarren fase batean analizatzailea *hitz ez-estandarrek tratatzeko* zabaldu zen. Hitz ez-estandarrek ezagutzeko, bi analizatzaile berri sortu ziren: aldaera dialektalak eta gaitasun-desbideratzeak lantzen dituenak, eta lexikorik gabeko hitzak lantzen dituen analizatzailea (*guesser*), bestea.

Hiru analizatzaile hauek (estandarra, aldaera/desbideratzeena eta lexikorik gabekoa) bata bestearen atzetik aplikatzen dira modu inkrementalean.

Ebaluazioari dagokionez analizatzaile morfologikoak, zuzentasunari begiratzen diogunean, % 99,3tik gorako neurriak lortzen ditu (Alegria *et al.*, 2003a). Hitzen % 79,83 dira anbiguo (tokenen % 65,86) batez beste 3,87 interpretazio dituztelarik. Ikuspuntu konputazionaletik segmentatzailearen berri Alegriaren (1995) eta Ezeizaren (2002) tesi-lanetan ematen da. Ikuspuntu linguistikoa, berriz, Urkiarenean (1997).

3. **Morfosintaxia.** Segmentatzailean lortutako morfemei buruzko informazio morfologikoa (eta hainbat kasutan sintaktikoa ere) *biltzea* eta *opti-*

²Morfema ez-askeak hiztegi-sarrera ez direnak dira, deklinabide-morfemak, erlazio-morfemak, atzizki-lexikalak, aurrizki-lexikalak, aspektu-morfemak, elipsia eta elkarketa-marra, hain zuzen ere.

mizatzea da analizatzaile morfosintaktikoaren helburua (Aduriz *et al.*, 2000; Arriola *et al.*, 2005). Informazioa *garbitu* edo *iragazi* egiten da zenbait informazio soilik aukeratu edo goratuz. Ezeizaren (2002) hitzen arabera, “segmentazioa gaingitu eta formaren barruko informazioa elaboratu behar da, hitzaren egituraren berri emateko eta horixe da analisi morfosintaktikoaren zeregina”. Morfosintaxiaren deskribapen sakona egiten da (Arriola *et al.*, 2005) lanean.

Analisi morfosintaktikoa ahalbidetzen duen *gramatika*, hitzaren egitura deskribatzen duen testuingururik gabeko gramatika bat da (Gojenola, 2000; Aduriz *et al.*, 2000). Gramatika honek morfemetatik lortutako informazioa konbinatu egiten du hitz-formaren interpretazio bakoitzeko ezaugarri-egitura bat emanaz. Baterakuntza erabiltzen du ezaugarri-egituren arteko bateragarritasuna egiaztatzeko. Programari dagokionez, PATR formalismoaren inplementazio bat aukeratu da (Shieber, 1986), bere sinpletasun eta malgutasunagatik.

4. **Hitz anitzeko unitate lexikalen (HAUL) tratamendua.** HAUL kontsideratzen dira hitz elkartuak, lokuzioak eta kolokazio murriztuak (Alegria *et al.*, 2004). Zehazkiago, EDBLn landu diren elementuak³, inolako barne-aldaketa edo morfosintaxi aldaketarik behar ez duten adierazpen finkoak (*in situ*, *a priori*, *e.a.*), esamolde deskonposagarri eta ez-deskonposagarriak, eta lexikalizatutako termino konposatuak izan dira. Atsotitzak, kolokazioak, makulu-hitzak eta alderaketak ez dira kontuan hartu. Entitate izendunentzako, datentzako eta zenbakientzako aparteko tratamendua egin da. Hitz anitzeko unitate lexikalak mugatzeko irizpideak (Ezeiza, 1997) lanean azaltzen dira.

EDBLn 2.270 hitz anitzeko unitate deskribatu dira. Hauetatik 1.300 inguru 4,7 milioitako corpus batean maiztasun handienarekin azaltzen direnak dira, gainontzekoak *Hiztegi Batutik*⁴ erauzi dira.

Hitz anitzeko unitateak identifikatu eta tratatzeko HABIL programa diseinatu eta inplementatu da. Horrek, lehen urrats batean testuko HAULak identifikatzen ditu EDBLko informazioa baliatuz, eta bigarren urrats batean dagokien analisi morfosintaktikoa esleitzen die. HABILek HAUL jarraituak eta ez-jarraituak lantzen ditu, osagaien ordena posible guztiak eta deklinabide murriztapenak kontuan hartzen ditu

³HAUL jarraitu batzuk segmentazioan eta morfosintaxian landu dira, adibidez, *etxe-egile*.

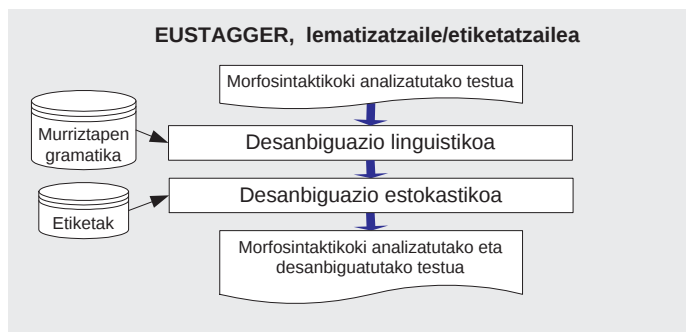
⁴<http://www2.euskaltzaindia.net/hiztegiatua>

eta, esan dugun bezala, unitate osoari dagokion analisi morfosintaktiko berria sortzen du.

IV.1.1.2 EUSTAGGER, lematizatzaile/etiketatzailea

Sarreran MORFEUSEk emandako analisia izanik, EUSTAGGER euskararako lematizatzaile/etiketatzailearen helburua, sarrerako hitz-forma bakoitzeko, *testuinguru horretan* dagokion lema eta etiketak ematea da. Hau da, MORFEUSEk, hitz-forma bati dagozkion analisi morfosintaktiko posible guztiak ematen ditu, eta aukera guztien artean testuinguruko informazioa begiratu-ta onartezinak diren interpretazioak baztertu, ezabatu egin behar ditu EUSTAGGERek, egokiarekin geratzen saiatuz. Hiru anbiguotasun morfosintaktiko mota bereizten dira (Ezeiza, 2002): kategoriari, morfema ez-askeei eta sintaxiari dagozkionak. Kategoriari dagokion anbiguotasuna, adibidez izen/aditz, aditz/adjektibo/adberbio... moduko anbiguotasuna, omen da korapilatsuen-a. Kategoria sinplea soilik kontuan hartuta (azpikategoria ez) testu-hitzen % 46-48 anbiguo dira (2,3-2,4 interpretazio hitzeko). Esan beharra dago, analisirako aukeratutako etiketa-sistemaren granularitateak eragin handia duela anbiguotasunean. Gainera, euskararen kasuan, kategoria mailako anbiguotasunari, atzizkien anbiguotasun morfosintaktikoa gehitzen zaio.

Aipatu berri dugun anbiguotasuna ebazteko, teknika ezberdinak konbi-natu dira EUSTAGGERen (ikus IV.3 irudia): hizkuntzan oinarria duten tek-nikak (sinbolikoak) eta teknika estatistikoak (enpirikoak):



IV.3 Irudia: EUSTAGGER, lematizatzaile/etiketatzailea.

- Desanbiguazio morfosintaktikorako erabili den murriztapen-gramatika (hizkuntza-ezagutza) 1.113 erregelek osatzen dute, eta hauek de-

finitzeko ataza honekiko independentea den maila anitzeko etiketa-sistema bat erabili da. Etiketa-sistema hau lau mailatan banatuta dago. Lehen mailan hitzaren gramatika-kategoria erabiltzen da (20 kategoria-etiketa), bigarrenean, aurrekoaz gain gramatika-azpikategoria erabiltzen da (45 etiketa), hirugarrenean, aurrekoari informazio interesgarria gehitzen zaio, adibidez deklinabide-kasua, eta azkenekoan, analizatzaile morfologikoak emandako informazioa gehitzen da. Desanbiguazio-gramatika multzoetan banatu da aipatu berri dugun etiketa-sistemaren arabera. Desanbiguazio linguistikoaren berri Aduriz-en tesi-lanean (Aduriz, 2000) ematen da, besteen artean (Aduriz eta Díaz de Ilarraza, 2003; Aduriz *et al.*, 1997).

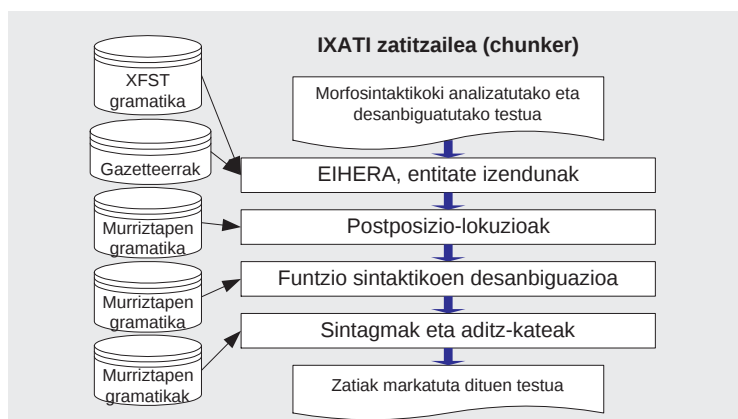
- Estatistika (corpusetan oinarritutako teknika) erabiltzen duen desanbiguaziorako moduluak lehen mailako Markov-en eredu ezkutuak ditu oinarri. Moduluan corpus handi batetik automatikoki erauzitako estatistikak erabiltzen dira hitzari dagozkion interpretazioen artean desanbiguazioa burutzeko.

Hainbat esperimenturen emaitzak aztertuta, desanbiguaziorako hurbilpenik egokiena metodoen konbinazioa dela ikusi da. Horrela, Markov-en ereduak soilik erabiliz bigarren mailan lortutako % 85,01eko desanbigutasun-doitasuna, metodo konbinatua erabiliz % 95,42ra igo da.

IV.1.1.3 IXATI zatitzailea

IXATI zatitzaileak testua ingelesez *chunk* deritzotenetan edo *kateetan* banatzea du helburu. Katea sintagma kategoriako zatia da eta sintaktikoki erlazionatutako hitzez osatuta dago. Horrela bada, testua kateetan zatitzea gainjartzen ez diren eta elkarrekin sintaktikoki erlazionaturik dauden hitz multzoak atzematean datza. Analisi-katean sintaktikoki erlazionaturiko hitz multzo kontsideratu ditugu: sintagmak eta aditz-kateak, postposizio-lokuzioak eta entitate izendunak. Horiek markatu ahal izateko, zatitzaileak dagoeneko morfosintaktikoki analizatutako eta desanbiguatutako testua jasotzen du.

- **Entitate izendunak (EIHERA).** Elementu hauek “*Message Understanding Conference (MUC)*” kongresuan honela definitu zituen Chinchorek (1998): entitate izendunak ezagutzen direnean, entitate-izenak (pertsonak, tokiak eta erakundeak), denbora-adierazpenak (datak eta orduak) eta zenbait zenbakizko espresio (portzentaajeak, diru-balioak...) identifikatu eta sailkatu egiten dira. Entitate izendunak ezagutu eta



IV.4 Irudia: IXATI zatitzailea (chunker).

sailkatzeko bi urratsetako metodo konbinatu bat erabiltzen du euskararako EIHERA tresnak (Alegria *et al.*, 2003b). Lehen urratsean eta informazio morfosintaktikoa erabiliz, XFST tresnako formatua jarraituz definitutako gramatika baten bidez entitate izendunak testutik erauzi egiten dira. Bigarren urrats batean, dagoeneko ezagutu diren entitate izendunak *pertsona*, *erakundea* eta *tokia* multzoetan sailkatzen dira testuinguruko informazioarekin eta *gazetteer*ekin⁵ definitutako zenbait heuristikok erabiliaz. Entitateok elementu bakarrekoak (adibidez, **Europar**, *tokia*) edo hitz anitzekoak (adibidez, **Europako Banku Zentralea**, *erakundea*) izan daitezke. EIHERA ebaluatu denean, sailkapena ongi egiten duela ikusi da (% 81,36ko doitasuna) baina ezagutza ez hain ongi (% 78,99ko doitasuna eta % 83,73ko estaldura).

- **Postposizio-lokuzioak.** Postposizio-lokuzioak (ikus deskribapen zabalak V. kapituluak) “perpausaren sintagmen arteko erlazio gramatikalak adierazten dituzten forma askeak direla” adierazten da EGLU-I (Euskaltzaindia, 1985) gramatika (adibidez, etxea-**ren barruan**, neska hon-**i buruz**...). IXATIn islapen-erregelak osatutako gramatikak erabiltzen dira egitura hauek ezagutzeko.

Postposizio-lokuzioen ebaluazioa egiteko, 2.279 postposiziotako lagin

⁵*Gazette*erra pertsona, erakunde, toki eta bestelako izenen zerrenda alfabetikoa da.

bat jaso da eta gramatika pasa ondoren, honako emaitzak lortu dira: % 88,73ko estaldura eta % 93,13ko doitasuna.

- **Funtzio sintaktikoen (FS) desanbiguazioa** (Aduriz, 2000). Kateko analisietako bakoitzak, gramatikak hala eskatzen duelako, funtzio sintaktiko bat (edo gehiago) izaten du esleituta. Funtzio sintaktiko hauek, esaldiaren azaleko sintaxiaren berri ematen dute, batzuetan funtzio tradizionalekin⁶ (@SUBJ, @OBJ...), eta besteetan sintagmak eta aditz-kateak lotzeko erabiltzen diren funtzioekin⁷ (@KM>, @-JADNAG_MP_SUBJ...). Funtzio sintaktiko hauetako batzuk EDBLtik datoz, beste batzuk, ordea, morfosintaxia lantzen denean esleitzen dira. Horrela, forma bati izan ditzakeen funtzio sintaktiko guztiak esleitzen zaizkio; beraz, batzuetan, hitzak FS bat baino gehiago baditu, hitza sintaktikoki anbiguo izango da. Adibidez, oso ohikoa da absolutibo kasuetan @SUBJ/@OBJ/ @PRED⁸ anbiguotasuna. Urrats honetan Murriztapen Gramatika erabiltzen da funtzio sintaktikoen desanbiguazioa egiteko.
- **Sintagmak eta aditz-kateak**. Zati horiek hainbat murriztapen-gramatika erabiliz ezagutzen dira, horretarako funtzio sintaktikoei buruzko informazioa baliatuz. Horrela, aditz-kateak ezagutzeko, aditz-erlazio etiketak⁹ (@+JNAG, @-JNAG, @+JLAG...) eta partikula batzuk (ezeztapenerako partikula, partikula modala...) erabiltzen dira. Sintagmak azalarazteko, berriz, honakoa onartzen da: funtzio modifikatzailea duen edozein hitz, sintaxi-funtzio etiketa nagusi bat duen (@SUBJ, @OBJ, @ZOBJ¹⁰) hitz edo hitz multzoren batera lotuko da (adibidez, [Nire (@KM>) txostena (@SUBJ)]_S luzea da. Are gehiago, sintaxi-funtzio etiketa nagusi bat duen hitz batek, berak bakarrik osa dezake sintagma (adibidez, [Nik (@SUBJ)]_S jaso dut). Beraz, hitz bakarreko edo hainbat hitzeko sintagmak topa ditzakegu. Era berean, aditz-kateak ere elementu bakarrekoak edo hainbat elementukoak izan daitezke. Azken kate hauetan, gainera, aditz-kate jarraituak eta ez jarraituak (adibidez, [Ez da]_{AK1} euririk [espero]_{AK2}) topa ditzakegu. Gehienez ere, hiru osagai dituzten aditz-kate ez-jarraituak ezagutu dira.

Sintagmak eta aditz-kateak harrapatzen dituen atala 24.000 hitzez osa-

⁶@SUBJ: subjektu; @OBJ: objektu.

⁷@KM>: Kasu-markadunaren modifikatzailea; @-JADNAG_MP_SUBJ: subjektu funtzioa duen aditz nagusi jokaturgabe mendekoa.

⁸@PRED: predikatiboa.

⁹@+JNAG: aditz nagusi jokaturgabe; @-JNAG: aditz nagusi jokaturgabea; @+JLAG: aditz laguntzailea.

¹⁰@ZOBJ: zehar-objektu.

tutako lagin batekin ebaluatu da, doitasuna eta estaldura neurtuz. Oro har, emaitza onak lortu dira. Adibidez, osagai bakarrek sintagmetan % 94,7ko doitasuna eta % 80,3ko estaldura lortu dira. Osagai bat baino gehiagoko sintagmetan, ordea, emaitzak ez dira hain onak izan (% 84ko doitasuna eta % 91,1eko estaldura). Arazo iturri nagusia sintagma koordinatuak izan dira.

IV.1.1.4 Mendekotasun-erlazioen ezartzailea

Esaldien egitura *dependentzia* edo *mendekotasun* izeneko hitzen arteko erlazioen bidez adierazi daitekeelakoaren ideia, mendekotasun-egituretan oinarria duten gramatika-teoria eta -formalismo multzoen azpian dago. Mendekotasun-erlazioa *gubernatzailea* (edo *burua*, *gurasoa*) izeneko hitz baten eta, *mendekoa* (edo *modifikatzailea*, *umea*) izeneko beste hitz baten artean ezarritako erlazio bitar asimetrikoa da (Lin, 1998). Erlazio horiek esaldiko hitz guztiak lotzen dituen mendekotasun-zuhaitz bat eratzen dute normalean. Esaldiko hitz batek mendeko anitz izan ditzake, baina hitz bakoitzak beste hitz bat soilik modifika dezake. Mendekotasun-zuhaitzaren erroa da inolako hitzik modifikatzen ez duen hitz bakarra, eta esaldiko burua dela esaten da. Mendekotasun-egiturak hitzen ordena librean duten eta, sintaxi-egitura ez-jarraituak dituzten hizkuntzetarako egokiak dira (Mel'čuk, 1988): euskararako egokiak dira.

Euskaraz hitzen artean eta zenbait zatiren artean (entitate izendunen artean, adibidez) ezartzen ditugu gobernatzaile-mendeko moduko erlazioak. Ezarritako erlazio ia guztiak sintaxi-mailakoak dira.

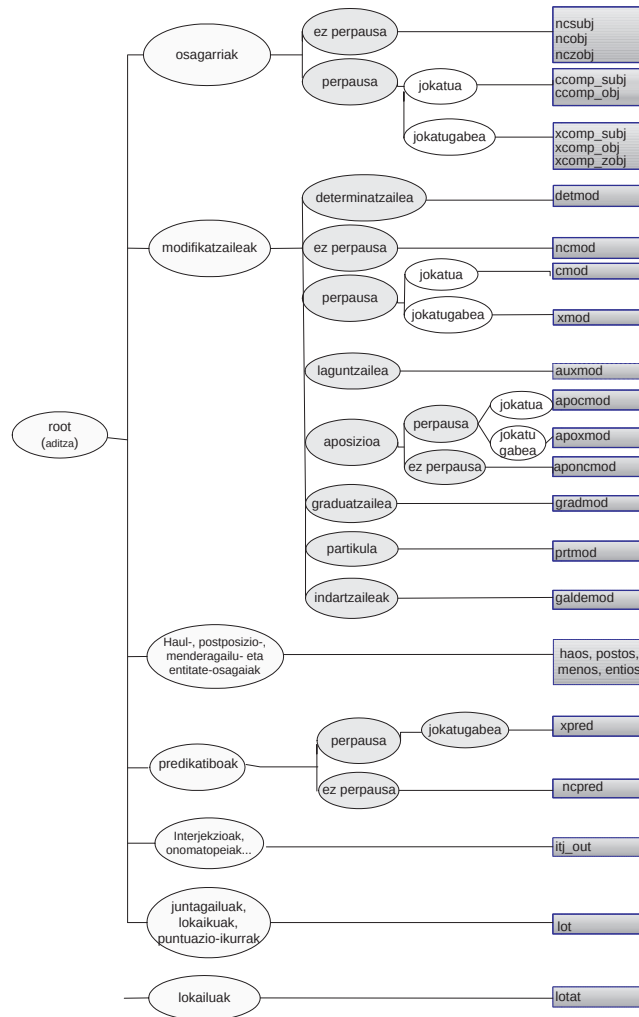
Analisi-kateko azken geruza hau bereziki garrantzitsua da VI. kapituluaren azalduko dugun *Saroi* tresnarako, eta hau dela eta, arreta berezia eskainiko diogu mendekotasun-erlazioen etiketatzeari.

Mendekotasun-erlazioak sintaxi-analizatzailearen bidez automatikoki ezartzen hasi aurretik, *mendekotasun-egituren eskema* bat definitu da. Eskema honetan azaltzen den etiketa-sistemak Carroll-ek aurkeztutako egituren du oinarria (Carroll *et al.*, 1998, 1999). Hala ere, zenbait desberdintasun daude, gure sisteman lexikalizatu gabeko argumentuak azal baitaitezke (adibidez, fonetikoki hutsa den *pro* argumentua, zeina *pro-drop*¹¹ izeneko hizkuntzetan azaltzen den). Etiketatze-sisteman definitutako etiketek IV.5 irudian azaltzen den hierarkia osatzen dute.

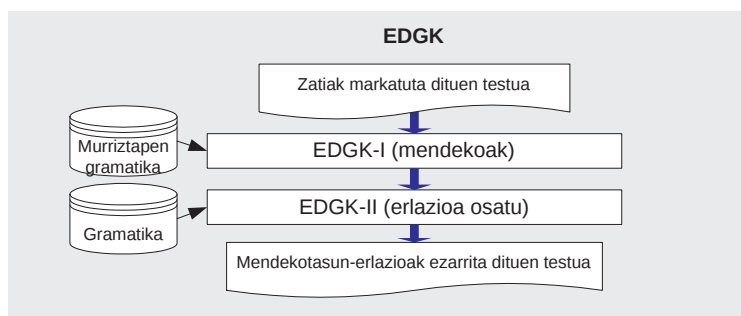
IXA taldean mendekotasun-erlazioak ezartzeko bi hurbilpenetan lanean dihardugu. Lehenean, hizkuntza-ezagutzan oinarria duenean, *Euskararako*

¹¹ *pro-drop* edo *pronoun dropped*: hizkuntza batean izenordainak ezaba daitezke pragmatika edo testuingurua erabiliz ondorioztatu daitezkeenean.

Dependentzia Gramatika Konputazionala (EDGK) (Aranzabe, 2008) izeneko gramatika erabiltzen da mendekotasun-erlazioak ipintzeko. Bigarrenetan, *Maltparser* (Nivre *et al.*, 2007) izeneko analizatzaile sintaktiko estatistikoa (corpusean oinarritua) euskararako egokitu da. Oraingoz, EDGK soilik integratu da katean, beraz, batez ere hau deskribatuko dugu.



IV.5 Irudia: Mendekotasun-erlazioak adierazten dituzten etiketak.



IV.6 Irudia: EDGK, mendekotasun-erlazioen ezartzailea.

EDGK

EDGK erabiliz, mendekotasun-erlazioak bi urratsetan ezarri dira. Lehen urratsean, Murritzapen Gramatika tresnak aplikatuko duen EDGK-I gramatika garatu da. Horrek, islapen-erregelen bidez, mendekoa den elementuari, gobernatzailearekiko duen sintaxi-erlazioa adierazten duen etiketa bat ezartzen dio. Etiketak sintaxi-erlazioaren izena edukitzeaz gain (ikus IV.5 irudian eskuineko zutabea), ">" edo "<" ikurraren bidez, gobernatzailea zein noranzkotan dagoen adierazten du. Murritzapen Gramatikak, ordea, mugak ditu: gramatikako erregeletako bakoitzak gobernatzaileari buruzko informazioa islatzen duen arren, testuan ezin da adierazi zein tokenek egiten duen gobernatzaile funtzioa (haren noranzkoa soilik). Hori egiteko, bigarren urrats batean, *Burubil* izeneko tresnak EDGK-II gramatika aplikatzen du.

EDGK-I sintaxi-analizatzaileak zatitzailearen irteera jasotzen du, 505 erregela ditu, eta horietako bakoitza garatzeko, eskuz desanbiguatutako corpus bat (24.000 hitz-forma) erabili du.

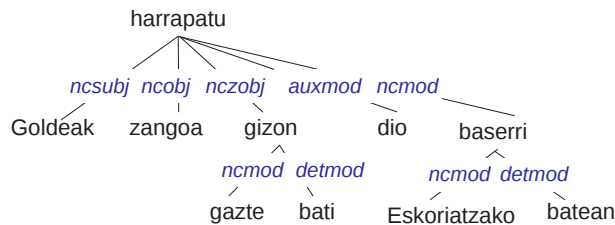
Adibide baten bidez ikusiko dugu mendekotasun-erlazioak nola ezar daitezkeen. Har dezagun (9) adibideko esaldia:

- (9) **Goldeak zangoa harrapatu dio gizon gazte bati Eskoriatzako baserri batean.**

Esaldi honi IV.1 taulan adierazitako mendekotasun-etiketak ipiniko lizkioke EDGK-I gramatikak. Taula honetan adierazten den informaziotik abiatuta, IV.7 irudiko mendekotasun-zuhaitza lortu beharko genuke.

Hitza	Mendekotasun-etiketa
Goldeak	NCSUBJ>
zangoa	NCOBJ>
harrapatu	
dio	<AUXMOD
gizon	<NCZOBJ
gazte	<NCMOD
bati	<DETMOD
Eskoriatzako	NCMOD>
baserri	<NCMOD
batean	<DETMOD

IV.1 Taula: EDGK-I sintaxi-analizatzailearen emaitza.



IV.7 Irudia: (9) adibideko esaldiaren mendekotasun-zuhaitza.

Adibidean, **baserri** hitzari **<NCMOD** etiketa ipinita, hitz horrek (baita bere azpiko **Eskoriatzako baserri batean** sintagmak ere) gobernatzaile duen aditzarekiko, **harrapatu** hitzarekiko, *perpauza ez den modifikatzailea* (NCMOD) mendekotasun-erlazioa duela adierazi nahiko genuke. Hori beharrean, erlazioa eta gobernatzailearen noranzkoa soilik adierazten dira, eta ez, gobernatzailea zein den.

Hori, EDGK-II gramatikak egiten du elementuen kategoria gramatikalak erabiliz. Adibidez, gramatikaren arabera, **baserri** hitza izena izanik baina ez leku-izena, eta gainera, ez dagoenez ez genitibo, ez lekuzko genitiboan deklinatuta, ezkerrera duen lehen aditzarekin (edo aditz trinkoarekin) egin beharko luke lotura. Aditz hori (**harrapatu**) da bere gobernatzailea.

Maltparser

Maltparser (Nivre *et al.*, 2007) analizatzaile sintaktiko estatistiko determinista da. Hizkuntzarekiko independentea da eta analizatzaile sintaktikoa indusitu dezake *treebank* bat oinarrian hartuta eta entrenamendurako datu

multzo mugatuak erabilia. *Maltparser* sistema, CoNLL 2007 lehiaketan¹² lehena geratu zen, 20 sistemen artean lehiatu ondoren.

Analizatzaile hau euskara lantzeko egokitu da (Bengoetxea eta Gojenola, 2007) *Maltixa* sortuaz, eta oinarrian erregela sintaktikoak ikasteko, 3LB *treebanka* erabili du (*treebankari* buruzko xehetasunak VII. kapituluaren emango dira). Analizatzailea ebaluatzeko, etiketa zuzenaren esleipena neurtzen duen neurria (*Labeled Attachment Score, LAS*) erabili da eta honen arabera, % 76,76ko emaitza lortu du. Analizatzaileak beti ipintzen du esaldiko hitz guztietan dependentzia-etiketaren bat eta beraz, mendekotasun-zuhaitza bakarra eta osoa ematen du.

Zuhaitz hori sortzeko, *Maltparser*rek analisi morfosintaktiko bakarra behar du tokeneko: sarrera guztiz desanbiguatua morfosintaxiari dagokionez. Ez du funtzio sintaktikorik erabiltzen; beraz, ez du sintaxi-anbiguotasunik.

IV.1.2 Analisi-katearen egokitasuna informazio ez-gramatikala lantzeko

Erroreen detekziorako, diagnosirako eta zuzenketarako erabiltzen diren teknikak eta sistemak errepasatu ditugu aurreko kapituluaren. Bertan aipatutako sistema ia guztiek nolabaiteko analisi linguistikoa egiten dute, eta horretarako erabiltzen dituzten analizatzaileak ez dira bereziki prestatuta egoten informazio ez-gramatikala lantzeko. Kapitulu hau guk erabiliko dugun analisi-katea azaltzen hasi dugu. Katea erabiltzen hasi aurretik, bi galdera egin dizkiogu, ordea, gure buruari:

1. Informazio gramatikalaren analisirako prestatuta dagoen analisi-katea *egokia* al da informazio ez-gramatikala analizatzeko?
2. Eta, erabiliko bagenu, analisi-kateak *egokitzapenik* beharko luke?

Jarraian bi galdera hauek erantzuten saiatuko gara.

Egokitasuna

Analizatzaile morfosintaktiko edo sintaktiko sendo bat garatzeak izugarriko lana ematen du. Behin hori edukita, informazio ez-gramatikalarekin lan egingo lukeen “beste” analizatzaile baten garapenarekin hastea ez da bideragarria izaten ikerkuntza-talde gehienentzat. Beraz, dagoenarekin lan egiten da gehienetan, erabilera hori egokia ote den ala ez planteatu gabe. Maritxalarrek bere tesi-lanean (Maritxalar, 1999) bigarren hizkuntzako ikasleen hizkuntza-ezagutza aztertzen duenean, analisi automatikorako aipatu berri

¹²<http://www.cs.jhu.edu/EMNLP-CoNLL-2007>

dugun hurbilpenari desegokia irizten dio. Bere proposamenean, ikasleen ikaskuntza-maila bakoitzaren tartehizkuntza finkatu beharko litzateke, eta hori egin ondoren, analisirako tresnetan informazio hori erabili. Horrela, hiztegiari buruz ari denean bere hitzetan, “ikaskuntza-maila bakoitzeko tresnek, hots, analizatzaileek, desanbiguatzaileek eta lematizatzaileek, mailari dago-kion hiztegia erabiliko lukete”. Tartehizkuntzan ezagutza zuzena zein ez-gramatikala gordetzen denez, analisisian bertan islatuta egongo litzateke *errorea*. Errorea duen analisiari marka bat ipini beharko litzaioke estandarretik zein mailatan desbideratzen den adierazteko. Aukera hori interesgarria iruditzen zaigun arren, eskatzen duen lan kopuruagatik baztertu egin dugu. Dagoeneko eskuartean daukagun analisi-katearekin lan egitea erabaki dugu, hori bai, egokitzapenik behar ote duen aztertu ondoren.

Egokitzapenen beharra ala ez

Analisi-katean egokitzapenik egin behar ote den edo ez ezagutzeko, bi aztergai izango ditugu ikerketa-helburu: erabilitako *analisi-teknika* eta aztertutako *fenomeno linguistikoak*.

Teknika

Teknikari dagokionez, gure analisi-kateak hurbilpen murriztailea erabiltzen duela esan dezakegu: analisi-aukera guztiak ematen ditu, eta gero batzuk baztertu. Horrela, erroreari dagokion analisisia interpretazioen artean egotea errazagoa da hurbilpen eraikitzailean baino.

Analisi-kateko geruzak aztertu ditugu testu ez-gramatikalen analisisian zein jokaera duten ezagutu nahian.

- Hitz-mailako analisisian honako hiru kasuak topatu ditugu:
 1. Hitz zuzena. Ohiko jokaera du EUSTAGGERek.
 2. Hitz okerra (adibidez, [xuhaitx](#)). MORFEUSEk hitza ezin duenez euskara estandarreko moduluaz analizatu, ez-estandar moduan analizatzen saiatzen da. Hitza aldaera dialektal edo gaitasun-desbideratzea bailitzan analizatzen saiatuko da, eta lortzen ez badu, hiru garren urrats batean lexikorik gabe analizatuko du. Beti lortuko du nolabaiteko analisisiren bat eta desanbiguatzaileak ere interpretazioen bat utziko du (MGren ezaugarria da hau). Utzitako interpretazioak erroreak detektatzeko beharrezkoak diren hizkuntza-ezaugarriak eduki ditzake, edo ez.

3. Hitza testuinguru horretan okerra da baina besteren batean zuzena. Demagun pertsona batek **etzean** idatzi duela **etxean** idatzi nahi zuenean **etzean itoginak dauzka** esaldian. Lehen hitza hiztegian azaltzen denez¹³ modu honetan analizatuko du MORFEUSEk, **etxean** hitz-formari dagokion analisirik eman gabe. Kasu honetan ezingo litzateke errorea detektatu analisi morfologikoan oinarrituta, semantika erabili beharko litzateke, **basamortuan** ezin baita **itoginik** egon.
- Desanbiguatzaile morfosintaktikoak informazio ez-gramatikarekin duen jokaera aztertu aurretik, erdaretan erroreak detektatzeko Murritzapen Gramatika erabiltzen duten zenbait sistema aztertuko ditugu. Norveгийн erroreak detektatzeko egin diren lanetan (Johannessen *et al.*, 2002; Hagen *et al.*, 2001) *erlaxatu* egin dituzte desanbiguazio erregela batzuk. *Erlaxatu* terminoa erabiltzen dute zenbait murritzapen *kendu* edo *baztertu* egiten baitira analisi egokia lortu ahal izateko. Zehazte aldera, lan horietan *erlaxatu* dituzten erregelak sintagma barruan determinatzailearen eta izenaren arteko genero- eta numero-komunztadura lantzen dituztenak dira. Katalanerako (Badia *et al.*, 2004) ere *erlaxazioa* erabili dute fenomeno bera, zatiko komunztadura, aztertze-ko. Suedierarako egin den lanean (Arppe, 2000), *erlaxazioa* erabili dela esaten da, landutako fenomeno linguistikoa zehaztu gabe.
- Euskarari dagokionez, ez dugu desanbiguazio-gramatika *erlaxatzeko* beharrik izan, berez gramatika *erlaxatua* baitago eta edozein testu analizatzeko gai baita, hau zuzena ala erroreduna izan. Har dezagun ***Zenbait gizonak etorri dira** esaldi erroreduna. Bertan, ***Zenbait gizonak** sintagma okerra da, **zenbait** zenbatzaile zehaztugabeak eta **gizonak** izen mugatuak ez baitute komunztatzen. Hala ere, testua analizatu eta desanbiguatu ondoren, IV.8 irudiko analisia lortzeko gai izan da EUSTAGGER inolako *erlaxaziorik* gabe (formatu sinplifikatuan). Hots, mugatasuna ez da kontuan izan desanbiguazioan¹⁴.
- IXATI zatitzaileak sintagmak mugatzeko funtzio sintaktikoak eta horien noranzkoak erabiltzen dituela esan dugu. Ez du numero-mugatasun komunztadurarik gertatzen ote den konprobatzen. Euskaraz, sintagmetan bestelako komunztadurarik eskatzen ez denez, desanbiguazio-gramatika *erlaxatua* dagoela kontsideratuko dugu. Erlaxazioak erroreak

¹³etzean: alfer-lurra, basamortua, lur antzua.

¹⁴Erlaxazio honek kalte egingo luke, adibidez, **Zenbait gizonak ikustera etorri da** esaldi zuzenaren analisisian **Zenbait gizonak** sintagma onartuko bailuke.

```

"<Zenbait >"<HAS_MAI >"
  "zenbait" DET DZG MG ZERO HAS_MAI @ID> %SIH
"<gizonak >"
  "gizon" IZE ARR ABS NUMP MUGM @OBJ %SIB
  "gizon" IZE ARR ABS NUMP MUGM @PRED %SIB
  "gizon" IZE ARR ABS NUMP MUGM @SUBJ %SIB

```

IV.8 Irudia: *Zenbait gizonak sintagma erroredunaren analisia.

detektatzea ahalbidetzen duen aldi berean, anbiguotasun handiagoa uzten du.

- Analisi-kateak erroreekin izan dezakeen jokabidearen inguruko hausnarketa egin dugunean, ez dugu ondorio garbirik atera funtzio sintaktikoen desanbiguazioari dagokionez. Gauza bera gertatu zaigu mendekotasun-erlazioak ipintzen dituen EDGK gramatikarekin (*Maltparserrek* ez du funtzio sintaktikoen beharrik dependentzia-erlazioak ipintzeko). Hori horrela izanik, hauen erlaxazioaren inguruko probak egitea otu zaigu. Esperimentu hauek VII. kapituluan deskribatuko ditugu.

Fenomeno linguistikoak

Fenomeno linguistikoei dagokienez, zenbait fenomeno aztertzeko analisi-katean egokitzapenak beharrezkoak lirakeela uste dugu, eta beste hainbaten azterketarako, aldiz, ez. Adibidez, sintagma barneko komunztadura aztertzeke ez dugula inolako *erlaxazioaren* beharrik ikusi dugu. Baina hitz-mailako zenbait errore detektatzeko EDBLn informazio osagarria gehitu beharko litzatekeela ere aipatu dugu (ikus *etzean* hitzarekin ipini dugun adibidea). Modu okerrean erabil daitezkeen egitura linguistikoak ia kontaezinak direnez, fenomeno bakoitzerako zein egokitzapen den beharrezkoa aurreikustea ezinezkoa da. Hori dela eta, egokitzapenen azterketa tesi-lan honetan aztertuko ditugun egituretara mugatuko dugu.

IV.2 Errepresentazioa: XML bidezko anotazio-amarauna

Analisi-kateko informazio linguistikoa errepresentatzeko, tesi-lan honetan IXA taldean definitutako XML bidezko anotazio-amarauna erabili dugu. Atal honetan amaraunak zertan datzan ikusiko dugu.

XML (*eXtensible Markup Language*, markatzeko lengoaia hedagarriaren ingelesezko siglen bidez adierazia), *World Wide Web Consortium (W3C)* erakundeak sortutako etiketa-lengoaia hedagarria da. SGML (*Standard Generalized Markup Language*) lengoiaren sinplifikazio eta egokitzapen bat da eta lengoaia zehatzen gramatikak definitzeko erabiltzen da. Horretaz gain, XML plataforma ezberdinen artean informazio egituratuaren elkar trukatetarako estandarra da eta datu-baseetan, testu-editoreetan, eta ia edozein ataza edo tresnetan erabil daiteke.

Teknologia sinplea izan arren, XMLren inguruan sortutako teknologiek tresna boteretsu eta aukera askotarikoa bihurtu dute. Gaur egun, paper garrantzitsua du XMLk informazioa konpartitzeko modu erraza, segurua eta fidagarria baita.

XMLn alde zehatzetik definitu den formatu bati buruzko oinarritzko definizio guztiak betetzen dituzten, eta beraz, edozein XML analizatzailek analizagarriak diren dokumentuei *ongi osatutako dokumentuak* deitzen zaie. Dokumentu *baliagarriak*, ongi osatuak egoteaz gain, kanpo-dokumentu batek (*DTDak*¹⁵ edo *XSchemak*¹⁶) definitutako egitura bete egin behar dute.

Estandarizazioaren bideari jarraituz, eta tresnetako sarrera/irteerak *ad hoc* sortuak izan daitezkeen ekidite aldera, IXA taldean hizkuntzaren tratamendurako tresnak integratzeko proposamen bat landu dugu (Artola *et al.*, 2000). Proposamen hori analisi-kateko hizkuntzaren prozesamendurako zenbait tresnek sortutako hizkuntza-etiketatzeko dokumentuak sortzeko, arakatzeko eta editatzeko balio duen ingurune batean gauzatu dugu. Helburu nagusia dokumentu etiketatuak kontsultatu, bistaratu eta aldatzeko ingurune malgu eta hedagarria sortzea da, beti ere errepresentaziorako eskema koherente eta orokorra jarraituz (Artola *et al.*, 2004a).

Tresnen artean hizkuntza-informazioa trukatu ahal izateko ongi formalizatutako proposamena nahi dugunez, TEI-P4ko¹⁷ gidalerroak eta bertan definitutako *ezaugarri-egitura motatuak* (Sperberg-McQueen eta Burnard, 2002) erabili ditugu. Ezaugarri-egiturak aukeratzeko arrazoiak hauek izan dira besteak beste: errepresentazio-ahalmen handia dute, hedagarritasunerako erraztasunak ematen dituzte eta landu beharreko informazio linguistikoaren konplexutasunarekin lan egiteko gai dira.

¹⁵Document Type Definition: XML edo SGML dokumentu baten egitura murriztapenak eta, beraz, dokumentua bera definitzen duen egitura. Dokumentuaren barnean onar daitezkeen elementuak, beraien atributuak eta aukeran atributu horien balio posibleak definitzen ditu DTDak.

¹⁶Schemak DTDen ordezkioak dira, hauen antza dute egituren definiziorako balio baitute baina, gainera, XML estandarra betetzen dute.

¹⁷<http://www.tei-c.org/P4X/DTD/>

IXA taldean markaketarako *eskema banatua* deritzona erabiltzea erabaki dugu. Eskema banatua edozein informazio linguistiko edo egitura adierazteko gai da elementu gutxi batzuk erabilita. Eredu banatuan (ingelesez *stand-off*, *sparse* edo *distributed annotation*) informazio linguistikoa sarrerako testutik aparteko dokumentuetan jasotzen da, eta testuaren eta analisiak gordetzen dituzten dokumentuen artean estekak ezartzen dira.

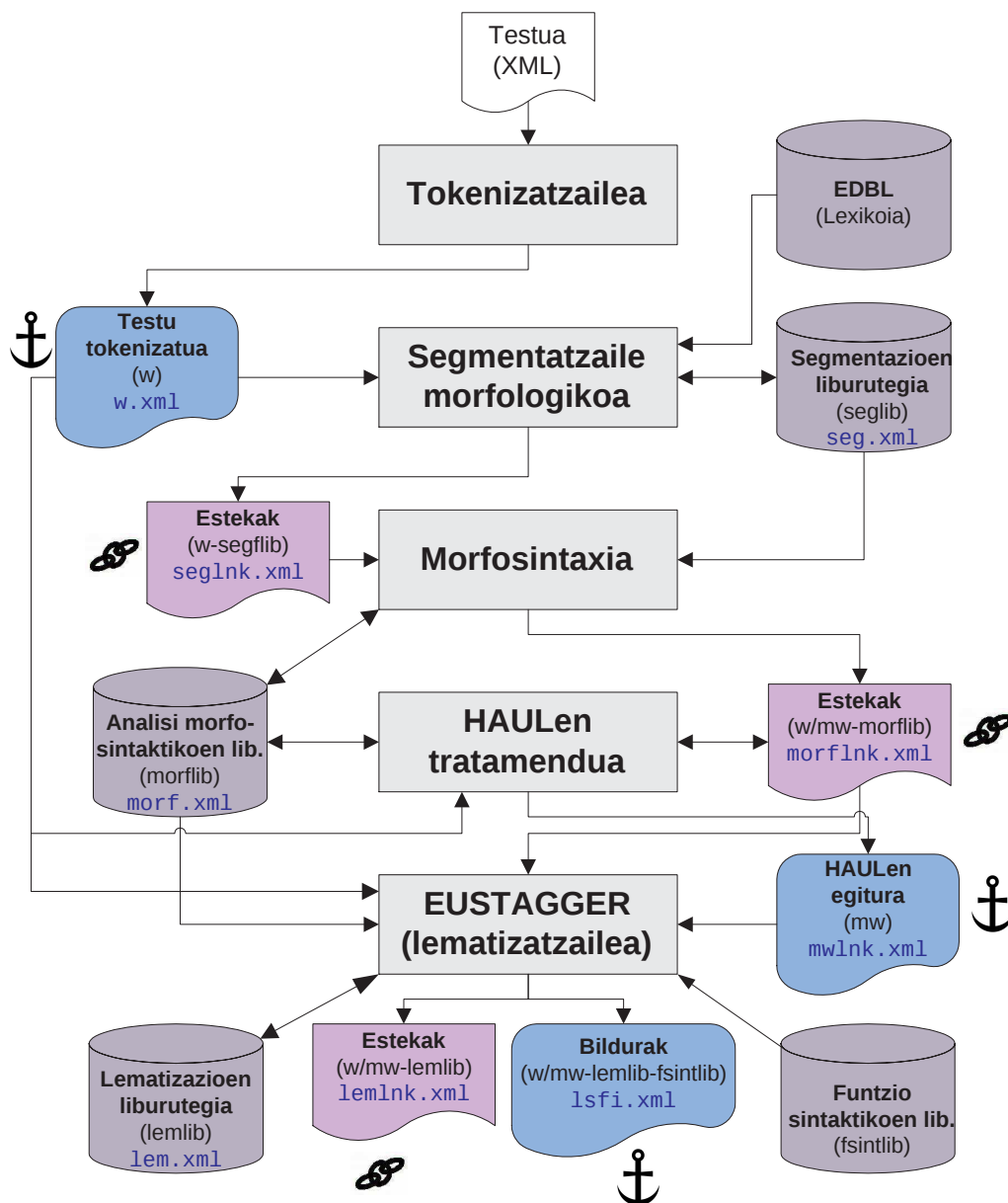
Analisi-katean azaldu dugun tresna bakoitzaren irteera XML dokumentu ezberdinekin osatuta dagoenez, *anotazio-amaraun* batekin lanean ari garela esan dezakegu. Amaraunean dokumentu mota hauek topa ditzakegu:

- **Aingurak:** elementu hauek sarrerako testuetan topa daitezkeen elementu fisikoak (*offset*¹⁸ adierazpenak edo *xptr*¹⁹ adierazpenak), aurreko anotazio-prozesuen ondorio diren elementuak (morfemak, unitate bakarreko eta unitate anitzeko tokenak, kateak...) eta azken hauen interpretazio linguistikoak dira. *Testu-aingurak* tokenizatu ondoren edo hitz anitzeko unitateak identifikatu ondoren lortutako elementuak dira. *Interpretazio-aingurak*, berriz, anotazio-elementuak edo aingura bereziak dira (informazio linguistikoa gehitu ahal izateko sortutako elementuak). Aingura bereziek aurreko prozesuetan sortutako interpretazio-identifikadoreen arteko bildura (*join*²⁰ TEIn) moduan lan egiten dute. Hauen artean hitz sekuentziak, sintagmak eta aditz-kateak, aingura berezi moduan sailka ditzakegu. Egiturazko anbiguotasuna teilakatzen diren ainguren bidez adierazten da.
- **Anotazio-estekak:** aingurak eta dagozkien analisiak lotzen dituzte anotazio-estekak. Anbiguotasuna esteka hauen bidez adierazten da, aingura bati hainbat interpretazio-analisi lotzea posiblea baita. Desanbiguaziorako, hauetako esteka bat zuzen moduan markatzea eta gainontzekoak oker moduan markatzea nahikoa da.
- **Informazio linguistikoa:** ezaugarri-egitura motatuak erabiltzen dira analisi-prozesuaren ondorioz sortutako mota ezberdinetako informazio linguistikoa biltegiratzeko. Adibidez, segmentazio morfologikoaren edo lematizazioaren emaitza modu honetan gordetzen da.

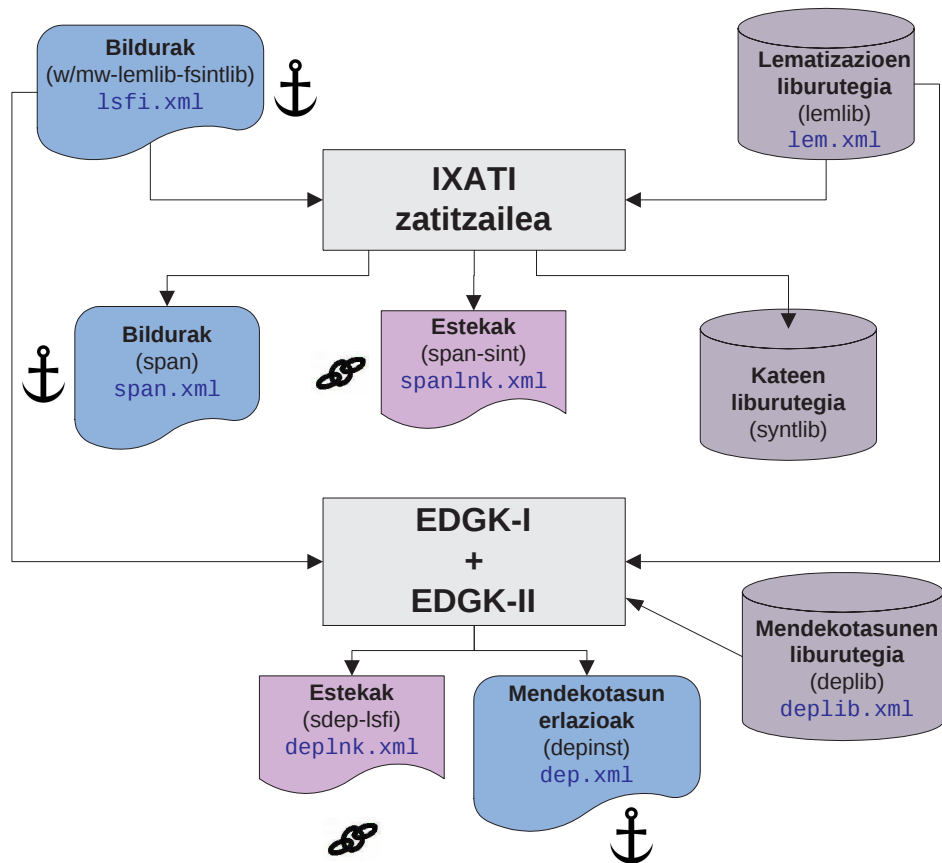
¹⁸Offset: denborazko edo lekuzko adierazpen erlatibo baten parte da. Adierazpeneko bi daten, edo bi denboren edo bi lekuren arteko desplazamendua eta noranzkoa adierazten ditu. IXA taldean, tokenen arteko karaktere-desplazamendua adierazteko erabiltzen da.

¹⁹xptr edo erakusle hedatua: uneko dokumentu bateko edo kanpoko dokumentu bateko kokaleku batera erakusle bat definitzen du.

²⁰Join: zatitua egon daitekeen testu-zatia adierazten du, bera osatzen duten osagaiei (jarraituak edo ez-jarraituak) erreferentzia eginez.



IV.10 Irudia: Dokumentu anitzeko anotazio-amarauna (1/2).



IV.11 Irudia: Dokumentu anitzeko anotazio-amarauna (2/2).

tema, testu bildumen XMLzko prozesamendurako tresna, erabili da. Gaur egungo LibiXaML bertsioak Unix-en lan egiten du eta Interneten dago eskuragarri <http://ixa.si.ehu.es/ixa/resources/libixaml/> helbidean.

**TESTUINGURU MUGATUKO
ERROREEN DETEKZIOA ETA
ZUZENKETA**

V. KAPITULUA

Patroien erabilera datetako eta postposizio-lokuzioetako erroreak detektatzeko eta zuzentzeko

V.1 Sarrera

Kapitulu honetan testuinguru sintaktiko lokalari erreparatuz detekta daitezkeen erroreak aztertuko ditugu. Zehazkiago, gehienez bospasei hitzeko leihoetan gertatzen direnak. Horiek, gogora dezagun, Atwell-en arabera (1987) ingelesezko testuetan eskuz detektatu dituzten errorearen % 55 dira. Hitz-leiho handiagoetan gertatzen diren erroreak aurrerago aztertuko ditugu VII. kapituluan.

Errore lokalen azterketarako aukeratu ditugun hizkuntza-egiturak *datak* eta *postposizio-lokuzioak* dira. Datek zein postposizioek, morfologia- eta sintaxi-mailako fenomeno aberatsak dituzte eta uste dugu testuinguru mugatuak aztertuz detekta daitezkeen errorearen adierazgarri kontsidera daitezkeela. Gainera, adituen arabera (ikus II. kapitulua), datetan maiztasun handiaz egiten ditugu erroreak. Postposizio-lokuzioen kasuan, egituraren izaera anbiguoak sortu du interesa guregan. Bi fenomeno hauek aukeratzekoan, gure ustez hauen artean dauden alde nabariak ere kontuan hartu ditugu:

- **Datak.** Egitura hauek *mugatutzat* joko ditugu orokorrean honelako segida zurrana osatzen baitute:
[leku-izena,]¹ urtea hilabetea eguna

¹[] ikurrekin adierazi nahi dugu datetan leku-izena aukerazkoa dela.

Egitura honetan kokatzen den esaldia da adibidez, [Hondarribian, 2007ko abenduaren 21ean](#). Datak *ez-anbiguoak* dira: urte baten ondoren hila-betea topatzen badugu, eta gero eguna, ia beti data izango da; eta ohikoa da *egitura berean bizpahiru errore* topatzea.

- **Postposizio-lokuzioak.** Ondorengo orrialdeetan definituko ditugu postposizio-lokuzioak. Horien adibidea da [jende-Ø arteAN sartu da](#) esaldiko “-Ø arteAN” egitura. Egitura honetan, postposizio-lokuzioekin lotutako *anbiguotasuna*, bai gramatikala bai semantikoa, aurreikus dezakegu, egiturak ez baitu zertan postposizio izan (adibidez, [hainbat lagun-Ø arteAN iritsi gabeak ziren](#) esaldian ez da postposizio); zentzu honetan *ez da mugatua*, eta gainera, datetan ez bezala, *egitura berean errore bakarra* gertatu ohi da.

Data-erroreak detektatu eta zuzentzeko, *Xerox Finite State Tool (XFST)* tresna erabili dugu (Karttunen *et al.*, 1997; Ait-Mokhtar eta Chanod, 1997) eta postposizio-lokuzioetan gertatzen diren erroreak detektatzeko, *Murritzapen Gramatika (MG)* (Karlsson *et al.*, 1995). Ez dugu arrazoi berezirik izan hizkuntza-egitura bakoitzari formalismo zehatz bat esleitzeko. Gure ustez, era berean erabil zitekeen MG datak lantzeko, eta XFST postposizio-lokuzioak aztertzeko. Azken finean, islapen-erregelak erabili ditugu erroreak etiketekin inguratzeko. Tresna desberdinak esperimintatzearen probatu ditugu, eta bide batez, XFSTk eta MGk erroreak detektatzeko eta zuzentzeko eskaintzen dituzten erraztasunak eta dituzten zailtasunak ezagutu ditugu.

Bi tresna hauek komunean duten ezaugarria da oinarrian egoera finituko makinak erabiltzen dituztela. Beesley eta Karttunen-en arabera (2003), “Compared to the alternatives, applications based on finite-state networks are usually smaller, faster, more elegant, and easier to maintain and modify”. Txikiagoak, azkarragoak, dotoreagoak eta aldatzeko eta mantentzeko errazagoak izateaz gain, liburu berean honako ontasunak gehitzen zaizkie egoera finituko mekanismoei:

- Egoera finituko sareen ezaugarri matematiko sinpleak erraz ulertzen dira eta honek sareak kudeatu eta konbinatzeko erraztasunak ematen dizkigu; baita malgutasuna ere.
- Konputazio aldetik eraginkorrak dira, adibidez, analisi morfologikoaren moduko atazetan. Prozesamendu-abiadura onak lortzen dira.
- Kasu gehienetan, egoera finituko sareetan informazio kopuru handia memoria txikian gorde daiteke, sareak konprimitu baitaitezke teknologia desberdinak erabilita.

Garapenaren ikuspuntutik, egoera finituko programazioa erazagutzailea da. Erabiltzaileek helburu-hizkuntzari buruzko informazioa kodetzen dute, algoritmoen kezkatu gabe.

Egoera finituko mekanismoak erabiltzearen ezaugarri komuna alde batera utzi eta MG eta XFST tresnak zertan ezberdintzen diren aipatuko dugu: Murritzapen Gramatika “egoera finituko murritzapen-gramatika” moduan definitzen du Tapanainen-ek (1999). Horrek, hitz hauekin definitzen ditu horrelako tresnak: “A finite-state constraint language is a language for specifying a disambiguation machine that discards alternative readings of syntactic elements by contextual tests that are expressed in terms of regular languages”. MG erabiltzen dugunean, ezin ditugu zuzenean automatak kudeatu. Erregelak idazten ditugu eta horiek tresnak berak egoera finituko makina bihurtzen ditu. Horretan da desberdina XFST. Aurreko aukeraz gain, geroago V.2.1 atalean ikusiko dugun moduan, automatak zuzenean kudea ditzakegu, adierazpen erregularren eta erlazio erregularren osagarriak kalkulatz, kateamenduak eta konposizioak eginez e.a. Ikuspuntu horretatik, XFST tresna MG baino boteretsuagoa dela iruditzen zaigu.

Kapitulu honetan azalpenak honela emango ditugu: V.2 atalean datetan gertatzen diren erroreak detektatzeko eta zuzentzeko XFSTren erabilera aztertuko dugu. V.3 atalean, berriz, bost postposizio-lokuziotako errorearen azterketa sakona egingo dugu MG tresna baliatuz.

V.2 Transduktoreak eta datak

XFST erabili dugu datetan gertatzen diren erroreak detektatzeko eta zuzentzeko (Gojenola eta Oronoz, 2000; Díaz de Ilarraza *et al.*, 2007b). Tresna boteretsu hori erabilita, daten analisi morfosintaktikoari aplikatuko zaizkion errore-patroi linguistiko konplexuak definitu ditugu. Patroi horien bidez, erroreak detektatu ditugu, geroago, beste zuzenketa-/sorkuntza-patroi batzuen bidez, data okerre dagozkien ordain zuzenak sortzeko.

Lehenago ere erabili izan da XFST errorearen detekzioarako. Hashemi *et al.* (2003) lanean (ikus III. kapitulua), tresna hau baliatuta sortutako transduktoreen arteko kenketa egiten da. Kenketa egiten da egitura linguistiko zuzenak eta okerrak deskribatzen dituen *gramatika zabal* bat konpilatuz lortutako transduktorearen eta egitura zuzenak soilik deskribatzen dituen *gramatika estuaren* bidez lortutako transduktoreen artean. Kenduran lortutako automatek egitura okerrak detektatzeko gaitasuna dute.

Erroreei dagozkien ordain zuzenen sorkuntzarako egin dugun lanaren

antzekoa deskribatzen du Karttunenek (2006): suomieraz “hizkiz” idatzitako zenbaki deklinatuak, “zenbakietara” itzultzen ditu, eta alderantziz.

Atal hau modu honetan antolatuko dugu: XFSTren oinarriak V.2.1 atalean azaldu ondoren, datetan gertatzen diren erroreen sailkapena egingo dugu V.2.2 atalean. Sistemaren ikuspegi orokorra V.2.3 atalean emango dugu; detekzioan eta zuzenketan egindako lanak deskribatuko ditugu V.2.3.1 eta V.2.3.2 ataletan, hurrenez hurren. Bukatzeko, zenbait emaitza erakutsiko ditugu V.2.4 atalean.

V.2.1 Erabilitako tresna: XFST

Dagoeneko esan genuen III. kapituluan *Xerox Finite State Tool (XFST)* tresnak (Karttunen *et al.*, 1997; Ait-Mokhtar eta Chanod, 1997) sarreratzat adierazpen erregularrak² jasotzen dituela eta hauek transduktore bihurtzen dituela. Zehatzagoak izateko, XFST erabilia, egoera finituko automata sinpleak edo egoera finituko transduktoreak defini daitezke.

Automata sinpleak eredu abstraktu ezagutzaileak dira (Chomsky, 1956): adierazpen erregularrak erabiliz definitutako egiturak testu analizatuan eza-gutuz gero, onartu egiten dituzte eta egitura betetzen ez duten analisiak, baztertu. Lengoaia erregularrak ezagutzen dituzte.

Transduktoreak sarrera eta irteera duten egoera finituko ereduak dira. Bi ikur multzo “parekatzen” dituzte, edo modu sinpleago batean ikusita, karaktere-kate bat irakurtzen dute eta beste bat sortzen dute. Adibidez, **a:b** kateen bikotean, **a** katea irakurriko litzateke, eta **b** katea sortu (alderantzizko prozesua ere zilegi da).

XFSTk egoera finituko kalkulua ahalbidetzen duen eragiketa multzo aberatsa du. Gojenolaren arabera (2000) “kalkulu honen ahalmen espresiboari esker XFST tresnan erregela multzoak mantentzeko eta aldatzeko erraztasunak emango dira, lengoaiaren tratamendurako aplikazioen garapena azkartuz”. Aipatutako eragiketak garatzeko, XFSTn eragile anitz erabil daiteke:

- Oinarrizko eragiketak: bildura ($A \mid B$), kateamendua ($A \bullet B$), itxidura (A^*) eta itxidura positiboa (A^+), aukerazko elementua ((A)) edo barruan egotea ($\$A$).

²Adierazpen erregular bat, notazio algebraiko bat da karaktere-kate multzo bat definitzeko erabiltzen dena. Beraz, lengoaia erregularrak adierazteko erabil daiteke, baita bilaketa-patroiak definitzeko ere. Transduktoreekin lanean ari garenean, “adierazpen erregular” terminoa baino egokiagoa da “erlazio erregular” erabiltzea, transduktoreek karaktere-kate multzoen arteko erlazioak ezartzen baitituzte. Hala ere, orokortzearen batzuetan lehen terminoa erabiliko dugu.

- Lengoaia erregularrekin soilik erabil daitezkeenak: osagarria ($\sim A$), eba-
kidura ($A \ \& \ B$) eta kenketa ($A - B$).
- XFSTn definitutako eragile konplexuak:
 - Murritzapena: $A \Rightarrow B _ C$. A bakarrik onartuko da ezkerretik B eta eskuinetik C duenean. Mota honetako adierazpenak debekuak adierazteko erabiltzen dira.
 - Ordezpena: $A \rightarrow B \parallel L _ R$. Erlazio bat sortzen du. L A R testuingurua duen katea L B R katearekin erlazionatu egiten da, eta A irakurri eta B sortzen du.
 - Parekatze luzeenaren ordezpena: $A @ \rightarrow B \parallel L _ R$. Lehengoaren antzekoa baina A parekatzeko aukera asko daudenean, eta bata bestearen azpikatea denean, beti hartuko du luzeena.
 - Markaketa. $A \rightarrow L \dots R$ eta $A @ \rightarrow L \dots R$. Ordezpenen antzeko eragiketak. A lengoaiako edozein instantzia irakurri eta A bera sortzen du baina L eta R kateekin inguratuta. Ezker eta eskuin testuinguruak zehatz daitezke.
 - Konposizioa: $[A .o. B]$. Bi erlazio hartuta, erlazio berri bat eman-
go du, Aren goiko lengoaia Bren beheko lengoaiarekin erlaziona-
tuz. Adibidez, $[[a|b]:c .o. c:d]$ konposizioaren emaitza $[[a|b]:d]$ da,
beste modu batera esanda, $a:d$ eta $b:d$ bikoteen multzoa.
 - Konposizio biguna edo malgua: $[A .O. B]$. Konposizio arrunta
desanbiguazio-lanetarako erabiltzen denean, murritzapenak kon-
posatu egiten dira hasierako kate anbiguoarekin. Kasuren batean
murritzapenen batek analisi guztiak ezabatuko balitu, emaitza
hutsa litzateke, eta hori ez da onargarria. Konposizio bigun honen
helburua arazo hori konpontzea da, kate hutsa lortuko litzatekeen
kasuetan, hasierako esaldia “berreskuratu” egiten baitu.

Eragile hauetako asko baliatuko ditugu datetan gertatzen diren erro-
reen detekziorako erregeletan baina, azken helburua erroreak automatikoki
etiketatzea dugunez, markaketarako eragilea izango da azken urratsean era-
biliko duguna. Aipatzekoa da eragile honek egitura erroredunaren hasiera
eta amaiera adierazten dituzten etiketak aldi berean ipintzen dituela. Era
honetan, Murritzapen Gramatikak duen arazo bat ekiditen da: amaierako
etiketarik ez duten hasierako etiketak (eta alderantziz) idaztea.

V.2.2 Datetan gertatzen diren erroreen sailkapena

Harluxet hiztegi entziklopedikoan, data honela definitzen da: *gertaera bat jatzotzen den eguna*. Definizioaren arabera, [Martxoaren 27a](#), [asteazkena](#), [1960ko urtarrila](#) eta [2007ko irailaren 7a](#) datak dira. Atal honetan *data osoak* deitu ditugunak landuko ditugu, hau da, urtea, hilabetea eta eguna dituzten datak (urtearen aurretik leku-izen bat eduki dezakete aukeran). Elementuren bat edo gehiago faltan dituzten egituretako erroreak ez ditugu landuko.

Daten idazkera ongi araututa dago euskaraz. Euskaltzaindiak araua eman du egitura honen inguruan (*37. araua: Data nola adierazi*) eta zenbait gramatika eta estilo-liburutan ere, argi adierazia dago datak idazteko modua (Zubiri eta Zubiri, 1995; Alberdi eta Sarasola, 2001). Hala ere, eta II. kapituluko II.4.3 atalean ikusi genuen moduan, oso ohikoak dira erroreak datetan. Aipatutako arauaren arabera, honakoak dira datak adierazteko forma hedatuenak:

«Data adierazteko formarik hedatuenak eta erabilienak honako hauek dira:

Durango(n), 1983ko martxoaren 7a(n)

Durango(n), 1983ko martxoak 7.

Dena den, bi horien artean ere, lehena da orain arte gehien erabili dena, bai *-n* kasu-marka leku izenari eta egunari ezarriaz, bai kasu-markarik gabe. Hori da, beraz, batez ere Euskaltzaindiak gomendatzen duena:

Durango, 1983ko martxoaren 7an

Durango, 1983ko martxoaren 7a.»

(Euskaltzaindia, *37. araua: Data nola adierazi*)

Egitura zuzen hauetatik aldentzen diren egitura oker anitz topatu dugu datekin osatutako corpus erroredun bat eskuz aztertu dugunean. Ohikoenak diren bederatzi errore motak aukeratu eta landu ditugu (ikus V.1 taula). Errore horietako batzuk zuzen-zuzenean lotzen zaizkie datei: datetan besterik ez dira gertatzen (0, 3, 4, 5 eta 6 erroreak). Beste zenbait, ordea, okerrak dira edozein testuingurutan (1, 2, 7³ eta 8 erroreak).

Arrazoiak aztertu ez ditugun arren, uste dugu zenbait erroreren (ikus V.1 taulan 4, 5, eta 6 erroreak) iturria hau dela: datetan agertzen diren

³Euskaltzaindiaren gomendioa da: “datari dagokion urtea adieraztean ez erabili *-garren* atzizkia (ez letraz eta ez puntuz). Salbuespena: urte-kopuruaren ondoren *urte* hitza idazten denean (eta orduan, nahitaez erabili *-garren* atzizkia).”

elementuek (urtea, hilabetea eta eguna) kasu-konbinazio zehatzetan deklinatuta agertu behar dute (adibidez, hilabetea genitibo kasuan deklinatuta azaltzen denean, egunak absolutibo kasuan deklinatuta egon behar du).

Erroreak errore-sailkapenari jarraituz kodetu ditugu. Gogora dezagun, ikerkuntza-taldean darabilgun erroreen sailkapenean (Aldabe *et al.*, 2005b), kategoria nagusiak irizpide linguistikoei jarraituta sailkatu ditugula, eta kategoria nagusi bakoitzaren azpian, irizpide deskriptiboa erabili dugula. Hau horrela izanik, datak *Nozioak* errore-kategoria nagusian sailkatu ditugu. Azpikategoria zehazteko erabili dugun kodeketa honakoa da:

KategoriaNagusia_Aldaketa_EzaugarriLinguistikoa_Non

Hau da osagai bakoitzaren azalpena:

- *Kategoria Nagusia*: erroreen sailkapeneko kategoria nagusia. Kasu honetan, DAT idatziko genuke, errorea datan gertatzen delako.
- *Aldaketa*: irizpide deskriptibotik zein errore mota den: elementuren bat ordenaz aldatu (OAL), gehitu (GEH), kendu (KEN) edo nahastu (NAH) ote den.
- *Ezaugarri Linguistikoa*: aldaketa jasan duen ezaugarria, linguistikoa (KAS, kasua, adibidez) edo ez (MAR, marra, adibidez).
- *Non*: zein tokitan gertatu den ezaugarriaren aldaketa. Daten kasurako, urtean (U), hilabetean (H) edo egunean (E). Komunztadura eza dagoen kasuetan posizio bikoteak idatziko ditugu. Adibidez, LE kodeak adieraziko luke Lekuaren eta Egunaren ezaugarriak ez datozela bat.

Adibidez, DAT_KEN_KAS_E kodeak [*2007ko abuztuaren 3](#) datan gertatzen den errorea deskribatuko luke. Hau da, DATan KENDu egin da KASua Egunean⁴. Datetako errore motak deskribatzeko erabiltzen dugun taulan (V.1 taula) erroreak modu honetan kodetu ditugu.

⁴Adibide honetan DAT_NAH_KAS_E kodea ere egokia litzateke, kasua *kendu* beharrean *nahastu* dela esan baitezakegu (eguna, [3](#), azaleko markarik gabeko absolutibo mugagabearen deklinatuta dagoela pentsa dezakegu).

Errore-kodea eta -deskribapena	Adibidea
0. DAT_NAH_KAS_LE Leku-izenak inesibo kasua izanik eguna adierazten duen zenbakiak inesibo kasurik ez izatea. Leku-izenak absolutibo kasua izanik, eguna adierazten duen zenbakiak absolutibo kasurik ez izatea.	*Donostia[n], 2007ko maiatzaren 27a[] *Donostia[], 2007ko maiatzaren 27a[n]
1. DAT_GEH_MAR_U Urtea adierazten duen zenbakia deklinatzeko marratxo erabiltzea.	*Donostian, 1995[-]eko maiatzaren 14an
2. DAT_NAH_MAI_H Hilabetea letra larriz adieraztea.	*1999ko [M]aiatzaren 2an
3. DAT_KEN_KOM_L Datetan aukerazko leku-izena erabiliz gero, ondoan komarik ez idaztea.	*Frantzia 1997ko [] maiatzaren 8an
4. DAT_KEN_KAS_E Hilabeteak genitibo kasua izanik, eguna adierazten duen zenbakiak absolutibo kasua izatea.	*Donostian, 1995eko maiatzaren 22[]
5. DAT_GEH_KAS_E Hilabeteak ergatibo kasua izanik eguna adierazten duen zenbakia deklinatuta egotea.	*1998.eko maiatzak 14[an]
6. DAT_NAH_KAS_H Hilabetea ez genitiboa ez absolutiboa ez den kasu batean deklinatuta egotea.	*Donostian, 1995eko maiatza[n] 14an
7. DAT_GEH_PUNT_U Urtea adierazten duen zenbakiaren ondoren -garren adierazten duen puntua idaztea. Salbuespena: urte-kopuruaren ondoren urte hitza idazten denean.	*Donostian 1997[.]eko maiatzan 28an
8. DAT_GEH_KAS_E 11 eta 31 zenbakiak absolutibo singular kasuan deklinatzea.	*1997-ko maiatzaren 31[a]

V.1 Taula: Datetan gertatzen diren errore mota ohikoenak (erroreak letra lodiz eta makoen artean markatu dira).

V.2.3 Sistemaren ikuspegi orokorra

Erroreen detekzio- eta zuzenketa-prozesurako sarrera, maila morfosintaktikoan analizatutako eta desanbiguatutako testua da. Irteeran, testu horretan gaizki erabilitako datek errore-etiketa pare bat izango dituzte errorea inguratzen, eta dagozkien ordain zuzenak sortuta egongo dira. Sistema (ikus V.1 irudia) egoera finituko bi transduktore multzorekin osatuta dago, bata erroreen detektziorako (ikus V.2.3.1 atala), eta bestea, berriz, data zuzenen sorkuntzarako (ikus V.2.3.2 atala). Testu analizatua prestatu behar da egoera finituko transduktore hauen sarrera izan dadin, eta horretarako, dokumentuen formatua aldatzen duten bi programa sortu ditugu.



V.1 Irudia: Sistemaren arkitektura nagusia.

Prozesua errazago uler dadin, adibide bat erabiliko dugu. Har dezagun ***1995eko maiatzaren 15** data okerra. Data ez-gramatikala da, genitibo kasuan deklinatutako hilabete baten ondoren, egunak ageriko kasu-markaren bat eraman behar duelako. Aipatutako data sarreraren hartuta, modulu hauek erabiliko dira data horretako erroreak detektatzeko eta zuzentzeko:

1. “XMLtik XFSTrako” iragazkia. Lehen urrats honetan, sarrerako XML fitxategiko informazioa (analisi morfosintaktiko desanbiguatua, *lem.xml* luzapeneko fitxategia), egoera finituko transduktoreen sarrerako for-

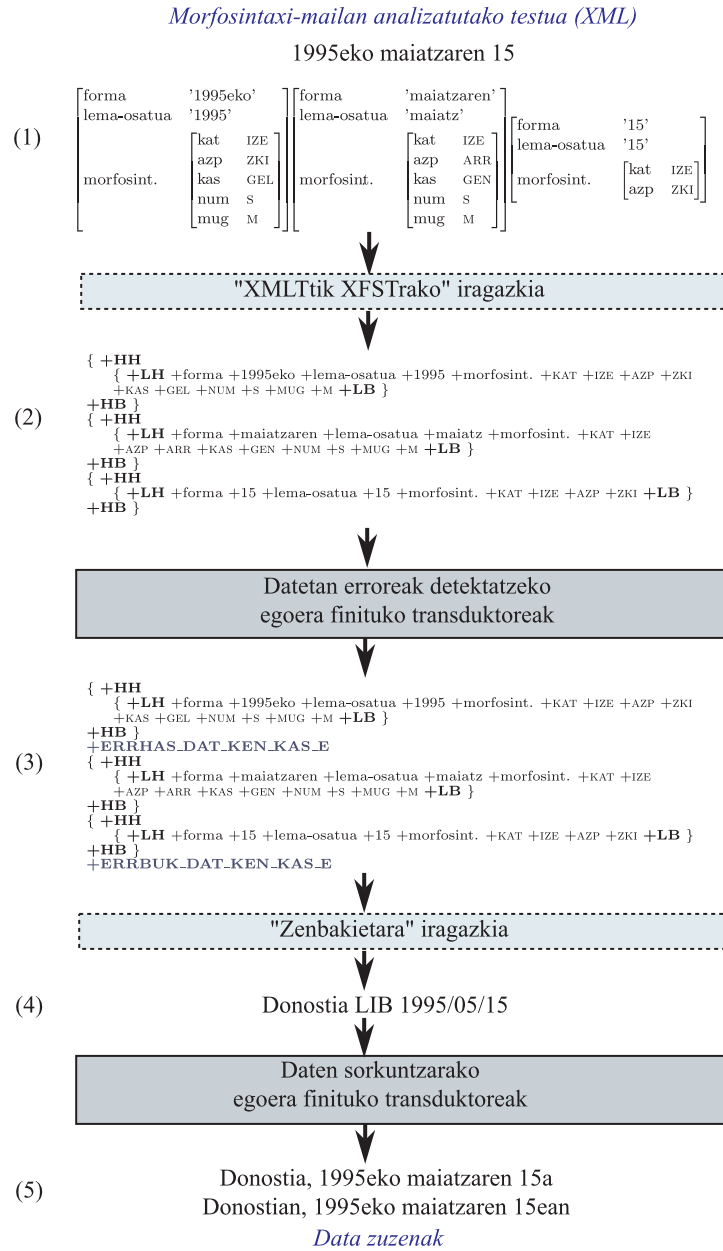
matura aldatzen da. Adibideko data erroredunaren analisi morfosintaktikoa XMLz biltegitratzeko erabiltzen den ezaugarri-egitura V.2 irudian “1” etiketarekin adierazia ikus dezakegu. Bertan, data osatzen duten elementuen kategoria gramatikala (kat), azpikategoria gramatikala (azp), kasua (kas), numeroa (num) eta mugatasuna (mug) adierazten dira⁵. Iragazki-programa hau pasa ondoren lortuko genukeen formatu sinplifikatua “2” etiketarekin erakusten da irudi berean⁶.

2. *Datetan erroreak detektatzeko egoera finituko transduktoreak*. Datetan gertatzen diren erroreak detektatu ahal izateko, egoera finituko bederatzi transduktore aplikatu dizkiogu era sekuentzialean sarrerako testuari (transduktore bat errore mota bakoitzeko). Transduktoreetako bakoitza aplikatu ondoren, hark detektatzen duen egitura linguistiko erroreduna etiketa pare batekin inguratuko da. Etiketek errore mota adierazteaz gain, errorearen hasiera (ERRHAS) eta bukaera (ERRBUK) markatzen dituzte. Transduktoreen aplikazioaren emaitza V.2 irudian “3” zenbakiarekin etiketatua ikus dezakegu. Transduktoreak aplikatzen goazen heinean, erroreak adierazteko erabiltzen diren etiketak gehitzen joango dira analisisian.
3. *“Zenbakietara” iragazkia*. Datetako erroreak automatikoki detektatu eta etiketatu ondoren (ohikoa da data berean errore bat baino gehiago topatzea), zuzenketa-prozesuari ekingo diogu. Zuzenketarako, (Otaegi, 2006) lanean sortutako transduktore multzoko azpimultzo bat erabiliko dugu. Lan horretan, euskara-ikasleei laguntza ematen zaie, digitu-segidak idatzita, dagozkien datak, orduak eta zenbakiak hizkiak erabiliz idatzita emanez⁷. Daten sorkuntzarako transduktore multzoa soilik erabili dugu data zuzenak sortzeko. *Zenbakietara* iragazkiak “urte/hilabete/egun” moduko digituzko adierazpenak lortzen ditu gutxienez errore batekin markatutako datetatik abiatuta. Dataren aurretik komaz jarraitutako leku-izen berezi bat balego, puntu honetan leku-izenaren lema eta komaren kokalekua adieraziko lukeen LIB etiketa genituzke (ikus V.2 irudian “4” etiketa).
4. *Daten sorkuntzarako egoera finituko transduktoreak*. Aurretik aipatu

⁵Analisisiko laburtzapenek honako esanahia dute: *Kategoria gramatikalak*: IZE = ize-na; *Azpikategoria gramatikalak*: ZKI = zenbakia, ARR = arrunta; *Kasuak*: GEL = lekuzko genitiboa, GEN = genitiboa; *Numeroa*: S = singularra; *Mugatasuna*: M = mugatua.

⁶Informazio morfosintaktikoari honako etiketak gehitu dizkiogu XFSTz definitutako adierazpen erregularren diseinua errazteko: HH (hitz-hasiera), HB (hitz-bukaera), LH (lematizazio-analisiaren hasiera) eta LB (lematizazio-analisiaren amaiera).

⁷<http://kantauri.eleka.net/neh> eta <http://sisx04.si.ehu.es/~iniebla001/idazlagun/>



V.2 Irudia: Sistemaren arkitektura nagusia eta tarteko emaitzak.

dugun moduan, datetan erroreak “zuzendu” ahal izateko, datak adierazten dituzten digitu-segidak testura aldatzen dituen egoera finituko transduktore multzoa baliatu dugu. Adibideko data erroredunarentzat sortutako ordain zuzenak V.2 irudian “5” zenbakiarekin adierazita ikus ditzakegu. Nahiz eta datak zuzen idazteko aukera gehiago egon, Euskaltzaindiak hobesten dituen formak sortzea erabaki dugu.

V.2.3.1 Detekzioa

Arestian esan dugun moduan, V.1 taulan deskribatutako bederatzi errore motak detektatu ahal izateko, XFST tresnak egoera finituko transduktore bihurtuko dituen bederatzi adierazpen erregular multzo definitu ditugu.

Erroreen deskribapenerako adierazpenak idazteko orduan esan dezakegu “mailaz-mailako erlaxazio” moduko bat erabili dugula: data batean gerta daitekeen errore bat definitzerakoan, gainontzeko erroreak aldi berean gerta daitezkeela kontuan hartu behar izan dugu. Datetan bizpahiru errore gertatzea ohikoa denez, errore-patroiak diseinatzerakoan egitura zuzena ez ezik, bere erabilera okerrak ere kontuan eduki behar ditugu. Adibidez, ondorengo lerroetan (ikus V.3 irudia), DAT_KEN_KAS_E erroreak definizioa egin dugu (hilabeteak genitibo kasu-marka daramanean egunak ezin du absolutibo kasuan egon). Errore-patroia definitu dugunean, data berean beste errore batzuk gerta daitezkeela kontuan izan dugu. Adibidez, hilabetea maiuskulaz idatzita egon daitekeela (ikus V.3 irudiko 5. lerroa) eta urtea marratxoa erabilita deklinatua egon daitekeela (ikus V.3 irudiko 7. lerroa).

DAT_KEN_KAS_E errorea bi urratsetan definitu dugu (ikus V.3 irudia). Lehenengo, hizkuntza-egitura erroreduna definitu dugu DAT_KEN_KAS_E patroia idatziz (1-10 definizioak): hilabeteak genitibo kasu-marka darama ongi edo gaizki (letra larriz) idatzita egonda ere (*HilaGen*, *HilaGenZuzena* edo *HilaGenMaiOkerra* da), eta jarraian, eguna adierazteko erabiltzen den zenbakia ez da deklinatuta agertzen (*ZenbakiEzDek*). Ondoren, ezkerretik eskuinera *parekatze luzeenaren ordezipena* eragilea (@->) erabiltzen da egitura erroreduna (“...” bidez adierazia) errore-etiketekin inguratze. Erregelaren aplikazioa mugatzeko, egitura erroredunaren ezker eta eskuin testuinguruak zehaz daitezke, erregelak datetan soilik aplikatzen direla ziurtatzeko. Kasu honetan, urtea da hilabetearen ezkerrean topatuko duguna. Urtea, ongi idatzita egon daiteke (*UrteaZuzena*), edo gaizki, marratxoa erabili delako lekuzko genitibo kasu-marka (-ko) gehitzeko (*UrteaOkerra*). Lehen esan dugun moduan, errorea deskribatzen duen patroiak alde aurretik beste zenbait errore gerta daitezkeela aurreikusten du.

```

1. define IzlgEkoEdoKo $[ "+kas" "+gel" ];
2. define Hila $[ "+lema-osatua" [ "+urtarril" | ["+otsail"|"Otsail"]
    ... | "+abendua" ] ];
3. define HilaGen Hila $[ "+kas" "+gen" "+num" "+s" ];
4. define HilaGenMai [ HilaGen & $[ "+has_mai" ] ];
5. define HilaGenMaiOkerra "+ERRHAS_DAT_NAH_MAI_H" HilaGenMai
    "+ERRBUK_DAT_NAH_MAI_H";
6. define UrteaZuzena [ Zenbakia & IzlgEkoEdoKo ];
7. define UrteaOkerra [ "+ERRHAS_DAT_GEH_MAR_U" ErroreaMarraUrtean
    "+ERRBUK_DAT_GEH_MAR_U";
8. define Urtea [ UrteaZuzena | UrteaOkerra ];
9. define HilaGen [ HilaGenZuzena | HilaGenMaiOkerra ];
10. define DAT_KEN_KAS_E HilaGen ZenbakiEzDek;
11. define MarkatuErrorea DAT_KEN_KAS_E [ ErroreaGenAbs ]
    @-> +ERRHAS_DAT_KEN_KAS_E ... +ERRBUK_DAT_KEN_KAS_E || Urtea _ ;

```

V.3 Irudia: Errore-patroia XFSTz definitzeko adierazpen erregularrak.

V.2.3.2 Zuzenketa

Aipatu dugun moduan (ikus V.2.3 atalean “Zenbakietara” iragazkia), date-tan gertatzen diren erroreak zuzentzeko, XFST erabiliz dagoeneko definituta zegoen transduktore multzo bat jaso, eta azpimultzo bat egokitu dugu. Egoera finituko transduktore hauek digitu-segidak jaso (urtea/hilabetea/eguna) eta hizki bidezko datak sortzen dituzte.

Euskaltzaindiak datak idazteko hobesten duen forman leku-izenaren eta egunaren kasu-markek bat etorri behar dute, edo absolutiboan, edo inesiboan. Beraz, egoera finituko transduktoreak erabiliko ditugu: i) datak absolutiboan komunztatzen sortzeko eta ii) datak inesibo kasuan komunztatzen sortzeko⁸. Beti bi ordain zuzen eman beharrean, absolutiboan zein inesiboan, kontuan hartzeko aukera litzateke, sorkuntza detektatu den erroreak arabera egitea. Adibidez, DAT_KEN_KAS_E errore motaren kasuan (mota honetako adibidea da, [Donostian, 2007ko irailaren 10](#), non eguna deklinatzea falta den) egokiagoa litzateke ordain zuzena inesiboan soilik ematea.

Aukerazkoa den leku-izena eta bere ondoko koma ez dira transduktoreak erabilita sortzen. Leku-izenaren lema testu analizatutik jasotzen da eta

⁸Gramatika honetan hilabete-zenbakiak dagoeneko deklinatuta dauden karaktere-ka-teetara itzultzen ditugu (adibidez, 3 zenbakia [martxoaren](#) bihurtzen da). Kontuan hartzeko beste aukera bat sorkuntza morfologikoa erabiltzea litzateke, eta hilabetea inesibo eta absolutibo kasuetan deklinatuta sortzea. Aukera hau beste tresna batekiko dependentziak ekiditeko baztertu dugu.

lexc (Beesley eta Karttunen, 2003) erabiliz dagokion kasuan deklinatzen da (absolutibo edo inesibo). Koma idatzi egiten da, lekua adierazten duen izen berezi bat existitzen dela (*LIB* etiketa V.2 irudian, 4. urratsean) egiaztatu eta gero.

```

1. define ItzuliUrtea [ "/"-> "ko" || - Zbkia Zbkia "/" ];
2. define Itzuli05Hila [ "0"5 "/" -> "maiatzaren " || "ko" - ];
3. define ItzuliHila [ Itzuli01Hila .o. Itzuli02Hila
    .o. ... .o. Itzuli12Hila ];
4. define ItzuliEguna [ [ .. ] -> "a" "n" || Zbkia - .#. ];
5. define Itzuli [ ItzuliUrtea .o. ItzuliHila .o.
    ItzuliEguna ];
6. define GehituEEgunean [ "a" "n" -> "e" ... ||
    [ [ "0" | 2 | 4 | 6 | 8 ] 1 | [ 1 | 3 | 5 | 7 | 9 ] "0" | 5 ] - ];
    ...
n-1. define Garbitu [ GehituEEgunean .o. GehituEUrtean ];
n. define DataSortu [ Itzuli .o. Garbitu ];

```

V.4 Irudia: Datak XFSTz sortzeko adierazpen erregularrak.

Jarraian, inesibo kasuan komunztatzen duten datak sortzeko erabilitako erlazio erregularrak azalduko ditugu. Sarrera “barra etzana” karakterearekin banatutako hiru zenbaki multzo dira. Zenbaki multzoek urtea, hilabetea eta eguna adierazten dituzte, hurrenez hurren. Lehen zenbaki multzoa (urtea) topatzen denean, lekuzko genitiboa adierazten duen “-ko” morfema eranstzen zaio (*ItzuliUrtea*, V.4 irudiko 1. erregela). Hilabeteak zenbakietatik karaktereetara itzultzeko, ordezen eragiketak erabiltzen dira. Ordezpenak, ezkerrean urtearen ondorengo “-ko” segida edukitzera (ikus, adibidez, V.4 irudiko 2. erregelan *Itzuli05Hila*) mugatzen dira (testuinguru hori izan ezean, ordezipena ez da gertatzen). Bukatzeko, inesibo singularra adierazten duen deklinabide-atzizkia (“-an”) egunari eranstzen zaio (*ItzuliEguna*, 4 zenbakidun definizioa). Islapen-erregeletan salbuespenak ere tratatzen dira ikus dezakegu: urtea edo eguna adierazten duten zenbakiak kontsonantez amaitzen direnean⁹ “-e” epentetikoa gehitzen zaio lekuzko genitiboari urtean (“-e” + “-ko” = “-eko”) eta inesiboari egunean (“-e” + “an” = “-ean”).

⁹V.4 irudiko 6. erregelan (*GehituEEgunean*) adibide bat dugu. Sarrerako zenbaki bikotea 1ez (baina ez *hamaikaz*, hots, 1, 21, 41...), *hamarrez* (10, 30, 50, 70 edo 90) edo 5ez amaitzen bada, -*ean* karaktere-katea gehituko zaio zenbakiari. Bestela, -*an*. Adibideak dira, 1ean, 2an, 3an, 4an, 5ean... Antzera definitu dugu *GehituEUrtean*.

Zuzenketak sortzeko modu hau erabilia (zenbakizko datetatik abiatuta) data batean gertatzen diren errore guztiak ez detektatu arren, guztiak zuzenduko direla bermatzen da. Adibidez, data batean 3 errore egongo balira, bakarria detektatzea nahikoa da guztiak zuzentzeko.

Daten zuzenketarako modulua OLHIA ingurune batean integratu nahiko balitz, segur aski egokiagoa litzateke errore bakoitzaren zuzenketa banan-banan egitea. Aproposa litzateke, ordea, gramatika-zuzentzaile batean.

V.2.4 Emaitzak

Ebaluaziorako, garapenerako eta probarako, 267 testudun corpora baliatu dugu. Testu-bilduman ikasleek idatzitako testuak eta *Euskaldunon Egunkari*ko testuak daude, guztira 500.000 hitz-forma. Bertatik, 658 esaldi erauzi ditugu automatikoki. Esaldien bilaketarako hilabete-izenak, urte-zenbakiak eta antzeko aztarnak erabili ditugu. Gako-hitz hauek beste egitura batzuekin konparatuta errazagoa egiten dute daten inguruko corpus berezitua bilatzea. Jasotako esaldi multzo horretan data zuzenak, data okerrak eta daten antzeko egiturak biltzen dira.

Behin eta berriro esan dugu, ebaluazioan garrantzi handia eman diogula alarma faltsurik ez sortzeari: sistemak errore moduan data zuzen eta daten antzekoak diren egiturak ez markatzeari.

Datuak bi multzotan banatu ditugu (ikus V.2 taula). Lehen multzoa garapenerako erabili dugu, eta bigarrena probarako. Data zuzenen proportzioa altuagoa da probarako corpusean garapenerako corpusean baino, modu honetan alarma faltsuak zehaztasun handiagoarekin probatuko ditugulako. Adierazgarria iruditzen zaigu probarako corpusean 2 errore dituzten egiturak (% 65,8), bakarria dutenak baino gehiago izatea (% 23,7).

Corpusa erroreekin etiketatu dugu, analizatu egin dugu, eta azkenik, erroreak detektatzeko egoera finituko transduktoreak pasa dizkiogu.

	Garapenerako corpora		Probarako corpora	
Elementu kopurua	411		247	
Data zuzenak	51		35	
Daten "antzeko" egiturak	263		173	
Data erroredunak	97		38	
Data erroredunak: errore 1	48	% 49,6	9	% 23,7
Data erroredunak: 2 errore	35	% 36,0	25	% 65,8
Data erroredunak: 3 errore	10	% 10,3	3	% 7,9
Data erroredunak: 4 errore	4	% 4,1	1	% 2,6

V.2 Taula: Garapenerako eta probetarako corpusak.

	Garapenerako corpusa		Probarako corpusa	
Elementu kopurua	411		247	
Erroreak	97		38	
Errore detektatu gabeak	4	% 4,1	3	% 7,9
Ongi detektatutako err.	93	% 95,9	35	% 92,1
Alarma faltsuak	2		4	
Alarma faltsuen tasa		% 0,48		% 1,61

V.3 Taula: Ebaluazioko emaitzak.

Ebaluazioaren emaitzak V.3 taulan ikus ditzakegu. Garapenerako corpusa errore-patroien garapen- eta birfintze-garaian azter daitekeenez, bigarren eta hirugarren zutabetako emaitzak sistemak egun lortzen dituen emaitzen goi-muga kontsidera ditzakegu. Garapenerako corpusean % 95,9ko estaldura¹⁰ eta % 97,8ko doitasuna¹¹ (93 errore ongi detektatuak/95 errore proposamen, hau da, 2 alarma faltsu) lortu ditugu.

Aldez aurretik ikusi gabeko 247 elementuko corpusean % 92,1eko estaldura lortzen da. Doitasunari dagokionez, sistemak ongi detektatzen ditu 39 proposamenetatik 35 errore (% 89,7). Alarma faltsuak probarako erabilitako elementu kopuruarekin zatitzen baditugu (4/247), estima dezakegu alarma faltsuen proportzioa testu errealean % 1,6 ingurukoa izango dela. Emaitzak onak diren arren, doitasuna hobetu nahi badugu datu kopuru handiagoarekin egin beharko genituzke probak.

Topatutako zenbait alarma faltsu eta beraien arrazoiak V.4 taulan ikus ditzakegu bakoitzaren zergatia azalduz:

- *Prozeduraren ondorio direnak*: alarma faltsu bat erabili dugun prozeduraren ondorioa da. Ebaluaziorako corpusean, datak eta antzekoak lerroz lerro zerrendatuta daude. Lerro bakoitzaren amaieran puntuazio-markarik ez dagoenez, eta analizatzaileak lerro-amaierak detektatzen ez dituenaz, datak lerro berean egongo bailiran analizatzen dira. Honen ondorioz, V.4 taulako lehen adibidean ikus dezakegun moduan, leku-izena hurrengo datarekin batera analizatzen da eta sistemak leku-izenaren ondoren koma falta dela adierazten du. Testu errealean mota honetako alarma faltsurik ez gertatzea aurreikusten dugu.
- *Desanbiguazio-akatsa*: taulako beste alarma faltsuan, *Primakov* hitz

¹⁰estaldura = ongi detektatutako erroreak/detektatu beharko lirakeen errore guztiak.

¹¹doitasuna = ongi detektatutako erroreak/(ongi detektatutako erroreak + alarma faltsuak).

ezezaguna leku-izen moduan analizatu delako detektatu da existitzen ez den errorea.

Zuzenketari lotutako emaitzei dagokienez, data bateko errore guztiak detektatu ez arren, guztiak zuzentzen direla esan dugu: nahikoa da data batean errore bat detektatzea bere ordain guztiz zuzenak sortzeko. Esan dezakegu zuzenketa % 100ean egiten dela ongi.

Adibidea	Errorearen arrazoia
1998ko abenduak 20. Bizkaiko → 1998ko abenduak 25.	Analisi morfologikoa egiten denean ez da lerro-amaiera detektatzen eta lehen lerroko <i>Bizkaiko</i> leku-izena bigarren lerroko datarekin batera analizatzen da. Kasu horretan koma faltako litzateke leku-izenaren ondoren.
Primakovek 1998ko irailaren 11n hartu zuen. . .	<i>Primakov</i> hitz ezezaguna leku-izentzat hartzen du.

V.4 Taula: Alarma faltsuak daten detekzioan.

Interesgarria iruditu zaigu aztertzea erroreen detekzioan anbiguitasun morfosintaktikoak izan dezakeen eragina. Horregatik, proba bat egin dugu erroreen detekzioa EUSTAGGERen desanbiguazio-mailaren arabera nola aldatzen den aztertzeko asmoz. Emaitzak V.5 taulan ikus ditzakegu. Sarrerako testua analizatzerakoan inolako desanbiguaziorik egin gabe eta testua lehen mailan (kategoria-mailako anbiguitasuna ebatzita) edo bigarren mailan (kategoria eta azpikategoria) desanbiguatuta, emaitzak ez direla aldatzen ikus dezakegu. Kasu-marka kontuan hartzen denean, ordea, (hirugarren mailan) emaitzek okerrera egiten dute. Detektatu ez diren erroreak beti dira mota berekoak: DAT_NAH_KAS_LE eta DAT_GEH_PUNT_U. Garapenerako corpusean 18 errore gutxiago detektatzen dira eta probarako corpusean 6 errore gutxiago. Errore gehien *-garren* adierazten duen puntuaren detekzioan gertatzen da (DAT_GEH_PUNT_U), daten definizioan urteak *lekuzko genitibo* kasu-marka edukitzea eskatzen baita eta analisi hau hirugarren mailan desanbiguatuta desagertu egiten da, errorea detektatzea galeraziz. Leku-izenaren eta egunaren kasu-markak berdin deklinatuta egon daitezen eskatzen duen errore-erregelak huts egiten du hirugarren desanbiguazio-mailan, leku-izenean absolutibo kasua daraman analisia desagertu egiten delako. Laugarren desanbiguazio-mailan (gogoratu, desanbiguazioa murriztapen gramatika soilik erabiliz egiten da) errore bat gutxiago detektatzen da. Alarma faltsuei da-

gokienez, ez da inolako aldaketarik egon desanbiguazio-maila aldatuta eta alarma faltsu kopurua mantendu egin da.

	Detektatutako erroreak	
	Garapenerako corpusa	Probarako corpusa
Desanbiguatu gabe	93	35
1.mailan desanb.	93	35
2.mailan desanb.	93	35
3.mailan desanb.	75	29
4.mailan desanb.	93	34

V.5 Taula: Anbiguotasunaren eragina datetako erroreak ezagutzerakoan.

Transduktoreak datetako erroreak detektatzeko eta zuzentzeko nola erabili ditugun azaldu dugu atal honetan. Ondoren, egindako esperimentuetan lortutako emaitzak eman ditugu (kasurik hoberenean % 95,9ko estaldura eta % 97,8ko doitasuna) eta alarma faltsuen zergatia azaldu dugu. Hurrengo atalean beste fenomeno linguistikoko bat, postposizio-lokuzioak, eta beste tresna bat, Murritzapen Gramatika, izango ditugu aztergai.

V.3 Murritzapen Gramatika eta postposizio-lokuzioak

Erroreen detektziorako maiz erabili da Murritzapen Gramatika (MG). Katalanerako (Badia *et al.*, 2004), suedierarako (Birn, 2000; Arppe, 2000) eta norvegierarako (Johannessen *et al.*, 2002) sistemetan erabili dute formalismo hau modu arrakastatsuan (ikus III.4 atala). MG erabiliz definitutako errore-patroiak morfosintaktikoki analizatutako testuei aplikatzen zaizkie. Analisisirako tresnak informazio zuzenarekin lan egiteko prestatuak egoten direnez, zenbait sistematan egokitzapenak, zehazkiago murritzapenen “erlazazioak” egin behar izan dituzte testu ez-gramatikalak analizatu ahal izateko (Johannessen *et al.*, 2002; Badia *et al.*, 2004).

Azterketa bibliografikoan ez dugu postposizio-lokuzioen fenomenoari zuzenean lotzen zaion lanik topatu. Chodorow *et al.* (2007) lanean, ordea, ingelesezko preposizioetan gertatzen diren erroreak detektatzeko ikasketa automatikoa erabiltzen duen sistema bat aurkeztu dute. Bai ingelesezko preposizioek, baita euskarazko postposizioek ere, lotura dute ezaugarri semantikoekin, baina gure ustez, euskarazko postposizioak konplexuagoak dira. Hasteko, bi hitzetan zehar banatzen dira, eta aldaera anitz edukitzeaz gain, elementuen arteko nolabaiteko komunztadura eskatzen dute.

Atal hau era honetan antolatuko dugu: lehenik, Murritzapen Gramatika-ren inguruko zenbait azalpen emango ditugu V.3.1 atalean. Postposizio-loku-

zio terminoarekin lotzen dugun kontzeptua definituko dugu V.3.2 atalean eta postposizio horien artean aztertuko ditugun bost postposizioak aukeratzeko prozesua deskribatuko dugu. Aukeratutako postposizio bakoitzaren analisi sakona egingo dugu V.3.3 atalean. Atal honetan, detektatu ditugun errore motak zehatz-mehatz deskribatuko ditugu eta lortutako emaitzak adierazi. Postposizioen zuzenketa nola egin dugun azaltzeko atala izango da V.3.4 zenbakia duena. Bukatzeko, lan hau guztia XUXENG euskararako gramatika-zuzentzailean integratzeko modua labur-labur azalduko dugu V.3.5 atalean.

V.3.1 Erabilitako tresna: CG2

Murritzapen Gramatikaren (Karlsson *et al.*, 1995; Voutilainen eta Tapanainen, 1993; Tapanainen, 1996) deskribapen labur bat egin dugu III. kapituluaren egoera finituko mekanismoak azaldu ditugunean. MG formalismoa implementatzen duen tresnaren 2. bertsioa erabili dugu (*Constraint Grammar 2, CG2*) atal honetan aurkeztutako lanerako.

Lau dira MGn erabilera ohikoena duten eragileak: bi lehenak islapen-erregelatan erabiltzen dira, eta azken biak, desanbiguazioan:

- MAP: islapen-eragile honek etiketa-sekuentzia bat *gehitzen* dio analisi-interpretazio bati, analisiak erregelan definitutako testuingurua betetzen badu. Interpretazio berean ezin dira bi etiketa-sekuentzia ipini.
- ADD: MAPen berdina baina nahi beste etiketa-segida *eslei* daitezke.
- REMOVE: helburu-etiketa duen interpretazioa *ezabatzen* du interpretazioak baldintza guztiak betetzen baditu. Token batean geratzen den azken interpretazioa ez du ezabatzen.
- SELECT: helburu-etiketa duen interpretazioa *aukeratzen* du, interpretazioak baldintza guztiak betetzen baditu.

Murritzapen Gramatika IXA taldean oso erabilera zabalekoa da eta ikerketa-lan ugariren oinarritzko tresna izan da (Arriola, 2000; Aduriz, 2000). Hau horrela izanik, formalismo honen inguruko informazio ugari topa deza-kegu, adibidez, Adurizen tesi-txostenean (2000) 19. orrialdetik aurrera.

Erroreak etiketekin markatzeko, MGk eskaintzen dituen islapen-erregelak erabiliko ditugu, MAP eragilea zehatzagoak izateko. Errore baten hasiera eta errorearen bukaera deskribatzeko gai bagara, murritzapen-gramatikak errorea mugatzeko eta, beraz, detektatzeko aukera ematen digu.

V.3.2 Postposizio-lokuzioak

V.3.2.1 Definizioa eta sailkapena

Postposizioak “perpausaren sintagmen arteko erlazio gramatikalak adierazten dituzten forma askeak direla” zehazten da EGLU-I gramatikan (Euskaltzaindia, 1985). IXA ikerketa-taldean postposizioen inguruan eginiko txostenean (Aduriz *et al.*, 2008), honakoa irakur dezakegu:

«Egungo ikuspegitik, bada, postposizioen artean bi mota nagusi bereiz daitezke: alde batetik lemari erantsirik ageri diren atzizkiak (10), eta bestetik, osagai nagusia elementu beregaina dutenak (11).

(10) **etxetik**

(11) **etxearen gainetik**»

Zabala eta Odriozolari (2004) jarraiki lehenengo multzoko postposizioei *atzizki-postposizioak* deituko diegu (*posposiciones sufixales*). Bigarren multzokoak, ostera, *postposizio-lokuzioak* direla esango dugu.»

Tesi lanean postposizio-lokuzioak landuko ditugu. Aduriz *et al.* (2008) lanean hau diote: “gramatikari tradizionalak zein modernoek postposizio (edo postposizio beregain) kategoria aitortzen diote askea den elementuari soilik. Aurreko sintagmak hartu ohi duen atzizki-postposizioa, ostera, postposizioaren osagarria dela aldarrikatzen dute. Erdaretako joerari erreparatuz, eta postposizio-lokuzioak ezagutzeko sistema errazte aldera, atzizki-postposizioak eta elementu beregainak duten multzo osoa unitate bakartzat hartuko dugu, elementu beregaina biluzik (**-en truke[]**, **-tik kanpo[]**) nahiz atzizki-postposizio batekin (**-i buru[z]**, **-en modu[ra]**) ageri. Beraz, hemendik aurrera, postposizio-lokuzioez ari garenean, sekuentzia osoari egingo diogu erreferentzia eta ez bakarrik elementu beregainari.”

Euskaltzaindiaren arabera:

«Postposizio diren neurrian kasu markaren bat hartzen dute (“nor” kasuari dagokion -Ø marka ere barne delarik). Kasu marka hau aldakaitza izan ohi da.

(12) **Elizaren bitartez**

(13) **Astearen buruan**

Ezin da esan *elizaren bitartean, *elizaren bitarteari edo *as-tearen buruaz. >>

(Euskaltzaindia (1985), 438. orrialdea).

Aipuaren amaieran aipatzen diren moduko erabilera okerrak dira aztertuko ditugunak. Erabilera oker horien deskribapena errazte aldera, nahiz eta unitate osoa postposizio-lokuzio kontsideratu, *postposizio-atzizki* (PA) eta *osagai beregain* (OB) terminoak zehaztuko ditugu tokenei erreferentzia egiteko. Postposizio-lokuzioak deskribatzen ditugunean V.6 taulako moduko eskemak erabiliko ditugu, non lehen elementua postposizio-atzizkia den (-aren¹²), bigarrena Osagai Beregainaren Lema (OBL) (*buru*), eta azkena, Osagai Beregainaren postposizio-atzizkia den (OBPA) (-an).

PA	OB	
PA	OBL	OBPA
Postposizio-Atzizkia	Osagai Beregainaren Lema	Osagai Beregainaren Postposizio-Atzizkia
aste-aren	buru	-an

V.6 Taula: Postposizio-lokuzioen osagaiak.

Erroreen deskribapenean hiru elementu hauek hartuko dute parte gehienetan, baina maiz, egitura gaizki erabili ote den jakiteko, postposizio-lokuzioaren inguruko hitzak eta hauen hizkuntza-ezaugarriak aztertu beharko ditugu. Hau esan ondoren, postposizio-lokuzioetan gertatzen diren erroreen etiketatze “orokor” bat deskribatuko dugu.

Postposizio-lokuzioen errore-etiketak aukeratzeko, datekin erabili dugun eskema bera erabiliko dugu:

Kategoria Nagusia_Aldaketa_EzaugarriLinguistikoa_Non

Kasu honetan, eremu hauek har ditzaketen balioak honakoak dira:

- *Kategoria Nagusia*: POST. Batzuetan, zehaztasuna irabaztearren etiketa honi postposizioaren OBa gehitu diogu. Adibidez, POSTAURREAN etiketak *aurrean* OBa duen postposizio-lokuzioa adieraziko luke.
- *Aldaketa*: Irizpide deskriptiboari jarraiki, ia kasu gehienetan postposizioekin ematen den *aldaketa*, postposizio-atzizkien *nahasketa* izaten da.

¹²“IXA taldean sorturiko analisi-tresnetan (...) determinatzailea eta atzizki-postposizioa morfema bakarrean landu ziren, sistema sinplifikatze aldera, nahiz eta euskararen izaera eranskariak jakitun ziren.” (Aduriz *et al.*, 2008)

- *Ezaugarri Linguistikoa*: Aldaketa jasaten duen ezaugarria kasua (KAS) izaten da maiz.
- *Non*: Ezaugarria postposizio-atzizkian (PA), osagai beregainaren leman (OBL), osagai beregainaren postposizio-atzizkian (OBPA) edo bestelakoren batean (BST) aldatzen da.

Errore bati dagokion etiketa zehaztea batzuetan zaila da, errorea deskribatzea ere hala izaten delako. Kasu horietan, adibidez POSTAURRE_1¹³ moduko etiketa orokorrak ipiniko dizkiegu errore motei, etiketak behetik goranzko ordenan zenbakituz.

V.3.2.2 Aukeraketa-prozedura

Lersundik bere tesi-lanean (Lersundi, 2005), 172 postposizio-lokuzio zerrendatu ditu, batzuk oso ohikoak, besteak oso erabilera urrikoak. Hauek guztiak landu ezinean, sakonean aztertuko ditugun bost postposizio aukeratu ditugu, ondorengo lerroetan zehaztu ditugun urratsak jarraituz:

1. **Ahalik eta corpus handienak (zuzena zein okerra) bildu ditugu postposizio-lokuzioen agerpenak aztertzeko.** Postposizio-lokuzioen agerpen kopurua testuetan ez da oso handia. Hau dela eta, hauek bilatzeko corpus handiak eskuz aztertzea ekidite aldera, maiztasuna aztertzeko metodo automatiko bat bilatu dugu. Testuetan postposizioak azkar eta erraz identifikatzeko, euskarazko postposizioak detektatzeko gramatika (Lersundi, 2005) baliatu dugu. Honek ongi erabilitako postposizioa detektatzen du, eta ez gaizki erabilitakoa, hori baita, hain zuzen ere, gure lanaren helburua. Gure ustea da, testu zuzenetan maiz erabiltzen diren postposizio-lokuzioak, testu okerretan ere maiz agertuko direla.

Corpus zuzenak *Euskaldunon Egunkariko* 2002 urteko testu guztiak biltzen ditu eta 8.207.919 hitz ditu. Erroreak izan ditzakeen corpusa biltzea zailagoa eta neketsuagoa izan da, beraz, txikiagoa da, eta 994.658 hitz ditu. Txikiagoa izanik, gaiei, egileei eta hizkuntza-mailari begiratzen badiegu, aberatsagoa dela esan dezakegu (ikus V.7 eta V.8 taulak).

2. **Osagai beregain izan daitezkeen hitz guztien lemak bilatu ditugu (172 lema, aztertutako postposizio-lokuzio bakoitzeko bat) bildutako bi cor-**

¹³Etiketa hau *aurre* OBLa lukeen postposizio-lokuzio bati ipiniko genioke.

Corpus erroreduna	
Euskaltegiak	126.447
EuskaraZ zerrenda	542.866
Euskara teknikoak	19.549
KBP	305.796
Guztira:	994.658

V.7 Taula: Postposizio-lokuzioen azterketan erabilitako corpus erroreduna.

Euskaltegiak		
Behe-maila (1-5)	AEK98	27.572
	Trintxerpe05	942
Maila ertaina (6-8)	Irale90	8.097
	Ilazki92	14.783
	AEK98	13.659
Goi-maila (9-12)	Ilazki92	13.717
	Ilazki95	25114
	Ilazki97	10434
	Trintxerpe05	2257
Guztira		126.447

V.8 Taula: Postposizio-lokuzioen azterketan erabilitako euskaltegiako corpus erroreduna ikasketa-mailaren arabera zehaztua.

pusetan (adibidez, **buruz**¹⁴ edo **aurre**), eta beraien testuinguruan erakutsi. Gero, karaktere-kate horietako bakoitza hitz-formetan bilatu du, eta ondoren, aurrerago aipatutako gramatikak postposizio moduan markatutakoak identifikatu ditugu. Hona hemen azterketa honetan topatutako kasu bitxi pare bat:

- *bat* karaktere-katea oso maiz agertzen da corpusetan (corpus erroredunean 12.008 aldiz), baina oso gutxitan postposizio-lokuzio moduan (adibidez, **zurekin batera**).
- Kontrakoa da *arabera* katearen kasua. Hitz-formetan agertzen denean (481 agerpen) ia beti postposizio-lokuzio baten parte izateko da (455 kasu corpus erroredunean).

3. Postposizio moduan markatutako egiturak maiztasunaren arabera sailkatu ditugu agerpen-kopuru handienetik txikienera (ikus V.9 taula).

¹⁴EDBLn **buruz** hitzak adberbio kategoriadun sarrera du. *Euskal Hiztegia*n ere “-i buruz” egiturak sarrera du bere horretan. Hau dela eta deskribatu dugu **buruz** lema bailitzan, eta ez, **buru** hitza instrumental kasuan deklinatuta.

	Corpus zuzena	Corpus erroreduna
1.	arte (10418 agerpen)	arte (1726)
2.	bitarte (3075)	bezala (738)
3.	barru (2870)	aurre (691)
4.	kanpo (2250)	gabe (671)
5.	buruz (2205)	alde (647)
6.	gora (2115)	inguru (565)
7.	esker (2040)	bide (460)
8.	oste (2011)	arabera (455)
9.	gisa (1916)	bat (285)
10.	aurre (1715)	bitarte (283)
11.	gabe (1483)	gero (277)
12.	arabera (1424)	zehir (254)
13.	erdi (1127)	gisa (232)
14.	geroztik (1083)	barru (219)
15.	ondo (1054)	kontra (214)
16.	inguru (804)	aurka (203)
17.	ondoren (794)	ondo (177)
18.	bide (705)	kanpo (128)
19.	zehir (636)	atze (126)
20.	begira (632)	bila (122)

V.9 Taula: Corpus zuzenean eta erroredunean agerpen kopuru handieneko 20 postposizio-lokuzioak.

Taulan erakutsitako maiztasunak aukeratutako testuen arabera dira, eta ez dute zertan orokortzeko modukoak izan. Corpus zuzena *Euskaldunon Egunkariko* testuekin osatuta dago eta honek, egunkari gehienek bezala, estilo bati jarraitzen dio. Estiloak, corpusean erabilitako postposizio-lokuzioetan eragina duela uste dugu. Antzeko fenomenoak gertatzen da aukeratutako corpus erroredunean. Erroreak izan ditzakeen corpusaren zati handi bat euskaraz idatzitako posta elektronikoko mezuekin osatuta dago (EuskaraZ zerrenda, 2000 eta 2001 urteetako 3049 mezu). Mezuetan ohikoa da norbaiten galderari “erantzutean” jasotako testua erantzunean jartzea. Jasotako testuan oker erabilitako postposizio bat bagenu, mezu horren erantzun guztietan agertuko litzateke. Corpusaren izaerak postposizioen agerpen kopurua baldintzatzen du, mezu bakarrean agertu arren, hainbat aldiz “zenbaitzen” baita.

Estiloak eta corpusaren izaerak emaitzetan eragin handia izan dezaketela ezagutu arren, gure ustez, corpus egokia eta zabala bildu dugu.

4. [Azkenik](#), landuko ditugun bost postposizio-egiturak aukeratu ditugu. Postposizio-lokuzioen erabilera okerrak testuetan bilatzea, dokumentatzea, automatikoki lantzea eta ebaluatzea, lan konplexua eta luzea

izan da. Hau izan da bost postposizio aukeratzeko arrazoia. Hautatutako postposizio-lokuzioek bi baldintza betetzen dituzte:

- Postposizio-lokuzioaren erabilera okerra, errorea, dokumentatuta dago gramatika-libururen batean, erabilera zuzenaren deskribapena egiten denean, batzuetan erabilera okerra ere, zehazten baita. Erabilitako iturriak V.10 taulan zerrendatu ditugu eta hauek jasotako postposizio-lokuzioen erabilera okerrak, definizio eta adibideekin, V.11 taulan laburtu ditugu¹⁵.
- Corpus zuzenean zein okerrean agerpen-maiztasun handieneko postposizio-lokuzioen artean dago. Kasu honetan salbuespen bat egin dugu *buruz* OBa duen postposizio-lokuzioarekin. Corpus zuzenetan postposizio moduan maiz agertzen den arren (5. posizioan agertzen da V.9 taulan), corpus erroredunetan ez da maizen erabiltzen diren 20 postposizioen artean agertzen.

Aukeraketa-prozesuaren amaieran, aurreko bi baldintzak betetzen dituzten honako OBak dituzten postposizio-lokuzioak aukeratu ditugu:

- *arte*
- *aurre*
- *bitarte*
- *buruz*
- *zehar*

Hurrengo atalean erroreak detektatzeko erabilitako erregela bat aztertuta ondoren (ikus V.3.3 atala), aipatu berri dugun postposizioetako bakoitza sakon aztertuko dugu.

¹⁵Taula honetan ematen diren azalpenak eta adibideak hitzez hitz jaso dira jatorrizko testuetatik.

IZ ₁ :	“Gramática didáctica del euskara” (Zubiri, 1994).
IZ ₂ :	“Euskal Gramatika Osoa” (Zubiri eta Zubiri, 1995).
IU:	Itziar Ugartebururen “Testu zuzenketak” ikastaroa http://www.ehu.es/PAT/zuzenketak/postposizioak.htm .
AI:	Euskal Herriko Unibertsitateko <i>azpidazki</i> zerbitzua http://www.ehu.es/azpikazki/ .
EGLU-I:	“Euskal Gramatika. Lehen Urratsak-I” (Euskaltzaindia, 1985).
EGLU-IV:	“Euskal Gramatika. Lehen Urratsak-IV (Juntagailuak)” (Euskaltzaindia, 1994).

V.10 Taula: Postposizio-lokuzioen erabilera okerrari buruzko informazio-iturriak.

Post.	Iturria	Errorea	Adibideak
ARTEAN	IZ ₁	Artean pluralean eta genitiborik gabe erabiltzea.	*LagunAK artean erosi diote.; Lagunen artean erosi diote.
	IZ ₂ , EGLU-IV	Juntaduretan, lehen junta-gaiak ezin du pluralean egon, bigarrenaren deklinabide-atzizki ezberdina duenean.	*Neskak eta mutilen artean dago.; Neska eta mutilen artean dago.
	IZ ₂	Artean mugagabeen erabiltzean lehen osagaia kasu-markarik gabe jartzen da. Hala ere, izena zenbatzaile batekin badoa, genitibo kasu-marka ezinbestekoa da mugagabeen ere. Pertsona-izenek, determinatzaileek eta izen-ordainek ezinbestekoa dute genitibo kasua.	*Jende asko artean banatu ditugu.; Jende askoren artean banatu ditugu.
ARTE	EGLU-I, IZ ₂	Lekuzko esanahia duten hitzekin “-raino” erabili behar da eta ez “arte”.	*Etxauri parkea arte; Etxauri parkeraino
BEZALA	AI		*Irakasle bezala aritu zen.; Irakasle aritu zen.
AZPIAN	IZ ₁	Postposizioaren aurreko hitzari artikulua jartzea.	*OheA azpian dago.; Ohe azpian dago.
BIDEZ	IU		*Hori, bi bektore bidez irudika daiteke.; Hori, bi bektoreREN bidez/bitartez irudika daiteke.
GAINEAN	IZ ₁	Markarik ez duten kasuetan artikulua jartzea.	*MahaiA gainean dago.; Mahai gainean dago.
ONDOAN	IZ ₁	Markarik ez duten kasuetan artikulua jartzea.	*AteA ondoan dago.; Ate ondoan dago.
ORDEZ	IU		*ganbera orde zalogunea duten hilerriak.; ganberaREN orde zalogunea duten hilerriak.
ZE HAR	IZ ₁	Zehar postposizioarekin genitibo erabiltzea inesiboaren orde.	*BasoaREN zehar etorri da.; Basoan zehar etorri da.

V.11 Taula: Dokumentatutako erroreak postposizio-lokuziotan.

V.3.3 Detekzioa

Postposizio-lokuzioetan gertatzen diren erroreak detektatzea ez da berehalako lana. Zailtasun handiak topatu ditugu egituron anbiguotasuna dela kausa, eta baliabide ugari erabili behar izan dugula lana aurrera eramanez. Ikus dezagun adibide moduan *buruz* hitz-forma osagaitzat duen errore mota bat detektatzeko erabilitako islapen-erregela (ikus V.5 eta V.7 irudiak). *Buruz* postposizio-lokuzioaren erabilera zuzenak “-ra/-i -Ø/-ko” eskema jarraitzen du. Hots, postposizio-atzizkian adlatibo (-ra) edo datibo (-i) marka du, eta *buruz* osagai beregaina markarik gabe, edo lekuzko genitibo (-ko) markarekin agertzen da. Adibidez, *oihanera buruz abiatu zen* edo *liburu honi buruz mintzatu gara* egiturak zuzenak dira (Euskaltzaindia, 1985), baina ez, **bidaia buruz hitz egin genuen* egitura. Azken adibideko moduko egitura okerrak detektatu nahi ditugu, eta horretarako, hasiera batean V.5 irudiko erregela definitu genuen:

```
LIST POSTPOSIZIOAK-BURUZ = "buruz";
MAP ({POSTBURUZ_NAH_KASEZDATALA_PA} TARGET IZE-DET-IOR IF (NOT 0 DAT-ALA)
      (1 POSTPOSIZIOAK-BURUZ + ADB));
```

V.5 Irudia: *Buruz*en erabilera oker bat deskribatzen duen islapen-erregela baten lehen hurbilpena.

Hau da, idatzi “{POSTBURUZ_NAH_KASEZDATALA_PA}” etiketa, baldin eta izena, determinatzailea edo izenordaina kategoriako uneko hitzak (TARGET IZE-DET-IOR) ez duen datibo edo adlatibo postposizio-atzizkirik (NOT 0 DAT-ALA), eta gainera, ondorengo hitza (1 posiziokoa), *buruz* adberbioa den (1 POSTPOSIZIOAK-BURUZ + ADB).

V.5 irudiko islapen-erregelak, V.6 irudiko analisisian ongi markatu du errorea, baina beste hainbat adibideren azterketak erakutsi digu erregela *orokorregia* dela: erakutsi dugun analisisian errorea detektatzen duen era berean ipin lezake errore-etiketa *bidaia buruz ari zitzaigun kontatzen* egitura zuzenean. *Buruz* hitza tartean duten eta postposizio ez diren esamolde/egitura zuzen ugari daude, eta egitura horiek ezin ditugu erroretzat jo. Honela landu ditugu ekidin nahi ditugun egiturak:

- Esamoldean aditza lagun duten egiturak (adibidez, *buruz erori/jo, buruz ikasi/hartu, buruz jokatu, buruz ari izan...*) BURUZ-ADI zerrendan bildu ditugu. Aditza, *buruz* hitzaren aurretik ala atzetik ager daiteke perpaus berean eta ez du zertan ondoan agertu (adibidez, *kanta bu-*

```

0. "<bidaia>"
   "bidaia"IZE ARR DEK ABS MG AORG {PostBURUZ_NAH_KASEZDATA_PA
   "bidaia"IZE ARR DEK ABS NUMS MUGM AORG {PostBURUZ_NAH_KASEZDATA_PA
   "bidaia"IZE ARR ZERO AORG {PostBURUZ_NAH_KASEZDATA_PA
   "bidaiatu"ADI SIN AMM ADOIN NOTDEK
1. "<buruz>"
   "buru"IZE ARR DEK INS MG
   "buruz"ADB ARR ZERO

```

V.6 Irudia: **bidaia buruz** egituraren analisi morfosintaktikoa, errore etiketekin.

ruz eta letraz letra ikasi du¹⁶, edo **ez ikasi testua buruz**). Aipatutako esamoldeak zuzenak dira, eta errorea detektatzera goazenean, ez daitezela inguruan ageri eskatu dugu V.7 irudiko erregelari. Horretarako, **BURUZ-ADI** zerrendako elementuak ez daitezela agertu *buruz* OBaren eskuineko edozein posiziotan (**NOT *1 BURUZ-ADI**), ezta ezkerreko edozein posiziotan ere (**NOT *-1 BURUZ-ADI**) zehaztu dugu.

- Aditza osagai ez duten gainontzeko esamoldeetan *buruz* hitzaren kidea haren ondoan azaltzen da eskuineko aldean (**buruz buru** eta ez ***buru buruz**). **BURUZ-GAIN** zerrendan bildutako elementuek ez dutela *buruz* hitzaren ondoan agertu behar (**NOT 2 BURUZ-GAIN**) zehaztu dugu V.7 irudiko erregelari.
- Futbolarekin lotutako zenbait egitura ere ekidin behar izan ditugu alarma faltsuen kopurua jaisteko. **baloi, gol, sare** eta antzeko lemak **BURUZ-FUTBOLA** zerrendan bildu ditugu eta analizatu beharreko testuetan agertzen direnean, alde batera utzi ditugu.
- Alarma faltsuak ekiditeko definitutako baldintza hauetaz gain, beste baldintza bat ezarri diogu erregelari: uneko hitza lexikorik gabe (*gues-serraren bidez*) analizatua ez izatea (**NOT 0 EZEZAGUNA**).

Ikusi dugun moduan, errore-patroiak definitzen ditugunean, postposizio erroredunaz gain, erroretzat jo daitezkeen egitura zuzenak izan behar ditugu kontuan, baita postposizio ez direnak ere. Alarma faltsuak ekiditeko ezarritako baldintzek, ordea, egitura erroredunen bat ez detektatzea eragin dezakete. Adibidez, ***baina hori buruz zerbait esatea diote** egitura erroreduna

¹⁶*kanta buruz hitz egin postposizio-egitura, okerra litzateke.

ez genuke detektatuko, **buruz...esan** esamoldea topatu dugulako. Hala ere, lehenago ere esan dugu, erabaki dugula egokiagoa dela egitura erroredun bat ez detektatzea, alarma faltsu bat sortzea baino.

```
LIST BURUZ-ADI = "erori" "jo" "ikasi" "hartu" "jokatu" "ari_izan" "esan" "errematatu"
                "jakin" "gogoratu";
LIST BURUZ-GAIN = "behera" "belarri" "buru" "gain" "gora";
LIST BURUZ-FUTBOLA = "falta" "baloi" "jokalari" "gol" "futbol" "sare" "zutoin" ...;
LIST POSTPOSIZIOAK-BURUZ = "buruz";
MAP ({PostBuruz_Nah_KasEzDataLA_PA} TARGET IZE-DET-IOR IF (NOT 0 DAT-ALA)
                                (NOT 0 EZEZAGUNA OR BURUZ-ADI OR BURUZ-FUTBOLA)
                                (1 POSTPOSIZIOAK-BURUZ + ADB)
                                (NOT 2 BURUZ-GAIN)
                                (NOT *1 BURUZ-ADI OR BURUZ-FUTBOLA)
                                (NOT *-1 BURUZ-ADI OR BURUZ-FUTBOLA);
```

V.7 Irudia: *Buruzen* erabilera oker bat deskribatzen duen islapen-erregela.

buruz postposizio-lokuzioa aztertu dugunean ikusi dugu postposizio-lokuzioen erabilera okerrak detektatzea ez dela berehalakoa. Postposizio bakoitzak zailtasun eta detekziorako behar desberdinak ditu, eta hau horrela izanik erabaki dugu kasu bakoitza banan-banan eta ahalik eta sakonen aztertzea. Lehenengo, aztertutako postposizio-lokuzio guztien erabilera zuzena taula batean (V.12 taulan) erakutsiko dugu¹⁷; ondoren, postposizio bakoitzerako detektatutako errore motak zerrendatuko ditugu, eta erroreak detektatu ahal izateko erabilitako baliabideak deskribatuko ditugu; eta azkenik, lortutako emaitzak erakutsiko ditugu.

V.3.3.1 Arte/artean

Postposizio korapilatsua da *arte* osagai beregainaren lema duena. Anbigutasun morfosintaktikoak eragiten dituen arazoez gain (adberbioa edo izena izan daiteke), postposizioa zuzen edo oker erabili ote den ezagutzeko, informazio semantikoa erabili behar izan dugu (postposizioa *lekua* edo *denbora* adierazten duten hitzekin erabili den ezagutu behar da zenbait kasutan).

Topatutako erabilera okerrak eta hauek detektatzeko egindako lana errazago aztertzearen, bi zatitan banatu dugu *arte* OBLa duen postposizioaren analisia. Lehenengo, osagai beregaintzat *artean* duten egiturak landuko ditugu, eta ondoren, OBa *arte* dutenak.

¹⁷Postposizio-lokuzioen erabilera zuzenaren deskribapena (Aduriz *et al.*, 2008) lanetik jaso dugu.

OBL	OBLren kategoria	Forma		Adibideak
ARTE	IZE	-Ø/-en	-an/-ra/-tik/-ko	Beste gauza askoren artean. Gazteen artetik gazteena. Jende artean. Hirurak arte ez ziren iritsi. Bilbora arteko trenak. Orain arte. Bihar arte. Atzo arte
		-abs	-Ø/-ko	
		-ra	-Ø/-ko	
		-Ø	-Ø	
AURRE	IZE	-Ø/-en	-an/-ra/-tik/-ko	Hormaren aurretik atera dituzte liburuak. Eliza aurrean ikusi ditut hizketan Hemendik aurrera...
		-tik	-ra	
BITARTE	IZE	-en	-z	Izaskunen bitartez lortuko dugu.
		-ra	-n	Etxera bitartean esango dizut.
		-a	-n	Ordubata bitartean etxean izango naiz.
		-Ø	-n	... meza bitartean...
		-en	-n	VI. eta VII. mendeen bitartean...
		-Ø	-Ø/-ko	... lau hektarea bitarte...
		-ak	-Ø/-ko	... seiak bitarte...
		-ra	-Ø/-ko	... etxera bitarte...
BURUZ	ADB	-ra	-Ø/-ko	Oihanera buruz abiatu ziren.
		-i	-Ø/-ko	Liburu honi buruz mintzatu gara.
				Mendiari buruz abiatu ziren.
ZECHAR	ADB	-n	-Ø/-ko	Mendian zehar ibili ginen.
		-Ø	-Ø/-ko	Hor zehar dabilta.
		-tik	-Ø/-ko	Hortik zehar dabilta.

V.12 Taula: Aukeratutako postposizio-lokuzioen erabilera zuzena.

Artean

Ilari Zubiriren arabera (Zubiri, 1994), *artean* bigarren osagaitzat duen postposizio-lokuzio batek ez du genitiboa ez den pluraleko postposizio-atzizkirik onartzen (ikus V.11 taulan dokumentatutako lehenengo errore mota). Gogorra dezagun erabiltzen duen adibidea (ikus (14) adibidean gaizki erabilitako egitura bat, eta dagokion ordain zuzena):

- (14) ***LagunAK artean erosi diote.**
Lagunen artean erosi diote.

Postposizio-lokuzioa osatzen duten bi osagaiei soilik begiratuz gero okertzat jotako egitura, zuzena izan daitekeela esan dezakegu, anbigua baita:

1. **Salbamendu taldeetako lagunak artean ez ziren heldu istripua gertatu zen tokira** esaldian *oraindik* iritsi gabeak zirela adierazi nahi da (“artean” *adverbio*).
2. **Lagunak artean ikusi zuen habia** esaldian, berriz, zuhaitzaren inguruan arituko ginateke (“artean” *izena*).
3. Azkenik, **Lagunak artean ospetsu egin dira eta erakusketa egingo dute** esaldian *artearen munduaren* inguruan arituko ginateke (“artean” *izena*).

Erakutsitako aukeretako batzuk, traketsak izan arren, zuzenak dira; beraz, ezin dira errore moduan markatu. Era berean, postposizio-lokuzioetatik bereizteko zailak dira. *Euskaldunon Egunkariko* corpusetik *artean* hitzaren lagin bat atera dugu eta 172 kasutatik 60 (% 30) ez dira postposizio. Anbigutasuna kasu gehienetan ezin dugu ebatzi, eta alarma faltsuak sortzeko arriskua ezagutzen dugun arren, esperimentuarekin aurrera jotzea erabaki dugu. Hori horrela izanik, erroreak detektatzeko garatu dugun gramatikan aipatutako errorea (ikus (14) adibidea) definitzen duen erregelak errorea markatu du ondorengo esaldian:

- (15) *? **Neska beste lau neskekin joan ziren eta haiek artean oso giro ona zeramaten.**

Kasu honetan ez gara gai idazlearen intentzioa aurreikusteko, eta hori dela eta, ezin dugu postposizioa gaizki erabili ote den esan. *Oraindik* zuten giro ona, ala *haien artean* zuten giro ona?

Aipatutakoaz gainera, *artean* osagaitzat duen postposizio-lokuzioaren inguruko errore gehiago topatu dugu (ikus errore-kasuistika V.13 taulan). Taula honetan eta landutako postposizioei dagozkien antzeko tauletan, deskribapenerako egitura bera erabili dugu: lehenengo zutabean, errore-adibideak zein iturritatik jaso ditugun adierazi dugu; bigarrenean zein hirugarrenean, egitura erroreduna deskribatu dugu; eta azkenekoan adibideak eman ditugu. Errore mota bakoitzeko islapen-erregela bat edo gehiago idatzi ditugu MG gramatikan.

Iturria	Errorea	Forma	Adibideak
IZ ₁	POSTARTEAN_NAH_NUMP_PA. Postposizio-atzizkia (PA) plurala izatea, eta gainera, ez genitiboa.	PLU <i>artean</i>	*LagunAK <i>artean</i> erosi diote.
IZ ₂ , EGLU- IV	POSTARTEAN_NAH_NUMP_BST. Juntaduretan, lehen jun- tagaiak ezin du pluraleko hondarkirik eduki bigarren osagaiaren deklinabide-atzizki desberdina duenean.	PLU JUNT <i>-en</i> <i>ar-</i> <i>tean</i>	*Neskak eta mutilen <i>artean</i> dago.
IZ ₂	POSTARTEAN_NAH_KAS_PA. OBa mugagabeen erabiltzean post- posizio-atzizkia markarik ga- bea (Ø) izan ohi da. Hala ere, izena zenbatzaile batekin badoa, genitiboa ezinbestekoa da mugagabeen ere. Pertsona- -izenek, determinatzaileek eta izenordainek ezinbestekoa du- te postposizio-atzizkia geniti- boan.	IZE <i>-Ø</i> <i>artean</i> (DET, IOR)	*Jende asko <i>artean</i> banatu ditugu.
Corpus erroreduna	POSTARTEAN_NAH_KAS_SOZ_PA. Postposizio-atzizkiak sozietibo marka izatea.	<i>-ekin</i> <i>artean</i>	*Poz-pozik pasa ni- tuen egunak, as- teak, beti lagunekin <i>artean</i> .
Corpus erroreduna	POSTARTEAN_NAH_KAS_INE_PA. PA inesibo izatea.	<i>-an</i> <i>artean</i>	*Kantabria eta Pa- lantzian <i>artean</i> da- goen herri txikia.
	POSTARTEAN_NAH_KAS_PA. PA ergatibo singularra edo absolu- tibo plurala izatea.	<i>-ak</i> <i>artean</i>	*Txalok eta Inerso- ren ohiuk <i>artean</i> ...
	POSTARTEAN_I. Genitiboaren erabileran ordena nahastea.	<i>-en -Ø</i> <i>artean</i>	*anai-arreben guzti <i>artean</i> .

V.13 Taula: *artean* OBa duen postposizioaren inguruko erroreak.

Erroreak detektatzeko definitu ditugun erregelekin lortutako emaitzak aurrerago ikusiko ditugu, baina aipagarriak iruditzen zaizkigu ebaluazioa egin ahal izateko eskuz egindako azterketan topatu ditugun zenbait adibide

(ikus (16), (17) eta (18) adibideak). Horietan “-Ø artean” patroia “zuzena” betetzen da, hau da, adibide hauek **esku artean** edo **jende artean** egituren parekoak dira sintaxiari begiratuta. Guk, ordea, errore kontsideratuko genituzke. Segur aski, sintaxia ez, baina semantika aztertu beharko genuke arrazoia topatzeko. Tesi-lan hau sintaxira mugatzen denez, egitura ustez erroredun hauek ez lantzea erabaki dugu.

- (16) ?...proposamen artean...
- (17) ?...sanitari sistemako hospital artean...
- (18) ?...beste arraza artean...

Bestalde, taulako zenbait deskribapen aztertzeraz jota, POSTARTEAN_NAH_NUMP_BST kodearekin etiketatutako errore-patroia alarma faltsuen iturri izan daitekeela susma dezakegu, juntagailu den elementuak (**eta**) perpausak ala sintagmak juntatzen dituen jakitea oso zaila baita. Honako adibidean (ikus (19) adibidea), gramatikak errorea markatuko luke erabilera zuzena izanik:

- (19) Irlandako bake prozesuan izan duen parte **hartzeak eta herritarren artean** lortu duen entzuteak...

Arte

Aztergai dugun postposizioarekin jarraituz, osagai beregaina *arte* duten postposizio-lokuzioaren erabilera okerrak zerrendatu ditugu V.14 taulan. Kasu honetan ere, errore mota bakoitza detektatzeko erregelak idatzi ditugu MG erabiliz. Taulan aipatzen diren egitura erroredunen inguruan zenbait zehaztapen egingo ditugu oraingoan ere:

- Orduak identifikatzea (POSTARTE_NAH_KAS1_PA). Errore mota hau zuzen-zuzenean ordukin lotuta dagoenez, zenbait gako-hitz erabili dugu aztertutako egiturak orduetara mugatzeko. *Ordu-aztarna* izena eman diogu hitz horien zerrendari eta bertan **goiz**, **eguerdi**, **arratsalde** eta antzekoak bildu ditugu. Ordu-aztarnarik agertzen ez denean, ezin dugu ziurtatu ordu bati buruz ari garenik, beraz, ez dugu errorearik markatuko. Fenomeno hau gertatzen da corpus erroredunean azaldu zaigun (20) adibideko esaldian. Dena dela, adibide honetan elipsia dagoenez ez dakigu orduaren inguruan ari diren (postposizioaren erabilera okerra), ala **12 porro erretze**ari buruz ari diren (postposiziorik ez, zuzena).

- (20) Porro bat erre nuen eta 12 arte etxean zorian egon nintzen.

Iturria	Errorea	Forma	Adibideak
Euskaltz. (35. araua)	POSTARTE_NAH_KAS1_PA (ORD). Orduak: ordu bata ergatiboan deklinatzea.	ordu bata <i>k</i>	*ordu bata <i>k</i> arte
IXA hiz- kuntzala- riak	POSTARTE_NAH_KASALA_PA. Adberbioak: adberbioak ad- latibo kasuan deklinatzea. Salbuespena: hur engora arte	-ra (ADB) arte	*atzora arte
Corpus erroreduna	POSTARTE_NAH_RAINO. -raino <i>vs</i> arte: lekua adierazten duten hitzekin, salbuespenak salbu, <i>arte</i> erabiltzea -raino deklinabide-kasua erabili beharrean.		*Etxauri parkea ar- te.
	POSTARTE ₁ . Euskarazkoa ez den egitura erabiltzea.	beste egun arte	*beste egun arte
	POSTARTE_NAH_KAS2_PA (ORD). Orduak: gainontzeko orduak ergatiboan ez deklinatzea.	not(-ak) arte	*Goizean hamar ar- te ohean.

V.14 Taula: *arte* OBa duen postposizioaren inguruko erroreak.

- **Lekuak identifikatzea** (POSTARTE_NAH_RAINO). Postposizio-lokuzio anitzen erabilera aldatu egiten da lekuari edo denborari buruz ari garenean. Adibidez, **gero arte** esan dezakegu, baina ez, ***etxea arte** (egokia goa da **etxeraino** baina *arte* erabiltzekotan **etxera arte** izan beharko luke)¹⁸. *Lekua* adierazten duten hitzak zeintzuk diren ezagutzeko, orain *arte* erabili ez dugun hizkuntza-baliabide baten beharra izan dugu: *genusak*. Lersundik bere tesi-lanean (Lersundi, 2005) Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa egin zuen eta bertatik *definizioen gune sintaktiko eta semantikoak —genusak—* erauzi zituen. Genusak **toki**, **leku** edo antzeko hitzak direnean (ikus (21) adibidea) lekua adierazten duen hitz baten aurrean gaude. Euskal Hiztegitik erauzitako 168 hitz eta Euskal Hiztegi Modernotik (Elhuyar, 2000) jasotako 242 hitz baliatu ditugu mota honetako errorearen detektziorako. Horiei, leku-izen bereziak (adibidez, **Hondarribia**) gehitu dizkiegu.

(21) legartza A0 iz. **toki** edo **eremu** legartsua.

- **Anbiguitasun semantikoa** ekiditea (POSTARTE_NAH_KAS3_PA eta POSTARTE_NAH_RAINO). *arte* hitzaren erabilera postposizioetara mugatze-

¹⁸EGLU-1 (395): “Has gaitzen lehenetik. Batak, -raino-k, lekua adierazten duela eta hainbestez leku izenekin gehienbat erabili behar dela esan izan da gramatika guztietan; bigarrenak, **arte**-k, aldiz, denbora adierazten duela eta denborazko adberbio eta sintagmekin erabili behar dela batik bat. [...] Baina bi atzizki horien gurutzaketak askotxo dira.”

ko, ondorengo hitzen antzekoak jarraian dituzten egiturak ekidin egin ditugu: **areto**, **museo** eta **erakusketa** izenak eta **hartu** aditza. *Arte* hitzaren aurretik **zazpigarren** hitza duten egiturak ere ekidin ditugu.

- **Orokortzea (POSTARTE_NAH_RAINO)**. Orokortu egin dugu “-raino *vs* arte” bereizketa. Jakin badakigu salbuespenak ugariak direla (adibidez, **gaurdaino** edo **Lezamara arte**), baina ezinezkoa egiten zaigu salbuespen guztiak kontuan edukitzea.

Emaizak

Aurreko atalean zehaztutako errore-patroiak islatzen dituzten erregelak idatzi ondoren, horiek corpus erroredunean eta corpus zuzenean aplikatu ditugu. Corpus erroredunean erabilitako esaldien kopurua V.15 taulan erakusten da, iturriaren arabera banatuta. II. kapituluaren esandakoa gogoratuko dugu: corpusak osagai beregainaren lemaen agerpenekin osatuta daude. Hots, i) ongi erabilitako postposizio-lokuzioekin, ii) gaizki erabilitakoekin eta iii) postposizio ez diren egiturekin osatuta daude. Postposizio bakoitzeko, *garapenerako* (G) erabilitako corpusaren esaldi kopurua (elementu guztietatik % 60), *probarako* (P) erabilitako kopurua (elementu guztietatik % 40) eta orotara erabilitako esaldi kopurua (T) zehaztu ditugu. Modu berean antolatatu dugu corpus zuzenaren banaketa erakusten duen V.16 taula. Corpus erroredunean *doitasuna* eta *estaldura* neurtuko ditugu eta corpus zuzenean *doitasuna* soilik (II. kapituluaren eman genuen aukeraketa honen arrazoia).

Probetarako elementu kopurua															
	arte			aurre			bitarte			buruz			zehar		
	T	G	P	T	G	P	T	G	P	T	G	P	T	G	P
Behe	22	13	9	6	4	2	5	3	2	6	4	2	4	2	2
Ertain	52	31	21	34	20	14	20	12	8	20	12	8	19	11	8
Goi	59	35	24	56	34	22	27	16	11	33	20	13	20	12	8
EuskaraZ	498	299	199	385	231	154	333	200	133	273	164	109	40	24	16
KBP	403	242	161	527	316	211	245	147	98	95	57	38	75	45	30
Eusk. tek.	33	20	13	31	19	12	36	22	14	24	14	10	12	7	5
Etab.	948	569	379	889	533	356	620	372	248	482	289	193	142	85	57
Guztira	2015	1209	806	1928	1157	771	1286	772	514	933	560	373	312	186	126

V.15 Taula: Garapenerako eta probetarako corpus erroredunak.

Probetarako elementu kopurua															
	arte			aurre			bitarte			buruz			zehar		
	T	G	P	T	G	P	T	G	P	T	G	P	T	G	P
Egunk.	12948	7769	5179	13549	8129	5420	6410	3846	2564	9058	5435	3623	2500	1500	1000

V.16 Taula: Garapenerako eta probetarako corpus zuzenak.

arte OBa duen postposizio-lokuzioarekin jarraituko dugu. Horrekin lortutako emaitzak V.17 taulan zehaztu ditugu. Corpus erroreduneko garapenerako zatian 20 errore ongi detektatu dira; beraz, desanbiguatu gabeko corpus erroredunean % 80ko estaldura lortu dugu. Detektatu ez ditugun egituren artean (16), (17) eta (18) adibideetan aipatutako egiturak daude. Probarako corpusean 5 errore geratu dira detektatu gabe, eta honen eraginez, estaldura % 28,6 da. Egitura horietako bi “bereziak” dira eta ez genituzke orokortuko detekzioarako erregeletan isla izan dezaten, errore mekanikoak direla uste baitugu (adibidez, *orduan arte* eta *denon artea hitzegitea* egiturak). Gainontzeko hiru egiturek, berriz, zenbait deskribapenen zehaztapena eskatzen dute. Har dezagun (22) adibidea. Bertan azaltzen den egitura, “-en artean”, zuzen moduan deskribatuta dago V.12 taulan. Hala eta guztiz ere, postposizioa gaizki erabili da. Postposizio-lokuzioen erabilera zehazten den gramatika-liburuetan, egitura hori postposizio-atzizkian pluraleko mugatasun ezaugarria duten hitzekin soilik dela zuzena zehaztu beharko litzateke¹⁹.

(22) *Mendiaren artean bezala biziko gara.

Zehaztu beharreko beste egitura bat (23) adibideari dagokiona da. Egitura zuzena litzateke, osagai beregainaren ondorengo hitzaren kategoria *aditza* ez balitz (adibidez, *erlazio hau bi entitateen arteko harremana ezartzeko hobesten da* egitura, zuzena da). Errorrea detektatzeko, beraz, postposizio-lokuzioaren ondorengo hitzaren kategoria zehaztu beharko litzateke.

(23) *Erlazio hau bi entitateen arteko hobesten da.

Doitasunari dagokionez, garapenerako corpusean % 100ekoa bada ere, probarako corpusean % 28koa da alarma faltsuen detekzioagatik (aurrerago aztertuko ditugu alarma faltsuak).

Corpus zuzenean, garapenerako corpusean zehazkiago, % 29ko doitasuna lortu da errore-etiketa jaso duten egituretatik 12 benetako erroreak baitira eta 29 alarma faltsuak. Kopuru altua dirudien arren, kontuan hartu beharra dago 7769 esaldirekin (“arte” forma dutenak, denak) egin dela proba. Honek % 0,37ko alarma faltsuen tasa ematen du (*alarma faltsuak/elementu kopurua*)²⁰. Probarako corpusean antzeko emaitzak lortu ditugu % 28ko doi-

¹⁹Segur aski, munduaren ezagutza suposatzen delako ez da zehazten ezaugarri hau (unitate anitzen *artean* ibiliko gara, ez bakarraren *artean*), baina datu hau erroreen detekzioarako ezinbestekoa da.

²⁰Tasa zehatza emateko, *alarma faltsuak / elementu kopurua - errore guztiak* formula erabili beharko genuke, baina horretarako corpuseko errore guztiak ezagutu behar dira. Ematen dugun neurria estimazio bat den arren, fidagarria dela uste dugu eta benetako baliotik oso hurbil dago.

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa
Elementu kopurua	1209	806	7769	5179
Erroreak	25	7	?	?
Detektatu gabeak	5	5	?	?
Ongi detektatuak	20	2	12	10
Alarma faltsuak	0	5	29	24
Estaldura	% 80	% 28,6	-	-
Doitasuna	% 100	% 28	% 29	% 28
Alarma faltsuen tasa			% 0,37	% 0,46

V.17 Taula: *Arte*. Emaitzak.

tasunarekin eta % 0,46ko alarma faltsuen tasarekin. Corpus zuzenean errore gutxi azaltzen da eta *arte* postposizioaren anbigutasun handiaren ondorioz, alarma faltsu asko dago. Alde positiboari begira, alarma faltsuen tasa baxu mantentzearen helburua ongi bete dugula esango dugu.

Alarma faltsuak	<i>Desanbiguatu gabeko corpus analizatua</i>				
	Corpus erroreduna		Corpus zuzena		Guztira
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa	
Anbigutasun morfologikoa	0	0	9 + 4 (duda)	12	25
Anbigutasun semantikoa	0	0	9	7	16
Sintaxia	0	5	3	5	13
Prozesu automatikoa	0	0	4	0	4
Guztira	0	5	29	24	58

V.18 Taula: *Arte*. Alarma faltsuak.

Alarma faltsuak

OBa *arte* duen postposizio-lokuzioaren emaitzetan bereziki aipatzekoa da sortzen duen alarma faltsuen (AF) kopurua. Emaitzak zehatz-mehatz aztertutitu arrazoiak bila (ikus V.18 taula) eta aurreikusi genuen moduan baieztatu dugu alarma faltsuen kopuru altuak zuzen-zuzenean duela zerikusia anbigutasun morfosintaktikoarekin eta anbigutasun semantikoarekin (58 alarma faltsuetatik 41ek dute lotura anbigutasunarekin):

- **Anbigutasun morfologikoa (25 AF)**. Analisiak desanbiguatu gabe erabiltzearen ondorioz sortu dira honako alarma faltsuak:

– *Nahasketa kategoria gramatikalean*

- * *artean* adberbio. Aipatua dugu kasu hau dagoeneko. *artean* hitz-forma bai *izen* bailitza baita *adberbio* bailitza analizatzen da. Lehen kasuan gure gramatikak postposizio okertzat jotzen du, informazio gehiagoren faltan ezinezkoa baita *adberbio* dela jakitea.
- * Erroreak detektatzeko erregeletako batean, “pertsona-izenek, izenordainek eta determinatzaileek derrigorrez genitiboan” deklinatuta egon behar dutela zehaztu dugu (Zubiri eta Zubiri, 1995). Kategoria gramatikal hauek dira zenbait alarma faltsuren iturburu. Adibidez, *hiri arteko* sintagman, *hiri* hitza, “hi” izenordaina datibo postposizio-atzizkiarekin deklinatuta bailitza analizatzen da eta alarma faltsua sortzen du. Beste zenbait kasutan, zenbakiak “determinatzaile” kategoriarekin agertzen zaizkigu, testuinguruaren arabera, hala ez direnean.
- *Nahasketa postposizio-atzizkian*. Leku izenek *genitiboan* deklinatzen direnean, “-n” marka jasotzen dute; hau da, *inesibo* kasuan deklinatzen direnean jasotzen duten marka bera. Horrela, *Malaga eta Ciudad Realen artean* sintagman, *Real* hitzari bi analisiak ematen zaizkio eta batek alarma faltsua sortzen du.
- *Ekain* (dudazkoa). Kasu honetan, *Ekainera arte* moduko egituretan errorea markatu egiten da. Hilabeteari buruz ariko bagina, egitura zuzena litzateke; *Ekain* haitzuloari buruz ariko bagina, ordea, errorea legoke, *Ekaineraino* behar luke.
- **Anbiguotasun semantikoa (16 AF)**. Sintaxitik haratago doan informazio semantikorik gabe, zaila da zenbait alarma faltsu ekiditea. Azalduak ditugu dagoeneko, *artearen munduarekin* zerikusia duten egiturak ekiditeko egindako ahaleginak. Hala eta guztiz ere, zenbait egitura erroretzat jo ditugu. Adibidez, *Modak eta estetikak artean duten eragina* testu-zatian ez dago egitura ekiditeko zuzenean *artearekin* lotutako hitzik (moda arte kontsideratzen ez bada behintzat).
- **Sintaxia (13 AF)**. *Juntagailua perpausak lotzen*. Aurreikusiak genituen juntagailu batek sintagmak edo perpausak bereizten ote dituen ezin jakinak sor zitzakeen arazoak (ikus (19) adibidea). Horren garrantzia indartu dute emaitzek.
- **Prozesu automatikoa (4 AF)**. Tokenizatzailearen akats baten ondorioz (hitzak gaizki moztu ditu) sortu dira zenbait alarma faltsu.

Anbiguotasunaren eragina

Berriro ere, eta datekin egin genuen modu berean, desanbiguazioak erroreen detekzioan duen eragina aztertzeko asmoz, corpus erroreduneko garapenerako atala maila desberdinetan desanbiguatu dugu. Erregelak desanbiguatu gabeko testuekin lan egiteko prestatu ditugu; beraz, emaitza onenak maila honetan espero dira. Desanbiguazio-mailaren araberrako emaitzen aldaketa erakusten du V.19 taulak. Testua lehen eta bigarren mailetan desanbiguatzaren denean, estaldura % 80tik (desanbiguatu gabe) % 64ra pasa da. Aldeak handia dirudien arren, kontuan izan behar da, adibide gutxi izanda, ehuneko hauek ez direla oso esanguratsuak, 4 erroretako aldea baitago soilik. Errore horiek ez detektatzearren arrazoia hau da: **artean** osagai beregainari adberbio gramatika-kategoria esleitzen dio EUSTAGGERek, baina erabilera zuzenaren deskribapenari jarraituz, erregeletan izen kategoriakoa izan dadin eskatu dugu. Hirugarren desanbiguazio-mailan, dagokion kategoria berreskuratu du 2 kasutan eta estaldurak berriro gora egin du. Doitasunaz esan, laugarren mailan izan ezik, gainontzekoetan % 100ean mantendu dela. Maila honetan alarma faltsu bat gertatu da: **Londres eta Dublinen artean konpartituta da-goela** esaldian errorea detektatu da, perpausa zuzena izanik. **Dublinen** hitz formak leku-izena izanik, inesibo eta genitibo markak daramatza anbiguo denean. Laugarren mailan inesibo kasua duen analisi ez-egokia soilik laga dio hitzari, eta hori izan da errorea markatzearren arrazoia.

	<i>Garapenerako corpora</i>		
	Corpus erroreduna		
	Errore detektatu gabeak	Errore detektatuak	Alarma faltsuak
Desanbiguatu gabe	5	20 (% 80)	0
1.mailan	9	16 (% 64)	0
2.mailan	9	16 (% 64)	0
3.mailan	7	18 (% 72)	0
4.mailan	7	18 (% 72)	1

V.19 Taula: *Arte*. Anbiguotasunaren eragina emaitzetan.

V.3.3.2 Aurre

Oso erabilera zabalekoa da *aurre* hitz-forma. Hiztegi Batuko bilatzailean²¹ *aurre* katea idatziz gero, 71 sarrera eta azpisarrera bueltatzen dizkigu sis-

²¹<http://www.euskaltzaindia.net/hiztegibatu/>

temak. Elhuyar hiztegian²² sarrera kopurua 116koa da. Sarrerak hainbat motatakoak dira: *aurre* aurrizkia duten hitzak (adibidez, *aurrekari*, *aurrei-kusi*...), aditz konposatuak (adibidez, *aurrera eraman*, *aurre egin*...), egitura eginak, eta bestelakoak. Aztertutako corpusetan, osagai beregainaren lema *aurre* duen postposizio-lokuzioaren erabilera ez da oso zabala, eta orokorrean, ongi erabiltzen da. Hau da, postposizio horrekin dugun helburu nagusia da alarma faltsuak ekiditea.

Postposizio-lokuzio honen inguruan gertatzen diren erroreak detektatu ahal izateko, aurreko atalean aipatutako baliabideez gainera, hitzen tasun semantiko bat, *bizidun/bizigabe* ezaugarria zehazkiago, erabili behar izan dugu. Azter dezagun lehenengo *aurre* OBko postposizioaren inguruan gertatzen diren errorearen zerrenda (ikus V.20 taula).

Iturria	Errorea	Forma	Adibideak
IZ ₂ (235 or.)	POSTAURRE_NAH_KASBIZØ_PA. Izen <i>bizidunekin</i> PAn markarik ez eramatea.	-Ø aurrean (BIZ+)	*Amaia aurrean; *zu aurrean
IZ ₂ (235 or.)	POSTAURRE_NAH_KAS_PA. Izenaren atzean beste hitzen bat (izenondoa, erakuslea, artikulua zehaztugabea) agertzen denean POa genitiboa ez izatea.	IZE not(-en) aurrean	*Etxe hori aurrean
Corpus erroreduna	POSTAURRE_NAH_KASINE_PA. PA inesiboa izatea.	-an aurrean	*umean aurrean
	POSTAURRE_NAH_KASADJØ_PA. PA adjektiboa izanik, markarik ez (-Ø) izatea.	-Ø (ADJ) aurrean	*arriskutsu aurrean
	POSTAURRE_1 <i>aurreran</i> erabiltzea.	aurreran	*kongresuaren aurreran

V.20 Taula: *aurre* OBa duen postposizioaren inguruko erroreak.

Lehenago egin dugun moduan, taulako zenbait erroreren inguruan zenbait aipamen egingo ditugu:

- *Bizidun/bizigabe* tasun semantikoa (POSTAURRE_NAH_KASBIZØ_PA). Ilari eta Entzi Zubiriren *Euskal Gramatika Osoan* hau diote postposizio-lokuzio honen inguruan: “normalean aurreko hitzak NOREN kasu-marka hartzen du, izen *bizigabeekin* singularrean markarik gabe erabil daitezkeelarik” (ikus Zubiri eta Zubiri, 1995, 235 orr.). Hortik ondorioztatu dugu izen *bizidunekin* genitiboa erabili behar dela: *etxe aurrean* zuzena litzateke baina ez **zu aurrean*. Hitz batek semantika-mailako tasun hau ote duen ezagutzeko, itzulpen automatikoko lanetarako

²²http://www1.euskadi.net/hizt_el

IXA taldean erabiltzen den ezaugarri semantikoekin aberastutako hiztegi bat erabili dugu (Díaz de Ilarraza *et al.*, 2002). Bertan, 15.100 hitz inguru bizidun/bizigabe (BIZ), hizkuntza (HZK), denbora (DEN), gaia (GAI) eta gailua (GLU) ezaugarriekin etiketatu dira. Guk, kasu honetarako, erdiautomatikoki etiketatutako bizidun ([BIZ+]) ezaugarria duten hitzak erauzi ditugu aipatutako hiztegitik (1.642 hitz) eta erroreak detektatzeko gramatikan erabili. Lan honetarako definitutako *bizidunak* zerrendan hauek bildu ditugu: semantikoki bizidun moduan etiketatutako hitzak (adibideak dira, *agnostiko*, *arradio*, *atleta*...), pertsona-izenak (adibidez, *Koldo*) eta pertsona adierazten duten izenordainak (adibidez, *hi*, *gu*, *zuek*...).

Errore mota hau detektatzeko erregelak erabiltzen duen leihoa osagai beregainaren hurrengo hitzeraino hedatu behar izan dugu, postposizio-lokuzioaren atzetik datorren hitzaren arabera, egitura zuzena izan baitaiteke. Adibidez, *ez da gauza bera Kovacevic aurrean izatea* perpau-sak erroreetarako zehaztutako egitura betetzen du, eta zuzena da.

- *Lekuak ekiditea* (POSTAURRE_NAH_KASINE_PA). Errore mota honen deskribapenerako idatzi dugun adibidean (**umean aurrean*) errorea dagoela uste dugun era berean, hizkuntza-egitura berbera duten beste esaldi batzuetan, adibidez, *baratzean aurrean letxugak daude eta atzean azak* esaldian, ez zaigu errorerik dagoenik iruditzen. Kasu honetan, gainera, ez dago postposiziorik, baina errore-detekzio posiblea ekiditeko beharra dugu. Egitura horiek erroredun moduan ez markatzeko, postposizio-atzizkia daramaten hitzek ez dezatela *lekurik* adierazi zehaztu dugu. Horretarako, *genusak* erabili ditugu.

Emaitzak

aurre postposizioaren inguruko errore-patroiak (ikus V.20 taula) espresuki *aurre* OBLarekin osatutako probarako zerrendetan aplikatu ondoren (ikus corpusen deskribapena V.15 taulan), V.21 taulako emaitzak lortu ditugu.

Emaitzei dagokienez, onak direla esan dezakegu. *aurre* OBLa duen postposizio-lokuzioaren garapenerako corpus erroredunean, corpora desanbiguatu gabe egonik eta erabilera okerrean kopurua txikia izanik, % 85,7ko estaldura eta % 100eko doitasuna lortu dira (ez dago alarma faltsurik). Probarako corpusean 771 esalditako multzo batean ez da errore 1 detektatu (horregatik atera da % 0ko doitasuna). Bertan agertzen den egitura (**zuloko aurrean*) ez zaigu garapenerako corpusean agertu, eta beraz, ez zegoen egitura detektatzeko erregularik. Corpus berean, alarma faltsu bakarra gertatu da: **umean*

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa
Elementu kopurua	1157	771	8129	5420
Erroreak	7	1	?	?
Detektatu gabeak	1	1	?	?
Ongi detektatuak	6	0	0	2
Alarma faltsuak	0	1	6	4
Estaldura	% 85,7	% 0	-	-
Doitasuna	% 100	% 0	% 0	% 33
Alarma faltsuen tasa			% 0,07	% 0,07

V.21 Taula: *Aurre*. Emaitzak.

aurrean egitura okerraren “-an aurrean” elementu berak dituen egitura zuzen batean errorea markatu da, ... **hori euskaratzean aurrean aipatutako bi proposamenak**. . . esaldian hain zuzen ere.

Corpus zuzena oso handia da eta aurreratu dugun moduan errore gutxi dago. Horregatik, alarma faltsuak sortzeko probabilitatea handiagoa da; % 33ko doitasuna besterik ez dugu lortu.

Alarma faltsuak

Corpusean egitura zuzenak erroretzat jo direnean (ikus V.22 taula), gehienetan akatsak semantikarekin izan du lotura (11 kasutatik 7 kasutan). Gainontzeko 4 egiturak erregelaren baten zehaztapen faltagatik gertatu dira. Adibidez, “-n aurrean” egituraren alarma faltsuen agerpena ekiditeko, semantika erabili dugula esan dugu dagoeneko, postposizio-atzizkia daraman hitzak tokiren bat adierazten ote duen aztertu baitugu. Corpus zabalagoak landu ditugunean, antzeko fenomeno bat topatu dugu. Horrela, **sailkapen nagusian aurrean geratu da** egitura zuzena da, postposizioduna izan gabe, eta gure gramatikak markatu egiten du. Egitura zuzenak ekidite aldera, lekuez gain, erabilera hau “aurrealdea” eta “atzealdea” duten beste zenbait kontzeptutan onartu beharko genuke (besteak beste, **zerrenda, lasterketa**. . .).

Anbiguitasunaren eragina

Garapenerako corpusa maila ezberdinetan desanbiguatzen denean (ikus V.23 taula) ongi detektatutako errore kopurua 5ekoa edo 6koa da. Diferentzia ***guren aurrean** egituraren detekzioan gertatzen da. Hitz-segida hori erroretzat jo dugu, idazleak **gure aurrean** esan nahiko lukeela suposatzen baita.

Alarma faltsuak	<i>Desanbiguatu gabeko corpus analizatua</i>				
	Corpus erroreduna		Corpus zuzena		Guztira
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa	
Semantika	0	0	4	3	7
Landugabekoak	0	1	2	1	4
Guztira	0	1	6	4	11

V.22 Taula: *Aurre*. Alarma faltsuak.

tugu. Era berean errorea izango litzateke idazleak **guren** hitz-forma erabili nahi izan balu (*rollizo*, *predilecto* hitzen itzulpena), modu honetan PA duen hitza biziduna eta deklinatu gabea izango litzatekeelako. Zuzena izan zitekeen, ordea, **guren** formak *gura* hitzaren genitiboa izan nahiko balu. Testua inolako desanbiguaziorik egin gabe analizatzen denean, egitura honetan ez da errore-etiketak ezartzen, baina bai desanbiguatzen denean.

	<i>Garapenerako corpusa</i>		
	Corpus erroreduna		
	Errore detektatu gabeak	Errore detektatuak	Alarma faltsuak
Desanbiguatu gabe	2	5	0
1.mailan	1	6	1
2.mailan	1	6	1
3.mailan	1	6	1
4.mailan	1	6	1

V.23 Taula: *Aurre*. Anbiguotasunaren eragina emaitzetan.

V.3.3.3 Bitarte

Postposizio honek osagai beregainaren postposizio-atzizkia instrumentala duenean (bitartez) “baliabide” interpretazio semantikoa du (Lersundi, 2005). Inesiboa duenean, berriz, denborazko balioa hartzen du edo espazio/denbora tartea adierazteko erabiltzen da. Azken forma (*bitartean*) gaztelaniaren interferentzia medio “kontrastea” adierazteko ere gaizki erabiltzen da EHUL-KU zerbitzuaren arabera (<http://www.ehu.es/ehulku/>, *bitartean* aholkua). Esaldi batean kontrasterik ote dagoen jakiteko oraindik ez dugu baliabiderik, beraz, kalko oker hau erabiltzen ote den jakitea ezinezkoa egiten zaigu. OB-PA instrumentalarekin lotutako erabilera okerrak landu ditugunean, *bidez* osagai beregaina duen postposizioa ere landu dugu, ia era berean erabiltzen baitira (ikus V.12 taula).

Bitarte OBL duten postposizio-lokuzioekin topatutako erabilera okerrak V.24 taulan laburtu ditugu. Ez dugu ohar berezirik egingo postposizio honen inguruan.

Iturria	Errorea	Forma	Adibideak
IXA hiz-kuntzalarriak	POSTBITARTEZ_NAH_Ø_PA. PAn markarik ez eramatea, OBa bitartez izanik.	-Ø bitartez	*Estitxu bitartez; *ekimen bitartez
Corpus erroreduna	POSTBITARTE_NAH_KASØ_PA. PAn markarik ez eramatea, OBa bitarte izanik.	-en bitarte	*horren bitarte
	POSTBITARTE_NAH_KASABS_OB. PAdun hitza aditza izanik, OBa bitartea izatea.	ADI bitartea	*prestatu bitartea

V.24 Taula: *bitarte* OBa duen postposizioaren inguruko erroreak.

Emaitzak

Garapenerako erabili dugun corpus erroredunean, postposizio honen inguruan gertatzen diren errore guztiak azaleratu egiten ditu gure gramatikak. Horrela, corpus horretan % 100eko estaldura lortzen da. Probarako corpusean, ordea, detektatu gabe geratu zaizkigun 2 postposizio-erabilera oker direla eta, % 66ko estaldura lortu dugu. Kasu batean, ... *bere abizena, bere aitaren abizena bitartez aldatu zuen*... adibidean, *bitartez* eta *ordez* postposizio-lokuzioak guztiz nahastu ditu egileak; bigarrenean, zehaztu ez dugun egitura bat agertu da, *epe garrantzitsu bitartean*.

% 90 eta % 100 balioen artean mantendu da doitasuna postposizio-lokuzio honen erabilera okerrak bi corpusetan detektatzerakoan. Emaitzak mantendu egin dira desanbiguazio-maila desberdinak erabili direnean.

Corpus zuzenean aipatzekoa da, lan honetan errore kopuru altuena postposizio-lokuzio honekin detektatu dela: 10 errore detektatu dira garapenerako corpusean eta 9 errore probarako corpusean. Euskaltzaindia (1985) gramatikaren arabera “-en bitartez” da OBa *bitartez* duen postposizio honetan onartzen den egitura bakarra. Hala ere, gaur egun oso zabalduta da “-Ø bitartez” egituraren erabilera *komunikazio-bidea* adierazten duten izen bizigabeekin. Adibidez, *Internet bitartez*, *SMS bitartez* eta *irudi bitartez* moduko egiturak. Arauei jarraituta espresio horiek okerrak dira.

Erroreak eskuz etiketatu ditugunean ikusi ahal izan dugu, postposizioen inguruan ez, baina denborazko mendeko perpausekin erroreak gertatzen direla (emaitzetan kontatu ez ditugun 6 errore). Horiek, “-ten/-tzen -Ø bitartean” eskemari jarraitzen dioten erroreak izaten dira maiz. Adibidez, **lagun*

guztiak dantzatzen ikusten ditut bitartean, ni hemen eserita esaldian du is-la aipatutako egiturak. Mendeko perpausen tratamendua tesi-lan honetako eginbeharretatik at geratzen denez, errore mota hau ez dugu landu.

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa
Elementu kopurua	772	514	3846	2564
Erroreak	7	6	?	?
Detektatu gabeak	0	2	?	?
Ongi detektatuak	7	4	10	9
Alarma faltsuak	0	0	1	0
Estaldura	% 100	% 66,6	-	-
Doitasuna	% 100	% 100	% 90	% 100
Alarma faltsuen tasa			% 0,02	

V.25 Taula: *Bitarte*. Emaizak.

Alarma faltsu bakarra gertatu da, kasu zehatz bat aurreikusi ez dugulako: **agintean iraun zuen bitarte hartan** egituran **bitarte** osagai beregainak ez du inesibo OBPArik, baina bai, bere ondorengo erakusleak.

Garapenerako corpus erroreduna anbiguotasun-maila desberdinetan analizatu dugunean, ez da inolako aldaketarik nabaritu (ikus V.26 taula).

	<i>Garapenerako corpusa</i>		
	Corpus erroreduna		
	Errore detektatu gabeak	Errore detektatuak	Alarma faltsuak
Desanbiguatu gabe	0	7	0
1.mailan	0	7	0
2.mailan	0	7	0
3.mailan	0	7	0
4.mailan	0	7	0

V.26 Taula: *Bitarte*. Anbiguotasunaren eragina emaitzetan.

V.3.3.4 Buruz

Osagai beregaina *buruz* duen postposizio-lokuzioaren inguruan aritu gara V.3.3 atalean erroreen detekziorako MGren erabilera erakutsi dugunean. Postposizio-lokuzio honi buruzko kasuistika V.27 taulan zehaztu dugu. Ohar bakarria egingo dugu *buruz* postposizioaren inguruan:

- **Alarma faltsuak ekiditea (POSTBURUZKO_1).** Errore-patroia zehaztu behar izan dugu, osagai beregainak lekuzko genitiboaz gain absoluti-boa hartzen duenean, egitura zuzena baita. Hau da, **zailtasunei buruzkoa izango da** edo **psikologiari eta kulturari buruzkoak izango dira** esaldiak zuzenak dira.

Iturria	Errorea	Forma	Adibideak
Corpus error.	POSTBURUZ_NAH_KASEZDATA_PA. PAn, ez datiborik, ez adlatiborik edukitzea.	not(-ra/-i) buruz	*katuaren buruz egin genuen.
	POSTBURUZKO_1. OBak lekuzko genitibo hartzen duenean, ondorengo hitza aditza kategoriakoa izatea.	-i buruzko ADI	*bolantari buruzko sortu izan diren istiluak
	PAk datiboaz gain lekuzko genitibo edukitzea.	-ko + -i buruz	*mundukori buruz

V.27 Taula: *buruz* OBa duen postposizioaren inguruko erroreak.

Emaitzak

Erroreen detekzioari begira esan dezakegu, alarma faltsuak kontuan izan gabe, postposizio-lokuzio honen detekzioaren inguruko emaitzak orain arteko okerrenak direla. Garapenerako erabilitako corpus erroredunean 5 errore detektatzen diren arren, 4 errore detektatu gabe geratu dira. % 55,55eko estaldura (9 erroretatik 5 detektatuta) eta % 71,4eko doitasuna (2 alarma faltsu) lortu dira. Esaldi erroredunak markatu gabe geratzearen arrazoiak, alarma faltsuak ekiditeko ipinitako baldintzan dago. Kasu batean (***kotxean gertatu zitzaiguni buruz hitz egin genuen**) postposizio-atzizkia daraman hitza ezezaguna delako (**zitzaiguni**); gainontzeko kasuetan, normalean erabilera zuzenekoak diren esamoldeetako bat agertzen delako: ***baina hori buruz zerbait esatea diote...** esaldi erroredunean **buruz esan** lokuzioa azaltzen da. Probarako erabilitako corpus erroredunean, % 71,4ko estaldura eta % 100eko doitasuna lortu ditugu. Detektatu gabe geratu zaizkigun bi esaldiek egitura bera dute, “-i buruzko *adberbioa*” egitura (adibidez, ***gizonak hitzei buruzko asko daki**).

Corpus zuzenean, *buruz* postposizio-lokuzioarekin oso errore gutxi detektatu dira eta adibide zuzen kopurua oso handia izanik, 2 eta 3 alarma faltsu gertatu dira garapenerako eta probarako corpusetan, hurrenez hurren.

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa
Elementu kopurua	560	373	5435	3623
Erroreak	9	7	?	?
Detektatu gabeak	4	2	?	?
Ongi detektatuak	5	5	3	2
Alarma faltsuak	0	0	2	3
Estaldura	% 55,5	% 71,4	-	-
Doitasuna	% 100	% 100	% 60	% 40
Alarma faltsuen tasa			% 0,15	% 0,13

V.28 Taula: *Buruz*. Emaizak.Alarma faltsuak

Buruz OBa duten hitzekin osatutako corpusean gertatutako alarma faltsuak arrazoi hauengatik markatu dira:

- **Analisi-errorea (1 AF)**. Kasu batean, **Menchu Gali buruz** sintagman, **Gal** abizena, **Galia** leku-izenaren aldaera bailtzen analizatu da. Kasu honetan postposizio-atzizkia ez da datiboa eta errorea markatzen da.
- **Anbigutasun semantikoa (3 AF)**. Postposizio-lokuzio honetan, osagai beregainaren lema semantikoki anbigua da. Gorputz atalarekin zerikusia duten perpausetan gertatu dira 3 alarma faltsu. Adibidez, **altueraz zein buruz txikiagoak direla** esaldian.
- **Landugabekoak (1 AF)**. Egitura zuzenak ekiditen aritu garenean, **buruz buruko** egitura landu dugu, baina **buruz burukako** egitura landu gabe utzi dugu. Azken egitura izan da alarma faltsu baten iturri.

Alarma faltsuak	<i>Desanbiguatu gabeko corpus analizatua</i>				
	Corpus erroreduna		Corpus zuzena		Guztira
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa	
Analisi-errorea	0	0	0	1	1
Anbig. semantikoa	0	0	2	1	3
Landugabekoak	0	0	0	1	1
Guztira	0	0	2	3	5

V.29 Taula: *Buruz*. Alarma faltsuak.

Anbiguotasunaren eragina

Anbiguotasunaren aldaketari dagozkion datuak V.30 taulan ikus ditza-kegu. Desanbiguatu gabeko testu-analizatuan detektatzen diren 5 erroretik, 3 errore detektatzera pasa gara 1, 2 eta 3 mailetan (4 errore 4. mailan²³). Errore gutxiago detektatzearen arrazoia, *buruz* hitzari esleitzen zaion gramatika-kategorian dago: adberbio kategoria galdu egiten du maila hauetan, eta beraz, detekzio-erregelak ez dira aplikatzen. Hori ekiditeko, desanbiguazio-gramatikan erregela espezifikoak idatz zitezkeen. Alarma faltsuei dagokien zutabea da emaitza hauetan bereziki aipatzeko beste kontu bat. 1 eta 2 mailetan, **erabaki** eta **komunikatua** hitzei *aditz* kategoria esleitzen zaie POSTBURUKO_1 erregelaren aplikazioa eraginez (ikus V.27 taula). Desanbiguazioa 3. mailan egiten denean, 4 errore gehiago markatzen dira honako egituretan: **arrantzari buruz**²⁴, **familiari buruz**²⁵, **kiolari buruz** eta **futbolari buruz**. Aurreko desanbiguazio-mailetan egitura horiei ez zaie errore-etiketarik esleitu, datibo markadun interpretazioa mantentzen zutelako. Desanbiguazioko 3. mailan egitura horiek guztiek duten analisi bakarra, datibo markarik gabekoa da, guztiak baitira zuzenak absolutibo mugagabea (ikus oin-oharrak). Hori horrela izanik, “not(-ra/-i) buruz” errore-patroia parekatzen dute. Laugarren mailan alarma faltsuen kopurua berriro jaisten da, datibodun interpreta- zioak mantendu egiten baitira.

	<i>Garapenerako corpora</i>		
	Corpus erroreduna		
	Errore detektatu gabeak	Errore detektatuak	Alarma faltsuak
Desanbiguatu gabe	4	5 (% 55,5)	2
1.mailan	6	3 (% 33,33)	3
2.mailan	6	3 (% 33,33)	3
3.mailan	6	3 (% 33,33)	7
4.mailan	5	4 (% 44,44)	1

V.30 Taula: *Buruz*. Anbiguotasunaren eragina emaitzetan.

²³Gogoratu 1., 2. eta 3. mailetan MG eta metodo estokastikoak erabiltzen direla desanbiguaziorako eta 4.enean MG soilik. Kategoria ezaugarria hartzen da kontuan 1. mailan, kategoria eta azpikategoria 2.enean, eta horietaz gain, kasua, 3.enean. Azken mailan, 4.enean, eskura dagoen informazio morfosintaktiko guztia erabiltzen da.

²⁴arrantzari: arrantzalea iparraldean.

²⁵familiari: ustez norbaiten lagun eta zerbitzari den deabrua.

V.3.3.5 Zehar

Erroreen detekzioari dagokionez arazo gutxien eman duen eta, aldi berean, emaitzei begiratuta emankorrena den postposizio-lokuzioa da oraintxe deskribatuko dugun hau. *Zehar* osagai beregaina duen postposizio-lokuzioaren erabilera okerrak V.31 taulan zerrendatu ditugu. Postposizio honekin gertatzen diren erroreak detektatzeko ez dugu baliabide berezirik behar izan. Egin dugun bakarra da atzetik marratxoa daramaten hitzak ekiditea. Adibidez, *zehar-galdera*, *zehar-lerro* edo *zehar-zentzu* modukoak baztertu ditugu.

<i>Desanbiguatu gabeko corpus analizatua</i>			
Iturria	Errorea	Forma	Adibideak
IZ ₁	POSTZEHAR_NAH_KASGEN_PA. PAn genitiboa izatea.	-ren zehar	*basoaren zehar etorri da
Corpus	POSTZEHAR_NAH_KASØ_PA. PAn markarik ez eramatea.	-Ø zehar	*espazio zehar ibili dira

V.31 Taula: *zehar* OBa duen postposizioaren inguruko erroreak.

Emaitzak

Emaitzetan, corpus guztietan % 100eko doitasuna eta %100eko estaldura lortu dira (ikus V.32 taula). Doitasuna maila guztietan baliorik hoberenean egon da, ez baita alarma faltsurik gertatu.

Anbiguotasun-maila aldatu denean (ikus V.33 taula) estaldura, ordea, aldatzen joan da eta kasurik okerreanean % 58,33ra iritsi da. Erroreak detektatu gabe geratu diren kasu guztietan, *zehar* osagai beregainak adberbio kategoria galdu duelako izan da.

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpusa	Proba corpusa	Garapen corpusa	Proba corpusa
Elementu kopurua	186	126	1500	1000
Erroreak	12	8	?	?
Detektatu gabeak	0	0	?	?
Ongi detektatuak	12	8	5	1
Alarma faltsuak	0	0	0	0
Estaldura	% 100	% 100	-	-
Doitasuna	% 100	% 100	% 100	% 100

V.32 Taula: *Zehar*. Emaitzak.

	<i>Garapenerako corpora</i>		
	Corpus erroreduna		
	Errore detektatu gabeak	Errore detektatuak	Alarma faltsuak
Desanbiguatu gabe	0	12 (% 100)	0
1.mailan	1	11 (% 91,66)	0
2.mailan	1	11 (% 91,66)	0
3.mailan	1	11 (% 91,66)	0
4.mailan	5	7 (% 58,33)	0

V.33 Taula: *Zehar*. Anbiguotasunaren eragina emaitzetan.

V.3.3.6 Laburbilduz

Aukeratutako bost postposizio-lokuzioei buruzko emaitzak V.34 taulan bildu ditugu. Corpus erroredunean garapenerako eta probarako corpusetan 60 eta 29 errore daude, hurrenez hurren.

	<i>Desanbiguatu gabeko corpus analizatua</i>			
	Corpus erroreduna		Corpus zuzena	
	Garapen corpora	Proba corpora	Garapen corpora	Proba corpora
Elementu kopurua	3884	2590	26679	17786
Erroreak	60	29	?	?
Detektatu gabeak	10	10	?	?
Ongi detektatuak	50	19	30	24
Alarma faltsuak	0	6	38	31
Estaldura	% 83,3	% 65,51	-	-
Doitasuna	% 100	% 76	% 44,11	% 43,63

V.34 Taula: Emaitzen laburpena.

Probarako corpusean lortutako doitasun-balioak laburbiltzen ditu V.35 taulak. Postposizio multzo osoa hartuz gero, % 53,75eko doitasuna lortu da, detektatutako 80 elementuetatik 37 alarma faltsuak baitira. Alarma faltsuen iturria honakoetan egon da: anbiguotasun morfologikoan (alarma faltsuetatik % 33,33), anbiguotasun semantikoan (% 34,66), sintaxi-anbiguotasunean (% 17,33), landu gabeko egituretan (% 8) eta prozesu automatikoko akatsetan (% 6,66). Gogoratu, horiek guztiak zehatz-mehatz deskribatu ditugula, postposizio-lokuzioetako bakoitzari buruzko emaitzak eman ditugunean. Alarma faltsu gehienak **arte** OB anbigua duen postposizioarekin gertatu direnez, kontaketa egin dugu postposizio hau alde batera utzita. Gainontzeko lau postposizioekin doitasun hobea lortu da (% 79,48, V.35 taulan bigarren errenkada) baina errore gutxiago detektatu dira. Sistema **bitarte**, **buruz** eta

zehir postposizio-lokuzioekin probatzen badugu, doitasuna % 90,62ra iristen da (hirugarren errenkada V.35 taulan).

Postposizio-lokuzioak	Doitasuna
arte, aurre, bitarte, buruz, zehir	% 53,75 (43/80)
aurre, bitarte, buruz, zehir	% 79,48 (31/39)
bitarte, buruz, zehir	% 90,62 (29/32)

V.35 Taula: Doitasunaren emaitzak probarako corpusetan.

Johannessen *et al.* (2002) egileen arabera, gramatika-zuzentzaile batean alarma faltsuen kopuru “onargarriak” ezin du % 30 baino handiagoa izan; hots, gutxienez alarmen % 70 benetako erroreak izan behar dute. Gure emaitzak muga horietan daude, nahiz eta balio horiek lortzeko, postposizio-lokuzio *emankorrenetara* murriztu behar izan dugun sistema.

MG erregelen kopurua postposizio batetik bestera asko aldatzen den arren (15 erregela behar dira **arte** postposizioa lantzeko eta 2 **zehir** lantzeko), hauen eraginkortasuna testuinguruko baldintzen konplexutasunaren eta anbiguotasunaren arabera da. Hau da, sistemak **arte** postposizio-lokuzioarekin lortu ditu doitasunari dagokionez emaitzarik okerrenak, postposizio horrek errorearen detekzioarako erregela kopuru handiena duen arren. Beste muturrean **zehir** postposizioa dugu, zeinak 2 erregelarekin doitasun hobereana lortzen duen. Azken emaitza postposizioaren anbiguotasun baxuagatik lortzen da. Orotara, 30 erregela definitu ditugu eta 5 postposizio-lokuzio.

Lan honek erakusten du postposizio-lokuzio guztiak lantzeko sistema bat eraginkortasunez garatzeko, aldezturik postposizioen anbiguotasunaren eta konplexutasunaren azterketa egitea komeni dela. Azterketa ez da zertan postposizioz postposizio egin, antzeko jokaera duten postposizio-lokuzioekin multzoak egin baitaitezke (adibidez, Aduriz *et al.* (2008) lanean, “-en -ko” postposizio-atzizki konbinazioarekin **aurrer**en multzo berean 50 postposizio landu dira, besteak beste, **aitzin**, **albo**, **alde**, **arte**, **atze**...). Guk 5 postposizio landu ditugu baina lana jokaera bera duten postposizio-multzoetara hedatzea uste dugu, landutako postposizioen kopurua handituz.

V.3.4 Zuzenketa

Postposizio-lokuzioen erabilera okerrak MG erabiliz testu analizatuan automatikoki etiketatu ondoren, ordain zuzenak emateko jarraitu dugun prozesua azaltzeari ekingo diogu. Zuzenketarako sorkuntza morfologikoa erabili

dugu. Sorkuntzarako moduluak, hitz baten lema, kategoria eta azpikategoria gramatikalak, numeroa eta kasua jasotzen ditu, eta izenak dioen moduan, hitz-forma zehatz bat sortu egiten du. Ikus dezagun hau guztia adibide baten bidez. Demagun **bidaiari** hitza sortu nahi dugula (bidaia datiboan). Honako datuak beharko genituzke sortzailearen sarreran: lema = *bidaia*, gramatika-kategoria = *izena*, azpikategoria = *arrunta*, numeroa = *singularra*, eta azkenik, kasua = *datiboa*. Datu hauekin (24) adibideko katea osatuko genuke eta sorkuntza egiten duen programan erabili.

(24) “bidaia” “[IZE][ARR]” “[NUMS]” “[DAT]”

Sorkuntzarako beharrezkoak ditugun datu ia guztiak errorea detektatzen dugunean lor ditzakegu. Jo dezagun, postposizio-lokuzioetan gertatzen diren erroreak detektatzen dituen gramatikak errore-etiketa ipini duela analizatutako egituretako batean. Adibidez, lehenago erakutsi dugun V.8 irudian, {POSTBURUZ_NAH_KASNOTDATALA_PA etiketa ipini dio **bidaia** hitzari. Etiketa duen interpretaziotik jasoko ditugu sorkuntzarako ezinbestekoak diren lema, kategoria eta azpikategoria gramatikalak eta numeroa. Hurrengo hitz-forma ere jasoko dugu analisitik, osagai beregaina agertzen deneko forma zehatza, postposizio-atzizki zehatz bat duena²⁶, emango baitugu ordain zuzenetan. **bidaia** hitzaren postposizio-atzizkia ezagutzea soilik falta zaigu sorkuntza egin ahal izateko. Hori ordea, ezin dugu analisitik jaso, hain zuzen ere, okerra delako.

```
0. "<bidaia>"
   "bidaia" IZE ARR DEK ABS MG AORG {POSTBURUZ_NAH_KASNOTDATALA_PA
   "bidaia" IZE ARR DEK ABS NUMS MUGM AORG {POSTBURUZ_NAH_KASNOTDATALA_PA
   "bidaia" IZE ARR ZERO AORG {POSTBURUZ_NAH_KASNOTDATALA_PA
   "bidaiatu" ADI SIN AMM ADOIN NOTDEK
1. "<buruz>"
   "buru" IZE ARR DEK INS MG
   "buruz" ADB ARR ZERO
```

V.8 Irudia: **bidaia buruz** egituraren analisi morfosintaktikoa, errore-etiketekin.

Postposizio-lokuzioen erabilera okerra detektatzeko garatu dugun murriztapen-gramatikan, aztertutako postposizio guztien erabilera zuzena deskribatu dugu ZUZ-SORKUNTZA definitutako atal batean. (24) esaldiari da-

²⁶ *buruz* postposizio-lokuzioaren kasuan ez dago aldaerarik baina *bitarte* postposizioaren kasuan, *bitartez*, *bitartean* edo *bitarteko* formak eduki ditzakegu aukeran.

gokion zuzenketa posiblea, adibidez, V.36 taulan erakusten den moduan definitu genuke.

PA	OBPA
-ALA	-ZERO
-DAT	-ZERO

V.36 Taula: Postposizio-lokuzioen zuzenketarako eskemaren adibidea.

Deskribapen honetatik jasoko ditugu PA zuzenak. Datu guztiak edukita, beharrezko ordain zuzenak sortuko ditugu: **bidaira buruz** eta **bidaiari buruz**.

V.3.5 Integrazioa XUXENG gramatika-zuzentzailean

IXA taldea aspaldi ari da XUXENG izeneko gramatika-zuzentzailearen oinarri izango diren errore-erregelen diseinuan eta proban. Erregelok Microsoft Word testu-prozesadorean integratzeko lanak taldeko proiektu batean hasi baziren ere (Otegi, 2003), azken urteotan erregela berrien integrazio-lanak Eleka²⁷ enpresak egin ditu.

XUXENG gramatika-zuzentzailearen lehen prototipoan (Otegi, 2003) informazio linguistikorik gabe detekta zitezkeen estilo- eta puntuazio-erroreak integratu ziren Word-en. Hala nola, puntu eta komaren ondorengo espazioak, esaldi luzeegiak, letra larriaren erabilera puntuazio-marken ondoren, parentesien eta antzekoen hasiera eta amaiera... Horiek integratzeko, Microsoft enpresako CGAPI izeneko API interfaze bat (*Aplikazioen Programaziorako Interfazea*²⁸) inplementatu behar izan zen. Bertan estilo- eta puntuazio-errore bakoitzeko azpiprograma bat garatu zen.

Bigarren prototipo batean (Otegi, 2006), informazio linguistikoa behar duten errorean integrazioa egin zen. Hain zuzen ere, postposizio-lokuzioetan gertatzen diren erroreak detektatzeko gramatikaren lehen bertsioa txertatu zen. Kasu honetan APIarekin egin beharreko lanari, hizkuntza-informazioa lortzeko beharrezko tresnak integratzeko lana gehitu behar izan zaio. Tokenizatzailea, analizatzaile morfosintaktikoa, desanbiguaziorako modulua eta errore sintaktikoak detektatzeko gramatika liburutegi dinamiko batean jarri dira. Prototipo honetan, Word testu-prozesadorean erabiltzaileak euskaraz idazten duenean, erroreren bat egiten badu, zuzentzaileak horren berri

²⁷IXA taldearekin elkarlanean dabilen enpresa da Eleka. Bere lanek ingeniarietza linguistikoa dute ardatz. <http://www.eleka.net>

²⁸API edo *Aplikazioen Programaziorako Interfazea* liburutegi batek eskaintzen duen funtzio eta prozedura multzoa da. Funtzio eta prozedura multzo hau beste software batek erabiltzen du.

emango dio. Ondoren, aurreko atalean aipatutako moduan sortutako postposizio-lokuzioaren ordain zuzenak erakutsiko zaizkio erabiltzaileari egokiena aukera dezan.

Atal honetan aurkeztu dugun lana XUXENG zuzentzailean integratzeko, nahikoa da hasierako gramatika azaldu berri dugunarekin ordezte.

V.4 Laburpena eta ondorioak

Atal honetan errore lokalak, bospasei hitzeko leihoetan gertatzen diren erroreak, detektatzeko eta zuzentzeko modua aztertutako dugu. Oinarrian egoera finituko mekanismoak erabiltzen dituzten bi tresna erabili ditugu bi hizkuntza-egitura tratatzeko: XFST erabili dugu datetan gertatzen diren erroreak lantzeko eta MG postposizio-lokuzioetakoak aztertzeko.

Erabilitako tresnei dagokienez, nahiz eta biek oinarrian egoera finituko mekanismoak erabili, XFSTk eta MGk ezaugarri desberdinak dituzte. XFSTk aukera ematen du hainbat eragile baliatuz automatik zuzenean kudeatzeko (konposatu, osagarria kalkulatu...). MG “egoera finituko murriztapen-gramatikak” definitzeko formalismoa da, eta helburu oso zehatza du: adierazpen erregularrak erabiliz egitura linguistikoak baztertzea edo mugatzea. Erroreen detekzio-prozesuan, bi baliabideok batez ere “markaketarako” erabili ditugunez, aipatutakoez gain, ez dugu erabilera-diferentzia handirik topatu²⁹. Hori bai, aipatzekoa da etiketak (gure kasuan errore-etiketak) ipintzeko modua: XFSTk egituraren hasiera eta amaiera mugatzeko etiketak aldi berean ipintzen ditu. MGk erregela-definizio bat behar du etiketa bakoitzeko eta horrekin egitura mugatzeko orduan akatsak gertatzeko aukerak handitzen dira. Erroreen zuzenketari dagokionez, MG eginkizun honetarako prestatuta ez dagoen bitartean, XFSTk sorkuntzarako erraztasun handiak ematen ditu eta zuzenketaren atazarako egokiagoa egiten du.

Landu ditugun hizkuntza-egiturak aukeratzerakoan, irizpide desberdinak erabili ditugu. Lehenago esan dugun moduan, euskararen inguruko adituek betetako itaunketa baten ondorioz, datetan erroreak maiztasun handiz gertatzen zirela jakin izan genuen. Egokiagoa litzateke corpus erroredun bat errore posible guztiekin etiketatzea eta agerpen-maiztasunak automatikoki erauzte, baina garestiegia da eskuzko lanari dagokionez. Postposizio-lokuzioak duten interesagatik landu ditugu batez ere, haien maiztasuna % lekoa dela estimatu baitugu.

²⁹XFSTren kasuan automataren *osagarria* erabil zitekeen egitura zuzenak definituta, egitura okerrak baztertzeko. Kasu honetan, ordea, zein errore zehatz gertatu den jakitea konplexua litzateke.

Datak eta postposizio-lokuzioak deskribatu ditugunean, haien arteko desberdintasunak aipatu ditugu (ikus V.1 atala). Horien artean emaitzei begiratuta ezaugarri garrantzitsuena anbiguotasuna da. Data *osoak* ez dira egitura anbiguoak. “urtea, hilabetea, eguna” segidak ia beti dira data, eta anbiguotasunik ezean, hauetako errorearen detekzio-tasa ez da era nabarrian aldatzen testua maila ezberdinetan desanbiguatzen denean. Kontrakoa da postposizio-lokuzioen kasua. Postposizio bakoitzaren arabera aldatzen den arren, orokorrean sintaktikoki eta semantikoki anbiguoak dira egitura hauek. Ondorio nabariak izan ditu ezaugarri honek postposizioak landu ditugunean. Anbiguotasunak sortzen dituen arazoak ebazten laguntzeko, hizkuntza-baliabide ugari eta ezberdinak behar izan ditugu, besteen artean, esamoldeen ekiditea eta hiztegieta definizioen *genusak* baliatuta erauzita-leku/denbora eta bizidun/bizigabe tasunak.

Baliabide hauek desanbiguatu gabeko testuaren gainean postposizio-lokuzioak detektatzeko erabili ditugu. Analizatuko testuaren desanbiguazio-maila aldatu dugunean, detektatutako errorearen emaitzek era nabarrian egin dute okerrera. Honen arrazoieta bat desanbiguazio-gramatikan erabiltzen den informazioan datza. Analisi-katea informazio zuzena analizatzeko prestatuta egoteak eragin negatiboa izan du postposizio ez-gramatikalak detektatzeko. Adibidez, informazio zuzenean oinarria duen desanbiguazio-gramatikak *buruz* osagai beregainarentzako *adberbio* kategoria aukeratzen du postposizio-atzizkia datiboa denean. Kasu honetan ***katuaren buruz** egituran ez dugu errorerik detektatzen, osagai beregainak *adberbio* kategoria galtzen duelako. Arazo horren aurrean, bi aukera proposatuko genituzke: anbiguo diren zenbait hizkuntza-egituretan errorearen detekzioa egiterakoan, testu analizatua desanbiguatu gabe erabili behar da edo, desanbiguazio-gramatika “erlaxatu” egin behar da gaizki osatutako egituretan errorearen detekzioa egin ahal izateko. Gure emaitzek erakusten dutenari kontrajarriaz, Birn-ek (2000) ezinbestekotzat jotzen du testu desanbiguatua erabiltzea “The relationship between disambiguation and grammar error detection is intricate. On the one hand, it is obvious that disambiguation is a prerequisite for any effort at precise error detection.”. Hori dela eta, desanbiguazio ondorengo testu “ez-egokia” rekin lan egiteko, beste proposamen bat egiten du: “take the disambiguator’s disturbed output as it is and write error detection rules based on that” (Birn, 2000). Gure ustez landuko den hizkuntza-egituraren arabera aldatzen dira bertan gertatzen diren errorearen detekzioarako desanbiguazio-beharrek eta erabili beharreko baliabideak.

Zuzenketari dagokionez, data eta postposizio-lokuzio okerren zuzenketa “orokorra” egin dugu. Hau da, egituran detektatutako errorea edozein izanda ere, ordain zuzen posible *guztiak* eman ditugu. Etorkizunean, zuzenketa

“bideratua” egiteko asmoa dugu. Hasierako egiturarekiko aldaketa gutxien dituzten ordain zuzenak emango ditugu lehenengo, eta gainontzekoak gero.

Emaitzei dagokionez, orokorrean egitura ez-anbiguoekin, datekin, lortu dira emaitza onenak; postposizio-lokuzioetan, emaitzak postposizioaren anbiguotasun semantikoaren eta sintaktikoaren arabera aldatu dira. Datekin probarako corpusean % 89,7ko doitasuna eta % 92,1eko estaldura lortu dira. Probarako corpusean postposizio-lokuzioekin probak egin ditugunean, emaitzak postposizioaren arabera aldatu dira. Osagai beregaina *bitarte* eta *zehar* duten postposizioekin oso esfortzu txikia eginez (4 eta 2 erregela), % 100eko doitasunak lortu dira (ikus kapitulu honetako taulak). Alderantzizkoa gertatu da, ordea, *arte* eta *buruz* postposizioen kasuan. Postposizio-lokuzio multzo osoa hartuz gero, % 53,75eko doitasuna lortu da. Alarma faltsu gehienak *arte* osagai beregain anbigua duen postposizioarekin gertatu direnez, kontaketa egin dugu postposizio hau alde batera utzita eta % 79,48ko doitasuna lortu dugu. Multzotik *aurre* postposizio-lokuzioa atereaz gero, % 90,62ko doitasuna lortu dugu. Esan beharra dugu, esku artean darabilgun adibide erroredunen kopurua txikia izanik, erakutsitako portzentajeak (doitasuna eta estaldura) ez direla oso esanguratsuak: adibide kopuru txiki batek portzentajeak asko aldatzen ditu. Emaitzak erakusteko ohiko modua delako erakutsi ditugu ebaluaziorako neurriak ehunekoetan.

Lana 5 postposizio-lokuzio zehatzetara mugatu dugun arren, uste dugu erraz heda daitekeela antzeko portaera duten postposizioetara (gogoratu, *aurre* postposizioaren antzeko portaera dutela *aitzin*, *albo*, *aldamen*, *alde*, *atze*... postposizio-lokuzioek “-en -ko” konbinazioarekin). Postposizio-lokuzio horietan errore berak gertatzen direla egiaztatuz gero, landutako postposizio-lokuzioen kopurua izugarri handituko litzateke lan handirik egin gabe.

Errorerik gabeko corpusak aztertu dira erroreen eta, batez ere, alarma faltsuen bila. Hori gure ustez, tesi-lan honen ekarpen aipagarria da, ez baitugu horrelako neurketarik maiz ikusi erroreen detekziorako sistemetan. Corpus zuzenak handiak dira, eta bertan topa daitezkeen egiturak maizago daude zuzen erabilita, gaizki erabilita baino. Hau da, erroreak ongi detektatzeko baino aukera handiagoa dago alarma faltsuak detektatzeko. Doitasuna definitzeko moduagatik, corpus zuzenetan ezaugarri honek gure kontra egiten du emaitzetan, baina sistema baten baliagarritasuna neurtzeko garrantzitsua iruditzen zaigu.

Alarma faltsuak topatu direnean, errore-erregelak birfindu egin dira egitura zuzenak ez markatzearren. Ekidin beharreko egitura orokorrak definitu ditugu, eta ez hitz-forma zehatzak. Azkenak erabilita, alarma faltsu kopurua asko jaitisiko litzateke baina, era berean, erregelak *gehiegi egokituko* lirateke testuetara (*overfitting*).

XFST eta MGrekin testuinguru lokala aztertuta detektatu daitezkeen erroreak landu daitezkeela egiaztatu dugu. Erroreak lokalak izanik, erregela berak erabili ditzakegu ordena desberdinetan antolatutako esaldietan erroreak detektatzeko. Adibidez, esaldi hauetan (*-ren zehar*) egitura okerra detektatzeko:

- (25) *Basoaren zehar joango gara amarekin.
- (26) *Joango gara basoaren zehar amarekin.
- (27) *Joango gara amarekin basoaren zehar.

Errorea detektatzeko aztertu behar diren elementuak auzokideak ez direnean, urrun geratzen direnean, aipatutako tresnak ez dira egokiak. Demagun, *subjektuaren eta aditzaren arteko komunztadura eza* detektatu nahi dugula (28)tik (33)erako adibideetan. XFST edo MG erabilita, errorearen detekzioarako erregela anitz definitu beharko litzateke, esaldian subjektuaren posizioa aditzarekiko aldatu egiten baita. Adibide hauek guztiek, ordea, mendekotasun-zuhaitz bera dute³⁰.

- (28) *Isilaldiak ez ditu bost edo sei segundo baino gehiago iraun behar.
- (29) *Isilaldiak ez ditu iraun behar bost edo sei segundo baino gehiago.
- (30) *Isilaldiak bost edo sei segundo baino gehiago ez ditu iraun behar.
- (31) *Ez ditu bost edo sei segundo baino gehiago iraun behar isilaldiak.
- (32) *Ez ditu iraun behar bost edo sei segundo baino gehiago isilaldiak.
- (33) *Bost edo sei segundo baino gehiago ez ditu iraun behar isilaldiak.

Kasu honetan, esaldiko ordena linealarekin baino, esaldiaren egiturarekin lan egitea egokiagoa iruditzen zaigu. XFSTk eta MGk aukera hau ematen ez dutenez, mendekotasun-zuhaitzen gainean patroierregelak aplikatzen dituen *Saroi* izeneko tresna sortu dugu. Hurrengo kapituluan (VI. kapituluan) tresna honen inguruan arituko gara. Tresna komunztadura-erroreak detektatzeko nola erabili dugun VII. kapituluan aztertuko dugu.

³⁰Adibide hau <http://www.ehu.es/PAT/zuzenketak/komunztadura.htm> helbidetik jasoa izan da, eta gero, eraldatua.

TESTUINGURU ZABALEKO ERROREEN DETEKZIOA

VI. KAPITULUA

***Saroi*: mendekotasun-zuhaitzetan informazioa kontsultatzeko tresna**

VI.1 Sarrera

Hizkuntzaren analisia egiten dugunean, zenbait kasutan aztertu beharreko egituretako osagaiak esaldiaren ordena linealean urrun kokatuta egon daitezkeela esan dugu. Gainera, euskaraz esaldiko elementuen ordena librea izanik, egitura bereko osagaiak ordena ezberdinetan ager daitezke. Kasu horietarako ikusi dugu, esaldi horiek ordena linealean aztertzea baino, haien *egiturazko ordena* edo analisi-zuhaitza aztertzea egokiagoa dela. Orain arte erabilitako tresnek (XFSTk eta MGk) egiturazko ordena lantzen ez dutenez, tresna propioa diseinatu eta eraiki dugu. Tresnak *Saroi* du izena eta bizpahiru urte ditu.

Izenaren aukeraketak badu bere zergatia. *Saroi*k “*Sintaxi ARboletan OI-narrizko bilaketak*” hitzen akronimoa izan nahi du. Gainera, hiztegieta¹ begiratzuz gero, saroia edo sarobeia, abereak eta abeltzainak gauez egoteko tokiaz gain, *basoa* edo *oihana* da. Gure ustez izen aproposa, besteren artean esaldiei dagozkien analisi-zuhaitz posibleen multzoak, *saroiak*, sortzen dituen tresnarentzat.

Hasiera batean *Saroi*ren sorrera testuinguru linguistiko zabaletan, lauzpabost hitz baino gehiagoko analisi-leihoetan, gertatzen diren erroreak detektatzeko eman bazen ere, laster konturatu ginen tresna orokorragoa izan zitekeela. *Saroi*ren helburu nagusia esaldi erreal bati dagozkion *mendeko-*

¹Ikus adibidez, Elhuyar Hiztegia (Elhuyar Fundazioa, 2006) eta Harluxet Hiztegi Entziklopedikoa (Klaudio Harluxet Fundazioa, 2001)

tasun-zuhaitzetan informazioa bilatzea edo kontsultatzea da. Horretarako, besteak beste honako zereginak hartu beharko ditugu gure gain:

- Anbiguitasun sintaktikoaren kudeaketa.
- Informazio linguistiko osagarria erabiliz hizkuntza-informazioaren abetasketa.
- Kontsulta-erregelak idazteko zehaztapen-lengoaia abstraktu eta erazagutzaille baten zehaztapena.
- Lortutako emaitzen aurkezpena.

*Saroi*k analizatzen duen analisi-zuhaitzen multzoa automatikoki, erdiautomatikoki edo eskuz sortua izan daiteke. Azken bi moduetan sortutako zuhaitz multzoei *treebank* deitzen zaie, eta hauei hizkuntzalaritza konputazionalan egotzi izan zaien ohiko erabilera baterako erabiliko dugu *Saroi*: fenomeno linguistiko zehatzen azterketarako². Zeregin horretarako, gramatika teorikoetan deskribatuta dauden fenomenoak *treebank*aren edo automatikoki sortutako corpusaren kontra egiaztatzen dira, kontsultak bide direla (ikus VI.2.4.1 atala). Kontsulta-erregelen bidez deskribatutako hizkuntza-fenomenoak erroreak direnean, *Saroi*k erreoreen detekzioarako balioko du, baina hau aplikazio bat besterik ez da izango, garrantzitsua, baina ez bakarra. Kapitulu honetan, beraz, *Saroi*ren ikuspuntu orokor bat emango dugu eta hurrengo kapituluan (VII. kapitulua) *Saroi*ren bi erabilera posible zehaztuko ditugu. Horien artean, guretzat garrantzitsuena den aplikazioa: komuntadura-erreoreen detekzioa.

*Saroi*ren erabiltzaileak, gehienetan, hizkuntzalariak izango dira; hizkuntza-ezaugarrien inguruan ezagutza zabala duten pertsonak, baina ez dutenak zertan jantziak egon informatika kontuetan. Hori horrela izanik, *Saroi* aplikazioak erabilerraza eta sinplea izan behar du, malgua eta ahal izanez gero interfaze intuitiboduna. Gainera, maila desberdinetako (morfologia, morfosintaxia, sintaxia...) hizkuntza-ezaugarri ahalik eta gehienak lantzeko gai

²Hizkuntzalaritza konputazionalan *treebank*en beste erabilera ohiko bat, parse-rren ikasketa eta ebaluazioa da. Ikasketari dagokionez, gaur egun metodo enpirikoak (estatistika eta ikasketa automatikoa erabiltzen dutenak) erabiltzen dituzten algoritmoak trebatzeko eta lortutako emaitzak ebaluatzeko ohikoa da *treebank*ak erabiltzea. Nahikoa da *Conference on Computational Natural Language Learning 2007* <http://www.cnts.ua.ac.be/con11/> moduko konferentzian egindako lana begiratzea, *treebank*en garrantziaz ohartzeko. Ebaluazioari dagokionez, bai metodo enpirikoekin zein sinbolikoekin sortutako analizatzaileak ebaluatzeko da egokia *treebank*a, hizkuntzalariak eskuz zuzenduta egoten baita.

behar du izan, kontsulten emaitzak hizkuntzalarientzat ahalik eta erabilgarrien izan daitezen. Hizkuntzalariei lana erraztearren, kontsultak zuzenean hizkuntza-egituren gainean, mendekotasun-zuhaitzen gainean, egingo dira eta ez beste zenbait sistematan egiten duten moduan (Bouma eta Kloosterman, 2002; Cassidy, 2002) dokumentu-egituren gainean.

*Saroi*ren helburu nagusia aztertu dugu, baita bere erabiltzaileen eskakizunak ere. Jarraian, *Saroi*ren egitura edo arkitektura aztertuko dugu.

VI.2 Arkitektura nagusia

Software-arkitekturak sistemaren egitura ematen du, softwarea zein modutan antolatzen den, eta sistemako osagaien betebeharrak zeintzuk diren deskribatzen du. Gainera, osagaien arteko elkarrekintzek sistemako eskakizunak betetzen dituztela ziurtatzen du (Shaw eta Clements, 1997).

Saroi sistemaren software-arkitektura inplizitua da: ez du estandar edo patroiz ezagunik jarraitzen, diseinu-atazan sortutako beharrak asetzeko garatu da. Softwarea antolatzeko moduari dagokionez, *objektuei zuzendutako arkitekturaren* bidetik doala esango genuke eta bertan erabiltzen diren ereduak modelatzeko *Unified Modeling Language (UML)* (Fowler eta Scott, 1997; Booch *et al.*, 1997) modelatze-lengoaia erabili dugu (ikus UMLri buruzko informazio zabalagoa VI.2.1.2 atalean).

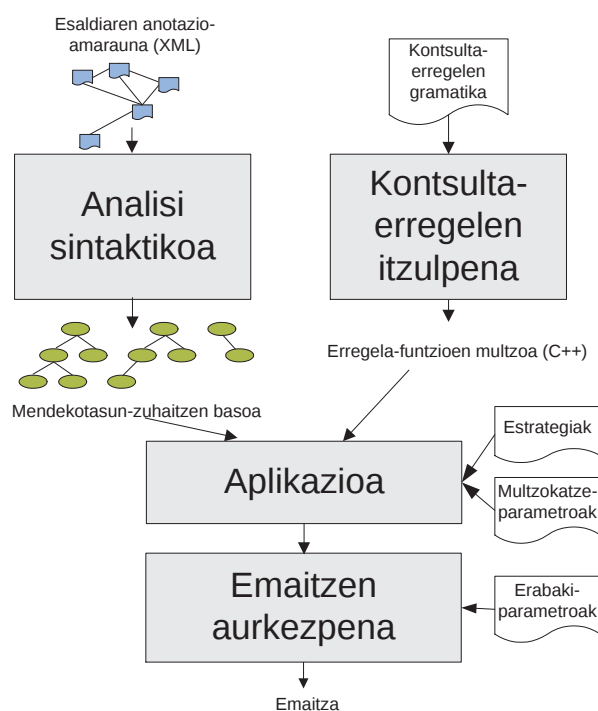
*Saroi*ko software-unitate bakoitzerako, honek beteko dituen funtzioak eta dituen helburuak nahiko argi bereizita daude. Jarraian, *modulu* (ikus VI.1 irudia) bakoitzaren betebeharrak labur-labur zerrendatuko ditugu:

- **Analisi sintaktikorako modulu**a. Modulu honek analisi-kateak sortutako anotazio-amarauna jasotzen du eta bertatik abiatuta, esaldi bakoitzari dagozkion mendekotasun-zuhaitz posible guztiak sortzen ditu. Mendekotasun-zuhaitzotan aurki dezakegun hizkuntza-informazioari, *abasketa*-prozesu baten bidez, informazio osagarria gehi dakioke.
- **Kontsulta-erregelen itzulpenarako modulu**a. Aldez aurretik eta *Saroi*ko bereziki definitu dugun kontsulta-lengoaia batean idatzitako erregela-gramatika jaso, eta erregela bakoitza helburu kodean³ idatzitako funtzio batera itzultzen du modulu honek.
- **Kontsulta-erregelen aplikaziorako modulu**a. Mendekotasun-zuhaitzak eta erregela itzuliak jaso ondoren, erregelak zuhaitzetan aplikatzen dituen modulu da hau.

³Helburu kodea C++ da kasu honetan.

- **Emaitzen aurkezpenarako modulua.** Aurreko modulua aplikatzearen ondorioz, erregeletan zehaztutako kontsultak betetzen dituzten zuhaitzak ezagutuko dira. Emaiza horiek zehaztu eta aurkeztu egiten ditu emaitzen aurkezpenarako moduluak.

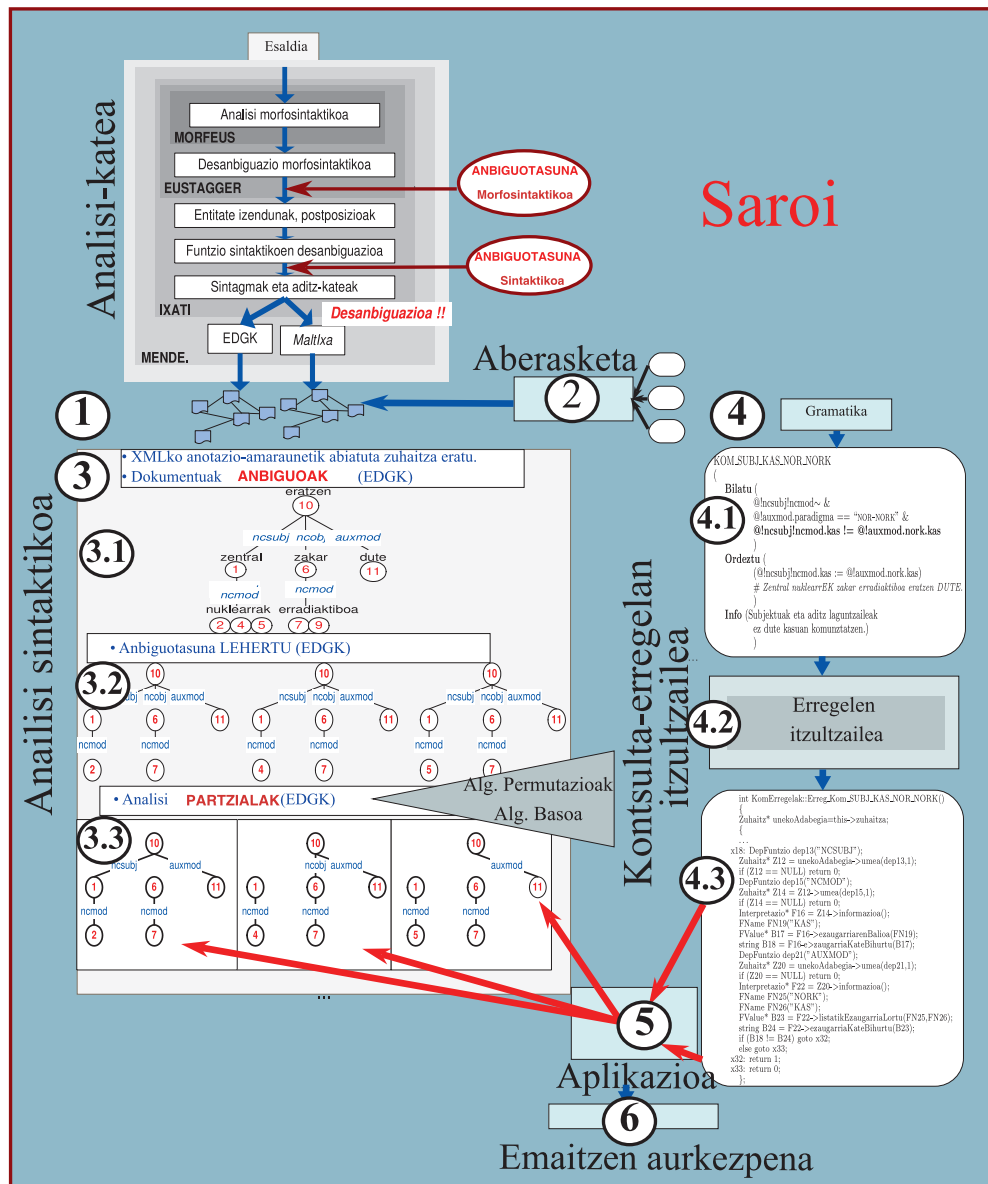
Aipatutako moduluen *erabilera-ordenari* dagokionez, analisiari eta erregelei dagozkien moduluak ez dira zertan modu sekuentzialean erabili. Modu guztiz independentean erabil daitezke, beren artean ez baitute inolako mendekotasunik. Erregelak zuhaitzetan aplikatzeko, ezinbestekoa da zuhaitzak eta erregela-itzuliak sortuta egotea. Gauza bera gertatzen da emaitzak aurkezteko moduluarekin, emaitzak aurkeztu ahal izateko, erregelak zuhaitzetan aplikatuak behar ditugu.



VI.1 Irudia: *Saroi*ren arkitektura nagusia.

*Saroi*ekin mendekotasun-zuhaitzetan kontsultak egiteko jarraitzen dugun prozesua, VI.2 irudian laburbildu dugu. Jarrai bitez irudiko zenbakiak, ondorengo zerrendakoekin batera:

1. Prozesua hasteko, *Saroik* analisi-katetik sortutako anotazio-amarauna irakurtzen du. Analisiak katetik *anbiguotasun morfosintaktikoarekin* eta *anbiguotasun sintaktikoarekin* jasotzen dira (irudian bunbuilo baten barruan adieraziak).
2. Komenigarria irudituz gero, anotazio-amarauneko dokumentuak hizkuntza-informazioarekin aberasten dira (aberasketarako erabiliko den informazioa *Saroiren* erabileraren arabera izango da).
3. Ondoren, *Saroiren* analisi sintaktikorako modulua erabilita, analisiak landu egingo dira. Modulu honekin zerikusia duten zenbait ezaugarri honakoak dira:
 3. 1. Anotazio-amaraunean mendekotasun-zuhaitza adierazten duen dokumentua anbigua da.
 3. 2. Anbiguotasun hori *lehertu* egin dugu, zuhaitz anbigua mendekotasun-zuhaitz ez-anbiguoetan banatuaz. Zuhaitz ez-anbiguoen multzo honi *basoa* (*saroia*) deituko diogu.
 3. 3. EDGK analizatzaile sintaktikoa erabiliz sortutako mendekotasun-zuhaitzak *partzialak* dira. *MaltIxarekin* sortutakoak ez, baina jasotzen duten tratamendua berbera da.
4. Bestaldetik, *Saroiren* kontsulta-erregelen itzulpenerako moduluak, mendekotasun-zuhaitzetan egin nahi diren kontsultak bilduta dauzkan gramatika bat jasotzen du. Gramatika horretako erregelak *Saroiren* kontsulta-lengoaian idatzita daude, eta zuhaitzetan aplika daitezkeen funtzioetara itzuli behar dira. VI.2 irudian atal honi dagokionez honakoak zehaztu ditugu:
 4. 1. *Saroiren* kontsulta-erregelaren adibidea.
 4. 2. Itzultzaileak sarreraren analisi lexikoa egiten du, eta honetan lortutako tokenak, analizatzaile sintaktikoari pasatzen dizkio. Sintaxiak zuzendutako itzulpen-eskema bat erabilita (Aho *et al.*, 1985), erregelako C++-en idatzitako funtzio bat lortzen da.
 4. 3. Kontsulta-erregelari dagokion funtzioa.
5. Kontsulta-erregelen aplikaziorako moduluak, erregelak mendekotasun-zuhaitzetan aplikatzen ditu.
6. Emaizten aurkezpenerako moduluak, emaitzak eskaintzen ditu.



VI.2 Irudia: Saroiaren eskema.

Modu oso eskematikoan deskribatutako moduluak (ikus VI.1 irudia), sakonean zehaztuko ditugu ondorengo ataletan. Horretarako, erabilitako ereduak eta algoritmoak zehaztu ditugu. VI.2.1 atalean analisi sintaktikoaren berri emango dugu. Segidan, VI.2.2 azpiatalean, mendekotasun-zuhaitzen gainean kontsultak egiteko lengoaia eta honen itzulpena aztertuko ditugu. Ondoren kontsulta-erregelen aplikaziorako moduluaren berri emango dugu VI.2.3 azpiatalean. Eta bukatzeko, VI.2.4 azpiatalean, *Saroi*ren emaitzak mugatzeko, zehazteko eta aurkezteko modua azalduko dugu.

VI.2.1 Analisi sintaktikorako modula

Analisi sintaktikorako moduluaren helburu nagusia, sarrerako esaldi bakoitzari dagozkion mendekotasun-zuhaitz guztiak eraikitzea da. Helburua lortzeko, sarreran XMLz kodetutako anotazio-amarauna erabiliko da, amarauneko dokumentuetan baitago zuhaitzak eraikitze beharrezkoa den informazio linguistiko guztia (ikus IV. kapitulua).

Saroi tresnak sarreran jasotzen duen XMLz kodetutako anotazio-amarauna bai automatikoki, baita erdi-automatikoki ere sor daiteke. Amarauna automatikoki sortzeko, euskararako analisi-katea erabiliko genuke (ikus IV kapituluko IV.1 atala). Analisiei eskuzko zuzenketa gehitzen diogunean, analisi erdi-automatikoa egin dela esaten da. Berrikuste-prozesurako bi anotazio-tresna erabili ditugu taldean: datu-base lexikalak, segmentatzaileak, analizatzaile morfosintaktikoak eta lematizatzaileak emandako emaitzak eskuz errepasatzeko, *Eulia* (Artola *et al.*, 2004b) tresna erabili dugu; dagoeneko desanbiguatutako dokumentuak erabiliz mendekotasun-erlazioak eskuz ezartzeko, berriz, *Abar-Hitz* (Díaz de Ilarraza *et al.*, 2004) tresna grafikoa erabiliko dugu. Tresna hauen bidez etiketatzean sor daitezkeen erroreak minimizatu egiten dira eta informazio *garbia* sortzen da.

Anotazio-amaraunak markaketarako *eskema banatua* deritzona (ingelesez *stand-off*) erabiltzen duela aipatu dugu dagoeneko. Bertan, informazio linguistikoa aparteko dokumentuetan jasotzen da (segmentazioa batean, morfologia beste batean, sintagmak hirugarren batean e.a.) eta informazioa lotzeko estekak erabiltzen dira. Guri, zuhaitzen gainean kontsultak egin ahal izateko, hizkuntza-informazio hau guztia elkarrekin egitura berean edukitzeak lana izugarri errazten digu. Helburu horrekin definitu eta garatu ditugu atal honetan azalduko ditugun eredu eta egiturak.

Ereduak eta egiturak definitzea erabaki aurretik, maila ezberdinetako hizkuntza-informazioa XML dokumentu bakarrean biltzearen aukera aztertu genuen. Hori erraz egin liteke *XSLT* (*Extensible Stylesheet Language Trans-*

formations) moduko lengoaia bat erabiliz (ikus VI.2.2.1 azpiatala) baina gure ustez hainbat desabantaila ditu aukera horrek:

- Sortutako dokumentuaren tamaina oso handia litzateke, honek dakarren espazio-galerarekin.
- Dokumentuak hierarkia-maila ezberdinetako informazioa bilduko luke, eta honen ondorioz kontsultak konplexuak lirateke.
- XMLz kodetutako anotazio-amaraunean mendekotasun-erlazioak tokenen artean baino, haien interpretazioen artean ezartzen dira. Hau horrela izanik, mendekotasun-erlazioak gordetzen dituzten dokumentuetan ez da zuhaitz bakarra errepresentatzen, n zuhaitz biltzen dira. Zuhaitzak era egokian banatzeko algoritmoa XML dokumentu baten gainean aplikatzea zailagoa litzateke. Gainera, zuhaitz horietan zenbait interpretazioen informazioa berdina denez, informazioa errepikatu egingo litzateke.

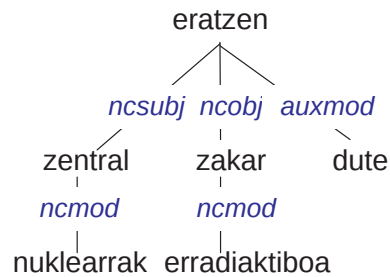
Arrazoi horiek direla eta, guk jarraitu dugun inplementazio-estrategia gure egitura eta eredu propioak diseinatu eta inplementatzea izan da. Jarraian hauen berri emango dugu. VI.2.1.1 atalean sarrera moduan jasoko ditugun XML dokumentuen berri emango dugu. Dokumentuetan topa dezakegun anbiguitasuna eta hori islatzen duen *basoa* egituraren nondik norakoak emango ditugu VI.2.1.2 atalean, ondoren, basoaren parte den *zuhaitza* egitura VI.2.1.3 atalean azaltzeko. Azkenik, basoko zuhaitzak errepresentatzeko erabilitako dokumentuak VI.2.1.4 atalean erakutsiko ditugu.

VI.2.1.1 Sarrerako dokumentuak

Analisi sintaktikorako moduluak, esaldi bati dagozkion mendekotasun-zuhaitzak eraiki ahal izateko, analisi-kateko askotariko informazioa erabiltzen du sarreran. Atal honetan zuhaitzak eraikitzeke erabiliko dugun informazioa eta hau biltzen duen XML dokumentu multzoa zehaztuko ditugu.

Demagun (34) adibideko esaldiari dagokion mendekotasun-zuhaitza sortu nahi dugula. Zuhaitz hori VI.3 irudian ikus dezakegu. Analisi-katea erabiltzaile lortutako anotazio-amaraunetik, honako dokumentuak erabiliko genituzke:

(34) *Zentral nuklearrak zakar erradiaktiboa eratzen dute.



VI.3 Irudia: **Zentral nuklearrak zakar erradiaktiboa eratzen dute* esaldiari dagokion mendekotasun-zuhaitza.

1. *Adabegietan* esaldiko token edo multitoken bakoitzari dagokion interpretazio bakarra gordetzen da, informazioa honako dokumentuetatik jaso ondoren:
 - *Tokenak* (*w.xml* dokumentua). Tokenizatzailaren emaitza, hau da, testuan bereizitako token bakunen zerrenda. Jatorrizko dokumentuko elementuei (hitzei, puntuazio-markei...) erreferentzia egiten zaie lehenago ere aipatu izan ditugun erakusle hedatuak (*<xptr >*) erabiliz.
 - *Multitokenak* (*mwlnk.xml*). Hitz anitzeko unitate lexikoen egitura deskribatzen duen dokumentua. Testuko multitokenen zerrenda.
 - *Analisi morfosintaktiko desanbiguatua* (*lem.xml*). Jatorrizko testuko token eta multitokenei dagozkien lematizazio desberdinen multzoa. Analisiak desanbiguatuta daude.
 - *“Token-analisi” estekak* (*lemlnk.xml*). Dokumentuko token eta multitokenen, eta beren lematizazioen arteko estekak. Lehenago esan bezala, anbiguotasuna token edo multitoken bati hainbat analisi lotuaz adieraz daiteke⁴. Zuhaitzeko adabegi bakoitzean analisi bakarra edukiko dugu, nahiz eta dokumentu honetan anbiguotasuna eduki.
 - *“Token-analisi-funtzio” estekak* (*lsfi.xml*). Token edo multitoken, lematizazio-analisi eta funtzio sintaktikoarekin osatutako hirukoteak dituen dokumentua. Token edo multitoken batek lematizazio

⁴Gogoratu, desanbiguazio-prozesuan ez dela anbiguotasun guztia ebazten.

berarentzat hainbat funtzio sintaktiko eduki ditzake esleituta; beraz, sintaxi-funtzioekiko anbiguotasuna adierazteko erabiltzen da dokumentu hau.

2. Adabegien arteko mendekotasun-erlazioak adierazten dituzten [arkuak](#) ezartzeko, zerrenda honetako XML dokumentuak erabili ditugu:

- *Dependentzia bikoteak (dep.xml)*. Zuhaitzetako adabegien artean gobernatzaile-mendeko erlazioak adierazteko, interpretazio bikoiteen zerrenda. Analisi-interpretazioak “*token, analisi, sintaxi-funtzio*” hirukoteen bidez adierazten dira eta mendekotasun-erlazio bat bi analisi-interpretazioen artean ezartzen da. Lehenago aipatutako *lsfi.xml* dokumentuko osagaiei egiten die erreferentzia dokumentu honek.
- *Dependentzia-etiketak (deplib.xml)*. Mendekotasun-etiketen bidez adierazten diren erlazioen (NCSUBJ⁵, NCMOD⁶...) deskribapena (ikus etiketa hauek IV. kapituluko IV.5 irudian).
- *Dependentzia-etiketa eta bikoteen arteko estekak (deplnk.xml)*. Lehenengo dokumentuan (*dep.xml*) ezarritako mendekotasunen eta dagozkien etiketen arteko estekak.

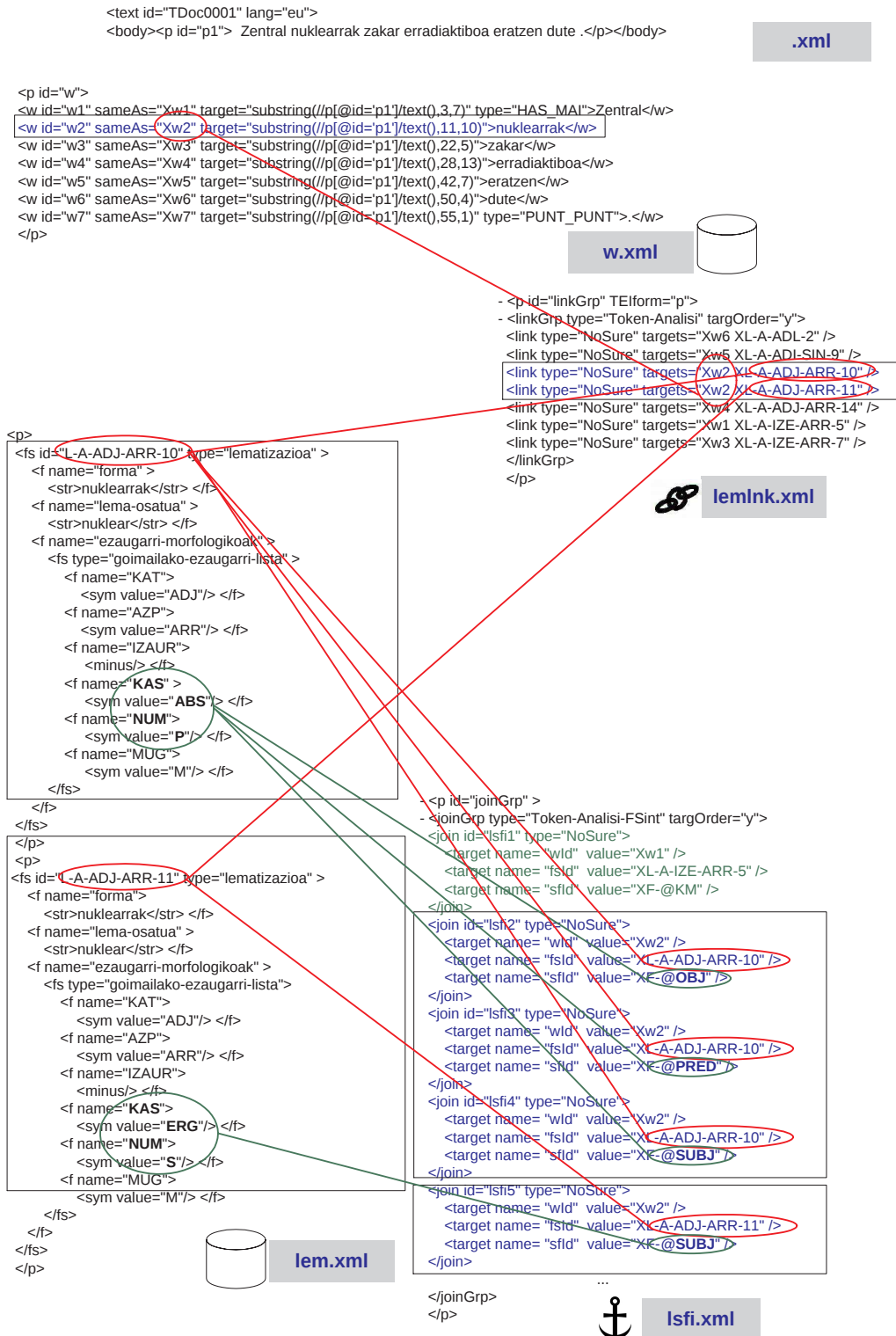
Analisi-amarauneko XML dokumentu hauek guztiek nolako informazioa eta zein modutan gordetzen duten VI.4 eta VI.5 irudietan azaltzen den adibidean ikus daiteke.

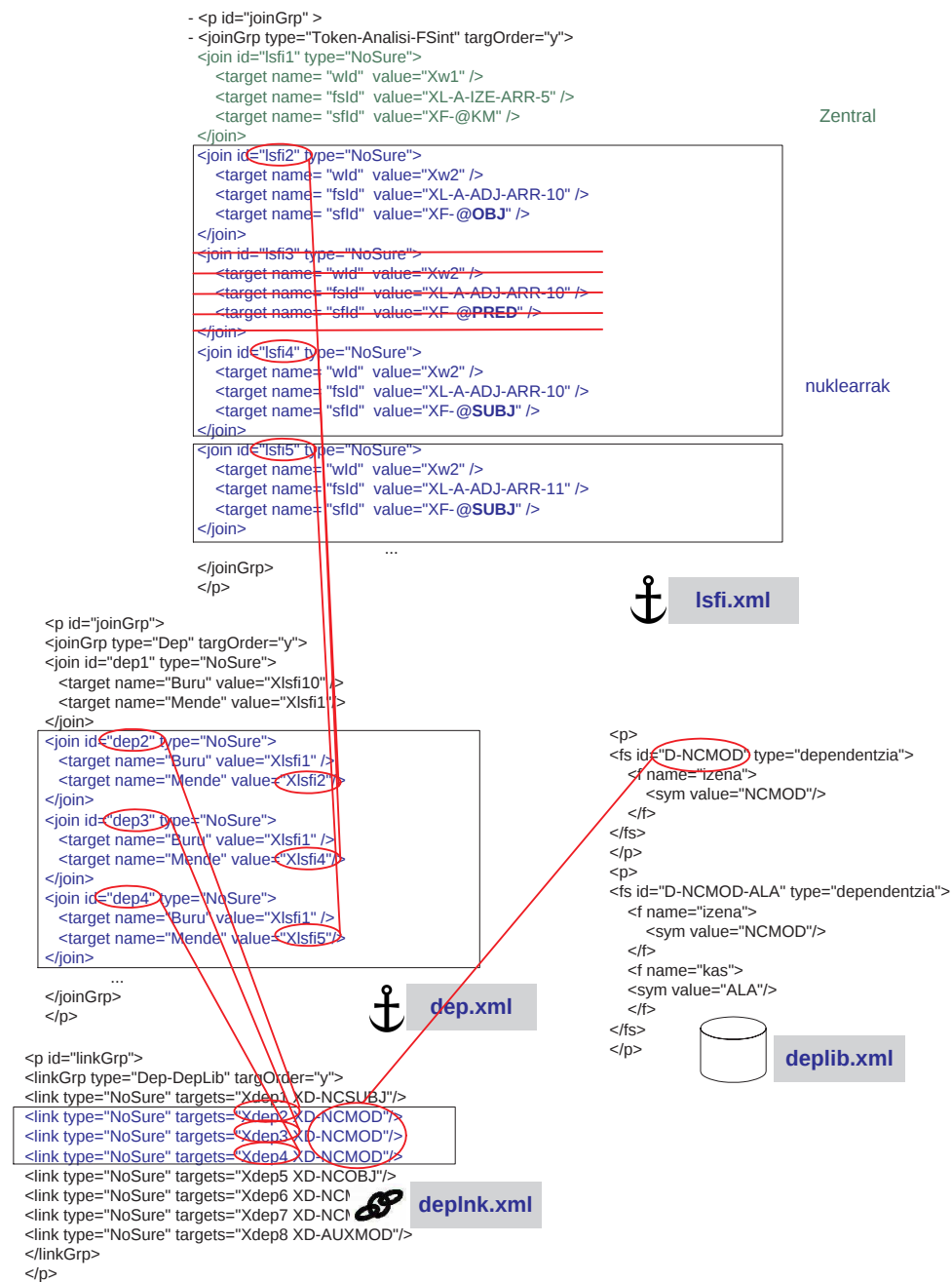
Mendekotasun-zuhaitzak eraikitzeke erabilitako XML dokumentuak ikusi ondoren, arrazoi ezberdinengatik behar ez ditugun dokumentuak aipatuko ditugu jarraian:

- *Segmentazioari eta morfosintaxiari* dagozkienak. Segmentatzaileak eskaintzen duen informazio linguistikoa morfosintaxiak *landu* egiten du. Era berean, EUSTAGGERek, informazio morfosintaktikoa *desanbiguatu* egiten du. Azkenean, tokenei buruzko informazio elaboratua eta desanbiguatua, lematizazioaren ondorengo fitxategietan topatzen da, gainontzekoak bazter daitezkeelarik tesiaren helburuetarako.
- *Sintagmei eta aditz-katei* dagozkienak. Unitate hauek mendekotasun-zuhaitzetan ez dute zentzurik, hauetan tokenen arteko erlazioekin lan egiten baita. Sintagma eta aditz-kateen osagaiak adabegi bereizietan agertzen dira mendekotasun-zuhaitzetan.

⁵NCSUBJ: *Non Clausal* SUBJECT (perpauza ez den subjektua).

⁶NCMOD: *Non Clausal* MODifier (perpauza ez den modifikatzailea).





VI.5 Irudia: XML anotazio-amaraunaren adibidea (2).

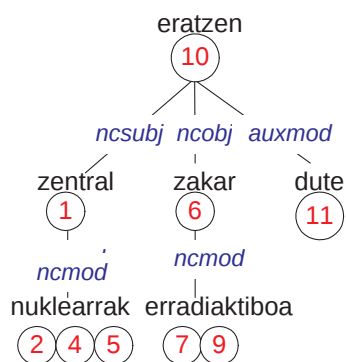
Anotazio-amaraunaren parte den (ikus IV. kapituluko IV.2 irudia) baina erabiltzen ez den hizkuntza-informazioa aztertu ondoren, garrantzi berezia duen dependentzia-erlazioen (*dep.xml*) dokumentuaren inguruan arituko gara hurrengo lerroetan. Dokumentu horrek, mendekotasun-erlazioen batean parte hartzen duten esaldiko token/multitokenen interpretazio-pareak lotzen ditu. Honako hiru ezaugarriengatik iruditzen zaigu aipagarria:

1. **Analisi-unitatea.** *Mendekotasun-erlazioak interpretazioen artean ezartzen dira eta ez tokenen artean.* Aukera honen zergatia hizkuntza guztiek duten *berezko anbiguotasunean*, eta hau adierazteko beharrean, dago. Adibidez, **Txakurrak ikusi ditu txoriak** esaldian **txakurrak** hitzak, bai subjektuaren papera, baita objektuaren papera ere, har dezake. Mendekotasun-zuhaitzean *hitza* izango balitz loturarako unitatea, ezin go genuke berezko anbiguotasun hau modu egokian adierazi, **Txakurrak** eta **ikusi** hitzen artean mendekotasun-erlazio etiketa bakarra ipintzeko aukera egongo bailitzake. Mendekotasun-erlazioak interpretazioen artean ezartzea beharrezkoa da, eta batzuetan interpretazio bakoitzarekin lotuko den mendekotasun-erlazioa desberdina izango da (aurreko kasuan, NCSUBJ kasu batean, eta NCOBJ⁷ bestean). Ezaugarri horrek zailtasunak gehitzen dizkio gure lanari.
2. **Anbiguotasuna.** Erlazioak interpretazioen artean ezartzen direnez, normalean esaldi baten analisiaren ondorioz mendekotasun-zuhaitz bat baino gehiago lortuko da. Berezko anbiguotasunari, analisi automatikoa erabiltzen den kasuetan EUSTAGGERen desanbiguazio-akatsen ondorioz sortutako anbiguotasuna gehitzen zaio. Azterketa helburu dugun *dep.xml* dokumentuan mendekotasun-erlazioetan parte hartzen duten interpretazio-pare guztiak adierazten dira zuhaitz berean erabil daitezkeen ala ez begiratu gabe. Token bakoitzeko interpretazio bakarra izango da aldioro mendekotasun-erlazio batean parte-hartzaile. Adibidez, VI.6 irudian ikus dezakegunez, **zentral** hitzak interpretazio bakarra du eta **nuklearrak** hitzak hiru. **zentral** hitzeko “1” zenbakidun interpretazioa⁸, **nuklearrak** hitzeko “2” zenbakidun interpretazioarekin, edo “4” duenarekin, edo “5” duenarekin lotu ahal izango da, baina “1-2”, “1-4” eta “1-5” pareak ez dira aldi berean gertatuko. Hori kontuan hartuta, VI.6 irudiko zuhaitzetik abiatuta, VI.7 irudiko sei zuhaitzak lortuko ditugu.

⁷NCOBJ: *Non Clausal Object* (perpauza ez den objektua).

⁸“1” zenbakia ipini diogu interpretazio honi, *lsfi.xml* dokumentuan *lsfi1* identifikadorea baitu. Gauza bera egin dugu gainontzeko interpretazioekin.

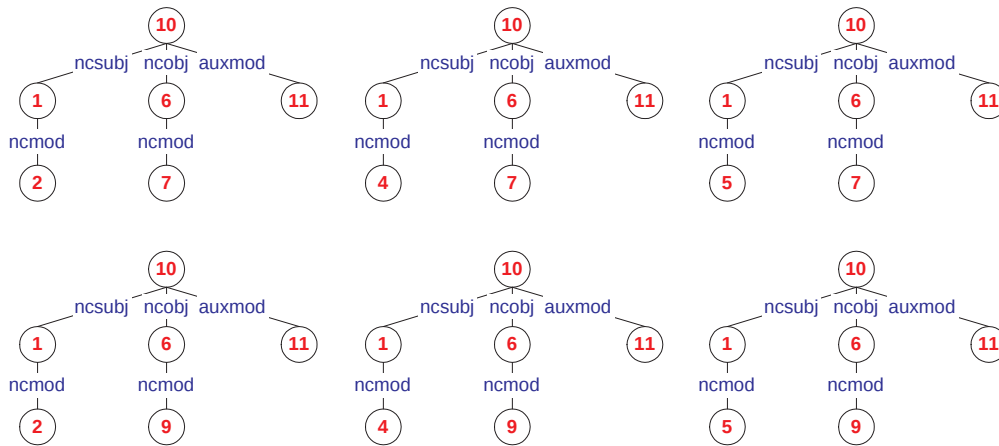
3. **Kontzeptuen nahasketa.** Mendekotasun-erlazioak adierazten dituen dokumentuan modu berean adierazten dira kontzeptuz desberdinak diren bi fenomeno: gobernatzaile batek n mendeko desberdin edukitzea eta n interpretazioko mendeko bakarra edukitzea. Adibideari begiratuta, **eratzten** gobernatzaileak hiru mendeko desberdin dituela adierazteko (bata subjektua, bestea objektua eta azkena, aditz laguntzailea) “10-1”, “10-6” eta “10-11” bikoteak erabiltzen dira *dep.xml* dokumentuan. Era berean adierazten da kontzeptuz desberdina den anbiguotasuna **zentral** eta **nuklearrak** artean, “1-2”, “1-4” eta “1-5” pareen bidez⁹. Beste dokumentu bat aztertu beharra dago interpretazioak zein hitzetakoak diren ezagutzeko, eta horrela, bereizketa egin ahal izateko.



VI.6 Irudia: Anbiguotasuna duen mendekotasun-zuhaitza.

Aipatu dugun moduan, *anbiguotasuna* eta honen tratamendua oso garrantzizkoak dira *Saroi* tresnan. Anbiguotasunaren eta mendekotasun-zuhaitzen multzoaren, *basoaren*, inguruan arituko gara hurrengo atalean.

⁹**zentral** hitzak NCMOD mendekotasun-erlazioa du **nuklearrak** hitzeko hiru interpretazioekin: **nuklearrak** hitzak *absolutibo-plurala* ezaugarri duen bi analisi-interpretazio ditu (bata @SUBJ eta bestea @OBJ funtzio sintaktikoekin) eta beste bat *ergatibo-singularra* ezaugarriekin (@SUBJ sintaxi-funtzioarekin).



VI.7 Irudia: Mendekotasun-zuhaitz anbiguotik eratorritako zuhaitz ez-anbiguoak.

VI.2.1.2 Basoa

Tresna baten sarrerako analisia (*input*, ingelesez) anbigua dela esango dugu berarentzat askotariko egitura linguistikoak eraiki ahal badira (Jurafsky eta Martin, 2000). Anbigutasuna lengoia naturalaren prozesamenduko maila guztietan agertzen da, bai maila morfosintaktikoan, baita sintaktikoan, semantikoan eta pragmatikoan ere; lehen esan dugun bezala, izatez baita hizkuntza anbiguo. Horrela, LNPko arazo nagusietarikoa da anbigutasuna.

Tesi-lan honetan ez dugu helburu sintaxi-mailako anbigutasuna ebaztea, hau da, ez dugu anbigutasunaren inguruko erabakirik hartu. Gure helburua, dokumentuetan modu inplizituan agertzen den anbigutasuna esplizitu egitea, azalaraztea, da. Alde praktikoari begira, anbigua den mendekotasun-zuhaitz batetik (ikus lehenago kokatuta dagoen VI.6 irudia), anbiguo ez diren behar adina mendekotasun-zuhaitz lortzen ditugu (ikus VI.7 irudia), mendekotasun-zuhaitzen *basoa* edo *saroia* sortuaz.

Anbigutasunaren inguruan aritu ondoren, sortu nahi dugun *basoaren* ezaugarri batzuk aipatuko ditugu ondorengo lerroetan. Basoa sortzen duen algoritmoa azalduko dugu geroago, eta bukatzeko, inplementaziorako erabiltako domeinu-eredua deskribatuko dugu.

Anbigutasuna

Samuelsson eta Wirén-en arabera (2000) lengoia “naturalaren” analiza-

tzaileen eta lengoia “artifizialaren” analizatzaileen (adibidez, konpiladoreen) arteko diferentzietako bat lengoia naturalaren egiturazko anbiguotasun itzelean dago. Esaldi bat analizatzea helburu dugunean, gramatika-erregela edo analisi-ekintza ugari izan ohi dira aplikagarriak, lengoia naturalaren izatezko anbiguotasuna aintzat hartu nahi badugu, behintzat.

Analisi-katea azaldu dugunean hitz-mailako anbiguotasunaren edo anbiguotasun morfologikoaren inguruan aritu gara, anbiguotasun hori ebaztea baita EUSTAGGER tresnaren helburu nagusia (Ezeiza, 2002). Anbiguotasun morfologikoari, zenbaitek *anbiguotasun gramatikal lokala* deitzen dio (Karls-son *et al.*, 1995; Aduriz, 2000). Guk *sintaxi-mailakoa* deitu duguna, berriz, *anbiguotasun gramatikal globala* deitzen dute. Hitzaren esparruan, kategoriarri, morfema ez-askeei eta sintaxiari dagozkien anbiguotasunak aztertu dira EUSTAGGERen. Morfema ez-askeen anbiguotasuna oso lotuta dago anbiguotasun sintaktikoarekin. Adibidez, **nuklearrak** hitzak bi irakurketa posible ditu, ADJEKTIBO-ABSOLUTIBO-PLURALA, objektu edo subjektu funtzioa bete dezakeena, eta ADJEKTIBO-ERGATIBO-SINGULARRA, subjektu funtzioa bete dezakeena (ikus VI.4 irudiko adibidean *lem.xml* eta *lsfi.xml* dokumentuak).

Beraz, sintaxia lantzerantz iritsi orduko, sintaxiarekin lotura estua duen hitz-mailako anbiguotasuna guztiz ebatzi gabe geratu da eta honi, gainera, sintaxi-mailako anbiguotasuna gehitu behar zaio (gogoratu **Txakurrak ikusi ditu txoriak** adibidea).

Anbiguotasun semantikoa polisemiaren edo anbiguotasun lexikalaren inguruan aritzen da, eta berau, Agirrerren (1999) eta Martinezen (2004) tesietan landu da. Polisemiaz gain, anaforaren tratamendua ere semantika-anbiguotasunaren alorrean koka dezakegu.

Pragmatika-mailako anbiguotasunari dagokionez, hau ebazteko denbora- eta espazio-testuinguruaren ezagutza eta ezagutza metatestuala behar dira.

Anbiguotasun morfosintaktikoa eta sintaktikoa, hortaz, gure lanean oso garrantzizkoak dira. Anbiguotasun semantikoa eta pragmatikoa, alderantziz, gure aztergaietatik at geratu dira.

Basoaren zenbait ezaugarri

Sintaxi-anbiguotasuna islatzen duen basoaren diseinua zehaztu aurretik, honakoak aipatuko ditugu basoari buruz:

- Analisi-katean ebatzi gabe geratzen den anbiguotasun morfosintaktikoaren eta ezarritako mendekotasun-erlazioen baitan dago basoan sor daitekeen zuhaitz kopurua eta hauen zuzentasuna.

- Basoko zenbait mendekotasun-zuhaitz esaldiak paragrafoan duen testuinguruan zentzugabeak dira. Une honetan testuinguruko informazioaren faltan gaudenez, ez gara gai esaldi zentzugabeak baztertzeko.
- Honezaz gainera, analisi automatikoa erabilia sortutako zuhaitzetan ez dira esaldiko elementu guztiak konektatuta egoten. Analisi-zuhaitz osoen faltan, zuhaitz partzialekin egin beharra daukagu lan.

Bistan da sintaxi-analisiak erabiltzen dituzten aplikazio gehienetarako egokiagoa litzatekeela ahalik eta mendekotasun-zuhaitz kopuru txikiena lortzea, eta zuhaitz hauek ahalik eta zuzenenak izatea. Tesi-lan honetan batez ere erroreen detekzioa landuko dugunez, tresnaren sarreran informazio ez-gramatikala izango dugu. Informazio zuzenerako prestatutako analisi-katea informazio ez-gramatikala aztertzeko erabilia, analisi asko eta era askotakoak lortuko ditugu. Horietatik abiatuta sortutako mendekotasun-zuhaitz guztiak aztertu beharko dira erroreen detekzioa ahalbidetuko duen informazioaren bila.

Algoritmoa

Atal honek, mendekotasun-zuhaitz anbiguo batetik, beharrezkoak diren adina mendekotasun-zuhaitz ez-anbiguo sortu ahal izateko erabili dugun algoritmoa zehaztea du helburu. Lehenengo azalpena modu ez-formalean egingo dugu, geroago, hau formalizatuz. Algoritmo honek anotazio-amaraunean dependentziak errepresentatzen dituen dokumentuaren (*dep.xml*) ezaugarriak (*analisi-unitatea*, *anbiguotasuna* eta *kontzeptuen nahasketa*) aintzat hartu behar ditu.

1. [Sarrera](#). Anbigua izan ohi den mendekotasun-zuhaitzari dagozkion *dep.xml* eta *deplnk.xml* dokumentuak izango ditugu abiapuntu. Gogorra dezagun, hitzen interpretazio pare posible guztiak azaltzen zaizkigula zuhaitz anbiguoari dagokion *dep.xml* dokumentuan, eta pare horiei guztiei dagozkien mendekotasun-erlazioak *deplnk.xml* dokumentuan. Prozesu osoa azaltzeko erabiliko dugun esaldia (35) adibidekoa da, eta lehenago ere aipatu dugu. Esaldi oker horri dagozkion dokumentuen edukia VI.1 taulan adierazten da.

(35) [*Zentral nuklearrak zakar erradiaktiboa eratzen dute.](#)

2. [Interpretazioen aukeratzea](#). Jasotako interpretazio multzotik zuhaitzetako bakoitzean erabiltzen diren interpretazioak aukeratzeko, bigarren

DEP.XML			DEPLNK.XML
Mendekotasun- erlazioaren identifikadorea	Gobernatzailearen identifikadorea (<i>lsfi.xml</i>)	Mendekoaren identifikadorea (<i>lsfi.xml</i>)	Mendekotasun- erlazioa (<i>dep-lib.xml</i>)
dep1	10 (<i>eratzen</i>)	1 (<i>zentral</i>)	NCSUBJ
dep2	1 (<i>zentral</i>)	2 (<i>nuklearrak</i>)	NCMOD
dep3	1 (<i>zentral</i>)	4 (<i>nuklearrak</i>)	NCMOD
dep4	1 (<i>zentral</i>)	5 (<i>nuklearrak</i>)	NCMOD
dep5	10 (<i>eratzen</i>)	6 (<i>zakar</i>)	NCOBJ
dep6	6 (<i>zakar</i>)	7 (<i>erradiaktiboa</i>)	NCMOD
dep7	6 (<i>zakar</i>)	9 (<i>erradiaktiboa</i>)	NCMOD
dep8	10 (<i>eratzen</i>)	11 (<i>dute</i>)	AUXMOD

VI.1 Taula: Adibideko mendekotasun-zuhaitz anbiguoari dagozkion *dep.xml* eta *deplnk.xml* fitxategiak.

taula bat erabiliko dugu (ikus VI.2 taula). Taula honen goiburukoan (taulan letra lodiz) esaldiko tokenak eta hauei dagozkien interpretazioak adieraziko ditugu *kontzeptu-nahasketa* deitu dugun ezaugarria lantzeko. Ez dira interpretazio guztiak adieraziko, mendekotasun-erlazioaren baten parte diren interpretazioak baizik. Informazio hori *dep.xml* eta *lsfi.xml* dokumentuetatik jasotzen dugu.

Taula balioekin betetzea dugu hurrengo urratsa. Hori egiterakoan gure helburua zuhaitz bakoitzean erabilgarriak diren interpretazioak soilik aukeratzea izango da: anbiguotasuna azalaraztea. Algoritmoaren deskribapenari dagokionez, taulako errenkada bakoitzean token bakoitzeko interpretazio bakarra aukeratzea da egin behar duguna. Aukeratu-tako interpretazioari “+” balioa ipiniko diogu, eta gainontzekoei “−” balioa. Honako metodoa erabiliko dugu taula betetzeko:

2. 1. Tokenak interpretazio bakarra badu, “+” balioa ipiniko dugu tokenaren zutabeari dagozkion errenkada guztietan. Adibidez, VI.2 taulako **zentral** tokenari dagokion zutabea modu honetan bete dugu.
2. 2. Tokenak n interpretazio baditu, n elementuko permutazioak egingo ditugu “+” balioa aldioro tokenaren interpretazio batean ipiniaz eta “−” balioa gainontzeko $n-1$ interpretazioetan. Errepikatuzko permutazio¹⁰ hauek ez dira askeak. Gainerako interpreta-

¹⁰Multzo bateko n elementuekin osa daitezkeen *errepikatuzko permutazioak* (hau da,

zioetako tokenak kontuan hartuta errepikatuko dira, interpretazioen arteko konbinazio posible guztiak lortu arte.

Adibideko interpretazioen arteko aukeraketak VI.2 taulan eginak daude. Aurreko paragrafoan adierazitakoa hobeto ulertzeko, har ditzagun **nuklearrak** eta **erradiaktiboa** tokenak. Lehenak hiru interpretazio ditu, eta beraz, “+ − −”, “− + −” eta “− − +” errepikatuzko permutazioak egin ditugu **erradiaktiboa** tokenak “+ −” mantendu duen bitartean. Aukera posible guztiak sortu ahal izateko, ordea, **nuklearrak** hitzeko hiru errepikatuzko permutazioak errepikatu egin behar dira, oraingoan, **erradiaktiboa** tokenean “− +” ipiniaz.

Konb.	Zentral w1	nuklearrak w2			zakar w3	erradiaktiboa w4		eratzten w5	dute w6
	1	2	4	5	6	7	9	10	11
1	+	+	−	−	+	+	−	+	+
2	+	−	+	−	+	+	−	+	+
3	+	−	−	+	+	+	−	+	+
4	+	+	−	−	+	−	+	+	+
5	+	−	+	−	+	−	+	+	+
6	+	−	−	+	+	−	+	+	+

VI.2 Taula: Konbinazio-taula baten adibidea.

Guk *konbinazio* deitu dugu aipatutako prozesuaren ondoren VI.2 taulan sortutako errenkada bakoitza. *Konbinazio* bakoitzak mendekotasun-zuhaitz ez-anbiguo bat adierazten du. *Konbinazio* hauek guztiek *konbinazio-lista* bat osatzen dute, hau da, esaldi bati dagozkion zuhaitz guztiak: *saroia* edo *basoa*.

Adibidean sei errenkada eta, ondorioz, sei zuhaitz sortuko dira. Kopuru hau aldeztu aurretik kalkula dezakegu, errepikatuzko permutazioei eta permutazio soilei dagozkien formulak erabiliz gero:

- Errepikatuzko permutazio baten elementu kopurua formula honen bidez kalkulatu dugu:

$$P_n^{k_1, k_2, \dots, k_r} = \frac{n!}{k_1! k_2! \dots k_r!}$$

elementu horien ordena bereizgarritzat duten azpimultzoak) $P_n^{k_1, k_2, \dots, k_r}$ formularekin adierazten dira. Horren arabera, lehenengo elementua k_1 aldiz errepikatuko da, bigarrena k_2 aldiz, hirugarrena k_3 aldiz... , eta r elementua k_r aldiz.

Gure adibidean, **nuklearrak** hitzean errepikatuzko permutazioa dugu eta formula aplikatuz gero, honakoa lortuko genuke:

$n=3$ (elementu-kopurua)

$k_1=1$ (“+” behin errepikatzen da)

$k_2=2$ (“−”, bi aldiz errepikatzen da)

$$P_3^{1,2} = \frac{3!}{1!2!} = 3$$

- Permutazio soilari dagokion formula honakoa da:

$$P_n = n!$$

Gure adibidean, **erradiaktiboa** hitzeko permutazioan aplikatzean

$$P_2 = 2! = 2 \text{ lortu dugu.}$$

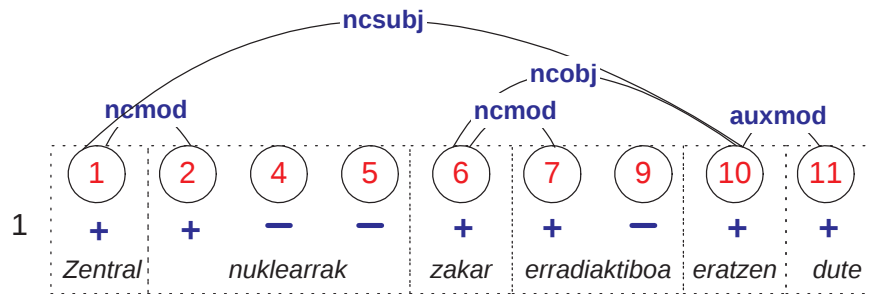
Permutazio bateko balioak kalkulatzeko beste permutazioa ezagutu behar denez, adibide zehatz honetan permutazio posibleen kopurua $2 * 3 = 6$ konbinaziokoa da. Orokortze aldera, permutazio edo errepikatuzko permutazio bakoitzari aplikatu beharko litzaioke egokitzen zaion formula, eta ondoren, lortutako emaitza guztien arteko biderkadura egin.

3. **Zuhaitzen eraikitzea.** Konbinazio horietako bakoitzari dagokion zuhaitza eraikitzeke VI.1 taulako mendekotasun-erlazioen zerrenda eta VI.2 taulako konbinazio bat (errenkada bat) erabiliko ditugu. Mendekotasun-erlazioko interpretazio bikote guztiak banan-banan konbinazioan begiratuko ditugu, eta *bi interpretazioen artean erlazioa ezarriko dugu, baldin eta soilik baldin, konbinazio horretan bi interpretazioek “+” balioa badute.*

Har dezagun, adibide bat jartzearen, VI.1 taulako **dep1 10 1** bikotea eta VI.2 taulako lehen errenkada. Errenkada honetako konbinazioan (ikus VI.8 irudia) bai 10 identifikadorea duen interpretazioak, bai 1 duenak “+” balioa dutela ikus dezakegu. Hori dela eta, interpretazioen artean arku bat marraztuko dugu, NCSUBJ mendekotasun-etiketa duena. Zerrendako bikote guztiekin prozedura bera jarraituko genuke zuhaitz osoa eratu arte. Modu honetan, taulako errenkada bakoitzeko mendekotasun-egituretan oinarritutako zuhaitz bat eraikiko dugu.

4. **Irteera.** Sarrerako zuhaitz anbiguotik abiatuta n mendekotasun-zuhaitz ez-anbiguo lortuko ditugu algoritmoa aplikatu ondoren.

Algoritmoari jarraituz zuhaitzak eraikitzerakoan, zenbait erabaki hartu behar izan ditugu topatutako arazoei aurre egiteko:



VI.8 Irudia: Basoa sortzeko algoritmoari jarraituz 1 konbinazioari dagokion mendekotasun-zuhaitza.

- Zenbaitetan zuhaitzean konektatu gabeko arkuak edo erro desberdineko azpizuhaitzak sor daitezke. Zuhaitzetako azpizuhaitz guztiak aztertzea erabaki dugu, bilatzen dugun egitura horietako edozeinetan (edo batzuetan) egon baitaiteke.

Modu ez-formalean deskribatu berri dugun prozesu hau 1 eta 2 zenbakidun algoritmoetan formalizatu dugu. Lehena, mendekotasun-zuhaitz ez-anbiguo guztiak sortzeko abiapuntua diren permutazioak sortzeko algoritmoa da (algoritmo ez-formalean 2. puntua edo *interpretazioen aukeratzea*). Bigarrenak, konbinazio bakoitza oinarrian hartuta dagokion mendekotasun-zuhaitz ez-anbigua nola eraikitzen den deskribatzen du (algoritmo ez-formalean 3. puntua edo *zuhaitzen eraikitzea*).

Algoritmoa 1 *Permutazioak* sortzeko algoritmoa.

Sarrera:

$HI-B \rightarrow$ hitza-interpretazio bikoteen multzoa.

$\{ \langle \langle w1 \ i1 \rangle, \langle w1 \ i2 \rangle, \langle w2 \ i4 \rangle \dots \rangle \}$

$MB-B \rightarrow$ mendekotasun-erlazioetako interpretazio bikoteen multzoa.

$\{ \langle dep1 \ i4 \ i1 \ i1 \rangle \}$

Irteera: $konbLista \rightarrow$ konbinazio-lista.

hasiera

$konbListaren$ goiburukoan hitza-interpretazio identifikadoreak idatzi ($HI-B$)

$konbListaren$ goiburukotik mendekotasun-erlazio baten parte ez diren interpretazioak baztertu ($MB-B$)

egin $hi \in HI-B$ **bakoitzeko** $\{ \langle w1 \ i1 \rangle \}$

hitzaren h identifikadorea lortu $\{w1\}$

h hitzaren kop interpretazio kopurua lortu $\{kop=2\}$

$konb$ konbinazio bat hasieratu ‘

baldin h aurreko hitzaren desberdina **orduan**

baldin $konbLista$ hutsa dago **orduan**

baldin $kop = 1$ **orduan** $\{Interpretazio \ bakararra\}$

h hitzaren interpretazioari “+” balioa eman $konb$ konbinazioan

$konb$ konbinazioa $konbListan$ erantsi

bestela $\{n \text{ interpretazio ditu}\}$

$konbListan$ $konb$ konbinazioa erantsi

$konbListako$ konbinazio guztiak kop aldiz errepikatu

$konbListan$ h hitzean “+” ikurrarekin eta behar adina “–” ikurrekin permutazioak egin

amaiera baldin

bestela $\{konbLista \text{ betetzen hasia dago}\}$

baldin $kop = 1$ **orduan** $\{Interpretazio \ bakararra\}$

egin $konbListako$ $konb$ **bakoitzeko**

h hitzaren interpretazioari “+” balioa eman $konb$ konbinazioan

$konb$ konbinazioa $konbListan$ ordeztu

amaiera egin($konb$ bakoitzeko)

bestela $\{n \text{ interpretazio ditu}\}$

$konbListako$ konbinazio guztiak kop aldiz errepikatu

$konbListan$ h hitzean “+” ikurrarekin eta behar adina “–” ikurrekin permutazioak egin

amaiera baldin

amaiera baldin

amaiera baldin

amaiera egin

amaiera

Algoritmoa 2 *Basoa* sortzeko algoritmoa.

Sarrera:

HI-B $\rightarrow$ hitza-interpretazio bikoteen multzoa (BASOArenak, zuhaitz guztienak)
MB-B $\rightarrow$ mendekotasun-erlazioetako interpretaziobikoteen multzoa (BASOA)
ME-E $\rightarrow$ mendekotasun-etiketen multzoa (BASOA)
konbLista $\rightarrow$ interpretazioen arteko konbinazio posible guztiak dituen lista (ikus 1 algoritmoa)

Irteera: *konbListari* dagozkion mendekotasun-zuhaitzekin osatutako *basoa*

```

hasiera {konbinazio bakoitzeko mendekotasun-zuhaitz bat sortuko da}
egin konbinazio  $\in$  konbLista bakoitzeko
  MB-Z sortu
  {MB-Z  $\rightarrow$  mendekotasun-erlazioetako bikoteen azpimultzoa
  (ZUHAITZA)}
egin menpBik  $\in$  MB-B BAKOITZEKO
  menpBikeko gobernatzailea den interpretazioa lortu
  menpBikeko mendekoa den interpretazioa lortu
  konbinazioan gobernatzaileari dagokion buruBal balioa lortu (“+” edo “-”)
  konbinazioan mendekoari dagokion mendekoBal balioa lortu (“+” edo “-”)
  {bi interpretazioen (sfLnk) arkua marraztuko da b.s.b. bi
  interpretazioek konbinazioan ‘+’ balioa badute}
baldin (buruBal = “+”) eta (mendekoBal = “+”) orduan
  konbinaziotik sortuko den zuhaitzari dagokion MB-Z egituraren goberna-
  tzailea eta mendekoa gorde
amaiera baldin
amaiera egin
  zuhaitza sortu MB-Z, ME-B eta HI-B erabilita
  zuhaitzaren erroa lortu
amaiera egin
amaiera

```

Domeinu-eredua

Mendekotasun-zuhaitzen basoak eta hauek lortu ahal izateko erabilitako baliabideek oinarrian dituzten egiturak, baita hauen arteko erlazioak ere, datu-eredu baten bidez adierazi ditugu. Datu-eredu hori hainbat klasetan adierazten da *objektuei zuzendutako paradigmatari* jarraituz. Objektuei zuzendutako analisia eta diseinua erabiltzearen arrazoia, erraz mantentzeko, sendoa den eta ongi definituta dagoen softwarea lortzea da.

Ereduaren diseinu kontzeptuala egiteko UML modelatze-lengoaia orokorra (*Unified Modeling Language*) erabili dugu. UML software-sistemak bistaratzeko, zehazteko, eraikitze eta dokumentatzeko lengoaia da (Larman,

2002). Lengoaiak hainbat anotazio-eredu eskaintzen du, objektuei zuzendutako paradigmari jarraituz sistema baten analisia eta diseinua egiteko. Horien artean guk *domeinu-ereduak* eta *klase-diagramak* erabiliko ditugu. *Domeinu-ereduan*, *Saroi* tresnaren klase kontzeptualak eta objektuak erakutsiko ditugu, baita beren arteko erlazioak ere. *Klase-diagrametan* klase kontzeptual horiek metodoekin osatuko ditugu, software-elementuak lortuaz.

Saroi sistemaren analisi sintaktikorako moduluaren domeinu-ereduaren zati bat VI.9 irudian erakusten da. Domeinu-eredu horretan klaseak hiru multzotan banatu ditugu haien abstrakzio-mailaren arabera:

- *Behe-maila*. *LibiXaML* klase-liburutegiko klaseak oinarritzko-mailan kokatu ditugu. Horiek dira zuzen-zuzenean datuekin, analisiekin eta analisisietako ezaugarriekin lan egiten duten klaseak. Gogora dezagun, klase-liburutegi honek analisi-kateko tresnen sarrera-analisiak ezagutzeko, eta emaitzak sortzeko beharrezkoak diren datu-mota eta eragiketak eskaintzen dituela. Anotazio-amaraunaren kudeaketa egiten du. *Saroi* tresna *LibiXaML* liburutegiaren bezeroa da, erabili egiten du, baina ez aldatu.

Analisi-dokumentuekin zerikusia duten klaseak topa ditzakegu multzo honetan: aingurekin zerikusia dutenak (adibidez, `joinXMLDoc`, `w`, `mw...`), estekekin (adibidez, `LnkXMLDoc` eta bere azpiklaseak) eta informazio linguistikoarekin (adibidez, `FS`, `LemXMLDoc...`)

- *Tarteko maila*. *LibiXaML* klase-liburutegiko informazioa jasotzeko, lantzeko eta azken finean, kudeatzeko, tarteko mailako klaseak garatu ditugu. *Saroi* tresnaren inplementaziorako geruza honen beharra ikusi dugu *LibiXaML* liburutegia aldatzeko ezintasunagatik. Gainera, maila honetan adierazten diren kontzeptuetan behe-mailakoetan baino abstrakzio-maila altuagoa dago.

Multzo honetan, *basoa* sortzeko beharrezko klaseak biltzen dira, `Konbinazioa`, `KonbinazioLista`, `MendekotasunenBikoteLista`, `MendekotasunenEtiketaLista...` Hauei erreferentzia zuzena egiten zaie aipatu berri ditugun 1 eta 2 zenbakidun algoritmoetan.

- *Goi-maila*. Multzo honetan abstrakzio-maila altueneko kontzeptuak bildu ditugu. Mendekotasun-zuhaitzen alorrean erabiltzen diren kontzeptuei dagozkien software-klaseak adierazten dira. Nolabait esateko, sintaxi-egiturak adierazteko mendekotasun-zuhaitzak ezagutzen dituen edozein pertsonak identifikatuko lituzkeen kontzeptu eta elementuak agertzen dira maila honetan.

Kontzeptu klaseen artean daude *Testua*, *Esaldia*, *Basoa*, *Zuhaitza* eta *zuhaitza* osatzeko beharrezkoak diren elementuak (zuhaitzak VI.2.1.3 atalean zehatzago aztertuko ditugu). Kontzeptuak zehazte aldera, *esaldia* puntutik punturako karaktere-katea kontsideratu dugula esatea garrantzitsua iruditzen zaigu. Esaldi bakoitzeko, *baso* bat izango dugu zenbait *mendekotasun-zuhaitz* ez-anbiguorekin osatutakoa.

Aipatu berri dugun domeinu-ereduari dagokion klase-diagrama ikusi nahi izanez gero, jo B eranskinera. Inplementazioari dagokionez, C++ programazio-lengoaia erabili dugu beste arrazoen artean hau delako *LibiXaML* liburutegiaren inplementazioan erabili den lengoaia.

Basoa eta bere eraikuntza-prozesua aztertu ondoren, berau osatzen duten zuhaitzak deskribatuko ditugu jarraian.



VI.2.1.3 Zuhaitza

Mendekotasun-zuhaitzak adierazteko eredua sakonean azaltzen hasi aurretik, horiek errepresentatzeko egindako aukeraren ezaugarriak (Nivre, 2005) aipatuko ditugu:

- **Geruza bakarrekoa.** Mendekotasun-egituretan oinarria duten zenbait teoria *geruza bakarrekoak* (ingelesez *mono-stratal*) dira eta sintaxi-adierazpen sinplea dute. Beste zenbaitek, *geruza anitzekoak* (ingelesez, *multi-stratal*), ezagutza semantikoa teoriaren parte kontsideratzen dute, eta ezagutza linguistikoa geruza ezberdinetan adierazten dute.

Gurean, oraingoz, morfosintaxiari eta sintaxiari dagokion ezagutza soilik adierazten da. Hala ere, erabiltzen dugun XMLzko anotazio-amarauna erabilita, ez litzateke arazorik egongo tokenaren edo multitokenaren identifikadoreari ezagutza semantikoa lukeen ezaugarri-egitura bat gehitzeko.

- **Adabegietan elementu lexikoak.** Nahiz eta teoria gehienek mendekotasun-erlazioak elementu lexikalen artean gertatzen direla adosten duten, elementu horiek modu desberdinetan definitzen dira. Gehienetan testuko hitz-formak izaten diren arren, batzuetan, lemak, morfemak edo bestelako hizkuntza-elementuak izan daitezke.

Hitzak, hitz anitzeko unitate lexikal jarraituak eta entitate izendunak dira adabegietan gordeko ditugun hizkuntza-elementuak. Erabiltako anotazio-amaraunari esker, elementu horiei buruzko informazio linguistiko osoa eskuragarri izango dugu.

- **Arkuetan gramatika-erlazioak.** Teoria ezberdinetan mendekotasun mota ezberdinak ezartzen dira. Horrela zenbaitzuetan SUBJ, OBJ eta NC-MOD moduko kategoria funtzionalak ipintzen diren bitartean, beste batzuetan, semantikara bideratutako rol motak erabiltzen dira, hala nola, *egile*, *jasale* e.a.

Geruza anitzeko teorietan bi erlazio motak erabiltzen diren arren, guk geruza bakarrean gramatika-erlazioak erabiliko ditugu (deskribapena egin genuen IV. kapituluan, IV.1.1.4 atalean).

- **Ordena linealik ez.** Tesnière-ren arabera (1959) mendekotasun-erlazioek *egiturazko ordena* dute, eta ordena hau hitz egiterakoan edo idazterakoan erabiltzen den *ordena linealaren* desberdina da. Hitzen

egiturazko ordena, hitzen arteko konexioak¹¹ ezartzen direneko ordena da. Gobernatzaille batek hainbat mendeko izan ditzakeenez, konexio horiek askotarikoak dira eta, beraz, egiturazko ordenak “zenbait dimentsio” ditu (alde praktikoari begira, bi dimentsio ditu, Tesnièrek *stemma*¹², eta guk *mendekotasun-zuhaitza* deitutakoak). Tesnièreren *egiturazko sintaxia* egiturazko ordenaren eta ordena linealaren artean existitzen diren erlazioetan oinarritzen da. Berak *stemma* deitzen duena eraikitzea, ordena lineala egiturazko ordena bihurtzea da. Berak dioenaren arabera:

«La verdadera frase es la «frase estructural» cuya frase lineal no es más que su imagen proyectada, aún de forma imperfecta y con todos los inconvenientes de aplanamiento que comporta esta proyección sobre la cadena hablada. »

(Tesnière, 1994)

Gure teorian mendekotasun-egituretako adabegiekin ez dute ordena linealik jarraitzen. Ordena hori, ordea, erraz lor daiteke adabegietako token eta multitokenek esaldian zein posizio duten adierazten duen dokumentua erabiliaz (*w.xml*). LNPko zenbait lan-alorretarako, itzulpen automatikorako adibidez, garrantzitsua iruditzen zaigu ordena lineala ere mantendu ahal izatea.

Erabilitako teoriaren ezaugarriak aztertu ondoren, errepresentazioaren inguruan arituko gara orain. Normalean, mendekotasun-zuhaitzak *orientatutako grafo etiketatuen* bidez adierazten dira eta sintaxi-analisi osoa eman nahi bada, grafoak *konektatuta* egon behar du. Grafo hauei Nivre-k (2005) bi murriztapen ezartzen dizkie. Bata, *gobernatzaille soilaren* murriztapena, zeinaren arabera adabegi bakoitzak gehienez buru bat duen. Bestea, *acyclicity* deitzen duena eta grafoak ziklorik ez edukitzea dakarrena. Bi murriztapen horiek direla eta, zuhaitza da mendekotasun-egiturak adierazteko graforik egokiena.

Tesi-lanean zuhaitzak adierazteko erabili dugun domeinu-eredua VI.10 irudian azaltzen da (klase-diagrama, B eranskinean). Zuhaitzak honako ezaugarriak ditu:

¹¹Konexioa, Tesnièreren arabera (1959): egiturazko konexioek hitzen artean mendekotasun-erlazioak ezartzen dituzte. Konexio batek normalean *gaineko* termino bat eta *azpiko* termino bat lotzen ditu.

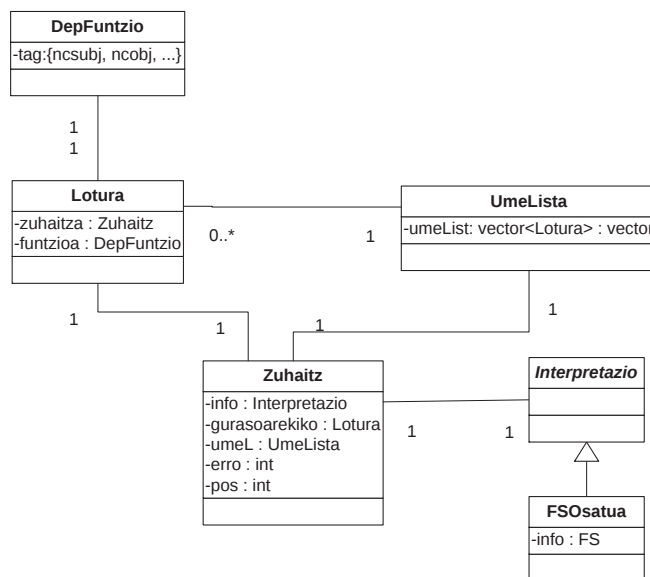
¹²*Stemma*: Nozio abstraktu baten adierazpen grafikoa da, esaldiaren egitura adierazten duen eskemaren adierazpena. Stemma konexioen hierarkia adierazten du.

- **Zuhaitzak.** Zuhaitzak eta hauen parte diren erroa, barne-adabegiak eta hostoak kontzeptu desberdinak diren arren, gure ereduan eta alde praktikoari begira, *zuhaitz* (**zuhaitz** kontzeptua domeinu-ereduan) kontsideratu ditugu. Zuhaitza modu errekursiboan definitu dugu.
- **Zuhaitzen (adabegien) arteko loturak.** Zuhaitz edo adabegi bakoitzak bere guraso bakarrarekiko (erroak ez du gurasorik) *lotura* (**Lotura** VI.10 irudian) bat du eta bere umeetako bakoitzarekiko (hostoek ez dute umerik) beste *lotura* bat. Hau da, *zuhaitz* bakoitzak *ume-lista* (**UmeLista**) edo *lotura* zerrenda bat du. Barne-adabegi diren *zuhaitzek*, *guraso*arekiko eta *umee*kiko erlazioa dutenez, *zuhaitzeko adabegiak* bi noranzkoetan korritu ahal dira: *gurasotik* *umera* edo *umetik* *gurasora*.
- **Dependentzia-erlazioak.** Zuhaitzeko *lotura* bakoitza *zuhaitz* (edo *adabegi*) batetik beste *zuhaitz* (edo *adabegi*) batera definitzen da, eta mendekotasun-erlazio batekin (**DepFuntzio**) etiketatuta dago.
- **Adabegietako informazio linguistikoa.** Zuhaitz bakoitzean (adabegi bakoitzean) token edo multitoken baten *interpretazio* bakar bati (**Interpretazio**) buruzko informazio linguistikoa gordetzen da, ezaugarri-egitura (**FS0satua**¹³) bat erabilita. Horretaz gain, beti izango dugu tokenaren edo multitokenaren posizioari buruzko informazioa eskuragarri anotazio-amaraunaren bidez.

Domeinu-ereduarekin zerikusirik ez duten, baina *zuhaitzaren* deskribapenerako interesgarriak iruditzen zaizkigun ezaugarri hauek aipatuko ditugu:

- **Ordenatu gabeko zuhaitza** da, beraz, *zuhaitz-adabegi* baten umeen artean ez dago inolako ordenarik. Adibidez, “**Gaur etorri da**” edo “**Etorri da gaur**” esaldiak adierazteko *zuhaitz* bera erabiliko genuke, hauek *egiturazko ordena* bera baitute.
- **Zuhaitz *n*-tarra da**, *zuhaitz-adabegi* batek bi ume baino gehiago izan baititzake.

¹³Ezaugarri-egitura ingelesez “feature structure” da eta “FS” moduan adierazi dugu. Gure kasuan, hau osatu dugunez, **FS0satua** deitu dugu.



VI.10 Irudia: Mendekotasun-zuhaitzen domeinu-eredua.

VI.2.1.4 Irteerako dokumentuak

Saroi tresnak zuhaitzak eraiki ahal izateko sarreran jasoko lituzkeen dokumentuen inguruan aritu gara VI.2.1.1 atalean. Dokumentu horietatik abiatuta hainbat mendekotasun-zuhaitz eraiki eta landu ondoren, zuhaitz horiek tresnaren irteeran nola errepresentatu ditugun adieraziko dugu.

Basoan sortzen diren mendekotasun-zuhaitz ez-anbiguoak bi modutan *errerepresentatzea* erabaki dugu, era *inplizituan* eta *esplizituan*.

- Era *inplizituan* (ikus VI.11 irudia), zuhaitzak anotazio-amaraunean definitutako ereduari jarraituz adieraziko ditugu. *Saroi* tresnaren sarreran mendekotasun-zuhaitzak errerepresentatzeko erabiltzen ziren bi dokumentuetan (*dep.xml* eta *deplnk.xml*) zuhaitz anbiguoak adierazten zen. *Saroi*ren irteeran anotazio bera erabiliko dugu, XML dokumentu berekin, baina zuhaitz ez-anbiguo bakoitza bere aldetik dokumentu batean adieraziz¹⁴. Errepresentazioa inplizitua dela esaten dugu, ez de-

¹⁴ *n* zuhaitz errerepresentatzeko *n* "*dep.xml*" dokumentu sortuko ditugu, zuhaitz bakoitzak bere mendekotasun-erlazio propioak edukiko baititu. Mendekotasun-erlazioei dagozkien etiketak zeintzuk diren adierazteko dokumentua *n* aldiz idaztea alferrikako lana irudituz gero, beraz, "*deplnk.xml*" dokumentua erabili dugu.

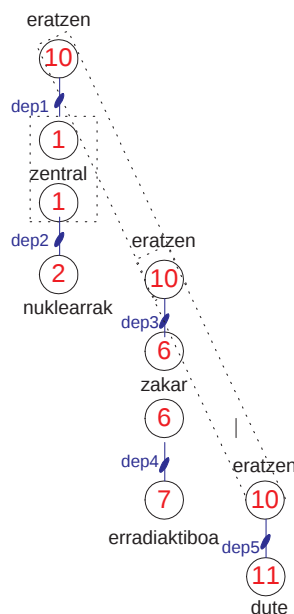
lako zuzenean adierazten adabegi bakoitzaren adabegi umeak zeintzuk diren. Zuhaitza “eraiki” egin behar da egitura osoa ezagutzeko.

```

dep.xml
<joinGrp type="Dep" targOrder="y">
  <join id="dep1" type="NoSure">
    <target name="Buru" value="Xlsfi10"/>
    <target name="Mende" value="Xlsfi1"/>
  </join>
  <join id="dep2" type="NoSure">
    <target name="Buru" value="Xlsfi1"/>
    <target name="Mende" value="Xlsfi2"/>
  </join>
  <join id="dep3" type="NoSure">
    <target name="Buru" value="Xlsfi10"/>
    <target name="Mende" value="Xlsfi6"/>
  </join>
  <join id="dep4" type="NoSure">
    <target name="Buru" value="Xlsfi6"/>
    <target name="Mende" value="Xlsfi7"/>
  </join>
  <join id="dep5" type="NoSure">
    <target name="Buru" value="Xlsfi10"/>
    <target name="Mende" value="Xlsfi11"/>
  </join>
</joinGrp>

deplnk.xml
<linkGrp type="depId-depLibId" targOrder="y">
  <link type="targets" Xdep1 XD-NCSUBJ"/>
  <link type="targets" Xdep2 XD-NCMOD"/>
  <link type="targets" Xdep3 XD-NCOBJ"/>
  <link type="targets" Xdep4 XD-NCMOD"/>
  <link type="targets" Xdep5 XD-AUXMOD"/>
</linkGrp>

```



VI.11 Irudia: Zuhaitz errepresentazio *inplizitua*.

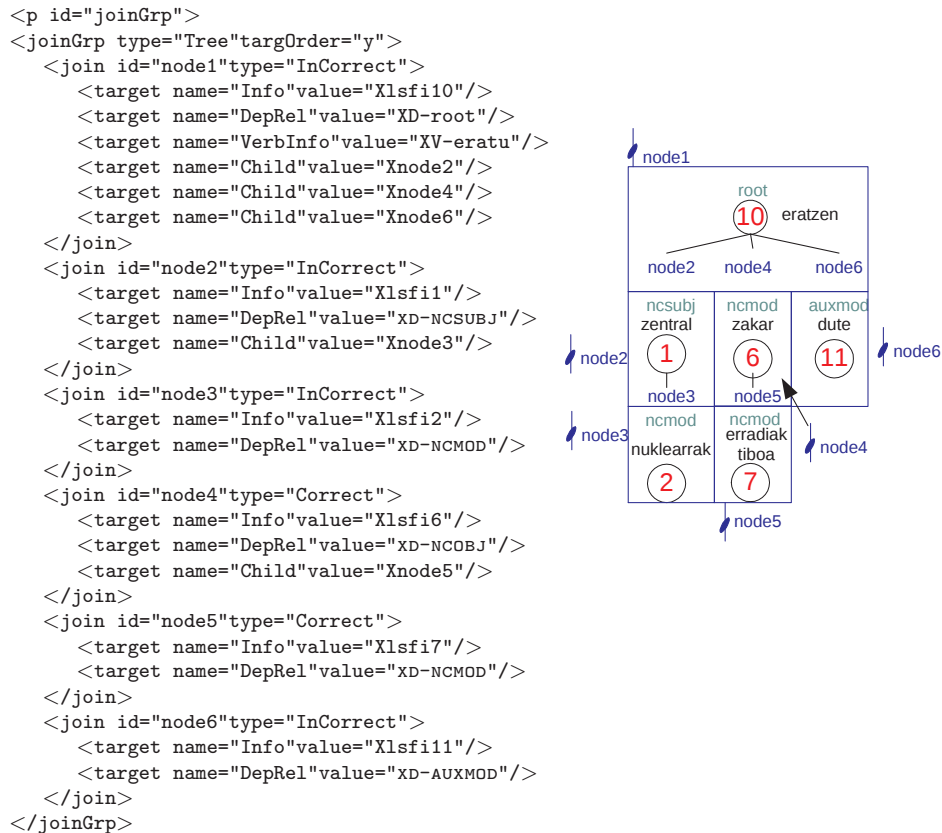
- Bestetik, zuhaitzaren egitura modu *esplizituan* errepresentatu nahi izan dugu (ikus VI.12 irudia¹⁵). Hots, adabegi bakoitzaren umeak

¹⁵Irudi honetan “*basoj.xml*” dokumentuaren zati bat besterik ez da agertzen. Dokumentuaren gainontzeko zatian informazio-elementuetako bakoitza zein XML dokumentutan topa daitekeen adierazten da *xptr* (*external pointer*) elementuen bidez. Adabegiko interpretazioari buruzko informazioa (*Info*) “*lsfi.xml*” luzapena duen dokumentuan topa daiteke, aditzari buruzko informazioa “*adi-lib.xml*” liburutegian, eta mendekotasun-etiketari dagokiena “*dep-lib.xml*” liburutegian.

zeintzuk diren eta zein mendekotasun-etiketaren bidez lotzen zaizkien, dokumentu berean adierazita (“*basoj.xml*” luzapena ipini diogu dokumentuari). Hainbat aukera aztertu ondoren, besteak beste TEIk *grafoak*, *sareak* eta *zuhaitzak* errepresentatzeko ematen dituen gidallerroak¹⁶, anotazio-amaraunean erabili izan ditugun *aingura bereziak* erabiltzea erabaki dugu. Gogoratu, elementu hauek aurreko prozesuetan sortutako interpretazio-identifikadoreak biltzeko erabiltzen direla eta hori da, hain zuzen ere, zuhaitzean egiten duguna. Zuhaitza adabegien bilduma egituratu bat da eta adabegi bakoitzari buruzko informazioa identifikadoreen bidez adierazten da: token/multitokenaren interpretazio-analisia (*Info*), mendekotasun-etiketa (*DepRel*), umeak (*Child*) eta aditzen kasuan aberasketa-informazioa (*VerbInfo*). Saroiaren aplikazioetako bat erroreen detekzioa izango denez, adabegi bakoitzeko *type* atributua erabil dezakegu adabegi hori egitura erroredun baten parte den (*type*=“*InCorrect*”) ala ez (*type*=“*Correct*”) adierazteko. Aingura bereziak, *join*-ak, erabili ditugunez, *LibiXaML* liburutegia aprobetxatu ahal izan dugu dokumentuak sortzeko.

Mendekotasun-zuhaitzak errepresentatzeko XML dokumentuak erabili ditugun arren, erabiltzaile arruntarentzat ez dira oso intuitibo edo esanguratsuak; zuhaitz grafikoak ikustea askoz ere samurragoa da. Gai honen inguruan arituko gara emaitza aurkezteko moduluan (ikus aurrerago VI.2.4 atala).

¹⁶<http://www.tei-c.org.uk/P5/Guidelines/GD.html>

VI.12 Irudia: Zuhaitz-errepresentazio *esplizitua*.

Aberasketa

Orain arte *Saroiren* analisi sintaktikorako moduluan deskribatu diren prozesuak orokorrak dira eta LNPko edozein aplikaziotan erabil daitezke. Hots, lortzen diren analisi-zuhaitzak, galdera-erantzunen sistemetan, itzulpen automatikoan edo sintaxi-mailako ezagutza behar duen edozein alorretan erabili ahal izango dira.

Aberasketa deitu dugun modulua erabiltzea aukerazkoa da eta bere orokortasuna erabiltzen duen ezagutzaren arabera da. Aberasketan, anali-

siak dakarren informazio linguistikoari interesatzen zaigun hizkuntza-informazio gehigarria gehitzen diogu. Askotarikoa izan daiteke gehitutako informazioa. Adibidez, galderak erantzuteko sistemetarako, EuskalWordNet (Agirre *et al.*, 2006) balia dezakegu eta informazio semantikoa gehitu. Gure helburuetako bat aditzarekiko komunztadura-erroreak detektatzea denez, aditz nagusiak eta aditz laguntzaileak aberastu ditugu. Horretaz gain, sintagmen arteko koordinazioa adierazi dugu kasu zehatz batzuetan.

Aditz laguntzaileetan komunztadurarekin zerikusi zuzena duten hizkuntza-ezaugarriak esplizitu egin ditugu. Agerian jarritako ezaugarriak aditz laguntzaileen paradigma edo saila (NOR, NOR-NORI, NOR-NORI-NORK, NOR-NORK) eta bere osagaietako bakoitzaren (NOR, NORI eta NORK) *kasua*, *numeroa* eta *pertsona* dira.

Aditz nagusiak aberasteko hiru iturri erabili ditugu:

- Eskuzko azterketa sakon baten ondorioz lortutako 100 aditz-eredu (Aldezabal, 2004).
- Corpusetatik automatikoki jasotako 2541 aditzi buruzko gauzatze-egiturak (Aldezabal *et al.*, 2003b).
- *Euskal Hiztegia* (EH) (Sarasola, 1996) sarrera duten aditz guztien (4016 aditz) azpikategorizazio-etiketak (DA, DU, DIO, ZAIO eta DA-DU...).

Aberasketa-prozesuari buruzko informazio sakona emango dugu VII. kapituluko VII.4.2 atalean.

VI.2.2 Kontsulta-erregelen itzulpenerako modulua

VI.2.2.1 Kontsulta-lengoaia linguistikoak

Hizkuntzarekin zerikusia duten aplikazioetan gero eta ohikoagoa da modu hierarkikoan etiketatutako testu- eta ahots-datuak edukitzea, eta horien gainean kontsulta-lengoiak baliatuz nolabaiteko kontsultak edo bilaketak egitea. Corpusetarako zehazki garatutako kontsulta-lengoiak datu linguistikoak arakatzen dituzte, eta zenbait kasutan, aldatu.

Hainbat ezaugarri eskatzen zaizkie kontsulta-lengoiak:

- **Abstrakzioa.** Kontsultak egiteko ez litzateke datuen errepresentazio zehatzari buruzko ezagutzarik behar. Abstrakzio-maila handiko lengoaia bat, erabiltzaileak duen eredu kontzeptualetik hurbilago legokeena, egokiagoa litzateke.

- **Erabilerraztasuna.** Kontsulta-lengoaiei erabiltzaileak era askotakoak izan daitezkeenez (hizkuntzalariak, ikertzaileak orokorrean, e.a.), lengoia hauek erabiltzeko errazak izan daitezen komeni da. Horrela, adierazpen erregularrak oso erabiliak dira eta sistema batzuek *itxidura* (“*” ikurraren bidez adierazia) moduko eragileak izaten dituzte nabigaziorako. Agindu sinple, ulerterraz eta ahaltsuak behar dira.
- **Adierazgarritasuna.** Lengoaia batek erabiltzaileari adierazi nahi dituen kontsulta guztiak adierazteko aukera eman behar dio. Horretaz gain, Taylor-en (2003) ustetan datuak transformatzeko eta integratzeko aukera eman behar du eta interfaze egoki bat eskaini. Kontsulta-lengoiaiek, *egitura hierarkiko eta sekuentzialetan bilaketak eta murriztapenak adierazteko gaitasuna* behar du. Bilaketetarako, zuhaitzen barrena *nabigazio bertikala* eta *nabigazio horizontala* ahalbidetu behar ditu, eta zuhaitzen ezaugarri linguistikoak aztertze baliabideak eskaini. Kontsulta-lengoiaien *eragile multzoak* eskainiko ditu aukera hauek.
- **Eraginkortasuna.** Corpus handietan erabili nahi bada, eraginkorra, eta ahal bada, optimizagarria izan behar du kontsulta-lengoaia batek.

Garatutako kontsulta-lengoaia gehienek betetzen dituzte, neurri batean edo bestean, ezaugarri hauek batzuk. Lai eta Bird-en lanean (2004) minimoetatik harantzago, kontsulta-lengoiaei eska dakizkiekeen ezaugarriak zehaztu dituzte. Egile hauek zuhaitzen errepresentaziorako osagai-egiturak erabiltzen dituzten bost kontsulta-lengoaia linguistiko¹⁷ aztertu dituzte, kontsulta-lengoaia berrerabilgarri eta erraz hedagarri batek bete beharko litzuzkeen eskakizunak zehazteko asmoarekin. Honakoak dira, egileen ustetan, kontsulta-lengoiaiek bete beharrekoak:

1. **Nabigazioa.** Zuhaitzen barrena nabigazio hierarkikoa eta sekuentziala ahalbidetu behar dute. Hau da, *nabigazio bertikala*, edo egileek *naguisitasun-erlazio* (ingelesez *dominance*) deitzen dutena, eta *horizontala* edo *aurrekotasunezkoa* (ingelesez, *precedence*). Gainera, *ez barne hartzea* (*non-inclusion*) moduko erlazioak adierazi behar dituzte.

¹⁷Honakoak dira egileek aztertutako *treebankak* eta kontsulta-lengoiaiek: 1) *Penn Treebank* (Marcus *et al.*, 1994) eta *Tgrep2* kontsulta-lengoaia (Rohde), 2) *TIGER* corpora (Brants *et al.*, 2002) eta *TIGERSearch* kontsulta-lengoaia (König eta Lezius, 2001), 3) *Emu* ahots datu-basea (Cassidy eta Harrington, 2001) eta bertako kontsulta-lengoaia, 4) *Penn-Helsinki Parsed Corpus of Middle English* corpora eta *CorpusSearch* kontsulta-lengoaia eta 5) *NITE XML Toolkit*-en (Heid *et al.*, 2004) parte den *NiteQL* kontsulta-lengoaia.

1. 1. Kontsultaren emaitza zuhaitza izan ohi da. Azpizuhaitzak auke-
ratu ahal izatea komenigarria litzateke.
1. 2. Zuhaitzak goitik behera eta ezkerretik eskuinera korritzeaz gain,
gainontzeko mugimenduak, eskuinetik ezkerrera eta behetik gora,
desiragarriak dira.
1. 3. Zuhaitzetako hostoetan, testuan, bilaketak egiten diren moduan,
barne-adabegietako etiketa-segidetan bilaketak egitea eskatuko
lukete. Adibidez, aditz baten ondoren agertzen diren sintagmak
topatzea.
2. **Itxidurak (*)**. Aztertutako kontsulta-lengoaiek oinarritzko erlazioetara-
ko (*nagusitasuna, aurrekotasuna...*) itxiduraren aukera dute. Hori oso
interesgarria da, *gobernatzaille* baten eta bere azpiko edozein mailata-
ko umeen arteko distantzia, edo maila bereko adabegien arteko dis-
tanzia ezagutu behar ez izateko. Orokorrean, edozein gramatika-teo-
riaren edozein erregela errekurtsibo adierazi ahal izatea nahiko lukete
artikuluaren egileek. Gainera, itxidura positiboa (+) oso erabilgarria
litzateke.
3. **Ordenatutako zuhaitzetatik haratago**. Anafora moduko fenomenoak
aztertzeko, zuhaitz batean esaldi bakarra adierazi beharrean, hainbat
esaldi adieraztea ideia ona litzateke, zuhaitzaren tamainagatik izango
ez balitz. Hierarkien ebakiduretan kontsultak egitea edo baso ordena-
tuetan kontsultak egin ahal izatea, desiragarria litzateke.
4. **Eguneratzeak**. Zuhaitzen corpusetan, osagaien berrantolatzeak, gober-
natzaileen txertatzeak eta etiketa-aldatzeak, ohiko eragiketak dira. Az-
pizuhaitzak zuhaitzean lekuz aldatzea maiz ez da eragiketa erraza, eta
adabegiak banan-banan mugitu behar izaten dira. Eragiketa hori erraz
egin ahal izatea eska daiteke.

Gaur egun, corpusak orokorrean, eta *treebank*ak ere bai, XML lengoaia erabiliz errepresentatzen dira eta modu horretan biltegitatu. Hori horrela izanik, logikoa litzateke XMLrekin erlasionatutako kontsulta-lengoaia erabiltzea datu horien gainean kontsultak egiteko. XML dokumentu edo dokumentu multzo birtual eta errealetatik, bai modu lokalean bai Internetetik datuak erauzteko, XQuery da W3C¹⁸ estandarretarako erakundeak proposa- tzen duen kontsulta-lengoaia. XQueryrekin lan egiten duen erakundeko az- pitaldeak zuzeneko harremana mantentzen du XSLTrekin (Clark, 1999) lan

¹⁸W3C edo *World Wide Web Consortium*, amaraunerako estandarrak definitzen dituen nazioarteko erakundea da.

egiten duen azpitaldearekin. Bi azpitaldeok dira XPathen (Clark eta DeRose, 1999) arduradun. XPath XSLTren funtsezko osagarria da eta XQueryn erabiltzen da. Hizkuntza-informazioa lantzeko kontsulta-lengoaia hauen erabilera aztertuko dugu azaletik.

XPath

XPathek testuinguru honetan duen eginbehar nagusia, XML dokumentuen zatiak aukeratzea da. XPathek XML dokumentu baten egitura logiko eta abstraktuarekin lan egiten du. Dokumentua zuhaitz moduan ikusten du eta bertan adabegiak aukeratzen ditu, haietara heltzeko bidea erakusten badiogu. Horretarako, zuhaitzeko puntu batetik hasten da, orokorrean errotik, eta zuhaitzean zehar adabegien bila mugitzen da aurrez definitutako ardatz¹⁹ batzuk erabiliz (adibidez, *child*-umea, *descendant*-ondorengoa, *parent*-gurasoa, *ancestor*-arbasoa, *following*-hurrengoa, *preceding*-aurrekoa. . .). Ardatz horien eta beren itxiduren bidez, zuhaitzean nabigazio bertikala eta horizontala bidera daitezke. Gainera, atributuak araka daitezke *atributu* (*attribute*) ardatzaren bidez. XPathen XML zuhaitzeko bideak zehazten dira, *kokapen-bide* izeneko adierazpenen bidez. *Kokapen-bide* bat *kokapen urratsez* osatuta dago. *Kokapen-urrats* bat ardatz batekin, adabegi-egiaztapen²⁰ batekin eta zero predikatu²¹ edo gehiagorekin osatuta dago. Ondorengo adibidean (36 zenbakidunean) *child* litzateke ardatza, *ncsubj* deituraz egindako adabegi-egiaztapena eta *[kas="erg"]* predikatua. Adibide honetako *kokapen-bideak* VI.13 irudiko bigarren zuhaitzean uneko adabegiaren *ume* den *subjektuaren* kasua *ergatiboa* ote den konprobatuko luke. Hori noski, XML dokumentuak zuhaitzaren egitura zuzen-zuzenean islatuko balu gertatuko litzateke.

(36) `child::ncsubj[kas='erg']`

Badirudi XPath kontsulta-lengoaia nahikoa izan daitekeela sintaxi-zuhaitzetan zehar nabigatzeko eta bertako elementuen ezaugarri linguistikoak aztertzeko. Bird *et al.* (2006) lanean, ordea, osagai-egituretan oinarri duten zuhaitzetan kontsultak egiteko zenbait eragiketa adierazgarri faltan bota dituzte. Horrela, XPath zenbait eragiketarekin hedatu dute, LPath izeneko kontsulta linguistikotarako lengoaia adierazgarri eta eraginkorra sortuaz.

¹⁹Ardatza = *axis*: urratsak zehaztutako adabegien eta uneko adabegiaren arteko erlazioa.

²⁰Adabegi-egiaztapena: urratsak zehaztutako adabegi(ar)en mota(k) eta deitura hedatua(k).

²¹Predikatua: urratsak zehaztutako adabegien multzoa areago murrizteko espresioak.

Nabigazio horizontalerako ardatz berriak garatu dituzte, hala nola, *aldameneko-hurrengo-senidea* (*immediate-following-sibling* ingelesez), *aldameneko-aurreko-senidea* (*immediate-preceding-sibling* ingelesez)... eta beren itxidurak, zuhaitzaren azterketa-eremua zehazteko funtzio bat eta zuhaitzetan ezkerren edo eskuinen dagoen adabegia zehazteko ardatzak.

Bouma eta Kloosterman-en lanean (2002) XPath XMLz kodetutako mendekotasun-zuhaitzetan kontsultak egiteko erabili dute. Nederlanderrako Alpino *Treebank*en²² topatu duten XML errepresentazioen konplexutasuna eta tamaina handia dela eta, XPathen erabilera zaila zenez, maila altuagoko lengoia bat garatu dute kontsultak errazago adierazi ahal izateko. Honek, ordea, ez ditu XPathen aukera guztiak eskaintzen eta, hau horrela izanik, XML dokumentuetako datuak XML dokumentu trinkoago batera transformatu dituzte. Azken dokumentu honetan *XPath* zuzenean erabil daiteke. *Treebank*aren tamaina txikituz, kontsultak eraginkorragoak bihurtu dituzte.

XQuery

XQuery (XML Query) W3C erakundeak XML dokumentuetan kontsultak egiteko estandar moduan proposatu zuen 2006 urtean (Boag *et al.*, 2006) nahiz eta 1999z geroztik definituta egon. XQuery kontsulta-lengoaiak SQL kontsulta-lengoia antza du eta XML dokumentuen edo datu-baseen galdetarako erabiltzen da. XML dokumentuetarako erabiltzen denean dokumentu-zatiak bueltatzen ditu, eta horrek emaitzan berriro kontsultak egin ahal izatea ahalbidetzen du (*compositionality* propietatea). XPath erabiltzen du adabegiak aukeratzeko eta XQueryk aukeratutako adabegiak lantzeko murriztapenak eta eragiketak eskaintzen ditu.

Kontsulta-lengoia linguistiko moduan, (Cassidy, 2002) lanean aztertu dute erabilpen-kasu zehatzen bidez. Kontsulta-lengoaiak analisi hierarkikorako edo nabigazio bertikalerako baliabide anitz eskaintzen dituela, baina murriztapen sekuentzialak zehazteko eskas geratzen dela ondorioztatu dute.

XSLT

XSLT (*Extensible Stylesheet Language Transformations*) lengoia (Clark, 1999) zeregin nagusia XML dokumentu bat beste dokumentu batean transformatea da. XSLT lengoia aipatu berri dugun XPath lengoia erabiltzen da XML zuhaitzeko adabegiak aukeratzeko eta dokumentuan zehar nabigatzeko. XSLTn transformazioak zehazten dituzten txantiloiak definitzen dira eta adabegiren bat zehaztutako patroi batekin parekatzen denean, txantiloia

²²<http://www.let.rug.nl/~vannoord/trees/>

aplikatzen da transformazioa aurrera eramateko. Transformazioa deskribatzen duten XMLz idatzitako espezifikazioei *XSLT estilo-orri* edo *script* deitzen zaie.

Bex *et al.* (2000) lanean XSLT transformaziorako ahalmen izugarriak dituen lengoaia moduan aurkeztu zuten eta XML kontsulta batzuk simulatzeko gai zela ikusi. XSLT ordea, erabiltzaile arruntarentzat erabiltzeko zaila dela ondorioztatu zuten. Programatzeko esperientziarik gabeko erabiltzaileentzat oso zaila litzateke konbinazioak eta taldekatzeko eragiketak beharko lituzketen kontsultak idaztea. Taylorren arabera (2003), XSLTk kontsulta-lengoaia bati eskatu beharreko funtzionalitate gehienak betetzen ditu, baina ez du datuekiko interfaze egokirik eskaintzen. Erabiltzaileak datuen errepresentazioari buruzko ezagutza eduki behar du kontsultak egiteko. Gure ustez, erabiltzaileak duen eredu kontzeptual linguistikotik hurbilago legokeen kontsulta-lengoaia behar da.

VI.2.2.2 *Saroiren* kontsulta-lengoaia

Saroiren kontsulta-lengoaia definitu genuenean gure helburu nagusia patroi edo egitura linguistiko bat mendekotasun-zuhaitzetan betetzen ote zen egiaztatzea zen. Helburu hori betetzeko, patroia kontsulta moduan definitzen dugu eta zuhaitzen basoan *bilatu* egiten dugu. Patroia betetzen duten mendekotasun-zuhaitzak markatu egiten dira, eta gero, inprimatu. Emaiztan mendekotasun-zuhaitz bat edo gehiago jasotzen dituela esan dezakegu.

Kontsultak edo bilaketak egiteko moduan pentsatzen hasi ginenean hauek XML dokumentuetako egiturari jarraituaz egin beharrean, zuzenean mendekotasun-zuhaitzen egitura kontuan hartuta egitea egokiagoa eta erabiltzailearentzat errazagoa zela iruditu zitzaigun. Kontsulta egiteko mendekotasun-zuhaitzaren egituran pentsatzea erabiltzailearen eredu kontzeptualetik hurbil dago, abstrakzio-maila altuagokoa da, eta beraz, lana errazten du.

Dokumentu-egitura baliatzen duten lengoaiak baztertuta (XPath, XQuery eta XSLT), kontsulta-lengoaia propioa zehaztea eta inplementatzea erabaki genuen. *Saroiren* kontsulta-erregelek mendekotasun-zuhaitzetan kontsultak egiteko eta zuhaitzeko ezaugarriak aldatzeko ahalmena izan zezaten zen gure lehen asmoa. Aldaketak egin ahal izateko aukera, ordea, etorkizunean lantzeko asmoa dugu.

Saroiren kontsulta-lengoaiak hizkuntzaren inguruko galderak egiteko era honetako lengoaiak bete beharko lituzketen ezaugarriak neurri handi batean betetzen dituela uste dugu:

Abstrakzioa

Kontsultak egin ahal izateko hizkuntza-informazioa gordetzen duen dokumentuaren egitura zehatza ezagutu behar izateak kontsultak zailtzen ditu. *Saroi* tresnaren kontsulta-lengoaian kontsultak egin ahal izateko erabiltzailearen eredu kontzeptualetik hurbil egon daitekeen egitura batean pentsatu dugu, mendekotasun-zuhaitzean. Kontsultak egiteko nahikoa da erabiltzaileak esaldiaren egitura, mendekotasun-zuhaitza, buruan edukitzea. Gure ustez aukera honek abstrakzio-maila altuagoa duenez, kontsulta egitea hein handi batean errazten du.

Erabilerraztasuna

Kontsultak errazak izan daitezen, agindu sinple eta ulerterrazak definitzen saiatu gara. Gure ustez, sortutako kontsulta-lengoaia intuitiboa da, zuhaitzak korritzeko zenbait eragile eta hizkuntzarekin lan egiten duen edozein pertsonak ezagut ditzakeen kontzeptuak besterik ez baitira ezagutu behar. Adibidez, kontsultak deskribatzeko erabiltzen den informazio linguistikoa honakoa da:

- Mendekotasun-zuhaitza gorantz edo beherantz korritzeko, mendekotasun-erlazioak ezagutu behar dira (NCSUBJ, NCMOD...).
- Zuhaitzean aukeratutako adabegian kokatu garenean, bertako tokenaren ezaugarriak aztertzeko nahikoa da ezaugarriaren izena ezagutzea (*kasua*, *numeroa*, *posizioa*...).

Adierazgarritasuna

Lengoaia batek erabiltzaileari adierazi nahi dituen kontsulta guztiak adierazteko aukera eman behar diola esan dugu. *Saroi*rekin helburu hau betetzen saiatu gara. Har dezagun, VI.3 taula. Bertan, hizkuntzalari batek *treebank*ean egin nahi lituzkeen zenbait galdera eta horiek *Saroi*ren kontsulta-lengoaia erabiliz nola gauzatuko lirakekeen ikus dezakegu.

Zuhaitzak modu hierarkikoan nahiz sekuentzialean nabigatzeko eta topatutako informazioa arakatzeko, hau da, kontsultak egiteko, honako eragile multzoa definitu dugu (ikus laburpena VI.5 taulan):

- *Nabigazio bertikalerako*, hau da, mendekotasun-zuhaitzak gurasoetatik umeetara eta alderantziz korritzeko, *gurasoa* (“!”), *umea* (“!”) eta *uneko adabegia* (“@”) eragileak definitu ditugu.

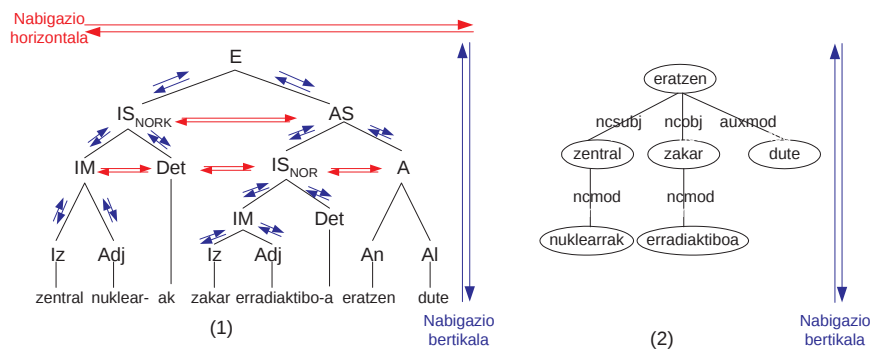
Kontsulta	Saroiko adierazpena
"Adierazi zein esalditan agertzen diren hiru izenlaguneko izenak."	@.kat="IZE" & @!ncmod# == 3
"Adierazi zein esalditan ez duten subjektuak eta NOR-NORI-NORK paradigmako aditz laguntzaile batek kasuan komunztatzen"	@!auxmod.Paradigma == "NOR-NORI-NORK" & @!auxmod.nork.kas != @!ncsubj.kas
...	...

VI.3 Taula: *Saroiren* kontsulta-lengoaiaz erantzun daitezkeen zenbait kontsulta.

- *Nabigazio horizontalean* bereizketa bat egin dugu. *Adabegietan egin daitezkeen nabigazio horizontalak* osagai-egituretan oinarritutako zuhaitzetan zentzu osoa du. Azter dezagun, adibidez, VI.13²³ irudiko (1) zuhaitza. Bertan, ohikoak izan daitezke nabigazio horizontala eskatzen duten kontsultak (ikus gezi gorriak): "lortu aditzaren aurreko NOR motako izen-sintagma". Mendekotasun-egituretan oinarritutako zuhaitzetan (VI.13 irudian (2) zuhaitza) kontsulta horrek ez du inolako zentzurik. Tokenek esaldian duten posizioa landu dugu eta token bat bestearen *aurretik* ("<"), *atzetik* (">") eta *n posizio atzetik* ("+"n") ote dagoen ezagutzeko eragileak definitu ditugu.
- *Adabegi bakoitzeko ezaugarri linguistikoak aztertze*ko ezaugarria (".") jasotzen duen eragilea, eta ezaugarria *existitzen* ote den ("~") begiratzeko duen eragile boolearra sortu ditugu.
- Horietaz mendekotasun-erlazioekin lan egiteko interesgarriak iruditu zaizkigun eragile batzuk ere definitu ditugu. Horiek mendekotasun-erlazio bereko *ume kopurua* ("#"), mendekotasun-erlazio bereko umeen artean *azkena* ("<"), edo *n-garrena* ("adibidez, hirugarrena (3)") dira. Ezaugarriekin bezala, mendekotasun-erlazioaren bat *existitzen* ote den ("~") begira daiteke.
- Kontsultetako *baldintzak konposatzeko*, eta ("&") nahiz *edo* ("|") eragileak daude eskuragarri.
- *Bi adabegitako ezaugarrien balioak konparatzeko*, *berdin* ("==") eta *desberdin* ("!=") konparaziozko eragileak ditugu.

Eragile multzo honen implementazioarekin gure beharrak asetu ditugu, baina erabilera orokorrago baterako *Saroiren* kontsulta-lengoaia eragile be-

²³E: esaldia; IS: izen-sintagma; A: aditza; AS: aditz-sintagma; IM: izen multzoa; Det: determinatzailea; Iz: izena; Adj: adjektiboa; An: aditz nagusia; Al: aditz laguntzailea



VI.13 Irudia: Osagai-egituretan (1) eta mendekotasun-egituretan (2) oinarritutako zuhaitzak eta nabigazioa.

rriekin zabaldu beharko genuke. Adibidez, une honetan *Saroi*ren kontsulta-lengoaia erabiliz egin ezin diren zenbait kontsulta eta arrazoia VI.4 taulan adierazi dira.

Kontsulta	Zergatia
"Adierazi zein esaldik ez duen <i>ikusi</i> lemako aditzik."	Ez-barne hartze eragilea definitu gabe.
"Adierazi subjektuarekin lotutako mendekotasun-erlazioak zeintzuk diren."	Oraingoz emaitza boolearra edo mendekotasun-zuhaitza ematen ditu.

VI.4 Taula: *Saroi*ren kontsulta-lengoaiaz erantzun ezin diren zenbait kontsulta.

Etorkizunean, eragile berriak definitzeko asmoa daukagu, baita informazioa *aldatzeko* aukera eskaintzeko asmoa ere.

Eraginkortasuna

Kontsulta-erregelen itzulpena egiten denean, *zirkuitulaburrak* (ingelesez, *short-circuit*) egiten dira (Aho *et al.*, 1985): adibidez, "&" eragileekin soilik osatutako baldintzetan, erregelako lehen baldintza betetzen ez bada, ez da kontsultarekin aurrera egiten; beraz, ez da adierazpen osoa ebaluatzen. Arestian aipatu dugun moduan, inplementaziorako aukeratutako programazio-lengoaia (C++) exekuzio denboran azkarrenetarikoa den eta memoria-konsumo txikienetarikoa duen programazio-lengoaia da (Prechelt, 2000). Horrek guztiak eraginkorrak egiten ditu kontsultak.

Eragilea	Semantika zuhaitzean	Adibidea
@	Uneko adabegia	
!	Mendekotasun-erlazio bati jarraituz zuhaitzean maila bat jaitzi. Umea lortu.	<i>@!ncsubj!ncmod.kas</i> (<i>Subjektu</i> mendekotasun-erlazioaren bidezko uneko adabegiaren umea lortu eta horren <i>modifikatzailea</i> den umearen kasua lortu.)
i	Mendekotasun-erlazio bati jarraituz zuhaitzean maila bat igo. Gurasoa lortu.	<i>@!auxmod.azpikat</i> (Uneko adabegiaren gurasoa eta haren azpikategoria lortu.)
!=	Balioak desberdinak	<i>@!ncsubj.kas != @.nork.kas</i> (<i>Subjektu</i> mendekotasun-erlazioaren bidez loturiko umearen kasua eta uneko adabegiaren “nork” ezaugarriko kasua desberdinak dira.)
==	Balioak berdinak	
>	Handiagoa	<i>@.posizioa > @!detmod.posizioa</i> (Uneko adabegiko tokenaren posizioa handiagoa da —atzetik dago— determinatzailearena baino.)
<	Txikiagoa	
:=	Ezkerraldeko aldagaiari eskuinaldeko balioa esleitu	<i>kas:=”erg”</i>
+	Adabegietako tokenen <i>posizioen arteko gertutasuna</i>	<i>@.posizioa + 1 == @!detmod.posizioa</i> (Uneko adabegiko tokenak bere determinatzailearen hurrengo posizioa du esaldian.)
.	Adierazitako Ezaugarriaren balioa lortu (kas, nork...)	<i>@.kat == “ADI”</i> (Uneko adabegiaren kategoria aditza da.)
~	Mendekotasun-erlazioa edo ezaugarria existitzen da	<i>@!detmod~</i> (Existitzen da uneko adabegiari <i>determinatzaile</i> mendekotasun-erlazioaren bidez lotzen zaion umeren bat.)
\$	Mendekotasun-erlazio bera duten umeen artean, azkena	<i>@!ncmod\$.kas == erg</i> (Uneko adabegiaren ume diren modifikatzaileetako azkenaren kasua ergatiboa da.)
(n)	Mendekotasun-erlazio bereko umeen artean, n-garrena	<i>@!ncmod(1)</i> (Uneko adabegiaren ume diren modifikatzaileen artean lehena aukeratu.)
#	Mendekotasun-erlazio bereko ume kopurua	<i>@!ncmod# == 1</i> (Uneko adabegiaren ume diren modifikatzaileen kopurua 1 da.)
&	Eta	<i>@.kat == “ADI” &</i> <i>@.lema-osatua == “eratu”</i> (Uneko adabegiko tokenaren kategoria aditza da eta lema “eratu”.)
	Edo	<i>@!auxmod.Paradigma == “NOR” </i> <i>@!auxmod.Paradigma == “NOR-NORI”</i> (Uneko adabegiaren umea den aditz laguntzailearen paradigma “NOR” edo “NOR-NORI” da.)

VI.5 Taula: Mendekotasun-zuhaitzak nabigatzeko eta arakatzeko *Saroir*en erregeletan erabil daitezkeen *eragileak*.

VI.2.2.3 Erregelak. Deskribapena eta adibidea

Kontsulta-lengoiaren helburua eta ezaugarriak ezagutu ondoren, kontsultak egin ahal izateko erabiliko diren erregelen deskribapenari ekingo diogu azpiatal honetan. Deskribapena atseginagoa izan dadin, adibideen bidez egitea ideia ona izan daitekeela otu zaigu eta ***zentral nuklearrak zakar erradiaktiboa eratzen dute** esaldi okerra hartuko dugu adibidetzat. Erroreen tratamendua tesi-lan honen helburuetako bat izanik, esaldian komunztadura-errorea detektatzeko eta zuzentzeko erregela aztertuko dugu. Esaldi zehatz horretan eta egitura bera duten hainbatetan errorea topatzeko, VI.14 irudiko erregela erabiliko genuke.

*Saroi*n definitutako erregelek antzeko sistematan erabiltzen diren erregelen antza dute (Knutsson *et al.*, 2001) eta hiru atalez osatuta daude.

- **Bilatu** (*detektatu* erroreen tratamenduan): mendekotasun-zuhaitzean bilatu nahi den egitura atal honetan deskribatzen da. Kasu honetan errorea deskribatzen duen patroia idatzi dugu.
- **Ordeztu** (*zuzendu* erroreen tratamenduan): mendekotasun-zuhaitzean aldaketaren bat egin nahi izanez gero, aldaketa atal honetan deskribatzen da. Erroreen detekziora ekarrita, errorea *zuzentzeko* burutu beharreko ekintza hemen deskribatu dugu.
- **Info** (*informazioa*): Erregela aplikatzen denean erakutsiko den informazio gehigarria idatziko dugu atal honetan. Erregelak errore bat detektatzen duenean, errorea deskribatzen duen mezu bat idatziko genuke hemen, erroreak *diagnosia* eginez, eta ondorioz, erabiltzailearekiko *feedbacka* landuaz.

Orokorrean mendekotasun-zuhaitzetan *Bilatu-Ordeztu* moduko eragiketak egin daitezkeela esan dezakegu (erroreen kasuan, *detektatu-zuzendu*). Banaketa hau bat dator II. kapituluaren II.2.2 atalean egin dugun *detekzioa*, *diagnosia* eta *zuzenketa* kontzeptuen definizioarekin. *Bilatu* eta *ordeztu* atalak jarraian aztertuko ditugu.

Bilatu

Erregelako *Bilatu* atalean errorea detektatu ahal izateko bete beharreko baldintza multzoa erakusten da. Geroago azalduko dugunez, erregelako atalen ordenak garrantzia du; beraz, kasu honetan erregela murriztaileena da lehen posizioan ipini duguna (hots, mendekotasun-erlazio horiek existitzen ez badira, gainontzeko baldintzak ez dira egiaztatuko).

```

KOM.SUBJ.KAS.NOR.NORK
(
  Bilatu (
    @!ncsubj!ncmod~ &
    @!auxmod.paradigma == "NOR-NORK" &
    @!ncsubj!ncmod.kas != @!auxmod.nork.kas
  )
  Ordeztu (
    (@!ncsubj!ncmod.kas := @!auxmod.nork.kas)
    # Zentral nuklearrEK zakar erradiaktiboa eratzen DUTE.
  )
  Info (Subjektuak eta aditz laguntzaileak
        ez dute kasuan komunztatzen.)
  )
)

```

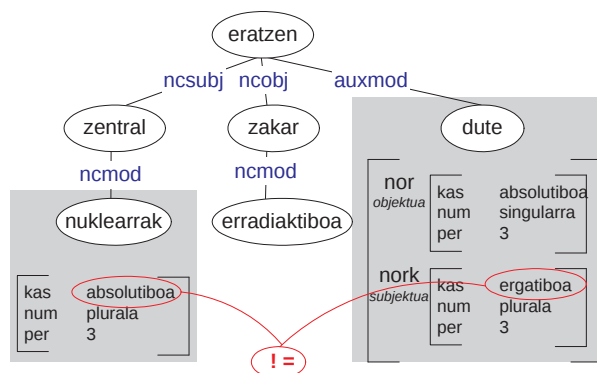
VI.14 Irudia: *Saroi*ren kontsulta-lengoaian definitutako komunztadura-erroreen detekziorako erregela bat.

Letra lodiz markatutako espresioak kasu-komunztadura aztertzen du eta honela parafraseatuko genuke: uneko adabegitik hasita (“@”), jaitsi zuhaitzean (“!”) “*ncsubj*” mendekotasun-erlazioarekin etiketatutako adarretik eta jaitsi berriro (“!”) “*ncmod*” mendekotasun-erlazioarekin etiketatutako adarretik. Subjektuaren eta aditz laguntzailearen arteko komunztadura-errore bat gertatuko da, subjektuaren kasua (“@!ncsubj!ncmod.kas”) eta aditz laguntzailean *nork* paradigmari dagokion komunztadura-marka daraman elementuaren kasua (“@!auxmod.nork.kas”) desberdinak direnean (“!=”). Erregela mendekotasun-zuhaitzari aplikatzeko modua VI.15 irudian aurkezten da. Irudiko subjektuaren (**zentral nuklearrak**) kasua (*absolutiboa*) eta aditz laguntzailean (**dute**) dagokion osagaiaren kasua (*ergatiboa*) bat ez datozeenez, errore bat detektatu dugu.

Ordeztu

Erregelaren *ordeztu* atalean zuzenketa egiteko proposamen bat egiten da: aditz laguntzailearen kasua subjektuari esleitzen badiogu, **zentral nuklearrEK zakar erradiaktiboa eratzen DUTE** esaldi zuzena lortuko dugu. Alderantzizko esleipena ez da posible, aditz laguntzailearen “nork” osagaiaren kasua ezin go baikenuke aldatu²⁴. Esleipenaren eragilearen (“:=”) erabileran sorkuntza morfologikoa erabiliko dugu hitz bera, kasu honetan subjektua, jasotako

²⁴Aditz laguntzailearen NOR osagaiak absolutibo kasuarekin egin behar du komunztadura beti; NORK osagaiak ergatiboarekin; eta azkenik, NORI osagaiak datiboarekin.



VI.15 Irudia: Erroreen *detekziorako* erregelaren adibidea.

kasuarekin sortzeko. Gogora dezagun, erregelen zuzenketa posible bakarria aurkeztu dugun arren, sor daitekeen ordain zuzenen kopurua izugarria dela.

Ordezte-atala inplementatu ez den arren, honen inguruan zenbait hausnarketa egin ditugu. Alde batetik, sortuko liratekeen proposamenen artean testinguruari hobekien egokitzen zaiona aukeratzeko moduan pentsatu dugu. Zuzenketa morfologikorako erabiltzen diren hainbat metodo erabil daitezke sintaxi-mailara egokituta. Batzuk heuristikoetan oinarritzen dira eta proposamen egokienak maila morfosintaktikoan aldaketa gutxien eskatzen dituztenak (Menzel, 1988) edo maila semantiko eta fonetikoan aldaketa (Genthial *et al.*, 1994) gutxien eskatzen dituztenak dira. Azterketa bibliografikoa egin genuenean beste zenbait metodok elementu erroredunaren testinguru sintaktiko/semantikoa aztertzen zutela ikusi genuen (ikus III. kapituluan, adibidez, (Golding eta Roth, 1996; Carlson *et al.*, 2001)). Antzeko teknika erabil daiteke sintaxi-mailan, zuzenketarako hautagaien artean aukerarik aproposena emateko.

Gogoratu, zuzenketarako proposamenak sortu eta egokiena aukeratu ondoren, sortutako zuhaitzean erroreak detektatzeko erregelak berriro pasatu behar direla, errore batzuen zuzenketak errore berriak sor baititzake.

Esaldi bati dagozkion mendekotasun-zuhaitz guztiak sortzeko modua aztertut dugu, baita horietan kontsultak egiteko modua eta erregelak ere. Erregela abstraktuak zuhaitzetan aplikatu ahal izateko, programazio-lengoiaren batera itzuli behar dira, eta hori da, hain zuzen ere, hurrengo atalean aztertut dugu.

VI.2.2.4 Itzulpen-prozesua

*Saroi*ren kontsulta-lengoaia abstraktuko erregelak exekuta daitezkeen aginduetara itzultzeko, itzultzaile bat definitu eta inplementatu dugu. Itzultzaileak *Saroi*ren goi-mailako lengoaiako erregela bakoitza C++ goi-mailako programazio-lengoaian idatzitako funtzio batera itzultzen du. C++ aukeratzeko arrazoiak besteak beste ondorengoak izan dira: gainontzeko moduluak (LibiXaML) hala inplementatuta daude, plataforma anitzetan erabil daiteke eta datu-egituretarako txantilo estandar interesgarriak eskaintzen ditu (*Standard Template Library* edo STL).

Itzultzailea sortzeko honakoak garatu ditugu: analisi lexikoa, analisi sintaktikoa eta bitarteko kodearen sorkuntza. Uste izatekoa da itzultzailearen sarreran jasoko ditugun erregelak semantikoki eta egituraz zuzenak izango direla; beraz, itzultzailean ez dugu analisi semantikorik eta errore-kontrolik egin. Kodearen optimizazioa eta behe-mailako kodearen sorkuntza C++eko konpiladorearen esku utzi ditugu; beraz, itzultzailearen *aurreko etapa* (*front end*) garatu dugula esan dezakegu. Jarraian analizatzaile lexikoaren eta analizatzaile sintaktikoaren inguruan arituko gara.

Analizatzaile lexiko batekin erregetako karaktere-kateak irakurtzen dira eta osagai lexikoen sekuentzia bat sortzen da. Analizatzaile lexikoaren zehaztapenerako eta garapenerako *flex*²⁵ (*Fast lexical analyzer generator*) tresna erabili dugu. Zehaztapenean erabilitako patroiak, sortutako osagai lexikoa eta zenbait adibide C.1 eranskinean erakusten dira.

Analizatzaile sintaktikoak lexikotik jasotzen ditu osagai lexikoak eta erregelei egitura hierarkikoa ematen die. Egitura hori deskribatzeko testuingururik gabeko gramatika bat (TGG aurrerantzean) definitu dugu. Hona hemen TGG horren deskribapen formalak:

*Saroi*ren testuingururik gabeko gramatika $G=(\Sigma, N, P, S)$ laukotearen bidez definitzen da, non osagaietako bakoitza honakoa den:

- Σ bukaerako ikurren multzoa. Kasu honetan, analisi lexikotik jasotako osagai lexikoei osatzen dute Σ .

$\Sigma = (\text{Multzo_izena}, \text{Erreg_izena}, (,), \text{Detekt}, \&, |, @, \text{Katea}, \text{Zbkia}, \$, \#, +, -, !, i, !=, >, >=, <, <=, \text{ncsubj}, \text{ncobj}, \text{nczobj} \dots \text{mota}, \text{kas}, \text{kat}, \text{nor}, \text{nork} \dots)$

- N , ez-bukaerako ikurren multzoa

$N = (\text{NAGUSIA}, \text{MULTZOAK}, \text{ERREGELAK}, \text{ERREGELA}, \text{DETEKT}, \text{DETEKT_OSAGK}, \text{D_OSAG}, \text{POS_ERAG}, \text{POS_ERAG}, \text{PLUS_MINUS}, \text{MURRIZ-}$

²⁵<http://flex.sourceforge.net>

TAPENA, ERLAZIOAK, NORANZKOA, EREMUA, KONP, ERLAZIOA, EREMUA)

- P: produkzio-erregelen multzoa, non erregelak horrelakoak diren:
 $u \longrightarrow v$ eta $u \in N, v \in (N \cup \Sigma)^*$

Produkzio-erregelak VI.6 taulan agertzen dira.

- S hasierako ikurra. Kasu honetan NAGUSIA.

Sarrerako erregelen egitura hierarkikoa guztiz lotuta dago erregelen semantikarekin. TGGko ez-bukaerako ikurrei ikurren semantikarekin zerikusia duten atributuak eta ekintza semantikoak gehituz gero, *sintaxiak zuzendutako itzulpen-eskema (SZIE)* (Aho *et al.*, 1985) sortzen da. Horren bidez definituko da erregelak helburuko lengoaiara itzultzeko modua. Kontsulta-erregelen itzulpenarako definitu dugun SZIEa ?? eranskinean azaldu dugu bere osotasunean. SZIEa erabiliz analizatzaile sintaktikoa sortzeko YACC²⁶ (*Yet Another Compiler-Compiler*) analizatzaile sintaktikoen sortzailea erabili dugu.

Analisi sintaktikorako definitutako testuingururik gabeko gramatikak gure erregelen sintaxi-egitura definitzen duela esan dugu. Kontsulta-erregelen egiturarekin erlazio zuzena duten zenbait kontzeptu definituko ditugu jarraian, TGGa bera eta kontzeptu hauek guztiak modelatzeko erabili dugun domeinu-eredua (ikus VI.16 irudia) errazago uler dezagun.

- *Multzoak*. Mendekotasun-zuhaitzetan kontsultak egiteko erregelak izendun multzotan biltzen dira, erabiltzaileak aukeratutako ezaugarriaren arabera: landu beharreko fenomenoak, informazio linguistikoaren iturria. . . Aukeratutako aplikazio-estrategiaren arabera, multzoak ordena batean edo bestean, elkarrekin edo modu independentean, aplikatu ahal izatea da gure helburua.
- *Erregelak*. Multzoetako bakoitzean nahi adina erregela definituko dugu. Bakoitza bere izenarekin eta aipatutako atalekin (*bilatu*, *ordeztu* eta *diagnosia-infoa*).
- *Detekziorako osagaiak* (*detekt_osagk*). Erregelak osagaiez osatuta daude. Gutxienez osagai bat (*detekt_osag*) izango dute edo, erregela konposatuen kasuan, bi osagai edo gehiago konjuntziozko (“&”) edo disjuntziozko (“|”) eragileekin lotuta.

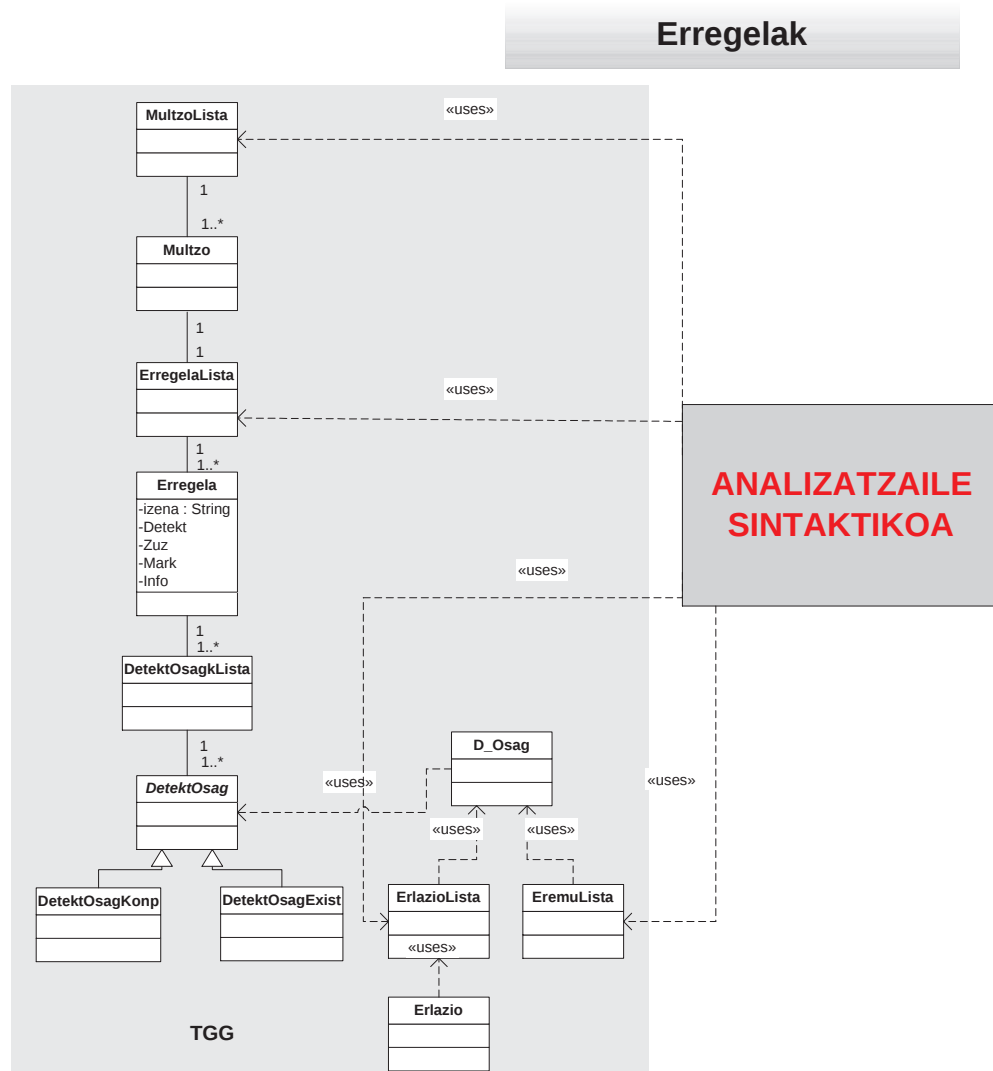
²⁶<http://dinosaur.compilertools.net>

```

NAGUSIA → MULTZOAK
MULTZOAK → Multzo_izena ERREGELAK MULTZOAK
           | Multzo_izena ERREGELAK
ERREGELAK → ERREGELAK ERREGELA
           | ERREGELA
ERREGELA → Erreg_izena ( DETEKT ZUZ MARK INFO)
DETEKT → Detekt ( DETEKT_OSAGK)
DETEKT_OSAGK → DETEKT_OSAGK & DETEKT_OSAG
              | DETEKT_OSAGK “|” DETEKT_OSAG
              | DETEKT_OSAG
DETEKT_OSAG → D_OSAG KONP D_OSAG
              | D_OSAG ~
              | ( DETEKT_OSAGK )
D_OSAG → @ ERLAZIOAK EREMUA POS_ERAG
        | Katea
        | KasuKonstantea
        | Zbkia
POS_ERAG → PLUS_MINUS Zbkia
          | €
PLUS_MINUS → +
          | -
MURRIZTAPENA → #
              | $
              | ( Zbkia )
              | €
ERLAZIOAK → ERLAZIOAK NORANZKOA ERLAZIOA MURRIZTAPENA
          | €
NORANZKOA → !
          | i
EREMUA → EREMUA . EREMUA
        | €
KONP → == | != | > | >= | < | <=
ERLAZIOA → ncsbj | ncobj | nczobj | ccomp_subj | ccomp_obj...
EREMUA → mota | kas | kat | nor | nork | ...

```

VI.6 Taula: Analisi sintaktikoaren oinarri den TGGaren produkzio-erregelak.



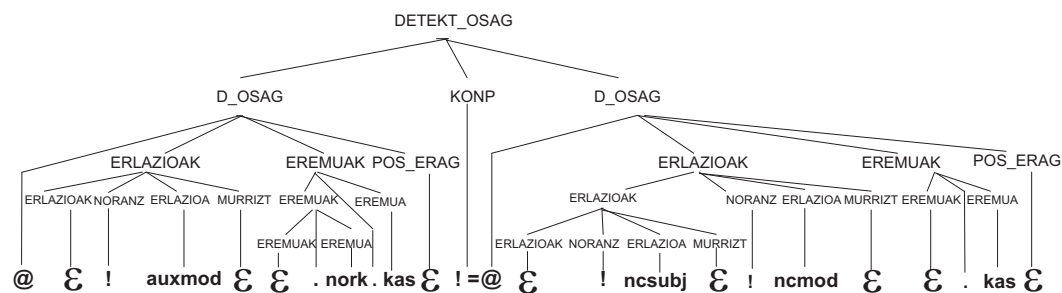
VI.16 Irudia: *Saroi*ren *erregelak* lantzeko domeinu-eredua.

- *Detekziorako osagaia* (*detekt_osag*). Osagai bakoitza, konparaziozko eragileren batekin (“==”, “!=”, “>”...) lotutako osagaiez edo existentzia adierazten duen eragilearekin (“~”) osatutako osagai batez osatuta dago.
- *Bidea definitzeko osagaia*. Baldintza batean zenbakiak, karaktere-ka-teak edo ezaugarri-bideak (adibidez, *@!ncsubj!ncmod.kas*) konpara daitezke. Ezaugarri-bide adierazpen bat honakoek osa dezakete: “@” eragileak, erlazio batzuek (“*!ncsubj!ncmod*”), eremu batzuek “kas”, eta aukeran, posizioa lantzeko espresio batek (“+1”, adibidez). Erlazioak eta eremuak aukerazkoak dira.
- *Erlazioak, eremuak eta zehaztapenerako elementuak*. Ezaugarri-bide batean hainbat elementu topa ditzakegula esan dugu. Erlazio-espresioak noranzkoarekin (“!” , “i”), erlazioarekin (“ncsubj, ncobj...”) eta murriztapenekin (“#”, “\$”, “*zbkia*”) osatuta daude. Hainbat erlazio-espresio idatz daitezke bata bestearen atzetik. Bidea adierazteko erlazio-espresioa idatzi ondoren, aztertu nahi ditugun eremuak zehatz ditzakegu, eta horretarako eremuaren izena (“mota, kas...”) idatziko dugu “.” eragileaz jarraituta. Azkenik posizioarekin lan egiteko, posizio erlatibo eragile bat (“+ 1”, adibidez) zehazteko aukera eman da.

Kontzeptua	TGG (ez-bukaerakoak)	Domeinu-eredua (klaseak)
Multzoak	NAGUSIA MULTZOAK	<i>MultzoLista</i> <i>Multzo</i>
Erregelak	ERREGELAK ERREGELA DETEKT	<i>ErregelaLista</i> <i>Erregela</i>
Detekziorako Osagaiak	DETEKT_OSAGK	<i>DetektOsagkLista</i>
Detekziorako Osagaia	DETEKT_OSAG	<i>DetekOsag</i> <i>DetektOsagKonp</i> <i>DetekOsagErist</i>
Bidea	D_OSAG	<i>D_Osag</i>
Erlazioak, eremuak zehaztapenerako elementuak	NORANZKOA ERLAZIOAK ERLAZIOA EREMUAK EREMUA POS_ERAG MURRIZTAPENA	<i>ErlazioLista</i> <i>Erlazio</i> <i>EremuLista</i>

VI.7 Taula: Kontsulta-erregelen itzulpenean erabilitako TGGaren eta domeinu-ereduaren arteko baliokidetasunak.

Detekziorako osagaien nondik norakoak hobeto ulertzeko, VI.17 irudian osagai baten analisi-zuhaitza irudikatu dugu. Bertan VI.14 irudiko *@!aux-mod.nork.kas != @!ncsubj!ncmod.kas* erregela osagaietan xehatu dugu.



VI.17 Irudia: Analisi-zuhaitza.

Itzultzailearen bitartez sarrerako lengoiako erregela bakoitzeko sortzen den C++-eko funtzioak *egiazkoa* balioa itzuliko du erregelak zehazten dituen baldintzak mendekotasun-zuhaitzean betetzen badira, eta *faltsua*, bestela. Sortutako kodea zuzen-zuzenean aplika daiteke mendekotasun-zuhaitzetarako definitu dugun egituran (ikus adibide bat ?? eranskinean). Sortutako kodean *zirkuitulaburrak* erabiltzen dira. Adibidez, konjuntzioa erabiltzen duten erregeletan lehen baldintzako (kasu honetan erlazio baten existentziari buruzko kontsulta), *@!ncsubj!ncmod~*, espresioaren ebaluazioak faltsua balioa ematen badu, ez dira gainontzeko azpibaldintzak ebaluatzen. Disjuntzioan, lehen baldintza egiazkoa bada, ez da besterik ebaluatzen. Horrela, mendekotasun-zuhaitzak azkar baztertzen dira, ez dira erregela osoak erabiltzen, eta ondorioz, eraginkortasunean hobetzen da.

VI.2.3 Kontsulta-erregelen aplikaziorako modulua

Kontsultak emaitzarik izango badu, ezinbestekoa da kontsulta-erregelak mendekotasun-zuhaitzetan aplikatzea. Erregelak multzoetan bilduko ditugu, estrategia bati jarraituz zuhaitzetan aplikatu, eta zuhaitzak baldintzak betetzen ote dituen erabakiko dugu. Horri guztiari buruz arituko gara atal honetan.

Kontsulta-erregelak multzoetan biltzeak hainbat abantaila ditu. Esate baterako, lehenetsunak ezar ditzakegu erregelen artean. Bestetik, erregela multzo bakar batzuk soilik aplika ditzakegu, eta ez guztiak, ordena zehatz bati jarraituz. Adibidez, lehenengo erregela murriztaileenak aplika ditzakegu, zuhaitzak azkarrago baztertuz eta *Saroi* tresnaren erabilera azkartuz.

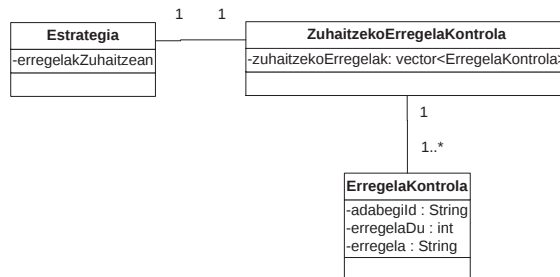
Multzoen definizio *estatikoa* egin daiteke, baita *dinamikoa* ere. Definizio dinamikoan erabiltzaileari aukera emango litzaioke parametro batzuen arabera kontsulta-erregelak multzokatzeko. Kontsulta-erregelak multzoetan biltzeko irizpide desberdinak erabil litezke:

- Lantzen duten *fenomeno linguistikoa* irizpide moduan erabiliz gero, erregelak kategoria gramatikalen arabera bil litezke (adibidez, aditzak multzo batean, izenak bestean. . .) edo mendekotasun-erlazioen arabera (adibidez, laguntzaileak aztertu nahi badira multzo batean, AUX-MOD etiketa duten erregelak, erlatiboazko egiturak aztertzeko XMOD eta CMOD etiketak dituztenak. . .). Beste edozein hizkuntza-ezaugarri erabiltzea ere posible litzateke.
- Fenomeno linguistiko zehatz bat lantzen ari garela, kontsulta-erregela batzuk beste batzuk baino *ziurtasun* handiagoarekin landuko direla aurreikus dezakegu. Ziurtasun-mailaren arabera bil ditzakegu erregelak eta ziurrenak lehenengo aplikatu.

Definitutako multzoen aplikazioa erabiltzaileak parametrizagarria izatea ere, interesgarria litzateke. Horrela, multzoen aplikazio *sekuentziala* (lineala) erabil genezake edo *erabiltzaileak erabakitako aplikazio-sekuentzia* aplikatu. Adibidez, baldintza gehien betetzea eskatzen duten multzoak lehenengo, geroago gutxiagokoak. . .

Multzokatzeta erabakita egonda, multzo horiek zuhaitzetan aplikatzeko hainbat aplikazio-estrategia defini daitezke. Errazena eta sinpleena guk erabili duguna da: “erregela guztiak zuhaitzetako adabegi guztietara aplikatzea”. Estrategia hori aurrera eramateko erabilitako domeinu-eredua VI.18 irudian ikus dezakegu. Bertan, **Estrategia** klasea erabiliko da erregelak zuhaitzetan aplikatzeko. Gero, zuhaitz bakoitzean aplikatutako erregela guztien emaitzak **ZuhaitzekoErregelaKontrola** klasean gordeko dira, horretarako zein zuhaitzetan zein erregela (ikus **ErregelaKontrola** klasea) aplikatu den gordeaz. Maiz erregeletan ezaugarri zehatz baten existentzia berma dadin eskatzen denez, benetan ez dira zuhaitzetako adabegi guztiak begiratuko, baldintzak betetzen dituztenak baizik. Modu honetan sistema eraginkorragoa izango da. Une honetan, erregela idazten duen pertsona arduratzen da existentzia baldintza bete dadin eskatzeaz (adibidez, kas~ baldintza-espresioak kasua existitu dadin eskatzen du). Etorkizunean, ezaugarri bati buruzko kontsulta egiten denean, ezaugarriaren existentzia automatikoki beti egiazta dadin lortu nahi dugu.

Lezius-ek (2002) bere tesi-txostenean dioenaren arabera, kontsultak eraginkorragoak egiteko estrategia bat kontsultak sinplifikatzea eta bilaketa-



VI.18 Irudia: *Estrategiak* kudeatzeko domeinu-eredua.

-espazioa murriztea da. Berak lehenengo kontsultak sinplifikatu egiten ditu adabegietako ezaugarriak soilik kontuan hartzeko (eman dezagun, kategoria soilik aztertzen duela, adibidez aditza, eta gainontzeko ezaugarriak baztertzen dituela). Ondoren, parekatze-hautagai diren corpuseko grafoetara (zuhaitzak) murrizten du bilaketa. Kasu horretan, aditzak dituzten zuhaitzak soilik aztertuko lituzke, eta gainontzekoak baztertu. Kontsulta sinplifikatuak zehaztapenak kenduta ditu, hau da, hasierako kontsultaren orokortzea da. Beraz, kontsulta sinplifikatuarekin lortutako zuhaitz multzoan egongo da hasierako kontsultaren erantzuna.

Beste estrategia posible bat, esaldi baten analitik sortutako mendekotasun-zuhaitz guztien artean baldintzak betetzen dituen zuhaitz bakarra aukeratzea da, gainontzekoak baztertuaz baldintzak betetzen ote dituzten aztertu gabe. Azkarragoa egingo litzateke kontsulta, baina emaitza bakarra izango litzateke eta, agian, ez egokiena.

Mendekotasun-erlazioak ere erabil daitezke estrategiak zehazteko eta kontsultak optimizatzeko. Esate baterako, ez luke zentzurik determinatzaile bati dagokion zuhaitz-adarrean subjektuekin lotutako erregelak aztertzea.

Demagun erregela multzoak definitu ditugula, ondoren aplikazio-ordena erabaki dugula, eta azkenik, strategiaren bati jarraituta erregelak zuhaitzetan aplikatu ditugula. Hurrengo urratsa zuhaitz batean erregela betetzen den edo ez *erabakitzea* litzateke.

Erabakitze-atazak erraza dirudien arren, azken finean erregela esaldian betetzen den edo ez erabakitzean datza, anbigutasun sintaktikoa dugunez (mendekotasun-zuhaitz bat baino gehiago esaldiko), aztertzen den fenomenoaren arabera erabaki beharko genuke erregela betetzen den ala ez. Adibidez, landu nahi den fenomenoak komunztadura-erroreen detekzioa denean, esaldi bateko *zuhaitz guztietan* bete beharko da komunztadura-eza. Gerta

liteke, esaldi anbiguo zuzen batean, errore-detekzioko erregelak analisi okerretako bat errore sintaktiko moduan markatzea, eta beraz, alarma faltsu bat ematea. Beste kasuren batean, erregelak definitutako fenomenoak *gutxienez zuhaitzetako batean* gertatzea nahikoa izango da. Esaldi berari dagozkion zuhaitzetan erabiltzaileak erregelaren aplikazio-proporzioa zehazteko aukera eduki zezan ideia ona litzateke. Laburtuz, esaldi batean erregela bat betetzen ote den erabakitzeke, dagozkion analisi-zuhaitz guztietan bete daiteke, bakarrean edo proportzio zehatz batean.

VI.2.4 Emaizten aurkezpenerako moduluak

Aurreko hiru ataletan *Saroi*ren hiru moduluak aztertu ditugu. Batek sarreran jasotako esaldiari dagokion anotazio-amarauna jaso, bertako sintaxi anbigutasuna bistaratu, mendekotasun-zuhaitzak eraiki eta aberastu egiten ditu. Beste batek, sarreran jasotzen duen erregela multzoa kode exekutagarria bihurtzen du. Azaldu dugun azken moduluak kontsulta-erregela exekutagarriak mendekotasun-zuhaitzetan aplikatzen ditu, estrategia bat jarraituz. Oraintxe aurkeztuko dugun moduluak, erregelak zuhaitzetan aplikatzearen ondorioz lortutako emaitzak bistaratuko lituzke.

Emaitzak erakusteko modua erabakitzeke, interesgarria iruditzen zaigu hizkuntzalariak *treebank*etan zein kontsulta mota egin nahiko lituzketen eza gutzea, beren erantzunen arabera diseinatutako baitugu aurkezpen-moduluak. Eskaeren analisia VI.2.4.1 atalean egingo dugu. Eskaera eta emaitza posibleak ezagutu ondoren, emaitzak bistartzeko interfaze posible baten berri emango dugu VI.2.4.2 atalean.

VI.2.4.1 Eskaeren analisia

Abeillé-ren arabera (2003), orain arte corpusetan oinarritutako hizkuntzalariak hitz zehatzak bilatuz azter zitezkeen fenomenoetara mugatu da. *Treebank*en sorkuntza eta erabilera zabaldu den heinean, hizkuntzalari konputazionalak gramatika-liburuetan gainetik pasatzen diren berezitasunak testu errealean topa ditzaketela konturatu dira. Horrela, hizkuntzalariak teoria edo hipotesi zehatz bati buruzko adibideak (eta maiztasunak) bila ditzakete *corpus* edo *treebank*etan. Psikolinguistek maiztasunak lantzen dituzte orokorrean eta gizakien lehentasunekin konparatzen dituzte. Soziolinguistek eta hizkuntzen historia lantzen dutenek, *treebank*ak erabiltzen dituzte zeregin honetarako. *Treebank*ak gainera, testuen sailkapenerako, hitzen adiera-desanbiguaziorako, eta testu eleanitzen lerrokatze eta indexaziorako erabiltzen dira.

Eskaera horiek aurrera eramateko ezaugarri anitz hartu behar dira kontuan. Besteak beste, helburu zehatz bakoitzerako zein corpus den egokien (tamaina, domeinua...), anotazio linguistikoa egiteko egokia ote den, etiketa-sistema erabilgarria ote den ala aldaketaren bat behar ote duen...

Gure kasuan, *Saroi*ren sarrerako dokumentuak bai *treebank*etik, baita analisi automatikotik jaso daitezkeenez, corpusaren domeinua eta tamaina kudeatzea erraza da. Erabiltzaileak aukeratu egiten ditu. Hizkuntzalaritza konputazionala da guk landu dugun alorra, eta alor honetan lan egiten duten zazpi hizkuntzalariri galdegin diegu *treebank*ean egin nahiko lituzketen kontsulten inguruan. Erantzunak asko eta askotarikoak izan dira eta horiek VI.8 taulan bildu ditugu²⁷. Kontsulta horiek denek duten ezaugarria da, kontsulta-baldintzak betetzen dituzten esaldiak ezagutu nahi izateaz gain, hizkuntzalariek esaldi horiei buruzko estatistikak ere ezagutu nahi izaten dituztela. Hizkuntzalariei fenomeno linguistiko zehatzen azterketarako baliagarriak izan daitezkeen kontsultak interesatzen zaizkie gehienbat.

Eskaera
<i>Honako egituradun esaldiak itzuli:</i> - Koordinazio-egiturak, erlatibozkoak, zehar-galderak eta konpletiboak. - Elipsia dutenak. - Aditz bakarreko esaldiak (aditz jakinekoak). - Aditz bat baino gehiagoko esaldietan, aditz jokatugabea duten esaldiak. - Postposizio-lokuzio jakin bat duten esaldiak, batzuetan aditz jakin batekin konbinatuta. - 1,2,3... modifikatzaile dituzten izen-sintagmak edo perpausak.
<i>Honako pareak dituzten esaldiak itzuli:</i> - Funtzio-konbinazio zehatzak. Adibidez, @SUBJ eta @ADLG. - Objektu-izen pareak. Adibidez, “zarata” izena objektua delarik, zein aditzekin agertzen da? egin? atera? - Subjektu-aditz pareak. Adibidez, “haizeak” ibili, egin, jo... . - Izen-adjektibo pareak. Adibidez, “lagun” min, fidel... . - Aditzak zein izen-sintagma edota postposizio-sintagma ditu inguruan?

VI.8 Taula: Hizkuntzalariek *treebank*ean egin nahiko lituzketen kontsultak.

Taulan (VI.8 taulan) zerrendatutako kontsultak kategoria gramatikalen ingurukoak (aditzak, izenak...), sintaxi-egituren ingurukoak (koordinazioa, erlatibozkoak...), mendekotasun-erlazioen ingurukoak (subjektu, objektu...) edo horien konbinazioekin osatutakoak dira. Horiek guztiek, anotazio-amaraunean isla dutenez, *Saroi*rekin erantzun daitezkeela uste dugu. Hala ere, baieztapen honek azterketa sakonago bat behar du.

²⁷Taulan, dagoeneko IXA taldean garatutako sistemekin erantzun ezin diren kontsultak soilik idatzi ditugu.

VI.2.4.2 Emaizten aurkezpena

Hizkuntzalariek *treebank*en gainean kontsultak egin ahal izateko *Saroirako* definitu dugun kontsulta-lengoaia konplexua ez den arren, interfaze grafiko baten erabilerak asko erraztuko luke kontsulta-lana. Honako ezaugarriak bete beharko lituzke interfaze grafiko horrek:

- *Treebanka* eta bertako informazio linguistikoa definitzeko erabilitako etiketa-sistema ezagutu gabe kontsultak egitea posiblea izan beharko luke: ezaugarriak (adibidez, KAT, KAS, NORK...) eta hauen baliok (adibidez, IZE, ADI, ABS, ERG...) interfazean zerrendatuta agertu beharko lukete.
- Ez litzateke beharrezkoa kontsulta-lengoiaren sintaxi zehatza eza gutzea. Interfazean egindako aukerek zehaztuko lukete kontsulta eta hauen aukeraketaren ondorioz automatikoki sortuko litzateke galdera dagokion kontsulta-lengoian.
- Braga eta Campi-ren arabera (2003):

«a visual interface to a query language is effective in supporting the intuitive formulation of queries when the basic graphical constructs of the language are close to the visual abstraction of the underlying data model. »

Hitz hauei kasu eginez, kontsulta-lengoian erabilitako eraikitzaile grafikoak eta azpian erabiltzen den datu-motaren abstrakzio grafikoak hurbilekoak izan behar dute interfazea eraginkorra izan dadin. Gure kasuan, hortaz, kontsultarako interfazean zuhaitzekin berehalako lotura bisuala luketen elementuak erabili beharko lirateke. Adibidez, adabegiak eta arkuak.

- Interfazearen bidez, kontsulta-lengoia onartzen dituen kontsulta guztiak, sinpleenetatik konplexuenetara, egin behar dira.

*Treebank*etan galderak egin nahiak kontsultak modu grafiko eta erosoan egiteko beharra sortu du. *VIQTORYA* (A Visual Query Tool for Syntactically Annotated Corpora) kontsultarako interfazeak *Tübingen German Treebank*ean egiten ditu kontsultak (Steiner eta Kallmeyer, 2002). Galdeketak *kuantifikatzaile*rik gabeko *lehen ordenako logikako* adierazpenak erabiliz egiten dira, eta hauek gero SQLra itzultzen dira, aurreprozesu batean *treebanka* datu-base erlazional batera aldatzen baita. *VIQTORYA* tresnaren

bidez, zuhaitzak ez diren egituren gainean egin daitezke kontsultak. Ezau-garri bera du *TIGERSearch* softwareak (Lezius eta König, 2000). *TIGERSearch* tresnaren bidez, *TIGER treebank*aren gainean egiten dira kontsultak eta corpusaren deskribapenerako lengoia batek, kontsulten optimizaziorako baliabideek eta emaitzak ikusteko *TIGERGraphViewer* tresnak osatzen du-te. *TIGERSearch* softwarearen erabiltzaile berrientzat, grafikoa den kontsul-ta-editore bat garatu da *TIGERin* izeneko (Voormann eta Lezius, 2002). Tresna horretan, kontsultak adabegiak eta haien arteko erlazioak konbina-tuz marraz daitezke. Nygaard eta Johannessen-en arabera (2004), *TIGERin* tresna ikusgarria da bere adierazgarritasunagatik eta aukera grafikoengatik. Aipatu berri ditugun ikertzaileak *SearchTree treebank*etan kontsultak egite-ko interfazearean egileak dira. Nahiz eta *SearchTree* ez zaion inolako *tree-bank*i lotzen, *Sofie Treebank*ean²⁸ kontsultak egiteko tresna moduan sortu zen eta *TGrep2* kontsulta-lengoia erabiltzen du oinarrian. *Netgraph* tresna-ren bidez (Mírovský, 2006), “*Prague Dependency Treebank 2.0*” (PDT2.0) kontsulta daiteke Interneten bidez eta modu grafikoan. Tresna bezero-zer-bitzari arkitektura batean oinarritzen da, eta kontsultak bai PDT2.0-ko maila analitikoan baita tektogramatikalean ere ahalbidetzen ditu. PDTan bilaketak egiteko erabiltzen diren tresnak dira *Tred* eta *Oraculum*. *Tred*²⁹ zuhaitzak editatzeko tresna den arren, kontsultarako ere erabiltzen da, baina ataza honetarako Perl programazio-lengoia erabiltzen denez, programatzai-leentzako lengoia dela esan dezakegu. *Oraculum* (Ljubopytnov *et al.*, 2002) bezero-zerbitzari arkitekturari oinarritzen da, tresna boteretsua da, baina ez da eroso, eta zenbait ikasleren proiektua izanik, ez du jarraipenik izan.

Kontsultarako interfaze grafiko batek eduki beharko lituzkeen ezauga-riak aztertu ondoren, eta gure lanerako erreferentzia izan daitezkeen lanen errepasoa egin eta gero, *Saroi*rako interfaze grafiko baten azaleko diseinua azalduko dugu. Ideia nagusiak emango ditugu soilik, inplementazioa etorki-zuneko lanerako utziko baitugu. Interfazean bi ataza nagusi garatuko lirate-ke: kontsulta eta emaitzaren bistaratzea.

Kontsulta-atala

Kontsulta zehaztu baino lehen, ezinbestekoa da corpora aukeratzea. Hone-tarako egokiena direktorio-egitura aztertu eta bertan fitxategi indibidualak zein direktorio osoak aukeratu ahal izatea litzateke.

Corpora aukeratu ondoren, kontsulta idazteari edo marrazteari ekingo

²⁸Sofie *treebank*a garapen-fasean dago eta Europako Iparraldeko *Treebank* Sareko kideen lanaren ondorioa da (daniera, nederlandera, estoniera, ingelesa. . .)

²⁹<http://ufal.mff.cuni.cz/pdt2.0/>

genioke. Aipatu ditugun sistema askotan egiten den moduan, gure ustez garrantzitsua da bai kontsultaren bertsio grafikoa baita testu-forma ere eskura edukitzea, eta biak editatu ahal izatea. Elementu grafikoek adabegi eta arku itxura edukitzeak intuitiboagoa egingo luke interfazea eta, noski, etiketa linguistiko eta hauen balioak erabilgarri edukiko genituzke uneoro.

Interfaze grafikoa VI.19 irudikoaren itxurakoa izango da. Interfazearen oinarritzko egitura bi ataletan banatzen da: goialdean kontsulta-atala eta azpialdean emaitza-atala. Kontsulta-ataleko eskuinaldeko kontsulta-erregela grafikoki irudikatzeke modua labur-labur azaldu dugu VI.9 taulan. Gure ustez kontsultak zuhaitz itxura edukitzeak asko errazten dio erabiltzaileari kontsulta-lana. Gainera, emaitzan jasoko dituzten mendekotasun-zuhaitzek azpizuhaitzen batean eskatutako egitura grafiko hori bera edukiko dute.

Kontsultan adabegietako ezaugarri linguistiko guztiei buruz galdetu ahal izango da. Ezaugarri batzuetan bilatutako balioa deskribatzen duen testua idatzi beharko da kontsultarako (adib. *lema*, *forma*...) eta beste zenbaitetan, berriz, balio posibleen enumerazioa edukiko dugu eskura. Ezaugarri morfologiko guztien zehaztapena *XML Schema* batean islatuta dago dagoeneko. Adabegien arteko konbinazioen bidez kontsulta konplexuagoak egin ahal izango dira emaitzaren bila.

Emaitza-atala

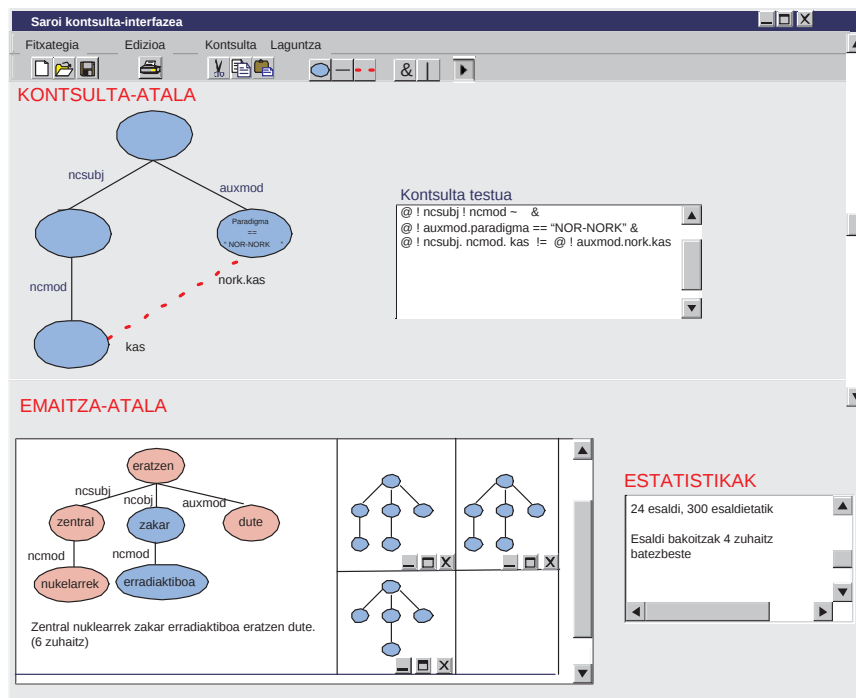
Kontsulta-erregelak mendekotasun-zuhaitzetan aplikatzearen ondorioz, lortuko dugun emaitza honakoek osatuko dute: mendekotasun-zuhaitza bera, bertan erregularik aplikatu ote den edo ez, eta lotuta duen informazio linguistiko guztia XML dokumentuetan bilduta.

Aipatu berri dugun informazio orokorretik abiatuta, hainbat iragazki defini daitezke emaitzak emateko. Proposamen bat egin dugu VI.19 irudian. Bertan, kontsultan definitutako baldintzak betetzen dituzten mendekotasun-zuhaitzen irudiak eta zuhaitzei dagozkien esaldiak agertzen dira. Normalean esaldi bakoitzaren analisiak hainbat zuhaitz sortuko ditu. Osatuena, adar gehien duena, erakutsiko dugu handien, beti ere, gainontzeko zuhaitzak eskura edukiko genituzkeelarik.

Mendekotasun-zuhaitzen bistaratzaile grafiko bat dagoeneko garatuta daukagu IXA ikerkuntza-taldean. ESALT du izena eta Abar-Hitz euskarazko *treebank*a eskuz etiketatzeke laguntza-tresnan (Díaz de Ilarraza *et al.*, 2004) integratuta dago. ESALTen sorkuntzarako oinarrian kontzeptu-mapen “*CM-ED*” editorea (Arruarte *et al.*, 2001) erabili genuen. ESALTen mendekotasun-zuhaitzak grafikoki alda daitezke etiketa berriak eta balioak gehituz, azpizuhaitzak bilduz edo zabalduz, adabegiak eta arkuak ezaba-

tuz edo gehituz. . . Bistaratzaila edozein aplikaziotik independentea da, eta beraz, *Saroi*ren interfazeen erraz integra daiteke.

Hizkuntzalariek *treebank*ean egin nahiko lituzketen kontsultak errepa-satu ditugunean, *zein* eta *zenbat* esaldik betetzen ote dituzten baldintzak jakin nahiko luketela ikusi dugu. Lehen galderaren erantzuna nola eman azaldu dugu aurreko paragrafoan. Zenbat esaldik betetzen duten baldintza erantzuteko, estatistikak emango lituzkeen atal bat gehitu beharko genuke interfazean. Oraingoz, estatistika sinpleak emango lirarteke. Adibidez, esaldi kopurua eta esaldiko zuhaitz kopurua.



VI.19 Irudia: *Saroi*ren kontsulta-interfazearen diseinua.

VI.3 Laburpena eta ondorioak

Atal honetan, mendekotasun-zuhaitzetan kontsulta linguistikoak egiteko diseinatu dugun *Saroi* tresnaren diseinua eta inplementazioa aztertu ditugu. Helburu orokorreko tresna da, ez dago fenomeno edo ataza bakar bati lotuta, eta lau moduluz osatuta dago: analisi sintaktikorako modulua, kontsulta-erregelen itzulpenerako modulua, kontsulta-erregelak mendekotasun-zuhaitzetan aplikatzen dituen modulua eta, bukatzeko, emaitzak aurkezten dituen modulua (gogoratu VI.1 irudia).

Analisi sintaktikorako modulua esaldi errealean mendekotasun-zuhaitzak kudeatzeko eta aukeratutako informazioarekin aberasteko gai da. Nahiz eta moduluaren sarrerako informazio linguistikoa partziala eta ez-osoia izan (mendekotasun-etiketak faltan, desanbiguazio okerrak...), *Saroi* gai da zuhaitzak eraikitzeko. Mendekotasun-zuhaitzok gainontzeko *Saroi*ren moduluak erabili gabe modu independentean eraiki daitezke.

Kontsultetarako erabiltzen dugun informazio linguistikoa nahikoa iruditzen zaigu garatu ditugun aplikazioetarako (ikus VII. kapitulua). Beste zenbait aplikaziotarako, beharrezko informazioa gehitzea erraza da, hala ere. Gainera, definitutako eragileek ahalbidetzen dituzten kontsulta guztiak aurrera eramateko gai da *Saroi*, bai *treebank*ean, baita analisi automatikoaren ondorioz sortutako zuhaitzetan ere.

Kontsultak mendekotasun-zuhaitzetan eta ez dokumentu-egituretan egiteak, abstrakzio-maila altua ematen dio *Saroi*ren kontsulta-lengoiari, eta programatzaile ez diren erabiltzaileentzako erabilgarri egiten du *Saroi*.

Emaitzak aurkezteko modua estandarra (XML) eta egokia iruditzen zaigu. Hala ere, diseinatutako interfaze grafikoaren inplementazioak asko hobetuko luke *Saroi*rekiko elkarrekintza.

*Saroi*ren ezaugarriei dagokionez, sendoa dela esan dezakegu, eta beraz, fidagarria. Erabilgarria dela uste dugu, erraz uler, ikas eta erabil daitekeen neurrian. Eraginkortasunari dagokionez, *Saroi* C++ programazio-lengoaian inplementatu da, exekuzio-denboran azkarrenetarikoa den eta memoria kontsumo txikienetarikoa duen programazio-lengoaian (Prechelt, 2000). Mantentzeko erraztasuna eta garraiagarritasuna ez ditugu bereziki landu, esan bezala, erabiltzailearen ikuspuntutik landu baitugu *Saroi*.

Hurrengo kapituluan (VII zenbakidunean), sakonago aztertu ditugu *Saroi*ren bi aplikazio: komuntadura-erroreen detekzioa eta hiztegi baten abesketa.

VII. KAPITULUA

***Saroiren* erabilera komunztadura-erroreak detektatzeko eta hiztegiak aberasteko**

VII.1 Sarrera

Esku artean dugun kapitulan, tesi-lan honetako helburu nagusiei ongi lotzen zaion *Saroiren* erabilera bat azalduko dugu: komunztadura-erroreen detekzioa. Tresnaren orokortasuna eraskustearren, kapituluko azken atalean (ikus VII.8 atala) *Saroi* Euskal Hiztegian aditzen informazioa aberasteko nola erabili dugun azalduko dugu.

Has gaitezen bada, mendekotasun-zuhaitzetan komunztadura-erroreak bilatzeko egin dugun lan enpirikoa azaltzen. Horretarako, lehendabizi, helburu eta metodo berberarekin sortutako txekierarako (Holan *et al.*, 1997), errusierarako (Mitjushin, 1996) eta frantseserako (Courtin *et al.*, 1991) sistemei gainbegiratu azkar bat emango diegu (hauek III. kapitulan aipatu ditugu eta A eranskinean sakondu). Lehenak, bereziki txekiera lantzeko sortutako *Robust Free Order Dependency Grammar* erabiltzen du. Txekierarako sistemak analisi sintaktikoa egiten du, eta analisisia egin ahal izateko trinkotasun falta deskribatzen duen metaerregelaren bat erabiltzen bada, esaldia modulu ebaluatzaile batera bidaltzen du. Esaldia analizatzen duen bitartean egiten du erroreen detekzioa. Ez da gure kasurako erabilgarria, guk analisisia egina baitugu dagoeneko, gure formalismo propioa erabiliz. Gauza bera gertatzen da errusierarako egindako lanean. “Meaning Text” izeneko eredu linguistikoan oinarria duen sistema batekin analizatzen dute esaldia, eta honi ezin zaionean egitura sintaktiko bat esleitu, *sintaktikoki konektatuta ez dagoenean*, egituraren errorea dagoela susmatzen dute. Berriro ere, analisisia eta

errore-detekzioa estu lotuta agertzen zaizkigu. Gure hurbilpenetik hurbilago dago frantseserako garatutako sistema. Bertan, mendekotasun-zuhaitzak eraikita daude eta baldintza-ekintza moduko erregelak aplikatzen zaizkie haiei, komunztadura-erroreen bila. Zuhaitzak eraikitze erabiltzen duten sistema PILAF da (Courtin, 1977) eta honek oso modu berezian erreprezentatzen ditu zuhaitzak. Sintaxia egiaztatze erregelak PROLOG programazio-lengoaia erabiliz idatzi dituzte. Guk analisi-metodo propioa dugunez, azken aukera hau ere, ez zaigu erabilgarria iruditzen.

Aztertu berri ditugun sistemetan erroreen detekzioa analisi-metodoaren dependente da. Guk analisia egin jasotzen dugu, eta sarrerako XML dokumentuak egokiak diren bitartean, berdin zaigu analisi-metodoa zein den (EDGKrekin eta *MaltIxarekin* egin dugu proba). Hau izan da, hain zuzen ere, komunztadura-erroreen detekzioa egin ahal izateko *Saroi* tresna garatzeko eta erabiltzeko arrazoiak.

Komunztadura-erroreen detekzioa VII.2 atalean azalduko dugu eta corpusetatik abiatuta hiztegietako aditzen azpikategorizazio-informazioa aberasteko proposamena VII.8 atalean egingo dugu.

VII.2 Komunztadura-erroreak

VII.2.1 Definizioa

Adituen arabera, euskaraz egiten diren erroreen artean maiztasun handienetakoa dutenak dira esaldiko elementuen arteko komunztadurarik ezarekin zerikusia dutenak (ikus II. kapituluaren aitaturako itaunketa). Errore horien azalpena eta azterketa hobeto uler dadin, *komunztaduraren* definizio bat ematen hasi dugu atal hau. Harluxet hiztegi entziklopedikoan honela definitzen da komunztadura:

«Esaldi berean loturik dauden bi hitzen artean gertatzen den kideko morfemen arteko adostasuna (...)»

Harluxet hiztegia (*komunztadura* hitzaren definizioa).

Esaldiko kideko morfemetan adostasun hori galtzen denean, *komunztadura-errore* bat gertatu dela esango dugu.

Euskarazko komunztadura-erroreen sailkapen bat egin dugu, adostu beharreko elementuak zein *testu-zatitan* kokatuta dauden begiratuta. Sailkapena, mota bakoitzeko adibide batekin lagundu dugu.

1. **Komunztadura esaldian.** Esaldia osatzen duten perpausen arteko komunztadura behar da. Adibidez, gaizki egiten da komunztadura perpaus nagusiaren eta mendekoaren artean honako kasuan:

(37) *Ile zuria duen gizona [SING.] etorri dira [PL.].

2. **Komunztadura perpausean.** Perpausaren barneko osagaien arteko komunztadura behar da. Aditzak, perpausean ager daitezkeen hiru sintagmekin (nor, nori eta nork sintagmekin) gorde behar du komunztadura. Subjektuaren eta aditzaren arteko komunztadura gaizki egin da adibide honetan:

(38) *Arranoak [SUBJ-PL.] hegan dator [SUBJ-SING.]

3. **Komunztadura sintagman.** Sintagma edo zatiko elementuen artean komunztadurak egon beharra du. Mugatasunari dagokion komunztadura ez da mantentzen honako adibide honetan:

(39) *Zenbait [DET-MG] umeak [IZE-MUG] etorri dira.

4. **Komunztadura aposizioan.** Aposizioaren¹ eta perpaus nagusiaren arteko komunztadura eskatzen da (hau araututa dago Euskaltzaindiako 109. arauan). Aposizioa aparteko sintagma gisa duen esaldi erroredun baten adibidea da honakoa:

(40) *Bere anaiak [sing.], Jon eta Ikerrek [pl.], etxea duela hiru hilabete utzi zuten.

Sailkapen honetan alde batera utzi ditugu aditzetako denbora-komunztadura, eta izen-sintagmaren eta adjektibo-sintagmaren artean gertatzen den numero-komunztadura (gizona ona da, baina gizonak onak dira).

Adostu beharreko hizkuntza-ezaugarriari dagokionez, gure inguruko hainbat hizkuntzatan *generoari*, *numeroari* eta *kasuari* dagokion lotura egin behar bada, euskaraz genero gramatikalik ez dagoenez, orokorrean numeroaren eta kasuaren inguruan egiten dugu lan. Sintagma da euskaraz kasu-komunztadurarik ez duen “testu-zati” bakarra, sintagmari kasu-marka behin soilik eransten baitzaio. Sintagman gertatzen diren komunztadura-erroreek, numero-mugatasunarekin dute zerikusia (ikus (39) adibidea).

¹Aposizioa, izen bati buruzko zehaztapen edo azalpen bat ematen duen elementua da; batzuetan sintagma berean txertatua, eta beste batzuetan koma arteko aparteko sintagma bezala; alegia tarteki gisa.

Aipatutako lau komunztadura motetan erreoren bila jardutea lan izugarria izateaz gain, ez da kasu guztietan bideragarria. Esaldien barruko komunztadura lantzeko, esaldiko perpausak ezagutu beharko lirateke. IXA taldean ataza honetan lanean dihardugun arren, ez dago esaldiko perpausak esplizituki markatzen dituen tresnarik. Batzuetan, ordea, mendekotasun-erlazioak ezartzen dituen gramatika gai da esaldiko mendeko perpausen eta perpaus nagusien arteko erlazioa ezartzeko². Kasu horietan esaldiko komunztadurak landu ahal izango ditugu. Antzekoa gertatzen da aposizioekin³. Laburtuz, perpausen eta sintagmen barruko osagaien arteko adostasuna sistematikoki aztertuko dugu. Esaldietan eta aposizioetan gertatzen dena, esaldiko elementuen arteko mendekotasun-erlazioa ezartzen denean soilik.

Mota bakoitzeko komunztadura-erreoren maiztasuna ezagutu nahi izan dugu, perpausako eta sintagmako komunztadura lantzea erabaki egokia ote den ezagutzeko asmoz. Horretarako, mota guztietako komunztadura-erroreak dituzten 64 esaldi jaso ditugu eta eskuz etiketatu. Corpus txiki honetako esaldiek iturri desberdinak dituzte: karrera amaierako proiektuak, DBHko fisika-ikasleen lanak, II. kapituluaren zerrendatutako euskaltegiak. . . Azterketaren ondorioz VII.1 taulako emaitzak jaso ditugu.

Testu-zatia	Errore kop.	%
Sintagman	5	% 7,8
Perpausaren barruan	59	% 92,2
Esaldian mendeko perpausaren eta perpaus nagusiaren artean	0	% 0
Guztira	64	% 100

VII.1 Taula: Komunztadura-erroreak testu-zatietan.

Perpausen barruko osagaien arteko komunztadura-erroreak direla ohikoena adierazten dute emaitzek (erreoren % 92,2 dira), eta horien atzetik, sintagma barruan ematen direnak. Ez dugu esaldiko edo aposizioeko elementuen arteko desadostasunik topatu. Kopuru horiek eta aipatu berri dugun perpausak identifikatzeko ezintasunak, gure erabakia indartzen dute: batez ere perpaus- eta sintagma-mailako erroreak landuko ditugu. Ondorengo lefroetan sakonago aztertuko dugu perpausaren barruan gertatzen den aditzaren eta subjektuaren/objektuaren/zehar-objektuaren arteko komunztadura.

²IXA taldeko analizatzaile sintaktikoetan CMOD, XMOD, CCOMP_SUBJ, XCOMP_SUBJ...etiketen bidez zehazten da mendeko perpausaren eta perpaus nagusiaren arteko erlazioa.

³APOCMOD, APOXMOD, eta APONCMOD etiketen bidez zehazten da aposizioaren eta perpaus nagusiaren arteko erlazioa.

VII.2.2 Aditzaren eta subjektuaren/objektuaren/zehar-objektuaren arteko komunztadura

Euskal aditzak, perpausean ager daitezkeen hiru sintagmarekin (nor, nori eta nork sintagmekin) gordetzen duela komunztadura esan dugu. Sintagma horien eta aditz jokatuaren arteko lotura, hiru sintagma horiei dagozkien morfemak aditzari erantsiaz gauzatzen da. Aditz mota edo paradigma, aditzaren *komunztadura-marka* (*agreement marker* erabiltzen da ingelesez) horiek definitzen dute:

«The structure of finite verb forms in Basque is rather complex. The simplest forms of intransitive verbs are monovalent and mark agreement with the subject (NOR). Intransitive verbs can also have bivalent forms marking agreement with an absolutive argument (the subject of the clause) and a dative argument (NOR-NORI). Finite transitive verb forms are minimally bivalent, marking agreement with an ergative argument (the subject of the clause) and an absolutive (the direct object) argument (NOR-NORK). In addition, there are trivalent forms that add agreement with a dative argument (NOR-NORI-NORK). »

(Hualde eta de Urbina, 2001).

Aditz laguntzaileko edo aditz trinkoko komunztadura-marken arabera, aditza NOR, NOR-NORI, NOR-NORK edo NOR-NORI-NORK *paradigmakoa* izan daitekeela esango dugu. Aipuan esaten den moduan, aditz-paradigmaren arabera erabakiko da komunztadura esaldiko subjektuarekin, objektuarekin edo zehar-objektuarekin egiten den. Ondorengo adibidean (ikus (41) adibidea), **badauka** aditz trinkoa NOR-NORK paradigmakoa da (*hark, hura*), eta ondorioz, subjektuak (**nukleoa**) ergatibo kasu gramatikala behar du eta objektuak (**eskape bat**) absolutibo kasua. Subjektuak ergatibo kasu-markarik ez duenez, aditzaren eta subjektuaren⁴ arteko komunztadura ezagatikoko errore baten aurrean gaude.

(41) *Oso arriskutsua da nukleoa eskape bat baldin badauka.

Batzuetan arazoak egoten dira komunztadura-erroreak eta *paradigma nahasketak* bereizteko. Adibidez, (42) esaldian, **zait** aditz laguntzaileak (*ho-*

⁴Semantika aldetik **nukleoa** sintagma da subjektua. Kasu gramatikalari begiratuta, sintaxiari, ergatiboa faltan duenez, objektua izan daitekeela esan dezakegu. Kasu honetan esaldian bi objektu genituzke, eta beraz, errorea.

ri, *niri*) ongi egiten du komunztadura zehar-objektuarekin (*niri*) eta subjektuarekin (*hori*), beraz, ez dago komunztadura-erroreirik. *Molestatu* aditzak, ordea, NOR-NORK eta NOR-NORI-NORK paradigmako aditz laguntzailerak soilik onartzen ditu (ikus *Euskal Hiztegia*n *molestatu* aditzari dagokion hiztegi-sarrera). Hots, paradigma desegokia (NOR-NORI) erabili da. Aditzen azpikategorizazioarekin dago guztiz lotuta paradigma nahasketa.

(42) **Niri HORI ez ZAIT molestutzen*.

Perpausetako komunztadura-erroreak zein paradigma nahasketak landu ditugu tesi-lan honetan. Azkenak lantzeko aditzen patroiei buruzko informazioa hiru iturri desberdinetatik (corpusetatik, Euskal Hiztegitik eta eskuzko analisitik) jaso dugu (ikus aditzen aberasketa VII.4.2 atalean).

VII.2.3 Komunztadura aldatzen duten fenomenoak

Perpausetako komunztadurarekin egin dugun sailkapen xehatua azaltzen hasi aurretik, ohiko komunztadura-eskema aldatzen duen fenomeno bat aipatuko dugu: *emendiozko juntadura*. *EGLU-IV* gramatikan, zera aipatzen da emendioaren inguruan:

«Sintagma gramatikalak *-nor*, *nori* eta *nork* sintagmak-juntatzeak, beraz, emendioa batuketatzat edo gehiketatzat hartzen denez, eragina izango du nahitaez aditzaren komunztaduran:

- (43) [Gizon handi horrek eta beste andreak] liburu bat eman dioTE haurrari
- (44) Gizon handi horrek [liburu bat eta koaderno bat] eman diZKio haurrari
- (45) Gizon handi horrek liburu bat eman diE [haurrari eta irakasleari]

Emendiozko juntadurak, beraz, aditza plural komunztadura hartzera behartzen duela esan daiteke oro har.»

(EGLU-IV, 49. orrialdea).

Numeroaren batuketaren fenomenoak landu ahal izateko, “eta” emendiozko juntagailuaren azaleko azterketa egin dugu. Juntagailu horrek aditz laguntzaile edo trinkoaren *numeroa* VII.2 taulan adierazten den moduan aldatzen du. Taula honetako kasuistika, *EGLU-IV*ko (ikus 49. orr.) honako baieztapenaren zehaztapena dela esan dezakegu:

«Esan beharrik ez da, juntatutako sintagma horietako bat plurala izatea aski da sintagma juntatuta nahitaez pluralizat hartzeko:

(46) [Liburu bat eta orri batzuk] aurkitu niTuen

Bestalde, numero-komunztaduraz gain pertsona-komunztadura ere egiten da euskaraz. Alde honetatik, juntagaietako bat pertsona izenordaina bada, zein pertsonatakoa den ikusi beharko da komunztadurari begira (...). Laburbilduz:

ni/gu eta X = 1.pertsona + plurala
zu / hi /zuek eta X = 2.pertsona + plurala »

X	Emend. junt.	Y	→	Z	Adibidea
sing.	eta	sing.	→	pl.	Jon eta Miren joan dira.
pl.	eta	pl.	→	pl.	Bilboko jokalariek eta Bermeoko arrantzaleek batzarra egin dute.
pl.	eta	sing.	→	pl.	Daviden gudari adoretuak eta bere anaia Simon...
sing.	eta	pl.	→	pl.	... negoziazioa eta akordioak direla
Ø	eta	sing.	→	pl.	Bere emazte eta alabari galdetu zien.
Ø	eta	pl.	→	pl.	Maite ditut lagun eta ahaideak.
Ø	eta	MG	→	sing. & pl.	Neska eta hainbat mutil ikusi dut/ditut
MG	eta	sing.	→	sing. & pl.	Makina bat jende eta Ana etortzen da/dira.

VII.2 Taula: Emendiozko juntadura eta numero aldaketa.

Aipuan aipatutako pertsona-komunztadura ere kontuan hartu dugu komunztadura-erroreak lantzeko. Hori bai, Euskaltzaindiak emandako araua zehaztu behar izan dugu, hor ez baita esaten, lehen erregelak lehentasuna duela bigarrenaren gainean. Hau da, lehen erregelako izen ordain bat bigarren erregelako batekin juntatzen denean, lehena aplikatzen da. Adibideak dira, “**zu** + **ni** → **gu**”, “**hi** + **ni** → **gu**”, “**zuek** + **ni** → **gu**” e.a.

Emendiozko juntadurak aditzari plural komunztadura eragiten diola esan dugu baina lege horri men egiten ez dioten egitura ugari topa daitezke EGLU-IV gramatikan. Besteak beste, kideko edo erreferente bereko izen sintagmak juntatzen direnean (ikus EGLU-IVko 50. orrialdeko (47) adibidea), partitiboan doazen bi izen-sintagma juntatzean (ikus EGLU-IVko 54. orrialdeko (48) adibidea), aditza aurretik duten zerrendatzetan... adizkia singularrean erabiliko da.

- (47) Hemen erabili den arreta eta zehaztasuna
- (48) Nork nahi du ordean etxerik eta lurrik erosi?

VII.2.4 Komunztadura lantzeko zailtasunak

Komunztaduraren inguruan gertatzen diren erroreak detektatzeari ekin diogunean, hainbat arazo topatu dugu. Batzuk, *hizkuntzaren ezaugarriek* eragindakoak, beste batzuk berriz, *hizkuntzaren tratamendu automatikoak eragindakoak*. Azpiatal honetan, arazo hauen berri emango dugu eta hartutako erabakiak aipatuko ditugu.

Hizkuntzaren ezaugarriek eragindako arazoak

Elipsiak eta *anbiguotasunak* hainbat buruhauste sortu dizkigute:

- **Elipsia.** Aditzarekin komunztadura egin behar duten kasu gramatikalak perpausean ageri ez direnean, *elidituta* daudenean, oso zaila egiten zaigu komunztadura ongi egin ote den jakitea. Hori oso argia da, adibidez, **etorri dira** perpausean. Azken adibidean ezinezkoa da errorerik gertatu ote den jakitea, ez baitugu nahikoa informaziorik perpausean.

Kasu gramatikalen elipsia dela eta, komunztadura-erroreen bilaketa egiteko abiapuntu, **aditza** (laguntzailea zein trinkoa) hartzea erabaki dugu. Aditz laguntzailea edo trinkoa (komunztadura-markak dituen) finkatuko dugu eta honetan adierazten diren kasuak —funtzio sintaktikoak— bilatuko ditugu perpausean. Horiek perpausean azaltzen ez direnean, ezin dugu komunztadura faltarik detektatu.

- **Anbiguotasuna.** Gizakia gehienetan gai da testuingurua kontuan hartuta hizkuntzaren anbiguotasuna ebazteko. Analisi automatikoan anbiguotasunaren ebazpenari informazio semantikoaren falta, eta gure kasuan, egitura okerrean erabilera gehitzen badizkiogu, anbiguotasuna areagotu egiten da. Erakutsitako adibidean, (49), meta-ezagutza erabiliz erroreduntzat sailka dezakegun esaldi bat dugu. Ezagutza honen faltan, esaldia sintaktikoki zuzena dela esango dugu: **Eguzki izpiek** “zerbait” kontrola dezakete **ordenagailuen bidez**.

- (49) ?Eguzki izpiek uretan nola erreflejatzen diren kontrolatzen dute ordenagailuen bidez.

Hizkuntzaren tratamendu automatikoak eragindako arazoak

Elipsiari eta anbiguotasunari, hizkuntza-informazio osoaren gabezia eta honek izan ditzakeen akatsak gaineratu behar dizkiogu hizkuntzaren tratamendu automatikoa egiten dugunean.

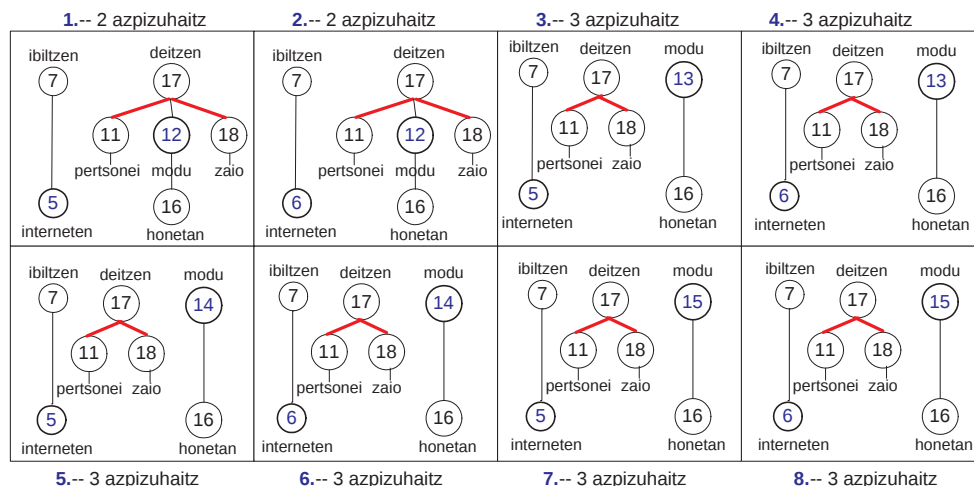
- **Analisi partzialak.** Erabiltzen ditugun analisiak bi zentzutan dira partzialak. Alde batetik, *ez ditugu esaldi bati dagozkion analisi-zuhaitz osoak lortzen*, hauen azpizuhaitzak baizik (ikus VII.1 irudia). Bestetik, analisi-zuhaitzen eraikuntza ahalbidetzen duten tresnek (EDGK-k eta *MaltIrak*) *ez dituzte euskarazko sintaxi-egitura guztiak osotasunean lantzen*, beraz, sarrerako esaldietan agertzen diren egitura batzuk (perpaus koordinatuak, aposizioak...) landu gabe geratzen dira. Bi fenomeno hauek lotuta egongo dira zenbaitetan, eta zuhaitz partzialak lortzeko arrazoia landu gabeko egituren agerpena izango da. Beste batzuetan, ordea, mendekotasun-erlazioa landuta egon arren, mendekotasun-etiketarik ez ezartzea gertatuko da.

Mendekotasun-zuhaitz “zatikatuak” erabiliz, erroreak detekta daitezkeen ziurtatzeko, analisiaren ondorioz lortzen diren azpizuhaitz guztietan egitura okerra bilatzea erabaki dugu. Har dezagun, (50) adibidea. Honi dagokion analisi sintaktikoa EDGKrekin egin ondoren, VII.1 irudiko 8 analisi partzialak lortu ditugu⁵ eta horietako bakoitzean bizpahiru azpizuhaitza. (50) adibideko esaldiaren kasuan, errorea detektatzeko behar dugun egitura sintaktikoa, analisi-zuhaitz guztietan agertzen da, beraz, errorea detektatu ahal izango dugu. Errore detektatzeko beharrezkoa den egitura honakoa da: “**pertsonei** sintagma **deitzen** aditzaren zehar-objektua da, eta **zaio** adizkia aditz beraren laguntzailea da”. Baldintza hauetan, *zehar-objektuaren eta aditz laguntzailearen arteko numero-komunztadura eza* detektatzeko gai gara. Beste batzuetan, ordea, behar dugun egitura ez da analisi guztietan azalduko eta errorea ezingo dugu detektatu.

(50) ***Interneten ibiltzen diren pertsoneri modu honetan deitzen zaio.**

Bigarren arazoari, dependentzia-gramatikan sintaxi-egitura guztiak ez lantzeari, ezin diogu oraingoz tesi-lan honetan irtenbiderik topatu. Analizatzaile sintaktikoen estaldura handitzen den heinean, landutako egituren kopurua handitzen joango da.

⁵Zuhaitzak erabiltzen dituzten interpretazioetan desberdintzen dira.



VII.1 Irudia: Analisi partzialen adibidea.

Komunztadura-erroreak detektatzeko lanari ekin aurretik, erabiliko ditugun analizatzaile sintaktikoak ebaluatuko ditugu. Horrela, analizatzaileek emandako analisiak *treebank*arekin konparatu dira, dependentzia-erlazio osoak (mendeko-gobernatzaile bikoteak) ongi ipini ote dituzten aztertuz. Hots, mendekoari etiketa egokia ezarri behar zaio, eta gainera, gobernatzaile zuzenarekin lotu behar da, mendekotasun-erlazioa ongi ipinita dagoela erabakitzeko. Baldintza hauetan, EDGK-k % 48ko doitasuna eta % 46ko estaldura lortu ditu. Gogora dezagun, *MaltIxak* esaldiko token guztiei esleitzen diela dependentzia-etiketa bat, beraz, estaldura eta doitasuna neurtzeak zentzu handirik ez duenez, etiketa zuzenaren esleipena neurtzeko (*Labeled Attachment Score*, *LAS*) neurria erabiltzen du. *MaltIxak* % 76,76ko LAS balioa lortzen du. Balio orokor hauek gure iritziz ez dira konparagarriak, sistemak ezaugarri ezberdinak baitituzte: i) *MaltIxak* puntuazio-markak etiketatzen ditu eta hauekin oso emaitza onak lortzen ditu (% 98,97ko estaldura eta % 98,76ko doitasuna, beraz, LAS balioa asko igotzen da); EDGK-k ez ditu etiketatzen, horrek bere emaitza orokorretan eragin negatiboa duelarik. ii) EDGK-k hitz batzuetan ipintzen ditu dependentzia-etiketak eta *MaltIxak* guztietan. Gainera, ez dakigu ebaluazioak *treebankeko* zati beraren kontra egin ote diren.

Dependentzia-etiketetako bakoitza bere aldetik hartuta konparatzea, ordea, zentzuzkoa da. Etiketa horietako batzuen estaldura eta doitasu-

na VII.3 taulan erakutsi nahi izan ditugu, analisi-zuhaitzak eraikitzeke abiapuntua zein dugun erakusteko asmoz. Komunztadurarekin zerikusi zuzena duten etiketen emaitzak eman ditugu soilik.

Sistema	EDGK		<i>MaltIxa</i>	
	Doitasuna	Estaldura	Doitasuna	Estaldura
Mendekotasun-erlazioa				
<i>Subjektua</i> (NCSUBJ)	% 61	% 37	% 68,79	% 66,51
<i>Objektua</i> (NCOBJ)	% 61	% 43	% 71,94	% 78,45
<i>Zehar-objektua</i> (NCZOBJ)	% 30	% 82	% 85,21	% 71,14
<i>Aditz laguntzailea</i> (AUXMOD)	% 84	% 94	% 95,65	% 96,09
Guztira	% 48	% 46	% 76,76 (LAS)	

VII.3 Taula: EDGKren eta *MaltIxa*ren ebaluazioa, komunztadurarekin lotura zuzena duten mendekotasun-erlazioei dagokienez.

- **Akatsak analisi automatikoan.** Hizkuntzaren analisi automatikorako sortutako analisi-katea hainbat modulekin osatuta dagoela ikusi dugu IV. kapituluaren. Modulu horietako bakoitzak errore-tasa bat du eta hau mugatua izan arren, analisisa moduluetatik zehar pasatzen doan heinean, akatsak metatu egiten dira. Analisi-akatsak alarma faltsuen kopuruaren hazkundean eragina izaten dute. Kasu batzuetan, esaldi zuzenak erroreduntzat jotzen dira analisi-moduluetan pilatu diren erroreen eraginez (ikus zehaztasun gehiagoren bila VII.7.3.2 atala). Kapitulu honetan zehar mendekotasun-zuhaitz “egoki” terminoa askotan erabili dugu. Zaila da esaldi erroredun bati dagokion mendekotasun-zuhaitzak nolakoa izan behar duen zehaztea, beraz, ezin dugu esaldiari dagokion zuhaitz zuzenaz hitz egin. Zuhaitz “egokia” deituko dugu, lortutako analisi-zuhaitzen artean ahalik eta akats kopuru txikiena duen hori.

VII.3 Komunztadura-erroreen sailkapena

Fenomeno konplexua da aditzaren eta kasu-gramatikalen arteko komunztadurarena. Hori horrela izanik, komunztadurarekin lotutako erroreak sailkatzeko erabiliko ditugun kodeak, orain arte erabili ditugunetatik zertxobait aldatzen dira. Forma hau dute:

KategoriaNagusia_Aldaketa_Non_EzaugarriLinguistikoa_Aditza

- *Kategoria Nagusia*: KOM, KOMunztadura errorea dela adierazteko.
- *Aldaketa*: irizpide deskriptiboari jarraituz, komunztaduraren kasuan, funtzio sintaktiko nagusiak (subjektua, objektua eta zehar-objektua)

edo haien ezaugarriak esaldian kendu (KEN), gehitu (GEH) edo nahastu (NAH) egin direla esango dugu. Kasu honetan funtzio horiek ordenaz aldatu direla esateak ez du zentzurik, euskaraz esaldiaren barneko elementuen artean ordena librea baita. Gainontzeko aldaketa motak zehazte aldera, zera esango dugu, perpausetik subjektua, objektua edo zehar-objektua,

- “kendu” egiten denean, edo agertzen ez denean, ez dagoela komunztadura-errerorik. *Elipsia* dago. Ez dugu ezaugarri hau duen errore-koderik topatuko. Kasu honetan, aditzeko komunztadura-marketan adierazten diren elementu guztiak (demagun, objektua eta subjektua) ez dira esaldian agertzen.
- “gehitu” egiten denean, aditzean agertzen ez den elementuren bat agertzen dela perpausean. Aditzeko komunztadura-marketan adierazten diren elementuak (demagun markak subjektua eta objektua eskatzen dituela) baino elementu gehiago agertzen dira esaldian (adibidez, zehar-objektua). Hau da (51) adibideko kasua: *nor-nork* paradigmako aditz laguntzailea duen perpaus batean, *zehar-objektu* funtzioa edukiko lukeen sintagma agertzen da:

(51) *Neskak (SUBJ) lagunari (ZOBJ) liburua (OBJ) eraman du
(*hark hura*).

Gure ustez, kasu hau zuzen-zuzenean dago lotuta *paradigma nahasketarekin*.

- “nahastu” egiten denean, aditzean agertzen diren elementuei dagozkien perpauseko elementuetako pertsonak (PER), numeroak (NUM) edo kasuak (KAS) ez du komunztadurarik egiten aditzarekin. Hau da komunztadura-erroreen iturri nagusia.
- *Non*: zein funtzio sintaktikotan gertatzen den aldaketa, subjektuan (SUBJ), objektuan (OBJ) edo zehar-objektuan (ZOBJ)⁶.
- *Ezaugarri Linguistikoa*: ezaugarri hau, noski, nahasketetan soilik erabiltzen da. Orduan, bat ez datozen ezaugarriak, kasua (KAS), numeroa (NUM) edo pertsona (PER) izan daitezke.
- *Aditza*: aditz laguntzailearen edo aditz trinkoaren paradigma zehazten da atal honetan. Aditzak NOR (N), NOR-NORI (N-NI), NOR-NORK (N-NK) edo NOR-NORI-NORK (N-NI-NK) paradigmatakoak izan daitezke.

⁶*Predikatiboarekin* gertatzen den komunztadura oso azaletik landu dugu.

Beraz, ezaugarri horiek guztiak kontuan hartuta, KOM_NAH_SUB-KAS_N-NK etiketak honakoa esan nahiko luke: KOMunztadura errorea, NAHasketa dago SUBjektuaren KASuan; aditz laguntzailea NOR-NORK (N-NK) aditz paradigmatikoa da. Etiketa horrekin kodetuta dago, adibidez, (52) perpausa.

(52) *Adimen Naturala eboluzioa du eta artifiziala berriz ez.

Lehenago ere aipatu dugu komunztadura-erroreak detektatzeko errege-lak definitzen hasi aurretik, komunztadura-erroren bat duten 64 esaldi eskuz aztertu genituela (66 errore topatu genituen). Lan honen helburua bi-koitza izan da. Alde batetik, esaldiak komunztadura-erroreen sailkapenean kokatzen saiatu gara eta kodeketa egokia ote den egiaztatu nahi izan dugu. Ariketa zaila izan den arren, esaldi guztientzako topatu dugu errore-etiketaren bat. Beste aldetik, “testu-zatietan” erroreen maiztasunak zeintzuk diren ezagutzeko aukera izan dugu (gogoratu VII.1 taula). Perpaus barruko erroreen azterketa sakonago bat egin dugu oraingoan. Errore horien (59 komunztadura-errore) ezaugarri eta maiztasunak VII.4 taulan zehaztu ditugu.

Aditza	Esaldiko elementuak		
	<i>Subjektua</i>	<i>Objektua</i>	<i>Zehar-objektua</i>
Komunztadura-marketan <i>kasua</i>	30	0	1
Komunztadura-marketan <i>numeroa</i>	17	4	3
Objektua partitiboan eta dagokion komunztadura-markan plurala	3		
Objektua agertu da aditz iragangaitzarekin	1		
Guztira	59		

VII.4 Taula: Aritzarekiko komunztadura-erroreen maiztasuna.

Taulan 47 errore (guztira dauden % 79,6) aditzaren eta subjektuaren komunztadurarik ezagatik gertatzen dira. Arrazoia, aski ezaguna da: euskaraz, absolutibo-plurala eta ergatibo-singularra hitzetan gauzatzen direnean, morfema bera hartzen dute (-ak). Hori dela eta, *kasuaren* eta *numeroaren* nahasketa detektatzen dituzten errege-lak maiz batera aplikatzen dira egitura berean⁷.

⁷Hemen egin dugun aukeraketa, kasuaren nahasketa eta numeroaren nahasketa bereiz lantzea izan da. Errore bakartzat jo daitezke eta honela definitu: *absolutibo-plurala* vs *ergatibo-singularra* nahasketa

VII.4 Informazio linguistikoaren aberasketa

Saroi tresnaren aberasketarako aukerazko moduluak (ikus VI. kapitulua) analisi-katetik jasotzen duen informazio linguistikoari informazio gehigarria eranstean dio. Uneko atazarako, komunztadura-erroreen detekziorako, ezinbestekoa iruditu zaigu *koordinazioa* esplizitu egitea eta *aditzak* aberastea.

VII.4.1 Sintagma-koordinazioa esplizitu egitea

Komunztadura aldatzen duten fenomenoaren artean *emendiozko juntadura* aipatu dugu, honek aditzen komunztadura pluralean egitera bultzatzen baitu. Perpausen zein sintagmen arteko koordinazioa oraindik ez dugu ikerkuntza-taldean landu, baina komunztadura ezaren detekziorako behar-beharrezkoa da gure ustez. Hori dela eta, *sintagmen arteko juntadura lantzeko* lehen hurbilpena egin dugu. Prozesua azalduko dugu, (53) adibideko esaldi zuzena erabiliz:

(53) **Bilboko jokalaria eta Bermeoko arraunlariak bilera egin dute.**

1. *Murritzapen-gramatika* bat garatu dugu VII.2 taulako kasuak biltzen dituen. Gramatikak `ETA_S_P_HAIEK` etiketa ipintzen dio sintagmak biltzen dituen “eta” orori. Etiketak, bildutako sintagmaren numeroak `P` (plurala) izan behar duela eta `HAIEK` pertsonarekin egin behar duela komunztadura adierazten du. Adibidez, aipatu berri dugun esaldia-ren analisiari gramatika aplikatu ondoren, VII.2 irudiko analisia⁸ lortu dugu. Bertan, **arraunlariak** hitzak `&NCSUBJ>` mendekotasun-etiketa du (subjektu funtzioa duen sintagma **Bilboko jokalaria eta Bermeoko arraunlariak** da) eta emendiozko juntagailuak `ETA_S_P_HAIEK` etiketa.
2. Bigarren fase batean, programa bat garatu dugu loturazko elementuak adierazten dituen `LOT` mendekotasun-etiketa ipintzeko eta beste zenbait etiketa *lekuz aldatzeko*. Adibidez, VII.3 irudian, **jokalaria** hitzari `&LOT>`⁹ etiketa ipini diogu juntagailuarekin duen lotura adierazteko. Asmo berarekin esleitu diogu **arraunlariak** hitzari `&<LOT`. Sortu dugun sintagma berriko (**Bilboko jokalaria eta Bermeoko arraunlariak**) emendiozko juntagailuraino (**eta**) garraiatu dugu `&NCSUBJ>` etiketa.

⁸% ikurrarekin hasitako etiketek sintagmak mugatzen dituzte. %SIH: sintagmaren hasiera, %SIB: sintagmaren bukaera, %SINT: osagai bakarreko sintagma, %ADIKATHAS: aditz-katearen hasiera, %ADIKATBU: aditz-katearen bukaera.

& ikurrarekin hasitakoak mendekotasun-etiketak (`&NCSUBJ`, `&NCMOD...`) dira.

⁹`&LOT>` etiketak mendeko elementu honek (**jokalaria**) bere gobernatzailea (**eta**) eskunera (“>”) duela adierazten du. `&LOT<` etiketak, gobernatzailea ezkerrean (“<”) duela.

```

"<Bilboko>"
  "Bilbo"IZE LIB GEL NUMS MUGM HAS_MAI @IZLG> %SIH &NCMOD>
"<jokalaririk>"
  "jokalaririk"IZE ARR ERG NUMS MUGM @SUBJ
"<eta>"
  "eta"LOT JNT EMEN @PJ {ETA_S_P_HAIEK
"<Bermeoko>"
  "Bermeo"IZE LIB GEL NUMS MUGM HAS_MAI @IZLG>
"<arraunlariak>"
  "arraunlari"IZE ARR ERG NUMS MUGM @SUBJ %SIH &NCSUBJ>
"<bilera>"
  "bilera"IZE ARR ABS NUMS MUGM @OBJ %SINT &NCOBJ>
  "bilera"IZE ARR ABS NUMS MUGM @PRED %SINT
  "bilera"IZE ARR ABS NUMS MUGM @SUBJ %SINT &NCSUBJ>
"<egin>"
  "egin"ADI SIN PART BURU NOTDEK @-JADNAG %ADIKATHAS
"<dute>"
  "*edun"ADL A1 NOR_NORK NOR_HURA NK_HAIEK-K @+JADLAG %ADIKATBU &<AUXMOD

```

VII.2 Irudia: Sintagmen koordinazioa esplizitu egitea. Adibidea.

```

"<Bilboko>"
  "Bilbo"IZE LIB GEL NUMS MUGM HAS_MAI @IZLG> %SIH &NCMOD>
"<jokalaririk>"
  "jokalaririk"IZE ARR ERG NUMS MUGM @SUBJ &LOT>
"<eta>"
  "eta"LOT JNT EMEN @PJ {ETA_S_P_HAIEK_2 &NCSUBJ>
"<Bermeoko>"
  "Bermeo"IZE LIB GEL NUMS MUGM HAS_MAI @IZLG>
"<arraunlariak>"
  "arraunlari"IZE ARR ERG NUMS MUGM @SUBJ %SIH &<LOT
"<bilera>"
  "bilera"IZE ARR ABS NUMS MUGM @OBJ %SINT &NCOBJ>
  "bilera"IZE ARR ABS NUMS MUGM @PRED %SINT
  "bilera"IZE ARR ABS NUMS MUGM @SUBJ %SINT &NCSUBJ>
"<egin>"
  "egin"ADI SIN PART BURU NOTDEK @-JADNAG %ADIKATHAS
"<dute>"
  "*edun"ADL A1 NOR_NORK NOR_HURA NK_HAIEK-K @+JADLAG %ADIKATBU &<AUXMOD

```

VII.3 Irudia: Sintagmen koordinazioa esplizitu egitearen ondorioz, zenbait mendekotasun-erlazioen garraiatzea. Adibidea.

3. Bukatzeko, analisi linguistikoa biltegitratzen duten 3 XML fitxategi aldatu ditugu. Dependentsiak adierazteko erabiltzen diren bi fitxategiak (*dep.xml* eta *deplnk.xml*) berriro sortu ditugu; eta, analisi morfosintaktiko aberastua gordetzen duen fitxategian (*lemAberas.xml*) emendiozko juntagailuari dagokion informazioa aldatu dugu: numeroa singularretik pluralera eta “haiek” pertsona gehitu diogu.

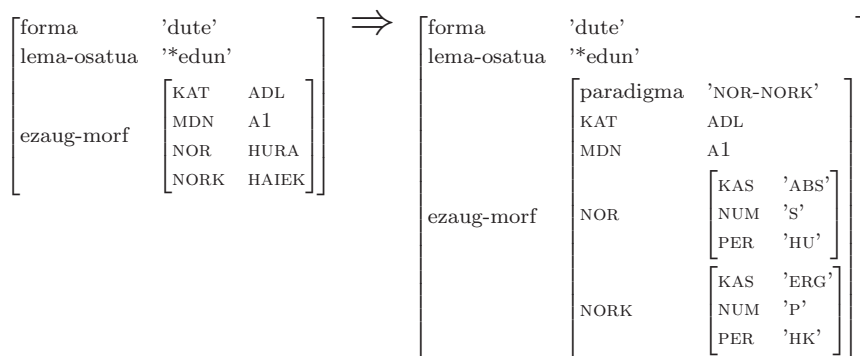
VII.4.2 Aditzen aberasketa

Aditza da, gure ustez, komunztadura-erroreen detekziorako oinarritzko elementua. Ahalik eta informazio gehien gehituko diegu bai aditz laguntzaileei baita aditz nagusiei ere, komunztaduraren egiaztatzea errazagoa izan dadin.

Aditz laguntzaileak

Aditz laguntzaileetan komunztadurarekin zerikusi zuzena duten hizkuntza-ezaugarriak esplizitu egin ditugu. Hain zuzen ere, honakoak erakutsi ditugu: aditz laguntzaileen paradigma edo saila (NOR, NOR-NORI, NOR-NORI-NORK, NOR-NORK) eta bere osagaietako bakoitzaren (NOR, NORI eta NORK) *kasua*, *numeroa* eta *pertsona*. Adibidez, **dute** aditz laguntzailean komunztadura-marketako ezaugarriak ageriko egin ditugu VII.4 irudiko eskuinaldeko ezaugarri-egituran¹⁰. Irudiaren ezkerreko aldean EUSTAGGERek **dute** hitza-erentzat emango lukeen analisia ikus dezakegu (*lem.xml* dokumentuaren zati bat erakusten da). Bertan, NOR osagaiaren balioa HURA da, eta NORK osagaiarena, HAIEK. Horrek ez du aditz laguntzaileak funtzio gramatikalekin komunztadurarik egiten ote duen edo ez automatikoki aztertzeiko nahikoa informaziorik ematen. Implizitu dagoen informazioa argi eta garbi erakutsi behar dugu. Horregatik egin ditugu zenbait ezaugarri esplizitu eskuin aldeko ezaugarri-egituran, eta adibidez, NOR adizki-osagaiari, *absolutibo* kasua (KAS=ABS), *singularra* numeroa (NUM=S) eta *hura* pertsona esleitu dizkiogu (PER=HU).

¹⁰Irudi honetan XML dokumentua aurkeztu ordez, ezaugarri-egiturak erakutsi ditugu intuitiboagoak direla uste baitugu.



VII.4 Irudia: Ezkerraldean **dute** aditzaren analisia. Eskuinaldean analisi bera *aberasketa*-prozesuaren ondoren.

Aditz nagusiak

IXA taldean hainbat urtetan zehar aditzetara eta beraien erabilerari buruz bildu den informazioa erabiltzea interesgarria iruditu zaigu aditzaren komunztadura eta paradigma nahasketak aztertzeke. Informazio-iturri aberatsak erabili ditugu; hala nola, eskuz egindako azterketa baten ondorioz bildutako aditz-ereduak (Aldezabal, 2004), aditzek corpus zehatz batean erakutsi dituzten gauzatzeak (Aldezabal *et al.*, 2003b) eta hiztegi bat. Aditzetara buruzko informazio hori guztia analizatutako aditzetan gehitu da (*lem.xml* dokumentuan), eta gero *Saroiko* erregelen bidez arakatu.

- *Eskuzko azterketaren ondorioz lortutako patroiak*. Aldezabalek bere tesi-lanean 100 aditzen eskuzko azterketa sakona egin zuen (Aldezabal, 2004). Aditz hauen azpikategorizazioa (pieza lexikalen gauzapeen sintaktikoa) egiterakoan, aditzen *alternantziak*¹¹, hau da, aditz beraren forma eta balio sintaktiko/semantiko (aurrerantzean *bss*) desberdinak aztertu zituen. *Rol tematikoekin* (adibidez, *egilea*, *kausa*, *gaia*...) eta rol horiek sintaxi-mailan gauzatzen dituzten *Kasuekin*¹² deskribatzen

¹¹ Alternantziaren deskribapen sakonagoa egiten da Aldezabalen tesian (Aldezabal, 2004), 89. orrialdean.

¹² Aldezabalek *Kasu* terminoa zentzu zabal batean erabiltzen du eta *Kasuen* artean a) Kasu gramatikalak nahiz gainontzeko deklinabide-kasuak, b) postposizioak (edo deklinabide-kasu konposatuak) (etxetik *at*, bidean *zihar*...) eta c) mendeko Kasu sinpleak (*-t(z)ean*, *-rako*...) nahiz konposatuak (duen *bitartean*, *duela*...), sartzen ditu. Hauen kopurua handia izanik, *Euskaldunon Egunkari*ko corpusean % 1etik gorako maiztasuna erakutsi dutenak soilik aukeratu ditu: absolutiboa (abs), ergatiboa (erg), inesiboa (ine), datiboa (dat), elatiboa (ela), instrumentala (ins), soziatiboa (soz), ablatiboa (abl) eta adlatiboa (ala).

da *bss* bakoitza. Tesi-txosten honetan egindako lana sintaxi-mailara mugatu dugunez, rol tematikoak alde batera utziko ditugu, eta aditz laguntzaile *mota* (iragankor datibogabea (DU), iragangaitz datibogabea (DA), ZAIO edo DIO) eta komunztadurarekin zerikusia duten Kasuak aztertuko ditugu.

Aditz hauei buruzko informazioa *XX mendeko euskararen corpus estatistikotik* (1900etik 1994ra bitarteko testuak) eta *Euskaldunon Egunkaririk* (1999ko urtariletik 2000ko maiatza bitarteko aleak) jaso zen eta datu-base batean biltegitatu.

Aldezabalek datu-basean aditz bakoitzaren azpikategorizazio-egiturak edo *bssak etorri* aditzaren deskribapenerako erabili den VII.5 taulako moduko egituretan¹³ antolatu ditu:

<i>bssak</i>	<i>abs</i>	<i>erg</i>	<i>dat</i>	<i>ela</i>	<i>ine</i>	<i>ins</i>	<i>soz</i>	<i>abl</i>	<i>ala</i>
etorri-1	✓	—	+	—	+	+	+	✓	✓
etorri-2	✓	—	—	—	+	+	+	✓	—
etorri-3	✓	—	—	—	✓	+	+	—	—
etorri-4	✓✓	—	—	—	+	—	—	—	—
etorri-5	✓	—	✓	—	+	+	+	+	—

VII.5 Taula: *etorri* aditzaren *bssak*.

Taulan errenkadetako bakoitzean *etorri* aditzaren *bss* bat azaltzen da. Egitura horietako batek adierazten duen Kasua azpikategorizatua denean (aditzak berezkoa du eta Kasu hori *bssa* zehazteko beharrezkoa da) “✓”¹⁴ ikurraz markatzen da. Kasua onartua denean “+” ikurraz markatuko da (ez dago azpikategorizatua baina ez da ezinezkoa *bss* horretan) eta “—” ikurraz onartzen ez denean. Aditzarekiko komunztadura egiten duten Kasuak dira guri bereziki interesatzen zaizkigunak, kasu ergatibo, absolutibo eta datiboak, hain zuzen ere. Aldezabalek *aurrera begirakoetan* aditzetan oinarrituta sintaxi-erroreak detektatzeko modu bat proposatzen zuen:

«Kasuei bakarrik erreparatuta (eta ez Kasu hauek duten balio semantikoari), “—” ikurra erabil daiteke, aditzaren

¹³Taula hau Aldezabalen tesi-txosteneko 177. orrialdean osorik azaltzen da.

¹⁴Absolutibo kasuarekin “✓✓” marka ager dakiguke. Egilearen arabera, predikatibotzat jotzen diren sintagmak kasu absolutibo mugagabea doazela esaten da. Ondorioz, esaldi batean bi absolutibo ager daitezke eta beharrezkoak dira. Adibidez, **Etxeak politak dira**.

balio semantiko guztietan azaltzen bada. Hala, beti iragan-gaitza den aditz batean (adib. *etorri*), ergatiboa okerra izango da (kausatibo morfologikoa ez bada); edota *ikusi* gisako aditz batean, Kasu adlatiboa okerra izango da. »

Adibideetara ekarrita, VII.5 taulan *etorri* aditzak ez lituzke **etorri du* edo **etorri dio* modukoak onartuko, bere *bss* guztietan “—” baitu ergatibo Kasuari dagokion zutabearen.

Komunztadura-erroreen detekziorako aditz-analisiak aberasteko asmoz, datu-baseko ehun aditzoi dagozkien *bss* guztiak jaso ditugu Kasuekin batera. Analizatutako aditza hauetako bat denean, Kasuren baterako *bss* guztietan “—” ikurra ote duen aztertzen da. Kasu horiek aditzaren analisisan gehitzen dira, aditza kasu horiekin *ez dela gertatzen* adieraziz. Analizatutako esaldian Kasu horietako bat azaltzen bada, gerta liteke errore baten aurrean egotea.

Aditzen patroiak eskuz aztertzearen abantaila nagusia jasotako informazioaren zuzentasuna eta sakontasuna da. Kontran, lan izugarria eskatzen du eta egindako adibide aukeraketaren arabera, egitura zuzen posible guztiak ez agertzea gerta liteke.

- *Corpusen azterketa automatikoa eginez lortutako aditzen gauzatze-egiturak*. Atutxaren tesi-proiektuan kokatutako lan batean (?), 2541 aditz aztertu dira corpusetan zein sintaxi-egituretan gauzatu diren ezagutzeko (367 hitz anitzekoak, adibidez, *tenk egin*). Aditz-zerrenda bat lortu da non aditz bakoitza bere testuinguruan agertzen den dagozkion sintaxi-mailako mendekoekin eta argumentuekin. Aditz bakoitza zein paradigmatako laguntzailek zein proportziotan laguntzen duen ere adierazten da.

Aditzek corpusean agertzen dituzten sintaxi-egiturak aztertzeke erabili den corpusa *Euskaldunon Egunkariko* 2000 eta 2001 urteko aleekin osatuta dago.

Komunztadura-erroreak detektatzeko, aditz laguntzaileen sailei buruzko informazioa (agerpen proportzioekin) baliatuko dugu. Adibidez, har dezagun berriro *etorri* aditza eta azter dezagun VII.6 taulan jasotako informazioa. Corpusetan *etorri da* moduko egiturak dira gehien erabiltzen direnak (NOR paradigma % 94ean); eta *etorri du* eta *etorri dio* modukoak oso gutxitan. Azkenaren modukoak analisisetan agertzen badira, errore kontsidera ditzakegula uste dugu.

Etorri		
Agerpenak guztira: 5649		
Kopurua-Du:	2	% 0,03
Kopurua-Da:	5331	% 94,37
Kopurua-Dio:	0	% 0
Kopurua-Zaio:	316	% 5,59

VII.6 Taula: [etorri](#) aditzak corpusean paradigma bakoitzeko aditz laguntzailerkin duen agerpen-proportzioa.

Corpusek aditzen erabilerari buruzko adibide mordoa eskaintzen dute eta aditz-egituren deskribapen zabalagoa eta osatuagoa ematen dute. Jasotzen den informazioa zuzena dela bermatzea, ordea, ez da berehalakoa. Maiz, eta lexikorik gabeko analisia dela eta, aditz ez direnak (adibidez, [konpontu](#)) horrela kontsideratzen dira. Gainera, informazioa guztiz automatikoki jasotzearen ondorioz, erroreak sor daitezke.

- *Euskal Hiztegiaren azterketa automatikoaren ondorioz lortutako patroiak*. Euskal Hiztegiaren (EH) (Sarasola, 1996) XML bertsio bat baliatu dugu hiztegitik automatikoki 4016 aditzi buruzko informazioa erauzteko. Sarrerez gain, aditz bakoitzeko laguntzaile mota adierazten duten etiketak jaso ditugu (DA, DU, DIO, ZAIO eta DA-DU...).

Hiztegiaren egiturari dagokionez, eta helburua aditzen deskribapena izanik, garrantzitsua iruditu zaigu honakoak aipatzea: i) Bi hitzez osatutako formak, bietako lehenean azpisarrera moduan kokatu dituela dio egileak, baina hau ez da beti gertatzen. Horrela, lehenago aipatu dugun [tenk egin](#) aditzak sarrera propioa duen bitartean, [adarra jo](#) aditza, [adarra](#) izenaren sarreran agertzen da. Inkoherentzia hori dela eta, zenbait aditzen informazioa automatikoki gaizki jasotzea eta ondorioz, aberasketa okerra egitea suposatuko luke; ii) Bestaldetik, hiztegiko sarrera bakoitzaren azpian aditz-adierak topatzen ditugu kronologikoki ordenatuta. Adiera bakoitzean, baina ez guztietan, hartzen duen laguntzaile mota adierazten da. Adibidez, [eratu](#) aditzak laguntzaile-markadun bi adiera ditu eta markarik gabeko bat. Aditz-laguntzailea duen lehen adierak, [konponduren](#) parekoa denak, DA laguntzailea hartzen du eta bigarrenak, [moldatu](#), [antolatu](#), [prestatu](#)-ren parekoa denak, berriz, DU. Aditza automatikoki prozesatu denean, DA-DU etiketa ipini dugu, nahiz eta semantikaz desberdinak izan, sintaxi-mailan ez baitute desberdintasun nabaririk.

EHren hitzaurrekoan aditz laguntzaile motak zehazten dira, eta zehaz-

tapenaren arabera, hiztegiko azpikategorizazio-etiketak eta corpusekoak desberdinak dira. Taula bat eraiki dugu (ikus VII.7 taula) balio-kidetasunekin. Lan honetan, bigarren zutabeko etiketak erabili ditugu.

Hiztegia	Corpusa
DA	NOR eta NOR-NORI
DU	NOR-NORK eta NOR-NORI-NORK
ZAIO	NOR-NORI
DIO	NOR-NORI-NORK

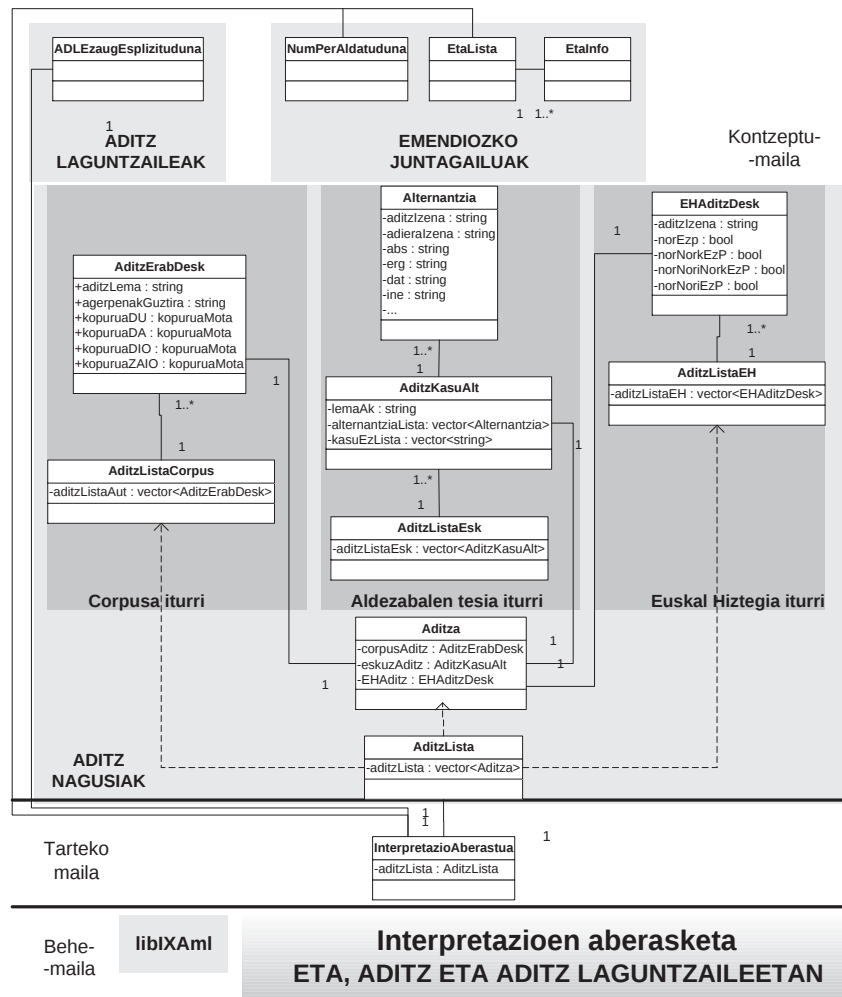
VII.7 Taula: Hiztegiko eta corpuseko etiketen arteko baliokidetasunak.

Aipatu berri ditugun hiru iturrietatik informazioa jaso eta aditzen analisietan gehitu ondoren, [etorri](#) aditzaren analisia VII.5 irudian agertzen den moduan geratuko litzateke.

<pre> <fs id="V-etorri-1" type="AdiInfo"> <f name="maiztasun-ezaugarriak"> <fs type="aditz-maiztasuna"> <f name="agerpenak"> <nbr value="5649"/> </f> <f name="NOR-maizt"><nbr value="94.37"/></f> <f name="NOR-NORK-maizt"><nbr value="0.03"/></f> <f name="NOR-NORI-NORK-maizt"><nbr value="0"/></f> <f name="NOR-NORI-maizt"><nbr value="5.59"/></f> </fs> </f> <f name="KasuEz" org="list"> <sym value="(ERG)"/> <sym value="(ELA_KONP)"/> </f> <f name="EH_hiztegi_def-EZ" org="list"> <sym value="Nor-Nork"/> <sym value="Nor-Nori-Nork"/> </f> </fs> </pre>	Corpusa
	Eskuzko lana
	EH hiztegia

VII.5 Irudia: [etorri](#) aditza aberasketa-prozesuaren ondoren.

Aberasketa-prozesuaren garapenerako diseinatu dugun domeinu-eredua VII.6 irudian dago. Kontzeptualizazio honetan, hiru abstrakzio-maila zehaztu dira. *Behe-mailan* berriro ere *LibiXaML* klase-liburutegiko klaseak kokatu ditugu eta *tarteko mailan* analisi aberastuari dagokiona. *Kontzeptu-mailako* klasetan aberasketa-iturrien arabera bereizketa argia egiten da.

VII.6 Irudia: Mendekotasun-zuhaitzak *aberasteko* domeinu-eredua.

VII.5 Gramatika

Saroi tresnaren kontsulta-lengoaia erabili dugu komunztadura-erroreak detektatuko dituen gramatika idazteko (ikus VI. kapituluaren, VI.2.2.2 atala). Erregelak idazteko abiapuntu moduan hizkuntzalari bati 85 esaldi errore-dunen ustezko mendekotasun-zuhaitz “egokia” marrazteko eskatu diogu. Zuhaitz horiek oinarri izanik idatzi ditugu errorearen detekzioarako erregelak, ez baitakigu automatikoki sortutako zuhaitzak nolakoak izango ote diren. Komunztadura-erroreak detektatzeko hurbilpen eta baliabide desberdinak erabili ditugu, eta horiek oinarri hartuta, erregelak multzoetan bildu ditugu.

Onorengo lerroetan, erregela multzoak azalduko ditugu labur-labur. Lantzen duten fenomenoak adibide batekin lagunduko dugu:

1. **Aditzen patroiak.** Aditzei aberasketa-prozesuan erantsitako informazioa oso baliagarria izan da errorearen detekzioarako. Multzo honetako erregelak duten abantaila nagusia zera da: nahikoa da mendekotasun-erlazio bakarra ongi ipintzea errorea detektatzeko. Aditz nagusia eta bere laguntzailea lotzen dituen mendekotasun-erlazioa ongi ipiniz gero, edozein zuhaitz kopuru sortuta ere, errorea detektatu egiten da. Besteen artean, honako adibideko esaldian detektatu dugu errorea multzo honetako erregelak erabilita:

(54) *Eta orduan nire amak esnatu zidan eta konturatu nintzen lotan nengoela.

Multzo honetako erregelen emankortasunak badu bere alde beltza. Aditzen patroietan hainbat alternantzia eta erabilera ez dira deskribatzen. Hori dela eta, erregela hauek alarma faltsu ugariren iturri izan dira. Adibidez, (55) adibideko erabilera ez dago Euskal Hiztegian bilduta¹⁵, eta errorea detektatu da esaldi zuzenean.

(55) Bazkaldu ondoren bilerek jarraituko dute.

Erregelak definitzeko orduan zailtasunak izan ditugu detektatutako *erroreen* eta *alarma faltsuen* artean neurri egokia lortzeko. Azal dezagun arazoa:

¹⁵Euskal Hiztegian **jarraitu** aditzari ZAIO saila egokitzen zaio soilik.

- Hiru iturritatik jaso dugu aditzek erabiltzen duten laguntzailearen paradigmari buruzko informazioa. *Euskal Hiztegitik* 4016 aditzi buruzkoa, corpusetatik 2541 aditzena eta eskuz egindako azterketatik 100 aditzi buruzkoa. Demagun, **esnatu zidan** egitura detektatzeko, hiru iturrietako informazioa bildu nahi dugula erregela berean. Kasu honetan, VII.7 irudiko moduko erregela definituko genuke. Erregela honetan uneko adabegia (@) aditza kategoriakoa bada (@.kat = "ADI") baina ez aditz-izena (@.adm != "ADIZE"), eta aditz laguntzailea NOR-NORI-NORK paradigmakoa badu (@!auxmod.Paradigma == "NOR-NORI-NORK"), errorea honako kasuan gertatuko dela adierazten da: 1) Hiztegian, aditz horri dagokion azpikategorizazio-etiketen artean ez bada agertzen NOR-NORI-NORK azpikategoria (@.NorNoriNorkEz~); 2) Corpusean aditz hori paradigma horrekin % 2ko maiztasuna baino txikiagoarekin agertzen da (@.NOR-NORI-NORK-maizt < 2); eta 3) Eskuzko azterketan adierazten da aditza ez dela datiboa kasuarekin agertzen (@.KasuEz == "DAT").

```

Erreg_KOM_AditzAzpikta_Nor_Nori_Nork_1
(
    Detekt (
        @.kat == "ADI" &
        @.adm != "ADIZE" &
        @!auxmod~ &
        @!auxmod.Paradigma == "NOR-NORI-NORK" &
        1. @.NorNoriNorkEz~ &
        2. @.NOR-NORI-NORK-maizt < 2 &
        3. @.KasuEz == "DAT"
    )
)

```

VII.7 Irudia: Komunztadura-erroreak detektatzeko “aditzen patroiak” erabiltzen dituen erregelaren adibidea.

Erregela hau oso *segurua* litzateke (3 baldintzak bete behar dira), eta beraz, alarma faltsuak ekiditeko oso egokia. Baina, bestalde, aztertutako aditz multzoa 100 aditzera murriztuko luke.

- Alarma faltsuen kopurua asko handitu gabe landutako aditz kopurua handitzeko asmoarekin, erregeletan bi baldintza bete daitezkeen eskatu dugu (hiztegiari eta corpusari dagozkien baldintzak). Baldintza horien pean egindako proban ez genuen, adibidez, (54) adibideko ***esnatu zidan** egitura erroreduna detektatu. Hiztegian, egitura hau zuzentzat jotzen da, kasu batzuetan zuzena baita (**ize-**

ba gelara sartu zen eta umea esnatu zidan esaldian¹⁶, adibidez). Erabilera proportzioei begiratzen badiogu, ordea, kasu gehiagotan izango da egitura okerra, zuzena baino. Hau da, segur aski erroreduna izan daitekeen egitura baten detekzioa galduko genuke, egitura zuzen gutxi batzuk ez detektatzearen truke. Kasu antzekoa gertatzen zaigu **ikusi zaio** moduko egiturekin.

- Erabaki bat hartu behar izan dugu, eta *errore detekzioaren* alde egin dugu: alarma faltsu gehiago edukiko ditugula jakin arren, ohikoak diren egitura horiek detektatu nahi izan ditugu. Horrela, informazio-iturri bakoitzeko, baldintza bakarreko erregela laukote bat (erregela bat paradigmako) idatzi dugu.

2. *Objektuaren edo zehar-objektuaren sorkuntza*. Komunztadura-erroreen sailkapena egin dugunean, *sorkuntza* kasuak aipatu ditugu. Kasu haueetan aditz laguntzaileko komunztadura-marketan agertzen ez den kasuren bat agertzen da esaldian. Hau da (56) adibidearen kasua. Aditz laguntzaileak (**du**) ez du *zehar-objekturik* eskatzen, baina esaldian *zehar-objektu* bat markatzen du EDGK-k (*erabiltzaileari*). Errore zehatz hau detektatzeko, VII.8 erregela erabiliko genuke.

- (56) **Erabiltzaileari sare batean egon daitekeen urruneko beste ordenagailuekin konektatzeko aukera ematen du.*

```
Erreg_KOMSOR_ZOBJ_1
(
    Detekt (
        @.kat == "ADI" &
        @!nczobj~ &
        @!auxmod~ &
        (@!auxmod.Paradigma == "NOR-NORK" |
        @!auxmod.Paradigma == "NOR")
    )
)
```

VII.8 Irudia: Komunztadura-erroreak detektatzeko “zehar-objektuaren sorkuntza” kontuan hartzen duen erregelaren adibidea.

3. *Komunztadura perpausaren barruan*. Multzo honetan biltzen da erregela kopuru handiena (34 erregela). Aztertutako esaldi erroredun gehie-

¹⁶Adibide hauetan, erabilera zuzena ote den jakiteko esaldian *objekturik* ote dagoen begiratu behar dela pentsa dezakegu. *Objektua* elidituta egon daiteke baina.

netan mota honetako egiturak topatu ditugu. Erregelok aditz laguntzaileko komunztadura-marketan agertzen diren ezaugarrien, eta dau-denean, esaldiko subjektuaren, objektuaren eta zehar-objektuaren arteko desadostasunak egiaztatzen dituzte. Adostasun faltak, *kasuaren* eta *numeroaren* ingurukoak izaten dira (maiz kasuaren eta numeroaren desadostasuna egiaztatzeko erregelak batera aplikatzen direla esan dugu). Multzo honetako erregela batekin detektatzen da dagoeneko aski ezaguna dugun esaldiko komunztadura-errorea (ikus (57) adibidea).

(57) *Zentral nuklearrak zakar erradiaktiboa eratzen dute.

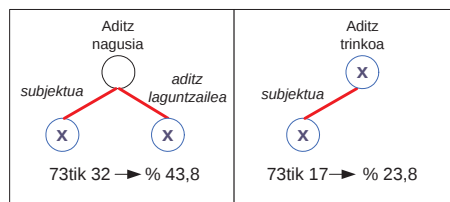
VII.9 irudiko erregelak, *subjektuaren* eta NOR-NORK (eta NOR-NORI-NORK) motako aditz laguntzaileen arteko adostasuna aztertzen du *kasuari* dagokionez. Erregela oso emankorra da eta 13 esalditan topatzen du errorea.

```
Erreg_KOMNAH_SUB_KAS_1
(
  Detekt (
    @.kat == "ADI" &
    @!ncsubj~ &
    @!ncsubj.kas~ &
    (@!ncsubj.kas == "ABS" |
    @!ncsubj.kas == "ERG") &
    @!auxmod~ &
    (@!auxmod.Paradigma == "NOR-NORK" |
    @!auxmod.Paradigma == "NOR-NORI-NORK") &
    @!ncsubj.kas != @!auxmod.nork.kas
  )
)
```

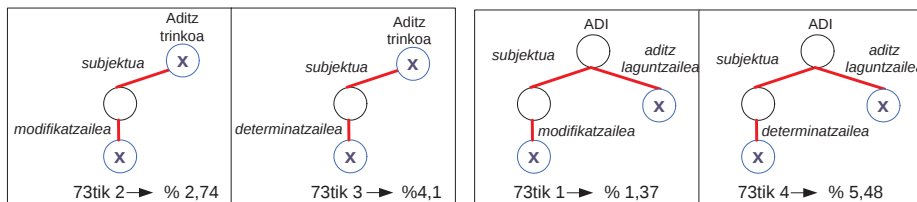
VII.9 Irudia: Komunztadura-erroreak detektatzeko “perpausaren barruko komunztadura” aztertzen duen erregelaren adibidea.

Multzo honetako erregelak definitzeko egin dugun azterketa baliatuz, euskarazko esaldietan komunztadura egiten duten esaldiko elementuak mendekotasun-zuhaitzetan zein modutan kokatzen diren ezagutu ahal izan dugu. Subjektuaren kasuan, ohikoenak diren egiturak VII.10 irudikoak dira¹⁷. Maiztasunean hauen atzetik, egitura hauen aldaerak datoz (ikus VII.11 irudia) zeintzuetan VII.10 irudiko egiturei determinatzaile edo modifikatzaile baten agerpena gehitzen zaien.

¹⁷Multzo honetan 73 esaldi kokatu dira. Irudietan egitura bakoitzaren azpian, egitura hori duen esaldi kopurua agertzen da.



VII.10 Irudia: Subjektuari dagokionez, komunztatzen duten elementuen kokapen ohikoena zuhaitzetan (kokapena X ikurra erabiltuta adierazi dugu).



VII.11 Irudia: Komunztatzen duten elementuen kokapen ohikoenaren aldaerak (kokapena X ikurra erabiltuta adierazi dugu).

4. **Komunztadura sintagman.** Sintagma barruko komunztaduran, bertako elementuen arteko *mugatasuna* eta *numeroa* izaten dira egiaztatu beharreko ezaugarriak (ikus (58) adibidea).

(58) *Liburuak eta [errebistak espezializatuak] gai zehatz baten inguruan idatziak daude.

5. **Partitiboa.** Pare bat erregela kokatu ditugu multzo honetan, eta hauek *Euskaltzaindiak* emandako 15. arauan (*Partitiboaren erabilera okerra*) dute funtsa. Arau honek zera dio hitzez hitz: "Erabil bedi partitiboa euskaraz, beti erabili izan den gisa, hau da, artikulurik gabe (*lagunik*, *lanik*, *ahuntzik*, e.a.) eta pluralik gabeko komunztadura eginez aditzarekin (*dugu*, *dute*, *genuen*, e.a.)". Informazio honetan oinarrituta detektatu dugu (59) esaldiko errorea.

(59) *Normalean ez dituzte ordenagailuetan destrozurik egiten.

6. **Heuristikoak.** Pertsona-komunztadurarekin dute zerikusia heuristikoen multzoan definitutako erregelek. Berezi samarra da (60) adibideko esaldian dagoen errorea markatzeko idatzi behar izan dugun erregela. Adibideko *ni* izen ordaina objektu bailitzan etiketatu da (ez du ergatibo

kasurik). Aditz trinkoaren arabera (**dakit**) objektuak 3. pertsona singularrarekin egin beharko luke komunztadura, beraz, pertsonak desberdinak izanik, komunztadura-erregela batek errorea markatu egin du. Gure hurbilpenaren kontra badoa ere, heuristiko honetan analisi linguistikoan gertatutako akatsa kontuan hartu behar izan dugu.

(60) *Ni ez dakit.

7. **Subjektu anitz, objektu anitz, zehar-objektu anitz.** Bukatzeko, alarma faltsu ugariren errudun nagusia den erregela multzoari buruz arituko gara. Gramatikalak diren perpausetan gehienez objektu, subjektu eta zehar-objektu bakarra agertu ohi da¹⁸. Ezaugarri horrek erroreak detektatzeko arrasto egokia ematen digu, funtzio horiek dituzten sintagma gehiago egonez gero, errorea genukeelako. Analisi automatikoa egiten denean, subjektu/objektu anbiguitasuna ebaztea oso zaila denez, oso ohikoa da funtzio horiek errepikatuta agertzea. Hori dela eta, esaldi erroredunetan zein zuzenetan, erregela hauek maiz aplikatzen dira. Horrela, esaldi erroredunak batzutan arrazoi ez-egokiagatik markatzen dira errore moduan, eta esaldi zuzenak erroredun jotzen dira alarma faltsuak sortuz (ikus emaitzak VII.7.2 atalean). Adibide moduan har dezagun (61) esaldi zuzena. Bertan EDGK-k hiru subjektu ezagutu ditu, “**Petrolio konpainiek**”, “**mozkin astronomikoak**” eta “**galera ekonomikoa**”; beraz, errorea markatu dugu.

(61) **Petrolio konpainiek mozkin astronomikoak eta kontsumitzailearentzat galera ekonomikoa ekarri dutela adierazi zuen.**

8. **Zenbait aditzen erregimena (gonbidatu, iguriki, itxaron...).** *Euskaltzaindiak* emandako izen bereko arauan (39.enean) du oinarri erregela multzo honek. Arauan, adibidez, **itxaron** aditzak NOR-NORI paradigma eskatzen duela objektua duenean, eta NOR-NORK objekturik gabe dagoenean, zehazten da. Erregelak definitzeko modu honetako baieztapenak erabili ditugun arren, erregela multzo hau ez da oso baliagarria izan. Ez dira erabilera inpertsonalak zehazten eta jasotako adibide multzoan ez dugu multzo honetan koka daitekeen adibiderik topatu. (62) adibideko egitura multzo honetan kokatuko litzateke.

(62) *Mikeli gonbidatu diote.

¹⁸Zenbait kasutan, adibidez erreferente bera dutenean, bi ager daitezke. Kasu hau da, **ni neu naiz hautagaia** esaldikoa.

VII.6 Aplikazio-estrategia eta errore-erabakia

Komunztadura-erroreak detektatzeko garatu dugun gramatikako erregelak analisi-zuhaitzetan aplikatzeko modu asko egon daitezke. Gure aukera honakoa izan da: hasiera batean, esaldi baten analisiaren ondorioz sortutako analisi guztietan, horietako azpizuhaitz guztietan eta adabegi guztietan aplikatuko dira erregela guztiak. Hori teorian horrela izan arren, praktikara eramaten dugunean, azterketa-lana *arindu* egiten da: zenbait baldintza betetzen dituzten adabegiak soilik aztertuko dira. Adibidez, erregeletako lehen baldintzan adabegiko tokenaren kategoria gramatikala zehazten da maiz. Kategoria hori ez duten adabegiak ez dira aztertzen.

EDGK erabiliz esaldi bat analizatzen dugunean, makina bat aldiz analisi-zuhaitz bat baino gehiago sortzen da. Analisi-zuhaitz horiek, gainera, ez dira osoak izaten, partzialak baizik. Esaldi bat aztertzerakoan hainbat zuhaitz sortzen badira, eta horiek zenbait azpizuhaitzetan banatzen badira, komunztadura-erroreen detekzioarako erregelak horietako zenbatetan aplikatu behar dira esaldia erroreduna dela erabakitzeko? Analisi-zuhaitzak sarri ez direla egokiak izaten ikusita, aukera kontserbadorea egin dugu: esaldi bati dagozkion analisi-zuhaitz **guztietan** aplikatua behar da erregela, esaldi horretan errorea dagoela esateko. Erabakia oso murriztailea dela ulertzen dugu, baina alarma faltsuak ekiditeko modua da. Egitura bat erroreduna dela esateko, oso ziur egon behar dugu hortaz. Hala ere, hurrengo atalean ikusiko dugun moduan, batzuetan alarma faltsu batzuk saihestezinak dira.

VII.7 Komunztadura-erroreen eta anbiguotasunaren arteko erlazioa aztertzeko metodologia. Emaizak

Errore-detekzioaren eta anbiguotasunaren artean lotura hertsia dago. Baiez-tapen hau testuinguru mugatuko erroreak detektatzeko egin genituen esperimentuek aski ongi egiaztatu zuten. Kasu hartan, anbiguotasun morfosintaktikoa izan genuen aztergai. Esku artean dugun egituraren anbiguotasun morfosintaktikoaz gain, sintaxi-mailako anbiguotasuna ere topatuko dugu. Hori dela kasua, berriro ere, *errore-anbiguotasun* erlazioa egiaztatzeko probak egin nahi baditugu, probak kopuruan izugarri handituko dira. Anbiguotasun morfosintaktikoko mailen (EUSTAGGER) eta anbiguotasun sintaktikoko mailen arteko konbinazio posible guztiekin egin beharko lirateke probak. Kasu honetan, ez zaigu bideragarria iruditu aurrekoan egin genuen moduan, komunztadura-erroreen inguruan egin beharreko proba guztiak corpus guztiekin egitea. Programak prestatzeak, esperimentuak aurrera eramateak eta

emaitzak aztertzeak, corpus handiak erabili nahi badira behintzat, lan handiegia emango liguke. Hau horrela izanik, esperimntua alderantziz egitea otu zaigu. Hainbat urratsek osatzen duten metodologia bat definitu dugu erreoren detekziorako egokiena den anbiguotasun-maila aukeratzeko. Anbiguotasun-maila hori aukeratuta, sistema ebaluatzeako proba nagusia harekin egin dugu. Horrek ez du esan nahi aurreko urratsetan ebaluaziorik egin ez denik, urratsetako bakoitzean egin diren proba guztiak ebaluatu dira, eta haien emaitzak erakutsi ditugu.

Urratsak azaltzen hasi aurretik, aztertuko ditugun anbiguotasun-mailak zeintzuk diren zehaztuko dugu:

- **Anbiguotasun morfosintaktikoa.** EUSTAGGERek hitz-mailako desanbiguazioa egiteko bi hurbilpen konbinatzen dituela esana dugu: gramatika bidezko desanbiguazioa Murritzapen Gramatika (MG) baliatuz, eta teknika estatistikoetan oinarritutako desanbiguazioa. Tresnari dei egiterakoan “maila” zehaztuz desanbiguazioa egiteko modua aukera dezakegu EUSTAGGERen. “Maila” bakoitzean erabiltzen diren metodo eta ezaugarriak VII.8 taulan erakutsi ditugu.

Maila	Metodoa	Desanbiguaziorako hizkuntza-ezaugarria
1.maila	MG + estokastikoa	kategoria
2.maila	MG + estokastikoa	kategoria + azpikategoria
3.maila	MG + estokastikoa	kategoria + azpikategoria + kasua
4.maila	MG	Analizatzaile morfologikoak ematen dituen guztiak

VII.8 Taula: EUSTAGGERen desanbiguaziorako aukerak.

- **Anbiguotasun sintaktikoa.** Ezaugarri morfosintaktikoen anbiguotasunaz gainera, sarrerako testu analizatueta *funtzio sintaktikoen* (subjektu, objektu, izenlagun, aditzlagun...) anbiguotasuna geratu ohi da hitz bakoitzeko interpretazioetako bakoitzean. Funtzio sintaktikoen anbiguotasuna ebazteko MG gramatika bat garatuta dago IXA taldean. Gramatika horretan erregelak hiru multzotan bildu dira lantzen duten funtzio sintaktikoaren arabera. Multzoetako bakoitzean honako funtzio sintaktikoak lantzen dira:
 - 1. multzoa: Sintagma barruko modifikatzaileei dagozkienak. Adibidez, eskuineko izenlaguna (@<IZLG), ezkerreko graduatzailea (@GRAD>)...
 - 2. multzoa: Aditzei dagozkienak. Adibidez, aditz konposatua (@ADILOK), aditz nagusi ez-jokatu mendekoa (@JADNAG_MP)...

- 3. multzoa: Funtzio sintaktiko nagusiak. Adibidez, subjektua (@SUBJ), objektua (@OBJ)...

Lan honetan sintagma gramatikalei dagozkien funtzioen eta bestelakoen arteko bereizketa egitea interesatzen zaigu. Horretarako, gramatikako erregela multzoak desanbiguazio sintaktikorako bi “mailetan” banatu ditugu:

- 2. mailan: 1. eta 2. multzoko funtzio sintaktikoak desanbiguatuko dira.
- 3. mailan: 3. multzoko funtzio sintaktikoak desanbiguatuko dira.

Morfosintaxi-mailako anbiguotasunak zein sintaxi-mailako anbiguotasunak erroreak detektatzeko izan dezakeen eragina aztertzearen, “desanbiguazio-konbinazioak” egiten dituzten programak prestatu ditugu. Horrela, proba **1-2** desanbiguazio-mailarekin egin dugula esaten dugunean, desanbiguazio morfosintaktikoan **1. maila** erabili dugula, eta desanbiguazio sintaktikoa **2. mailan** egin dugula adierazten dugu. Azalpen hauek eman ondoren, egindako probak urratsez urrats azalduko ditugu VII.7.1, VII.7.2 eta VII.7.3 ataletan, hurrenez hurren.

VII.7.1 Desanbiguaziorako aukera egokienaren bila

Aipatutako konbinazio guztiak (8 konbinazio posible) aplikatu dizkiegu 10 esaldi erroreduni eta hauetako bakoitzari dagokion zuzenketa bati, haien arteko erlazioa aztertzeako asmoz. VII.9 taulan proba honen berri eman dugu jarraian azalduko dugun informazioa erakutsiz.

Esaldietako bakoitza desanbiguazio-mailetako bakoitzarekin aztertu dugunean, analisisian zenbat analisi-zuhaitz sortu den jaso dugu (**Z** zutabearen azpian agertzen da datua) eta zuhaitzetatik zenbatetan lortu den errorea detektatzea (**E** zutabearen). Gogoratu, esaldi baten analisi-zuhaitz guztietan aplikatu behar dela erroreak detektatzeko gutxienez erregelaren bat, esaldia erroreduna dela erabakitzeke. Gainera, gehienetan, analisi-zuhaitzak partzialak izango dira. Esaldi bakoitzeko lortu den zuhaitz kopurua analizatzeaz gain, datu honen azpian batez beste zuhaitzetako bakoitza zenbat azpizuhaitzetan sakabanatu den adierazi dugu letra etzana erabiliz. Esaldia-aren mendekotasun-zuhaitz guztietan komunztadura detektatzeko erregelak emaitza positiboa eman duenean, errorea detektatu denean, zenbakiak hizki lodiz idatzi ditugu.

Datu hauek guztiak analizatuta, desanbiguazio-konbinazio “egokiena” topatu nahi izan dugu. Konbinazio “egokiak” honako baldintzak bete behar ditu:

1. Ahalik eta esaldi erroredun gehienetan detektatu behar du errorea.
2. Ahalik eta esaldi zuzen gutxienetan eman behar du alarma faltsua.
3. Ahalik eta zuhaitz gutxien sortu behar du, exekuzioa azkarragoa izan dadin.

Komunztadura-erroreen detekzioarako erregelak zuhaitzetan aplikatu ondoren lortutako emaitzak xehatuta erakutsi ditugu VII.9 taulan, eta laburtuta, VII.10 taulan ($\sqrt{\quad}$ ikurra analisi horretan errorea detektatu dela adierazten du). Emaitzetatik ondorio hauek atera ditugu:

1. **1-2** eta **2-2** konbinazioek beti ematen dute haien artean berdina den emaitza. Gauza bera gertatzen da **1-3** eta **2-3** konbinazioekin. Konbinazio horietan sortzen den analisi-zuhaitz kopurua desanbiguazio morfosintaktikoa 3. mailan egiten dutenekin konparatuz (**3-2** eta **3-3**) handia izaten da. Erroreen detekzioari dagokionez, kasu guztietan 5 erre detektatzen dira eta ez da alarma faltsurik ematen. Desanbiguazio morfosintaktikorako lehen eta bigarren mailak erabiltzen dituzten konbinazioak (hots, **1-2**, **2-2**, **1-3** eta **2-3**) sortzen duten zuhaitz kopuru altuagatik baztertu ditugu.
2. **4-2** eta **4-3** konbinazioekin ere orokorrean zuhaitz ugari sortzen da (emaitzak aldakorrek dira, 9. eta 10. esaldiekin zuhaitz gutxiago ematen baitute). Bi kasuetan 6 esalditan detektatzen da errorea, eta **4-2** konbinazioan, alarma faltsu bat detektatu da. **4-2** eta **4-3** aukerak ere, baztertu egin ditugu sortzen duten zuhaitz kopuru altuagatik.
3. Guztien artean hasieran jarritako baldintzetara hobekien egokitzen diren aukerak **3-2** eta **3-3** dira. Analisi-zuhaitz gutxi sortzen dute eta 10 esalditatik 6 esalditan errorea topatzen da errorea (alarma faltsu bakarra 3. esaldian).

Xehetasunak dituen taularen (VII.9 taula) analisi bat eginda, albo-ondorio batzuk atera ditzakegu. Labur-zurrean emango ditugu:

1. **Zuhaitz-egitura minimoa egotea nahikoa da errorea detektatzeko.** Esaldiaren analisi automatikoaren ondorioz sortutako mendekotasun-zuhaitzak patroian deskribatutako egitura minimoa badu, errorea detektatu

egiten da (ikus 1, 2, 3, 9 eta 10 esaldiak). Zuhaitz kopurua, eta hauetan azpizuhaitz kopurua, handia izan daiteke baina errorea detektatzeko gai izango da sistema (ikus 3 esaldia).

Erroreen detekziorako behar dugun mendekotasun-erlazioa ezartzen ez bada, ordea, *Saroi* ez da gai izango errorea detektatzeko (ikus 2, 4, 5 eta 6 esaldiak). Esandakoaren adierazgarri da 6 eta 7 esaldien kasua. Azken esaldian zehar-objektua markatzeko gai izan da EDGK, eta errorea detektatu dugu. Beste kasuan, 6 esaldian, datibo kasu-markaren faltagatik (*elipsia*) ez da zehar-objektua markatu, eta beraz, ezin izan da errorerik adierazi.

2. [Analisi-zuhaitz askorekin, detekzioa bideraezina](#). Batzuetan, analisi-zuhaitz gehiegi sortzen dira (taulan Z++ etiketarekin adierazi dugu fenomeno) eta errore detekzioa bideraezina bihurtzen da (ikus 8. esaldia eta 3. esaldiko kasu batzuk).
3. [Aditz-ereduen erabilerak, emaitza onak](#). Aberasketa-prozesuan jaso dugun aditz-azpikategorizazio informazioan aditzaren alternantzia eta adiera guztiak kontuan hartu badira, modu errazean detekta daitezke erroreak (ikus 9 esaldia). Analisi-zuhaitz kopuruak kasu honetan ez du axola handirik, gehiegizkoa ez bada, noski.
4. [Emendiozko juntaduraren tratamendua ezinbestekoa](#). Errore batzuk detektatzeko, ezinbestekoa da emendiozko juntaduraren tratamendua (ikus 10 esaldia).

Lehen urrats honetan, desanbiguaziorako konbinazio egokienak **3-2** eta **3-3** direla ondorioztatu dugu. Hots, desanbiguazio morfosintaktikorako MG eta teknika estokastikoak konbinatzen dituen aukera, eta desanbiguaziorako informazio gehien (kategoria + azpikategoria + kasua) erabiltzen duen aukera (zenbaki bikoteko lehen **3** zenbakiak adierazten du hau). Oraindik ez dugu argi ikusi funtzio sintaktikoak desanbiguatzeiko aukerarik egokiena zein den (ikus bigarren zenbakia, **2** edo **3**). Zalantza hau argitzeko emango dugu hurrengo urratsa.

	Anbiguotasun-maila															
	1-2		2-2		3-2		4-2		1-3		2-3		3-3		4-3	
	Z	E	Z	E	Z	E	Z	E	Z	E	Z	E	Z	E	Z	E
1. *Helbideak itxura ezberdinak dituzte baina normalena ondokoa da. ZUZENA: Helbideek...dituzte.	9	9	9	9	2	2	45	45	4	4	4	4	1	1	20	20
	1 3 1	0	1 3 1	0	1 2 1	0	1 15 1	15	1 2 1	0	1 2 1	0	1 1 1	0	2 10 2	0
2. *Egunero jende asko etortzen dira. ZUZENA : ...etortzen da.	8 1,75 8 1,75	2 0	8 1,75 8 1,75	2 0	2 1 2 1	2 0	2 1 2 1	2	4 1,75 3 1,66	1 0	4 1,75 3 1,66	1 0	1 1 1 1	1 0	1 1 1 1	1 0
3.*Gas naturala, petroleoa baino zailtasun gutsiago ditu ateratzeko eta baita gordetzeko ere. ZUZENA: *Gas naturalak, petroleoak...	256	256	256	256	16	16	64	64	18	18	18	18	2	2	0	0
	2,75 Z++	Z++	2,75 Z++	Z++	2 16	16	2 144 1,75	108	2,66 162 2,3	98	2,66 162 2,3	98	2 2	2	0 8 1,5	6
4. *Ze itxura du web orri bat? ZUZENA: ...batek?	48 1,6 24 1,6	28 14	48 1,6 24 1,6	28 14	12 2 6 2	4 2	36 2,2 18 2,2	25 12	12 1,66 12 1,66	2 2	12 1,66 12 1,66	2 2	4 2 4 2	0 0	4 2,2 8 2,2	2 2
5. *Normalean ez dituzte ordenagailuetan destrozurik egiten. ZUZENA: Normalean ez dute...	9 1 9 1	3 0	9 1 9 1	3 0	1 1 1 1	0 0	18 1 18 1	6 0	9 1 9 1	3 0	9 1 9 1	3 0	1 1 1 1	0 0	18 1 18 1	6 0
6. *Ezaugarri hau onartzen duten datu-base natiboak, eskemarekiko independenteak deitzen zaie. ZUZENA: ...datu-base natiboei...	96 1,75 32 2,75	0	96 1,75 32 2,75	0	96 1,6 48 2,66	0	192 1 64 2,75	0	12 1,66 12 2,66	0	12 1,66 12 2,66	0	16 1 16 2,5	0	16 1 16 2,5	0
7. *Interneten ibiltzen diren pertsonen modu honetan deitzen zaie. ZUZENA: pertsonen...deitzen zaie.	16 1,75 16 1,75	16	16 1,75 16 1,75	16	2 1 2 1	2	2 1 2 1	2	12 1,66 12 1,66	12	12 1,66 12 1,66	12	2 1 2 1	2	2 1 2 1	2
8. *MOCAP teknikak, tradizionalak, 3D modelatze arloan aplikazio profesionalentzat zuzendua egon da,... (25 hitz) ZUZENA: ...teknika...egon da...	Z++	Z++	Z++	Z++	48	0	Z++	Z++	Z++	Z++	Z++	Z++	48	0	Z++	Z++
9. *Eta orduan nire amak esnatu zidan eta konturatu nintzen lotan nengoela eta ez zela ezer gertatu. ZUZENA: ...esnatu ninduen...	96 1,25 72 1,25	96	96 1,25 72 1,25	96	27 1,33 27 1,33	27	8 2,25 4 2,25	8	96 1,25 72 1,25	96	96 1,25 72 1,25	96	9 2 9 2	9	8 2,25 4 2,25	8
10.*Astoa eta behia zeramatela. ZUZENA: ...zeramatatela.	72 1 72 1	72	72 1 72 1	72	72 1 72 1	72	54 1 54 1	54	32 1 32 1	32	32 1 32 1	32	32 1 32 1	32	24 1 24 1	24

VII.9 Taula: Komunztadura-erroreak eta anbiguotasuna.

Anbiguotasun maila								
	1-2	2-2	3-2	4-2	1-3	2-3	3-3	4-3
1.	✓	✓	✓	✓	✓	✓	✓	✓
1. ZUZENA	-	-	-	✓	-	-	-	-
2.	-	-	✓	✓	-	-	✓	✓
2. ZUZENA	-	-	-	-	-	-	-	-
3.	✓	✓	✓	✓	✓	✓	✓	✓
3. ZUZENA	Z++	Z++	✓	-	-	-	✓	-
4.	-	-	-	-	-	-	-	-
4. ZUZENA	-	-	-	-	-	-	-	-
5.	-	-	-	-	-	-	-	-
5. ZUZENA	-	-	-	-	-	-	-	-
6.	-	-	-	-	-	-	-	-
6. ZUZENA	-	-	-	-	-	-	-	-
7.	✓	✓	✓	✓	✓	✓	✓	✓
7. ZUZENA	-	-	-	-	-	-	-	-
8.	Z++	Z++	-	Z++	Z++	Z++	-	Z++
8. ZUZENA	Z++	Z++	-	Z++	Z++	Z++	-	Z++
9.	✓	✓	✓	✓	✓	✓	✓	✓
9. ZUZENA	-	-	-	-	-	-	-	-
10.	✓	✓	✓	✓	✓	✓	✓	✓
10. ZUZENA	-	-	-	-	-	-	-	-
Erroreak erroredunetan	5	5	6	6	5	5	6	6
AF zuzenetan	0	0	1	1	0	0	1	0

VII.10 Taula: Laburpen-taula. Komunztadura-erroreak eta anbiguotasuna.

VII.7.2 Sintaxiko desanbiguazio-maila egokienaren bila, mendekotasun-gramatika erlaxatuz

Bigarren urrats honetan egindako esperimentuaren helburua aipatu berri duguna da: desanbiguazio morfosintaktikoa 3. mailan mantenduaz, sintaxiaren desanbiguaziorako mailarik egokiena zein den erabakitzea. Helburu horrekin, aurreko esaldi multzoarekin disjuntua den esaldi multzo handiago bat erabili dugu: 75 esaldi jaso ditugu eta haiei dagozkien ordain-zuzenen artean bana (75 esaldi zuzen).

Analisirako prozedura aurreko urratsean jarraitutakoaren berdina izan da. Esaldi horiek guztiak **3-2** eta **3-3** desanbiguazio-konbinazioak erabiliz analizatu ditugu. Esaldiak analizatzen gindoazen heinean, fenomeno berri bat analizatzearen beharra sumatu dugu: *mendekotasun-erlazioak esleitzen dituen EDGK gramatikaren erlaxazioa*.

Gramatika erlaxatzeko beharra erabilitako esaldien azterketatik suertatu zaigu. Har dezagun, (63) adibideko esaldia. Esaldian, subjektua **nik** sintagmak osatzen duela esango genuke (ez du objekturik, aditz laguntzaileak ez baitu eskatzen). Dependentziak ezartzen dituen gramatikak, ordea, aditz laguntzailea NOR (edo NOR-NORI) izanik, ergatibo kasudun sintagma bati ez dio subjektu dependentziarik inoiz esleitu¹⁹. Ezingo genuke inoiz mota honetako komunztadura-errorerik detektatu. Arazo hau konpontzeko, EDGK

¹⁹Aditz laguntzaileak NOR-NORK edo NOR-NORI-NORK behar du izan, EDGK-k ergatibo kasudun sintagmari NCSUBJ etiketa ipintzeko.

gramatika *erlaxatu* behar izan dugu. Hau da, **nik** hitzari &NCSUBJ> etiketa ipintzen dion islapen-erregelatik, esleipena aditz laguntzailearen zenbait paradigmatar mugatzen duen baldintza kendu behar dugu. Paradigma edozein izanda ere, dependentzia-etiketa ipin dadin nahi dugu. *Erlaxazio*arekin kontu handiz ibili behar dugu, baldintzen erlaxazioak, etiketa “gehiegiren” esleipena ekar baitezake (subjektu anitz, objektu anitz...), eta alarma faltsuak sor daitezke.

(63) *Nik ez nago konforme.

Urrats honetan proba, beraz, lau konbinazio posibleekin egin dugu: **3-2** eta **3-3** desanbiguaziorako konbinazioak, mendekotasun-gramatika *erlaxatuta* probatu ditugu lehenengo; eta konbinazio berak, EDGK gramatika *erlaxatu gabe* gero.

Esaldi erroredunak

Esaldi erroredunekin (75 esaldi) egindako proben emaitzak VII.11 taulan laburtu ditugu. Hala ere, komenigarria iruditzen zaigu E eranskinen E.1 taulan jasota dauden datu xehatuak aztertzea, bertan, landutako egituren konplexutasuna ikus baitaiteke. Kasu zehatz honetarako (75 esaldi erroredunentzako) konbinazio onena **3-2-Erlaxatua** dela dirudi % 56ko asmatze-tasarekin. Era berean, zuhaitz gehien sortzen duen konbinazioa da (24,08 zuhaitz sortzen ditu batez beste). 49 kasutan (% 65,33) konbinazioaren batek esaldia erroreduna dela agerrarazi du, baina 27 kasutan soilik (% 36) egon dira konbinazio guztiak ados egiaztapen honekin. 26 kasutan ez da inolako konbinaziorikin errorea detektatu. Konbinazioen artean adostasun falta honek gaian ikertzen jarraitzeko beharra erakusten du.

	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
Errorea markatu da zuhaitz guztietan	42	40	36	34
Asmatze-tasa	% 56	% 53,33	% 48	% 45,33
Batez besteko zuhaitz kopurua	24,08	10,32	18,09	9,49
Konbinazio guztietan errorea detektatu da	27 (% 36)			
Konbinazioen batean asmatu da	49 (% 65,33)			
Inolako konbinaziok ez du errorea detektatu	26 (% 34,66)			

VII.11 Taula: Laburpen-taula (75 esaldi erroredun). Komunztadura-erroreak eta anbiguotasuna.

Errorea detektatu ez denean, arrazoia aztertu egin dugu eta E eranskineko E.2 taulan zerrendatu. Zehaztu beharra dago, errorea ez dela detektatu kontsideratuko dugula, inolako konbinazioarekin detektatu ez denean (26 esalditan gertatu da). Jarraian, aipatutako arrazoiak laburtu ditugu:

1. **Mendekotasun-erlazioa gaizki ipini da** (26 esaldietatik 20 esaldietan, % 76,92). Etiketa honen azpian kokatu ditugun esaldietan kasuistika zabala da. Batzuetan, mendekotasun-erlazioa ipini da, baina hitz okerrarekin lotu da. Beste batzuetan subjektu-objektu anbiguotasuna dela eta, ez da lortu zuhaitz *guztietan* errore detekziorako erregela aplikatzea. Zenbait kasutan, ez da errorearen detekziorako behar den mendekotasun-erlazioa ipini edo etiketa okerra ipini da (ikus datu zehatzak, esan bezala, E eranskineko E.2 taulan). Adibidez, (64) adibideko komunztadura-errorea detektatu gabe geratu da, subjektua markatzen duen mendekotasun-etiketarik esleitu ez delako. Analisi-katetik dator akatsa. EDGK ez da **honek**²⁰ hitzari subjektu etiketarik ezartzeko gai izan, aldez aurretik ez delako adierazi hitz horrek sintagma osatzen duela.

(64) *Honek algoritmo baten bidez forma berriak asmatzeko gai da

2. **Aditz modala ez da lotuta analizatu**²¹. (26 esaldietatik 4 esaldietan, % 15,38). Har dezagun (65) adibidea, bertako **nahi du** aditz modaleko osagaietako bakoitza bere aldetik analizatu da. Hots, **nahi** hitza izena baiditza analizatu da. Hori horrela izanik, mendekotasun-erlazio okerrak ezarri dira (objektu etiketa ipini zaio **nahi** hitzari) eta komunztadura egiten duten elementuak oker kokatu dira mendekotasun-zuhaitzean.

(65) *Hirugarrena orriaren lekua edo mota esan nahi du.

3. **Baldintzak ez betetzeagatik, ez da HAUL moduan analizatu** (26 esaldietatik 2 esaldietan, % 7,69). Adibideko esaldian, **balio** hitzari ez zaio “balio izan” analisirik egokitu, NOR paradigmako aditzaz lagunduta agertzen delako.

(66) *Baina ez zen ezertarako balio.

²⁰Gogoratu, **honek** hitza subjektua dela adieraziko dugu soilik, EDGK gramatika erlatu dugulako. Bestela, aditz laguntzaile horri subjektua absolutibo kasuan deklinatuta legokio.

²¹Adibideetako aditzak gaur egun HAUL moduan analizatzen dira; beraz, ez litzateke arazo hau sortuko, ongi desanbiguatzen badira, behintzat.

Kasu hauetan, komunztadura-errorea detektatu ez denean, arrazoiak jasotako analisisian egon dela ondorioztatu dugu. Hots, tesi-lan honen zereginetatik at geratzen diren arrazoiengatik gertatzen dira akatsak. Hutsune nagusia, mendekotasun-erlazioen ezarpenean dagoela ikusi dugu. Mendekotasun-erlazioetan oinarrituta eraikitzen ditugun analisi-zuhaitzak batzuetan “ez-egokiak” direnez, zaila da komunztadura-erroreen detekzioarako sortutako erregelen doitasuna neurtzea. Sarrerako mendekotasun-zuhaitza “egokia” bada (ez du zertan osoa izan), errorea detektatu egiten da. Analisi sintaktikoa hobetzen den heinean, komunztadura-erroreen detekzioarako erregelen emaitzak hobetuko direla uste dugu.

Esaldi erroredunak vs haien zuzenketak

Aipatu berri dugun esperimentuarekin aurrera joan ahala, alde zuzenetik gertatzen zen baita “susmo” baieztatzen joan zaizkigu. Esaldiari dagozkion analisi-zuhaitzetan errorea detektatzea lortzen denean, batzuetan mendekotasun-etiketak gaizki ipini direlako izaten da (adibidez, subjektu bat baino gehiago lotu zaizkiolako aditz berari), eta ez arrazoi egokiagatik. Hots, batzuetan analisiak okerrak direlako detektatzen da errorea (ikus komunztadura-erroreak detektatzeko orduan, hizkuntzaren tratamenduan topa ditzakegun arazoen inguruko hausnarketa VII.2.4 atalean). Fenomeno hori baieztatze-ko asmoarekin, 75 esaldi okerre dagokien ordain zuzen banarekin errepikatu da esperimentu osoa. Azken esperimentu honen inguruko emaitzen laburpen bat eman dugu VII.12 taulan (emaitza xehatuak E eranskinen E.3 taulan topa daitezke).

	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
Errorea markatu da zuhaitz guztietan	23	16	18	15
Batez besteko zuhaitz kopurua	24,78	9,37	17,91	8,47
Konbinazio guztietan errorea detektatu da	9 (% 12)			
Inolako konbinaziotan ez da errorea markatu	47 (% 62,66)			

VII.12 Taula: Laburpen-aula (75 esaldi zuzen). Komunztadura-erroreak eta anbiguotasuna.

Taulako emaitzen arabera, errore gutxien markatzen dituzten, eta beraz, esaldi zuzenekin emaitza onenak ematen dituzten konbinazioak, **3-3-Erlaxatua** eta **3-3-Erlaxatu gabe** dira. 9 esaldietan (esaldien % 12) konbinazio guztiak bat datoz alarma faltsua ematen. Esaldi zuzenean errorea

ez markatzearekin ados daude konbinazio guztiak kasuen % 62,66n (47 esalditan). Gainontzekoetan, konbinazioen batek esaldi zuzenean erroreren bat dagoela adierazi du.

Esaldi okerrekin eta haien zuzenketekin lortutako emaitzak alderatuta (ikus E eranskineko E.4 taula), kasuistika hau topatu dugu:

- Batzuetan ez da errorerik detektatzen esaldi erroredunean, ezta dagoen zuzenketan ere.
- Beste zenbaitetan, kasu ideala ematen da. Esaldi erroredunean errorea detektatzen da, eta zuzenketan ez.
- Azkenik, hainbatetan, analisi automatiko okerraren eraginez, esaldi erroredunean errorea detektatzen da, baina baita zuzenean ere. Azken hauek ez ditugu “ongi detektatutako erroretzat” jo. Esaldi batean errorearen detekzio “erreala” eman dela esango dugu, esaldi erroredunean errorea detektatu bada, eta dagokion ordain zuzenean errorerik markatu ez bada.

Datu zehatzak ematera jota, VII.13 taulan kasuistika hau konbinazio bakoitzean nola gauzatzen den adierazi dugu (gogoratu 75 esaldi erroredun eta 75 zuzen aztertu direla). Erroreren detekzio “erreala” kalkulatzeko, *esaldi erroredunean errorea markatua* (EE) duen esaldi kopuruari, *alarma faltsu kopurua* kendu diogu. Neurri hori kontuan hartuta, kasurik hoberenean 75etik 24 esaldietan (% 32) detektatu da errorea, eta kasurik okerrenean 18 kasuetan (% 24). Esaldi zuzenetan erroreak detektatzeko arrazoiak hauek izan dira (arrazoiak kopururik gabe emango ditugu, baina batez besteko portzentaje batekin, erabilitako konbinazioaren arabera kopurua aldatu egingen baita):

- Mendekotasun-erlazioa gaizki ipini da (% 73 alditan).
- Aditz baten erabilera zuzena aditz laguntzaile zehatz batekin ez da agertzen *Euskal Hiztegian* edo corpusean (% 7).
- Desanbiguazio morfosintaktikoa gaizki egin da (% 7).
- Aurrekoen konbinaketak (% 13).

Zenbakiei soilik begiratuz gero, desanbiguaziorako konbinazio egokiena **3-3-Erlaxatua** dela ondorioztatu dugu. Hots, honako baldintzak biltzen dituen aukera:

	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
Esaldi erroredunear errorea markatuta (EE)	42	40	36	34
Esaldi zuzenean errorea markatuta (alarma faltsua, AF)	23	16	18	15
Erroreen detekzio erreala (EE - AF)	19 (% 25,33)	24 (% 32)	18 (% 24)	19 (% 25,33)

VII.13 Taula: Sintaxi-anbiguotasuna eta erlaxazioa. Emaitzak.

- **Analisi morfosintaktikoaren desanbiguazioa:** desanbiguaziorako Murriztapen Gramatika eta teknika estokastikoak konbinatzen dituenak, tokenen kategoria, azpikategoria eta kasua kontuan hartuaz.
- **Sintaxiaren desanbiguazioa:** funtzio sintaktiko nagusiak (subjektu, objektu, zehar-objektu) desanbiguatzen dituenak. Paradoxa bat egon daiteke kasu honetan, komunztadura-erroreak detektatzeko funtzio sintaktiko nagusiak anbiguo behar ditugula pentsa baitezakegu: errorea dagoenean ez dago argi zein den sintagma bati hobekien egokitzen zaion funtzioa (subjektua edo objektua, adibidez). Hala ere, hauek desanbiguatzea hobe dela adierazten digute zenbakiak. Arrazoiaren bila analisi-zuhaitzen kopurura joan behar dugu. Desanbiguaziorik egin ezean kopurua oso handia da eta zuhaitz guztietan detekziorako erregelak aplikatzea zaila da.
- **Dependentzia-erlazioak:** gramatika *erlaxatua* erabiltzen duena. *Erlaxazioak* hainbat erroren detekzioa ahalbidetzen duela ikusi dugunez, aukera hau hobe izateak zentzua du.

Behin eta berriro esaten ari gara, mendekotasun-erlazioak gaizki ipintzen direnez, erroreak detektatzea zaila egiten zaigula. Baieztapen bat ere egin dugu, analisi sintaktikoa hobetzen den heinean, komunztadura-erroreen detekziorako erregelen emaitzak hobetuko direla esan dugu. Azken puntua baieztatze beharra sortu zaigu, eta horretarako, proba masiboa bi modutan egin dugu:

1. Alde batetik, EDGKz gain, *MaltIxarekin* egin ditugu probak corpus erroredunetan.
2. Bestetik, euskarazko *treebank*ean, ustez analisi-zuhaitz perfektuak dituen zuhaitz-bankuan, aplikatu ditugu erroreen detekziorako erregelak.

Corpus erroredunean, zein neurritan detektatzen diren erroreak aztertuko dugu (alarma faltsuak ahaztu gabe); *treebank*ean berriz, alarma faltsuen kopuruari ipiniko diogu arreta (erroreak ahaztu gabe).

Bi corpus horiek, erroreduna zein arrunta (zuhaitz-bankua), garapenerako eta probarako azpicorpusetan banatu ditugu. Dena dela, gure susmoa egiaztatzen bada, hots, komunztadura-erroreen detekziorako erregelen arrakasta mendekotasun-zuhaitzen “egokitasunaren” arabera dela, probak garapenerako atalean soilik egingo ditugula erabaki dugu. *Saroi*ren erregelak huts zertan egiten duten jakin ezinean, ez ditugu aldatuko. Beraz, ez du zentzurik izango beste proba bat egiteak.

VII.7.3 Aukeraketa eginda, azken probak

VII.7.3.1 Corpus erroreduna

Egiaztatu berri dugun moduan, esaldi batean egon litezkeen komunztadura-erroreak detektatzeko oso garrantzitsua da analisi-zuhaitzak akats gutxi-koak izatea. Horregatik, azken probak analizatzaile sintaktiko desberdinekin egitea erabaki dugu. Corpus erroreduneko esaldiak EDGK eta *MaltIta* analizatzaile sintaktikoekin analizatu ditugu; eta ondoren, esaldi bakoitzeko sortutako mendekotasun-zuhaitzen gainean egitura erroredunak deskribatzen dituzten *Saroi*ren patroierregelak aplikatu ditugu. Atal honetan, analizatzaile sintaktikoen ezaugarriak aztertuko ditugu lehenengo, corpus erroreduna gero, eta bukatzeko, lortutako emaitzak erakutsiko ditugu.

Analizatzaile sintaktikoak

Analisi-katean integratuta dagoen analizatzaile sintaktikoa EDGK da eta honek dituen ezaugarrien artean, tesi-lan honetarako garrantzitsuak diren hiru ezaugarri gogoratuko ditugu:

- Analisi partzialak egiten ditu, hots, ez ditu esaldi bati dagozkion analisi-zuhaitz osoak lortzen, azpizuhaitzak baizik; eta gainera, ez ditu euskarazko sintaxi-egitura guztiak lantzen.
- Analisi-kateak uzten duen anbiguotasun morfosintaktiko eta sintaktiko guztiarekin egiten du lan. Beraz, esaldi batekin analisi-zuhaitz ugari sortzen ditu.
- Egiturak definitzeko erregelak erabiltzen dituenaz, hizkuntza-ezagutzan oinarritutako teknika erabiltzen du, teknika sinbolikoa.

Analizatzaile sintaktiko honi kontrajarriaz, oraindik analisi-katean integratu ez den *MaltIxa* analizatzaile sintaktiko estatistikoa erabili dugu. Honako ezaugarriak ditu:

- Esaldiko hitz guztiei esleitzen die dependentzia-etiketaren bat: analisi-zuhaitz oso bakarra eman ohi du.
- Tresnak sarreratzat morfosintaktikoki guztiz desanbiguatutako sarrera behar du. Ez du funtzio sintaktikorik erabiltzen.
- Corpusetan oinarritutako teknika erabiltzen du, teknika enpirikoa.

Analisi-katean anbigutasun morfosintaktikoa guztiz ebatzi gabe dagoez eta gainera artean *MaltIxa* sortu gabe zegoenez, EDGK erabiltzea erabaki genuen hasiera batean.

Corpusa

Corpus erroreduna orokorra dela esan genuen, hots, ez da euskara ikasleen testuetara mugatzen. Euskaltegiako maila ertainetako eta goi-mailetako ikasleen testuez gain, karrera amaierako proiektuak eta euskara teknikoko unibertsitate-ikasleen testuek osatzen dute. Corpusaren osaera VII.14 taulan ikus dezakegu (guztira 7992 hitz).

Corpus erroreduna		
	Garapenerako corpusa (% 60a)	Probarako corpusa (% 40a)
Hitz kopurua	4995	2997

VII.14 Taula: Corpus erroredunaren banaketa ebaluaziorako.

Ondorengo atalean erakutsiko ditugun emaitzak probarako corpusarekin lortutakoak dira. Probarako erabili dugun testu-bilduma 267 esaldik osatzen dute, eta horietan 4995 hitz biltzen dira. Hau da, batez beste 18,7 hitzeko esaldi errealekin egin dugu lan. Orokorrean esaldiak luzeak dira (luzeenak 65 hitz ditu) eta inola ere ez laborategikoak. Egitura konplexuak eta batzuetan oso traketsak erabiltzen dituzte. Corpus honetan hizkuntzalari batek 45 komunztadura-errore etiketatu ditu, errorea hasiera eta amaiera adierazten dituzten etiketekin mugatuaz (orokorrean hizkuntzalariak errorearen eremutzat perpausa etiketatu du).

Ebaluazio-emaitzak

Corpus erroreduna analizatzaileetako bakoitzarekin analizatu ondoren komunztadura-erroreen inguruan lortutako emaitzak VII.15 taulan laburtu ditugu. Taula honetako zutabeetan, lortutako doitasuna (D) eta estaldura (E) eman ditugu errenkadetan adierazitako kasu guztietarako. Horien arteko oreka aztertzearren, *f-score*²² (F) balioa ere adierazi dugu.

		Sistemak asmatu	AF	Sistemak detektatu	D	E	F
Erregela guztia	M	28	81	109	% 25,68	% 62,22	36,35
	E	17	53	70	% 24,28	% 37,77	29,56
	M & E	11	31	42	% 26,19	% 24,44	25,28
	M E	33	103	136	% 24,26	% 73,33	36,45
KOMBIK_SUBJ kenduta	M	26	59	85	% 30,58	% 57,77	39,99
	E	14	35	49	% 28,57	% 31,11	29,78
	M & E	10	21	31	% 32,25	% 22,22	26,31
	M E	29	73	102	% 28,43	% 64,44	39,45
KOMBIK_OBJ kenduta	M	22	55	77	% 28,57	% 48,88	36,06
	E	12	39	51	% 23,52	% 26,66	24,99
	M & E	8	19	27	% 29,62	% 17,77	22,21
	M E	26	75	101	% 25,74	% 57,77	35,61
KOMBIK_SUBJ eta KOMBIK_OBJ kenduta	M	21	33	54	% 38,88	% 46,66	42,41
	E	10	19	29	% 34,48	% 22,22	27,02
	M & E	8	10	18	% 44,44	% 17,77	25,38
	M E	23	42	65	% 35,38	% 51,11	41,81
Errore kopurua corpusean							45
Esaldi kopurua corpusean							267
Hitz kopurua corpusean							4995
Batez besteko hitzak esaldiko							18,7

VII.15 Taula: Komunztadura-erroreen detekzioa *MaltIxa* eta EDGK erabilita.

Proba hauek egin ditugu:

- Lehenengo, komunztadura-erroreak detektatzeko erregela guztia aplikatu dizkiegu corpuseko esaldien analisi-zuhaitzei. Emaitzak lau modutan eman ditugu:

1. *MaltIxa*rekin (**M**) lortutako emaitzak.
2. EDGKrekin (**E**) lortutako emaitzak.

²²*f-score* balioa formula honekin kalkulatu dugu:

$$f\text{-score} = \frac{2 * (doitasuna * estaldura)}{doitasuna + estaldura}$$

3. *MaltIxa*rekin **eta** EDGKrekin lortutakoak (**M & E**). Kasu honetan esaldian errorea detektatu dela esango dugu, bai *MaltIxa*rekin baita EDGKrekin ere (biekin) lortutako analisi-zuhaitzetan errorea detektatu bada. Sistemak alarma faltsua eman duela esango dugu, bi analizatzaileekin lortutako mendekotasun-zuhaitzetan, egitura erroredunik egon gabe, errorea markatu bada.
4. *MaltIxa*rekin **edo** EDGKrekin lortutakoak (**M | E**). EDGK-k sortutako zuhaitzean edo *MaltIxa*k analizatutakoan, bietakoren batean errorea detektatzea nahikoa da, errorea markatu dela kontsideratzeko (alarma faltsuekin berdin).

Emaitzei dagokienez, argi dago hobekuntzarako tarte handia daukagula. Doitasuna % 24,26 eta % 26,19 tartean mantendu da. Doitasun hoberena, espero izatekoa den moduan, **M & E** aukerarekin lortu da (bi analizatzaileek lortutako zuhaitzetan errorea detektatzen bada, errorea ziurragoa da). **M & E** kasu honetan, ordea, estaldura asko jaisten da (% 24,44). Orokorrean, bai doitasuna baita estaldura ere begiratuta, *MaltIxa* edo **M | E** aukerak direla egokienak dirudi (*MaltIxa*ren kasuan f-score balioa 36,35 da eta **M | E** kasuan, 36,45).

- Susmatzen genuen moduan, bai EDGK, baita *MaltIxa* erabilia ere, komunztadura-errorea detektatzeko gehien aplikatzen diren erregelak KOMBIK_SUBJ eta KOMBIK_OBJ dira. Horrek esan nahi du, esaldi erroredunetan, baina baita zuzenetan ere, oso ohikoa dela mendekotasun-zuhaitzean aditz berari gutxienez bi subjektu edo bi objektu lotzea. Hori analisi sintaktikoaren akatsa izan ohi da eta alarma faltsu ugari-
ren iturri. Gure susmoa da, analisi okerren maiztasuna dela eta, onura baino kalte gehiago egiten dutela erregela hauek. Beraz, hiru proba egin ditugu susmo hau baieztatzeko. Bigarren errenkadan, kontaktak errepikatu ditugu aditz berarekin lotutako bi subjektu detektatzen dituen erregela kenduta (KOMBIK_SUBJ). Hirugarrenean, erro berari lotuta bi objektu daudenean errorea dagoela adierazten duen erregela kendu dugu (KOMBIK_OBJ). Eta azkenekoan, bi erregela horiek kenduta egin ditugu kontaktak.

Azken kasu honetan, hain zuzen ere, lortu ditugu emaitzarik hobere-
nak. **M & E** kasuan % 44,44ko doitasuna lortu dugu, hori bai, atal horretako estaldura okerrenaren truke (% 17,77a, eta f-score balioa 25,38). Bi neurriak kontuan hartuta, berriro ere, *MaltIxa*k emaitza hobeak ematen dituela dirudi, % 38,88ko doitasuna eta % 46,66ko estaldura (f-score balioa 42,41).

Azterketatik atera ditzakegun hainbat albo-ondorio honakoak dira:

- Emaitez erakutsitakoaren arabera, analizatzaileak hobetzen jarraitu beharra daukagu, erroreak detektatzeko sistemak, erroreak detektatu ez ezik, analizatzaile sintaktikoen akats edo ahultasun guztiak azalarazten baititu. Hala ere, esan beharra daukagu, analizatzaileak garapen-fasean daudela.
- Badirudi, bi analizatzaile sintaktikoak konbinatzea aukera ona dela esaldien analisi-zuhaitzen “zuzentasuna” hobetzeko, eta horrela erroreak doitasun altuagoarekin detektatzeko. Etorkizunean planteatutakoei aukera da bi analizatzaileen integrazioa. Adibidez, hitz bateko interpretazio bati mendekotasun-etiketa bat esleia diezaiokegu, soilik eta baldin soilik, bi analizatzaileek horrela adierazten badute. Hala ere, aukera hori bideragarria izango da, analisi-katearen aurreko urratsetako analisi-akatsak konpontzen direnean. Bestela, analisi-akatsak, mendekotasun-erlazio oker bera esleia dadin eragin dezakete, eta gure kasuan, alarma faltsua.

Analisi-akatsak konpondu ahal izateko, lehenengo identifikatu egin behar dira, eta hori, *MaltIx*arekin zein EDGKrekin sortutako mendekotasun-zuhaitzetan sortzen diren alarma faltsuak aztertuaz egin daiteke. Akatsak azalarazteko, alarma faltsuen azterketa **M & E** aukerarekin eta alarma faltsu gehienek iturri diren erregelak kenduta (KOMBIK_SUBJ eta KOMBIK_OBJ erregelak) egiteak ematen du logikoena. Baldintza horiekin, VII.16 taulan erakutsitako esaldi zuzenetan markatu du sistemak errorea. Sistemak emandako 10 alarma faltsuetatik, kasuen erdian (% 50) desanbiguazio morfosintaktikoan akats bat gertatu da. Taulan ikus daitekeen moduan, aditz laguntzailea NOR paradigmakoa izanik, subjektuak *ergatibo-singularra* interpretazioa jaso du, *absolutibo-plurala* interpretazioa mantendu beharrean. Gure ustez, adibide zehatz hauetarako modu errazean konpon litezke desanbiguazio-akatsak, NOR paradigmako aditzek subjektua absolutiboan deklinatuta eskatzen baitute. Gainontzeko kasuak honela labur ditzakegu: 3 kasutan (% 30) aditz-azpikategorizazioan erabilera bat falta da corpusean edo EH hiztegian. Kasu batean postposizio-lokuzioa ez detektatu izanagatik datibo postposizio-atzizkia daraman hitzari, zehar-objektu mendekotasun-etiketa esleitu zaio. Eta azken kasuan, esaldia oso luzea izanik (49 hitz) bi sistemek errore desberdina markatu dute. Ikusten dugun moduan, alarma faltsuak gertatzeko arrazoiak, egiten ditugun proba guztietan berberak dira.

Azterketa komuna egin ondoren, analizatzaile sintaktikoetako bakoitza bere aldetik aztertu dugu sistemako alarma faltsuak zein erregelarekin eta

AFaren arrazoa	Kop	Adibidea
Desanbiguazio morfosintaktikoa (KOMNAH_SUB_KAS)	5	Honetaz gain, emaitzak (NCSUBJ>-ERG-S) ez ziren (NOR_HAIEK) alfabetikoki ateratzen. Urtero... analisiak (NCSUBJ>-ERG-S) egiten dira (NOR_HAIEK). ... baserriak (NCSUBJ>-ERG-S) krisialdi larri batetan daude (NOR_HAIEK) murgildurik. ... ez zeuden (NOR_HAIEK) klase bakarreko morroiak (<NCSUBJ-ERG-S) ... Kanpoan lan egiten duten morroiak (NCSUBJ>-ERG-S) ere bazeuden (NOR_HAIEK).
Aditzen azpikategorizazioa	3	Kimika nahiko gogorra da baina neukan zientzia grina baretu dit . Corpusean <i>baretu</i> ez da NOR-NORI-NORK laguntzaileekin agertzen. Hainbat faktore daude berezitasun hauek galtzen laguntzen dutenak . EH hiztegian <i>lagundu</i> ez da NOR-NORK laguntzaileekin agertzen. ... beste arazotxo bat sortzen zaie . Corpusean <i>sortu</i> NOR-NORI laguntzaileekin % 1,3tan agertzen da. Muga % 2n ipini dugu.
Postposizio-lokuzioa ez detektatzea	1	Napoleonen maitasunari (NCZOBJ>) buruzko istorio asko esan dira (NOR).
Arrazoi ezberdinak (49 hitzeko esaldia)	1	... sekulako kontraesanak (NCSUBJ-ERG-S) sortzen dira ... denok (NCSUBJ-ERG-PH) dakigunez (NORK_GUK NR_HURA).

VII.16 Taula: *MaltIxa*ko (M) eta EDGK-ko (E) alarma faltsuak.

zein arrazoiengatik sortzen ote diren ezagutzearren. *MaltIxa* baliatuz sortutako analisi-zuhaitzak erabili ditugunean, 81 alarma faltsu sortu ditu komunztadura-erroreen detekziorako erregelen aplikazioak; EDGK-k sortutakoekin 53. Kopuruak ongi ulertzeko, honakoa azaldu beharra daukagu: *MaltIxa*k sarreran jasotako esaldi guztietarako sortzen ditu mendekotasun-zuhaitzak eta EDGKrekin anbiguotasuna lantzen denez, sortutako analisi-zuhaitzen kopuru izugarria dela eta, ez da beti emaitzarik lortzen. Hots, *MaltIxa*rekin alarma faltsu gehiago sortzen dira, zuhaitz gehiago aztertzen direlako. Konparaketa “garbia” ez den arren, VII.17 taulan kopuru batzuk eman ditugu. Azterketa honetan, aurrekoan ez bezala, erregela guztiak hartu ditugu kontuan. Oso ohikoa da, aditz batekin lotutako bi subjektu detektatzen dituen erregela aplikatzen denean (KOMBIK_SUBJ), subjektuetako baten kasua edo numeroa duen erregela ere (KOMNAH_SUBJ_KAS) aldi berean aplikatzea (kasu gehienetan subjektutzat jotako bat ez da subjektua, beraz ez ditu hizkuntza-ezaugarri egokiak). Kasu honetan, eta antzekoetan (bi objektu), lehen erregela soilik kontatu dugu.

MaltIxa eta EDGK-k analizatutako esaldietan sortzen diren alarma faltsuen errepaso honetan, kopuruak baino garrantzitsuagoak iruditzen zaizkigu arrazoiak, alarma faltsu horiek iturria non duten ezagututa, ekiditeko aukera baitugu. Alarma faltsuei dagokienez bi analizatzaile sintaktikoak batera aztertu ditugunean nabaritutako alarma-iturri diren hizkuntza-fenome-

Erregela	Azalpena	Kop. M	Kop. E
KOMBIK.SUBJ	Aditz berarekin bi subjektu lotu dira.	28	22
KOMBIK.OBJ	Aditz berarekin bi objektu lotu dira.	31	15
KOMADITZ.AZPIKAT	Aditz-azpikategorizazioaren arabera, aditzari ez zaio ongi egokitzen idatzitako aditz laguntzailea.	5	4
KOMNAH.SUBJ.KAS	Subjektuaren kasua okerra da.	8	7
KOMNAH.OBJ.NUMCMOD	Objektuaren numeroa okerra da. Objektua mendeko perpaus jokatu batek osatzen du.	4	-
KOMSOR.OBJ	Objektu bat markatu da, aditz laguntzaileak eskatzen ez duenean.	3	3
KOMSOR.ZOBJ	Zehar-objektu bat markatu da, aditz laguntzaileak eskatzen ez duenean.	1	2
KOMNAH.SUB.KAS.XCOMP	Subjektuaren kasua okerra da. Subjektua mendeko perpaus ez jokatu batek osatzen du.	1	-
Guztira		81	53

VII.17 Taula: *MaltIxan* (M) eta EDGKn (E) alarma faltsuak zein erregelarekin sortzen diren.

noez gain (desanbiguazio morfosintaktiko okerra, aditzen azpikategorizazioa eta postposizio-lokuzioen falta), beste fenomeno batzuk topatu ditugu. Jarraian, interesgarriak iruditu zaizkigunak zerrendatuko ditugu labur-labur, fenomenoetako bakoitza *MaltIxak* (M) edo *EDGK*-k (E) analizatutako dependentzia-zuhaitzetan topatu ote dugun adieraziz:

- *HAULak ez detektatzeagatik* (M eta E). Hitz anitzeko unitate lexikoak errepresentaziorako XML amaraunean oraindik integratuta ez daudenez (lantzen dira, baina haiek errepresentatzeko modua definitu gabe dago²³), ez dauzkagu eskuragarri. Hori dela eta, HAULen osagaiek jasotzen dituzten analisiak ez dira egokienak HAUL direnerako. Adibidez, (67) esaldian, **horrez gain** hitz bikotea, unitate bakar moduan analizatu beharrean, **gain** hitzak izen kategoria jasotzen du, eta geroago, NCSUBJ dependentzia-etiketa.

(67) **Horrez gain** (NCSUBJ>), **bizimodu baten ikurrina** (NCSUBJ>) **gertatzen da**.

- *Aditz modala lotuta ez analizatzeagatik* (M eta E). Arazoa lehenago ere aipatu dugu eta HAULen tratamenduarekin guztiz lotuta dago. HAULak landu ezean, **nahi**, **behar**... eta antzeko aditz modalak, bananduta analizatzen dira eta izen kategoria jasotzen dute, beraz, batzuetan

²³Arazo bera daukagu postposizio-lokuzioekin.

esaldiko subjektua edo objektua bailiran analizatzen dira (ikus (68) eta (69) adibideak). Arazoa HAULak analisi-katean eta amaraunean integratzen direnean konponduko da.

- (68) Beste batzuk (NCSUBJ>-dira), bainatzeko asmotan hurbiltzen direlako hondartzara, arreta handiz ibili behar (NCSUBJ>-dira) dira.
- (69) Dirua (NCOBJ>) behar (NCOBJ>) zutenean.

- *Aposizioa ez markatzeagatik* (M eta E). Egitura hauek errepresentatzeko IXA taldean mendekotasun-etiketa bat definitu dugun arren, APONCMOD, oraindik ez da gehiegi landu. Hori dela eta, (70) moduko esaldietan, aposizioan dagoen elementuari, zein esaldi nagusikoari, subjektu funtzioa esleitzen zaie. Gure ustez, analisisa ez da okerra, bi osagai horiek “kontzeptu” bera adierazten baitute. Etorkizunean planteatzeko kontua da, aposizio kasuetan komunztadura-erroreen detekzioarako erregeletan bi subjektu/objektu edo zehar-objektu onartzea.

- (70) Bazegoen... beste era (<NCSUBJ) bat: neskametza (<NCSUBJ).

- *Bikoizketa, erreferente bera izanagatik* (E). Adibidean ikus dezakegunaren arabera (ikus (71)), **ni** izenordainari NCSUBJ dependentzia-etiketa ezarri zaio, baina baita **neu** bere indartzaileari ere. Esaldiko bi osagai hauek erreferente bera dute, beraz, indartzaileari dagokion dependentzia-etiketarik ipintzen ez bazaio (GALDEMOD), komunztadura-erroreen detekzioarako erregeletan fenomeno hau islatu beharko dugu.

- (71) Ni (NCSUBJ>) neu (NCSUBJ>) ere, neurri honen alde naiz.

Antzeko fenomenoak gertatzen da (72) adibidean. Kasu honetan **gizona** izenak APONCMOD etiketa beharko luke, **hau** objektuaren aposiziotzat jotzen baita.

- (72) 1769an jaiotako gizon (NCOBJ>) bat dugu hau (<NCOBJ)

- *Predikatiboa eta subjektua nahasteagatik* (M eta E). Segur aski desanbiguazio okerra dela kausa, batzuetan predikatiboari subjektu dela adierazten duen mendekotasun-etiketa esleitu zaio (ikus (73) adibideak).

- (73) Morroia (NCSUBJ>) umea (NCSUBJ>) edo heldua izan zitekeen.

- *Enumerazioa* (M eta E). Zerrendetako osagaiak unitate bakarrean biltzen ez badira, osagaietako bakoitzean errepikatzen da mendekotasun-etiketa (ikus (74) adibidea), KOMBIK.SUBJ eta KOMBIK.OBJ erregelen aplikazioa eraginez.

(74) **Lehenengo sailkapenean Hondarrabikoa (NCSUBJ>), Kontxa eta Ondarreta (NCSUBJ>) sartzen dira.**

- *Koordinazioa* (M eta E). Zerrenden inguruan deskribatu berri dugun arazoa, koordinazioarekin errepikatu egiten da (ikus (75) adibidea).

(75) **Liburu (NCSUBJ>) eta informazio gehienak (NCSUBJ>) erderaz daude. . .**

- *Hitzak nahasteagatik*. Komunztadura-errorerik ez duen arren, (76) adibideko esaldi erroredunean, **basapisti** hitza gaizki erabili da. Gure ustez, idazleak **basati** hitza erabili nahi izan du. Nahastuta erabili duen hitzaren gramatika-kategoriaren aldaketak, errore baten detekzioa ahalbidetu du (komunztadura-errorea ez den arren, bi objektu markatu dira).

(76) ***Txakur (NCOBJ>) basapisti (NCOBJ>) bat enkargatu dugu.**

VII.7.3.2 Corpus zuzena

Aurreko atalean, analizatzaile sintaktikoen analisi-akatsak komunztadura-erroreen detekziorako erregelen ebaluazioa eragozten dutela ikusi dugu. Bestela esateko, komunztadura-erroreak detektatzeko erregelek maiz analizatzaileen akatsak azalarazten dituzte, gaizki eratutako mendekotasun-zuhaitzak detektatzen dituzte, erroreak detektatu beharrean. Komunztadura-erroreak detektatzen dituen gramatikaren doitasuna probatzeko modua, erregelak euskarazko *treebank*ean, teoriarik analisi-zuhaitz perfektuen bankuan, aplikatzea dela otu zaigu.

Euskarazko *treebank*a

Euskarazko *treebank*a (Aduriz *et al.*, 2003b) 3LB proiektuaren barnean garatu zen. Proiektuak maila morfosintaktiko eta sintaktikoan eskuz etiketatutako corpusak sortu zituen katalanerako, euskararako eta gaztelararako (Palomar *et al.*, 2004). Katalanerako eta gaztelararako corpusak osagai-egituretan oinarrituta analizatu ziren eta euskarazkoa mendekotasun-egiturak oinarri hartuta.

3LB *treebanka* bere euskarazko bertsioan 50.223 tokenek osatzen dute, *Abar-hitz* tresna erabilia etiketatu zen eskuz, eta gero tratamendu automatikorako egokitu zen (Bengoetxea eta Gojenola, 2007). Egokitzapenaren ondorioz, mendekotasun-zuhaitzak lehenago ere aipatu dugun “CoNLL 2007” kongresuan erabilitako formatu tabulatuaren lortu ziren. Gure beharretarako, azken formatua programa baten bidez *Saroik* sarreran erabiltzen duen XML amaraunera bihurtu dugu.

Corpus honetan ez dago inolako anbiguotasun sintaktikorik. Mendekotasun-zuhaitz *bakarra* dugu esaldiko, eta emaitzak lortu ahal izateko, horietako bakoitzean komunztadura-erroreen detekzioarako erregelak aplikatzea besterik ez dugu egin.

Treebankeko esaldiek *XX. mendeko euskararen corpus estatistikoa*²⁴ (Aduriz *et al.*, 2003a) eta *Euskaldunon Egunkarian* dute iturburua. Corpusa, berriro ere, bi zatitan banatu dugu: garapenerako bata eta probarako bestea (ikus VII.18 taula). Corpus erroredunean egin genuen modu berean, erregelak aldatzeko beharrik ikusten ez bada, lehen testu-bildumarekin soilik egingo dugu proba.

Corpus zuzena		
	Garapenerako corpusa (% 60a)	Probarako corpusa (% 40a)
Esaldi kopurua	1906	1271

VII.18 Taula: *Treebankaren* banaketa ebaluazioarako.

Emaitzak

Esperimentu honetarako 1906 esalditako *treebankaren* azpimultzo bat erabili dugu (hasiera batean *garapenerako corpus* moduan definitu genuena). Erregelak aplikatu ondoren, 161 esalditan detektatu da errorea, hau da, corpusaren % 8,45 (ikus VII.19 taula). Alarma faltsuen artean bereizketa egin dugu, batzuk gure ustez ezin baitira halakotzat jo:

- Ustezko alarma faltsu hauetatik 90 *alarma faltsu ez kontsideratzea* erabaki dugu honako arrazoiagatik: ez dira sistemaren gehiegizko markaketak izan, batzuk *treebanka* etiketatze hartutako erabakiengatik sortu baitira (bi subjektu aditz berari lotuta, adibidez), eta beste batzuk, etiketatze-akatsengatik. Anotazio-akatsak direla esan dezakegu.

²⁴Corpus honek lehen *EEBS-Egungo Euskararen Bilketa-lan Sistematikoa* zuen izena.

Horiek kontuan hartuz gero, zuhaitz-bankua ebaluatzen arituko ginateke, eta ez erroreen detekziorako sistema.

- Detektatutako gainontzeko 63 erroreak, berriz, *alarma faltsuak* dira, eta hauek, *treebank* osoaren % 3,305 osatuko lukete. Horietatik 58 kasutan (errorea aditzei buruzko informazio faltagatik gertatu da, eta ez erroreen detekziorako erregelen hutsengatik. Aditzaren erabilera zehatz bat bai corpusean, baita hiztegian ere konprobatzen denean, alarma faltsuen portzentajea izugarri jaisten da.

Sistemak adierazitako erroreak	Kop.	AF-etatik portzentajea	Treebanketik portzentajea
Alarma faltsu kontsideratuko ez ditugunak	90	% 55,9	% 4,72
Alarma faltsu kontsideratuko ditugunak	63	% 39,13	% 3,305
Esaldi erroredunak	8	% 4,97	
Guztira	161		% 8,45

VII.19 Taula: Alarma faltsuak.

Alarma faltsuek azterketa sakonago bat merezi dutela iruditzen zaigu, zenbait fenomeno hizkuntzaren ikuspuntutik interesgarriak direla uste baitugu. Horien zerrendapen bat egin dugu ondorengo lerroetan, fenomenoak corpusean topatutako adibidearekin lagunduaz. Lehenengo, alarma faltsu kontsideratuko ez ditugun kasuak zerrendatuko ditugu, gero, alarma faltsutzat jo ditugunak.

Alarma faltsutzat jo ez ditugun kasuak (90 adibide, % 55,9)

Ondoren zerrendatuko ditugun kasuetan, erroreak markaketaren arrazoia erabilitako corpusaren etiketatze linguistikoa dago eta ez guk landu beharreko erroreen detekziorako erregetan. Batzuetan, etiketatze-erabakien ondorioz gertatzen dira markaketak; besteetan, etiketatze-erroreen ondorioz. Kasu guztietan, etiketatzea aldatuz gero, errege-erregelak ez lirateke aplikatuko. Hona hemen alarma faltsu *faltsuen* inguruko azalpena:

- *Treebanka etiketatze moduagatik sortutakoak* (41 kasu). Aditzari lotuta gutxienez bi subjektu edo bi objektu topatu ditugu multzo honetako esaldietan. *Treebankean* hizkuntzalariek *elipsia* errepresentatu nahi izan zutenean, tratamendu konputazionalerako egokia ez zen aukera egin zuten. Elipsiarekin eskuz etiketatutako zuhaitz haiek automatikoki landu ahal izateko, adabegien arteko lotura berriak egin

behar izan dira eta urrats honetan sortu dira etiketatze okerrak. Har dezagun (77) esaldiari lotutako zuhaitza (ikus VII.12 irudia). Bertan, **asmatu** aditzaren subjektutzat hartzen dira **nagusiek** eta **haurrek** hitzak (azken honek aditz eliditua du). Hurrengo urrats batean *tree-banka* etiketatze modua aldatu da, eta orain, elipsiak saihestu egi-ten dituztenez (ikus xehetasunen bila (Aranzabe, 2008) tesi-txostena) aurrerantzean ez da funtzio gramatikalen bikoizketarik gertatuko.

(77) (...) **nagusiek** **asmatu** ditugu **umekeriak**, ez **haurrek**.



VII.12 Irudia: Elipsia, aditzari lotuta bi subjektu.

- *Etiketatzeko-erroreen ondorioz sortutakoak* (49 kasu). Jarraian zerrendatuko ditugun kasuetan mendekotasun-erlazioa gaizki etiketatu da, edo akatsen bat dago analisi morfosintaktikoan:

- *Objektuaren sorkuntza, subjektuarekin nahastu izanagatik* (17 kasu): NOR eta NOR-NORI paradigmako aditz laguntzaileak dituzten perpausetan, objektu moduan etiketatutako elementuren bat agertu zaigu. Kasu gehienetan, perpausak inbertsonalak dira eta nahastuta, hizkuntzalariak, osagairen bat subjektu moduan markatu beharrean, objektutzat jo du. Adibidez, (78) adibidean, **badi-ra** aditz laguntzailea duen **aztertu** aditzari, objektu funtzioa duen osagai bat esleitu zaio (**bakuoloak** eratzeko **gaitasuna**, **inhibizioa**, **etab.**).

(78) Hala ere, protoplasmaren zenbait ezaugarri fisiko eta kimiko (**bakuoloak** eratzeko **gaitasuna**, **inhibizioa**, **etab.**) aztertzen **badira** bakarrik uler daitezke.

- *Subjektu-objektu nahasketagatik* (6 kasu).
- *Absolutibo plurala - ergatibo singularra nahasketagatik* (18 kasu). Kasu hauetan, aurrekoetan (subjektu-objektu nahasketa) gertatzen ez den bezala, subjektua ongi etiketatu da, baina analisi

morfologikoa okerra da. Hau da (79) adibidearen kasua, zeinetan subjektuak absolutibo-plurala analisia jaso duen (**du** aditz laguntzailearekin komunztadura ongi egiteko ergatibo-singularra analisia beharko luke izan).

(79) **Hernaniko udalak (NCSUBJ>) eskubide osoa du erabaki hori hartzeko.**

- *Bestelako etiketatze erroreengatik* (6 kasu): hitzak errepikatuta, elementuak gaizki lotuta. . .
- *Mugatu-zero* nahasketagatik (2 kasu). Kasu hauetan sintagmako azken elementua ez den hitz bati atzizki mugatu bat esleitu zaio. Adibidez, **ekintza zuzena** sintagman, **ekintza** hitzari *absolutibo-mugatu* analisi ez zuzena eman zaio. Euskaraz, kasua eta numeroa sintagmako azken elementuak eraman ohi dituenetz, adibideko kasuan eta antzekoetan, analisi automatikoa egitean **ekintza** hitzari *zero* etiketa esleitzen zaio.

Alarma faltsuak (63 adibide, % 39,13)

Egitura zuzena izanik, errorea detektatu da zerrendatutako ondorengo kasuetan. Aipatzekoa da aditzen erabilerari buruzko informazio faltagatik gertatzen den alarma faltsuen kopurua. Alarma faltsuen % 92 arrazoi honengatik gertatzen da.

- *Aditz-azpikategorizazioa* (58 kasu). Aditz nagusia ohikoa ez duen paradigma bateko aditz laguntzaile batekin agertzen denean, errorea detektatzen da esaldi zuzenetan ere. Mota honetako alarma faltsuak gerta ez daitezten, azpikategorizazio-informazioa osatu beharko litzateke, adibidez, corpusetako informazioa jasotzeko testu-bilduma zabalagoa erabiliz, edo beste hiztegiren bat baliatuz. *Treebank*ean agertu zaizkigun eta *Euskal Hiztegia*n topatu ez ditugun aditzen gauzatze batzuk zerrendatu ditugu VII.20 taulan. Hiztegia corpusetan topatutako informazioarekin aberasteko bidea egon daitekeela ikusi dugu azterketa hau egin dugunean. Ideia horri eutsiko diogu berriro VII.8 atalean.
- *Pertsona-komunztadura* (2 kasu). Aditz laguntzaileko komunztadura-marketan erakusten den pertsonaren eta perpausean agertzen denaren artean pertsona komunztadura zuzena gertatu ez arren, zenbait komunztadura zuzenak dira euskaraz. Corpusean agertu zaigun adibidean (ikus (80) adibidea) **gara** aditz trinkoan “lehenengo pertsona

Aditza	Treebankean azpikat. agerpena	Adibidea	Agerpen kopurua	EHn
jarraitu	NOR-NORK	XVII. mendearen erdialdera hainbat ikerlarik kamera ilunetik eratorritako arazo desberdinetan sakontzen jarraitzen dute.	11	ZAIO
jarraitu	NOR-NORI-NORK	"... PSE-EE benetako ahalegi-na ari dela egiten PPri jarraitzen dion politikatik urruntzeko..."	2	ZAIO
eman	NOR-NORI	Duela 25 milioi urte, aldi berri bati eman zitzaion hasiera;	3	DU
oztopatu	NOR-NORK	Argia ondo sartzen da inguru-ne garden batean, baina gorputz opako batek oztopatu egiten du bere jarraipen lerrozuzena...	3	DA
erantsi	NOR-NORK	"badago gehiengo sozial aurrera-koi bat osatzeko aukera", erantsi zuen.	3	DIO
egin	NOR-NORI	Sinesgaitz egiten zitzaion oro...	1	DU
flotatu	NOR-NORK	Errealitatea, fotografikoki, jada-nik agortu eta erreportajea "hil delako" sentimenduak flotatzen du.	1	DA
frogatu	NOR-NORI	... behin baino gehiagotan fro-gatu bait zaizu legearen araeu-rakoa zela.	1	DU
gainezkatu	NOR-NORI	... arteriak eta zainak adrenalina-z gainezkatzen zitzaizkiolarik.	1	DU
hautsi	NOR-NORI	Punduok pindatzen lapitaren punta hautsi zaiola profita-tuaz,...	1	DU
igarri	NOR-NORI	Ozta igartzen zaio lehenago zu-ria izandakoa dela.	1	DIO
lagundu	NOR-NORK	... konponbidea bilatzen lagun-tzen duelako".	1	DIO
marraztu	NOR-NORI	Irribarre triste labur bat ma-rraztu zitzaion ezpain bazterre-tan...	1	DU
oihukatu	NOR	... onartu zuen une batean "Go-ra ETA" oihukatu zela...	1	DU
tartekatu	NOR-NORI	... Rosaren irudia tartekatu zi-tzaion berriro...	1	DU

VII.20 Taula: EH hiztegia falta diren azpikategorizazio-etiketak

plurala” eskatzen zaio subjektuari, baina honek (*jendea*) “hirugarren pertsona singularra” du. Subjektuko hitzak, ordea, multzoa adierazten duenez, “gu” ere barnean hartzen du, komunztadura zuzena eginez.

(80) *jendea* [*3. perts., sing.*] *gara* [*1. perts., pl.*].

Pluraleko lehenengo eta bigarren pertsonetan komunztadura erlaxatu beharko litzateke, hirugarren pertsonarekin komunztadura zenbait kasutan (multzoa adierazten den kasuetan, adibidez) onar zezan.

- *Juntadura-marratxoaren erabilera* (kasu 1). Emendiozko juntadurak, aditzari plural komunztadura eragiten diola esan dugu. (81) adibideko esaldian, marratxoak (“-”) emendiozko juntagailuaren papera beteko balu, esaldia zuzena litzateke. Analisi kateak *EAJ-EA* karaktere-katea unitate bakarra bailitzan eta numero singularrarekin analizatu du. Hori horrela, subjektuak aditzarekin komunztadura okerra egiten duela adierazi du.

(81) *EAJ-EA* [*sing.*] *udalaren erabakia kritikatu dute* [*pl.*].

Semantikarik erabili ezean marratxodun hitzetan “batuketa” adierazten duten hitzak, bestelakoetatik bereiztea (adibidez, *duda-muda*) oso zaila da. Beraz, alarma faltsu hau ekiditea ere, hala da, gaitza.

- *Zenbatzaileen komunztadura* (2 kasu). Euskaltzaindiaren arabera zenbatzailedun sintagmekin aditzak numero komunztadura egiten duenean, “bi aukerak zilegi dira, bai singularrekoa bai pluralekoa (...). Zenbakiak direnean, ordea, askoz ere arruntagoa da pluraleko komunztadura, singularrekoa ere ontzat eman behar den arren (JAGONET, 2000/07/19ko galdera)”²⁵. Tesi-lan honetan ez da zenbatzaileekin tratamendu berezirik egin, eta hori dela eta, (82) adibidean errorea markatu da, subjektu funtzioa egiten duen sintagma (*9729 milioi pezeta*) singularra baita, eta aditz laguntzaileak (*baitziren*) plurala eskatzen baitu.

(82) *Igoera handia da, 1996an 9729 milioi pezeta inbertitu baitziren.*

Alarma faltsu hauek ekiditeko, zenbatzailedun sintagmetarako aditzarekiko komunztadura singularrean eta pluralean baimendu beharko litzateke.

²⁵Ikus ere Gramatika Batzordearen *Euskal Gramatika: Lehen Urratsak* liburukia (1985), 205-208 orr.

Erroreak (8 adibide, % 4,97)

Etiketatzailer automatikoaren erroreak alde batera utzita, “corpus zuzenean” idazleek egindako erroreak ere topatu ditugula esan beharra dugu. Batez ere, sintagma-mailako komunztaduran:

- *Sintagma-mailakoak* (6 kasu). Honako adibideetan topatu dugu errorea: *garrantzia handia dauka, *beste pastoral bat (2 alditan), *pastoral bakarra, *adatsa luzea eta *sendotasuna honen arrazoia.
- *Alokutiboa* (kasu 1). Ondorengo adibidean, idazleak adizki alokutiboa nahastu egin du: *ez huen etorri erabili du, ez zuan etorri beharrean (ikus Euskaltzaindiaren 14. araua²⁶).
- *Komunztadura-errorea subjektuaren eta aditzaren artean*. Honako adibidean topatu dugu komunztadura-errorea.

(83) Joan den mende-amaierako eta mende honen hasierako zientzilari askok zeraren eritzikoak ziren (...)

Esperimentu honetan, *Saroï*ekin *treebank*ean alarma faltsuak detektatzear gainera, bi helburu bete ditugula ikusi dugu:

1. *Treebank*eko etiketatze-akatsak detektatu ditugu.
2. Aditzen ereduetan erabilerari buruzko hutsuneak topatu ditugu.

²⁶ *Izan*-en ez-oraineko NOR bakarreko 3. pertsonako adizkera alokutiboentzat, 2. pertsonako indizea atzizki gisa dakarten formak hobestea (aurrizki gisa daukatenak baino): adibidez, *zen* → *zuan*, eta *ez huen*.

VII.8 Hiztegi aberasketa

VII.8.1 Sarrera

Aurreko atalean *Euskal Hiztegiak* aditzen azpikategorizazioaren inguruan hutsune batzuk dituela ikusi dugu. Hutsune horiek betetzeko, hots aditzei buruzko informazioa aberasteko, proposamen bat egingo dugu atal honetan (Díaz de Ilarraza *et al.*, 2007a). Aberasketarako proposamenak, azpikategorizazioarekin lotutako alarma faltsuen kopurua gutxitzen lagun lezakeela uste dugu, baina hala ere, oraingoz ez dugu erabili. Azpimarratu nahi dugu, gurea proposamen bat besterik ez dela. Ez dago gure esku EH hiztegiaren inguruko erabakirik hartzea. Gure helburua, topatutako ahultasunak aurkeztea da, baita erroreen detekzioarekin zerikusirik ez duen *Saroi* tresnaren aplikazio bat erakustea ere.

Hiztegiak oinarritzko baliabideak dira. Informazio lexikoaren iturri aberatsak, gainera. Haatik, denbora luzea behar da hiztegiak sortzeko eta batzuetan ez dituzte hizkuntzaren erabileran ematen diren aldaketak islatzen.

Lan ugari garatu dira hiztegiak automatikoki aberasteko helburuarekin. Horiek lantzen dituzten alderdiak ugariak dira, baita iturritzat hartutako edo sortutako baliabideak ere. Adibidez, Gelbukh *et al.* (2004) egileek aurkeztutako lanean, hitz-konbinazioekin osatutako hiztegi bat aberastu egin zuten mendekotasun-egituretan oinarritutako analizatzaile sintaktiko bat erabiliz. “*Prague Dependency Treebank*” ikasketa automatikoko teknikak erabiliz txekiera hizkuntzako aditzen azpikategorizazio-egiturak ikasteko erabili zuten Zeman eta Sarkar (2000) lanean. Briscoe eta Carroll-ek (1997) agerpen-maiztasunak erakusten dituen aditz-azpikategorizazio egituren hiztegi bat sortu zuten estatistikan oinarritutako analizatzaile bat erabilita. “*BulTreeBank*” izeneko beste *treebank* bat erabili zuten Simov eta Osenova-k (2004) lexikoi semantiko bat sortzeko. Kilgarri-ek (1997) hitzen maiztasunak corpus batetik erauzi zituen eta *Longman Dictionary* hiztegian gehitu.

Gure kasuan, helburu dugun hiztegia lehenago ere aipatu dugun *Euskal Hiztegia* (EH) (Sarasola, 1996) da. Hauxe 1996 urtean sortu zen. Orduz geroztik Euskaltzaindiak hitz batzuen forma estandarrei buruzko erabakiak hartu dituenaz, 2003 urtean hiztegiko sarreraren estandarizazio lan bat egin zen (Arregi *et al.*, 2003). Bestalde, corpusetan hizkuntzaren aldaketak hiztegietan baino hobeto islatzen direnez, testuetatik abiatuta hiztegiko aditz-informazioan topatutako hutsuneak betetzeko lanari ekin diogu.

Hiztegia aberasteko prozesua ahalik eta automatikoena izan dadin saiatu gara, eskuzko lana minimora murrizteko asmoa dugularik; ahal izanez

gero, emaitzen azterketa soilik egingo dugu eskuz. Erabili ditugun baliabideak honakoak izan dira: aipatu berri ditugun corpusetik automatikoki erauzitako aditzen gauzatze-egiturak eta EHren azterketa automatikoaren ondorioz lortutako aditz-patroiak; eta bestetik, euskararako dependentzia *treebanka*. Baliabide hauek guztiak *Saroï* erabilia kudeatu ditugu. Baliabideak osagarriak direla uste dugu: i) corpusak gaur egun egiten den hizkuntzaren erabilera islatzen du, ii) hiztegia eskuzko lan linguistiko izugarria egin ondoren garatu zen eta iii) *treebankak* bi ikuspuntuak biltzen ditu.

Hiztegiko aditzak bi modutan aberasteko proposamenak egingo ditugu:

1. **Azpikategorizazio-informazioa aberastuaz.** Corpuseko eta hiztegiko aditzen artean azpikategorizazioari buruzko desadostasunak topatu ditugunean, egoki iruditzen zaigun azpikategorizazio-informazioa hiztegian gehi dadin proposatu dugu.
2. **Aditzak gehituaz.** Corpusean topatu ditugun eta hiztegian topatu ez ditugun aditzak hiztegian gehi daitezen proposamena egin dugu, aditz horiek corpusetik aparte *treebank*ean ere agertzen badira.

Hiztegia aberasteko informazioarekin sortutako aditz-zerrendak hizkuntzalari bati erakutsi zaizkio errepasa ditzan. Proposamenak egiteko hiztegitik erauzitako informazioa erabili dugunez, *feedback* prozesu baten aurrean gaudela esan dezakegu: EH hiztegia erabili dugu, EH hiztegia aberasteko.

Lana garatzeko baliabideak ezagunak ditugunez, esperimenduak aurkezteari ekingo diogu. Hauek garatu aurretiko prestakuntza-lana aurkeztuko dugu VII.8.2 atalean. Geroago, VII.8.3 eta VII.8.4 atalak esperimenduak azaltzeko erabiliko ditugu. Azkenean, ateratako zenbait ondorio laburbilduko ditugu VII.9.2 atalean.

VII.8.2 Aurreprozesua

Honenbestez, jatorri ezberdina duten datu linguistikoekin osatutako hiru baliabide ditugu eskuartean: a) corpusetik erauzitako aditz multzo bat aditz bakoitzeko laguntzaile mota adierazita duelarik, b) informazio bera duen baina hiztegitik jaso dugun beste aditz multzo bat, eta azkenik, c) euskara estandar eta zuzeneko *treebanka*.

Aditz multzoak modu honetan osatuta daude: corpusetik 2.541 aditz lortu ditugu eta hiztegitik 4.016 aditz. Guztira 5.264 aditz desberdin. Zenbakiak erakusten dutenaren arabera, aditz batzuk bi zerrendetan agertzen diren bitartean, beste batzuk bakarrean agertzen dira:

- Corpusean soilik agertzen diren aditzek 1.248 aditzeko multzoa osatzen dute. Honi “*Corpusean soilik*” edo “*CS*” izena ipini diogu eta honelakoekin osatuta dago:
 - Kazetaritzako idatz-estiloan agertzen diren aditzak baina, oraindik EHn agertzen ez direnak. Adibidez, **amnistiatu**, **desmilitarizatu**, **estraditatu**, **klonatu**...
 - Oker idatzitako aditzak: **aiaptu**, **azpinmarratu**...
 - Ez sarrera, ezta azpisarrera moduan agertzen ez diren hitz anitzeko aditzak. Adibidez, **atsekabe hartu**, **kolokan egon**...
- Hiztegietan soilik agertzen diren aditzekin 2.723 aditzeko multzo bat osatu dugu. Multzoa “*Hiztegian soilik*” edo “*HS*” izendatu dugu. Multzo honetako aditzak EH hiztegia honela markatuta daude:
 - Gutxi erabiliak edo gaur egun gutxi erabiliak (G.er. edo G.g.er.): **ogitu**, **indartsutu**, **desarratu**, **urgoitu**...
 - Euskalkietan erabiltzen diren aditzak (Bizk., Ipar., Sart., Naf.): **haurridetu** (Ipar.), **deboilatu** (Ipar.), **ibitu** (Bizk. batez ere), **kokolotu** (Bizk.), **komekatu** (Naf.)...
 - Hizkuntza-maila altukoak diren aditzak (Goi): **hatsanditu**, **minztuldu**, **ukaldikatu**...
- Azkenik, hiztegian zein corpusean agertzen diren aditzekin 1.293 elementuko multzoa osatu dugu eta “*Bietan*” edo “*B*” izendatu dugu.

Atal honetan azaldukotakoa laburbiltzen du VII.13 irudiak.



VII.13 Irudia: Corpuseko eta Euskal Hiztegiko aditzak.

VII.8.3 Corpuseko eta EH hiztegiko aditzen arteko inkoherentziak

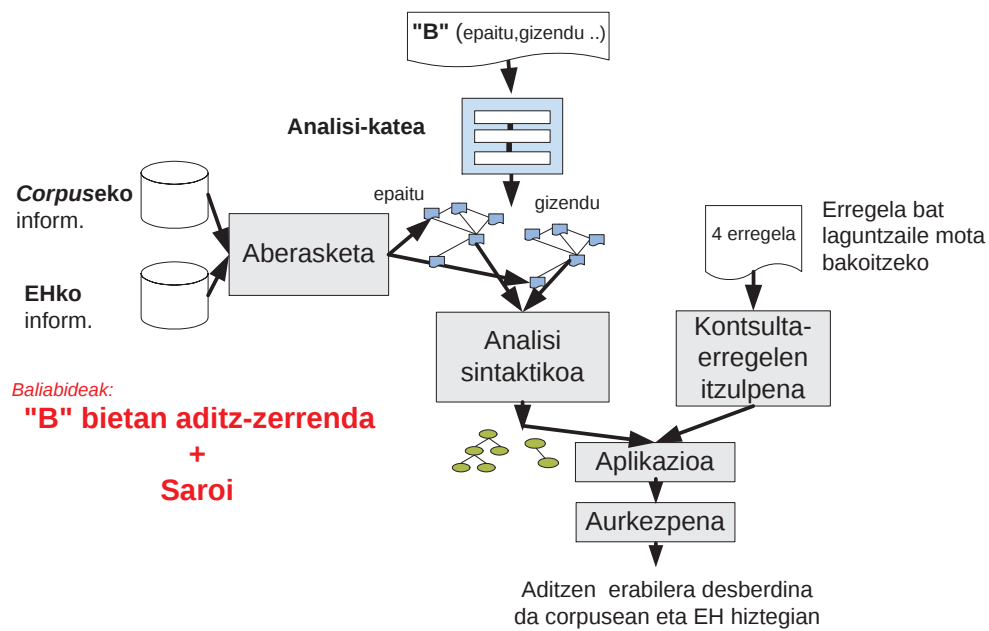
Esperimentu honetarako erabilitako baliabideak, “*Bietan*” aditz-zerrenda eta *Saroï* dira. Esperimentuaren helburu nagusia, corpusean agertu den aditz-azpikategorizazioari buruzko informazioaren eta hiztegian agertzen diren arteko inkoherentziak topatzea da. Gure iduriko, inkoherentzia ematen da corpusean eta hiztegian aditz bakoitzerako adierazten diren laguntzaileak desberdinak direnean. Adibidez, inkoherentzia gertatzen da **zauritu** aditza corpusean NOR motako laguntzaileekin agertzen bada baina DU etiketa badiu hiztegian (gogora dezagun DU etiketak NOR-NORK eta NOR-NORI-NORK paradigmak adierazten dituela).

VII.8.3.1 Esperimentua

Ikus dezagun urratsez-urrats inkoherentzien bila jarraitutako prozesua (prozesua grafikoki VII.14 irudian adierazi dugu):

1. Lehenengo, corpusean zein hiztegian (“*B*” aditz-zerrenda) agertzen diren aditzak analizatu ditugu IV. kapituluan azaldutako analisi-katea erabiliz. Ondorioz, *Saroï*ren sarreran erabiliko den anotazio-amaraun lortu dugu zerrendako aditz bakoitzeko.
2. Geroago, aditzak corpusetik eta hiztegitik jasotako informazioa erabiltuta aberastu ditugu. Aberasketa-prozesua amaitutakoan, aditz bakoitzak lematizazioko dokumentuan aditzei buruzko informazio guztia izango du (ikus atal honetako VII.5 irudia).
3. Bukatzeko, lau erregela —aditz laguntzaile mota bakoitzeko bat— aplikatu dizkiogu dagoeneko analizatuta eta aberastuta dagoen aditz-zerrendari, horretarako *Saroï* erabiliz. VII.15 irudian NOR motako aditz laguntzaileetan inkoherentziak topatzeko erregela ikus dezakegu. *Saroï*k mendekotasun-zuhaitzak analizatzen dituenaz, aditz bakoitzari laguntzaile bat lotu zaio zuhaitz *artifizial* bat sortuz.

VII.15 irudiko erregela modu honetan parafrasea dezakegu: zuhaitz batek erregela betetzen duela adieraziko dugu, baldin eta uneko adabegiaren (“@”) kategoria gramatikala (@.kat) aditza (ADI) den, aditzak 4 baino agerpen gehiago dituen corpusean, eta NOR motako aditz laguntzaile batekin agertzen den 50 baino handiagoa den proportzioan. Informazio horren guztiaren iturria corpora da. Gainera, aditz berarentzako hiztegitik jasotako informazioak aditza ez dela NOR motako



VII.14 Irudia: Corpuseko eta EH hiztegiko aditzen arteko inkoherentziak.

```

Erreg_KOM_Inkoherentziak_Nor_1
(
  Detekt (
    @.kat == "ADI" &
    @.agerpenak >4 &
    @.NOR-maizt >50 &
    @.NorEz~
  )
)

```

VII.15 Irudia: NOR paradigmako aditz laguntzaileak dauzkaten aditzetan inkoherentziak topatzeko erregela.

laguntzaileekin agertzen esaten du (@.NorEz~). Corpusetik erauzitako datuek (aditza agertzen den aldietatik erdia baino gehiagotan NOR motako laguntzaile batekin agertzen da) eta hiztegitik jasotakoak (aditza ez da NOR motako laguntzaileekin agertzen) kontrako informazioa ematen dute, beraz, inkoherentzia topatu dugu. Erregelari ez da aditzaren lehen edo formari buruz galdetzen eta honek erregela orokorra egiten du laguntzaile mota bera duten aditz guztietarako.

Corpusen 4 alditan baino gehiagotan agertzen diren aditzak aztertzeak badu bere arrazoia. Baldintza hau erabilita, oker idatzitako eta analisi-erroreen ondorioz aditz moduan markatutako hitzen azterketa ekidin nahi dugu. Honezaz gain, inkoherentzia argia izan dadin aditzak laguntzaile mota horrekin agertu behar du agerpen guztien aldien erdia baino gehiagotan (% 50tik gora aldiz).

Aipatutako VII.15 erregelarekin NOR paradigmari lotutako inkoherentziak topatu ditugu honako aditzetan: **urritu**, **zauritu**, **bueltatu**, **gutxitu**. . . Hauek guztiak corpusen NOR motako laguntzaile batekin agertu dira eta hiztegiari DU azpikategorizazio-etiketa dute. Honako aditzak NOR-NORK motako laguntzailearekin idatzi dira testuetan eta parentesi artean adierazitako etiketarekin EHn: **begiztatu** (DA), **erantsi** (DIO), **ilundu** (DA). . . NOR-NORI motakoak dira **zilegitu** eta **zor izan** aditzetan agertutakoak (biak DU sailekoak hiztegiari). Inkoherentzia hauen inguruko datuak hurrengo atalean zehaztu ditugu.

VII.8.3.2 Emaizak

1.293 aditzeko zerrendatik, 53k (% 4) aditz laguntzaileekin zerikusia duten inkoherentziak dauzkate.

- 45 kasutan (% 84,9) NOR motako inkoherentziak daude.
- 6 alditan (% 11,3) NOR-NORK motako inkoherentziak topatu ditugu.
- 2 alditan soilik (% 3,77a) markatu ditugu NOR-NORI motako diferentziak.
- Ez dugu NOR-NORI-NORK paradigmarekin lotura duen inkoherentziarik aurkitu.

NOR motako inkoherentzien proportzio altuaren arrazoia aski ezaguna dugu. Esaldi inbertsonaletan azaltzen den alternantzia honek ez du hiztegiari islarik.

Topatutako inkoherentzietan kasuistika hau lortu dugu:

- Kasuetatik 36tan (% 67,9) hiztegi-sarreran aditzaren alternantziaren bat falta da (inbertsonala, inkoatiboa...). Kasu honetan, corpusean topatutako alternantzia-egiturak eta dagozkien adibideak gehi daitezke hiztegian.
- 11 kasutan (% 20,7) aditza hiztegian eta corpusean desberdin erabili da. Kasu horiek aditzen egungo erabilera eta erabilera-aldaketen arrazoiak aztertzeke erabil daitezke.
- 5 errore (% 9,4) identifikatu ditugu hiztegian. Kasuak banan-banan eskuz aztertu ditugu eta hiztegi-sarreran laguntzaile mota bat zehazten bazen ere, adibideetan beste batzuk adierazten zirela ikusi dugu.
- Kasu bakarrean (% 1,9) nahiz eta hitz-forma berdina izan, corpusean eta hiztegian erabilitako adierak desberdinak direla ikusi dugu.

Saroi erabilita 1.293 aditzeko zerrenda batean 53 aditz (% 4) inkoherentziadun moduan markatu dira. Hizkuntzalari batek aurkeztutako proposamen guztiak benetako inkoherentziak direla baieztatu du, beraz, emaitza fidagarriak lortu ditugu. Inkoherentzia hauek hiztegirako aditz-alternantzia eta aditz-erabilera berriak proposatzeko erabili dira. Gainera, corpusak hizkuntza-informazio eta hizkuntza-erabilera iturri moduan duen garrantzia erakutsi dute.

VII.8.4 EH hiztegian sarrerak gehitzen

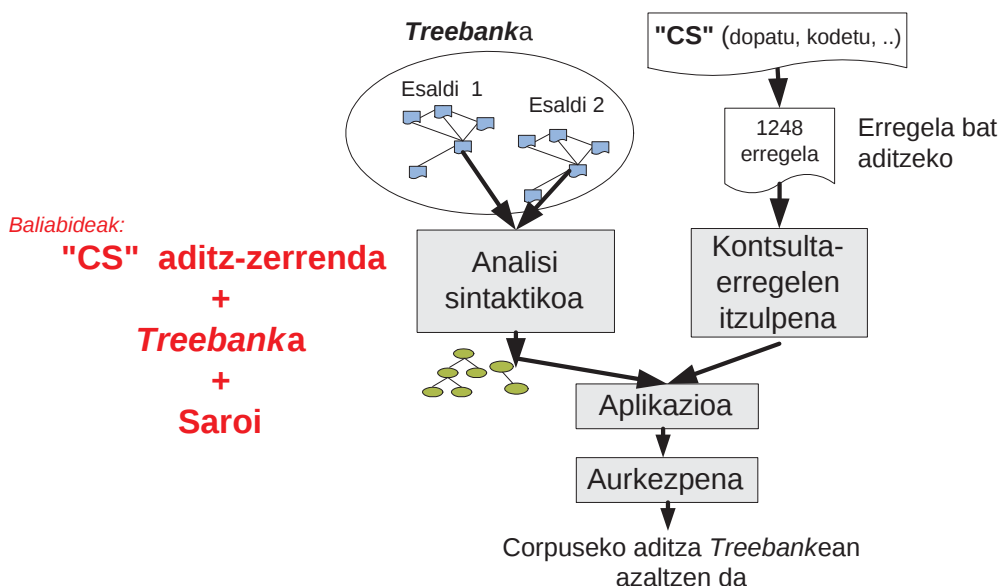
Bigarren esperimentuaren helburua, EH aditz-sarrera berriekin aberastea da. Horretarako, “*Corpusean soilik*” (CS) aditzen zerrenda eta *treebanka* erabili ditugu.

Aditz bat EH hiztegian gehitzeko proposatuko dugu, baldin eta corpusean agertzeaz gain, *treebank*ean ere agertzen bada. *Treebanka* eskuz etiketatutenez, hizkuntza-informazio zuzenaren eredutzat jotzen da eta aurkeztu dugun helbururako nahikoa bermea ematen duela uste dugu. *Treebank*en abantaila da bertako datuek sintaxi-analizatzaileekin lortuko liratekeen datuek baino zarata txikiagoa dutela (Zeman eta Sarkar, 2000).

VII.8.4.1 Esperimentua

Corpusean eta *treebank*ean tokia duten, baina hiztegian agertzen ez diren aditzak prozesu hau jarraituz lortu ditugu (prozesuaren eskema VII.16 irudian):

1. *Treebankean* aditz-lema zehatzen bila ari garenez, lehenengo, automatikoki VII.17 irudiko erregelaren antzeko erregela bat sortu dugu corpusean bakarrik agertzen den aditz bakoitzeko (1.248 aditz).
Irudiko erregelaren arabera, honako baldintzak betetzen dituzten mendekotasun-zuhaitzak markatuko dira: ADI gramatika-kategoria du, “*ados etorri*” lema du eta corpusean 10 aldiz baino gehiago agertzen da.
2. Sortutako erregelak *treebankean* aplikatu dira *Saroi* erabilita. *Treebankean* erregelen bidez markatutako aditzak dira EH hiztegian sartzeko hautagaiak direnak.



VII.16 Irudia: EH hiztegian sarrerak gehitzen.

VII.8.4.2 Emaitzak

Esperimentu honen emaitzak VII.21 taulan zehaztu ditugu. Lehen zutabea proposatutako aditzak ikus daitezke. 2. zutabeak corpuseko agerpen kopurua erakusten du eta 3.enak, berriz *treebankean* erregelak zenbat aldiz markatu diren. *Treebankari* dagokion zutabea bi ataletan banatu dugu: a) alde batetik, literatura-testuak (*XX. mendeko euskararen corpus estatistikoa* osatzen dutenak) bereiztu ditugu, eta bestetik, b) kazetaritza-testuak. Bereizketa hau testu motak aditzen erabileran garrantzirik ote duen aztertzeko

```

Erreg_ADI_ados_etorri
(
    Detekt (
        @.kat == 'ADI' &
        @.lema == 'ados etorri' &
        @.agerpenak >10
    )
)

```

VII.17 Irudia: *Treebank*ean “ados etorri” lemadun aditza agertzen ote den aztertzen duen erregela.

erabiliko dugu etorkizunean. Azken zutabeak (Proposatu?) aditu batek aditza hiztegian gehitzeko proposatzen ote duen eta proposamenerako arrazoia zehazten ditu.

Hizkuntzalariak corpusean soilik agertzen diren aditzak onartzeko edo baztertzeko, era askotako arrazoiak erabili dituzte:

- Aditz-hautagaia onartzeko arrazoiak (O), ondorengoak dira:
 - 1-O. Aditza corpusean topatu ondoren, lau hiztegitan bilatu da. Hiztegiok honakoak dira: a) *Euskalterm*, “Terminologia Banku Publikoa”, b) *Hiztegia 3000* c) *Elhuyar Hiztegia* eta d) *Hiztegi Batua*. Hautagaia den aditza hiztegiotatik gutxienez bitan agertzen bada, proposatu egiten da.
 - 2-O. Aditz-hautagaia ez da corpuseko forma berarekin agertu hiztegiko inolako sarreratan, baina bai antzeko forma batekin azpiserretako batean. Adibidez, ez dugu **kaltetu** aditza hiztegian topatu, baina bai esanahi bera duen **kalte egin** aditza **kalte** sarreraren azpian. Corpusean topatutako forma proposatzen da, aipatutako hiztegietan forma horrekin agertzen baita.
 - 3-O. Hautagaia den aditza hiztegian agertzen da baina ez da forma hobetsia. Adibidez, **planteatu** aditza EH hiztegian *esp.*²⁷ etiketarekin agertzen da eta bere ordezt **ezarri** proposatzen da. Gai-nontzeko hiztegietan, ordea, **planteatu** forma agertzen da hiztegi-sarrerara egoki moduan. Hori dela eta, hizkuntzalariak corpusean topatutako forma proposatu du hiztegi-sarreratzat.

²⁷esp.: espainolekoa soilik den hitz berria, edo espainolak eraginiko hitz edo adiera.

1. Aditza	2. Corpusa	3. Treebanka		4. Proposatu?	
<i>Hiztegia ez</i>	<i>Agerpenak</i>	<i>XX. mendekoan agerpenak</i>	<i>Kazetaritzakoan agerpenak</i>	<i>Proposatu?</i>	<i>Arrazoia</i>
baloratu	>50	1	2	Baztertu	1-B
blokeatu	>50	0	5	Onartu	1-O
diseinatu	>50	0	1	Onartu	1-O
exijitu	>50	1	2	Duda	D
inbertitu	>50	0	4	Onartu	1-O
hitzartu	>50	0	7	Onartu	2-O
hitzeman	>50	0	2	Onartu	2-O
kaltetu	>50	0	1	Onartu	2-O
justifikatu	>50	0	1	Onartu	1-O
planteatu	>50	2	3	Onartu	3-O
menperatu	>50	3	0	Baztertu	2-B
afiliatu	>10	0	1	Onartu	1-O
berdintsu izan	>10	0	1	Duda	D
deskubritu	>10	4	0	Baztertu	1-B
erlazionatu	>10	1	0	Onartu	1-O
errebindikatu	>10	0	2	Onartu	1-O
errekurritu	>10	0	3	Duda	D
finantzatu	>10	0	6	Onartu	1-O
kargugabetu	>10	0	1	Onartu	1-O
kartzelaratu	>10	1	0	Onartu	1-O
kolaboratu	>10	0	1	Onartu	1-O
komentatu	>10	1	0	Duda	D
konplikatu	>10	1	0	Onartu	1-O
lanpetu	>10	1	0	Baztertu	3-B
ingresatu	>10	0	2	Baztertu	1-B
inkomunikatu	>10	0	6	Onartu	1-O
inkulpatu	>10	0	1	Baztertu	4-B
inspiratu	>10	1	0	Onartu	1-O
integratu	>10	1	1	Onartu	1-O
konprometitu	>10	0	1	Onartu	1-O
kotizatu	>10	0	1	Onartu	1-O
kriminalizatu	>10	0	1	Duda	D
merkaturatu	>10	0	5	Onartu	1-O
praktikatu	>10	1	0	Onartu	1-O
profitatu	>10	1	0	Onartu	1-O

VII.21 Taula: EH hiztegia gehitzeko hautagai diren aditzak.

- Aditz-hautagaia baztertzeko arrazoiak (B) hauek dira:
 - 1-B. Esanahi bereko beste forma bat hobesten da aztertutako hiztegietan.
 - 2-B. Hiztegiean ez da sarrera moduan agertzen, beste aditzen baten aldaera moduan baizik.
 - 3-B. Hiztegiean ez da aditz kategoriarekin agertzen. Adjektibo kategoriarekin agertzen da.
 - 4-B. Ez da inolako hiztegitan agertzen.
- Dudazko (D) aditzak nahiz eta corpusean agertu, ez dira aipatutako lau hiztegietan agertu.

Bigarren esperimenturako bi atalase-balio definitu ditugu. Lehenengo, aditza corpusean 50 alditan edo gehiagotan ager dadin eskatu dugu, gero atalase-balio hori 10 baliora jaitsi dugularik. VII.21 taulako datuak VII.22 taulan laburbildu ditugu. Taula horien azterketatik ondorio honetara irits gaitezke: corpusean agerpen asko edukitzeak ez du proposamenaren onarpena bermatzen. Datuetatik erauzi dezakegun informaziotik, adibidez, corpusean 50 alditan baino gehiagotan agertzen diren aditzetatik % 72,72 onartzen dela eta % 18,18 baztertzen dela ikus dezakegu. Bigarren multzoan (corpusean 10 agerpen baino gehiago) % 66,66 onartzen da eta % 16,16 baztertu. Bigarren multzoan, baztertutako aditzen proportzioa txikiagoa da, baina aditz gehiagorekin dago duda. Gure susmoa da, dudazko moduan markatutako aditzetako asko onartuko lirakeela onarpenerako baldintzak hain zorrotzak ez balira. Datuak ikuspuntu honi jarraituta analizatuz gero, bigarren multzoan etekin handiagoa ateratzen da.

	Aditzak	Onartua	Baztertua	Dudazkoa
Corpusean >50 aldiz + <i>treebank</i> ean	11	8 (% 72,72)	2 (% 18,18)	1 (% 9,09)
Corpusean >10 aldiz + <i>treebank</i> ean	24	16 (% 66,66)	4 (% 16,16)	4 (% 16,66)

VII.22 Taula: EH hiztegiean gehitzeko aditzen onartze-tasak.

Bi esperimentuetan proposatutako aditz-zerrendak erraz heda daitezke. Adibidez, inkoherentzien bila ari garela, corpusean eskatzen dugun agerpen kopurua jaitsi dezakegu (adibidez 50etik 25era) eta modu honetara inkoherentzia gehiago topatu. Gauza bera gertatzen da aditz-sarrerekin. Corpusean 50 edo 10 agerpen beharrean, 5 eskatuz gero, hiztegiean gehitzeko proposatutako aditz-zerrenda segur asko luzeagoa litzateke.

VII.9 Laburpena eta ondorioak

Lotura argia duten arren, arras desberdinak dira *Saroï* kapitulu honetan aurkitu dizkiogun aplikazioak. Hau horrela izanik, ondorio komunak ematea zaila suertatu zaigu, eta hauen inguruan esan beharrekoak bi ataletan banatu ditugu.

VII.9.1 Komunztadura-erroreen detekzioari dagokionez

Esaldiko funtzio gramatikalen eta aditzaren arteko komunztadura aztertze-ko, maiz esaldiko ordena linealean urruti kokatuta dauden elementuak begiratu behar izaten dira. *Saroï* tresnak ordena lineala utzi eta mendekotasun-zuhaitza aztertzen du. Beraz, egokia da bospasei hitz baino gehiagoko testuinguruetan hizkuntza-egiturak bilatzeko.

Saroï erabiliz komunztadura-erroreak detektatzeko erregelak idazten hasi aurretik, komunztadura eta berau aldatzen duten fenomenoak aztertu ditugu. Hori egin dugunean, atazaren zailtasunaz ohartu gara, eta zailtasun hau analizatzaile morfologikoak eta sintaktikoak areagotu egiten dutela egiaztatu dugu. Lana errazteko asmoarekin baliabide anitz erabili eta sortu ditugu. Adibidez, sintagmen arteko koordinazioa esplizitu egin dugu eta haren ezaugarri linguistikoak aldatu (pluraltasuna adierazten duen marka, adibidez). Horretaz gain, aditz laguntzaileen deskribapena osatu egin dugu eta aditz nagusia aberastu. Azken atazarako, informazio-iturritzat eskuz aztertutako azpikategorizazio-ereduak, corpusetatik jasotako gauzatze-egiturak eta *Euskal Hiztegitik* jasotako aditz-deskribapenak erabili ditugu. Hainbat dokumentazioren irakurketaren, eta batez ere, corpus erroredunaren azterketaren ondorioz, komunztadura-erroreen detekziorako gramatika bat osatu dugu.

Gramatika hasiera batean EDGK baliatuz sortutako mendekotasun-zuhaitzetan probatu nahi izan dugu. EDGK-k sortutako analisi-zuhaitzetan anbigutasun morfologikoa eta sintaktikoa geratzen dira, eta nahiz eta *Saroï* horiek kudeatzeko gai den, erroreen detekziorako desanbiguazio-konbinaziorik egokiena zein den ezagutu nahi izan dugu. Gramatika zein desanbiguazio-konbinazioarekin analizatutako zuhaitz-basoetan aplikatu behar den erabakitzea, ez da berehalakoa izan. Erabakia urratsez-urratseko metodologia bati jarraituz hartu dugu. Lehen urratsean, analisi morfosintaktikoaren desanbiguaziorako aukeren eta funtzio sintaktikoen desanbiguaziorako aukeren arteko konbinazioak egin ditugu, eta horien artean emaitzarik onenak lortzen dituzten bi konbinazioak aukeratu ditugu. Horietan, desanbiguazio morfologikorako mailarik egokiena aukeratu da, baina ez funtzio sintakti-

koen desanbiguaziorako egokiena den aukera. Hori bigarren urrats batean egitera gindoazenean, mendekotasun-erlazioak ipintzen dituen gramatika *erlaxatzeko* beharraz konturatu gara. Gramatika erlaxatu ezean, zenbait errore ezin da detektatu. Lehen urratsen aukeratutako konbinazioak 75 esaldi erroredunetako corpus batean probatu ditugu, EDGK gramatika *erlaxatuta* lehenengo, eta *erlaxatu gabe* gero. Maiz, esaldiari dagozkion mendekotasun-zuhaitzak gaizki eratuta daudelako —mendekotasun-erlazioak gaizki ipintzen direlako— detektatzen dira erroreak. Fenomeno hau baieztatzearren, esaldi zuzenekin egin dugu proba. Emaitzek gure susmoa baieztatu dute, eta gainera, azken probak egiteko erabiliko dugun desanbiguazio-konbinaziorik egokienaren berri eman digute: 3-3-Erlaxatua. Hots, EDGK dependentzia-gramatika erlaxatua erabiltzen duena, desanbiguazio morfologikoa 3. mailan egiten duena eta funtzio sintaktiko nagusiak desanbiguatzen dituen.

Analisi-zuhaitzak “ez egokiak” direnez, zaila da komunztadura-erroreen detekziorako sortutako erregelen doitasuna neurtzea. Maiz, erregelek, analisiaren beraren akatsak azalarazten dituzte, erroreak agerrarazi beharrean. Analisi sintaktikoa hobetzen den heinean, emaitzak hobetuko direla uste dugu. Baieztapen hau frogatzeko, alde batetik, EDGKz gain *MaltIxa* analizatzaile sintaktiko estatistikoa erabili dugu corpus erroredunetan proba masiboak egiteko. Eta bestetik, euskarazko *treebank*ean ere, analisi-zuhaitz “perfektuen” bankuan, egin ditugu probak.

EDGK-k zuhaitz partzialak lortzen ditu. *MaltIxa*k zuhaitz bakar bat zeinetan token guztiak (puntuazio-markak barne) lotuta dauden. EDGK-k sarreran anbiguotasuna onartzen du eta *MaltIxa*k sarrera guztiz desanbiguatua erabiltzen du (funtzio sintaktikorik gabekoa). Beraz, EDGK eta *MaltIxa* ez dira haien artean konparagarriak. Bi analizatzaileak corpus erroredun bera analizatzeko erabili ditugu, eta haiekin hainbat konbinazio egin dugu. Proba hauen arabera, *MaltIxa*rekin emaitza hobeak lortzen dira. Tresnak, ordea, sarrera guztiz desanbiguatua behar izatearen alde txarra du. Azken proban, bi tresnekin sortutako alarma faltsuek analisi-katearen akatsen berri eman digute (desanbiguazio morfologikoa, aditzen azpikategorizazioa...).

Esan bezala, komunztadura-erregelen doitasuna neurtzeko probak, euskarazko zuhaitz-bankuan egin ditugu. Hau egin dugunean, etiketatze-erabakien edo etiketatze-akatsen ondorioz sortutako mendekotasun-zuhaitz akas-dunak identifikatu ditugu. Bestalde, alarma faltsu “erreal” gutxi batzuk (% 3,305) detektatu ditugu, eta horien azterketa sakona egin dugu.

VII.9.2 Aditzen aberasketari dagokionez

Corpusak eta *treebanka* erabiliz hiztegiak aditz-informazioarekin aberasteko bi aukera aztertu ditugu. Esperimentuak honakoak topatzeko garatu dira: alde batetik, dagoeneko hiztegian agertzen den, baina testuetatik jasotakoarekin bat ez datorren aditz-informazioa jasotzeko; eta bestetik, corpusean eta *treebankean* agertzen diren eta hiztegian agertzen ez diren aditzak topatzeko. Lan honek hiztegiak aberasteko corpusen eta *treebanken* erabilera erakusten du.

Existitzen diren baliabideak berrerabiliz, corpusetik emaitzak lortzeko eta hiztegia aberasteko egin dugun eskuzko lana minimoa izan da. Ezaugarri hau interesgarria iruditzen zaigu, hiztegiak garatzeko egin behar izaten den lana oso gogorra eta denbora luzekoa izan ohi baita.

Hiztegia aditz-sarrerekin aberasteko garatutako esperimentuko VII.21 emaitza-taula aztertuz gero, corpuseko agerpen kopurua alde batera utzita, kazetaritzako testuekin osatutako *treebankean* 4 alditan baino gehiagotan agertzen diren aditz guztiak proposamen moduan onartzen direla ikus dezakegu. *Treebankaren* zati honek euskararen egungo erabilera eta hizkuntzaren zuzentasuna konbinatzen ditu. Aditzen onarpen-tasak argi uzten du informazio-iturri gisa *treebankek* duten baliagarritasuna.

Erabili dugun metodologia edozein hizkuntzarako orokorra dela uste dugu. Hala ere, eskakizun pare bat bete behar da: a) hizkuntza horretarako hizkuntza-baliabide egokiek eskuragarri egon behar dute, eta b) hizkuntza-informazioak *Saroï*rako zehaztutako sarrera-espezifikazioak bete behar ditu. Hau da, mendekotasun-zuhaitzak XML erabiliz errepresentatuta behar dira.

Gure ustez, atal honetan aditzentzat egin dugun lana erraz heda daiteke hiztegiko gainontzeko hitzetara. Esate baterako, izen, adjektibo, adberbio eta gainontzeko kategoriatako hiztegi-sarrerak, kategoriari, adibideei edo erabilera-eremuei buruzko informazioarekin aberas daitezke. Testu-bilduma aldatuz gero, domeinu zehaztutako hitzak (adibidez, kirola, politika, ekonomia...) gehi daitezke hiztegian. Era berean, euskalkien arabera sailkatutako testuak erabilita, informazio honekin aberastuko genuke hiztegia.

WordNet-eko informazioa *Saroï*ko aberasketa modulua erabilita analisisetan gehituz gero, hiztegiak corpusetatik erauzitako informazio semantikoarekin aberas daitezke. Etorkizunean, gai honen inguruan zenbait esperimentu garatzeko asmoa dugu.

ONDORIOAK ETA
ZABALDUTAKO
IKERLERROAK

VIII. KAPITULUA

Ondorioak eta etorkizuneko lanak

Tesi-proiektu honetan *euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideak garatzeko* lehen pausoak eman ditugu eta, aldi berean, errorearen tratamendu automatikoan oinarrizko azterketa egin dugu. Helburu hau bete dugula uste dugu, hainbat fenomeno linguistikoren erabilera okerrak detektatu baititugu. Honako hizkuntza-egituretan jarri dugu nagusi arreta: datetan, postposizio-lokuzioetan eta aditzaren eta subjektuaren, objektuaren eta zehar-objektuaren arteko komunztaduran. Zuzenketan ere lehen urratsak egin ditugu, gaizki erabilitako daten eta postposizio-lokuzioen ordain zuzenak emateko gai diren sistemak sortu baititugu. Jakin badakigu, hizkuntza-erabiltzaileok oker erabil ditzakegun egituren multzoa handia dela, eta hauen izaera askotarikoa. Tesi honekin errorearen tratamenduaren alorrean egin beharreko lana ez da amaitzen, hasi besterik ez da egin. Hala ere, euskaraz sintaxi-mailan egin daitezkeen erroreak automatikoki detektatzea eta zuzentzea bideragarria dela frogatu dugu, horretarako azterketa sakona eginez.

Sarrerako kapituluan definitutako helburuak lortzeko egin dugun bidean, hizkuntzaren tratamendu automatikoari (edo *lengoaia naturalaren prozesamenduari*) lotzen zaizkion hiru arazo nagusi uneoro izan ditugu bidelagun: *estaldura*, *anbiguasuna* eta *errepresentazio aberatsa*. Estaldura osoa lortzea ezinezkoa da hizkuntzaren analisi zein errorearen tratamendu automatikoan. Estaldura ahal izan dugun neurrian landu dugun arren, errorearen tratamenduan doitasunari eman diogu lehentasuna, tentuz ibili beharra baitago erabiltzaileek hizkuntzari buruz duten ezagutza dudatan ipin ez dezaten. Anbiguasun morfosintaktikoa eta sintaktikoa sakon aztertu ditugu, anbiguasun-maila ezberdinen erabilerak emaitzetan duen eragina aztertu baitugu.

Gainera, *Saroik* anbiguotasuna modu egokian lan dezan lortu dugu. Hizkuntzaren errepresentazio aberatsa IXA taldean definitutako XML bidezko anotazio-amarauna erabiliz ongi biltzen da, gure ustez.

Azken kapitulu honetan, proiektuan ateratako ondorioak zehaztuko ditugu VIII.1 atalean. Gero, VIII.2 atalean, egindako ekarpenak zerrendatu ditugu. Azken atala, VIII.3 atala, irekitako ikerlerroak eta lanean egin daitezkeen hobekuntzak azaltzeko erabiliko dugu.

VIII.1 Ondorio nagusiak

Erroreen tratamendu automatikoa ataza zaila da. Jamesen (1998) arabera “a covert error is superficially well-formed but does not match intentions: it is right by chance. . . the learner is not saying what he means, even though what he is saying is in impeccable English”. Aipu horrek dio, egitura bat gaizki erabilia dagoela ezagutzea zaila dela idazlearen intentzioa ezagutu gabe, azaleko forma zuzena izanik, *ezkutuko errore* baten aurrean egon baikeitezke. Erroreen detekzioarako abiapuntua, egitura bat erroreduna ote den ezagutzea, beraz, ez da argia batzuetan, ezta gizakiontzat ere. Errore hauek detektatzea zaila delarik, azaleko forman, testuan gure kasuan, aztarna uzten duten erroreak automatikoki detektatzeko lana hartu dugu. Horretarako, askotariko informazio linguistikoa (morfologikoa, sintaktikoa, semantikoa, . . .) aztertu eta integratu dugu erabilitako sistemetan. Ezagutza linguistikoa aberatsa eta tresna egokiak eskura izan arren, konplexua suertatu da, orokorrean, errore sintaktikoak detektatzea.

Erroreak automatikoki detektatzeko eta zuzentzeko egin dugun saiakerak honetan, lehenengo, eman beharreko pausoak zehaztu ditugu sarrerako kapitulan (I. kapitulan). Pauso horiek ematerakoan landu ditugun gaiak lotuta, ondorio batzuk atera ditugu. Orain, horiek zehazteko unea heldu da (ondorengo zerrendapenetan, urrats haiekin lotutako gaiak letra etzanez adierazi ditugu, eta jarraian, ateratako ondorioak idatzi ditugu). Ondorioek, tesi-lanaren garapenarekin eta ebaluazioarekin dute lotura batez ere, eta horrela multzokatu ditugu VIII.1.1 eta VIII.1.2 azpiataletan, hurrenez hurren.

VIII.1.1 Garapena

Erroreak detektatzeko eta zuzentzeko sistemen garapenarekin zerikusia duten ondorioen artean batzuek, analisi-katearekin eta anbiguotasunarekin dute zerikusia. Gainontzeko ondorioek, lotura zuzena dute detekzio-prozesuarekin:

- *Analisi-katea*: analisi-geruzetan zehar metatzen diren analisi-akatsek emaitzak “zikintzen” dituzte. Informazio ez-gramatikala analizatzeko erabili dugun analisi-kateak haren garapenean informazio gramatikala baliatu du. Ezaugarri horren garrantziaz ohartuta, analisi-katearen egokitasuna aztertu dugu IV. kapituluaren ondorengo kapituluetan (V. eta VII. kapituluetan) gramatikek sarreratzat jasotzen duten analisiaren garrantzia egiaztatu dugu lortutako emaitzak aztertu ditugunean. Emaitzetan eragin handia izan dute sarrerako analisisetako akatsek: besteak beste, oker esleitutako kategoria gramatikalek, gaizki ebatzitako anbiguitasunek, etiketatu gabeko mendekotasun-erlazioek, edo errepikatutako mendekotasun-erlazioek. Batzuetan, errorea analisi-akatsen erruz detektatu dela egiaztatu dugu, *detekzio faltsua* gertatu da. Arazo horren jakinaren gainean egon arren, erabaki bat hartu dugu: *erroreen detekziorako gramatika ez dugu sarrerako analisisetara eta haien akatsetara egokitu; hots, ez ditugu sarrerako akatsak gramatikara eramanez*. Horrek baditu bere abantailak: *Saroik* sarreran erabiltzen duen analizatzailea aldatuz gero (egin daiteke, eta egin dugu EDG-Kren ordez *MaltIta* (Bengoetxea eta Gojenola, 2007) analizatzailea erabiltzea), ez dago erroreen detekziorako gramatika aldatzeko beharrik. Beste aukerarik ere badago, ordea. Birn-en (2000) ustea da, erroreen detekziorako erregelak sarrerako tresnek egiten dituzten akatsetan oinarrituta definitu behar direla. Metodo enpirikoen bidez corpus erroredun analizatuetatik erroreen detekziorako erregelak ikasten direnean (ikasketa automatikoan, adibidez), hein batean hori bera egiten da: gramatika sarrera akasdunera egokitzen da. Azken aukera honetan, sarrerako datuak aldatu ahala, erroreen detekziorako gramatika ere aldatu beharko litzateke.
- *Anbiguitasuna*: erroreen detekzioan eragin bereziki handia du anbiguitasunak eta eragin hau landutako sintaxi-egitura bakoitzeko aztertu behar da. Tesi-lan honetan sakon aztertu dugu erroreen detekzioaren eta anbiguitasunaren arteko harremana, eta sintaxi-egitura bakoitzean harreman hori aldatzen dela egiaztatu dugu. Egitura mugatu eta zurrunetan, datetan, anbiguitasunak ez du eragin handiegirik erroreen detekzioan. Uste baino anbiguoagoa suertatu zaizkigun egituretan, postposizio-lokuzioetan, emaitza hobeak lortzen dira testuak desanbiguatzea gabe erabiltzen direnean desanbiguatzen direnean baino. Hori, besteak beste, desanbiguoazio-gramatikak postposizio-lokuzio zuzenei dagozkien egiturak landuta dituelako izan daiteke. Komunztadura-erroreen detekziorako, aurrekoaren kontrara, egiturak ahalik eta

gehien desanbiguatzeak emaitza hobeak ematen ditu, bai morfosintaxi- baita sintaxi-mailan ere. Gure ustez, azken fenomenoaren arrazoa linguistikoa baino gehiago prozedurazkoa izan daiteke. Desanbiguatutako gabeko egiturak erabilia analisi-zuhaitz gehiegi sortzen da, eta zaila da guztietan erreoreen detekzioarako patroia betetzea.

Ukaezina da erreoreen detekzioan anbiguitasunak duen eragina. Egingadako probetatik ondoriozta dezakegu, ez dagoela erreoreen detekzioari dagokionez egitura guztietarako balio duen desanbiguatutako-aukerarik. Aztertu nahi den egitura bakoitzak anbiguitasunaren azterketa propioa eskatzen du. Errore gramatikalek desanbiguatutakoan eragina dute eta anbiguitasunak erreoreen detekzioan eragina du. Birnek (2000) anbiguitasunaren inguruan ondoko baieztapena egiten du: desanbiguatutako, erreoreen detekzio egoki baterako aurrebaldintza da. Ez gaude ados baieztapen horrekin, gure emaitzek kontrakoa adierazten baitute.

- *LNP teknikak/tresnak*: hizkuntza-egitura desberdinak lantzeko, teknika ezberdinak erabili behar dira, errore mota bakoitzarentzako teknika/tresna bat egokiagoa baita. Hautatu ditugun hizkuntza-egiturak lantzeko egokiagoak iruditu zaizkigu teknika sinbolikoak (hizkuntzari buruzko informazioa erregeletan kodetzen dutenak) teknika enpirikoak (corpus handiak eta estatistika/ikasketa automatikoa erabiltzen dutenak) baino. Hori dela eta, tesi-lan honetan lehenak erabili ditugu. Horrek ez du teknika enpirikoen erabilera baztertzen, hitzen nahasketak, gabeziak edo eransketak lantzeko, adibidez, egokiagoak izan baitaitezke. Gainera, zuzenketa egiten denean, teknika enpirikoak oso aproposak dirudite ordain zuzenen artean aukeraketa egiteko. Gramatika-zuzentzaileei gainbegiratua eman diegunean, detektatutako erreoreen eta erabilitako tekniken artean harremana topatu dugu: horrela, sintagma-mailako erreoreak detektatzeko patroierregelak erabili ohi dira (MG, XFST formalismoekin, adibidez), esaldi-mailakoak lantzeko analisi-zuhaitzak (TGGak, mendekotasun-zuhaitzak...) eta hitz zuzenen erabilera nahasketak tratatzeko, teknika sinbolikoak.

Teknika sinbolikoek corpusetan oinarritutako tekniken aurrean hedagarritasunaren arazoa dute, baina era berean, doitasun handiagoa lortzen dute emaitzetan, iturri fidagarrietako (gramatikak, hiztegiak...) informazioa erabiltzen baitute. Gainera, teknika sinbolikoetan erregelek adituek eskuz definitzen dituztenez, modu egokian landu daiteke bereziki hizkuntzen irakaskuntzaren alorrean garrantzitsua den berrelikadura.

Testuinguru mugatuko erroreak detektatzeko eta zuzentzeko, dagoe-neko sortuta zeuden tresnak erabili ditugu: datetan gertatzen diren erroreak lantzeko XFST, eta postposizio-lokuzioetan gertatzen direnetarako, MG. Testuinguru zabalekoak lantzeko, berriz, mendekotasun-zuhaitzekin lan egiteko tresna egokirik topatu ez dugunez, guk sortu dugu *Saroi* tresna. Hots, errore mota ezberdinekin lan egiteko tresna ezberdinak erabili ditugu. Ez al legoke teknikarik/tresnarik errore mota guztien tratamendurako balioko lukeenik? Gure ustez errorearen tratamendu aproposenerako ez. Segur aski, teknika berarekin, demagun patroien erabilerarekin, errore multzo handi bat landu ahal izango genuke, baina ez egitura guztietarako modu egokienean. Adibidez, bibliografian irakurritakoaren arabera, hitz zuzenen nahasketa lantzeko egokiagoa da ikasketa automatikoa erabiltzea. Are gehiago, teknika beraren barruan ere, hizkuntza-egitura bakoitza lantzeko egokiagoa izaten da tresna edo hurbilpen bat gainontzekoak baino. Gogora dezagun, adibidez, erregeletan oinarritutako teknikak (sinbolikoak) erabili arren, komunztadura lantzeko analisi-zuhaitzak erabili ditugula, eta postposizio-lokuzioak lantzeko ez dela beharrezkoa izan. Praktikotasunari begira errore mota bakoitzerako tresna ezberdina erabiltzea askotan ez da bideragarria, eta adibidez, errorearen tratamenduan egingdako lana zuzentzaile edo OLHIA sistema batean integratu nahiko balitz, segur aski teknika bakarra aukeratu beharko litzateke.

- *Ezagutza linguistikoa*: errore sintaktikoak lantzeko, batzuetan ezagutza sintaktiko soila ez da nahikoa. Esaldi-mailako informazioaz gain, semantika, testuingurua, eta are gehiago, munduari eta erabiltzaileari buruzko ezagutza ere beharko genituzke. Aipatu berri dugun eta modu automatikoan lor daitekeen informazio ahalik eta aberatsena eskura edukita, errore mota batzuk detekta daitezke; beste batzuk, ordea, informazio horrekin guztiarekin ere, ezingo dira detektatu erabiltzailearen *intentzia* ezagutu gabe (gogora ditzagun *ezkutuko erroreak*).

Aurkeztutako lanean, datetako erroreak detektatzeko analisi morfo-sintaktiko hutsa erabili dugu. Postposizio-lokuzioetan gertatzen diren erroreak detektatzeko, berriz, hitzen tasun semantikoak behar izan ditugu (osagaiek tokia adierazten ote duten, bizidunak ote diren...), baita esapideei buruzko informazioa ere. Hots, hiztegieta informazioa eta haietatik automatikoki jasotako genusak baliatu ditugu.

Komunztadura-erroreen detektziorako analisi sintaktiko osoagoa erabili dugu (mendekotasun-zuhaitzak). Gainera, sintagmen arteko koordina-

zioa esplizitu egin dugu, aditz laguntzaileen komunztadura marketako ezaugarriak (kasua, numeroa, pertsona) azaleratu ditugu, eta aditzak azpikategorizatu ditugu informazioa hiru iturritatik jasota (eskuz, automatikoki corpusetik eta Euskal Hiztegitik).

Laburtuz, tratatu nahi den egitura bakoitzeko, zein hizkuntza-baliabide behar ditugun aztertu dugu eta hauek egokitu egin ditugu. Askotariko ezagutza linguistikoa eskura izatea oso garrantzitsua irizten dugu erroreak automatikoki detektatzeko eta zuzentzeko, honek ezagutzaren integrazioaren aldetik lana zaildu egiten duen arren.

VIII.1.2 Ebaluazioa

Azkenik, ondorioekin amaitzeko, ebaluazioarekin lotutakoak laburtu ditugu:

- *Corpus zuzenak vs corpus erroredunak*: erroreak detektatzeko, corpus zuzenak, corpus erroredunak bezain garrantzitsuak dira. Esana dugu, guretzat erroreak detektatzea bezain garrantzitsua dela zuzenak diren egiturak ez erroretzat jotzea (ez alarma faltsurik sortzea). Horregatik, gramatikak garatzeko eta ebaluatzeko corpus erroredunak erabiltzeaz gain, zuzenak ere erabiltzea ezinbestekoa da. Egindako esperimentuetan, azken horien erabilerak asko hobetu du garatutako gramatiken doitasuna baina, era berean, emaitzen batezbesteko balioak okertu egin ditu, gramatikak corpus zuzenetan aplikatzean errore gutxi eta alarma faltsu ugari markatzen baitira. Gure ustez, corpus erroredunak egokiak dira sistemak erroreak detektatzeko duen gaitasuna neurtzeko, eta corpus zuzenak, berriz, alarma faltsu kopurua neurtzeko. Datetako eta postposizio-lokuzioetako erroreekin *Euskaldunon Egunkaria* erabili dugu corpus zuzen moduan; komunztadura-erroreen kasuan, euskarazko *treebanka* (Aduriz *et al.*, 2003b).

Corpusak modu egokian bildu, kodetu eta etiketatu behar izan ditugu haietatik ahalik eta etekin handiena ateratzeko. Bilketa-prozesua testu arruntekin egiten dena baino neketsuagoa da, iturri espezializatuetara jo behar izaten baita (euskaltegietara, ikasleen testuetara e.a.).

- *Erroreak eta alarma faltsuak*: oreka zaila da. Erroreen eta alarma faltsuen detekzioaren arteko konpromisoa (*trade-off*) dago. Hots, errore gehiago detektatzen ditugunean, alarma faltsu gehiago azaltzen dira. Alarma faltsuen kopurua jaitsi nahi dugunean, detektatutako errore kopurua ere beherantz joan ohi da. Gramatikak errore instantzia asko

eta alarma faltsu gutxi detektatzeko fintzea lan gogorra izan da, eta horrek ezinbestean corpus zuzenak zein erroredunak behar izan ditu.

VIII.2 Ekarpenak

Ondorengo lerroetan, tesi-lan honek erroreen tratamendu automatikoari orokorrean, eta euskarazko erroreen analisiari zehazkiago, egindako ekarpenak laburbildu ditugu:

- Datetako, postposizio-lokuzioetako eta aditz-komunztaduraren inguruko erroreak automatikoki detektatu ditugu, eta emaitzak zehatz-mehatz erakutsi ditugu. Lehen bi egituretan, gainera, zuzenketa automatikoa egin dugu. Probarako corpusetan (erroredunetan eta arruntetan) aipatutako egiturekin lortutako emaitza onenak VIII.1 taulan laburtu ditugu.

	Estaldura	Doitasuna
Datak	% 92,1	% 89,7
Postposizio-lokuzioak		
<i>arte, aurre, bitarte, buruz, zehar</i>	-	% 53,75
<i>aurre, bitarte, buruz, zehar</i>	-	% 79,48
<i>bitarte, buruz, zehar</i>	-	% 90,62
Komunztadura-erroreak		
<i>MaltIxa</i>	% 46,66	% 38,88
<i>MaltIxa & EDGK</i>	% 17,77	% 44,44
<i>MaltIxa EDGK</i>	% 51,11	% 35,38

VIII.1 Taula: Emaitzak

Datekin emaitza onak lortu ditugu, % 90 inguruko estaldura eta doitasuna lortu baititugu. Egitura horiek zurrinak eta ez-anbiguoak dira. Landutako hizkuntza-egituraren konplexutasuna eta anbiguotasuna eta, beraz, erabili beharreko tresnen konplexutasuna handitzen joan den heinean, postposizio-lokuzioekin adibidez, emaitzek okerrera egin dutela ikus daiteke taulan. Emaitzetan analisi-katearen estaldura faltak, anbiguotasun morfologikoak zein semantikoak, eta geruzaz geruza metatutako analisi-akatsek eragin zuzen-zuzena izan dute. Hala eta guztiz ere, eta lanerako abiapuntua konplexua zela ikusirik (anbiguotasun handia eta analisi desegokiak) postposizio-lokuzioekin lortutako emaitzak onak direla esan dezakegu.

Komunztadura-erroreen kasuan, sistemaren itogunea, analisi-zuhaitzak eraikitzeke erabili ditugun analizatzaile sintaktikoetan egon da. EDGK

erregelatan oinarritutako analizatzaile sintaktikoak % 48ko doitasuna eta % 46ko estaldura ditu, analisi-katetik heredatutako anbiguotasun morfosintaktikoarekin eta sintaktikoaren egiten du lan, eta gainera, zuhaitz partzialak sortzen ditu. Horiek guztiak kudeatu behar ditu *Saroi* kontsultarako tresnak. *MaltIxa* analizatzaile sintaktiko estatistikoak sarrera ez-anbigua jasotzen du eta mendekotasun-zuhaitz osoak sortzen ditu (%76,76ko LAS balioa du). EDGK eta *MaltIxa* haien artean konparagarriak ez diren arren, biek mendekotasun-erlazioak ipintzerakoan akatsak egiten dituzte. Hori dela eta, zaila da komunztadura-erroreen detekzioarako erregelen doitasuna neurtzea. Gure sistemak maiz mendekotasun-zuhaitzetako akatsak azalarazten ditu, komunztadura-erroreak detektatu beharrean. *MaltIxa* eta EDGK erabili ditugu corpus erroredun bereko esaldiak analizatzeko. Bi sistemak esaldi berak modu berean analizatzearekin ados jarri direnean, % 44,44ko doitasuna lortu da baina estaldura baxua (% 17,77). Bi neurrien batez bestekoa kontuan hartuta, aukerarik onena analisirako *MaltIxa* erabiltzea dela dirudi, % 38,88ko doitasuna eta % 51,11ko estaldura lortzen baitira. Analisi-kateak eta analizatzaile sintaktikoek hobera egiten duten heinean, emaitzak hobetuko direla aurreikusten dugu.

Tesi-lanean egindako ekarpen garrantzitsua iruditzen zaigu, ebaluazioko emaitzak argi eta garbi erakustea. III. kapituluaren amaierako tauletan azaldu genuen moduan (ikus III.5, III.6 eta III.7 taulak) ez da oso ohikoa gramatika-zuzentzaileetan emaitzak ageriko egitea.

Zenbakiak alde batera utzita, tesi honetan datetako erroreen, postposizio-lokuzioetako erroreen eta komunztadura-erroreen deskribapen osoa eta zehatza egin dugu, topatutako adibideak kasuz kasu zerrendatuz.

- **Erroreen detekzioaren eta anbiguotasunaren arteko erlazioa sakon aztertut dugu.** Gainera, komunztadura-erroreen detekzioaren azterketa eraginkortasunez egin ahal izateko metodologia bat definitu dugu.
- ***Saroi* tresna garatu dugu.** *Saroi*k euskarazko *treebank*ean (Aduriz *et al.*, 2003b), baita automatikoki eratutako analisi-zuhaitzetan ere, informazio linguistikoa (egiturak zein ezaugarriak) bilatzeko aukera ematen du. Tesi-txosten honetan azaldutako lanean *komunztadura-erroreak* detektatzeko, eta corpusetan eta Euskal Hiztegian aditzen azpikategorizazioaren inguruan dauden inkoherentziak topatzeko erabili dugu. Tresnaren aplikazio gehiago ere bururatu zaizkigu eta etorkizunean egin daitezkeen lanetan zehaztu ditugu (ikus hurrengo atala). Tresna erabilerraza eta malgua da, eta maila desberdinetako (mor-

fologia, morfosintaxia, sintaxia...) hizkuntza-ezaugarriekin lan egiteko gai da. Kontsultak mendekotasun-zuhaitzen gainean egiten ditu, eta ez XPathek edo XSLTk egiten duten moduan dokumentu-egituren gainean (Bouma eta Kloosterman, 2002; Cassidy, 2002); beraz, hizkuntzalarientzako lan-tresna egokia eta erabilgarria izan daitekeela iruditzen zaigu. Gainera, tresna sendoa da, anotazio-amaraunean definitutako formatua mantentzen den bitartean, analizatzaile sintaktikoa alda baitezakegu (*MaltIta* eta EDGK erabili ditugu).

- *Euskal Hiztegia* aditzen informazioarekin aberasteko proposamena egin dugu (Díaz de Ilarraza *et al.*, 2007a). Alde batetik, hiztegian eta aukeratutako corpus batean agertzen diren aditzak jaso ditugu, eta haien arteko azpikategorizazio-inkoherentziak topatu ditugu. Bestetik, corpusean eta *treebank*ean agertzen diren zenbait aditz, oraindik EHN agertzen ez direnez, hiztegian txerta daitezen proposatu dugu. Lan honek, hiztegiak aberasteko corpusen eta *treebank*en erabilera erakusten du. Baliabideak berrerabiliz, eta erregela gutxi definituz, aditz multzo handiak automatikoki aztertu ditugu (5264 aditz desberdin erabili ditugu; 2541 corpusetatik jaso ditugu eta 4016 hiztegitik).
- Erroreak detektatzerakoan egitura zuzenak landu behar direla ondorioztatu dugu. Erroreen detekziorako erregelak garatzeko, hizkuntza-arauei buruzko ikuspegia aldatu egin behar da. Hori izaten da lan honetan laguntza eman duten hizkuntzalari baten ustez, erroreen tratamenduan egin beharreko gauzarik zailenetarikoa: egitura linguistikoa bat, demagun postposizio-lokuzioa, ispiluaren beste aldetik begiratzea. Maiz entzun ditugu modu honetako iruzkinak: “*sailkapenean aurrean geratu da* perpaus zuzenean ez dago postposizio-lokuziorik. Ez zenuke landu behar”. Zaila egiten da batzuetan ulertzea, egitura zehatz bat landu nahi denean (adibidean *-n aurrean*, *-ren aurrean* ordeztuz), ez dagoe-la modurik jakiteko postposizio-lokuzioa den ala ez. Ondorioz, egitura zuzena izanik, kasu honetan postposizioa ez bada ere, erroretzat ematea ekidin behar da.
- Erroreen tratamenduak LNPko hainbat alorri informazioa eskain diezaiokeela ondorioztatu dugu. Gizakiok munduari buruz dugun ezagutzak eragin zuzena du egiten dugun hizkuntzaren erabileran, eta oraindik ere, ezagutza abstraktu hori guztia ez dugu lortu formalizatzea. Gizakiok, egituraren bat zuzena ala okerra den jakiten dugu, nahiz eta arrazoia ematen ez jakin (“horrela delako” esaten dugu). Adibidez, bi hitzek pertsona multzoa adierazita ere, zergatik da okerra **gu*

artean¹ eta zuzena **jende artean**? Segur aski, zentzu desberdina dutelako. Hitzen semantika beharrezkoa dugu, baita ezagutza handiagoa ere. Semantikaren alorrak errorean tratamenduan mesede handia egin lezake, baina, gure ustez, errorean tratamenduak ere ekarpenak egin diezazkioke semantikaren alorrari. Erroreak detektatzeko egin beharreko hausnarketek, maiz kontzeptu multzoak beren semantikaren arabera multzokatzera bideratu gaitu. Adibidez, tokiek *aurreko* eta *atzeko* aldeak dituztenez, **etxean aurrean sukaldea dago eta atzean gelak** perpausa zuzena litzateke. Tokiak soilik kontuan hartu ditugunez **sailkapenean aurrean geratu da** egitura zuzena markatu egin du sistemak. Alarma faltsu horri esker **sailkapen**, **zerrenda** eta antzeko kontzeptuek tokia adierazten duten kontzeptuekin nolabaiteko lotura semantikoa dutela ezagutu dugu. Erroreen tratamenduak gainontzeko hizkuntza-alorrak aberats ditzake, hausnarketaren bidez “suposatzen” dugun baina espizitu egiten ez dugun informazioa azalarazten laguntzen digulako.

- **Erroreen deskribapena zehaztu dugu.** Hizkuntzalaritza deskribatzaileraren alorrak errorean detekzioetik eta diagnostiketik etekinak atera ditzakeela uste dugu, sintaxi-egituren erabilera zuzenen eta okerren deskribapenak hizkuntza-egiturak hobeto ulertzen lagun lezakeen aldetik. Egitura sintaktikoen deskribapenak maiz informazio sintaktiko soila erabiliz egiten dira eta, ikusi dugun moduan, batzuetan hori ez da nahikoa. Zenbait egitura okerrekin zuzentzat jotako patroiak betetzen dituzte. Izakiok berezkoa dugun metainformazioak isla izan behar du hizkuntza-egituren deskribapenetan fenomenook zehaztasun osoz deskribatu ahal izateko. Horrela, gramatika-liburuetan egiten diren deskribapenak osoak eta onargarriak izanik ere, testuetan agertzen diren egituren ebaluazio enpirikoak eginez, deskribapenak argitu egiten dira, salbuespenak detektatzen dira eta murriztapenak zorrozten dira.

Bukatzeko, eta errorean detekzioarekin eta zuzenketarekin lotura zuzenik ez badu ere, honako ekarpena egin dugu:

- **Euskarazko *treebank*ean etiketatze-akatsak topatu ditugu.** *Treebank*a komunztadura-erroreak ebaluatzeko erabili dugunean, hainbat errore topatu ditugu. Batzuetan etiketatze-erroreak izan dira (hizkuntzalari-etiketatzailerak subjektua eta objektua nahastu egin dituzte, adibidez); beste batzuetan, etiketatze-prozesuan hartutako erabakien ondorioa (elipsia etiketatzeko moduagatik, adibidez). Kasu urri batzuetan,

¹Kasu honetan kategoria gramatikalak ez du garrantzirik, gure ustez.

benetako erroreak ere topatu ditugu. *Treebanka* automatikoki *txukundu* egin dugula esan dezakegu.

VIII.3 Etorkizuneko lanak

Jarraian aipatuko ditugun hiru gaiak jorratzea interesgarria iruditzen zaigu tesi-lan honetan hasitako bidean aurrera jo nahi bada: euskarazko erroreen tratamendua, *Saroiren* moduluen hobekuntza eta *Saroiren* erabilerak. Ondorengo hiru azpiatalek hurrenez hurren lantzen dituzte gai horiek.

VIII.3.1 Erroreen tratamendua

Tesi-lana amaitutzat jo ondoren, oraindik ere lantzeke geratzen den egitura erroredunen multzoa handia da. Hizkuntza-fenomenoak tratatzen jarraitu nahi dugu, eta horretarako, fenomeno bakoitzerako teknika eta baliabide egokienak bilatu beharko ditugu. Epe ertainean honakoak egin daitezkeela uste dugu:

- Gure ustez, guztiz bideragarria da erroreak detektatzeko sortu ditugun sistemak, hiru errore motak detektatuko lituzkeen sistema bakar batean konbinatzea. Horretarako, analisi morfosintaktikoa egiten denean datetako eta postposizio-lokuzioetako erroreak detektatuko eta zuzenduko dira XFST eta MG direla medio, eta gero ezarriko dira mendekotasun-erlazioak, komunztadura lantzeko. Testuinguru mugatuko erroreen detekzioa mendekotasun-etiketak ipini aurretik egin behar da, etiketak ongi ipintzeko aukerak handiagoak izan daitezen.
- Testuinguru mugatuko erroreen zuzenketa automatikoa diseinatu eta inplementatu dugu baina egitekoen artean laga dugu zuzenketarako sistema horien ebaluazioa. Hurrengo urratsa, testuinguru zabaleko erroreen zuzenketa lantzea da, eta gure kasuan, *Saroi* erabiliko genuke horretarako (ikus hurrengo atala, VIII.3.2 atala).
- Datekin eta komunztadura-erroreekin egindako lana XUXENG gramatika-zuzentzailean integratu nahi dugu (postposizio-lokuzioen detekzioa integratuta dago). Egokitzen fase baten ondoren, integrazio lan bera egin nahi dugu IXA taldean euskararen irakaskuntzarako garatu ditugun tresnetan (Aldabe *et al.*, 2005d, 2006) eta ariketen sorkuntzarako ArikIturri sisteman, adibidez (Aldabe *et al.*, 2006).

VIII.3.2 *Saroiren* moduluen hobekuntza

Saroi tresna sendoa eta erabilgarria dela uste dugun arren, beti egin dezake hobera. Atal honetan, *Saroi* osatzen duten moduluetan egin daitezkeen hobekuntzak zerrendatuko ditugu.

- **Analisi sintaktikorako modula**
 - Une honetan *Saroiren* sintaxi-analisirako moduluak sarrera anbigua *lehertu* egiten du eta sarrera horri dagozkion zuhaitz posible guztiak ematen ditu. Hori beharrezkoa da erroreak detektatzeko, sarrerako esaldia okerra denean ez baitakigu zein zuhaitzetan topatuko dugun gaizki eratutako egitura. Zenbait aplikaziotarako, edo azkartasuna helburu dugunean, ordea, *analisi-zuhaitz bakarra* nahikoa izango da. Esaldiko zuhaitz bakarra lortzeko, anbigutasun morfologikoa eta sintaktikoa aldeztu aurretik, analisiak *Saroiren* analisirako modulura iritsi aurretik ebatzi beharko lirateke. Adibidez, sarrera guztiz desanbiguatua hartzen dute (Nivre eta Scholz, 2004; Nivre *et al.*, 2007; Bengoetxea eta Gojenola, 2007) lanetan eta sarrera horri algoritmo deterministak aplikatzen dizkiote. Sarrerako analisiak automatikoki sortuak direnean anbiguoak izan ohi direnez, sortutako zuhaitzen arteko *desanbiguazioa* egin beharko litzateke. Analizatutako testuak zuzenak direnean, desanbiguazioa egiteko erregelak *treebank*etik ikasi ahal lirateke. Analizatzaile sintaktiko estatistikoek aspalditik erabiltzen dituzte desanbiguaziorako eredu estatistikoak *analisi onena* lortzeko (Charniak, 2000; Collins, 1997).
 - Zuhaitz-egituran beheko mailako adabegietatik goragoko mailetara informazioa goratzeko aukera gehitzea interesgarria litzateke. Adibidez, komunztadura-erroreen detekziorako erregeletan komunztatzen duten elementuek zuhaitz-egituretan kokaleku desberdina dutela ikusi dugu. Hala, ***zentral nuklearrak zakarra eratzen dute** esaldian subjektuaren modifikatzailean bilatzen dugu kasua (*@!ncsubj!ncmod.kas*), eta ***zentralak zakarra eratzen dute** esaldian subjektuan bertan bilatuko genuke (*@!ncsubj.kas*). Adabegietan zehar informazioa goratuz gero, bi erregelen ordez, bakarra (azkena) erabiliko genuke bi egituretan errorea detektatzeko.
 - Analisiak informazio semantikoarekin (sinonimoekin, hiperonimoekin, hiponimoekin...) aberats daitezke, edo entitate izendunei buruzko informazioarekin. Horrela, *Saroi* informazioaren

erauzketarako sistemetan, galdera-erantzun sistemetan eta itzulpen automatikoko sistemetan erabil liteke.

- Komunztadura-erroreen detekziorako sistema XUXENG gramatika-zuzentzailean integratu nahi bada, segur aski, anotazio-amarunaren eta *LibiXaML* liburutegiaren erabilera baztertu beharko da. Eraginkortasunaren truke, XML bidezko hizkuntza-informazioaren errepresentazio konplexuari uko egin beharko diogu. Nahikoa izango da *Saroiko* klase-diagrametan *LibiXaML* erabiltzen duten eragiketak identifikatzea eta aldatzea.

- **Kontsulta-erregelen itzulpenerako modulua.**

- Erregeletan *ordeztu* eta *info* atalak inplementatu behar dira. *Ordeztu* atala inplementatuz gero, sorkuntza morfologikoa dela medio zuhaitzetan ezaugarri morfosintaktikoak aldatzeko eta, nahi izanez gero, hasierakoaren aldaerak liratekeen esaldiak sortzeko aukera genuke. Erroreen detekziora ekarrita, erroreen zuzenketa egingo genuke (ikus VI. kapituluko VI.2.2.3 atala).
- Kontsulta-lengoiaren adierazgarritasuna hobetzearen, *Saroiko* eragile multzoa eragile gehiagorekin zabaltzea interesgarria litza-teke. Adibidez, ez barne-hartze eragilearen edo “*” eta “+” eragileen inplementazioarekin (ikus eragile hauen deskribapena VI. kapituluan). Hala ere, esan, tesi-lan honetan beharrezkoak izan ditugun patroiak definitzeko, dagoeneko inplementatuta dagoen eragile multzoa nahikoa izan dugula.

- **Kontsulta-erregelen aplikaziorako modulua.**

- Hainbat ideia azaldu ditugu VI. kapituluan aplikazio-estrategien inguruan eta erregelak mendekotasun-zuhaitzetan betetzen ote direneko erabakiaren inguruan. Une honetan, estrategia eta erabaki bakarra dugu inplementatuta. Etorkizunean, eta *Saroik* izango duen erabilerak horrela eskatzen badu, estrategia eta erabaki gehiago inplementatu beharko dira.

- **Emaitzen aurkezpenarako modulua.**

- Diseinatu dugun interfazearen inplementazioa eginez gero, *Saroi*, hizkuntzalaritzari buruzko ezagutza minimo batzuk dituen edozein ordenagailu-erabiltzailek erabili ahal izango luke.

VIII.3.3 *Saroiren erabilerak*

Saroi, bere osotasunean edo aukeratutako moduluak, aplikazio hauetan integra daitezkeela uste dugu:

- **Galdera-erantzun sistemetan eta itzulpen automatikoan.** Azken urteotan zuhaitzen konparaketa (*tree matching*) erabiltzen hasiak dira galdera-erantzun sistemetan (Punyakanok *et al.*, 2004) eta itzulpen automatikoan (Eisner, 2003). Galdera-erantzun sistemei dagokienez, Tanev *et al.* (2006) egileen arabera TREC² 2004 konferentzian emaitzarik hoberenetarikoa, galderaren egitura sintaktikoa eta erantzun posibleen egitura sintaktikoa konparatzen dituzten sistemek lortzen dituzte. Zenbait galdera moten erantzun egokiena topatzeko, erantzuneko eta galderako elementuek era berean lotuta agertu behar dute dagozkien mendekotasun-zuhaitzetan (Punyakanok *et al.*, 2004). Beste aldetik, itzulpen automatikorako sistema estatistikoak elkarren itzulpena diren esaldi pareekin entrenatzen dira. Itzulpen hauek ez dira beti *literalak* izaten eta Eisner-ek (2003) zuhaitzen egituraren dauden *desberdintasunetatik* ikasteko eredu bat proposatzen du. Itzulpena bera egiteko baino, ebaluaziorako ere erabiltzen da zuhaitzen konparaketa. Itzultzaile automatikoen emaitzen kalitatea ebaluatzeko metodo bat aurkezten da (Owczarzak *et al.*, 2007) lanean. Haien arabera, ebaluazioa egiteko zenbait neurrik (BLEU, NIST e.a.) itzulpenaren ondorio den esaldia eta erreferentziazko esaldia karaktere-kate mailan konparatzen dituztenez, zuzenak izan daitezkeen sintaxi-aldaerak zigortu egiten dituzte. Haiek, konparaketa dependentzia-zuhaitzen mailan egitea proposatzen dute. *Saroik* ez du zuhaitzak konparatzeko aukerarik oraindik ematen, baina proposamen bat egingo dugu helburu bera-ekin erabilia izan dadin: esaldietako bat, galdera-erantzun sistemetan, galdera adibidez, *Saroiko analisi sintaktikorako modulua* erabiliz analizatuko dugu, galderari dagokion mendekotasun-zuhaitza XMLz era *esplizituan* errepresentatua lortuaz. Hemendik abiatuta eta *XSLT* erabilita, XML dokumentuan deskribatzen den egituraren isla duen *Saroiko* kontsulta-lengoaian idatzitako erregela bat lortuko dugu. Azkenik, kontsulta-erregela erantzun posibleen mendekotasun-zuhaitzetan aplikatuko dugu definitutako egiturari hobekien egokitzen zaion zuhaitzaren —erantzunaren— bila. Itzulpen automatikoan metodo bera erabil dezakegu itzulpena egiteko; ebaluaziorako, erreferentziazko

² *Text REtrieval Conference*, <http://trec.nist.gov>

zuhaitza sortuta egongo litzateke eta bertatik aterako genuke kontsulta-erregela.

- **Sintaktikoki etiketatutako corpusen (*treebank*) kontsultan.** Zuhaitz-bankuko hizkuntza-informazioa kontsultatzeko balia liteke *Saroi*, VI. kapituluan erakutsi ditugun moduko galderak exekutatzuz. Adibidez, *Abar-Hitz* mendekotasun-erlazioak eskuz ezartzeko tresna grafikoan (Díaz de Ilarraza *et al.*, 2004) integra liteke, hizkuntzalariek etiketatu nahi duten egituraren antzekoak *treebankean* lehenago nola etiketatu dituzten azter dezaten.
- **Irakaskuntzarako sistemetan.** Ingurune hauetan erabil liteke *Saroi*, kontsulten bidez hizkuntzaren egiturak ezagutzen laguntzeko. Horretaz gain, aukera anitzeko galderak sortzeko ere baliagarria izan liteke *Saroi*. Corpusetan esaldi erroredunak detektatuko genituzke *Saroi* erabilita (komuntadura-erroreak dituztenak, adibidez) eta tresna berarekin zuzenketa posibleak sortuko genituzke. Horrela, ikasleak lau aukeren artean esaldi erroreduna haren ordain zuzenetatik bereizi beharko luke, adibidez.

Bibliografia

- Abeillé A. Introduction. In Abeilleé A., editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, 2003.
- Aduriz I. *EUSMG: Morfoloġiatik sintaxira Murritzapen Gramatika erabiliz. Euskararen desanbiguazio morfoloġikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak*. Doktoretza-tesia, Filologia eta Historia-Geografia Fakultatea. UPV-EHU, Gasteiz, 2000.
- Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., eta Urizar R. A framework for the automatic processing of Basque. *Proceedings of the Workshop on Lexical Resources for Minority Languages*, Granada, Spain, 1998.
- Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar M., Sarasola K., eta Urkia M. A word-grammar based morphological analyzer for agglutinative languages. *COLING, 18th International Conference on Computational Linguistics*, 1–7, Universität des Saarlandes, Saarbrücken, Germany, 2000. Morgan Kaufmann.
- Aduriz I., Aldasoro E., Aldezabal I., Aranzabe M., Arriola J.M., Ceberio K., Estarrona A., Iruskietia M., Lersundi M., Pociello E., Uria L., eta Urizar R. Euskarazko postposizio-lokuzioen tratamendu konputazionala. Barne-txostena, IXA Taldea. Lengoia eta Sistema Informatikoak Saila. Euskal Herriko Unibertsitatea., 2008.
- Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., eta Sarasola K. Construcción de un corpus etiquetado sintácticamente para el Euskera. *Revista de la Asociación Española para el Procesamiento del Lenguaje Natural*, 2002a.

- Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., Sarasola K., eta Urizar R. The design of a digital resource to store the knowledge of linguistic errors. *DRH2002 (Digital Resources for the Humanities)*, Edimburgh, 2002b.
- Aduriz I., Aranzabe M., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M., eta Uria L. A cascaded syntactic analyser for Basque. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing: 5th International Conference CICLing2004, Seoul, Korea, February 15-21*, 2945 lib. of *Lecture Notes in Computer Science*, 124–134. Springer-Verlag GmbH, 2004. ISBN 3-540-21006-7.
- Aduriz I., Aranzabe M., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Proceedings of Corpus Linguists 2003*, Lancaster, UK, 2003a.
- Aduriz I., Aranzabe M., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing. In Andrew Wilson P.R. eta Archer D., editors, *Corpus Linguistics Around the World.*, 56 lib. of *Book series: Language and Computers*, 1–15, Netherlands, 2006. Rodopi.
- Aduriz I., Aranzabe M., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Garmendia A., eta Oronoz M. Construction of a Basque dependency treebank. *TLT 2003. Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden, November 14-15 2003b.
- Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., eta Maritxalar M. Morphosyntactic disambiguation for Basque based on the Constraint Grammar formalism. *Proceedings of Recent Advances on NLP (RANLP)*, 282–288, Tzigov Chark, Bulgary, 1997.
- Aduriz I. eta Díaz de Ilarraza A. Morphosyntactic disambiguation and shallow parsing in computational processing of basque. *Inquiries into the lexicon-syntax relations in Basque*. University of the Basque Country, 2003.

- Agirre E. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala*. Doktoretza-tesia, University of the Basque Country, Donostia, 1999.
- Agirre E., Aldezabal I., eta Pociello E. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordNet. *GOGOA aldizkaria*, 237–266, 2006.
- Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Urkia M., Maritxalar M., eta Sarasola K. XUXEN: A spelling checker/corrector for Basque based on two-level morphology. *Proceedings of ANLP'92*, 119–125, Povo Trento, 1992.
- Agirre E., Gojenola K., Sarasola K., eta Voutilainen A. Towards a single proposal in spelling correction. In Boitet C. eta Whitelock P., editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, 22–28, San Francisco, California, 1998a. Morgan Kaufmann Publishers.
- Agirre E., Gojenola K., Sarasola K., eta Voutilainen A. Towards a single proposal in spelling correction. Barne-txostena, University of the Basque Country (UPV/EHU), 1998b.
- Agirre E. eta Rigau G. Word sense disambiguation using conceptual density. *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
- Aho A.V., Sethi R., eta Ullman J.D. *Compilers: Principles, Techniques, and Tools*. Reading, Mass.: Addison-Wesley, 1985.
- Aït-Mokhtar S. eta Chanod J.P. Incremental finite-state parsing. *Proceedings of the fifth conference on Applied Natural Language Processing*, 72–79, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
- Alberdi X. eta Sarasola I. *Euskal estilo libururantz. Gramatika, estiloa eta hiztegia*. Euskal Herriko Unibertsitatearen argitalpen zerbitzua, 2001.
- Aldabe I., Aldezabal I., Aranzabe M., Arrieta B., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Oronoz M., Otegi A., eta Uria L. Euskarazko erroreen sailkapena ERROREAK eta DESBIDERATZEAk datu-baseetan. Barne-txostena UPV/EHU / LSI / TR 27-2005, University of the Basque Country, Informatika Fakultatea, Donostia, 2005a.

- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Niebla I., Oronoz M., eta Uria L. The use of NLP tools for Basque in a multiple user CALL environment and its feedback. *TAL & ALAO workshop. Proceedings of the 13th Conference Sur Le Traitement Automatique des Langues Naturelles*, 2 lib., 815–824, Leuven, Belgium, 2006.
- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Oronoz M., eta Uria L. Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, 10(2):47–60, 2005b.
- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Oronoz M., Uria L., eta Amoros L. IRAKAZI: a web-based system to assess the learning process of Basque language learners. *Proceedings of EuroCALL 2005*, Cracow, Poland, 2005c.
- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Oronoz M., Uria L., eta Amoros L. Learner and error corpora based computational systems. In Walinski J. K.K..S.G.R., editor, *PALC 2005: Practical Applications in Language and Computers*. Frankfurt am Main: Peter Lang, 2005d.
- Aldezabal I. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoa baliatuz*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, 2004.
- Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: a general lexical basis for the automatic processing of Basque. *IRCS Workshop on linguistic databases.*, Philadelphia, USA, 2001.
- Aldezabal I., Aranzabe M., Arrieta B., Maritxalar M., eta Oronoz M. Toward a punctuation checker for Basque. *ATALA Workshop of Punctuation*, 2003a.
- Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Oronoz M., eta Sarasola K. Application of finite-state transducers to the acquisition of verb subcategorization information. *Natural Language Engineering*, 9(01):39–48, March 2003b.
- Aldezabal I., Gojenola K., eta Oronoz M. Combining chart-parsing and finite state parsing. *Proceedings of the European Summer School in Logic, Language and Information (ESSLLI) Student Session*, Utrecht, The Netherlands, August 16-20 1999.

- Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktoretza-tesia, Informatika Fakultatea. UPV-EHU, uztaila 1995. Kepa Sarasola eta Xabier Artola, UPV-EHUko irakasleen zuzendaritzapean eginiko tesia.
- Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., eta Urizar R. Representation and treatment of multiword expressions in Basque. In Tanaka T., Villavicencio A., Bond F., eta Korhonen A., editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, 48–55, 2004.
- Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., eta Urizar R. Robustez y flexibilidad de un lematizador/etiquetador. *VIII Simposio Internacional de Comunicación Social*, Santiago de Cuba, 20-24 Enero 2003a.
- Alegria I., Arrieta B., Díaz de Ilarraza A., Izagirre E., eta Maritxalar M. Using machine learning techniques to build a comma checker for Basque. *Proceedings of Coling-ACL*, 1–8, Sydney. Australia, 2006.
- Alegria I., Balza I., Ezeiza N., Fernández I., eta Urizar R. Named entity recognition and classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI*, Madrid, Spain, 2003b.
- Alegria I., Ceberio K., Ezeiza N., Soroa A., eta Hernandez G. Spelling correction: from two-level morphology to open source. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, 2008.
- Alexopoulou A. Aproximación al tratamiento del error en la clase de E/LE desde la perspectiva del análisis de errores. *Estudios de Lingüística Aplicada*, 23(041):101–125, Julio 2005. Universidad Nacional Autónoma de México.
- Ansa O., Arregi X., Arrieta B., Ezeiza N., Fernandez I., Garmendia A., Gojenola K., Laskurain B., Martínez E., Oronoz M., Otegi A., Sarasola K., eta Uria L. Integrating NLP tools for Basque in text editors. *Proceedings of the 1st International Workshop on Proofing Tools and Language Technologies*, Patras, Greece, 2004.
- Aranguren J., Arrarats I., Goikoetxea A., eta Zabalondo B. *Berria. Estilo liburua*. Andoain, 2006.
- Aranzabe M.J. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, 2008.

- Arppe A. Developing a Grammar Checker for Swedish. *Proceedings from the 12th Nordiske datalingvistikdager*, Department of Linguistics, Norwegian University of Science and Technology (NTNU), December 9-10 2000. Nordgard.
- Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., García E., Lascurain V., Sarasola K., Soroa A., eta Uria L. Semiautomatic conversion of the *Euskal Hiztegia* Basque dictionary to a queryable electronic form. *T.A.L. journal*, 44(2):107–124, 2003.
- Arrieta B., Díaz de Ilarraza A., Gojenola K., Maritxalar M., eta Oronoz M. A database system for storing second language learner corpora. *Learner corpora workshop. Corpus linguistics 2003*, 16 lib., 33–41, Lancaster, UK, 2003.
- Arriola J.M. *Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikala-ren eskuratze automatikoari begira. Aditz-adibideen analisisa Murritzapen Gramatika baliatuz, azpikategorizazioaren bidean*. Doktoretza-tesia, Filologia eta Historia-Geografia Fakultatea. UPV-EHU, 2000.
- Arriola J.M., Esparza I., Ezeiza N., Gojenola K., eta Sologastoa A. Anlizatzaile morfosintaktikoa. Barne-txostena, EHU/UPV, 2005.
- Arruarte A., Elorriaga J., eta Rueda U. A templated based concept mapping tool for computer-aided learning. In Okamoto K.J.K.E. R. Hartley, editor, *IEEE International Conference on Advanced Learning Technologies. IEEE Computer Society*, 309–312, 2001.
- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Hernández G., eta Soroa A. A class library for the integration of NLP tools: Definition and implementation of an abstract data type collection for the manipulation of SGML documents in a context of stand-off linguistic annotation. *Fourth International Conference on Language Resources and Evaluation*, Lisbon (Portugal), 2004a.
- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., eta Soroa A. A proposal for the integration of NLP tools using SGML-tagged documents. *Second International Conference on Language Resources and Evaluation*, Athens (Greece), 2000.
- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sologastoa A., eta Soroa A. Eulia: a graphical web interface for creating, browsing and editing linguistically annotated corpora. *Proceedings of the Fourth Interna-*

- tional Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, Portugal, 2004b.
- Atwell E. How to detect grammatical errors in a text without parsing it. *Proceedings of The 3rd Conference of the European Chapter of the Association for Computational Linguistics*, 38–45, University of Copenhagen, Copenhagen, Denmark, 1987.
- Badia T., Gil A., Quixal M., eta Valentín O. NLP-enhanced error checking for Catalan unrestricted text. *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*, 1919–1922, Lisbon, Portugal, 2004.
- Banko M. eta Brill E. Scaling to very very large corpora for natural language disambiguation. *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 26–33, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- Beck J. eta Woolf B. High-level student modeling with machine learning. In Gauthier G., Frasson C., eta VanLehn K., editors, *Proceedings of the Fifth International Conference on Intelligent Tutoring Systems. Lecture Notes in Computer Science*, 1839 lib., 584–593, 2000.
- Becker M., Bredenkamp A., Crysmann B., eta Klein J. Annotation of error types for german news corpus. *Proceedings of the ATALA Workshop on Treebanks*, Paris, France, 1999.
- Beesley K.R. eta Karttunen L. *Finite State Morphology*. CSLI Studies in Computational Linguistics, 2003.
- Bender E.M., Flickinger D., Oepen S., Walsh A., eta Baldwin T. Arboretum: Using a precision grammar for grammar checking in CALL. *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy, 2004.
- Bengoetxea K. eta Gojenola K. Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. *Actas del XXIII Congreso de la SEPLN*, Sevilla, Spain, 2007.
- Bernth A. Easyenglish: a tool for improving document quality. *Proceedings of the fifth conference on Applied natural language processing*, 159–165, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

- Bex G.J., Maneth S., eta Neven F. A formal model for an expressive fragment of XSLT. *Lecture Notes in Computer Science*, 1861:1137–1151, 2000.
- Bigert J. Probabilistic detection of context-sensitive spelling errors. *Proceedings of LREC-04*, 5 lib., 1633–1636, Lisbon, Portugal, 2004.
- Bigert J., Kann V., Knutsson O., eta Sjöbergh J. Grammar checking for Swedish second language learners. *CALL for the Nordic Languages*, 33–47. Copenhagen Studies in Language 30, Copenhagen Business School., Samfundslitteratur, 2005.
- Bigert J. eta Knutsson O. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. *Romand 2002 (Robust Methods in Analysis of Natural language Data)*, Frascati, Italy, 2002.
- Bird S., Chen Y., Davidson S.B., Lee H., eta Zheng Y. Designing and Evaluating and XPath Dialect for Linguistic Queries. *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, Atlanta, Georgia, April 3–8 2006.
- Birn J. Detecting grammar errors with Lingsoft's Swedish grammar-checker. *Proceedings from the 12th Nordiske datalingvistikkdager*, Department of Linguistics, Norwegian University of Science and Technology (NTNU), December 9-10 2000. Nordgard.
- Boag S., Chamberlin D., Fernández M.F., Florescu D., Robie J., eta Siméon J. *XQuery 1.0: An XML Query Language*, 2006. URL <http://www.w3.org/TR/xquery>.
- Bolioli A., Dini L., eta Malnati G. JDII: Parsing Italian with a Robust Constraint Grammar. *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, 1003–1007, 1992.
- Bolshakov I.A. An experiment in detection and correction of malapropisms through the web. *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19*, 3406 lib. of *Lecture Notes in Computer Science*, 803–815. Springer, 2005. ISBN 3-540-24523-5.
- Booch G., Rumbaugh J., eta Jacobson I. *The Unified Software Development Process*. Addison Wesley, 1997.

- Bouma G. eta Kloosterman G. Querying dependency treebanks in XML. *Proceedings of the third international conference on Language Resources and Evaluation, LREC 2002*, Gran Canaria. Spain., 2002.
- Braga D. eta Campi A. A graphical environment to query XML data with XQuery. 31–40, December 2003.
- Brants S., Dipper S., Hansen S., Lezius W., eta Smith G. The TIGER Treebank. *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria, 2002.
- Brehony T. eta Ryan K. Francophone stylistic grammar checking (FSGC) using Link Grammars. *Computer Assisted Language Learning*, 7(3):000–000, 1994.
- Brill E. Transformation-based error-driven learning and Natural Language Processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.
- Briscoe T. eta Carroll J. Automatic extraction of subcategorization from corpora. *Proceedings of the fifth conference on Applied Natural Language Processing*, 356–363, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
- Bustamante F.R. eta León F.S. Gramcheck: a grammar and style checker. *Proceedings of the 16th conference on Computational linguistics*, 175–181, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
- Carlberger J., Domeij R., Kann V., eta Knutsson O. A Swedish grammar checker. Unpublished. Submitted 2002 Association for Computational Linguistics, 2002.
- Carlson A.J., Rosen J., eta Roth D. Scaling up context-sensitive text correction. In Hirsch H. eta Chien S., editors, *Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference (IAAI-01)*, 45–50, Menlo Park CA, august 7–9 2001. American Association for Artificial Intelligence.
- Carroll J. Parsing. In Mitkov R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.
- Carroll J., Briscoe T., eta Sanfilippo A. Parser evaluation: a survey and a new proposal. *1st International Conference on Language Resources and Evaluation*, 447–454, Granada, Spain, 1998.

- Carroll J., Minnen G., eta Briscoe T. Corpus annotation for parser evaluation. *EACL 99 post-conference workshop on Linguistically Interpreted Corpora (LINC)*, 35–41, Bergen, Norway, 1999.
- Cassidy S. XQuery as an annotation query language: a use case analysis. *Proceedings of the third international conference on Language Resources and Evaluation, LREC 2002*, Gran Canaria. Spain., 2002.
- Cassidy S. eta Harrington J. Multi-level annotation in the Emu speech database management system. *Speech Communication. Special issue on speech annotation and corpus tools*, 33(1–2):61–77, 2001.
- Charniak E. A parser with something for everyone. In King M., editor, *Parsing Natural Language*, 117–149, London, 1983. Academic Press.
- Charniak E. A Maximum-Entropy-Inspired parser. *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- Chinchor N. Overview of MUC-7. *Proceedings of the 7th Message Understanding Conference (MUC-7)*, 1998.
- Chodorow M. eta Leacock C. An unsupervised method for detecting grammatical errors. *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, 140–147, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- Chodorow M., Tetreault J., eta Han N.R. Detection of Gramatical Errors Involving Prepositions. *4th ACL-SIGSEM workshop on Prepositions*, Prague, 2007.
- Chomsky N. Three models for the description of language. *IRE Transactions on Information Theory*, 2:113–124, 1956.
- Chomsky N. *Aspects of the theory of syntax*. MIT Press, Cambridge, Massachusetts, 1965.
- Chomsky N. *Lectures on Government and Binding*. Foris Publications, The Netherlands, 1981.
- Clark J. *XSL Transformations (XSLT)*, 1999. URL <http://www.w3.org/TR/xslt>.

- Clark J. eta DeRose S. *XML Path Language (XPath)*, 1999. URL <http://www.w3.org/TR/xpath>.
- Collins M. Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, 16–23, Morristown, NJ, USA, 1997. Association for Computational Linguistics.
- Copestake A. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA, 2002.
- Corder P.S. Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics*, 5(4):147–160, 1991. Traducción al español: "Dialectos idiosincrásicos y análisis de errores. Juana Muñoz Liceras (Ed.) 1991, La adquisición de las lenguas extranjeras, Visor, Madrid, pp. 63–77.
- Courtin J. *Algorithmes pour le traitement interactif des langues naturelles*. Doktoretza-tesia, USMG INPG, 1977.
- Courtin J., Dujardin D., Kowarski I., Genthial D., eta Strube de Lima V.L. Towards a complete detection/correction system. *International Conference on Current Issues in Computational Linguistics*, 158–173, 1991.
- Cucerzan S. eta Brill E. Spelling correction as an iterative process that exploits the collective knowledge of web users. *Proceedings of Empirical Methods in Natural Language Processing*, 293–300, 2004.
- Dale R. Symbolic approaches to Natural Language Processing. In Dale R., Moisl H., eta Sommers H., editors, *Handbook of Natural Language Processing*. Marcel Dekker Inc, 2000.
- DeSmedt W. Herr Komissar: An ICALL conversation simulator for intermediate German. In Holland V., Kaplan J., eta Sams M., editors, *Intelligent Language Tutors: Theory Shaping Technology*, 371–381. NJ: Lawrence Erlbaum Associates Inc., 1995.
- Díaz de Ilarraza A. *Diseño de un Módulo de Interacción Tutor-alumno para un sistema inteligente de enseñanza de la programación*. Doktoretza-tesia, Informatika Fakultatea, Euskal Herriko Unibertsitatea, 1990.
- Díaz de Ilarraza A., Garmendia A., eta Oronoz M. Abar-Hitz: An annotation tool for the Basque dependency treebank. *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*, 251–254, Lisbon, Portugal, 2004.

- Díaz de Ilarraza A., Gojenola K., eta Oronoz M. Design and development of a system for the detection of agreement errors in Basque. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing: 6th International Conference CICLing2005, Mexico City, Mexico, February 13-19*, 3406 lib. of *Lecture Notes in Computer Science*, 793–803. Springer-Verlag GmbH, 2005. ISBN 3-540-24523-5.
- Díaz de Ilarraza A., Gojenola K., eta Oronoz M. Reusability of a corpus and a treebank to enrich verb subcategorisation in a dictionary. *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP07)*, 280–284, Borovets, Bulgaria, 27-29 September 2007a.
- Díaz de Ilarraza A., Gojenola K., eta Oronoz M. Detecting Erroneous Uses of Complex Postpositions in an Agglutinative Language. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, 18–22 August 2008.
- Díaz de Ilarraza A., Gojenola K., Oronoz M., Otaegi M., eta Alegria I. Syntactic error detection and correction in date expressions using finite-state transducers. *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP07)*, Postdam, Germany, 14-16 of September 2007b.
- Díaz de Ilarraza A., Maritxalar M., eta Oronoz M. Reusability of NLP tools for detecting rules and contexts when modelling language learners knowledge. *Proceedings of the Conference on Recent Advances in NLP (RANLP97)*, 342–348, Tzigov Chark (Bulgary), 1997.
- Díaz de Ilarraza A., Maritxalar M., eta Oronoz M. An implemented inter-language model for learners of Basque. In Sake Jager J.N. eta van Essen A., editors, *Language Teaching and Language Technology*, 149–166, Lisse, 1998a. Swets and Zeitlinger.
- Díaz de Ilarraza A., Maritxalar M., eta Oronoz M. IDAZKIDE: an intelligent CALL environment for second language acquisition. *Proceedings of a one-day conference Natural Language Processing in Computer-Assisted Language Learning*, UK, May 1999. Centre for Computational Linguistics, UMIST, in association with EUROCALL, a special ReCALL publication.
- Díaz de Ilarraza A., Maritxalar M., Oronoz M., eta Maritxalar A. Integration of NLP tools in an Intelligent Computer Assisted Language Learning environment for Basque: IDAZKIDE. *Proceedings of Natural Language Processing and Industrial Applications*, Moncton, Canada, 1998b.

- Díaz de Ilarraza A., Mayor A., eta Sarasola K. Semiautomatic labelling of semantic features. *International Conference of Computational Linguistics (COLING)*, Taipei, Taiwan, 2002.
- Domeij R., Knutsson O., Carlberger J., eta Kann V. Granska- an efficient hybrid system for Swedish grammar checking. *Proceedings of the 12th Nordiska Datorlingvistikdagarna (NoDaLiDa)*, 49–56, Trondheim, Norway, 1999.
- Domeij R., Knutsson O., eta Severinson Eklundh K. Different ways of evaluating a Swedish grammar checker. *Proceedings of The Third International Conference on Language Resources and Evaluation*, 262–267, Las Palmas, Spain, 2002.
- Dulay H., Burt M., eta Krashen S. *Language Two*. Oxford University Press, 1982.
- Edge J. *Mistakes and Correction*. Longman, London, 1989.
- Egunkaria. *Estilo liburua*. Egunkaria, Lasarte-Oria, Gipuzkoa, 1995.
- Eisner J. Learning non-isomorphic tree mappings for machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Supporo, Japan, 2003.
- Elhuyar. *Euskal Hiztegi Modernoa*. Elkar, 2000.
- Elhuyar Fundazioa. *Elhuyar Hiztegia. Euskara-Gaztelania, Castellano-Vasco. 3. argitaralpena*. Elhuyar Fundazioa, Usurbil, 2006.
- Emi I., Uchimoto K., eta Isahara H. The overview of the SST speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. *Proceedings of LREC-04*, 4 lib., 1435–1439, Lisbon, Portugal, 2004.
- Engineering I.J.S. ISO/IEC FCD 9126-1: Information Technology - Software product quality - part 1: Quality model. Barne-txostena, 2000.
- Euskaltzaindia. *Euskal Gramatika. Lehen Urratsak-I*. Euskaltzaindia, Burlata, 1985.
- Euskaltzaindia. *Euskal Gramatika. Lehen Urratsak-IV (Juntagailuak)*. Euskaltzaindia, Bilbo, 1994.

- Evans L.L.D. eta Gates D. The ALICE system: A workbench for learning and using language. *CALICO journal*, 9(1):27–56, 1991.
- Ezeiza N. EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 1997.
- Ezeiza N. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua*. Doktoretza-tesia, University of the Basque Country, Donostia, 2002.
- Flickinger D. On building a more efficient grammar by exploiting types. *Natural Language Engineering (Special Issue on Efficient Processing with HPSG)*, 6(1):15–28, 2000.
- for Standardization/ International Electrotechnical Commission I.O. ISO/IEC 14598-5:1998 standard - information technology – software product evaluation – part 5: Process for evaluators. Barne-txostena, 1998.
- Foster J. *Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English*. Doktoretza-tesia, University of Dublin, Trinity College, Dublin, Ireland, 2005.
- Foster J. Treebanks gone bad: parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition*, 2007.
- Foster J. eta Vogel C. Good reasons for noting bad grammar: Constructing a corpus of ungrammatical language. *Pre-Proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, 151–152, Tübingen, Germany, 2004a.
- Foster J. eta Vogel C. Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, 21(3-4):269–291, 2004b.
- Fowler M. eta Scott K. *UML Distilled: Applying the Standard Object Modelling Language (Object Technology)*. Addison Wesley, 1997.
- Gartzia J. *Kalko Okerrak. Ikasmaterialen Aholku Batzordea, estilo-liburua-ren bigarren atala*. Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, Vitoria Gasteiz, 2005.
- Gelbukh A., Sidorov G., Han S.Y., eta Hernández-Rubio E. Automatic enrichment of very large dictionary of word combinations on the basis of

- dependency formalism. *Lecture Notes in Artificial Intelligence*, (2972): 430–437, 2004.
- Genthial D., Courtin J., eta Menezo J. Towards a more user-friendly correction. *COLING-94*, Tokyo, Japan, 1994.
- Gojenola K. *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta errorearen tratamenduan*. Doktoretza-tesia, Informatika Fakultatea. Euskal Herriko Unibertsitatea, Donostia, 2000.
- Gojenola K. eta Oronoz M. Corpus-based syntactic error detection using syntactic patterns. *NAACL-ANLP00, Student Research Workshop*, Seattle, USA, April 30 2000.
- Gojenola K. eta Sarasola K. Aplicación de la relajación gradual de restricciones para la detección y corrección de errores sintácticos. *Actas del X congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, (SEPLN)*, 4 lib., Córdoba, Spain, 1994.
- Golding A.R. A Bayesian hybrid method for context-sensitive spelling correction. In Yarovsky D. eta Church K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, 39–53, Somerset, New Jersey, 1995. Association for Computational Linguistics.
- Golding A.R. eta Roth D. Applying Winnow to Context-Sensitive Spelling Correction. *Proc. 13th International Conference on Machine Learning*, 182–190. Morgan Kaufmann, 1996.
- Golding A.R. eta Roth D. A Winnow-Based Approach to Context-Sensitive Spelling Correction. *Machine Learning*, 34(1-3):107–130, 1999.
- Granger S. English language corpora: Design, analysis and exploitation. Amsterdam, 1993. Rodopi.
- Guinovart F.J.G. Aportaciones a la metodología de evaluación de los sistemas de verificación automática de la sintaxis. *Procesamiento del Lenguaje Natural*, (19):7–13, 1996a.
- Guinovart F.J.G. *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*. Doktoretza-tesia, Universidade de Santiago de Compostela, 1996b.

- Güvenir H. Drill and practice for Turkish grammar. In Swartz M.L. eta Yazdani M., editors, *Proceedings of the NATO Advanced Research Workshop "The Bridge to International Communication: Intelligent Tutoring Systems for Foreign Language Learning"*, NATO ASI Series F: Computer and Systems Sciences, 275–291, Berlin, 1992. Springer-Verlag.
- Hagen K., Johannessen J.B., eta Lane P. Some problems related to the development of a grammar checker. *Proceedings of NODALIDA '01. 13th Nordic Conference on Computational Linguistics*, Uppsala, May 20-22 2001.
- Hashemi S.S., Cooper R., eta Andersson R. Positive grammar checking: A finite state approach. *Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing 2003, Mexico City, Mexico, February 16-22*, 2588 lib. of *Lecture Notes in Computer Science*, 635–646. Springer, 2003. ISBN 0302-9743.
- Heid U., Voormann H., Milde J., Gut U., Erk K., eta Padó S. Querying both time-aligned and hierarchical corpora with NTX search. *Fourth Language Resources and Evaluation Conference*, Lisbon, Portugal, 2004.
- Helfrich A. eta Music B. Design and evaluation of grammar checkers in multiple languages. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 2 lib., Saarbrücken, Germany, 2000.
- Holan T., Kuboň V., eta Plátek M. A prototype of a grammar checker for Czech. *Proceedings of the fifth conference on Applied Natural Language Processing*, 147–154, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
- Holland V.M., Maisano R., eta Alderks C. Parsers in tutors: what are they good for? *CALICO journal*, 11(1):28–46, 1993.
- Hualde J.I. eta de Urbina J.O. *A grammar of Basque*. 2001.
- Iparraldeko euskaltzainak. Euskararen batasunari buruz “h” letraren auzia dela-ta. *Euskera agerkaria*, 23, 1978.
- I.V.A.P. *IVAPeko estilo-liburua*. I.V.A.P. Instituto Vasco de Administración Pública, 2005.
- IXA taldea eta Elhuyar Fundazioa. Testu-corpusak: ezaugarriak, eraketa eta tresnak. *Hizkuntza, komunikazioaren eta teknologiaren garaian. IVAP aldizkaria.*, 2007.

- Izumi E., Uchimoto K., Saiga T., Supnithi T., eta Isahara H. Automatic error detection in the Japanese learners' English spoken data. *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 145–148, Morristown, NJ, USA, 2003. Association for Computational Linguistics. ISBN 0-111-456789.
- James C. *Errors in Language Learning and Use*. Applied Linguistics and Language Study. Macquarie University, Sydney, 1998.
- Jaynes E.T. Information theory and statistical mechanics I. *Physical Review*, 106:620–630, 1957.
- Jensen K. PEG: The PLNLP English grammar. In Jensen K., Heidorn G.E., eta Richardson S.D., editors, *Natural Language Processing: The PLNLP Approach*, 29–45. Kluwer, Boston, 1993.
- Johannessen J.B., Hagen K., eta Lane P. The performance of a grammar checker with deviant language input. *Proceedings of the 19th international conference on Computational linguistics*, 1–8, COLING, Taipei, Taiwan, 2002. Association for Computational Linguistics.
- Jurafsky D. eta Martin J.H. *Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, USA, 2000. ISBN 0-13-095069-6.
- Kann V. CrossCheck - a grammar checker for second language writers of Swedish. Barne-txostena, Nada, KTH, 2002.
- Karlsson F. SWETWOL: A comprehensive morphological analyser for swedish. *Nordic Journal of Linguistics*, 15(1):1–45, 1992.
- Karlsson F., Voutilainen A., Heikkilä J., eta Anttila A. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Prentice-Hall, Berlin, 1995.
- Karttunen L. Numbers and Finnish Numerals. In *A Man of Measure Festschrift in Honour of Fred Karlsson on his 60th Birthday, a special supplement to SKY Journal of Linguistics*, 19:407–421, 2006.
- Karttunen L., Gaál T., eta Kempe A. Xerox finite state tool. Barne-txostena, Xerox Research Centre Europe, 1997.
- Kilgarrieff A. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155, 1997.

- Klaudio Harluxet Fundazioa. *Harluxet Hiztegi Entziklopedikoa*. Elkarlanean, S.L. Argitaletxea, 2001.
- Knutsson O., Carlberger J., eta Kann V. An object-oriented rule language for high-level text processing. *NoDaLiDa '01 - 13th Nordic Conference on Computational Linguistics*, Uppsala, Sweden, May 21-22 2001.
- König E. eta Lezius W. The TIGER language – a description language for syntax graphs – formal definition. Barne-txostena, IMS, University of Stuttgart, 2001.
- Koskenniemi K. *Two-level Morphology: a general computational model for word-form recognition and production*. University of Helsinki, Helsinki, 1983.
- Kukich K. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439, December 1992.
- Kumar A. eta Nair S. An artificial immune system based approach for English grammar checking. In L.N. de Castro F.V.Z. eta H.Knidel, editors, *ICARIS 2007, Lecture Notes in Computer Science*, number LNCS 4628, 348–357, 2007.
- Lai C. eta Bird S. Querying and updating treebanks: A critical survey and requirements analysis. *Proceedings of the Australasian Language Technology Workshop*, 2004.
- Larman C. *UML y Patrones. Una introducción al análisis y diseño orientado a objetos y al proceso unificado*. Pearson. Prentice Hall, 2 edition, 2002.
- Leech G. The importance of reference corpus. <http://www.uzei.org/modulos/usuariosFtp/conexion/archivos59A.pdf/>, 2002.
- Lersundi M. *Ezagutza-base lexikala eraikitzeke Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoa: definizio-patroiak, eratorpenak eta postposizioak*. Doktoresetia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Leioa, 2005.
- Levin L. eta Evans D. ALICE-chan: A case study in ICALL theory and practice. In Holland V., Kaplan J., eta Sams M., editors, *Intelligent Language Tutors: Theory Shaping Technology*. NJ: Lawrence Erlbaum Associates Inc., 1995.

- Levison M., Lessard G., eta Walker D. A multi-level approach to the detection of second language learners errors. *Literary and Linguistic Computing*, 15(3):313–322, 2000.
- Lezius W. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Doktoretza-tesia, IMS, University of Stuttgart, December 2002. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.
- Lezius W. eta König E. Towards a search engine for syntactically annotated corpora. *KONVENS 2000 Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung Sprachkommunikation*, 113–116, Berlin, Germany, 2000. VDE-Verlag GmbH. ISBN 3-8007-2564-9.
- Lin D. Dependency-based evaluation of MINIPAR. *1st International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.
- Ljubopytnov V., Němec P., Pilátová M., Reschke J., eta Stuchl J. Oraculum, a system for complex linguistic queries. *Proceedings of SOFSEM 2002, Student Research Forum*, Milovy, 2002.
- Mangu L. eta Brill E. Automatic rule acquisition for spelling correction. *Proceedings of the 14th International Conference on Machine Learning*, 187–194. Morgan Kaufmann, 1997.
- Marcus M., Kim G., Marcinkiewicz M.A., MacIntyre R., Bies A., Ferguson M., Katz K., eta Schasberger B. The Penn treebank: Annotating predicate argument structure. *Proceedings of the ARPA Human Language Technology Workshop*, 1994.
- Marcus M. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, 1980.
- Maritxalar M. *MUGARRI: Bigarren hizkuntzako ikasleen hizkuntz ezagutza eskuratzeko sistema anitzeko ingurunea*. Doktoretza-tesia, Informatika Fakultatea. Euskal Herriko Unibertsitatea, Donostia, 1999.
- Maritxalar M., Díaz de Ilarraza A., eta Oronoz M. From psycholinguistic modelling of interlanguage in second language acquisition to a computational model. *Proceedings of Computational Natural Language Learning (CoNLL97). In conjunction with ACL'97/EACL'97*, 50–59, Madrid, Spain, 11-12th July 1997.

- Martinez D. *Supervised Word Sense Disambiguation: Facing Current Challenges*. Doktoretza-tesia, University of the Basque Country, Donostia, 2004.
- Matthews C. Grammar frameworks in intelligent CALL. *CALICO*, 11(1): 5–27, 1993.
- McCord M.C. Slot grammar: A system for simpler construction of practical natural language grammars. In Studer R., editor, *Natural Language and Logic: Proc. of the International Scientific Symposium, Hamburg, FRG*, 118–145. Springer, Berlin, Heidelberg, 1990.
- Mel'čuk I. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.
- Menzel W. Error diagnosing and selection in a training system for second language learning. *COLING-88*, 414–419, 1988.
- Menzel W. eta Schröder I. Constraint-based diagnosis for intelligent language tutoring systems. *Proceedings of the IT&KNOWS Conference at the IFIP '98 Congress*, Wien/Budapest, 1998.
- Menzel W. eta Schröder I. Error diagnosis for language learning systems. *Special edition of the ReCALL journal*, 20–30, 1999.
- Mírovský J. Netgraph: a tool for searching in Prague Dependency Treebank 2.0. *Proceedings of The Fifth International Treebanks and Linguistic Theories conference*, 211–222, Prague, Czech Republic, 1st and 2nd December 2006.
- Mitjushin L. An agreement corrector for Russian. *Proceedings of the 16th conference on Computational linguistics*, 776–781, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
- Mokhtar S.A., Chanod J.P., eta Roux C. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(3):121–144, 2002.
- Moré J., Climent S., eta Oliver A. A grammar and style checker based on Internet searches. *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*, 1931–1934, Lisbon, Portugal, 2004.

- Naber D. *A Rule-Based Style and Grammar Checker*. Doktoretza-tesia, Technische Fakultät. Universität Bielefeld, 2003.
- Nerbonne J. Computer-Assisted Language Learning and Natural Language Processing. In Mitkov R., editor, *The Oxford Handbook of Computational Linguistics*, 670–698. Oxford University Press, 2003.
- Nivre J. Dependency grammar and dependency parsing. Barne-txostena, Växjö University: School of Mathematics and Systems Engineering, 2005.
- Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., eta Marsi E. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.
- Nivre J. eta Scholz M. Deterministic dependency parsing of English text. *Proceedings of COLING 2004*, 64–70, Geneva, Switzerland, August 23-27 2004.
- Norrish J. *Language learners and their errors*. Macmillan Press, London, 1983.
- Nygaard L. eta Johannessen J.B. SearchTree - a user-friendly treebank search interface. In Kübler S., Nivre J., Hinrichs E., eta Wunsch H., editors, *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 04)*, 183–190, Tübingen, December 10-11 2004.
- Oflazer K. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Comput. Linguist.*, 22(1): 73–89, 1996. ISSN 0891-2017.
- Örvar Kárasón. Detecting grammatical errors with memory-based learning, 2005.
- Otaegi M. Datak, orduak eta zenbakiak euskaraz. Barne-txostena, University of the Basque Country, 2006.
- Otegi A. Estilo kontuak eta puntuazio ikurren erroreak detektatzeko sistema. Barne-txostena, Euskal Herriko Unibertsitatea, 2003.
- Otegi A. Zuzentzaile sintaktikoa Word-en integratzeko liburutegi baten sortkuntza. Barne-txostena, Euskal Herriko Unibertsitatea, 2006.

- Owczarzak K., van Genabith J., eta Way A. Labelled dependencies in machine translation evaluation. *Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics*, 104–111, Prague, June 2007.
- Paggio P. Spelling and grammar correction for Danish in SCARRIE. *Proceedings of the sixth conference on Applied Natural Language Processing*, 255–261, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- Palomar M., Civit M., Díaz de Ilarraza A., Moreno L., Bisbal E., Aranzabe M., Ageno A., Martí M.A., eta Navarro B. 3LB: Construcción de una base de árboles sintáctico-semánticos para el Catalán, Euskera y Castellano. *Actas del XX Congreso de la SEPLN*, 2004.
- Peterson J.L. Computer programs for detecting and correcting spelling errors. *Commun. ACM*, 23(12):676–687, 1980. ISSN 0001-0782.
- Prechelt L. An empirical comparison of C, C++, Java, Perl, Python, Rexx and Tcl. *IEEE Computer*, 33(10):23–29, 2000.
- Punyakanok V., Roth D., eta tau Yih W. Mapping dependencies trees: An application to question answering. *Proceedings of AI & Math 04*, 2004.
- Ravin Y. Grammar errors and style weaknesses in a text-critiquing system. In Jensen K., Heidorn G.E., eta Richardson S.D., editors, *Natural Language Processing: The PLNLP Approach*, 65–76. Kluwer, Boston, 1993.
- Riley M., Craven L., eta Olsen M.B. Customer-focused evaluation of a grammar checker. *Proceedings of the 1st International Workshop on Proofing Tools and Language Technologies*, Patras, Greece, 2004.
- Rohde D. *TGrep2 User Manual*.
- Samuelsson C. eta Wirén M. Parsing techniques. In Dale R., Moisl H., eta Sommers H., editors, *Handbook of Natural Language Processing*. Marcel Dekker Inc, 2000.
- Sarasola I. *Euskal Hiztegia*. Kutxa Gizarte- eta Kultur-Fundazioa, Donostia, 1996.
- Schmidt P., Garnier S., Sharwood M., Badia T., Díaz L., Quixal M., Ruggia A., Valderrabanos A.S., Cruz A.J., Torrejon E., Rico C., eta Jimenez J. ALLES: Controlled language tools' and information extraction tools for

- CALL applications. *Proceedings of InSTIL/ICALL2004, NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy, June 2004a.
- Schmidt P., Garnier S., Sharwood M., Badia T., Díaz L., Quixal M., Ruggia A., Valderrabanos A.S., Cruz A.J., Torrejon E., Rico C., eta Jimenez J. ALLES: Integrating NLP in ICALL applications. *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*, 1919–1922, Lisbon, Portugal, 2004b.
- Schneider D.A. eta McCoy K.F. Recognizing syntactic errors in the writing of second language learners. *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics (COLING-ACL)*, 2 lib., Montreal, Canada, 1998.
- Shaanan K.F. Arabic GramCheck: a grammar checker for Arabic: Research articles. *Software: Practice and Experience*, 35(7):643–665, 2005. ISSN 0038-0644.
- Shaw M. eta Clements P. A field guide to boxology: Preliminary classification of architectural styles for software systems. *Proceedings of COMPSAC97. 21st International Computer Software and Applications Conference*, 6–13, August 1997.
- Shieber S.M. *An Introduction to Unification-Based Approaches to Grammar*. Number 4. CSLI Lecture nOTES, Stanford, 1986.
- Si-Qing C. eta Luomai X. GRAMMAR-DEBUGGER: A parser for Chinese EFL learners. *CALICO journal*, 8(2):63–75, 1990.
- Simov K. eta Osenova P. A treebank-driven approach to semantic lexicon creation. *Proceedings of TLT04*, Tuebingen, Germany, 2004.
- Sjöbergh J. Chunking: an unsupervised method to find errors in text. *Proceedings of NODALIDA 2005*, Joensuu, Finland, 2005.
- Sjöbergh J. The Internet as a normative corpus: Grammar checking with a search engine. Barne-txostena TRITA-CSC-TCS 2006:3, School of Computer Science and Communication, the Royal Institute of Technology, Stockholm, Sweden, 2006. URL <http://dr-hato.se/research/internetnorm.pdf>.

- Sjöbergh J. eta Knutsson O. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. *Proceedings of RANLP 2005*, 506–512, Borovets, Bulgaria, 2005.
- Sleator D. eta Temperley D. Parsing English with a Link Grammar. *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands and Durbuy, Belgium, August 1993.
- Sperberg-McQueen C. eta Burnard L. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, 2002.
- Starlander M. eta Popescu-Belis A. Corpus-based evaluation of a French spelling and grammar checker. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 268–274, 2002.
- Steiner I. eta Kallmeyer L. VIQTORYA - a Visual Query Tool for Syntactically Annotated Corpora. *Proceedings of the third international conference on Language Resources and Evaluation, LREC 2002*, 1704–1711, Gran Canaria. Spain., May 2002.
- Tanev H., Kouylekov M., Magnini B., Negri M., eta Simov K. Exploiting linguistic indices and syntactic structures for multilingual question answering: ITC-IRST at CLEF 2005. *Lecture Notes in Computer Science*, 4022/2006:390–399, 2006.
- Tapanainen P. *The Constraint Grammar parser CG-2*. Publications of the University of Helsinki, 27, Helsinki, 1996.
- Tapanainen P. *Parsing in two frameworks: finite-state and functional dependency grammar*. Doktoretza-tesia, University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts., 1999.
- Taylor C.L. *XSLT as a Linguistic Query Language*. Doktoretza-tesia, Department of Computer Science and Software Engineering. The University of Melbourne, November 17 2003.
- Teixeira Martins R., Hasegawa R., Volpe Nunes M.D.G., Montilha G., eta Osvaldo Novais De Oliveira J. Linguistic issues in the development of ReGra: A grammar checker for Brazilian Portuguese. *Natural Language Engineering*, 4(4):287–307, 1998. ISSN 1351-3249.
- ten Hacken P. Book review of errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues, 2008.

- Tesnière L. *Éléments de syntaxe structurale*. Paris, 1959.
- Tesnière L. *Elementos de sintaxis estructural*. Biblioteca Románica Hispánica, 1994.
- Tomita M. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems*. Kluwer Academic Publishers, 1986.
- Tsiriga V. eta Virvou M. Integrating machine learning methods throughout the temporal extent of a web-based student model. In Stephanidis eta Jacko, editors, *Human-Computer Interaction: Theory and Practice (Part II)*, *Proceedings of the 10th International Conference on Human Computer Interaction (HCI 2003)*, 2 lib., 475–479, 2003.
- Ugarteburu I. EHUKo zuzenketak: mailak eta motak. *Senez*, 27, 2004.
- Urkia M. *Euskal morfologiaren tratamendu informatikorantz*. Doktoretza-tesia, Filologia eta Historia-Geografia Fakultatea. UPV-EHU, uztaila 1997. Miren Azkarate, UPV-EHUKo irakaslearen zuzendaritzapean eginitako tesia.
- Vandeventer A. *Syntactic Error Diagnosis in the context of Computer Assisted Language Learning*. Doktoretza-tesia, Université de Geneve, Geneve, 2003.
- Voormann H. eta Lezius W. TIGERin - grafische eingabe von benutzeranfragen für ein baumbank-anfragewerkzeug. In Busemann S., editor, *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken, 2002.
- Vosse T. Detecting and correcting morpho-syntactic errors in real texts. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 111–118, 1992.
- Voutilainen A. eta Tapanainen P. Ambiguity resolution in a reductionistic parser. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, 394–403, Morristown, NJ, USA, 1993. Association for Computational Linguistics. ISBN 90-5434-014-2.
- Wagner J., Foster J., eta van Genabith J. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 112–121, 2007.

- Weinberg A., Garman J., Martin J., eta Merlo P. A principled-based parser for foreign language tutoring in German and Arabic. In Holland V., Kaplan J., eta Sams M., editors, *Intelligent Language Tutors: Theory Shaping Technology*, 23–44. NJ: Lawrence Erlbaum Associates Inc., 1995.
- Zabala I. eta Odriozola J.C. Los complejos posposicionales en vasco. In Gaztelu E.P., Zabala I., eta Gracia L., editors, *Las fronteras de la composición en lenguas románicas y en vasco.*, 281–315, 2004.
- Zeman D. eta Sarkar A. Learning verb subcategorization from corpora: Counting frame subsets. *2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, May 31–June 2 2000.
- Zock M. Computational linguistics and its use in real world: the case of Computer Assisted-Language Learning. *Proceedings of COLING 1996 the 16th conference on Computational Linguistics*, 2 lib., 1002–1004, Copenhagen, Denmark, 1996.
- Zuazo K. *Euskalkiak, herriaren lekukoak*. Elkar, 2003.
- Zubiri I. *Gramática didáctica del euskara*. Didaktiker, Bilbo, 1994.
- Zubiri I. eta Zubiri E. *Euskal Gramatika Osoa*. Didaktiker, Bilbo, 1995.

Lengoaia eta Sistema Informatikoak Saila



Informatika Fakultatea

**EUSKARAZKO ERRORE SINTAKTIKOAK
DETEKTATZEKO ETA ZUZENTZEKO
BALIABIDEEN GARAPENA: DATAK,
POSTPOSIZIO-LOKUZIOAK ETA KOMUNZTADURA**

Eranskinak

Maite Oronoz Anchordoquik

Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 2008ko urria.

A. ERANSKINA

Gramatika-zuzentzaileak eta OLHIA sistemak

Erroreak detektatzeko eta zuzentzeko teknikak aztertu ditugunean (ikus III kapitulua), teknika hauek batez ere gramatika-zuzentzaileetan eta OLHIA sistemetan erabiltzen direla aipatu dugu.

Eranskin honetan helburua, kapitulu hartan aipatutako sistemetako bakoitzaren laburpen bat eskaintzea da. Laburpenetan, tesi-lan honetarako interesgarria iruditu zaigun informazioa bildu dugu, beraz, informazio osatuagoren bila iturrietara joatea gomendatzen dugu.

Lehen atalean, A.1 atalean, hainbat hizkuntzatarako aztertutako gramatika-zuzentzaileak zerrendatuko ditugu. Ondoren, A.2 atalean, Ordenagailuz Lagundutako Hizkuntzen Irakaskuntza adimenduneko (ingelesez, ICALL edo *Intelligent Computer Assisted Language Learning*) sistemetan erroreak detektatzeko eta zuzentzeko erabiltzen diren teknikak aipatuko ditugu.

A.1 Gramatika-zuzentzaileak

Sintaxi-mailako erroreak detektatzeko teknikak III kapituluko III.2.1.2 atalean aztertu ditugu. Atal honetan teknika horiek erabiltzen dituzten gramatika-zuzentzaileen errepasoari ekingo diogu.

- Katalan hizkuntzako testu libreak tratatzeko ortografia- eta gramatika-zuzentzaile bat egin dute (Badia *et al.*, 2004) lanean. Sistema azaleko analizatzaile morfosintaktiko batean oinarrituta dago eta hau egokitu behar izan dute gaizki eratutako testuak onar ditzan. Analizatzaile honetako modulu anitz Murritzapen Gramatika erabilia inplementatu dira, baita errorearen detekzioarako soilik erabiltzen dituzten

zenbait modulu ere. Egin dituzten aldaketen artean garrantzitsuenak, desanbiguazio morfologikorako eta sintaktikorako erregelak *erlaxatzea* izan da. Desanbiguatzaileak testu zuzenekin kenduko lituzkeen zenbait interpretazio utzi egiten ditu. Landutako erroreak izen-sintagma barruko komunztadura-erroreak, hitzen ordena okerra, aditza eta subjektuaren arteko komunztadura eza eta aditzen azpikategorizazio-erroreak dira, baina ez dute hauen inplementazioari buruzko datu gehiegi ematen. Errore konplexuagoak (semantika eta pragmatikarekin lotutako zenbait errore) domeinu zehatzerako sortutako moduluarekin lantzen dituzte. *ALLES* urrutiko irakaskuntzarako ikasketa sistema batean (Schmidt *et al.*, 2004a, b) eta PrADo idazleentzako errore-zuzentzaile batean erabiltzen dute gorago aipatutakoa. Sistemak eduki ditzakeen arazoen artean, *gehiegizko detekzioa* aipatzen dute eta alarma faltsuak iragaziko dituen iragazki batean lanean hasi dira. Ez dute oraindik inolako probarik egin, beraz, ez dute emaitzik ematen.

- Suedierarako *Grammatifx* gramatika-zuzentzailea garatzeko ere erabili zen MG (Birn, 2000; Arppe, 2000). Lingsoft enpresak garatutako zuzentzaile honek oinarrian bi mailatako morfologia jarraitzen duen analizatzaile morfologikoa du. Desanbiguazio morfosintaktikoa, beraien jabetzakoa den *Swedish Constraint Grammar*ekin (SWECEG) egiten dute, baita azaleko sintaxia ere. Errore mota desberdinak lantzeko teknika desberdinak erabili dituzte. Sintaxi-erroreak detektatzeko, murriztapen-gramatika (Birn, 2000); puntuazio-erroreak eta zenbaki formatu okerretako erroreak detektatzeko, adierazpen erregularrak; eta azkenik, hitz zehatzetan ematen den estilo desegokia lantzeko, bi mailatako lexikoa (Karlsson, 1992) zabaldu egin da. Murriztapen-gramatikari dagokionez, landutako 26 errore motetarako 650 erregela erabili dira. Lan honetan izen-sintagmen barruko komunztadura eta aditz-kateetan koherentzia landu dira, batez ere. Egileen arabera, aukeratutako errore multzoak hasiera batean oso sinplea dirudien arren, doitasun handia lortzen da emaitzetan (% 70 inguru aukeratutako errore mota bakoitzeko), baita landutako erroreak soilik neurtuta estaldura handia ere (% 85a). Nahiz eta errore gutxi eta sinpleak detektatu, hauen markaketa zuzena izatea oso garrantzitsua da azken erabiltzailearentzako. Diagnosian testuan susmagarria den hitzaren forma, susmagarri izatearen arrazoi linguistikoak eta egitura berean azaltzen diren beste hitzak deskribatzen dira. Azken atalean, zuzenketarako sorkuntza morfologikoaren erabilera proposatzen dute.

- *Grammar Checker for Norwegian (NGC)* (Johannessen *et al.*, 2002) norvegierarako gramatika-zuzentzailean egindako lana, suedierarako sortutako zuzentzailean oinarrituta dago. Sistema analizatzaile morfologiko batekin, MGrekin sortutako desanbiguatzaile morfologiko batekin eta erroreak detektatzeko modulu batekin osatuta dago. Desanbiguatzaileak bere lana egiteko, jasotzen duen informazioa zuzena dela suposatzen du eta hau *zuzentzaile* batean ezin denez inoiz suposatu, erlaxatu egin dute anbiguotasun handiagoa utziaz. Erroreak detektatzeko 700 erregela inguru idatzi dituzte eta hauekin % 75eko doitasuna lortu dute. Estaldura ez dute kalkulatu, baina detektatutako errorearen % 30a alarma faltsuak dira. Sintagma barneko komunztadura, hitzen ordena, esaldi berean aditz gehiegi egotea eta antzeko fenomenoak landu dituzte. Zuzentzailea ama hizkuntza norvegiera duten erabiltzaileentzako sortuta dago eta erabiltzaile *ez-estandarrekin*, ikasle txinatar eta eslaviarrekin, probatu dutenean, hauen errorearen % 85-90 detektatu gabe geratu da, NGCn landugabeko erroreak baitziren.
- Suedierarako egoera finituko sistema bat garatu dute Hashemi *et al.*-ek (2003). Sistema lau modulurekin osatuta dago: *hiztegiaren kontsulta*, *gramatika*, *analizatzailea* eta *errore-bilatzailea*. Lexikoak 160.000 hitz-forma inguru ditu eta Xerox-eko *Finite State Lexicon Compiler* erabiltzaile, egoera finituko transduktore baten bidez eraikita dago. Honek karaktere-kate bat jasotzen du eta gramatika-etiketa eta hitz-formaren ezaugarriekin etiketa bat itzultzen ditu. *Gramatika*, informazio positiboa duten bi gramatikekin osatuta dago. Alde batetik, *gramatika zabala* izendunak, egitura linguistiko zuzenak eta okerrak deskribatzen ditu. *Gramatika estuak* egitura gramatikalak soilik adierazten ditu modu zehatzean. Egitura hauek guztiak XFST tresnak erabiltzen duen idazkeran adierazten dira, adierazpen erregularren bidez. Hauek konpilatzean, egoera finituko transduktoreak lortzen dira. *Analizatzaileak*, aurretik datorren informazioari etiketa berriak gehitzen dizkio eta analisirako erabilitako erregelak ordenatu egiten ditu zenbait anbiguotasun ebazteko. *Erroreen bilatzailea* aparteko modulu bat da eta bertan topa ditzakegun automatak, *gramatika zabalarri*, *gramatika estua kenduz* lortzen dira. Horrela lortutako makinek, egitura erroredunak detektatzeko gaitasuna dute. Sistemako moduluak katean aplikatzen dira eta bi fasetan. Lehenengoan, gramatika-erroreak detektatzea da helburua. Bigarren fasean, berriz, lehenengo fasean sortutako alarma faltsuak *blokeatzea*, gehiegi sortzen baitira. Sistemak % 92ko estaldura eta % 45eko doitasuna lortzen du izen-sintagmetako komunztadurarako.

dura-erroreak eta aditz-forma finitu eta ez-finituekin gertatzen diren erroreak detektatzen saiatzen denean. Sistema hau umeen idazlane-tan eman daitezkeen gramatika-erroreak detektatzera bideratuta dago bereziki. Testu hauetan errore-moten maiztasuna eta banaketa desberdina izaten da, helduek idatzitako testuekin konparatzen baditugu.

- Bere tesi-lanean, Naber-ek (2003) ingeleserako kode irekiko estilo- eta gramatika-zuzentzaile bat garatu du. Zuzentzaileak oinarrian *Extensible Markup Language*-n (XML) eta Python programazio-lengoaian idatzitako patroierregelak ditu eta hauek analizatzaile morfologiko eta azaleko sintaxia egiten duen analizatzaile batekin analizatutako testu bati aplikatzen zaizkio. Analizatzailea Brill-en analizatzaile estatistikoaren aldaera da. Egileak gramatikaren diseinua eta mantentzeko eta zabaltzeko erraztasuna hobetzeko asmoz, objektuei zuzendutako programazioa erabili nahi izan du eta hori dela eta, analizatzaile morfologikoa eta zuzentzailea berriro¹ Python programazio-lengoaia erabiliz implementatu ditu. Sistema sintagma-mailako erroreak detektatzeko gai da, baina zailtasunak ditu izen-sintagma edo aditz-sintagmatik haratago doazen erroreak detektatzeko. Naberrek ez du estaldura edo doitasun neurririk ematen.
- *EasyEnglish* (Bernth, 1997) IBMn sortutako tresna da. Erabiltzaileari, ingelesez idatzitako testu argiagoak eta sinpleagoak sortzen laguntzen dio. Hau da, gramatika-zuzentzaile bat baino gehiago da. Gramatika-erroreak adierazteaz gain, idatzitakoan topatutako egiturazko sintaxi-anbiguotasunak eta idatziaren konplexutasunaren berri ematen dio erabiltzaileari, eta anbiguotasunak ebazteko proposamenak egin. Azkenean, *argitasun-indize* bat ematen du eta indize honek barruti baten barnean egon behar du dokumentu bat IBMkoa bailitzan argitaratua izan dadin. Gramatika-zuzentzaileek dituzten ezaugarriez gain, *hizkuntza kontrolatuko* ezaugarriak erabiltzen ditu *EasyEnglishek*. Hizkuntza kontrolatua definitzen denean, gramatikaren eta hiztegiaren aldetik murriztapen batzuk ezartzen dira sortu nahi den hizkuntza ulertzeko eta itzultzeko erraza izan dadin. Sarrerako testua analizatzeko, *English Slot Grammar (ESG)* izeneko erabiltzen dute (McCord, 1990). Honek sare-egitura moduan adierazitako analisi-zuhaitzak sortzen ditu. Analisi-zuhaitz hauei patroierregelak aplikatzen zaizkie anbiguotasunak eta gramatika-erroreak detektatzeko. Egileek ez dute adierazten

¹Brill-en analizatzailea C programazio-lengoaian implementatuta dago. Naber-ek Perl-en implementatutako zuzentzaile bat erabili zuen.

erregela hauek nolakoak diren eta zein modutan adierazten duten informazioa. Gramatika-zuzentzailea atalean sintaxi-arazoak, lexiko-arazoak eta puntuazio-arazoak lantzen dituzte. Sintaxi-arazoak, sintaxi-analisiarekin mendekotasun handiena dutenak dira eta besteak beste, koordinazioa, ezezko bikoitzak, esaldi ez-osoak e.a. lantzen dira. Lexiko-arazoetan ortografia-hutsegiteak eta hitz errepikatuak lantzen dituzte; puntuazio-arazoen artean, berriz, lotutako sintagmen arteko koma falta, komak biltzea, puntuazio-marken falta eta antzekoak jorratzen dituzte. Doitasun eta estaldura onak lortzen dituzte, % 81 eta % 87a hurrenez hurren, baina hizkuntza kontrolatua lantzen ari direla kontuan izan behar da.

- Lehenago ere, suedierarako beste sistema batzuk aipatu ditugu. Suedierarako azaleko sintaxia erabiltzen duen *Granska* gramatika-zuzentzaile hibridoa (Carlberger *et al.*, 2002; Domeij *et al.*, 1999) aipatuko dugu oraingoan. Honek, patroi-erregelatan eta estatistikan oinarritutako teknikak konbinatzen ditu eraginkortasuna eta sendotasuna lortzeko asmoz. Patroi-erregelak, datu estatistiko batzuk jarraituz aplikatzen dira. Erregelak ez dira kategoria-etiketa guztietan aplikatzen, batzuetan soilik aplikatzen dira. Optimizazio honek 6 aldiz azkarragoa egiten du erregelen aplikazioa. Errorea detektatu ondoren, zuzenketa posibleak sailkatuta eskaintzen dira eta zuzenketa egin ondoren, sortutako esaldia berriro aztertzen da, zuzenketak errore gehiago sortu ote dituen ikusteko. Egileek argi eta garbi erakusten dituzte erreoren detekzioarako eta zuzenketarako erregelak eta hauen aplikazioarako algoritmoak. Erregelak, Markov-en ezkutuko eredua erabiltzen duen analizatzaile morfologiko batekin analizatutako testu bati aplikatzen zaizkio eta erregela horietako batzuk, *laguntza-erregelak* deiturikoek, azaleko sintaxia egiten dute. Erroreak detektatzeko 180 erregela, 60 laguntza-erregela eta 110 onartzeko erregela (salbuespenak argitzeko erregelak) ditu erregela-gramatikak. Sistemak, komunztadura-erroreak eta suedieraz errore tipikoak diren gaizki banatutako hitz-konposatuak lantzen ditu, batez ere. % 53ko doitasuna eta % 52ko estaldura lortu dituzte.
- *GramCheck* proiektuaren barnean Bustamante eta León-ek (1996) gaztelararako eta grekorako gramatika- eta estilo-zuzentzaile bat garatu zuten. GramCheck egiaztatzailean, detekzioaz eta diagnosiaz gain, zuzenketa ere egiten da analisisa-transferentzia-sintesia ziklo baten bidez. Lan honetan Europar Batasunean hizkuntza-ingeniaritzako apli-

kazioak garatzeko ALEP plataforma erabili da. GramCheck gramatika-egiaztatzailean, ALEPerako garatutako baterakuntzan oinarritutako eta ezaugarri motatuak dituen formalismoaren hedapen batzuk erabili dira. Hedapen hauek (Constraint Solvers, CS) zenbait eragiketa boolear eta erlazional ahalbidetzen dituzten PROLOG kode zatiak dira. Erroreen detekzioarako, *ezaugarrien erlaxazioaren* eta *erroreei aurre hartzearen* teknikak konbinatzen dira. Erlaxazioa CS-ak erabiliz egin da eta erroreei aurre hartzeko erregela esplizituak erabili dira. Erroreen diagnosia, CSak erabiliz eta heuristika erabiltzen duen teknika bat erabiliz egiten da. Erroreen zuzenketa egiteko, informazio okerra duten egitura linguistikoak “zuzen” moduan definitu den gatzelerara transferitzen dira itzulpen automatikoko tekniken antza duen teknika bat erabiliaz. Kasu honetan, transferentzia eta sintesia egitura okerrekin soilik egiten da. Landutako erroreei dagokienez, sintagma barruko eta arteko erroreak landu dituzte: generoa eta numeroa, *dequeismoa*, *queismoa*. . . Corpusetako errore gramatikalen % 28,2a detektatu eta zuzendu dituzte eta estilo-erroreen % 15,5a. Erroreen gainontzeko % 56,3a landu gabe geratu da.

- Vosse-k (1992) nederlanderako ortografia- eta gramatika-zuzentzaile bat garatu du. Sistemak hitz mailako zuzenketa egiten du batez ere, baina esaldi-maila ere lantzen du. Esaldien analisirako, TGG hedatu bat eta Tomitaren algoritmoaren hedapen bat erabiltzen dira. Analizatzaile honetan, ez da analisirik egingo, instantziazioren bat ezin bada aurrera eramanez edo konstanteren baten balioa topatzen ez bada (esaldi erroreduna). Kasu honetan, errorea detektatzen da baina ez da zuzenketarik proposatzen, ezta errorea non gertatu den adierazten duen informaziorik ematen ere. Hau lortzeko, sistemak porrot egin beharrean, analisia lor dadin ahalbidetu behar da, okerra gertatu deneko tokian marka bat ipinita. Beraz, nahiz eta testuan zuzenean ez aipatu, gramatikan nolabaiteko *erlaxazioa* erabiltzen duela iruditzen zaigu, bat ez datozen atributuak egonda ere, analisia egitea lortzen baita. Analisia amaitu denean, zuzentzaileak markak aztertuko ditu eta komunztadura derrigortuko du ezaugarriak hedatuaz. Sintaxi-erroreek gramatika-egiturak aldatzen dituzteneko kasuetarako, egitura okerrak deskribatzen dituzten errore-erregelak gehitu ditu Vosse-gramatikan. Errore-erregela hauei, errore-pisuak gehitzen dizkio eta gero, lortutako analisi-zuhaitz guztiak, errore-pisu horien baturaren arabera ordenatzen ditu pisu txikiena duenetik hasita handiena duenera arte. Metodo honekin analisi kopuru izugarria sortzen da bai erroredun esaldietara-

ko baita esaldi zuzenetarako ere, erregela zuzen eta oker guztiak probatzen baitira. Honek anbigutasun oso handia gehitzen dio dagoe-neko hizkuntzak duenari. Errore tipografikoak, ortografikoak (gaizki banatutako hitz-konposatuak barne) eta morfosintaktikoak landu ditu Vossek. Azken hauen artean, erlaxazioaren bidez komunztadura-erroreak landu dituela ikusi dugu. Hau guztia, modu ez elkarreragilean eta nahiko geldo lan egiten duen idazketarako-laguntza tresna batean integratu du. Emaitei dagokionez, esaldi motzekin oso emaitza onak ematen ditu eta alarma faltsu gutxi. Esaldi luzeak dituzten testuekin, berriz, analizatzaileak arazoak ematen ditu anbigutasunari eta denborari dagokionez. Testu handiekin, Vossen sistemak 11 benetako errore eta 12 alarma faltsu detektatu ditu, hots, % 47,8ko doitasuna lortu du.

- *JDII* italieraz egiten diren sintaxi-erroreak detektatzeko gai den sistema da (Bolioli *et al.*, 1992). Oinarrian baterakuntza formalismoa jarraitzen duen DIMACheck sintaxi-analizatzailea du eta berau *murriztapen ahulak*, *murriztapen sendoak* eta *erabiltzailearen eragile eta funtzioak* izeneko mekanismo linguistikoekin aberasten dute. Nahiz eta detektatu ditzaketen errorearen zerrendapen bat egin, ez da oso argi geratzen nola egiten duten sintaxi-erroreen detekzio automatikoa. Maila teorikoan geratzen direla dirudi, eta ez dute inolako emaitzarik ematen.
- *ReGra* gramatika-zuzentzaileak (Teixeira Martins *et al.*, 1998) Brasilen hitz egiten den portuges hizkuntzako gramatika-erroreak salatzen ditu. REDATOR² izeneko testu-editorean integratuta dago eta bi estrategia jarraitzen ditu erroreak lantzeko. Testuinguruaren mendekoak diren baina *aurrez definitutako patroierregelekin* erraz detekta daitezkeen erroreak erregela lokal eta heuristikoekin lantzen dira. Detekzio leiho zabalagoa duten erroreentzako, adibidez komunztadura-erroreen detektziorako, analizatzaile sintaktiko sendoa erabiltzen da. Analizatzaile honek oinarrian TGG bat du eta murriztapenen *erlaxazioa* erabiltzen da gramatikalak ez diren esaldiekin lan egiteko. Analizatzaile honetan, testuetan maizen agertzen diren egiturak zehazten dituzten erregelak dira lehen erabiltzen direnak. Gramatika-zuzentzaile hau 37 milioi hitzetako corpus batean probatu da. Corpora egunkari, aldizkari eta zuzendu gabeko ikasleen testuekin osatuta dago. Lantzen dituen erroreari dagokionez, sintagmako buruaren eta modifikatzaileen arteko

²Itautec/Philco S.A. enpresako tresna

numero eta genero komunztadura, aditz-kateko buruaren eta subjektuaren arteko komunztadura eta komak aipatzen dira. *ReGra* gramatika-zuzentzaileak % 45,6ko estaldura eta % 31,21ko doitasuna lortzen ditu.

- Arabiera hizkuntza semitikoa da eta morfologia eta sintaxia aberatsak ditu. Arabiera Moderno Estandarra lantzeko sortu da *Arabic GramCheck* gramatika-zuzentzailea (Shaan, 2005). SICStus Prolog erabiliz inplementatu da eta arabieran ematen diren gramatika-errore arruntenak detektatzeko gai da, arazoa deskribatzen du eta zuzenketa-rako proposamenak ematen ditu. Gramatika-zuzentzaileak bi zati ditu: arabierarako analizatzaile morfologikoa eta gramatika-egiaztatzaile batekin hedatutako sintaxi-analizatzailea. Analizatzaile morfologikoa, analisisa egiten duen moduluak, lexiko-mailako desanbiguazioa egiten duen moduluak eta ezaugarriak erauzten dituen moduluak osatzen dute. Sintaxi-analizatzailea, baterakuntzan oinarrituta dago eta bertan sintaxi-erroreak detektatzeko *erlaxazioa* erabiltzen da. Sintagma barruko komunztadurak landu dituzte batez ere *Arabic GramCheck*en, hots, generoa, aditzaren kasua. . . . Ebaluaziorako 100 esaldi motz eta sinpleak (batez beste 4 hitzetakoak) erabili dira, 10 gramatikalak eta 90 ez-gramatikalak. Corpus honekin ez da harritzekoa emaitza onak lortzea: % 96,6ko doitasuna eta % 95,5eko estaldura lortzen dituzte.
- *Erlaxazioaren* teknika erabiltzean, sintaxi-analizatzaileak analisi posible ugari eman ohi ditu, esaldi bat analizatzeko erregela erlaxatu ga-beak eta erlaxatuak erabiltzen baitira. Ingelesa moduko hizkuntzetan, osagaien ordena nahiko finkoa da eta osagai horien arteko komunztadura, eskasa. Euskara, errusiera eta latina moduko ordena askeko eta komunztadura aberatseko hizkuntzetan sor daitekeen analisi kopurua gehiegizkoa izan daiteke. Hau dela eta, Gojenola eta Sarasola-k (1994) (Gojenola, 2000) ezaugarrien erlaxazio mailakatua proposatzen dute. Metodo honetan, maila bakoitzean murriztapen multzo bat erlaxatuko litzateke, erlaxazio-konbinazioak murriztuaz. Hasiera batean, analisisa gramatika osoarekin egiten saiatzen dira. Analisisirik lortzen ez bada, lehenengo murriztapenen multzoa erlaxatzen da. Hala ere, analisisirik lortzen ez bada, bigarren murriztapenen multzoa erlaxatzen da e.a. Maila bakoitzean, errore mota ezberdinak landuko dira, konputazio-koste eta zailtasun gutxienetik hasita, handienera joaz. Euskarako teknika hau aplikatu denean, gramatika idazteko, inolako teoria sintaktikorekin loturarik ez duen eta PATR-II-rekin bateragarria den

formalismo sintaktiko erazagutzaile bat erabili da. Gainera, mailetako bakoitzean zein murriztapen aplikatzea den derrigorrezkoa, eta zein murriztapen erlaxa daitekeen adierazten duen modulu bat sortu da. Erlaxa daitekeen erregela bakoitza aipatzen denean, eman beharreko errore-mezua adierazten da. Sintaxi-analisia, behetik goranzko analizatzaile sinple batekin egin da. Analizatzaile honek, nahiz eta murriztapen bat erlaxagarria izan, murriztapena betetzen ez denean soilik erlaxatzen du (ez beti), horrela sor daitekeen analisi kopurua txikituz. Zuzenketari dagokionez, ordezeko zuzenen artean ordena ezartzeko orduan, errore kopuru minimoa (erlaxatutako murriztapen minimoa) duen analisia dela egokiena ikusi da. Baldintza hau betetzen duten analisi anitz egonez gero, anbiguitasuna ebazteko modua aztertu beharko da edo aukera guztiak aurkeztu. Emaitei dagokionez, 53 esaldirekin erlaxazio-maila ezberdinak aplikatu ondoren, ondorio hontara iritsi dira egileak: maila baxuko erroreak dituzten esaldietan, erlaxazio-maila anitz erabiltzeak emaitza hobeak ematen ditu. Errore konplexuetarako maila asko erabiltzeak, denbora asko eskatzen du tratamendua errepikatzen baita. Errore mota bakoitza detektatzeko maila egokiena zein den aztertzea izango litzateke egokiena. Ez dute sistemaren estaldurarik eta doitasunik aipatzen.

- *Critique*, gramatika- eta estilo-zuzentzaile bat da (Ravin, 1993), zeinek azpian PLNLP³ English Grammar (PEG) (Jensen, 1993) izeneko estaldura zabaleko gramatika duen. 25 gramatika-errore landu dituzte (komunztadura numeroan, aditz-forma okerra, puntuazioa. . .) eta estiloa desagokia egiten duten 85 fenomeno (konplexutasun handiegia, luzera handiegia, formaltasun gutxi. . .). Erroreak detektatzen direnean, dagozkien ordain zuzenarekin ordeztzen dira eta estilo-kontuetan, berriaz, arazoa deskribatu egiten da kode bat erabilita. Bi kasutan azalpen zabal bat jaso daiteke eta, hala eskatuz gero, esaldirako analisi-zuhaitza lortzea ere posiblea da. Inplementazio aldetik ere ongi bereizten dira errorea eta estiloa desagokia egiten dutenak. Gramatika-erroreak analisi sintaktikoa egiten denean detektatzen dira. Errorea dagoenean analisi osoa egitea ezinezkoa denez, gramatikako baldintzak *erlaxatu* egiten dira analisia lortu arte. Estilo kontuak analisi sintaktikoa egin ondoren detektatzen dira horretarako sortutako 300 erregelekin.

³PLNLP: *Programming Language for Natural Language Processing* LNPrako bereziki sortutako programazio-lengoaia da. Honekin batera, garapenerako tresnak eta analisi eta sintesi osagaiak sortu zituzten ikerlari talde batek.

Microsoft©-eko gramatika-zuzentzaileak Critique sisteman oinarrituta daude (Riley *et al.*, 2004).

- Paggio-k (2000), Eskandinavian testu-edizioari dagokionez sortzen diren beharrak asetzen saiatzen den *SCARRIE* proiektuaren barnean, daniera hizkuntzarako gramatika-zuzentzailea garatu zuen. Danierarako atalean, ohiko zuzenketa ortografikoa eta gramatika-zuzenketa konbinatzen dira. Ortografia-mailako zuzenketa, egunkariekin osatutako corpus batetik erauzitako 251.000 hitz-forma dituen hiztegi bat erabiliz egiten da. Honi hitz anitzeko terminoak dituen zerrenda bat gehitu zaio. Ordain zuzenak emateko, ordezkapenak egiten dituzte, eta horretarako, trigramen eta trifonemen analisisian oinarritzen den algoritmo bat erabiltzen dute. Testuinguruaren mendekoak diren erroreak detektatzeko (sintaxi-mailakoak), testua Tomita-ren algoritmoan (Tomita, 1986) oinarritzen den analizatzaile sintaktiko batekin analizatzen dute. Honi erroreak detektatzeko modulu bat gehitu zaio. Modulu honetan, TGG hedatu bat erabiltzen da, erregeletan errore-pisuak eta errore-mezuak dituen. Errorea baterakuntza egitean ezaugarriak bat etortzen ez direlako gertatzen bada (adib. komunztadura-erroreetan), ezaugarria ez da kontuan hartzen (*erlaxazioa*) eta hiztegian ordezeko bat bilatzen da. Errorea egitura ez-osoia izateagatik gertatzen bada (adib. aditzaren forma nahastu da eta aditz-sintagmak ez du bururik), errore-mezu bat ematen da. Lehendabizi, egitura zuzenak definitzeko erregelak aplikatzen dira, eta gero, erroreak deskribatzen dituztenak. Landutako erroreak honakoak dira: hitz arruntetan eta konposatuetan ematen direnak, izen-sintagmen barneko komunztadura eta aditz-kateetan ematen direnak dira. Sistemaren ebaluazioan estaldura egokia lortu da (% 58,1) baina, doitasuna oso baxua (% 20,6). Honen arrazoia hitz asko sistemarentzako ezezagunak izatea da (izen bereziak). Hitz horiek kenduta % 34,9ko doitasuna lortuko litzatekeela diote egileek.
- Foster eta Vogel-ek (2004b) *chart*-etan oinarritutako behetik goranzko (*bottom-up*) analizatzaile bat erabiltzen dute sarrerako esaldia analizatzeko. Analizatzailea sintaxi-analisi bat lortzeko gai ez bada, adibidez esaldia gaizki osatuta dagoelako, *errore-gramatika* bat erabiltzen dute esaldiaren analisi bat lortzen saiatzeko. Sintaxi-analizatzailearen ondoren aplikatutako errore-gramatika, 1113 erregelekin osatutako testuingururik gabeko gramatika da. Errore-erregelak gramatika arruntetik eratorri dira baina testu erroredunetako informazioa kontuan hartuta. Gramatika hau, berreskuratze-algoritmo sendo batek aplikatzen du.

Honek *chartean* errore-erregelari dagozkien ertzak gehitzen ditu banan-banan eta analisirako behetik goranzko algoritmoari dei egiten dio analisirik lortzen ote den aztertzeke. Hau errore-erregela guztiekin egiten da baina aldiro errore-erregela bakarria erabiltzen denez, bi errore dauzkaten esaldiak ezin dira analizatu. 50 esaldirekin egin da proba (esaldiek 4-35 hitz inguru dituzte) eta hauetatik % 84ean analisisia lortu da.

- AbiWord⁴ testu-prozesadore irekiak, bere 2.4. bertsioan, *esteka-gramatika* bat erabiltzen du ingeleserako gramatika-zuzentzailean. Funtzionamendu zehatza ez da aipatzen, baina tresnaren egileek diotenaren arabera, zuzentzaileak esteka-gramatika oinarri duen analizatzaile sintaktiko bat du azpian. Honek esaldi bat analizatzen ez duenean, okerra dela suposatzen da. Ez da aipatzen, ez zein errore lantzen den, ezta zein emaitza lortzen dituzten ere.
- Brehony eta Ryan-ek (1994) sarrerako esaldia *esteka-gramatika* bat erabiliaz aztertu ondoren, postprozesua egokitu egiten dute frantses hiztunek ingelesez egiten dituzten estilo-desegokitasunak detektatzeko. Detektatu nahi den fenomeno mota bakoitzeko hiztegi definizioa aldatu egin behar da, esaldiak onartuak izan daitezen. Aurrekoan bezala, kasu honetan ere, ez da ez landutako errore motei, ezta lortutako emaitzei buruzko informaziorik ematen.
- Txekierarako sortutako zuzentzaile gramatikalak (Holan *et al.*, 1997) oinarrian *mendekotasun-zuhaitzak* eraikitzen dituen analizatzaile bat du eta 3 moduluk osatzen dute: analisi morfologikoa eta lexikoa egiten duen modulua, gramatika-egiaztatzailea eta ebaluatzailea. Gramatika-egiaztatzailean sarrerako esaldia jaso eta analizatu egiten da honako hiru emaitza posible lortuz:
 1. Analisisia arrakastatsua da eta ez da inkonsistentziarik topatu sintaxiari dagokionez.
 2. Analisisia arrakastatsua da baina sortutako analisi-zuhaitz guztiek gutxienez inkonsistentzia bat dute, beraz, emaitza modulu ebaluatzaileari pasako zaio.
 3. Analisisia ez da posible izan (segur aski gramatika osoa ez izateagatik).

⁴<http://www.abisource.com>

Analisia egiteko sistema honek azpian *Robust Free Order Dependency Grammar* izeneko formalismoa jarraitzen duten metaerregelen osatutako gramatika du. Erregela hauen bidez ondoz ondoko harreman sintaktikorik ez duten egiturak eta egitura sintaktiko okerrak deskribatzen dira, baita egitura zuzenak ere. Inkonsistentziaren bat deskribatzen duen metaerregelaren bat erabiltzen bada esaldia analizatzerakoan, orduan eraikitako zuhaitzean ikur negatibo bat txertatzen da. Erregela hauetan murriztapenen *erlaxazioa* egiteko aukera dago. Analisia hiru fasetan egiten da:

1. *Ondoz ondoko harremanak eta positiboak (positive projective)*. Analizatzaile estandar batek sarrerako esaldia inolako ikur negatiborik ez duen eta ondoz ondoko harremanak mantentzen dituen sintaxi-zuhaitz baten bidez adierazi dezakeen aztertzen du. Zuhaitzak ikur negatiboa izango du, analisia egin ahal izateko murriztapenen bat erlaxatzeko beharra egon bada edo *erroreak aurreikusten dituen erregela* bat aplikatu baldin bada. Ikur negatiborik gabeko zuhaitzak dituzten esaldiak zuzenak dira.
2. *Ondoz ondoko harremanik ez eta positiboa & ondoz ondoko harremanak eta negatiboa (positive nonprojective & negative projective)*. Fase honetan ikur negatiboak edo ondoz ondokoak ez diren harremanak dituzten analisi-zuhaitzak bilatzen dira. Horrelako zuhaitzen bat existitzen bada, ebaluazio-moduluak erabakiko du errorerik ote dagoen edo ez.
3. *Ondoz ondoko harremanik ez eta negatiboa (negative nonprojective)*. Fase honetan ondoz ondoko sintaxi-harremanik ez, eta aldi berean ikur negatiboak dituzten analisi-zuhaitzak bilatzen dira. Horrelakorik topatzen bada, errore-mezu bat itzuliko da.

Faseetan egindako analisiak sistema moteldu egiten du eta gainera maiz, esaldi bererako aldaera asko sortzen dira. Aldaera kopurua murrizteko eta era berean sistema azkartu ahal izateko, aukera ezberdinak erabili dituzte egileek. Hala nola, ambiguitasun handia sortzen duten laburdurak dagozkien forma osoarekin ordezteak, leiho txikiko zenbait egituretan errorea modu lokalean bilatzea zuhaitza osoaren analisiarekin hasi aurretik, zuhaitzaren egitura aldatzen ez duten hitzak ezabatzea. . . Azkenean, eraginkortasun ezaren arrazoia modifikatzaile askeen eta adjuntuen kopurua zela ondorioztatu zuten egileek. Hau dela eta, analisirako gramatika mailaka antolatzea erabaki zuten. Lehen mailatan hurbileko osagaiak lantzen dituzten erregelak egongo dira eta

azkenekoetan, leiho zabalagokoak. Egileek sistema prototipo moduan aurkezten dute, inolako ebaluaziorik egin gabe. Ez dute zein errore mota landu nahi duten deskribatzen.

- Errusieran hitzak forma ez-zuzenetan erabiltzea oso ohikoa da eta fenomeno honi komunztadura-errore deitzen diote. Errore hauek zuzendu ahal izateko *mendekotasun-zuhaitz* moduan gauzatzen diren mendekotasun-egiturak erabiltzen ditu Mitjushin-ek (1996). Analizatu beharreko esaldia gaizki osatuta dagoenean eta, beraz, ezin zaionean egitura sintaktiko bat esleitu, sistemak aldaketa minimo batzuk egiten ditu egitura osatu ahal izateko. Aldaketa horiek zuzenketa posible gisa aurkezten dira. Hots, ez da diagnosirik egiten, zuzenean sarrerako esaldia zuzendu egiten da. Esaldiak *sintaktikoki konektatuta* daudela esaten dute, ongi osatutako mendekotasun-zuhaitz bat osatu ahal bada. *Sintaktikoki ez-konektatuta egotearen maila C* zenbaki batekin definitzen dute. C konektatutako segmentuen kopurua da, eta horrela, C=1 denean, egitura sintaktiko osoa sor daiteke. Komunztadura-erroreen kasuan, lehenengo hitz bat aldatzen da eta esaldia analizatzen saiatu. Posiblea ez bada, bi hitz aldatzen dira e.a. esaldiaren analisia lortu arte. Hitzak, hitz beraren *aldaera multzo* batetik ateratako aldaera batekin ordeztzen dira. Aldaerak hitzaren numeroa, pertsona eta antzeko ezaugarri morfologikoak aldatuta lortzen dira. Mendekotasun-zuhaitza eraikitzeke, behetik goranzko algoritmo bat erabiltzen da egitura eraikitzeke erregela batzuekin batera. Oinarrian dagoen informazio linguistikoa urria denez, maiz gertatzen da sistemak esaldi zuzenei dagozkien zuhaitzak ezin eraiki izana. Hau dela eta, C-ren balioa berriro kalkulatu dute. Honela, gaizki eratutako esaldien % 90arentzako zuzenketa egokia lortu da. Zuzentzaile honek informazio *positiboa* soilik erabiltzen du.
- Courtin *et al.*-ek (1991) frantsesez idatzitako testuetan lexiko- eta sintaxi-mailako erroreak detektatu eta zuzendu ahal izateko lana egin dute. Zuzenketari dagokion lana III kapituluko III.3.1 atalean azaldu dugu eta honakoa, sintaxian egin duten lanari eskainiko diogu. Komunztadura-erroreak landu dituzte batez ere, *mendekotasun-zuhaitzei* baldintza-ekintza moduko erregelak aplikatuaz. Mendekotasun-zuhaitza eraikitzeke erabiltzen den sintaxi-analizatzailea PILAF (Courtin, 1977) sistemarako garatutakoa da eta honako moduko erregelak ditu gramatikan:

(84)

Burua * mendekoa := pisuen zerrenda

Adib.:

subc * det := -32⁵

subc * adjq := -16, -15

Pisuak mendekoak buruaren azpian duen posizioa adierazten du, non pisu negatibo batek ezkerrean kokatzen dela adierazi nahi duen eta pisu positiboak, eskuineko aldean. Formalismoa sinplea da baina ez du aipatutako posizioaz gain testuinguruaren inguruko inolako kontrolik ahalbidetzen. Erregela hauek zuhaitz baten eraketa ahalbidetzen dute baina adierazpen-muga handiak dituzte. Muga hauek direla eta, koherentziarik gabeko zuhaitz-egitura franko sortzen dira. Arazo hau konpontzeko, kategoria lexiko bakoitzeko, zein zuhaitz-egitura diren posibleak zehazten duten *konfigurazioak* gehitu dira. Esaldi bati dagozkion mendekotasun-egiturak lortu ondoren, sintaxi-egiaztatzailearen lana hasten da. Sintaxi-egiaztapenerako moduluak bi egitura erabiltzen ditu: esaldi batekin lotutako mendekotasun-zuhaitza eta *sintaxia egiaztatze*ko erregelak. Sintaxia egiaztatze

- CLAWS (*Constituent Likelihood Automatic Word-tagging System*), estatistika oinarri duen analizatzaile morfologikoa da. Ez du analisi osoa egiten, Markov-en eredu ezkutua erabiltzen du hitz bati etiketa-klase bat esleitzeko. Atwell-ek (1987), etiketatzailea alda daitekeela dio, gramatika-erroreak detektatu ahal izateko. Sistemak erabilitako teknika

⁵Mendekoa den determinatzailea subjektuaren ezkerrean.

III.2.2.1 atalean azaldu dugu. Sistemak errore dezente omen du baina abiaduraz azkarra omen da, eta ezer baino hobe, egilearen arabera.

- Funtsean aurrekoaren ideia bera jarraitzen duen baina informazio linguistikoa gehitzen duen sistema hibrido baten berri ematen dute Bigert eta Knutsson-ek (2002). Egileek hitz zuzen erabilera nahasketak eta hitz-konposatuen banaketa okerraren ondorioz sortutako erroreak detektatzen dituzte. Lan honetan ideia nagusia honakoa da: etiketa morfosintaktiko sekuentzia arraroak errorea izan daitezkeela suposatzen dute. n-gramaren maiztasun txikiaren arrazoietakoa bat datu-urritasuna izan daiteke. Beraz, etiketa maiztasun handiagoko batekin ordeztzen da eta berriro kalkulatzeko n-gramaren maiztasuna. Hau oraindik txikia bada, errore bat detektatu da.

Markov-en eredu ezkutua erabiltzen duen analizatzaile bat erabili dute n-gramak ikasteko eta corpora morfologia-mailan analizatu eta desanbiguatu dute. Informazio hau soilik jaso duen metodo ez-gainbegiratu bat erabili dutenean, erroreen detekzio-maila ona lortu dute, baina alarma faltsu kopuru handia. Sintagmei buruzko informazioa gehitu diotenean, alarma faltsuen kopurua asko jaitsi da. Sintagmak identifikatzeko, erregetan oinarritutako sistema bat erabili dute. Ezagutza linguistikoa gehitzeak onura izugarria suposatu du errore detekzioaren doitasunari dagokionez. Ez dute erroreak detektatzeko atazari buruzko ebaluazio-emaitzarik ematen.

- Aurreko puntuan azaldutako lanaren antzekoa egiten du Sjöbergh-ek (2005) unitate moduan osagai sintaktiko sinpleak (izen-sintagmak, aditz-kateak...) erabiliaz. Testu berri bat zatitzaile batekin analizatzen du eta hori zuzenak diren zatiekin konparatu. Zati-sekuentzia arraroak errore kontsideratzen dira. Zati mota desberdinak gutxi direnez eta testu zuzenetan agertzen direnez, datu urritasunaren arazoa txikiagoa da, baita alarma faltsuena ere. Alarma faltsu gutxi detektatzearekin batera, errore gutxi detektatzen dira. Hauen artean, koma toki okerrean ipintzea edo ez ipintzea, erroreak hitzen ordenan, hitz zuzenen nahasketak e.a. daude. Teknika hau oso merkea omen da, etiketatu gabeko corpora eta zatitzaile bat edukitzea nahikoa baita. Aurkeztutako lanean emaitzak ez dira argi azaltzen.
- *Estatistika* oinarri hartuta, hiztegi zehatz baten *erabilera ez-egokia* detektatzen dute Chodorow eta Leacock-ek (2000). Horretarako, *Test of English as a Foreign Language, (TOEFL)-eko* idazlanak jasotzen dituzte eta ALEK (*Assessing Lexical Knowledge*) izeneko sistema esta-

tistikoa aplikatu. Ikerketa honen helburu nagusietariko bat, ebaluatu nahi duten hitz bakoitzeko errore-corpusa biltzearen lana ekiditea dela diote egileek. Hau lortzeko, ALEK ingeleseko corpus orokor batean entrenatzen dute. Gero, erabilera okerrak topatzeko, hitz okerraren testuinguru lokalean topatutako aztarnak, erabilera okerrari dagokion eredia, eta hitz horrentzako testu zuzenetan topatutako eredia konparatzen dituzte. Estatistikaren alderdiari dagokionez, *mutual information* teknika erabili dute oso gutxitan gertatzen diren hitzen bigramen eta kategoría gramatikalen trigramen konbinazioak topatzeko. Proba 20 helburu hitzekin egin dute, 30 milioi hitzetako corpus zuzenetik abiatuta lortu dituzte erabilera zuzenei dagozkien ereduak eta erabilera okerra detektatzeko proba 8000 esaldirekin egin dute. Beharrezkoak ez diren hitzen, falta diren hitzen, aldatu behar diren hitzen eta gaizki erabilitako hitzen ondorioz sortutako erroreak detektatzen ditu ALEK sistemak. Hitz zuzenekin ematen diren erabilera nahasketak ere detektatzen ditu. % 80 inguruko doitasuna lortzen dute eta % 20ko estaldura, baina alarma faltsu dezenterekin.

- Suedieraz idatzitako testuetan erroreak detektatzeko eskuzko lanik behar ez duen teknika aurkeztu dute Sjöbergh eta Knutsson-ek (2005) beraien lanean. Corpus erroreduna lortzeko, idatzizko testuetan automatikoki sortzen dituzte erroreak, eta hau egin ahala, testuak erroreekin markatzen dituzte. Markatutako eta TnT etiketatzailer estatistikokoaren bidez kategoría gramatikal mailan etiketatutako testu hauek, *erroreen ikasketan entrenatzeko* corpus moduan erabiltzen dira. Ikasketarako, errore bakoitzerako modulu bat entrenatu dute zuzenketak proposatu ahal izateko. Ikasketarako, *transformazioan oinarritutako ikasketa* (Brill, 1995), fnTBL, metodoa erabili dute. Honek, erraz ulertzen diren erregelak ikasten eta sortzen ditu. Metodoa suediera, alemaniera eta antzeko hizkuntzetan ohikoak diren gaizki zatitutako hitz-konposatuekin eta hitzen ordena okerrarekin probatu dute, baina teknika edozein errore motetara hedagarria dela diote. Eskuz sortutako erregelekin lortutako emaitzen antzekoak lortzen dira, baina doitasun baxuagoarekin eta estaldura handiagoarekin. Emaitzak ez dituzte modu estandarrean ematen.
- Aurreko kasuan bezala Örvar Kárasen-ek (2005) gaizki zatitutako hitz-konposatuekin egiten du lan, baina kasu honetan islandierarako eta *memorian oinarritutako ikasketa*⁶ erabiliz. Alarma faltsuekin arazoak

⁶Memory-based Learning

ditu eta hauen kopurua jaisteak detektatutako errore kopurua ere jais-tea suposatzen du. Azkenean, alarma faltsu kopuru onargarri bat lor-tzen denean, errorean % 65a detektatzen du.

- Ingelesa bigarren hizkuntza moduan ikasten duten ikasle japoniarren ahozko etorria transkribatu ondoren, ikasketa automatikoa erabiltzen dute Izumi *et al.*-ek (2003) gramatikako eta lexikoko erroreak detekta-tzeko. Transkribatutako milioi hitzeko ikasle-corpusa, eskuz etiketatu dute errore bakoitzerako gramatika-kategoria, gramatika-lexiko siste-ma eta forma zuzendua idatziz. Erroreak bi multzotan banatu dituzte egileek. Alde batetik, *ez-egite* erroreak daude zeintzuetan beharrez-ko hitza falta den eta hori adierazten duen etiketa bat ipintzen den. Bestetik, berriz, *ordezkatze*-erroreak daude, eta hauetan, hitz okerra etiketa batzuen artean txertatzen da zuzenketaren orduan bere forma zuzenarekin ordezteko.

Ikasketa automatikorako, *entropy handieneko eredua* (Jaynes, 1957) erabili dute. Teknika hau, datuen banaketa probabilitatea estimatze-ko erabiltzen da eta ezer ez denean ezaguna, datuen banaketak aha-lik eta uniformeena izan behar duela dio. Hau gertatzen denean, en-tropiaren balio maximoa lortzen da. Aipatutako artikuluan, *ez-egite* erroreekin eta *ordezkatze*-erroreetako bakoitzarekin bi metodo proba-tu dituzte. Sinplifikatzeko, lehenengoan, errore-kategoria bakoitzeko errorea *gertatzen da*, *ez da gertatzen* klaseetan sailkatzen dute. Biga-rren metodoan, n erroreetarako egiten dute proba eta gertatzeaz gain, errorea zein kategorietakoa den estimatzen saiatzen dira. Zuriuneak eta hitzak klase hauetan sailkatzeko, sailkatzaileari ematen zaizkion testuinguru-ezaugarriak 23 dira. Hauen artean, aurreko eta ondorengo bi hitzak, hitzen konbinazioak eta ondorengo hitzaren lehen eta az-ken hizkiak daude. Teknika honekin, % 30eko estaldura lortzen dute eta % 50eko doitasuna. Corpusa esaldi zuzenekin eta era artifizialean sortutako erroreekin aberastuta, % 80eko doitasuna lortzen dute, es-taldura mantenduaz.

- Hitz zuzenen erabilera nahasketak desanbiguazio-ataza moduan ikus-tea erraza denez, errore hauek detektatzeko ikasketa automatikoa eta *nahasketa multzoak* erabiltzea ohikoa da. *Nahasketa multzoetan* de-sanbiguazio lanak egiteko sailkatzaile Bayesiarrak eta erabakitze-ze-rrendak erabili zituen Golding-ek (1995) bere lanean. Geroago, Carl-son *et al.*-en lanean (2001) SNoW sailkatzailea eta eguneraketetarako *Winnow* erregela erabiltzen dituen *ikasketa automatikoko* teknika bat

probatu zuten. Testu eta ezaugarri multzo berak erabiliz bi teknika hauek konparatu direnean, (Golding eta Roth, 1996), *Winnnow* erabiltzen duen sistemak errendimendu hobeak erakutsi du. Azken sistema honek ezaugarri kopuru oso handia eta dagozkion pisuak ikasteko gaitasun handia du eta gaitasun honek performantzia asko hobetzen omen du (Mangu eta Brill, 1997; Golding eta Roth, 1999).

- Ezaugarri eta pisu multzo ilunak lortu beharrean, ulergarriak diren erregela sinple gutxitan harrapatzen dute testuetako ezagutza linguistikoa Mangu eta Brill-ek (1997). *Nahasketa multzoen* desanbiguazioa dute helburu hauek ere, baina aldi berean, era errazean uler daitekeen ezagutza jaso nahi dute. Honetarako, *transformazioan oinarritutako ikasketa* erabiltzen dute erregelak ikasteko. Adibidez, *piece*, *peace* nahasketa multzoarentzako honako moduko erregelak ikastea lortu dute “Change *peace* to *piece* if the previous word is *every*”. Zenbait esperimendu egin dituzte eta honako ondorioetara iritsi dira: Mangu eta Brill-en (1997) lanean, baldintza beretan, *Winnnow* erabiltzen duen sistemarekin konparatuz gero, emaitza okerragoak lortu dira. Ikasketarako leiho txikiagoa erabiltzen badute eta analizatzaile morfologiko bat gehitu, *Winnnow*ekin lortutako emaitzak lortzera iristen dira. Hori bai, ikasitako informazioa tamainan askoz txikiagoa da eta ulertzeko askoz ere errazagoa.
- Kumar eta Nair-en (2007) lanean corpusetan oinarritzen den teknika bat aurkezten dute hurbilpen berritzaile batekin: immunologia-sistema artifizialeko (*Artificial Immune System (AIS)* ingelesez) oinarriak erabiltzen dituzte gramatika-erroreak detektatzeko. Erroreak testuetako patogenoak bailiran tratatzen dituzte, eta antigorputz detektatzaileak eraikitzen dituzte hauek detektatzeko.

Gizakion immunologia-sistema linfozito izeneko zelula immunologiko batzuetan oinarritzen da. Zelula hauek batez ere B eta T motatakoak dira eta patogenoen patroi antigenoak ezagutzeko gai diren molekula hartzaileak dituzte.

Egileek AISeko algoritmoak hizkuntzaren tratamendu automatikora egokitu dituzte. Haienezako, corpora esaldi zuzenen multzo batek osatzen du eta zelula multzo bezala ikus dezakegu. Corpora analizatu egiten da eta bertako hitzen kategoria-gramatikalen bigramekin, trigramekin eta tetragramekin egiten da lan. Detektatzaileak corpusean aurkitzen ez ditugun egiturak dira (n-gramak) eta errore moduan kontsidera daitezke (corpusean agertzen ez diren egitura zuzenak ere, erro-

retzat jotzen dira). Antigenoak (errore gramatikalak) harrapatzeko, antigorputz detektatzaileak sortu behar dira. Hauek, ausaz sortzen dira eta gero corpus zuzenaren kontra probatzen dira. Corpusean agertzen ez diren n-gramak, antigorputz detektatzaile kontsideratzen dira. Erroreak dituen beste corpus batean detektatzaileak, patogenoekin (errore gramatikalekin) parekatzen direnean, errorea detektatzen da. Teknika honekin detekta daitezkeen zortzi errore mota topatu dituzte soilik, eta gramatika-liburu batetik ateratako esaldiak erabili dituzte probarako, hots, laborategiko esaldi motzak. Landutako errorearen artean subjektuaren eta aditzaren arteko komunztadura, aditzaren denboraren aukeraketa, artikulua gaizki erabiltzea e.a. daude. Artikuluak AISi buruzko informazio asko ematen du, ez ordea, hau errorearen detekziora aplikatzeko moduari eta emaitzei buruz.

A.2 OLHIA sistemak

Gramatika-zuzentzaile asko hizkuntzen irakaskuntzarako sistemen parte dira. OLHIA sistemetan errorearen detekzioa egiten duen modulua sistemaren zatia izaten da soilik, eta sistema batzuetan garrantzia handia ematen zaion bitartean, beste batzuetan ez da errorearen inolako detekziorik egiten. Beste ataza batzuei, edukien kudeaketari, irakaste-estrategiei eta batez ere, ikaslearen ezagutza eta berarekin sistemak eduki dezaketen berrelikadurari ematen zaio garrantzia.

OLHIA sistemei eskainitako atal honetan erroreak lantzen dituzten zenbait sistema aurkeztuko ditugu.

- Kanadan garatutako *CLEF* sistemak (*Computer assisted Learning Exercises for French*), 62 gai, 62 gramatika-azalpen eta mota desberdineta-ko 250 ariketatik gora dauzka bilduta (hutsuneak betetzeko ariketak, aukera anitzeko galderak...). Sistemak oinarritzko frantsesa irakastea du helburu. Vandeventer-ek (2003) bere tesi-lanerako sistema hau probatu zuenean, *patroi-erregelen* erabilera zentzuduna egiten duela iritzi zion. Probarako, ama-hizkuntza frantsesa zuen erabiltzaile bat ipini zuen erantzun okerrak ematen eta sistemak ematen zuen berrelikadura neurtzeaz gain, sistemaren mugak gainditzen saiatu zen. Azken hau oso zaila suertatu zitzaion, sistemak oso ingurune mugatua baitu.
- *VINCI* sisteman (Levison *et al.*, 2000) *patroi-parekatzearen* eginbehera, aurrekoan bezala, sistemak sortutako emaitza zuzena ikasleak emandakoarekin konparatzea da. Eginkizun honetarako, bi mailako algorit-

mo bat aplikatzen da zeinek hurbileko karaktere-kateen patroiparekatzea erabiltzen duen. Modu honetan, hitzen txertatze-ezabatzeak, hitzen ordenari dagozkion erroreak, errore fonologikoak e.a. lantzen dituzte. Ez dute sistema ebaluatu ote duten adierazten.

- Aurreko atalean aipatu berri dugun suedierarako Granska gramatika-zuzentzailea, hizkuntza-ikasleen erabilerarako egokitua izan da *CrossCheck*⁷ izena jasoaz (Kann, 2002). Erroreen detekzioarako teknika hibridoa azaldu berri dugunez, erabiltzaile berri hauentzako egindako egokitzapenaren inguruan arituko gara orain. Granska, ikasleek idatzitako corpus batean probatu denean, erroreen % 30a soilik detektatu du. Honen zergatia, ikasleek egin ditzaketen erroreen berezitasunean eta maiztasunean egon daiteke. Hau dela eta, ikasleek bereziki egin ditzaketen erroreen sailkapen bat egin da eta errore zehatz hauek deskribatzen dituzten *errore-erregelak* gehitu zaizkio jatorrizko Granskaren errore-gramatikari. Honetaz gain, ikasleek aldezturik ezagutu ezin diren erroreak sor ditzaketenez, *probabilitatea eta heuristikoak* erabiltzen dituzten algoritmoak aplikatu dira ikasle-corpusetan errore hauek detektatzeko.
- *Grim* (Bigert *et al.*, 2005), suedieraz idatzi nahi duten idazleentzako hizkuntzaren ikasketarako ingurune bat da. Bertan, teknika ezberdinak erabiltzen dira errore mota ezberdinak detektatzeko. Ingurune honetan, *Granska* gramatika-zuzentzailea hibridoa, hitz zuzenen erabilera nahasketak detektatzeko erabiltzen den *ProbGranska* algoritmo probabilitistikoa (Bigert eta Knutsson, 2002), eta azkenik, ikasketa automatikoa erabiltzen duen *SnalGranska* tresna (Sjöbergh eta Knutsson, 2005), integratuta daude edo integratuko dira (hauek guztiak A.1 atalean aipatu ditugu). Hiru metodo erabiltzean, estaldura eta zehaztasuna hobetzen dira eta erabiltzaileari berrelikadura modu ezberdinetan ematea posiblea da. Emaitzak xehatuak ematen dituzte, eta ez, modu estandarrean (estaldura eta doitasuna zehaztuaz).
- Ameriketako Estatu Batuetako armadako ikerketarako institutuak (ARI), hizkuntzen ikasketarako *BRIDGE* eta bere ondorengoa den *MILT* sistemak garatu ditu (Weinberg *et al.*, 1995; Holland *et al.*, 1993). Alemaniera, arabiera eta espainiera (MILTen) erakusten duten sistema hauetan, erroreen detekzioari eta diagnosiari dagokion atala sistema osoaren zati bat besterik ez da. Sistema hauek erabiltzen duten anali-

⁷www.csc.kth.se/tcs/projects/xcheck/index-en.htm

zatzaileak, oinarria Chomsky-ren (1981) *government and binding* teorian dauka. Ingurune honek sistemaren eraikuntzarekiko independenteak diren formulak erabiltzen ditu. Formula hauek parametrizagarriak diren moduluetan multzokatuak daude. Analizatzaileako erabilitako algoritmoari dagokionez, Prolog programazio-lengoaia erabiliz implementatutako laburtze-desplazamendu algoritmo eraldatu bat erabiltzen dute. Sistemak modu *erlaxatuan* egiten du beti lan zenbait errore detektatuz. *Interruptore* edo *marka* batzuk ditu eta hauek *on* edo *off* ipinita ezaugarrien erlaxazioa kontrolatzen da, errore gehiago edo gutxiago detektatu ahal izateko. Alemanierako tutorean, komunztadura-erroreak detektatu dira batez ere (subjektu-aditza, determinatzaile-izen, aditz-preposizioa...), erroredun esaldiak (64 esaldi) % 95,3 ongi detektatuz eta errore-mezu egokia emanez. Esan beharra dago probetarako erabilitako esaldiak ez direla errealak.

- *Grammar Debugger* (Si-Qing eta Luomai, 1990) ingelesa ikasten duten ikasle txinatarrentzako sortutako analizatzailea da. Oinarrian *Parsifal* (Marcus, 1980) analizatzailea erabiltzen du zeinetan erregelak pakeetan multzokatzen diren, eta gero, lehentasun baten arabera ordenatzen diren. Erregelak baldintza-ekintza modukoak dira. *Parsifal* analizatzailea ez da gai esaldi erroredunak analizatzeko eta asmo honekin Charniak-ek (1983) *Paragram* analizatzailea sortu zuen. *Grammar Debugger* eta *Paragram* erregelak *erlaxatzeko* moduan desberdintzen dira. *Grammar Debugger*en, erregeletako baldintza atala bitan zatitzen da: oinarrizkoa eta bigarren mailakoa. Erregela bat aplikatzeko, oinarrizko baldintza nahitaez bete behar da, bigarren mailako baldintzak aukerazkoak direlarik. Sarrerako testuak bigarren mailako baldintza betetzen badu, sarrera zuzena da; betetzen ez badu, ordea, analisisa egiten da, eta aldi berean errore-mezu bat ematen da. Baldintzetako bigarren mailako atala da, beraz, erlaxatzen den zatia. Sistema honek aditza-subjektu komunztadura-erroreak, determinatzailearen aukeraketa eta aditzaren moduari dagokion erroreak lantzen ditu.
- Menzel eta Schröder-en sistema (Menzel eta Schröder, 1998, 1999) erroreen diagnosirako erabilia izan da zenbait OLHIA sistemetan. Sistema honek *analisi ezabatzailea* erabiltzen du. Analisisan sarrerako esaldiaren adierazpen guztiz ambigua egiten da egitura posible guztiak sortuz, eta gero hauen artean zenbait egitura *ezabatu* egiten dira. Ezabaketarako, *mendekotasun-egituretan* zenbait murriztapen bete daitezzen eskatzen da. Murriztapen hauek betetzen ez dituzten egiturak

ezabatu egiten dira bakar batekin geratu arte. Analisia egiturazko desanbiguazio lan bihurtzen da, zeinak murriztapen-asetze teknikak erabiltzen dituen. Analizatu beharreko esaldia erroreduna denean ez da posible izango murriztapen guztiak aldi berean asetzea. Murriztapen horiek partzialki betetzen badira, ordea, analisiren bat lor dezakegu. Murriztapenetako bakoitzari 0 eta 1en arteko pisu bat ematen zaio murriztapena hausteak duen garrantzia adieraziz. 0 pisua duten murriztapenak, gogorak, ezin dira inoiz hautsi. $0 < \text{pisua} \leq 1$ artekoak, komunztadura adierazten dutenak adibidez, errore-kasuak adierazteko balio dute eta esaldi erroredunak analizatzeko hautsi daitezke. Hau gertatzen denean, errore-mezu bat igortzen da. Erroredun esaldi bakarrarekin egin dute proba. Esaldiak errorerik ez edo bakarra duenean esaldia beti analizatzen da ongi. Bi errore baino gehiago dituenean, kasuen % 90ean. Hiru errorerekin % 60ean, eta gehiagorekin, ez da analisirik lortzen.

- *ICICLE (Interactive Computer Identification and Correction of Language Errors)* sistemak (Schneider eta McCoy, 1998), behetik goranzkoa den eta chart-ak erabiltzen dituen analizatzaile sintaktiko bat du oinarrian. Honek errore sintaktikoak detektatzen ditu ASL-ren (*American Sign Language*) hiztun osoen idatzietan ingelesa bigarren hizkuntza moduan ikasten dutenean. Gorrek errore bereziak egiten omen dituzte, eta errore zehatz horiek detektatzea du helburu sistema honek. Besteak beste, aditzaren eta subjektuaren arteko numero-komunztadura, determinatzailea ez idaztea edo oker idaztea... Sistemak *erre-produkzioak* edo *mal rules* deiturikoak erabiltzen ditu. Lan honek, lehenago aipatutako Foster-en lanarekin duen diferentzia nagusia, errore multzo desberdina lantzeaz gain, honakoa da: ICICLE sisteman, erregela arruntak eta erroreak deskribatzen dituzten erregelak batera aplikatzen dira analisisian, eta gerta daiteke, esaldi zuzenak aztertzeko errore-erregelak aplikatzea eta hauek esaldia okertzat ematea. Erregelak aplikatzerakoan garrantzi handia du ikasle-ereduan gordetako informazioak, lehendabizi ikaslea momentu horretan ikasten ari den egiturei buruzko erregelak aplikatzen baitira. Ebaluazioa egiteko, 79 esaldi erroreduneko eta 101 esaldi zuzeneko corpusak erabili dituzte. Corpuseko esaldietan perpausak eskuz banatu dituzte eta errore ortografikoak zuzendu egin dituzte. Esaldi erroredunen % 56a esperotako errorearekin markatu du sistemak, % 29ari ez dio analisirik eman eta % 29an ez du errorerik markatu.

- Hizkuntza-ikasleentzako *Arboretum* tutoreak (Bender *et al.*, 2004) ikasleen testuen analisirako *the English Resource Grammar (ERG, (Flickinger, 2000)* gramatika erabiltzen du LKB analizatzaile eta sortzailearekin batera (Copestake, 2002). Testu hauetako erroreen detekziorako, diagnosirako eta zuzenketarako, gramatika ICICLE sisteman aipatu ditugun *mal rules* deritzanekin hedatu dute. Errore-erregela hauen bidez, linguistikoki okerra den sarrerako testuarentzako adierazpen semantiko zuzena sortzen da. Ongi eratutako adierazpen semantiko honetatik abiatuta, Arboretum sistema erroreari buruzko diagnosia egiteko gai da, baita erroreari dagozkion ordain zuzenak sortzeko ere. Forma zuzenak sortzea konplexua suertatzen da, bai sorkuntzan, bai analisisian, anbigüotasun handia sortzen delako. Arazo honi aurre egiteko, *sorkuntza lerrokatua* izeneko estrategia garatu dute. Matthews-ek (1993) eta Menzel eta Schröder-ek (1999) diote, gramatika bati *errore-produkzioak* gehitzea ez dela OLHIA sistemetan gramatika-zuzenketa egiteko modu egokia, ikasleek egin ditzaketen errore posible guztiak aurreikustea ia ezinezkoa baita; erlaxazioaren alde egiten dute. Ez dute erroreen detekzioari buruzko emaitzik aipatzen.
- *ALICE* (Evans eta Gates, 1991), OLHIA sistemak garatzeko tresna multzo bat da. Bere osagaien artean, i) sintaxi-erroreen detekziorako, analisi morfologikorako eta morfologia-paradigmen sorkuntzarako LNP tresnak, ii) testu, bideo eta audio corpusa eta iii) irakasleei OLHIA sistemak konfiguratzeko lagunduko dien tresna multzo bat ditu. Bi prototipo sortu dira ALICEekin: espainierarako bata, eta japonierarako, bestea. Erroreen detekzioari dagokionez, analisi sintaktikoa Tomita analizatzailearekin egiten da eta erroreen detekziorako bi ezagutza-base erabiltzen dira: espainierarako *errore-gramatika* bat esaldi erroredunak lantzeko gai dena, eta berrelikadurarako txantiloia multzo bat. Erregelak baterakuntzan oinarritzen dira eta erregela bakoitza testuingururik gabeko egitura batek eta ekuazio multzo batek osatzen dute. Ekuazioek, ezaugarri-egituraren erroreari buruzko informazioa gehitzen dute eta errorea detektatuta dagoenean berrelikadura sortzeko txantiloia bat lortzen da (adib. "*x1*" *aditzak ez du* "*x2*" *subjektuarekin* "*x3*" *denboran komunztatzen*). Txantiloia esaldiko hitzekin betetzen da.

ALICE-chan (Levin eta Evans, 1995), ALICEren japonierarako bertsioa da eta hizkuntza ikasten hasten ari diren ikasleentzako bereziki prestatuta dago. Sarreraren analisia egiten denean, gramatika-erregela bereziak aplikatzen dira testu erroreduna landu ahal izateko.

B. ERANSKINA

Saroi. Klase-diagramak

Eranskin honetan, *Saroi* tresna garatu ahal izateko diseinu-fasean definitu ditugun klase-diagramak erakutsiko ditugu.

Saroi objektuei orientatutako ikuspegi bati jarraituz garatu da eta bere inplementaziorako erabili dugun programazio-lengoaia C++ izan da. Aukeraketa neurri batean *LibiXaML*-k baldintzatu du lengoaia horretan inplementatuta baitago. Honetaz gain, C++ erabiltzeak programazio-lengoaia honetarako garatutako *The Standard Template Library (STL)*-ren erabilera ahalbidetu digu. STL txantiloien estandarren liburutegia, klase-txantiloien eta algoritmoen bilduma generikoa da, eta programatzaileari baliabide oso erabilgarriak eskaintzen dizkio. Besteak beste, zenbait datu-egitura programatzailearen eskura jartzen ditu dagozkien metodo eta algoritmoekin.

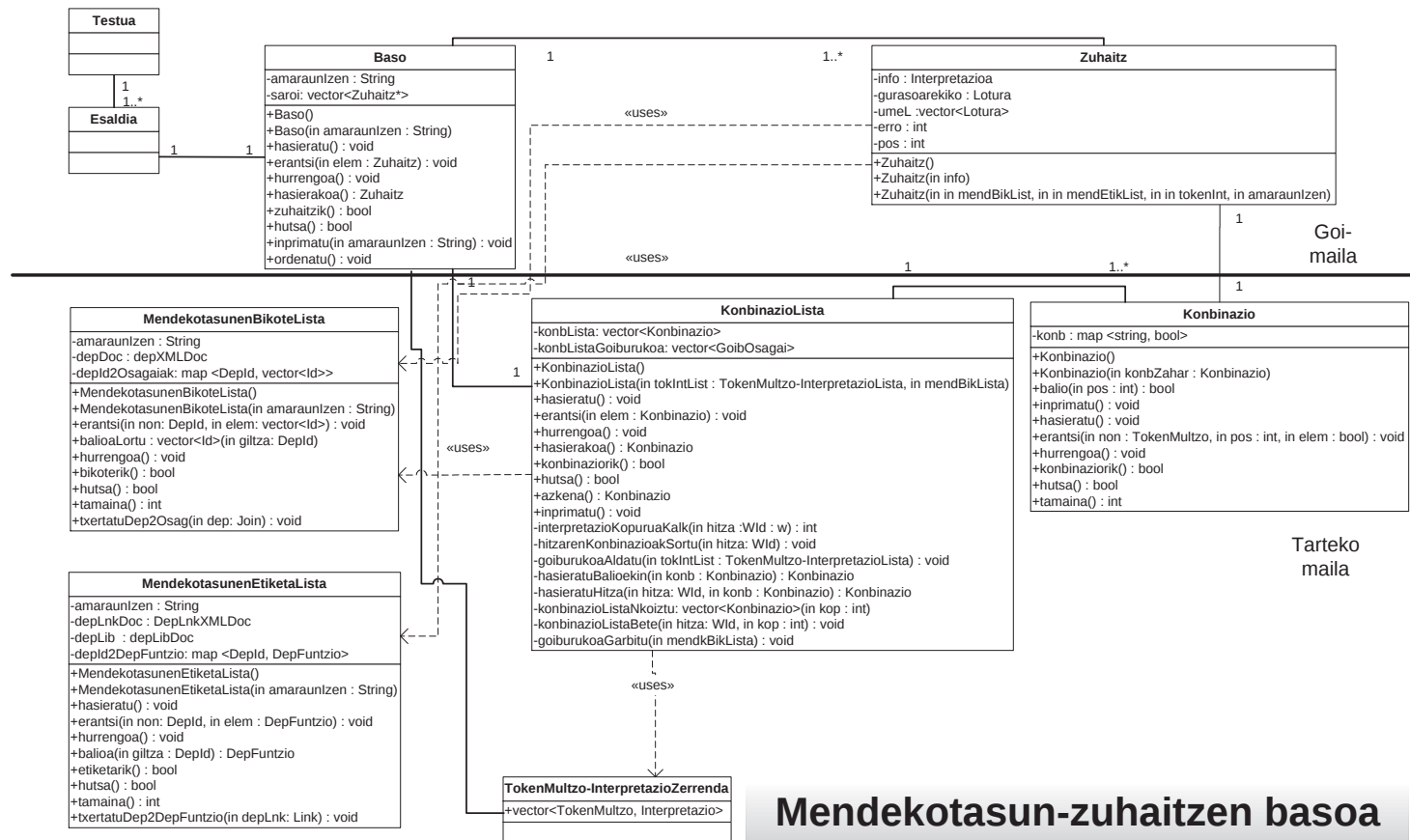
Ereduaren diseinu kontzeptuala egiteko UML modelatze-lengoaia orokorra¹ erabili dugu. UML software-sistemak bistaratzeko, zehazteko, erakitzeko eta dokumentatzeko lengoaia da (Larman, 2002). Lengoaia honek hainbat anotazio-eredu eskaintzen du objektuei orientatutako paradigma-ri jarraituz sistema baten analisisa eta diseinua egiteko. Horien artean guk *domeinu-ereduak* eta *klase-diagramak* erabiliko ditugu. *Domeinu-ereduan*, *Saroi* tresnaren klase kontzeptualak eta objektuak erakutsiko ditugu, baita beraien arteko erlazioak ere. Esku artean ibiliko ditugun kontzeptu anitz orokorrak dira eta mendekotasun-zuhaitzen alorrean ohikoak. *Klase-diagrametan* kontzeptu-klase horiek metodoekin osatuko ditugu, software-elementuak lortuaz.

Klase-diagramei eskaini diogun eranskin honetan, honako klase-diagra-

¹Unified Modeling Language

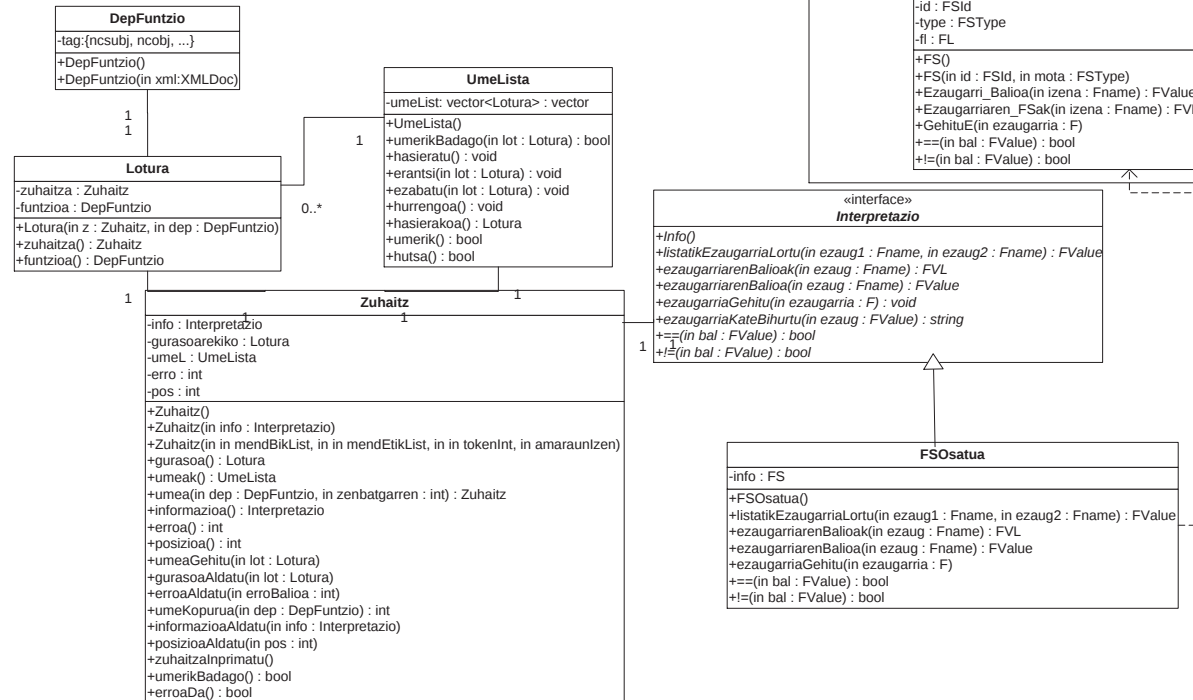
mak erakutsiko ditugu:

1. Analisi sintaktikorako modulua.
 1. 1. *Basoa* edo *Saroi*a (B.1 irudia).
 1. 2. *Zuhaitza* (B.2 irudia).
 1. 3. *Aberasketa* (B.3 eta B.4 irudiak).
2. Erregelen itzulpenerako modulua (B.5).

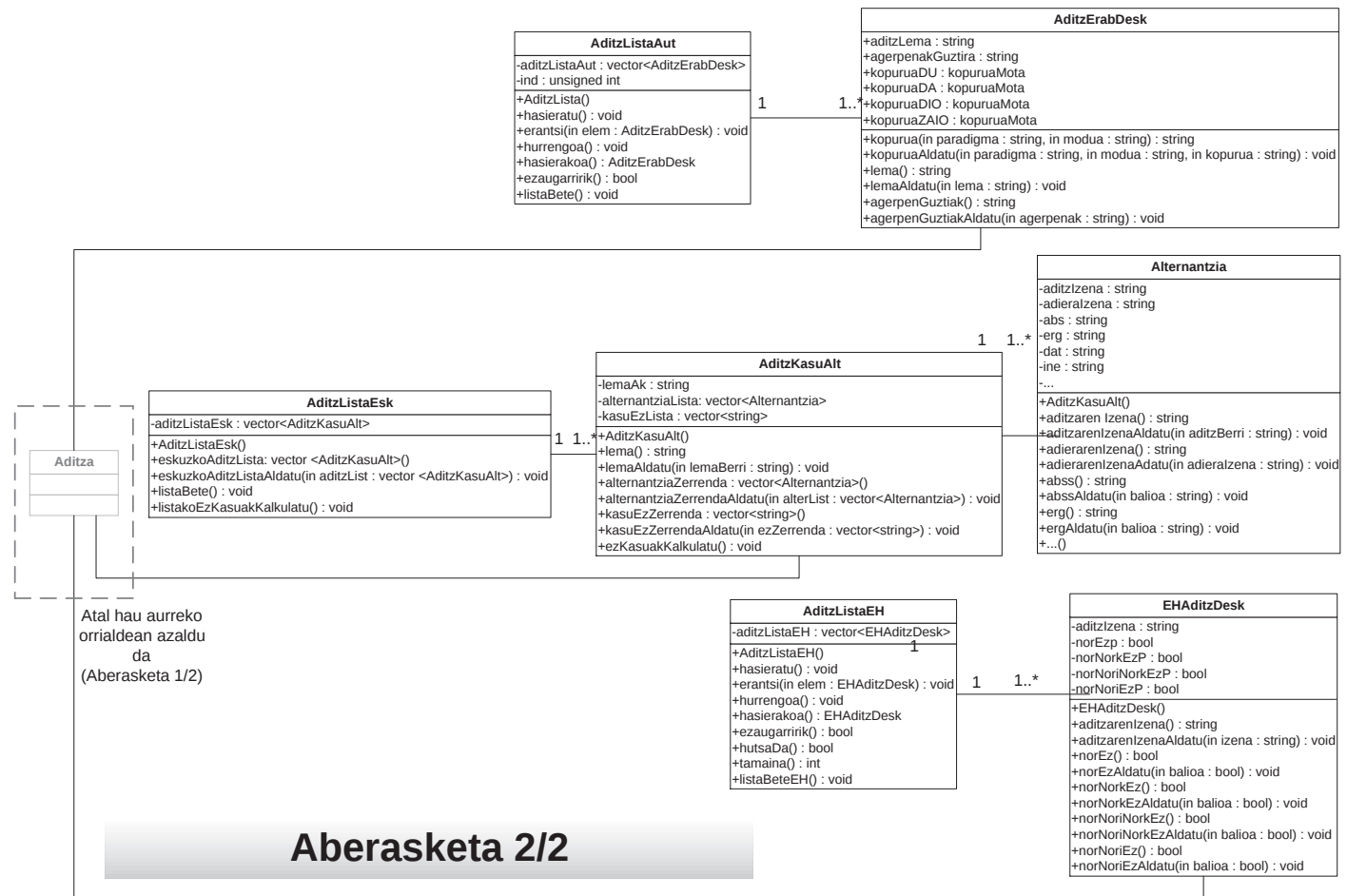


B.1 Irudia: Mendekotasun-zuhaitzen *basoa* (sarioa) inplementatzeko erabilitako klase-diagrama.

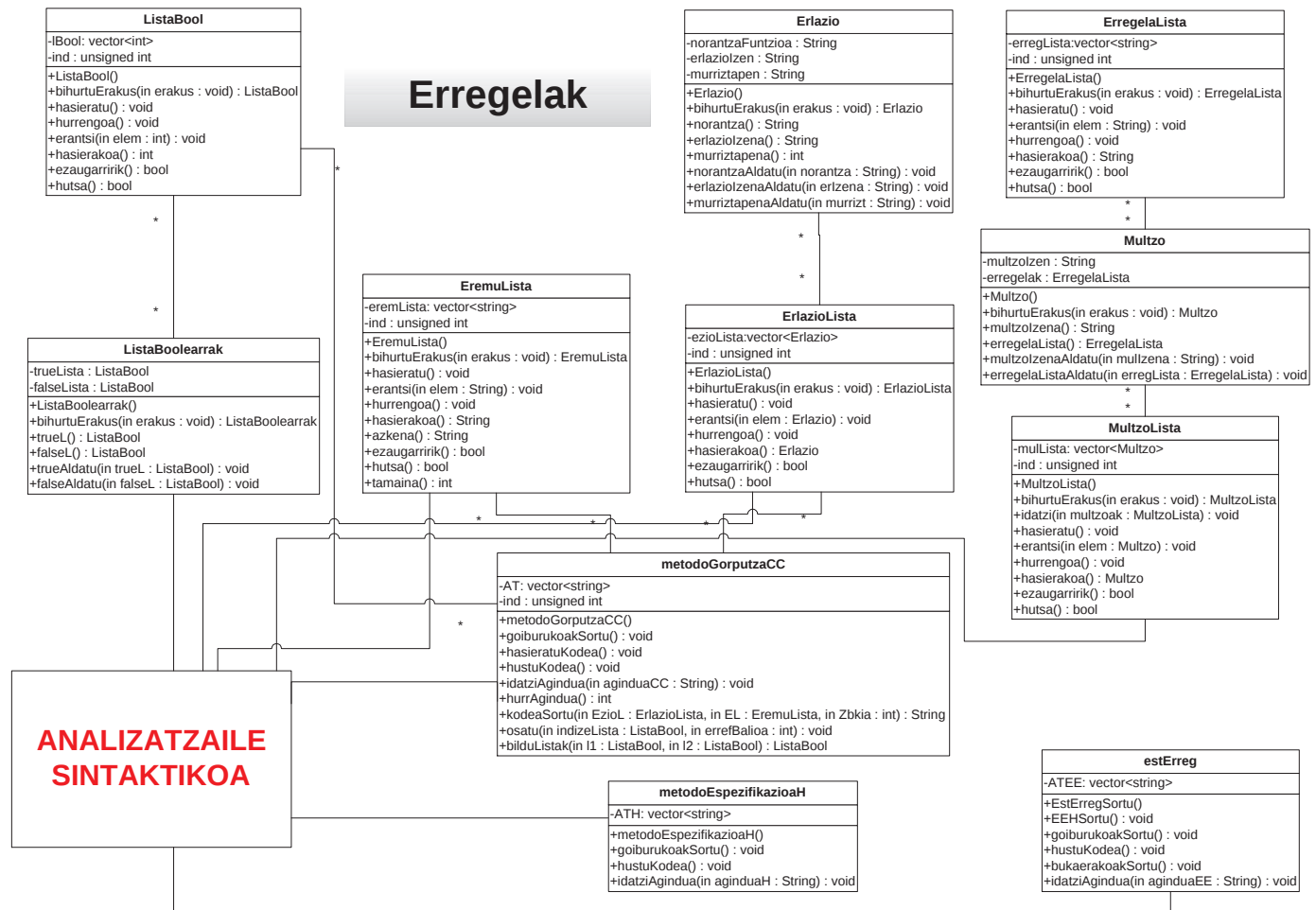
Mendekotasun-zuhaitza



B.2 Irudia: Mendekotasun-zuhaitza implementatzeko erabilitako klase-diagrama.



B.4 Irudia: Aditz nagusien aberasketarako eta aditz laguntzaileen ezaugarriak esplizitu egiteko implementazioan erabilitako klase-diagrama (2/2).



B.5 Irudia: Erregelaren itzulpeneko moduluaren implementaziorako erabilitako klase-diagrama.

C. ERANSKINA

Saroi. Erregelen itzulpena

Eranskin honetan, *Saroi*ren erregelak C++ programazio-lengoaian idatzita-ko koderaz itzultzeko sortu dugun itzultzailearen berri emango dugu. Lehenengo, sortu dugun *analizatzaile lexikoa* deskribatuko dugu C.1 atalean. Gero, *analizatzaile sintaktikoa* sortzeko definitu ditugun *testuingururik gabeko gramatika (TGG)* eta honetan oinarritutako *sintaxiak zuzendutako itzulpen-eskema (SZIE)* (Aho *et al.*, 1985) azalduko ditugu C.2 atalean.

C.1 Analisi lexikoa

Analisi lexikoan sarrerako karaktere-kateak irakurtzen dira eta emaitza moduan osagai lexikoen sekuentzia bat sortzen da. Analizatzaile sintaktikoak osagai hauek erabiliko ditu egitura-analisia egiteko.

Horretaz gain, analizatzaile lexikoak batzuetan bigarren mailako funtzioak betetzen ditu, hala nola, sarreratik oharrak, zuriuneak, tabuladoreak eta antzekoak ezabatzea.

Ondorengo taulan (C.1 taula), sintaxi-analizatzaileak jasoko dituen osagai lexikoak, patroiekin bat egiten duten sarrerako karaktere-kateen adibide batzuk eta osagaia deskribatzen duen patroia erakusten dira.

Analizatzaile lexikoen zehaztapenerako eta garapenerako *flex* (*Fast lexical analyzer generator*)¹ tresna erabili dugu. *flex* tresna *lex* tresna ezagunaren oso antzekoa da eta eskanerrak, testuetan osagai lexikoak topatzen dituzten programak, sortzen ditu. C programazio-lengoaian idatzita dago eta Vern Paxson-ek sortu zuen 1987an.

¹<http://flex.sourceforge.net>

Osagai lexikoa	Lexemen adibidea	Patroiaren ² deskribapena
multzo_izena	Multzo_KOM1_ERREGIMENA	Multzo_KOM[a-zñA-ZÑ—+*_0-9]*
erreg_izena	Erreg_KOMNAH_SUB	“Erreg_KOM”[a-zñA-ZÑ—+*_0-9]*
Detekt	Detekt	Detekt
erlazioa	ncsubj, ncobj2	ncsubj[0-5]* ncobj[0-5]* nczobj[0-5]* detmod ncmmod ... DETMOD ...
eremua	kas, lema, Paradigma	[a-zñA-ZÑ0-9]*
kasu_konstantea	ABS, ERG, DAT	ABS ERG DAT ELA_KONP ...
zbkia	0, 27	[0-9]+
oharra	\\	\\ [\\t]*[^\\n]*
katea	“ADI”, “NOR-NORK”	“[a-zñA-ZÑ]*[a-zñA-ZÑ—0-9+*_]”
zuria		“ “ \\ t \\ n
par_ireki	(	(
par_itxi	)	)
unekoa	@	@
azkena	\$	\$
kopurua	#	#
eta	&	&
edo		
behera	!	!
gora	i	i
puntu	.	.
existitzen_da	~	~
desberdin	!=	!=
berdin	==	==
handiago	>	>
handberdin	>=	>=
txikiago	<	<
txikberdin	<=	<=
esleipena	:=	:=
gehi	+	+

C.1 Taula: *Saroiko* erregelen analisi lexikoan erabiltzen diren osagai lexikoak.

C.2 Erregelen analisi sintaktikoa. TGG eta SZIE

Aurreko atalean *Saroi*ren erregelak helburu-kodera (C++ programazio-lengoaian idatzitako kodera) aldatzen dituen itzultzailearen zati bat ikusi dugu: analizatzaile lexikoa. Oraingoan, analizatzaile lexikoaren ondoren datorren atalaren inguruan arituko gara, analizatzaile sintaktikoaren inguruan. Analizatzaile sintaktikoak analizatzaile lexikotik jasotzen ditu osagai lexikoak eta erregelei egitura hierarkikoa ematen die. Egitura hori deskribatzeko *testuingururik gabeko gramatika* (TGG) bat erabili dugu.

TGG honetako ez-bukaerako ikurrei semantikarekin zerikusia duten atributuak eta ekintza semantikoak gehituz gero, *sintaxiak zuzendutako itzulpen-eskema* (SZIE) (Aho *et al.*, 1985) sortzen da. Honen bidez zuzen-zuzenean definituko da erregelen itzulpena helburu-kodera. Noski, SZIEa ez da nahikoa helburuko kodea sortzeko, hau aplikatuko duen programaren bat behar dugu eta hori, gure kasuan YACC³ (*Yet Another Compiler-Compiler*) analizatzaile sintaktikoen sortzailea da.

Ondorengo lerroetan labur-labur emango dugu aipatutakoen berri.

C.2.1 Testuingururik gabeko gramatika (TGG)

Nahiz eta gramatikak *teorian* eredu sortzaileak izan, batzuetan beste zenbait eredu (adibidez, automatak) baino argiagoak izan daitezke, besteak beste, programazio-lengoaiei edo erregelen sintaxi-egitura modu txukunean adierazteko.

C.2.1 taulan *Saroi*ko erregelen sintaxi-egitura definitzen duen TGGaren definizio formala eman dugu. Definizio formal honen parte den produkzio-erregelen multzoa C.3 taulan agertzen da. Gure ustez, definizio hau auto-deskribatzailea da eta osagaiak neurri handi batean *Saroi*ri eskaini diogun kapituluko VI.2.2.4 atalean azaldu dira; beraz, ez ditugu sakonago aztertuko.

C.2.2 Sintaxiak zuzendutako itzulpen-eskema (SZIE)

Sintaxiak zuzendutako itzulpen-eskema TGGaren gainean definitzen da. TGGa hezurduratzat hartuta, gramatikako erregela bakoitzari ekintza-sekuentzia bat gehitzen zaio, eta analisi sintaktikoa egiten den bakoitzean, ekintza hauek exekutatu egiten dira. Ekintza horiek dira, ekintza semantikoak, itzulpen-prozesua zehazten dutenak. Ekintzei dagokien atalean, ez-bukaerako bakoitzarekin lotuta dagoen klase bateko (ERREGELAK, ERLAZIOAK,

³<http://dinosaur.compilertools.net>

Saroiren testuingururik gabeko gramatika $G=(\Sigma, N, P, S)$ laukotearen bidez definitzen da, non osagaietako bakoitza honakoa den:

- Σ bukaerako ikurren multzoa. Kasu honetan, analisi lexikotik jasotako osagai lexikoek osatzen dute Σ .
 $\Sigma = (\text{Multzo_izena}, \text{Erreg_izena}, (,), \text{Detekt}, \&, |, @, \text{Katea}, \text{Zbkia}, \$, \#, +, -, !, i, !=, >, >=, <, <=, \text{ncsubj}, \text{ncobj}, \text{nczobj} \dots \text{mota}, \text{kas}, \text{kat}, \text{nor}, \text{nork} \dots)$
- N , ez-bukaerako ikurren multzoa
 $N = (\text{NAGUSIA}, \text{MULTZOAK}, \text{ERREGELAK}, \text{ERREGELA}, \text{DETEKT}, \text{DETEKT_OSAGK}, \text{D_OSAG}, \text{POS_ERAG}, \text{PLUS_MINUS}, \text{MURRIZTAPENA}, \text{ERLAZIOAK}, \text{NORANTZA}, \text{EREMUAK}, \text{KONP}, \text{ERLAZIOA}, \text{EREMUA})$
- P : produkzio-erregelen multzoa, non $u \longrightarrow v$ eta $u \in N, v \in (N \cup \Sigma)^*$
 Produkzio-erregelak C.3 taulan agertzen dira.
- S hasierako ikurra. Kasu honetan hasierako ez-bukaerakoa NAGUSIA da.

C.2 Taula: Analisi sintaktikoaren oinarri den TGGaren definizio formala.

$\text{EREMUAK} \dots$) metodoak exekutatzen dira gure kasuan, baita zenbait atributuren kalkuluarekin zerikusia duten ekintzak ere.

Ezinbestekoa izaten da SZIEtan gramatikako osagai sintaktikoekin lotutako atributuen erabilera. Bukaerako ikurretan (Σ -ko elementuetan) ez da atributu-kalkulurik egiten, analisi lexikoaren fasean jasotzen baitute atributu-balioa. Ez-bukaerakoetan, berriz, erregela semantikoen bidez kalkulatzeko dira balioak.

Gure kasuan atributu sintetizatuak erabili ditugu, hau da, ezkerreko aldeko ez-bukaerakoetako atributuek eskuinaldeko elementuetatik jasotzen dituzte balioak. Analisi-zuhaitza eginez gero, umeetako atributuetako balioak, dauden moduan edo aldatuta, gurasoen atributuetan jasotzen dira.

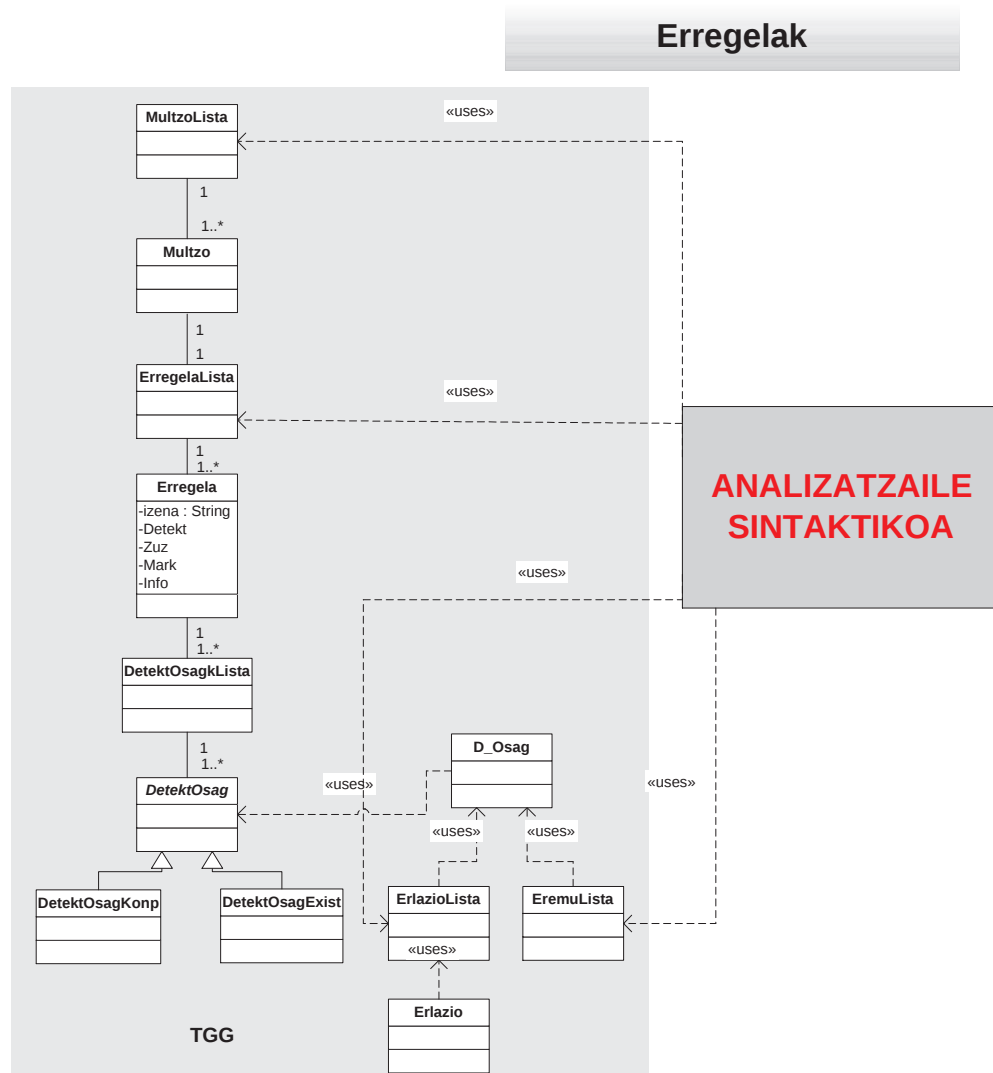
Ondorengo lerroetan SZIEa erakusten da. Bertako zenbait ez-bukaerakotan oharrak idatzi dira hobeto uler daitezen. Batzuetan ohar hauek zuzeneko erlazioa dute inplementaziorako erabilitako datu-egiturarekin. Kasu hauetan, C.1 irudiko klaseren bati egiten zaio erreferentzia.

```

P: NAGUSIA → MULTZOAK
MULTZOAK → Multzo_izena ERREGELAK MULTZOAK
           | Multzo_izena ERREGELAK
ERREGELAK → ERREGELAK ERREGELA
           | ERREGELA
ERREGELA → Erreg_izena ( DETEKT ZUZ MARK INFO)
DETEKT → Detekt ( DETEKT_OSAGK)
DETEKT_OSAGK → DETEKT_OSAGK & DETEKT_OSAG
              | DETEKT_OSAGK “|” DETEKT_OSAG
              | DETEKT_OSAG
DETEKT_OSAG → D_OSAG KONP D_OSAG
              | D_OSAG ~
              | ( DETEKT_OSAGK )
D_OSAG → @ ERLAZIOAK EREMUAK POS_ERAG
        | Katea
        | KasuKonstantea
        | Zbkia
POS_ERAG → PLUS_MINUS Zbkia
          | €
PLUS_MINUS → +
          | -
MURRIZTAPENA → #
              | $
              | ( Zbkia )
              | €
ERLAZIOAK → ERLAZIOAK NORANZKOA ERLAZIOA MURRIZTAPENA
          | €
NORANZKOA → !
          | i
EREMUAK → EREMUAK . EREMUA
        | €
KONP → == |!= |> |>= |< |<=
ERLAZIOA → ncsubj | ncobj | nczobj | ccomp_subj | ccomp_obj ...
EREMUA → mota | kas | kat | nor | nork | ...

```

C.3 Taula: Analisi sintaktikoaren oinarri den TGGaren produkzio-erregelak.



C.1 Irudia: *Saroiren erregelak* lantzeko domeinu-eredua.

```

# Lotutako klasea → MultzoLista
# Erregela-fitxategia hainbat erregela multzorekin osatuta dago.
# Multzoetako erregelak analizatu direnean, sortutako kodea datu-egituretatik
dagozkien fitxategietara hustuko da.
# Multzoko erregela bakoitzeko funtzio bat sortuko da:
- komErregelak.h fitxategian → Funtzioen espezifikazioak.
- komErregelak.cc fitxategian → Funtzioen gorputza edo inplementazioa.

# Multzoko erregelak estrategia zehazten duen fitxategian agertuko dira:
- estrategiaErregelak.h → Estrategien espezifikazioa.
- estrategiaErregelak.cc → Estrategien gorputza.

```

NAGUSIA → MULTZOAK

```

{
# Metodoen espezifikazioak dituen fitxategia sortu (KomErregelak.h)
Goiburukoak_sortu_buruko_H_fitxategian();
Hustu_kodea_buruko_H_fitxategian();

# Metodoen kodea/gorputza duen fitxategia sortu (KomErregelak.cc)
Goiburukoak_sortu_gorputzeko_CC_fitxategian();
Hustu_kodea_gorputzeko_CC_fitxategian();

# Multzo bakoitza zein erregelarekin osatuta dagoen idatzi
Erregela_multzoak_osatu(MULTZOAK.lista);

# Erregelak aplikatzeko, aukeratutako estrategia aplikatzeko,
# fitxategia idatzi (estrategiaErregelak.h)
Estrategia_erregelak_sortu_EE();
Goiburukoak_sortu_estrategia_EE_fitxategian();
Hustu_kodea_estrategia_EE_fitxategian();
}

```

```

# Lotutako klasea → Multzo
# Multzoak izenaren bidez identifikatzen dira (Multzo_izena) eta erregelekin osatuta
daude.

```

MULTZOAK → **Multzo_izena** ERREGELAK MULTZOAK

```

{
MULTZOAK0.lista = Bildu_multzoak( bikotea(Multzo_izena.katea,
ERREGELAK.lista),
MULTZOAK1.lista);
}
|Multzo_izena ERREGELAK
{

```

```

MULTZOAK0.lista = Bildu_multzoak( bikotea(Multzo_izena.katea,
                                             ERREGELAK.lista),
                                   Lista_hutsa());
}

```

```
# Lotutako klasea → ErregelaLista
```

```

ERREGELAK → ERREGELAK ERREGELA
{ERREGELAK0.lista = Bildu_erregelak( ERREGELA.izena,
                                       ERREGELAK1.lista);}

|ERREGELA
{ERREGELAK0.lista = Bildu_erregelak(ERREGELA.izena,
                                       Lista_hutsa());}

```

```

# Lotutako klasea → Erregela
# Erregelak izenaren bidez identifikatzen dira (Erreg_izena) eta lau atal dituzte
(DETEKT, ZUZ, MARK, INFO).
# Guk DETEKT soilik garatu dugu.
# Erregelen izenak erabilita, zera idatziko da:
- Espezifikazio-fitxategiari dagokion egiturari, funtzioaren espezifikazioa
(gero "komErregelak.h").
- Gorputz-fitxategiari dagokion egiturari, funtzioaren izena eta hasierako agindua
(gero "komErregelak.cc").
- Estrategia-fitxategiei dagozkien egituretan, beharrezko aginduak
(gero estrategiaErregelak.h eta estrategiaErregelak.cc).

```

```
ERREGELA → Erreg_izena
```

```

{
  ERREGELA.izena=Erreg_izena.balioa;
  Idatzi_agindua_H_buruan("int " + Erreg_izena.balioa + "());

  Idatzi_agindua_CC_gorputzean("int KomErregelak::" + Erreg_izena.balioa + "());
  Idatzi_agindua_CC_gorputzean("{}");
  Idatzi_agindua_CC_gorputzean("Zuhaitz* unekoAdabegia = this→zuhaitza;");
  Idatzi_agindua_CC_gorputzean("{}");

  Idatzi_agindua_EEn("if ( erregelak." + Erreg_izena.balioa + "() == 1)");
  Idatzi_agindua_EEn("{}");
  Idatzi_agindua_EEn("erroreaDU = erroreaDU + 1");
  Idatzi_agindua_EEn("erregela=" + Erreg_izena.balioa + ";");
  Idatzi_agindua_EEn("erregelaZerrenda = erregelaZerrenda + erregela;");
  Idatzi_agindua_EEn("{}");
}

```

```
( DETEKT ZUZ MARK INFO)
{
  Idatzi_agindua_CC_gorputzean("}");
}
```

```
# Erregelen detekziorako atala "Detekt"hitzarekin hasiko da eta hainbat osagai
izango ditu barruan.
```

```
DETEKT  $\longrightarrow$  Detekt ( DETEKT_OSAGK)
{
  Osatu(DETEKT_OSAGK.true_lista, Hurr_Agindua());
  Osatu(DETEKT_OSAGK.false_lista, Hurr_Agindua()+1);
  Idatzi_agindua_CC_gorputzean("}");
  Idatzi_agindua_CC_gorputzean("return 1;");
  Idatzi_agindua_CC_gorputzean("return 0;");
}
M  $\longrightarrow$   $\epsilon$ 
{ M.erref := Hurr_Agindua();}
```

```
# Lotutako klasea  $\longrightarrow$  DetektOsagkLista
# Detekziorako erregelak "&"(eta) eta "|"(Edo) eragilekin konposatu daitezke.
# Adibidea:
@!ncsubj!ncmod~ &
(@!auxmod.paradigma == ‘‘NOR-NORK’’ |
(@!auxmod.paradigma == ‘‘NOR-NORI-NORK’’)) &
@!ncsubj!ncmod.kas != @!auxmod.nork.kas
```

```
DETEKT_OSAGK0  $\longrightarrow$  DETEKT_OSAGK1 & M DETEKT_OSAG
{
  Osatu (DETEKT_OSAGK1.true_lista, M.erref);
  DETEKT_OSAGK0.false_lista:= Bildu (DETEKT_OSAGK1.false_lista,
  DETEKT_OSAGK1.true_lista);
  DETEKT_OSAGK0.true_lista:= DETEKT_OSAGK1.true_lista;
}
|DETEKT_OSAGK1 “|” M DETEKT_OSAG
{
  Osatu (DETEKT_OSAGK1.false_lista, M.erref);
  DETEKT_OSAGK0.true_lista := Bildu (DETEKT_OSAGK1.true_lista,
  DETEKT_OSAGK1.true_lista);
  DETEKT_OSAGK0.false_lista:= DETEKT_OSAGK1.false_lista;
}
|DETEKT_OSAG
{
  DETEKT_OSAGK0.true_lista:= DETEKT_OSAGK1.true_lista;
```

```

DETEKT_OSAGK0.false_lista:= DETEKT_OSAG.false_lista;
}

```

```

# Lotutako klasea → DetektOsagKonp
# Erregelak konparaziozko ikurren batekin bereizitako bi alde edo osagai eduki
ditzake.
# Adibidea: @!ncsubj!ncmod.kas != @!auxmod.nork.kas
# Dagokion kode-zatia idatziko da.

```

```

DETEKT_OSAG → D_OSAG1 KONP D_OSAG2
{
  DETEKT_OSAG.true_lista := Hasi_Lista(Hurr_Agindua());
  DETEKT_OSAG.false_lista := Hasi_Lista(Hurr_Agindua() + 1);
  Idatzi_agindua_CC_gorputzean( "if" + D_OSAG1.ald + KONP.katea + D_OSAG2.ald );
  Idatzi_agindua_CC_gorputzean( "else goto " );
  Idatzi_agindua_CC_gorputzean( "}" );
  Idatzi_agindua_CC_gorputzean( "{" );
}

```

```

# Lotutako klasea → DetektOsagExist
# Edo osagai baten existentziari buruz galde dezakegu.
# Adibidea:
@!ncsubj!ncmod~ &

```

```

|D_OSAG ~
{
  DETEKT_OSAG.true_lista := Hasi_Lista(Hurr_Agindua());
  DETEKT_OSAG.false_lista := Hasi_Lista(Hurr_Agindua() + 1);
  If (eremurikEz){ # @!detmod
    Idatzi_agindua_CC_gorputzean( "if" + D_OSAG.ald + "!= NULL goto " );
    Idatzi_agindua_CC_gorputzean( "else goto " );
  }
  else { # @!detmod.kas
    Idatzi_agindua_CC_gorputzean( "if" + D_OSAG.ald + "!= "" goto " );
    Idatzi_agindua_CC_gorputzean( "else goto " );
  }
  Idatzi_agindua_CC_gorputzean( "}" );
  Idatzi_agindua_CC_gorputzean( "{" );
}

```

```
# Edo osagaietan ordena ezar daiteke parentesiak erabilita.
# Adibidea:
(@!auxmod.paradigma == 'NOR-NORK' |
 @!auxmod.paradigma == 'NOR-NORI-NORK') &
```

```
| ( DETEKT_OSAGK )
{
  DETEKT_OSAG.true_lista := DETEKT_OSAGK.true_lista;
  DETEKT_OSAG.false_lista := DETEKT_OSAGK.false_lista;
}
```

```
# Lotutako klasea  $\rightarrow D\_Osag$ 
# Detekzio-osagaiek, '@' ikurra edukiko dute, ondoren nahi adina mendekotasun-
-erlazio, gero nahi adina eremu (baleude) eta gero eragileren bat (balego).
# Adibidea: @!ncsubj!ncmod.kas
```

```
D_OSAG  $\rightarrow$  @ ERLAZIOAK EREMUAK POS_ERAG
{
  azkenEtik=Kodea_sortu(ERLAZIOAK.lista, EREMUAK.lista, zbkia);
  if (POS_ERAG.balioa != "")
  {
    D_OSAG.ald := azkenEtik + POS_ERAG.balioa;
  }
  else
  {
    D_OSAG.ald := azkenEtik;
  }
}
```

```
# Edo, kate batekin osatuta egon daiteke. Adibidea: 'NOR-NORK'
```

```
| Katea
{
  string ald=Lortu_balio_aldagaia();# B8
  Idatzi_agindua_CC_gorputzean("string" + "ald" + "=" + Katea.balioa + ";");
  # FValue B8="nor-nork";
}
| KasuKonstantea
{
  string aldKonst=Lortu_balio_aldagaia();# B8
  Idatzi_agindua_CC_gorputzean("string" + "aldKonst" + "=" + KasuKonstantea.balioa + ";");
  # FValue B8="abs";
}
```

```

|Zbkia
{
string aldZbkia=Lortu_balio_aldagaia();# B8
Idatzi_agindua_CC_gorputzean("int"+ aldZbkia + "=" + Zbkia.balioa +";");
# FValue B8="3";
}

```

```

# POS_ERAG esaldiko hitzen arteko distantzia neurtzeko erabiltzen da.
# Adibidea: @!ncsubj.pos == @.pos + 1

```

```

POS_ERAG → PLUS_MINUS Zbkia
{
POS_ERAG := PLUS_MINUS. balioa Zbkia.balioa;
# "+ 1"edo "- 2";
}
|∈
{
POS_ERAG := "";
}

```

```

PLUS_MINUS → +
{
PLUS_MINUS.balioa ="+";
}
|-
{
PLUS_MINUS.balioa ="-";
}

```

```

# Erlazioei ipin dakizkieken murriztapenak honakoak dira:
- Kopurua. Adibidea: @!ncmod#
- Azkena. Adibidea: @!ncmod$
- n-garrena, mendekotasun-erlazio berekoen artean. Adibidea: @!ncmod(2)

```

```

MURRIZTAPENA → #
{MURRIZTAPENA.mota = 100;}
|$
{ MURRIZTAPENA.mota = 0;}
|( Zbkia )
{ MURRIZTAPENA.mota = Zbkia.balioa;}
|∈
{ MURRIZTAPENA.mota = "";}

```

```
# Lotutako klasea → ErlazioLista # Nahi adina mendekotasun-erlazio defini
daitezke.
# Erlazio bakoitza noranzkoarekin (! edo <), erlazio-izenarekin (ncsubj, ncmmod ...)
eta murriztapenarekin osatuta egongo da (balego).
# Adibidea: @!ncmod#
```

```
ERLAZIOAK0 → ERLAZIOAK1 NORANZKOA ERLAZIOA MURRIZTAPENA
{
  ERLAZIOAK0.lista = Bildu_erlazioak( hirukotea(NORANZKOA.funtzioa,
                                                    ERLAZIOA.katea,
                                                    MURRIZTAPENA.mota),
                                     ERLAZIOAK1.lista);
}
| ∈
{
  ERLAZIOAK0.lista = Lista_hutsa();
}
```

```
NORANZKOA → !
{NORANZKOA.funtzioa="umea";}
| i
{NORANZKOA.funtzioa="gurasoa";}

```

```
# Lotutako klasea → EremuLista
# Nahi adina eremu (kas, mota, paradigma, ...) defini daitezke erlazio batean.
# Eremua zehazteko ‘.’ ikurra erabiliko da. # Erlazio bakoitza noranzkoarekin (!
edo <), erlazio-izenarekin (ncsubj, ncmmod ...) eta murriztapenarekin osatuta egongo
da (balego).
# Adibidea: @!ncmod.kas
```

```
EREMUAK0 → EREMUAK1 . EREMUA
{
  EREMUAK0.lista = Bildu_eremuak(EREMUA.katea, EREMUAK1.lista);
}
| ∈
{
  EREMUAK0.lista = Lista_hutsa();
}
```

```
KONP → ==
{KONP.katea="==";}
| !=
```

```

{KONP.katea= "!=";}
|>
{KONP.katea= ">";}
|>=
{KONP.katea= ">=";}
|<
{KONP.katea= "<";}
|<=
{KONP.katea= "<=";}

```

Lotutako klasea $\longrightarrow$ <i>Erlazio</i>
--

```

ERLAZIOA  $\longrightarrow$  ncsubj
    {ERLAZIOA.katea= "ncsubj";}
|ncobj
    {ERLAZIOA.katea= "ncobj";}
|nczobj
    {ERLAZIOA.katea= "nczobj";}
|ccomp_subj
    {ERLAZIOA.katea= "ccomp_subj";}
|ccomp_obj
    {ERLAZIOA.katea= "ccomp_obj";}
|ccomp_zobj
    {ERLAZIOA.katea= "ccomp_zobj";}
|xcomp_subj ...

EREMUA  $\longrightarrow$  mota
    {EREMUA.katea= "mota";}
|kas
    {EREMUA.katea= "kas";}
|kat
    {EREMUA.katea= "kat";}
|nor
    {EREMUA.katea= "nor";} ...

```

SZIEaren eraginez, sarrerako erregela bakoitzeko C++ programazio-lengoan idatzitako kodea lortuko dugula esan dugu. C.2 irudian adibide bat ikus dezakegu. Ezkerreko aldean komunztadura-errore bat detektatzeko idatziko genukeen erregela agertzen da. Eskuinekoan, berriz, interpretatzailearen emaitza erakusten da. Emaitzan erabiltzen diren metodoak *Zuhaitza* klasekoak dira. Jauziak erabiltzen dira sortutako funtzioak eraginkorragoak izan daitezen.

```

KOM_SUBJ_KAS_NOR_NORK
(
  Detekt (
    @!ncsubj!ncmod~ &
    @!auxmod.paradigma == "NOR-NORK" &
    @!ncsubj!ncmod.kas != @!auxmod.nork.kas
  )
  int KomErregelak::Erreg_Kom_SUBJ_KAS_NOR_NORK()
  {
    Zuhaitz* unekoAdabegia=this->zuhaitza;
    {
      DepFuntzio dep2("NCSUBJ");
      Zuhaitz* Z1 = unekoAdabegia->umea(dep2,1);
      if (Z1 == NULL) return 0;
      DepFuntzio dep4("NCMOD");
      Zuhaitz* Z3 = Z1->umea(dep4,1);
      if (Z3 == NULL) return 0;
      if (Z3 != NULL) goto x10;
      else goto x33;
    }
    {
      x10: DepFuntzio dep6("AUXMOD");
      Zuhaitz* Z5 = unekoAdabegia->umea(dep6,1);
      if (Z5 == NULL) return 0;
      Interpretazio* F7 = Z5->informazioa();
      FName FN10("Paradigma");
      FValue* B8 = F7->ezaugarriarenBalioa(FN10);
      string B9 = F7->ezaugarriaKateBihurtu(B8);
      string B11 = "NOR-NORK";
      if (B9 == B11) goto x18;
      else goto x33;
    }
    {
      x18: DepFuntzio dep13("NCSUBJ");
      Zuhaitz* Z12 = unekoAdabegia->umea(dep13,1);
      if (Z12 == NULL) return 0;
      DepFuntzio dep15("NCMOD");
      Zuhaitz* Z14 = Z12->umea(dep15,1);
      if (Z14 == NULL) return 0;
      Interpretazio* F16 = Z14->informazioa();
      FName FN19("KAS");
      FValue* B17 = F16->ezaugarriarenBalioa(FN19);
      string B18 = F16->ezaugarriaKateBihurtu(B17);
      DepFuntzio dep21("AUXMOD");
      Zuhaitz* Z20 = unekoAdabegia->umea(dep21,1);
      if (Z20 == NULL) return 0;
      Interpretazio* F22 = Z20->informazioa();
      FName FN25("NORK");
      FName FN26("KAS");
      FValue* B23 = F22->listatikEzaugarriaLortu(FN25,FN26);
      string B24 = F22->ezaugarriaKateBihurtu(B23);
      if (B18 != B24) goto x32;
      else goto x33;
    }
    {
      x32: return 1;
      x33: return 0;
    }
  };
)

```

C.2 Irudia: *Saroi*ren itzultzailearen sarrerako erregela eta irteeran lortutako kodea.

D. ERANSKINA

Saroi. Document Type Definition (DTDa)

Tesi-lan honetan ezagutza linguistikoa errepresentatzeko erabili dugun anotazio-amarauneko dokumentuen sintaxia edo egitura, ondorengo *Document Type Definition* edo DTD eskemak deskribatzen du.

```
<!ELEMENT TEI.2 (teiHeader, text)>
<!ELEMENT teiHeader (fileDesc, encodingDesc, profileDesc,
                    revisionDesc)>

<!ELEMENT fileDesc (titleStmt, publicationStmt, sourceDesc)>
<!ELEMENT titleStmt (title?, author?)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT publicationStmt (publisher, idno?)>
<!ELEMENT publisher (#PCDATA)>
<!ELEMENT idno (#PCDATA)>
<!ELEMENT sourceDesc (bibl?)>
<!ELEMENT bibl (#PCDATA)>

<!ELEMENT encodingDesc (fsdDecl*)>
<!ELEMENT fsdDecl EMPTY>

<!ELEMENT profileDesc (language)>
<!ELEMENT language (#PCDATA)>
```

```

<!ELEMENT revisionDesc (change*)>
<!ELEMENT change (date, respStmt, item)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT respStmt (name, resp)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT resp (#PCDATA)>
<!ELEMENT item (#PCDATA)>

<!ELEMENT text (body)>
<!ELEMENT body (p*)>
<!ELEMENT p (#PCDATA|xptr|w|linkGrp|joinGrp|fs)*>
<!ELEMENT xptr EMPTY>
<!ELEMENT w (#PCDATA|m)*>
<!ELEMENT m (#PCDATA)>
<!ELEMENT linkGrp (link*)>
<!ELEMENT link EMPTY>
<!ELEMENT joinGrp (join*)>
<!ELEMENT join (target*)>
<!ELEMENT fs (f*)>
<!ELEMENT f (EMPTY|(fs|str|nbr|sym|plus|minus)*)>
<!ELEMENT str (#PCDATA)>
<!ELEMENT nbr EMPTY>
<!ELEMENT sym EMPTY>
<!ELEMENT plus EMPTY>
<!ELEMENT minus EMPTY>
<!ELEMENT target EMPTY>

<!ATTLIST fsdDecl  fsd  CDATA #REQUIRED
                  type CDATA #REQUIRED>
<!ATTLIST language id ID #REQUIRED>
<!ATTLIST text id    ID #REQUIRED
              lang    CDATA #IMPLIED
>
<!ATTLIST p      id    ID #IMPLIED>
<!ATTLIST xptr id     CDATA #REQUIRED
              doc     CDATA #REQUIRED
              from     CDATA #IMPLIED
              to       CDATA #IMPLIED
>

```

```

<!--ATTLIST w      id      ID      #REQUIRED
                  sameAs CDATA #REQUIRED
                  type   CDATA #IMPLIED
                  etik   CDATA #IMPLIED
                  target CDATA #IMPLIED
                  rend   CDATA #IMPLIED
                  orig   CDATA #IMPLIED
                  status CDATA #IMPLIED
                  targetsynced CDATA #IMPLIED
                  edited CDATA #IMPLIED
>

<!--ATTLIST m      id      ID      #REQUIRED
                  target CDATA #IMPLIED
                  rend   CDATA #IMPLIED
>

<!--ATTLIST linkGrp type      CDATA #REQUIRED
                  targOrder CDATA #REQUIRED
                  targType  CDATA #IMPLIED
>

<!--ATTLIST link      id      ID      #IMPLIED
                  type   CDATA #IMPLIED
                  targets CDATA #REQUIRED
>

<!--ATTLIST joinGrp type      CDATA #REQUIRED
                  desc   CDATA #IMPLIED
                  targOrder CDATA #REQUIRED
                  targType  CDATA #IMPLIED
>

<!--ATTLIST join id      ID      #REQUIRED
                  type   CDATA #IMPLIED
>

<!--ATTLIST target      name CDATA      #REQUIRED
                  value  CDATA #REQUIRED

```

```

>

<!ATTLIST fs      id      CDATA    #IMPLIED
                  type     CDATA    #IMPLIED
>
<!ATTLIST f       id      ID       #IMPLIED
                  name     CDATA    #REQUIRED
                  type     CDATA    #IMPLIED
                  org       CDATA    #IMPLIED
>
<!ATTLIST nbr     value    CDATA    #REQUIRED>
<!ATTLIST sym     value    CDATA    #REQUIRED>

```

Dependentzia-zuhaitzak errepresentatzeko `joinGrp` eta `join` elementuak erabili ditugu batez ere (ikus adibideak VI kapituluan). `join` elementuak aurreko prozesuetan sortutako interpretazio-identifikadoreak biltzeko erabiltzen dira; beraz, egokiak dira adabegien arteko loturak errepresentatzeko. Gogora dezagun, mendekotasun-zuhaitz ez-anbiguoak bi modutan adierazten genituela, era *implizituan* eta *esplizituan*. Zuhaitz mota bakoitzak bere berezitasunak ditu. Bereztasun horiek zehazteko, `join` elementuekin lotura zuzena duten `target` elementuen atributuak erabili ditugu. Zuhaitza era implizituan gordetzen duen XML dokumenturako atributuen balioak D.1 taulan zehaztu ditugu. Era esplizituan errepresentatzeko, berriz, D.2 taulan definitutako enumerazioa erabili dugu.

```

<!ATTLIST target  name (Buru | Mende) #REQUIRED
                  value CDATA #REQUIRED
>

```

D.1 Taula: Zuhaitz *implizituaren* DTD-zehaztapena.

```

<!ATTLIST target name (Info|DepRel|VerbInfo?|Child*) #REQUIRED
                  value CDATA #REQUIRED
>

```

D.2 Taula: Zuhaitz *esplizituaren* DTD-zehaztapena.

E. ERANSKINA

***Saroi.* Komunztadura-erroreen detekzioa. Emaitzak.**

Komunztadura-erroreen detekzioa ebaluatzeko egin ditugun esperimentuan lortutako emaitzak tesi-txosten nagusian erakustea ezinezkoa denez (ikus VII. kapitulua), emaitza horiek eranskin honetara ekartzea erabaki dugu.

Emaiza eta analisi horiek ondorengo tauletan erakutsiko ditugu:

1. Desanbiguazio morfosintaktikoa 3.mailan mantenduaz, sintaxiaren anbigutasunarekin eta mendekotasun-erlazioen erlaxazioarekin probak egin ditugu. Erroredunak diren 75 esaldi jaso ditugu eta honako konbinazioekin egin ditugu probak.

- 3-2-Erlaxatua: sintagmekin eta aditzekin zerikusia duten funtzio sintaktikoak desanbiguatu dira. Mendekotasun-erlazioak ipintzen dituen EDGK gramatika erlaxatu da.
- 3-3-Erlaxatua: funtzio sintaktikoetan sintagmekin eta aditzekin eta zerikusia duten funtzio sintaktikoak eta funtzio sintaktiko nagusiak desanbiguatu dira. Mendekotasun-erlazioak ipintzen dituen EDGK gramatika erlaxatu da.
- 3-2-Erlaxatu gabe: sintagmekin eta aditzekin zerikusia duten funtzio sintaktikoak desanbiguatu dira. Mendekotasun-erlazioak ipintzen dituen EDGK gramatika ez da erlaxatu.
- 3-3-Erlaxatu gabe: funtzio sintaktikoetan sintagmekin eta aditzekin eta zerikusia duten funtzio sintaktikoak, eta funtzio sintaktiko nagusiak.

siak desanbiguatu dira. Mendekotasun-erlazioak ipintzen dituen EDGK gramatika ez da erlaxatu.

Esaldien analisisian honako datuak eman ditugu konbinazio bakoitzeko: zenbat analisi-zuhaitz sortu den (**Z**), horiek batez beste zenbat azpizuhaitz eduki dituzten (parentesi artean), eta hasierako zuhaitzetatik zenbatetan aplikatu den erroreen detekziorako erregelaren bat. Analisisi-zuhaitz *guztietan* komunztadura-erroreen detekziorako erregelaren bat detektatu bada, esaldia urdinez markatu dugu, bertako errorea detektatutzat jotzen baitugu. Datu hauen guztien berri E.1 taulan eman dugu.

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
1. Honek algoritmo baten bidez forma berriak asmatzeko gai da.	2Z (1) 0E	1Z (1) 0E	2Z 0E	1Z (1) 0E
2. Sarean dabilen datu mugimenduak entzuten dituzte.	12Z (1,33) 6E	6Z(2) 0E	12Z (1,33) 6E	6Z (2) 0E
3. Bestalde badaude beste hari mota bat RDSI izenekoak.	0Z	0Z	0Z	0Z
4. Ez dituela lehengairik behar energia elektrikoa ateratzeko.	8Z (2) 8E	4Z (2) 4E	8Z (2) 8E	4Z (2) 4E
5. Adimen Naturala eboluzioa du eta artifiziala berriz ez.	6Z (1) 6E	2Z (1) 1E	4Z (1) 4E	1Z (1) 1E
6. Abiaduraren norantzan egindako indarra abiadura haundituko du.	192Z (1) 160E	64Z (2) 32E	192Z (1) 160E	64Z (2) 32E
7. Azken biak zerikusi haundia daukate.	4Z (1) 4E	1Z (1) 1E	4Z (1) 4E	1Z (1) 1E
8. Hackerrak nolabaiteko fama bat dute baina ez dira gaiztoak beraien atakeetan.	2Z (3) 2E	1Z (3) 0E	2Z (3) 0E	1Z (3) 0E
9. Besteen ordenadoreetan sartzen dabilena edo birusak bidaltzen duena da.	12Z (2) 6E	6Z (2) 0E	12Z (2) 6E	6Z (2) 0E
10. Makila bolari eragin diolako indarra.	4Z (1) 4E	4Z (1) 4E	4Z (1) 4E	4Z (1) 4E
11. Nik ez nago konforme.	1Z (1) 1E	1Z (1) 1E	1Z (1) 0E	1Z (1) 0E
12. Nik ez noa.	1Z (1) 1E	1Z (1) 1E	0Z	0Z
13. Zer da nabigatzaileak?.	1Z (1) 1E	1Z (1) 1E	1Z (1) 1E	1Z (1) 1E
14. Ordenagailuetan dagoen informazio guztia orrialde horietan daude antolatuak.	12Z (1,66) 12E	6Z (1,66) 6E	12Z (1,66) 12E	6Z (1,66) 6E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
15. Erabiltzaileari sare batean egon daitekeen urruneko beste ordenagailuekin konektatzeko aukera ematen du.	4Z (2) 4E	2Z (2) 2E	4Z (2) 2E	2Z (2) 0E
16. Oso arriskutsua da nukleoa eskape bat baldin badauka.	24Z (2) 24E	6Z (2) 6E	24Z (2) 12E	6Z (2) 6E
17. Modema beti gure ordenagailura konektaturik egon behar du.	3Z (2) 0E	3Z (2) 0E	3Z (2) 0E	3Z (2) 0E
18. Hirugarrena orriaren lekua edo mota esan nahi du.	5Z (2) 0E	4Z (2) 0E	5Z (2) 0E	4Z (2) 0E
19. Beraien asmoa erabat suntsitzaileak dira.	3Z (1,33) 1E	1Z (1) 1E	3Z (1,66) 1E	1Z (1) 1E
20. Ni ez dakit.	2Z (1) 2E	1Z (1) 1E	1Z (1) 1E	1Z (1) 1E
21. Ni gizonak ikusi ditut.	4Z (1) 4E	1Z (1) 1E	2Z (1) 2E	1Z (1) 1E
22. UEUk, heuskal unibertsitatea definizioan, hiru puntu definitzen du.	3Z (1,3) 3E	2Z (1,5) 0E	3Z (1,3) 3E	2Z (1,5) 0E
23. Ziberespazioan dabilzan pertsonen dagokion izena da.	192Z (2,5) 192E	64Z (2,5) 64E	48Z (2) 0E	16Z (2) 0E
24. Energia mota honi, energia jariakor deritze.	1Z (2) 0E	1Z (2) 0E	1Z (2) 0E	1Z (2) 0E
25. Multzo bakoitzaren barnean dauden erregelak ezaugarri amankomunak dituzte noski.	12Z (1,66) 12E	6Z (1,6) 3E	12Z (2) 6E	6Z (2) 3E
26. Honela, izen sintagma guztiak balio berdinak arabera sailkatuak izango ditugu eta esan dezakegu sintagma guztiak definizio antzerakoa izango dutela.	96Z (2,66) 48E	48Z (2,66) 48E	96Z (2,66) 96E	48Z (2,66) 48E
27. Hizkuntzalaritza teoriko nahiz konputazionalan aditza kontsideratzen da buru eta honi atxikitzen zaizkio funtzioa ezberdinak beteko dituzten perpauseko elementuak.	24Z (3,3) 20E	8Z (3,5) 6E	24Z (3) 20E	8Z (3,5) 6E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
28. Orain datozen bi erregelek ergatiboak subjektu bezala markatuko ditu.	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E
29. Deklinabide kasu bezala, ezaugarri hauek hartu ditugu bezeroak.	2Z (1) 2E	1Z (1) 1E	2Z (1) 2E	1Z (1) 1E
30. Honi, irailerako asignatura bat utzi izanak batzen badiogu, ez zigun denbora eman proiektua irailean aurkezteko.	16Z (3) 16E	4Z (3) 4E	8Z (3) 8E	2Z (3) 2E
31. Tesia Aitor Soroa Etxabe egina da, eta bere zuzendaria Xabier Artola Zubillaga da.	144Z (2) 132E	3Z (2) 3E	6Z (2) 6E	3Z (2) 3E
32. Bertako jendea ni begira zeunden.	36Z (1) 0E	18Z (1) 0E	18Z (1) 0E	18Z (1) 0E
33. Baina ez zen ezertarako balio.	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E
34. Nik isilik gelditu nintzen.	3Z (1) 3E	3Z (1) 3E	3Z (1) 0E	3Z(1) 0E
35. Gizonak niregana gehiago hurbildu zen eta indarrez bultzatu zidan.	18Z (1) 18E	9Z (1) 9E	18Z (1) 18E	9Z(1) 9E
36. Gu oso urduri jarri ginen eta bakoitza bere etxeko ibilbidea hartu zuen.	108Z (2,66) 81E	27Z (2,66) 27E	54Z (2,66) 54E	27Z (2,66) 27E
37. Orduan, guk etxera joan ginen eta udaltzainak ijituaren atzetik.	9Z (1) 9E	6Z (1) 6E	9Z (1) 0E	6Z(1) 0E
38. Baina Talaiakoak nere aldean zeuden beti bezala eta defenditu zidaten.	24Z (2) 24E	24Z (2) 24E	24Z (2) 24E	24Z(2) 24E
39. Egun batean diskusio handi bat zegoenean, ni irabazi nuen.	12Z (2) 6E	6Z (1) 0E	6Z (2) 0E	6Z(1) 0E
40. Nire ametsean, nik ohean nengoen lotan, eta beti bezala, nire amak zazpi eta erdietan esnatu zidan eskolara joateko.	36Z (3,33) 36E	24Z (3) 24E	36Z (3,33) 36E	24Z(3) 24E
41. Nik ohetik altxatu nintzen eta nire urtebetetzean oparitutako arropak jantzi nituen.	108Z (1,66) 108E	72Z (1,75) 72E	108Z (1,66) 108E	72Z(1,75) 72E
42. Amak eskua sorbaldan jarri zidan esanez gau hartan nire osabak hil zela.	48Z (2,33) 48E	8Z (2,5) 8E	48Z (2,33) 48E	8Z(2,5) 8E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
43. Nik harrituta geratu nintzen hori entzutean.	3Z (2) 0E	3Z (2) 0E	3Z (1) 0E	3Z(1) 0E
44. Hamaiketakoa eman ziguten; sagarra bat guztiei.	6Z (A) 3E	6Z (1) 3E	3Z (1) 0E	3Z(1) 0E
45. Gizonak esaten zidan gezurti galanta bat nintzela eta nik bakarrik nengoela parking horretan.	64Z (1,43) 32E	16Z (1,43) 0E	52Z (1,84) 52E	13Z(2,8) 0E
46. Nik negarrez nengoen, arropa puskatua nuen, ilea guztiz despeinatua.	4Z (1) 0E	4Z (1) 0E	4Z (1) 0E	4Z(1) 0E
47. Bapatean nire gelako bati deitu egin nion ikusi niolako, baina ez zidan kasurik egiten.	3Z (3) 3E	3Z (3) 3E	3Z (3) 3E	3Z(3) 3E
48. Eta txarrena, nik ametsa hori bakarrik ametsgaiztoa zela pentsatu nuela, baina egitan pasatu zen, eta hori ezin da aldatu.	Z++	48Z (2,25) 48E	Z++	48Z(2,75) 48E
49. Doitasun askoz ere hobe erakutsi dute eskuz sortutako erregelak erabiltzen dituzten teknikak.	96Z (2,66) 16E	64Z (2,75) 0E	96Z (2,66) 16E	64Z (2,75) 0E
50. Eta eragin aundiak edukiko du balio ekologiko garrantzitsua duten eremu natural batzuetan.	2Z (2) 2E	1Z (2) 1E	2Z (2) 2E	1Z (2) 1E
51. Azpiegitura berri honek Altsasu eta Gipuzkoa artean dauden errepideak bikoiztuko du.	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E
52. Modu honetan ingurugiroak jasatzen duen eraginak gutxituko eta azpiegiturako ugalketa ekidingo lirateke Altsasuko dermioan.	12Z (2) 12E	2Z (2) 2E	12Z (2) 12E	2Z (2) 2E
53. Autobide berri hau eragin aundia eduki dezake.	4Z (1) 4E	1Z (1) 1E	4Z (1) 4E	1Z (1) 1E
54. Han eman zenuten adierazpenak ikusi ondoren, komunikatu honek atera genuen.	54Z (1,33) 54E	36Z (1,5) 36E	54Z (1,33) 54E	36Z (1,5) 36E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
55. Guri ere bururatu zaizkigu antzeko adierazpenen bat.	8Z (2) 4E	8Z (2) 4E	4Z (2) 0E	4Z (2) 0E
56. Horrela Garraxi Irratia alternatiba txiki bat izan nahi du gaur egun komunikazio bideak menperatzen duten enpresentzat.	24Z (3) 12E	6Z (3) 0E	24Z (3) 12E	6Z (3) 0E
57. Beti izan dugu sekzio batzuk gutxi gora behera finkoak izan direnak denboraldi batzuetan.	2Z (2) 2E	1Z (2) 0E	2Z (2) 2E	1Z (2) 0E
58. Horrialde honek hemen zegoen.	9Z (1) 9E	9Z (1) 9E	9Z (1) 9E	9Z (1) 9E
59. Baina Nafarroako Gobernuaren proiektua ikusi ondoren, Altsasuko Udaletxea proposatu zuen proiektua.	36Z (1,66) 15E	6Z (1,5) 3E	18Z (1,66) 12E	6Z (1,5) 3E
60. Autobide berria Altsasuko dermioaren gainean izugarrizko inpaktuak eragingo du.	2Z (2) 2E	1Z (2) 0E	2Z (2) 0E	1Z (2) 0E
61. Kontutan hartu behar da gobernuko teknikariak, pribatuen behintzat, bikoizketa posiblea dela baieztatu dutela.	32Z (4,75) 0E	8Z (1,75) 0E	16Z (4,75) 0E	8Z (1,75) 0E
62. Altsasuko bizilagunak orain ez duten aisialdirako leku lasai bat izan dezaten.	2Z (3) 0E	2Z (3) 0E	2Z (3) 0E	2Z (3) 0E
63. Hainbat bilera egin ondoren, Altsasuko udala eta Zelai kaleko auzokideek akordio batera ailegatu dira.	18Z (2,66) 10E	4Z (3) 0E	18Z (3,33) 6E	4Z (3,5) 0E
64. Aztertu behar ziren gaiaren artean Zelai kaleko egitasmo berria zegoen.	32Z (1,75) 0E	16Z (1,75) 0E	16Z (1,75) 0E	16Z (1,75) 0E
65. Sindikatuek ez du baztertzen beste sektoreekin protesta bateratuak egitea.	3Z (2) 3E	3Z (2) 3E	3Z (2) 3E	3Z (2) 3E
66. Petrolio konpainiek "mozkin astronomikoak" eta kontsumitzailearentzat "galera ekonomikoak" ekarri duela adierazi zuen idatzi batean.	24Z (1) 24E	6Z (1) 6E	24Z (1) 24E	6Z (1) 6E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
67. Amoratuek, ohe batean etzanik, herstuki besarkatzen dira.	6Z (2) 6E	6Z (2) 6E	6Z (2) 6E	6Z (2) 6E
68. Espainiak krisi politikoa eta ekonomikoa sufritzen ari da.	36Z (1) 36E	4Z (1) 4E	36Z (1) 36E	4Z (1) 4E
69. Baina gure mutilarentzat prozesu hura naskagarri zirudien.	16Z (1) 16E	8Z (1) 0E	16Z (1) 0E	8Z (1) 0E
70. Jospin, Alderdi Sozialistaren hautesgaia, portzentaia handiena lortu du.	24Z (1) 24E	4Z (1) 0E	24Z (1) 24E	4Z (1) 0E
71. Hori eztiarekin zerikusirik ez dauka.	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
72. Beste multzoak hizkuntza nazionala dela sinisten dute.	8Z (2,25) 0E	4Z (2,25) 0E	8Z (2,25) 0E	4Z (2,25) 0E
73. Inportazioak, aldiz, inflakzioaren igoeraren arriskua ekarriko digute.	16Z (2) 16E	16Z (2) 16E	16Z (2) 16E	16Z (2) 16E
74. Mutila izugarritzko haizea nabaritu zuen.	18Z (1,3) 15E	6Z (1,5) 3E	18Z (1,3) 15E	6Z (1,5) 3E
75. Bazkaldu ondoren bilerak jarraituko dute.	4Z (1) 4E	2Z (1) 2E	4Z (1) 4E	2Z (1) 2E
Esaldi erroredunean errorea markatuta	42	40	36	34

E.1 Taula: Komunztadura-erroreen detekzioa. Emaitzak.

2. Komunztadura-erroreen detektziorako erregelak mendekotasun-zuhaitzetan aplikatu ditugunean kasu hoberenean 26 kasutan ez dela errorerik detektatu ikusi dugu. Arrazoiak banan-banan aztertu ditugu eta ondorengo taulan (ikus E.2) laburtu ditugu. Gogora dezagun, taula honetan azaltzen diren esaldietan inolako konbinazioarekin ez dela errorerik detektatu.

Testuan errorea ez detektatzeko arrazoiak zehaztu ditugun arren hemen-txe laburbilduko ditugu berriro (ikus informazio zehatzagoa VII kapituluko VII.7.2 atalean):

1. Mendekotasun-erlazioa gaizki ipini da (19 kasu).
2. Aditz modala ez da lotuta analizatu, eta beraz, gaizki lotu zaio bere mendekoari (4 kasu).
3. Desanbiguazio morfosintaktikoa gaizki egin da (2 kasu).

Esaldia	Arrazoa
1. Honek algoritmo baten bidez forma berriak asmatzeko gai da	Honek - (NCSUBJ) - da → “Honek” hitzak ez du %SINT etiketarik, beraz, ez du NCSUBJ etiketarik.
2. Sarean dabilen datu mugimenduak entzuten dituzte.	dabilen - (CMOD) -mugimenduak → “dabilen” hitza “datu”-rekin lotzen du, “mugimenduak” hitzarekin lotu beharrean.
3. Bestalde badaude beste hari mota bat RDSI izenekoak.	0 mendekotasun-etiketa.
6. Abiaduraren norantzan egindako indarra abiadura haundituko du.	indarra - NCSUBJ - haundituko eta du - AUXMOD - haundituko → “Indarra” NCOBJ eta NCSUBJ. Zuhaitz guztiek NCSUBJ ez.
9. Besteen ordenadoretan sartzen dabilena edo birusak bidaltzen duena da.	birusak - NCOBJ - bidaltzen → “Birusak” hitzari NCOBJ eta NCSUBJ. Zuhaitz guztiek NCSUBJ EZ.
17. Modema beti gure ordenagailura konektaturik egon behar du.	modema - NCSUBJ - egon eta egon - AUXMOD? - behar_du → “egon behar du” behar ez den bezala etiketatzen da.”behar du” ez da lotzen.
18. Hirugarrena orriaren lekua edo mota esan nahi du.	Hirugarrena - NCSUBJ - esan eta esan - AUXMOD? - nahi_du → “esan nahi du” behar ez den bezala etiketatzen da. “nahi du” ez da lotzen.
24. Energia mota honi, energia jaria- kor deritze.	mota - NCZOBJ- deritze → Ez dago “mota”ri NCZOBJ ipintzeko dep. erregelarik.
32. Bertako jendea ni begira zeunden.	Bertako jendea hitzari NCSUBJ eta NCOBJ. Zuhaitz guztiek NCSUBJrik ez.
33. Baina ez zen ezertarako balio.	“balio” ez du “balio_izan” aditz moduan analizatu.
39. Egun batean diskusio handi bat zegoenean, ni irabazi nuen.	ni - NCSUBJ - irabazi eta irabazi - AUXMOD - nuen → “ni” hitzari NCOBJ eta NCSUBJ. Zuhaitz guztiek NCSUBJ EZ.
43. Nik harrituta geratu nintzen hori entzutean.	nik - NCSUBJ - geratunik - NCSUBJ - geratu eta geratu - AUXMOD - nintzen → “harrituta” gitza aditz moduan analizatzen du eta ondorioz “ni”, “harrituta”rekin lotzen du.
44. Hamaiketakoa eman ziguten; sagarra bat guztiei.	bat - DETMOD - sagarra → “bat”-i ez dio inolako etiketarik ipintzen.
46. Nik negarrez nengoen, arropa puskatua nuen, ilea guztiz despeinatua.	nik - NCSUBJ - nengoen → “nik” hitzari ez dio inolako dependentzia etiketarik ipintzen.
49. Doitasun askoz ere hobea erakutsi dute eskuz sortutako erregelak erabiltzen dituzten teknikak.	teknikak - NCSUBJ - erakutsi → “teknikak”, “sortutako” aditzarekin lotzen du. Dependentzia gramatika ez dago prestatuta subjektua 3 aditz haratago lotzeko.
55. Guri ere bururatu zaizkigu antzeko adierazpenen bat.	bat - DETMOD - adierazpenen eta adierazpenen - NCSUBJ - bururatu eta bururatu - AUXMOD - zaizkigu → “adierazpen” hitzari ez dio mendekotasun erlaziorik ipintzen.
<i>Jarraipena hurrengo orrialdean</i>	

Esaldia	Arrazoia
56. Horrela Garraxi Irratia alternatiba txiki bat izan nahi du gaur egun komunikazio bideak menperatzen duten enpresentzat.	Irratia - NCSUBJ - izan_nahi_du → “izan_nahi_du” ez da lotzen.
59. Baina Nafarroako Gobernuaren proiektua ikusi ondoren, Altsasuko Udaletxea proposatu zuen proiektua.	udaletxea - NCSUBJ - proposatu eta proposatu - AUXMOD - zuen → “udaletxea” objektu bailitzan analizatzen du.
61. Kontutan hartu behar da gobernuko teknikariak, pribatuan behintzat, bikoizketa posiblea dela baieztatu dutela.	teknikariak - NCSUBJ - baieztatu eta baieztatu - AUXMOD - dutela → Zuhaitza zatituegia. Koma arteko esaldiak mendekotasun-zuhaitza zati gehiagoetan banatzen du.
62. Altsasuko bizilagunak orain ez duten aisialdirako leku lasai bat izan dezaten.	bizilagunak - NCSUBJ - izan eta izan - AUXMOD - dezaten → “bizilagunak” hitzak ez du mendekotasun etiketarik.
63. Hainbat bilera egin ondoren, Altsasuko udala eta Zelai kaleko auzokideek akordio batera ailegatu dira.	eta ¹ - NCSUBJ - ailegatu eta ailegatu - AUXMOD - dira → Dependentsia gramatikan ez dira loturazko elementuak landu.
64. Aztertu behar ziren gaiaren artean Zelai kaleko egitasmo berria zegoen.	Berezia. aztertu_behar_ziren - CMOD - gaiaren_arte → Oso berezia. Analisi automatikoan ez da ez postposiziorik, ezta aditz kate konplexurik lotzen.
71. Hori eztiarekin zerikusirik ez dauka.	hori - NCSUBJ - dauka → “hori” hitzak ez du dependentsia etiketarik.
72. Beste multzoak hizkuntza nazionala dela sinisten dute.	multzoak - NCSUBJ - sinisten eta sinisten - AUXMOD - dute → “multzoak” hitzak ez du inolako dependentsia etiketarik.
74. Mutila izugarritzko haizea nabaritu zuen.	mutila - NCSUBJ - nabaritu eta nabaritu - AUXMOD - zuen → “mutila” eta “haizea”-k subjektu eta objektu etiketak dituzte. Zuhaitz guztietan ez da erregela aplikatzen.

E.2 Taula: Komunztadura-erroreak ez detektatzeko arrazoiak.

3. Eranskinetako lehen taulan azaldu dugun esperimentu bera errepikatu dugu E.3 taulan, baina, kasu honetan, E.1 taulako esaldietako bakoitzaren ordain zuzen batekin. Maiz, esaldi zuzenetan errorea detektatzen da. Gehienetan, mendekotasun-erlazioak gaizki ipintzen direla, eta beste zenbaitetan, aditzen azpikategorizazio sailkapena okerra delako.

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
1. Hau algoritmo baten bidez forma berriak asmatzeko gai da.	2Z (1) OE	1Z (1) OE	2Z (1) OE	1Z (1) OE
2. Sarean dabiltsan datu mugimenduak entzuten dituzte.	12Z (1,33) 6E	6Z (1,33) OE	12Z (1,33) 6E	6Z (1,33) OE
3. Bestalde badago beste hari mota bat RDSI izenekoa.	OZ	OZ	OZ	OZ
<i>Jarraipena hurrengo orrialdean</i>				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
4. Ez duela lehengairik behar energia elektrikoa ateratzeko.	8Z (2) 8E	4Z (2) 4E	8Z (2) 8E	4Z (2) 4E
5. Adimen Naturalak eboluzioa du eta artifizialak berriz ez.	6Z (1) 6E	2Z (1) 0E	4Z (1) 4E	1Z (1) 0E
6. Abiaduraren norantzan egindako indarrak abiadura haundituko du.	192 (1,66) 160E	64Z (1,75) 32E	192 (1,66) 160	64z (1,75) 32E
7. Azken biek zerikusi haundia daukate.	2Z (1) 2E	1Z (1) 0E	2Z (1) 0E	1Z (1) 0E
8. Hackerrek nolabaiteko fama bat dute baina ez dira gaiztoak beraien atakeetan.	2Z (3) 2E	1Z (3) 0E	2Z (3) 0E	1Z (3) 0E
9. Besteen ordenadoretan sartzen dabilena edo birusak bidaltzen dituen da.	12Z (2) 6E	6Z (2) 0E	12Z (2) 6E	6Z (2) 0E
10. Makilak bolari eragin diolako indarra.	4Z (1) 0E	4Z (1) 0E	4Z (1) 0E	4Z (1) 0E
11. Ni ez nago konforme.	2Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
12. Ni ez noa.	2Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
13. Zer dira nabigatzaileak ?.	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
14. Ordenagailuetan dagoen informazio guztia orrialde horietan dago antolatua.	12Z (1,66) 0E	6Z (1,66) 0E	12Z (1,66) 0E	6Z (1,66) 0E
15. Erabiltzaileari sare batean egon daitekeen urruneko beste ordenagailuekin konektatzeko aukera ematen dio.	4Z (2) 4E	4Z (2) 4E	4Z (2) 2E	4Z (2) 2E
16. Oso arriskutsua da nukleoak eskape bat baldin badauka.	36Z (2) 36E	18Z (2) 18E	36Z (2) 36E	18Z (2) 18E
17. Modemak beti gure ordenagailura konektaturik egon behar du.	3Z (2) 0E	3Z (2) 0E3	3Z (2) 0E	3Z (2) 0E
18. Hirugarrenak orriaren lekua edo mota esan nahi du.	5Z (2) 0E	4Z (2) 0E	5Z (2) 0E	4Z (2) 0E
19. Beraien asmoak erabat suntsitzaileak dira.	3Z (1,33) 0E	1Z (1) 0E	3Z (1,66) 0E	1Z (1) 0E
20. Nik ez dakit.	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
21. Nik gizonak ikusi ditut.	2Z (1) 1E	1Z (1) 0E	2Z (1) 1E	1Z (1) 0E
22. UEUk, heuskal unibertsitatea definizioan, hiru puntu definitzen ditu.	3Z (1,33) 3E	2Z (1,5) 0E	3Z (1,33) 3E	2Z (1,5) 0E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
23. Ziberespazioan dabiltzan pertsonen dagokien izena da.	192Z (2,5) 0E	64Z (2,5) 0E	48Z (2) 0E	16Z (2) 0E
24. Energia mota honi, energia jariakor deritzo.	2Z (1) 0E	2Z (1) 0E	2Z (1) 0E	1Z(2) 0E
25. Multzo bakoitzaren barnean dauden erregelek ezaugarri amankomunak dituzte noski.	6Z (1,66) 6E	3Z (1,66) 0E	6Z (1,66) 0E	3Z (1,66) 0E
26. Honela, izen sintagma guztiak balio berdinak arabera sailkatuak izango ditugu eta esan dezakegu sintagma guztiek definizio antzerakoa izango dutela.	12Z (2,66) 12E	6Z (2,66) 6E	12Z (2,66) 12E	6Z (2,66) 6E
27. Hizkuntzalaritza teoriko nahiz konputazionalan aditza kontsideratzen da buru eta honi atxikitzen zaizkio funtzio ezberdinak beteko dituzten perpauseko elementuak.	24Z (3,33) 16E	8Z (3,5) 4E	24Z (3,33) 12E	8Z (3,5) 4E
28. Orain datozen bi erregelek ergatiboak subjektu bezala markatuko dituzte.	3Z (1,66) 0E	3Z (1,66) 0E	3Z (1,66) 0E	3Z (1,66) 0E
29. Deklinabide kasu bezala, ezaugarri hauek hartu ditu bezeroak.	2Z (1) 2E	1Z (1) 0E	2Z (1) 0E	1Z (1) 0E
30. Honi, irailerako asignatura bat utzi izana batzen badiogu, ez zigun denbora eman proiektua irailean aurkezteko.	8Z (3) 8E	4Z (3) 4E	4Z (3) 4E	2Z (3) 2E
31. Tesia Aitor Soroa Etxabek egin da, eta bere zuzendaria Xabier Artola Zubillaga da.	144Z (2) 132E	3Z (2) 3E	6Z (2) 6E	3Z (2) 3E
32. Bertako jendea ni begira zegoen.	36Z (1) 0E	18Z (1) 0E	18Z (1) 0E	18Z (1) 0E
33. Baina ez zuen ezertarako balio.	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E
34. Ni isilik gelditu nintzen.	6Z (1) 3E	3Z (1) 0E	3Z (1) 0E	3Z (1) 0E
35. Gizona niregana gehiago hurbildu zen eta indarrez bultzatu ninduen.	18Z (1) 9E	18Z (1) 9E	18Z (1) 9E	18Z (1) 9E
36. Gu oso urduri jarri ginen eta bakoitzak bere etxeko ibilbidea hartu zuen.	108Z (2,66) 81E	27Z (2,66) 27E	54Z (2,66) 54E	27Z (2,66) 27E
Jarraipena hurrengo orrialdean				

Esaldia	Emaizak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
37. Orduan, gu etxera joan ginen eta udaltzainak ijituaren atzetik.	18Z (1) 9E	18Z (1) 9E	9Z (1) 0E	9Z (1) 0E
38. Baina Talaiakoak nere aldean zeuden beti bezala eta defenditu ninduten.	24Z (2) 24E	24Z (2) 24E	24Z (1) 24E	24Z (2) 24E
39. Egun batean diskusio handi bat zegoenean, nik irabazi nuen.	6Z (2) 0E	6Z (1) 0E	6Z (2) 0E	6Z (1) 0E
40. Nire ametsean, ni ohean nengoen lotan, eta beti bezala, nire amak zazpi eta erdietan esnatu ninduen eskolara joateko.	72Z (3,33) 24E	24Z (3) 12E	36Z (3,33) 12E	24Z (3) 12E
41. Ni ohetik altxatu nintzen eta nire urtebetetzean oparitutako arropak jantzi nituen.	216Z (1,66) 144E	72Z (1,75) 0E	108Z (1,66) 108E	72Z (1,75) 0E
42. Amak eskua sorbaldan jarri zidan esanez gau hartan nire osaba hil zela.	48Z (2,33) 48E	4Z (2) 4E	48Z (2,33) 48E	4Z (2) 4E
43. Ni harrituta geratu nintzen hori entzutean.	6Z (2) 0E	3Z (2) 0E	6Z (2) 0E	3Z (2) 0E
44. Hamaiketako eman ziguten; sagar bat guztiei.	12Z (2) 6E	6Z (2) 3E	6Z (2) 0E	3Z (2) 0E
45. Gizonak esaten zidan gezurti galanta nintzela eta ni bakarrik nengoela parking horretan.	16Z (1,25) 8E	4Z (1,25) 0E	16Z (1,75) 16E	4Z (2,75) 0E
46. Ni negarrez nengoen, arropa puskatua nuen, ilea guztiz despeinatua.	4Z (1) 0E	4Z (1) 0E	4Z (1) 0E	4Z (1) 0E
47. Bapatean nire gelako bati dei egin nion ikusi nuelako, baina ez zidan kasurik egiten.	36Z (2,5) 27E	9Z (2) 9E	36Z (2,5) 27E	9Z (2) 9E
48. Eta txarrena, nik amets hori bakarrik ametsgaiztoa zela pentsatu nuela, baina egitan pasatu zen, eta hori ezin da aldatu.	Z++	48Z (2,25) 48E	Z++	48Z (2,75) 48E
49. Doitasun askoz ere hobe erakutsi dute eskuz sortutako erregelak erabiltzen dituzten teknikak.	48Z (2,66) 8E	32Z (2,75) 0E	48Z (2,66) 8E	32Z (2,75) 0E
Jarraipena hurrengo orrialdean				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
50. Eta eragin aundia edukiko du balio ekologiko garrantzitsua duten eremu natural batzuetan.	2Z (2) 2E	1Z (2) 0E	2Z (2) 0E	1Z (2) 0E
51. Azpiegitura berri honek Altsasu eta Gipuzkoa artean dauden errepideak bikoiztuko ditu.	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E	3Z (1,66) 3E
52. Modu honetan ingurugiroak jasatzen duen eragina gutxituko eta azpiegiturako ugalketa ekidingo lirateke Altsasuko dermioan.	4Z (2) 4E	2Z (2) 2E	4Z (2,5) 4E	2Z (2) 2E
53. Autobide berri honek eragin aundia eduki dezake.	2Z (1) 2E	1Z (1) 0E	2Z (1) 0E	1Z (1) 0E
54. Han eman zenituzten adierazpenak ikusi ondoren, komunikatu hau atera genuen.	108Z (1,33) 108E	36Z (1,5) 0E	108Z (1,33) 0E	36Z (1,5) 0E
55. Guri ere bururatu zaizkigu antzeko adierazpen batzuk.	2Z (1) 0E	2Z (1) 0E	2Z (1) 0E	2Z (1) 0E
56. Horrela Garraxi Irratiak alternatiba txiki bat izan nahi du gaur egun komunikazio bideak menperatzen dituzten enpresentzat.	4Z (3) 2E	2Z (3) 0E	4Z (3) 2E	2Z (3) 0E
57. Beti izan ditugu sekzio batzuk gutxi gora behera finkoak izan direnak denboraldi batzuetan.	2Z (2) 2E	1Z (2) 0E	2Z (2) 2E	1Z (2) 0E
58. Horrialde hau hemen zegoen.	18Z (1) 0E	9Z (1) 0E	18Z (1) 0E	9Z (1) 0E
59. Baina Nafarroako Gobernuaren proiektua ikusi ondoren, Altsasuko Udaletxeak proposatu zuen proiektua.	36Z (1,66) 15E	6Z (1,5) 3E	18Z (1,66) 12E	6Z (1,5) 3Z
60. Autobide berriak Altsasuko dermioaren gainean izugarritzko inpaktua eragingo du.	2Z (2) 0E	1Z (2) 0E	1Z (2) 0E	1Z (2) 0E
61. Kontutan hartu behar da gobernuko teknikariek, pribatuan behintzat, bikoizketa posiblea dela baieztatu dutela.	32Z (4,75) 0E	8Z (4,75) 0E	16Z (4,75) 0E	8Z (4,75) 0E
<i>Jarraipena hurrengo orrialdean</i>				

Esaldia	Emaitzak			
	Erlaxatua		Erlaxatu gabe	
	3-2	3-3	3-2	3-3
62. Altsasuko bizilagunek orain ez duten aisialdirako leku lasai bat izan dezaten.	4Z (3) 0E	8Z (4,75) 0E	4Z (3) 0E	4Z (3) 0E
63. Hainbat bilera egin ondoren, Altsasuko udala eta Zelai kaleko auzokideak akordio batera ailegatu dira.	18Z (2,66) 10E	4Z (3) 0E	18Z (3,33) 6E	4Z (3,5) 0E
64. Aztertu behar ziren gaien artean Zelai kaleko egitasmo berria zegoen.	32Z (1,75) 0E	16Z (1,75) 0E	16Z (1,75) 0E	16Z (1,75) 0E
65. Sindikatuak ez du baztertzen beste sektoreekin protesta bateratuak egitea.	3Z (2) 0E	3Z (2) 0E	3Z (2) 0E	3Z (2) 0E
66. Petrolio konpainiek "mozkin astronomikoak" eta kontsumitzailearentzat "galera ekonomikoa" ekarri dutela adierazi zuten idatzi batean.	24Z (1) 24E	6Z (1) 6E	24Z (1) 24E	6Z (1) 6E
67. Amaratuek, ohe batean etzanik, herstuki elkar besarkatzen dute.	12Z (2) 6E	6Z (2) 0E	6Z (2) 0E	6Z (2) 0E
68. Espainia krisi politikoa eta ekonomikoa sufritzen ari da.	72Z (1) 72E	4Z (1) 4E	144Z (1,75) 36E	4Z (1) 4E
69. Baina gure mutilarentzat prozesu hark naskagarria zirudien.	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E	8Z (1) 0E
70. Jospinek, Alderdi Sozialistaren hautesgaiak, portzentaia handiena lortu du.	24Z (1) 24E	4Z (1) 0E	24Z (1) 24E	4Z (1) 0E
71. Horrek eztiarekin zerikusirik ez dauka.	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E	1Z (1) 0E
72. Beste multzoek hizkuntza nazionala dela sinisten dute.	8Z (2,25) 0E	4Z (2,25) 0E	8Z (2,5) 0E	4Z (2,5) 0E
73. Inportazioek, aldiz, inflakzioaren igoeraren arriskua ekarriko digute.	16Z (2) 8E	8Z (2) 0E	16Z (2) 8E	8Z (2) 0E
74. Mutilak izugarrizko haizea nabaritu zuen.	6Z (1) 3E	3Z (1) 0E	6Z (1) 3E	3Z (1) 0E
75. Bazkaldu ondoren bilerek jarraituko dute.	2Z (1) 2E	2Z (1) 2E	2Z (1) 2E	2Z (1) 2E

E.3 Taula: Komunztadura-erroreen detekzioa esaldi zuzenetan.
Emaitzak

4. Ondorengo taulan (E.4 taulan), eranskin honetako E.1 eta E.3 tauletako emaitzak laburtzen dira. Esaldi erroredunetako bakoitzari eta dagokion ordain zuzen bati dagozkion datuak parez pare idatzi dira. Honen helburua, erroreen detektziorako sistemaren *fidagarritasuna* neurtzea izan da. Hots, zenbat kasutan detektatzen den errorea arrazoi egokiarekin, eta zenbatetan aurreikusi ez dugun arrazoiren batengatik.

Esaldia	Emaitzak				
	Erlaxatua		Erlaxatu gabe		Errorea?
	3-2	3-3	3-2	3-3	
1.	-	-	-	-	
1.ZUZ.	-	-	-	-	
2.	-	-	-	-	
2.ZUZ.	-	-	-	-	
3.	-	-	-	-	
3.ZUZ.	-	-	-	-	
4.	✓	✓	✓	✓	ERR! Akatsa: Dependentsia etiketatze okerra. Bi objektu aditz okerrari lotuta.
4.ZUZ.	✓	✓	✓	✓	
5.	✓	-	✓	✓	ERR! Akatsa: Analisi morfosintaktikoa. Subjektua <i>absolutiboan</i> aditz laguntzailea NOR-NORK sailekoa izanik.
5.ZUZ.	✓	-	✓	-	
6.	-	-	-	-	
6.ZUZ.	-	-	-	-	
7.	✓	✓	✓	✓	ERR! Akatsa: Dependentsia etiketatze okerra. Bi subjektu aditz berarekin lotuta, batak analisi okerra.
7.ZUZ.	✓	-	-	-	
8.	✓	-	-	-	ERR! Akatsa: Dependentsia etiketatze okerra. Subjektu okerra.
8.ZUZ.	✓	-	-	-	
9.	-	-	-	-	
9.ZUZ.	-	-	-	-	
10.	✓	✓	✓	✓	ERR!!
10.ZUZ.	-	-	-	-	
11.	✓	✓	-	-	ERR!
11.ZUZ.	-	-	-	-	
12.	✓	✓	-	-	ERR!
12.ZUZ.	-	-	-	-	
13.	✓	✓	✓	✓	ERR!
13.ZUZ.	-	-	-	-	
14.	✓	✓	✓	✓	ERR!
14.ZUZ.	-	-	-	-	
15.	✓	✓	-	-	ERR! Akatsa: Dependentsia etiketatze okerra. Zehar objektua aditz okerrarekin lotzen du eta honen aditz laguntzaileak ez du zehar objekturik onartzen.
15.ZUZ.	✓	✓	-	-	
Jarraipena hurrengo orrialdean					

Esaldia	Emaizak				
	Erlaxatua		Erlaxatu gabe		Errorea?
	3-2	3-3	3-2	3-3	
9.	-	-	-	-	
9.ZUZ.	-	-	-	-	
10.	✓	✓	✓	✓	ERR!!
10.ZUZ.	-	-	-	-	
11.	✓	✓	-	-	ERR!
11.ZUZ.	-	-	-	-	
12.	✓	✓	-	-	ERR!
12.ZUZ.	-	-	-	-	
13.	✓	✓	✓	✓	ERR!
13.ZUZ.	-	-	-	-	
14.	✓	✓	✓	✓	ERR!
14.ZUZ.	-	-	-	-	
15.	✓	✓	-	-	ERR!
15.ZUZ.	✓	✓	-	-	Dependentzia etiketatze okerra. Zehar objektua aditz okerrarekin lotzen du.
16.	✓	✓	-	✓	ERR!
16.ZUZ.	✓	✓	✓	✓	Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta eta hitz berean subjektu/objektu analisiak.
17.	-	-	-	-	
17.ZUZ.	-	-	-	-	
18.	-	-	-	-	
18.ZUZ.	-	-	-	-	
19.	-	✓	-	-	ERR!
19.ZUZ.	-	-	-	-	
20.	✓	✓	✓	✓	ERR!
20.ZUZ.	-	-	-	-	
21.	✓	✓	✓	✓	ERR!!
21.ZUZ.	-	-	-	-	
22.	✓	-	✓	-	ERR!
22.ZUZ.	✓	-	✓	-	Dependentzia etiketatze okerra. Hiru subjektu aditz berarekin lotuta eta baten analisiak erreareari dagozkio.
23.	✓	✓	-	-	ERR!
23.ZUZ.	-	-	-	-	
24.	-	-	-	-	
24.ZUZ.	-	-	-	-	
25.	✓	-	-	-	ERR!
25.ZUZ.	✓	-	-	-	Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta eta baten analisiak erreareari dagozkio.
26.	-	✓	✓	✓	ERR!
26.ZUZ.	✓	✓	✓	✓	Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta eta baten analisiak erreareari dagozkio. + Aditzaren azpikategorizazioa. "izango" DU okerra omen.
27.	-	-	-	-	
27.ZUZ.	-	-	-	-	
28.	✓	✓	✓	✓	ERR!!
Jarraipena hurrengo orrialdean					

Esaldia	Emaizak				
	Erlaxatua		Erlaxatu gabe		Errorea?
	3-2	3-3	3-2	3-3	
28.ZUZ.	-	-	-	-	
29. 29.ZUZ.	✓ ✓	✓ -	✓ -	✓ -	ERR! Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta.
30. 30.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓	ERR!! Aditzaren azpikategorizazioa. “batu” DIO okerra omen.
31. 31.ZUZ.	- -	✓ ✓	✓ ✓	✓ ✓	ERR! Dependentzia etiketatze okerra. Subjektu anitz aditz berarekin lotuta.
32. 32.ZUZ.	- -	- -	- -	- -	
33. 33.ZUZ.	- -	- -	- -	- -	
34. 34.ZUZ.	✓ -	✓ -	- -	- -	ERR!!
35. 35.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!!
36. 36.ZUZ.	- -	✓ ✓	✓ ✓	✓ ✓	ERR!! Dependentzia etiketatze okerra. Objektua markatu da aditz laguntzailea NOR paradigmakoa izanik.
37. 37.ZUZ.	✓ -	✓ -	- -	- -	ERR!!
38. 38.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓ ✓	ERR!! Analisi morfosintaktikoa. Subjektua <i>erga-tiboan</i> aditz laguntzailea NOR sailekoa izanik.
39. 39.ZUZ.	- -	- -	- -	- -	
40. 40.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!!
41. 41.ZUZ.	✓ -	✓ -	✓ ✓	✓ -	ERR!!
42. 42.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓	ERR!! Dependentzia etiketatze okerra. Aditz berarekin lotutako hiru hitzetan objektu eta subjektu etiketa aldi berean.
43. 43.ZUZ.	- -	- -	- -	- -	
44. 44.ZUZ.	- -	- -	- -	- -	
45. 45.ZUZ.	- -	- -	✓ ✓	- -	ERR! Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta.
46. 46.ZUZ.	- -	- -	- -	- -	
47. 47.ZUZ.	✓ -	✓ ✓	✓ -	✓ ✓	ERR!! Dependentzia etiketatze okerra. Bi objektu aditz berarekin lotuta.
<i>Jarraipena hurrengo orrialdean</i>					

Esaldia	Emaizak				
	Erlaxatua		Erlaxatu gabe		Errorea?
	3-2	3-3	3-2	3-3	
48. 48.ZUZ.	z++ z++	✓ ✓	z++ z++	✓ ✓	ERR!!
49. 49.ZUZ.	- -	- -	- -	- -	
50. 50.ZUZ.	✓ ✓	✓ -	✓ -	✓ -	ERR!! Dependentzia etiketatze okerra. Objektua, subjektu moduan etiketatu da.
51. 51.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓ ✓	ERR! Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta eta baten analisiak erreareari dagozkio.
52. 52.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓ ✓	ERR! Dependentzia etiketatze okerra. Objektua markatu da aditz laguntzailea NOR paradigmakoa izanik.
53. 53.ZUZ.	✓ ✓	✓ -	✓ -	✓ -	ERR! Dependentzia etiketatze okerra. Bi subjektu aditz berarekin lotuta.
54. 54.ZUZ.	✓ ✓	✓ -	✓ -	✓ -	ERR! Dependentzia etiketatze okerra. Objektua, subjektu moduan etiketatu du (objektu moduan ere bai).
55. 55.ZUZ.	- -	- -	- -	- -	
56. 56.ZUZ.	- -	- -	- -	- -	
57. 57.ZUZ.	✓ ✓	- -	✓ ✓	- -	ERR! Dependentzia etiketatze okerra.
58. 58.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!
59. 59.ZUZ.	- -	- -	- -	- -	
60. 60.ZUZ.	✓ -	- -	- -	- -	ERR!!
61. 61.ZUZ.	- -	- -	- -	- -	
62. 62.ZUZ.	- -	- -	- -	- -	
63. 63.ZUZ.	- -	- -	- -	- -	
64. 64.ZUZ.	- -	- -	- -	- -	
65. 65.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!!
66. 66.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓ ✓	ERR!! Dependentzia etiketatze okerra. Hiru subjektu aditz berarekin lotuta eta baten analisiak erreareari dagozkio.
Jarraipena hurrengo orrialdean					

Esaldia	Emaitzak				
	Erlaxatua		Erlaxatu gabe		Errorea?
	3-2	3-3	3-2	3-3	
67. 67.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!
68. 68.ZUZ.	✓ ✓	✓ ✓	✓ -	✓ ✓	ERR! Dependentzia etiketatze okerra. Aditz berarekin lotutako hiru hitzetan objektu eta subjektu etiketa aldi berean.
69. 69.ZUZ.	✓ -	- -	- -	- -	ERR!
70. 70.ZUZ.	✓ ✓	- -	✓ ✓	- -	ERR!! Dependentzia etiketatze okerra. Hiru subjektu aditz berarekin lotuta eta baten analisiak erroreari dagozkio.
71. 71.ZUZ.	- -	- -	- -	- -	
72. 72.ZUZ.	- -	- -	- -	- -	
73. 73.ZUZ.	✓ -	✓ -	✓ -	✓ -	ERR!!
74. 74.ZUZ.	- -	- -	- -	- -	
75. 75.ZUZ.	✓ ✓	✓ ✓	✓ ✓	✓ ✓	ERR!! Aditzaren azpikategorizazioa. “jarraitu” ZAIO okerra omen.

E.4 Taula: Komunztadura-erroreen detekzioa. Emaitzak (2).

5. Azken taula hau analizatu ondoren ateratako emaitzak E.5 taulan labur ditzakegu. Taula honetan, erroreen detekzio “erreala” kalkulatu dugu. Detekzioa erreala dela erabaki dugu, esaldi erroredunean errorea detektatu bada (EE) eta dagokion orain zuzenean ez bada errorerik detektatu. Datu horiek kontuan hartuta, komunztadura-erroreen detekziorako desanbiguazio konbinaziorik egokiena 3-3-Erlaxatua da.

	3-2 Erlaxatua	3-3 Erlaxatua	3-2 Erlaxatu gabe	3-3 Erlaxatu gabe
Esaldi erroredunean errorea markatuta (EE)	49	46	39	37
Esaldi zuzenean errorea markatuta (alarma faltsua, AF)	24	17	18	15
Erroreen detekzio erreala (EE - AF)	25	29	21	22
Erroreak detektatu gabe	50	46	54	53

E.5 Taula: Sintaxi-anbiguotasuna eta erlaxazioa. Emaitzak.