

Euskal Herriko Unibertsitatea/ Universidad del País Vasco



Lengoaia eta Sistema Informatikoak Saila
Departamento de Lenguajes y Sistemas Informáticos

Matxin

Erregeletan oinarritutako itzulpen automatikoko
sistema baten eraikuntza
estaldura handiko baliabide linguistikoak berrerrabiliz

Aingeru Mayor Martinez

Informatikan Doktore titulua eskuratzeko aurkezturiko

Tesia

Donostia, 2007ko uztaila

Euskal Herriko Unibertsitatea/ Universidad del País Vasco



Lengoaia eta Sistema Informatikoak Saila
Departamento de Lenguajes y Sistemas Informáticos

Matxin

Erregeletan oinarritutako itzulpen automatikoko
sistema baten eraikuntza
estaldura handiko baliabide linguistikoak berrerabiliz

Aingeru Mayor Martinezek Arantza Díaz de Ilarraza eta Kepa Sarasola zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2007ko uztaila

Aingerutxo beltzak*

Pintzel atzerritarrarekin nire lurrean pintore jaioa.
Jarraikitzen zaren horrenbeste pintore zaharren norabidea pintorea.
Birjina izan edin zuria, aingerutxo beltzak pinta iezazkidazu, ere zerura
doazela beltz on guztiak.
Maitasunarekin pintak pintorea
zer haren kolorea erdeinatzen duzu, badakizu zeruan ere Jainkoa nahi
dituela?
Logelaren santuen pintorea, arima baduzu gorputzean, pintatzea zure
koadroetan beltzetatik ahaztu zinen?
Elizak pintatzen edun, aingerutxo ederrak pintatzen dituzu, baina inoiz ez
erabaki zinen aingeru beltz bat pintatu izatera.
(Andoni Matxin)

* *Matxin1.0* es-eu itzultzaile automatikoarekin itzultitako kanta:

Angelitos negros

Pintor nacido en mi tierra con el pincel extranjero.
Pintor que sigues el rumbo de tantos pintores viejos.
Aunque la virgen sea blanca, pintame angelitos negros,
que también se van al cielo todos los negritos buenos.
Pintor que pintas con amor
¿por qué desprecias su color, si sabes que en el cielo también los quiere Dios?
Pintor de santos de alcoba, si tienes alma en el cuerpo
¿por que al pintar en tus cuadros te olvidaste de los negros?
Siempre que pintas iglesias, pintas angelitos bellos, pero nunca te acordaste de pintar un
ángel negro.
(Antonio Machín)

Eskaintza

A Olatz: eres el faro que en la noche me ilumina, el puerto que me resguarda en la tormenta, el mar en el que me sumerjo.
Eres mi mar, mi dulce mar. Te amo.

Laiari: ene printzesa, hemendik ilargiraino eta kanarietaraino eta kanbritxeraino maite zaitut. Zurekin bizitako abenturetan zenbat hazi naizen!! Eta zenbat lagun egin ditugun: Pupupi, Wendy, Peter, Charlie, Matilda, Baccicin, Blancaflor, Mariquilla, antxumeak, Violeta, Olimpia, Ulises, sirenak...

Alairi: zure begi handiek nire bihotza laztantzen dute. Horren txikia izanik, nola da posible hainbeste maitasuna sortzea?

A mi ama: me lo diste todo, lo más grande, la vida.

A mi aita: me regalaste el mundo, llevándome de la mano, cantándome mil canciones.

Josune eta Ugori: betidanik elkarrekin, hurbil edo urrun, ederra da geure bidea konpartitzen jarraitea.

Nire ilobei: Aiora, Aroa, Nahia, Sofia, zoragarriak zarete.

Nire familiako beste guztiei: Atzio, Zoryana, Ainhoa, Ruben, M. Puy eta Pedro. A mi tía Marisa.

Lukari: gure etxekoen zaindari.

Koadrilako lagunei: Agus, Manzi, Gari, bakoitza bere bidean, aurrera, beti aurrera.

Biki eta Arani: Sanfran ta Unzaenen, milaka oroitzapen, zuekin elkarbizitza zoragarria izan zen.

Sustraiako kideei: Ainhoa, Beltza, Gorka, Larraitx, Maite, Miren, Marisol, Rosa, hain da ederra jakitea zuekin nire bihotza biluz dezakedala!

A Carlos: me mostraste los senderos de la ternura y me dejaste caminar.

Inmari: maitagarria zara.

Leireri: hor zaude beti, hemen nago ni.

A Luis y Eva: besitos p'al sur de vuestro amigo euskaldun.

A Corrado, Valeria, Loredana...: grazie a voi l'Italia é sempre nel mio cuore.

A Josemi, Vega y Jose: os quiero. Si pudiesemos acercar un poco en el mapa Donostia y Bilbao!

A Jose: qué bien me han venido esas llamadas, estilo sueco, para ir a currar a tu casa.

Davidi: zure masajeek tesiak gogortutako gorputz hau bigundu dute.

Tanari eta Artaunsoroko beste lore guztiei.

Marian eta Benjari: beti laguntzeko prest, zuen etxea milaka koloreetako ortzadarra da.

A Luis, Sonia, Cristina, Ramon y toda la gente que trabaja por humanizar los nacimientos.

Josetxori eta mundu hobeago batekin amets egiten duten guztiei.

Nire lagunak ere badiren Laiaren laguntxoei: Nora, Julia, Ane, Sara, Angela, Martin, Stella, Alaia, Santi...

Munduko ume guztiei, eta ume izan zirela ahaztu ez duten helduei.

Unibertso osoko matxinatuei.

Zoriontsuak izaten saiatzen diren guztiei.

Maite ditudan guztiei,
maite nauten guztiei.

Esker onak

Tesi honen bidai luze-luzean laguntza eman didaten guztiei nire eskerrik beroenak eman nahi dizkiet.

Ixakide guztioi, zuek guztiok egindako lanak izan baitira puzzle baten modura *Matxin* eraikitzeko erabili ditugun piezak.

Kepa eta Arantzari, tesi-lan hau zuzentzeagatik. Keparen umore eta animoei esker ikerketaren bide-beltzak bide-gorriak bihurtu dira.

Mikel, Gorkari eta Iñakiri, *Opentrad* proiektuan zuek egindako lanak *nirea* sentitzen nuen *Matxin* sistema *gurea* bihurtu duzuelako. Tesi honek ez luke erdia ere balioko zuen lana gabe. Gorka, mila esker zure pazientziagatik.

Ordenagailuen aurkako borrokan lagundu didazuenei: Kike, arazo tekniko guztientzako soroslea; Mikel Astiz, ixapazengaitk; Aitor Soroa, latexengatik; Aitor Sologaitza eta Oier, ixaemelengatik; Aitzol, David, Gregorio, Maite eta Nerea, edozein motatako zalantzak argitzen saiatzeagatik.

Nornorinork, baldintza eta subjuntibo y ole, eta guzti horiek ulertzen saiatzen lagundu didazuenei: Mikel, Maxux, Izaskun, Eli, Larraitx,...

A Lluís, Jordi y Eli, por los sudores para mejorar *Freeling*.

Aitziber eta Eneke, preposizioen itzulpenarako elkarrekin egindako lanarengatik.

Koldo eta Montse Hermori, erregelak eta gramatiken inguruan izandako solasaldiengatik.

Xabier Artolari, akronimo eta letra-zopa horietan ordena pixka bat jartzeagatik.

Saroiri, zure postedizio-lanekin kalkulaturako emaitzak opari ederrena izan direlako niretzat tesiaren azken txanpan.

Aitor Maritxalarri, IXA taldean sartu nintzenean egindako harrera eta emandako babesagatik.

Montseri, palan jolastu genituen partida ahaztezin haiengatik. Aupa Pippi!

Mikel Quintiaini, gure eztabaida filosofikoengatik.

Izaskuni, lankide eta lagun izateagatik.

Bertoli, zure istorioengatik. Egin kasu eta ez jarraitu denbora galtzen.

Maxuxi, biok aldi berean basamortua zeharkatzean, oasi bat izan direlako niretzat otorduetan zurekin partekatutako istorioak, parrak eta negarrak.

A Bienve, porque tu sonrisa es la mejor bienvenida cada mañana.

Iñakiri, tesiaren idazketa-garaian zugandik jasotako laguntza eta babesa ezinbestekoak izan zaizkidalako tesi hau bukatzeko. Zuk emandakoa ordaintezina da: zure begietako irriparra, zure arimako lasaitasuna... eta errespetua. Bihotz-bihotzez mila esker.

Gaien aurkibidea

Eskaintza	i
Esker onak	iii
Aurkibidea	v
Irudien zerrenda	ix
Taulen zerrenda	xiii
Akronimoak eta laburdurak	xv
I Sarrera	1
I.1 Aurkezpen orokorra	2
I.2 Ingeniaritza linguistikoa	3
I.3 Itzulpen automatikoa	4
I.4 Euskararen prozesamendurako estrategia	4
I.5 Tesi-lanaren egitura eta argitalpenak	6
II Itzulpen automatikoa, euskara eta software librea	11
II.1 Itzulpen automatikoa	12
II.1.1 Oinarrizko kontzeptuak	12
II.1.2 Historia	21
II.1.3 Erregeletan Oinarritutako IA	27
II.1.4 Corpuetan oinarritutako IA	33
II.1.5 Hibridazioa	40
II.1.6 Itzultzaileentzako tresnak	43
II.1.7 Gaur egungo erronkak	46
II.2 Itzulpen automatikoa eta euskara	48
II.2.1 Erregeletan oinarritutako sistemak	48
II.2.2 Corpuetan oinarritutako ikerketa-lanak	49

	II.2.3	Itzulpenerako laguntza-tresnak	51
	II.2.4	Sorkuntza automatikoa	52
II.3		Itzulpen automatikoa eta kode irekiko software librea . .	55
	II.3.1	Software librea vs. kode irekiko softwarea	55
	II.3.2	Kode irekiaren abantailak	57
	II.3.3	IAko kode irekiko sistema libreak	59
II.4		Laburpena	62
III		<i>Matxin</i> itzulpen-sistemaren teknologia	63
III.1		Sistemaren bilakaera	64
	III.1.1	Hasierako diseinua eta prototipoak	64
	III.1.2	<i>Matxin</i> teknologia OpenTrad proiektuan	68
	III.1.3	<i>Matxin</i> 1.0	69
III.2		Ezaugarriak	70
	III.2.1	Algoritmoak eta datuak banatzea	71
	III.2.2	Berrerabilgarritasuna	73
	III.2.3	Elkarreragingarritasuna eta estandarizazioa . .	73
	III.2.4	Kode irekia	74
III.3		Mugak	75
III.4		Arkitektura orokorra	76
III.5		Esaldiaren itzulpena prozesatzeko datu-egitura	79
	III.5.1	Egitura sintaktikoak: osagaiak vs. mendekotasu- nak	80
	III.5.2	Esaldia, chunka eta nodoa	82
	III.5.3	Datu-egituraren formatua	84
III.6		Datuen formatua	84
	III.6.1	Hiztegiak	84
	III.6.2	Preposizioen hiztegia eta hautapen-erregelak . .	88
	III.6.3	Corpusetik erauzitako informazioa	89
	III.6.4	Datu-egitura manipulatzeko erregelak	89
	III.6.5	Aditz-kateen transferentziarako formalismoa . .	95
III.7		Itzulpen-prozesua: adibide bat	96
III.8		Laburpena	97
IV		Analisia	103
IV.1		Analisiaren sakonera	104
	IV.1.1	Analisi morfologikoa	104
	IV.1.2	Analisi sintaktikoa	105
		IV.1.2.1 Analisi sintaktiko partziala	105
		IV.1.2.2 Analisi sintaktiko osoa	105

IV.2	<i>FreeLing</i>	106
IV.3	Mendekotasun-analizatzailea	108
IV.3.1	Oinarrizko mendekotasun-analizatzailearen erai- kuntza	109
IV.3.2	<i>Txala</i> , <i>FreeLing</i> eko mendekotasun-analizatzailea	112
IV.3.3	Mendekotasun-analizatzailearen mugak	114
IV.4	Laburpena	116
V	Transferentzia	119
V.1	Transferentzia lexikala	120
V.1.1	Transferentzia lexikalaren modulua	122
V.1.2	Lexikoi elebiduna	127
V.2	Transferentzia estrukturala	135
V.2.1	Espainiera eta euskararen arteko dibergentziak .	135
V.2.2	Transferentzia estrukturalaren moduluak	140
V.2.3	Preposizio eta funtzio sintaktikoen transferen- tziaren deskribapena	153
V.2.4	Aditz-kateen transferentziaren deskribapena . .	164
V.3	Laburpena	168
VI	Sorkuntza	169
VI.1	Sorkuntza sintaktikoa	170
VI.1.1	Hitz-hurrenkera euskaran	171
VI.1.2	Chunk barruko sorkuntza sintaktikoa	172
VI.1.3	Chunken arteko sorkuntza sintaktikoa	176
VI.2	Sorkuntza morfologikoa	181
VI.2.1	Moduluaren deskribapena	181
VI.2.2	Euskararako prozesadore morfologikoa	182
VI.2.3	Euskararako prozesadore morfologikoaren egoki- tzapena	186
VI.3	Laburpena	187
VII	Baliabide linguistikoen berrerabilpena	189
VII.1	Lexikoi elebidunaren eraikuntza	189
VII.1.1	Helburuak eta estrategia	191
VII.1.2	Berrerabilitako baliabideak	193
VII.1.3	Hiztegien egokitzapena	197
VII.1.4	Hiztegiak beste baliabideekin bateragarri bihur- tzea	200
VII.1.5	Lexikoiaren osaketa	205

VII.1.6	Lexikoia XML formatura bihurtzea	209
VII.2	Ezaugarri semantikoen etiketatze erdi-automatikoa . . .	209
VII.2.1	Etiketatzeta hiztegiko informazioa erabiliz	211
VII.3	Laburpena	222
VIII	Ebaluazioa	223
VIII.1	Itzulpen automatikoaren ebaluaziorako neurriak	223
VIII.1.1	Giza-ebaluaziorako neurriak	224
VIII.1.2	Neurri automatikoak	227
VIII.1.3	Neurri automatikoen argiak eta itzalak	230
VIII.1.4	Edizio-distantzia	233
VIII.2	<i>Matxin</i> sistemaren ebaluazioa	235
VIII.2.1	Ebaluazioaren diseinua	235
VIII.2.2	<i>Matxin</i> sistemaren emaitzak	240
VIII.2.3	Bi sistemen konparazioa: <i>Matxin</i> vs <i>Matrex</i> . . .	241
VIII.2.4	<i>Bleu</i> neurriaren baliagarritasunari buruzko haus- narketa	243
VIII.3	Erroreen analisisa	246
VIII.4	Laburpena	249
IX	Ondorioak eta etorkizunerako lanak	251
IX.1	Tesi-lan honen ekarpenak	251
IX.2	Ondorioak	253
IX.3	Etorkizunerako lanak	254
	Bibliografia	256
	Eranskinak	276
A	Itzulpen-adibideak	277
B	Parole kategoria-sistema	289
C	EDBL kategoria-sistema	297

Irudien zerrenda

I.1	Baliabideak, tresnak eta aplikazioak	5
II.1	Gizakiak lagundutako itzulpen automatikoa	14
II.2	Itzulpen automatikoaren ikerketaren gora-beherak	21
II.3	Vauquois-en triangelua	28
II.4	Itzulpen zuzeneko metodoen eredua	29
II.5	Interlinguan oinarritutako eredua	30
II.6	Transferentzian oinarritutako eredua	30
II.7	<i>Ariane</i> sistemaren arkitektura	33
II.8	Ordenagailuek lagundutako giza-itzulpena	44
II.9	<i>MultiMeteo</i> sistemaren emaitza euskaraz	54
II.10	<i>Apertium</i> sistemaren arkitektura	61
III.1	Arkitektura orokorraren eskema	77
III.2	Osagai- eta mendekotasun-egiturak. Adibidea	81
III.3	<i>Matxin</i> sistemako zuhaitz-egitura. Adibidea.	81
III.4	Gure sistemako egituratik mendekotasun-egiturara bihurketa	82
III.5	Mendekotasun-zuhaitza: esaldia, chunkak eta nodoak	83
III.6	Moduluen arteko datu-egituraren XML formatuaren DTDa .	85
III.7	Datu-egitura manipulatzeko erregelen XML formatuaren DT- Da	90
III.8	Datu-egitura manipulatzeko erregelen adibideak	91
III.9	Aditz-kateen transferentziarako gramatika. Adibideak	96
III.10	Analisi fasearen irteera	98
III.11	Transferentzia lexikalaren irteera	99
III.12	Transferentzia estrukturalaren irteera	100
III.13	Sorkuntza fasearen irteera	101
IV.1	Vauquois-en triangelua analisi posibleetarako	106
IV.2	Chunkak konbinatzeko eragiketa motak	111
IV.3	Mendekotasun-zuhaitzaren eraikuntza. Adibidea.	113

V.1	Transferentzia moduluen arkitektura	120
V.2	Transferentzia lexikala: moduluaren arkitektura	122
V.3	Lexikoian kontsultarako erregelak	123
V.4	Transferentzia lexikaleko beste eragiketarako erregelak . . .	125
V.5	Transferentzia lexikala: sarrera eta irteera.	126
V.6	Sekzio nagusiaren formatua. Adibideak	130
V.7	Paradigmak. Adibideak	131
V.8	Kategoria itxien sekzioaren formatua. Adibideak	133
V.9	Puntuazio ikurren sekzioaren formatua. Adibideak	135
V.10	Transferentzia estrukturalaren arkitektura	141
V.11	Intrachunk mugimenduetarako erregelak	143
V.12	Nodoen ezabaketarako erregela	144
V.13	Intrachunk transferentzia estrukturala. Adibidea	145
V.14	Preposizio eta funtzio sintaktikoen transferentziarako erregelak	146
V.15	Preposizio eta funtzio sintaktikoen transferentzia. Adibidea .	146
V.16	Pertsona-informazio osatzeko erregelak	147
V.17	Interchunk mugimenduetarako erregelak	148
V.18	Aditz-chunkeko pertsona-informazioen osaketa	149
V.19	Chunken ezabaketarako erregela	149
V.20	Interchunk eragiketak. Adibidea	150
V.21	Aditz-kateen transferentziarako erregelak	151
V.22	Aditz-kateen transferentzia. Adibidea	152
V.23	Egokitzen eragitetarako erregelak	152
V.24	Preposizioen hiztegia eta hautapen-erregelen formatua	158
V.25	Azpikategorizazio-patroien formatua	161
V.26	Azpikategorizazio-patroien erabilera: 1. adibidea	162
VI.1	Sorkuntza: arkitektura	170
VI.2	Sorkuntza sintaktikoa intrachunk. Adibidea.	173
VI.3	Adjektiboen koordinazioaren prozesaketarako erregelak . . .	174
VI.4	Kokapen-atributuaren (<i>loc</i>) osaketarako erregelak	175
VI.5	Aurrekotasun-erregelak	177
VI.6	Postposizioaren informazioa chunketik azken nodora pasatze- ko erregela	177
VI.7	Sorkuntza sintaktikoa interchunk. Adibidea.	179
VI.8	Ordena erlatiboa ebazteko erregelak	180
VI.9	<i>lmi</i> atributuaren osaketarako erregela	182
VI.10	Sorkuntza morfologikorako antolaketa (I)	183
VI.11	Sorkuntza morfologikorako antolaketa (II)	184
VI.12	Sortzaile morfologikoari egiten zaion deia	184

VI.13	Hiztegi morfologikoaren formatuaren adibidea	188
VII.1	Hiztegien prozesaketa	194
VII.2	<i>Elhuyar</i> hiztegiaren hiru fitxategietako adibideak	195
VII.3	<i>Euskalterm</i> bankuko zenbait sarrera	196
VII.4	Entitateen zerrenda elebiduneko zenbait sarrera	197
VII.5	Hiztegi egokitua. Adibideak	199
VII.6	Hitz anitzeko sarrerak markatuta	204
VII.7	Lehenengo ordainaren hautaketa. Adibideak	208
VII.8	<i>Elhuyar</i> hiztegiko kategoria itxien zenbait adibide	209
VII.9	Etiketatzeko automatikoa eta eskuzko etiketatzea	219
VII.10	Etiketatzeko automatikoa eta iterazio kopurua	220
VIII.1	Kalitate-eredu bateko atributuak	225
VIII.2	Edizio-distantziaren kalkuluaren traza.	235
VIII.3	Postediziorako testu-fitxategia. Adibidea	237
A.1	Interneteko web orrien itzulpena. Egunekeo berriak.	284
A.2	Interneteko web orrien itzulpena. Salmentak.	285

Taulen zerrenda

II.1	EOIA, AOIA eta IAE paradigmnen oinarritzko prozesuak . . .	40
III.1	Kokapen-bideen adibideak	92
III.2	Ekintzen adibideak	94
IV.1	Analisi partzialaren emaitza	109
V.1	Anbigutasun kontzeptualaren adibideak	121
V.2	Chunk moten baliokidetzak	127
V.3	Sekzio nagusia. Adibideak	129
V.4	Paradigmak. Adibideak	131
V.5	Kategoria itxiak. Adibideak	132
V.6	Puntuazio ikurrak. Adibideak	134
V.7	Preposizio hiztegia eta hautapen-erregelak	157
V.8	Preposizio hiztegiaren datuak	159
V.9	Azpikategorizazio-patroiak	162
VII.1	Sarrerren fitxategiaren kopuruak kategoriaka	196
VII.2	Sarrerren egokitzapena	198
VII.3	Ordainen garbiketa	199
VII.4	Kategoria-sistemen baliokidetzak	201
VII.5	Sarrera sinpleen kategorietan desbiderapenak	203
VII.6	Hitz anitzeko sarrerren kategoriaren esleipen erregelak	203
VII.7	Moldatu beharreko ordainak	206
VII.8	Hiztegi semantikoa	207
VII.9	Eskuzko etiketatzea [$\pm$ bizidun] ezaugarriarentzat	216
VII.10	Genus eta erlatureen informazioa erabiliz etiketatzearen 1. iterazioa	217
VII.11	Sinonimiaren informazioa erabiliz etiketatzearen 1. iterazioko zenbait adibide	218
VII.12	Genus eta erlatureen informazioa erabiliz etiketatzearen 2. iterazioko zenbait adibide	218

VII.13	Etiketatzeko automatikoa eta iterazio kopurua	219
VII.14	Etiketatzeko automatikoaren doitasuna	220
VII.15	Etiketatzeko automatikoaren estaldura	221
VII.16	Etiketatzeko automatikoaren estaldura corpusean	221
VII.17	Beste ezaugarri semantiko batzuen etiketatzeko datuak . . .	221
VIII.1	Ebaluaziorako erabilitako jatorrizko testuen datuak	236
VIII.2	Hitzetan oinarritutako edizio-distantzia. Adibideak	240
VIII.3	Hitz-erro eta atzizkietan oinarritutako edizio-distantzia. Adibideak	240
VIII.4	<i>Matxin</i> sistemaren emaitzak	241
VIII.5	<i>Matrex</i> sistemaren emaitzak	242
VIII.6	<i>Matxin</i> vs <i>Matrex</i>	243
VIII.7	<i>Bleuren</i> emaitzak	244
VIII.8	Mendekotasun-analizatzailearen erroreak. Adibideak	247
VIII.9	Tagger-aren erroreak. Adibideak	247
VIII.10	Funtzio sintaktikoen esleipenean erroreak. Adibideak	248
VIII.11	Hautapen lexikalean erroreak. Adibideak	248
VIII.12	Preposizio eta funtzio sintaktikoen itzulpenen erroreak. Adibideak	248
VIII.13	Hitz anitzeko terminoekin erroreak. Adibideak	248
A.1	Laborategiko adibideak	278
A.2	Itzulpen-adibideak. <i>Consumer</i> corpusa	279
A.3	Itzulpen-adibideak. Administrazio-testua	280
A.4	Itzulpen-adibideak. Legeak	281
A.5	Itzulpen-adibideak. Eskuliburua.	282
A.6	Itzulpen-adibideak. Chat.	283
A.7	Itzulpen-adibideak. Kanta bat	286
A.8	Itzulpen-adibideak. Literatura	287

Akronimoak eta laburdurak

Euskaraz:

AA:	Adimen Artifiziala
AEB:	Ameriketako Estatu Batuak
AH:	Abiapuntu-Hizkuntza
AOIA:	Adibideetan Oinarritutako Itzulpen Automatikoa
COIA:	Corpusetan Oinarritutako Itzulpen Automatikoa
EB:	Europar Batasunak
EDBL:	Euskararen Datu Base Lexikala
EOIA:	Erregeletan Oinarritutako Itzulpen Automatikoa
HAUL:	Hitz Anitzeko Unitate Lexikala
IA:	Itzulpen Automatikoa
IAE:	Itzulpen Automatiko Estatistikoa
IL:	Ingeniaritza Linguistikoa
IM:	Itzulpen-Memoria
LNP:	Lengoaia Naturalaren Prozesamendua
PS:	Preposizio-Sintagma
SM:	Sorkuntza Morfologikoa
SS:	Sorkuntza Sintaktikoa
TE:	Transferentzia Estrukturala
TL:	Transferentzia Lexikala
SESB:	Sobietar Errepublika Sozialisten Batasuna
XH:	Xede-Hizkuntza

Erdaraz:

BLEU:	BiLingual Evaluation Understudy
BSD:	Berkely Software Distribution
CBMT:	Corpus Based Machine Translation
EBMT:	Example Based Machine Translation

FAHQT:	Fully Automatic High Quality Translation
FOSS:	Free Open Source Software
FSA:	Finite State Automata
FST:	Finite State Transducer
GNU:	GNU's Not Unix
GPL:	General Public License
HAMT:	Human-Aided Machine Translation
HTER:	Human-targeted Translation Edit Rate
IE:	Information Extraction
IR:	Information Retrieval
KBMT:	Knowledge Based Machine Translation
LGPL:	Lesser General Public License
MAHT:	Machine-Aided Human Translation
MRD:	Machine Readable Dictionary
MT:	Machine Translation
MTD:	Machine Tractable Dictionary
UNL:	Universal Networking Language
POS:	Part-of-Speech
PP:	Prepositional Phrase
RBMT:	Rule Based Machine Translation
SGML:	Standard Generalized Markup Language
SMT:	Statistical Machine Translation
SOV:	Subject-Object-Verb
SVO:	Subject-Verb-Object
WER:	Word Error Rate
PER:	Position-independent Error Rate
TER:	Translation Edit Rate
XFST:	Xerox Finite-State Tool
XML:	Extensible Markup Language
XSL:	Extensible Stylesheet Language
XSLT:	XSL Transformations

Erakundeak:

ALPAC:	Automatic Language Processing Advisory Committee
DARPA:	Defense Advanced Research Projects Agency
EHU:	Euskal Herriko Unibertsitatea
EITB:	Euskal Irrati Telebista
FSF:	Free Software Foundation
GALE:	Global Autonomous Language Exploitation
GETA:	Groupe d'Etudes pour la Traduction Automatique

IAMT:	Internatinal Association for Machine Translation
INM:	Instituto Nacional de Meteorología
ISLE:	International Standards for Language Engineering
MIT:	Massachusetts Institute of Technology
NCLT:	National Centre for Language Technology
NIST:	National Institute of Standards and Technology
OSI:	Open Source Initiative
PAHO:	Pan American Health Organization
PRHLT:	Pattern Recognition and Human Language Technology
PRST:	Pattern Recognition and Speech Technology
TALP:	Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla
TAUM:	Traduction Automatique de l'Université de Montréal
UPC:	Universitat Politècnica de Catalunya
UPV:	Universidad Politécnica de Valencia
W3C:	World Wide Web Consortium

I. KAPITULUA

Sarrera

*Euskara, jalgi hadi plazara.
Lengoajetan ohi hintzan estimatze gutitan;
orai aldiz hik behar duk ohorea orotan.
Euskara, jalgi hadi mundura.
Kontrapas. Bernat Etxepare, 1545*

*-Altxa dezatela eskua euskara maite dutenak!
Kostatu zaie mugitzen hasia, baina, azkenerako, den-denek jaso dute eskua,
lehen egunean behintzat irakaslea pozik edukitzea komeni delakoan edo.
Irakasleak, ordea, errietan hasi zaie:
-Euskara maite izatea, hori bai txorakeria!
Euskara ez da maitatu behar, euskara erabili egin behar da, eta kito!
Kurrin-kurrin vitae. Joxerra Gartzia, 2004.*

Euskara erabili. Euskaraz hitz egin. Euskaraz irakurri eta idatzi. Jolastu, lan egin, amestu. Euskaraz gozatu, maitatu. Euskaraz bizi.

Unesco erakundeak argitaratutako galtzeko arriskuan dauden munduko hizkuntzen atasean (Wurm, 2001) aitortzen da Hego Euskal Herrian euskara galzorian dagoela, eta larriki galzorian Ipar Euskal Herrian.

Horregatik euskarak bizirautea nahi badugu, euskara erabili behar du-gu, euskarak plazara jalgi behar du. Eta horrek, gaur egun, informazioaren gizartearen plazara jalgi behar duela esan nahi du. Informazioaren aro honetan euskarak, beste hizkuntzen parean, nazioarteko komunikaziorako eta ulermenerako tresna izan behar du.

Hizkuntza desberdinen arteko harreman horretan, gaur egungo mundu globalizatuaren Babelgo dorrean, itzulpen automatikoak (IA) garrantzi handia hartzen du eta, zenbaitetan, nahitaezko bitartekaria bihurtuko da. Eus-

karari dagokionez, itzulpen automatikoa beharrezkoa izango da, alde batetik, euskaldunek erdal hizkuntzetako idatziak ulertzeko, eta bestetik, euskarazko produkzioa ez-euskaldunenagana iritsi ahal izateko.

Gainera, Euskal Herrian bizi ditugun errealitate elebidunetan, itzulpenen eskaera handiari aurre egiteko, itzultzaileentzako laguntza-tresnak oinarritzko lanabes izango dira, ekoizpena areagotzeko eta kostuak gutxitzeko.

1.1 Aurkezpen orokorra

Tesi-lan honetan *Matxin* aurkezten dugu, euskararekin lan egiten duen lehenengo itzulpen automatikoko sistema. Edo, umilagoak izanik, euskararekin lan egiten duen IAKo sistema baten eraikuntzaren lehenengo urratsak, burututako lana hasiera baino ez delako.

Ingeniaritza linguistikoko ikerketa-lan hau IXA¹ taldearen barruan garatua izan da, talde honek euskararen prozesamendurako diseinatutako estrategiaren barruan.

Garatu dugun IAKo sistema erregeletan oinarritutakoa da, transferentzia sintaktiko sakona egiten duena. Sistemaren diseinua modularra da, algoritmoak eta datuak banatuz. Hainbat tresna eta baliabide linguistiko berrerabili ditugunez, sistemako datuen, erregelen eta moduluen arteko datu-fluxuaren formatuen estandarizazioa beharrezkoa izan da, elkarreragingarritasuna bermatzeko.

Itzulpen-prozesua hiru fasetan banatzen da: analisia, transferentzia eta sorkuntza. Analisi-faserako analisi sintaktiko partziala ematen duen pakete bat berrerabili eta analisi osoa emateko hedatu dugu. Transferentzia- eta sorkuntza-faseak, berriz, baliabide linguistikoak eta modulu batzuk berrera-biltzen badituzte ere, *Matxin* sistemarentzat diseinatu ditugu.

Transferentzia bi urratsetan egiten da: transferentzia lexikala eta transferentzia estrukturala. Transferentzia lexikalean estaldura handiko hainbat hiztegi berrerabiliz eraiki dugun lexikoi elebiduna kontsultatzen da. Transferentzia estrukturalean, besteak beste, postposizioen desanbiguazioa eta aditz-kateen transferentzia burutzen da.

Sorkuntza ere bitan banatzen da: sintaktikoa eta morfologikoa. Sorkuntza sintaktikoan hitzak eta sintagmak ordenatzen dira eta izen-sintagmetan azken hitzari pasatzen zaio postposizio-informazioa. Sorkuntza morfologikoak hitzen formak lortzen ditu, euskararako prozesadore morfologikoa berrerabili dugula.

¹<http://ixa.si.ehu.es>

Sistemaren arkitektura abiapuntu- eta xede-hizkuntzetatik independentea izateko diseinatu izan bada ere, gaur egun inplementatutako *Matxin 1.0* prototipoak espainieratik euskarara itzultzen du, zenbait lan burutu ditugula ingelesetik euskarara itzultzeko.

Matxin 1.0 prototipoa erabilgarri dago Interneten², kode irekiko software libre bezala banatzen delarik³.

I.2 Ingeniaritza linguistikoa

Lengoaia naturalaren prozesamenduaren (LNP) helburua hizkuntzaren azterketa da, konputagailuak erabiliz teoria linguistikoen zatiak modelatzeko. Ingeniaritza linguistikoak (IL), 90eko hamarkadan jaio zena, LNPko arazo berak aztertzen ditu, baina ikuspuntu praktikoa hartuz. Ingeniaritzak arreta jartzen du ezagutza zientifikoaren erabilera praktikon, eta bere helburua irizpide batzuen arabera taxuz jokatu duten tresnak garatzea da (Cunningham, 1999). ILaren arloan burutzen den ikerketan edozein teoria linguistikoa onar daiteke, emaitza hobeagoetara badarama.

Cunningham-ek (1999) horrela definitzen du ILa:

Language Engineering is the discipline or act of engineering software systems that perform tasks involving processing human language. Both the construction process and its outputs are measurable and predictable.

LNPko hasierako jostailuzko sistemak gainditzeko asmoarekin sortu zen ILaren arloa, industriak mundu errealeko testuen gainean lan egiteko gai ziren tresna eraginkorrak behar zituelako. Baliabide lexikal zabalak berreskuratzea ekarri zuen testu errealak prozesatzeko gai ziren sistemak garatu ahal izatea.

Zenbait faktore indartu dute ILaren arloa (Cunningham, 1999):

- Konputagailuetako hardwarean izan den aurrerabideak aukera ematen du azkarragoak diren prozesadoreak eta memoria handiagoa kudeatzeko, konputazio-zama handiko prozesuak garatzea bideragarria dela.
- Neurri erreala duten eta Internet saretik eskuragarriak diren iturri linguistikoen kopuruaren handitzea: hiztegiak, corpusak...
- LNPko teknologiaren heldutasunak aukera ematen du doitasun handiko aplikazioak sortzeko.

²<http://www.opentrad.org/demo/>

³<http://matxin.sourceforge.net/>

- Software-ingeniaritzaren aurrerapenak: osagaietan oinarritutako softwarea, diseinu-patroien erabilpena...

Tesi-lan hau, beraz, ingeniartza linguistikoaren arloan kokatzen da. Ondoko irizpide metodologiko hauek jarraitu ditugu lanaren bideragarritasuna eta kalitatea ziurtatzearren:

- Sendotasuna. Sistemak funtzionatu behar du, ez soilik laborategiko baldintzetan, baizik eta egoera errealetan, lan-karga handiekin eta sarrera okerrekin ere.
- Eraginkortasuna. Sistemak denbora onargarrian bete behar ditu bere eginkizunak.
- Berrerabilgarritasuna. Sistema eraikitzeke saiakera handia egin dugu, ahalik eta gehien aurretik garatutako baliabideak eta tresnak berrerabiltzeko. Bestalde, gure asmoa da tesi-lan honetan garatutako moduluak aurrerantzean sistema konplexuagoetan txertatu ahal izatea

1.3 Itzulpen automatikoa

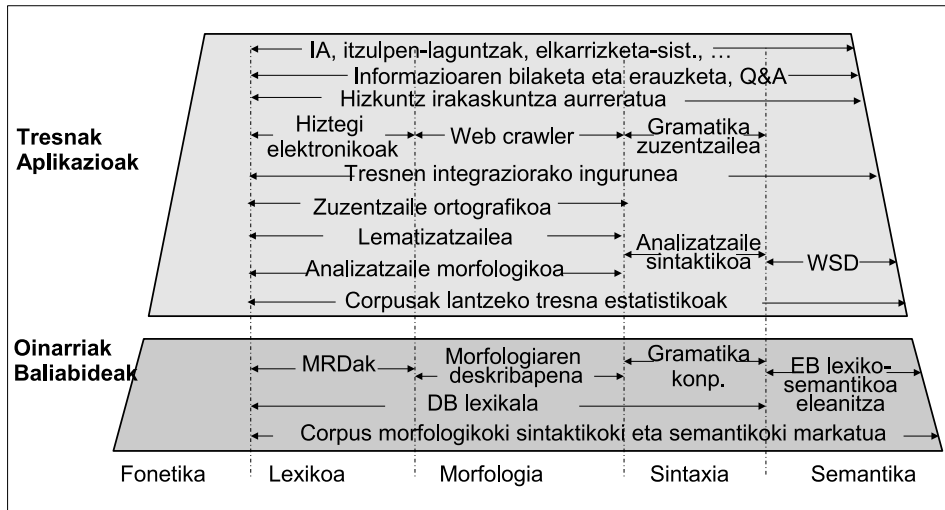
Itzulpen automatikoa (IA; *Machine Translation*, *MT*) sistema informatikoak erabiliz hizkuntza batetik beste batera itzulpena burutzen duen prozesuari esaten zaio.

Itzulpen automatikoa ingeniartza linguistikoaren arloan kokatzen da. Ez da zientzia teoriko bat, baizik eta metodo linguistiko eta konputazionalen aplikazio praktikoa. IAKo sistemek eredu linguistiko desberdinak integratzen dituzte, eredu horiek sistemaren beharretara eta eskakizun konputazionalak egokituz.

...the history of MT shows, to me at least, the truth of two (barely compatible) principles that could be put crudely as 'Virtually any theory, no matter how silly, can be the basis of some effective MT' and 'Successful MT systems rarely work with the theory they claim to. (Wilks, 1989)

1.4 Euskararen prozesamendurako estrategia

Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldeak LNPa-ren gaineko ikerketa-lana du helburu, euskararen gaineko ikerketa aplikatua



Irudia I.1: Baliabideak, tresnak eta aplikazioak

xede nagusia delarik. IXA taldea 1987an sortu zen eta taldearen lehenengo ideia IAKo sistema bat sortzea izan zen. Baina atariko azterketa bat egin ondoren, taldeak urteetan jarraitu duen estrategia hartzea erabaki zuen (Sarasola, 2000), produktuen eraikuntzan ondoko ordenari jarraituz (ikus I.1 irudia)

1. Oinarri eta baliabide linguistikoak. Hizkuntzaren tratamendu automatikorako ezinbestekoa den azpiegitura. Baliabide lexikalak eta corpusen bildumak dira lehen urratsak. Tresnak eta aplikazioak eraikitzeke oinarria izango dira.
2. Tresnak. Adituentzat eta programa-garatzailentzat. Produktu berriak garatzeko balio dute.
3. Aplikazioak. Informatikan eta linguistikan ezagutzarik suposatzen ez zaien erabiltzaile arruntentzako produktuak.

Oinarri eta baliabide linguistikoak eta ikerketa-lanak nahitaezkoak dira tresnak eta aplikazioak garatzeko baina, era berean, azken hauek ere neurri handi batean ikerketa-lana eta oinarri berrien eraikuntza erraztuko dute. Hiru produktu mota horiek era koordinatu eta osagarrian heldu behar dira, ahalik eta etekin handiena ateratzeko. Diseinatutako tresnak estaldura handikoak eta orokorrak izatea garrantzitsua da oso, haien gainean tresna aurreratuagoak garatu nahi badira.

Bestalde LNPan bost ikerketa-arlo bereiz daitezke, erabilitako ezagutza linguistikoaren motaren arabera: fonetika, lexikoa, morfologia, sintaxia eta semantika. IXA taldearen lana testuen tratamenduan kokaten da, ahotsaren prozesaketan lanik apenas ez delarik egin.

20 urteetako esperientziak, tartean zenbait aplikazio komertzial garatuz, taldeak hartutako estrategiaren zuzentasuna frogatua uzten du. Eta, orain bai, sortutako oinarritzko baliabide eta tresnei esker, itzulpen automatikoko aplikazio bati aurre egitea posible da.

Hori da, hain zuzen, tesi-lan honetan aurkezten duguna: euskararekin lan egiten duen itzulpen automatikoko sistema baten eraikuntzaren lehenengo urratsak. Euskararekin lan egiten duen *lehenengo* IAko sistema.

1.5 Tesi-lanaren egitura eta argitalpenak

Ondokoa da tesi-lan honen egitura:

- I. kapitulua. Sarrera
- II. kapitulua. Itzulpen automatikoa, euskara eta software librea.
Kapitulu honetan IAren inguruko sarrera bat aurkezten dugu: oinarritzko kontzeptuak, historia, IArako estrategiak, gaur egungo ikerketa-leerroak eta etorkizuneko erronkak. Euskararekin IAren arloan egindako ikerketa-lan urriak aztertuko ditugu, baita azken urteetan IAren aplikazioetan hainbesteko garrantzia hartzen ari den software librearen garapena.
- III. kapitulua. *Matxin* itzulpen-sistemaren teknologia.
Hemen azaltzen ditugu eraiki dugun sistemaren ezaugarriak eta mugak, transferentzian oinarritutako arkitektura orokorra, sistemak prozesatzen dituen datuen egitura eta haietan oinarritutako moduluen arteko datu-fluxuen formatua, eta baliabide linguistikoen formatua, hiztegiarena zein erregelena. Garrantzi handia eman diogu formatu estandarrei jarraitzeari.
- IV. kapitulua. Analisia.
Gure sistemaren fase honetarako erabiltzen dugu *FreeLing* analizatzailerik aurkezten dugu, espainiera-euskara itzulpenerako beharrezkoa den mendekotasun-egitura lortzeko analizatzaile horri egin diogun hedapena azalduz.

- V. kapitulua. Transferentzia.
Abiapuntu- eta xede-hizkuntzen artean dauden dibergentziei aurre egiteko beharrezkoak diren moduluak aurkezten ditugu. Eginkizun hau bi mailatan burutzen da, transferentzia lexikala eta estrukturala bereiziz. Zailtasun handienak aditz-kateen eta preposizio eta funtzio sintaktikoen transferentzia estrukturalan ematen dira. Arazo horiek ebazteko egindako hausnarketak eta konponbideak sakonki azaltzen ditugu.
- VI. kapitulua. Sorkuntza.
Hemen euskararen sorkuntzarako modulua aurkezten dugu. Alde batetik sorkuntza sintaktikoan hitzen eta osagaien ordenatzea, eta bestetik postposiziodun hitzen sorkuntza morfologikoa nola burutzen dugun aurkezten dugu. Horretarako IXA taldean garatutako prozesadore morfologikoa gure sisteman berrerabiltzeko nola prozesatu dugun azaltzen dugu..
- VII. kapitulua. Baliabide linguistikoen berrerabilpena.
Nahitaezkoa izan da *Matxin* sistema eraikitzeko, aurretik garatutako baliabide lexikalak berrerabiltzea. Kapitulu honetan azalduko dugu ze lanak egin behar izan ditugun *Elhuyar* hiztegitik eta *Euskalterm* banku terminologikotik sistemaren lexikoa sortzeko eta nola sortu dugun euskarazko ezaugarri semantikoen hiztegia modu erdi automatiko bidez *Euskal Hiztegiaren* definizioak erabiliz.
- VIII. kapitulua. Sistemaren ebaluazioa.
IAko sistemak ebaluatzeko neurriak aztertu ondoren, gure sistema ebaluatzeko erabiliko dugun *edizio-distantzia* deskribatzen dugu. Burutu dugun ebaluazioaren detaileak azaldu ondoren, lortutako datuak argitaratzen ditugu, errorearen analisia eginez.
- IX. kapitulua. Ondorioak eta etorkizunerako lanak. Kapitulu honek tesi-lan honetako ekarpenak laburtzen ditu, gure lanaren ondorioak azpimarratuz, eta etorkizuneko ikerketa-bideak azalduz.

Tesi-lan honekin lotutako argitalpenen berri ematen dugu ondoko lerro-tan, argitalpen bakoitza dagokion kapituluarekin lotuz. Argitalpen guztiak IXA taldeko web orritik jaso daitezke⁴.

- II. kapitulua:

⁴<http://ixa.si.ehu.es/Ixa/Argitalpenak>

- Díaz de Ilarraza A., Mayor A., Sarasola K.
Adibideetan Oinarritutako Itzulpen Automatikoa (AOIA): azterketa bibliografikoa eta euskararekin lan egiteko proposamenak.
Senez aldizkaria. 25.zbkia. ISBN: 847086287-1, 17-35 orr. 2002.
- Díaz de Ilarraza A., Mayor A., Sarasola K., Loinaz M., Chevreau K. , Coch J.
MultiMateok euskaraz ere badaki.
Elhuyar Zientzia eta Teknologia. 170. zbkia. 42-49 orr. 2001.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K.
Itzulpen automatikoa IXA taldean.
Ordenagailuz Lagundutako itzulpena. EHuko Udako Ikastaroak.
 Donostia, Euskal Herria. 2005.
- Corbí-Bellot A.M. , Forcada M.L. , Ortiz-Rojas S, Pérez-Ortiz J.A., Ramírez-Sánchez G., Sánchez-Martínez F., Alegria I., Mayor A., Sarasola K.
An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain
Proceedings of the EAMT2005. Poster session. Budapest, Hungary. 2005.
- III. kapituluua:
 - Díaz de Ilarraza A., Mayor A., Sarasola K.
Reusability of wide-coverage linguistic resources in the construction of an English-Basque Machine Translation System
IAI Working Paper No. 36. Hybrid Approaches to Machine Translation. Institut of Applied Information Sciences, Saarbruecken. 1999. <http://www.iai.uni-sb.de/iaien/iaiwip>
 - Díaz de Ilarraza A., Mayor A., Sarasola K.
Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System.
Proceedings of the MT 2000. University of Exeter, United Kingdom. 2000.

- Díaz de Ilarraza A., Mayor A., Sarasola K.
Inclusión del par castellano-euskara en un prototipo de traducción automática multilingüe.
Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies. Jaén, Spain. 2001.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K., Forcada M., Ortiz S., Padró L.
An Open Architecture for Transfer-based Machine Translation between Spanish and Basque.
Proceedings of the MT Summit X. Workshop: OSMaTran, Open-Source Machine Translation. Phuket, Thailand. 2005.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K.
Transfer-based MT from Spanish into Basque: reusability, standardization and open source.
Proceedings of the Cicling 2007. LNCS 4394. Mexico. 2007.
- IV. kapitulua:
 - Atserias J., Comelles E., Mayor A.
TXALA un analizador libre de dependencias para el castellano
Actas del XXI Congreso de la SEPLN XXI. Demo session. Granada, Spain. 2005.
- V. kapitulua:
 - Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K.
An FST grammar for verb chain transfer in a Spanish-Basque MT System.
Proceedings of the FSMNLP Finite-State Methods and Natural Language Processing. Poster session. Helsinki. 2005.
- VII. kapitulua:
 - Díaz de Ilarraza A., Mayor A., Sarasola K.
Building a Lexicon for an English-Basque Machine translation

System from Heteogeneous Wide-Coverage dictionaries.
Proceedings of the MT 2000. University of Exeter, United Kingdom. 2000.

- Díaz de Ilarraza A., Lersundi M., Mayor A., Sarasola K.
Etiquetado semiautomático del rasgo semántico de animicidad para su uso en un sistema de traducción automática.
Actas del XVI Congreso de la SEPLN. Vigo. 2000.
- Díaz de Ilarraza A., Lersundi M., Mayor A., Sarasola K.
Semiautomatic labelling of semantic features.
Proceedings of the 19th International Conference on Computational Linguistics, Coling. Taipei, Taiwan. 2002.

II. KAPITULUA

Itzulpen automatikoa, euskara eta software librea

Gero, esan zuten: “Ea, eraiki ditzagun hiri bat eta zerurainoko dorre bat. Ospetsu egingo gara eta ez gara munduan zehar barreiatuko”. Jaitsi zen Jauna gizasemeak eraikitzen ari ziren hiria eta dorrea ikustera, eta esan zuen: “Hara, guztiak herri bat bera dira, hizkuntza bat berekoak. Hau beren lehenengo lana badute, ez zaie aurrerantzean egin nahiko duten ezer ezinezko gertatuko. Jaitsi eta nahas diezaiegun, bada, hizkuntza, elkar uler ez dezaten”.

Biblia. Hasiera 11

*Yo soy del sur, tú eres del norte,
no hablamos el mismo idioma,
pero haremos que no importe.*

Bebe

Itzulpen automatikoak (IA) historia luzea du dagoeneko, konputagailue-na baino luzeagoa ere. Hala ere, arlo honetan, euskararen bidea hasi berria da.

Kapitulu honetan historia horien berri emateaz gain, gaur egungo erronkak aurkeztuko ditugu. IAren inguruko oinarritzko kontzeptuak eta estrategia nagusien ezaugarriak azalduko ditugu: erregeletan oinarritutako hurbilpen tradizioaletatik, gaur egun IAren arloan optimismoa berpiztu duten corpusetan oinarritutako estrategietaraino. Estrategia desberdinen abantailak baliatzen dituzten sistema hibridoen gaurkotasuna azpimarratuko dugu eta azken urteetako itzulpen-laguntzarako tresna arrakastatsuak ere aurkeztu.

Tesi-lan honetan aurkezten dugun itzultzaile automatikoa euskararekin lan egiten duen lehenengo sistema erabilgarria bada ere, azken urteetan eus-kara tratatzen duten IAren inguruko arloan abiatu dira zenbait ikerketa.

Kapitulu honetan gainbegiratu bat emango diegu: itzultzaileentzako tresnen garapena, corpus elebidunak parekatzeko lanak, itzulpen-memorien garapena, metodo estatistikoekin ikerketak eta sorkuntza automatikoan egindako lanak.

Kapitulua bukatzeko, gaur egun gero eta zabalduagoa den kode irekiko software librearen ideia geurera ekarriko dugu, bere ezaugarri eta onurak agertuz, eta IAren alorrean oraindik urriak diren kode irekiko sistemak eza-gutaraziz.

II.1 Itzulpen automatikoa

A computer is a device that can be used to magnify human productivity. Properly used, it does not dehumanize by imposing its own Orwellian stamp on the products of the human spirit and the dignity of human labor but, by taking over what is mechanical and routine, it frees human beings for what is essentially human. Translation is a fine and exacting art, but there is much about it that is mechanical and routine and, if this were given over to a machine, the productivity of the translator would not only be magnified but his work would become more rewarding, more exciting, more human.

(Kay, 1997)

Itzulpen automatikoaren inguruko oinarritzko informazioa zenbait liburutik aurki daiteke (Hutchins eta Somers, 1992; Somers, 2003; Trujillo, 1999; Nirenburg, 1987; Arnold *et al.*, 1994). Arlo honetara hurbiltzeko oso gomendagarriak dira John Hutchins idazlearen artikuluko orokorrak, bere web orrian¹ aurki daitezkeenak. Sakondu nahi duenarentzat lasterbiderik ederra da Hutchinsek berak prestatutako eta Interneten publiko dagoen *Machine Translation Archive* biltegi elektronikoa², non itzulpen automatikoaren eta itzulpenenerako laguntza-tresnen inguruan argitaratutako artikuluko gehienak aurki daitezkeen. Gainera liburuen, aldizkarien, kongresuetako akten eta bestelako baliabideen erreferentziak ere ematen dira. IAri buruzko sarrera bat euskaraz (Abaitua, 2002) lanean aurki dezakegu.

II.1.1 Oinarritzko kontzeptuak

Itzulpen hitzak itzultzearen ekintza eta ondorioa adierazten du, hau da, hizkuntza batean esan edo idatzi dena beste hizkuntza batean esan edo idaztea (Sarasola, 1996).

¹<http://ourworld.compuserve.com/homepages/WJHutchins>

²www.MT-Archive.info

Historian egin ziren lehenengo itzulpenak Bibliarenak izan ziren, hebraieratik grezierara, bere hizkuntza galduta zuten judutarrek liburu sakratua ulertu ahal izateko. K.a.ko III. mendean burutu zuten itzulpen hau 72 jakitunek Ptolomeo Filadelfo-k aginduta. Hala ere, ezagutzen den lehenengo literatura-itzultzailearen izena Lucius Livius Andronicus da, K.a. 240. urtean Homeroren Odisea itzuli zuena grezieratik latinerara erromatar ezjakinentzat. Harez gero itzultzaileek lan asko egin dute, bai testuak itzultzen, bai itzulpenaren teoriak garatzen. Itzulpenen automatizazioa aspaldiko ametsa bada ere, soilik ordenagailuen agerpenarekin hasi da errealitate bihurtzen.

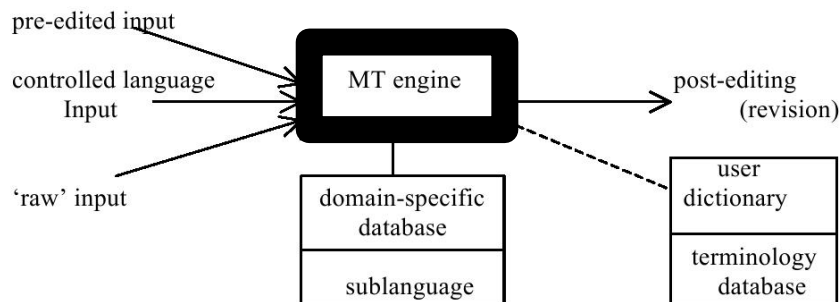
Itzulpen Automatikoa (IA; *Machine Translation, MT*) esaten zaio sistema informatikoak erabiliz hizkuntza batetik beste batera itzultzen duen prozesuari, eta prozesu hori aztertzen duen ikerketa-arloari ere.

Konputagailuak itzulpenarako erabiltzen direnean bi sistema mota bereizi behar ditugu:

- Itzulpen automatikoko sistemak. Testu osoak modu guztiz automatikoan itzultzen dituzten sistemak. Helburua kalitate handiko itzulpenak egitea bada ere, sistemen irteera ez ohi da egokia izaten eta gizakien esku hartzea beharrezkoa da jatorrizko testua egokitzen, itzulitako testua zuzentzen, edo prozesuan zehar ordenagailuarekin elkarrekintzan. Sistema hauek izendatzeko, gizakiak lagundutako itzulpen automatikoa (*Human-Aided Machine Translation, HAMT*) terminoa erabiltzen da. Itzulpen automatikoko sistema batek esaldi osoak hartzen ditu sarreran, eta beste hizkuntzan dagozkion esaldi osoak ematen ditu. Hiztegi bidezko itzulpenari, hau da, abiapuntuko testuaren hitzak xede-hizkuntzako hitzekin ordezkatzeko soilari ez zaio itzulpen automatikoa esaten.
- Itzulpen-laguntzarako tresnak. Itzultzaileei laguntza linguistikoa ematen dieten sistemak: hiztegiak, gramatikak eta, gaur egun batez ere, itzulpen memoriak (aurretik itzulitako testuekin eratutako datu-baseak). Horrelakoetan ordenagailuz lagundutako giza-itzulpenaz (*Machine-Aided Human Translation, MAHT*) hitz egiten da.

Esan bezala, IAren azken helburua, itzulpen-prozesu osoaren automatizazioa da. (Hutchins, 2005a) lanetik ateratako II.1 irudian ikus dezakegun bezala. IAko sistema erdian kokatzen da, eta *kutxa beltz* bat balitz bezala, sarreran testu bat jaso eta irteeran itzulitako testua ematen du.

Batzuetan sistemek emandako itzulpenak dauden bezala uzten dira, errebisatu gabe, itzulpen gordin gisa, itzulpenaren hurbilpena edo zirriborro bat



Irudia II.1: Gizakiak lagundutako itzulpen automatikoa

nahikoa denean. Baina irteeraren kalitatea ziurtatu behar denean, sistema-ren emaitzak giza-itzultzaileek errebisatzen dituzte. Postedizioa deitzen zaio horri, eta beharrezkoa izaten da bi hizkuntzak ezagutzen dituzten itzultzaileek egitea. Sistemak ematen duen irteeraren kalitatea txarra bada, edo postedizioko lanetan trebatua ez dagoen itzultzaile batek egiten badu, hutsetik itzultzea baino denbora gehiago eraman dezake postedizioko lanak.

Itzulpenen kalitatea hobe daiteke jatorrizko testuan murrizpenak ezarri. Azpihizkuntza batera mugatutako testuak itzultzeko sistemak eraiki daitezke, eremu berezi baterako edo dokumentu mota zehatz baterako.

Sistamarako sarrera gordina (*raw*) izan daiteke, baina itzulpenetan eman daitezkeen hainbat errore saihas daitezke sarrerako testuak prestatuz: aurre-edizioa burutuz edo hizkuntza kontrolatua erabiliz.

Aurre-edizioan ondoko lanak egiten dira: izen bereziak eta hitz anitzeko terminoak identifikatu, hitzen kategoria markatu, testua puskatan antolatu, sintagmen mugak seinalatu, aurrizki eta atzizkiak markatu, esaldi luzeak esaldi txikiagotan moztu...

Beste aukera bat abiapuntuko testuan hizkuntza kontrolatua erabiltzea da, anbiguitasunak eta konplexutasuna murrizteko. Horretarako, sarrerako testuen hiztegia eta egitura gramatikalak mugatzen dira, testu horiek idazteko orduan erregela zorrotzak jarraitu behar direla: hitz anbiguoak ez erabiltzea, esaldi bakoitzean mintzagai bakarra egotea, esaldi motzak idaztea, koordinazioa saihestea...

Edozein kasutan itzulpen automatikoari aurre egiteko estrategiak bitan bana daitezke:

- Erregeletan oinarritutako itzulpen automatikoa (EOIA; *Rule Based Machine Translation*, *RBMT*). Sistema hauetan beharrezkoa den eza-

gutza linguistikoa adituek kodetzen dute erregeletan. Hiru estrategia desberdin daude: itzulpen zuzena, transferentzian oinarritutakoa eta interlingua bidezko hurbilpena.

- Corpusetan oinarritutako itzulpen automatikoa (COIA; *Corpus based Machine Translation*, CBMT). Sistema enpirikoak dira, non itzulpena burutzeko ezagutza corpusetatik jasotzen den. Bi estrategia bereiz daitezke: adibideetan oinarritutako IA (AOIA; *Example Based MT*, *EBMT*) eta itzulpen automatiko estatistikoa (IAE; *Statistical MT*, *SMT*).

Azken urteetako ikerketa gehienak corpusetan oinarritutako hurbilpen enpirikoen inguruan egin dira, eta hainbatentzat erregeletan oinarritutako estrategia aspaldian zaharkitua gelditu da. Baina gaur egungo sistema komertzial guztiak erregeletan oinarrituta dira Boitet-ek (2001) esaten duen bezala:

Whatever is said in conferences, the reality of current MT is that all commercial systems are rule-based, and that their linguistic resources, lexical, syntactical and semantic, are by and large hand-crafted.

Etorkizuneko IA ziur aski estrategia ezberdinen konbinazioa egitetik sortuko diren sistema hibridoen eskutik etorriko da.

Bestalde, hizketaren IAK, gaur egun, interes handiak biltzen ditu, aplikazio posible ugari baitaizka. Estaldura handiko sistemak egiteko oso urrun gaudenez, arlo honetan eremu zehatzetarako sistemak garatzen ari dira (hotelen eta kongresuetarako erreserbak, mediku kontsultak, militarrentzako domeinuak...). Hizketaren itzulpenetarako metodo asko erabiltzen dira: erregeletan zein corpusetan oinarritutako IA, hizketaren ezagutza, ezagutza datu-baseak...

IAko sistemen kalitatea

Lortutako sistema automatikoak, oraingoz, perfektuak izatetik oso urrun daude.

What has been achieved is the development of programs which can produce “raw” translations of texts in relatively well-defined subject domains, which can be revised to give good-quality translated texts at an economically viable rate or which in their unedited state can be read and understood by specialists in the

subject for information purposes. In some cases, with appropriate controls on the language of the input texts, translations can be produced automatically that are of higher quality needing little or no revision. (Hutchins eta Somers, 1992)

IAren hasierako kalitate osoko itzulpenen asmo handinahiak alde batera utziz, gaur egungo helburua egoera zehatzetan erabilgarriak diren itzulpenak lortzea da.

Fully automatic systems capable of producing idiomatic texts comparable to human translation are no longer the goal of MT research. (...) In favourable conditions, limited-domain systems which are far from perfect can be and are being used successfully and cost-effectively. Of course, everyone wants to see improved quality, but it is not expected in the near future. (Hutchins, 1995b)

Argi dago, gaur egungo IAko sistemak iraganekoak baino askoz hobeak eta azkarragoak direla (ordenagailuak ere azkarragoak direlako) eta erakunde handiek dirua aurrezten ari direla IAko sistemek egindako itzulpenen zirriborroekin. Baina badirudi azkeneko 20 urteetan IAko sistemen kalitatea ez dela asko hobetu. Hutchins-ek (2003) historian zehar garatutako IAko sistema komertzialak aztertu ostean, haien emaitzen kalitatea konparatu du hurrengo ondorioekin:

It does seem that overall there has been definite progress since the mid 1960s and very probably since the early 1970s. What is more uncertain is whether and where there have been improvements since the early 1980s.

Izan ere, IAk zailtasun linguistiko handiei aurre egin behar die (Hutchins, 2006b): anbiguotasuna, kolokazioak, egitura konplexuak, elipsia, koordinazioa, aditz-esapideak, egitura ez-konposizionalak eta ez-jarraituak, egitura elebidunen dibergentziak, erreferentziak, anafora, estilo desberdintasunak edo arazo semantikoak. Errore gehienak esaldi luzeak itzuli nahi direnean ematen dira.

Gaur egun, hala ere, badirudi optimismorako aukera badagoela, zenbaiten ustez corpusetan oinarritutako paradigmeekin IAren kalitateak jauzi handia egin dezakeelako:

There is now also optimism for the future of the field and for achieving even better quality. The major reason for this change

has been a paradigm shift away from linguistic/rule-based methods towards empirical/data-driven methods in MT. (Och, 2006)

Harago joan da Alex Waibel, International Centre for Advanced Communication Technology erakundeko zuzendaria. Elkarrizketa batean³, hurrengo urteetan itzulpen teknologietan eztanda bat emango dela baieztatzen du eta IA noraino irits daitekeen galderaren aurrean ondokoa esatera ere ausartzen da:

There is no reason why they should not become as good, if no better, than humans.

Hala ere, IAren historiak erakutsi du horrelako iragarpenak ausartegiak direla eta optimismoaren aurrean jarrera eszeptikoa zentzuzkoagoa delakoan gaude.

MT is a promising technology. It has been a promising technology for more than 30 years. But promises do not materialize. (Bonet, 2006)

IA vs. giza-itzultzaileak

Iraganean giza-itzultzaileek ez zuten begi onez ikusten IAren ikerketak eman zezakeen emaitza. Dena dela IAKo ikertzaile askok uste zuten IAren helburua giza-itzultzaileak ordezkatzeko zela, partzilaki baten ere. Baina gaur egun egoera aldatu da

But now it is apparent that MT and human translation can and will co-exist in relative harmony. (Hutchins, 2005b)

Testu bat itzultzeko orduan askotan hartu behar dira kontutuan anbiguitasunak sortzen dituzten linguistikatik kanpoko baldintzak: asmoa, egoeraren testuingurua edo testuinguru soziala, historikoa, kulturala... Gainera ñabardurak antzeman behar dira, munduaren ezagutza behar da eta sen ona erabili behar da. Horregatik giza-itzultzaileek soilik itzuli ahal izango dituzte argitaratzeko kalitatea beharko duten testu konplexuak. Hori bai, haien lana errazteko tresna informatikoak izango dituzte eskura.

Hainbat kasutan, berriz, IA izan daiteke irtenbide egokiena eta zenbaitetan bakarra: dokumentazio tekniko kopuru handien itzulpen errepikakor eta aspergarriak (eskuliburu teknikoak, administrazioko testuak...), testuen

³<http://businessnetwork.smh.com.au/articles/2006/06/29/5104.html>

itzulpen-zirriborroak, telebistako azpitituluen itzulpena denbora errealean, web orrien itzulpena linean, posta elektronikoren itzulpena... Eskuzko itzulpenaren bidez egingarriak ez diren egoeretan, IAK ez ditu itzultzaileak ordezkatzeko, ez-itzultzea baizik (Gerber, 2001).

IAren erabilerak

Bi erabilera nagusi bereizi izan dira IAKo sistemetan, asimilazioa (*assimilation*) eta zabalkundea (*dissemination*):

- Asimilaziorako IA. Oinarrizko ulermenerako balio duten itzulpen zaka-rrak ematen ditu. Emandako itzulpenak ez dira onak, ezin dira errebi-satu kalitatezko itzulpen bat lortzeko, baina irakurleak informazioaren muina edo behintzat gaia uler dezake. Sistema azkarrek dira.

Gaur egun oinarrizko ulermenerako IA Internetetik nabigatzeko eta informazio bilaketarako (helburu militar, ekonomiko edo zientifikoe-kin) erabiltzen da nagusiki. Erabiltzaileak zenbat eta gutxiago ezagutu abiapuntuak hizkuntza, orduan sistema gero eta erabilgarriagoa izango da. Itzultzaileek sistema hauek emandako itzulpena iradokizun bezala erabil dezakete.

Mota honetako IAKo sistema merke asko egon arren, erabilgarri dau-den hizkuntz bikoteen kopurua dagoen beharra baino askoz ere txi-kiagoa da. Adibide gisa Europako Batasunean dauden 380 hizkuntza bikoteetatik (kontutan hartuta soilik 20 hizkuntza ofizialak), Euro-par Komisioaren *EC-Systran* sistemak soilik 18 bikote dauzka (Petrits *et al.*, 2001).

Hizkuntza bikote berri baten lehenengo bertsio komertziala garatzeko gutxienez 40 urte/pertsona behar direla estimatzen da (Boitet, 2001).

- Zabalkunderako IA. Kalitatezko itzulpenak ematen ditu, baina emai-tza argitaragarria dela ziurtatzeko postedizioa beti beharrezkoa da. Gehienetan ere sarrera kontrolatzen da, aurre-edizioa burutuz edo hiz-kuntza kontrolatuak erabiliz.

Eremu zehatz baterako sistema berezituak izaten dira, emaitza oso onekin eta oso erabilgarriak. Erabiltzen da eskuliburuak, publizitatea edo dokumentazio tekniko itzultzeko (txosten handiak, errepikako-rrak eta maiztasun handiko eguneraketa dituztenak). Oso baliagarria da itzulpena hizkuntza anitzetara egin behar denean. Mota honetakoa da *Météo* meteorologi-iragarpenak itzultzeko sistema, 30 milioi hitz

urtero itzultzen dituen, orrialde bakoitzak soilik minutu bateko errebisioa behar duena (Boitet, 2001). Sistema adituak diren neurrian, oso onak dira bere arloan, baina soilik bere arloan, kalitatea eta estaldura alderantziz proportzionalak baitira.

Hizkuntza bikote berri baten lehenengo bertsio komertziala garatzeko gutxienez 100 urte/pertsona behar direla kalkulatu da (Boitet, 2001).

Bi erabilera nagusi klasiko horiez gain, gaur egun elkartrukerako (*interchange*) eta datu-baseen atzipenerako (*database access*) aplikazioak ere gero eta garrantzi handiagoa ari dira hartzen, batez ere Internet-en hazkundearekin (Hutchins, 2004):

- Elkartrukerako IA. Hizkuntza desberdinak erabiltzen dituzten pertsonen arteko komunikazioa (gutunen, posta elektronikoaren edo telefonoaren bidezkoa) ahalbidetzen du. Itzulpenaren kalitatea ez da garrantzizkoa, lehenetsia mezua ulertzea baita.
- Datu-basen atzipenerako IA. Datu-base batean erabiltzaileak ulertzen ez duen hizkuntza arrotzean idatzita dagoen informazioa bilatzen laguntzen du. Gaur egun Interneten bilaketak egiteko eta web orriak atzitzeko erabiltzen da.

IAko sistemen konplexutasuna

IAko sistemak hiru kategorietan bana daitezke: erabilera pertsonalekoak, profesionalak eta enpresa mailakoak. *IAMT Certification Initiative* ekimenak (Hutchins, 2000) kategoriatan horietako ezaugarriak definitu zituen sistemak sailkatzeko helburuarekin:

- Oinarrizko maila: hiztegiak handienak 50.000 sarrera baino gutxiago ditu, edo ez dauka erabiltzaileak hiztegia zabaltzeko erraztasunik, edo perpaus bakarreko esaldi sinpleak soilik itzultzen ditu. Ez da egokia itzultzaile profesionalentzat, eta ez du kalitate oneko itzulpenik ematen. Nagusiki testuen muina ulertzeko edo posta arrunterako erabiltzen da.
- Maila estandarra (edo profesionalen maila): hiztegiren batek 50.000 sarrera baino gehiago ditu, erabiltzaileak hiztegia zabaltzeko erreztasunak ematen ditu eta ez da mugatzen perpaus bakarreko esaldi sinpleen

itzulpenera. Itzultzaileentzako sistema hauek ez dute argitaratzeko kalitaterik ematen zuzenketarik gabe, baina baldintza egokietan kostua merketzen dute.

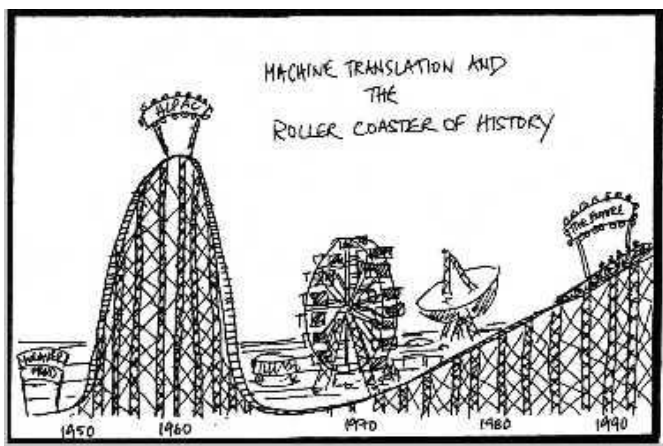
- Maila aurreratua (edo enpresa maila): hiztegiak txikiak 75.000 sarrerak baino gehiago ditu, hiztegia erabiltzaileak zabaltzeko erreztasunak ematen ditu, ez da mugatzen perpaus bakarrek esaldi sinpleen itzulpena eta bezero-zerbitzaria arkitektura eskaintzen du. Prozesaketa handietarako sistemak dira hauek: dokumentu errepikakorrak, teknikoak, administrazioak...

IAren onurak

Itzulpen automatikoak ekar ditzakeen onurak oso modu argian laburbiltzen ditu Hutchins-ek (2005a) ondoko galderari erantzuna ematen dionean: zergatik erabili nahiko genituzke ordenagailuak itzulpen lanetarako?

- Badago itzultzeko beharra duen material asko, eta giza-itzultzaileek ezin dute denarekin.
- Dokumentu teknikoak oso aspergarriak dira giza-itzultzaileentzat eta ez zaie atsegin itzultzea.
- Korporazio handientzat itzulpen teknikoetan garrantzitsua da terminologia koherentzia erabiltzea, hau da, termino berak beti itzulpen bera izatea. Konputagailuek hala egiten dute, baina giza-itzultzaileek aldatzeko joera izaten dute.
- Itzulpen automatikoko tresnek itzulpen-prozesua azkar dezakete.
- Kalitate handiko itzulpena ez da beti beharrezkoa.
- Itzulpen automatikoak eta itzultzaileentzako tresnek kostuak murrizteko balio dezakete.

IAren inguruko oinarritzko kontzeptuak aurkeztu ondoren, hurrengo atalean bere historiaren gertaera garrantzitsuenen berri ematen dugu.



Irudia II.2: Itzulpen automatikoaren ikerketaren gora-beherak

II.1.2 Historia

John Hutchins ikerlariari zor diogu IAren historia zehaztasunez jaso izana. Atal honetan azalduko ditugu historia horretako mugarri aipagarrienak, bere idatzietan (Hutchins, 1995a, b; Hutchins eta Somers, 1992) eta (Abaitua, 2002) lanean oinarrituz.

IAren historiaren irudikapen grafiko umoretsu bat II.2 irudian ikus dezakegu (Arnold *et al.*, 1994).

Konputagailuaren aurretik

XVII. mendean Descartes eta Leibniz pentsalariak kode numeriko unibertsalean oinarritutako hiztegien gauzatzeari buruzko teoriak proposatu zituzten, *hizkuntza unibertsalaren* aldeko mugimenduaren barruan.

1933an bi patente ezberdin erregistratu ziren. Frantziar George Artsouni ikerlariak hitz baten beste hizkuntza bateko baliokidea ematen zuen paper bandako biltegitratze-dispositiboa diseinatu zuen. Sobietar Errepublika Sozialisten Batasunean (SES) Petr Smirnov-Troyanskii ikerlariak itzulpenak egiteko prozedura mekaniko bat diseinatu zuen, hiru faseetan banatzen zena: abiapuntuko hitzen analisi logikoa oinarritzko formak eta funtzio sintaktikoen sekuentzia lortzen zuena, sekuentzi horien transformazioa xede-hizkuntzako sekuentzietan, eta sekuentzia horien bihurketa hitzen formetan.

1. aroa: 1949-1954. Aitzindariak

1949an Ameriketako Estatu Batuetan (AEB) Warren Weaver-ek, Rockefeller Fundaziokoak, ordenagailuen asmakuntza eta urte gutxitara itzulpen automatikoaren ideiarekin berri eman zuen metodo ezberdinak proposatuz: gerra garaiko teknika kriptografikoen erabilera, analisi estatistikoak, Shannon-en informazio teoria eta hizkuntzaren azpiko logika eta ezaugarri unibertsalak.

I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text. (Weaver, 1949)

1951an Yehoshua Bar-Hillel, Massachusetts-eko Teknologia Institutuko (MIT), IAan aritzeko arduraldi eskusiboa izan zuen lehenengo ikerlari kontratatua izan zen. Urtebete geroago IAren lehenengo sinposiuma antolatu zuen eta, bertan, besteren artean ondoko gaiak eztabaidatu ziren: sintaxia, hizkuntza kontrolatuak, azpihizkuntzak eta gizakien esku hartzearen beharra (aurre-edizioa eta postedizioa).

2. aroa: 1954-1966. Optimismoa eta desilusioaren hamarkada

1954an Leon Dostert-ek IBMrekin batera Georgetown-eko Unibertsitatean burutu zuen proiektu baten emaitza IAko programa baten lehenengo erakustaldia izan zen. Arretaz aukeratutako errusierazko 48 esaldi ingelesera itzuli ziren, 250 hitzetako hiztegia eta 6 erregela gramatikal erabiliz.

Erakustaldiak AEBetako aurrekontuetan partida garrantzitsuak (gehienak Defentsako Ministeriotik) IAren ikerketarako bideratzea eragin zuen. Aldi berean SESBean eta beste hainbat tokitan ere IAko proiektuak martxan jarri ziren. Arrazoi militarrengatik AEBetan egindako ikerketa gehienak errusieratik ingelesera itzultzeko izan ziren, eta Sobiet Batasunean, berriaz, ingelesetik errusierara itzultzeko.

Ikerketa bi lerro oso ezberdinetan polarizatu zen: *indar hutsa* (*brute-force*) deitutako hurbilpen enpirikoak, metodo estatistikoak erabiliz epe motzean sistemak eraikitze asmoa zutenak; eta *perfekzionistak* (*perfectionist*) deitu zitzaizkien hurbilpen teorikoak, oinarritzko linguistikan ikerketa eginez, epe luzerako soluzioak helburu zutenak. Egindako ikerketak eragin handia izan zuen bai linguistika konputazionalan, bai adimen artifizialean.

Georgetown Unibertsitatean *Systran* (Gachot, 1989) sistemaren oinarriak ezarri ziren, eta Texas-eko *Linguistics Research Center* ikerketa-gutunean *Metal* sistemarenak (Bennett eta Slocum, 1985; Schneider, 1995).

Hasierako optimismoak eragin zuen kalitate handiko itzulpen automatiko osoa (*Fully Automatic High Quality Translation, FAHQT*) delakoaren helburua planteatzea. Konputagailuen garapenak eta linguistika formaleko aurrerapenak, bereziki sintaxiaren arloan, itzulpen-sistemen kalitatean jauzi handi bat laster izango zela pentsatzera eramane zuten. Baina konplexutasuna eta arazo linguistikoak azalarazten ziren neurrian, desilusioa hazi egin zen pixkanaka.

1960an Bar-Hillelek kalitate handiko itzulpen guztiz automatikoaren ideia zalantzan jarri zuen. Giza-itzultzaileek lortutako emaitzen parekoak lortzeko, ezagutza semantiko eta pragmatiko zabala gehitu beharko litzatekeela esan zuen, eta IAn helburu txikiagoak jartzeko gomendioa eman zuen.

3. aroa: 1966-1976. ALPAC txostena eta bere ondorioak

1966an *ALPAC (Automatic Language Processing Advisory Committee)* izeneko komiteak, IArengo egoera eta etorkizunerako aukerak baloratzeko 1964an eratua, itzal handiko txostena argitaratu zuen. Bertan IA giza-itzulpena baino motelagoa, txarragoa eta bi aldiz garestiagoa zela baieztatzen zuen, eta ez zela aurreikusten IA epe laburrean erabilgarria izan zitekeenik.

There has been no machine translation of general scientific text, and none is in immediate prospect. (ALPAC, 1966)

ALPAC txostenak itzultzaileentzako tresnak garatzeko eta LNPrako oinarritzeko ikerketekin jarraitzeko gomendioa ematen zuen. Ikerlari askok txostena gaitzetsi bazuten ere, IArako inbertsioak murriztu ziren, hamarkada batez AEBetako ikerketen bukaera suposatuz.

Hala ere, Kanadan eta Europan ikerketak aurrera jarraitu zuen, baina oso behar ezberdinekin: eleaniztasunak administrazioan sortzen zuen itzulpenen eskaera handiari aurre egitea, batez ere.

1970an AEBetako Aireko Armadan Peter Toma ikerlari hungariarrak garatutako *Systran* errusiera-ingeles sistema instalatu zuten. Logos Corporation enpresak bere IArako lehenengo saiakera ere egin zuen hegazkinen eskuliburuak itzultzeko ingeles-vietnamera *Logos* sistema garatuz (Scott, 1989, 2003).

1976an Kanadan TAUM taldeak (*Traduction Automatique de l'Université de Montréal*) *Météo* sistema (Chandioux, 1976; Philippe Langlais eta Lapalme, 2005) martxan jarri zuen, eguneroko meteorologi iragarpenak ingelesetik frantsesera itzultzeko. Arloko hiztegi eta sintaxi mugatuek eragin zuten sistemaren kalitate handia. Aurreko urteetan agertuak ziren ere azpihizkuntzekin lan egiten zuten beste bi sistema: *Titus* Frantzian eta *Cult* Chinan.

1976an ere, Europan, Europar Batasunak (EB) bere egoitzetan *Systran* ingelesa-frantsesa IAKo programa instalatzea erabaki zuen.

Urte horietako ikerketa ahalegin nagusiak interlinguazko hurbilpenetan egin ziren. Grenobleko Unibertsitatean *Ceta* sistema garatu zuten pibot lengoaia bat erabiliz (Boitet, 1988). Antzeko eredua jarraitu zuten Texasen *Metal* sistema eraikitzeke. SESBean interlingua bidezko sistemetan burutzen ari zen ikerketak anbizio handiagoa izan zuen, Mel'cuk-en *meaning-text* teoriak eraginda (Mel'cuk, 1988; Polguère, 1998). Baina hurbilpen hauek arazo handiekin egin zuten topo, bereziki abiapuntu-hizkuntzako sarrera-testuen azaleko formen informazio galera, xede-hizkuntzako formen hautaketa lagun zezakeena. Interlingua bidezko hurbilpenaren etorkizuna zalantzan jarri zen eta orokorrean onartu zen xumeagoa zen transferentzian oinarritutako saiakerak posibilitate gehiago eskaintzen zituela.

4. aroa: 1977-1988. Sistema komertzialen agerpena

Itzulpen automatikoak arlo askotan aurrera egin zuen: hainbat sistema erabilgarri agertu ziren, IAKo sistemen merkatua handitu egin zen eta IAren inguruan ikerketa-ildo ugari zabaldu ziren.

Sistema komertzialen artean *Systran* izan da inoiz izandako aplikazio orokorretarako diseinatutako sistema operazionalik arrakastatsuenak. Hitz anitzeko unitateen hiztegi handia eta itzulpen-moduluen kodeketa eraginkorra dira bere arrakastaren arrazoi nagusiak. Beste zenbait sistema egonkor ere garatu ziren: alemanera eta ingeles artean itzultzeko *Logos* eta *Metal*, azkeneko hau hasieran interlingua bidezko hurbilpenarekin eta geroago transferentziazko eredua jarraituz. Ingurune zehatzetarako *Pan American Health Organization* (PAHO) erakundeak espainiera-ingeles *Spanam* itzulpen zuzeneko sistema eta ingelesa-espainera *Engspam* transferentziazko sistemak eraiki zituen (Léon, 1985). *Smart Corporation* bezalako hainbat enpresa, berriz, neurrira eraikitako sistemen eraikuntzan trebatu ziren. Jarduera komertzial handiena Japonian izan zen, Japoniako indar ekonomiko-teknologikoari esker, ingeles eta japonieraren arteko sistema asko garatu zirela: *Pensee*, *Meltran*, *As-Transac*, *Hicats*, *Atlas*...

Ikerketan hiru lerro nagusitu ziren: transferentzian oinarritutako sistema aurreratuen eraikuntza, interlingua bidezko sistema mota berrien garapena eta adimen artifizialeko tekniken ikerketa.

Grenoblen GETA taldeak (*Groupe d'Etudes pour la Traduction Automatique*) bere *Ariane* sistema garatzen hasi zen (Vauquois eta Boitet, 1985), transferentziazko sistemen paradigma bezala hartua izan dena (II.1.3 atalean aurkezten dugu sistema honen deskribapena). Sistema honek beste proiektu

askotan eragina izan zuen, adibidez, Kiotoko Unibertsitatean Makoto Nagao ikerlariaren zuzendaritzapean eraikitako *Mu* sisteman (Nagao, 1985). Saarbrück-en, Alemanian, *Susy* transferentzia bidezko sistema eleanitza garatu zuten, modulartasun oso handikoa eta teknika linguistiko oso ezberdina erabiltzen zituena (Maas, 1984). Europar Batasunak bertako hizkuntza ofizial guztien artean itzuliko lukeen IA programa eleanitza garatzeko ikerketa-proiektu handia abiatu zuen, *Eurotra*, baina ez zen prototipo operatiborik sortu (Johnson *et al.*, 1985; Maegaard, 1995).

Utrecht-en, Holandan, BSO konpainiak *DLT* (*Distributed Language Translation*) interlingua bidezko sistema garatu zuen, interlingua bezala Esperanto erabiliz (Witkam, 1987). Holandan ere, Eindhoven-en, Philips konpainiak *Rosseta* proiektua garatu zuen, interlinguazko adierazpenetarako Montague-ren gramatika erabiliz eta gramatiken itzulgarritasuna aztertuz (Leermakers eta Rous, 1986).

Japonian ikerketa handia egin zen, bai interlingua bidezko sistemetan eta ezagutzen oinarritutako esperimentuetan. *CICC* proiektua martxan jarri zen sistema eleanitz bat eraikitze asmoarekin.

Ezagutzen oinarritutako hurbilpenek indarra hartu zuten, azpimarraztekoa izanik Carnegie Mellon unibertsitatean garatutako *Kant* prototipoa (Carbonell *et al.*, 1992).

5. aroa: 1988-1995

80ko hamarkadaren bukaera arte IAKo ikerketan izandako erregela linguistiko nagusitasuna hautsi zen corpusetan oinarritutako metodo berrien agertzearekin. Hurbilpen horietan ez da erregela sintaktiko edo semantikorik erabiltzen, itzulpenak egiteko informazioa corpus handietatik jasotzen dela.

Alde batetik 1988an IBMko talde batek metodo estatistikoak soilik erabiltzen zituen *Candide* sistema eraginkorraren emaitzak argitaratu zituen (Brown *et al.*, 1990), IAren ikerketako lehenbiziko garaietan emaitza onik lortu ez zuten teknika estatistiko erabilera berpiztuz. Onargarriak ziren emaitzak hobetzeko asmoz metodo estatistiko konplexuagoak erabiltzeaz gain, IBMko taldeak informazio linguistiko minimoa erabiltzea proposatu zuen.

Bestetik Japonian adibideetan oinarritutako hurbilpena erabiltzen hasi zen 1984an Makoto Nagao ikerlariak proposatutakoari jarraituz.

Beste berrikuntza bat hizketaren itzulpenean egindako ikerkuntza izan zen, non hizketaren ezagutza eta sintesia, itzulpen-moduluekin integratzen zen. *C-Star* proiektuan (Woszczyna *et al.*, 1998), Japoniako ATR ikerzentroa, AEBko Carnegie-Mellon Unibertsitatea eta Alemaniako Karlsruhe Uni-

bertsitatea elkarlanean, telefono bidezko denbora errealeko itzulpen-sistema bat garatzen hasi ziren, ingelesa, japoniera eta alemanerako, kongresuetako izen-emateak egiteko elkarriketen eremura mugatuta. *Verbmobil* proiektu erraldoian (Wahlster, 1993), Alemaniako gobernuaren finantziazioarekin, enpresen arteko negoziaketetarako baliagarri litzatekeen hizketa-laguntza bati ekin zioten.

Hala ere, erregeletan oinarritutako ikerketak ere aurrera jarraitu zuen. *Eurotra* proiektuan egindako ikerketaren emaitza *PaTrans* transferentzian oinarritutako sistema da (Hansen, 1994). Carnegie Mellon Unibertsitatean ezagutzen oinarritutako hurbilpena eta sarrera kontrolatua konbinatzen dituen *Catalyst* sistema eraiki zuten. *Pangloss* interlingua bidezko sistema Southern California, New Mexico State eta Carnegie Mellon Unibertsitateetan garatu zen (Nirenburg, 1995).

Urte haietako garapen arrakastatsuen itzultzaile profesionalentzat pentsatutako *itzultzailearen lan-estazioak* izan ziren. Hizkuntza kontrolatuetan eta eremu zehatzeko sistemetan ere aurrerapen handiak egin ziren eta softwarearen lokalizazioak ere indarra hartu zuen. Ordenagailu pertsonaletarako IAko programen salmentak izugarri handitu ziren eta Interneten IAko zerbitzuak linean eskaintzen hasi ziren.

6. aroa: 1995-2006

Aldaketa nabarmenak eman dira azken urteetan IAren arloan, Interneten hedapenak eragin handia izan duela sistemen garapenean.

Ikerketa gehienak itzulpen automatiko estatistikoaren arloan kokatu dira, eta nahiz eta oraindik ez den sistema komertzialetan txertatu, badira horretan ari diren enpresak, besteak beste, Language Weaver eta Google. Interes handia dago hibridazioan, bereziki sistema estatistikoetan informazio morfologikoa edo sintaktikoa erabiliz, eta erregeletan oinarritutako sistemak metodo estatistikoekin konbinatuz.

Enpresen eta merkatuen globalizazioak produktu eta zerbitzuen lokalizazioaren beharra dakar, hau da, ez soilik itzulpena, baizik eta tokian tokiko egokitzapen kultural eta linguistikoa. Itzulpena informazioaren kudeaketaren beste osagai bat gisa integratzen da.

IA ebaluatzeko metodologietan berrikuntzak (ikus VIII.1 atala) oso garrantzitsutzat jotzen dira, bereziki IBMk proposatutako *Bleu* sistema (Papineni *et al.*, 2002). AEBetako NIST⁴ (*National Institute of Standards and Technology*) erakundeak IAko sistemen ahalmenak neurtzeko ebaluazio ire-

⁴www.nist.gov

kiak antolatzen ditu, ikerlerroak bideratzeko kontribuzio handia suposatzen dutenak.

Itzulpen automatikoaren egungo erronkak eta etorkizunerako aukerak II.1.7 atalean aztertuko ditugu.

IAko sistemen belaunaldiak

Itzulpen automatikoaren historian garatutako sistemak hiru belaunaldietan banatu izan dira (Hutchins, 1995b):

- *Lehenengo belaunaldia*: ALPAC txostena argitaratu zen arte itzulpen zuzenaren metodoa erabiltzen zuten sistemak.
- *Bigarren belaunaldia*: ALPAC txostenaren ostean erregeletan oinarritutako transferentzia eta interlingua zeharkako metodoak erabiltzen zuten sistemak.
- *Hirugarren belaunaldia*: Corpusetan oinarritutako hurbilpen berriak erregeletan oinarritutako metodoekin konbinatzen dituzten sistemak.

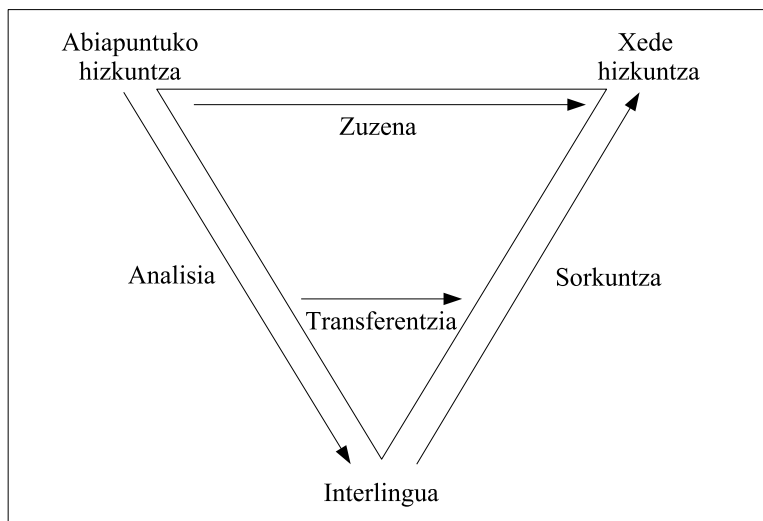
Hurrengo ataletan lehenengo eta bigarren belaunaldiko erregeletan oinarritutako estrategiak zein hirugarren belaunaldiko corpusetan oinarritutakoak eta sistema hibridoetan estrategia desberdinak konbinatzeko moduak aurkezten ditugu.

II.1.3 Erregeletan Oinarritutako IA

Erregeletan oinarritutako IAKo sistemetan (EOIA; *Rule Based Machine Translation, RBMT*) adituek ezartzen dituzte itzulpena egiteko beharrezkoak diren erregelak. Horrek itzulpenaren bi hizkuntzetan adituak diren linguisten esfortzu handia eskatzen du.

Sistema hauek ondoko zailtasunei aurre egin behar diete (Hutchins, 2006b):

- Erregela gramatikalen konplexutasuna: aurreikusi ezinezko erregelen arteko elkarrekintza, estaldura ez osoa... Oso zaila da testuinguru guztiak estaltzen dituzten erregelak zehaztea.
- Hiztegien konplexutasuna eta hautapen-murrizpenak.
- Kolokazioak, aditz-esapideak, izenordainak, anafora...



Irudia II.3: Vauquois-en triangela

- Egitura konplexuak: esaldi luzeak, koordinazioa...
- Arazo semantikoak

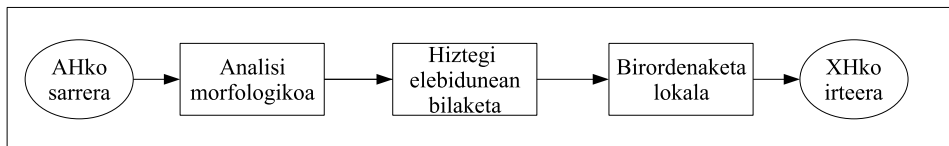
Esan bezala, arazo horietako asko eremu zehatzetako edo hizkuntza kontrolatuetako sistemekin minimiza daitezke.

EOIA modu arrakastatsuan frogatua izan da sistema komertzial eta ikerketa-proiektu askotan. (Bond *et al.*, 2005) lanean azpimarratzen dute guretzat oso garrantzitsua izango den abantaila: EOIAko sistema bateko osagaiak (hiztegiak, gramatikak, analizatzaileak eta sortzaileak) berrerabil daitezke. Honek esan nahi du IAKo sisteman burututako lanak beste proiektu batzuei onura ekarriko diela, eta beste proiektu batzuetako lanek IAKo sistema hobetuko dutela.

Erregeletan oinarritutako sistemetan itzulpen-prozesua hiru fasetan banatzen da: analisia, transferentzia eta sorkuntza.

Erabiltzen den informazio edo ezagutza linguistikoaren arabera hiru estrategia nagusi bereizi ohi dira: itzulpen zuzena, transferentziazko metodoak eta interlingua bidezko sistemak.

Estrategia horiek ulertzeko, Vauquois-ek (1976) lehen aldiz erabilitako diagrama ikus dezakegu (II.3 irudia). Vauquoisen triangeluak estrategia bakoitzean zenbat analisi eta sorkuntza burutu behar den eta abiapuntu- eta xede-adierazpideen artean dagoen antzekotasuna erakusten du. Itzulpen zu-



Irudia II.4: Itzulpen zuzeneko metodoen eredua

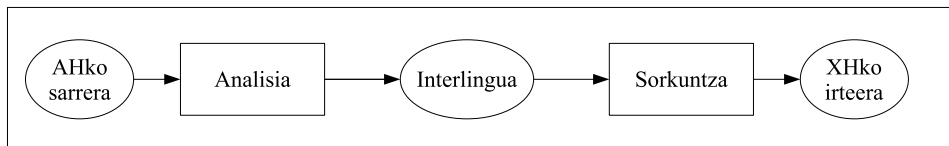
zena oso azaleko analisia eta sorkuntza egiten den muturrean kokatzen da, interlingua bidezko sistemak transferentzia beharrik ez daukan beste muturrean, eta bien artean transferentzia bidezko hurbilpenak.

Itzulpen zuzeneko sistemak

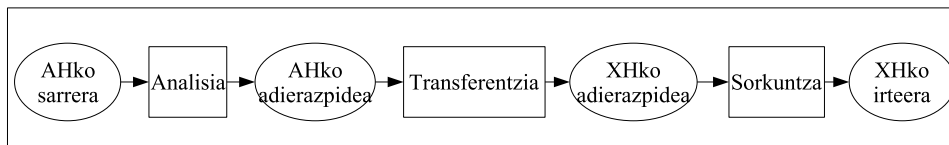
Abiapuntu-hizkuntzatik (AH) xede-hizkuntzara (XH) zuzenean itzultzen da, analisi sintaktiko edo semantikorik egin gabe. Analisi morfologikoa egiten da hitzen formatatik hiztegi elebidunean bilatuko den lema lortzeko. Batzuetan, xede-hizkuntzako hitzen ordenari dagozkion egokitzapenak ere egiten dira, baina testuinguru oso lokala kontuan harturik. Lehenengo belaunaldiko aspaldiko sistemetan erabiltzen zen estrategia hau, eta gaur egun erabilgarria da oso gertu dauden hizkuntza bikoteentzat, hitzez hitz baliokideak lortzea ia nahikoa denean. Hurbilpen zuzenaren eredua II.4 irudian ikus daiteke.

Interlingua bidezko sistemak

Sistema hauetan AHko esaldia analizatzen da lengoaia neutral baten adierazpidea lortuz, hortik XHko esaldia sortzeko. Transferentziaren urratsa sahies-ten da, horren truke analisia eta sorkuntza moduluen konplexutasunak gora egiten duela. Estrategia honek sarrerako testuaren analisi sakon eta osoa eskatzen du, abiapuntu- eta xede-hizkuntzetatik independentea den adierazpide kontzeptuala lortzeko. Interlingua deritzon adierazpide horrek XHko testua sortzeko behar den informazio guztia bildu behar du. Interlingua- ren helburua lexiko eta egitura unibertsala lortzea da, hau da, testuaren *esanahia* adieraztea. Horregatik sistema hauetan esan daiteke, itzuli aurretik, testua *ulertu* egiten dela. Zailtasun handiko erronka da interlingua egoki baten espezifikazioa zehaztea, espezifikazio horretatik edozein hizkuntzatako esaldia eratortzeko. Sistema eleanitzetan estrategia honek abantaila nabar- men bat dauka, hizkuntza berri bat gehitzeko bi modulu berri soilik eraiki behar direlako. Hizkuntza kopurua n bada, moduluen kopurua $2 * n$ izan- go da. Oztoporik handiena, berriz, interlingua definitzearen zailtasuna da. Interlinguazko eredua II.5 irudian ikus daiteke.



Irudia II.5: Interlinguan oinarritutako eredua



Irudia II.6: Transferentzian oinarritutako eredua

Ezagutzen oinarritutako IA (*Knowledge Based MT*, *KBMT*), interlinguazko hurbilpenen artean kokatzen da. Sistema hauetan, sakoneko ulermena lortzeko, eremu bateko ezagutza semantiko eta pragmatiko zabala erabiltzen da, eremu horretako kontzeptuei buruz arrazoitzeko gaitasunarekin. Eremu bereziko sistemak izaten dira, mundu errealearen ezagutza osoa adieraztea oso konplexua baita. Adimen artifizialeko teknikak ere erabiltzen dituzte.

Transferentziazko sistemak

Sistema hauek itzuli behar den AHko testu-zatia analizatzen dute tarteko adierazpide bat lortzeko. AHko testuaren adierazpidea dagokion XHko adierazpidera transferitzen da, hortik XHko testua sortuko delarik. Tarteko adierazpide horiek hizkuntza bakoitzerako espezifikoak izango dira.

Transferentziazko sistemen artean mota desberdinak egongo dira, analisia eta transferentzia gauzatzen diren sakontasun mailaren arabera (analisi sintaktiko partziala, analisi sintaktiko osoa eta analisi semantikoa). Oro har, zenbat eta abstraktuagoa izan transferentziazko adierazpidea, orduan eta errazagoa da transferentziazko modulu egokia eraikitzea. Sistema eleanitz batean hizkuntza berri bat gehitzean, hizkuntza horretarako analisiaren eta sorkuntzaren moduluak eraikitzeaz gain, sistemako beste hizkuntza bakoitzarentzat norabide bakoitzeko transferentzia-modulua beharko da. Hizkuntza kopurua n bada, moduluen kopurua $n*(n+1)$ izango da. Transferentzian oinarritutako eredua II.6 irudian ikus daiteke.

Hiru transferentzia mota desberdin bereiz daitezke: sintaktikoa, semantikoa eta konbinatua. Transferentzia sintaktikozko sistemetan AHko egitu-

ra sintaktikoak XHko egiturak bihurtzen dituzten zuhaitzetik zuhaitzerako transformazioak burutzen dira. Transferentzia semantikozko sistemek testuaren esanahiaren adierazpidea eraikitzen dute, AHren menpekoa dena. Sistema hauen transferentziak nagusiki predikatuen itzulpena burutzen du, hainbatetan transferentzia estrukturalerako erregelak erabiltzen badira ere.

Transferentzian oinarritutako sistemen artean *Eurotra*, *Metal* eta *Ariane* aurkitzen dira, azkeneko hau transferentzian oinarritutako beste proiektu askotarako eredu izan dena.

Trujillo-k (1995) transferentzian oinarritutako IA sistemetarako ezaugarri komenigarriak proposatzen ditu, lehenengo lauak (Rich, 1983) lanetik egokituak direla:

- Adierazpen-egokitasuna (*representational adequacy*): ezagutza mota guztiak adierazteko gaitasuna.
- Inferentzia-egokitasuna (*inferential adequacy*): hizkuntzen arteko erlazio berriak aurretik daudenetatik eratorri ahal izatea.
- Inferentzia-eraginkortasuna (*inferential efficiency*): hizkuntzen arteko ezagutzak erlazio berrien eraikuntza gidatzea.
- Eskuratzeko-eraginkortasuna (*acquisitional efficiency*): hizkuntzen arteko ezagutza erraz zabaltzea.
- Deklaratibotasuna (*declarativity*): ezagutza linguistikoaren deskribapena sistemak erabiliko duen algoritmoen inplementazio zehatzetatik independentea izatea. Horrek sistema ulertzea, aldatzea eta araztea (*debugging*) errazten ditu.
- Monotonia-legea (*monotonicity*): transferentzia-erregela baten eragina ez litzateke blokeatu behar beste erregela bat gehitzean.
- Itzulgarritasuna (*reversibility*): transferentzia-modulu bera erabiliz bi norabideetan itzuli ahal izatea.
- Modulartasuna: transferentzia-modulua banatzea, ahalik eta zehatzen, analisi- eta sorkuntza-moduluen gramatika eta diseinu zehatzetik.
- Gardentasuna: transferentzia-erregelen egiturak hizkuntz bien arteko erlazioen patroiak islatu beharko lituzke. Beraz, problemak ez lituzke transferentziarako formalismoak sortu behar, baizik eta fenomenoek konplexutasunak.

- Lokaltasuna (*locality*): transferentzia-erregelak aplikatzen diren testuingururik txikienari buruzkoak izango dira
- Uniformetasuna: adierazpidea ondo definitua izan behar du
- Koherentziaren egiaztapena: ezagutzaren deskribapeneko inkoherentziak, balio okerren esleipena eta egitura-deskribapen okerrak lehenbailehen identifikatu behar dira

Transferentziazko sistemen adibidea: *Ariane*

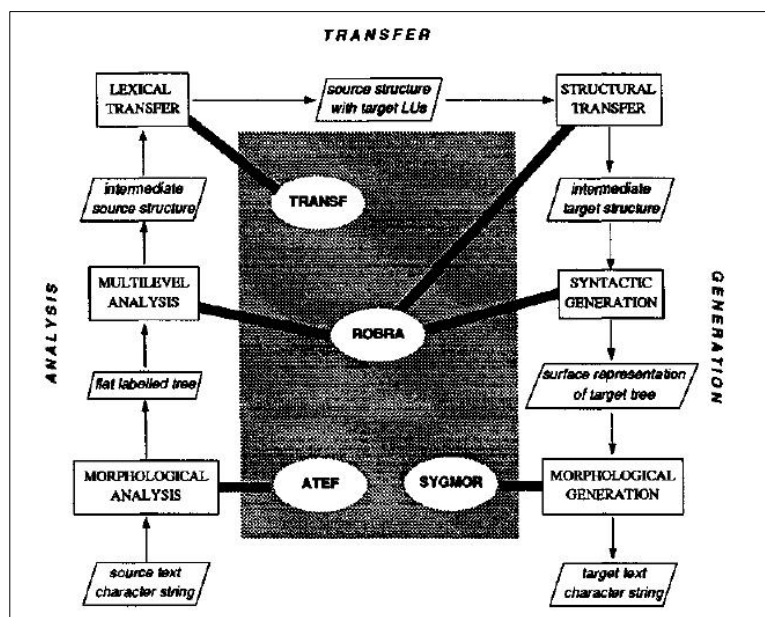
GETA taldeak garatutako *Ariane* sistemaren ezaugarriek transferentzian oinarritutako sistemen adibide on bat bihurtu dute. Horregatik, eta gure lanerako izan duen eraginagatik, sistema honen ezaugarriak eta arkitektura laburki deskribatuko ditugu (Vauquois eta Boitet, 1985; Hutchins eta Somers, 1992) lanetan oinarrituz.

Sistemaren diseinua guztiz modularra da eta linguistikan oinarrituta. Fase bakoitzeko ezagutza linguistikoa eta programak guztiz banatuta daude. *Arianer*en modulu bakoitzeko oinarri konputazionala produkzio-sistema bat da: adierazpen linguistikoa duen datu-egitura, erregela-multzoa eta interpreteta. Modulu bakoitzerako badago erregelen idazketa formalismo zehatz bat. Formalismoak eta algoritmoak, hein handi batean, hizkuntzatik independenteak dira. Gainera analisi-modulua xede-hizkuntzatik independentea da, eta sorkuntza-modulua abiapuntuko hizkuntzatik.

Sistemaren arkitektura II.7 irudian ikus dezakegu (Hutchins eta Somers, 1992). Analisia eta sorkuntza modulu morfologiko eta sintaktikoetan banatzen dira, eta transferentziak ere bi urratsetan burutzen da, lexikala eta estrukturala. Datu-egitura maila anitzekoa da, informazio linguistiko azalekoa eta sakona adierazten dituela, mendekotasun-erlazioak eta osagaien egiturak konbinatuz. Erregelen idazketarako lau formalismo erabiltzen dira: *Atef*, *Robra*, *Trans* eta *Sygmor*.

Atef analisi morfologikoan erabiltzen da eta karaktere kateak antolatzen ditu etiketatutako zuhaitz lau batean. *Transek*, transferentzia lexikalean, zuhaitz-egituretako nodoen etiketak aldatzen ditu. *Sygmor*, sorkuntza morfologikoan, etiketatutako zuhaitzak karaktere-kateak bihurtzen ditu egoera finituetako automata determinista bat erabiliz.

Robra da formalismorik ahaltsuena. Maila anitzeko analisisian, transferentzia estrukturalen eta sorkuntza sintaktikoan erabiltzen da. Aukera ematen du zuhaitz-egitura oso konplexuak sortzeko eta zuhaitz-transdukzioak burutzeko, hau da, egitura horiek manipulatz besterik ez batzuk sortzeko. Bertako erregelen ezkerreko aldeak datu-egiturarekin egiaztatatu behar den



Irudia II.7: Ariane sistemaren arkitektura

kidetze-patroia (*pattern match*) agertzen da, non zuhaitz-egitura bat eta baldintza multzo bat adierazten diren. Eskuineko aldean eraiki behar den zuhaitz-egitura berria deskribatzen da, non ondokoak adierazten diren: kopiatu behar den informazioa, gehitu behar den informazio berria eta beste eragiketa konplexuago batzuk.

II.1.4 Corpusetan oinarritutako IA

Corpusetan oinarritutakoak hurbilpen enpirikoak dira, non ezagutza linguistikoa itzulpen-corpusetatik jasotzen den. Sistema hauen garapena posible izan da oso corpus handiak eta baliabide konputazional ahaltsuak eskuragarri egon direnean. Sistema hauek abantaila nabarmen bat dute: hizkuntza pare baterako teknologia garatu ondoren, oso modu erraz eta azkarrean gara daiteke beste hizkuntza-pare baterako, hizkuntza-pare berrirako corpus nahiko handia baldin badago. Corpusetan oinarrituz bi estrategia daude: adibideetan oinarritutako metodoak eta teknika estatistikoak

Adibideetan Oinarritutako itzulpen automatikoa

Adibideetan oinarritutako itzulpen automatikoa (AOIA; *Example Based Machine Translation, EBMT*) burutzen duten sistemek aldeztu aurretik itzulitako adibideen datu-basea edo corpusa erabiltzen dute. Sarrera berri bat itzultzeko sistemak adibideen datu-basearekin parekatzen du eta egokiak diren adibideak erauzten ditu. Ondoren, adibide hauek konbinatzen ditu itzulpen zuzena lortzeko.

AOIAren ideia Makoto Nagaok aurkeztu zuen, *adibideetan oinarritutako inferentziaren bidezko itzulpen automatikoa edo analogia printzipioaren bidezko IA* izendatu zuela (Nagao, 1984):

Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference.

Bere esaldi horretan Nagaok adibideetan oinarritutako estrategiaren hiru ataza nagusiak identifikatzen ditu: esaldi-zatiak adibide errealeko datu-base batekin parekatzea, hauei dagozkien itzulpen-zatiak identifikatzea eta zati hauek birkonbinatzea itzulitako testua emateko.

Ondoko adibidean (Sato eta Nagao, 1990), (1) esaldiaren itzulpena lortzeko (2) eta (3) esaldietatik zati egokiak hartzen dira, emaitza (4) esaldia izanik:

- (1) He buys a book on international politics.
- (2) He buys a notebook. /
Kare wa noto o kau.
- (3) I read a book on international politics. /
 Watashi wa kokusai seiji nitsuite kakareta hon o yomu.
- (3) Kare wa kokusai seiji nitsuite kakareta hon o kau.

AOIAren oinarria corpus elebidun parekatua da, hau da, testu bat bere itzulpenarekin, non bi testuak prozesatuak izan diren bi hizkuntzetako segmentu berdina erlazionatzeko. Inplementazio gehienetan parekatutako segmentua esaldia da. Hala ere, AOIAren ahalmenaren funtsa esaldia baino txikiagoak diren testu zatien ustiapenean dago (Cranias *et al.*, 1994).

Zenbat eta adibide gehiago egon datu-basean, orduan eta emaitza hobeak lortzen dira orokorrean, baina badago muga bat, eta zenbaitetan sistemak txarrera egiten dute adibideak gehitzean (Sumita eta Iida, 1991). Arrazoa da adibide berdinek itzulpen desberdinak dituztenean gatazka sortzen dela.

Adibideak eraginkorrak izateko zenbait sistematan antzeko adibideak konbinatzen dira eta adibide *orokortu* bakar gisa gordetzen dira. Brown-ek (2000) adibideak tokenizatzen ditu baliokidetzak klaseak (pertsona izenak, herrien izenak, datak...) erakusteko, adibideetako zatien ordezkari klase horiek jarritz.

Itzuli beharreko abiapuntu-hizkuntzako testua hartu eta hobekien kide-tzen (*matching*) den adibidea (edo adibide multzoa) aurkitzeko distantzia edo antzekotasun-neurri bat erabiltzen da. Adibideak gordeta dauden moduan arabera aurre egiten zaio bilaketarako estrategiari: karaktere-kateetan, hitzetan, etiketatutako hitzetan edo zuhaitz-egituretan oinarrituta.

Behin adibide multzo bat bere itzulpenekin eskuratu dela, prozesuko urratsik zailenak datoz: hizkuntzen arteko parekatzea eta itzulpen-zatien birkonbinatzea.

Hizkuntzen arteko parekatzeko erlazionatutako itzulpenetako zein zatia jatorrizko testuko zein zatiei dagozkien identifikatzen da. Horretarako hiztegi elebiduna, sinonimoen zerrendak eta analizatzaile morfologikoak erabili daitezke, *Pangloss* sisteman bezala (Brown, 2000; Nirenburg *et al.*, 1994; Brown, 1996).

Azkenik identifikatutako zatia modu egokian birkonbinatzen dira irteera gramatikal bat emateko.

Adibideetan oinarritutako metodoak erabiltzen hasi zirenean, aldarrikatzen ziren abantaila guztiak lortu ez badira ere, badirudi ondokoak konprobatu direla (Somers, 1999):

- Adibideak benetako hizkuntzakoak direnez hizkuntza errealean erabiltzen diren egiturak estaltzen ditu, erabiltzen ez direnak baztertuz.
- Ezagutza linguistikoa errazago aberats daiteke, soilik adibide gehiago gehitzen.
- Gramatika konplexurik ez dagoenez gutxiagotzen dira erregelen gatazkak eta interakzioak.
- Egitura-mantentzearen arazoa saihesten da.
- Hizkuntza bikote berri baterako AOIA sistema bat azkar garatzea posible izan daiteke, soilik parekatutako corpus elebidunean oinarrituta.

Hau oso interesgarria izanik analizatzailea edo hiztegiak ez dituen hizkuntza baterako.

Adibideetan oinarritutako estrategia erregeletan oinarritutakoa baino egokiagoa izan daiteke itzulpen-erregelen osaketa zaila denean, erregela orokorrak fenomeno bat zehazki deskribatu ezin duenean, kasu berezia adierazten duelako edo itzulpena konposizionalki burutu ezin denean (Sumita eta Iida, 1991).

EBMT is certainly here to stay, not as a rival to rule-based methods but as an alternative, available to enhance and, sometimes, replace it. (Somers, 1999)

Hori bai, zalantzan jarri izan da AOIAk itzulpen-prozesu osoa burutu zezakeenik, eta zentzu horretan saiakera gutxi egin dira.

Itzulpen automatiko estatistikoa

Gaur egun itzulpen automatiko estatistikoa (IAE; *Statistical Machine Translation, SMT*) gehien ikertzen ari den IArako estrategia da, horren erakusgarria izanik azken urteetan teknika estatistikoei buruzko argitalpen kopuruaren igoera izugarria.

Hala ere, datu estatistikoak erabiltzea ez da ideia berria, IAren ikerketaren hasierako garaian proposatu baitzen, baina ez zuen arrakastarik izan garai haietako ordenagailuek ez zutelako hurbilpen honen kalkuluak egiteko behar den ahalmen konputazionalik. 80.eko hamarkadan, berriz, hizketaren prozesamenduaren arloan teknika estatistikoen arrakastaren eraginez, IBMko taldeak *Candide* sistema eraiki zuen (Brown *et al.*, 1990), ikerlariak harritu zituzten emaitza onak lortu zirela. Ingeleseztan eta frantseseztan Kanadako Legebiltzarreko debateak jasotzen dituen *Hansard* corpus handia erabili zuten, hizkuntza bakoitzean 800.000 hitz dituen.

IA estatistikoaren ideia hauxe da: ordenagailuak automatikoki testu elebidun paralelo handietatik ateratako datu estatistikoak erabilia, itzulpenak egiten ditu, hiztegi elebidunik edo informazio linguistiko espliziturik erabili gabe.

IAEren funtsa corpus elebidun handiak esaldika parekatzea da, gero itzulpen-ereduko eta hizkuntza-ereduko probabilitateak erauztea eta bukaeraren probabilitate horiek erabilia itzulpenak lortzea xede-hizkuntzaren eredu oinarritzat hartuta.

Ezagutzaren iturri elebakarrak (hizkuntza-eredua) eta elebidunak (itzulpen-eredua) banatzen dira eta itzulpenaren probabilitatea emateko konbinatzen dira (Brown *et al.*, 1990):

- Hizkuntza-eredua. XHko kate bat XHko baliozko esaldi bat izateko probabilitateak ematen ditu. Hitzetako bigrama edo trigrametan oinarritzen da.
- Itzulpen-eredua. XHko kate bat AHko kate baten itzulpena izateko probabilitateak ematen ditu. Parekatze estatistikoen eredueta oinarritzen da, eta hainbat informazio erabiltzen ditu: abiapuntu- eta xede-hitzen aldibereko agerpen maiztasuna, hitz horiek azaltzen diren esaldien luzera, hitzen esaldi barruko posizioa... Probabilitate guztien kalkulua esaldika parekatutako corpusaren gainean egiten da.

IA estatistikoaren oinarriak eta egungo aurrerapenak ederki laburbiltzen dira (Och, 2005) lanean. IAEan ondoko ataza nagusiak burutzen dira:

1. Corpusaren bilketa eta prestaketa. Corpora bildu eta prestatu behar da: zenbaitetan aurreprozesaketa, tokenizazioa eta normalizazioa burutu behar dira, corpus elebiduneko fitxategiak parekatu behar dira eta, azkenik, esaldi edo chunken⁵ parekatzea burutu behar da.

Ikertzaileek corpus handiagoek emaitza hobeagoak ematen dituztela baieztatzen dute.

How much data is needed? Answer: MORE. (Och, 2005)

Emaitza optimoak lortzeko ereduak ahalik eta datu kopuru handienarekin entrenatu behar dira: ahal izanez gero, hizkuntza-ereduak ehun bilioika hitzekin eta itzulpen-ereduak ehun milioika hitzekin (Och, 2006). Honetarako, noski, oso baliabide konputazional handiak behar dira.

2. Modelatzea. Abiapuntu-hizkuntzako esaldi bat (f) itzultzeko, hurrengo formula maximizatzen duen xede-hizkuntzako esaldia (e) bilatu behar da:

$$\hat{e}(f) = \operatorname{argmax}_e \operatorname{Pr}(e|f).$$

$\operatorname{Pr}(e|f)$ formulak e XHko esaldia f AHko esaldiaren itzulpena izateko probabilitate-banaketa adierazten du. Baina probabilitate-banaketa hori ezezaguna denez, hurbiltzen duen eredu bat kalkulatzeko da:

$$\hat{e}(f) = \operatorname{argmax}_e \operatorname{Pr}(e|f) \approx \operatorname{argmax}_e p(f|e) * p(e),$$

non $p(f|e)$ itzulpen-eredua den eta $p(e)$ hizkuntza-eredua

⁵Chunkak sintagmak edo sasi-sintagmak dira

3. Entrenamendua (edo parametroen estimazioa). Corpora emanda eredu-parametroei balio zehatzak esleitzen zaizkie.
4. Bilaketa (edo deskodeketa). Hauxe da, zehazki, itzulpen-prozesua, non abiapuntu-eraldiarentzat itzulpen hoberena aurkitzeko, itzulpen-eredu eta hizkuntza-eredu probabilitateak maximizatzen dituen XHko katea bilatzen den. Bilaketa-espazioa oso handia denez bilaketa-algoritmoetan heuristikoak erabiltzen dira. Itzulpen-probabilitateak corpus oso handi batetik erauzten direnez, AHko hitz edo sintagma-rako XHko parekideak emateko aukera asko izango dituzistemak.

IA estatistikoaren helburua erabaki optimoak egitea da, hau da, abiapuntu-hizkuntzako esaldi bat emanda, itzulpen onena zein den erabakitzea, hurbilpen estatistikoa ondoko ekuazioarekin adieraz daitekeela (Ney, 2001):

$$IA = \text{Modelatze linguistikoa} + \text{Erabaki-teoria estatistikoa}$$

Hasieran teknika estatistikoak hitzetan oinarritzen baziren ere, gaur egun, nagusiki IAEa n-grametan oinarrituta (*phrase-based*⁶) egiten da. N-gramek berrordenatze lokala harrapatzen dute eta, gainera, testuinguru lokala desanbiguaziorako erabilgarria da. Corpus oso handiekin, batzuetan, esaldi osoak ere harrapa daitezke. N-grametan oinarritutako hurbilpenaren arazoa n-grama posible guztiak biltegitratzeko espazio itzelaren beharra da. Sintaxian oinarritutako sistema estatistikoekin zenbait ikerketa ere burutzen ari dira.

Corpus on eta handi bat eskuragarri baldin badago, IA estatistikoa hurbilpen azkarra eta merkea da, hizkuntatik independenteak diren datu estatistikoen erauzketa eta IAKo prozesuak automatikoki egiten direlako.

Hori bai, corpora ez bada oso handia, eta eremu bakar bateko testuak soilik baditu, sistema eremu horretarako espezifikoak izango da, hau da, entrenatu ez den beste eremu bateko testuak itzultzean emaitza okerragoak emango ditu.

IAE hutsa inplementatzen zuten hasierako sistemen emaitzak ez ziren oso onak. Baina erakutsi izan da informazio linguistiko sinplea sartuz gero sistemaren doitasuna hobetzen dela eta, gaur egun, n-grametan oinarritutako teknikak erabiliz eta corpus handigoak baliatuz, metodo estatistikoen emaitzak asko hobetu dira.

Baina, gaur egun, eskuragarri dauden corpus elebidun paraleloen kalitatea eta tamaina ez dira behar direnerako nahikoak.

⁶Metodo estatistikoetako sintagma (*phrase*) ez dator bat sintagma linguistikoarekin, sintagma edozein n-grama (hitz-sekuentzia) izan daitekeela. Horregatik guk sintagma hitza erabili beharrean n-grama erabiliko dugu

In our experience, a lack of well aligned parallel text is the norm, rather than the exception. (Bond *et al.*, 2005)

Zenbait hizkuntzentzat corpus elebidun erraldoiak lor daitezke (adibidez arabiera-ingelesa eta txinera-ingelesa) eta horiek prozesatzeko azpiegitura sofistikatuak behar dira. Hain zuzen, IAEren erronkarik handiena datu kopuru erraldoiak biltzea eta prozesatzea da,

Beste erronka bat *bigarren ordenako sormena* deiturikoa da (*second-order creativity*), hau da, corpusean ez dauden itzulpenak sortzeko edo corpusean dauden itzulpen posibleak egokiak ez direla antzemateko ahalmena (Marcu eta Melby, 2006).

Hobekuntza esanguratsua ekar dezakete eredu estatistikoetan ezagutza linguistikoa erabiltzea (WordNet, Tree Bank paraleloak...) eta eredu estatistiko konplexuagoak garatzea (Och eta Ney, 2001).

Garrantzi handikoa da prozesaketa morfologikoaren erabilera, batez ere, morfologikoki aberatsak diren hizkuntzetan, baina ingelesera itzultzen duten sistemen eraikuntzan izandako nagusitasunak baztertu du morfologia aberats sortzearen problema konplexuen azterketa, non ez den soilik testuinguru lokala kontutan hartu behar. (Koehn eta Monz, 2006) laneko esperimentuan frogatu dute sistema estatistikoak erregeletan oinarritutako sistema komertzialen atzean gelditzen direla morfologia aberatsa duten hizkuntzetara itzultzean.

Paradigmen antzekotasunak eta desberdintasunak

Ondoko lerrootan paradigma desberdinen arteko desberdintasunak eta antzekotasunak azaltzen ditugu (Hutchins, 2006b; Och, 2005):

- IA estatistikoa *kutxa beltz bat da*, ezin da jakin sistemak nola egiten duen lan kasu partikularretan, zergatik batzuetan duen arrakasta eta besteetan ez. Erregeletan eta adibideetan oinarritutako estrategietan, berriz, erregelak eta prozedurak azter eta alda daitezke. Hori bai, aldaketek ustekabeko ondorioak ekar ditzakete.
- Erregeletan oinarritutako teknikak eta metodo estatistikoak elkarren aurkako hurbilpenak diruditen arren, gero eta maizago gehitzen ari dira erregelak sistema estatistikoetan: hasieran morfologia, gero n-gramak eta orain ere zuhaitz sintaktikoak.
- Adibideetan oinarritutako IAn adibideak zuhaitz-egituretan gorde eta tratatzen badira, bere prozesuak eta erregeletan oinarritutakoenak an-

	EOIA	AOIA	IAE
Aurre-prozesaketa (Datu-basea)	Hiztegiak Gramatikak	Parekatzea Analisia	Parekatzea (Analisi estatistikoa)
Analisia	Morfologia Analisi sakona Analisi semantikoa	Morfologia Azaleko analisia	(Morfologia)
Interfazea	Adierazpen egituratua	Patroiak (Zuhaitzak)	Hitzak/N-gramak
Transferentzia (muina)	Transfer. lexikala Zuhaitzen transdukzioa	Kidetzea (Zuhaitzen bihurtzea) Erauztea	Kidetzea Itzulpen-eredua
Interfazea	Adierazpen egituratua	Kateak (Zuhaitzak)	Kateak
Sintesia	Sorkuntza	Birkonbinazioa	Hizkuntza-eredua

Taula II.1: EOIA, AOIA eta IAE paradigmnen oinarritzko prozesuak

tzekoak bihurtzen dira. Erregeletan oinarritutako IAko erregelak ere datu-base elebidunetatik erator daitezke.

- IA estatistikoa eta adibideetan oinarritutakoa hurbilpen enpirikoak badira ere, AOIAk garrantzia ematen dio adibideetatik ikasteari, eta IAEk, berriz, erabaki hoberen egiteari.

Hasierako IA estatistikoa hitzetan oinarritzen zen eta adibideetan oinarritutakoa, berriz, hitz multzoetan. Baina n-grametan oinarritutako IA estatistikoa (*phrase based MT*) burutzen denean, desberdintasun hau desagertzen da. AOIA eta IAE antzekoagoak dira corpusa eta itzuli beharreko sarrera analizatzen direnean. Syntaxian oinarritzen direnean baliokidetzat hartzen dira.

Ikertzaile askok adibideetan oinarritutako hurbilpena IA estatistikoa-ren barruan ikusten dute gaur egun.

II.1 taulan paradigma desberdinen oinarritzko prozesuak erkatzen dira, (Hutchins, 2006b) lanean oinarrituta. Lan horretan ere Hutchins-ek baieztatzen du estrategia desberdinen emaitzen arteko konparazioak egitea zaila dela, neurtzeko metodoak desberdinak direlako, baina adostasun orokorra dagoela *intuitiboki* erregeletan oinarritutako IAren emaitzak doiagoak eta ulergarriagoak direla.

II.1.5 Hibridazioa

AMTA 2006 kongresuan itzulpen automatiko hibridoari buruzko panelaren aurkezpenak honela zioen:

In recent years, statistical machine translation has made great strides, example-based approaches have pushed forward, but systems based on linguistic knowledge have not been abandoned. In fact, we have seen the pendulum swing back a little towards the latter through the inclusion of linguistic knowledge in statistical systems and vice versa. Much current research in machine translation is neither based purely on linguistic knowledge nor on statistics, but includes some degree of hybridization. (Cavalli-Sforza eta Lavie, 2006)

Argi dago estrategia bakoitzak bere abantailak dituela eta hibridazioa continuum bat bezala ikus daitekeela, non erregeletan edo corpusetan oinarritutako osagaiak modu desberdinetan nahas daitezkeen (Habash, 2006).

Beraz, sistema hibridoek estrategia bakoitzeko metodo onenak har ditzakete (Hutchins, 2006b):

- EOIA analisirako (morfologia, mendekotasunak).
- AOIA esamoldeetarako (*idioms*), kolokazioetarako eta egitura ez-konposizionaletarako.
- EOIA eta AOIA egitura elebidunen dibergentzietarako.
- IAE maiztasun handieneko XHko formen aukeraketarako eta hizkuntza-eredua XHko irteera ebazteko.
- Ezagutzan oinarritutako IA arazo semantiko eta erreferentzietarako.

Eisele-ek (2006) aipatzen duen bezala, estrategia desberdinak integra daitezke haien abantailak konbinatzeko, baina atazaren, hizkuntzen, eremuaren edo testu motaren propietate zehatzek eragina handia izango dute hurbilpen onena erabakitzeko.

Erregeletan eta corpusetan oinarritutako estrategiak konbinatzean arretaz aritu behar da konbinazio horretatik jaso daitekeelako *bi mundu horietako onena edo... bi mundu horietako txarrena* (Habash, 2006). Estrategiak konbinatzean ondoko erronkei egin behar zaie aurre: sintagma linguistikoa eta metodo estatistikoetako n-grama (*phrase*) ez dute zertan bat etorri behar, lotzen diren estrategia desberdinetako osagaiak adierazpide desberdinak erabil ditzakete, hibridatzean sistemaren konplexutasuna gehiagotzen da...

Aukerarik sinpleena kutxa-beltzeko motor anitzak konbinatzea izan daiteke, baina badirudi integrazio finagoek emaitza hobeagoak izan ditzaketela,

lan gehiago eskatzen badute ere. Ondoko lerrootan hibridaziorako zenbait aukera aurkezten ditugu.

Erregeletan oinarritutako IA eta metodo estatistikoak

Erregeletan oinarritutako sistema bat sistema estatistiko batekin bi modutan osa daiteke:

- Sistema nagusiak emandako hainbat hipotesiren artean aukeraketa egiteko.
- Babes-sistemaren papera jokatzuz, sistema nagusiak itzultzen lortzen ez duenerako.

Gainera erregeletan oinarritutako sistemen baliabide linguistikoen erakuntzan teknika estatistikoak erabil daitezke (Seneff, 2006):

- Itzulpen-lexikoiak eraikitzeke parekatze estatistikoa erabili.
- Gramatika eta lexikoien garapenerako IA estatistikoko tresna interaktiboak erabili.

Systran erregeletan oinarritutako sistemak, adibidez, teknika estatistikoak erabiltzen hasi da hiru modutan: corpus parekatutik terminologia-erazketa masiboan, adiera-hautapenerako hizkuntza-ereduaren erabileran, eta erregelak eta estatistiken integrazioan (Senellart eta White, 2006).

IA estatistikoa eta ezagutza linguistikoa

Hasierako IA estatistiko hutsaren emaitza eskasak ikusirik, laster hasi zen datu estatistikoez gain, ezagutza linguistikoa erabiltzen, emaitzak hobetuz. Euskara bezalako hizkuntza baterako, adibidez, ezagutza morfologikoa erabiltzea nahitaezkoa da.

Bestalde, sistema estatistikoaren deskodetzailea erregeletan oinarritutako sistema baten emaitzarekin ere elika daiteke (Eisele, 2006).

Erregeletan oinarritutako IA eta adibideetan oinarritutakoa

AOIA erregeletan oinarritutako itzulpen automatikoa burutzen duten sistemen lagungarri gisa erabili izan da hainbat modutan:

- AOIA kasu bereziatarako, erregeletan oinarritutako hurbilpena zaila suertatzen denean. Adibidez, erregeletan oinarritutako *ATR* sistemak

(Sumita eta Iida, 1991) adibideetan oinarritutako modulua erabiltzen du japonierazko izen-sintagma konplexuen itzulpenerako. Japonieraz *N1 no N2* motako izen-sintagmak normalean ingelesezko *N2 of N1* motakoei dagozkie, baina salbuespen asko daude non beste preposizio batzuekin itzultzen den (*for, in, 's*).

- Adibideetan oinarritutako transferentzia. Abiapuntu-hizkuntzako sarrerak analizatzen dira modu tradizionalen, transferentzia-modulua adibideetan oinarritzen da, eta ondoren xede-hizkuntzako irteera modu klasikoan sortzen da.

Motor-anitzeko sistemak

Zenbait ikerketek hibridazioaren bidea jarraitu beharrean, motor-anitzeko sistemak eraikitzeari ekin diote, Horretarako estrategia desberdinetako sistemek emandako irteerak konbinatzen dira hizkuntza-eredu estatistiko bat erabiliz.

Adibidez *Pangloss* sisteman (Brown, 1996; Nirenburg *et al.*, 1994) hiru modulu paraleloan egiten dute lan: adibideetan oinarritutakoa, ezagutzen oinarritutakoa eta erregeletan oinarritutakoa. Modulu bakoitzak sarrerako zati ezberdinentzat itzulpen posibleak ematen ditu, diagrama (*chart*) batean konbinatuz. Ondoren eredu estatistiko batek diagrama horretako biderik hoberena zein den erabakitzen du eta hori izango da sarrerako esaldiaren itzulpena adieraziko duen irteera.

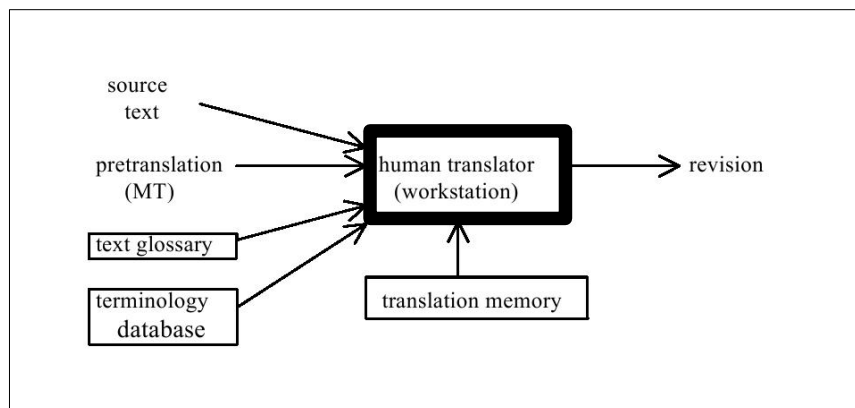
Hala ere, (Och eta Ney, 2001) lanean aipatzen da IAKo estrategia ezberdinen irteera konbinatzeak ez duela hobekuntza handirik lortzen.

Itzulpen automatikoari aurre egiteko estrategia desberdinak aurkeztu ditugu, haien ezaugarriak azalduz, eta hibridazioaren bidea erakutsiz.

Erabat automatiko den itzulpena egiteko ageri diren mugak direla-eta, itzulpenetan lagungarri izateko tresnak garatzea izan da indartu den arlo garrantzitsu bat. Ondoko atalean aurkezten ditugu itzultzaileentzako tresna horiek.

II.1.6 Itzultzaileentzako tresnak

Itzulpen automatiko osoaren kalitatea oso ona ez denean, ez da egokia itzultzaile profesionalentzat, haien gogoia ez baita makina batek ematen dituen



Irudia II.8: Ordenagailuek lagundutako giza-itzulpena

emaitza traketsak zuzentzen aritzea. Itzultzaile profesionalek eta itzulpen-agentziek, itzulpen automatiko hutsa baino, haien lanean laguntza emango dieten tresnak behar dituzte.

Ordenagailuek lagundutako giza-itzulpenean (*Machine Aided Human Translation, MAHT*), funtsa ez da IAKo sistema bat, giza-itzultzailea baizik, bere lan-estazioan eskura izango dituela ordenagailuak emandako tresnak eta baliabideak, II.8 irudian (Hutchins, 2005a) ikusten dugun bezala.

Itzultzailearentzat bereziki prestatutako lan-estazioek (*translator workstation*) ondoko itzulpenerako laguntza-tresnak (*translation support tools*) eskaintzen dituzte:

- Testu-edizio eleanitza errazten duen hainbat leiho era koordinatuan kudeatzen dituen interfazea.
- Hiztegi elektronikoa: sarrera lexikalen hiztegi elebidun edo eleanitzak, kontsultarako mekanismoarekin.
- Terminologia kudeatzeko sistema: datu-base terminologiko eleanitza sortzeko, mantentzeko eta bilaketak burutzeko softwarea.
- Corpus elebidunak parekatzeko tresnak.
- Itzulpen-memoretan (IM, *Translation Memory, TM*) oinarritutako sistema: aldeztatik itzulitako testuen bildumak sortzeko eta erabiltzeko softwarea, parekatzeko, kontsultarako eta bilketaren kudeaketarako aukerarekin.

- Itzulpen automatiko osoko sistemak, nahi izanez gero, eta nahi den zatian, erabiltzeko.

Itzulpen-memoriak

Itzulpen-memoria itzulitako testuen bildumari zein corpus hori kudeatzen duen sistemari esaten zaio

Itzulpen-memoriek⁷ aurretik itzulita dauden testuak datu-base batean gordetzeko aukera ematen dutenez, funtsezkoa corpus parekatuen (bitestuak ere deituta) kudeaketa da. Testu bat itzultzeko, memorian dauden adibideetan zati berdina edo antzekoa bilatu eta itzultzaileari proposatzen dio. Hau da, esaldi berri baten itzulpenari ekiterakoan, sistemak memorian bilaketa azkarra egiten du, eta bertan esaldi hori edo antzekorik lehendik itzulita dagoen bilatzen du. Sistemak aurkitzen dituen testu-zatien artean itzultzaileak hautatu edo osatzen ditu, itzulpen egokia emateko. Horrek aurretik itzulitako testu-puska handiak berriz itzuli behar izatea saihesten du.

Itzulpen-memoriak, askotan, adibideetan oinarritutako IArekin lotu dira. Bien ideia nagusia aurretik egindako itzulpenen adibideak berrerabiltzea bada ere, IMA giza-itzultzaileentzako tresna interaktiboa da eta AOIA, berriz, itzulpen automatikoko teknika. Esaldi berri bat itzultzeko IMak aurkitutako adibideak eskaintzen dizkio erabiltzaileari, berak erabaki dezan haiekin zer egin; hauxe adibideetan oinarritutako IArekin hasierako urratsa besterik ez da.

Hain zuzen ere, hauxe da IMen ahulezi nagusia, ez duela aurkitutako testuak konbinatzen laguntzen. Gainera, IMek zailtasunak izaten dituzte *fuzzy matching* deitutakoarekin: adibide desegoki asko proposatzen dituzte, edo adibide erabilgarri asko ez ditu aurkitzen. Adibide erredundanteek ere arazoak ematen dituzte.

Itzultzaileentzako lan-estazioen etorkizuna

Gaur egun, gizakiak lagundutako itzulpen automatikoa (*HAMT*) eta konputagailuek lagundutako giza-itzulpena (*MAHT*) bateratzen ari dira. Lan-estazioek IArako osagaiak eskaintzen dituzte, eta IAKo sistemek itzulpen-memoriak erabiltzen dituzte. Etorkizunean bi hurbilpenen integrazioa emango omen da (Hutchins, 2006a).

⁷Itzulpen-memoriak kudeatzen dituzten pakete asko daude erabilgarri. Zaharrenak eta finkatuena Trados enpresaren *Translation Workbench*, STAR-en *Transit*, Atril-en *Déjà Vu* eta Logos-en *Wordfast* dira. Baina beste enpresa asko lehian ere sartu dira: SDL (*SDLX*), Multilizer, Terminotix (*LogiTerm*), MultiCorpora (*MultiTrans*) IBM (*TranslationManager*), Eurolang (*Optimizer*)...

Etorkizuneko itzulpen-tresna ondoko ezaugarriak beteko dituen sistema elkarreragile mota berri bat izango da (Macklovitch eta Valderrábanos, 2001): giza-itzultzaileek gidatua izango da, ikasketa automatikoko teknikak erabiliko ditu aldez aurretik itzulitako testuei etekina hobeto ateratzeko, kalitatearen berme eta kontrolerako tresnak izango ditu (ortografia, gramatika eta estilo zuzentzaileak, terminologiaren koherentzia egiaztatzeko tresnak...). eta aukera emango du IAko sistemak erabiltzeko, giza-itzultzaileen gainbegiraketarekin. Sistema horiek itzultzailearen jardunetik ikasteko gai izan beharko dira, bere beharretara egokituz.

II.1.7 Gaur egungo erronkak

Aurkezten ari garen itzulpen automatikoaren panorama orokor honi bukaera eman nahi diogu, gelditu gabe aurreratzen ari den arlo honetako erronkak aipatuz.

Egungo IAren ikerketan lerro nagusi hauek azpimarra daitezke (Hutchins, 2005b, 2006b):

- Erregeletan oinarritutako sistema klasikoen hobekuntza, corpusetan oinarritutako hurbilpenetan (adibideetan oinarritutakoak eta estatistikoak) ikerkuntza eta estrategia desberdinak erabiltzen dituzten sistema hibridoaren eraikuntza.
- Teknika berrien (sare neuronalak, prozesaketa paraleloa...) eta sistema elkarreragileetan ikerketa.
- Hizketaren itzulpen automatikoa, ikerketa arlorik berritzaileena dena.
- Eremu zehatzetako eta testu informalak (mezu elektronikoak, chat-en elkarizketak...) itzultzeko sistemen garapena.
- Hizkuntza gutxituetarako sistemen garapena.
- IA eta itzulpen-memorien integrazioa.
- Lexikoien eskuratzekoak suposatzen duen itoguneari aurre egitea, azkenaldian bildu diren corpus eta baliabide lexikal erraldoiak erabiliz, eta Interneten erabilera baliabide nagusiena bezala.
- IAko osagaien berrerabilpena.

Etorkizunean itzulpen automatikoa sistema orokorretan integratuko da, itzultzaileentzako tresnekin batera eta informazio-bilaketarako sistemekin.

Internet-zerbitzuetako erabiltzaileak informazioa bilatzen du, edozein hizkuntzatan dagoela (hemen itzulpena ez da xedea, tresna baizik) eta hori lortzeko hainbat teknologia modu gardenean bildu beharko dira: dokumentuen berreskurapena (*Information Retrieval, IR*), informazio-erazketa (*Information Extraction, IE*), laburpen-sistemak (*Summarization*) eta itzulpen automatikoa.

2001. urtean ospatu zen *Towards a Road Map for MT* workshop-ean itzulpen automatikoaren etorkizuneko bilakaeran hainbat norabide seinalatu ziren (Gerber, 2001; Schütz, 2001; Boitet, 2001; Farwell eta Helmreich, 2001):

- Itzulpen-sistemen sarreran murrizpenak jartzea, emaitza arrakastatsuak bideratzen dituztelako: hizkuntz kontrolatua erabiltzea, azpihizkuntza batean bereiztea edo testu teknikoak itzultzea komunikazioaren testuinguruko alderdiak neurri txikian agertzen direlako.
- Hiztegien eta sarreraren formatua XML izatea, espezifikazio egokia ematen duelako eta berrerabilpenerako.
- Iturburu irekiko (*open source*) baliabide linguistikoen garapen kooperatiboa, enpresa eta instituzio handiek ez dutelako interes handirik gaur egungo IA sistemetan sartuta ez dauden hizkuntz gehienetan.
- Ikasketa-metodoak erabiltzea, bai erregelak zein adibideetako patroiak erazteko.
- Forma linguistikoen manipulazio soila alde batera uztea, eta testuaren egitura osoa, testuinguruaren informazioa eta hizlariaren asmoak kontutan hartzea, diskurtsoa baztertu gabe.
- IA guztiz abandonatzea eta sorkuntza eleanitza zabaltzea, UNL (*Universal Networking Language*) bezalako hizkuntzaren adierazpide formal batean oinarrituta.

Itzulpen automatikoaren laburpen orokor honen ondoren guri bereziki interesatzen zaizkigun bi alderdi aztertuko ditugu. Alde batetik euskara barne hartzen duten itzulpen automatikoaren arloko ikerketa eta, bestetik, kode irekiko itzultzaile automatikoena.

II.2 Itzulpen automatikoa eta euskara

Txapela buruan ta ibili munduan.

Herrikoia

Euskararekin itzulpen automatikoaren arloan orain dela gutxi arte oso ikerketa urriak burutu badira ere, azkeneko urteetan zenbait saiakera abian jarri dira, haien berri ondoko ataletan ematen dugula.

II.2.1 Erregeletan oinarritutako sistemak

Matxin sistema

Tesi-lan honetan aurkezten dugun transferentzia sakoneko sistemaren lehen-go prototipoak ingelesetik euskarara izen- eta preposizio-sintagmak itzultzen zituen (Díaz de Ilarraza *et al.*, 1999). Gaur egun, espainieratik euskarara esaldi osoak itzultzen dituen prototipoa (Alegria *et al.*, 2007) euskararekin lan egiten duen lehenengo IAKo sistema erabilgarria da. Sistema Interneten probatu daiteke⁸, eta kode irekiko software libre bezala banatzen da⁹.

ATS-euskara prototipoa

AutomaticTrans (AT) enpresak¹⁰, Eusko Jaurlaritzaren finantziazioarekin, espainieratik euskarara itzultzeko sistema baten bideragarritasun-azterketa egin zuen (AutomaticTrans, 2003; Abaitua, 2003) baina proiektua bertan behera gelditu zen.

AT enpresaren ATS teknologiak espainiera eta katalanen arteko hizkuntz hurbilak itzultzeko gaitasuna erakutsi du¹¹. Itzulpen-memoriak eta erregelak uztartzen ditu. Erregela guztiak aplikatzen dira eta itzulpen posibleen artean onena hautatzen da ponderazio-sistema bat erabiliz. Patroiak ere erabiltzen ditu, non hitzak aldagaiekin ordeztzen diren.

Baina ATS teknologia hurbilak diren hizkuntzentzat aproposa izan arren, zailtasun handiak izan ditzake urrun dauden bi hizkuntzen arteko itzulpena automatizatzeko.

⁸<http://www.opentrad.org/demo>

⁹<http://matxin.sourceforge.net>

¹⁰<http://www.automatictrans.es>

¹¹*El Periódico* egunkariak ATS sistema erabiltzen du eguneroko edizioaren itzulpena egiteko.

II.2.2 Corpusetan oinarritutako ikerketa-lanak

PRHLT eta PRST taldeak

PRHLT¹² ikerketa taldeak eremu zehatzetarako (txosten teknikoak, hotelen zerbitzuak...) hizketaren zein testuaren itzulpen automatikoaren inguruko ikerketa burutzen du. IAko sistemen oinarria egoera finituetako transduktoreak (*Finite-State Transducers, FST*) eta transduktore estatistikoak dira, eredu horiek itzulpenen corpusetatik automatikoki ikasten direlarik. PRHLT taldeak garatutako tekniketaz baliatuta eta PRST¹³ taldearekin batera, euskararekin lan egiteko saiakerak egin dituzte.

(Pérez *et al.*, 2005) lanean espainieratik euskarara hizketaren itzulpenarako abiatutako ikerketa aurkezten dute, *Euskal Turista* (ET) deituriko corpora erabiliz. Emaizak espainieratik ingelesera egindako experimentuekin konparatu ondoren, euskararako ataza zailagoa dela aitortzen dute flexio handiagatik eta hitzen ordenagatik. Horregatik itzulpen-eredua hobetzeko parekatze-metodoa aldatzea eta informazio linguistikoa gehitzea nahitaezkoa ikusten dute.

(Pérez *et al.*, 2006) lanean, berriz, IAko sistema estatistiko baterako egokia den eguraldi-iragarpenen eremuko corpus elebidun baten garapena aurkezten dute, corpora etiketa morfosintaktikoekin (hitzen lehen eta post-posizio-informazioa) aberastu behar izan dutela espainiera eta euskararen arteko itzulpenaren zailtasunengatik, Saiakera honetako emaitzak ere eskasak direla aitortzen dute.

Beste esperimentu batean (Nevado *et al.*, 2004) IA estatistikoan oinarritutako bi segmentazio teknika automatiko elebidun desberdin aplikatzean, eu-es bikoterako ataza, ca-es eta en-es bikoteetarako baino askoz zailagoa dela ondorioztatu dute.

Matrex sistema

NCLT-MT¹⁴ taldeak garatutako *Matrex* datuek gidatutako IAko (*Data-Driven MT*) sistema hibridoak adibideetan oinarritutako IA zein teknika esta-

¹²Pattern Recognition and Human Language Technology. Universidad Politécnica de Valencia-ko (UPV) Departamento de Sistemas Informáticos y Computación (DSIC) saileko ikerlariek, Instituto Tecnológico de Informática (ITI) institutukoek eta Departament d'Informàtica de Sistemes i Computadors (DISCA) sailekoek osatuta. <http://www.iti.upv.es/groups/prhlt>

¹³Pattern Recognition and Speech Technology. EHUko Elektriika eta Elektronika Saila <http://grah.ehu.es/PRST>

¹⁴Machine Translation at the National Centre for Language Technology. Dublin City University, Ireland. <http://www.nclt.dcu.ie/mt>

tistikoak erabiltzen ditu. Sistemaren modulu garrantzitsuenak hauek dira: hitzen parekatzailea (*GIZA++*), chunkerra, chunk-parekatzailea eta deskodetzailea (*decoder*). Deskodetzaileak itzulpen bat bilatzen du jatorrizko corpus lerrokatuaz gain eraikitako hitz eta chunk lerrokatuak erabiliz.

IXA taldearekin elkarlanean *Matrex* sistema ingelesetik euskarara itzultzeko egokitu da (Way *et al.*, 2006). Euskararen chunken ordena librea tratatu ahal izateko. euskararako chunkerrak IXA taldean garatutako *EusMG* analizatzailea erabiltzen du, *EusMGk* emandako morfemen informazioa string arruntekin ordezkatu eta egokitzen direla ingeleserako erabilitako chunkerraren emaitzarekin konparagarria izan ahal izateko.

Erabili den corpora software eskuliburuen itzulpen-memoria batetik erauzitakoa da, parekatutako 270.000 itzulpen-esaldiekin (3,2 miloi hitz ingelesez eta 2'9 miloi euskaraz). Egokitutako sistema hitzetan oinarritutako (*word-based*) sistema estatistiko batekin eta n-grametan oinarritutako (*phrase-based*) beste batekin konparatu dute, *Matrexek* emaitza onenak lortuz.

Gaur egun, sistema espainieratik euskarara itzultzeko ere egokitu da, lortutako emaitzak *Matxin* sistemarekin konparatu ahal izan ditugula (ikus VIII.2.3 atala).

DELI taldea

DELI taldeak¹⁵ edizio digitaleko eta web semantikoko arloan lan egiten du eta, besteak beste, corpus elebidunen parekatze eta prozesaketan ikerketa burutu du.

LEGEBiDUNA proiektuan (Abaitua, 1997) espainiera-euskara administrazioko testuen corpus elebidunen ustiapenaren aukerak aztertu dituzte, itzulpen eta edizio inguruneen eraikuntzarako. Corpora automatikoki prozesatu dute, itzulpen-memoriak sortzeko, bi bertsioetan baliokideak diren itzulpen unitateak identifikatzen dituzten SGML¹⁶ etiketa deskribatzaileekin markatuz.

Administrazioko testu elebidunen idazketarako *BiGentor* editorearen prototipoa (Casillas, 2000) ere eraiki dute, dokumentu egituratuen prozesaketa eta itzulpen-memorien ustiapena konbinatuz.

Zenbait proiektutan, berriz, dokumentu eleanitzen fluxuaren kudeaketarako prozedurak aztertu dituzte, formatu ezberdinak erabiliz: TEI/XML formatua XML-Bi proiektuan (Abaitua *et al.*, 2001) eta XLIFF eta TMX formatua X-Flow proiektuan (Jacob *et al.*, 2003).

¹⁵<http://www.deli.deusto.es>

¹⁶Standard Generalized Markup Language. <http://www.w3.org/MarkUp/SGML>

Dokumentu eleanitzen berreskuratze lanak ere burutu dituzte DELI taldean (Labrador *et al.*, 2003).

II.2.3 Itzulpenerako laguntza-tresnak

Ametra proiektua¹⁷

*Ametra*¹⁸ proiektua (Ortíz *et al.*, 2003) itzulpen-memoriako sistemen eremuan kokatzen da, eta bere xedea ordenadorez lagundutako itzulpen-erreminta sortzea da, *Euskal Herriko Aldizkari Ofizialeko* corpusean oinarrituta eta funts desberdinetako teknikak uztartuz:

- Hizkuntzaren azterketarako teknikak: lematizazioa eta azterketa morfosintaktikoa, segmentazioa eta corpus elebidunen parekatzea, hitz berrien sintesia, eta digituen, izen propioen eta siglen identifikazioa.
- Metodo estatistikoak: hitzen parekatze estatistikoetan oinarritutako segmentazio elebiduna, hitzen parekatze estatistikoa hobetearren hizkuntz informazioa bereganatzea, segmentu laburren itzulpen estatistikorako software eraginkorra garatzea eta eremu murrizteko testuetarako itzulpen estatistiko hutsa.
- Itzulpen-memoriarako garatutako teknikak: esaldien, segmentuen eta hitzen indexazioa, baliokidetzak barne, eta segmentuen eta segmentu multzoen errepikakortasunean oinarritutako berreskuratzea.

Gainera, ondoko tresnak garatu dituzte: baliokidetzak sistematikoen glosarioak (terminologikoak, berreskurapenekoak, itzulpenerakoak), konkordantzien bistaratzailerak eta datu-base guztiak (itzulpen-memoriak, glosarioak) egindako itzulpen berriez osterak elikatzea. Erabilitako teknika guztiak itzultzailearentzat aurrerapen nabarmena dakarren erraminta bakarrean uztartu dituzte.

Hiztegixa hiztegi-kontsultarako tresna

Tesi-lan honetan IAKo sistemaren transferentzia lexikalean burutzen den hiztegi elebidunaren kontsultarako burututako lanak (Díaz de Ilarraza *et al.*, 2000b) beste aplikazio baterako berrerabili dugu IXA taldean, itzultzaileek

¹⁷ *Ametra* proiektuan corpusetan oinarritutako ikerketa-lanak ere burutzen dira, eta beraz aurreko atalean ere koka genezakeen.

¹⁸ AMETRA proiektuko zentru teknologikoak EHUko Bilboko Zientzia Fakultatea eta Institut Tecnològic d'Informàtica de Valencia dira.

egin beharreko bilaketa-lanena¹⁹ oso baliagarria den tresna bat garatuz: *Hiztegia*.

Bost hiztegi elektroniko integratzen ditu: *Elhuyar* eu-es/es-eu eta *Morris* eu-en/en-eu hiztegi elebidunak, *Euskal Hiztegia* eta *Euskal Hiztegi Modernoa* euskarazko hiztegi elebakarrak, eta euskarazko *Sinonimoen Hiztegia*. Gainera, Euskaltzaindiaren euskara estandarren hitz zerrenda ere integratzen du. Hiztegi hauek banaka kontsulta daitezke edo euskarazko hitz bat sartu eta hiztegi guztietan bilatu.

Sisteman beste zenbait baliabide ere integratu ditugu: euskararako zuzentzaile, erabiltzaileak euskara estandarra ez den hitz bat idazten badu, hitz horren karaktere segida antzekoa duten hitz zuzenak emateko; *Egunkaria* corpusean bilaketak egiteko aukera, hitz baten erabilera bere testuinguruan ikusi ahal izateko; eta euskarazko lematizatzailea, edozein hitz formaren lema ematen duena.

Sistema honen baliagarritasuna ondoko ezaugarrietan oinarritzen da: ingurune bakar batean hiztegi anitzen (eta beste baliabideen) integrazioa; eraginkortasuna, bilaketak oso azkarrak dira; eta sistemaren interfazea erabiltzeko erraztasuna.

Tresna honen baliagarritasuna IXA taldeko ikertzaileen eguneroko erabilerarekin erabat frogatua izan da. Horregatik prototipo honen interfaze grafikoa duen bertsioa ere eraiki da. Hala ere, hiztegien egile-eskubideengatik tresna hau ezin da ez zabaldu ez publiko egin.

Euskalbar tresnabarra

*Euskalbar*²⁰ *Firefox* nabigatzaileko tresnabarra da, euskera-espainiera itzulpenak egiten laguntzen duena, Interneten dauden hiztegi ofizialak eta beste zenbait baliabide kontsultatuz: *Euskalterm*, *Elhuyar*, *Hiztegia3000*, *Euskaltzaindia*, *ItzuL* posta zerrendaren artxiboa, *Harluxet*, *Mokoroo*, *ZT Corpusa*, *XUXENweb* zuzentzaile ortografikoa eta *Matxin* itzultzaile automatikoa.

11.2.4 Sorkuntza automatikoa

MultiMeteo sistema

(Coch eta Chevreau, 2001; Schütz, 2001) lanetan sorkuntza eleanitz elkarrreragilea IArean alternatiba izan daitekeela planteatzen da. Hori dela eta, tesi-lan honen barruan, sorkuntza automatikoan ikertzeko eta sakontzeko

¹⁹Itzultzaileek bere lan-orduen %60a baino gehiago ematen omen dute hiztegieta eta terminologia-ikerketako lanetan bilatzen (Hutchins eta Somers, 1992).

²⁰<http://www.interneteuskadi.org/euskalbar>

asmoz, *MultiMeteo* eguraldi iragarpenen sorkuntza automatikorako sistemaren euskararako egokitzapenaren diseinu eta inplementazio lanak burutu genituen.

MultiMeteo proiektua abian jarri zen Frantziako Meteorologi-Zerbitzuak bultzatuta, lau hizkuntzetarako (frantsesa, ingelesa, espainiera eta alemanera) sistema garatuz (Chevreau *et al.*, 1999).

Meteorologia-behatokietatik eta sateliteetatik jasotako datu numeriko guztiak eredu matematiko konplexuen bidez prozesatzen dira, eguraldiaren bilakaera simulatzen duten iragarpen-meteorologikoetarako datu-matrizeak sortuz. Datu hauetatik sistemak automatikoki testuaren zirriborro bat sortzen du zuzenketa-lanetan ariko den meteorologoaren ama hizkuntzan. Meteorologoak, interfaze egoki baten bitartez, testuan zuzenketak egiten ditu eta, egindako aldaketak kontuan hartuta sistemak testuak hizkuntza guztietan sortzen ditu.

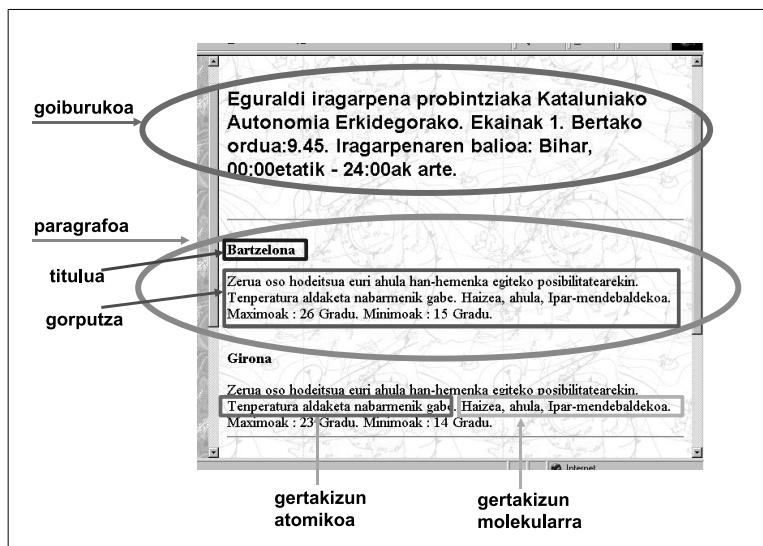
Teknika honen abantailak ondokoak dira: bizkortasuna (hizkuntza bakoitzean testu bakoitza sortzeko 2 segundo inguru behar dira, giza-itzultzaile batek 10 minutu inguru behar duen bitartean); sortutako testuen kalitate handia; mantentzeko eta moldatzeko erraztasuna; eta azkenik, meteorologoek sistema onartzen dutela, ez dielako lanpostua kenduko, hizkuntza arrotzetan idazten lagunduko baizik.

Sistemaren sorkuntza-motoreak bi modulu dauzka: planifikazioa eta gauzatzea. Planifikazioan, sarrerako datuetan oinarrituta, testuaren edukia zehazten da: gertaerak (egoera meteorologiko bati edo egoeraren eboluzio bati lotutako objektu kontzeptualak) eta horien arteko loturak. Gertaerak interlingua bat erabiliz adierazten dira; horrela deskribapena hizkuntzetatik independentea izatea lortzen da.

Gauzatze-moduluan ezagutza-base linguistikoa erabiltzen da, eta lortutako kontzeptuak hizkuntza bakoitzean linguistikoki gauzatzen dira, esanahi-testu teorian (Mel'cuk, 1988; Polguère, 1998) oinarrituta. Bost etapatan banatzen da: aurre-denotazioa, semantika, sintaxi sakona, azaleko sintaxia eta morfologia.

2000. urtean Espainiako Instituto Nacional de Meteorología (INM) eta Lexiquet enpresak akordio batetara iritsi ziren sistema lau hizkuntza gehiagotara zabaltzeko: nederlandera, katalana, galegoa eta euskara. Euskarara zabaltzeaz (Díaz de Ilarraza *et al.*, 2001b) IXA taldea eta UZEI terminologia-zentroa arduratu ziren.

Euskararako egokitzapenerako espainierarako gauzatze-modulua oinarritzat hartu zen. Esaldien egitura gehienak aldatu behar izan ziren eta morfologia-mailako atzizkiekin lan berezia egin. Aurretik azterketa linguistikoa burutu genuen, ondoren ezagutza-basean sartu beharreko informazioaren



Irudia II.9: *MultiMeteo* sistemaren emaitza euskaraz

azterketa eta diseinua, gero gertaera bakoitzerako adibide adierazgarri baten informazioa sartu eta probatu, eta azkenik, gertaera-mota bakoitzerako aukera guztiak sartu eta proba masiboak egin.

Egindako lana INMren sisteman integratu zen, urtebetez egunero-egunero estatu espainoleko komunitateetako eguraldi iragarpenak eskaintzen zituela bere web helbidean (adibide bat ikus daiteke II.9 irudian). Hala ere, eta arrazoirik eman gabe, INMek zerbitzu hau eskaintzeari utzi zion.

II.3 Itzulpen automatikoa eta kode irekiko software librea

*Libre te quiero, como arroyo que brinca de peña en peña, pero no mía.
Pero no mía ni de Dios ni de nadie
ni tuya siquiera.*

Agustín García Calvo

Azken urteetan kode irekiko software libreak hedapen handia izan du. oso sistema arrakastatsuak garatu direla: *Linux* sistema eragilea, *Mozilla-Firefox* web nabigatzailea, *Apache* web zerbitzaria, *OpenOffice* ofimatika-paketea...

Baina itzulpen automatikorako erabiltzen diren sistema gehienek, dohain erabiltzekoak direnak barne, kode itxia eta jabetuna dute. Eta, 2005.eko MTSummit X kongresuaren barruan antolatu zen *Open-Source Machine Translation* workshop-eko deialdian esaten zen bezala, hau bereziki harrigarria da, ikusita IAn lan egiten duten talde asko finantziazio publikoa dutela.

Commercial, government and academic groups spread throughout the world spend a lot of effort and money into developing the perfect [MT] system. Their efforts are seldom leveraged against or in support of each other. (...)

It is high time for the language technology industry to join forces and establish guidelines for the development of the 'Linux' of machine translation. (Beninatto, 2003)

Atal honetan kode irekiko software librearen inguruko terminoak argitu ondoren, mota honetako softwareak eskaintzen dituen abantailak aztertzen ditugu. Azkenik, IA burutzen duten kode irekiko sistema urrien berri ematen dugu.

II.3.1 Software librea vs. kode irekiko softwarea

Software librea eta *kode irekiko softwarea* terminoen inguruan eztabaida be-roa dago software mota hauen garatzaileen artean (Stallman, 2002b). Software librearen eta kode irekiaren mugimenduak ados daude praktikan. Desberdintasunak oinarritzko printzipiotan daude.

Software librea

Software librea (*free software*) libreki erabilia, kopiatua, aztertua, moldatua eta birbanatua izan daitekeena da.

Richar Stallman-ek, 1984an GNU (*GNU's Not Unix*) proiektuan ²¹ lanean hasi eta urtebete beranduago Free Software Foundation²² (FSF) elkar-tea sortu zuen. Mugimendu honek emandako software librearen definizioaren arabera (Stallman, 2002a), softwarea *librea* da soilik ondoko askatasunak bermatzen dituen:

1. Programa egikaritzeko askatasuna edozein xedearekin.
2. Programa aztertzeke eta norberaren beharretara egokitzeko askatasuna.
3. Programa birbanatzeko askatasuna, ondokoaren onerako.
4. Programa hobetzeko eta hobekuntzak publikoak egiteko askatasuna, komunitate osoaren onerako.

1 eta 3 askatasunak bete ahal izateko, iturburu-kodea atzitzea nahitaezkoa da, hau da, kode irekia izan behar du.

Software librearen mugimenduak askatasunaren alderdi filosofiko eta etikoak kontutan hartzen ditu. Mugimendu honentzat software jabeduna etikaren aurkakoa da, gizakien artean konpartitzearen debekua lege naturalen aurka doalako.

Software librea ez da doako softwarearekin (*freeware*) nahastu behar, honek ez dituelako aldatzeko eta birbanatzeko eskubideak bermatzen.

Kode irekiko softwarea

1998an Open Source Initiative²³ (OSI) sortu zen bi helburuekin: iturburu-kodea konpartitzearen onura praktikoei garrantzi handiagoa ematea, eta software-enpresen artean kontzeptu horretara hurbiltzea. Mugimendu honetan kode irekiko software (*open source software*) terminoa nahiago dute.

Software librearen aldean, kode irekiaren filosofiaren muina guztiz teknikoa da: kodea konpartitzean, sortuko den programa software jabeduna baino kalitate hobeagokoa izango da.

Baina kode irekiaren mugimenduarentzat *kode irekia* terminoak ez du soilik adierazten kodea ikus daitekeela. Mugimendu honek dekalogo bat du, sistema batek kode irekikoa izateko bertako baldintzak bete behar dituelarik (OSI, 2007). Baldintza horiek software librearen lau askatasunen guztiz baliokideak dira.

²¹<http://www.gnu.org>

²²<http://www.fsf.org>

²³<http://www.opensource.org>

Kode irekiko software librea

Software librea eta *kode irekiko softwarea* terminoen erabileraren inguruan dagoen eztabaida gairatuz, guk gaur egun zabaltzen ari den beste termino bat erabiliko dugu, *kode irekiko software librea* (*Free Open Source Software*, *FOSS*) (Wong eta Sayo, 2004).

Hala ere, termino hau askotan errepikatu behar denerako luzea denez, *kode irekia* edo *librea* erabiliko ditugu esanahi berdinarekin.

Kode irekiko software librearen banaketarako zenbait lizentzia-mota erabiltzen dira: GNU GPL²⁴, LGPL²⁵ eta BSD²⁶ adibidez; haien artean desberdintasun handiak egon daitezke.

II.3.2 Kode irekiaren abantailak

Given enough eyeballs, all bugs are shallow. (Raymond, 2001)

Itzulpen automatikoko sistema bat hutsetik sortzea baliabide handiak behar ditu eta karga handiko lana da. Kode irekiko sistema librean garapenak egoera hau aldatzen du ondokoak ahalbidetzen dituelako:

- Sistema hobetzea komunitate ireki batean, non edonork parte har dezakeen, guztien onerako.
- Sisteman modulu berriak integratzea, adibidez itzulpen automatikoko estrategia desberdinak hibridatzeko.
- Sistema egokitzea erabilera zehatzetarako.
- Sisteman hizkuntza berriak integratzea, hizkuntza bikote berrietarako sistema berriak eraikitzea baino askoz merkeagoa eta azkarragoa bihurtuz.

²⁴<http://www.gnu.org/licenses/gpl.html> (*General Public License*) da erabiliena. Egileak *copyright* eskubideak mantentzen ditu eta birbanaketa eta aldaketa baimentzen ditu. Hori bai, aldatutako bertsioa guztiak GNU GPL lizentziarekin mantendu beharko dira.

²⁵<http://www.gnu.org/licenses/lgpl.html> (*Lesser General Public License*) lizentzia GPLren bertsio murriztua da eta kode librearen erabilera programa jabetunetan baimentzen du.

²⁶<http://www.opensource.org/licenses/bsd-license.php> (*Berkely Software Distribution*) motako lizentziak softwarearen moldaketa eta birbanaketa baimentzen dute, jatorrizko egileasuna akreditatuz. GPLrekin ez bezala, BSD softwarearen birbanaketa edozein beste lizentziarekin egin daiteke, baita lizentzia jabetunekin.

Kode irekiko IAko sistema libreekin jende gehiagok erabili ahal izango luke teknologia hau. Gainera, instituzio eta enpresek sistemak beren beharretara egokitu ahal izango lituzkete, datu linguistiko berriak garatzen (hiztegiak, erregelak, corpusak) edo beste paketeekin integratzen. Hori bai, kode irekiak enpresaren ereduan aldaketa handia dakar, beste zerbitzu mota bat eskaini behar dela: aholkularitza, sistemen egokitzapenak, datu linguistikoen garapena, sistema orokorretan integrazioa...

Kode irekiko sistemak bereziki garrantzitsuak izan daitezke erabiltzaile gutxiko hizkuntzentzat, hizkuntza hauek ez direlako sistema komertzialen helburu.

Kode irekiko IAko sistemek abantaila hauek izateko berebiziko garrantzia du datuen estandarizazioak, nahitaezkoa dena elkarreragingarritasunerako eta elkartrukerako.

Kode irekiaren eragina IArako estrategien arrakastan

(Bond *et al.*, 2005) lanean oso hipotesi interesgarria eskaintzen dute kode irekiko garapena eta IAren inguruan, metodo estatistikoen gaur egungo arrakasta argi dezakeena.

Erregeletan oinarritutako IAren inguruko ikerketan oztopo handiena hasierako martxan jartzeak behar duen esfortzua da. IARI dagozkien arazo zehatzei aurre egin baino lehen, beharrezkoa da analisi-, transferentzia- eta sorkuntza-moduluak, eta hiztegi elebakar eta elebidunak eraikitzea. Erregeletan oinarritutako sistema gehienak kode irekiaren ideia zabaldu aurretik eraiki zirenez, garatutako moduluak eta lexikoiak jabetun mantendu dira, osagai horiek ez direla sistema desberdinen artean elkartrukatu, ez konparatzeko, ezta sitemak hobetzeko ere.

IA estatistikoaren ikerketan, berriz, kodea eta algoritmoak partekatu izan dira, eta hasieratik eskuragarri egon dira kode irekiko oinarritzko sistema estatistikoak. Corpusak ere eskuragarri izan dira, lexikoiak baino askoz murrizpen gutxiagorekin, nahiz eta soilik ikerketarako izan. Honek IA estatistikoa oso bizkor aurreratzea eragin du, sistema bat eraikitzeke ordurarte garatutako gainean egin ahal izan delako, probatutako metodo onenak erabiliz.

Garapen-metodologiaz gain, argi dago bi estrategien arteko desberdintasun handiena dela IA estatistikorako ezagutza linguistikoa, ia eskuzko lanik gabe, corpusetatik jasotzen dela, eta erregeletan oinarritutako IArako, berriz, ezagutza hori, lexikoietan eta gramatiketan, eskuz kodetu behar dela. Baina honek, hain zuzen, erregeletan oinarritutako itzulpen automatikoa kode irekiaren onuradun handiagoa bihurtzen du, kode irekiko komunitate

batean, hizkuntzalarien ekarpenek, lexikoietan zein erregeletan, erregeletan oinarritutako sistemaren emaitzen hobekuntzan zuzenean eragingo luketee-lako. Gainera, sistemaren algoritmoak hobetu eta modulu berriak integratu ahal izango lirateke, sistema bateko moduluak beste batean erabili, sistema baten gainean hizkuntza berrietarako sistemak eraiki...

II.3.3 IAKo kode irekiko sistema libreak

Hizkuntza guztietarako kalitate egokia izango duten IAKo sistemak ahal-bidetuko duten gakoetako bat, kode irekiko baliabide linguistikoen garapen kooperatiboa izango da Boitet-en (2001) ustez. Baina, esan bezala, oraindik, IAKo kode irekiko software gutxi garatu da.

Tesi-lan honetan aurkezten dugun *Matxin* sistemaz gain, zuzenean erabil daitezkeen itzulpen automatiko osoa burutzen duten bi sistema orokor, soilik, aurkitu ditugu:

- *Apertium*²⁷. Hizkuntza hurbilak (gaztelera, katalana, galiziera eta okzitaniera) itzultzen dituen azaleko transferentziazko sistema.
- *Logos Open Source Machine Translation*²⁸. Logos sistema komertzialaren kode-irekiko bertsioa. en/de-fr/it/sp/pt hizkuntza bikoteetarako itzulpenak eskaintzen dituena.

Beste proiektu batzuetan lanean ari dira, baina oraindik ez dute sistema erabilgarririk garatu, adibidez, *Traduki*²⁹ eta *GPLTrans*³⁰.

Transferentzia semantikokoan oinarritutako sistema bat garatzen ari da *Deep Linguistic Processing with HPSG Initiative (DELPH-IN)* ekimenak eskaintzen duen azpiegituran oinarrituta (Bond *et al.*, 2005). Norvegiar-ingelesa³¹ eta japoniera-ingelesa prototipoak garatu omen dituzte dagoeneko.

Itzulpen automatiko estatistikoa burutzeko hainbat sistema libre garatu dira. Sistema hauek, hala ere, ezin dira zuzenean itzultzeko erabili, aurretik itzultzea nahi den hizkuntza-bikotearen corpusa bildu eta prestatu behar baita:

- *Moses*³². N-grametan oinarritutakoa, C++ lengoiaz implementatua. *Pharaoh*³³ deskodetzaile ezagunaren hedapena da.

²⁷<http://apertium.sourceforge.net>

²⁸<http://212.218.86.59/LogOSMaTran.do>

²⁹<http://traduki.sourceforge.net>

³⁰<http://www.translator.cx>

³¹<http://www.emmtee.net>

³²<http://www.statmt.org/moses>

³³<http://www.isi.edu/licensed-sw/pharaoh>

- *Phramer*³⁴. N-grametan oinarritutako sistema hedagarria, Javaz inplementatua.
- *Open A.I. Kit* (Walker, 2005) IAEaren osagai nagusien implementazioa.
- *PSMT*³⁵. Prolog-ez egindako implementazio sinple bat, dagoeneko erabilera praktiko gutxi duena.

Badira ere IAren arloko kode irekiko beste aplikazio gutxi batzuk:

- *Linguaphile*³⁶. Hitzez hitzezko itzulpen sistema. Hiztegia bilatu aurretik, hitzak lematizatzen ditu. 56 hizkuntza desberdinen arteko itzulpena ematen duela estaten dute, euskara barne, baina euskararako soilik asteko egunak eta hilabeteak itzultzen ditu.
- *Carabao Machine Translation DIY Kit*³⁷. Edozein hizkuntzatarako interlingua oinarritutako IAko sistema propioa garatzeko tresna, hiztegiak eta erregelak eraikitze aukera ematen duena.
- *libtranslate*³⁸. Internet bidez itzulpenak ematen dituzten zerbitzuak (*Babel Fish*, *Google Language Tools*, *SYSTRAN*...) erabiliz, hizkuntza desberdinen arteko testuak eta web orriak itzultzeko paketea.

Itzulpen automatiko osoa burutzen duten *Apertium* eta *Logos Open Source Machine Translation* sistemak sakonago aurkeztu ditugu ondoko lerroetan.

Apertium

Apertium (Corbí-Bellot *et al.*, 2005) plataforma hasiera batean hizkuntza hurbilen artean (espainiera, katalana edo galegoa) itzultzeko sortu bazen ere, berriki hizkuntza urrunagoen artean (ingelesa-katalana) itzultzeko hedatzen ari dira.

Plataformak ondoko baliabideak eskeintzen ditu: hizkuntzatik independentea den IAko sistema bat, hizkuntza-bikote batentzako IAko sistema eraikitze beharrezkoa diren datu linguistikoak kudeatzeko tresnak, eta gero eta hizkuntza-bikote gehiagoentzako datu linguistikoak.

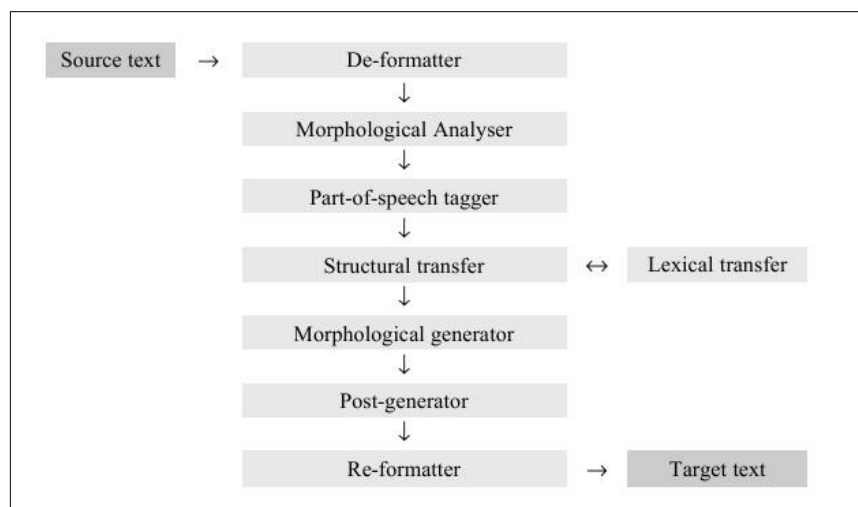
³⁴<http://www.utdallas.edu/~mgo031000/phramer>

³⁵<http://psmt.sourceforge.net>

³⁶<http://linguaphile.sourceforge.net>

³⁷<http://ai-depot.com/nolimits/aboutcarabao.htm>

³⁸<http://www.nongnu.org/libtranslate>

Irudia II.10: *Apertium* sistemaren arkitektura

IAko sistemak azaleko transferentzia burutzen du, urrats desberdinetan banatuta: desformatatzea, analisi morfologikoa, *part-of-speech* desanbiguazioa, azaleko transferentzia estrukturala, transferentzia lexikala, sorkuntza morfologikoa eta birformatatzea. Sistemaren arkitektura II.10 irudian ikus daiteke (Corbí-Bellot *et al.*, 2005).

Sistemak teknologia desberdinak erabiltzen ditu: egoera finituetako transduktoreak prozesaketa lexikal guztientzako (analisi eta sorkuntza morfologikoa eta transferentzia lexikala), Markov ereduak *part-of-speech* etiketatzeke, eta etapa anitzetako egoera finituetan oinarritutako *chunking* transferentzia estrukturalerako.

Sistemaren datu linguistikoak XMLn oinarritutako formatu estandarretan kodetuta daude eta banatzen diren tresnekin konpilatzen dira sistemak erabiliko duen abiadura-handiko formatura bihurtzeko.

Apertium arkitektura Transducens taldeak garatu du Alacanteko Unibertsitateko Departament de Llenguatges i Sistemes Informàtics sailan.

Logos Open Source Machine Translation

OpenLogos (edo *LogOSMaTran*) *Logos* sistema jabetun komertzialaren kode irekiko bertsioa da.

Logos (Scott, 2003), 30 urteetan garatu izan dena, IAKo sistema komertzialen artean handienetako eta ahaltsuenetako bat da. Sistemak erregeletan oinarritutako estrategia itzulpen-memorien erabilerarekin konbinatzen du.

Formatu desberdinetako dokumentuak prozesatzen ditu, hizkuntza desberdinetara oso denbora laburrean itzuliz. Emaizak kalitate handia eskaintzen du, eta behar izanez gero, giza-itzultzaile batek errevisa dezake. Denbora eta kostu handiak aurrezten ditu, beste ezaugarrien artean, jatorrizko dokumentuaren formatua mantentzen duelako.

Ingelesatik eta alemaneratik, frantsesera, italierara, espainierara eta portugesez itzulpenak ematen ditu. Hala ere, arkitekturak hizkuntz desberdinetako baliabideak integratzeko plataforma orokor bat eta oinarritzko sistema eskaintzen du.

Edozein testu motatarako eta arlotako itzultzaile orokorra da. Arlo baterako berezitu nahi bada, sistemak errazten du erabiltzaileak terminologia eta hiztegiak egokitu ahal izatea, eta erabiltzailearen hiztegia sortzea ere.

II.4 Laburpena

Kapitulu honetan itzulpen automatikoaren sarrera bat eskaini dugu. Oinarritzko kontzeptuak eta IAreneko historia laburbildu ondoren, IArako estrategia nagusien ezaugarriak azaldu eta estrategia desberdinak konbinatzen dituzten sistema hibridoak aurkeztu ditugu, gaur egungo erronkak nabarmenduz.

IAreneko inguruko arloan euskararekin burututako ikerketa lan urriek errepaso eman diegu, hortan gure aportazioa den *Matxin* IAreneko sistema mugarri esanguratsua izanik: euskara barne hartzen duen lehenengo itzulpen automatikoko sistema erabilgarria.

Bukatzeko kode irekiko software librearen ezaugarriak eta abantailak, eta IAreneko alorrean oraindik ugariak ez diren kode irekiko sistema libreak aztertzen ditugu, honetan ere, hurrengo kapituluan aurkezten dugun *Matxin* sistema aintzindaria dela.

III. KAPITULUA

Matxin itzulpen-sistemaren teknologia

<i>matxinada</i>	<i>Agintaritzaren aurkako herri-altzamen- edo jazzarraldia.</i>
<i>matxinatu</i>	<i>Aginpidea duenaren aurka altxatu.</i>
<i>matxingorri</i>	<i>Marigorringo (occinella septempunctata). Deabruari ematen zaion izena.</i>
<i>matxino</i>	<i>Ola langilea. Burdina lantzeko tresna, goitik behera higitzen den masa astun batez osatua.</i>
<i>matxinsalto</i>	<i>Intsektu jauzikaria, berde horixka, otiaren antzekoa baina alorretan kalterik egiten ez duena.</i>
<i>matxin-sukalde</i>	<i>Etartetik ez ateratzeko joera duen gizona.</i>
<i>matxin-transleixion</i>	<i>Teknologi berrienak, betiko kutsuarekin.</i>

Euskal Hizteria, 1972.

Kapitulu honetan, *Matxin* sistemaren bilakaera, ezaugarriak eta mugak zehaztu ondoren, sistemaren deskribapen orokor bat aurkeztuko dugu. Diseinatutako arkitektura erregeletan oinarritutako transferentzia sintaktiko sakoneko sistemena da, non prozesaketa analisia, transferentzia eta sorkuntza faseetan banatzen den. Sistemak analisi sintaktiko osoa behar du, mendekotasun-analisia barne. Transferentzia-fasea bi mailatan burutzen da: lexikala eta estrukturala. Sorkuntza ere bi urratsetan banatzen da: sintaktikoa eta morfologikoa.

Sistemak prozesatzen duen datu-egitura eta moduluen arteko datu-fluxuen formatua azaldu ondoren, sistemak erabiltzen dituen datuak deskribatuko ditugu: baliabide linguistikoak (hiztegiak eta corpusetik erauzitako

informazioa), datu-egitura manipulatzeko erregelak eta aditz-transferentzia-rako gramatika.

III.1 Sistemaren bilakaera

IXA taldeak oinarri linguistikoak ezarri eta konplexutasun ertaineko tresnak garatu ondoren, tresna aurreratuen inguruko ikerketari ekin zion, eta itzulpen automatikoaren (IA) erronkari aurre egitea 1998. urtean erabaki zuen, asmo horrekin tesi-lan hau abian jarri zela. Helburua itzulpen automatikoko sistema eleanitza eraikitzea zen, itzulpen-bikoteetako hizkuntza bat, beti, euskara (eu) izango zela, eta espainiera (es) edo ingelesa (en) bestea, Hego Euskal Herrian itzulpenak burutzeko eskaera handiena duten hizkuntzak direlako.

Matxin sistema izan da lan horren lehen fruitua. Bere garapen-prozesua modu inkrementalean egin da, eta emandako pausuak atal honetan aurkeztuko ditugu.

III.1.1 Hasierako diseinua eta prototipoak

Lehenengo prototipoen diseinuari ekin aurretik, itzulpen automatikoaren inguruko bibliografia irakurri eta erabilgarri izan zitezkeen tresnak aztertu genituen, II kapituluan azaldutakoari oinarria jarriz.

Azterketa horiek eraikiko genuen sistemaren oinarritzko ezaugarriak finkatzea ekarri zuten: espainieratik eta ingelesetik euskarara itzulpen-prozesu osoa burutuko zuen erabilera orokorreko sistema hibridoa, lehenengo hurbilpena erregeletan oinarritutakoa izanik.

Lehenengo prototipoak ingelesetik euskarara izen- eta preposizio-sintagmak itzultzen zituen. Ondoren espainieratik euskarara itzultzen zuen prototipoa inplementatu genuen, esaldi mailara jauzia emanez.

Bibliografia eta tresnen azterketa

Lehenengo lana itzulpen automatikoari buruz zegoen bibliografia zabala aztertzea izan zen, gure eginbeharrerako komenigarrien izango zen estrategia erabakitzeke.

Horrekin batera euskara, espainiera eta ingeleseko prozesaketa morfologiko eta sintaktikorako erabilgarri zeuden tresnak eta IAKo sistema batean berrerabiltzeko ematen zituzten aukerak aztertu genituen.

Euskararako tresnak IXA taldean garatutakoak ziren: EDBL, prozesadore morfologikoa...

Espainierarako eta ingeleserako baliabideak beste talde batzuek garatutakoak izan beharko ziren eta, hasierako hurbilpenerako behintzat, dohainekoak. Analizatzaileak aurkitu genituen, baina ez sortzaileak.

Ingeleserako *EngCG* (*Constraint Grammar Parser of English*) analizatzaile morfosintaktikoa (Voutilainen, 1995) aztertu genuen, Internet bidez erabil zitekeena, Lingsoft enpresaren web orritik¹. Gaur egun ere, analizatzaile honen hedapena den *EngCG-2* erabilgarri dago Connexor enpresaren webgunetik².

Espainierazko analisirako UPC unibertsitatean garatutako *Maco*, *Tacat* eta *Relax* moduluak biltzen zituen paketea (Carmona *et al.*, 1998) aztertu genuen, ikerketan erabiltzeko baimena lortu genuela. Gaur egun, pakete horien hedapena den *Freeling* analizatzailea (Atserias *et al.*, 2006) dago, kode irekiko software libre moduan banatzen dena³.

Hiztegi dagokionez ikerketarako hiru aukera genituen erabilgarri: *Elhuyar* espainiera-euskara eta euskara-espainiera (Elhuyar, 2000), *Morris* ingelese-euskara eta euskara-inglese (Morris, 1998) eta *Euskal Hiztegia* euskarazko hiztegi elebakarra (Sarasola, 1996).

Garai hartan corpus elebidun gutxi zegoen eskuragarri, soilik administrazioetako buletin ofizialak, corpusetan oinarritutako IAKo estrategietan erabiltzeko nahikoa ez zena.

Erabaki estrategikoak

Bibliografiaren eta tresna eskuragarrien azterketa egin ondoren, eraiki nahi genuen sistemaren ezaugarriak ezarri genituen:

- Sistemak itzulpen prozesu *osoa* automatikoki burutuko zuen, bukaerari post-edizioa egin ahal izateko aukera eskainita. Bestela esanda, ez genuen egin nahi itzulpen-laguntzarako tresna bat, itzulpen automatikorako sistema bat baizik.
- Sistema *hibridoa* izatea, erregeletan oinarritutako estrategia adibideetan oinarritutako metodoekin eta teknika estatistikoekin konbinatuz. Lehenengo hurbilpenerako transferentzian oinarritutako ereduari jarraituko genion, corpusetan oinarritutako estrategiak aurrerago aztertuko genituzkeela, corpus elebidun eskuragarrien kopurua handituz gero.

¹<http://www2.lingsoft.fi/cgi-bin/engcg>

²<http://www.connexor.com/demo/syntax>

³<http://garraf.epsevg.upc.es/freeling>

- Sistema *leanitza*, euskaraz gain, espainiera eta ingelesa barne hartuko lituzkeena. Lehenengo hurbilpenetarako xede-hizkuntza euskara izatea erabaki genuen, kontrako norabidea aurrerago lantzeko utziz, ingeleserako eta espainierarako sortzailerik eskura ez genuelako. Gainera, etorkizunean beste hizkuntzak integratzeko gaitasuna eduki beharko luke sistemak.
- *Erabilera orokorreko* sistema izatea, nahiz eta, itzulpenen kalitatea hobetzeko, geroago sistema eremu zehatz baterako bereiztea aurreikusi genuen.
- Eraikitako sistema software librean inplementatuko zen.
- Softwarea garatzeko *prozesu inkrementala* egitea, bide horretan zenbait prototipo eraikiz. Hasieran sistemaren hezurdura sortuko genuen eta aurrerago osatu eta estaldura eta doitasuna handitu.

IAko lehenengo prototipoa: en-eu, izen- eta preposizio-sintagmen itzulpena

Eraiki genuen lehenengo prototipoa ingelesetik euskarara izen- eta preposizio-sintagmak itzultzen zituen (Díaz de Ilarraza *et al.*, 1999, 2000c).

Prototipo hori eraikitzeko ondoko moduluak garatu genituen, inplementazioa Perl programazio-lengoaiaz burutu zela:

- Sintagma barruko mendekotasun-analizatzailea, *EngCG* ingeleserako analizatzailearen emaitzatik sintagma barruko mendekotasun-zuhaitza sortzen duena.
- Transferentzia lexikala, lexikoi elebidunean bilaketa egiten duena.
- Transferentzia estrukturala, ingelesezko artikulua eta preposizioak euskarara itzultzean sortzen diren balio lexikalik gabeko unitate lexikalak ezabatzen dituen, beren postposizioen informazioa sintagmari pasatuz.
- Sintagma barruko sorkuntza sintaktikoa, sintagmen barruko hitzen ordena ebatzen duena, sintagmaren postposizio-informazioa sintagmako azken hitzari pasatuz.
- Sorkuntza morfologikorako informazioaren egokitzea, sortzaile morfologikok behar duen formatura bihurtzen duena.

Ingeles-euskara lexikoi elebiduna sortzeko, *Morris* hiztegiaren gainean prozesaketa konplexu bat burutzeaz gain (Díaz de Ilarraza *et al.*, 2000b), kategoria itxietako (determinanteak, preposizioak...) hitz erabilienak eskuz kodetu genituen.

Itzulpenean ezaugarri semantikoen beharra ikusi genuen eta arlo horretan azterketa bat egin ondoren ezaugarri semantikoak etiketatzen zituen prozesu erdi-automatiko bat diseinatu eta inplementatu genuen (Díaz de Ilarraza *et al.*, 2000a, 2002b).

Lortutako prototipo honen emaitzak oso itxaropentsuak izan ziren.

Itzulpenerako laguntza-tresna baten eraikuntza

Transferentzia-faserako eraiki genuen hiztegien kontsulta azkarrerako modulutik abiatuta itzultzaile eta hizkuntzalarientzat beste zereginetarako ere oso erabilgarri izango zen interfaze bat garatu genuen: *Hiztegi*. Honek 5 hiztegi-tako informazioa kontsultatzeko eta corpusean bilaketak egiteko aukerak emateaz gain, lematizatzailea eta hitz zuzenen proposatzailea integratzen ditu.

Beste estrategien azterketa

Erregeletan oinarritutako metodoen mugak gainditzeko, eta etorkizunean sistema hibrido bat eraikitzeko asmoarekin, adibideetan oinarritutako estrategia aztertzeari ekin genion (Díaz de Ilarraza *et al.*, 2002a).

Gainera, zenbait kasutan itzulpen automatikoaren alternatiba edo osagarri izan daitekeen sorkuntza automatikoak eskaintzen dituen aukerak aztertzeko, *MultiMeteo* sorkuntza eleanitzeko sistemaren euskararako ego-kitzapenerako (Díaz de Ilarraza *et al.*, 2001b) diseinua eta inplementazioa burutu genuen.

IAko bigarren prototipoa: es-eu, esaldi osoen itzulpena

Espainiera-euskara itzulpena garatzeko prototipoari heldu genion eta sintagma mailatik esaldi mailara jauzia eman genuen (Díaz de Ilarraza *et al.*, 2001a), sistemaren modulartasunari esker, ingelesa-euskararako egindako lan gehiena berrerabiliz.

Ondoko modulu berriak garatu genituen:

- Sintagma barruko eta sintagmen arteko mendekotasun-analizatzailea, espainierazko analizatzailearen emaitzatik mendekotasun-zuhaitza sortzen duena. Hasieran *Maco*, *Tacat* eta *Relax* moduluak (Carmona

et al., 1998) biltzen zituen espainierazko analizatzailea erabiltzen genuen eta, ondoren, pakete horien hedapena den *Freeling* analizatzailea (Carreras *et al.*, 2004), bertan integratu dutela gurean oinarritutako mendekotasun-analizatzailea (Atserias *et al.*, 2005).

- Aditz-kateen transferentzia estrukturala, egoera finituetako teknologia erabiliz (Alegria *et al.*, 2005a).
- Sintagmen arteko sorkuntza sintaktikoa, sintagmen ordena ebazten duena.

Gainera, transferentzia moduluertarako espainiera-euskara erregelak prestatu eta *Elhuyar* espainiera-euskara hiztegiaren egokitzapena burutu genuen.

Lehenengo prototipoan garatutako euskararako sorkuntza morfologikorako modulu eta datu berberak erabili ziren.

Konplexutasunari aurre egiteko sistema birdiseinatu genuen. Objektuei zuzendutako diseinu berria egin genuen, modulartasuna hobetuz, eta datuen eta programen arteko banaketa egokiagoa egin genuen.

III.1.2 Matxin teknologia OpenTrad proiektuan

2005.ean OpenTrad⁴ proiektua martxan jarri zen. Proiektu horren helburua estatu espainiarreko hizkuntza nagusietarako kode irekiko itzulpen automatikoko sistemak sortzea zen. Horretarako abiadura handiko eta kode irekiko itzulpen automatikoko bi gailu edo motore garatu ziren: *Apertium*, gertuko hizkuntza bikoteentzat (espainiera-katalana, katalana-espainiera, espainiera-galegoa eta galegoa-espainiera), transferentzia sintaktiko partziala baino egiten ez duena (Corbí-Bellot *et al.*, 2005); eta *Matxin*, elkarrengandik urrunago dauden hizkuntz bikoteentzat (espainiera-euskara), transferentzia sintaktiko osoa burutzen duena (Alegria *et al.*, 2005b).

Proiektuaren berrikuntza teknologiko aipagarrienak honako hauek ziren: estatu espainiarreko hizkuntzen arteko itzulpen automatikorako kodea eta datu linguistikoak estandarizatzea eta irekitzea (software librea); elkarreragingarritasuna, modulartasuna eta integrazteko gaitasuna; itzulpen-abiadura handia web-orriak itzuli ahal izateko nabigatu bitartean; eta espainiera-euskara (es-eu) bikoterako itzulpen automatikoko lehenengo sistema sor-

⁴<http://www.opentrad.org>

tzea. Lau unibertsitate eta hiru enpresen arteko elkarlanaren emaitza izan zen proiektu hau⁵.

OpenTrad proiektuko es-eu moduluak eraikitze aurretik garatutako *Matxin* prototipoa hartu eta birdiseinatu eta birprogramatu genituen IXA taldeko hainbat hizkuntzalari eta informatikariok elkarlanean.

Ondoko ataza nagusiak burutu genituen:

- Sistemak prozesatzen dituen datu-egituren diseinu berria.
- Estandarizazioa. Moduluen arteko datu-fluxuaren, baliabide linguistikoek eta erregelen formatuak XML lengoaiak definitu genituen. Hiztegien formatua proiektuko espezifikaziotara egokitu genuen eta hiztegien kontsultarako *Apertium*eko egoera finituetako transduktoreak erabili.
- Preposizio eta funtzio sintaktikoen itzulpenaren desanbiguaziorako moduluaren eraikuntza.
- Datu linguistikoek hedapena eta egiaztapena.
- *Freelingen* hobekuntza, batez ere mendekotasun-analizatzailearen eraikuntzarekin lotuta.
- Kodearen berrinplementazioa C++ lengoaiari.
- Erroreen arazketa eta zuzenketa.
- Banaketa kode irekiko lizentzia batekin.

Proiektu honetan garatutako prototipoa (Alegria *et al.*, 2007) Interneten erabilgarri dago⁶.

III.1.3 *Matxin* 1.0

Opentrad proiektua bukatu ondoren, *Matxin* sistemaren erroreak aztertu eta zuzenketa ugari egin ditugu. Sistema 2007.ko martxoan Eusko Jaurlaritzak

⁵Ikerketa-taldeak hauek dira: Euskal Herriko Unibertsitateko IXA, Alacanteko Unibertsitateko Transducens, Vigoko Unibertsitateko Linguistika Informatikoko Mintegia eta Kataluniako Unibertsitate Politeknikoko TALP. Enpresa arduraduna Eleka Ingeniaritza Linguistikoa da, Elhuyar Fundazioaren zein Galiziako Imaxin Software enpresaren laguntzarekin. Espainiako Industria, Turismo eta Merkataritza Ministerioaren laguntzaz garatu da proiektua.

⁶<http://www.opentrad.org/demo>

antolatutako deialdi batean aurkeztua izan da, harrera onarekin, eta zuzenketa guztiak dituen bertsioa *Matxin* 1.0 bertsio zenbakiarekin banatu egin dugu⁷. *Matxin* sistemaren hobekuntzak, orain, abian jarri den EurOpenTrad proiektuaren barruan egingo dira.

Matxin 1.0 sistemako moduluek erregelen idazketarako formalismo desberdinak erabiltzen dituzte. Tesi honetan berrikuntza garrantzitsu bat proposatzen dugu, haren inplementazioa oraindik burutu ez duguna: erregelen idazketarako formalismo bateratu bat erabiltzea; horrela interpretatzaile bakar batek erregela multzo desberdinak aplikatu ahal izango ditu itzulpenaren faseetan zehar. Hobekuntza honek sistemaren modulartasuna eta hizkuntzekiko independentzia hobetu, arazketan eta erroreen zuzenketa lagundu, eta hedagarritasuna izugarri erraztuko du.

Matxin sistemaren deskribapenean formalismo hau erabiltzen dugu, formalismo bakar bat erabiltzeak tesiaren irakurketa erraztuko duelako. Erregelak adierazteko formalismo horretan erabiltzen den informazio guztia *Matxin* 1.0 bertsioaren erregelek duten informazio bera da. Hala ere, formalismoa etorkizunean ahalmen handiagoko eragiketak ere burutu ahal izateko diseinatu dugu.

III.2 Ezaugarriak

Matxin sistemaren diseinuak *Ariane* eta *Apertium* sistemen ezaugarrietatik edaten du. *Ariane* 80. hamarkadan garatu eta transferentzia sakoneko sistemen eredia bihurtu zen (ikus II.1.3 atala). *Apertium* hizkuntza hurbilen artean itzultzeko kode irekiko sistema da (ikus II.3.3 atala).

Matxinek *Ariane* sistemaren arkitektura eta beste hainbat ezaugarri bereak egiten ditu. Itzulpen-prozesua hiru faseetan banatzen da (analisi, transferentzia eta sorkuntza) eta fase bakoitza hainbat modulutan, modulu bakoitzak ataza zehatz bat burutuz. Sistemaren diseinua guztiz modularra⁸ da eta itzulpen-prozesuko ataza linguistikoek gidatu dute moduluen arteko banaketaren diseinua. Algoritmoak eta datuak banatuta egoteaz gain, formalismo bat diseinatu dugu algoritmoetan datuak manipulatzeko erregelek idazteko eta erregela horiek algoritmoetatik kanpora atera ditugu. Beste helburu bat izan da modulu elebakarrak modulu elebidunetatik ahalik eta

⁷<http://matxin.sourceforge.net>

⁸Modulartasunaren funtsa ebatzi beharreko problema azpiproblemetan banatzea da, algoritmo sinpleekin ebatzi daitezkeen problema bakunak lortuz. Gainera algoritmo sinple horiek inplementatzen dituzten programa txikiak, moduluak, antzeko problemak ebazteko sistemak eraikitze berrirerabili ahal izango dira. Programazio modularrak algoritmoetako erroreak arazten eta zuzentzen laguntzen du, modulu horiek banaka egiazta daitezkeela.

independenteak izatea. Moduluen arteko komunikaziorako interfaze egokia finkatzeak berezko garrantzia duenez, datuen egitura arreta handiz diseinatu dugu.

Apertium sistemaren joerari jarraituz eta berarekin bidea elkarrekin eginenez, estandarizazioa helburu nagusia dugu, baliabide linguistiko guztiek eta moduluen arteko datuen fluxuak XMLn oinarritutako formatu estandar bati jarraitzen diotela. Itzulpen-abiadura handia lortzeko egoera finituetako transduktoreak erabiltzen ditugu. Sistemaren kodea irekia izatea da beste helburu garrantzitsu bat.

Sistema diseinatuta dago erdal hizkuntzetatik euskarara itzultzeko eta arkitektura abiapuntu-hizkuntzatik independentea izateko. Abiapuntu-hizkuntza berri bat sistemari gehitzeko, analisiaren modulua gehitu behar da eta transferentziaren modulueta datu linguistikoak prestatu. Sistemaren xede-hizkuntza, berriz, euskara izateko eraiki dugu momentuz, baina etorkizunean beste hizkuntzetara hedatzea nahi dugu.

III.2.1 Algoritmoak eta datuak banatzea

Jakina denez, sistema informatikoen garapenean, eta are gehiago LNParen arloko ataza konplexuetan, algoritmoak eta datuak banatzea biziki garrantzitsua da. Ezagunak izan arren, horren arrazoiak azalduko ditugu, askotan zailtasunek bultzatzen baitute hain argiak diruditen metodologia-arauak apurtzera:

- Informatikariek garatu ahal dituzte algoritmoak eta hizkuntzalariek datu linguistikoak.
- Erroreen arazketa (*debugging*) errazten da, algoritmoetan edo datuetan dauden erroreak ezberdintzen laguntzen duelako, eta horrela datuetan dauden erroreengatik itzulpen okerrak egiten direnean datuak zuzentzen konpon daitezke.
- Moduluen berrerabilpena aiseago egin daiteke.
- Itzuliko dituzten hizkuntzetatik independenteak izango diren IAko sistemak diseinatu ahal izateko haxe da bidea.

Algoritmoetatik banatzen diren datuak deklaratiiboak izatea gure helburu ere bada. Deklaratibotasuna (*declarativity*) honela definitzen da: ezagutza linguistikoaren deskribapena sistemak erabiliko duen algoritmoen inplementazio zehatzetatik independentea izatea. Informazio deklaratiiboa hainbat

programa ezberdinek modu ezberdinetan erabil dezaketen informazio neutroa da.

Hala ere, sistemak eraikitzean eraginkortasuna eta deklaratiotasunaren arteko konpromisoa bilatu behar da, (Arnold *et al.*, 1994) lanean esaten den bezala:

Despite these advantages of declarativity there is a strong temptation to introduce non-declarative characteristics. (...)

Thus, though declarativity is a generally accepted goal, it is a property that systems have in different degrees, and it is not even generally agreed what the correct compromise between efficiency and declarativity is.

Gure sistema diseinatzean hizkuntzalariek landu beharreko informazio guztia algoritmoetatik atera dugu, modu errazean aldaketak egin ahal izateko, algoritmoak ukitu gabe.

Informazio hori batzuetan deklaratiho hutsa da, eta beste batzuetan informazio prozeduralarekin nahastuta dago. Horren arabera algoritmoetatik ateratako datuak bi multzotan banatu ditugu:

- Datu linguistikoak. Informazio deklaratihoa dute. Bi datu mota bereiz ditzakegu sisteman:
 - Hiztegiak.
 - Corpusetik erauzitako datuak. Postposizioak desanbiguatzeke erabiltzen dira.
- Erregelak. Informazio linguistikoa eta prozedurala erabiltzen dute. Eginbeharreko atazaren arabera, erregelak kodetzeko bi formalismo desberdin diseinatu ditugu:
 - Itzulpen-prozesuko datu-egitura manipulatzeko erregelak. Formalismo bateratu bat diseinatu dugu itzulpen-prozesu osoan zehar beharrezkoak diren manipulazioak definituko dituzten erregelak idazteko. Erregeletako informazioa ez da guztiz deklaratihoa, datu-egituraren gainean irakurketak, aldaketak, gehitzeak, ezabaketak eta sistemako funtzioei deiak ere adieraz ditzakeelako.
 - Aditz-transferentziarako gramatika. Oso prozesu konplexua den aditz-kateen transferentzia estrukturalari aurre egiteko ez da nahikoa datu-egitura manipulatzeko erregelek duten ahalmena. Horregatik, ataza hori burutzeko, gramatika berezi bat diseinatu

behar izan dugu egoera finituetako transduktoreen sintaxia erabiliz (Beesley eta Karttunen, 2003).

III.2.2 Berrerabilgarritasuna

Kode-zatiak edo baliabide linguistikoak, aldaketekin edo gabe, beste ataza, sistema edo aplikazio baterako berriz erabiltzeari berrerabilgarritasuna (*reusability*) esaten zaio. Garrantzitsua da software-osagaiek ondoko ezaugarriak betetzea berrerabilpenerako egokiak izateko: moldaerrazak, malguak, parametrizagarriak, orokorrak, sinpleak, modularrak, egonkorak, hedagarriak, sendoak eta zuzenak.

Garapen-denbora laburtzeko⁹, eta beste batzuek aurretik egindako lana alperrik ez errepikatzeko, berrerabilgarritasuna izan da guretzat sistemaren eraikuntzaren gakoa. Ahalik eta gehien berrerabili nahi izan ditugu aurretik eraikita zeuden moduluak (espainierazko analizatzailea, euskararako sortzaile morfologikoa, desformatatzailea eta birformatatzailea...) eta sortutako baliabide linguistikoak (hiztegiak, corpusak...).

Gure helburua ez da soilik beste batzuek garatutakoa guk berrerabiltzea. Gure lanetik sortutako moduluak eta datu linguistikoak beste ikertzaile batzuek ere beste helburu batzuekin berrerabilgarri izan daitezen eraiki ditugu: espainierazko mendekotasun-analizatzailea, preposizioen hiztegia, sortzaile morfologikoa kode irekian...

Berrerabiltzeak dakarren ondorioetako bat baliabideen eta moduluen heterogeneotasuna denez, modulu berrerabilien arteko elkarrragingarritasuna ziurtatzea nahitaezkoa da.

III.2.3 Elkarrragingarritasuna eta estandarizazioa

Sistema heterogeneo baten barruan oso desberdinak diren zenbait modulu daudenean ataza komun batean elkarrekin lan egiteko, elkarrragingarritasuna (*interoperability*) ezinbestekoa da. Elkarrragingarritasuna esaten zaio sistemak eduki behar duen gaitasunari datuak eta prozesuak moduluen artean elkarri pasa ahal izateko, formatu berdinak irakurtzeko eta idazteko eta protokolo berak erabiltzeko. Elkarrragingarritasunak informazioa berrerabiltzeko aukera maximizatzen du.

Matxin sisteman elkarrragingarritasuna ziurtatzeko oinarritzkoa izan da moduluen arteko datuen fluxuak, baliabide linguistikoek eta datu-egitura

⁹Gure sistema eraikitzeak, hutsetik hasi izan bagina, eta (Boitet, 2001) lanean aipatutako estimazioak zuzenak badira, 40 urte/pertsona beharko genukeen lehenengo bertsio komertziala garatzeko.

manipulatzeko erregelen formalismoak formatu estandar bati jarraitzea, aukeratu dugun formatu estandarra XML izanik.

XML¹⁰ markaketa-metalengoaia hedagarria da lengoaia zehatz desberdinak definitzeko aukera ematen duena. W3C¹¹ erakundeak garatu zuen eta SGML¹² lengoiaren azpimultzo sinplifikatua da. Informazio-sistema desberdinen arteko datu egituratuen elkartrukea erraztea da bere helburu nagusia.

Datu-egiturak manipulatzeko erregelen formalismoa XPath¹³ lengoiaian oinarritzen da. XPath XML dokumentu bateko zatiak atzitzeko lengoaia da, XSLT¹⁴ eta XPointer¹⁵ lengoaiak erabiltzeko diseinatua. XPath lengoiaren gainean XQuery¹⁶ lengoaia eraiki da.

Bestalde, aditz-kateen transferentziarako egoera finituetako transdukto-reen alorrean zabalduea den XFST¹⁷ tresnaren sintaxia erabili dugu.

III.2.4 Kode irekia

Matxin teknologia software librearen lizentziekin banatzeko sortu dugu, kode irekiaren onurak aprobeztatu ahal izateko. Modu honetan beste garatzaile batzuek ere hobekuntzak eta hedapenak egin ditzakete sisteman: inplementazioan egon daitezkeen akatsak zuzendu, beste modulu batzuk integratu, beste estrategiekin konbinatu, erabilera zehatzetarako egokitu, beste hizkuntza-bikoteetarako hedatu...

Modu honetan euskara eta itzulpen automatikoa uztartuko dituzten sistemak eraikitze bidea erraztuko delakoan gaude. Gainera, modu honetan, teknologia hau jende gehiagok erabili ahal izateaz gain, enpresentzat ere erakargarria izango da behar zehatzetarako egokitu ahalko dutelako kode irekiko sistema, datu linguistiko berriak garatuz edo beste aplikazio batzuetan integratuz.

¹⁰ *Extensible Markup Language*. <http://www.w3.org/XML>

¹¹ *World Wide Web Consortium*. <http://www.w3.org>

¹² *Standard Generalized Markup Language*. <http://www.w3.org/MarkUp/SGML>

¹³ *XML Path Language*. <http://www.w3.org/TR/xpath>

¹⁴ XSLT (*XSL Transformations*) XML dokumentuak transformatzeko lengoaia da, XML-rako estilo orrien lengoaia den XSLren (*Extensible Stylesheet Language*) parte bezala erabiltzeko diseinatua izan dena. <http://www.w3.org/TR/xslt>

¹⁵ *XML Pointer*, XML dokumentu bateko zatiak identifikatzeko lengoaia da, estekak sortzeko helburuarekin. <http://www.w3.org/TR/WD-xptr>

¹⁶ XQuery XML galdeketa lengoaia da, iturburu desberdinetako mota desberdineko informazioa berreskuratzeko eta interpretatzeko. <http://www.w3.org/TR/xquery>

¹⁷ *Xerox Finite-State Tool*. <http://www.cis.upenn.edu/cis639/docs/xfst.html>

III.3 Mugak

Gure sistemak itzulpen-kalitateari dagokionez erregeletan oinarritutako beste sistemen mugak ditu. Muga horiei aurre egiteko neurri batean, gure sistema corpusetan oinarritutako teknikekin hibridatzea ikertu nahi dugu modu desberdinetan.

Ondokoak dira, gaur egun, *Matxin* sistemak dituen muga nagusiak:

- Analizatzailearen mugak. Esaldi luzeak itzuli nahi direnean arazo handiak sortzen dira, batez ere koordinazioa edo menpeko perpausak agertzen direnean. Errore hauen iturri nagusia analisi-fasean dagoenez, funtsezkoa da espainierazko analizatzailea hobetzea, agian metodo estatistikoekin konbinatuz.
- Anbiguotasunaren itogunea. Sistemak duen beste arazo handi bat anbiguotasuna da, itzulpen posibleen artean aukeraketa zuzena egitea oso zaila delako kontuan hartu behar baitira ezagutza sintaktikoa, semantikoa, pragmatikoa, estilistikoa eta munduaren ezagutzaren arteko elkarrekintza konplexuak. Hautapen lexikalean ematen den anbiguotasuna ebazteko, gaur egun soilik postposizioen hautaketarako erabiltzen ditugu desanbiguazio teknikak. Beste hitzetarako desanbiguazio-teknikak aztertzea da gure etorkizun hurbileko erronka.
- Xede-hizkuntza. Sistema abiapuntu-hizkuntzatik independentea da, espainieratik eta ingelesatik itzultzeko probatu dugula, baina euskarara itzultzeko diseinatu izan da, eta horretarako soilik izan da probatua.
- Alderanzgarritasuna. Gure sistemaren moduluak ez dira alderanzgarriak.
- Predeterminazioa. Aukera on bat baino gehiago egon daitezkeenean, sistemak bakar bat ematen du eta honek itzulpenetan oso estilo uniforme eragingo du, aldaerarik gabekoa. Hau abantaila dela planteatu daiteke testu teknikoak itzultzean, baina errealitatean emaitza pobretzen ditu.
- Egituraren kontserbazioa. Sistemak askotan abiapuntu-hizkuntzako egitura mantentzen du. Horren ondorioz, zenbaitetan, itzulpenak oso literalak izaten dira.
- Esaldien itzulpena. Sistemak esaldi bakoitza itzultzen du, inguruko esaldiak kontutan hartu gabe.

III.4 Arkitektura orokorra

*Matxin*en arkitektura transferentzian oinarritutako eredu klasikoan oinarritzen da, itzulpen-prozesua hiru fasetan banatzen delarik: analisisa, transferentzia eta sorkuntza.

Erdal hizkuntzako testuen analisi sintaktiko partziala lortzen duen analizatzaile bat berrerabili eta analisi osoa emateko hedatu dugu.

Transferentzia- eta sorkuntza-faseak, berriz, baliabide linguistikoak eta modulu batzuk berrerabiltzen badituzte ere, *Matxin* sistemarentzat diseinatu ditugu. Itzulpen-prozesuko egoera adierazteko modulu guztiek datu-egitura bera erabiltzen dute, III.5 atalean azaltzen dena. Datu linguistikoen eta erregelen formatuak III.6 atalean deskribatzen dira.

Sistemaren arkitekturan (III.1 irudia) ohiko hiru fase horiei desformatatzailea eta birformatatzailea, eta postediziorako moduluak gehitu zaizkie. *Apertium* sistemako desformatatzailea eta birformatatzailea berrerabili ditugu jatorrizko dokumentuen formatua mantentzeko. Sistemaren diseinuan postedizio-moduluaren beharrak kontutan hartu ditugu, modulu hori oraindik inplementatu ez badugu ere.

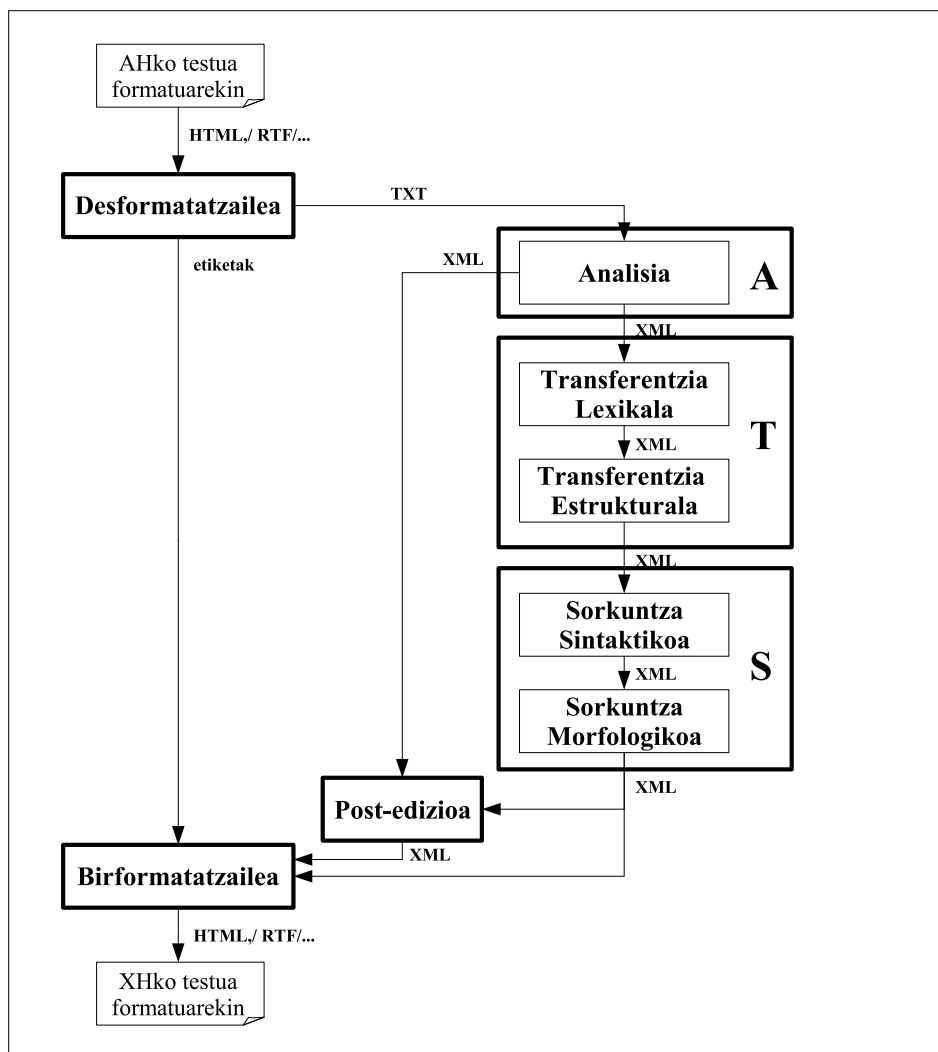
Desformatatzailea

Itzulpen-prozesuaren aurretik desformatatzaile modulua erabiltzen dugu, RTF eta HTML dokumentuen formatua itzulpenaren emaitzan ere mantendu ahal izateko

*Apertium*eko desformatatzailea eta birformatatzailea berrerabili ditugu, baina egokitu behar izan ditugu gure sistemaren prozesaketa konplexuagoa delako. *Apertium*en, hizkuntza hurbilak itzultzean, abiapuntuko hitzen ordena ez da itzulpenean kasik aldatzen, eta beraz, etiketa guztiak bere lekuan mantentzen dira. Gure kasuan, hitzen ordena aldatzen denez, dokumentuko etiketak eta testuan duten kokapenaren erreferentzia gorde eta prozesatu behar da.

Desformatatzaileak jatorrizko fitxategitik bi fitxategi sortzen ditu: testu hutsa duen bat, itzultzerantz bidaltzen duena, eta dokumentuko etiketak dituen beste bat, etiketa bakoitzak testuan duen kokapena gordetzen duena.

Bukaeran birformatatzaileak itzulpen-prozesuaren emaitza eta etiketen fitxategia erabiltzen ditu, dokumentuaren formatua berreskuratzeko.



Irudia III.1: Arkitektura orokorraren eskema

Analisia

Analisi sintaktiko partziala burutzen duen analizatzaile sendo bat berrerabili dugu, zenbait hobekuntza eta egokitzapen burutu behar izan direla. Gure sistemak analisi sintaktiko osoa behar duenez, mendekotasun-analisia egiten duen modulua garatu dugu.

Analizatzailearen sarrera desformatatzailetik datorren testu hutsa da eta bere irteerak *Matxin* sistemaren datu-fluxuaren formatuari jarraitzen dio. Hala izango ez balitz, formatu-bihurketa egin beharko litzateke.

Analizatzaileak nodo bakoitzaren lehenengo karakterea testuan duen posizioa ematen du, itzulpen-prozesuaren bukaeran dokumentuen formatua berreskuratu ahal izateko,

Analizatzailearen ezaugarriak eta mendekotasun-analisia lortzeko burututako lanak IV kapituluan deskribatuko ditugu.

Transferentzia

Transferentziaren helburua abiapuntu-hizkuntzako testuaren adierazpide abstraktua xede-hizkuntzako adierazpidea bihurtzea da. Bi urratsetan egiten da: aurretik transferentzia lexikala eta ondoren estrukturala.

Transferentzia lexikalaren funtsa hizkuntza bateko unitate lexikal bat beste hizkuntzan dagokionarekin ordezkatzeta da, lexikoi elebidunean kontsulta eginez. Lexikoia eraikitzeke estaldura handiko hiztegi elebidun bat eta zenbait hiztegi terminologiko berrerabili ditugu. Lexikoia aberastu da hiztegi elebaker batetik modu erdi-automatikoan erauzitako ezaugarri semantikoaren informazioarekin.

Transferentzia estrukturalak abiapuntu-hizkuntzako egitura xede-hizkuntzarako egokia den egitura bihurtzen du. Hori lortzeko egitura horretako zenbait elementutako informazioa beste batzuei pasatu beharko zaie (adibidez, erdal hizkuntzetako artikuluen mugatasuna edo numeroa, euskarazko sintagman postposizioa izango duen hitzari) eta elementu batzuk ere desagertzen dira (adibidez, artikulua eta preposizioak). Gainera bi prozesu oso konplexu burutu behar dira: preposizio eta funtzio sintaktikoen transferentzia eta aditz-kateen transferentzia.

Transferentzia-fasearen deskribapen sakonagoa V kapituluan irakur daiteke.

Sorkuntza

Sorkuntzaren helburua transferentzia fasean lortutako egituratik abiatuta xede-hizkuntzako testua ematea da. Bi mailatan burutzen da: sintaktikoa

eta morfologikoa.

Sorkuntza sintaktikoak hitzak eta sintagmak xede-hizkuntzako sekuentzia zuzen batean ordenatzen ditu, transferentziatik datorren egituratik abiatuta. Ordenatu ondoren izen-sintagmetan postposizio-informazioa sintagma-ko azken elementuari pasatzen zaio.

Azkenik, sorkuntza morfologikoa egiteko berrerabili dugun euskarazko prozesadore morfologikoak postposizio-informazioa duten hitzetatik hitzen formak lortzen ditu. Postposizio-informaziorik ez dagoenean ez da beharrezkoa, hitzaren forma lema bera delako.

Sorkuntza-fasea sakonago azalduko dugu VI kapituluan.

Postedizioa

Postedizioko moduluaren diseinu osoa eta inplementazioa etorkizuneko lanetarako utzi badugu ere, sistemaren arkitekturan eta datu-egituraren diseinuan kontuan hartu dugu.

Postedizioko moduluaren helburua interfaze grafiko bat eskaintzea da, egindako itzulpena editatu ahal izateko. Horretarako, interfazeak abiapuntu-testua, egindako itzulpena eta bi testuen arteko elementuen (esaldiak, sintagmak eta hitzak) loturak erakutsi beharko lituzke.

Birformatatzailea

Birformatatzaileak bi fitxategi jasotzen ditu: desformatatzaileak emandako etiketen fitxategia eta itzulpen-prozesutik datorrena. Prozesu osoaren emaitza izango den fitxategian idazteko orduan, itzulpena duen fitxategiko elementu bakoitzaren kokapena etiketen fitxategian bilatzen da, posizio horri dagozkien etiketak emaitza-fitxategian gehitzeko.

III.5 Esaldiaren itzulpena prozesatzeko datu-egitura

Gure sistemak prozesatuko duen datu-egitura esaldien egitura sintaktikoan oinarritzen da, osagaien eta mendekotasunen informazioa erabiliz. Datu-egiturak III.5.2 atalean azalduko ditugun hiru objektu-mota erabiltzen ditu: esaldia, chunka eta nodoa. Objektu hauen arteko mendekotasun-erlazioak zuhaitz-egituran adierazten dira.

Diseinatu dugun datu-egitura hau transferentzia- eta sorkuntza-faseetako zenbait moduluek prozesatuko dute, eta moduluen arteko komunikaziorako erabiliko da.

Eragingarritasuna bermatzeko XMLn oinarritutako formatu bat diseinatu dugu, esaldi bakoitzaren itzulpenerako beharrezkoa den informazio guztia edukitzeko ahalmenarekin.

III.5.1 Egitura sintaktikoak: osagaiak vs. mendekotasunak

Egitura sintaktikoa bi modutan adieraz daiteke (Civit, 2003; Calvo eta Gelbukh, 2006), osagaiekin (*constituents*) edo mendekotasunekin (*dependencies*):

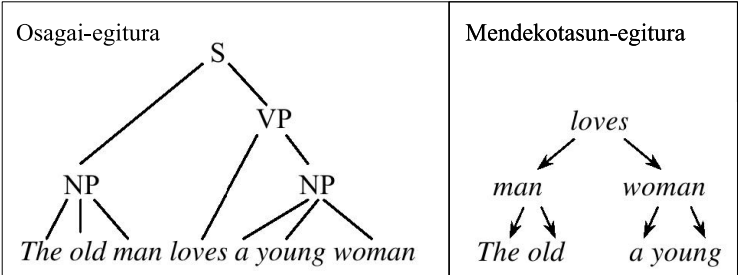
- Osagaien hurbilpenean, esaldiaren egitura deskribatzeko, hitzak osagitan multzokatzen dira, eta osagai bakoitzaren mota zehazten da, gehienetan osagaiko hitz nagusiaren arabera dena. Bi elementu mota erabiltzen dira: amaierakoak (hitzak) eta ez-amaierakoak (osagaiak). Gune-modifikatzaile erlazioa ez da esplizituki azaltzen, baina ez da zaila lortzea; nahikoa da osagai bakoitzeko gunea markatzea.
- Mendekotasunen hurbilpenean, hitza beste hitz baten menpekoa edo modifikatzailea da. Mendekotasun-analisiak amaierako elementuak (hitzak) soilik erabiltzen ditu, haien arteko gune-modifikatzaile erlazioak adieraziz.

III.2 irudian¹⁸ ikus ditzakegu (Calvo eta Gelbukh, 2006) lanean *The old man loves a young woman* esaldirako egituraren adierazpide sinplifikatuak bi formalismoetan. Lan horretan esaten den bezala, ez dago adostasunik formalismorik onena zein den erabakitzeko orduan. Dirudienez hurbilpen bakoitzak bere abantailak ditu ataza desberdinetan, eta badira ere bien abantailak konbinatzeko egitura hibridoak (Brants *et al.*, 1999; Coheur *et al.*, 2004). Mendekotasun-erlazioak eta osagaiak konbinatzen dituzten egiturak itzulpen-sistemetan erabili izan dira, adibidez, *Ariane* sisteman.

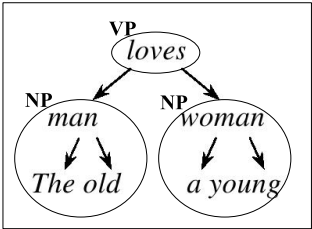
Guk ere, gure sistemarentzat, transferentzia sakoneko itzulpen automatikoko sistemarentzat beharrezkoak diren bi formalismoen ezaugarriak biltzen dituen egitura sintaktiko hibrido bat proposatzen dugu: osagaiak multzokatzen dituen mendekotasun-egitura.

Gure egituran, (Coheur *et al.*, 2004) lanaren hurbilpenean bezala, osagaiak etiketatzen dira eta osagai bakoitzeko hitzen arteko eta osagaien arteko mendekotasun-erlazioak ere adierazten dira. III.3 irudian aurreko adibiderako gure sistemak erabiliko lukeen egitura ikus dezakegu.

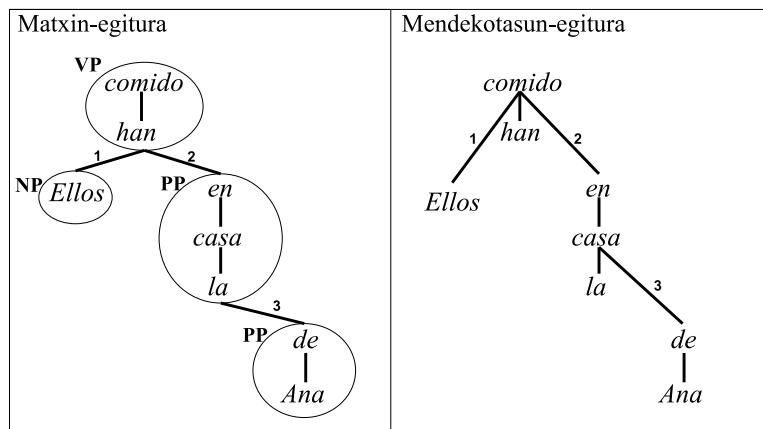
¹⁸S: Esaldia(*Sentence*); NP: Izen-sintagma (*Noun phrase*); VP: Aditz-sintagma (*Verb phrase*);



Irudia III.2: Osagai- eta mendekotasun-egiturak. Adibidea



Irudia III.3: *Matxin* sistemako zuhaitz-egitura. Adibidea.



Irudia III.4: Gure sistemako egituratik mendekotasun-egiturara bihurketa

Egitura sintaktiko honetatik mendekotasun-egitura *tipikoa* lortu nahiko bagenu, egin beharreko prozesua ez litzateke zaila. III.4 irudian¹⁹, adibidez, nahikoa litzateke bi osagaien arteko lotura bakoitza, bi osagai horietako erroak diren hitzen arteko lotura bihurtzea (irudian 1 eta 2 loturak). Diferentia da osagai gurasoa preposizio-sintagma bat denean; horrelakoetan osagai umearen erroa gurasoaren preposizioaren azpiko nodoarekin lotu beharko litzateke (irudian 3 lotura).

III.5.2 Esaldia, chunka eta nodoa

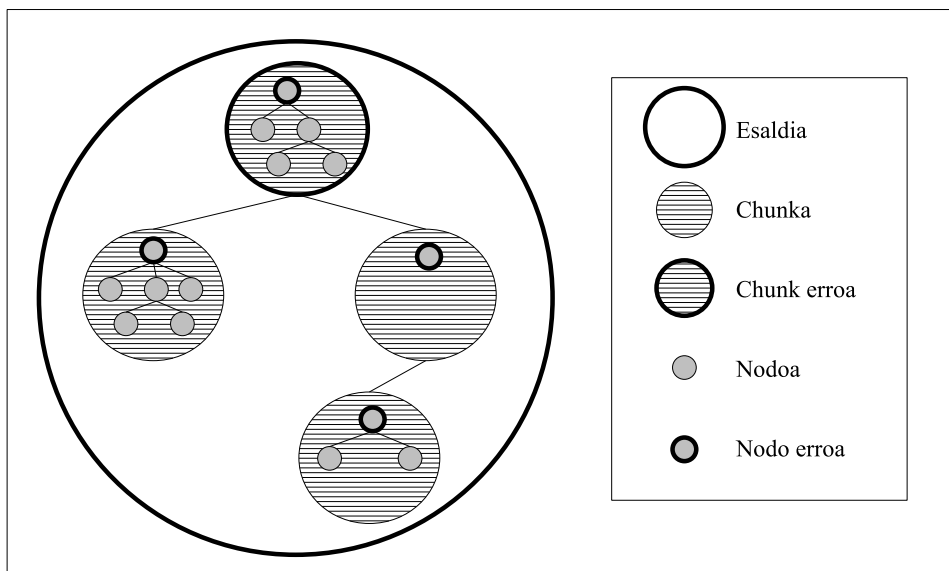
Diseinatu dugun egitura sintaktikoa hiru objektu-mota desberdinekin adierazten da:

- Esaldia. Itzulpenerako unitatea da; sarrerako testua esaldietan banatu ondoren, esaldiak banan bana itzultzen dira, beste esaldiak kontutan hartu gabe.
- Chunka. Osagai bat adierazten duen sasi-sintagma da²⁰. Abney-k (1991) horrela definitzen du:

The typical chunk consists of a single content word surrounded by a constellation of function words, matching a fixed template.

¹⁹PP: preposizio-sintagmam (*prepositional phrase*)

²⁰Literaturan osagaiak hainbat izenekin deituak izan dira: *chunks*, *clusters*, *chains* eta *segments*. Guk *chunk* hitza erabiliko dugu.



Irudia III.5: Mendekotasun-zuhaitza: esaldia, chunkak eta nodoak

Civit-ek (2003) definizio hori zabaltzen du, chunkak sintagma ez-errekurtsiboak direla aipatuz, hau da, bere barnean beste sintagmarik ez dutela. Gainera definizioa ulertzeko chunk bezala hartzen dituzten segmentuak zehazten ditu:

En nuestra propuesta, entendemos por chunks los segmentos nominales (incluidos los adjetivos), los segmentos preposicionales y los segmentos verbales. (...)

Además (...) otros elementos que no quedan incluidos en ninguno de los chunks anteriores, como son los relativos, las formas no personales del verbo, y las conjunciones.

Gure sisteman chunkek garrantzi handia dute prozesamendua errazteko, modulu bakoitzak maila bakar batean lan egiten duelako, chunk barruan edo chunken artean.

- Nodoa. Hitz bat edo hitz anitzeko unitate bat adierazten du.

Objektu mota bakoitzak bere atributuak izango ditu, informazio linguistikoa eta formatu-informazioarekin.

Datu-egituran objektu bakoitzaren mende beste zenbait objektu egongo dira, mendekotasun-zuhaitz bat osatuz (III.5 irudia):

- Esaldi bakoitzaren menpe chunk erroa dago. Chunk erro bat baino gehiago ager daitezke, analisisian chunk horien arteko lotura ezin izan bada ebatzi.
- Chunk bakoitzaren menpe chunk horretako nodo erroa eta chunk horren menpeko chunkak daude.
- Nodo bakoitzaren menpe nodo horren azpiko nodoak daude.

III.5.3 Datu-egituraren formatua

Eragingarritasuna eta estandarizazioa helburu, diseinatu dugun datu-egituraren formatua XMLn oinarrituta dago. DTDan itzulpen-prozesuko hiru elementu nagusiak (esaldiak, chunkak eta nodoak), beren atributuak eta mendekotasun-erlazioak deskribatzen dira.

XML formatuan elementu batek beste bat bere barnean duenean, horrek esan nahi du mendekotasun-egituran bere azpian dagoela. Beraz, esaldi batek bere barnean izango du mendekotasun-zuhaitzean azpian duen chunka (edo bat baino gehiago prozesaketan zehar chunk horien arteko lotura ezin izan bada ebatzi). Chunk batek bere barnean izango du chunkaren erroa den nodoa, eta chunk horren azpian dauden chunkak. Azkenik, nodo batek bere barnean izango ditu zuhaitz-egituran bere azpian dauden nodoak. Elementuen atributuek informazio linguistikoa edo formatukoa adierazten dute.

DTDaren definizioa III.6 irudian dago, non hiru elementu-motetako atributuak definitzen diren. Datu-egituraren formatuko atributuen erabilera hobeto ulertzeko III.7 atalean adibide baten itzulpen-prozesua deskribatuko dugu.

III.6 Datuen formatua

Matxin sistemak erabiltzen dituen datuen mota desberdinak adierazteko, ondoko ataletan deskribatzen ditugun formatuak erabiltzen dira: hiztegiarena, preposizioen hautapen-erregelena, corpusetik erauzitako informazioarena, datu-egitura manipulatzeko erregelena eta aditz-transferentziarako gramatikarena.

III.6.1 Hiztegiak

Transferentzia- eta sorkuntza-faseetan estaldura handiko baliabide linguistikoak berrerrabiliz sortu ditugun hiztegiak erabiltzen dira: lexikoi elebiduna

```

<!ELEMENT SENTENCE (CHUNK+)>
<!--Esaldia-->
<!--ATTLIST SENTENCE
ord      CDATA #IMPLIED <!--zenbagarren esaldia testuan-->
ref      CDATA #IMPLIED <!--XHko esaldiari dagokion AHkoa-->
alloc    CDATA #IMPLIED <!--1.hizkiaren posizioa testuan-->
-->
<!ELEMENT CHUNK (NODE, CHUNK*)>
<!--Chunka-->
<!--ATTLIST CHUNK
ord      CDATA #IMPLIED <!--zenbagarren chunka esaldian-->
ref      CDATA #IMPLIED <!--XHko chunkari dagokion AHkoa-->
alloc    CDATA #IMPLIED <!--1.hizkiaren posizioa testuan-->
type     CDATA #IMPLIED <!--mota-->
si       CDATA #IMPLIED <!--funtzio sintaktikoa-->
focus   CDATA #IMPLIED <!--fokua-->
spost    CDATA #IMPLIED <!--postposizioa-->
spostref CDATA #IMPLIED <!--postposizioa zuen nodoaren ref-->
spostalloc CDATA #IMPLIED <!--postposizioa zuen nodoaren alloc-->
prep     CDATA #IMPLIED <!--preposizioa-->
trans    CDATA #IMPLIED <!--iragankortasuna-->
subper   CDATA #IMPLIED <!--subjektuaren pertsona-->
objper   CDATA #IMPLIED <!--objektuaren pertsona-->
datper   CDATA #IMPLIED <!--zeharkako objektuaren pertsona-->
headlem  CDATA #IMPLIED <!--nodo erroaren lema-->
headpos  CDATA #IMPLIED <!--nodo erroaren kat. eta azpikat.-->
headsem  CDATA #IMPLIED <!--nodo erroaren inf. semantikoa-->
leafpos  CDATA #IMPLIED <!--nodo hostoaren kat. eta azpikat.-->
length   CDATA #IMPLIED <!--luzera, nodo kopurua-->
-->
<!ELEMENT NODE (NODE*)>
<!--Nodoa-->
<!--ATTLIST NODE
ord      CDATA #IMPLIED <!--zenbagarren chunka esaldian-->
ref      CDATA #IMPLIED <!--XHko chunkari dagokion AHkoa-->
alloc    CDATA #IMPLIED <!--1.hizkiaren posizioa testuan-->
form     CDATA #IMPLIED <!--forma-->
lem      CDATA #IMPLIED <!--lema-->
mi       CDATA #IMPLIED <!--informazio morfologikoa-->
pos      CDATA #IMPLIED <!--part-of-speech: kat. eta azpikat.-->
suf      CDATA #IMPLIED <!--atzizkiaren informazioa-->
det      CDATA #IMPLIED <!--mugatasuna-->
num      CDATA #IMPLIED <!--numeroa-->
per      CDATA #IMPLIED <!--pertsona-->
loc      CDATA #IMPLIED <!--kokapen-informazioa-->
lmi      CDATA #IMPLIED <!--lemaren segmentazio morfologikoa-->
sem      CDATA #IMPLIED <!--informazio semantikoa-->
post     CDATA #IMPLIED <!--postposizioa-->
spost    CDATA #IMPLIED <!--menpeko postposizioa-->
vpost    CDATA #IMPLIED <!--aditz-postposizioa-->
prep     CDATA #IMPLIED <!--preposizioa-->
-->

```

Irudia III.6: Moduluen arteko datu-egituraren XML formatuaren DTDa

eta euskararako hiztegi morfologikoa,

Hiztegi horiek *Apertium* (Forcada *et al.*, 2006) proiektuaren espezifikazioari jarraitzen dioten XMLn oinarritutako formatuan kodetuta daude. Proiektu horretako konpiladore batek hiztegi horiek fitxategi bitarrak bihurtzen ditu, oso azkar prozesatuko diren egoera finituetako transduktoreetan oinarritutako adierazpide bat sortuz.

Formatu horrekin kodetutako hiztegi bakoitza hainbat bloketan banatzen da:

- `<sdefs>`: sinboloen definizioen (`<sdef>`) multzoa.
- `<pardefs>`: paradigmaren definizioen (`<pardef>`) multzoa.
- `<section>`: hitzen definizioen (`<e>`) multzoa.

Hitz mota desberdinak multzokatzeko hainbat sekzio egon daitezke hiztegi batean. Hitz bakoitzaren definizioan ondoko elementuak agertzen dira:

- `<e>`: sarrera (*entry*)
- `<p>`: itzulpen bikotea (*pair*)
- `<l>`: ezkerreko elementua (*left*)
- `<r>`: eskuineko elementua (*right*)
- `<s>`: sinbolo bati erreferentzia. Sinboloak *sdefs* atalean definitzen dira.
- `<j>`: elementuetako informazioen juntagailua (*joiner*). Hiztegia irakurtzean “+” ikurrarekin ordezkatzen da.
- `<par>`: paradigma baten erreferentzia. Paradigmek hiztegiaren edukia trinkotzen dute eta haien definizioak *pardefs* atalean definitzen dira, hitzen definizioetako elementu berak erabiliz.

Ondoko ataletan formatu honi jarraitzen dioten hiztegi morfologikoaren eta lexikoi elebidunaren adibide batzuk erakutsiko ditugu.

Hiztegi morfologikoa

Hiztegi hau sorkuntza morfologikoan erabiltzen da. Hona hemen hiztegi morfologikoaren *section* blokeko sarrera bat eta erabiltzen duen paradigmaren definizioa:

```

<e><p> <l>etxe</l>
      <r>etxe<s n='pos' />[IZE][ARR]</r></p>
      <par n="I_31"/>
</e>

<pardef n="I_31">
  <e><p> <l>a</l>
      <r><s n='post' />+[DEK][ABS][NUMS]</r></p></e>
  <e><p> <l>aren</l>
      <r><s n='post' />+[DEK][GEN][NUMS]</r></p></e>
  ...
</pardef>

```

Bikoteko ezkerreko elementuak postposizioak itsatsiko diren hitzaren sasi-lema (*etxe*) adierazten du, eta eskuinekoa hitzaren lema (*etxe*) eta kategoria eta azpikategoria²¹ (*[IZE][ARR]*). Horrez gain, bikote bakoitzerako informazioa zabaltzen duen paradigma bat ematen da eta hainbat sarrerek paradigma bera erabiltzen dutenez hiztegiko informazioa trinkotzen lortzen da. Paradigmako ezkerreko elementuan postposizioaren forma ematen du (*a/aren*) eta eskuineko elementuan informazio morfologikoa (*+ [DEK][ABS][NUMS] / + [DEK][GEN][NUMS]*). Ondokoa da, beraz, aurreko adibidean ematen den informazioa:

```

etxea      ↔  etxe<pos>[IZE][ARR]<post>+[DEK][ABS][NUMS]
etxearen   ↔  etxe<pos>[IZE][ARR]<post>+[DEK][GEN][NUMS]

```

Gure sistemak euskararen hiztegi morfologikoa sorkuntzarako erabiltzen duenez, eskuinetik ezkerrera aplikatuko du.

Lexikoi elebiduna

Lexikoi elebiduna transferentzia lexikalean erabiltzen da, abiapuntu-hizkuntzako unitate lexikal bakoitzari dagokion xede-hizkuntzako ordaina emateko.

Paradigmen jatorrizko erabilera *Apertium*en analisia eta sorkuntza kontrolatzea bada ere, gure sisteman paradigmak ere lexikoi elebidunean informazio morfologikoaren transferentzia egiteko erabiltzen ditugu, hiztegiaren edukia trinkoagoa bihurtuz.

Adibide gisa, ikus dezagun lexikoi elebiduneko sarrera bat, eta erabiltzen duen paradigma:

```

<e><p> <l>beso</l>
      <r>musu<s n='sem' />[BIZ-]</r></p>
      <par n='NC_IZE_ARR' />
</e>

```

²¹Euskararako kategoria-sistema C eranskinean ikus daiteke.

```

<pardef n='NC_IZE_ARR'>
  <e><p> <l><s n='mi' />NCFP000</l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMP]</r></p></e>
  <e><p> <l><s n='mi' />NCFS000</l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMS]</r></p></e>
  <e><p> <l><s n='mi' />NCFP00D</l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMP]
      <s n='suf' />txo[ATZ][IZE][ARR]</r></p></e>
  <e><p><l><s n='parol' />NCFP00D</l>
    ...
</pardef>

```

Bikoteko ezkerreko elementuak espainierazko hitzaren lema (*beso*) adierazten du, eta eskuinekoak euskarazko lema (*musu*) eta beste informazio batzuk, adibide honetan biziduntasunaren informazio semantikoa (*[BIZ-]*). Gainera paradigma bat ematen da, non ezkerreko elementuaren *Parole* formatuko informazio morfologikoaren²² arabera eskuineko elementuetarako hainbat informazio ematen diren: kategoria eta azpikategoria (*pos*), informazio morfologikoa (*mi*), atzizkien informazioa (*suf*)...

Hauxe da aurreko adibidean ematen den informazioa:

beso<mi>[NCFP000]	↔	musu<pos>[IZE][ARR]<num>[NUMP]<sem>[BIZ-]
beso<mi>[NCFS000]	↔	musu<pos>[IZE][ARR]<num>[NUMS]<sem>[BIZ-]
beso<mi>[NCFP00D]	↔	musu<pos>[IZE][ARR]<num>[NUMP]<suf>txo[ATZ][IZE][ARR]<sem>[BIZ-]

Lexikoi elebiduna eraiki dugu espainieratik euskarara itzultzeko, beraz ezkerretik eskuinera aplikatuko da.

III.6.2 Preposizioen hiztegia eta hautapen-erregelak

Transferentzia estrukturean preposizio eta funtzio sintaktikoen itzulpenearako eskuz eraiki dugun preposizioen hiztegi honetan *Apertium*eko espezifikazioan egokitzen bat burutu dugu, hiztegian hautapen-murritzpenak definitu ahal izateko. Horretarako sarrera bakoitzeko itzulpen bikotean *cond* elementu bat gehitu dugu, non hautapenerako baldintza eta baldintza horren mota (*baladin* edo *bsb*) adierazten diren:

```

<e><p> <l>de</l>
  <r>[INS]+[GEL]</r>
  <cond type='bsb'>./[@headsem='[GAI+]']</cond>
</p></e>
<e><p> <l>de</l>
  <r>[INS]</r>

```

²²Espainierako *Parole* etiketa multzoa B eranskinean ikus daiteke.

```

<cond type='baldin'>./[@headsem='[BIZ+]' ] ||
                        [@headpos='[IZE][IZB]'] </cond>
</p></e>

```

Adibidean ikusten dugunez *de* preposizioak *[INS]+[GEL]* postposizioekin itzuliko da (*de madera* → *egurrezko*) baldin eta soilik baldin (*bsb*) chunkaren erroa gai bat bada (*[GAI+]*), eta *[GEN]* postposizioarekin (*de Mikel* → *Mikelen*) baldin (*baldin*) chunkaren erroa izaki bizidun bat (*[BIZ+]*) edo pertsona-izen berezi bat (*[IZE][IZB]*) bada.

Erregeletako baldintzak XPath-en sintaxia (ikus III.6.4 atala) erabiltzen dute: desanbiguatze postposizioa duen uneko chunka “.” ikurrarekin adierazten da, eta bere gurasoarena “..” ikurrarekin. Baldintzetan eragile logikoak erabil daitezke.

III.6.3 Corpusetik erauzitako informazioa

Transferentzia estrukturalen postposizioak desanbiguatu ahal izateko, hautapen-erregelak dituen preposizioen hiztegiak gain, corpusetik erauzitako informazioa erabiltzen dugu: aditzen azpikategorizazioa (ikus V.2.3 atala). Informazio hori XMLn oinarrituta diseinatu dugun formatu batekin kodetzen dugu, hurrengo adibidean ikus dezakegun bezala:

```

<verb>filtratu
  <pattern ord=1>
    <tr>DU</tr>    <sub>ERG</sub>    <post>ABS    </post></pattern>
  <pattern ord=2>
    <tr>DU</tr>    <sub>ERG</sub>    <post>ABS-ALA</post></pattern>
  <pattern ord=3>
    <tr>DA</tr>    <sub>ABS</sub>    <post>          </post></pattern>
  <pattern ord=4>
    <tr>DIO</tr>   <sub>ERG</sub>    <post>ABS    </post></pattern>
  ...
</verb>

```

Aditz (*verb*) bakoitzerako patroi (*pattern*) posibleak ematen dira, corpusen duten maiztasunaren araberrako ordenaren (*ord*) informazioarekin. Patroi bakoitzean iragankortasuna (*tr*), subjektuaren postposizioa (*sub*) eta aditz-katearen osagaien postposizioen zerrenda (*post*) agertzen da. Adibidez *filtratu* aditzerako lehenengo patroian aditzaren iragankortasuna *DU* da, subjektuaren postposizioa ergatiboa (*ERG*) eta osagai bakarraren postposizioa absolutiboa (*ABS*).

III.6.4 Datu-egitura manipulatzeko erregelak

Matxin sistemaren transferentzia eta sorkuntzako moduluetan (ikus V eta VI kapituluak) datu-egitura manipulatzeko beharrezkoak diren eragiketa guz-

<!ELEMENT	rule	(match, actions)>	<!--Erregela-->
<!ATTLIST	rule		
	id	CDATA	#REQUIRED
>			
<!ELEMENT	match	(def+)>	<!--Kidetze-patroia-->
<!ELEMENT	def>		<!--Elementu-definizioa-->
<!ELEMENT	actions	(act+)>	<!--Ekintzen multzoa-->
<!ELEMENT	act>		<!--Ekintza-->

Irudia III.7: Datu-egitura manipulatzeko erregelen XML formatuaren DTDa

tiak algoritmoetatik banatu ditugu. Horretarako, eta *Ariane* sistemak erabiltzen duen produkzio-sistemaren antzera (ikus II.1.3 atala), egitura konputazionalak bere barnean hartzen ditu: datu-egitura, erregela multzo bat eta erregela horiek aplikatuko dituen interpretatzailea.

Hizkuntzalariak zehaztu dituzte datu-egituran bilatu behar diren konfigurazioak eta, konfigurazio horiek aurkitzen badira, burutu beharreko al-daketak. Erregelen idazketan, teorian, informazio deklaratihoa soilik idatzi beharko balitz ere, praktikan eta eraginkortasunagatik, prozedurazkotzat har dezakegun informazioa ere erabiltzen da (datu-egiturako elementuen ezabaketa, kanpo-funtzioei deiak...). Adibideak III.2 taulan ikus daitezke.

Ez dago erregela multzo bakar bat, itzulpen-prozesuko modulu bakoitzak bere erregela multzoa dauka eta hizkuntzalariak azpimultzo bakoitza noiz aplikatzen den erabakitzen du²³.

Erregelak definitu dugun formalismoan idazten dira eta erregelen aplikazioaz interpretatzailea arduratzen da.

Zuhaitzen transdukziorako formalismoa

Erregelen idazketarako XMLn oinarritutako formatu bat definitu dugu (ikus DTDa III.7 irudian). Erregela (*rule*) bakoitzak bi atal dauzka:

- **<match>**: kidetze-patroia (*pattern match*). Datu-egituran bilatu behar den konfigurazioa adierazteko, konfigurazio horretako elementuaren (edo elementu multzoaren) definizioa (*elem*) ematen da, elementu bakoitzaren definizioan aurretik definitutako elementuen erreferentzia erabil daitezkeela.

²³Horrek ere zalantzan jartzen du ezagutza linguistikoaren eta algoritmoen arteko banaketa, baina itzulpen automatikoko prozesuak oso konplexuak dira, eta problemari aurre egiteko beharrezkoa da hainbat modulutan banatzea eta modulu horietako bakoitzak bere erregela multzoa izatea.

```

<rule id='c_n_post_up'>
  <match>
    <def> C := //CHUNK[@type!='adi-kat']      </def>
    <def> N := C/NODE//NODE[@post]           </def></match>
  <actions>
    <act> C/@post := N/@post                  </act></actions>
</rule><!------->
<rule id='c_c_spost_del'>
  <match>
    <def> C1 := //CHUNK[not(./NODE)][./@spost] </def>
    <def> C2 := C1/CHUNK                      </def></match>
  <actions>
    <act> C2/@spost:= C2/@spost                </act>
    <act> delete (C1)                          </act></actions>
</rule><!------->

```

Irudia III.8: Datu-egitura manipulatzeko erregelen adibideak

- **<actions>**: ekintza-multzoa. Ekintza (*act*) bat edo gehiago agertuko dira. Ekintzetan kidetze-patroian definitutako elementuen erreferentziak erabiltzen dira.

Erregela identifikatzeko *id* atributua erabiltzen da, arazketan erregelen aplikazioaren traza lagungarri izango delako.

Kidetze-patroiak elementuak definitzeko XPath²⁴ lengoaiaren sintaxia-ri jarraitzen dioten espresioak erabiltzen ditu. Interpretatzaileak XPath-eko espresio horiek ebaluatuz datu-egituratik konfigurazioa osatzen duten elementuen erreferentziak jasotzen ditu, eta elementu horien gainean erregelen zehaztutako ekintzak burutzen ditu. III.8 irudian zenbait erregela ikus ditzakegu.

Kidetze-patroien sintaxia

Erregeletako kidetze-patroian bilatu nahi den konfigurazioa osatzen duten elementuen (edo elementu multzoen) definizioak agertzen dira.

Elementuen definizioan bi atal agertzen dira “:=” ikurraz banatuta: elementuaren (edo elementu multzoaren) identifikadorea²⁵ eta elementua (edo

²⁴<http://www.w3.org/TR/xpath>

²⁵Elementuaren identifikadoreen izenak adierazgarriak izateko, ondoko konbentzioa jarraitzen dugu: C ikurra datu-egiturako chunkak adierazteko eta N ikurra nodoak adierazteko.

	Kokapen-bidea	Azalpena
(1)	C/NODE	C elementuaren umeak diren nodoen multzoa
(2)	/SENTENCE	Zuhaitzaren erroaren azpian dagoen esaldia
(3)	/SENTENCE/CHUNK	Zuhaitzaren erroaren azpian dagoen esaldiaren umeak diren chunken multzoa
(4)	//CHUNK	Zuhaitzaren erroaren azpian dauden ondorengo chunk guztiak (hau da, chunk guztien multzoa)
(5)	C//NODE	C elementuaren ondorengo nodoen multzoa
(6)	C/NODE//NODE	C elementuaren umeak diren nodoen ondorengo nodoen multzoa
(7)	C/@num	C elementuaren <i>num</i> atributua
(8)	//*	Zuhaitzeko elementu guztien multzoa
(9)	//*[@ord]	<i>ord</i> atributua duten elementuen multzoa
(10)	C/NODE[@num]	C elementuaren umeak diren nodoen multzoa, <i>num</i> atributua dutenak
(11)	//CHUNK[@type!='adi-kat']	<i>type</i> atributua duten chunken multzoa, atributu horretako balioa <i>adi-kat</i> ez dela
(12)	//*[@num='1' or @num='2']	<i>num</i> atributua duten elementuen multzoa, atributu horretako balioa 1 edo 2 izanik
(13)	//NODE[@det] [@num]	Nodo guztien artean, <i>det</i> eta <i>num</i> atributuak dituztenen multzoa
(14)	C/NODE//NODE[@det] [1]	C elementuaren umeak diren nodoen <i>det</i> atributua duen lehenengo ondorengoa
(15)	C//NODE[@ord=count(C//NODE)]	C elementuaren ondorengo nodoa, <i>ord</i> atributua duena, bere balioa izanik C elementuaren ondorengo nodoen kopuruaren berdina

Taula III.1: Kokapen-bideen adibideak

elementu multzoa) hautatzeko kokapen-bidea, XPath lengoaiaz definituta²⁶.

Kokapen-bidea (*location path*) elementu-multzo bat hautatzen duen espresioa da, “/” ikurraz banatutako kokapen-urratsen (*location step*) sekuentzia bat adierazten duena. Kokapen-bideen adibideak eta bere esanahia III.1 taulan ikus daitezke²⁷.

Kokapen-bidea erlatiboa izango da espresioaren ezkerrean hasierako-testuingurua izango den elementu bat agertzen bada (1). Absolutua izango da, berriz, “/” ikurrarekin hasten bada (2), orduan hasierako testuingurua zuhaitzaren erroa izango dela (datu-egiturako elementu erroa zuhaitzaren

²⁶Kokapen-bideen sintaxia deskribatzeko ondoko lerrotan erabiltzen dugun terminologia XPath lengoaiaren oinarrituta dago, baina sinplifikatu dugu eta zenbaitetan kontzeptuen izenak moldatu ditugu, adibidez *Matxin* sistemaren datu-egiturako *nodoen* eta XPath lengoaiako *nodoen* artean nahasmenak ez sortzeko.

²⁷Atal honetako azalpenetan parentesien arteko zenbakiekin adieraziko ditugu taula horretako adibideak.

erroaren azpian dago).

Kokapen-urratsak ezkerretik eskuinera osatzen dira, urrats bakoitzak testuinguru-elementu bati dagozkion elementuen multzoa hautatzen duela. Multzo horretako elementu bakoitza hurrengo urratserako testuingurua izango da. Kokapen-urrats batean ardatza, elementu-mota eta predikatuak (0, 1 edo gehiago) agertzen dira.

Ardatz batek testuinguru-elementuaren eta kokapen-urratsean hautatutako elementuen arteko erlazio hierarkikoa zehazten du:

- `"/"`: umeak (3)
- `"/"`: ondorengoak, hau da, umeak eta umeen umeak, eta abar (4,5,6)
- `"@"`: atributua (7)

Ardatzaren ondoren hautatu nahi diren elementuen mota edo izena agertzen da. Elementu guztiak hautatu nahi direnean `*` ikurra erabiliko da (8.9).

Hautatu nahi diren nodoen multzoa gehiago murrizteko predikatuak erabiltzen dira, hautatutako nodoek betebeharreko baldintza kortxeteen artean idatziz (10,11). Eragiketa logikoak adieraz daitezke *and*, *or* eta *not* eragileak ere erabiliz (12), *and* eragiketa logikoa ere hainbat predikatu jarraian emanez adieraz daitekeela (13).

Predikatuan zenbaki bat kortxeteen artean idatziz elementuen posizioa adierazten da (14). Funtzio konplexuagoak ere erabil daitezke, adibidez, `count()` hautatutako elementu-kopurua adierazteko (15).

XPath lengoaiak eskaintzen dituen oinarritzko tresnak soilik aurkeztu ditugu, baina kokapen-bideak adierazteko XPath lengoaiaren ahalmen guztia erabil genezake.

Ekintzak

Ekintzek oinarritzko eragiketak eta kanpo-funtzioei deiak egikaritzen dituzte, kidetze-patroian definitutako elementuen identifikadoreak erabiliz. Ekintzen adibideak III.2 taulan ikus ditzakegu²⁸.

Oinarritzko eragiketak ondokoak dira:

- `“:=”`: esleipena (1,2)
- `delete()`: ezabaketa, (3,4)
- `concat()`: kateaketa, (5)

²⁸Atal honetako azalpenetan taulako adibideei parentesien arteko zenbakiekin erreferentzia egiten zaie.

Kokapen-bidea	Azalpena
(1) <code>E/@ref := E/@ord</code>	E multzoko elementu bakoitzaren <i>ref</i> atributuari E elementu beraren <i>ord</i> atributuaren balioa esleitu
(2) <code>C/@num := N/@num</code>	C multzoko chunk bakoitzaren <i>num</i> atributuari dagokion N nodoaren <i>num</i> atributuaren balioa esleitu
(3) <code>delete(N)</code>	N multzoko nodoak ezabatu
(4) <code>delete(N/@lem)</code>	N multzoko nodoen <i>lem</i> atributua ezabatu
(5) <code>N/@loc := concat(N/@loc, '[1]')</code>	N multzoko nodo bakoitzaren <i>loc</i> atributuari atributu horren balioa eta '[1]' katearen kateaketa esleitu
(6) <code>\$Target := &dictlook(N/@lem,N/@mi)</code>	N multzoko nodo bakoitzarentzat, <i>\$Target</i> aldagaian jaso <i>&dictlook</i> funtzioaren emaitza, sarrera bezala nodoaren <i>lem</i> eta <i>mi</i> atributuak dituen
(7) <code>&getattribs(N,\$Target)</code>	N multzoko nodo bakoitzarentzat <i>&getattribs</i> funtzioari deia, sarrera bezala nodoaren erreferentzia eta <i>\$Target</i> aldagaia dituen. Funtzio horrek <i>\$Target</i> aldagaian zehazten diren balioak dagozkien N nodoko atributuetan idatziko ditu
(8) <code>\$Orders := &precedence({N/@loc},\$Rule)</code>	<i>&precedence</i> funtzioari deia, sarrerako parametroak izanik N multzoko nodo guztien <i>loc</i> atributuaren balioa eta <i>\$Rule</i> aldagaia, funtzioaren irteera <i>\$Orders</i> aldagaian gordetzen dela
(9) <code>&getOrders({N/@ord},\$Orders)</code>	<i>&getOrders</i> funtzioari deia, sarrerako parametroak izanik N multzoko nodo guztien <i>ord</i> atributuen erreferentzia eta <i>\$PrecRule</i> aldagaia. Funtzio horrek <i>\$Orders</i> aldagaian zehazten diren balioak dagozkien N multzoko nodoen <i>ord</i> atributuan idatziko ditu

Taula III.2: Ekintzen adibideak

Gainera, ekintzetan kanpo-funtzioak (“&” ikurrarekin) deitu ahal izango dira, eta aldagaiak (“\$” ikurrarekin) erabiltzeko aukera ere ematen da, adibidez funtzioek itzultzen duten balioa gordetzeko (6,7).

Ekintza batean elementu-identifikadoreak elementu multzo bat adierazten badu, orduan multzo horretako elementu guztientzat errepikatzen da ekintza (1-7).

Ekintza berean identifikadoreak adierazten duen multzoko elementu guztien balioak erabili nahi badira, giltzekin (“{”, “}”) adierazten da (8,9).

Interpretatzailea

Interpretatzaileak kidetze-patroian agertzen diren elementu multzoen erreferentziak jasotzen ditu, elementuen definiziotan agertzen diren XPath lengoaiako kokapen-bideak erabiliz.

Ondoren, jasotako elementuen eta elementu multzoen erreferentziak erabiliz, interpretatzaileak ekintzak burutzen ditu, datu-egituran eginbeharreko irakurketak, idazketak eta ezabaketak burutuz, eta beharrezkoa denean kanpo-funtzioei deiak eginez.

III.6.5 Aditz-kateen transferentziarako formalismoa

Aditz-kateen transferentziarako gramatikak XFST²⁹ tresnaren sintaxia erabiltzen du. Gramatikako erregelek, modu sekuentzialean, aditzen transferentzia burutzeko beharrezkoa den informazio guztia duen karaktere-katearen gainean, adierazpen erregularrak erabiltzen dituzten baldintzapeko ordezkapenak egiten dituzte (Beesley eta Karttunen, 2003). Ordezkapen guztiak egin ondoren aditz-katearen egitura berria lortzen da.

Gramatikaren erregeletan erabiliko diren sinboloak definitu ondoren sarrrerako karaktere-katea moldatuko duten gramatikako erregelak kodetzen dira. Erregelek *baldintzapeko ordezkapen-eragilea* (*conditional replacement operator*) erabiltzen dute, ondoko formatua duena:

```
[Patroia->Ordezkapena || EzkerrekoTestuingurua _ EskuinekoTestuingurua]
```

Patroia dagoen lekuan *Ordezkapena* idatziko da, baldin eta *Ezkerreko-Testuingurua* eta *EskuinekoTestuingurua* betetzen badira. Adierazpen erregularrak erabil daitezke testuinguruak adierazteko. Gramatikako erregelen adibideak III.9 irudian ikus daitezke.

Aditz-kateen transferentziarako modulua V.2.4 atalean azalduko dugu.

²⁹ *Xerox Finite-State Tool*. <http://www.cis.upenn.edu/cis639/docs/xfst.html>

```

!Definizioak
define Perts      [ %0 | 1 | 2 | 3 ];
define Num        [ %0 | s | p ];
define Sub        [ "[" {sub} Perts Num "]" ];

!Ordezkapen erregelak
[ {AspN}->"EZBU"      || ?* [VMIP|VMII]      ?* XXX [{AR}|{P3}] ?* _ ?* ]
[ {AspN}->"ADOIN"     || ?* [VMSP|VMSI|VMM] ?* XXX [{AR}|{P3}] ?* _ ?* ]

[ {Per}->"etorri"     || ?* [{.venir.}]          ?* _ ?* ]
[ {Per}->"jarraitu"   || ?* [{.seguir.}|{.continuar.}] ?* _ ?* ]

[ {DA}   -> "DU"      || ?*                _ ?* XXX {P} ?* PerifEdun ?* ]
[ {ZAI0} -> "DIO"     || ?*                _ ?* XXX {P} ?* PerifEdun ?* ]
[ {DU}   -> "DA"      || ?*                _ ?* XXX {P} ?* PerifIzan ?* ]

[ {DenMod}->"[A3]"    || ?*   VMSP                ?* XXX ?* _ ?* ]
[ {DenMod}->"[B2]"    || ?* [ VMIC | VASI ]          ?* XXX ?* _ ?* ]

```

Irudia III.9: Aditz-kateen transferentziarako gramatika. Adibideak

III.7 Itzulpen-prozesua: adibide bat

Matxin sistemaren arkitektura eta datu-egitura hobeto ulertzeko itzulpen-prozesua burutzen duten moduluen lana eta beren arteko datu-fluxuen formatua erakutsiko dugu ondoko adibidea erabiliz: *Un tribunal niega los derechos constitucionales a los presos políticos.*

Prozesuaren traza ikusteko, moduluen arteko komunikaziorako *Matxin* sisteman diseinatu dugun datu-egiturarako XML formatuaz gain, informazio hori errazago ulertzeko helburuarekin diseinatu den XSLT estilo-orri batek emandako irudia erakutsiko dugu.

Analisi-fasearen irteera III.10 irudian ikus dezakegu. Analizatzaileak AHko testuarentzat emandako informazioan elementu bakoitzaren ordena *ord* atributuan ikusten da, eta nodoetan jatorrizko testuan nodoaren lehenengo hizkia okupatzen duen posizioa *alloc* atributuan. Bestelako atributuak sarre-rako chunk eta nodoen informazio linguistikoa adierazten dute. Informazio morfologikoak (*mi*) *Parole* notazioari jarraitzen dio (ikus B eranskina).

Transferentzia lexikalaren ondoren (III.11 irudia) honako aldaketak ikus ditzakegu: *ord* atributuan zegoena *ref* atributuan gordetzen da jatorrizko testuaren ordenaren erreferentzia erabili ahal izateko postedizioan. *ord* eta *form* atributuak desagertzen dira, sorkuntza-fasean XHko testurako ordena berria kalkulatu eta XHko hitzen forma eman beharko delako. Lexikoi elebidunean kontsulta eginez AHko nodoetako *lem* eta *mi* atributuen trans-

ferentzia lexikalak XHko *lem*, *pos det*, *num* eta *sem* atributuen balioak eman ditu. Chunketako *type* atributua ere itzuli egin da.

Transferentzia estrukturalean (III.12 irudia) balio lexikalik gabeko nodoak (preposizioak eta artikulua) desagertu dira, zeukaten mugatasunaren, numeroaren eta preposizioaren informazioa chunkari pasatuz. Gainera, desagertu diren nodo horien erreferentziak (*ref* eta *alloc*) chunkari ere pasatuko zaizkio (*postref* eta *postalloc* atributuetan) postedizioan eta birformatatzaillean beharrezkoak izango direlako. Preposizio eta funtzio sintaktikoak itzuli dira, chunketako postposizioa (*post*) lortuz. Aditz-kateen transferentzia burutzean, egitura berri bat duen chunka sortzen da, nodo bakoitzari esleitzen zaiola ordenaketarako beharrezkoa duen informazioa (*loc*), eta sorkuntza morfologikorakoa (*post*).

Sorkuntzaren ondoren (III.13 irudia) kalkulaturako ordena berria *ord* atributuan agertzen da. Chunkeko azken hitzari sorkuntza morfologikoa burutzeko postposizio-informazioa *post* pasatu zaio eta informazio horrekin lotura zuten desagertutako nodoen erreferentziak. *form* atributuan nodo bakoitzaren xede-hizkuntzako hitz-forma jasotzen da, nodoak postposizio-informazioa bazuen prozesadore morfologikoak emandako forma, eta bestela zuzenean lemaen informazioa.

Adibide honetarako sistemak ematen duen irteera hauxe da: *Auzitegi batek eskubide konstituzionalak ukatzen dizkie preso politikoei*

III.8 Laburpena

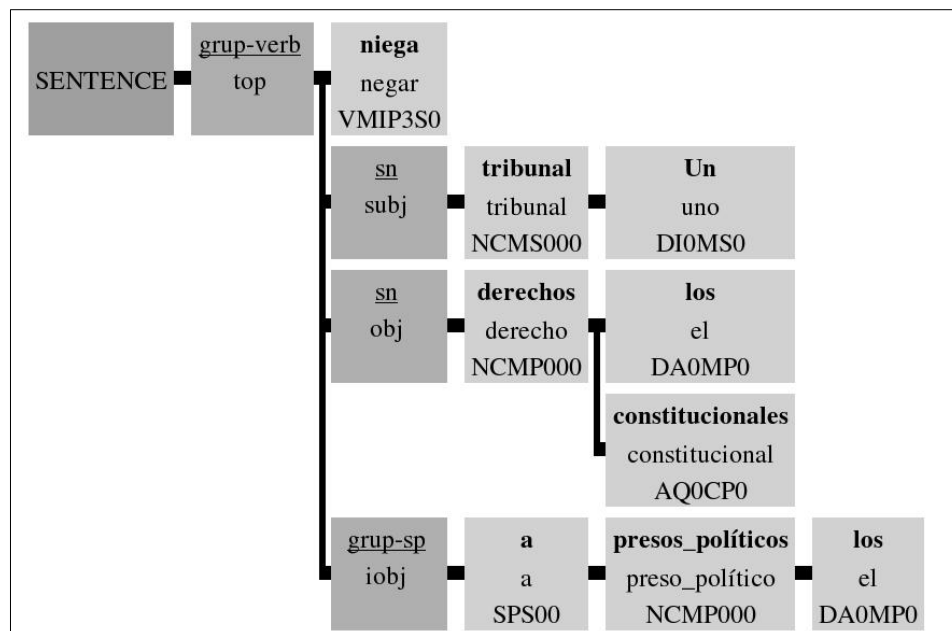
Kapitulu honetan *Matxin* teknologiaren nondik norakoak eta sistemaren deskribapen orokor bat egin dugu.

Sistemaren bilakaera laburki kontatu ondoren bere ezaugarriak eta mugak azaldu ditugu. Sistemaren arkitektura orokorra aurkeztu dugu eta itzulpena prozesatzeko datu-egitura eta sistemak erabiltzen dituen datuen formatua deskribatu ditugu.

Azkenik, itzulpen-prozesu osoa nola burutzen den hobeki ulertzeko adibide bat eman dugu.

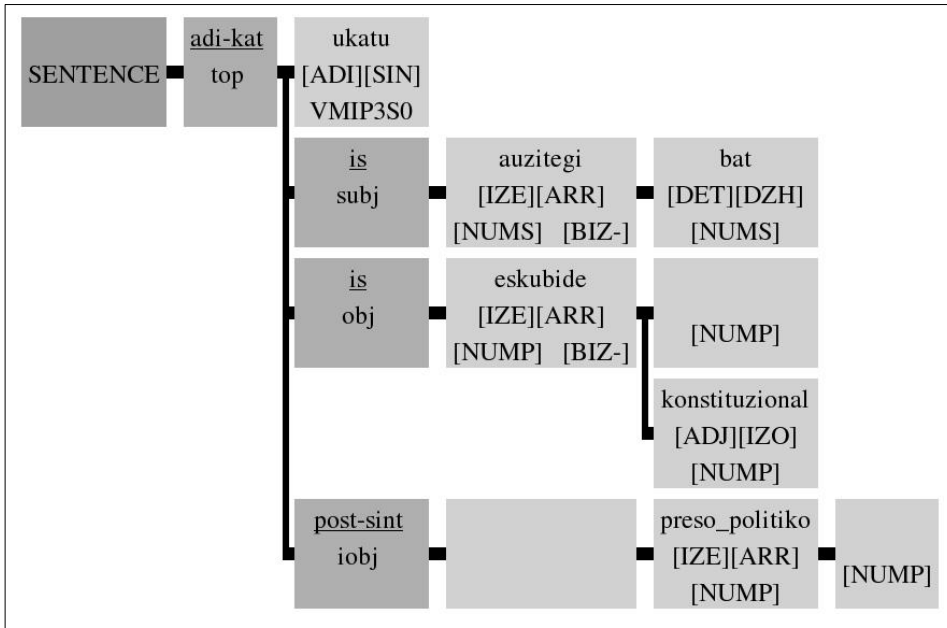
Ondoko hiru kapituluetan sistemaren hiru fase nagusien zehaztasunak sakonki aurkezten ditugu: analisia, transferentzia eta sorkuntza.

Un tribunal niega los derechos constitucionales a los presos políticos



```
<SENTENCE ord='1' alloc='0'>
  <CHUNK ord='2' alloc='12' type='grup-verb' si='top'>
    <NODE ord='1' alloc='12' form='niega' lem='negar' mi='VMIP3S0'>
      <CHUNK ord='1' alloc='0' type='sn' si='subj'>
        <NODE ord='2' alloc='3' form='tribunal' lem='tribunal' mi='NCMS000'>
          <NODE ord='1' alloc='0' form='Un' lem='uno' mi='DI0MS0'>
            </NODE>
          </CHUNK>
        <CHUNK ord='3' alloc='18' type='sn' si='obj' focus='true'>
          <NODE ord='2' alloc='22' form='derechos' lem='derecho' mi='NCMP000'>
            <NODE ord='1' alloc='18' form='los' lem='el' mi='DA0MP0'>
              <NODE ord='3' alloc='31' form='constitucionales'
                lem='constitucional' mi='AQ0CP0'>
                </NODE>
              </CHUNK>
            <CHUNK ord='4' alloc='48' type='grup-sp' si='iobj'>
              <NODE ord='1' alloc='48' form='a' lem='a' mi='SPS00'>
                <NODE ord='3' alloc='54' form='presos_políticos'
                  lem='preso_político' mi='NCMP000'>
                  <NODE ord='2' alloc='50' form='los' lem='el' mi='DA0MP0'>
                    </NODE>
                  </NODE>
                </CHUNK>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </NODE>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
```

Irudia III.10: Analisi fasearen irteera

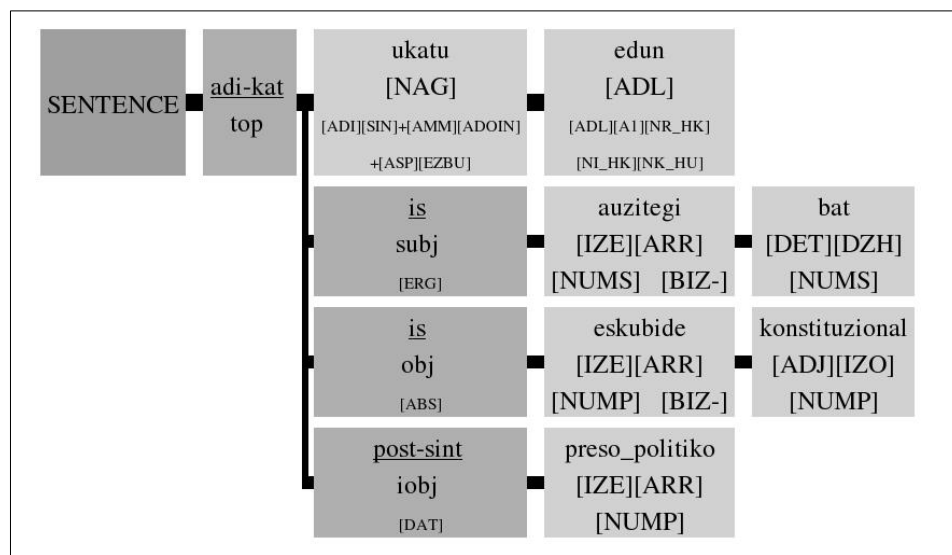


```

<SENTENCE ref='1' alloc='0'>
  <CHUNK ref='2' alloc='12' type='adi-kat' si='top'>
    <NODE ref='1' alloc='12' lem='ukatu' mi='VMIP3S0' pos='[ADI][SIN]' />
    <CHUNK ref='1' alloc='0' type='is' si='subj'>
      <NODE ref='2' alloc='3'
        lem='auztegi' pos='[IZE][ARR]' num='[NUMS]' sem=[BIZ-]>
        <NODE ref='1' alloc='0' lem='bat' pos='[DET][DZH]' num='[NUMS]' />
      </NODE>
    </CHUNK>
    <CHUNK ref='3' alloc='18' type='is' si='obj' focus='true'>
      <NODE ref='2' alloc='22'
        lem='eskubide' pos='[IZE][ARR]' num='[NUMP]' sem='[BIZ-]'>
        <NODE ref='1' alloc='18' det='[MUGM]' num='[NUMP]' />
        <NODE ref='3' alloc='31'
          lem='konstituzional' pos='[ADJ][IZO]' num='[NUMP]' />
        </NODE>
      </CHUNK>
    <CHUNK ref='4' alloc='48' type='post-sint' si='iobj'>
      <NODE ref='1' alloc='48' prep='a'>
        <NODE ref='3' alloc='54'
          lem='presio_politiko' pos='[IZE][ARR]' num='[NUMP]'>
          <NODE ref='2' alloc='50' det='[MUGM]' num='[NUMP]' />
        </NODE>
      </NODE>
    </CHUNK>
  </CHUNK>
</SENTENCE>

```

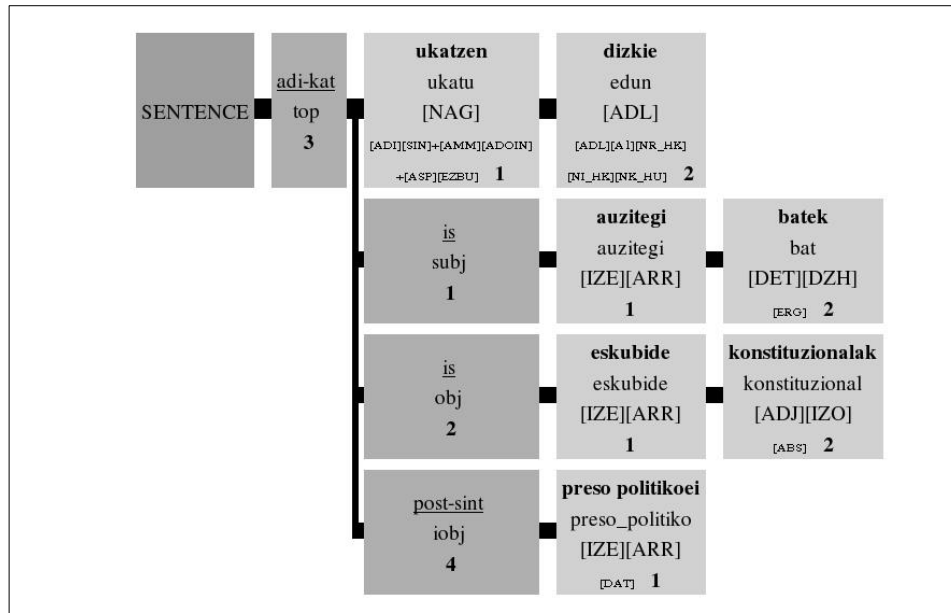
Irudia III.11: Transferentzia lexikalaren irteera



```

<SENTENCE ref='1' alloc='0'>
  <CHUNK ref='2' alloc='12' type='adi-kat' si='top'>
    <NODE ref='1' alloc='12' lem='ukatu' loc='[NAG]'
      post='[ADI][SIN]+[AMM][ADOIN]+[ASP][EZBU]''>
    <NODE ref='1' alloc='12' lem='edun' loc='[ADL]'
      post='[ADL][A1][NR_HK][NI_HK][NK_HU]''/>
    </NODE>
  <CHUNK ref='1' alloc='0' type='is' si='subj' num='[NUMS]' post='[ERG]''>
  <NODE ref='2' alloc='3' lem='auzitegi' pos='[IZE][ARR]' num='[NUMS]''>
    <NODE ref='1' alloc='0' lem='bat' pos='[DET][DZH]' num='[NUMS]''/>
  </NODE>
</CHUNK>
<CHUNK ref='3' type='is' alloc='18' postref='1' postalloc='18'
  si='obj' focus='true' det='[MUGM]' num='[NUMP]' post='[ABS]''>
  <NODE ref='2' alloc='22' lem='eskubide' pos='[IZE][ARR]' num='[NUMP]''>
    <NODE ref='3' alloc='31'
      lem='konstituzional' pos='[ADJ][IZO]' num='[NUMP]''/>
  </NODE>
</CHUNK>
<CHUNK ref='4' alloc='48' postref='1,2' postalloc='48,50'
  type='post-sint' si='iobj' det='[MUGM]' num='[NUMP]' post='[DAT]''>
  <NODE ref='3' alloc='54'
    lem='preso_politiko' pos='[IZE][ARR]' num='[NUMP]''/>
  </NODE>
</CHUNK>
</CHUNK>
</SENTENCE>
  
```

Irudia III.12: Transferentzia estrukturalaren irteera



```
<SENTENCE ord='1' ref='1' alloc='0'>
  <CHUNK ord='3' ref='2' type='adi-kat' alloc='12' si='top'>
    <NODE ord='1' ref='1' alloc='12' form='ukutzen' lem='ukatu'
      pos='[NAG]' post='[ADI][SIN]+[AMM][ADOIN]+[ASP][EZBU]'>
      <NODE ord='2' ref='1' alloc='12' form='dizkie'
        lem='edun' pos='[ADL]' post='[ADL][A1][NR_HK][NI_HK][NK_HU]'>
      </NODE>
    <CHUNK ord='1' ref='1' alloc='0' type='is' si='subj' post='[ERG]'>
      <NODE ord='1' ref='2' alloc='3' form='auzitegi'
        lem='auzitegi' pos='[IZE][ARR]' num='[NUMS]'>
      <NODE ord='2' ref='1' alloc='0' form='batek'
        lem='bat' pos='[DET][DZH]' num='[NUMS]' post='[ERG]'> </NODE>
      </CHUNK>
    <CHUNK ord='2' ref='3' alloc='18' type='is' si='obj'
      focus='true' det='[MUGM]' num='[NUMP]'
      post='[ABS]' postref='4' postalloc='18'>
      <NODE ord='1' ref='2' alloc='22' form='eskubide'
        lem='eskubide' pos='[IZE][ARR]' num='[NUMP]'>
      <NODE ord='2' ref='1,3' alloc='31,18' form='konstituzionalak'
        lem='konstituzional' pos='[ADJ][IZO]' num='[NUMP]' post='[ABS]'>
      </NODE> </NODE> </CHUNK>
    <CHUNK ord='4' ref='4' alloc='48' type='post-sint' si='iobj' det='[MUGM]'
      num='[NUMP]' post='[DAT]' postref='1,2' postalloc='48,50'>
      <NODE ord='1' ref='1,2,3' alloc='54,48,50' form='preso politikoei'
        lem='preso_politiko' pos='[IZE][ARR]' num='[NUMP]' post='[DAT]'>
      </NODE> </CHUNK> </CHUNK>
  </SENTENCE>
```

Auzitegi batek eskubide konstituzionalak ukatzen dizkie preso politikoei

Irudia III.13: Sorkuntza fasearen irteera

IV. KAPITULUA

Analisia

*(...) eta gu negar gasen mende gaudenok
oker omen gaude.
Ez dute uzten maitatzen gauean
ez agurrik ez negarrik gauean
ez zergatik ez norarik gabe
sirenotsa garraixi bakarra gauean.
Hertzainak*

Analisia itzulpen-prozesuaren lehenengo fasea da transferentzian oinarritutako sistemetan eta bere emaitza abiapuntu-testuaren errepresentazio abstraktua izango da. Gure kasuan, erdal hizkuntzetatik (espainiera zein ingelesetik) euskarara itzuli nahi dugula, berrerabilpena helburua izanik ez du zentzurik behar dugun analizatzailea guk geuk eraikitzeak, dagoeneko hizkuntza horiek analizatzeko tresna sendoak egon badaudelako. Hala ere, ikerketa hau burutzen ari zen garaian, analisi partziala soilik ematen zuten analizatzaileak zeuden eskuragarri, eta guk, transferentzia sakona burutu ahal izateko, mendekotasun-erlazioak eta funtzio sintaktikoak ematen dituen analisia behar dugu. Beraz, berrerabili dugun analizatzaile partziala hedatzea beharrezkoa izan zaigu.

Kapitulu honetan analisiaren inguruko zenbait kontzeptu finkatzeaz gain, *Matxin* espainiera-euskara sistemarako berrerabili dugun *FreeLing* analizatzailea eta mendekotasun-erlazioak ebazteko garatu dugun hedapena aurkeztu ditugu. Kapituluaren bukaeran analizatzaileak oraindik dituen mugak aztertzen ditugu.

IV.1 Analisiaren sakonera

Hizkuntzaren analisia hainbat mailatan egin daiteke (Allen, 1995): fonologikoa, morfologikoa, sintaktikoa, semantikoa, pragmatikoa, diskurtsoarena eta munduaren ezagutza.

Gure kasuan, momentuz behintzat, analisi morfologikoa eta sintaktikoa soilik erabiliko ditugu. Testu idatzien itzulpenaz ari garelako fonologiaren beharrik ez dugu, eta beste ezagutzak, analisia burutzeko duten zailtasunagatik, baztertu ditugu.

Morfologia hitzaren egiturari dagokio eta sintaxia esaldiaren egiturari. (Allen, 1995) lanetik jasotzen dugu zein den ezagutza morfologikoa eta sintaktikoaren muina:

- Ezagutza morfologikoa: hitzen osaketari dagokio. Morfemak dira hitzak osatzen dituzten oinarrizko unitateak.
- Ezagutza sintaktikoa: esaldi zuzenak osatzeko hitzak nola konbinatzen diren, hitz bakoitzak esaldian zein funtzio betetzen duen, eta nola sintagma bestea sintagma batzuen parte diren aztertzen du.

IV.1.1 Analisi morfologikoa

Analisi morfologikoaren zeregina hitz bakoitzaren interpretazio desberdin posibleen morfema-osaketa eta kategoria gramatikala ezagutzea da. Morfemak hitzak osatzeko osagaiak dira: lemak (hitz-erroak), morfema gramatikalak eta hizkiak (aurrizkiak eta atzizkiak).

Hiztegi morfologikoetan morfema posibleak definitzen dira, eta zenbait sistemetan forma flexionatu guztiak. Sarrera bakoitzean forma, lema, kategoria sintaktikoa, ezaugarri sintaktikoak (numeroa, mugatasuna, kasua, pertsona...) eta beste zenbait informazio kodetzen da.

Analisi morfologikoak ondoko prozesaketak burutu ditzake:

- Flexionatutako hitz-formetan lemaren identifikazioa. Formak erregularrak (*amar:amé*) edo irregularrak (*tener:tuve*) izan daitezke.
- Eratorpena ezagutzea (*rápida:mente:rápido+mente*).
- Konposatutako hitzak banatzea: (*dáselo:da+se+lo; del:de+el*);
- Hitz anitzeko unitate lexikalak ezagutzea (*dar la cara*).
- Hiztegiko informazioarekin ezagutu ezin daitezkeen hitzak prozesatzea. Askotan haien flexio gramatikalak identifikatuz bere kategoria ere asma daiteke.

Testuingururik gabe hitz-forma baterako zenbait analisi morfologiko eman daitezke, baina testuinguru batean analisi-aukerak murriztu egiten dira. Desanbiguatzaile edo etiketatzaile morfosintaktikoa (*part-of-speech tagger*) esaldi bateko hitz baten analisia lortzeko erabiltzen da. Desanbiguazioa burutzeko hitzaren testuingurua kontuan hartzen da.

IV.1.2 Analisi sintaktikoa

Esaldi baten egitura sintaktikoak esaldiko hitzen arteko loturak erakusten ditu (Allen, 1995): nola biltzen diren hitzak chunketan¹, noiz modifikatzen duen hitz batek beste hitz bat, eta zeintzuk diren esaldian berebiziko garrantzia duten hitzak. Gainera egitura honek chunken arteko erlazio motak identifikatu ditzake. Adierazpen sintaktiko gehienak zuhaitz gisa aurkezten dira, non hitz edo chunken arteko erlazioak azaltzen diren.

Analisi sintaktikoaren helburua hitzak esaldiko osagai nagusietan multzokatzeko eta osagai horiek etiketatzea da (Civit, 2003).

IV.1.2.1 Analisi sintaktiko partziala

Analisi sintaktiko partzialetan (edo azalekoetan) testuen analisi gramatikala ez da osoa, bere helburua esaldi barruko chunken sekuentzia identifikatzeko da. Analizatzaile partzialak *chunker* izenarekin ezagutzen dira eta baliabide linguistiko mugatuak erabiltzen dituzte: *part-of-speech* (POS) kategoriak, ezaugarri morfologikoak (numeroa, generoa,...) eta batzuetan azpikategorizazioa. Bi ezaugarri bete ohi dituzte: sendotasuna eta eraginkortasun konputazionala (abiadura).

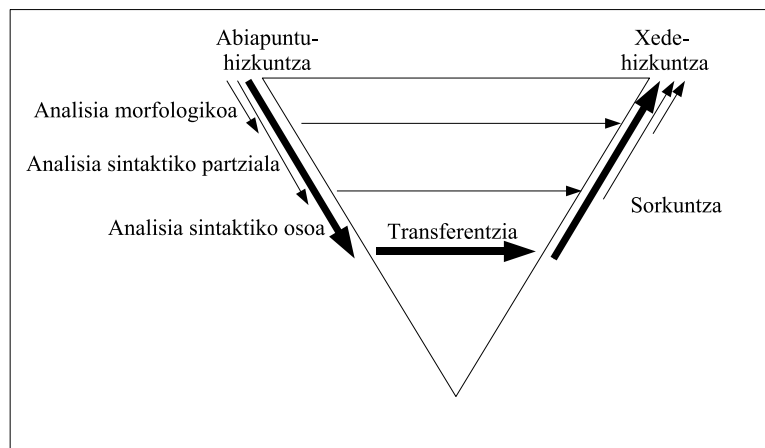
Analizatzaileen sendotasunaren definizioa (Ait-Mokhtar *et al.*, 2002) lanetik jasotzen dugu: mundu errealeko testuen analisi baliagarriak emateko gaitasuna. Analisi baliagarria zuzena (behintzat partzialki) eta aplikazio baterako erabilgarria denari esaten zaio, ondoko bi eskakizunak bete behar direla:

- Behintzat analisi bat ematea.
- Ematen diren analisien kopurua mugatua izatea.

IV.1.2.2 Analisi sintaktiko osoa

Analisi sakonago bat ere lor daiteke, funtzio sintaktikoak eta hitzen edo chunken arteko mendekotasun-erlazioak identifikatuz. Mendekotasun-analizatzailearen emaitza zuhaitz-egitura baten bidez adierazten da.

¹Chunk terminoaren definizioa III.5.2 atalean eman dugu



Irudia IV.1: Vauquois-en triangelua analisi posibleetarako

Espainieratik euskarara itzuli ahal izateko, bi hizkuntzen arteko distantzia handia denez, hitz bakoitzaren informazio morfologikoa lortzeaz eta chunkak identifikatzeaz gain, beharrezkoa da hitzen arteko eta chunken arteko mendekotasun-informazioa, eta funtzio sintaktikoen informazioa ere. Hau da, analisi sintaktiko osoa behar da (ikus IV.1 irudia, non gezi lodiez adierazten diren *Matxin* sistemaren fase desberdinak).

Gure sistemarako behar dugun analizatzailea guk geuk eraikitzea hasieratik baztertu genuen, berrerabilpena buruan, zentzurik ez duelako egina dagoen lana berregitea. Hala izanik, garatutako analizatzaile bat erabiltzea izan da gure erabakia, *FreeLing* aukeratuz.

IV.2 *FreeLing*

Kataluniako UPC² Unibertsitateko TALP³ taldeak espainierarako garatutako *FreeLing* (Atserias *et al.*, 2006) analizatzailea izan da aukeratu dugun tresna bi arrazoiengatik: batetik software librea delako, eta bestetik, aukera ematen digulako, Opendrad bezalako proiektuen barnean, UPCko taldearekin elkarlanean aritzeko, analizatzailea gure beharretara egokituz.

²Universitat Politècnica de Catalunya. <http://www.upc.edu>

³Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla.
<http://www.talp.upc.edu>

FreeLing kode irekiko liburutegi librea⁴ da, zenbait hizkuntzatarako oinarritzko LNPko zerbitzuak eskaintzen dituen: testuaren segmentazioa hitzetan, analisi morfologikoa, desanbiguatzaile morfosintaktikoa, analisi sintaktikoa, entitateen ezagutza...

FreeLing analizatzailearen oinarria (Carmona *et al.*, 1998) lanean aurkeztutako *Maco+* analizatzailea da. Berrerabilgarritasuna eta eraginkortasuna helburu, analizatzailea bezero-zerbitzari arkitektura batera egokitu dute, LNPko aplikazioetan erraz integratu ahal izateko (Carreras eta Padró, 2002). Analizatzailearen diseinuak errazten du erabiltzailearen beharretarako egokitzapena.

*FreeLing*en lehenengo bertsioak (Carreras *et al.*, 2004) espainiera, katalan eta ingeleserako ondoko funtzionalitateak eskaintzen zituen: tokenizazioa eta esaldien segmentazioa, analisi morfologikoa, desanbiguatzaile morfosintaktikoa (*POS tagger*), analizatzaile sintaktiko partziala (*chunker*) osagaie-tan oinarrituta, hitz anitzeko unitateen ezagutza, izen berezien ezagutza eta zenbait adierazpenen ezagutza (data eta orduak, moneta- eta zenbaki-adierazpenak).

*FreeLing*ek hitzen informazio morfologikoa adierazteko erabiltzen dituen etiketak *Eagles*⁵ taldeak europar hizkuntz guztietako lexikoi eta corpusak markatzeko *Parole*⁶ etiketa multzoan oinarritzen da (Leech eta Wilson, 1996). *FreeLing*en espainierarako etiketa multzoa B eranskinean ikus daiteke⁷.

Analizatzailearen espainierarako hiztegi morfologikoak 71.000 forma inguru dauzka, eskuz kodetu dituztela kategoria itxietako lema guztiak eta kategoria irekietako maiztasun handieneko 5.000 lema. Analizatzaile morfologikoak estaltzen ditu kategoria itxietako token guztiak eta testu libreko kategoria irekiko tokenen %80a. Hala ere, hitz ezezagunak ere tratatzen ditu. Sistemaren abiadura⁸ analizatzaile morfologikoa eta desanbiguatzailea egikaritzean, 2.300 hitz/segundoko da.

Analizatzailearen lehenengo bertsioak (Carreras *et al.*, 2004) chunken identifikazioa burutzen bazuen ere, ez zituen mendekotasun-erlazioak ezagutzen, ezta funtzio sintaktikoak esleitzen, transferentzia sakonerako nahi-

⁴*FreeLing* GNU Lesser General Public License (LGPL) lizentziarekin banatzen da. <http://garraf.epsevg.upc.es/freeling>

⁵Expert Advisory Group on Language Engineering Standards. <http://www.ilc.cnr.it/EAGLES/home.html>

⁶Preparatory Action for Linguistic Resources Organization for Language Engineering

⁷*FreeLing*en etiketa multzoa eguneratua ondoko web-orrian aurki daiteke: <http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html>

⁸Pentium4 2,8 GHz prozesadore batean.

taezkoak direnak.

Horregatik, tesi-lan honen barruan, *FreeLing*ek ematen duen azaleko analisitik mendekotasun-erlazioak eta funtzio sintaktikoak ebazteko modulu bat diseinatu eta garatu genuen (ikus IV.3 atala), modulu honetan oinarrituta UPCko taldeak eraiki zuela *Txala* mendekotasun-analizatzailea (Atserias *et al.*, 2005).

*FreeLing*en hurrengo bertsioetan (Atserias *et al.*, 2006) mendekotasun-analizatzailearen diseinua hobetu dute (ikus IV.3.2 atala). Gainera, errega gehiago kodetu dituzte mendekotasun-analisiaren kalitatea hobetuz. Analizatzaileak irteera hainbat formatutan eskaintzen du, tartean *Matxine*k prozesatzen duen datu-fluxuaren XML formatua (ikus III.5.3 atala).

IV.3 Mendekotasun-analizatzailea

Ingelesaren mendekotasun-analisiari buruz hainbat lan aurki daitezke (Briscoe *et al.*, 2006; By, 2004; Mel’cuk, 1988), baina espainierarako oso mendekotasun-analizatzaile gutxi daude, besteak beste, *Connexor*⁹ enpresaren analizatzaile komertziala eta berriki garatutako *Diluct* analizatzailea (Calvo eta Gelbukh, 2006), kode irekikoa izango omen dena.

FreeLing analizatzailea erabiltzea erabaki genuenean, mendekotasun-erlazioak lortzeko sistema hedatzea beharrezkoa izango zela aurreikusi genuen.

Horretarako, (Coheur *et al.*, 2004) lanean aurkezten duten hurbilpenean bezala, analisi partzialetik abiatuta, chunken arteko loturak eta chunk barruko nodoen artekoak ebazten dituen modulu bat diseinatu eta garatu dugu.

Tesi-lan honetako helburua analizatzaileen arloan sakontzea ez denez, gure xedea ez da izan mendekotasun-analizatzaile ezin hobe bat guk garatzea. Behar hau sortu zitzaigunean, nahikoa genuen hurbilpen simple bat eraikitzearekin, haren emaitzak baliatuz (nahiz eta zenbaitetan guztiz onak ez izan) sistema bere osotasunean lan egin zezan ahalbidetzen zuena. Hori bai, sendotasuna buruan, gure helburua zen mundu errealeko testuetarako beti analisi bat ematea, bat eta bakarra, aplikazioaren hurrengo moduluek prozesatu ahal izango dutena.

Garatutako modulua sinplea izanik eta analisirako formalismo jakin bat erabiltzen ez badu ere, bere baliagarritasuna frogatuta gelditu da; chunken arteko erlazioak ebazten dituen gure modulua egokituz UPCko taldeak *Txala* mendekotasun-analizatzailea eraiki duelako, *FreeLing*en azkeneko bertsioek barne dutena.

⁹<http://www.connexor.com>

Chunk mota	Hitz forma	Lema	Inf.morfologikoa
sn	Testigos presenciales	testigo presencial	NCMP000 AQ0CP0
grup-verb	han relatado	haber relatar	VAIP3P0 VMP00SM
subord	que	que	CS
sn	el ejército israelí	el ejército israelí	DA0MS0 NCMS000 AQ0CS0
grup-verb	destruyó	destruir	VMIS3S0
sn	las viviendas	el vivienda	DA0FP0 NCFP000
sp-de	de diez activistas palestinos	de 10 activista palestino	SPS00 Z NCCP000 AQ0MP0
grup-sp	en Jenin	en Jenin	SPS00 NP00000

Taula IV.1: Analisi partzialaren emaitza

IV.3.1 Oinarrizko mendekotasun-analizatzailearen eraikuntza

Esan bezala, mendekotasun-erlazioak ebazteko beharraren aurrean, gure helburua ahalik eta modu sinpleenean oinarrizko modulu eraginkor bat garatzea izan zen.

Modulu horren sarrera analisi partziala da, eta lortu nahi dugun irteera esaldiaren egitura da, non chunken arteko loturak eta chunk barruko nodoen artekoak, eta chunken funtzio sintaktikoak adierazten diren.

Analisi partzialak esaldiko hitz bakoitzerako bere forma, lema eta informazio morfologiko desanbiguatua emateaz gain, hitzak chunketan multzokatzen ditu, chunka bere motarekin etiketatuz.

IV.1 taulan ikus dezakegu *Testigos presenciales han relatado que el ejército israelí destruyó las viviendas de diez activistas palestinos en Jenin* esaldia-
rentzat burututako analisi partziala¹⁰.

Hortik analisi osoa lortzeko hiru prozesu burutu behar ditugu:

- Chunk barruko nodoen arteko mendekotasunak ebatzi.
- Chunken arteko mendekotasunak ebatzi.

¹⁰Chunk mota adierazteko etiketak: *sn*: sintagma nominal; *grup-verb*: grupo verbal; *subord*: elemento de subordinación; *grup-sp*: sintagma preposicional; *sp-de*: sintagma preposicional con preposicion *de*.

Informazio morfologikoa adierazteko *Parole* etiketa multzoa erabiltzen da ¹¹ :
NCMP: nombre común masculino plural; *VAIP3P*: verbo auxiliar indicativo presente 3ª plural...

- Funtzio sintaktikoak eta fokoa identifikatu.

Chunken barruko nodoen arteko mendekotasunak

Chunk bakoitzeko nodoen arteko mendekotasunak ebazteko *top-down* analizatzaile sinple bat eraiki dugu.

Gramatikako erregelek, chunk motaren arabera, chunkaren erroa den nodoa bilatzen dute, eta ondoren bere azpiko nodoak, eta horien azpikoak... chunkeko nodo guztiak jasoko dituen zuhaitz-egitura sortu arte.

Chunk barruko nodoetako hitzen kategoria posibleen kopurua oso mugatua denez, eraiki dugun gramatika oso sinplea da (30 erregela) eta emaitza onak lortzen ditu.

Chunken arteko mendekotasunak

Esaldiko chunken arteko mendekotasun-erlazioak adierazten dituen zuhaitz-egitura eraikitzeke, ordea, diseinatu dugun analizatzaileak, chunkak behin bakarrik eta modu sekuentzialean, ezkerretik eskuinera, aztertzen ditu. Modu horretan zenbait esaldi konplexuren analisi zuzena egitea ezinezkoa bada ere, gure apustua sinpletasunaren aldekoa izan da.

Beraz, analisi partzialean jasotako chunkak konbinatzen dituzten erregela sintaktikoak kodetu ditugu, ezkerretik eskuinera lan egiten duen estrategia bat erabiliz.

Erregela-aplikazio bakoitzean chunk berri bat gehitzen da zuhaitzean, eta chunk berria zuhaitzean non jarriko den ebazteko bi erakusle erabiltzen dira: *Gunea* eta *Azkena*. *Guneak* identifikatzen du zein den, ordura arte sortutako zuhaitzean, chunk berriarekin konbinatzeko orduan, kontutan hartu beharreko chunka. *Azkenak* zuhaitzean gehitutako azken chunka identifikatzen du.

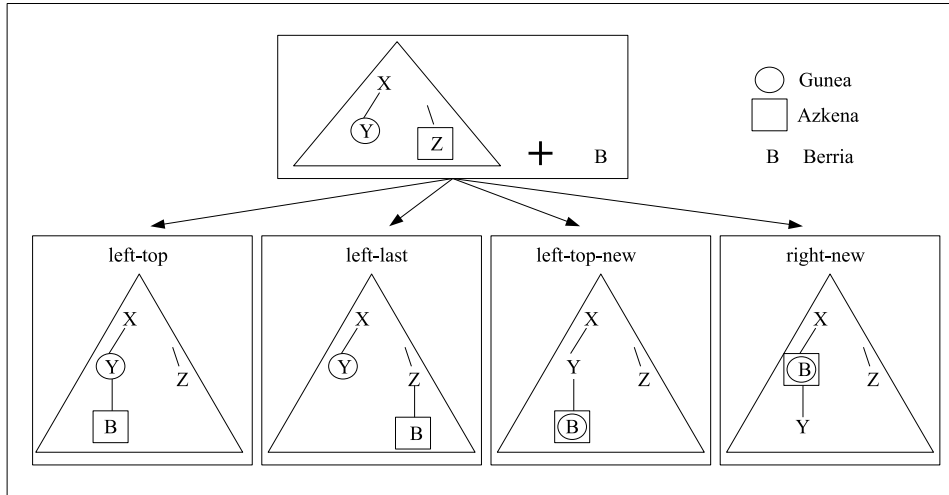
Erregelen formatua ondoko da:

GunearenMota	BerriarenMota	Eragiketa
--------------	---------------	-----------

Adibide gisa, ikus ditzagun erregela batzuk:

sn	grup-verb	right-new
grup-verb	sn	left-top
grup-verb	grup-sp	left-top
grup-verb	sp-de	left-last
grup-verb	subord	left-top-new
subord	sn	left-top-new

non:



Irudia IV.2: Chunkak konbinatzeko eragiketa motak

- *GunearenMota*: ordura arte osatutako zuhaitzean *Gunea* den chunkaren mota. Esan bezala, *Gunea* da kontutan hartu behar den chunka, chunken sekuentziako hurrengoarekin, hau da, *Berriarekin*, konbinatzeko.
- *BerriarenMota*: chunken sekuentziako hurrengoaren mota.
- *Eragiketa*: ordura arte osatutako egitura eta chunk berria konbinatuko diren modua. Ondoko eragiketak onartzen dira (ikus IV.2 irudia):
 - *left-top*: chunk *Berria* zuhaitzeko *Gunearen* azpian gehitzen da. Zuhaitz berriaren *Gunea* lehengo bera izango da.
 - *left-last*: chunk *Berria* zuhaitzeko *Azkenaren* azpian gehitzen da. Zuhaitz berriaren *Gunea* lehengo bera izango da.
 - *left-top-new*: Chunk *Berria* zuhaitzeko *Gunearen* azpian gehitzen da. Chunk *Berria* izango da zuhaitz berriaren *Gunea*.
 - *right-new*: Chunk *Berria* zuhaitzeko *Gunearen* gurasoaren azpian gehitzen da eta *Gunea* bere guraso azpitik kendu eta *Berriaren* azpian gehitzen da. Zuhaitzeko *Guneak* ez badu gurasorik, hau da, zuhaitzeko erroa bada, orduan chunk *Berria* bihurtzen da zuhaitzeko erroa, eta *Gunea* *Berriaren* azpian gehitzen da. Chunk *Berria* izango da zuhaitz berriaren *Gunea*.

Chunk berri bat zuhaitzean konbinatzeko erregularik ez bada aurkitzen, orduan *left-top* eragiketa burutzen da. Guztira 32 erregela kodetu ditugu. IV.3 irudian ikus daiteke nola aplikatzen diren erregelak chunken arteko mendekotasunak ebazteko *Testigos presenciales han relatado que el ejército israelí destruyó las viviendas de diez activistas palestinos en Jenin* esaldia-
rentzat.

Urrats bakoitzean eraikitako zuhaitz-egitura ikusten dugu, non *Gunea* borobil batekin seinalatzen den eta *Azkena* lauki batekin.

Funtzio sintaktikoak eta fokoa

Espainieratik euskarara itzultzerakoan subjektua eta objektua bereiztea funtsezkoa da. Espainieraz subjektua eta objektua, gehienetan, preposiziorik gabe azaltzen dira eta, beraz, bereizketa hori burutzea zaila da. Zein den subjektua eta zein objektua jakiteko bi irizpide erabil daitezke: posizionala (subjektua-aditza-objektua da lehenesten den ordena) eta komunztadura (numeroan eta pertsonan subjektua eta aditzaren artean).

Funtzio sintaktiko nagusiak, subjektua eta objektua, etiketatzeko hurbilpen sinple bat inplementatu dugu, chunken mendekotasun-zuhaitza eraiki ondoren, ondoko bi heuristikoak erabiliz:

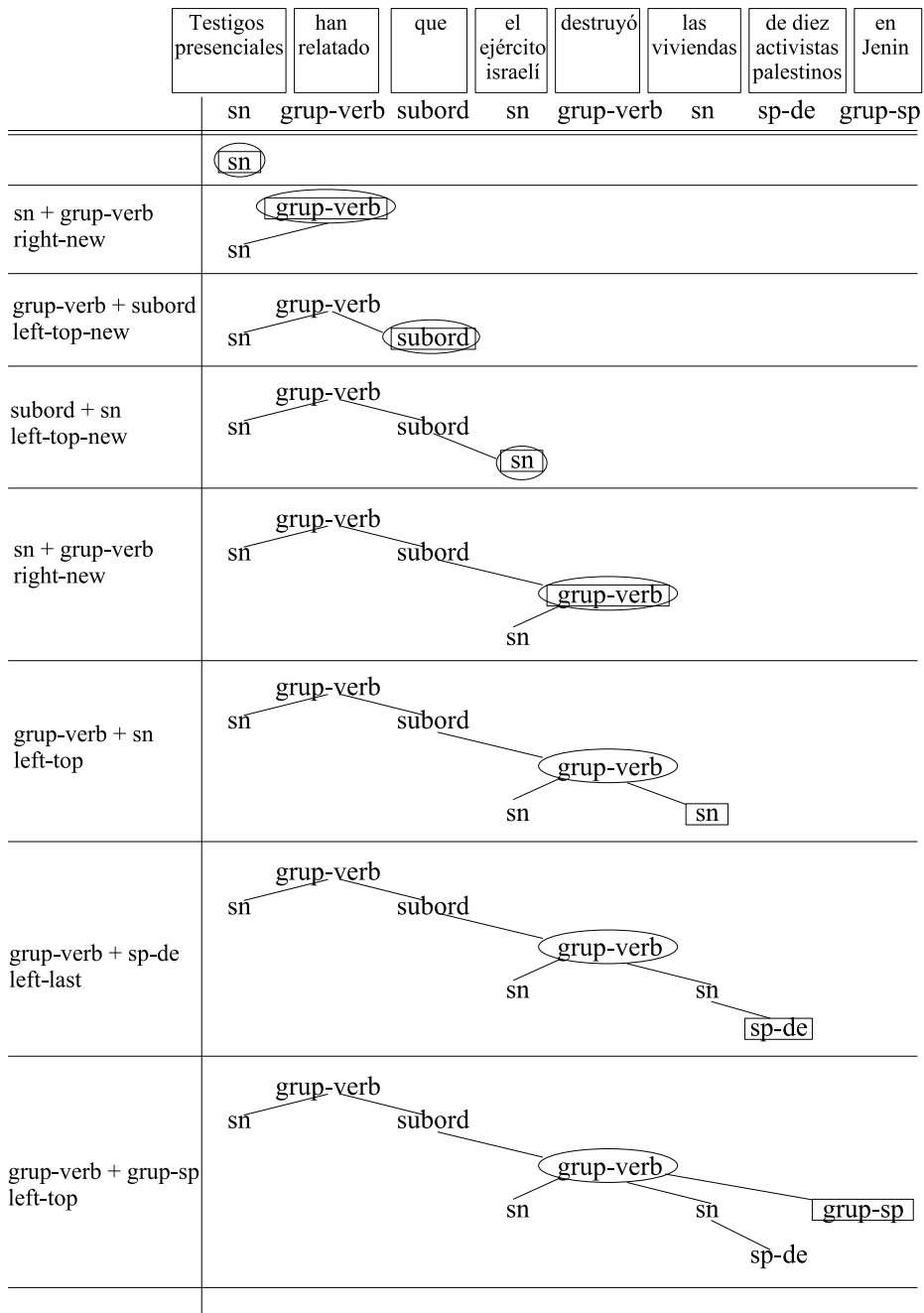
- Subjektua: izen sintagma bat (*sn*) bere gurasoa den aditz-sintagma (*grup-verb*) baino lehenago azaltzen bada esaldian eta bien pertsonak eta numeroak bat badatoz, izen sintagma subjektua da.
- Objektua: izen sintagma bat (*sn*) bere gurasoa den aditz-sintagma (*grup-verb*) eta gero azaltzen bada esaldian edo lehenago egonda bien pertsonak eta numeroak ez badatoz bat, izen sintagma objektua da.

Fokoa zein den identifikatzea beharrezkoa da, euskarazko sorkuntza sintaktikoan sintagmak modu egokian ordenatu ahal izateko. Espainieraz fokoa aditzaren eskuinean dagoen lehenengo sintagma da.

Gure adibidean *Testigos presenciales* eta *el ejército israelí* chunkei subjektu-etiketa esleituko zaie; *las viviendas* chunkari, berriz, objektu-etiketa. Fokoaren etiketa, berriz, *que* eta *las viviendas* chunkek jasoko dute.

IV.3.2 Txala, FreeLingeko mendekotasun-analizatzailea

UPCko taldeak, gure moduluetan oinarrituta, ondoko hedapenak burutu zituen *FreeLing*en mendekotasun-análisisa emateko (Atserias *et al.*, 2005):



Irudia IV.3: Mendekotasun-zuhaitzaren eraikuntza. Adibidea.

- Chunk barruko nodoen arteko erlazioak zuzenean lortzearen analizatzaille partzialaren gramatika (eta analizatzaile bera) egokitu zuten, erregela bakoitzean gunea den elementua markatuz.
- Chunken arteko erlazioak lortzeko guk diseinatutako moduluaren berriplementazio bat burutu zuten, *Txala* izenekoa.
- Funtzio sintaktikoak esleitzeko prozedura inplementatu dute, subjektu eta objektuaz gain funtzio sintaktiko gehiago etiketatuz (objektu ez-zuzena, izenaren modifikatzailea, osagarri zirkunstantziala...). Heuristikoen kopurua handitu dute eta heuristikoetan aditzaren mota (adibidez, mugimenduzkoa den edo ez) erabiltzen da. Horretarako beharrezkoak diren aditz-multzoak definitu dituzte.

Aurrerago, eta *OpenTrad* proiektuaren barruan es-eu itzultzaile automatikoaren beharrei aurre egiteko, chunken arteko mendekotasun-analizatzailean erregela kopurua handitzeaz gain, diseinuan zenbait aldaketa burutu dituzte, *Freelingen* ondoko bertsioetan gauzatu direnak (Atserias *et al.*, 2006).

Mendekotasun-analizatzaile berriak *bottom-up* analisi bat burutzen du, hasierako gure ezkerretik eskuinerako korritze sinplea baino hobe dena, noski. Erregeletan bi eremu berri agertzen dira: etiketa bat, bi chunk konbinatzean sortzen den zuhaitza zenbaitetan berretiketatu behar delako; eta lehentasun-balio bat.

Erregelak lehentasun handienekotik txikienera aplikatzen dira, alboko chunk bikoteak hartuz. Lehentasun handieneko erregela betetzen duen alboko chunk bikotea bilatzen da eta, aurkitzekotan, erregela aplikatzen da bi chunkak konbinatuz. Chunk bikote bat baino gehiago aurkituz gero, ezkerretik eskuinera aplikatzen da erregela.

Chunken arteko mendekotasunak ebazteko gramatikak 60 erregela ditu eta funtzio sintaktikoak identifikatzekoa 30 erregela.

Diseinu honek emaitza hobeak lortzen ditu. Aipatu behar da, hala ere, erregelak lehentasun-balioaren arabera aplikatzen direnez, erregelen arteko interferentziak sortzen direla eta horrek gramatikaren zuzenketa eta hedapena zaildu egiten duela.

IV.3.3 Mendekotasun-analizatzailearen mugak

Gaur egun *Matxin* sistemak sortzen dituen errore gehienak analizatzaileak eragindakoak dira (ikus VIII.3 atala). Gainera, analisi-fasean ematen diren erroreek itzulpen-prozesu osoari eragiten diote.

*FreeLing*ek esaldi sinpleak nahiko ongi analizatzen baditu ere, esaldi luze eta konplexuetan errore asko sortzen ditu, mendekotasun-egitura oso nahasiak eraikiz.

Mendekotasun-analizatzailearen zailtasun handienak loturazko elementuekin sortzen dira, batez ere koordinazioetan eta komak agertzen direnean:

- Komak.

Esaldi barruan komek funtzio desberdinak bete ditzakete: izen-sintagmak koordinatzea, (*el perro, el gato y el ratón*), esaldiak koordinatzea (*Vine, ví, vencí*), diskurtso-markatzaileak mugatzea (*Han pedido, además, lo que querían*)...

Analizatzaileak ez ditu analisi egokiak ematen komak prozesatzean. Hobetzen ez den bitartean, arazoari neurri batean aurre egiteko estrategia bat erabaki dugu, esaldi konplexuak tratatu ahal izateko: komak esaldi banatzaileak izango dira. Horrela, ondokoa bezalako esaldi bat analizatzeko orduan, komek zatituko dute esaldia 6 azpiesaldi sortuz, banan bana itzuliko direnak: *Horas antes de producirse la votación, el ministro de Asuntos Intergubernamentales, Michael Chong, ha dimitido de su cargo, sin renunciar a su acta parlamentaria, para poder abstenerse de votar.*

- 1- Horas antes de producirse la votación
- 2- el ministro de Asuntos Intergubernamentales
- 3- Michael Chong
- 4- ha dimitido de su cargo
- 5- sin renunciar a su acta parlamentaria
- 6- para poder abstenerse de votar

Azpiesaldi bakoitza bere aldetik itzuliko da besteak kontutan hartu gabe. Honek itzulpenean erroreak emango ditu azpi-esaldi berrien arteko komuntadurak galtzen direlako (adibidez 2 eta 3 azpi-esaldiek subjektuaren funtzioa betetzen dute, eta aditz iragankorra denez, postposizio ergatiboa hartu beharko lukete).

Baina askoz okerragoa litzateke esaldi osoaren analisitik abiatzea, analisi horrek azpi-esaldi desberdinetako chunkak nahasten dituelako, zentzurik gabeko zuhaitz-egitura ematen.

- Koordinazioa.

Gaur egun analizatzaileak zenbait egitura koordinatu ondo analizatzen ditu. Adibidez:

el perro y el gato comen

```
comen
|_ y
   |_ el perro
   |_ el gato
```

Baina beste askotan analizatzaileak ez du asmatzen. Adibidez *el perro come pan en casa y el gato bebe* esaldirako analisia guztiz nahasten da:

el perro come pan en casa y el gato bebe

```
comen
|_ el perro
|   |_ bebe
|   |_ pan
|   |_ en casa
|_ y
   |_ el gato
```

Zailtasun horiei ezin zaie aurre egin soilik erregela berriak gehituz. Analizatzailearen diseinuan sakoneko aldaketak egin beharko lirateke egitura konplexu horiek modu egokian analizatu ahal izateko.

Horretarako, aurretik, loturazko elementuak, puntuazio ikurrak barne, zuhaitzean nola sartzen diren erabaki beharko litzateke. Zenbait lanetan, adibidez, puntuazio ikurrak ez dira zuhaitzean sartzen (By, 2004). Koordinazioa errepresentatzeko ere arazoak izaten dira, nagusiki bi irtenbide proposatu izan direla (Civit, 2003): koordinazioko osagai bat gune bezala hartzea, edo osagai berri bat ezartzea koordinatutako osagaiak bere barne hartzen dituen.

Analizatzailearen kalitatea hobetzeko aztertu beharreko beste aukera bat sistema estatistiko batekin konbinatzea izango litzateke.

IV.4 Laburpena

Kapitulu honetan itzulpen automatikoko sistemaren analisi-fasean egin beharreko azterketak eta burututako lanak aurkeztu ditugu. *Matxin* espainiera-euskara sistemarako *FreeLing* analizatzaile partziala berrerabili dugu, baina transferentzia sakonerako beharrezkoa den mendekotasun-egitura ebazteko hedapen sinplifikatu bat garatu behar izan genuen. Eraikitako moduluak (batez ere chunken arteko mendekotasunak ebazten dituen) bere fruituak eman ditu: hasiera batean gure sistema martxan jarri ahal izateko nahitaezko elementua izan zen eta, ondoren, *Freeling*eko garatzaileek eurak ere gure

moduluan oinarritu ziren *Freelinger*ako mendekotasun-analizatzailea eraikitzeko.

V. KAPITULUA

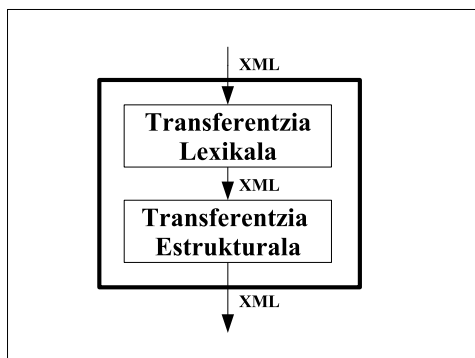
Transferentzia

- Andereño, supergosea daukat!
Demagun andereñoak barru-barrutik bizi duela euskararen kalitatearekiko kezka, eta
honako zuzenketa hau egiten diola haurrari:
- Ez da esaten “supergosea daukat”, “supergose naiz” baizik.
Demagun ikasleak “supergose naiz” zotinka esaten asmatzen duela, eta andereñoak, sari
gisa, ogitarteko bat ematen diola.
Demagun andereñoak gau hartan lo lasai egiten duela, bi ekintza on egin izanak
kontzientzia lasaiturik.
Kurrin-kurrin vitae. Joxerra Gartzia, 2004.

Transferentziaren helburua abiapuntu-hizkuntzako testuaren adierazpide abstraktua xede-hizkuntzako adierazpidea bihurtzea da. Transferentzia bi mailatan burutzen da: lexikala eta estrukturala (ikus V.1 irudia). Kapitulu honetan transferentziako bi maila horietan burutzen diren prozesuak deskribatuko ditugu aurkitu ditugun zailtasunak eta zailtasun horiei aurre egiteko hartutako erabakiak azalduz.

Transferentzia lexikalaren funtsa hizkuntza bateko osagai lexikal bakoitzaren beste hizkuntzan dagokion osagaiarekin ordezkatzeko da. Prozesu hau zailtzen duten anbiguotasunak aurkeztuko ditugu, moduluaren diseinuaren zehaztasunak eta erabiltzen den lexikoi elebidunaren formatua eta erabilpena aztertuz.

Transferentzia estrukturalak abiapuntu-hizkuntzatik datorren egitura xede-hizkuntzarako egokia den egitura bihurtzen du. Espainiera eta euskara hizkuntzen arteko dibergentzia sintaktikoak deskribatu ondoren ataza hau burutzeko moduluen banaketa aurkeztuko dugu, sakonki deskribatuko di-



Irudia V.1: Transferentzia moduluen arkitektura

tugula preposizio eta funtzio sintaktikoen transferentziarako modula eta aditz-kateen transferentziarakoa.

V.1 Transferentzia lexikala

Transferentzia lexikalaren funtsa hizkuntza bateko osagai lexikal bat beste hizkuntzan dagokion osagaiarekin ordezkatzeta da.

Termino baliokide bakarra dagoenean prozesua sinplea da, baina zailtasunak sortzen dira anbiguitasun lexikalak ematen direnean, hau da, termino batek xede-hizkuntzan zenbait itzulpen posible dituenean. Orduan itzulpen bat edo beste hautatzeko testuingurua aztertu behar da.

Lokuzio terminologiko edo esamolde idiomatikoen (*idiom*) itzulpenak ere aipamen berezia merezi du, itzulpen-automatikorako zailtasun handia suposatzen duelako.

Anbiguitasun lexikalak

Transferentzian anbiguitasun lexikalak eman daitezke, hau da, abiapuntu-hizkuntzako hitz bat (anbigua ez dena) xede-hizkuntzako hitz desberdinekin itzul daitekeela.

Anbiguitasun lexikal gehienak dibergentzia kontzeptualetan dute oinarri. Dibergentzia kontzeptual horiek hizkuntza bateko hitz batek adierazten duen kontzeptu bakar bat, beste hizkuntzako zenbait kontzeptuei, eta beraz zenbait hitzei, dagokienean gertatzen dira. Oso ohikoa den fenomeno hau itzulpenarako arazorik handiena izan ohi da. Adibide gisa, zenbait hitzetan ematen den anbiguitasun hau V.1 taulan ikus daiteke.

Espainieraz	Euskaraz
tomate	tomate zalaparta (iskanbila)
tráfico	salerosketa zirkulazio
techo	sabai muga bizitoki
pájaro	txori zakur, azeri (maltzur)

Taula V.1: Anbigutasun kontzeptualaren adibideak

Badira ere dibergentzia estilistiko eta gramatikalek sortutakoak. Dibergentzia estilistikoak xede-hizkuntzako hitzaren hautaketa testu mota edo erregistroaren arabera egin behar denean ematen dira eta gramatikalak testuinguru gramatikalak baldintzatzen duenean, baina ez dira sarri gertatzen.

Matxin sisteman, gaur egun, anbigutasun lexikalak ebazteko estrategia simple bezain eraginkorra erabiltzen dugu: lexikoi elebiduna eraikitzeke erabilitako hiztegi elebiduneko lehenengo adieraren lehenengo ordaina jasotzen dugu soilik, ordain horrek gehienetan itzulpen egokia ematen duelako, lexikografoek eskuz eraikitako hiztegi elebidunean ordainak, intuitiboki izanda ere, maiztasunaren arabera ordenatuta daudelako.

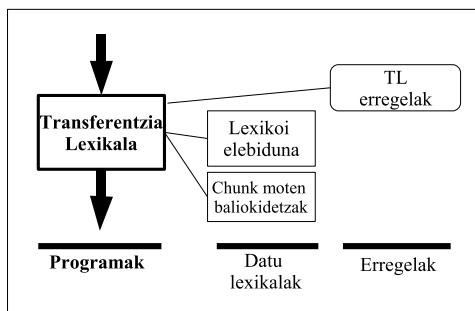
Gainera, lexikoian sarrera bakoitzarentzat jasotzen den ordaina itzulpen-lanetan egokiena izatearren, prozesaketa automatiko batez lehenengoa ez den beste bat hautatzen da zenbait kasutan (ikus VII.1.5 atala).

Estrategia honen emaitzak onak badira ere, gure etorkizuneko lanetan itzulpenaren hautapen lexikalerako desanbiguazio teknikak aztertzeari heldu beharko diogu. Horretarako sistemaren lexikoi elebidunean hiztegiko ordain guztiak jaso beharko dira.

Lokuzio terminologikoak

Lokuzioa forma sintaktiko eta esanahi ihartua duen hitz-multzoa da, bere itzulpena ez dela zertan bere osagaien itzulpenen konbinazioa izan (*irse a pique* → *hondoratu*).

Lokuzioak analisi-fasean identifikatu ezker, transferentzia lexikalean hitz bakarreko terminoen modura bilatzen dira lexikoian. Arazoa sortzen da batzuetan lokuzioaren osagaiak posizio ez jarraituetan datozenean (*se fue irremisiblemente a pique*), horrelakoetan zaila baita osagaien artean erlazioa aurkitzea.



Irudia V.2: Transferentzia lexikala: moduluaren arkitektura

V.1.1 Transferentzia lexikalaren modulua

Transferentzia lexikalaren funtsa lexikoi elebidunean abiapuntu-testuko lemei dagozkien itzulpenak bilatzea da, beste eragiketa batzuk burutzen bada ere. Moduluaren arkitektura V.2 irudian ikus daiteke.

Moduluak burutzen duen prozesaketa ulertzeko ondoko adibidea erabili dugu: *Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos*. Esaldiaren egitura transferentzia lexikalaren aurretik eta ondoren V.5 irudian ikus dezakegu.

Lexikoian kontsulta

Gure sistemako lexikoian bilaketak egiteko, abiapuntu-hizkuntzako lema eta informazio morfologikoa¹ erabiltzen da. Datu horiekin bat datorren sarrera bat aurkitzen bada, dagokion xede-hizkuntzako ordainaren lema, kategoria eta azpikategoria morfosintaktikoak², eta beste informazio batzuk jasotzen dira (ikus V.1.2 atala).

Transferentzia lexikala ez da nodo guztietarako egin behar. Ondoko kasu hauetan ez da egiten:

- Aditz-chunketan, erroa ez diren nodoetan (aditz laguntzaileak, perifrastikoak, izenordain atonoak...). Nodo horiek aditz-chunk osoaren transferentzia estrukturalen prozesatuko dira (ikusi V.2.4 atala).
- Preposizioak dituzten nodoetan. Preposizio eta funtzio sintaktikoak transferentzia estrukturalen prozesatuko dira, testuinguruaren informazioa kontutan harturik (ikus V.2.3 atala).

¹Espainierako *Parole* etiketa multzoa B eranskinean ikus daiteke.

²Euskararako kategoria-sistema C eranskinean ikus daiteke.

```

<!-- Aditz-kateen chunketako erroaren transferentzia lexikala>
<rule id='C/N_tl_adikat'>
  <match>
    <def> C := //CHUNK[@type='adi-kat']    </def>
    <def> N := C/NODE                      </def></match>
    <actions>
      <act> Target := &dictlook(N/@lem,N/@mi) </act>
      <act> &getattribs(N,Target)           </act></actions>
</rule><!------->
<!-- Aditz-kateak ez direnen nodoetako transferentzia lexikala>
<rule id='C/N_tl_adikat_ez'>
  <match>
    <def> C := //CHUNK[@type!='adi-kat']  </def>
    <def> N := C//NODE[@mi!~(S|Z|W)]      </def></match>
    <actions>
      <act> Target := &dictlook(N/@lem,N/@mi) </act>
      <act> &getattribs(N,Target)           </act> </actions>
</rule><!------->

```

Irudia V.3: Lexikoian kontsultarako erregelak

- Zifrak, datak eta orduak dituzten nodoetan. Hauetan ez dago transferentziaren beharrik, sorkuntza lematik zuzenean egingo delako.

Beraz, V.3 irudiko erregeletan adierazten den bezala, transferentzia lexikala ondoko kasuetan burutzen da:

- Aditz-kateetako nodo erroan.
- Aditz-kateak ez diren chunketako nodoetan, preposizioa, zifra, data edo ordua direnetan ezik.

Lexikoian kontsulta egitean jasotako informazioa nodoaren atributuetan sartzen da.

V.5 irudiko adibidean ikus dezakegu. aditz-chunkaren erroa ez den nodoan, (*estar*) aditz perifrastikoa duena, eta *a* preposizioa duen nodoan ez dela lexikoian kontsultarik egiten.

Beste nodo guztietarako kontsulta lexikoi elebidunean egiten da: berdin hitz bakarrekoak zein hitz anitzeko unitateak (*partido político*). Analizatzaileak emandako lema eta informazio morfologikoa (*asociación/NCFP000*) erabiltzen da bilaketa egiteko, eta lexikoian aurkitzen bada, euskarazko ordainaren lema, kategoria eta azpikategoria, eta beste informazio batzuk jasotzen dira (*elkarte/[IZE]/[ARR]/[NUMP]/[BIZ-]*).

Lexikoia numeroa (*[NUMP]*) lortzen du espainierazko informazio morfologikoa duen etiketatik (*NCFP000*). Determinante artikuluen transferen-

tzian (*el*) lexikoia ez du ordain bezala lemarik ematen, mugatasun eta numeroaren informazioa (*[MUGM]* *[NUMS]*) baizik.

Eragiketa osagarriak

Transferentzia lexikaleko modulu honetan, lexikoian kontsultaz gain, beste eragiketa batzuk burutzen dira V.4 irudiko erregelak aplikatuz:

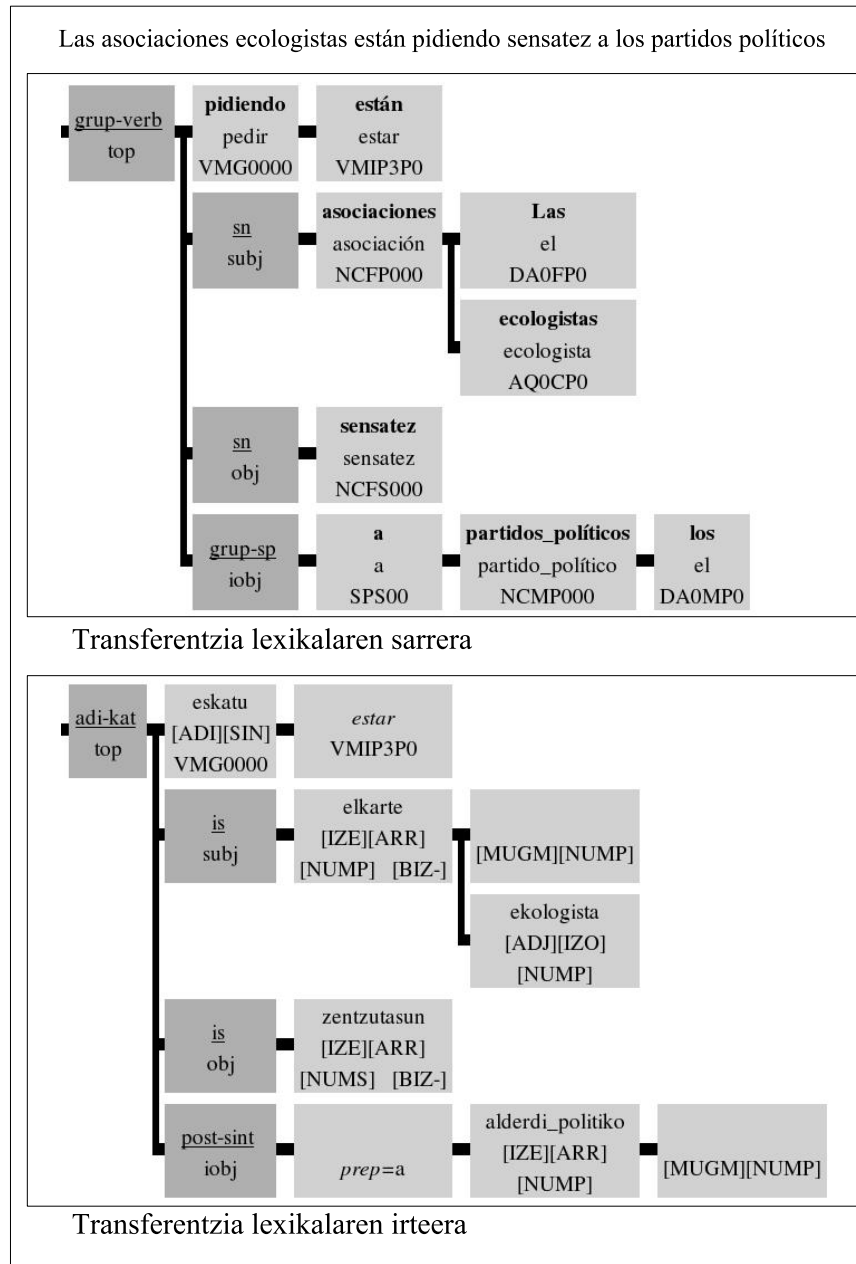
- Datu-egituraren elementu guztien (esaldiak, chunkak eta nodoak) ordena adierazten duen *ord* atributua ezabatu eta bere balioa *ref* atributuan gordetzen da, postedizioan xede-elementu bakoitza abiapuntu-testuko zein elementuren itzulpena den erreferentzia izateko. Ordena berria sorkuntza fasean kalkulatu da.
- Chunkaren espainierazko mota adierazten duen *type* atributua euskaraz dagokion balioarekin ordezkatzeko da. Horretarako, 32 baliokidetza kodetu dira (haietako batzuk V.2 taulan ikus daitezke). V.5 irudiko adibidean chunk guztien motak (azpimarrarekin adierazten dira) itzuli dira: (*grup-verb* → *adi-kat*, *sn* → *is*, *grup-sp* → *post-sint*).
- Nodoetan espainierazko forma duen *form* atributua ezabatzen da, atributu horretan itzulpen-prozesu osoaren bukaeran lortuko den euskarazko forma idatziko dela.
- Preposizioak dituzten nodoetan *lem* atributuan dagoen lemarren balioa *prep* atributuan gordetzen da, *lem* atributua ezabatuz. *prep* atributuaren balioa chunkari pasako zaio chunkaren postposizioa desanbiguatzeke.
- Aditz-kateetan burua ez den nodoetan *lem* atributuan dagoen lemarren balioa (V.5 irudiko adibidean *estar*) *vlem* atributuan gordetzen da, *lem* atributua ezabatuz. *vlem* atributua aditz-kate osoaren transferentzia estrukturalean erabiliko da.

```

<!-- Elementu guztien ord atributuaren ezabaketa>
<rule id='E_tl_ord_ref'>
  <match>
    <def> E := /*[@ord]                </def></match>
  <actions>
    <act> E/@ref := E/@ord              </act>
    <act> delete(E/@ord)                </act></actions>
</rule><!------->
<!-- Chunk motaren itzulpena>
<rule id='C_tl_type'>
  <match>
    <def> C := //CHUNK                  </def></match>
  <actions>
    <act> C/@type := &chunkLook(C/@type) </act></actions>
</rule><!------->
<!-- Nodoen forma duen form atributuaren ezabaketa>
<rule id='N_tl_form'>
  <match>
    <def> N := //NODE                  </def></match>
  <actions>
    <act> delete(N/@form)              </act></actions>
</rule><!------->
<!-- Preposizioen lem atributua ezabatu eta prep atributuan gorde>
<rule id='N_tl_prep'>
  <match>
    <def> N := //NODE[@mi='SPS00']      </def></match>
  <actions>
    <act> N/@prep := N/@lem             </act>
    <act> delete(N/@lem)               </act></actions>
</rule><!------->
<!-- Aditz-kateetan erroa ez den nodoen lem atributua ezabatu>
<!-- eta vlem atributuan gorde>
<rule id='N_tl_adikat_hostoak'>
  <match>
    <def> C := //CHUNK[@type='adi-kat'] </def></match>
    <def> N := C/NODE//NODE            </def></match>
  <actions>
    <act> N/@vlem := N/@lem             </act>
    <act> delete(N/@lem)               </act></actions>
</rule><!------->

```

Irudia V.4: Transferentzia lexikaleko beste eragiketetarako erregelak



Irudia V.5: Transferentzia lexikala: sarrera eta irteera.

Espainieraz	Euskaraz
sn	is
s-adj	adjs
sadv	adbs
sp-de	post-izls
grup-sp	post-sint
grup-verb	adi-kat
verb-pass	adi-pas
Fc	p-koma
F-no-c	p-ez-koma
F-term	p-buka
...	...

Taula V.2: Chunk moten baliokidetzak

V.1.2 Lexikoi elebiduna

Sistemaren lexikoi elebidunak abiapuntu-hizkuntzako unitate lexikal baten-tzat analisia emandako informaziotik xede-hizkuntzako sorkuntzan beharko den informazio guztia eman behar du.

Gure sistemako lexikoiaren diseinua hauxe da:

Espainiera	
lem	mi

Euskara											
lem	pos	suf	loc	per	num	det	lmi	post	spost	vpost	sem

Ondoko lerrootan sarrera bakoitzean kodetutako informazioa azalduko dugu, eta eremuen definizioak hobeto ulertzeko V.3 eta V.5 tauletan zenbait adibide eskaintzen dira.

Espainierazko aldean eremu hauek agertzen dira:

- *lem (lemma)*: lema. Hitz anitzeko unitatea bada, lemaren forma kano-nikoa erabiltzen da, hitzen artean beheko marra jartzen dela (adibidez, *partido_político*)
- *mi (morphological information)*: informazio morfologikoa. Analizatzailearen *Parole* notazioa³ erabiltzen da, non etiketaren lehenengo bi hiz-kiek kategoria eta azpikategoria adierazten duten.

Euskarazko aldean ondoko eremuak kodetzen dira (sarrera guztiek ez dute eremu guztietako informazioa):

³Espainierako *Parole* etiketa multzoa B eranskinean ikus daiteke.

- *lem* (*lemma*): lema. Zenbait determinantetarako ez da lemarik ematen.
- *pos* (*part-of-speech*): kategoria eta azpikategoria morfosintaktikoak⁴.
- *suf* (*suffix*): atzizkiaren informazioa (*-txo*, *-tzar*...). Soilik izen eta adjektiboetan.
- *loc* (*location*): kokapenari buruzko informazioa, euskarazko hitzen ordena ebazteko erabiliko da. Izenaren aurrekoa (*[AUR]*) edo atzekoa (*[ATZ]*) izan daiteke. Soilik adjektibo eta determinanteetan.
- *per* (*person*): pertsona, ondoko balioak har ditzake: *[NI]*, *[ZU]*, *[HU]*, *[GU]*, *[ZK]* eta *[HK]*⁵. Soilik izenordainetan.
- *num* (*number*): numeroa. Singularra (*[NUMS]*) edo plurala (*[NUMP]*) izan daiteke.
- *det* (*determination*): mugatasun informazioa. Mugagabe (*[MG]*) edo mugatua (*[MUGM]*) izan daiteke. Soilik determinante eta izenordainetan.
- *lmi* (*lemma's morphological information*): lemaren informazio morfologikoa. Sorkuntza morfologikoa egiteko beharko da. Soilik hiztegian ematen den ordainaren lema hitz-forma bat denean. Adibidez, *abismal* sarreraren ordaina den *leizeko* hitz-formaren informazio morfologikoa hauxe izango da: *leize**[IZE]**[ARR]* + *[DEK]**[NUMS]* + *[DEK]**[GEL]*.
- *post* (*postposition*): postposizioa. Nodo hori dagoen chunkari pasako zaio. Soilik kategoria itxiko zenbait sarreretan.
- *spost* (*subordination postposition*): menpeko postposizioa. Nodo hori dagoen chunkaren azpiko chunkari pasako zaio. Soilik kategoria itxiko zenbait sarreretan.
- *vpost* (*verb postposition*): aditz-postposizioa. Aditz horren fokua den osagaiari pasatuko zaio. Soilik zenbait aditzetan.
- *sem* (*semantic information*): informazio semantikoa, momentuz ondoko balioak izan ditzake: *[BIZ+]* (biziduna), *[BIZ-]* (ez biziduna), *[HZK+]* (hizkuntza), *[DEN+]* (denbora), *[GAI+]* (gaia) eta *[GLU+]*

⁴Euskararako kategoria-sistema C eranskinean ikus daiteke.

⁵Erdal hizkuntzetatik euskarara itzultzeko hika ez dugu erabiliko, singularreko 2. pertsona beti *zu* pertsonarekin itzuliko dugula.

l	r				par
lem	lem	loc	lmi	vpost	
adherir	itsatsi				VM_ADI_SIN
adhesivo	itsaskor	[ATZ]			AQ_ADJ_ARR
chalado	ero				VM_IZE_ARR
asf_sea	hala_bedi				RG_ADB_ARR
abismal	leizeko	[AUR]	(1)		AQ_ADJ_ARR
ablandamiento	biguntze		(2)		NC_IZE_ARR
preguntar_por	galdetu			[INS]	VM_ADI_SIN
caer_en_manos _de	erori			[GEN]++ eskutan	VM_ADI_SIN
(1) leize[IZE][ARR]+[DEK][NUMS]+[DEK][GEL]					
(2) bigundu[ADI][SIN]tze[ATZ][IZE][ARR]					

Taula V.3: Sekzio nagusia. Adibideak

(gailua). Ezaugarri semantikoak preposizio eta funtzio sintaktikoen transferentzian, eta sorkuntza morfologikoan erabiltzen dira.. Soilik zenbait izenetan.

Sarrera bakoitzeko (lema bat eta kategoria bat) ordain bakarra jaso da lexikoi elebidunean. Hautaketa lexikalerako desanbiguazio-teknikak erabili-ko balira, lexikoiak sarrera bakoitzeko adiera desberdinen ordainak eman beharko lituzke.

Lexikoi elebiduna eraikitzeke, *Elhuyar* hiztegia, *Euskalterm* banku terminologikoa, entitateen zerrenda elebidunak eta eskuz kodetu diren kategoria itxien hiztegia erabili dira. Iturri horietako informazio guztiarekin ego-kitze-prozesu bat pasa ondoren, *Matxin* sistemaren lexikoi elebiduna sortu da. Bere eraikuntza VII.1 atalean azaltzen da.

Lexikoi elebiduna XML formatuan kodetuta dago, *Opentrad* proiektua- ren espezifikazioaren arabera (ikus III.6.1 atala). Lexikoia ondoko ataletan deskribatuko ditugun lau sekziotan banatuta dago: sekzio nagusia, kategoria itxiena, puntuazio ikurrena eta erabiltzailearen hiztegia duen sekzioa.

Sekzio nagusia

Sekzio nagusiak kategoria irekiak biltzen ditu. Lexikoiaren sekzio handiena da: 62.000 sarrera ditu. Sarreraren adibideak V.6 irudian eta V.3 taulan ikus daitezke.

Sarreretako ezkerreko elementuan espainierazko lema agertzen da, eta eskuineko elementuan euskarazko lema eta zenbait sarreretan kokapenaren informazioa (*loc*), lemaaren informazio morfologikoa (*lmi*) edo aditzarekin

```

<section id="nagusia" >
  <e><p> <l>adherir</l>
    <r>itsatsi</r> </p>
    <par n='VM_ADI_SIN' /> </e>
  <e><p> <l>adhesivo</l>
    <r>itsaskor
      <s n='loc' />[ATZ]</r></p>
    <par n='AQ_ADJ_ARR' /> </e>
  <e><p> <l>as\ 'i_sea</l>
    <r>hala_bedi</r> </p>
    <par n='RG_ADB_ARR' /> </e>
  <e><p> <l>abismal</l>
    <r>leizeko
      <s n='loc' />[AUR]</r></p>
      <s n='lmi' />leize[IZE][ARR]+[DEK][NUMS]+[DEK][GEL]</r></p>
    <par n='AQ_ADJ_AUR' /> </e>
  <e><p> <l>caer_en_manos_de</l>
    <r>erori
      <s n='vpost' />[GEN]++eskutan</r> </p>
    <par n='VM_ADI_SIN' /> </e>
  ...
</section>

```

Irudia V.6: Sekzio nagusiaren formatua. Adibideak

lotutako postposizioa (*vpost*). Sekzio honetako sarreretan ez dira mugatasuna (*det*), postposizioa (*post*) eta mendeko postposizioa (*spot*) informaziorik ematen.

Adibidez, *abismal* sarrerarako *leizeko* ordainaz gain, *AUR* kokapenaren informazioa ematen da, adjetibo hori izenaren aurretik joango dela adierazten duena, eta ordainaren osaketa morfologikoa (*leize*[IZE][ARR]+[DEK][NUMS]+[DEK][GEL]).

Gainera, bikote bakoitzerako paradigma bat (*par*) ematen da. Paradigma desberdin bakoitzean espainierazko elementuaren informazio morfologikoari (*Parole* notazioan) euskarazko elementuaren zer informazio dagokion zehazten da: kategoria eta azpikategoria eta zenbaitetan numeroa (*NCFP000* → [NUMP]) eta atzizkien informazioa (*NCFP00D* → *txo*[ATZ][IZE][ARR]). Paradigmen formatua V.7 irudiko eta taulako V.4 adibideetan ikus dezakegu.

Paradigmaren identifikadorea *pardef* elementuaren *n* atributuaren balioa da, bi hizkuntzetako kategorien korrespondentzia adierazten duena. Batzuetan bi hizkuntzetarako kategoria baliokidea agertuko da (adibidez, *adherir* → *itsatsi* bi hizkuntzetan aditza da, *VM_ADI_SIN*⁶), baina beste zenbait pa-

⁶VM: Verb Main; ADI_SIN: Aditz Siplea

```

<pardef n='VM_ADI_SIN'>
  <e><p> <l><s n='mi' />VMG0000 </l>
    <r><s n='pos' />[ADI][SIN]</r> </p></e>
  <e><p> <l><s n='mi' />VMIC1P0 </l>
    <r><s n='pos' />[ADI][SIN]</r> </p></e>
  ... </pardef>
<pardef n='NC_IZE_ARR'>
  <e><p> <l><s n='mi' />NCFP000 </l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMP] </r></p></e>
  <e><p> <l><s n='mi' />NCFS000 </l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMS] </r></p></e>
  <e><p> <l><s n='mi' />NCFP00D </l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMP]
      <s n='suf' />txo[ATZ][IZE][ARR]</r></p></e>
  ... </pardef>
<pardef n='AQ_ADJ_ARR'>
  <e><p> <l><s n='mi' />AQFN000 </l>
    <r><s n='pos' />[ADJ][ARR]</r> </p></e>
  <e><p> <l><s n='mi' />AQFP000 </l>
    <r><s n='pos' />[ADJ][ARR]
      <s n='num' />[NUMP] </r> </p></e>
  ... </pardef>
<pardef n='VM_IZE_ARR'>
  <e><p> <l><s n='mi' />VMP00SF </l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMS] </r> </p></e>
  <e><p> <l><s n='mi' />VMP00PF </l>
    <r><s n='pos' />[IZE][ARR]
      <s n='num' />[NUMP] </r> </p></e>
  ... </pardef>

```

Irudia V.7: Paradigmak. Adibideak

pardef	l	r		
	mi	pos	num	suf
VM_ADI_SIN	VMG0000	[ADI][SIN]		
	VMIC1P0	[ADI][SIN]		
NC_IZE_ARR	NCFP000	[IZE][ARR]	[NUMP]	
	NCFS000	[IZE][ARR]	[NUMS]	
	NCFP00D	[IZE][ARR]	[NUMP]	txo[ATZ][IZE][ARR]
AQ_ADJ_ARR	AQFN000	[ADJ][ARR]		
	AQFP000	[ADJ][ARR]	[NUMP]	
VM_IZE_ARR	VMP00SF	[IZE][ARR]	[NUMS]	
	VMP00PF	[IZE][ARR]	[NUMP]	

Taula V.4: Paradigmak. Adibideak

l		r				
lem	mi	lem	pos	loc	det / num / per	...
aquel	DD0FP0	hura	[DET][ERKARR]	[ATZ]	[MUGM]	
el	DA0MS0				[MUGM] [NUMS]	
tanto	DI0MS0	horrenbeste	[DET][DZG]	[AUR]	[MG]	
alguno	DI0MS0				[MG]	
mi	DP1CSS	nire	[IOR][PERARR]		[MUGM] [NUMS]	
yo	PP1FP0	gu	[IOR][PERARR]		[GU]	
que	PR0CN0					
cuyo	PR0MP0	zeinen	[DET][IZL]		[MG]	
como	PR0000					
que	CS					
si	CS					
entonces	CS	beraz	[LOK][LOT]			

r			
...	lmi	post	spost
aquel	hura[DET][ERKARR][NUMS]		
el			
tanto	horrenbeste[DET][DZG][MG]		
alguno		[GEN]++bat	
mi	ni[IOR][PERARR][NUMS][NI] +[DEK][GEN]		
yo	gu[IOR][PERARR][NUP][GU]		
que			+n[ERL][MEN][ERLT]
cuyo			+n[ERL][MEN][MOS]
como			+n[ERL][MEN][MOS] ++bezala
que			+la[ERL][MEN][KONP]
si			ba[ERL][MEN][BALD]+
entonces			

Taula V.5: Katgoria itxiak. Adibideak

radigmetan katgoria desberdina izango du hizkuntza bakoitzak (adibidez, *chalado*→*ero* espainierazko sarreraren aditz partizipioa da eta euskarazko ordainean izen arrunta, *VM_IZE_ARR*⁷).

Katgoria itxien sekzioa

Sekzio honetan determinanteak, izenordainak eta loturazko elementuak aurkitzen dira, guztira 480 sarrera. Adibideak V.8 irudian eta V.5 taulan ikus daitezke.

Katgoria itxietan ezin dira paradigmak erabili informazioa trinkotzeko, ordaineko eremuak ebazteko sarreraren informazio morfologikoaz gain lema ere kontutan hartu behar delako (adibidez: *alguno* eta *tanto* sarreretan).

Azpimarratu behar dugu sarreraren abiapuntu-hizkuntzako lema eta infor-

⁷VM: Verb Main; IZE_ARR: Izen Arrunta

```

<section id="Kategoria_itxiak">
  <e><p> <l>aque<s n='mi' />DD0FP0</l>
    <r>haiek
      <s n='pos' />[DET][ERKARR]
      <s n='det' />[MUGM]
      <s n='num' />[NUMP]
      <s n='loc' />[ATZ]
      <s n='lmi' />hura[DET][ERKARR][NUMS]</r> </p></e>
  <e><p> <l>el<s n='mi' />DA0MS0</l>
    <r>
      <s n='det' />[MUGM]<s n='num' />[NUMP]</r> </p></e>
  <e><p> <l>tanto<s n='mi' />DI0MS0</l>
    <r>horrenbeste
      <s n='pos' />[DET][DZG]
      <s n='det' />[MG]
      <s n='loc' />[AUR]</r> </p></e>
  <e><p> <l>alguno<s n='mi' />DI0MS0</l>
    <r>
      <s n='det' />[MG]
      <s n='post' />[GEN]++bat</r></p></e>
  <e><p> <l>mi<s n='mi' />DP1CSS</l>
    <r>nire
      <s n='pos' />[DET][IZL]
      <s n='det' />[MUGM]
      <s n='num' />[NUMS]
      <s n='lmi' />ni[IOR][PERARR][NUMS][NI]+[DEK][GEN]
      </r></p></e>
  <e><p> <l>yo<s n='mi' />PP1FP0</l>
    <r>gu
      <s n='pos' />[IOR][PERARR]
      <s n='det' />[MG]
      <s n='per' />[GU]</r> </p></e>
  <e><p> <l>que<s n='mi' />PR0CN0</l>
    <r>
      <s n='vpost' />+n[ERL][MEN][ERLT]</r></p></e>
  <e><p> <l>que<s n='mi' />CS</l>
    <r>
      <s n='vpost' />+la[ERL][MEN][KONP]</r></p></e>
  <e><p> <l>cuyo<s n='mi' />PR0MP0</l>
    <r>zeinen
      <s n='pos' />[DET][IZL]
      <s n='mi' />[MG]
      <s n='vpost' />+n[ERL][MEN][MOS]</r></p></e>
  <e><p> <l>como<s n='mi' />PR0000</l>
    <r>
      <s n='vpost' />+n[ERL][MEN][MOS]++bezala</r></p></e>
  <e><p> <l>si<s n='mi' />CS</l>
    <r>
      <s n='vpost' />ba[ERL][MEN][BALD]+</r></p></e>
  <e><p> <l>entonces<s n='mi' />CS</l>
    <r>beraz
      <s n='pos' />[LOT][LOK]</r></p></e>
  ...
</section>

```

l		r	
lemma	mi	lemma	pos
.	Fp	.	Fp
,	Fc	,	Fc
¿	Fia		
?	Fit	?	Fit

Taula V.6: Puntuazio ikurrak. Adibideak

mazio morfologikoa agertzen direla, eta ez hitzaren forma. Adibidez, *aquellas* hitzarentzat *aquel/DD0FP0* lema eta informazio morfologikoa erabiltzen dira lexikoian kontsulta egiteko.

Ordainetan lema, kategoria eta azpikategoria, kokapena, numeroa, mugatasuna, lemaren informazio morfologikoa, postposizioa, menpeko postposizioa eta aditz-postposizio eremuak ager daitezke, baina sarreraren arabera eremu horietako batzuk soilik agertuko dira. Adibidez, *aquellas* hitza adierazten duen *aquel/DD0FP0* sarrerarentzat ordainaren formaz gain (*haiek*), bere kategoria eta azpikategoria ematen dira (*[DET]/[ERKARR]*) eta mugatua (*[MUGM]*) eta plurala (*[NUMP]*) dela, eta izenaren atzetik kokatzen dela (*[ATZ]*) adierazten duen informazioa. Gainera, ordainaren osaketa morfologikoa ematen da (*hura[DET]/[ERKARR]/[NUMS]*).

Sarrera batzuetarako ordainean ez dago informazio lexikorik, eta, beraz, ez da lema eta kategoriaren informaziorik ematen (adibidez: *el*, *alguno*, *que*, *como* eta *si* sarreretan).

Loturazko zenbait elementutan, soilik, *spost* menpeko postposizioa eman-go da (adibidez: *como*; *como quieren* → *nahi duten bezala*); Eremu horretako balioa, batzuetan, aurrizki bat izan daiteke (*si*; *si quieren* → *nahi badute*).

Puntuazio ikurren sekzioa

Puntuazio ikurren sekzioko sarrera gehienetarako lema mantentzen da, galdera eta harridura hasiera ikurretan (*¿,i*) ezik, non, ikur horiek euskaraz desagertzen direnez, ordaina hutsa den.

Sekzio honetan 23 sarrera kodetu dira. V.9 irudian eta V.6 taulan ikus ditzakegu adibide batzuk.

Erabiltzailearen hiztegia duen sekzioa

Sekzio honek sistemaren erabiltzaileak gehitutako sarrerak bilduko ditu, aurreko sekzioetako formatu bera jarraituz. Hiztegiko sarrerak gehitzeko tresnen diseinua eta inplementazioa etorkizuneko lanetarako utzi dugu.

```

<e><p> <l>entonces<s n='mi' />CS</l>

<section id="Puntuazio_ikurrak">
  <e><p> <l>.<s n='mi' />Fp</l>
    <r>.<
    <s n='pos' />[Fp]</r></p></e>
  <e><p> <l>,<s n='mi' />Fc</l>
    <r>,<
    <s n='pos' />[Fc]</r></p></e>
  <e><p> <l>><s n='mi' />Fia</l>
    <r></r></p></e>
  <e><p> <l>?<s n='mi' />Fit</l>
    <r>?<
    <s n='pos' />[Fit]</r></p></e>
  ...
</section>

```

Irudia V.9: Puntuazio ikurren sekzioaren formatua. Adibideak

V.2 Transferentzia estrukturala

Transferentzia estrukturalak abiapuntu-hizkuntzatik datorren egitura xede-hizkuntzarako egokia bihurtzen du. Tipologikoki hurbilak diren hizkuntzen artean egituren arteko alde txikia izango den bitartean, urrun dauden artean alde handia egongo da, prozesua zailduz. Inguruko erdal hizkuntzen eta euskararen artean desberdintasun sintaktiko handiak ematen dira.

Dibergentzia handi horiei aurre egiteko beharrezkoa da analisi sakon batetik abiatzea, non osagaien arteko mendekotasunak eta funtzio sintaktikoak identifikatu diren.

V.2.1 Espainiera eta euskararen arteko dibergentziak

Atal honetan itzulpena egiteko transferentzia estrukturalan kontuan hartu ditugun espainiera eta euskararen arteko dibergentziak aztertuko ditugu: alde batetik, postposizioen ingurukoak eta, bestetik, aditz-kateetan ematen direnak.

Zenbait sistematan hitzen eta sintagmen ordenan ematen diren dibergentziak transferentzia estrukturalan tratatzen dira⁸. Espainieraz eta euskaraz elementuen ordena hain desberdina denez, euskarazko hitzen ordena ez dugu transferentzia estrukturalan ebatziko, baizik eta euskarazko sorkuntza sintaktikoan.

⁸Adibidez, es-en sistematan espainieraz, normalean, izenaren ondoren kokatzen diren adjektiboen posizioa aldatzen da, ingelesez izenaren aurrean jartzen direlako.

Euskarazko postposiziotara itzultzen diren espainierazko elementuak

Euskara hizkuntza eranskaria da, postposizioak erabiltzen dituen. Badira bi postposizio mota: postposizio-atzizkiak⁹ (*etxean*) eta postposizio konposatuak (*etxearen aurrean*). Gainera, euskaraz badira kasu gramatikalak (aditzaren subjektua, objektu zuzena eta zeharkakoa) markatzen dituzten atzizkiak (*nor*, *nori* eta *nork* kasuak). Gure sisteman postposizioak eta kasu gramatikalak modu berean tratatzen dira. Beraz, hemendik aurrera, laburtzeko asmoarekin, postposizio deituko diegu guztiei.

Espainierazko zenbait elementuren itzulpenean euskarazko postposizioak erabiltzen dira (batzuetan, gainera, espainierazko elementu lexikal horiek euskaraz desagertzen dira):

- Artikuluak. Euskarara itzultzean mugatasun eta numeroaren informazioa ematen dute baina osagai lexikal bezala desagertzen dira (*el gato* → *katua*; *los gatos* → *katuak*).
- Menpeko konjuntzioak. Haien azpian dagoen aditz-sintagmari ematen diote menpekotasun-morfema. (*pregunta* / *cuándo* / *han llegado* → *noiz* / *iritsi diren* / *galdetzen du*). Batzuetan, artikuluen modura, haien osagai lexikala desagertzen da (*dice* / *que* / *han llegado* → *iritsi direla* / *esaten du*).
- Preposizioak eta funtzio sintaktikoak. Espainierazko preposizioak osagai lexikala bezala desagertzen dira eta chunken azkeneko hitzari itsatsiko zaizkion postposizioekin itzultzen dira (*en nuestra casa* → *gure etxean*). Postposizioekin ere esplizitu egiten dira subjektu eta objektu funtzio sintaktikoak, nahiz eta espainieraz funtzio horiek adierazteko preposiziorik ez egon (*el gato ha bebido leche* → *katuak esnea edan du*). Beraz, preposizio-eza (edo preposizio hutsa) kasu gramatikala adierazten duen postposizio batekin itzuliko da.

Aditz-kateak

Espainierak eta euskarak oso aditz-sistema ezberdina dute, bi hizkuntzetako egiturak oso desberdinak direlako. Ondoko ataletan, espainierazko eta euskarazko aditz-kateen ezaugarriak deskribatu ondoren, beren arteko itzulpena nola egin daitekeen azalduko dugu.

⁹Postposizio-atzizkiak gramatika klasikoetan *deklinabide-atzizki* deitu izan dira.

Espainierazko aditz-kateak

Espainierazko aditzek forma jokatuak (indikatibozko, subjuntibozko eta agintezko paradigmak) eta forma ez-jokatuak (infinitiboa, partizipioa eta gerundioa) dituzte. Forma jokatuetan ondoko informazioa duten morfemak daude: boza, modua, denbora, aspektua, pertsona eta numeroa. Forma ez-jokatuak, berriz, forma bakar bat dute.

Espainierazko aditz-formak bi kategoriatan banatzen dira: forma sinpleak, hitz bakar batez osatuta (*compro*); eta forma konposatuak, bi hitzez osatuta, *haber* adizki laguntzaile jokatu eta aditz nagusiaren partizipioa (*he comido*).

Aditz multzo txiki batek aditz-kate perifrastikoetan aspektuaren modifikatzailearen funtzioa betetzen du eta bere informazio semantikoa galtzen du (*he tenido que ir, voy a ver, debes trabajar, ando buscando...*). Unitate bakar bezala jokatzen duten perifrasietan ondoko elementuak agertzen dira:

1. Aditz perifrastikoaren forma sinple edo konposatua	<i>he tenido que</i>
2. Aditz nagusi ez-jokatu (infinitiboa edo gerundioa)	<i>comer</i>

Aditz iragankorrak boz aktiboan (*Luis compra unos libros*) edo pasiboan (*unos libros son comprados por Luis*) joka daitezke. Perpaus aktiboaren objektu edo osagarri zuzena (*unos libros*) subjektu bihurtzen da esaldi pasiboan, eta esaldi aktiboaren subjektua (*Luis*) esaldi pasiboan objektu agente bihurtzen da, *por* preposizioaz lagunduta (*por Luis*). Boz pasiboa izan daiteke perifrastikoa edo erreflexua. Boz pasibo perifrastikoa *ser* aditzarekin eraikitzen da (*unos libros son comprados por Luis*). Boz pasibo erreflexua *se* morfemarekin eraikitzen da (*se vende piso*).

Izenordain atonoak, edo klitikoak, aditzaren aurretik (banatuta) edo atzetik (itsatsita) agertzen dira: *me, te, se, nos, os, lo, los, la, las, le, les* (*se me caen, meteselas*).

Euskarazko aditz-kateak

Euskarazko adizkiak bi multzotan sailka daitezke: adizki jokatuak eta ez-jokatuak. Adizki jokatuak trinkoak edo analitikoak¹⁰ izan daitezke. Adizki

¹⁰Euskarazko gramatiketan adizki analitikoak izendatzeko adizki perifrastikoak erabiltzen da. Guk analitiko terminoa erabiliko dugu espainierazko perifrastiko kontzeptuarekin ez nahasteko.

trinkoak oso aditz gutxik dituzte: *izan*, *ukan*, *eduki*, *egon*, *etorri*, *joan*, *eraman*, *ekarri* eta *jakin*. Adizki trinkoetan informazio osoa forma bakar batean bilduta dago (*noakizu*, *dakartzat*).

Aditz guztiak (baita adizki trinkoak dauzkatenak ere) era analitikoan joka daitezke. Adizki analitikoetan (*joaten naiz*, *jango nituzke*) bi zati bereiz daitezke:

- Aditz nagusiaren forma (*joaten*, *jango*). Esanguraz gain aspektuari buruzko informazioa ematen du. Aditz nagusiak hainbat forma izan ditzake: aditz-oina (*eros*), burutua (*erosi*), ez-burutua (*erosten*) eta geroaldia (*erosiko*).
- Aditz laguntzailearen forma (*naiz*, *nituzke*). Pertsona, denbora eta modua adierazten du. Lau aditz laguntzaile daude euskaraz: *izan* eta **edun* indikatiborako, **edin* eta **ezan* subjuntiborako. Aditz iragan-korretan (osagarritzat objektu bat hartzen duten aditzak) **edun* eta **ezan* aditz laguntzaileak erabiltzen dira (*jan dut*); aditz iragangaitzetan (objekturik ez dutenak), berriz, *izan* eta **edin* laguntzailea (*joan naiz*).

Adizki jokatueta agertzen diren atzizki nagusiak ondokoak dira: komunztadura markatzailea, pluralgilea eta lehenaldiko edo subjuntibozko atzizkia. Aditz iragangaitzek komunztadura markatzen dute subjektuarekin (*ni nator*) eta batzuetan datiboarekin (*ni zuri natorkizu*). Aditz iragankorrek, berriz, behintzat bi argumentuekiko komunztadura adierazten dute, ergatiboa subjektuarekin eta absolutiboa objektu zuzenarekin (*nik oparia ekartzen dut*). Gainera, batzuetan, datiboaren komunztadura ere adierazten dute (*nik zuri oparia ekartzen dizut*).

Espainiera-euskara aditz-kateen itzulpena

Espainierazko aditz-kateak eta euskarazko beren itzulpenak oso desberdinak dira elementuei eta ordenari dagokienez. Ikus ditzagun aditz-kateen inguruan gertatzen diren fenomeno bakoitzerako itzulpenean egin beharreko egokitzapenak:

- Forma sinpleak. Aditzaren denboraren eta aspektuaren arabera, euskarazko aditz trinko (*voy* → *noa*) edo analitiko batez (*iré* → *joango naiz*, *compro* → *erosten dut*) itzul daitezke.
- Forma konposatuak. Aditz analitiko batez itzultzen dira (*he comprado* → *erosi dut*). Zenbaitetan espainierazko aditzaren aspektua eta

denbora adierazteko elementu berri bat, *izan* aditza, behar da (*hube comprado* → *erosi izan nuen*).

- Perifrasiak. Espainierazko aditz perifrastikoetan eta euskarazko itzulpenetan elementu desberdinak agertzen dira. Itzulpenak modu sistematikoan egin ahal izateko, euskarazko itzulpenean elementu berdinak dituzten espainierazko perifrasiak multzotan bildu ditugu.

- 1. multzoa (P1). Perifrasitik perifrasirako itzulpen arrunta. Multzo honek biltzen ditu infinitibozko perifrasi gehienak¹¹ (*poder, deber, querer, tener que, haber de, terminar de, dejar de, estar a punto de, empezar a, comenzar a, echar a, romper a, ir a*) eta gerundiozko perifrasiak (*acabar, andar, continuar, estar, ir, llevar, seguir, terminar, venir*).

Hona hemen multzo honetako perifrasi baten itzulpenaren adibidea: *hemos tenido que utilizar* → *erabili behar izan dugu*. Euskarazko itzulpenean ondoko elementuak agertzen dira: aditz nagusia (*erabili*), perifrastikoa¹² (*behar*) eta zenbaitetan adizki hutsa (*izan*), aspektu desberdinekin espainierazko aditzaren denboraren arabera, eta aditz laguntzailea (*dugu*).

Aditz laguntzailea zein izango den, aditz perifrastikoaren iragan-kortasunak ebatziko du: *nahi, behar, jarraitu, eraman, amaitu* aditz perifrastikoekin **edun* (edo **ezan*) aditz laguntzailea erabiliko da, eta *hasi, ari, egon, joan, etorri* aditzekin *izan* (edo **edin*). Bestalde, *ahal* aditz perifrastikoa erabiltzen denean, euskarazko aditz nagusiaren iragankortasunak ebatzen du laguntzailea: iragankorra bada **edun* (edo **ezan*) eta iragangaitza denean *izan* (edo **edin*).

- 2. multzoa (P2). Infinitibozko perifrasiatarako (*soler, deber de, acabar de, acostumar a*) euskarazko itzulpenean moduzko partikula bat (*ohi, omen, berri*) hartzen duena (*suelen cenar tarde* → *berandu afaldu ohi dute*).
- 3. multzoa (P3). Aditz perifrastiko bakarra du (*volver a*), bere itzulpenean moduzko partikula bat erabili beharrean adberbio bat (*berriz*) hartzen duena (*vuelven a gritar* → *berriz oihukatzen dute*).

¹¹Multzo honetan bada perifrastikoa ez den aditz bat, *querer*, baina bere itzulpena multzo honetako perifrastikoen modura egiten denez, bertan jarri dugu.

¹²Perifrastikoa deituko diogu espainierazko aditz perifrastikoaren baliokideari.

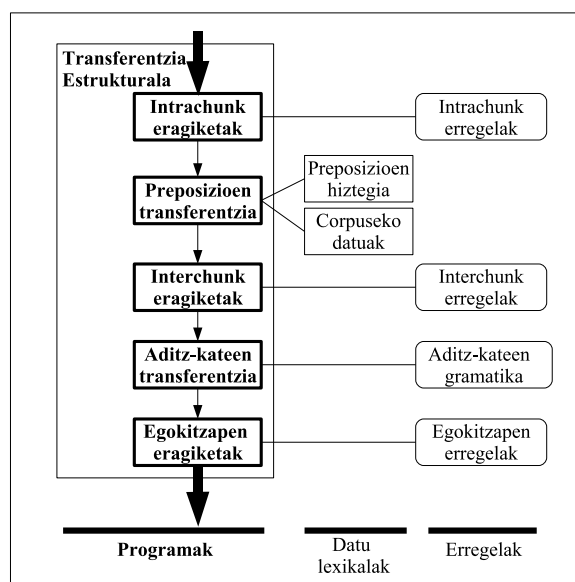
- 4. multzoa (P4). Aditz perifrastikoak (*llegar a, acertar a, alcanzar a*) egitura ez perifrastiko batekin itzultzen direnean (*no llega a entender* → *ez du ulertzen*).
- Forma ez-jokatuak. Espainierazko forma ez-jokatuak honela itzul daitezke euskaraz (lehenengo hurbilpen batean orokortasunari erreparatu diogu):
 - Infinitiboa: aditz-izen batez (*cantar es saludable* → *kantatzea osasungarria da*).
 - Gerundioa: inesiboa daraman aditz-izen batez (*vive cantando* → *kantatzen bizi da*).
 - Partizipioa: partizipio burutua gehi *-ta* moduzko atzizkia (*estamos cansados* → *nekatuta gaude*).
- Komunztadura-markatzaileak. Euskarazko aditzen komunztadura-markatzaileak erabakiko dira objektuei eta subjektuari buruzko informazioa eta aditzaren pertsona eta numeroa erabiliz.
- Boz pasiboa. Espainierazko boz pasiboan dauden aditzak, boz aktiboan egongo balira bezala itzul daitezke euskaraz (*los libros han sido comprados por María* → *Mariak liburuak erosi ditu*). Horretarako aditzaren subjektu komunztaduraren markatzailea ebazteko espainierazko subjektu pasiboa (*por María*) erabiliko da, eta objektuarena erabakitzeko objektu pasiboa (*los libros*).
- Izenordain atonoak. Izenordain atonoek objektuei buruzko informazioa gehitzeko erabiliko dira (*me los traen* → *ekartzen dizkidate*), edo aditz erreflexiboa sortzeko (*me rompo* → *hausten naiz*).

Espainiera eta euskararen arteko egitura-dibergentzia garrantzitsuenak aztertu ondoren, itzulpen-prozesuan dibergentzia horiei aurre egiteko transferentzia estrukturalako moduluen diseinua deskribatuko dugu.

V.2.2 Transferentzia estrukturalaren moduluak

Espainierazko egituratik euskarazko egiturara transferentzia burutzeko prozesua hainbat modulutan banatu dugu. Ataza bakoitza maila batean egiten du lan: chunk barruko nodoen artean edo esaldi barruko chunken artean.

Gure sistemako transferentzia estrukturala ondoko urratsetan burutzen da (ikus V.10 irudia):



Irudia V.10: Transferentzia estrukturalaren arkitektura

1. intrachunk eragiketak
2. preposizio eta funtzio sintaktikoen transferentzia
3. interchunk eragiketak
4. aditz-kateen transferentzia
5. egokitzapen-eragiketak

Ondoko ataletan deskribatuko ditugun modulu horien lana hobeto ulertzeko, V.1.1 atalean erabilitako *Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos* esaldia itzultzeko burututako prozesaketa erakutsiko dugu.

Intrachunk eragiketak

Intrachunk transferentzia estrukturalan chunk objektuaren barruko bi eragiketa mota burutzen dira:

- Informazio-mugimenduak. V.11 irudiko erregelek hainbat atributu nototik chunkera kopiatzen dituzte, hurrengo moduluetak eragiketetan chunkean izatea beharrezkoa izango delako.

Aditz-kateak ez diren chunketan mugatasunaren (*det*), numeroa (*num*) eta pertsonaren (*per*) informazioak nodoetatik chunkera kopiatzen dira. Informazio horiek beharrezkoak dira chunkaren postposizio-atzizkiaren sorkuntza morfologikorako, eta chunk hori subjektua, objektua zuzena edo zeharkako objektua bada, informazio hori interchunk mugimenduetan aditz-chunkari pasatuko zaio aditz-kateen transferentziarako beharrezkoa delako.

Aditz-chunketan aditz nagusiaren lema chunkera pasatzen da (*headlem*), eta aditz-katearen menpe dauden chunketan nodoetatik chunkera kopiatuko da preposizioaren informazioa (*prep*) eta beste zenbait informazio: izenaren informazio semantikoa (*headsem*), nodo nagusiaren kategoria (*headpos*) eta zenbakiak dituzten nodo hostoen informazioa (*leafpos*). Informazio guzti hori beharrezkoa izango da preposizio eta funtzio sintaktikoen transferentziarako.

Interchunk mugimenduetan ere mendekotasun-informazioa (*spost*) nototik chunkera pasatzen da, ondoren azpiko chunkari pasatzeko.

V.13 irudiko adibidean ikus dezakegu mugitu den informazioa: aditz-chunkean aditz nagusiaren lema (*eskatu*); subjektuan balio lexikala galdua duen artikularen (*las*) mugatasuna (*[MUGM]*), edozein nodoren numeroa (*[NUMP]*) eta izenaren (*elkarte*) informazio semantikoa (*[BIZ-]*); objektuan *zetzutasun* izenaren informazio semantikoa (*[BIZ-]*); eta zeharkako objektuan artikularen (*los*) mugatasuna (*[MUGM]*) eta numeroa (*[NUMS]*).

- Nodoen ezabaketa. Euskaraz informazio lexikalik ez duten nodoak ez dira xede-hizkuntzako egituraren behar eta behar zen informazioa mugitu denez ezaba daitezke.

V.12 irudiko erregelak informazio lexikorik ez duten determinante artikulak (*el, la...*), preposizioak, eta hainbat konjuntzio (*que, cuando...*) ezabatzen dituzte, zeukaten informazio esanguratsua chunkari pasatu zaiola aurreko urratsean.

V.13 irudiko adibidean ikus dezakegu hiru nodo ezabatu direla: AHko testuan *las* eta *los* artikulak zituzten nodoak eta *a* preposizioarena

```

<!-- Aditz-kateak ez diren chunketan, nodotik chunkera kopiatzen da:>
<rule id = 'C/N_te_det'> <!-- Mugatasunaren informazioa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@det][1] </def></match>
    <actions> <act> C/@det := N/@det </act></actions>
  </rule><!------->
<rule id = 'C/N_te_num'> <!-- Numeroaren informazioa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@num][1] </def></match>
    <actions> <act> C/@num := N/@num </act></actions>
  </rule><!------->
<rule id = 'C/N_te_per'> <!-- Pertsonaren informazioa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@per] </def></match>
    <actions> <act> C/@per := N/@per </act></actions>
  </rule><!------->
<rule id = 'C/N_te_prep'> <!-- Preposizioaren informazioa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@prep] </def></match>
    <actions> <act> C/@prep := N/@prep </act></actions>
  </rule><!------->
<rule id = 'C/N_te_spost'> <!-- Mendekotasun-informazioa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@spost] </def></match>
    <actions> <act> C/@spost := N/@spost </act></actions>
  </rule><!------->
<rule id = 'C/N_te_headsem'> <!-- Informazio semantikoa>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@sem] </def></match>
    <actions> <act> C/@headsem := N/@sem </act></actions>
  </rule><!------->
...
<!-- Aditz-kateak diren chunketan, nodotik chunkera kopiatzen da:>
<rule id = 'C/N_te_headlem'> <!-- Aditz nagusiaren lema>
  <match>
    <def> C := //CHUNK[@type='adi-kat'] </def>
    <def> N := C//NODE[@lem] </def></match>
    <actions> <act> C/@headlem := N/@lem </act></actions>
  </rule><!------->

```

Irudia V.11: Intrachunk mugimenduetarako erregelak

```

<!-- Lema hutsik duten nodoak ezabatu>
<rule>
  <match>
    <def> N := //NODE[not(@lem)] </def></match>
  <actions>
    <act> delete(N) </act></actions>
</rule>

```

Irudia V.12: Nodoen ezabaketarako erregela

Preposizio eta funtzio sintaktikoen transferentzia

Preposizio eta funtzio sintaktikoen transferentzian burutzen den prozesaketa chunk mailakoa da. Beharrezkoa den nodoetako informazioa aurreko urratsean mugitu da chunketara.

Modulu honetan bi egoera desberdin tratatzen dira:

- Aditz-kate baten osagaiak diren chunketako preposizio eta funtzio sintaktikoen transferentzia. Hautapen-erregelak dituen preposizioen hiztegia eta azpikategorizazio-patroien estrategia erabiltzen dira, aditz-chunkaren eta bere osagaietako informazioarekin. Prozesaketa honetan ere aditz-chunkaren iragankortasuna ebazten da.
- Izenen modifikatzailea den chunk baten preposizioaren transferentzia. Kasu honetan soilik hautapen-erregelak dituen preposizioen hiztegia erabiltzen da, preposiziodun chunkaren eta modifikatzen duen guraso-chunkaren informazioarekin.

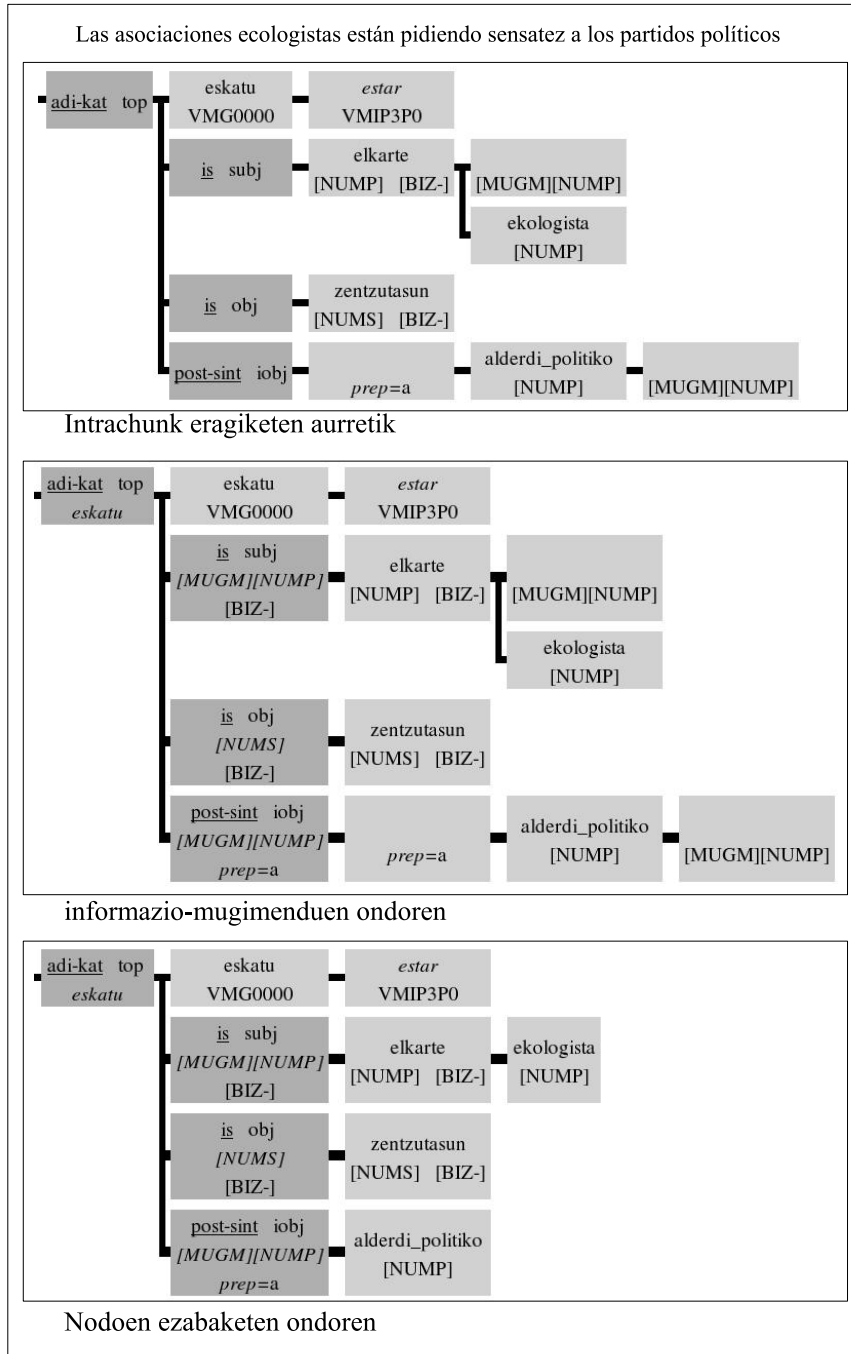
Modulu honetako egoera bakoitzerako aplikatzen diren erregelak V.14 irudian ikus daitezke. Erregeletan deitzen den *&preptrans* funtzioak burutzen duen prozesaketa V.2.3 atalean deskribatuko dugu.

Modulu honen sarrera eta irteera, gure adibiderako, V.15 irudian ikus dezakegu (irudia argiagoa izateko nodoak ez dira erakusten, urrats honetan ez delako haien informazioa aldatzen).

Interchunk eragiketak

Interchunk transferentzia estrukturalean chunketan dagoen zenbait informazio osatzen da, chunken artean informazioa mugitzen da eta chunk batzuk ezabatzen dira:

- Pertsona-informazioaren osaketa. Subjektu, objektu edo zeharkako objektua diren chunketan pertsona-informazioa ez badago, osatzen da,



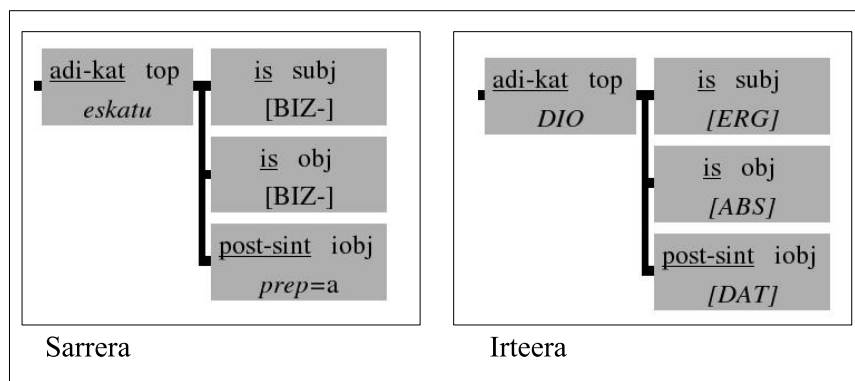
Irudia V.13: Intrachunk transferentzia estrukturala. Adibidea

```

<!-- Izenen modifikatzailea baten preposizioaren transferentzia>
<rule id = 'C/C_te_preptransf_adikat'>
  <match>
    <def> C1 := //CHUNK[@type!='adi-kat'] </def>
    <def> C2 := C/CHUNK[@prep] </def></match>
    <actions>
      <act> C2/@post := \&preptrans[C1,C2] </act></actions>
    </rule><!------->
<!-- Aditz-kate baten osagaien preposizio eta>
<!-- funtzio sintaktikoen transferentzia>
<rule id = 'C/C_te_preptransf_adikat_ez'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat'] </def>
    <def> C2 := C/CHUNK </def></match>
    <actions>
      <act> TransPost := \&preptransf(C1,{C2}) </act></actions>
      <act> C1/@trans := \&getTrans(TransPost) </act></actions>
      <act> {C2/@post}:= \&getPosts(TransPost) </act></actions>
    </rule><!------->

```

Irudia V.14: Preposizio eta funtzio sintaktikoen transferentziarako erregelak



Irudia V.15: Preposizio eta funtzio sintaktikoen transferentzia. Adibidea

```

<!-- Subjektu edo objektuek pertsona informaziorik ez dutenean:>
<!-- Numero singularra edo numerorik ez duenean, 3.sing ezarri (HU)>
<rule id='C_te_per_hu'>
  <match>
    <def> C := //CHUNK[(@si='subj') or (@si='obj') or(@si='iobj')]
              [not(@per)]
              [@num='[NUMS]' or not(@num)]          </def>
    <actions>
      <act> C1/@per := '[HU]' </act>></actions>
  </rule><!------->
<!-- Numero plurala duenean, 3.pl ezarri (HK)>
<rule id='C_te_per_hk'>
  <match>
    <def> C := //CHUNK[(@si='subj') or (@si='obj') or(@si='iobj')]
              [not(@per)]
              [@num='[NUMP]']                      </def>
    <actions>
      <act> C1/@per := '[HK]'                      </act>></actions>
  </rule><!------->

```

Irudia V.16: Pertsona-informazio osatzeko erregelak

aditz-transferentziarako beharrezkoa izango delako (ikus V.16 irudiko erregelak): chunkaren numeroa singularra bada edo numerorik ez bada, singularreko 3. pertsona (*[HU]*) esleituko da; chunkaren numeroa plurala bada, pluraleko 3. pertsona (*[HK]*).

- Chunketik chunkera informazio-mugimenduak. hurrengo modulueta-rako beharrezkoak izango diren hainbat atributu chunketik chunkera kopiatzen dituzte V.17 irudiko erregelek.

Aditz-kateari bere azpian dauden subjektuaren, objektuaren eta zeharkako objektuaren pertsona-informazioak pasatzen zaizkio *subper*, *objper* eta *datper* atributuetan.

Bestalde, izen-sintagmak koordinatuetan juntagailua duen chunkak jasotzen diren postposizio-informazioa (*post*) koordinatutako chunkei pasatzen die eta mendekotasun-postposizioa (*spost*) duten chunkek, bere azpiko chunkari.

- Aditz-chunkeko pertsona-informazioen osaketa. Aditz-transferentzian *subper*, *objper* eta *datper* atributu guztiak beharrezkoak dira. Beraz, baten bat falta bada (subjektua, objektua edo zeharkako objektua ez zegoelako) *[00]* balioa esleitzen zaio falta den atributuari (ikus V.18 irudiko erregelak).
- Chunken ezabaketa. Intrachunk eragiketetan lema hutsa zuten nodoak

```

<!-- Pertsonaren informazioa subjektutik aditz-chunkera pasa>
<rule id='C/C_te_subper'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']      </def>
    <def> C2 := ./CHUNK[@si='subj']          </def></match>
  <actions>
    <act> C1/@submper= C2/@per                </act></actions>
  </rule><!------->
<!-- Pertsonaren informazioa objektu zuzenetik aditz-chunkera pasa>
<rule id='C/C_te_objper'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']      </def>
    <def> C2 := ./CHUNK[@si='obj']           </def></match>
  <actions>
    <act> C1/@objper:= C2/@per                </act></actions>
  </rule><!------->
<!-- Pertsonaren informazioa zeharkako objektutik aditz-chunkera pasa>
<rule id='C/C_te_datper'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']      </def>
    <def> C2 := ./CHUNK[@si='iobj']          </def></match>
  <actions>
    <act> C1/@datper:= C2/@per                </act></actions>
  </rule><!------->

<!-- Koordinatutako sintagmei juntagailutik postposizio-inform. pasa >
<rule id='C/C_te_emen'>
  <match>
    <def> C1 := //CHUNK[@type='emen-is']      </def>
    <def> C2 := ./CHUNK                      </def></match>
  <actions>
    <act> C2/@post := C1/@post                </act></actions>
  </rule><!------->
<!-- Menpekotasun-informazioa azpiko chunkera pasa >
<rule id='C/C_te_inter_spost'>
  <match>
    <def> C1 := //CHUNK[@spost]               </def>
    <def> C2 := ./CHUNK                     </def></match>
  <actions>
    <act> C2/@spost := C1/@spost              </act>
  </rule><!------->

```

Irudia V.17: Interchunk mugimenduetarako erregelak

```

<!-- Subjektuaren pertsona-informaziorik ez zegoenez, [00] ezarri>
<rule id='C_te_subper'>
  <match>
    <def> C := //CHUNK[@type='adi-kat']][not(@subper)] </def>
    <actions>
      <act> C/@subper := '[00]' </act></actions>
    </match><!------->
<!-- Objektu zuzenaren pertsona-informaziorik ez zegoenez, [00] ezarri>
<rule id='C_te_objper'>
  <match>
    <def> C := //CHUNK[@type='adi-kat']][not(@objper)] </def>
    <actions>
      <act> C/@objper := '[00]' </act></actions>
    </match><!------->
<!-- Zeharkako objektuaren perts-informaziorik ez zegoenez, [00] ezarri>
<rule id='C_te_datper'>
  <match>
    <def> C := //CHUNK[@type='adi-kat']][not(@datper)] </def>
    <actions>
      <act> C/@datper := '[00]' </act></actions>
    </match><!------->

```

Irudia V.18: Aditz-chunkeko pertsona-informazioen osaketa

```

<!-- Nodorik gabeko chunkak ezabatu>
<rule id='C_te_delete'>
  <match>
    <def> C := //CHUNK[not(./NODE)] </def>
    <actions>
      <act> delete(C) </act></actions>
    </match><!------->

```

Irudia V.19: Chunken ezabaketarako erregela

ezabatu ziren, batzuetan chunkak nodorik gabe utziz. Nodorik ez duten chunk horiek ezabatzen dira (ikus V.19 irudiko erregela), XHko egi-turan ez direlako behar, chunk horiek duten beharrezko informazioa dagokien chunkei pasako zaiela. Eragiketa hau burutuko da, adibidez, konjuntzio bat (adibidez: *que*, *cuando*) duten chunketan, aurretik in-trachunk transferentzia estrukturalan nodoa ezabatu delako

V.20 irudian ikus dezakegu *Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos* adibidearekin burututako prozesaketa. Aditz-katearen menpe dauden subjektu eta objektuen pertsona-informazioa duten *per* atributuak egokitu ondoren, aditz-chunkera kopiatu dira ([HK], [HU], [HK]). Adibide honetarako ez da beste mugimendurik egin,


```

<!-- Aditz-chunketako nodoen transferentzia>
<rule id='C/N//N_verbtransf'>
  <match>
    <def> C := //CHUNK[@type='adi-kat']          </def>
    <def> N1:= C/NODE                            </def>
    <def> N2:= C/NODE//NODE                      </def></match>
  <actions>
    <act> Input = concat ( '#' , {N2/@lem,N2/@mi},N1/@lem,N1/@mi, '/',
                          C/@subper,C/@objper,C/@datper, '/',
                          C/@trans, '/',N1/@lem,'\$') </act>
    <act> delete(N1)                             </act>
    <act> delete(N2)                             </act>
    <act> Output := &verbtransf(Input)            </act>
    <act> &verbchunkconstruct(C,Output)          </act></actions>
</rule><!------->

```

Irudia V.21: Aditz-kateen transferentziarako erregelak

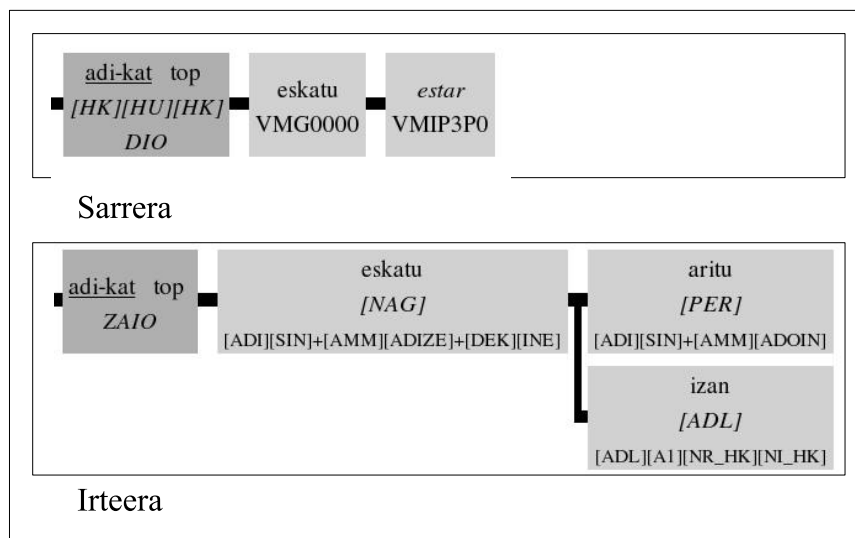
berrirako hiru nodo ematen dira “/” ikurrarekin banatuta: lehenengo nodoa (*eskatu*) erroa izango da eta besteak (*aritu* eta *izan*) bere azpikoak. Nodo bakoitzarentzat “<” eta “>” ikurren arteko informazioa sorkuntza sintaktikoan nodoen ordena ebazteko erabiliko den *loc* atributuan gordetzen da, eta bestelako informazioa *post* atributuan, sorkuntza morfologikorako.

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos adibiderako prozesaketa V.22 irudian ikus daiteke.

Egokitzapen-eragiketak

Transferentzia estrukturala bukatzeko bi moldaketa egin behar dira (ikus V.23 irudiko erregelak):

- Aditz-kateen transferentzian, aditz-katearen iragankortasuna aldatu bada, subjektuaren kasua egokitu behar da. Gure adibidean [ERG] ezarrita zuen subjektua duen chunkari [ABS] esleitzen zaio.
- Mendekotasun-morfemarik izango balitz aditz-chunkean, informazio hori chunkeko aditz-laguntzailea duen nodoari pasatzen zaio *post* atributuan duen informazioari gehituz.



Irudia V.22: Aditz-kateen transferentzia. Adibidea

```

<!-- Aditza iragankorra bada eta subjektuan ABS bazegoen, ERG ezarri>
<rule id='C/C_te_abs_erg'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']
              [(@trans='DU')or(@trans='DIO')] </def>
    <def> C2 := C1/CHUNK[@si='subj'][@post='ABS'] </def></match>
  <actions>
    <act> C2/@post := 'ERG' </act></actions>
</rule><!--
<!-- Aditza iragangaitza bada eta subjektuan ERG bazegoen, ABS ezarri>
<rule id='C/C_te_erg_abs'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']
              [(@trans='DA')or(@trans='ZAI0')] </def>
    <def> C2 := C1/CHUNK[@si='subj'][@post='ERG'] </def></match>
  <actions>
    <act> C2/@post := 'ABS' </act></actions>
</rule><!--
<!-- Aditz-chunkak menpekotasun-inform. badu, aditz-laguntzaileari pasa>
<rule id='C/N_te_spost_adl'>
  <match>
    <def> C := //CHUNK[@type='adi-kat'][@spost] </def>
    <def> N := C/NODE[@loc='ADL'] </def></match>
  <actions>
    <act> N/@post := concat(N/@post,C/@spost) </act></actions>
</rule><!--

```

Irudia V.23: Egokitzapen eragitetarako erregelak

V.2.3 Preposizio eta funtzio sintaktikoen transferentziaren deskribapena

Preposizioen itzulpena ataza zaila eta garrantzitsua da IAko sistema batean. Preposizio-sintagmak (PS; *Prepositional Phrase*, *PP*) ondo itzultzea erronka handia da, preposizioak ez baitira modu sistematikoan itzultzen:

Even between two closely related languages such as English and French, locative prepositions of even simple sentences do not seem to be translated from one language to the other in a clearly systematic and coherent way. (Japkowicz eta Wiebe, 1991)

Bi hizkuntza hurbilen arteko preposizio mota bat itzultzea zaila bada, erraz uler daiteke distantzia handiko bi hizkuntzen artean preposizio mota guztiak itzultzea oso eginkizun konplexua dela.

Ondoko ataletan itzulpen automatikoaren arloan preposizioak itzultzeko erabili izan diren estrategien azterketa, *Matxin* sistemarako diseinatu duguna eta gure sistemarako saiatu ditugun beste estrategia batzuk aurkeztuko ditugu.

Preposizioak itzultzeko estrategien azterketa

Preposizioen itzulpenari buruzko erreferentzia gutxi aurki ditzakegu literaturan, eta azken hamarkadan ez da batere aldatu Trujillo-k (1995) deskribatzen zuen egoera:

The specific aspects of preposition translation are in most cases left as a matter to be resolved during system development, and in the rare cases when explicit mention is made of this issue only a brief indication (with a few notable exceptions) is made about how this problem is tackled. (...)

Descriptions of the translation of prepositions are even more scant than those of transfer MT systems in general. This may be due to the complexity of the problem or to the lack of a research tradition in this area.

Antzeko zerbait, ere, adierazten zuen John Hutchins-ek *LinguistList* postaz-zerrendan 1998an¹³:

There are few treatments specifically devoted to prepositional phrases in MT. However, this problem area will have been treated in the context of valency, case grammar, and general syntactic and

¹³<http://www.linguistlist.org/issues/9/9-1332.html>

semantic analysis – particularly for languages such as Japanese, German and Russian

Preposizio-sintagmen itzulpenari aurre egiten zaion lanetan bi hurbilpen proposatzen dira:

- Testuinguruan oinarritutakoa (Mamidi, 2004; Trujillo, 1992a, b; Alam, 2004; Naskar eta Bandyopadhyayn, 2006). Preposizio baten itzulpena hautatzeko ondoko elementu hauek hartu behar dira kontuan: modifikatutako elementua (kanpoko argumentua), preposizioa bera eta elementu modifikatzailea (barneko argumentua). Testuinguruko elementu horien informazio semantikoa erabiltzen bada lexikoian kodetuta egon daitezke, edo *WordNet* bezalako kanpo-iturri batetik jaso ahal da.

Estrategia hau transferentzia bidezko hurbilpenetan erabiltzen da. Egokia da distantzia handia duten hizkuntzetarako, baina soilik hizkuntza pare bakar baterako eta norabide batean itzulpena egiteko balio du.

- Kontzeptuan oinarritutakoa (Mamidi, 2004; Trujillo, 1992a; Nubel, 1996; Japkowicz eta Wiebe, 1991). Metodo hau abstraktuagoa da: bere helburua preposizioaren egitura semantikoaren maila ertain batera iristea da, eta horretarako preposizioen/postposizioen analisi sakon bat burutu behar da. Preposizio-sintagmak kasu sakonekin eta predikatu unibertsalekin, edo adiera idealekin (*ideal meanings*) eta kontzeptualizazioekin adieraz daitezke.

Interlingua bidezko hurbilpenetan erabiltzen den kontzeptuan oinarritutako eredu hori. egituraren eta kulturaren aldetik hurbilak diren hizkuntzen artean itzultzeko egokiagoa da, bi norabideetako sistema eleanitzetarako aproposa izanik, itzulgarria delako.

Sistema desberdinetako preposizioen itzulpenerako metodoak aztertu ditugu, gure sistemarako hoberen egokitzen den estrategia diseinatzeko. *Sys-tran*, *Metal* eta *Eurotra* sistemen berri (Trujillo, 1992a) lanetik jaso dugu, eta ingelesa-bengalera sistema batena (Naskar eta Bandyopadhyayn, 2006) lanetik.

- *Sys-tran*. Preposizioak itzultzeko preposizioaren aurretik edo atzetik dagoen izen edo aditzaren markagailu semantikoa aztertzen du, baina ez du izen eta preposizioen semantikaren teoria koherenterik.
- *Metal*. Preposizioen itzulpena hautatzeko preposizioaren barne argumentua den izenaren semantikan oinarritzen da. Horretarako lexikoi

elebidunean AHko preposizio bakoitzaren itzulpenarekin batera, preposizioaren barne-argumentuak bete behar dituen murrizpen semantikoak kodetzen diren erregelak ematen dira. Markatzaile semantikoak sinple eta orokorrak dira, bere erabilera minimizatzen saiatzen dela. Ahal den guztietan soilik informazio sintaktikoa erabiltzen da. Soilik preposizioaren barne-argumentua den izenaren markatzaileak aztertzen dira, eta, batzuetan, horrekin ez da nahikoa.

- *Eurotra*. Preposizio-sintagmen preposizioak aditzarekin batera itzultzen dira. Desanbiguatzeko erregelek preposizioaren barneko eta kanpoko argumentuen ezaugarriak erabiltzen dituzte.
- Ingelesa-bengalera IA sistema (Naskar eta Bandyopadhyayn, 2006). Guretzat sistema hau interesgarria da bengalera eratorpena eta postposizio-hitzak daudelako, euskaraz bezala. Testuinguruan oinarritutako hurbilpena jarraitzen dute postposizio egokia aukeratu ahal izateko, preposizio sintagmako buru den izenaren informazio semantikoa *WordNet*en bilatzen dutela.

Preposizioen eta funtzio sintaktikoen transferentzia *Matxin* sisteman

Matxin sisteman preposizioak eta funtzio sintaktikoak itzultzeko bi estrategia nagusi erabiltzen ditugu. Alde batetik, testuinguruan oinarritutako hurbilpenetan bezala, preposizioaren itzulpena hautatzeko kanpo- eta barne-argumentuetako informazio lexiko, sintaktiko eta semantikoa erabiltzen da, preposizioen hiztegiko hautapen-erregelak aplikatuz. Bestetik, aditz-kate baten azpiko preposizio-sintagmen preposizio guztiak batera ebazten saiatzen gara, corpusetik erauzitako azpikategorizazio patroiak erabiliz.

Aditzen modifikatzaileak diren preposizio eta funtzio sintaktikoen transferentzia burutzeko hurrengo ataletan sakonago azalduko ditugun hiru urrats ematen dira:

1. Preposizioen hiztegia eta hautapen-erregelak. Hiztegian preposizio baterako ematen diren postposizio posibleetako baten erregelako baldintza betetzen bada, postposizio hori hautatuko da. Postposizio bat baino gehiagoko erregeletako baldintzak betetzen badira, postposizio multzo hori hurrengo urratsera pasako dira, eta ez bada postposizio bat ere hautatu erregelekin, erregelen baldintzek baztertu ez dituzten postposizio posibleak pasako dira.
2. Azpikategorizazio-patroiak. Aditz-chunk bakoitzerako aditzaren lema eta bere azpiko chunketako preposizio bakoitzerako aurreko urratsean

emandako postposizio-multzoak corpusetik erauzitako azpikategorizazio-patroiekin kontrastatzen dira. Postposizio gehien desanbiguatzen duen patroia hautatzen da. Zenbait postposizio desanbiguatu gabe geldi daitezke, haien balio posibleen multzoak hurrengo urratsera pasatzen direla.

3. Hiztegiko ordena. Desanbiguatu ez diren postposizio posibleen multzo bakoitzetik, sistemako hiztegiko ordenan aurrena dagoena hautatzen da.

Izen-sintagmak modifikatzen dituzten preposizioak modu sinpleagoan ebazten dira, soilik hiztegia eta hautapen-erregelak erabiltzen direla. Ebatzi gabe gelditzen diren preposizioetarako hiztegiko aurrenekoa ere hautatzen da.

Hurrengo ataletan aipatutako hiru urratsak sakonki azaldu ditugu: preposizioen hiztegia eta hautapen-erregelak, azpikategorizazio-patroien metodoa eta hiztegiko ordean.

Preposizioen hiztegia eta hautapen-erregelak

Preposizioen hiztegia eskuz eraiki dugu, preposizio bakoitzari (preposizio hutsa barne) dagozkion euskarazko posposizioak definituz.

Adibidez *de* preposizioaren itzulpen posibleak hauek dira:

lo pintan <i>de verde</i>	→	<i>berde-z</i>	(instrumentala)
la casa <i>de madera</i>	→	<i>egur-rezko</i>	(instrumentala gehi lekuzko genitiboa)
la casa <i>del monte</i>	→	<i>mendi-ko</i>	(lekuzko genitiboa)
la casa <i>de Mikel</i>	→	<i>Mikel-en</i>	(genitiboa)
viene <i>de casa</i>	→	<i>etxe-tik</i>	(ablatiboa)

Preposizio batek itzulpen posible bat baino gehiago duenean, haietako bat hautatu behar da. Horretarako, preposizio bakoitzarekin bere itzulpena izan daitezkeen postposizio posibleak kodetu dira, ahal denean postposizio horri lotutako hautapen-erregela batekin

Hautapen-erregelek postposizioak hautatu edo baztertu egin ditzakete, bi motatako baldintzak erabiliz:

- Baldintza nahikoa (*baldin*, $\Rightarrow$). Baldintza betetzen bada postposizioa hautatzen da, eta ez bada betetzen postposizioa ez da baztertzen.

Preposizioa	Baldintza	Mota	Postposizioa
de	./[@headpos='ADJ'] or ./[@headsem='DEN']	⇔	[INS]
de	./[@headsem='GAI']	⇔	[INS]+[GEL]
de	./[@headsem='BIZ+'] and ./[@type!='adi-kat']	⇒	[GEN]
de	(./[@type!='adi-kat'] and ./[@headsem='BIZ-']) or (./[@type='adi-kat'] and ./[headlem='izan']) or (./[@type!='adi-kat'] and ./[@headpos='Z']) or ./[@leafpos='Z'] or ./[@headpos='W']))	⇒	[GEL]
de_dentro_de	./[@type='adi-kat']	⇔	[GEN]++barrutik
de_dentro_de	./[@type!='adi-kat']	⇔	[GEN]++barruko

Taula V.7: Preposizio hiztegia eta hautapen-erregelak

- Baldintza beharrezko eta nahikoa (*baldin eta soilik baldin*, *bsb*, ⇔). Baldintza betetzen bada postposizioa hautatzen da, eta ez bada betetzen postposizioa baztertzen da.

Hautapen-erregelatan postposizioen bat behintzat hautatu bada, hurrengo urratsera hautatutako postposizioak pasako dira, eta ez bada postposiziorik hautatu baztertu ez diren postposizioak hurrengo urratsera pasako dira.

Erregelek ondoko informazioa erabiltzen dute:

- Desanbiguatu nahi den postposizioa duen chunkeko nodoen informazio lexiko, sintaktiko eta semantikoa (lema, kategoria eta azpikategoria, bizidun ezaugarria...). Erabiltzen den informazioa chunkaren atributuetan dago, intrachunk mugimenduetan kopiatu baita.
- Postposizioa duen chunkaren gurasoa den chunkaren informazioa.

Erregelen adibideak V.7 taulan eta V.24 irudian ikus daitezke. Adibidez, *de* preposiziorako postposizio instrumentala hautatuko da baldin eta soilik baldin postposizioa duen chunkaren gunea adjektibo bat bada (*lo pinto de verde* → *berdez margotzen dute*), eta genitiboa hautatuko da, berriaz, baldin postposizioa duen chunkaren gunea bizidun ezaugarri semantikoa duen izen bat bada eta bere gurasoa aditz-katea ez bada (*la casa de la amiga* → *lagunaren etxea*). Bestalde, *de_dentro_de* preposizio konposatua *[GEN]++barrutik* postposizioarekin itzuliko da bere gurasoa aditz-kate bat bada (*sale de_dentro_de casa* → *etxearen barrutik ateratzen da*), eta *[GEN]++barruko* postposizioarekin bere gurasoa ez bada aditz-kate bat (*la puerta de_dentro_de casa* → *etxearen barruko atea*).

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos adibiderako hiru chunketako preposizio edo funtzio sintaktikoak ebatzi behar dira:

```

<e><p> <l>de</l>
      <r>[INS]</r>
      <cond type='bsb'>
        ./[@headpos='[ADJ]'] or ./[@headsem='[DEN]'] </cond> </p></e>
<e><p> <l>de</l>
      <r>[INS]+[GEL]</r>
      <cond type='bsb'> ./[@headsem='[GAI+]'] </cond> </p></e>
<e><p> <l>de</l>
      <r>[GEN]</r>
      <cond type='baladin'> ./[@headsem='[BIZ+]']
                           and ./[@type!='adi-kat'] </cond> </p></e>
<e><p> <l>de</l>
      <r>[GEL]</r>
      <cond type='baladin'>
        ( ./[@type!='adi-kat'] and ./[@headsem='[BIZ-]'] ) or
        ( ./[@type='adi-kat'] and ./[headlem='izan'] ) or
        ( ./[@type!='adi-kat'] and
          ( ./[@headpos='[Z]'] or ./[@leafpos='[Z]'] or
            ./[@headpos='[W]']
          )
        )
      </cond> </p></e>
<e><p> <l>de_dentro_de</l>
      <r>[GEN]++barrutik</r>
      <cond type='bsb'> ./[@type='adi-kat'] </cond> </p></e>
<e><p> <l>de_dentro_de</l>
      <r>[GEN]++barruko</r>
      <cond type='bsb'> ./[@type!='adi-kat'] </cond> </p></e>

```

Irudia V.24: Preposizioen hiztegia eta hautapen-erregelen formatua

	Preposizio kop.	Postposizio kop.	Erregela kop.
Preposizio hutsa	1	12	10
Preposizio sinpleak	18	53	26
<i>ez anbiguoak</i>	5	5	0
<i>anbiguoak</i>	13	48	26
Preposizio konposatuak	333	397	53
<i>ez anbiguoak</i>	287	287	16
<i>anbiguoak</i>	46	110	37
Guztira	352	462	89

Taula V.8: Preposizio hiztegiaren datuak

- *Las asociaciones* duen chunkaren preposizio hutsarentzat ez da postposizio bat ere erregelekin hautatzen, asko baztertzen direla. Baztertu ez diren *[ABS]* eta *[ERG]* postposizioak pasako dira hurrengo urratse-ra.
- *sensatez* duen chunkaren preposizio hutsarentzat aurrekoaren berdina gertatzen da, *[ABS]* eta *[ERG]* postposizioak pasatuz.
- *a los partidos políticos* duen chunkaren *a* preposizioarentzat ez dago postposiziorik hautatzen duen erregelarik, baina *[INE]* baztertzen da, eta, beraz, *[DAT]*, *[ABS]*, eta *[ALA]* pasatzen dira.

Azpimarratu behar dugu funtzio sintaktikoaren informazioa ezagutzen badugu, hau da, objektu zuzena edo zeharkakoa dela, postposizio hautatzeko erregela bat kode genezakeela, objektu zuzenari absolutiboaren postposizioa (*[ABS]*) esleituz eta zeharkakoari datiboarena (*[DAT]*). Baina ezin dugu hau egin, gaur egun analizatzaileak objektuei buruz ematen duen funtzio sintaktikoaren informazioa askotan okerra delako.

Preposizioen hiztegian 352 preposizio kodetu ditugu: preposizio hutsaz gain, 18 preposizio sinple (*a*, *ante*, *bajo*...) eta 333 preposizio konposatu (*a base de*, *a beneficio de*, *a cambio de*, *en honor de*...). Preposizio bakoitzerako postposizio bat edo gehiago kodetu dira, guztira 462, horietako 89tan kodetu direla erregelak. Datu guztiak V.8 taulan ikus daitezke

Azpikategorizazio-patroiak

Postposizioen desanbiguaziorako euskarazko azpikategorizazio-patroiak ere erabil daitezke. Patroi horiek (Aldezabal *et al.*, 2002) lanetik jaso ditugu,

non *Euskaldunon Egunkaria* izeneko egunkariko corpus batetik (milioi bat hitz) erauzi dituzten¹⁴.

Azpikategorizazio-patroi horietan datu hauek ematen dira: aditzaren lema, subjektuaren postposizioa, beste osagaien posposizioen zerrenda, eta konbinazio horri dagokion aditzaren iragankortasuna. Aditz bakoitzaren patroia desberdinak maiztasunaren arabera ordenatuta agertzen dira.

Patroietan itzulpenean eragin handia izan dezakeen errore bat detektatu dugu: batzuetan iragangaitzak diren aditzak iragankor gisa agertzen dira eta iragankor direnak iragangaitz gisa. Corpusetik erauzitako patroiak euskarazko aditzen iragankortasunaren informazioarekin zuzendu ditugu, informazio hori *Euskal Hiztegia* (Sarasola, 1996) hiztegi elebakarra prozesatuz jaso dela (Arriola, 2000).

V.9 taulan eta V.25 irudian ikus ditzakegu *eskatu* aditzaren maiztasun handieneko patroiak. Patroi bakoitzean lau eremu agertzen dira:

- *ord*: patroia maiztasunaren arabera ordena.
- *tr*: iragankortasuna (*transitivity*). Lau balio izan ditzake: *DA* eta *ZAIO* aditz iragangaitzetan, eta *DU* eta *DIO* aditz iragankorretan.
- *sub*: subjektuaren postposizioa. Ergatiboa (*ERG*) izango da aditz iragankorrekin, eta absolutiboa *ABS* iragangaitzekin.
- *post*: beste postposizioen zerrenda, “-” ikurraz banatuta.

Posposizioak desanbiguatzeko subjektua alde batera utzi eta aditzaren menpe dauden beste chunketako postposizio posibleen konbinazioak aztertzen ditugu, konbinazio horietako batekin bat datorren maiztasun handieneko azpikategorizazio-patroia bilatzeko. Guztiz bat datorren patroirik ez badago, elementu komun gehien dituen patroia jasoko dugu. Chunk bakoitzari patroia horretan dagokion postposizioa esleituko zaio, subjektuari patroia horretan subjektuarentzat agertzen dena, eta aditzari patroiak ematen duen iragankortasuna. Patroi zehatza ez bada aurkitzen, zenbait chunken postposizioak desanbiguatu gabe geldi daitezke. Postposizio horiek hurrengo urratsetan ebatziko dira.

¹⁴Corpus horretan agertzen diren 1.412 aditz-lema desberdinetatik, agerpen-maiztasun aldetik esanguratsuak diren 640 aukeratu dituzte. Corpusaren azaleko analisia burutu ondoren, eraikitako gramatika baten bidez postposizio bakoitza bere aditzarekin lotu dute (%87ko doitasuna eta %66ko estaldura). Horren ostean aditz bakoitzarekin agertzen diren postposizio-konbinazio posibleak jaso dituzte, konbinazio bakoitza bere maiztasunarekin. Patroi horietan agertzen diren adjuntuak, hau da aditzak azpikategorizatzen ez dituen elementuak, metodo estatistikoak erabiliz baztertu dituzte. Patroietan subjektuaren postposizio-informazioa gehitzen da, nahiz eta corpusean subjektua eliptikoa izan.

```

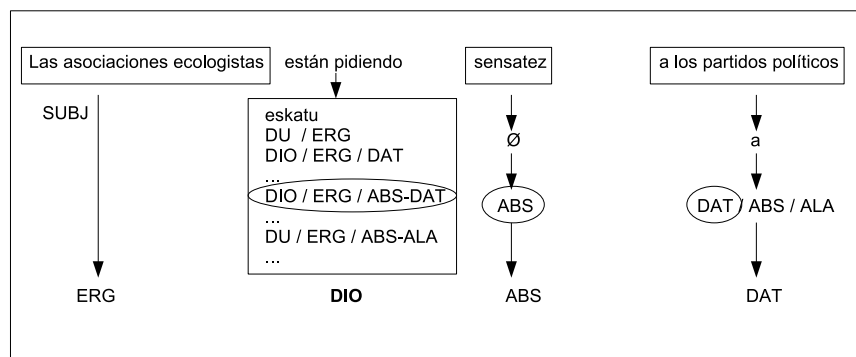
<verb>eskatu
  <pattern ord=1> <tr>DU</tr>
                  <sub>ERG</sub>
                  <post></post>   </pattern>
  <pattern ord=2> <tr>DIO</tr>
                  <sub>ERG</sub>
                  <post>DAT</post> </pattern>
  <pattern ord=3> <tr>DIO</tr>
                  <sub>ERG</sub>
                  <post></post>   </pattern>
  <pattern ord=4> <tr>DU </tr>
                  <sub>ERG</sub>
                  <post>ABS</post> </pattern>
  <pattern ord=5> <tr>DU </tr>
                  <sub>ERG</sub>
                  <post>INE</post> </pattern>
  <pattern ord=6> <tr>DIO</tr>
                  <sub>ERG</sub>
                  <post>ABS</post> </pattern>
  <pattern ord=7> <tr>DIO</tr>
                  <sub>ERG</sub>
                  <post>ABS-DAT</post> </pattern>
  ...
  <pattern ord=32> <tr>DU</tr>
                  <sub>ERG</sub>
                  <post>ABS-INE</post> </pattern>
  ...
  <pattern ord=79> <tr>DU</tr>
                  <sub>ERG</sub>
                  <post>ABS-ALA</post> </pattern>
  ...
</verb>

```

Irudia V.25: Azpikategorizazio-patroien formatua

eskatu			
Ordena	Iragankortasuna	Subjektua	Postposizioak
1	DU	ERG	DAT
2	DIO	ERG	
3	DIO	ERG	
4	DU	ERG	ABS
5	DU	ERG	INE
6	DIO	ERG	ABS
7	DIO	ERG	ABS-DAT
...			
32	DU	ERG	ABS-INE
...			
79	DU	ERG	ABS-ALA
...			

Taula V.9: Azpikategorizazio-patroiak



Irudia V.26: Azpikategorizazio-patroien erabilera: 1. adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos esaldia itzultzeko subjektuaren postposizioaz gain beste bi chunketako postposizioak desanbiguatu behar dira (ikus V.26 irudia): preposizio hutsa duen bat eta *a* preposizioa duen bestea. Preposizio hutsa duen chunkarentzat aurreko urratsetik *ERG* eta *ABS* postposizioak eman dira, eta *a* preposizioa duenarentzat *DAT*, *ABS* eta *ALA*. Postposizio horien konbinazio posibleentzat *eskatu* aditzaren patroiekin bat datozenak bilatzen dira (*ABS-DAT* eta *ABS/ALA*) bilatzen dira, maiztasun handienekoa (*ABS-DAT*) hautatuz. Beraz, *DIO/ERG/ABS-DAT* patroia jasotzen da, subjektuari ergati-boa (*ERG*), aditzari *DIO* iragankortasun-informazioa eta aditzaren beste osagaiei, preposizio hutsa zuenari absolutiboa (*ABS*) eta *a* preposizioa zuenari datiboa (*DAT*).

Hiztegiko ordena

Aurreko urratsetan desanbiguatu ez diren postposizio posibleen multzo bakoitzetik, sistemako hiztegiko ordenan aurrena dagoena hautatzen da. Hiztegiaren ordena hizkuntzalari batek ebatzi du, postposizioak ordenatuz intuitiboki testuen itzulpenetarako egokitasunaren arabera.

Preposizioak eta funtzio sintaktikoak itzultzeko beste estrategia batzuk

Preposizioak eta funtzio sintaktikoen transferentziarako beste estrategia batzuk ere aztertu ditugu, estrategia horien arteko konbinazio posibleak probatuz.

Matxin sistemarako eskuz kodetutako preposizioen hiztegia eta hautapen-erregelak erabili beharrean, ondokoak hiztegiak erabil ditzakegu:

- Preposizioen hiztegia hautapen-erregularik gabe.
- Corpusetik erauzitako preposizioen baliokidetzak erabili hiztegi bezala. Metodo estatistikoekin corpus handi batetik jaso dugu preposizio bakoitzaren probabilitate handieneko postposizioa.

Gainera, postposizioak desanbiguatzeko beste aukera bat aztertu dugu: aditz-izen-postposizio hirukoteak erabiltzea. Hirukote horiek corpusean aditz bakoitzarekin izen bakoitza zein posposizioarekin agertzen den adierazten dute, eta erauzi dira *Euskaldunon Egunkaria* corpus periodistikotik, 14 miloi hitz analizatuz. Hirukote hauekin subjektua ez diren chunken postposizioak desanbigua daitezke. Horretarako, aditzaren lema, chunkeko buruaren lema eta chunkeko preposiziorako preposizioen hiztegian emandako postposizio posible bakoitzarekin osatutako hirukoteak corpusetako hirukoteen artean bilatzen ditugu, maiztasun handienekoa jaso, eta hirukote horretako postposizioa hautatuz.

Estrategia guzti hauek banaka eta haien arteko konbinazio posibleak probatu ditugu eta badirudi preposizioen hiztegiak eta hautapen-erregelak azpikategorizazio-patroiekin batera emaitza onenak lortzen dituztela.

Estrategia desberdinen eta haien arteko konbinazioen ebaluazio kuantitatiboa diseinatu dugu eta ebaluatzeko prozesuan aurkitzen gara une honetan, ebaluazioaren emaitzek hartutako erabakiak konfirmatuko edo zalantzan jarriko dituztela.

V.2.4 Aditz-kateen transferentziaren deskribapena

Espainierak eta euskarak oso aditz-sistema ezberdina dute eta haien arteko itzulpena prozesu konplexua da aditz-kateen egitura oso bestelakoa delako bi hizkuntzetan.

Egoera finituetako teknologia egokia izan daiteke prozesu hori bideratzeko, homogeneousunari zein eraginkortasunari begira, eta horregatik soluzio bat diseinatu dugu egoera finituetako transduktoreak (*Finite State Transducer*, *FST*) erabiliz (Alegria *et al.*, 2005a). *XFST*¹⁵ tresnarekin inplementatu dugu, bere malgutasuna eta ahalmena frogatuta daudelako eta ondo dokumentatutako tresna delako (Beesley eta Karttunen, 2003). Azkartasunari begira emaitza ezin hobeak lortu ditugu (250 aditz-kate/segundo), baina *XFST* tresna ez da kode irekikoa.

Beraz, itzulpen-sistema osoa kode irekian zabaldu nahi dugunez, *FSA*¹⁶ software libreko tresna erabiltzea erabaki genuen. Baina *XFST* rako idatzitako gramatika *FSA* ren syntaxira egokitzeko arazo asko gaudituz ondoren, eraginkortasun arrazoiengatik bide hau baztertu behar izan genuen.

Honen aurrean, eta egoera finituetako transduktoreak kudeatzen dituen kode irekiko tresna eraginkor bat sortu arte, ondoko erabakia hartu dugu: aditzen transferentziarako gramatika *XFST* en syntaxi estandarrean oinarritzea, gramatika honetako erregelak adierazpen erregular bihurtzen dituen modulu bat inplementatzea, eta aditz-kateen transferentzia-moduluak adierazpen erregular horiek irakurri eta aplikatzea. Inplementatu dugun modulu honek aditz-kateak abiadura onargarri batekin itzultzen ditu: 50 aditz-kate/segundo.

Aditz-kateen transferentziarako sortu dugun *FST* gramatikaren erregelek jasotako sarreraren gainean zenbait transformazio burutzen dituzte, beharrezkoa den irteera emanez. Erregelen formatua III.6.5 atalean azaldu dugu.

Gramatika aplikatzean burutzen den prozesua hobeto azaltzeko adibide bat erabiliko dugu: *Te habré tenido que regalar aquellos libros*.

Sarrera

Gramatikarako sarrera izango den karaktere-katea ondoko informazioa dauka:

- Espainierazko aditz-katearen nodo guztien informazio morfologikoa.
tú[PP2CS000] haber[VAIF1S0] tener[VMP00SM] que[CS] opari_egin[VMN0000]

¹⁵ *Xerox Finite-State Tool*. <http://www.cis.upenn.edu/~cis639/docs/xfst.html>

¹⁶ *Finite State Automata*. <http://www.let.rug.nl/~vannoord/Fsa>

- Subjektuaren, objektu zuzenaren eta zeharkako objektuaren informazio morfologikoa: [00] [HK] [00].
[00] etiketek adierazten dute ez dagoela subjekturik eta zeharkako objekturik.
- euskarazko aditz nagusiaren iragankortasuna: [DU].
[DU] etiketak adierazten du aditza iragankorra dela.
- Espainierazko katearen aditz nagusiari dagokion euskarazko forma: `opari_egin`.

Kate horren hasiera ("#") eta bukaera ("\$") ikur bereziekin markatzen dira, gramatikako erregelak aplikatu aurretik. Hau da adibiderako sarrera:

```
# tú[PP2CS000] + haber[VAIF1S0] + tener[VMP00SM] + que[CS] + opari_egin[VMN0000]
// [00] [HK] [00] // [DU] // opari_egin $
```

Erregelak

Gramatikako erregela guztiek *baldintzapeko ordezkapen eragilea* (*conditional replacement operator*) erabiltzen dute, ondoko formatua duena:

```
[ Patroia -> Ordezkapena || EzkerrekoTestuingurua _ EskuinekoTestuingurua ]
```

Patroia katearen lekuan *Ordezkapena* jarriko da, baldin eta *EzkerrekoTestuingurua* eta *EskuinekoTestuingurua* betetzen badira. Adierazpen erregularrak erabil daitezke testuinguruak adierazteko.

Gramatikako erregelak hiru multzotan antolatu ditugu, aditz-kateak itzultzeko definitu diren hiru urrats hauekin lotuta daudenak:

- mota identifikatzeko eta eskema gehitzeko erregelak
- atributuak ordezkatzeko erregelak
- garbiketa egiteko erregelak

Mota identifikatzeko eta eskema gehitzeko erregelak

Erregela hauek espainierazko aditz-katearen mota identifikatzen dute eta mota horri dagokion euskarazko aditz-katearen eskema gehitzen dute muga ikur berezi baten (`==>`) ondoan:

```
[ $ -> Muga EuAditzKateEskema $ || # EsAditzKateMota _ ]
```

Badira mota honetako sei erregela, bakoitza espainierazko aditz-kate mota bati dagokiona: aditz ez-jokatuak (EJ), aditz arruntak ez-perifrastikoak (EP), eta definitu diren perifrastikoen lau multzoak: P1, P2, P3, P4 (ikus V.2.1 atala).

Euskararako bost eskema desberdin daude aditz-kateetan. Espainieraz baino multzo bat gutxiago dago, aditz arruntek eta aditz arruntekin itzultzen diren perifrastikoen P4 multzokoek euskaraz egitura bera dutelako. Hauek dira euskararako eskemak:

```
EJ)    AspN
EP/P4) AspN          / AdM AspM / AdL DenMod Nr Ni Nk
P1)    AspN / Per AspP / AdM AspM / AdL DenMod Nr Ni Nk
P2)    AspN / Prt      / AdM AspM / AdL DenMod Nr Ni Nk
P3)    AspN / Adb       / AdM AspM / AdL DenMod Nr Ni Nk
```

Eskema bakoitzean aditz-katearen motaren identifikadorea (EJ, EP/P4, P1, P2 eta P3) eta dagozkien balioekin ordezkatu beharko diren atributuen izenak¹⁷ agertzen dira.

Adibidearekin jarraituz, lehenengo erregela batek identifikatzen du espainierazko aditz-katea P1 perifrastikoen multzokoa dela, eta mota horri dagokion eskema gehitzen dio:

```
# tú[PP2CS000] + haber[VAIF1S0] + tener[VMP00SM] + que[CS] + opari_egin[VMN0000]
// [00][HK][00] // [DU] // opari_egin $
==> P1)
<NAG> AspN / Per <PER> AspP / AdM AspM / AdL DenMod Nr Ni Nk $
```

Atributuak ordezkatzeko erregelak

Erregela hauek euskarazko eskemako atributuak dagozkien balioekin ordezkatzeko dituzte, espainierazko sarreran dauden edo euskarazko eskeman jada ordezkatu diren zenbait balioen arabera. Erregela batzuek atributuak balio hutsekin ordezkatzeko dituzte, hau da, atributuak ezabatzen dituzte.

Atributuak ordezkatzeko erregelak eredu honi jarraitzen diote:

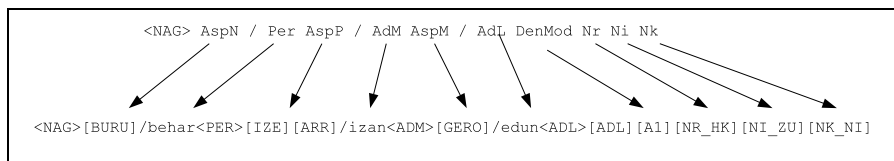
```
[ EusAtributua -> EusBalioa ||  ?* Balio1 ?* Balio2 ?* _ ?* Balio3 ?* ]
```

Ordezkapen hori egiteko ezkerreko eta eskuineko testuinguruetan agertu beharreko balioak adierazpen erregularrekin kodetzen dira. Edozein karakteretako sekuentzia (0,1 edo n) adierazteko *?** erabiltzen da.

¹⁷AspN: aditz nagusiaren aspektua; Per, AspP: aditz perifrastikoa eta bere aspektua; Prt: partikula; Adb: adberbio; AdM, AspM: aditz modala eta bere aspektua AdL, DenMod, Nr, Ni, Nk: aditz laguntzailea, bere denbora eta moduaren morfema eta nor, nori eta nork komuntadura informazioa.

Badira 125 erregela: 16 ordezkapen-erregela aditz nagusiaren aspektu-
rako (*AspN*), 27 erregela aditz perifrastiko (*Per*) eta bere atributuetara-
ko (*AspP*, *Prt*, *Adb*), 7 erregela aditz modal (*AdM*) eta bere aspekturako
(*AspM*), 12 erregela aditz laguntzaile (*AdL*) eta bere denbora eta modua
(*DenMod*) atributuetarako, 30 erregela nor (*Nr*), nori (*Ni*) eta nork (*Nk*)
atributuetarako, 23 erregela izenordain atonoen tratamendurako, eta beste
moldaketak burutzeko beste 13 erregela.

Adibidean zenbait erregelek euskarazko eskemaren atributuak bere baliokin ordeztzen dituzte.



Aurreko irudian zenbait balio laburtuta jarri ditugu hobeto ikusteko. Ordezkapen horien ondoren katea horrela gelditzen da.

```
# tú[PP2CS000] + haber[VAIF1S0] + tener[VMP00SM] + que[CS] + opari_egin[VMN0000]
// [00][HK][00] // [DU] // opari_egin $
==> P1)
<NAG>[ADI][SIN]+[AMM][PART]+[ASP][BURU] /
      behar<PER>[IZE][ARR] /
izan<ADM>[ADI][SIN]+[AMM][PART]+[ASP][GERO] /
edun<ADL>[ADL][A1][NR_HK][NI_ZU][NK_NI]
```

Garbiketa egiteko erregelak

Azkenik lau erregelekin ezabatzen da soberazko informazioa (hasiera, bukara eta muga ikurrak, aditz-kate motaren identifikadorea eta espainierazko informazioa).

Irteera

Gramatika aplikatu ondoren lortuko den irteera, sarrerari dagokion euskarazko aditz-katearen nodoen zerrenda izango da, beharrezkoa den informazioarekin:

- nodo horien arteko ordena ebazteko, eta
- nodo bakoitzaren sorkuntza morfologikoa burutzeko

Ondokoa da adibiderako sistemaren hurrengo modulura pasako den irteera:

```

                                [DU] &
opari_egin<NAG>[ADI] [SIN]+[AMM] [PART]+[ASP] [BURU] /
                                behar<PER>[IZE] [ARR] /
izan<ADM>[ADI] [SIN]+[AMM] [PART]+[ASP] [GERO] /
                                edun<ADL>[ADL] [A1] [NR_HK] [NI_ZU] [NK_NI]

```

Irteeran “/” ikurrarekin banatutako elementu bakoitzaren informazioa- rekin nodo bat eraiki beharko da, irteeran ematen den lehenengoa erroa izango dela, eta besteak bere azpiko nodoak.

“<” eta “>” ikurren arteko informazioa (<NAG>, <PER>, <ADM> eta <ADL>) nodoen *loc* atributuan gordeko da ordena erabakitzeke. Beste informazioa *post* atributuan gordeko da sorkuntza morfologikoa burutzeke.

Adibide honetarako sistemak emango duen itzulpena hauxe izango da: *Liburu haiek opari egin behar izango dizkizut.*

V.3 Laburpena

Kapitulu honetan transferentzia fasearen bi azpifaseak (lexikala eta estrukturala) burutzeke aurkitu ditugun zailtasunak eta zailtasun horiei aurre egiteko hartutako erabakiak azaldu ditugu.

Transferentzia lexikalaren moduluaren zehaztasunak aztertu ditugu eta moduluak erabiltzen duen lexikoi elebiduna formatua eta erabilpena deskribatu ditugu.

Transferentzia estrukturala burutzeke espainiera eta euskara hizkuntzen arteko dibergentziak deskribatu ondoren, transferentzia estrukturalako lau moduluen detaileak sakonki aztertu ditugu: intrachunk transferentzia estrukturala preposizio eta funtzio sintaktikoen transferentzia, interchunk transferentzia estrukturala eta aditz-kateen transferentzia.

Abiapuntu-hizkuntza aldatu nahiko bagenu, adibidez ingelesatik euskara, hizkuntza-bikote horretarako transferentzia-faseko datu linguistiko berriak (lexikoi elebiduna, preposizioen hiztegia, azpikategorizazio-patroiak, transferentziako erregelak eta aditz-kateen transferentziarako gramatika) sortu beharko genituzke, programak aldatu gabe.

VI. KAPITULUA

Sorkuntza

*Eta horrela bizitzen gera
sortuz ta sortuz gure aukera.*

Xabier Lete

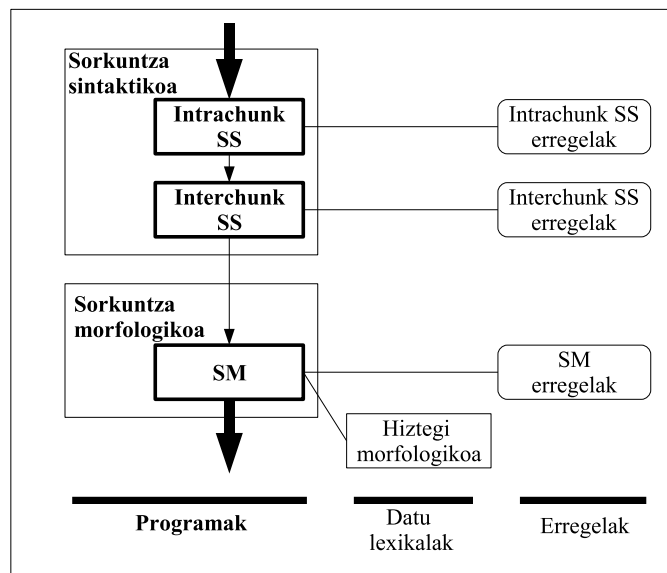
Kapitulu honetan gure sistemako azken fasea deskribatuko dugu: sorkuntza. Transferentzia-fasean abiapuntu-hizkuntzako testuaren egiturari dago-kion xede-hizkuntzako egitura lortu ondoren, sorkuntzaren helburua egitura horretatik xede-hizkuntzako testua ematea da. Sorkuntza bi mailatan burutzen da: sintaktikoa eta morfologikoa.

Sorkuntza sintaktikoak hitzak eta sintagmak xede-hizkuntzako sekuentzia zuzen batean ordenatzen ditu. Euskarazko hitz-hurrenkera deskribatu ondoren, sorkuntza sintaktikoa burutzen duten bi moduluen zehaztasunak emango ditugu. Modulu batek chunk barruko hitzak ordenatzen ditu eta besteak esaldiko chunkak.

Sorkuntza morfologikoak elementu lexikal etiketatuak interpretatzen ditu hitzen formak emateko. *Matxin* sistemak berrerabiltzen duen prozesadore morfologikoa aurkeztuko dugu, kode irekikoa izan dadin egin dugun ego-kitzapena deskribatuz. Gainera, moduluak prozesadore morfologikoa nola erabiltzen duen ere azalduko dugu.

Sorkuntza fasea abiapuntu-hizkuntzatik independentea izateko eraiki dugu. Beraz, modulu honek ezer aldatu gabe prozesatu ahal izango luke beste hizkuntza baterako analisia eta transferentzia moduluen irteera, datu-fluxuen formatua jarraitzen badu.

Gure sistemako sorkuntza-fasearen moduluen arkitektura VI.1 irudian ikus dezakegu.



Irudia VI.1: Sorkuntza: arkitektura

VI.1 Sorkuntza sintaktikoa

Sorkuntza sintaktikoak (SS) hitzak eta osagaiak xede-hizkuntzako sekuentzia zuzen batean ordenatzen ditu.

Ordenazioa burutzeko transferentziatik datorren egitura sakoneko zuhaitza erabiltzen da, egitura hori oso egokia izanik ordenazioa bi mailatan egiteko: chunk bakoitzeko nodoak ordenatzen dira alde batetik (SS intrachunk), eta esaldiko chunkak bestetik (SS interchunk).

Sorkuntza sintaktikoaren irteera mendekotasun-zuhaitz bera izango da, non nodo eta chunkei behar duten ordena esleitu zaien. Gainera, postposizio-informazioa duten chunketan informazio hori chunkaren azkeneko nodoan kopia-tuko da, hurrengo azpifasean nodo horren sorkuntza morfologikoa burutzeko.

Hizkuntza hurbilen artean itzulpena egiten duten IAko zenbait sistemetan abiapuntu-hizkuntzaren esaldian berrordenazio lokal batzuk eginda, lortzen da xede-hizkuntzako ordena egokia. Inguruko hizkuntzetatik (espainieratik edo latinetik datozen beste hizkuntzetatik, zein ingelesetik) euskarara itzultzeko orduan, ordena guztiz desberdina da, batzuetan alderantzizkoa izateraino.

Horregatik, gure sistemako sorkuntza sintaktikoa, hau da, hitz eta osa-

gaien ordenazioa, abiapuntu-hizkuntzatik guztiz independentea burutu behar dugu, mendekotasun-zuhaitz egituran dagoen informazio elebakarra erabiliz.

Ordenazio posible bat baino gehiago dagoenean, egokia den bat ematea rekin nahikoa da. Osagaiak ordenatzeko erregelek ez dituzte aukera posible guztiak emango, baizik eta bat bakarrik, orokorrean egokiena dena.

Hurrengo ataletan, euskarazko hitz-hurrenkerari buruzko zenbait azalpen eman ondoren, sorkuntza sintaktikorako bi urratsak aurkeztuko ditugu: chunk barrukoa eta chunken artekoa.

VI.1.1 Hitz-hurrenkera euskaran

Testu egokiak lortzeko, osagaien antolaketan lagungarriak izan daitezkeen zenbait ideia eskaintzen dira (Zabala, 2000) lanean¹, atal honetan laburbilduko ditugunak. Gure proposamena lan horretan oinarritzen da.

Hizkuntzen ezaugarrietariko bat bere osagaien arteko hurrenkeraren askatasun-maila da. Adibidez, ingelesak oso ordena zurruna du perpausen osagaiak antolatzeko. Euskarak, nahiz eta osagaien barruko antolakuntza zurruna izan, perpausen osagaiak antolatzeko askatasun handia du. Adibidez lau osagai dituen ondoko perpausak 24 hurrenkera onartzen ditu: *aitak / haurrari / sagarra / eman dio* (Euskaltzaindia, 1987).

Hitzen arteko ordena osagaien barruan

Euskaraz sintagmaren barruko ordenaketa oso zurruna da. Izen-sintagmaren barruan informazio nagusia izenaren inguruan kokatzen da. Hurbilen adjektiboak egongo dira eta gero determinanteak. Adjektibo eta determinanteak, motaren arabera, izenaren aurrean edo atzean koka daitezke. Adjektiboa adberbio batek modifikatzen badu, adberbioa adjektiboaren aurrean kokatuko da, eta beti izenaren aurrean, nahiz eta adjektiboa izen ondoan agertu.

Aditz-katearen barruko ordena ere finkoa da, baiezkotako eta ezezko esaldietarako desberdina.

Osagaien arteko ordena

Euskarazko perpausaren ordena oinarritzakoa edo neutroa osagaien burua bukaeran duena da, hau da, SOV² edo Subjektua-Objektua-Aditza (latinetik datozen hizkuntzak edo ingelesa, berriz, SVO motakoak dira).

¹Gai honi buruzko hausnarketa sakona eta azterketa bibliografiko zabal bat (Erdozia, 2006) lanean aurki daitezke. Bestalde, (Aldezabal *et al.*, 2003) lanean hitz-hurrenkeraren azterketa masiboa corpusean aurkezten dute.

²Subject-Object-Verb

Aditzaren ezkerrean doan osagaia *galdegaia* da, esaldian azpimarratzen dena (adibidez, *Hori Peruk esan du*). *Mintzagaia*, berriz, informazio nagusia zeri buruzkoa izango den, testuingurua, finkatzen duen osagaia da, eta galdegaiaren ezkerrean jartzen da, perpausaren hasieran (*Hori Peruk esan du*). Aditzaren atzean jartzen dugun informazioa bigarren plano batean geratzen da (*Hori Peruk esan du gaur*).

Hala ere, ohikoa da osagai luzeak aditzaren ondoren kokatzea (Hidalgo, 1996). Baina, osagaien luzera baino, haien izaera da erabakigarriagoa: izen-sintagmak hanpatzeko modurik egokiena aditzaren ezkerrean kokatzea da, eta perpaus osagarriak (konpletiboak) eta koordinaturiko izen-sintagmak aditzaren eskuinean kokatzea egokiagoa edo ezinbestekoa da (*Uste dut garbi dagoela eskualde honetako organismoak azkarrago ugaltzen direla*).

Erabateko adostasuna dago aditza aurreratuz oro har testu arinagoak lortzen direla. Edonola, badira aditza bukaera uztera bultzatzen gaituzten faktoreak, adibidez, mendeko perpaus askotan: izenaren ezkerrean kokatzen diren erlatibozko perpausetan, aditz nagusiaren ezkerrean kokatzen diren mendeko perpaus adberbialetan...

VI.1.2 Chunk barruko sorkuntza sintaktikoa

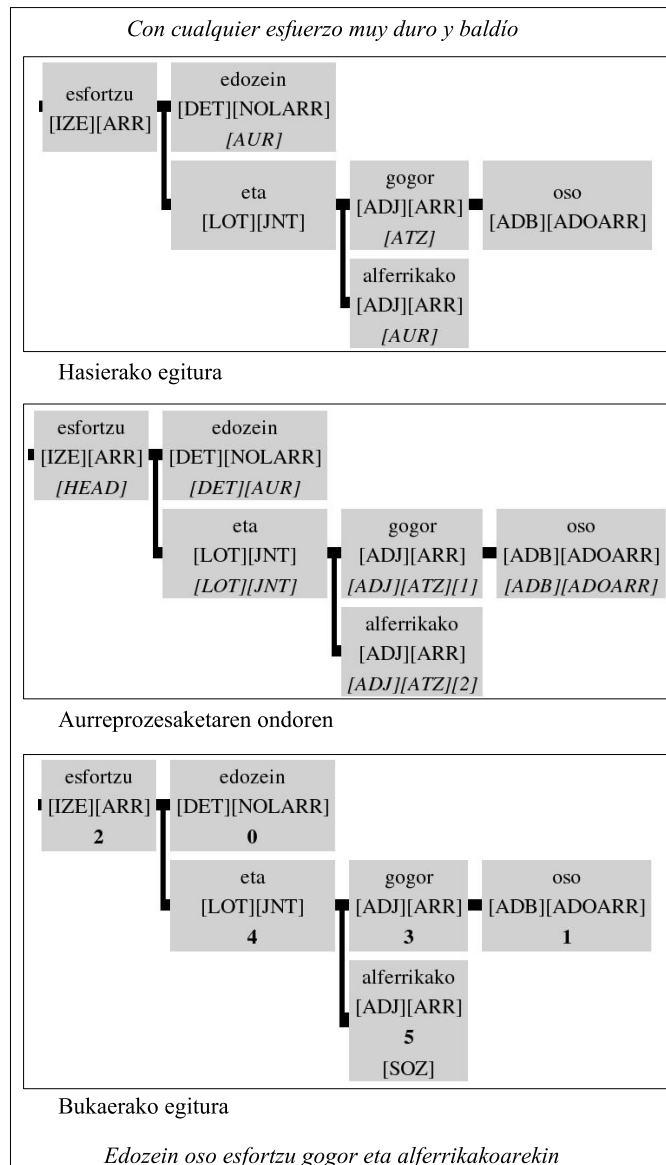
Chunk barruko sorkuntza sintaktikoa burutzen duen moduluak nodoak ordenatzen ditu chunk barruan eta postposizio-informazioa duten chunketan (orokorrean aditz-kateak ez direnak) postposizio-informazioa chunkaren azkeneko nodoari pasatzen dio.

Euskaraz sintagma bakoitzaren elementuak modu jakin eta zurrun batean ordenatzen direnez, chunk mota bakoitzerako ordena hori aurrekotasun-erregelakin kodetu dugu.

Hurrenkera ebazten duten aurrekotasun-erregelak aplikatu ahal izateko aurreprozesaketa bat egin behar da, nodoetan dagoen informazioa egokitze-ko: koordinatutako adjektiboak prozesatu eta kokapen-atributuaren (*loc*) osatu.

Ordenaketan kokapen-atributua *loc* erabiltzen da sarrera gisa eta ordena *ord* lortzen da.

Ondoko ataletan modulu honetan burutzen diren atazen deskribapena hobeto ulertzeko, *Con cualquier esfuerzo muy duro y baldío* espainierazko testuaren itzulpen-prozesuan burututako prozesaketa erabiliko dugu, VI.2 irudian ikus dezakeguna. Irudi horretan ikus ditzakegu transferentzia-fasetik datorren datu-egitura, aurreprozesaketa burutu ondorengo eta bukaerakoa. Nodo bakoitzean lema eta POS informazioaz gain, *loc* kokapen-atributuaren informazioa agertzen da letra etzanaz. Bukaerako egituraren ere nodoaren



Irudia VI.2: Sorkuntza sintaktikoa intrachunk. Adibidea.

```

<!-- Bi adjektiboen kokapen-atributuak desberdina badira, biak ATZ>
<rule id='N/NN_lot_aur_atz'>
  <match>
    <def> N1 := //NODE[@pos='[LOT][JNT]'] </def>
    <def> N2 := N1/NODE[@pos='[ADJ][ARR]'][@loc='[AUR]'] </def>
    <def> N3 := N1/NODE[@pos='[ADJ][ARR]'][@loc='[ATZ]'] </def></match>
  <actions>
    <act> N2/@loc := '[ATZ]' </act></actions>
</rule><!------->
<!-- Adjektibo bakoitzari zenbaki bat jarri>
<rule id='N/NN_loc_1_2'>
  <match>
    <def> N1 := //NODE[@pos='[LOT][JNT]'] </def>
    <def> N2 := N1/NODE[@pos='[ADJ][ARR]'] [1] </def>
    <def> N3 := N1/NODE[@pos='[ADJ][ARR]'] [2] </def></match>
  <actions>
    <act> N1/@loc := concat(N1/@loc, '[1]') </act>
    <act> N2/@loc := concat(N2/@loc, '[2]') </act></actions>
</rule><!------->

```

Irudia VI.3: Adjektiboen koordinazioaren prozesaketarako erregelak

ordena duen *ord* atributuaren balioa eta *post* postposizio-informazioarena agertzen dira.

Koordinatutako adjektiboen prozesaketa

Adjektiboen koordinazioa (*eta*, *edo*) ematen denean, adjektiboen kokapen-atributuaren prozesaketa bat egin behar da. Koordinatutako bi adjektiboen kokapen-atributuak (*loc*) desberdinak badira, izenaren aurrekoa (*AUR*) bat eta atzekoa (*ATZ*) bestea, bi adjektiboak atzean jarriko dira. VI.2 irudiko adibidean *alferrikako* adjektiboaren *loc* atributuari *[ATZ]* esleitu zaio.

Gainera koordinatutako bi adjektiboak bereiztu behar dira, bat lokailuaren aurrean eta bestea atzean joango direlako. Horretarako koordinazio-ko lehenengo adjektiboaren *loc* atributuan *[1]* zenbakia gehituko zaio, eta bigarrenari *[2]*. Zenbaki horrekin, ere, aurrekotasun-erregelak aplikatzean koordinatuta agertzen diren adjektiboak eta koordinatu gabe agertzen direnak bereiztu nahi dira. VI.2 irudian *gogor* adjektiboari *[1]* eta *alferrikako* adjektiboari *[2]* esleitzen zaie.

Prozesaketa hau VI.3 irudiko erregelek gidatzen dute..

Kokapen-atributuaren osaketa

```

<!-- Determinanteetan kategoria ([DET]) eta kokapen-atributua>
<rule id='N_loc_det'>
  <match>
    <def> N := //NODE[@pos=~'[DET]'] </def></match>
    <actions>
      <act> N/@loc := concat('[DET]',N/@loc) </act></actions>
    </rule><!------->
<!-- Adjektiboetan kategoria ([ADJ]) eta kokapen-atributua>
<rule id='N_loc_adj'>
  <match>
    <def> N := //NODE[@pos=~'[ADJ]'] </def></match>
    <actions>
      <act> N/@loc := concat('[ADJ]',N/@loc) </act></actions>
    </rule><!------->
<!-- Chunkaren burua den nodoan [HEAD] balioa>
<rule id='N_loc_head'>
  <match>
    <def> N := //CHUNK/NODE </def></match>
    <actions>
      <act> N/@loc := '[HEAD]' </act></actions>
    </rule><!------->
<!-- Beste nodo guztietan POS atributuak duen balioa>
<rule id='N_loc_pos'>
  <match>
    <def> N := //NODE[not(@loc)] </def></match>
    <actions>
      <act> N/@loc := N/@pos </act></actions>
    </rule><!------->

```

Irudia VI.4: Kokapen-atributuaren (*loc*) osaketarako erregelak

Ondoren, VI.4 irudian agertzen diren erregelak aplikatzen dira *loc* kokapen-atributua osatzeko. Determinante eta adjektiboetan *loc* atributuan bere kategoria gehitzen da (*[DET]* edo *[ADJ]*). Chunkaren burua den nodoan *loc* atributuak (*[HEAD]*) balioa hartzen du. Eta azkenik *loc* atributuan baliorik ez duten nodoetan, kategoria eta azpikategoria (*pos* atributua) kopiatzen da *loc* atributuan.

VI.2 irudian ikus dezakegu *loc* atributuan egindako aldaketak: *esfortzu* izenean *[HEAD]* esleitu zaio, *edozein* determinantean *[DET]* eta *gogor* eta *alferrikako* adjektiboetan *[ADJ]* gehitu, eta *oso* adberbioari kategoria eta azpikategoria (*pos*) kopiatu.

Aurrekotasun-erregelak

Aurreprozesaketa burutu ostean nodoen ordena ebazten dugu aurrekotasun-erregelak erabiliz. Aurrekotasun-erregela desberdin bat aplikatuko da (ikus VI.5 irudia) chunk mota bakoitzerako:

- baiezeko aditz-kateak
- ezezeko aditz-kateak
- bestelako chunkak: izen-sintagmak eta adjektibo-sintagmak.

Horretarako, *&precedence* kanpo-funtzioari dei egiten zaio. Funtzioaren sarrera chunkaren nodo guztien *loc* atributua eta chunk mota horren aurrekotasun-erregela da, eta irteera nodoen ordena, nodo bakoitzaren *ord* atributuan gordeko dena.

Nodoak ordenatu ondoren, chunkaren postposizio-informazioa chunkaren azken nodoan kopiatzen da VI.6 irudiko erregela aplikatuz, hurrengo fasean nodo horren sorkuntza morfologikoa burutu ahal izateko.

VI.2 irudian ikus dezakegu nola gelditzen den datu-egitura adibide horretarako, non *ord* atributua kalkulatu den eta chunkaren azkeneko hitzari postposizio-informazioa, kasu honetan *[SOZ]*, pasatu zaion.

Adibide honen itzulpen-prozesu osoa bukatzean, lortutako itzulpena ondokoa izango da: *Edozein oso esfortzu gogor eta alferrikakoarekin*.

VI.1.3 Chunken arteko sorkuntza sintaktikoa

Euskarazko perpauseko chunken arteko ordena oso librea bada ere, badira zenbait gomendio chunkak ordenatzeko. Sistemako erregeletan kodetzeko, guk kasu bakoitzerako aukera posible guztien artean orokorrean egokiena deritzoguna aukeratu dugu. Jarraitutako irizpideak hauek dira:

```

<!-- Aditz-kateetarako aurrekotasuna>
<rule id='C//N_prec_adi'>
<rule>
  <match>
    <def> C := //CHUNK[@type='adi-kat'] </def>
    <def> N := C//NODE </def></match>
  <actions>
    <act> {N/@ord} := &precedence( {N/@loc},
      '[ADB]<[HEAD]<[PER]<[PRT]<[ADM]<[ADL]' ) </act>></actions>
</rule><!------->
<!-- Ezezko aditz-kateetarako aurrekotasuna>
<rule id='C//N_prec_adiez'>
<rule>
  <match>
    <def> C := //CHUNK[@type='adi-kat-ez'] </def>
    <def> N := C//NODE </def></match>
  <actions>
    <act> {N/@ord} := &precedence( {N/@loc},
      '[PRT]<[ADL]<[ADB]<[HEAD]<[PER]<[ADM]' ) </act>></actions>
</rule><!------->
<!-- Beste chunk guztietarako aurrekotasuna>
<rule id='C//N_prec_is'>
<rule>
  <match>
    <def> C := //CHUNK[@type='is' or @type='post-izls'
      or @type='post-sint' or @type='adjs'] </def>
    <def> N := C//NODE </def></match>
  <actions>
    <act> {N/@ord} := &precedence( {N/@loc},
      '[DET][AUR]<[ADB][ARR]<[ADJ][AUR]
      <[ADJ][AUR][1]<[LOT][JNT]<[ADJ][AUR][2]
      <[Z]<[IZE][IZB]<[IZE][ARR]<[HEAD]
      <[ADJ][ATZ]<[ADJ][ATZ][1]<[LOT][JNT]
      <[ADJ][ATZ][2]<[DET][ATZ]' ) </act>></actions>
</rule><!------->

```

Irudia VI.5: Aurrekotasun-erregelak

```

<!-- Aditz-kateak ez diren chunketan postposizioa chunketik azken nodora>
<rule id='C/N_post_down'>
<rule>
  <match>
    <def> C := //CHUNK[@type!='adi-kat'] </def>
    <def> N := C//NODE[@ord=count(C//NODE)] </def></match>
  <actions>
    <act> N/@post := C/@post </act></actions>
</rule><!------->

```

Irudia VI.6: Postposizioaren informazioa chunketik azken nodora pasatzeko erregela

- Subjektua aditz-katearen ezkerrean (adibidez, *Nik dakart*).
- Galdegaia, hau da, fokoa duen osagaia, aditz-katearen ondoan, eta:
 - ezkerrean aditz-katea baiezkoa bada (*Nik ogia dakart*)
 - eskuinean aditz-katea ezezkoa bada (*Nik ez dakart ogia*).
- *ez* partikula ezezko aditz-katearen ezkerrean, bere ondoan (*Nik ez dakart*).
- Aditz-kate baten azpiko beste osagaiak aditz-katearen eskuinean (*Nik ogia dakart etxera*).
- Aditz-katea ez den osagai baten umeak beti bere ezkerrean (*Etorri den umea*).

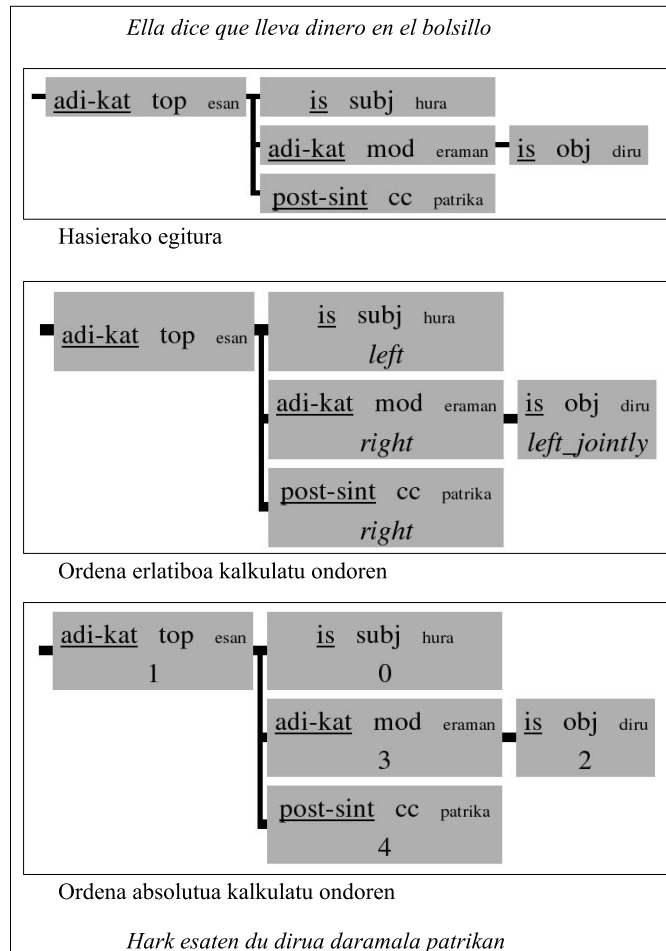
Chunken arteko ordena ebazteko ezin dira aurrekotasun-erregelak erabili, chunk barruko nodoekin bezala, chunken konbinazio posibleak ugari direlako. Horregatik beste estrategia bat erabili behar izan dugu. Ordenaketa bi urratsetan ebazten dugu: aurretik guraso-ume chunk-bikote bakoitzerako ordena erlatiboa ebatzi eta ondoren chunk guztien ordena absolutua.

Prozesaketa honen deskribapena ilustratzeko espainierazko *Ella dice que lleva dinero en el bolsillo* testuaren itzulpen-prozesuaren adibidea erabiltzen dugu. VI.7 irudian hiru egitura erakusten dira: aurreko modulutik datorrena, ordena erlatiboa kalkulatu ondorengoa eta ordena absolutuak kalkulaturakoa. Irudian chunk bakoitzerako bere mota (*type*) eta funtzio sintaktikoaz (*si*) gain, chunkeko buruan dagoen nodoaren lema erakusten da.

Ordena erlatiboaren ebazpena

Chunk bakoitza bere gurasoarekiko duen ordena erlatiboa (*relord*) ebazteko arestian aipatutako irizpideei jarraitzen zaie, VI.8 irudiko erregelak aplikatuz.

Erregeletan gurasoaren mota (*type*) eta chunk horren zenbait atributu hartzen dira kontuan: mota (*type*), informazio sintaktikoa (*si*) eta fokoa (*focus*). Horien arabera lortzen da ordena erlatiboa (*relord*), lau balio desberdin har ditzakeena: ezkerrean (*left*), eskuinean (*right*), ezkerrean elkarrekin (*left-jointly*) eta eskuinean elkarrekin (*right-jointly*). Azken biek chunka gurasoaren ondoan kokatu behar dela adierazten dute. Baldintzak betetzen dituen erregularik ez bada aurkitzen, bukaerako erregela lehenetsia (*left*) aplikatzen da.



Irudia VI.7: Sorkuntza sintaktikoa interchunk. Adibidea.

```

<!-- Subjektua aditzaren ezkerrean>
<rule id='C/C_relord_adi_sub'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']    </def>
    <def> C2 := C1/CHUNK[@si='subj']        </def></match>
  <actions>
    <act> C2/@relord := 'left'                </act></actions>
  </rule><!------->
<!-- Fokoa duen chunka aditzaren ezkerrean elkarrekin>
<rule id='C/C_relord_adi_focus'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']    </def>
    <def> C2 := C1/CHUNK[@focus='true']    </def></match>
  <actions>
    <act> C2/@relord := 'left-jointly'      </act></actions>
  </rule><!------->
<!-- Fokoa duen chunka ezezko aditzaren eskuinean elkarrekin>
<rule id='C/C_relord_adi_adiez_focus'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat-ez'] </def>
    <def> C2 := C1/CHUNK[@focus='true']    </def></match>
  <actions>
    <act> C2/@relord := 'right-jointly'     </act></actions>
  </rule><!------->
<!-- Ez partikula aditzaren ezkerrean elkarrekin>
<rule id='C/C_relord_adiez_ez'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat-ez'] </def>
    <def> C2 := C1/CHUNK[@type='ez']        </def></match>
  <actions>
    <act> C2/@relord := 'left-jointly'     </act></actions>
  </rule><!------->
<!-- Aditzen azpiko beste chunk guztiak aditzaren eskuinean>
<rule id='C/C_relord_adi_bestela'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat']    </def>
    <def> C2 := C1/CHUNK[@focus!='true']
                  [@si!='subj'][@type!='ez'] </def></match>
  <actions>
    <act> C2/@relord := 'right'              </act></actions>
  </rule><!------->
<!-- Aditzak ez diren chunken azpiko chunkak gurasoaren ezkerrean>
<rule id='C/C_relord_bestela'>
  <match>
    <def> C := //CHUNK[not(@relord)]        </def></match>
  <actions>
    <act> C/@relord := 'left'                </act></actions>
  </rule><!------->

```

Irudia VI.8: Ordena erlatiboa ebazteko erregelak

Ordena absolutuaren ebazpena

Zuhaitzeko chunkak aurreordenan korritzen dira, eta guraso-ume chunk-bikoteen ordena erlatiboaren (*relord*) informazioa erabiliz, chunken sekuentzia ordenatua kalkulatzeko da.

Horretarako pila bat erabiltzen da, pilaren elementu bakoitzean guraso den chunk baten indizea eta bere menpekoekin sortzen doan ordenazioa gordetzen dela.

Hortik chunk bakoitzaren *ord* atributuari sekuentzia horretan duen ordena absolutua esleitzen zaio.

VI.7 irudian ikus dezakegu chunk bakoitzak *ord* atributuari esleitutako ordena absolutuaren balioa. Adibide horretarako itzulpen-prozesu osoaren emaitza ondokoa da: *Hark esaten du dirua daramala patrikan*.

VI.2 Sorkuntza morfologikoa

Sorkuntza morfologikoaren (SM) helburua xede-hizkuntzako hitzen formak sortzea da, horretarako elementu lexikal etiketatuak interpretatu behar dituela.

Sorkuntza morfologikoa burutzeko IXA taldean garatutako prozesadore morfologikoa (Alegria *et al.*, 1996) berrerabili dugu. Prozesadore horren oinarria EDBL euskararen datu-base lexikala (Agirre *et al.*, 1995) da³

VI.2.1 Moduluaren deskribapena

Soilik postposizio-informazioa duten hitzak prozesatuko dira sortzaile morfologikoarekin, besteetan hitzaren forma lemaen berdina izango delarik. Aditz-chunketan nodo guztiek sorkuntza egiteko informazioa izango dute eta beste chunketan soilik azken elementuak. Zenbakiekin, datekin eta hitz ezezagunekin sorkuntza morfologiko berezia egiten da.

Sorkuntza morfologikoa burutzeko *lmi* atributua erabiltzen da, lemaen informazio morfologikoa duena. Lexikoi elebidunetik jasotako lema osaketa morfologikorik baldin badu, osaketa horren informazioa, *lmi* atributuan gordeta egongo da. Osaketa morfologikorik ez duten lemek, *lmi* atributua hutsik izango dute eta atributu hori lemaen balioarekin (*lem*) eta kategoria eta azpikategoriarekin (*pos*) osatzen da, VI.9 irudiko erregela aplikatuz

EDBLren diseinu-arrazoiengatik eta arrazoi linguistikoengatik (zenbait determinanteetan), euskararako prozesadore morfologikoak numero eta postposizioaren informazioa modu desberdinetan antolatzen ditu:

³Euskararako kategoria-sistema C eranskinean ikus daiteke.

```

<rule id='N_lmi'>
  <match>
    <def> N := //NODE[@post][not(@lmi)]      </def></match>
    <actions>
      <act> N/@lmi := concat(N/@lem, N/@pos)  </act></actions>
</rule><!------->

```

Irudia VI.9: *lmi* atributuaren osaketarako erregela

- Postposizioa eta numeroa morfema bakarrean.
 $etxearen \rightarrow etxe[IZE][ARR]+[GEN][NUM]$
- Aurretik numeroaren morfema eta ondoren postposizioaren morfema.
 $etxeke \rightarrow etxe[IZE][ARR]+[NUMS]+[GEL]$
- Numeroa beharrean mugagabe marka.
 $honetan \rightarrow hau[DET][ERKARR][NUMS]+[MG]+[INE]$

Beraz nodoetan dagoen informazioa antolatu behar da hitzaren kategoria, azpikategoria eta postposizioaren arabera. Antolaketa hau burutzeko erregelak VI.10 eta VI.11 irudietan ikus daitezke, sortzaile morfologikorako informazio guztia *gen* atributuan gordetzen dela

Ondoren prozesadore morfologikoa erabiltzen da, *gen* atributuan informazioa duten nodo guztietarako &sortzailea funtzioari deituz (ikus VI.12 irudiko erregela).

VI.2.2 Euskararako prozesadore morfologikoa

Atal honetan sorkuntza burutzeko erabiltzen dugun euskararako prozesadore morfologikoa aurkezten dugu, kode librekoa izateko burutu dugun egokitza-penarekin.

Berrerabili dugun IXA taldean garatutako *Morfeus* euskararako prozesadoreak (analizatzailea eta sortzailea dena) bi mailatako morfologiaren eredu konputazionalari⁴ jarraitzen dio eta (Alegria, 1995; Alegria *et al.*, 1996; Alegria eta Urkia, 2002) lanetan sakonki deskribatzen da.

⁴Morfologia konputazionalaren bi mailatako eredua Koskeniemi-k (1983) proposatu zuen. Onarpen zabala izan zuen ondoko ezaugarriengatik: aplikagarritasuna, morfotaktika eta morfofonologia bereiztea, erregelen deklaratiotasuna eta programen eta datu linguistikoen arteko banaketa. Flexio handiko hizkuntza eranskarien kasuan bi mailatako morfologia formalismorik hedatuena bihurtu da (Oflazer, 1994; Sproat, 1992; Antworth, 1990).

```

<!-- Antolaketa: lmi + post num>
<rule id='N_gen_post_num'>
  <match>
    <def> N := //NODE[(@pos=~/[IZE][ARR]|[ADJ]|[DET]/ and
      @pos!~/[ERKARR]|[ERKIND]|[DZH]/ and
      @post=~/[ABS|DAT|DES|ERG|GEN|INS|SOZ/ )
      or (@pos='[DET][ERKARR][NUMS]' and
      @post=~/[DAT|ERG/ )
      or (@pos=~/[DET][ERKARR][NUMP]|[DET][ERKIND]|[IOR]/ and
      @post=~/[ABS|DAT|ERG|INE/ )
    </def></match>
    <actions>
      <act> N/@gen := concat(N/@lmi, '+', N/@post, N/@num) </act></actions>
    </rule><!------->
<!-- Antolaketa: lmi + num + post >
<rule id='N_gen_num_post'>
  <match>
    <def> N := //NODE[(@pos=~/[IZE][ARR]|[ADJ]|[DET]/ and
      @pos!~/[ERKARR]|[ERKIND]|[DZH]/ and
      @post=~/[ABL|ABU|ABZ|ALA|BAN|DESK|GEL|INE|MOT/ )
      or (@pos=~/[IZE][IZB]|[IZE][LIB]/ and
      @post=~/[ABL|ABU|ALA|GEL|INE/ )
      or (@pos='[DET][ERKARR][NUMS]' and
      @post=~/[ABU|ABZ|ALA|BAN|DESK|GEL|INS|MOT/ )
      or (@pos=~/[DET][ERKARR][NUMP]|[DET][ERKIND]|[IOR]/ and
      @post=~/[ABL|ABU|ABZ|ALA|BAN|DESK|GEL|INE|MOT/ )
      or (@pos='[DET][DZH]' and
      @post=~/[ABL|ABU|ALA|GEL|INE/ )
    </def></match>
    <actions>
      <act> N/@gen := concat(N/@lmi, '+', N/@num, '+', N/@post) </act></actions>
    </rule><!------->
<!-- Antolaketa: lmi + post MG>
<rule id='N_gen_post_mg'>
  <match>
    <def> N := //NODE[(@pos=~/[IZE][ARR]|[ADJ]|[DET]/ and
      @pos!~/[ERKARR]|[ERKIND]|[DZH]/ and
      @post=~/[PAR|PRO/ )
      or (@pos=~/[IZE][IZB]|[IZE][LIB]/ and
      @post=~/[ABS|ABZ|BAN|DAT|DES|DESK|
      ERG|GEN|INS|MOT|PAR|PRO|SOZ/ )
      or (@pos='[DET][DZH]' and
      @post=~/[ABS|DAT|ERG|SOZ/ )
    </def></match>
    <actions>
      <act> N/@gen := concat(N/@lmi, '+', N/@post, '[MG]') </act></actions>
    </rule><!------->

```

Irudia VI.10: Sorkuntza morfologikorako antolaketa (I)

```

<!-- Antolaketa: lmi + MG + post >
<rule id='N_gen_mg_post'>
  <match>
    <def> N := //NODE[(@pos='[DET][ERKARR][NUMS]' and
                        @post=~'/ABL|INE/' )
    </def></match>
    <actions>
      <act> N/@gen := concat(N/@lmi, '+'[MG]+' , N/@post)
    </act></actions>
  </rule><!-- Antolaketa: lmi + post >
  <rule id='N_gen_post'>
    <match>
      <def> N := //NODE[(@pos='[DET][ERKARR][NUMS]' and
                          @post=~'/ABS|SOZ|DES|GEN/' )
              or (@pos='[DET][DZH]' and
                  @post=~'/ABZ|BAN|DES|DESK|GEN|INS|MOT|PAR|PRO/' )
      </def></match>
      <actions>
        <act> N/@gen := concat(N/@lmi, '+', N/@post)
      </act></actions>
    </rule><!-- Antolaketa: lmi + num + post >
    <rule id='N_gen_bestela'>
      <match>
        <def> N := //NODE[@post][not(@gen)]
      </def></match>
      <actions>
        <act> N/@gen := concat(N/@lmi, '+', N/@num, '+', N/@post)
      </act></actions>
    </rule><!--

```

Irudia VI.11: Sorkuntza morfologikorako antolaketa (II)

```

<rule id='N_sortzaile'>
  <match>
    <def> N := //NODE[@gen]
  </def></match>
  <actions>
    <act> N/@forma := &sortzailea(N/@gen)
  </act></actions>
</rule><!--

```

Irudia VI.12: Sortzaile morfologikoari egiten zaion deia

Hitzak ezagutzeko nahikoa da azaleko formaren baliozko adierazpide lexikalak aurkitzea. Eta sorkuntzak, alderantziz, adierazpide lexikal bati dagokion azaleko forma bilatzen du.

Bi mailatako morfologiaren osagai garrantzitsuenak ondokoak dira: lexi-koia, jarraitze-klaseak eta erregela fonologikoen multzoa.

Lexikoiaren ezagutza linguistikoa EDBL Euskararen Datu-Base Lexikaletik⁵ jasotzen da. Lexikoiak 60.000 sarrera inguru dauzka, 154 azpilexikoietan multzokatuta. Sarrera bakoitzean morfema bat definitzen da, lema edo hizkia izan daitekeena, ondoko informazioarekin:

- azpilexikoia
- forma kanonikoa eta bi mailatako forma (bigarrenetan diakritikoak erabiltzen dira erregela morfofonologikoen aplikazioa kontrolatzeko)
- jarraitze-klasea
- kategoria eta azpikategoria
- informazio morfologikoa: numeroa, mugatasuna, pertsona, denbora...
- informazio morfosintaktikoa
- gainerako informazioa

Jarraitze-klaseak morfotaktika deskribatzen duten azpilexikoi multzoak dira. Jarraitze-klase bat egokitzen zaio lexikoan morfema bakoitzari, eta bertan biltzen diren osagaiak dira morfema horren ondoren ager daitezkeen morfema guztiak eta bakarrak. Beraz, jarraitze-klaseak hitz batean ager daitezkeen morfemen arteko konbinaketa posibleak definitzeko mekanismoaren oinarri dira.

Erregela fonologikoen morfemak konbinatzean agertzen diren aldaketa morfologiko, fonologiko eta ortografikoak adierazten dituzte. Erregelek hiru eragiketa-mota burutzen dituzte: karaktere bat gehitu, ezabatu eta aldatu

Estaldura eta sendotasuna hobetzeko analizatzaile estandarrari bi modulu gehitu zaizkio: aldaera linguistikoen analizatzailea eta lexikorik gabeko analizatzailea, hitz formak ezagutu ditzakeena haien lema lexikoian ez badago ere.

Erregelak transduktore lexikaletan konpilatzen dira. Transduktore lexikala (Karttunen, 1994) egoera finituetako automata bat da, flexionatutako

⁵Datu-base hau diseinatua izan da formalismo linguistikoekiko neutrala, malgua, irekia eta erabiltzeko erraza izateko (Agirre *et al.*, 1995). Linguistek etengabe eguneratzen dute datu-basea, eta behar denerako esportatu egiten da.

azaleko formak eta forma lexikalak mapatzen dituen. Euskarazko analizatzaile/sortzaile morfologikoa Xerox enpresan garatutako XFST⁶ transduttore lexikalak erabiltzen inplementatu da (Alegria *et al.*, 1997), prozesu morfologikoa oso azkarra bihurtuz (milaka hitz segundoko).

VI.2.3 Euskararako prozesadore morfologikoaren egokitzapena

Baina, gure apustua itzulpen sistemarako kode irekiko teknologia erabiltzea izan da eta XFST liburutegia software jabetuna da, liburutegi horren baliokide librerik oraingoz ez dagoelarik. Gure hautua *Apertium*eko hiztegien formatua eta transduktoreak erabiltzea izan da. Horretarako prozesadore morfologikoaren egokitzapena burutu behar izan dugu *Apertium*eko formatuak zenbait muga dituelako:

- Morfotaktikan ezin da ziklorik egin, eta beraz mugatu behar da hitzen morfema kopurua.
- Ezin da erabili erregela fonologikorik.

Egokitzapenean ondoko urrats hauek burutu ditugu:

1. Euskararen ahalmen sortzailea mugatzeko XFST erregela batzuk definitu ditugu, bi mailatako erregela morfofonologikoen bidez. Erregela hauek hitzen osaketan bi genitibo baino gehiagorik ezin dela egon ezartzen dute.
2. Aurreko erregelak hartzen dituen analizatzaile/sortzaile morfologiko bat eraiki dugu, XFST tresnak erabilita. Eraikitako FSTa, noski, kode itxikoa da.
3. *Apertium*eko formatuan aldaketa fonologikorik onartzen ez denez, aldaketa fonologikorik eragin ditzaketen lehen bukaerak identifikatu eta zerrenda bat prestatu dugu, bi mailatako erregela morfofonologikoak begiraturata.
4. Euskaraz dagoen jarraitze-klase bakoitzeko (hau da, flexio-eredu bakoitzeko) eta lema-bukaera mota bakoitzeko lema posibleen multzoa osatu dugu.
5. Multzo bakoitzetik lema bat hautatu da.

⁶Xerox Finite-State Tool. <http://www.cis.upenn.edu/~cis639/docs/xfst.html>

6. Hautatutako lemetarako forma flexionatu posible guztiak sortu ditugu 2. urratsean eraikitako transduktorearekin.
7. Sortutako forma bakoitzeko sasi-lema eta flexioa banatzeko prozesaketa automatiko bat burutu da:
 - Sasi-lema: lemaren forma kanonikoa izaten da, baina batzuetan ez da bere osotasunean agertzen (*erakuts*, *erakusten*; *literatura*, *literaturen*).
 - Flexioa: atzizki edo atzizki-segida, batzuetan fonologiarengatik txertatu behar diren karaktereekin ere bai.

Horrela, hautatutako lema bakoitzeko (eta, beraz, lemen multzo bakoitzeko) flexio posible guztiak eta dagozkien informazio morfologikoak lortzen dira.

8. *Apertium* formatura bihurtu da informazio guztia.

Prozesadore morfologikorako egokitu dugun formatua VI.13 irudian ikus daiteke.

Bi mailatako morfologia egokitzeko lan hau euskararen zuzentzaile ortografiko librea eraikitzeke ere erabili izan da, *hunspell*⁷ programan oinarrituta.

VI.3 Laburpena

Kapitulu honetan *Matxin* sistemaren sorkuntza-fasea azaldu dugu.

Sorkuntza sintaktikoa eta sorkuntza morfologikoa burutzeko moduluak azaltzeaz gain, euskararako hitz-hurrenkera eztabaidatu eta berrerabili dugun sortzaile morfologikoa eta berau kode irekikoa izan dadin egin dugun egokitzapena azaldu ditugu.

Kapitulu honekin *Matxin* sistemaren deskribapena bukatu dugu eta ondoko kapituluan sistemak erabiltzen dituen baliabide linguistikoen prestaketa azalduko dugu.

⁷<http://hunspell.sourceforge.net>

```

<e><p> <l>literatur</l>
  <r>literatura<s n="IZE"/><s n="ARR"/> </r></p>
  <par n="I_31"/>
</e>
<e><p><l>eraku</l>
  <r>erakutsi<s n="ADI"/><s n="SIN"/> </r></p>
  <par n="A4_9"/>
</e>
...
<pardef n="I_31">
  <e><p> <l>a</l>
    <r><j/><s n="DEK"/><s n="ABS"/><s n="MG"/> </r></p></e>
  <e><p> <l>a</l>
    <r><j/><s n="DEK"/><s n="ABS"/><s n="NUMS"/> </r></p></e>
  <e><p> <l>agatik</l>
    <r><j/><s n="DEK"/><s n="MOT"/> </r></p></e>
  <e><p> <l>aren</l>
    <r><j/><s n="DEK"/><s n="GEN"/><s n="NUMS"/> </r></p></e>
  <e><p> <l>arentzat</l>
    <r><j/><s n="DEK"/><s n="GEN"/><s n="MG"/>
      <j/><s n="DEK"/><s n="PRO"/><s n="MG"/> </r></p></e>
  <e><p> <l>adun</l>
    <r>dun<s n="ATZ"/><s n="ADJ"/><s n="IZO"/> </r></p></e>
  <e><p> <l>adunarentzat</l>
    <r>dun<s n="ATZ"/><s n="ADJ"/><s n="IZO"/>
      <j/><s n="DEK"/><s n="GEN"/><s n="NUMS"/>
      <j/><s n="DEK"/><s n="PRO"/><s n="MG"/> </r></p></e>
  ...
</pardef>
<pardef n="A4_9">
  <e><p> <l>ts</l>
    <r><j/><s n="AMM"/><s n="ADOIN"/> </r></p></e>
  <e><p> <l>tsi</l>
    <r><j/><s n="AMM"/><s n="PART"/> </r></p></e>
  <e><p> <l>tsiko</l>
    <r><j/><s n="AMM"/><s n="PART"/>
      <j/><s n="ASP"/><s n="BURU"/> </r></p></e>
  <e><p> <l>tsiko</l>
    <r><j/><s n="AMM"/><s n="PART"/>
      <j/><s n="ASP"/><s n="GERO"/></r> </p></e>
  <e><p> <l>tsiko</l>
    <r><j/><s n="AMM"/><s n="PART"/>
      <j/><s n="DEK"/><s n="NUMS"/>
      <j/><s n="DEK"/><s n="GEL"/></r> </p></e>
  <e><p> <l>sten</l>
    <r><j/><s n="AMM"/><s n="ADOIN"/>
      <j/><s n="ASP"/><s n="EZBU"/> </r></p></e>
  <e><p> <l>tsaraz</l>
    <r>arazi<s n="ATZ"/><s n="ADI"/><s n="SIN"/>
      <j/><s n="AMM"/><s n="ADOIN"/></r></p></e>
  ...
</pardef>

```

Irudia VI.13: Hiztegi morfologikoaren formatuaren adibidea

VII. KAPITULUA

Baliabide linguistikoen berrerabilpena

Hitz bat aurkitzen duzun bakoitzean mundu zati berri bat azaltzen zaizu. Ez du axola hitz hori hiztegieta dagoen edo ez. Hiztegia, intsektua kolezio batetan bezala dahortua eta orratzez zeharkatua, betirako lotua, bere esanahiarekin eta bere etimologiarekin, baina hila. Baina hitzak bere bizia du, hitza esaten duen pertsona bakoitzak arnasa ematen dio.

Ainhoari gutunak. Joseba Sarrionandia, 1990.

Baliabide linguistikoek hizkuntzaren teknologietarako nahitaezko azpiegitura osatzen dute. Aplikazio errealerarako baliabide linguistikoak zabalak eta sendoak izan behar dira.

Gure apustua estaldura handiko baliabideak berrerabiltzea izan da, baliabide horiek itzulpen automatikoko sistemarako erabilgarri izan daitezen beharrezkoak diren prozesaketak burutuz.

Atal honetan baliabide linguistikoen prestaketarako burutu ditugun bi lan aurkezten ditugu: lexikoi elebidunaren eraikuntza eta ezaugarri semantikoaren etiketatze erdi-automatikoa.

VII.1 Lexikoi elebidunaren eraikuntza

Lexikoiaren eraikuntza lengoia naturalaren prozesamenduko (LNP) funtsezko ataza da, lexikografia konputazionalen testu klasiko guztietan aitor-tzen den bezala, hainbat lanetan (Briscoe, 1991; Zernik, 1991) aipatu izan dela ere lexikoiak LNPrako sistemen itogune nagusia izatera pasatu direla.

Lexikoiak behar duten ezagutza eskuratzeko bi metodo erabili izan dira (Wilks *et al.*, 1988): demo-hurbilpena (*demo approach*), non lexikoi txiki eta

aberatsa eskuz eraikitzen den, eta liburu-hurbilpena (*book approach*), hiztegi-tan dagoen ezagutza ordenagailuak prozesu dezakeen formatu itzul-tzeko metodoak garatzen duena, hizkuntzaren estaldura handiena lortzeko asmoz. Liburu-hurbilpena erabiliko da, beraz, hamar edo ehun milaka sarre-ra lexikal behar dituzten aplikazio errealetarako lexikoi aberatsak eraikitze-ko (Walker eta Zampolli, 1989). Sarrera kopuru handi hori duten lexikoiak erai-kitzeko, beraz, berebiziko garrantzia du baliabide lexikalen berrerabiltzea.

Ildo beretik, Dorr eta Klavans-ek (1995) aipatzen dute funtsezko be-tebeharra dela jostailuzko hiztegiak alde batera uztea eta neurri handiko lexikoiak eraikitzea. Hainbat egileren ustez IAKo sistemak garatzeko orduan atazarik zailenetakoa da hiztegiaren eraikuntza (Pai *et al.*, 2004), larrien, garestien eta denbora gehien kontsumitzen duena (Walker eta Zampolli, 1989). Hizkuntza bikote berri baterako hiztegiaren garapenak baliabideen %30a kontsumitzen duela estimatu izan da (Dillinger, 2001).

Hiztegien garapenerako badaude bost iturri nagusi (Dillinger, 2001): giza aditu-
en ezagutza, corpus paraleloak, paperezko hiztegiak, hiztegi ordenagai-luz irakurgarriak (*machine-readable dictionaries, MRD*) eta hiztegi sarean atzigarriak. Walker eta Zampolli-k (1989) aipatzen dute MRDak direla itu-
rririk aberats eta baliagarriena.

MRDek lexikografoen urteetako esfortzuarekin ezagutza lexikal eta se-mantiko ugari biltzen dute. Baina MRDak gizakiek erabiltzeko hiztegiak dira jatorriz eta beraien prozesaketa automatikoan arazo larriak sortzen di-ra ondoko arrazoiengatik (Boguraev eta Briscoe, 1989; Alshawi *et al.*, 1989; Dillinger, 2001; Ide eta Veronis, 1994):

- Edukia eta formatuaren informazioa nahastuta agertzen da. Datuen antolaketak irizpide lexikografiko eta tipografikoak jarraitzen dituen-
ez, efektu tipografikoek sarrera lexikalen antolaketa zein egitura logikoa adierazten dute, eta gainera etiketa tipografikoen esanahia desberdina izan daiteke testuinguruaren arabera.
- Formatu arauak ez dira zorrotzak eta zenbaitetan sarrera lexikal ba-tetik bestera aldatzen dira.
- Lexikografoek egindako erroreak ere agertzen dira noizbait, antzema-teko eta zuzentzeko oso zailak direnak.
- Informazioa kontraesankorra izan daiteke edo guztiz osatu gabea.

Arriola-k (2000) dioenez, MRDetatik informazioa eskuratzea ez da no-lanahiko lana, bertsio elektronikoen paperean inprimaturik dauden arazo

berberak dituztelako, eta batez ere informazioaren parte handi bat formalizatu gabe dagoelako, ez delako sendoa edo akatsak dituelako. Horregatik tratamendu konputazionala bideratzeko aurreprozesaketa bat beharrezkoa izango da.

Hau da, gizakientzat baliagarriak diren MRDak hiztegi ordenagailuz prozesagarriak (*machine-tractable dictionary*, *MTD*) bihurtu behar dira (Wilks *et al.*, 1988). Horretarako garatutako metodoek MRD hiztegian esplizituki zein inplizituki kodetuta dagoen informazioa ulertu behar dute, formatu, egitura eta edukia banatuz, MRD hiztegiari egitura homogeneoa emateko. Gainera, agertutako erroreak zuzendu behar dira. Sortuko den sistemaren beste baliabideak eta lexikoia bateragarriak direla ziurtatzea ere beharrezkoa da.

Prozesaketa automatikoan modu egokian tratatu daitezkeen hainbat eta hainbat sarrera zuzenean jaso daitezke, baina tratamendu berezia dutenak eskuz aztertu eta zuzendu beharko dira (Carrol eta Grover, 1989; Dillinger, 2001) lanetan aipatzen den bezala.

Atal honetan *Matxin* sistemaren espainiera-euskara lexikoi elebidunaren eraikuntzarako burututako lana azaltzen dugu, (Díaz de Ilarraza *et al.*, 2000b) lanean aurkeztu genuena. Lehenbizi lexikoia eraikitzeke orduan eza-rritako helburuak eta jarraitutako estrategia planteatuko ditugu. Erabilitako hiztegia eta beste iturriak aurkeztu ondoren prozesaketan burutako urratsak deskribatuko ditugu: hiztegiaren egokitzapena, hiztegiak bateragarri bihurtzea, lexikoia osaketa eta lexikoia XML formatura bihurtzea.

VII.1.1 Helburuak eta estrategia

Gure sistemak testu errealekin lan egingo duenez, geurean liburu hurbilpena nahitaezkoa da. Sarrera asko izango dituen lexikoi aberatsa eraikitzea izan da gure helburua, ahalik eta esfortzu txikienarekin, horretarako berrerabilpena izan behar duela gure estrategiaren gakoa.

Lexikoia eraikitzeke badugu ustiatu dezakegun oso baliabide erakargarria: *Elhuyar* espainiera-euskara hiztegiaren (Elhuyar, 2000) bertsio elektronikoa, 60.000 sarrera baino gehiago dituen. Sarrera kopuru hori hutsetik itzulpen sistemarako eskuz kodetzea suposatuko lukeen lan itzela burutu beharrean, hiztegiaren MRDa *Matxin* sistemarentzat erabilgarri izateko egokitu dugu.

Sistemak beharko dituen kategoria itxien¹ informazioa ezin da prozesaketa automatikoen bidez hiztegi elebidunetik eratorri, hiztegian ez delako

¹Kategoria itxiak: determinanteak, izenordainak, interjekzioak, loturazko elementuak eta preposizioak.

IAko sistemak behar duen informazio guztia: mugatasuna, numeroa, pertsona, postposizioen informazioa... Itzulpenenerako hain garrantzizkoa izango diren kategoria itxien baliokidetzak elebidunak eskuz kodetu behar izan ditugu. Eskuzko lan hau lorgarria da, kategoria itxietako sarrerak, guztira, ez direlako 600era iristen.

Gainera, hiztegiaren euskarazko ordainak ezaugarri semantikoen informazioarekin aberastu ditugu. Ezaugarri semantikoak modu erdi-automatikoan etiketatzeko egindako lana (Díaz de Ilarraza *et al.*, 2000a, 2002b) VII.2 atalean azalduko dugu.

Horrez gain, itzulpen automatikoko prozesuaren emaitzaren kalitatea handitzeko, eta batez ere hitzen adiera-desanbiguazioa burutzen ez den bitartean, mesede handia egingo du hitz anitzeko unitate lexikalak (HAUL) itzultzeko ahalmena izatea. Zenbaitetan HAUL horiek hitzez hitz modu egokian itzultzea ezinezkoa da, eta besteetan, posible izanda ere, zaila izan daiteke hitz bakoitzeko itzulpen egokia lortzea hitzen adiera desanbiguazioa ez bada egiten. Horregatik berebiziko garrantzia du *Elhuyar* hiztegiak ematen dituen lexia eta lokuzioez baliatzea, eta sarean atzigarri dagoen *Euskalterm* terminologia banku publikoa. eta corpusetatik erauzitako entitateen zerrenda elebidunak erabiltzea.

Iturri desberdin guzti horietako informazioa MTD bihurtzeko ondoko prozesaketak burutu dira:

1. Egitura homogeneoa lortzeko egokitzapena.
2. *Matxin* sistemaren beste baliabideekin bateragarri egitea. Espainierazko sarrerak eta beren kategoriak *Freeling* analizatzailearekin egiaztatu dira, eta euskarazko ordainak, berriz, sorkuntzarako erabiliko den euskararen datu-base lexikalean (EDBL) oinarritzen den *Morfeus* analizatzailea. Gainera kategoria-sistemak bateratu behar izan dira.

Prozesaketa horretatik bi fitxategi lortu ditugu:

- Lexikoi elebiduna, transferentzia lexikalean behar den informazio guztiarekin.
 - Lexikoian agertzen diren espainierazko hitz anitzeko unitate lexikalen (HAULak) zerrenda, analisi-fasean HAUL horiek identifikatu ahal izateko.
3. Lexikoaren osaketa. Sistemak erabiliko duen lexikoa osatzeko ondoko lanak burutu dira:

- Aurreko urratsetan sortutako hiztegiaren ordainen ordena moldatzea.
 - Euskarazko ordainetan ezaugarri semantikoen informazioa gehitzea.
 - Kategoria itxien hiztegitxoa, eskuz kodetutakoa, integratzea.
4. Lexikoi elebiduna XML formatu estandarrera bihurtzea, *Opentrad* proiektuaren espezifikazioak jarraituz.

Prozesaketaren eskema VII.1 irudian ikus daiteke. Lexikoi elebidunaren diseinua V.1.2 atalean deskribatu dugu.

VII.1.2 Berrerabilitako baliabideak

Matxin sistemaren es-eu lexikoia eraikitzeko lau iturri erabili ditugu. Baliabiderik garrantzitsuena *Elhuyar* es-eu hiztegiaren bertsio elektronikoa izan da, bere sarrera eta hitz anitzeko azpisarrerekin. Lexikoia aberastu dugu sarean publiko dagoen *Euskalterm* terminologia banku publikoaren HAULak eta corpusetatik erauzitako entitateen zerrenda elebiduna gehituz. Azkenik zenbait hizkuntzalariek eskuz kodetu beharreko kategoria itxien hiztegitxoa osatu dute. Esan beharra dago hizkuntzalariek ere lexikoi osoaren egokitasuna bermatzeko lana burutu dutela.

Zenbait sarrera iturri desberdinetan ager daitezke, batzuetan ere ordain desberdinekin, sarrera horien ordainak iturrien lehentasunari jarraituz ordenatzen direla: aurretik *Elhuyar* hiztegia, gero *Euskalterm* eta azkenik entitateen zerrenda.

Ondoko ataletan erabilitako baliabideak deskribatuko ditugu.

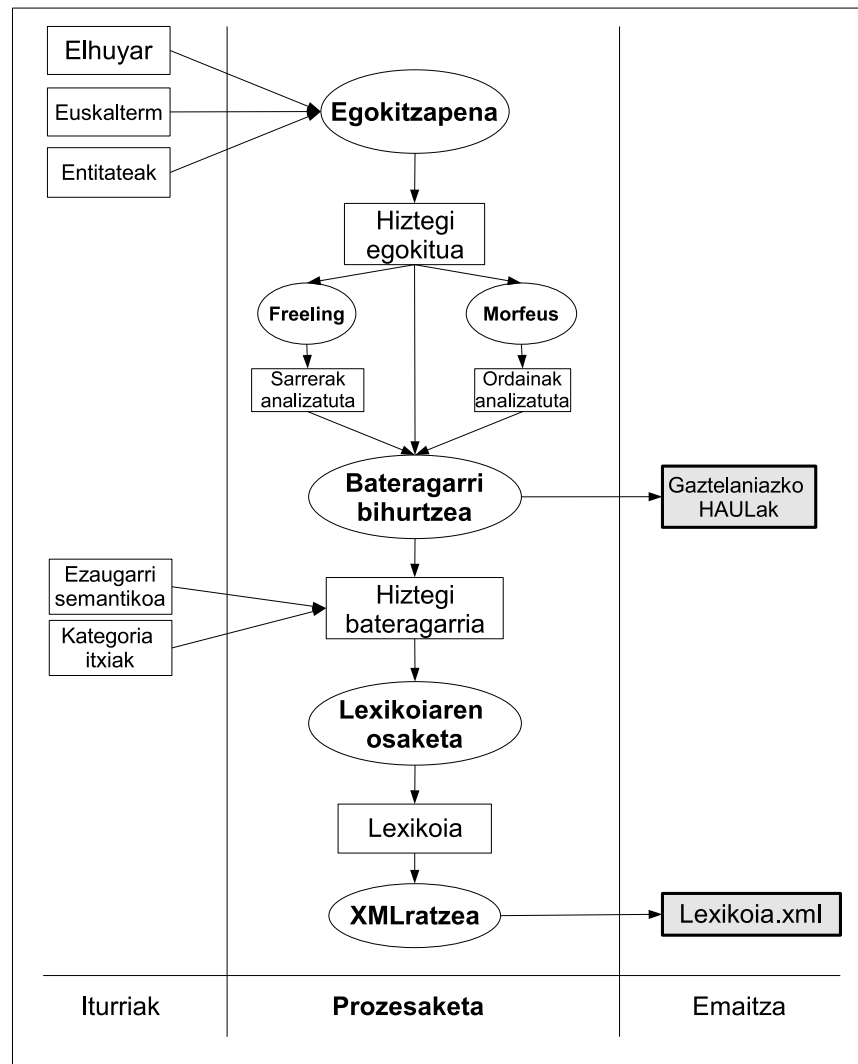
Elhuyar espainiera-euskara hiztegia

Elhuyar espainiera-euskara hiztegi berria (Elhuyar, 2000) erabili da, Euskaltzaindiaren *Hiztegi Batua* eta gomendioak jarraitzen dituen. Hiztegi hau Interneten erabilgarri dago².

Sarrera bakoitzerako, haren ordainak eta adibideak ematen ditu, baita azpisarrerak (lexiak zein lokuzioak) eta beste hainbat informazio ere.

Lexiak dira lexikalizatu gabeko unitate funtzional bat osatzen duten hitzen elkarketak, multzoaren esanahia haren elementuek dituzten esanahitik atera daitekeela (adib: *abeto de Douglas*). Lokuzioak, berriz, elementuen baturaren esanahi desberdin duten hitz multzoak dira (adib: *apretar el paso*).

²http://www1.euskadi.net/hizt_el/indice_e.htm



Irudia VII.1: Hiztegiak prozesaketa

```
// Sarreren fitxategia
aaronita    adj./s.      aarondar
ababol      s.m.        Bot. mitxoleta (Papaver rhoeas)
abacero, -a s.          janari-saltzaille
abacial     adj.        abade-, abadesa-, abadia- (hitz-elkarteetan)
abadejo     s.m.        Zool. bakailao (Gadus morhua)
abadejo     s.m.        Zool. abadira (Pollachius pollachius)
abadengo, -a adj.       abadearen, abadesaren, abadiaren; abade- (hitz-elkarteetan)
abalanzar   v.pnrl.     oldartu; (-en) gainera joan
abalanzar   v.tr.       balantza berdindu
abalanzar   v.tr.       bota, jaurti, egotzi

// Lexien fitxategia
a beneficio de      -en alde
a lo bestia         zakarki, konturik gabe, gogortasunez; astoaren moduan
abeto de Douglas    Douglas izei (Pseudotsuga menziesii)
aguja ferroviaria   (trenbide-)orratz

// Lokuzioen fitxategia
ahorcar los libros  ikasteari utzi, ikasketak utzi
ni un alma          inor(txo) ere ez; arimarik ere ez, kristaurik ere ez
despertar el apetito gosetu, gosea egin
apretar el paso     azkarrago/bizkorrago ibili
... y santas pascuas ... eta kito
tener mono de       -en premia izan, -en beharra izan
estar chiflado por  -ekin txora-txora eginda egon, -ekin txoratuta egon
```

Irudia VII.2: *Elhuyar* hiztegiaren hiru fitxategietako adibideak

Guk erabili dugun hiztegi elektronikoaren bertsioa hiru testu fitxategitan banatuta dago: sarreren fitxategia, lexiena eta lokuzioena. Fitxategi horien adibideak VII.2 irudian ikus daitezke.

Sarreren fitxategiko lerro bakoitzean sarrera baten adiera bana agertzen da ondoko formatuarekin:

```
Sarrera  Kategorial1[/Kat2/...] InformSemantikoa Ordain1[, Ordain2,...]
```

Lexien eta lokuzioen fitxategiaren lerroen formatua hauxe da:

```
Sarrera  Ordain1[, Ordain2,...]
```

Sarrera arrunten fitxategian 62.215 lerro daude, lexien fitxategian 4.783 eta lokuzioenean 6.657. Guztira 73.665 lerro.

Sarrera arrunten fitxategiko kopuruak kategoriaka VII.1 taulan ikus ditzakegu. Kontutan hartu behar da lerro bakoitzean kategorია bat edo gehiago ager daitezkeela (adib: *aaronita adj./s.*).

Euskalterm terminologia-banku publikoa

Euskalterm hiztegi terminologikoa UZEIko hiztegi guztietako fitxak biltzen ditu, gaiaren arabera sailkatuta. Euskaltzaindiaren arauen arabera eta normalizazio-batzordeetan hartutako erabakien arabera eguneratzen da. Termi-

Kategoria	Kopurua
izena	38.502
adjektiboa	13.343
aditza	11.215
adberbioa	758
izenordaina	260
determinantea	247
preposizioa	101
interjekzioa	95
loturazko elementua	39

Taula VII.1: Sarrerren fitxategiaren kopuruak kategoriak

// HAUL seguruak		
acta de nacimiento	jaiotza-agiri	Zuzenbidea Zuzenbide zibila Administrazioa
arreglos florales	lore-lanak	Psikologia. Hezkuntza Hezkuntza Lanbide Heziketa
colegio de abogados	abokatuen elkargo	Zuzenbidea Zuzenbide prozesala
consejo nacional	kontseilu nazional	Politika, soziologia, antropologia Politika, orokorra
corte suprema	gorde goren	Politika, soziologia, antropologia Politika, orokorra
jefe de gabinete	kabineteburu	Psikologia. Hezkuntza Hezkuntza
horas hábiles	lanorduak	Zuzenbidea Zuzenbide administratiboa
sitio web	web gune	Informatika Internet Antolamendua, erakundeak
// HAUL ez seguruak		
a la izquierda	ezkerretara	Industria Teknologia mekanikoa
a voces	ahotsetara	Arteak Musika
campo de trabajo	lan bortxatuen esparru	Politika, soziologia, antropologia
casco urbano	hiriaren erdialde	Eraikuntza eta hirigintza Hiriaren egitura
en juego	jokoan	Kirola, jokoak, aisia Kirolak Errugbia
estado actual	uneko egoera	Informatika
la negativa	ezezkoa	Filosofia
lugar visible	toki ageri	Garraioak Zirkulazioa
Mi PC	Ordenagailua	Informatika Bulegotika Orokorrak

Irudia VII.3: *Euskalterm* bankuko zenbait sarrera

noen itzulpena espainieraz, ingelesez edota frantsesez ematen du. *Euskalterm* hiztegia sarean atzigarri dago³.

Gure erabilerarako, *Euskalterm* bankuan espainieraz dauden HAUL guztiak Internetetik jaitsi ditugu, beren euskarazko itzulpenarekin: guztira 75.860 termino. *Elhuyar* hiztegian agertzen diren terminoak baztertu ditugu, hiztegiro itzulpen badugulako.

Termino horietako asko ez dira HAUL seguruak, hau da, beti ez dute unitate lexikal bezala funtzionatu eta, beraz, itzulpen desberdina izan dezakete testuinguruaren arabera (ikusi VII.3 irudiko adibideak). Horregatik beharrezkoa izan da eskuz errebisatzea, soilik seguruak direnak erabiltzeko.

Jasotako terminoen kopurua oso handia denez denak eskuz errebisatze-ko, lehenengo fase batean jasotakoan parte bat bakarrik eskuz errebisatzea eta erabiltzea erabaki dugu, corpusean maiztasun handiena dutena, alegia.

³www1.euskadi.net/euskalterm

Canal de Navarra	Nafarroako Ubide
Corea del Sur	Hego Korea
Gobierno Vasco	Eusko Jaurlaritza
Hospital de Navarra	Nafarroako Ospitale
Protocolo de Kyoto	Kyotoko Protokolo

Irudia VII.4: Entitateen zerrenda elebiduneko zenbait sarrera

Horretarako termino horiek Google⁴ bilatzailean duten agerpen-maiztasuna neurtu dugu eta, horren arabera sailkatu: 5.000 agerpen baino gehiago dutenak soilik errebisatu dira eskuz, hau da, 2.800 termino (%3,7).

Errebisatzeko orduan seguruak diren HAULak aukeratzeaz gain, zenbait zuzenketa egin dira eta, gaur egun *Matxin* sistemaren lexikoian ordain bakar bat jasoko denez, eskuzko lana errazteko ordainen artean bakar bat, egokiena dena, aukeratu da, besteak baztertuz.

Hizkuntzalariak egin dute errebisio lan hau eta emandako 2.800 terminoetatik 1.720 onartu dira eta 1.080 baztertu HAUL seguruak ez zirelako.

Entitateak

Beste lan baterako IXA taldean corpusetik erauzitako entitateak erabili ditugu. Corpora 2002. urteko *Egunkariarena* da, eta bertan agerpen kopuru minimo bat (10 aldiz) gainditzen zuten toki eta erakunde bezala identifikatutakoak jaso dira. Jasotako entitate guzti horiei izen bereziaren kategoria eta azpikategoria automatikoki esleitu zaie, *Matxin* sistemaren notazioa erabiliz.

Guztira 910 termino erauzi dira. Fitxategi honetako adibide batzuk VII.4 irudian ikus daitezke.

VII.1.3 Hiztegien egokitzapena

Hiztegietak informazioa, erabili ahal izateko, prozesu automatiko baten bidez egokitu dugu, egitura homogeneoa emanaz.

Prozesu horrek soberazko informazioa ezabatzen du, zuzenean prozesatu ezin diren lerroak baztertzen ditu, geroago eskuz lantzeko, eta sarrera edo ordain anitz adierazten denean banatu egiten ditu.

Sarreraren egokitzapena

Sarreretan aurkitzen diren homografo zenbakia, arlo semantikoaren eremua eta erabiliko ez dugun bestelako informazioa ezabatzen da.

⁴www.google.com

Sarrera	Sarrera zuzenduak	Kopurua
tumbón, ona	tumbón tumbona	11.506
abotagarse; abotargarse	abotagarse abotargarse	241
cardíaco, -a; cardiaco, -a	cardíaco cardíaca cardiaco cardiaca	22
Aezkoa{(Valle de)}	Aezkoa Valle de Aezkoa	29
Aguinaga {Señorío de}	Aguinaga Señorío de Aguinaga	25

Taula VII.2: Sarrerren egokitzapena

Kategoria itxiak dituzten sarrerak (determinanteak, izenordainak, preposizioak eta konjuntzioak; guztira 1.400 sarrera) baztertzen dira, eskuz kodetuko ditugulako.

Gainera, preposizio batekin edo *que* konjuntzioarekin bukatzen diren HAULak (*a favor de* → *alde izanik*), eskuz aztertzeko bereizten dira (2.900 kasu).

Sarrera anitzen banatzean ondoko kasuak eman daitezke:

- Sarrera batek kategoria bat baino gehiago duenean, kategoria bakoitzarekin lerro bat sortzen da kategoria bakoitzarekin.
- Sarrera batek parentesiak, giltzak, koma edo puntua eta koma erabiliz sarrera anitz adierazten duenean, bakoitzerako sarrera desberdin bat sortzen da (ikus VII.2 taula).

Ordainen egokitzapena

Ordainetan, berriz, soberazko zuriguneak kentzeaz gain, hauxe izan da egin-beharrekoa:

- Ordain anitz adierazten denean, parentesiak edo “/” ikurra erabiliz, bakoitzerako ordain bat sortu
- Parentesien artean agertzen den bestelako informazioa ordainetik kendu.

VII.3 taulan agertzen dira prozesaketa honetako adibideak eta prozesaketa mota bakoitza zenbat ordainei aplikatzen zaien.

Sarrera	Ordaina	Ordain moldatua	Kopurua
arramblar	hondarrez bete (errekek, ibaiek, etab.ek)	hondarrez bete	25
abacial	abadia- (hitz-elkarteetan)	abadia-	1323
ababol	mitxoleta (Papaver rhoeas)	mitxoleta	5028
abandonar	(bakarrik) utzi	utzi	962
abandonado	[pueblo, casa, etc.] huts	huts	1098
comisaria	polizi(a)-etxe	polizia-etxe polizi etxe	29
acantilado	malkar(tsu)	malkartsu malkar	2.586
abalarzar	(-en) gainera joan	-en gainera joan gainera joan	145
acuchillador	(zoru-)leuntzaile	zoru-leuntzaile leuntzaile	264
abanderizar	bandotan/banderizatan antolatu	bandotan antolatu banderizatan antolatu	4.748

Taula VII.3: Ordainen garbiketa

aaronita	adj.	aarondar
aaronita	s.	aarondar
abacero	s.	janari-saltzaile
abacera	s.	janari-saltzaile
abadejo	s.m	bakailao
abadejo	s.m	abadira
abalarzar	v.pnrl.	oldartu, -en gainera joan, gainera joan
abalarzar	v.tr.	balantza berdindu
abalarzar	v.tr.	bota, jaurti, egotzi
aguja ferroviaria		trenbide-orratz, orratz
tener mono de		-en premia izan, -en beharra izan

Irudia VII.5: Hiztegi egokitua. Adibideak

Karaktere bereziak (“∅”, “+”, “-”...) dituzten ordainak eskuz errebisatzeke bereizi dira. Burututako lehenengo errebisioan hiztegi egokitan marra (“-”) duten zenbait ordain onartzea erabaki dugu:

- Hitz-elkarteetako lehenengo hitza duten ordainak: ordainaren azkeneko ikurra “-” da (2.232 sarrera, adib: *abade-*)
- Postposiziodun aditzak: aditz baten aurretik postposizio atzizki edo postposizio-konposatua dagoenean (440 kasu, adib: *-en alde agertu*).

Lexikoiaren eraikuntzarako prozesaketako lehenengo ataza honen emaitza da hiztegi egokitua, non lerro bakoitzean sarrera bakoitzaren kategoria bakarreko adiera bat egongo den, dagozkion ordain guztiekin (ikus VII.5 irudiko adibideak).

VII.1.4 Hiztegiak beste baliabideekin bateragarri bihurtzea

Eraikitzen dugun lexikoi elebiduna sistemaren beste baliabide lexikalekin bateragarria izango dela ziurtatu behar izan dugu.

Iako sistemaren transferentzia-fasean lexikoi elebidunean bilaketak burutzeko *Freeling* analizatzailarekin lortutako espainierazko lema eta kategoria erabiliko dira. *Freeling*ek hitzen informazio morfologikoa adierazteko erabiltzen dituen etiketak *Eagles* taldeak europar hizkuntz guztietako lexikoi eta corpusak etiketatzeiko *Parole* etiketa multzoan oinarritzen da⁵.

Transferentzian lortutako ordaina sorkuntza fasean postposizio informazioziorik badu, ordainaren lema eta beharrezkoa den informazio morfologikoarekin euskararen datu-base lexikalean (EDBL) oinarritzen den prozesadore morfologikoak hitzaren forma emango du. *EDBL*ek hitzen informazio morfologikoa euskararentzat egokitutako kategoria-sistemaren arabera adieratzen du⁶.

Beraz, analizatzailarekin eta sortzaile morfologikoarekin bateragarri izan behar dute hiztegiko sarrerek eta ordainek.

Elhuyar hiztegiko sarrera arruntetatik eratorritako lerroetan hiztegiko kategoria agertzen da, eta entitateen lerroetan *Matxin* sistemako lexikoiaren etiketatzea. Lokuzioek, lexiek eta *Euskalterm*eko terminoek, berriz, ez dute oraindik kategoriarik.

Hiztegietako informazioa *Matxin* sistemako baliabideekin bateragarria izateko beherago deskribatzen diren ondoko urratsak burutu dira:

1. *Elhuyar* hiztegitik datozen lerroetan dagoen kategoria. *Matxin*en erabiltzen diren kategoria-sistemetara egokitu:
 - Espainierazko sarrerentzat *Parolerena*.
 - Euskarazko ordainentzat *EDBLrena*.
2. Sarreren eta ordainen analisisia:
 - Espainierazko sarrerak *Freeling* analizatzailarekin.
 - Euskarazko ordainak *Morfeus* prozesadore morfologikoarekin.
3. Sarreren prozesaketa, kategoria eta azpikategoria ebazteko.
4. Sarreretan dauden espainierazko HAULen zerrenda lortzea analisirako.
5. Ordainen prozesaketa, ebazteko:

⁵Espainierako *Parole* etiketa multzoa B eranskinean ikus daiteke.

⁶Euskararako kategoria-sistema C eranskinean ikus daiteke.

- Kategoria eta azpikategoria.
- Segmentazio morfologikoa, ordaina lema ez baina hitz-forma bat denean.

Jarraian banan bana azalduko ditugu.

Kategoriaren egokitzapena *Elhuyar* hiztegitik datozen lerroetan

Elhuyar hiztegiko lerroetan hiztegi horretako kategoria-sistema erabiltzen da, eta azpikategoria ez da ematen. Lerro horietako sarrerari *Freeling*eko kategoria baliokidea ezartzen diogu eta ordainari EDBLko baliokidea. Kategoria-sistema horien arteko kategoria-baliokidetza VII.4 taulan ikus daiteke.

	<i>Elhuyar</i>	<i>Freeling</i>	EDBL
<i>Kategoria irekiak</i>			
Izena	s./s.m./s.f./m./f.	N	[IZE]
Adjetiboa	adj.	A	[ADJ]
Adverbioa	adv.	R	[ADB]
Aditza	v.intr./v.tr./v.pnrl.	V	[ADI]
Interjekzioa	interj.	I	[ITJ]
<i>Kategoria itxiak</i>			
Izenordaina	pron.	P	[IOR]
Determinantea	det.	D	[DET]
Loturazkoa	conj.	C	[LOT]
Preposizioak	prep.	P	

Taula VII.4: Kategoria-sistemen baliokidetza

Kategoria irekien etiketak soilik ezarriko dira, kategoria itxietako sarre-
rak eskuz kodetuko direlako, hiztegiaren egokitzapenean baztertu direla.

Sarreraren eta ordainen analisia

Sarrera guztiak *Freeling*ekin analizatzen dira. Ondoko lerrootan ikus ditza-
kegu, adibidez, *aduladora* hitzerako *Freeling*ek ematen dituen 2 analisiak,
analisi posible bakoitzerako lema, informazio morfologikoa duen etiketa eta
probabilitate bat ematen dituelarik:

```
aduladora
  adulator AQ0FS0 0.52686
  aduladora NCFS000 0.47314
```

Ordainak prozesatzeko, *Morfeus* euskarazko prozesadore morfologikoare-
kin analizatzen dira. Ordainak HAULak baldin badira, edo hitz-elkarketak,
azkeneko hitza soilik analizatzen da. Euskarazko prozesadore morfologikoak

ematen duen analisiaren informazioa ondoko adibidean ikus dezakegu, non *urdin* hitzerako bost analisi lortu diren:

<i>urdin</i> [ADI] [SIN]+[AMM] [ADOIN]	Aditza
<i>urdin</i> [ADJ] [IZO]	Adjektiboa
<i>urdin</i> [ADJ] [IZO]+[DEK] [ABS] [MG]	Adjektiboa mugagabea
<i>urdin</i> [IZE] [ARR]	Izena
<i>urdin</i> [IZE] [ARR]+[DEK] [ABS] [MG]	Izena mugagabea

Sarreraren prozesaketa, kategoria eta azpikategoria ebazteko

Espainierazko sarreraren kategoria eta azpikategoria ebazteko hiru multzo daude:

- Entitateak: lexikoian agertu behar den kategoria eta azpikategoria ezarrita dituzte.
- *Elhuyar* hiztegitik datozen sarrera sinpleak: badute kategoria, baina ez azpikategoria.
- *Elhuyar* hiztegitik datozen HAULak: ez dute kategoria eta azpikategoriaren informaziorik.

Sarrera sinpleen kategoria eta azpikategoria zeintzuk izango diren erabakitzeke, kasu hauek bereiztu dira:

- Sarrera letra larriz badago izen bereziaren etiketa (*NP*) esleituko zaio (2.644 sarrera).
- *Elhuyar* hiztegian emandako kategoria *Freelingek* emandako analisi batekin bat badator, kategoria hori mantentzen da eta analisi horretako azpikategoria jaso (71.798 sarrera).
- *Elhuyar*reko kategoria ez badator bat *Freelingeko* analisisiekin edo kategoriarik esleitua ez badu, *Freelingek* ematen duen analisisirik probableenaren kategoria eta azpikategoria hartzen dira. Desbiderapen hauen adibideak eta kopuruak VII.5 taulan ikus daitezke.

Hitz anitzeko sarreraren kategoria eta azpikategoriaren ebazpenean kontuan hartzen dira *Freelingek* emandako hitz anitzeko terminoaren lehenengo bi hitzen eta azkenekoaren kategoria, azpikategoria eta lema. Erabiltzen diren erregelak eta adibideak VII.6 taulan ikus daitezke. Adibidez, lehenengo hitza gerundio bat bada (*VMG*), orduan HAULA adverbio bat izango da; preposizioekin (*SP*) hasten diren HAULak adverbioztzat hartuko dira,

<i>Elhuyar</i>	<i>Freeling</i>	Adibidea	Kopurua
A	NC	vengador	1.942
A	RG	nunca	15
N	AQ	anónima	1.247
N	RG	realmente	9
N	VM	embobado	278
N	W	lunes	29
R	AQ	sucio	68
...			203

Taula VII.5: Sarrera sinpleen kategorietan desbiderapenak

eta determinante batekin (D) hasten direnak, bigarren hitza izena bada, kategoria hori hartuko da; puntuazio ikurrekin (F) hasten diren HAULak interjekzioak (I) izango dira. Erregela hauekin ebazten ez diren HAULak adberbio (RG) gisa etiketatuko dira.

1. hitza	2. hitza	Azken hitza	Esleitutako kategoria	Adibidea
		que	CS	así es que
el	año/día		RG	el año pasado
VMG			RG	dando tumbos
NC			NC	abejorro común
AQ	NC		NC	falso techo
AQ			AQ	fácil de cocer
VM			VM	echar una cana al aire
SP			RG	con el culo al aire
D	NC		NC	los años cuarenta
F			I	¡que te den morcilla!
...			RG	

Taula VII.6: Hitz anitzeko sarreren kategoriaren esleipen erregelak

Espainierazko HAULen zerrendaren osaketa

Lexikoi elebiduneko espainierazko sarrera HAULa bada, prozesuak sarre-ra hori *Freeling*eko lokuzioen fitxategian gehitzeko prestatzen du. Horrela, *Freeling*ek analisi-fasean HAUL moduan identifikatuko ditu, transferentzian HAUL gisa itzuli ahal izateko.

Prozesaketa horretan bi lan burutzen dira: HAULean flexiona daitezkeen hitzak markatzea eta analisisa egiterakoan HAUL osoko informazio morfolo-gikoa zein hitzetik jasoko den erabakitzea.

HAULen zerrendaosatzean, deklina daitezkeen hitzak markatzeko kasu hauek bereizten dira:

- Lehenengo hitzaren forma eta lema desberdinak badira, ez da ezer

Brigadas_Rojas	Brigadas_Rojas	NP00000
dos_o_tres	dos_o_tres	RG
<jornada>_de_huelga	jornada_de_huelga	\$1:NC
<acceso>_<directo>	acceso_directo	\$1:NC
<pender>_de_un_hilo	pender_de_un_hilo	\$1:VM I
su_majestad	su_majestad	\$2:NC I

Irudia VII.6: Hitz anitzeko sarrerak markatuta

markatzen (adib: *lágrimas de cocodrilo*). Pluralean diren HAUL hauek soilik pluralean ezagutu beharko dira testuak analizatzean.

- Bestela, lehenengo bi hitzak izena eta adjektiboa badira, biak markatzen dira, biak singular edo pluralean agertu ahalko direlako testuetan (adib: <asno> <salvaje> de Nubia).
- Bestela, lehenengo hitza izena bada, markatzen da, singular edo pluralean agertu ahalko delako testuetan (adib: <vaca> de raza pirenaica).
- Bestela, lehenengo hitza aditza bada markatzen da, aditza flexionatu ahalko delako testuetan (adib: <ver> las orejas al lobo).

Sortzen den lokuzioen fitxategiaren lerro bakoitzean ondoko eremuak agertuko dira:

HAULaMarkatua HAULA HAULarenInformazioMorfologikoa

HAULen forma kanonikoa erabiliko da, zeinean hitzen artean “_” ikurra jartzen den. HAULEan flexiona daitezkeen hitzak “<” eta “>” ikurrekin markatzen dira. Bigarren zutabeko HAULaren forma kanonikoa analizatzailerak emango duen lema da. HAULaren informazio morfologikoan HAULEko hitzak ezin badira deklinatu, *Parole* etiketa osoa kodetuta agertzen da, eta HAULEko hitzen bat deklinatu ahal bada, orduan kategoria eta azpikategoria kodetzen dira, bestelako informazio morfologikoa hitzeko zein analisitatik jasoko den seinalatuz.

Adibide batzuk VII.6 irudian ikus daitezke *Freelingek jornadas de huelga* analizatzean, *jornada* lehenengo (\$1) hitzaren *NC* etiketarekin bat egiten duen analisiaren informazio osoa ezarriko dio HAUL osoari, besteak beste pluraltasunaren informazioa.

Ordainen prozesaketa

Ordainetan kategoria eta azpikategoriaz gain, batzuetan leman ere egin beharko dira moldaketak, eta gainera ordaina lema ez baina hitz-forma bat

denean sorkuntzarako beharrezkoa izango den segmentazio morfoloikoa jasoko beharko da, eta adjektiboaren kasuan kokapen-informazioa ere bai (*AUR* edo *ATZ*).

Horretarako zenbait tratamendu bereizi dira:

- Hiztegiko kategoria adberbio edo interjekzioa denean, inoiz deklinatzen ez direnez, eta beraz, ordain horiek sortzaile morfoloikora ez direnez inoiz bidaliko, ez da beharrezkoa euskarazko analisiarekin egiaztatzea. Adberbio direnean *[ADB]/[ARR]* ezarriko zaie, eta interjekzioetan *[ITJ]*.
- Hiztegian emandako lema eta kategoria bat badatoz euskarazko analizatzaileak emandako analisiren batekin, hiztegiko kategoria hori esleitzen zaio analisiak emandako azpikategoriarekin.
- Euskarazko lema ezin izan bada analizatu, EDBLrako hitz ezezaguna delako, hiztegiko kategoria mantentzen da, eta kategoria horri dago-kion azpikategoriarik ohikoena ezartzen zaio. Ordain hauek sorkuntza morfoloikoa egiteko lexikorik gabeko sorkuntza burutu beharko da.
- Hiztegiko lema edo kategoria ez badatoz bat prozesadoreak emandako analisiarekin, kasu bakoitzean prozesu automatikoak moldaketa desberdin bat egiten du (ikus VII.7 taula). Horretarako kasuistikaren araberako erregelak programatu dira.

Bestalde, postposizio-atzizkiak edo postposizio konposatuak agertzen diren ordainetan (*-en alde agertu*) postposizioaren forma bere etiketarekin ordezkatzeko da, beste eremu batean gordez (*-en alde agertu* $\rightarrow$ *agertu* / *[GEN]++alde*).

VII.1.5 Lexikoiaren osaketa

Fase honetan, honaino lortu den hiztegi bateragarrian informazio semantikoa gehitzen da, ordainen ordena moldatzen da eta kategoria itxien hiztegia bertan integratzen da.

Ezaugarri semantikoen hiztegia

Hiztegiko euskarazko ordainak ondoko ezaugarri semantikoen informazioarekin aberastu ditugu: biziduntasuna (*BIZ*), hizkuntza (*HZK*), denbora (*DEN*), gaia (*GAI*) eta gailua (*GLU*). Guztira 15.143 etiketa gehitu dira eta ezaugarri semantiko bakoitzerako kopuruak eta adibideak VII.8 taulan

Desbiderapena	Hiztegia	Analizatzailea	Kategoria	Kop.
<i>Izenak</i>				
Pluralean	oihuak[IZE]	oihu[IZE][ARR] +[DEK][ABS][NUMP]	[IZE][ARR]	770
Absolutiboarekin	traizioa[IZE]	traizio[IZE][ARR] +[DEK][ABS][NUMS]	[IZE][ARR]	327
Izen berezia/arrunta	Epifania[IZE]	Epifania[IZE][ARR]	[IZE][IZB]	102
Lema desberdina	anafilaxi[IZE]	anafilaxia[IZE][ARR]	[IZE][ARR]	972
Adjektiboa	hori[IZE]	hori[ADJ][IZO]	[IZE][ARR]	3.000
Aditza partizipioan	salatu[IZE]	salatu[ADI][SIN] +[AMM][PART]	[IZE][ARR]	1.024
Aditza part. pluralean	edadetuak[IZE]	edadetu[ADI][SIN] +[AMM][PART] +[DEK][ABS][NUMP]	[IZE][ARR]	25
Izena eta adjetiboa	ezpain-harro[IZE]	harro[ADJ][IZO]	[IZE][ARR]	289
Genitiboarekin	Andaluziako[IZE]	Andaluziako[ADJ][IZL] +[DEK][NUMS] +[DEK][GEL]	[IZE][ARR]	1.279
Atzizkiduna	biguntze[IZE]	bigundu[ADI][SIN] tze[ATZ][IZE][ARR]	[IZE][ARR]	2.975
<i>Adjektiboak</i>				
Graduarekin	luzeegi[ADJ]	luze[ADJ][IZO] +[GRA][GEHI]	[ADJ][IZO]	124
Absolutiboarekin	nekazaria[ADJ]	nekazari[ADJ][IZO] +[DEK][ABS][NUMS]	[ADJ][IZO]	20
Izena deklinatu gabe	leize[ADJ]	leize[IZE][ARR]	[ADJ][IZL]	5.676
Izena genitiboarekin	absoluziozko[ADJ]	absoluzio[IZE][ARR] +[DEK][INS][MG] +[DEK][GEL]	[ADJ][IZL]	4.328
Adberbioa	erruz[ADJ]	erruz[ADB][ALGARR]	[ADJ][IZL]	265
Aditza	ireki[ADJ]	ireki[ADI][SIN] +[AMM][ADOIN]	[ADJ][IZO]	848
Aditza erlatibozkoa	funtsezkoa den[ADJ]	izan[ADL][AI][NR_HU] +n[ERL][MEN][ERLT]	[ADJ][IZL]	1.333
Aditza partizipioan	aterata[ADJ]	atera[ADI][SIN] +[AMM][PART] +ta[ERL][MEN][MOD]	[ADJ][IZO]	22
Atzizkiduna	eraigarri[ADJ]	eraitsi[ADI][SIN]garri [ATZ][ADJ][IZO]	[ADJ][IZO]	669
<i>Aditzak</i>				
Atzizkiduna	geldiarazi[ADI]	gelditu[ADI][SIN] arazi[ATZ][ADI][SIN] +[AMM][PART]	[ADI][SIN]	410
Aurritziduna	birlandatu[ADI]	ber[AUR]landatu [ADI][SIN] +tu[AMM][PART]	[ADI][SIN]	126

Taula VII.7: Moldatu beharreko ordainak

Etiketa	Ezaugarria	Kopurua	Adibideak
[BIZ+]	biziduna	1.642	adiskide, agnostiko, amerikano, arkume
[BIZ-]	ez biziduna	10.664	afari, agonia, aho_sabai, berdintasun
[HZK+]	hizkuntza	1.386	Asturiera, Bielorrusiera, Esperanto, Euskara
[DEN+]	denbora	22	arratsalde, aste, asteburu, eguerdi
[GAI+]	gaia	98	aluminio, manganeso, margarina, nylon
[GLU+]	gailua	20	automobil, hegazkin, satellite, traktore

Taula VII.8: Hiztegi semantikoa

ikus daitezke. Ezaugarri bakoitzak duen “+” edo “-” ikurra ezaugarriaren balio positiboa edo ez adierazten du.

Ezaugarri semantiko horien informazioa preposizioen hiztegiko hautapen-erregelatan erabiliko da. Biziduntasunaren etiketa ere sorkuntza morfologikoan erabiliko da.

Ezaugarri semantikoak modu erdi-automatikoan erauz ditzakegu (ikus VII.2 atala). Horrela egin dugu biziduntasuna (*BIZ*) ezaugarriarentzat. Bestalde hizkuntza (*HZK*) ezaugarria hizkuntzen zerrenda elebidunetatik automatikoki jaso dugu, eta beste ezaugarriak eskuz kodetu ditugu haien kopurua txikia delako.

Lehenengo ordainaren hautaketa

Gaur egun, *Matxin* sisteman hautapen lexikalerako desanbiguazio-teknikak erabiltzen ez direnez, lexikoian ordain bakarra jasoko da.

Hori bai, hautapen lexikalerako, sinple izanik ere, eraginkorra den ondoko estrategia erabiltzen dugu: lehenengo adieraren lehenengo ordaina jasotzen dugu. Ordain horrek gehienetan itzulpen egokia ematen du, lexikografoek eskuz eraikitako hiztegi elebidunean ordainak, intuitiboki izanda ere, maiztasunaren arabera ordenatzen dituztelako.

Lexikoian jasotzen den ordaina itzulpenetan oraindik are egokiagoa izan dadin, zenbaitetan ez dugu hiztegiko lehenengoa jaso, prozesaketa automatiko batek hautatzen duen lehenengoa ez den beste ordain bat baizik:

- Adiera bateko ordain bat hautatuko da baldin eta lehenengo adierako lehenengo ordaina baino antzekoagoa bada abiapuntu-hizkuntzako sarrerarekin⁷. Adibidez, VII.7 irudian ikusten dugun *abandonar* sarrerarako *abandonatu* ordaina jasoko da.

⁷Bi hizkuntzatako hitzak antzekoak izan daitezke formaz eta esanahiz sustraikideak (*cognate*) direlako, edo hizkuntza horietako batek besteari mailegutan hartu diolako.

abandonar	1. utzi, alde bat utzi, alde batera utzi, bertan behera utzi 2. abandonatu, utzi, laga
abrasar	1. kiskali, txingartu, erraustu, zizpildu, erre 2. erre, ihartu, zimeldu 3. erre, kiskali, sutan eduki 4. oso bero egon, bero-bero egon 5. beroz erretzen egon, beroz erretzen ari izan
cesta	1. saski, otzara, otarre 2. saski 3. xistera, zesta
cama	1. ohe; ohatze 2. etzaleku, etzauntza, azpiak; ohatze, ohantze 3. sabelaldi, umaldi, kumaldi

Irudia VII.7: Lehenengo ordainaren hautaketa. Adibideak

- Lehenengo adierako lehenengo posizioan ez dagoen ordain bat hautatuko da baldin eta lehenengoa baino maizago agertzen bada hurrengo adieretan. VII.7 irudiko *abrasar* sarrerarako *erre* ordaina jasoko da, *kiskali* baino adiera gehiagotan agertzen delako. Honek askotan ordain posibleen artean hiperonimo bat hautatzea eragiten du, eta hau ona izango da gure helburua oinarritzko ulermenerako sistema bat erakitzea delako.

Zenbaitetan prozesaketa automatiko honekin hautatutako ordaina *lagun gaiztoa* da eta itzulpena burutzeko ez da egokia. Adibidez, *cesta* sarrerarako automatiko hautatutako ordaina *zesta* da, *cesta* hitzarekin *saskia* ordainak baino antz handiagoa duelako; eta *cama* sarrerarako *ohatze* hautatzen da, *ohe* baino gehiagotan agertzen delako.

Horregatik automatikoki aldatutako sarrerak eskuz errebisatu dira, egokiak ez diren aldaketak baztertuz. Automatikoki aldatutako 2.280 sarreretatik, 670 aldaketa baztertu dira. Beraz, hiztegian aldatutako sarreraren kopurua 1.610 da.

Estrategia honen emaitzak onak badira ere, gure etorkizuneko lanetan itzulpenaren hautapen lexikalerako desanbiguazio teknikak aztertzeari heldu beharko diogu. Horretarako sistemaren lexikoi elebidunean hiztegiko ordain guztiak jaso beharko dira.

lo	det.	ø
la	det.	-a
los	det.	-ak
mi	det.	nire, ene; neure
alguno, -a	det.	(bat ere)...-ik; inor, ezer
bastante	det.	nahiko, aski
aquello	pron.	-a, hura (erlatibozko perpausen aurrekari gisa)

Irudia VII.8: *Elhuyar* hiztegiko kategoria itxien zenbait adibide

Kategoria itxien hiztegia

Kategoria itxien hiztegia hizkuntzalariek eskuz kodetua izan da, itzulpeneko beharrezkoa den informazio morfologikoa ez delako hiztegietan agertzen. Lan hori osatzeko, hizkuntzalariek hartu duten abiapuntua, besteren artean, *Elhuyar* espainiera-euskara hiztegiaren prozesaketan baztertzen diren kategoria itxietako lerroen informazioa izan da (ikus VII.8 irudiko adibideak).

Kategoria itxien hiztegi horretan determinanteak, izenordainak eta konjuntzioak sartu dira, guztira 480 sarrera. Preposizioen itzulpenak dituen berezitasunengatik preposizioak ez dira transferentzia lexikalean itzultzen, transferentzia estrukturalean baizik, eta horregatik ez dira lexikoi elebidunean kodetzen, preposizioen hiztegian baizik.

Kategoria itxien hiztegian kodetutako sarreraren zenbait adibide III kapituluko V.5 taulan ikus daitezke.

VII.1.6 Lexikoia XML formatura bihurtzea

Aurreko prozesaketan lortzen dugun lexikoia gure itzulpen-sistemak behar duen informazio guztia dauka, baina testu-hutsean kodetuta dago.

Lexikoi hau, *Apertium*en espezifikazioari jarraitzen dion XML formatura bihurtzen da. Eraitza *Matxin* sistemaren lexikoi elebiduna da. Bere formatua eta adibideak V.1.2 atalean azaldu ditugu. Gaur egun, lexiko elebidunean, guztira, 62.500 sarrera kodetuta daude.

VII.2 Ezaugarri semantikoen etiketatze erdi-automatikoa

Lengoaia naturalaren prozesamenduko sistemetan ezaugarri semantikoen informazioa garrantzi handikoa da hainbat egile azpimarratzen duten bezala (Sheremetyeva eta Nirenburg, 2000; Pustejovsky, 2000). Horren erakuslea da *Matxin* itzulpen automatikoko sisteman biziduntasun ezaugarriaren beha-

rra, adibidez preposizioen itzulpenean gertatzen den anbigutasuna postposizioarenren ebazpenean (*ir a la casa* → *etxera joan* vs *dar a la madre* → *amari eman*) edota sorkuntza morfologikoan leku-kasuetan atzizkiaren osaketarako sortzen den anbigutasuna (*ir a la casa* → *etxera joan* vs *ir al médico* → *medikuarengana joan*).

IXA taldeak garatutako euskararen datu-base lexikalak (EDBL) 75.000 gorako sarrerak ditu, horietatik 28.000 inguru izen arruntak izanik, baina ez du momentuz ezaugarri semantikoen informaziorik ematen, eta oso garestia litzateke eskuzko etiketatzea burutzea.

Horren aurrean, izen arrunten ezaugarri semantikoak modu automatikoan etiketatzeko saiakera egin dugu. Lehenengo hurbilpen batean bizidun ezaugarriaren informazioa corpusetik automatikoki erauztea saiatu gara oso emaitza urriekin. Horregatik saiakera ezberdin bat egin dugu, (Chodorow *et al.*, 1985) lanean aurkeztutako lanaren estrategiari jarraituz: hiztegi elebarrereko definizioetatik atera daitezkeen hitzen arteko azaleko erlazio semantikoen erabiltzea. Genus, erlature espeziifiko eta sinonimiaren propietateak eta erlazio semantiko horien informazioa erabilita, hitz gutxi batzuk eskuz etiketatu ondoren, ezaugarri semantikoen etiketatze automatikoa burutu dugu, emaitza oparoa lortuz (Díaz de Ilarraza *et al.*, 2000a, 2002b).

Etiketatzeko automatikorako saiakera antzuak corpuseko informazioa erabiliz

Ezagutza semantikoa corpusetik erauzten saiatu gara bi modu ezberdinetan: batean leku-kasuak erabili ditugu eta bestean genitibo lokatiboa. Erabilitako corpusa *Euskaldunon Egunkaria* egunkaritik jasotakoa da⁸.

Euskarazko leku-kasuetan (inesiboa, ablatiboa, adlatiboa eta bukaerazko adlatiboa) *gan* partikula agertzen da, deklinatutako lemak bizidun ezaugarria duenean (adib: *medikuarengan*)⁹. Saiakera honetan leku-kasuekin deklinatutako hitzen atzizkiak *gan* partikula baldin badu maiztasun handiarekin, hitz horri [+bizidun] etiketa esleitzen zaio, bestela [-bizidun] etiketa ezarri. Kasu horietan deklinatutako 38.836 agerpen jaso ditugu, 2.366 izen ezberdin izanik. Baina metodo honekin etiketa daitezkeen izenen kopurua oso txikia da: 2366 hitz (75 [+] eta 2291 [-]). Fidagarritasun minimo bat eskatzen ba-

⁸Erabilitako corpus honek 1.267.453 hitz dauzka, hortik 311.901 izen arruntak eta haue-tatik 7.219 izen ezberdin aurkitzen dira.

⁹Honek bi salbuespen ditu: izaki ez bizidunei metaforikoki bizidunen ezaugarria ematen zaienean (adib: *itsasoarengan dut esperantza*), eta ohiz kanpo izaki bizidunek ez bizidunen atzizkiak erabiltzen dituztenean (*hiru emakumetatik bat, gizonetik gizonera alde handia dago*).

diogu, adibidez lema bakoitzerako leku-kasuetan deklinatutako 5 agerpen, emaitzak urriagoak dira: 767 hitz (3 [+] eta 764 [-])

Genitibo lokatiboarekin deklinatuta agertzen diren hitzei [-bizidun] etiketa ezarri zaie, [+]bizidun] ezaugarria duten lemek ez baitute kasu hori onartzen¹⁰. Metodo honekin soilik 1.544 hitz etiketatzea lortu dugu, oso kopuru txikia. Lortutako emaitzak oso urriak izan direnez, metodo hauek alde batera uztea erabaki dugu.

VII.2.1 Etiketatze erdi-automatikoa hiztegiko informazioa erabiliz

Hitzen arteko azaleko erlazio semantikoak ezaugarri semantikoen etiketatze erdi-automatikoa egiteko erabil daitezke.

Estrategia honen aurrekaria (Chodorow *et al.*, 1985) lanean aurkeztutakoa dugu: eskuz etiketatutako hitzen ezaugarri semantikoa haien hiponimoei heredatzea genusen informazioa erabiliz, automatikoki etiketatzen den hitz bakoitzerako erabiltzaileari kontsulta egiten zaiola onartzeko edo ez¹¹.

Gure metodoak genusaren informazioaz gain, erlature espezifikoena eta sinonimiarena ere erabiltzen du, eta eskuz eginbeharreko lan bakarra hasierako hitz gutxi batzuk etiketatzera mugatuko da.

Hitzen arteko azaleko erlazio semantikoak

Gure prozesaketarako erabiltzen dugun hitzen arteko azaleko erlazio semantikoen informazioa (Agirre *et al.*, 2000) lanean lortutakoa da. Lan horretan *Euskal Hiztegia* (Sarasola, 1996) hiztegi elebakarraren definizioetako informazioa erabiliz, hitzen arteko erlazio semantikoa metodo automatikoekin erauzten dute¹².

¹⁰Hemen errore-iturri bat ager daiteke euskarazko analizatzaileak *ko* atzizkia erabiltzen duen kasu banatzailearekin deklinatutako hitzen analisi posibleen artean genitibo lokatiboarena ere ematen duelako.

¹¹Lan horretako helburu bat taxonomia bat sortzea da, baina horrela eraikitako hierarkien kalitatea zalantzan jartzen dute (Ide eta Veronis, 1993) lanean egindako ebaluazio kualitatiboan, erauzitako informazioaren %55a baino gehiago nahasia delarik. Horren arrazoiak bi motakoak dira: alde batetik hiztegiaren informazioa osatu gabekoa dela, eta bestetik terminoen arteko bereizketa eza eta termino eta kontzeptuen arteko bat-bat erlazio eza. Lan horretan ere azaltzen dute nola arazo horiei aurre egin ahal zaien hainbat hiztegietatik erauzitako informazioa erabiliz, informazioan agertzen diren arazoak %6raino jaitsiz.

Hala ere gure helburua ez da taxonomia bat sortzea, eta eginbehar horretan sortzen diren arazoez ez dute eraginik izango ezaugarri semantikoak etiketatzeko orduan.

¹²Oso emaitza onak lortu dute sinonimia eta hiperonimiarentzat (estaldura: %96,85; doitasuna: %96,47) eta 17 erlature espezifiko definitu dituzte (estaldura: %87,24; doitasuna: %100).

Hiztegiko definizioek patroia ezberdinei jarraitzen dizkiete. Definizioan agertzen diren hitzak sarrera lexikalaren sinonimoak, genus edo erlature espezifikoa izan daitezke. Smith eta Maxwell-en (1973) arabera hiru modu daude sarrera lexikal bat definitzeko:

- Sinonimo baten bidez: sarrera lexikalaren esanahi berdina izango du. *akabatu*. *Bukatu* (sinonimoa), *amaitu* (sinonimoa).
- Definizio klasiko baten bidez, *genus + differentia*: Sarrera lexikalaren esanahia definitzeko erabiltzen dira termino generiko (genus) bat, eta termino generiko horren azpian dauden beste sarrera lexikaletatik bereizten duten ezaugarriak (differentia). Sarrera lexikala eta genusen arteko erlazioa hiperonimiarena da, hau da, genusa sarreraren hiperonimoa da, eta sarrera lexikala genusaren hiponimoa da. *aireontzi*. *Hegalda daitekeen* (differentia) *zernahi ibilgailu* (genus).
- Erlature espezifikoen bidez: Sarrera lexikala eta bere definizioa hitz bat lotzeko bide sintaktikoak dira. Sarrera lexikala eta definizioaren gunearen arteko erlazio semantikoa determinatzen du. *ezpara*. *Tabanidae familiako intsektuei* (erlazionatutako terminoa) *ematen zaien izena* (erlaturea).

Etiketatzeko erdi-automatizazio metodoa

Hiztegian agertzen diren izen arruntak etiketatzeko 16.380 izenen 26.461 adieren definizioak erabiltzen ditugu. Horietatik 14.569 genus eta erlatureen informazioa erabiltzen dute, eta 11.892 sinonimiarena.

Metodoak ondoko urratsei jarraitzen dizkie:

1. Sarrera batzuen eskuzko etiketatzea.
2. Etiketatzeko automatikoa genus eta erlature espezifikoen informazioa erabiliz.
3. Etiketatzeko automatikoa sinonimiaren informazioa erabiliz.
4. Bigarren eta hirugarren urratsen errepikapena.

Izen bakoitzaren etiketatze automatikoaren fidagarritasuna kalkulatzeko, izen hori etiketatzeko erabiltzen diren adieren genus, erlature edo sinonimoek zuten fidagarritasuna kontuan hartzen da:

$$izenFid = \frac{\sum adieraFid}{adieraKop}$$

non *izenFid* izenaren fidagarritasuna den, *adieraFid* etiketatzeko erabilako adiera bakoitzaren genus, erlatore edo sinonimoaren etiketaren fidagarritasuna eta *adieraKop* izen horren adiera kopurua.

Etiketatzeko-prozesuan hiru etiketa erabiltzen ditugu:

- + : hitz horren adiera guztiek ezaugarri semantiko horren balio positiboa dute
- : hitz horren adiera guztiek ezaugarri semantiko horren balio negatiboa dute
- ? : hitz horren adiera batzuk ezaugarri semantiko horren balio positiboa dute eta beste batzuk negatiboa (adibidez, *aireko* hitza biziduna da *izaki* adierarako, baina bizigabea *hegazkin* adierarako)

Eskuzko etiketatzea. Gure metodoaren lehenengo urratsa hiztegian maiztasun handieneko genus eta erlatore espezifikoak ezaugarri semantiko baterako eskuz etiketatzea da. Etiketatu hitz horietatik hiztegiko beste hitzen ahalik eta kopuru handien etiketa inferituko da, hitzen arteko azaleko erlazio semantikoak erabiliz.

Eskuz etiketatzean hitz horien adiera guztiak hartzen dira kontutan eta, beraz, bat ezarriko zaie fidagarritasunaren balioan adiera guztietako informazioak bat baldin badatoz.

Arestian aipatutako hiru etiketak erabiltzeaz gain, informazioak ematen ez duten erlatoeei [x] etiketa berezi bat jarriko zaie, haiengandik informazioak ezin dela inferitu seinaltatzeko (adibidez *izena*, *mota* edo *banako* erlatoeak).

Etiketatzeko genus eta erlatoeen informazioa erabiliz. Eskuz etiketatutako hitzen etiketatetik hiztegiko beste hitzen etiketa inferi daiteke, genus eta erlatoeen informazioa erabiliz.

Inferentzia hori posiblea da hiperonimia erabiliz, genusak ezartzen duen erlazioan atributu semantikoak heredatzen direlako. Hau da, *langile* hitza [+bizidun] ezaugarria badu, bere hiponimo guztiek, edo beste modu batean esanda, *langile* hitza genus bezala duten hitz guztiek (adib: *basomutil*, *ehule*, *antzarazain*), bere ezaugarri semantiko bera izango dute, kasu honetan [+bizidun]. Propietate trantsitiboa betetzen denez, hitz batek genusetik heredatutako ezaugarri semantikoak heredatzen dituzte bere hiponimoek ere.

Hainbat erlatore espezifikoetatik ere definitzen duten adieraren ezaugarri semantikoa ondoriozta daiteke. Adibidez *nolakotasuna* erlatoea duten hitzek [-bizidun] ezaugarria izango dute, *aitatasuna* kasuan bezala.

Prozesaketa automatikoan (ikusi 1. algoritmoa) izen arrunt bakoitzarentzat genus eta erlatoeren informazioa duen adiera bakoitza aztertzen da, genusak definitzen duen hierarkia hiperonimikoan etiketatuta dagoen genus edo erlatore bat aurkitu arte, adiera horri etiketa hori ezarriz. Sarrera lexikal beraren adiera ezberdinek etiketa kontraesankorrak jasotzen badituzte, zalantza hori adierazten duen [?] etiketa esleitzen zaio. Adiera ezberdinak koherenteak badira, hau da, etiketa bera jasotzen badute, etiketa hori izenari esleitzen zaio, etiketak duen fidagarritasuna kalkulatur.

Jakina da hiztegien definizioetan zikloak daudela, bereziki kontzeptuak hierarkiaren goiko partean daudenean (Ide eta Veronis, 1993). Beraz, prozesaketan ziklo bat ematen bada bilaketa eten egingo da, adiera hori etiketatu gabe geldituz.

Etiketatzeari sinonimiaren informazioa erabiliz. Genus eta erlatoeren bidez lortutako etiketatzea sinonimia erlazioaren bitartez beste hitzetara zabal daiteke. Sinonimia erlazioan ezaugarri semantikoak konpartitzen dira, beraz adiera baten ezaugarri semantikoa ezagut dezakegu bere sinonimoaren etiketa ziur baldin badakigu. Hemen ere propietate trantsitiboa betetzen da, beraz sinonimoaren sinonimoa erabil dezakegu.

Prozesu automatikoan (1. algoritmoaren antzera), aurreko urratsean jasotako informazioa, izen berriak etiketatzeko erabil dezakegu sinonimiaren bidez. Gainera, adiera batzuetan genus edo erlatoeren informazioari esker etiketatutako izenen fidagarritasun handiagoa lor daiteke beste adieretan duten sinonimia informazioa erabiliz. Sinonimiarekin jasotako etiketak genusarekin jasotakoa berresten badu fidagarritasuna handiagotzen da. Sinonimiarekin jasotako etiketa ez badator bat aurretik jasotakoarekin, anbiguitasuna adierazten duen [?] etiketarekin seinalatuko dugu.

Prozesu automatikoaren errepikapena. Sinonimiarekin lortutako informazioa orain genus eta erlatoeren bidez heredatu daiteke beste hainbat hitzetara. Horretarako prozesu automatikoa hasieratik errepikatzen da, aurreko iterazioan guztiz etiketatuta izan diren izenen (fidagarritasunerako 1 balioa dutenak) informazioa erabiliz, beste adierak etiketatzen saiatzeko.

Prozesu hau behin eta berriz errepika daiteke, hitz etiketatu gehiago lortuz eta fidagarritasuna handiagotuz. Iterazio berri bakoitzean adiera gehiago etiketatzen direnez, izen bakoitzaren adieren artean inkoherentzia gehiago ere sortuko dira, [?] etiketarekin markatzen den hitz anbiguen kopurua ere handituz.

Prozesu iteratibo honek modu logaritmikoan hobetzen ditu emaitzak ite-

Algoritmoa 1 GenusekinEtiketatu

```

// sarrera: izena eta zerrenda (errekurtsiboki aztertutako genus zerrenda)
// irteera: izena.etik eta izena.fid (fidagarritasuna)

fidBatura = 0, adieraKop = 0
egin izenaren adiera guztietarako
  adieraKop ++
  baldin definizioan genus (edo erlatore) informazioa dagoen orduan
    baldin genus etiketatuta orduan
      adiera.etik = genus.etik
      adiera.fid = genus.fid
      fidBatura = fidBatura + adiera.fid
    bestela
      zerrenda = zerrenda + izena
      baldin genus  $\notin$  zerrenda orduan
        (adiera.etik, adiera.fid) = GenusekinEtiketatu (genus, zerrenda)
      amaiera baldin
    amaiera baldin

    baldin izena.etik = null orduan
      izena.etik = adiera.etik
    bestela
      baldin izena.etik  $\neq$  adiera.etik orduan
        izena.etik = [?]
      amaiera baldin
    amaiera baldin
  amaiera baldin
amaiera egin

baldin izena.etik = [?] orduan
  izena.fid = 1
bestela
  izena.fid = fidBatura / adieraKop
amaiera baldin

itzuli (izena.etik, izena.fid)

```

Izena	[±bizidun]	Maiztasuna	Genus edo erlatorea
nolakotasun	-	531	erlatorea
pertsona	+	377	genus
multzo	-	362	erlatorea
txikigarri	x	213	erlatorea
zati	-	230	erlatorea
gai	-	202	genus
tresna	-	188	genus
landare	-	187	genus
egoera	-	177	erlatorea
...			
buru	?	54	genus
...			

Taula VII.9: Eskuzko etiketatzea [±bizidun] ezaugarriarentzat

razio kopuruaren arabera, hainbat iterazio eginez gero hortik aurrera prozesua errepikatzeak ez du ekarpenik egiten.

Adibidea [±bizidun] ezaugarriarentzat

Genus eta erlature bezala maizen erabilitako 100 hitzak eskuz etiketatu dira [±bizidun] ezaugarriarentzat (ikus VII.9 taula) eta ondoren prozesu automatikoa burutu da

VII.10 taulan 1. iterazioaren genus eta erlatureen informazioa erabiliz egindako etiketatze automatikoa ikus dezakegu, genus duten adieren informazioa dagoen zutabearen eskuz etiketatu diren hitzak beltzez ikus ditzakegula:

- *babesgarri* hitzak [-] etiketa jasotzen du eskuz etiketatua zen bere adiera bakarreko erlatureari esker, 1 fidagarritasuna lortuz.
- *armadurarentzat* bere hiru adieretatik (denak genus edo erlatureen informazioarekin) bi adierek ematen dute etiketa berdina [-], fidagarritasunerako 0,66 balioa lortuz ($fid = \frac{1+1}{3} = 0,66$).
- *ama* izenak guztira bost adiera dauzka, haietatik hiru genus edota erlatureek informazioa dutela. Hiru adiera horietatik, bi etiketatuta daude [+] balioarekin; beraz, modu koherentean etiketatzen dute *ama* hitza 0,4 fidagarritasun balioa lortuz ($fid = \frac{1+1}{5} = 0,4$).
- *zinismo* hitza [-] etiketa jasotzen du, bere bi adieren genusak etiketatuta daudelako. Hori bai, haietako batek ez du 1 fidagarritasuna eta, beraz, 0,87 fidagarritasuna lortzen da ($fid = \frac{1+0,75}{2} = 0,87$).

Izena	Adiera kop.	Genus kop.	Genus duten adieren informazioa	Etik.	Fid.
babesgarri	1	1	(zer [-]1)	[-]	1
armadura	3	3	(multzo [-]1)(babesgarri [-]1)(soineko[])	[-]	0.66
ama	5	3	(emakume [+]1)(animalia [+]1)(eme[])	[+]	0.4
iturburu	3	1	(aterabide[])	[]	0
gertaera	1	1	(gauza [-]1)	[-]	1
giltzape	2	1	(toki [-]1)	[-]	0.5
espetxe	2	2	(eraikuntza [-]1)(leku [-]1)	[-]	1
adiskide	1	1	(pertsona [+]1)	[+]	1
adiskidetzako	1	1	(lagun[])	[]	0
apio	2	2	(jateko[]) (landare [-]1)	[-]	0.5
filosofia	2	2	(jakintza [-]1)(multzo [-]1)	[-]	1
ikusgune	2	1	(gune [-]1)	[-]	0.5
jarrera	2	2	(era [-]1)(ikusgune[-]0.5)	[-]	0.75
zinismo	2	2	(filosofia[-]1)(jarrera[-]0.75)	[-]	0.87

Taula VII.10: Genus eta erlatoreen informazioa erabiliz etiketatzearen 1. iterazioa

VII.11 taulan 1. iterazioaren sinonimiaren informazioa erabiliz egindako etiketatze automatikoa ikus dezakegu:

- *iturburu* hitza, bere adieretako batek zuen genus informazioarekin etiketatzerik izan ez duena, [-] etiketa lortzen du orain sinonimiaren informazioa erabiltzen duten bere bi adieretako bati esker, $0,06$ fidagarritasuna lortuz ($fid = \frac{0,2}{3} = 0,06$).
- *ama* hitza [+] balioarekin etiketatua izan da genusaren informazioari esker, baina sinonimia erabiltzen duten bere bi adieren sinonimoak [-] etiketa dutenez, adieren araberako anbiguotasuna ematen da eta [?] etiketa esleitzen zaio.
- *gertakuntza*, *lagun* eta *jateko* sinonimia informazioa duen adiera bakarra dute, eta horri esker etiketatzen dira.
- *giltzape* eta *ikusgune* hitzek, genusari esker [-] etiketa zuten eta sinonimia erabiltzen duen adierak etiketa koherentea jasotzen du, aurreko etiketa mantenduz eta fidagarritasuna handiagotuz.
- *ikusgune* hitzaren fidagarritasuna horrela kalkulatzen da: $fid = \frac{1+0,33}{2} = 0,66$

Prozesua errepikatzen da, aurretik genus eta erlatoreen informazioa erabiliz, eta ondoren sinonimiarena. Horretarako aurreko iterazioan guztiz etiketatuta izan diren izenen informazioa erabiltzen da (1 fidagarritasuna lortu

Izena	Adiera kop.	Sinonimo kop.	Sinonimiaren informazioa	Etik.	Fid.
iturburu	3	2	(etorki[]) (hasiera[-]0.20)	[-]	0.06
ama	5	2	(iturburu[-]0.06)(jatorri[-]1)	[+] → [?]	1
gertakuntza	1	1	(gertaera[-]1)	[-]	1
lagun	1	1	(adiskide[+]1)	[+]	1
jateko	1	1	(janari[-]1)	[-]	1
giltzape	2	1	(espetxe[-]1)	[-] → [-]	1
ikusgune	2	1	(ikuspen[-]0.33)	[-] → [-]	0.66

Taula VII.11: Sinonimiaren informazioa erabiliz etiketatzearen 1. iterazioko zenbait adibide

Izena	Adiera kop.	Genus kop.	Sinonimiaren informazioa	Etik.	Fid.
armadura	3	3	(multzo[-]1)(babesgarri[-]1)(soineko[-]1)	[-]	1
adiskidetzako	1	1	(lagun[+]1)	[+]	1
apio	2	2	(jateko[-]1)(landare[-]1)	[-]	1
ikusgune	2	2	(gune[-]1)	[-]	0.5
jarrera	2	2	(era[-]1)(ikusgune[-]0.5)	[-]	0.75
zinismo	2	2	(filosofia[-]1)(jarrera[-]0.75)	[-]	0.87

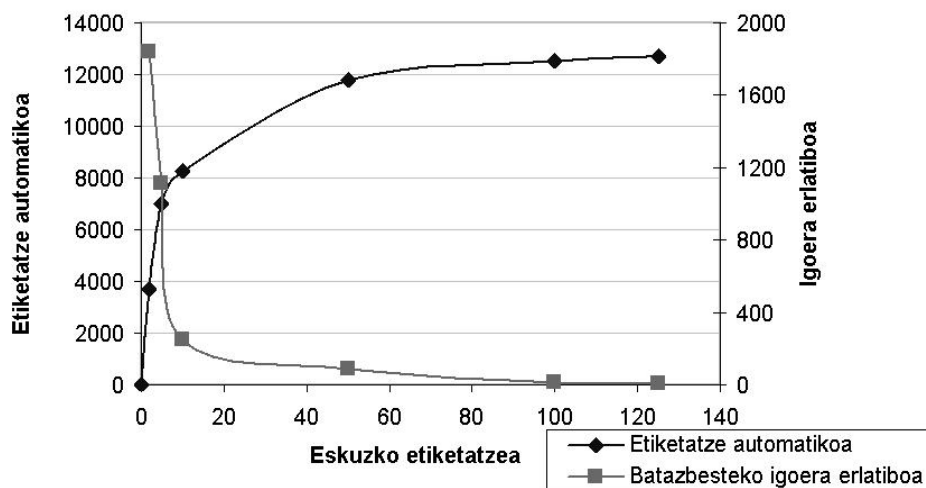
Taula VII.12: Genus eta erlatoreen informazioa erabiliz etiketatzearen 2. iterazioko zenbait adibide

dutenak: *babesgarri*, *adiskide*, *filosofia*, *ama*, *gertakuntza*, *lagun*, *jateko* eta *giltzape*, guztiz etiketatuta izan ez direnen adierak etiketatzeko. VII.12 taulan ikusten den bezala, 2. iterazio honetan *armadura*, *adiskidetzako* eta *apio* hitzetarako, aurretik etiketatzea lortu ez ziren adierak iterazio honetan etiketatzen dira, haien genusak aurreko iterazioan sinonimiaren informazioarekin guztiz etiketatu direlako. *ikusgune*, *jarrera*, eta *zinismo* berriz, aurreko iterazioan bezala etiketatzea lortzen da genusaren informazioarekin. Ondoren sinonimiaren informazioa erabiliz etiketatzen da, 2. iterazioa bukatuz. Prozesua nahi adina aldiz errepika daiteke.

Metodoaren eraginkortasunaren optimizazioa

Zenbait proba burutu ditugu [$\pm$ bizidun] ezaugarriaren etiketatzerako, eskuz etiketatutako hitzen kopuruaren eragina eta prozesua errepikatzen den iterazio-kopuruaren eragina aztertzeko.

Lehenik genus edo erlatore bezala gehien erabilitako 2, 5, 10, 50, 100, 125 eta 150 hitzak eskuz etiketatuz, (prozesu osoa 3 aldiz errepikatzen dela) VII.9 grafikoan ikus ditzakegun datuak jasotzen ditugu. Genus eta erlatore bezala maiztasun handienekin agertzen diren lehenengo 5 terminoak produktiboe-



Irudia VII.9: Etiketatzeko automatikoa eta eskuzko etiketatzea

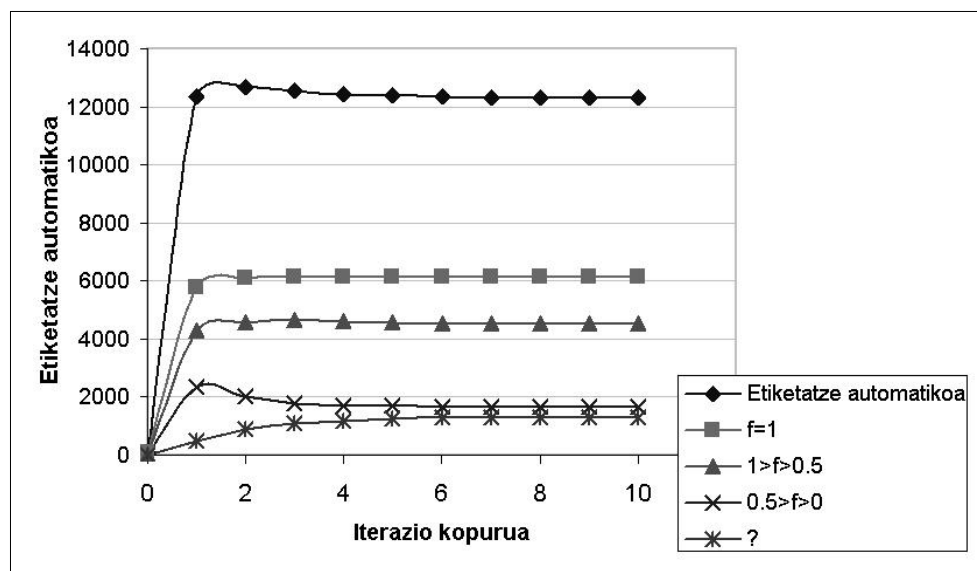
Iterazioak	Etik. autom.	f=1	$1/2 f/0.5$	$0.5 f/0$	?
1	12353	5764	4259	2330	474
2	12676	6095	4570	2011	897
3	12534	6132	4621	1781	1074
4	12434	6132	4590	1712	1175
...					
8	12308	6132	4513	1663	1301
9	12308	6132	4513	1663	1301

Taula VII.13: Etiketatzeko automatikoa eta iterazio kopurua

nak dira, eskuz etiketatutako bakoitzetik 1.000 baino gehiago etiketatzen direla automatikoki. Eskuz etiketatutako hurrengoekin emankortasuna jaisten doa, puntu batetik aurrera eskuz etiketatutako bakoitzetik soilik 7 termino etiketatzen direla automatikoki. Igoera erlatiboaren datuek adierazten dute eskuzko etiketatzearen hitz berriek batez beste zenbat hitz etiketatzen duten prozesu automatikoan.

Prozesua optimoa izateko zenbat iterazio burutu behar diren erabakitzeke, 1 eta 10 bitarteko iterazio kopuruarekin probatu dugu, eskuz 100 hitz etiketatu ondoren. VII.13 taulan eta VII.10 grafikoan ikus ditzakegu lortutako datuak.

Adiera guztietan etiketatutako izenen kopurua 3. iteraziotik aurrera mantentzen da, baina 1 fidagarritasuna duten etiketen datuak ez dira egonkortzen 8. iterazioa arte, informazio gehiago izatean kontraesan gehiago sor-



Irudia VII.10: Etiketatzeko automatikoa eta iterazio kopurua

fid=1	1>fid>0,5	0,5>fid>0	Guztira
%100	%100	%94	%99,2

Taula VII.14: Etiketatzeko automatikoaren doitasuna

tzen direlako, etiketa anbiguenen kopurua handituz. Emaizak modu horretan egonkortu direnean esan dezakegu eskuzko etiketatze horretarako eskuragarri zegoen informazio guztia erabili dela, emaitza zehatzenak lortuz.

Beraz, prozesua iterazio kopuru txiki batean egonkortzen da, exekuzio koste handirik gabe. $[\pm\text{bizidun}]$ ezaugarriarako etiketatze automatikoak 10 iterazio burutuz, 11 minutu eta 33 segundotan¹³.

Ebaluazioa

Etiketatzeko automatikoaren doitasuna eta estaldura aztertu ditugu, 100 hitz eskuz etiketatuta eta prozesua 8 iterazioekin burutuz.

Etiketatzeko automatikoaren doitasuna kalkulatzeko $[\pm\text{bizidun}]$ ezaugarriarako etiketatutako hitzen %1eko lagin bat hartu dugu eta eskuz errebisatu, VII.14 taulan ikus ditzakegun emaitzak lortuz.

Hasiera batean fidagarritasun osoa zuten etiketak soilik erabiltzea pen-

¹³CPU denbora Sun Sparc 10 makina batean, memoriako 512 Mb-ekin, 360 Mhz-tara.

fid=1	1>fid>0,5	0,5>fid>0	guztira
6.312	4.513	1.663	12.308

Taula VII.15: Etiketatze automatikoaren estaldura

	Corpuseako hitz agerpenak	Corpuseko izen ezberdinak
Guztira	311.901	7.219
Etiketatuak	(%68,2) 212.887	(%47,6) 3.434
[+]	17.408	356
[-]	195.479	3.078

Taula VII.16: Etiketatze automatikoaren estaldura corpusean

tsatuta bagenuen ere, emaitzak ikusi ondoren prozesuan lortzen diren etiketa guztiak erabiltzea erabaki dugu, %99,2ko doitasuna lortuz. Diseinatu eta inplementatu dugun prozesu automatikoa benetan eraginkorra da

Hiztegiko izenen %75,14ko estaldura lortzen da $[\pm\text{bizidun}]$ ezaugarriaren etiketatze automatikoan, 16.308atik 12.308 (ikus VII.15 taula). Batez beste eskuz etiketatutako hitz bakoitzarentzat, automatikoki 123 etiketatu dira

Etiketatzeko honen estaldura testu errealean ere kalkulatu dugu. Horretarako *Euskaldunon Egunkariaren* corpusa erabili dugu. VII.16 taulan ikus dezakegun bezala, testuan izenen agerpenen %68,2 etiketatu dira, corpuseko izen arrunt ezberdinen % 47,6a. Hau da, eskuzko oso etiketatze-lan txikia-rekin, corpusean agertzen diren 3 izenetatik 2 etiketatu dira.

Etiketatzearen orokortzea

Prozesuaren eraginkortasuna ikusita, beste ezaugarri semantikoetara zabal dezakegu metodoa. Horretarako inplementazioa egokitua izan da etiketatze automatikoa egin ahal izateko edozein ezaugarri semantikotarako egindako eskuzko etiketatzetik abiatuta.

Momentuz $[\pm\text{gizaki}]$ eta $[\pm\text{konkretu}]$ ezaugarritarako probatu dugu. Ezaugarri semantiko desberdinetarako jasotako datuak VII.17 taulan ikus ditzakegu.

Prozesua hobetzeko interesgarria izan zitekeen sarrera bakoitzaren adiera

Etiketatzeara	$[\pm\text{bizidun}]$	$[\pm\text{gizaki}]$	$[\pm\text{konkretu}]$
[+]	1.643	1.118	7.611
[-]	10.665	10.684	1.143
guztira	12.308	11.802	8.754

Taula VII.17: Beste ezaugarri semantiko batzuen etiketatzearen datuak

ezberdinen corpusean erabileraren distribuzio estatistikoari buruzko datuak kontutan hartzea, baina informazio hori ez daukagu.

VII.3 Laburpena

Aplikazio errealetarako baliabide linguistikoak zabalak eta sendoak izan behar direnez, gure apustua estaldura handiko baliabideak berrerabiltzea izan da. Kapitulu honetan *Matxin* sistemarako beharrezkoa diren baliabide linguistikoen prestaketarako burutu ditugun bi lan aurkeztu ditugu: lexikoi elebidunaren eraikuntza eta ezaugarri semantikoen etiketatze erdi-automatikoa.

Lexikoi elebiduna eraikitzeke erabilitako hiztegia eta beste iturriak aurkeztu ondoren prozesaketan burutako urratsak zehatz-mehatz deskribatu ditugu: hiztegiaren egokitzapena, hiztegiak bateragarri bihurtzea, lexikoiaren osaketa eta lexikoia XML formatura bihurtzea.

Preposizioen itzulpenean gertatzen den anbiguotasuna postposizioaren ebazpenean edota sorkuntza morfologikoan leku-kasuetan atzizkiaren osaketarako sortzen den anbiguotasuna ebazteko beharrezkoa den ezaugarri semantikoen etiketatze erdi-automatikoa ere aurkeztu dugu. Horretarako hiztegi elebazarreko definizioetatik atera daitezkeen hitzen arteko azaleko erlazio semantikoen informazioa erabilita, hitz gutxi batzuk eskuz etiketatu ondoren, ezaugarri semantikoen etiketatze automatikoa burutu dugu, emaitza oparoa lortuz.

VIII. KAPITULUA

Ebaluazioa

*Baga, biga, higa, laga, boga, sega,
zai, zoi, bele, harma, tiro, pun!*

Herrikoia

Kapitulu honetan itzulpen automatikoaren ebaluazioaren ikuspegi orokor bat emango dugu, IAre giza-ebaluaziorako neurriak eta neurri automatiko garrantzitsuenak deskribatuz. Azkeneko urteetan oso aktiboa izan da ikerketa arlo hau. Gaur egun gehien erabiltzen diren ebaluazio neurri automatiko arrakastatsuen ahuleziak aztertu ondoren, gure sistemaren irteeraren kalitatea neurtzeko erabili dugun metrika azalduko dugu: edizio-distantzia.

Ondoren, *Matxin* sistemaren irteeraren ebaluazioaren datuak aurkeztu eta interpretatuko ditugu, eta gure sistemaren emaitzak corpusetan oinarritutako *Matrex* sistemaren emaitzekin konparatu ahal izango ditugu.

Azkenik, sistemaren erroreen analisisa egingo dugu.

VIII.1 Itzulpen automatikoaren ebaluaziorako neurriak

Itzulpen automatikoaren ebaluazioak bi helburu izan ditzake (Goutte, 2006): ebaluazio erlatiboa, sistema desberdinak konparatzeko, eta ebaluazio absolutua, sistemaren jardueraren neurri absolutua lortzeko.

IAko sistema baten kalitatearen ikuspegi on bat lortzeko, testuinguruan oinarritutako kalitate-eredu baten atributu guztiak hartu beharko lirake kontuan, hasieran atributu bakoitza hoberen neurtzen duen metrikarekin ebaluatu eta gero atributu guztietako emaitzak konbinatuz. Kalitate-eredu-

ko atributuen multzoa taxonomia batean antola daiteke, VIII.1 irudian ikus dezakegun ISLE¹ proiektuan egin zuten bezala (Hovy *et al.*, 2002):

Hala ere, gaur egun IAko sistemen ebaluazioak, batez ere, irteerako testuaren kalitateaz (egokitasunaz eta jariotasunaz) arduratzen dira.

IAko sistemen irteeraren ebaluazioa konplexua da, lengoaia naturala anbigua delako eta bere egitura konplikatua delako. Oso zaila da itzulpen bat noiz den zuzena estimatzea, edo zuzena izatetik hurbil edo urrun dagoen. Guztiz desberdinak diren bi esaldi, baliokideak izan daitezke, eta detaile txiki batean soilik desberdinak diren beste bik esanahi guztiz desberdina izan. Itzulpenen ebaluaziorako eskuzko metodoak zein automatikoak proposatu eta erabili izan dira.

Ondokoak izan daitezke ebaluazio-neurri baten ezaugarri desiragarriak (Hovy *et al.*, 2002): definitzeko erraza, argia eta intuitiboa, giza-epaiekin korrelazio altua duena, prestatzeko eta aplikatzeko merkea, eta ahal bada automatikoa.

Baina, gaur egun, ebaluazio-neurri bat aukeratzeko orduan dilema bat sortzen zaigu (Eisele, 2006):

- Eskuzko ebaluazioa esanguratsua da, baina baita garestia, aspergarria, eta ez-berrerabilgarria ere.
- Ebaluazio automatikoa errepikagarria eta objektiboa da, baina ezin da ziurtatu bere emaitzak zuzenak direnik.

Hurrengo ataletan ebaluazio-neurri desberdinak aztertuko ditugu, dilema horren aurrean gure sistema ebaluatzeko orduan hartu dugun bidea erabakitzeko.

VIII.1.1 Giza-ebaluaziorako neurriak

Itzulpen automatikoaren ebaluazio-irizpiderik ideala erabiltzailearen onarpena litzateke, baina sistemen garapenerako problematikoa da, garestia, motela, aldakorra eta subjektiboa baita (Och, 2005). Gainera, ezin da berrerabili.

Human evaluations of MT are extensive but expensive. (Papineni *et al.*, 2002)

Itzulpenak eskuz ebaluatzeko gehien erabilitako neurriak egokitasuna (*adequacy*) eta jariotasuna (*fluency*) izan dira, ebaluatzaileei eskala batean puntuatzea eskatzen zaiela (King, 1996):

¹*International Standards for Language Engineering.*

ISLE proiektuko taxonomia:

- Erabileraren testuinguruaren ezaugarri nagusien sailkapena:
 - itzulpenaren helburua
 - erabiltzailearen ezaugarriak
 - sarrerako testuen ezaugarriak
- IAko softwarearen kalitate-ezaugarrien sailkapena:
 - Barne ezaugarriak:
 - * itzulpen-estrategia
 - * baliabide linguistikoak
 - * erabiltzeko modua
 - Kanpo ezaugarriak:
 - * funtzionaltasuna (*functionality*):
 - egokitasuna (*suitability/adequacy*),
 - zuzentasuna/jariotasuna (*accuracy/fluency*),
 - elkarreragingarritasuna (*interoperability*),
 - betetzea (*compliance*)
 - segurtasuna (*security*)
 - * fidagarritasuna (*reliability*)
 - * erabilgarritasuna (*usability*)
 - * eraginkortasuna (*efficiency*):
 - denbora
 - baliabideak
 - * mantenigarritasuna (*maintainability*)
 - * eramanagarritasuna (*portability*)
 - * kostua

Irudia VIII.1: Kalitate-eredu bateko atributuak

- Egokitasunak itzulpenak abiapuntu-testuaren esanahi bera adierazten duenentz neurtzen du, eta horretarako ebaluatzaileari abiapuntu-testua edota erreferentzia-itzulpena ematen zaio.
- Jariotasunak, berriz, itzulpena gramatikalki zuzena den ala ez neurtzen du, eta ebaluatzaileari ez zaio abiapuntu-testua edo erreferentzia-itzulpenik ematen. Jariotasuna neurtzen den modura, zenbaitetan ere zuzentasuna, ulergarritasuna eta argitasuna neurtzen dira.

Badira, halere, eskuz ebaluatzeko beste neurri batzuk:

- Irakurgarritasuna, itzulitako testua irakurtzeko erraztasuna adierazten du. Zenbait metrika erabil izan dira:
 - Irakurketa-denbora (Slype, 1979): ebaluatzaile batek IAko sistemak itzulitako dokumentuak eta lengoaia naturaleko testuak irakurtzeko hitzak/minutuko proportzioak neurtu eta konparatzen dira; zenbat eta hurbilago izan biak, hobeagoa izango da itzulpenaren kalitatea.
 - *Cloze test* (Slype, 1979): ebaluatzaileei eskatzen zaie itzulpenean aurretik ezabatutako zenbait hitz asmatzea.
- Postedizio-kostua. Sistemak ematen duen irteera itzulpen onargarri bihurtzeko egin beharreko lana. Neurri honek, zabalkunderako IAko sistemak ebaluatzeko erabiltzen denean, ez du kosturik, berez, postedizioa egin egin behar delako (Boitet *et al.*, 2006). Metrika desberdinak erabil daitezke:
 - Denbora (Dabbadie *et al.*, 2002). Posteditatzeko denbora zenbat eta txikiago izan, sistemaren itzulpenaren kalitate orokorra handiagoa izango da.
 - Tekla-sakatzeen kopurua (Dabbadie *et al.*, 2002). Postediziorako interfazeak zenbatzen dituen tekla-sakatzeen kopurua zenbat eta txikiagoa izan, orduan eta kalitate handiagoa.
 - Edizio-ekintzen kopurua (Boitet *et al.*, 2006). Postediziorako interfazeak zenbatzen ditu mota bakoitzeko edizio posibleak (txertatzea, ezabaketa, ordezen lokala eta ordezen globala), mota bakoitzari koefiziente bat esleitzen zaio, eta horrela kostua kalkulatu da. Metrika hau *Mètèò* sistema ebaluatzeko 80. hamarkadatik erabili izan da.

- Edizio-distantzia (Przybocki *et al.*, 2006; Snover *et al.*, 2006). Postedizioa egin ondoren, sistemaren irteera eta posteditatutakoaren arteko distantzia kalkulatu da. Gure sistema ebaluatzeko erabili dugun metrika hau zehatzago deskribatuko dugu VIII.1.4 atalean.

VIII.1.2 Neurri automatikoak

Neurri automatikoen ebaluazioak merkea, azkarra, errepikagarria eta objektiboa² izan beharko luke eta bere emaitzek giza-ebaluazioekin korrelazio handia izan beharko lukete (Dabbadie *et al.*, 2002).

IAren kalitatea modu automatikoan epaitzeko, sistemaren itzulpena eta giza-itzulpen onak (erreferentzia-itzulpenak) konparatzea izan ohi da bidea, hipotesia izanik irteeraren kalitatea ona izango dela erreferentzia-itzulpenen antzekoa bada, eta txarra oso desberdina bada.

The closer a machine translation is to a professional human translation, the better it is. (Papineni *et al.*, 2002)

Erreferentzia-itzulpen zuzen bakar bat ez da nahikoa ebaluaziorako, itzulpen honetatik oso desberdina den beste itzulpen zuzen bat edo batzuk egon daitezkeelako. Beraz, soluzioa erreferentzia-itzulpen multzo bat erabiltzea izan daiteke (Popescu-Belis *et al.*, 2002). Hala ere, orokorrean, erreferentzia gutxi edukitzen da eskuragarri, gehienetan bat bakarrik (Giménez eta Amigó, 2006). Metrika hauek gizakiek sortutako erreferentzia-itzulpenak behar dituztenez, zenbaitentzat automatikoak baino, erdi-automatikoak dira (Hamon eta Rajman, 2006).

Edozein kasutan, gertutasuna neurtzeko antzekotasun-metrika bat beharrezkoa da. Metrika hauek bi familiatan sailka daitezke: kateen parekatzean (*string matching*) oinarritutakoak, eta n-grametan oinarritutakoak.

Kateen arteko parekatze-teknikak

Kateen arteko parekatze-teknikak sistemaren irteeratik abiatuta erreferentzia-itzulpena lortzeko edizio-eragiketa kopuru minimoaren kalkuluan oinarritzen dira:

- *WER* (Nießen *et al.*, 2000), hitzen errore-proportzioa (*Word Error Rate*). Txertatzea, ezabaketa eta ordezipena dira edizio posibleak. Edizio

²Objektiboak esan nahi du itzulpen berak beti puntuazio bera jasotzen duela.

kopuru minimoa zati erreferentzia-esaldiaren luzera da kalkulatzeko modua. Gogor zigortzen du hitz ordenan desberdintasunak egotea.

WER_g neurria *WER* neurriaren aldaera bat da balio posibleak 0 eta 1 heinean mantentzen dituen. Horretarako edizio kopurua zati edizio eta parekatzeen batuketa kalkulatzen du.

mWER neurriak erreferentzia anitz erabiltzen ditu, *WER* balioa kalkulatzeko erreferentziazko esaldirik antzekoenarekin.

- *PER* (Leusch *et al.*, 2003), posiziotik independentea den errore-proporzioa (*Position-independent Error Rate*). *WER* bezalakoa da, baina parekatzeko orduan ez du kontutan hartzen hitzen ordena.

mPER neurriak erreferentzia anitz erabiltzen ditu, erreferentziazko esaldirik antzekoenarekin lortutako *PER* balioa emanez.

- *TER* (Snover *et al.*, 2006), itzulpenaren edizio-proporzioak (*Translation Edit Rate*). *WER* metrikaren hobekuntza bat da, non edizio posibleak dira txertatzea, ezabaketa, ordezkpena eta hitz multzo baten mugimendua. Erreferentzia anitzak erabiltzen ditu. Erreferentzia antzekoenarekin edizio kopuru minimoa zati erreferentzien batezbesteko luzera kalkulatzen du.

N-grametan oinarritutako teknikak

Informazio-berreskuratze (*Information Retrieval, IR*) arloan erabilitako teknikak dira. Kateen arteko parekatze-teknikak baino korrelazio altuagoa dute giza-epaiekin. Gaur egun, hauexek dira erabilienak IArene ebaluazioan, batez ere *Bleu* metrika.

Teknika hauek n-gramen doitasunean oinarritzen dira, hau da, erreferentzia-itzulpenean agertzen diren sistemaren irteeraren n-gramen proportzioan. Hitzen n-gramen doitasuna n-ren hainbat baliotarako kalkula daiteke eta modu desberdinetan konbinatu:

- *Bleu* (Papineni *et al.*, 2002). *Bleu* (*BiLingual Evaluation Understudy*) neurriak n-gram-en doitasunen (n=1..4) batezbesteko geometrikoa bider laburtasun-isun bat kalkulatzen du.

*Bleu*k moldatutako n-gramen doitasuna (*modified n-gram precision*) kalkulatzen du, hau da, erreferentzia-itzulpenen batean agertzen diren sistemaren itzulpenaren n-gramen kopurua zati sistemaren itzulpenaren hitz kopurua. Erreferentzietako n-grama bat sistemaren itzulpena-

ren n-grama batekin parekatzen denean, ez da berriz erabiltzen. *Bleu* metrikaren heina 0 eta 1 (edo 0 eta 100) tartean dago.

Moldatutako n-gramen doitasunak erreferentziak baino luzeagoak diren sistemaren itzulpenak zigortzen ditu, baina laburragoak direnak ez. Horregatik, laburtasun-isuna (*brevity penalty*) faktore bat erabiltzen da.

Zenbait lanetan *Bleu* metodoa azkarra, merkea eta hizkuntzatik independentea dela eta giza-ebaluazioekin korrelazio handia duela baieztatu da (Papineni *et al.*, 2002; Doddington, 2002).

- *NIST* (Doddington, 2002). *Bleuren* moldaketa den neurri honek n-gramen doitasunen batezbesteko aritmetikoa kalkulatzeko du, luzera isun batekin ere, eta n-gramen doitasunei corpuseko maiztasunaren araberrako pisu bat esleitzen die, maiztasun gutxienerako n-gramei garrantzi handiagoa emateko, informatiboagoak direlakoan.
 - *WNM* (Babych, 2004). *Bleuren* beste moldaketa bat da. N-gramen doitasunei pisu bat esleitzen die, n-gramen maiztasunaren arabera.
 - *F-measure* (Melamed *et al.*, 2003). Doitasun (*precision*) eta estalduraren (*recall*) batezbesteko harmonikoa da. IAre irteera eta erreferentziaren arteko parekatze maximo (*maximum matching*) bat bilatzen du, n-grama luzeen parekatzeak hobesten dituenak. Parekatze maximo-ko parekatutako hitzen kopurua zati itzulpenaren luzerarekin kalkulatzeko da doitasuna. Estaldura, berriz, zati erreferentziaren luzera.
- F-measure* neurriak *Bleu* baino korrelazio handiagoa du (Melamed *et al.*, 2003), baina ez beti, batez ere segmentu txikientzat (Blatz *et al.*, 2003). Horrez gain, doitasuna eta estaldura ondo uler daitezkeen neurriak dira, interpretazio intuitiboa dutenak.
- *Meteor* (Lavie *et al.*, 2004). Neurri honek *F-measure* hobetzen du, prozesaketa linguistikoa egiten duelako. Lematizatutako hitzak parekatzen ditu, formak parekatu beharrean, eta *Wordneten* sinonimia begira dezake. Gainera, batezbesteko harmonikoan pisu handiagoa ematen dio estaldurari, modu honetan giza-epaiekin korrelazioa hobetzen delako.

Beste motako teknikak

Erreferentzia-itzulpenak erabiltzen ez dituzten beste hainbat neurri automatiko ere proposatu izan dira:

- Entitate izendunen (*named entity*) itzulpena (Reeder *et al.*, 2001). IA-ko sistemaren irteerako entitate izendunak (pertsona-, leku- eta erakunde-izenak, datak, moneta-balioak,...) konparatzen dira erreferentziatzeko corpus batean aurretik identifikatutakoekin.
- *X-Score* (Hamon eta Rajman, 2006). Oraindik esperimentala den metrika honek gramatikaltasuna neurtzen du, testuan dagoen oinarritzko informazio-linguistikoaren banaketa corpus bateko informazio beraren banaketarekin konparatuz. Xede-hizkuntzako corpus bat analizatzen da, esaldiko osagaien mendekotasun sintaktiko mota bakoitzaren maiztasunak lortuz. Mendekotasun mota batzuk erabiltzen dituen formula erabiliz *X-Score* kalkulatzeko da, frogatu delarik formula horrek ondo iragartzen duela gizakiek ebaluatutako jariotasun-emaiztasun sailkapena. Ondoren *X-Score* hizkuntzaren adierazpena sistemak emandako itzulpenari aplikatzen zaio.
- Itzuli gabeko hitzen kopurua (Dabbadie *et al.*, 2002). Metrika honek itzulpenaren kalitatea hiztegiaren tamainarekin lotzen du. Test honen emaitza soilik ez dute balio IAko sistema baten irteeraren kalitatea neurtzeko.
- Ikasketa automatikoko teknikak. Giza-epaiekin korrelazio altua duen metrika bat datuetatik ikasteko saiakerak egiten hasi dira (Kulesza eta Shieber, 2004).
- *IQ* (Giménez eta Amigó, 2006), metrika desberdinen konbinazioa. Ebaluazio automatikoko metrikek ez dute ondo harrapatzen IAko sistemak ezagutza linguistikoa gehitzean lortutako hobekuntza, gehienek kontutuan hartzen duten informazio linguistikoa soilik maila lexikalekoa delako. Alderdi linguistiko guztiak kontutuan hartuko lukeen metrika bat diseinatzea oso zaila denez, (Giménez eta Amigó, 2006) lanean IAren kalitatearen aspektu partzialak ebaluatzen dituzten metrika desberdinak erabiltzen dituzte, haien irteera *IQ* izeneko neurri bakar batean konbinatuz.

VIII.1.3 Neurri automatikoen argiak eta itzalak

Neurri automatikoen artean n-gramen doitasunean oinarritutakoak izan dute arrakasta handiena, giza-epaiekin korrelazio altuena dutelako. Ebaluazio merkea, azkarra, errepikagarria eta objektiboa ematen dutenez, IAren ikerketa eta garapenaren zikloa azkar dezakete.

Haien artean *Bleu* da, gaur egun, gehien erabilitako ebaluazio neurria. Metrika honek, azken urteetan IAren aurrerapena gidatu du, batez ere IA estatistikoaren garapenean, *Bleuren* emaitzetan oinarritzen baita sistemetan aldaketa inkrementalen ebaluazioa eta sistemen optimizazioa, errore proportzio minimorantz entrenatuz (Och, 2003).

Love them or hate them, Bleu and Nist are the metrics that are currently most widely used. (Goutte, 2006)

(Callison-Burch *et al.*, 2006) lanean esaten da IAren komunitatea gehiegi fidatzen dela *Bleu* metrikari: kongresuetako artikuluek itzulpen-kalitatean hobekuntzak aurkeztu dituzte, *Bleu* emaitza hobeagoen berri emanez, baina itzulpen-adibide bat ere erakutsi gabe; edo sistemak konparatzen dira *Bleu* erabiliz, sailkapen hauek eskuzko ebaluazioen bidez egiaztatu gabe. Erabilera hauek guztiak egiten dira giza-epaiekin korrelazio altua dutela³ suposatzen delako, zenbaitetan frogatu den bezala (Papineni *et al.*, 2002; Doddington, 2002).

Hala ere, zalantzak hasi dira sortzen *Bleu* eta beste n-grametan oinarritutako metriken inguruan. Alde batetik, IAko sistema desberdinen kalitate erlatiboa konparatzeko baliagarria bada ere, *Bleu* adierazten duena interpretatzea oso zaila da.

What does a BLEU score of 0,016 mean? (Turian *et al.*, 2003)

Bestetik, (Callison-Burch *et al.*, 2006) lanean frogatu dute *Bleuren* hobekuntza bat ez dela nahikoa itzulpen kalitatearen hobekuntza islatzeko, eta ez dela beharrezkoa *Bleu* hobetzea itzulpen kalitatearen hobekuntza nabarmenak lortzeko. Horren arrazoia *Bleuren* ahulezietan datza: *Bleu* puntuazio bera duten milioika hipotesi posible daude, *Bleu* puntuazio bera lortzen duten arren guztiz desberdinak direnak; gainera, *Bleuk* ez ditu sinonimoak edo aldaera lexikalak parekatzen eta, beraz, erreferentzietan agertzen ez diren hipotesiko hitzak edozein hitzarekin ordezkari daitezke, *Bleu* puntuazioan ezer eragin gabe.

Sistemaren itzulpenak erreferentzia-itzulpenekin korrelazio altua izateak kalitate hobeagoa adierazten badu ere, teoriarik ezin dugu bermatu korrelazio baxua izateak kalitate txarra adierazten duela, ez bada itzulpen zuzen posible guztiak eskuragarri daudela (de Gispert Ramis, 2006).

³Korrelazioak erabiltzean kontuz ibili behar da. (Turian *et al.*, 2003) lanean aurkeztutako esperimentuen datuek erakusten dute, IAko sistemen kalitatearen giza-epaien arteko korrelazioa baxua dela. Beraz, IAko sistemen kalitatearen neurri automatikoen eta giza-epaien arteko korrelazioa ere baxua izan beharko da.

Zenbait lanetan (Callison-Burch *et al.*, 2006; Koehn eta Monz, 2006) frogatu dute *Bleu* ez duela giza-epaiekin uste zen bezain korrelazio altua. Adibidez, 2005eko NIST ebaluazio-kanpainaren datuetatik (Lee eta Przybocki, 2005) hauek azpimarra ditzakegu:

- Giza-ebaluazioko sailkapenean 1. posizioan gelditutako sistema, *Bleu*-rekin 6. sailkatu zen.
- IA estatistikoko sistemek *Bleu* puntuazio altuagoa jasotzen dute, n-grametan oinarritzen ez diren sistemek baino, erregeletan oinarritutako sistemek puntuazio oso baxuak lortzen dituztela⁴.
- Iako sistema onenen itzulpenak gizakien itzulpen-kalitatetik oso urrun badaude ere, bere puntuazioak giza-itzulpenetatik oso hurbil daude, eta urte gutxi barru Iako sistemen itzulpenek giza-itzulpenak baino puntuazio hobeagoak jasoko dituzte metrika hauekin (Giménez *et al.*, 2005).

(Boitet *et al.*, 2006) lanean oso kritikoak dira *Bleu* bezalako n-grametan oinarritutako neurriekin, desegokiak eta garestiak direla baieztatuz, nagusiki bi arrazoiengatik:

- Uste denaren kontra, ez dute itzulpenaren kalitatea neurtzen, erreferentzia-itzulpenetako n-gramen antzekotasuna baizik.
- Erreferentzia-itzulpenak prestatzearen kostua oso handia da. Egia da gero behin eta berriz erabili daitekeela inolako kosturik gabe, baina corpus berri baten gainean ebaluatu nahi bada, berriz sortu behar dira erreferentziak eskuz. Horrek eramaten ditu garatzaileak sistemen hobekuntzak corpus jakin baterako soilik egitera.

Bestalde, garrantzitsuena litzateke *Bleuren* zein erabilerak diren egokiak (edo bestelako n-grametan oinarritutako ebaluazio-metrikenak), eta zeintzuk ez bereiztea (Callison-Burch *et al.*, 2006):

- Erabilera egokiak:

⁴Gainera, corpus jakin batean entrenatutako sistema estatistiko bat erregeletan oinarritutako sistema batekin konparatu nahi bada, corpus horretako testuak itzultzen, beharrezkoa litzateke erregeletan oinarritutako sistema corpus horretarako afinatzea (*tuning*) eta orain arte burututako ebaluazio gehienetan hori ez da egin. Afinatze hori hiztegiekin egin daiteke (adibidez sistemarentzat ezezagunak diren corpus horretako hitzen hiztegi berezitu bat sortuz) eta agian gramatikekin ere. Beste zenbait parametro ere findu daitezke: hiztegi desberdinen lehentasuna, letra larrien tratamendua...

- Sistema baten aldaketa inkrementalen jarraipena.
 - Estrategia antzekoak erabiltzen dituzten sistemen konparatzea.
 - IA estatistikoaren parametroen balioen optimizazioa.
- Erabilera desegokiak:
 - Estrategia desberdinak erabiltzen dituzten sistemak konparatzea.
 - Sistemen puntuazioak konparatzea hizkuntza bikotea, erreferentzia kopurua edo n-gramen tamaina desberdina denean.
 - Metrikak ondo modelatzen ez dituen itzulpenaren aspektuen hobekuntzak identifikatzen saiatzea.
 - Test-corpusean gutxitan agertzen diren hobekuntzak monitorizatzea.

VIII.1.4 Edizio-distantzia

Egokitasun eta jariotasunari buruzko giza-epaiak, garestia izateaz gain, subjektiboak dira eta sistema baten aurrerapena neurtzeko ez dira nahiko zehatzak. Gainera, oso informazio erabilgarri gutxi ematen dute IAren irteeran gaizki dagoenari buruz.

Bestalde, *Bleu* bezalako neurri automatikoak ez dira efektiboak IAko sistemen itzulpenaren kalitate absolutua neurtzen, edo erabiltzaile batentzat itzulpenak zenbateraino diren baliagarriak erakusten. Estrategia desberdinetako sistemak konparatzeko ere ez dute balio.

AEBetako NIST⁵ erakundeak hainbat urtetan IAko sistemen ebaluazioak koordinatu ditu, ebaluazio neurri automatiko eta errepikagarriak erabiliz. DARPA⁶ agentziaren GALE⁷ programaren barruan, n-grametan oinarritutako neurrien eragozpenei aurre egiteko asmoarekin, ikerketa militarretan erabilgarria dela frogatu den metrika berri bat erabiltzea aztertzen ari dira: edizio-distantzia.

Edizio-distantzia modu honetan definitzen da (Przybocki *et al.*, 2006):

Giza-editore batek IAko sistema baten itzulpenean eginbeharreko moldaketa kopurua, editatutako bertsioak abiapuntuko testuaren esanahi osoa izan dezan idazkera ulergarrian.

⁵National Institute of Standards and Technology. www.nist.gov

⁶Defense Advanced Research Projects Agency. www.darpa.mil

⁷Global Autonomous Language Exploitation

Moldaketa posibleak dira banakako hitzen txertatzea, ezabaketa, ordezena eta hitz multzoen mugitzea. Bi hitz txertatzea (edo ezabatzea edo ordeztea) bi edizio dira eta hainbat hitz jarraituren mugitzea, berriz, edizio bakar bat. Puntuazio-ikurrak hitz arrunten modura tratatzen dira. Edizio-distantziak edizio kopurua zati moldatutako itzulpenaren hitzen kopurua kalkulatzen du.

Edizio-distantzia (Snover *et al.*, 2006) lanean aurkeztutako *HTER*⁸ (*Human-targeted Translation Edit Rate*) neurrian oinarritzen da.

Ebaluaziorako neurri intuitibo hau sistemaren itzulpenen kalitatea modu errealistan neurtzen du, itzulpenak zenbateraino diren baliagarriak erakutsiz (Goutte, 2006).

(Przybocki *et al.*, 2006; Snover *et al.*, 2006) lanetan aurkeztutako esperimentuetatik edizio-distantziari buruzko ondorio interesgarriak lortzen dira. Edizio-distantziak *Bleu* eta *Meteor* metrikek baino korrelazio altuagoa du giza-epaiekin. Gainera, korrelazio altuagoa izaten da giza-epai baten eta edizio-distantziaren artean, bi giza-epairen artean baino. Beraz, IArene irteeraren kalitatea neurtzeko gizakiak erabili behar badira, badirudi hobe dela gizakiak edizio-distantzia kalkulatzeko irteera editatzen aritzea, epai subjektiboak egiten baino.

Postedizio horrek duen kostua kontutan hartzekoa da: 500 eta 800 bitarteko hitz/orduko. Baina ez da beharrezkoa ebaluazio bakoitzean erreferentzia guztiak berriz sortzea; sistema gehienetarako ebaluaziotik ebaluaziora itzulpen gehienak ez direlako aldatzen, eta soilik aldatu diren itzulpenetarako sortu beharko dira editatutako erreferentzia berriak. Zabalkunderako erabiltzen diren sistemak ebaluatzeko, metrika honen kostua minimoa da, berez postedizioa egin behar delako. NIST erakundeak antolatzen dituen ebaluazio-kanpainetarako, berriz, postedizioaren kostua handia da.

Mota honetako metodoek etorkizun laburrean onarpen handia izango dutela planteatzen da (de Gispert Ramis, 2006) lanean, oinarri sendoetan ezarritako kalitatearen emaitza absolutua emateaz gain, erreferentzia-itzulpen berriak sortzen dituztelako, itzulpen-erroreak automatikoki identifikatzeko eta sailkatzeko baliagarriak izango direnak.

Zenbaketa automatizatzeak metrika honen erabileran izugarri laguntzen du. Gainera, (Snover *et al.*, 2006) lanean garatutako softwarea publikoki eskuragarri dago⁹.

VIII.2 irudian ikus dezakegu adibide bat: “*no coma alimentos con muchas calorías.*” esaldirako sistemak ematen duen itzulpena (*Hyp*) eta postedita-

⁸*HTER TER* neurriaren aldaera bat da.

⁹<http://www.cs.umd.edu/~snover/tercom/>

```

Ref: ez itzazu jan elikagaiak kaloria askorekin .
Hyp: elikagaiak ez ditzan jan kaloria askorekin .

REF:   ez ITZAZU jan   elikagaiak   kaloria askorekin .
HYP:  @ ez DITZAN jan [ elikagaiak ] kaloria askorekin .
EVAL:      S
SHFT: 1              1              1
TER Score: 28.57 (2/7)

```

Irudia VIII.2: Edizio-distantziaren kalkuluaren traza.

tutakoaren (*Ref*) arteko edizio-distantzia kalkulatzeko programaren traza. Hitz batean ordezipena (*S*) egin da (*ditzan* $\rightarrow$ *itzazu*), eta hitz bat (*elikagaiak*) lekuz mugitu da (*SHFT*). Beraz, 7 hitzetatik (puntu ere kontatzen da) 2 edizio burutu dira, edizio-distantziaren emaitza 28.57 izanik.

VIII.2 *Matxin* sistemaren ebaluazioa

Gure sistema ebaluatzeko sistemaren irteeraren kalitateari begiratuko diogu. Edizio-distantziaren metrika erabili dugu, eskuzko lana behar badu ere, lortutako emaitzek sistemak emandako irteeraren erabilgarritasunaren berri ematen digulako.

Gaur egun duen nagusitasunagatik *Bleuren* emaitzak ere kalkulatu ditugu, ez sistemaren kalitateari buruz hausnartzeko, baizik eta *Bleuren* emaitzen baliagarritasunari buruz hausnartzeko.

Matxin sistemak emandako itzulpen-adibideak A eranskinean ikus daitezke.

VIII.2.1 Ebaluazioaren diseinua

Ebaluazioa edizio-distantzia neurriaren kalkuluan oinarritu dugu, eta zenbait diseinu-erabaki hartu behar izan ditugu ebaluazioaren inguruko elementuei buruz: erabiliko diren esaldi-kopurua eta zein corpusetatik jasoko diren, editore-kopurua eta bere profila, postediziorako ingurunea eta jarraibideak, eta edizio-distantzia kalkulatzeko modua.

Testu-motak eta tamainak

Ebaluazioa burutzeko testuak bi corpus desberdinetatik jaso ditugu

- *Eitb*: hizkuntza orokorreko kazetaritza-corpusa. Euskal Irrati Telebistaren (EITB) albistegietako berriak biltzen ditu.

	<i>Eitb</i>	<i>Cons</i>
Esaldi kopurua	50	50
Token kopurua	682	741
Hitz kopurua	612	626
Ikur kopurua	70	115

Taula VIII.1: Ebaluaziorako erabilitako jatorrizko testuen datuak

- *Consumer*: kontsumoaren arloko testuen corpora. Eroski enpresaren *Consumer* aldizakarian argitaratutako artikuluak biltzen ditu.

Bi corpus horietako bakoitzetik 5 eta 25 bitarteko hitz kopurua duten 50 esaldi aukeratu dira ausaz ((Snover *et al.*, 2006) laneko esperimenduak ere 100 esaldiekin egin dituzte). Esaldi horietako datuak VIII.1 taulan ikus daitezke, non token kopurua hitzen eta ikurren kopuruaren batura den, edizio-distantziak token guztiak hartzen dituelako kontutan.

Editorea

Edizio-distantziaren emaitzak kalkulatzeko postedizio-lanak editore elebidun bakar batek egin ditu.

Postedizioa egiteak duen kostuagatik editore bakar bat erabili dugu, baina gure asmoa aurreragoko esperimenduetan editore bat baino gehiago erabiltzea da, datuen fidagarritasuna handitzeko eta editoreen arteko emaitzen koherentzia aztertzeke. Hala ere, (Snover *et al.*, 2006) lanean baieztatzen da praktikan editore bakar bat nahikoa izan daitekeela.

Editore elebidunak egin du postedizio-lana, IAKo sistemaren irteera editatzeko jatorrizko testua erabili ahal izateko. Nazioarteko ebaluazio kanpaina gehienetan, giza-ebaluazioa egiten denean, ebaluatzaile elebazarrek egin ohi dute. Itzulpenaren zuzentasuna edo egokitasuna neurtzeko edo posteditatzeko, erreferentzia-itzulpen bat ematen zaio bere epaia edo edizioa burutzeko. Hau, ebaluazioan, erroreen iturria izan daiteke, erreferentzia-itzulpenean erroreak badaude edo informazioa galdu bada.

Editorea testuen zuzenketa-lanetan aritua izan arren, hauxe izan da bere lehenengo esperientzia IAKo sistema baten irteera editatzen.

Postediziorako ingurunea

Postedizio lanetarako ez dugu interfaze berezirik erabili. Etorkizuneko ebaluazioetarako garatutako interfaze bat¹⁰ erabiltzeko edo guretzako egokia

¹⁰NISTek postediziorako garatutako interfazea Interneten eskuragarri dago: <http://www.nist.gov/speech/tests/gale/2007/software.htm>

* 1 (x)	*****
ES	Holanda aconseja recluir a las aves de corral desde marzo.
EU	Holandak sartzea aholkatzen die martxotik korralearen hegaztiei.
PE	Holandak martxotik korraleko hegaztiak sartzea aholkatzen du.
* 2 (x)	*****
ES	Detenido en Madrid un ex policía argentino acusado de tortura.
EU	Madrilen polizia ohi bat torturaren argentinak akusatua geldituta.
PE	Madrilen polizia argentinak ohi bat atxilotuta torturaz akusatua.
* 3 ()	*****
ES	Heras apelará la suspensión de dos años por dopaje.
EU	Heras-ek 2 urtez zintzilikatzea apelatuko du dopaje-agatik.
PE	Heras-ek 2 urtez zintzilikatzea apelatuko du dopaje-agatik.

Irudia VIII.3: Postediziorako testu-fitxategia. Adibidea

den bat garatzeko asmoa dugu.

Ebaluazio honetarako testu-editore sinple bat erabili dugu. Testu-fitxategi bat prestatu dugu non itzultako esaldi bakoitzerako ondokoak agertzen diren:

- Jatorrizko esaldia (ES).
- Sistemak itzultako esaldia (EU).
- Sistemak itzultako esaldia, bere gainean moldaketak egiteko (PE).

Testu-fitxategi honen formatuaren adibidea VIII.3 irudian ikus dezakegu, non editoreak dagoeneko 2. esaldia moldatzen bukatu duen.

Editoreak paperean zein sarean erabilgarri dauden hainbat hiztegi kontsulta ditzake.

Postediziorako jarraibideak

Postediziorako jarraibideak NIST erakundeak GALE ebaluazio-programarako prestatutakoetan¹¹ oinarritzen dira, gure beharretarako egokitu ditugula,

¹¹<http://www.nist.gov/speech/tests/gale/2007/doc.htm>

nagusiki gurean postedizio elebiduna egin delako.

Editoreak itzulpen automatikoko sistemaren irteera moldatu behar du abiapuntu-testuaren esanahi osoa edukitzeko eta ulergarria eta gramatikalki zuzena izateko, ahalik eta aldaketa gutxien egiten.

Edizio-aldaketa kopurua minimizatzea helburu bat bada ere, lehentasuna du moldatutako itzulpenak esanahi zuzena edukitzeak, eta ulergarria eta gramatikalki zuzena izateak.

Sistemaren irteera moldatu behar da ondoko lau arauak betez:

1. Abiapuntu-testuaren *esanahi bera* izatea.
2. *Ulergarria* izatea.
3. *Gramatikalki zuzena* izatea.
4. *Ahalik eta edizio-aldaketa gutxien* egitea.

Aurreko arauak bete behar dira ondoko oharrek kontutan hartuz:

- Ez da aldatu behar zuzena den ezer, soilik beste aukera bat gogokoagoa delako.
- Sintagmen ordena aldatu behar da soilik esanahia zuzentzeko edo ulergarria izateko beharrezkoa bada.
Adibidez, *El amigo ha llevado la manzana* esaldirako *Lagunak sagarra eraman du* eta *Eraman du lagunak sagarra* onargarriak dira. Baina *Lleva el libro y vuelve a casa* esaldirako *Liburua etxera darama eta itzultzen da* ez da onargarria.
- Izenen modifikatzaile diren elementuak, nahiz eta itzulpenean funtzio horrekin ez ageri, esaldia ulergarria bada onartzen dira.
Adibidez, *Come alimentos con muchas calorías* esaldirako *Jaten ditu elikagaiak kaloria askorekin*
- Sinonimoak onargarriak dira, esanahia zuzena bada.
- Abiapuntuko esaldia anbigua denean, aukera posible guztiak onargarriak dira.
Adibidez, *Trae pan de la tienda* esaldirako *Dendako ogia dakar* eta *Ogia dakar dendatik*.
- Informazioa falta denean esanahia betetzen duen sintagma motzena (edo hitz soila) gehitu behar da.

- Puntuazio-ikurrek, datek, zenbakiek eta abarrek euskararen arauak jarraitu behar dituzte.
- Akronimoak (edo laburdurak) eta beren forma hedatuak baliokideak dira.
Adibidez, *EHU* eta *Euskal Herriko Unibertsitatea*.
- Zenbakien idazkera digituz edo hitzez baliokideak dira
Adibidez, *2007* eta *bi mila eta zazpi*.
- Sistemak ezagutzen ez dituen hitzak deklinatzean lema eta atzizkiaren artean jartzen duen marratxoak ontzat emango da.
Adibidez, *Zigortu dute dopaje-agatik* edo *Putin-ek esan zuen*.

Azpinarratu behar dugu edizio-distantzia kalkulatzeko eginbeharreko posteditzioaren helburua ez dela itzulpenak argitaratze-mailako kalitatea lortzea, nahikoa izanik abiapuntu-testuaren esanahi osoa edukitzea eta ulergarria eta gramatikalki zuzena izatea.

Edizio-distantziaren kalkulua: hitzetan edo hitz-erro eta atzizkietan oinarrituta

Edizio-distantzia neurria kalkulatzeko (Snover *et al.*, 2006) lanean garatutako softwarea erabili dugu, Interneten eskuragarri dagoena¹².

Kalkulua bi modutan egin dugu:

- Hitzetan oinarrituta. Edizioak zenbatzeko hitz osoei begiratzten zaie (hitzak eta puntuazio ikurrak).
- Hitz-erro eta atzizkietan oinarrituta. Honetarako sistemaren itzulpenetan eta posteditatutako esaldietan erro-bilaketa (*stemming*) egiten da, hitz bakoitzean erroa eta atzizkia banatzeko. Modu honetan lortzen den emaitza informatiboagoa da. Gainera, gure sistemaren emaitzak hizkuntza ez eranskarietara itzultzen duten sistemen ebaluazio-emaitzekin konparatzeko egokiagoa da, euskarazko atzizki gehienen beste hizkuntza horietako baliokideak hitzak izaten direlako (adibidez, preposizioak).

Ikus dezagun adibide batez bi modutan egindako kalkuluen arteko diferentzia. Horretarako “*ellos van a los cursos*.” jatorrizko esaldiaren hainbat itzulpen posibleek lortzen duten edizio-distantziaren emaitza kalkulatu

¹²<http://www.cs.umd.edu/~snover/tercom/>

Sistemaren itzulpena	Posteditatua	Ed-dist
haiek ikastaroetara doaz .	haiek ikastaroetara doaz .	0.00 (0/4)
haiek IKASTAROEI doaz .	haiek IKASTAROEETARA doaz .	25.00 (1/4)
haiek BIDEETARA doaz .	haiek IKASTAROEETARA doaz .	25.00 (1/4)
haiek BIDEEI doaz .	haiek IKASTAROEETARA doaz .	25.00 (1/4)

Taula VIII.2: Hitzetan oinarritutako edizio-distantzia. Adibideak

Sistemaren itzulpena	Posteditatua	Ed-dist
hai ek ikastaro etaera doaz .	hai ek ikastaro etaera doaz .	0.00 (0/6)
hai ek ikastaro EI doaz .	hai ek ikastaro ETAERA doaz .	16.67 (1/6)
hai ek BIDE etaera doaz .	hai ek IKASTARO etaera doaz .	16.67 (1/6)
hai ek BIDE EI doaz .	hai ek IKASTARO ETAERA doaz .	33.33 (2/6)

Taula VIII.3: Hitz-erro eta atzizkietan oinarritutako edizio-distantzia. Adibideak

dugu, hitzetan zein hitz-erro eta atzizkietan oinarrituta. Demagun itzulpen posible hauek ematen direla: *haiek ikastaroetara doaz. / haiek ikastaroei doaz. / haiek bideetara doaz. / haiek bideei doaz.*

Itzulpen guzti horietarako posteditzioaren emaitza “*haiek ikastaroetara doaz.*” da.

Hitzetan oinarritutako edizio-distantziaren kalkulua VIII.2 taulan ikus daiteke. Lehenengo itzulpenean ez da moldaketarik egin, eta besteetan egin-dako moldaketek ematen duten edizio-distantziaren emaitza berdina dute, nahiz eta azkeneko aukera beste biak baino okerragoa izan.

Hitz-erroetan eta atzizkietan oinarritutako edizio distantzia kalkulatzeko da, berriz, emandako itzulpenetako eta moldatutako erreferentzietako hitzetan erroa eta atzizkia banatu ondoren (ikus VIII.3 taula). Ikus dezakegu azkeneko aukera gehiago zigortzen dela, ez baitu asmatu ez hitz-erroaren itzulpenean, ezta esleitutako postposizio-atzizkian. Itzulpenaren kalitatearen ebaluazio zehatzagoa ematen duela esan dezakegu.

VIII.2.2 *Matxin* sistemaren emaitzak

Aurreko atalean diseinatutako ebaluazioa *Matxin* sistemarentzat burutu dugu, eta emaitzak VIII.4 taulan ikus ditzakegu.

Bertan hitzetan eta hitz-erro eta atzizkietan oinarritutako emaitzak agertzen dira, erabilitako bi corpusetako 50na esaldientzat. Edizio-distantziaren emaitza (*Ed-dist*) edizio-kopurua (*Ed*) zati token-kopurua (*Tok*) kalkulatzeko da. Edizio kopurua edizio-mota guztien batuketa da: txertatzea (*Tx*), ezabaketa (*Ez*), ordezketa (*Ord*) eta mugimendua (*Mug*). Taulan ere ematen da mugimenduetan mugitutako token-kopurua (*MgT*).

		Tx	Ez	Ord	Mug	MgT	Ed	Tok	Ed-dist	
Hitzekin	Eitb	20	13	147	35	47	215	532	40.41	42.00
	Cons	27	30	152	50	56	259	594	43.60	
Hitz-erro eta atzizkiekin	Eitb	37	21	136	60	93	254	727	34.94	37.06
	Cons	50	50	149	66	87	315	804	39.18	

Taula VIII.4: *Matxin* sistemaren emaitzak

Egindako edizio gehienak, diferentzia handiz, ordezenak dira eta ondoren mugimenduak.

Hitzetan oinarritutako emaitza 42.00koa da, hau da, 100 tokenetatik (hitzak eta puntuazio ikurrak) 42 edizio burutu direla.

Hitz-erro eta atzizkietan oinarritutakoa, berriz, 37.06koa da, hau da, 100 tokenetatik (hitz-erro, atzizki eta puntuazio ikurrak) 37 edizio burutu direla. Modu honetan puntuazio hobeagoa lortzen da, zenbaitetan hitz-erroan errorerik badago, atzizkia zuzena delako, edo alderantziz.

Eitb corpus periodistikoaren adibideekin emaitzak hobeagoak dira, *Cons* corpuseko esaldietan domeinu zehatz bateko egitura sintaktiko eta terminologia berezitu gehiago agertzen delako.

Matxin sistemarekin lortutako emaitzen interpretazio sinplifikatua hau da: sistemak ematen duen itzulpena zuzentzeko editore batek 10 hitzetatik 4 aldatu behar ditu.

Emaitza hauek ez dira oso onak, noski, gure sistemaren garapena hasierako urratsetan baitago. Hala ere, emaitzen balorazioa egiteko interesgarria da beste sistemek edizio-distantzia neurriarekin lortutako emaitzekin konparatzea. (Snover *et al.*, 2006; Przybocki *et al.*, 2006) lanetan NIST MTEval04 kanpainan arabiera-ingeles atazarako aurkeztutako hainbat sistemen irteeraren edizio-distantzia kalkulatu dute. Kalitate onenetako sistemen emaitzak %26ra iristen dira, eta kalitate txarrenetakoak %47an gelditzen dira. Beraz, gure sistemaren emaitza ebaluazio-kanpaina horretan aurkeztutako sistemen kalitateen erdibidean kokatzen da.

Matxin sistemarekin lortutako zenbait itzulpen-adibide A eranskinean ikus daitezke.

VIII.2.3 Bi sistemen konparazioa: *Matxin* vs *Matrex*

Opari polita izan da guretzat espainieratik euskarara itzultzen duen beste sistema bat ebaluatu ahal izatea, *Matxin* lortzen duen emaitza baloratu ahal izateko. Dublinen garatutako *Matrex* corpusetan oinarritutako estrategiak erabiltzen dituen sistema (ikus II.2.2 atala), IXA taldearekin elkar lanean, ingelesetik eta espainieratik euskarara itzultzeko egokitu izan da, IXA tal-

		Tx	Ez	Ord	Mug	MgT	Ed	Tok	Ed-dist	
Hitzekin	Eitb	35	101	205	27	28	368	512	71.87	64,92
	Cons	47	55	187	60	71	349	602	57.97	
Hitz-erro eta atzizkiekin	Eitb	32	173	289	60	68	554	726	76.31	62.96
	Cons	46	112	145	84	113	387	780	49.61	

Taula VIII.5: *Matrex* sistemaren emaitzak

dean garatutako euskararako hainbat tresna erabiliz. Sistema entrenatu da *Consumer* aldizkariko corpuseko 50.000 esaldiekin (975.000 hitz espainieraz eta 785.000 euskaraz), esaldi kopuru hau txikia izanik konparatzen badugu, adibidez, oso erabilia den Europako Parlamentuaren aktak dituen 1.000.000 esaldietako corpusarekin edo (Och, 2006) lanean emaitza optimoak lortzeko proposatzen diren kopuruekin: hizkuntzaren ereduak ehun bilioika hitz eta itzulpenaren ereduak ehun milioika hitz. Tamaina horretako corpusak euskararako ez daude gaur egun, epe erdira ere ez dira egongo, eta ziur aski inoiz ez dira izango.

Matrex sistemaren emaitzak

Matrex sistemarako edizio-distantzia kalkulatzeko postedizioa *Matxin* sistemaren ebaluaziorako egindako baldintza berdinetan burutu da: editore bera, jarraibide berdinak eta esaldi berdinen itzulpenak.

Edizio-distantziaren emaitzak VIII.5 taulan ikus ditzakegu. Logikoa denez, *Consumer* corpuseko esaldietarako emaitza askoz hobe da, horixe baita entrenatu den corpora, domeinukoa.

Bestalde, *Consumer* corpuseko esaldietarako emaitza hitz-erro eta atzizkietan oinarritutako emaitza hobeagoa bada (57.97 vs. 49.61), ez da berdina gertatzen *Eitb* corpuseko esaldiekin (71.87 vs. 76.31), non emaitzak okerragoak diren egindako errore kopuru handia proportzioan gehiago handitzen delako hitz-erro eta atzizkietan banatzerakoan.

Azpimarratzeko beste datu bat da egindako edizio gehienak ordezipenak eta ezabaketa direla (Snover *et al.*, 2006) lanetako esperimentuetan bezala, non IA estatistikoko sistemak ebaluatu diren. Honek adieraz dezake IA estatistikoko sistemen erroreak zuzentzeko edizio mota erabilienak bi horiek direla. Aurreko atalean ikusi dugun bezala *Matxin* erregeletan oinarritutako sistemarentzat, berriz, edizio-mota erabilienak ordezipena eta mugimendua dira.

	<i>Hitzekin</i>		<i>Erro/atzizkiekin</i>	
	<i>Matxin</i>	<i>Matrex</i>	<i>Matxin</i>	<i>Matrex</i>
Eitb	40.41	71.87	34.94	76.31
Cons	43.60	57.97	39.18	49.61

Taula VIII.6: *Matxin* vs *Matrex**Matrex* eta *Matxin* sistemen emaitzen konparazioa

Bi sistemen emaitzak VIII.6 taulan bildu ditugu, konparazioa errazago egiteko. *Eitb*ko corpuseko esaldientzat *Matxin* sistema *Matrex* baino askoz hobeagoa da, hitz-erro eta atzizkietan oinarritutako emaitzetan *Matxin*en edizio kopurua *Matrex*enaren erdia baino txikiagoa izanik (34.94 vs. 76.31). *Consumer* corpuseko esaldietan, *Matxin*en irteera ere hobeagoa da, nahiz eta diferentzia hain handia ez izan (39.18 vs. 49.61). Esan beharra dago, gainera, *Matrex* sistema *Consumer* corpusarekin entrenatu den bezala, *Matxin* sistemarentzat ere corpus horretarako fintze-lana egin ahalko litzatekeela, bere emaitzak hobetuz.

Konparazio honetatik atera beharreko lehengo ondorioa da *Matxin* erregeletan oinarritutako sistema metodo estatistikoak erabiltzen duen *Matrex* sistema baino hobeagoa dela, diferentzia handiarekin. Honek *Matxin* sistemaren emaitzak berrirakurtzera eramaten gaitu, edizio-distantziaren 40ko emaitza hain txarra ez dela ondorioztatuz.

Edozein kasutan, *Matrex* sistemaren irteeraren kalitatea hobe daiteke, corpus handiagoekin entrenatuz eta hainbat doikuntza eginez. Ikustekoa da zenbateko hobekuntza lor daitekeen.

Etorkizuneko lanetarako uzten dugu, bi sistemen irteerak kualitatiboki konparatzea, sistema bakoitzak zertan asmatzen duen eta zertan kale egiten duen aztertuz, eta hibridaziorako aukerak ikertzea.

VIII.2.4 *Bleu* neurriaren baliagarritasunari buruzko hausnarketa

Matxin eta *Matrex* sistemak *Bleu*ekin neurtu ditugu, lortutako emaitzek *Bleu* neurriaren baliagarritasuna aztertzeko aukera ematen digulako.

Ebaluazioa *Eitb* eta *Consumer* corpusetako esaldiekin burutu da, corpus bakoitzetik 1500 esaldi ausaz aukeratu direla, espainierazko esaldi bakoitzarentzat euskarazko erreferentzia-itzulpen bakar bat izanik, honek *Bleu* emaitzaren fidagarritasuna gutxitzen duena.

Bleu neurriarekin lortutako emaitzak VIII.7 taulan ikus daitezke, haien ondoan hitzetan oinarritutako edizio-distantziaren datuak jarri ditugula. Edizio-distantziaren emaitza txikiagoek kalitate hobeagoa adierazten dute.

	<i>Bleu</i>		<i>Ed-dist</i>	
	Matxin	Matrex	Matxin	Matrex
Eitb	9.30	9.02	40.41	71.87
Cons	6.31	8.03	43.60	57.97

Taula VIII.7: *Bleuren* emaitzak

Bleuren emaitza txikiagoek, berriz, itzulpenaren kalitate okerragoa adierazi beharko lukete.

Bleuren emaitzei begiratzen badiegu, *Eitb*ko corpuseko esaldietan *Matxin* *Matrex* baino pixka bat hobeagoa da (9.30 vs. 9.02). Baina *Consumer* corpuseko esaldietan, hau da, *Matrex* entrenatutako corpus bereko esaldietan, *Matxin*en emaitza *Matrex*ena baino okerragoa da (6.31 vs. 8.03).

Datu hauetatik sistemen irteeraren kalitateari buruzko ondorioak aterako bagenu, entrenatu den domeinuaren testuetarako *Matrex*en itzulpenen kalitatea *Matxin*en itzulpenena baino hobeagoa dela esango genuke.

Baina eskuzko postedizioan oinarritutako edizio-distantziarekin jasotako emaitzek ondorio horiek guztiz baliogabetzen ditu. *Matxin*en irteeraren kalitatea *Matrex*ena baino askoz hobeagoa da *Matrex* entrenatu ez den domeinuko esaldietan (40.41 vs. 71.87), eta *Matrex* entrenatutako domeinuko esaldietan distantzia ez bada hain handia, *Matxin* ere dezente hobeagoa da (43.60 vs. 57.97).

Beraz, *Bleu* neurriaren emaitzek, erreferentzia batekin, sistemaren itzulpenen eta erreferentzi-itzulpenen arteko distantzia adierazten badute ere, egiaztatu dugu emaitza horietatik ezin dela irteeraren kalitateari buruzko ondorioak ateratu, eta ezin direla emaitza horiek erabili estrategia desberdinetako bi sistema desberdinak konparatzeko.

*Bleu*ri buruzko hausnarketa hau kritika daiteke erreferentzia bakarra erabili dugulako. Kontua da erreferentzia gehiagorekin frogatu nahiko bagenu, erreferentziak eskuz sortu beharko genituzkeela, eta hori edizio-distantzia kalkulatzeko egin-beharreko postediziorako lana baino askoz ere garestiagoa litzateke. Hala ere, (Callison-Burch *et al.*, 2006) lanean frogatu dute arazo bera erreferentzi anitzekin ere gertatzen dela.

Bleuren mugak, gaur egun, agerian jarri badira ere, eta erabilera egoiak eta desegokiak seinalatu badira ere, IAren arloan kasu askotan *Bleu* erabiltzen da zentzu handirik gabe. Erabilera horiek jarraitzen diote guk izendatutako *Todo a 100* filosofiari: *Bleu* erabiltzen dut, jakin gabe zertarako balioko zaidan, merkea delakoan.

Eta filosofia horretan bi oker egiten dira:

- Zenbaitetan *Bleu* erabiltzeak ez du ezertarako balio. Adibidez siste-

ma baten ebaluazio absolutua emateko edo estrategia desberdinetako sistemak konparatzeko, *Bleuren* emaitzetatik ezin dugu inolako ondorioarik atera.

- *Bleuk* kosturik ez duela esaten da, merkea dela, baina *Bleu* modu *fidagarrian* erabiltzeko 3 erreferentzia ematea beharrezkoa da, eta corpus elebidun gehienetan erreferentzia bakarra egoten da. Horregatik, *Bleu* erabiltzeko, erreferentziak eskuz sortu behar dira, eta eginbeharreko lan hori oso garestia da.

Horregatik premiazkoa da *Bleuren* erabilera zentzugabeak alboratzea, soilik egokia den kasuetan erabiliz eta serioski aztertuz duen kostua.

Matxin sistemaren garapenean, adibidez, ez dugu *Bleu* erabiliko kalitatea absolutua neurtzeko edo beste motatako sistemaekin konparatzeko. Aztertzekoa da sistemaren garapenean hobekuntza inkrementalak neurtzeko baliagarria zaigun, eta baliagarria izanez gero, erreferentzi anitzak sortzearen kostua zenbatekoa den. Horren arabera baloratu beharko dugu edizio-distantzia baino komenigarriagoa den edo ez.

Edizio-distantziaren erabileraren ondorioak

Itzulpen automatikoaren ebaluaziorako edizio-distantzia erabiltzearen erabakia oso positibotzat hartzen dugu.

Lortutako emaitzek IAko sistemen irteeraren kalitateari buruzko ebaluazio fidagarria eman digu. Bai sistemen ebaluazio absolutua, jakin ahal izateko irteera zenbaterainoko erabilgarria den, eta baita sistemen ebaluazio erlatiboa ere, sistema desberdinen artean konparatu hala izateko.

Gainera edizio-distantzia kalkulatzeko eginbeharreko postedizioa ez da hain garestia izan. Gure kasuan esperimentu guztietarako (200 esaldien postedizioa, 100 *Matxin* sistemarekin itzulitakoak eta beste 100 *Matrex* sistemarekin) 6 ordu eta 30 minutu behar izan dira soilik, edizio-abiadura 350 hitz/ordukoa izan dela, hau da, 2 esaldi/minutuko.

(Snover *et al.*, 2006; Przybocki *et al.*, 2006) lanetako esperimentuetan 500 eta 800 hitz/orduko abiadura aipatzen dute, gurea baino handiagoa, suposa dezakegula postediziorako ingurune baten erabilerak produktibitatea hobetzen duela.

Etorkizuneko ebaluazioetarako argi dugu edizio-distantzia erabiliko dugula. Postediziorako ingurune baten erabilera aztertzeko asmoa dugu, eta aukera izanez gero, postedizioa burutzeko pertsona bat baino gehiago erabili nahiko genuke, esaldi-kopurua ere handituz.

VIII.3 Erroreen analisia

Erroreen iturburua aztertu dugu, zein modulutan sakondu behar dugun de-
tektatzeko eta etorkizuneko hobekuntzak proposatzeko. Ondokoak dira erro-
reen iturburu nagusiak, sortzen duten errore-kopuruaren arabera ordenatu-
ta:

- Analisia.

Itzulpen-prozesurako beharrezkoa den espainierazko testuaren anali-
si sintaktiko osoa burutzean sortzen dira errore gehienak. Analisisian
zenbait errore-mota ematen dira:

- Mendekotasun-analizatzailea. Modulu honetan erroreak sortzen
dira mendekotasun-zuhaitza eraikitzean, batez ere loturazko ele-
mentuekin eta *PP-attachment* delakoarekin. Erroreen adibideak
eta eragiten dituzten itzulpen okerrak VIII.8 taulan ikus daitezke.
- Etiketatzaile morfosintaktikoa. Hitz bakoitzaren analisi morfolo-
giko posibleen artean zenbaitetan desanbiguazioa gaizki egiten
da. Adibideak VIII.9 taulan ikus ditzakegu.
- Funtzio sintaktikoen esleipena. Zenbaitetan ez da funtzio sintak-
tiko egokia esleitzen eta horrek, batzuetan, erroreak eragiten di-
tu itzulpenetan. Horregatik, zenbait modulutan erabilgarria den
informazio hori ez erabiltzea erabaki dugu, funtzio sintaktikoen
esleipena zuzenagoa izan arte. Errore mota honetako adibideak
VIII.10 taulan ikus daitezke.
- Hautapen lexikala. Hitz baterako ematen den itzulpena lexikoi elebi-
duneko sarrera horrek hiztegian duen lehenengo ordaina da eta zenbai-
tetan, honela, itzulpen okerra aukeratzen da. Adibideak VIII.11 taulan
ikus daitezke.
- Preposizio eta funtzio sintaktikoen itzulpena. Postposizioen desanbi-
guazioan gertatzen diren erroreen adibideak VIII.12 taulan ikus dai-
tezke.
- Hitz anitzeko terminoak. Gure sistemaren hiztegietan hitz anitzeko
termino asko dauzkagun arren, beste asko sistemak ez ditu harrapatzen
esaldi okerrak emanez. Adibideak VIII.13 taulan ikus daitezke.
- Beste errore batzuk, neurri txikiagoan, gertatzen dira: hitzen orde-
nan, aditzen transferentzian hiztegiarekin (hainbat sarrera ez egotea

	Okerra	Zuzena
Jatorrizkoa	El perro come pan en casa y el gato bebe	
Analisia	come _ El perro _ bebe _ pan _ en casa _ y _ el gato	y _ come _ El perro _ pan _ en casa _ bebe _ el gato
Itzulpena	Edaten du txakurrak ogia jaten du etxean eta katua	Txakurrak ogia jaten du etxean eta katuak edaten du
Jatorrizkoa	La inscripción a los cursillos ha sido correcta	
Analisia	ha sido correcta _ La inscripción _ a los cursillos	ha sido correcta _ La inscripción _ a los cursillos
Itzulpena	Inskripzioa ikastaroei zuzena izan zaie	Ikastaroetan inskripzioa zuzena izan da
Jatorrizkoa	Las farmacias dan fichas con consejos sobre enfermedades	
Analisia	dan _ Las farmacias _ fichas _ con consejos _ sobre enfermedades	dan _ Las farmacias _ fichas _ con consejos _ sobre enfermedades
Itzulpena	Farmaziek fitxak ematen dituzte aholkuekin gaixotasunen gainean	Farmaziek gaixotasunen gaineko aholkuak dituzten fitxak ematen dituzte
Jatorrizkoa	He venido con Juan y Pedro.	
Analisia	He venido _ con Juan _ y _ Pedro	He venido _ con _ y _ Juan _ Pedro
Itzulpena	Joanekin etorri naiz eta Petri.	Joan eta Petriekin etorri naiz.

Taula VIII.8: Mendekotasun-analizatzailearen erroreak. Adibideak

	Okerra	Zuzena
Jatorrizkoa	¿Qué nota tiene?	
Analisia	nota notar VMIP3S0	nota nota NCFS000
Itzulpena	Du zerk nabari du?	Zer nota du?
Jatorrizkoa	Yo como manzanas	
Analisia	como como CS	como comer VMIP1S0
Itzulpena	Sagarrak ni	Nik sagarrak jaten ditut

Taula VIII.9: Tagger-aren erroreak. Adibideak

	Okerra	Zuzena
Jatorrizkoa	Me llamo Juan	
Analisia	Juan sn obj	Juan sn pred
Jatorrizkoa	He visto a Juan	
Analisia	Juan sn iobj	Juan sn obj

Taula VIII.10: Funtzio sintaktikoen esleipenean erroreak. Adibideak

Jatorrizkoa	Itzulpen okerra	Zuzena
Realizarán una <i>audición</i> de todas las <i>composiciones</i> .	<i>Osaketa</i> guztien <i>entzute</i> bat egingo dute.	<i>Konposizio</i> guztien <i>entzunaldia</i> egingo dute.
Abrirán el <i>próximo curso</i> .	<i>Bide hurbila</i> irekiko dute.	<i>Hurrengo ikasturtean</i> irekiko dute.

Taula VIII.11: Hautapen lexikalean erroreak. Adibideak

Jatorrizkoa	Itzulpen okerra	Zuzena
Están <i>en</i> el colegio de Rojas.	Rojalesen ikastetxean daude.	Rojaleseko ikastetxean daude.
Llegan a todos los aeropuertos <i>del</i> mundo.	Munduaren aireportu guztietara iristen dira.	Munduko aireportu guztietara iristen dira.
Esa botella es <i>de</i> cristal.	Botila hori kristaletik da.	Botila hori kristalezkoa da.
El proceso <i>de</i> paz continúa.	Bakearen prozesuak jarraitzen du.	Bake-prozesuak jarraitzen du.
El Ayuntamiento subvencionará <i>el alquiler</i> a los inmigrantes.	Udalak <i>alokairua</i> dirulaguntza emango die etorkinei.	Udalak <i>alokairurako</i> dirulaguntza emango die etorkinei.

Taula VIII.12: Preposizio eta funtzio sintaktikoen itzulpenean erroreak. Adibideak

Jatorrizkoa	Itzulpen okerra	Zuzena
Hay muchos <i>casos de siniestralidad</i> .	<i>Ezbehar-kopuruaren kasu</i> asko dago.	<i>Ezbehar kasu</i> asko dago.
Debe introducir una <i>dirección de correo electrónico</i> .	<i>Posta elektronikoaren norabide</i> bat sartu behar du.	<i>Helbide elektroniko</i> bat sartu behar duzu.

Taula VIII.13: Hitz anitzeko terminoekin erroreak. Adibideak

edo ordain okerrak ematea), sorkuntza morfologikoan, aditzarekin komunztadura okerrak, zenbaki eta laburduren itzulpenetan eta forma-tua-arazoak (kakotxak, maiuskulak...)

VIII.4 Laburpena

Azkeneko urteetan itzulpen automatikoaren ebaluazioaren arloa oso aktiboa izan da.

Kapitulu honetan IAren giza-ebaluaziorako neurriak eta neurri automatiko garrantzitsuenak deskribatu ditugu eta gaur egun gehien erabiltzen diren ebaluazio neurri automatiko arrakastatsuen ahuleziak aztertu ondoren, edizio-distantzia zehazki azaldu dugu, hori baita gure sistemaren irteeraren kalitatea neurtzeko erabili dugun metrika.

Matxin sistemaren irteeraren ebaluazioaren datuak aurkeztu ondoren, metodo estatistikoetan oinarritutako *Matrex* sistemarekin egindako konparazioa jakitera eman dugu, *Matxin* sistemaren emaitzak diferentzia handiz hobeagoak direla erakutsiz.. Azkenik, sistemaren erroreen analisia egin dugu.

IX. KAPITULUA

Ondorioak eta etorkizunerako lanak

*Soy feliz, soy un hombre feliz,
y quiero que me perdonen
por este día los muertos de mi felicidad*

Silvio Rodríguez

Kapitulu honetan, egindako lanari buruzko gogoetak bildu ditugu. Tesi-lan honen ekarpen garrantzitsuenak aurkeztu ondoren, egindako lanetik ateratako ondorioak azalduko ditugu. Bukatzeko etorkizunerako lanen berri emango dugu.

IX.1 Tesi-lan honen ekarpenak

Tesi honetan burututako lanak emandako fruituak ondoko lerrootan laburbilduko ditugu:

- *Matxin* euskarara itzultzeko erregeletan oinarritutako transferentzia sakoneko sistemaren diseinua. Aurretik sortua ziren hainbat baliabide berrerabili ditugunez, estandarren erabilera behar-beharrezkoa izan da sistemaren kalitatea bermatu ahal izateko. Sistemaren arkitekturak posible egingo du erdal hizkuntzetatik euskarara itzultzeko sistema berriak implementatzea.
- *Matxin* espainieratik euskarara asimilaziorako itzulpen automatiko osoa eskaintzen duen prototipoaren implementazioa, kode irekiko software librean. Sistema hau bi zentzutan da aitzindari:

- Euskararekin lan egiten duen lehenengo IAko sistema erabilgarria da.
- Software librean eraikitako IAko lehenengo sistemetako bat da.
- *Matxin* sistemaren eraikuntzan ondoko lanak azpimarratu nahi ditugu, berritzaileak direlako:
 - Mendekotasun-analizatzailearen eraikuntza. *Freeling* analizatzailearen hedapena bideratu dugu transferentzia sakonerako beharrezkoa den mendekotasun-analisia lortzeko. Guk egindako lanean oinarrituta *Freeling* analizatzailearen garatzaileek *Txala* izeneko mendekotasun-analizatzailea inplementatu dute eta bere sisteman integratu, espainierarako kode irekian lehena izanik.
 - Preposizio eta funtzio sintaktikoen transferentziarako moduluak. Modulu honen eraikuntzarako postposizioen desanbiguaziorako zenbait teknika aztertu eta aplikatu ditugu.
 - Aditz-kateen transferentzia. Ataza konplexu honi aurre egiteko, bi hizkuntzetan oso desberdinak diren aditz-kateen transferentziarako modulu eraginkorra eraiki dugu.
 - Sorkuntza sintaktikoa abiapuntu-hizkuntzatik independentea izateko hitzen eta sintagmen ordenaketarako moduluen diseinua eta inplementazioa.
 - Sorkuntza morfologikoa burutuko duen prozesadore morfologikoaren egokitzapena, kode irekiko softwarea izateko.
 - Sistemako lexikoia sortzeko estaldura handiko hiztegien berrera-bilpenerako prozesaketa automatikoa. Lexikoian informazio semantikoa gehitzeko, ezaugarri semantikoen etiketatze erdi-automatiko bat diseinatu dugu.
- Sistemako erregelen idazketarako formalismo bateratuaren proposamena. *Matxin*1.0 sistemako moduluek erregelen idazketarako formalismo desberdinak erabiltzen dituzte eta formalismo bakar bat erabiltzea aurrerapauso garrantzitsua izango da sistemaren garapenean.
- *Matxin* sistemaren ebaluazioa. Edizio-distantzia neurria erabiliz sistemak %30 eta %40 bitarteko emaitza lortzen du. Metodo estatistikoetan oinarritutako *Matrex* sistemarekin konparazioa egin ahal izan dugu, *Matxin* sistemaren emaitzak diferentzia handiz hobeagoak izanik.

- *Bleu* ebaluazio-neurriaren erabilerari buruzko hausnarketa, azken bi urteetan argitaratutako hainbat esperimenteren ondorioa egiaztatu dugula: *Bleu*k ez du balio estrategia desberdinetako sistemak konparatzeko
- Itzulpen automatikoaren inguruko oinarritzko ezagutzaren aurkezpena euskaraz. Gure lanerako egin behar izan dugun azterketa bibliografiko sakonetik jaso dugun IAri buruzko panoramaren laburpena arlo honetan lanean diharduten eta arituko diren euskal ikertzaileentzat oinarritzko erreferentzia izatea espero dugu.

IX.2 Ondorioak

Tesi honetarako ezarritako helburua ondoko ezaugarriak beteko lituzkeen itzulpen-sistema baten diseinua eta inplementazioa zen: itzulpen prozesu osoa automatikoki burutzea; erregeletan oinarritutako estrategian oinarritzea, aurrerago sistema hibrido bat eraikitzeke asmoarekin; eleanitza izatea, xede-hizkuntza euskara izanik; erabilera orokorreko sistema izatea; eta software librean inplementatzea.

Gure helburuak sobera bete ditugu, *Matxin* espainiera-euskara itzulpen automatikoko sistema publikoki erabilgarri ere badagoela. Atazaren konplexutasuna kontuan hartuta, sistemaren itzulpenetan lortutako kalitatea positiboki baloratzen dugu, oinarritzko ulermenerako balio duelako, zabal-kunde-sistema batek izan behar duen zuzentasunetik urrun badago ere.

Egindako lana aitzindaria da, bai euskararekin lan egiten duen lehenengo IAko sistema erabilgarria delako eta bai software librean eraikitako lehenengo IAko sistemetako bat delako.

Azken hamarkadako joera, IAren arloan, corpusetan oinarritutako estrategiak erabiltzea izan da, erregeletan oinarritutako lanak gutxiesten. Euskarara itzultzeko sistema bat eraikitzerakoan ondoko bi baldintzei erreparatu behar zaie:

- Euskararako gaur egun dauden corpusak (eta etorkizunean egongo direnak ere) mugatuak dira.
- Euskara hizkuntza eranskaria da.

(Koehn eta Monz, 2006) laneko esperimenteretan frogatu dute sistema estatistikoak erregeletan oinarritutako sistema komertzialen atzean gelditzen direla morfologia aberatsa duten hizkuntzetara itzultzean. Gainera, teknika estatistikoek corpus erraldoiak behar dituzte emaitza onargarriak lortzeko.

Beraz, euskarara automatikoki itzultzeko, hurbilpen estatistikoen erabilerak zailtasun handiekin topo egiten du.

Etorkizuna hibridazioaren bidetik etorriko da, eta gaur egun dauzkagun datuekin, aurreikus dezakegu hibridazio horietan erregeletan oinarritutako teknikak pisu handia izango dutela.

IX.3 Etorkizunerako lanak

Tesi-lan honen hasieran esan dugun bezala, aurkeztu duguna IAko sistema baten eraikuntzaren lehenengo urratsak dira, burututako lana hasiera baino ez delako.

Aurrera begira bide asko daude zabalik eta interesgarrienak labur-labur aurkeztuko ditugu, hiru norabidetan multzokatuz: *Matxin* sistemako moduluak hobekuntzak, teknika eta estrategia berrien ikerketa eta ataza berriei aurre egitea.

Matxin es-eu sistemaren hobekuntzak

Matxin sisteman hiru hobekuntza-bide nabarmendu ditzakegu:

- Sistemako erregelen idazketarako formalismo bateratuaren inplementazioa. Horrek sistemaren garapenean egin nahiko ditugun lan guztietan eragina handia izango du, izugarri erraztuko dituelako, alde batetik, hobekuntzak egiteko erregelen moldaketa eta gehitze-lanak eta, bestetik, sistemaren hedapena edo beste atazetarako moduluen berreskualdatzea.
- Itzulpenen kalitatea hobetzeko sistemako moduluak eta datu linguistikoetan erroreen azterketa eta zuzenketa.
- Espainerarako analizatzailearen hobekuntza. Berezko garrantzia du, batez ere, mendekotasun-analizatzaileak loturazko elementuekin dituen arazoei aurre egitea, horretarako modulu ahaltsuago bat berriztatzea beharrezkoa izan daitekeela.

Teknika eta estrategia berrien ikerketa

Hasieratik aurreikusita bagenituen ere, orain arte heldu ez dizkiegun beste hainbat teknika eta estrategia erabiltzea beharrezkoa da gure sistemak dituen mugak gainditzeko:

- Hitzen adiera-desanbiguaziorako teknikak. Hautapen-lexikalerako desanbiguazio tekniken erabilera eta sisteman modu eraginkorrean integrazteko aukera ikertu behar ditugu. Bide horretan, dagoeneko, aurretikoa den urrats bat egin dugu: lexikoian adiera-anitzerako formatua diseinatu eta prestatu.
- Domeinu zehatzak. Gure sistema domeinu zehatzetarako egokitzeko bideak aztertu behar ditugu, frogatu izan baita hau dela itzulpenen kalitatea hobetzeko bidea. Oso erabilgarria izan daiteke, adibidez, administrazio-testuen domeinurako bereiztea es-eu sistema.
- Corpusetan oinarritutako tekniken ikerketa. Adibideetan oinarritutako estrategiek eta metodo estatistikoek eskaintzen dituzten aukerak aztertu eta probatu behar ditugu. Horretarako, gaur egun, zenbait aplikazio-eremutan, erabilgarriak izan daitezkeen tamaina ertaineko corpusak eskura ditugu.
- Hibridazioa. Erregeletan oinarritutako gure sistema corpusetan oinarritutako estrategiekin nola konbinatu ikertu behar dugu. Estrategia desberdinek zertan asmatzen duten eta zertan egiten duten kale aztertzea beharrezkoa da hibridaziorako bide posibleak diseinatu ahal izateko. Oso argi daukagu euskarara kalitate handiz itzuliko duten etorkizuneko sistemek estrategia guztien alde onak bilduko dituztela.

Ataza berriei aurre egitea.

Badira abiatu dugun lanaren jarraipen zuzena izango diren hainbat lan:

- Postediziorako interfazearen diseinua eta inplementazioa.
- Hiztegia aberasteko tresnen inplementazioa.
- Beste abiapuntu-hizkuntzetarako sistemen inplementazioa. *Matxin* sistema abiapuntu-hizkuntzatik independentea izateko diseinatu dugu, eta zailtasun handirik gabe heldu ahal izango diogu ingelesetik euskarara itzuliko duen sistemaren eraikuntzari. Horretarako:
 - Ingeleseko analizatzailea bat berrerabili beharko genuke.
 - Transferentzia-faserako modulu berdinak erabiliko genituzke, ingelesa-euskara hiztegi bat prozesatuz lexikoia eraikiko genukeela eta erregelak hizkuntza bikote honetarako kodetu beharko genituzkeela.

- Sorkuntza fase osoa dagoen bezala erabiliko litzateke, moduluak eta datu linguistikoak.

Ingelesetik euskarara itzuliko lukeen sistemaren erabilgarritasuna askoz handiagoa izango litzateke, oinarritzko ulermenerako ingelesezko web-orriak itzultzeko jende askok erabiliko lukeelako.

- Beste norabiderako sistemaren diseinua. Euskaratik erdal hizkuntzetera itzultzeko sistemaren eraikuntza *Matxin* sisteman oinarritzeko aukerak sakonki aztertu beharko lirateke, sistema euskarara itzultzeko diseinatua izan baita.
- Euskara barne hartzen ez duen sistemaren inplementazioa. *Matxin* sistemaren ahalmena aztertu ahal izango genuke, adibidez, espainieratik urruna den beste hizkuntza batera, adibidez quechuara.

Kapitulu honekin egin dugun lanaren kontakizuna amaitzen dugu. Esan bezala, itzulpen automatikoaren mundu liluragarriCCi honetan bide anitz zabaltzen zaizkigu. Euskararekin lan egiten dugunez, baliabideak urriak izango dira, eta beraz, poliki-poliki joan beharko dugu aurrera. Poliki, baina gelditu gabe.

Umiltasunez aitortzen dugu guk egindako lana hasiera baino ez dela; harrotasunez onartzen dugu gure ekarpenek bide asko zabaltzen dituztela.

Itzulpen automatikoa eta euskararekin lan egiten eta egingo dutenentzat guk egindako lana erabilgarria izatea espero dugu.

Hori izan da gure asmoa. Hori da gure gogoa.

Bibliografia

- Abaitua J. Itzulpengintza Automatikoa: hamar orduko sarrera, Urtarrila 2002. URL <http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/mt10h.eu>. DUko Itzulpengintza eta UEU/EHUko Hizkuntza-teknologia graduondoa.
- Abaitua J. El proyecto legebiduna. URL <http://www.serv-inf.deusto.es/abaitua/konzeptu/lege2dun.htm>. 1997.
- Abaitua J. Resultados del estudio de viabilidad esp-eus (28-02-2003). Technical report, AutomaticTrans, 2003. URL <http://www.slideshare.net/JosebaAbaitua>.
- Abaitua J., Casillas A., Diedrich T., Domínguez A., Garay A., Isasi C., Jacob I., Madariaga I., Martínez R., and Ramírez J.L. XML-Bi: procedimientos para la gestión de flujo documental multilingüe sobre TEI/XML. *Procesamiento del Lenguaje Natural*, 27:291–292, 2001.
- Abney S. *Principle-Based Parsing: Computation and Psycholinguistics*, chapter Parsing by Chunks, pages 257–278. Kluwer Academic, Boston, 1991.
- Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M., Martínez D., Sarasola K., and Urizar R. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. In *Proceedings of the EURALEX*, 2000.
- Agirre E., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., and Sarasola K. Different issues in the design of a general-purpose Lexical Database for Basque. In *First Workshop on Applications of Natural Language to Databases*, pages 299–313, Versailles, June 1995.

- Aït-Mokhtar S., Chanod J., and Roux C. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(3):121–144, 2002. ISSN 1351-3249.
- Alam Y.S. Decision trees for sense disambiguation of prepositions: Case of over. In Moldovan D. and Girju R., editors, *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 52–59, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.
- Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., and Sarasola K. Learning argument/adjunct distinction for basque. In *ACL'2002 SigLex Workshop on Unsupervised Lexical Acquisition*, 2002.
- Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K., and Zabala I. Hitz-hurrenkeraren azterketa masiboa corpusean. Barne txostena, UPV/EHU, 2003.
- Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak*. PhD thesis, EHU, 1995.
- Alegria I., Artola X., and Sarasola K. Improving a robust morphological analyser using lexical transducers. In Mitkov R. and Nicolov N., editors, *Recent Advances in Natural Language Processing*, volume 136 of *Current Issues in Linguistic Theory (CILT)*, pages 97–110. John Benjamins, 1997.
- Alegria I., Artola X., Sarasola K., and Urkia M. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203, 1996.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and Sarasola K. An FST grammar for verb chain transfer in a Spanish-Basque MT system. In *Proceedings of the FSMNLP Finite-State Methods and Natural Language Processin. Poster session*, Helsinki, 2005a.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and Sarasola K. Transfer-based mt from spanish into basque: reusability, standardization and open source. In *Cicling 2007. Proceedings. LNCS 4394.*, pages 374–384, Mexico, 2007.
- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K., Forcada M., Ortiz S., and Padró L. An open architecture for transfer-based Machine Translatio between Spanish and Basque. In *Proceedings of the MT Summit X. Workshop: OSMaTran, Open-Source Machine Translation.*, Phuket, Thailand, 2005b.

- Alegria I. and Urkia M. *Morfologia konputazionala. Euskararen morfologia-ren deskribapena*. UEU, 2002. ISBN 84-8438-034-3.
- Allen J. *Natural Language Understanding*. Benjamin Cummings Publishing Company, second edition, 1995.
- ALPAC. Languages and machines: computers in translation and linguistics. Technical report, National Academy of Sciences, National Research Council, 1966.
- Alshawhi H., Boguraev B., and Carter D. Placing the dictionary on-line. In *Computational lexicography for natural language processing*, pages 41–63. Longman Publishing Group, White Plains, NY, USA, 1989. ISBN 0-470-21187-3.
- Antworth E.L. *PC-KIMMO: a two-level processor for morphological analysis*. Number 16. Summer Institute of Linguistics, Dallas, 1990.
- Arnold D., Balkan L., Meije S., Humphreys R., and Sadler L. *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, 1994. ISBN 1855542-17x.
- Arriola J.M. *EUSKAL HIZTEGIAREN Azterketa eta Egituratzeari Ezagutza Lexikalaren Eskuratzeari Autoimatikoari Begira*. PhD thesis, Filologia eta Historia-Geografia Fakultatea, Euskal Herriko Unibertsitatea (EHU), Donostia, 2000.
- Atserias J., Casas B., Comelles E., González M., Padró L., and Padró M. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006.
- Atserias J., Comelles E., and Mayor A. Txala un analizador libre de dependencias para el castellano. In *Actas del XXI Congreso de la SEPLN XXI. Demo session*, pages 455–456, 2005.
- AutomaticTrans. Traducción automática euskara. guía para las pruebas del ATS. Technical report, Eusko Jaurlaritza, 2003.
- Babych B. Weighted N-gram model for evaluating Machine Translation output. In *Proceedings of the CLUK '04. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, pages 15–22, Birmingham, UK, 2004.

- Beesley K.R. and Karttunen L. *Finite-State Morphology*. (CSLI-SCL) Center for the Study of Language and Information - Studies in Computational Linguistics, 2003.
- Beninatto R. The automated translation manifesto: A vision for automated translation and its role in communications. In *Proc. of the 23th Internationalization and Unicode Conference*, Prague, 2003. URL <http://www.common senseadvisory.com>.
- Bennett W.S. and Slocum J. The lrc machine translation system. In *Computational Linguistics*, volume 11, 1985.
- Blatz J., Fitzgerald E., Foster G., Gandrabur S., Goutte C., Kulesza A., Sanchis A., and Ueffing N. Confidence estimation for Machine Translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2003. Summer Workshop Final Report.
- Boguraev B. and Briscoe T. Introduction. In Boguraev B. and Briscoe T., editors, *Computational Lexicography for Natural Language Processing*, pages 1–40. Longman Publishing Group, White Plains, NY, USA, 1989. ISBN 0-470-21187-3.
- Boitet C., Bey Y., Tomokiyo M., Cao C., and Blanchon H. IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT-06*, pages 23–30., Kyoto, Japan, 2006.
- Boitet C. Bernard vauquois’ contribution to the theory and practice of building mt systems: a historical perspective. In *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Pittsburgh, Pennsylvania, 1988.
- Boitet C. Four technical and organizational keys to handle more languages and improve quality (on demand) in MT. In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Bond F., Oepen S., Siegel M., Copestake A., and Flickinger D. Open Source Machine Translation with DELPH-IN. In *Proceedings of the MT Summit X. Workshop: OSMaTran, Open-Source Machine Translation.*, Phuket, Thailand, 2005.

- Bonet J. Present and future of Machine Translation in the Directorate-General for translation of the European Commission. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.
- Brants T., Skut W., and Uszkoreit H. Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop*, pages 69–76, Paris, France, 1999.
- Briscoe, Carroll J., and Watson R. The second release of the rasp system. In *Proceedings of the COLING/ACL 2006*, Sydney, Australia, 2006.
- Briscoe T. Lexical issues in Natural Language Processing. In Klein E. and Veltman F., editors, *Natural Language and Speech*, pages 39–68. Springer-Verlag, 1991.
- Brown P.F., Cocke J., Pietra S.A.D., Pietra V.J.D., Jelinek F., Lafferty J.D., Mercer R.L., and Roossin P.S. A statistical approach to Machine Translation. *Comput. Linguist.*, 16(2):79–85, 1990. ISSN 0891-2017.
- Brown R. Example-Based Machine Translation in the Pangloss system. In *Proceedings of the COLING-96: The 16th International Conference on Computational Linguistics (COLING-96)*, pages 169–174, Copenhagen, 1996.
- Brown R. Automated generalization of translation examples. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pages 125–131, 2000.
- By T. English dependency grammar. In *Workshop on Recent Advances in Dependency Grammar (Coling 2004)*, Ginebra, Suiza, 2004.
- Callison-Burch C., Osborne M., and Koehn P. Re-evaluating the role of BLEU in Machine Translation research. In *Proceedings of EAACL-2006*, 2006.
- Calvo H. and Gelbukh A.F. Diluct: An open-source spanish dependency parser based on rules, heuristics, and selectional preferences. In *NLDB*, pages 164–175, 2006.
- Carbonell J.G., Mitamura T., and Nyberg E.H. The kant perspective: a critique of pure transfer (and pure interlingua, pure statistics, ..). In *Proceedings of the Fourth International Conference on Theoretical and*

Methodological Issues in Machine Translation (TMI-92), Empiricist vs. rationalist methods in MT, pages 225–235, Montreal, 1992.

Carmona J., Cervell S., Márquez L., Martí M., Padró L., Placer R., Rodríguez H., Taulé M., and Turmo J. An environment for morphosyntactic processing of unrestricted spanish text. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, 1998.

Carreras X., Chao I., Padró L., and Padró M. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.

Carreras X. and Padró L. A flexible distributed architecture for natural language analyzers. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain, 2002.

Carrol J. and Grover C. Derivation of a large computational lexicon for English from LDOCE. In *Computational lexicography for natural language processing*, pages 117–133. Longman Publishing Group, White Plains, NY, USA, 1989. ISBN 0-470-21187-3.

Casillas A. *Explotación de corpus paralelos alineados para el desarrollo de entornos de composición de documentos estructurados bilingües*. PhD thesis, Ingeniaria Fakultatea, Deustuko Unibertsitatea, Bilbao, 2000.

Cavalli-Sforza V. and Lavie A. Description of panel: Hybrid Machine Translation, why and how? In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, AMTA 2006. Visions for the Future of Machine Translation*, Cambridge, Massachusetts, USA, 2006.

Chandioux J. Météo: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. *META*, 21, pages 127–133, 1976.

Chevreau K., J., García-Moya J., and Alonso M. Generación multilingüe de boletines meteorológicos. In *Proceedings of the SEPLN*, Lleida, Spain, 1999.

- Chodorow M.S., Byrd R.J., and Heidorn G.E. Extracting semantic hierarchies from a large on line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA, 1985. Association for Computational Linguistics.
- Civit M. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*. PhD thesis, Universidad de Barcelona, 2003.
- Coch J. and Chevreau K. Interactive multilingual generation. In *CICLing '01: Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, pages 239–250, London, UK, 2001. Springer-Verlag. ISBN 3-540-41687-0.
- Coheur L., Mamede N., and Bés G.G. From a surface analysis to a dependency structure. In *Workshop on Recent Advances in Dependency Grammar (Coling 2004)*, Ginebra, Suiza, 2004.
- Corbí-Bellot A.M., Forcada M.L., Ortiz-Rojas S., Pérez-Ortiz J.A., Ramírez-Sánchez G., Sánchez-Martínez F., Alegria I., Mayor A., and Sarasola K. An open-source shallow-transfer Machine Translation engine for the romance languages of Spain. In *Proceedings of the EAMT2005. Poster session*, Budapest, Hungary., 2005.
- Cranias L., Papageorgiou H., and Piperidis S. A matching technique in Example-Based Machine Translation. In *Proceedings of Coling 1994*, pages 100–104, Kyoto, Japan, 1994.
- Cunningham H. A definition and short history of language engineering. *Journal of Natural Language Engineering*, 1999.
- Dabbadie M., Hartley A., King M., Miller K.J., Hadi W.M.E., Popescu-Belis A., Reeder F., and Vanni M. A hands-on study of the reliability and coherence of evaluation metrics. In *Proceedings of the Workshop: Machine Translation evaluation, human evaluators meet automated metrics. LREC-2002, Third International Conference on Language Resources and Evaluation*, pages 8–16, 2002.
- de Gispert Ramis A. *Introducing Linguistic Knowledge into Statistical Machine Translation*. PhD thesis, Universitat Politècnica de Catalunya, 2006.
- Díaz de Ilarraza A., Lersundi M., Mayor A., and Sarasola K. Etiquetado semiautomático del rasgo semántico de animicidad para su uso en un sis-

- tema de traducción automática. In *Actas del XVI Congreso de la SEPLN*, Vigo, 2000a.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Reusability of wide-coverage linguistic resources in the construction of an English-Basque Machine Translation system, 1999. URL <http://www.iai.uni-sb.de/iaien/iaiwip/>. IAI Working Paper No. 36. Hybrid Approaches to Machine Translation.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Building a lexicon for an English-Basque Machine Translation system from heterogeneous wide-coverage dictionaries. In *Proceedings of the MT 2000*, Exeter, United Kingdom, 2000b.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Reusability of wide-coverage linguistic resources in the construction of a multilingual Machine Translation system. In *Proceedings of the MT 2000*, Exeter, United Kingdom, 2000c.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Inclusión del par castellano-euskara en un prototipo de traducción automática multilingüe. In *Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies*, Jaén, Spain, 2001a.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Adibideetan oinarritutako itzulpen automatikoa (AOIA): azterketa bibliografikoa eta euskararekin lan egiteko proposamenak. *Senez aldizkaria*, (25):17–35, 2002a.
- Díaz de Ilarraza A., Mayor A., Sarasola K., Loinaz M., Chevreau K., and Coch J. MultiMeteok euskaraz ere badaki. *Elhuyar Zientzia eta Teknologia*, (170):42–49, 2001b.
- Díaz de Ilarraza A., Mayor A., and Sarasola K. Semiautomatic labelling of semantic features. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002b. Association for Computational Linguistics.
- Dillinger M. Dictionary development workflow for MT: design and management. In *Proc. of the MT Summit VIII*, pages 83–87, Santiago de Compostela, Spain, 2001.
- Doddington G. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. In *Proceedings of the HLT 2002: Human*

- Language Technology Conference*, pages 138–145, San Diego, California, 2002.
- Dorr B.J. and Klavans J. Introduction. In Nirenburg S., editor, *Special Issue on Machine Translation: Building Lexicons for Machine Translation 2. Machine Translation, 9*, pages 151–153. Kluwer, 1995.
- Eisele A. Improving Machine Translation quality via hybrid systems and refined evaluation methods. Presentation in November 2006 at IST Event 2006, Workshops and Networking Sessions, Multilingualism and Language Technology a Challenge for Europe, Helsinki, 2006.
- Elhuyar. *Elhuyar Hiztegia*. Elhuyar Hizkuntz Zerbitzuak, 2000. ISBN 84-95338-08-4.
- Erdozia K. *Euskal hitz hurrenkerak azterketa psikolinguistiko eta neurolinguistiko bidez*. PhD thesis, EHU, 2006.
- Euskaltzaindia. *Euskal Gramatika Lehen Urratsak V. (Mendeko Perpausak - 1)*. Euskaltzaindia, 1987.
- Farwell D. and Helmreich S. Toward pragmatics-based Machine Translation. In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Forcada M.L., Bonev B.I., Rojas S.O., Ortiz J.A.P., Sanchez G.R., Martínez F.S., and Rosell M.G. Documentación del sistema de código abierto open-trad apertium de traducción automática de transferencia sintáctica superficial. Technical report, Departament de Llenguatges i Sistemes Informàtics. Universitat d’Alacant, 2006.
- Gachot D.A. The systran renaissance. In *Proceedings of the MT Summit II*, Munich, Germany, 1989.
- Gerber L. Working toward success in Machine Translation. In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Giménez J. and Amigó E. IQMT: a framework for automatic Machine Translation evaluation. In *Proceedings of the LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 685–690, Genoa, Italy, 2006.

- Giménez J., Amigó E., and Hori C. Machine Translation evaluation inside QARLA. In *Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation, IWSLT 2005*, Pittsburgh, PA, USA, 2005.
- Goutte C. Automatic evaluation of Machine Translation quality. Presentation at the European Community. Available at <http://www.xrce.xerox.com/Publications/Attachments/2006-002/MTeval.pdf>, 2006.
- Habash N. Hybrid Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, AMTA 2006. Visions for the Future of Machine Translation. Panel: Hybrid Machine Translation, why and how?*, Cambridge, Massachusetts, USA, 2006.
- Hamon O. and Rajman M. X-score: automatic evaluation of Machine Translation grammaticality. In *Proceedings of the LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 155–160, Genoa, Italy, 2006.
- Hansen V. Patrans - a mt system: development and implementation of and experiences from a mt-system. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 114–121, Columbia, Maryland, USA, 1994.
- Hidalgo V. Euskal joskeraren mito kaltegarriak (i). In *Hizpide*, number 35, pages 39–51. 1996.
- Hovy E., King M., and Popescu-Belis A. An introduction to MT evaluation. In *Proceedings of the Workshop: Machine Translation evaluation, human evaluators meet automated metrics. LREC-2002, Third International Conference on Language Resources and Evaluation*, pages 1–7, 2002.
- Hutchins J. Current commercial Machine Translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, 17(1-2), Jan-Dec 2005a.
- Hutchins J. Current developments in Machine Translation: an overview of progress and future prospects. Presentation in January 2006 at the University of Hamburg, Germany, 2006a.
- Hutchins J. Future prospects in Machine Translation usage and research. Presentation in February 2006 at the University of Leeds, UK, 2006b.

- Hutchins J. Machine Translation: a brief history. In E.F.K.Koerner and R.E.Asher, editors, *The concise history of the language sciences: from the Sumerians to the cognitivists*, pages 431–445. Pergamon, 1995a.
- Hutchins J. A new era in Machine Translation research. In *Aslib Proceedings 47*, pages 211–219, London, England, 1995b.
- Hutchins J. The IAMT certification initiative and defining translation system categories. In *Proceedings of Fifth EAMT Workshop*, Ljubljana, Slovenia, 2000.
- Hutchins J. Has Machine Translation improved? In *Proc. of the MT Summit IX*, pages 181–188, New Orleans, USA, 2003.
- Hutchins J. Machine Translation and computer-based translation tools: what’s available and how it’s used. In Bravo J.M., editor, *A new spectrum of translation studies*, pages 13–48. Univ. Valladolid, Valladolid, 2004.
- Hutchins J. Computer-based translation in Europe and North America, and its future prospects. In *JAPIO 20th anniversary*, pages 156–160, Tokyo, Japan, 2005b.
- Hutchins W. and Somers H.L. *An introduction to Machine Translation*. London: Academic Press, 1992. ISBN 0-12-362830-X.
- Ide N. and Veronis J. Refining taxonomies extracted from machine-readable dictionaries, 1993.
- Ide N. and Veronis J. Machine Readable Dictionaries: What have we learned, Where do we go. In *Proceedings of the post-coling94 international workshop on directions of lexical research*, pages 137–146. Calzolari and C. Guo, 1994.
- Jacob I., Abaitua J., Labrador J.D., and Quintana F. X-Flow: Gestión de flujo de contenidos multilingües sobre XLIFF y TMX. *Procesamiento del Lenguaje Natural*, 31:319–320, 2003.
- Japkowicz N. and Wiebe J. A system for translating locative prepositions from english into french. In *Meeting of the Association for Computational Linguistics*, pages 153–160, 1991.
- Johnson R., King M., and des Tombe L. Eurotra: a multilingual system under development. *Computational Linguistics*, 11(2-3):155–169, 1985.

- Karttunen L. Constructing lexical transducers. In *Proc. of COLING'94*, pages 406–411, 1994.
- Kay M. The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3–23, 1997. ISSN 0922-6567.
- King M. Evaluating natural language processing systems. *Communications of the ACM*, 39(1):73–79, 1996. ISSN 0001-0782.
- Koehn P. and Monz C. Manual and automatic evaluation of Machine Translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics.
- Koskenniemi K. *Two-level morphology: a general computational model for word-form recognition and production*. Number 11. University of Helsinki, 1983.
- Kulesza A. and Shieber S.M. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation, TMI-2004*, pages 75–84, Baltimore, Maryland, USA, 2004.
- Labrador J.D., Abaitua J., Jacob I., , Hernández F.Q., and Araolaza G. El sistema SARE-Bi de catalogación y recuperación de documentos multilingües. In *JOTRI 2003. II Jornadas de Tratamiento y Recuperación de Información*, pages 96–103, Leganés (Madrid, España), 2003.
- Lavie A., Sagae K., and Jayaraman S. The significance of recall in automatic metrics for MT evaluation. In *Proceedings of the 6th conference of the Association for Machine Translation in the Americas, AMTA 2004*, pages 134–143, Washington, 2004.
- Lee A. and Przybocki M. NIST 2005 Machine Translation evaluation official results. Technical report, NIST, 2005.
- Leech G. and Wilson A. EAGLES recommendations for the morphosyntactic annotation of corpora. Technical report, EAGLES Expert Advisory Group on Language Engineering Standards, Istituto di Linguistica Computazionale, Pisa, Italy, March 1996.
- Leermakers R. and Rous J. The translation method of rosetta. *Comput. Transl.*, 1(3):169–183, 1986. ISSN 0084-0709.

- Léon M.V..M. Spanam and engspan: machine translation at the pan american health organization. *Computational Linguistics*, 11(2-3), 1985.
- Leusch G., Ueffing N., and Ney H. A novel string-to-string distance measure with applications to Machine Translation evaluation. In *Proceedings of the MT Summit IX*, pages 240–247, New Orleans, USA, 2003.
- Maas H.D. The mt system susy. In King M., editor, *Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial*, pages 209–246, Edinburgh, 1984. Edinburgh University Press.
- Macklovitch E. and Valderrábanos A.S. Rethinking interaction: The solution for high-quality MT? In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Maegaard B. Eurotra, history and results. In *MT Summit V Proceedings*, Luxembourg, 1995.
- Mamidi R. Disambiguating prepositions for Machine Translation using lexical semantic resources. In *Proceedings of the 'National Seminar on Theoretical and Applied Aspects of Lexical Semantics' organized by Centre of Advanced Study in Linguistics*, Hyderabad, 2004.
- Marcu D. and Melby A.K. Data-driven Machine Translation: a conversation with linguistics and translation studies. In *Proceedings of the AMTA 2006: 7th Conference of the Association for Machine Translation in the America. Invited talk*, Cambridge, Massachusetts, USA, 2006.
- Melamed I., Green R., and P.Turian J. Precision and recall of Machine Translation. In *Proceedings of the HLT-NAACL 2003: Conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, 2003.
- Mel'cuk I.A. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York, 1988.
- Morris M. *Morris Student Hiztegia*. Klaudio Harluxet Fundazioa, 1998.
- Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn A. and Banerji R., editors, *Artificial and Human Intelligence*, pages 173–180. Elsevier Science Publishers, Amsterdam: North-Holland, 1984.

- Nagao M. Structural transformation in the generation stage of mu japanese to english machine translation system. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 200–223, New York, USA, 1985.
- Naskar S.K. and Bandyopadhyay S. Handling of prepositions in english to bengali Machine Translation. In *Proceedings of the EACL workshop on Prepositions*, Hyderabad, 2006.
- Nevado F., Casacuberta F., and Landa J. Translation memories enrichment by statistical bilingual segmentation. In *Proceedings of the 7th conference of the Association for Machine Translation in the Americas (AMTA)*, Lisbon, Portugal, 2004.
- Ney H. Stochastic modelling: from pattern classification to language translation. In *Proceedings of the ACL-EACL 2001 workshop Data-driven Machine Translation*, Toulouse, France, 2001.
- Nießen S., Och F.J., Leusch G., and Ney H. An evaluation tool for Machine Translation: fast evaluation for MT research. In *Proceedings of the LREC-2000: Second International Conference on Language Resources and Evaluation*, pages 39–45, Athens, Greece, 2000.
- Nirenburg S. The pangloss mark iii machine translation system. Technical report, CMU, 1995.
- Nirenburg S., Beale S., and Domashnev C. A full-text experiment in Example-Based Machine Translation. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 78–87, Manchester, England, 1994.
- Nirenburg S. *Machine Translation: Theoretical and Methodological Issues*. Cambridge University Press, 1987.
- Nubel R. Knowledge sources for the disambiguation of prepositions in Machine Translation. In *Proceedings of the FIMTP96*, 1996.
- Och F. Minimum error rate training in statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japa, 2003.
- Och F.J. Statistical Machine Translation: foundations and recent advances. tutorial. In *Proceedings of the MT Summit X*, Phuket, Thailand, 2005.

- Och F.J. Challenges in Machine Translation. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.
- Och F.J. and Ney H. What can Machine Translation learn from speech recognition? In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Oflazer K. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148, 1994.
- Ortíz D., García-Varea I., Casacuberta F., Lagarda A., and González J. On the use of statistical Machine Translation techniques with a memory-based translation system (AMETRA). In *Proceedings of the ninth Machine Translation Summit*, New Orleans, Louisiana, USA, 2003.
- OSI. The Open Source definition, 2007. URL <http://www.opensource.org/docs/definition.php>.
- Pai K., Shirai S., and Nakaiwa H. Automatic construction of a transfer dictionary considering directionality. In *MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources*, Geneva, Switzerland, 2004.
- Papineni K., Roukos S., Ward T., and Zhu W. BLEU: a method for automatic evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.
- Pérez A., Casacuberta F., Torres M., and Guijarrubia V. Finite-state transducers based on k-TSS grammars for speech translation. In *Proceedings of the Finite-State Methods and Natural Language Processing Conference (FSMNLP 2005)*, pages 270–272, Helsinki, Finland, September 2005.
- Pérez A., Torres I., Casacuberta F., and Guijarrubia V. A Spanish-Basque weather forecast corpus for probabilistic speech translation. In *Proceedings of the 5th SALT MIL Workshop on Minority Languages*, Genoa, Italy, 2006.
- Petrits A., Braun-Chen F., García J.M.M., Ross C., Sauer R., Torquati A., and Reichling A. The Commission’s MT system: Today and tomorrow. In *Proc. of the MT Summit VIII*, Santiago de Compostela, Spain, 2001.

- Philippe Langlais S.G. Thomas Leplus and Lapalme G. From the real world to real words: the meteo case. In *Proceedings of the 10th EAMT conference "Practical applications of machine translation"*, pages 166–175, Budapest, 2005.
- Polguère A. *La Theorie Sens-Texte*. Dialangue, 8-9, Université du Québec à Chicoutimi, Canada, 1998.
- Popescu-Belis A., King M., and Benantar H. Towards a corpus of corrected human translations. In *Proceedings of the Workshop: Machine Translation evaluation, human evaluators meet automated metrics. LREC-2002, Third International Conference on Language Resources and Evaluation*, pages 17–21, 2002.
- Przybocki M., Sanders G., and Le A. Edit distance: a metric for Machine Translation evaluation. In *Proceedings of the LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 2038–2043, Genoa, Italy, 2006.
- Pustejovsky J. Syntagmatic processes. In *Handbook of Lexicology and Lexicography*. de Gruyter, 2000.
- Raymond E.S. *The Cathedral & the Bazaar*. O'Reilly, 2001.
- Reeder F., Miller K., Doyon J., and White J. The naming of things and the confusion of tongues: An MT metric. In *Proceedings of the MT Summit VIII*, pages 55–59, Santiago de Compostela, Spain, 2001.
- Rich E., editor. *Artificial intelligence*. McGraw-Hill, 1983.
- Sarasola I. *Euskal Hiztegia*. Gipuzkoako Kutxa, 1996.
- Sarasola K. Strategic priorities for the development of language technology in minority languages. In *Proceedings of Workshop on "Developing language resources for minority languages: re-useability and strategic priorities"*. *Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000.
- Sato S. and Nagao M. Toward Memory-Based translation. In *Proceedings of Coling 1990*, pages Vol. 3, 247–252, Helsinki, Finland, 1990.
- Schneider T. The metal system. status 1995. In *Proceedings of the MT Summit V*, Luxembourg, 1995.

- Schütz J. Blueprint for MT evolution - Reflections on elements of style. In *Proceedings of the Workshop MT 2010 – Towards a Road Map for MT hosted by MT Summit VIII*, Santiago de Compostela, Spain, 2001.
- Scott B. The Logos model: An historical perspective. *Machine Translation*, 18(1):1–72, 2003. ISSN 0922-6567.
- Scott B.E. The logos system. In *Proceedings of the MT Summit II*, Munich, Germany, 1989.
- Seneff S. Combining interlingua with SMT. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, AMTA 2006. Visions for the Future of Machine Translation. Panel: Hybrid Machine Translation, why and how?*, Cambridge, Massachusetts, USA, 2006.
- Senellart J. and White J.S. First strategies for integrating hybrid approaches into established systems. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, AMTA 2006. Visions for the Future of Machine Translation. Panel: Hybrid Machine Translation, why and how?*, Cambridge, Massachusetts, USA, 2006.
- Sheremetyeva S. and Nirenburg S. Towards a universal tool for NLP resource acquisition. In *Proceedings of the LREC2000, 2nd International Conference on Language Resources and Evaluation*, 2000.
- Slype G.V. Critical methods for evaluating the quality of Machine Translation. Technical report, European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk, 1979.
- Smith R.N. and Maxwell E. An English dictionary for computerized syntactic and semantic processing systems. In *Proceedings of the 5th conference on Computational linguistics*, pages 303–322, Morristown, NJ, USA, 1973. Association for Computational Linguistics.
- Snover M., Dorr B., Schwartz R., Micciulla L., and Makhoul J. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas, 2006*, 2006.
- Somers H. Review article: Example-Based Machine Translation. *Machine Translation*, 14(2):113–157, 1999. ISSN 0922-6567.

- Somers H., editor. *Computers and translation. A translator's guide*, volume 35. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2003.
- Sproat R. *Morphology and computation*. The MIT Press, 1992.
- Stallman R.M. Free Software definition. In Gay J., editor, *Free Software, Free Society: Selected Essays of Richard M. Stallman*. 2002a.
- Stallman R.M. Why Free Software is better than Open Source. In Gay J., editor, *Free Software, Free Society: Selected Essays of Richard M. Stallman*. 2002b.
- Sumita E. and Iida H. Experiments and prospects of Example-Based Machine Translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 185–192, Berkeley, California, 1991.
- Trujillo A. Locations in the Machine Translation of prepositional phrases. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 13–20, Montreal, Canada, 1992a.
- Trujillo A. Spatial lexicalization in the translation of prepositional phrases. In *Meeting of the Association for Computational Linguistics*, pages 306–308, 1992b.
- Trujillo A. *Lexicalist Machine Translation of Spatial Prepositions*. PhD thesis, University of Cambridge, Cambridge, 1995.
- Trujillo A. *Translation engines: techniques for Machine Translation*. Springer Verlag, 1999.
- Turian J.P., Shen L., and Melamed I. Evaluation of Machine Translation and its evaluation. In *Proceedings of the MT Summit IX*, New Orleans, USA, 2003.
- Vauquois B. Automatic translation - a survey of different approaches. In Nirenburg S., Somers H., and Wilks Y., editors, *Readings in Machine Translation*. MIT Press, 1976. Published July 2003.
- Vauquois B. and Boitet C. Automated translation at grenoble university. *Comput. Linguist.*, 11(1):28–36, 1985. ISSN 0891-2017.

- Voutilainen A. A syntax-based part of speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–164, Dublin, 1995.
- Wahlster W. Verbmobil: translation of face-to-face dialogs. In *MT Summit IV: International Cooperation for Global Communication. Proceedings*, pages 127–135, Kobe, Japan, 1993.
- Walker D.J. The Open A.I. KitTM: general machine learning modules from Statistical Machine Translation. In *Proceedings of the MT Summit X. Workshop: OSMaTran, Open-Source Machine Translation.*, Phuket, Thailand, 2005.
- Walker D. and Zampolli A. Foreword. In Boguraev B. and Briscoe T., editors, *Computational Lexicography for Natural Language Processing*, pages xiii–xiv. Longman, 1989.
- Way A., Stroppa N., K.Sarasola, and Groves D. Example-Based Machine Translation of the Basque language. In *Proceedings of the 4th International Conference on Language Resource and Evaluation (LREC)*, Boston, USA, 2006.
- Weaver W. Translation. In Nirenburg S., Somers H., and Wilks Y., editors, *Readings in Machine Translation*. MIT Press, 1949. Published July 2003.
- Wilks Y. More advanced Machine Translation. In *International Forum of Translation Technology IFIT'89*, Oslo, 1989.
- Wilks Y., Fass D., ming Guo C., Mcdonald J.E., Plate T., and Slator B.M. Machine Tractable Dictionaries as tools and resources for Natural Language Processing. In *Proceedings of the 12th conference on Computational linguistics*, pages 750–755, Morristown, NJ, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3.
- Witkam T. Interlingual mt : an industrial initiative. In *MT Summit*, pages 135–140, Japan, 1987.
- Wong K. and Sayo P. *Free/Open Source Software. A general introduction*. UNDP-APDIP, 2004. URL <http://www.iosn.net/foss/foss-general-primer/>.
- Woszczyna M., Broadhead M., Gates D., Gavaldá M., Lavie A., Levin L., and Waibel A. A modular approach to spoken language translation for large

domains. In *Proceedings of the Machine translation and the information soup: third conference of the Association for Machine Translation in the Americas, AMTA '98*, pages 31–40, Langhorne, PA,USA,, 1998.

Wurm S.A. *Atlas of the World's Languages in Danger of Disappearing*. Unesco, 2001.

Zabala I. Hitz-hurrenkera euskara tekniko-zientifikoan. In *Ekaia*, number 12, pages 143–166. 2000.

Zernik U. Introduction. In Zernik U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 1–26. LEA, 1991.

A. ERANSKINA

Itzulpen-adibideak

Eranskin honetan *Matxin* sistemaren ahalmena eta mugak erakusteko mota desberdinetako adibideen itzulpenak aurkezten ditugu:

- Laborategiko adibideak (A.1 taula).
- *Consumer* corpuseko adibideak (A.2 taula).
- Administrazio-testu bat (A.3 taula).
- Lege-estua bat (A.4 taula).
- Eskuliburu bateko testu bat (A.5 taula).
- Chat bateko zati bat (A.6 taula).
- Rock abesti bat (A.7 taula).
- Literatura klasikoko zati bat (A.8 taula).

Gainera itzultzaileak formatudun testuaren itzulpena nola burutzen duen formatua mantentzen ondoko irudietan ikus dezakegu:

- Notizien web orri bat (A.2 irudia).
- Salmenten web orri bat (A.2 irudia).

Ella pide un favor. Ella nos pide un favor. Ella nos pide unas monedas. Ella nos las pide. Ella rompe el plato. Ella lo rompe. Ella nos lo rompe. Ella nos los rompe. Se nos rompe. Se rompe. El acuerdo ha sido roto por los representantes. Ellos comen. Ellos están comiendo. Ellos tienen que comer. Ellos han empezado a comer Ellos tienen que comerse las patatas. Ellos han dejado de fumar. Nos están quitando nuestros derechos. Habrán tenido que devolverles las entradas. Vuelven a caer. Yo le doy el dinero a Juan. Yo voy a casa. Yo veo a Juan. Los políticos piden que demos tiempo al tiempo.	Hark mesede bat eskatzen du. Hark mesede bat eskatzen digu. Hark moneta batzuk eskatzen dizkigu. Hark eskatzen dizkigu. Hark platera hausten du. Hark hausten du. Hark hausten digu. Hark hausten dizkigu. Hausten da. Hausten da. Akordioa ordezkariak hautsi dute. Haiek jaten dute. Haiek jaten ari dira. Haiek jan behar dute. Haiek jaten hasi dira. Haiek patatak jan behar dituzte. Haiek erretzeari utzi diote. Gure eskubideak kentzen ari zaizkigu. Sarrerak itzuli behar izango dizkiete. Berriz erortzen dira. Nik dirua ematen diot Joani. Ni etxera noa. Nik Joan ikusten dut. Politikariak eskatzen dute pazientzia izan dezagula.
--	---

Taula A.1: Laborategiko adibideak

Macarrones, albóndigas o carne guisada deben presentarse con salsas distintas, no siempre con las mismas.	Makarroiak, haragiek edo haragi prestatuak salsa desberdinekin aurkeztu behar dute, beti berak ez.
Las organizaciones ecologistas (que muchos ciudadanos no asocian a las ONG) y las de integración de marginados, por su parte, no suscitan tan mayoritaria aprobación.	Erakunde ekologistak (Hiritar askok ONG-ERA ez dutela elkartzen) Eta baztertuen integratzeen, haren parteagatik, honenbeste onartze gehienena ez dute sortzen.
En cuanto al funcionamiento de las ONG, destacan que la gestión debe ser eficaz y transparente.	ONG-EN funtzionamenduari dagokionez, azpimarratzen dute gestioak izan behar duela garden eta eraginkorra.
Los poderes públicos fomentarán la educación sanitaria, la educación física y el deporte.	Botere publikoek osasun-hezkuntza sustatuko dute, gorputz-hezkuntza eta kirola.
Como se sabe, muchos clientes habituales estudian minuciosamente las tarifas y optan por la que más conviene a su tipo de uso, cara a que su economía doméstica se vea lo menos afectada posible por este necesario gasto logístico.	Jakin, ohizko bezero asko minuciosamente estudiatzen dituzte tarifak eta aukeratzen dute gehiago erabileraren haren tipoari bat datorkion, haren etxeko ekonomia gutxienez ikus dadila itxurati posiblea beharrezko gastu logistiko hau.
La disponibilidad horaria media en las escuelas privadas es de 11 horas y 15 minutos por cada día, casi 2,5 horas más de media que en las públicas.	Orduen libre egote erdia eskola pribatuetan 11 orduren da eta 15 minutu egun bakoitzean, publikoetan orduak gehiago erdia ia 2,5.
La megafonía retumba y la visibilidad es mejorable.	Megafoniak burrunbatzen du eta ikuspena da mejorable-a.
Tres de cada diez afirmaron que algunos tramos de montaña y carretera resultan peligrosos, especialmente en A Coruña, Lugo y León.	3/10k baieztatu zuten mendiaren tarte eta errepidea direla arriskutsuak, Coruña es- pecialmente, eta Leon Lugo.
El saco de dormir resulta muy útil, ya que los albergues ofrecen muchas plazas en el suelo.	Lozakua oso baliagarria da, ostatuek plaza asko eskaintzen dutelako lurzoruan.
Sin embargo, se comprueba que las capitales más caras en la media general del precio del agua (Barcelona, Murcia y Valencia) y las más baratas (Vitoria, Logroño, Burgos, Santander y San Sebastián), lo son prácticamente en los cuatro supuestos testados.	Hala ere, egiaztatzen du kapital garesti gehiago uraren prezioaren erdi orokorra (Bartzelona, eta Valentzia murciartzten du) Eta merke gehiago (Gasteiz, Logroño, Burgos, Santander eta Donostia), prácticamente dira 4 eman testamentua eginda.
Así y todo, los grupos ecologistas advierten de que la primera medida en el control del consumo de este preciado líquido se debe adoptar en nuestra propia casa, utilizando el agua estrictamente necesaria, y no dejando los grifos abiertos.	Hala ere, talde ekologistek aholkatzen dute 1 neurriak likido hau preziatuaren kontsumoaren kontrolean gure etxe bera adoptatu behar duela, hertsiki beharrezko ura, erabili eta ez utzi txorrotak irekiak.
El precio del metro cúbico (muy caro en todos los casos) apenas sufre variaciones en función de la cantidad consumida.	Metro kubikoaren prezioa (Kasu guztietan oso garestia) Apenas aldaketak sufritzen ditu zenbakarritasun suntsituaren arabera.

Taula A.2: Itzulpen-adibideak. *Consumer* corpusa

Estimada profesora o profesor: Me dirijo a Vd para informarle sobre la forma de hacer llegar a UNIQUAL el fichero con la solicitud y curriculum con los méritos alegados en la recientemente finalizada convocatoria de complementos retributivos adicionales, tal y como le adelanté en el escrito de 21 de febrero de 2007. Como se indicaba en el citado escrito, el objeto de enviar el fichero es agilizar el proceso de evaluación por parte de UNIQUAL, y para ello se solicita que el fichero esté en formato pdf o word. El contenido del fichero que se remita debe ser idéntico al documento registrado como solicitud. En caso de discrepancia, prevalecerá el documento registrado. Se ha habilitado un acceso a un servidor en el que se deben depositar los ficheros. La clave de acceso para cada persona es la clave LDAP asignada a cada profesora o profesor, y el nombre del fichero DEBE SER OBLIGATORIAMENTE el número de DNI (sin puntos y sin la letra del NIF). Por ejemplo, 12345678.pdf, donde 12345678 es el DNI de la persona. El plazo para depositar el fichero en este servidor se abre el jueves 26 de abril y terminará el jueves 10 de mayo. Quedo a su disposición para cualquier comentario o aclaración que estime oportuna. Atentamente.	Irakasle estimatua edo irakaslea: Vd-i zuzentzen diot informatzeko formaren gainetik egin izatera iristea UNIQUAL-ERA eskaerarekin fitxategia eta curriculum-a meritu alegatuekin berriki retributivo osagarri gehigarrien dei bukatua, aurreratu edun 2007ko otsailaren 21ko idazkian. Erakusten izan zitatu idazkia, fitxategia bidaltzeko objektua bizkortzea da UNIQUAL-EN ALDETIK balioztatzearen prozesua, eta harentzat eskatzen da fitxategia formatu pdf-en egon dadila edo word. Fitxategiaren igor dezan neurridunak berdin-berdina izan behar du eskaera dokumentu miatu. Desadostasun, dokumentu miatua gailenduko da. Sarbide bat prestatu ditu zerbitzari baten-gana utzi behar duten fitxategiak. Pasahitza pertsona bakoitzarentzat klabe LDAP zaio esleituta irakasle bakoitzari edo irakaslea, eta fitxategiaren izenak NAHI-TAEZ izan behar du NANAREN numeroa (Punturik gabe eta letrarik gabe NIF-EN). Adibidez, 12345678.pdf, 12345678 pertsonaren NANA izan. Fitxategia uzteko epea zerbitzari honengan apirilaren 26an, Osteguna irekitzen da eta bukatuko du martxoaren 10an, Osteguna. Edozein iruzkinentzat haren antolamenduari edo argibidea estima dezan egokia barea. Begirunez
---	---

Taula A.3: Itzulpen-adibideak. Administrazio-testua

Otras Disposiciones Justicia, Empleo y Seguridad Social RESOLUCIÓN de 23 de febrero de 2007, del Director de Trabajo y Seguridad Social, por la que se dispone el registro, publicación y depósito del Convenio Colectivo de la Asociación de Centros de Actividades en Medio Rural (conocidos como Granjas Escuelas), para 2005-2007 (código de convenio n.º 8602215). ANTECEDENTES El día 10 de enero de 2007 se ha presentado en esta Dirección, la documentación preceptiva referida al convenio antes citado, suscrito con fecha 28 de noviembre de 2006, por la representación de las Empresas y las Organizaciones Sindicales, en representación de las y los trabajadores. FUNDAMENTOS DE DERECHO Primero. La competencia prevista en el artículo 90.2 de la Ley del Estatuto de los Trabajadores, Real Decreto Legislativo 1/1995, de 24 de marzo (BOE de 29-03-1995) corresponde a esta autoridad laboral de conformidad con el artículo 17.f) del Decreto 315/2005, de 18 de octubre (BOPV de 27-10-2005) por el que se establece la estructura orgánica y funcional del Departamento de Justicia, Empleo y Seguridad Social, en relación con el Decreto 39/1981, de 2 de marzo (BOPV de 02-04-1981) y con el Real Decreto 1040/1981, de 22 de mayo (BOE de 06-06-1981) sobre registro de convenios colectivos. Segundo. El convenio ha sido suscrito de conformidad con los requisitos de los artículos 85, 88, 89 y 90 de la referenciada Ley del Estatuto de los Trabajadores. En su virtud, RESUELVO: Primero. Ordenar su inscripción y depósito en el Registro de Convenios Colectivos de Euskadi, con notificación a las partes. Segundo. Disponer su publicación en el Boletín Oficial del País Vasco. En Vitoria-Gasteiz, a 23 de febrero de 2007.	Beste Disposiciones batzuk Justizia, erabiltzea eta Gizarte-segurantzaren 2007ko otsailaren 23ko ERABAKIA, eta Gizarte-segurantzaren Director De Trabajo-ren, jartzen den erregistroa, argitalpena eta Convenio Colectivo De La Asociación De Centros De Actividades-en gordailua Erdian Landatarra (Granjas Escuelas ospetsuak), 2005-2007 (N hitzarmenaren kodea.8602215 °). AURREKARIA 2007ko urtarilaren 10an Norabide hone-tan aurkeztu da, hitzarmenari lehenago zitatuta halabeharrezko dokumentazio kon-tatua, egunarekin 2006ko azaroaren 28an suscrito-a, Empresas-en eta Organizacio-nes Sindicales irudikapenagatik, eta langi-leak. ZUZENBIDEKO OINARRIAK 1. Gaian Ley De El Estatuto De Los Trabajadores-en 902, previsto lehia 1, Real Decreto Legislativo martxoaren 24ko (29-03-1995 BOEA) Lan-administrazio-ko agin-taritzak honi ordaintzen dio gaiaren arabera 17.f) 315, Dekretuaren urriaren 18ko (27-10-2005 EHAA-A) Ezartzen den Justizia-sailaren egitura funtzional eta organikoa, erabiltzea eta Gizarte-segurantzaren, Dekre-tuarekin 39, zerikusian martxoaren 2ko (02-04-1981 EHAA-A) Eta Errege Dekre-tuarekin 1040, martxoaren 22ko (06-06-1981 BOEA) Hitzarmen kolektiboen erre-gistroaren ginetik. 2. Hitzarmena baldintzen arabera gaieta-tik 85, izenpetu dute 88, 89 eta Ley De El Estatuto De Los Trabajadores referenciar-tuaren 90. Haren bertutean, Konpontzen dut: 1. Haren inskripzioa eta gordailua Regis-tro De Convenios Colectivos De Euskadi-n ordenatzea, parteei jakinarazpenarekin. 2. Boletín Oficial De El País Vasco-n haren argitalpena jartzea. Vitoria-gasteiz, 2007ko otsailaren 23ari.
--	---

Taula A.4: Itzulpen-adibideak. Legeak

Cuidado de su lente gran angular: 1) Sumerja la lente en agua dulce durante al menos 20 minutos después de cada uso y déjela secar antes de guardarla. Nunca deje que el agua salada se seque sobre la lente. 2) Siga las instrucciones correspondientes para la limpieza de la lente. En primer lugar, cepille o sople el polvo suelto, la arena o los residuos de la superficie de la lente. Luego, limpie la lente con una solución de agua jabonosa suave y seque con un paño de algodón limpio. No use agentes limpiadores ni productos químicos. 3) Al guardar o transportar la lente, manténgala siempre en el estuche para transporte proporcionado para evitar rayaduras. 4) Nunca permita que la lente quede expuesta a la luz solar directa durante períodos prolongados. Esto podría calentar la lente y dañar el sello impermeable. 5) Nunca salte al agua con la cámara ni con la lente. Primero entre en el agua y pídale a alguien que le alcance su equipo. 6) Siempre utilice el cordón de seguridad para evitar que la lente se caiga o se pierda. 7) Elimine el aire que pueda haber quedado atrapado entre la lente y el puerto de la lente de la cámara. 8) Guarde la lente en un lugar fresco y seco. 9) No desarme la lente, de lo contrario, la garantía carecerá de validez.	Haren angeluhandi lentearen ardura: 1) Lentea sar dezan ur gezan gutxienez 20 minutu erabilera bakoitzaren ondoren eta baino lehen utz beza lehortzea gorde baino lehen. Inoiz ez utz dezan gazia ura lentearen gainean lehor dadila. 2) Instrukzio urgazleak jarraiki bitza lentearen garbitasunarentzat. Lehen, eskuila beza edo hauts deslaia putz egin bedi, harea edo lentearen azalaren hondarrak. Gero, lentea garbi dezan ur suabe xaboi-dunaren konponbide batekin eta kotoi garbiaren oihal batekin lehor dadin. Ez dezan erabil garbitzaile agenteak ere produktu kimikoak. 3) Gordetzea edo garraiatzea lentea, beti manten beza garraia beza kutxatilan portzionatuta rayadura-ak saihesteko. 4) Inoiz ez utz diezaion lentea geldid dadila erakutsita aldi luzatuez zuzeneko eguzkiaren argiari. Honek lentea berotu ahal izango luke eta mintzea zigilu iragazgaitza. 5) Inoiz ez uraren antzekoa izan dadin ganberarekin ere lentearekin. Urean eta norbaiti eska biezaio atzeman diezaion haren taldea lehen. 6) Beti segurtasunaren lokarria erabil dadin saihesteko lentea eror dadin edo gal bedi. 7) Gelditu airea elimina dezan harrapatuta ganberaren lentearen portu lente eta artean. 8) Lentea gorde beza leku lehor eta fresko batean. 9) Lentea ez dezan desarma, de lo contrario, bermeak baliotasunaren ez edukiko du.
---	---

Taula A.5: Itzulpen-adibideak. Eskuliburua.

Prometeo: pero te advierto q eso, en un chat, es difícil de encontrar	Prometeo: baina q aholkatzen dizut hori, chat batean, difícil-a da aurkitu izatera
LamasMala: ke lastiman mi coraxonnnnn?	LamasMala: ke-k nire coraxonnnnn min ematen dute?
Oscar29: por mi parte seria indiscrecion preguntarte la edad por eso no lo hago (ahi va eso)	Oscar29: indiscrepcion galdetu nire parte serioagatik adinak horretatik ez dut egiten (Ahi hori doa)
Lluna: tengo uno menos ke tu	Lluna: bat dut gutxiago ke zure
Pau: joder oscar	Paue: oscar-a arraioa
Oscar29: a es verdad joder	Oscar29: egia da arraioa
Pau: que educacion tienes tio	Paue: educacion-ek tio-a duzula
Lluna: pero mis niños tienen 19 y 21	Lluna: baina nire haurrak tienen 19 eta 21
Prometeo: a ver, q alguien le haga un poco de caso al Sevillano...	Prometeo: ikustera, q norbait pixka bat egin dakion kasuaren Sevillarrari ...
Oscar29: ante todo la educacion	Oscar29: ororen gainetik educacion
Sevillano52: gracias Prometeo por tu interés!!!	Sevillano52: zure interesagatik Prometeo graziak
Oscar29: jos jovencitos	Oscar29: jo gaztetxoak
LamasMala: wenas sevillano	LamasMala: wenas sevillarra
Lluna: ya te digo	Lluna: jadanik esaten zaitut
Lluna: son unos yogurines	Lluna: yogurines batzuk dira
Sevillano52: pero estoy abocado al fracaso más tremendo buaaaaa	Sevillano52: baina gozo-lehorra nago porrot beldurgarri gehiagori buaaaaa
Sevillano52: hola La más	Sevillano52: La gehiago kaixo
Prometeo: Jueves 26... y la vida sigue iguaaaal (8)	Prometeo: 26an, Osteguna ... eta bizia iguaaaal jarraitzen da (8)
LamasMala: no llores	LamasMala: ez dezazun negar egin
LamasMala: no vale la pena	LamasMala: ez merezi du
Oscar29: yo tengo problemas oculares	Oscar29: nik arazo okularrak ditut
Sevillano52: anda que no!!! estoy absolutamente obviado por el personal femenino como pa no llorá!!! jejeje	Sevillano52: pa pertsonal femeninoagatik ez llorá jejeje-a ez zaitetz ibil ezer ere nago saihestuta
Prometeo: me reafirmo: hace un suicidio colectivo? xddd :P	Prometeo: berresten naiz: suizidio kolektibo bat egiten du?Xddd : P
Lluna: ke te pasa?	Lluna: ke pasatzen da?
Irene: si quereis ver mis videos y fotos o a mi por la webcam	Irene: quereis ezan ikustea nire video-ak eta fotoak edo nire webcam
Sevillano52: estoy casi a punto de apuntarme Prometeo	Sevillano52: ia nago Prometeo apuntatu nire burua zorian
LamasMala: y pake kieres tu una femina eh!!	LamasMala: eta pake-k zure femina bat kieres-tzen duzu eh!!
Prometeo: diox, necesito un café... p	Prometeo: diox-a, kafe bat behar dut ... p
Oscar29: cada vez q defeco un troncho del tamaño de una caja de leche me lloran los ojos por q puede ser	Oscar29: aldiko begiak negar egiten dizkide izan ahal da q-agatik q-ak esnearen kaxa baten tamainaren zuztar bat kaka egiten dut
Sevillano52: no sé si eso o pintar la bicicleta de amarillo chillón jejeje	Sevillano52: ez dakit hori edo pintatzea hori oihulariaren bizikleta jejeje-a
Oscar29: (soy un cerdo)	Oscar29: (Txerri bat naiz)
Lluna: has ido al okulista???	Lluna: okulista-ra joan zara???

Taula A.6: Itzulpen-adibideak. Chat.



Irudia A.1: Interneteko web orrien itzulpena. Egunekeo berriak.

eBay: Guitarras, guitarra y guitarra electrica en eBay.es - Mozilla Firefox

Enxategia Editatu Ikusi Joan Lastar-markak Tresnak Laguntza

http://instrumentos-musicales.listings.ebay.es/Guitarras_W00Q0fromZR40Q0sacatZ3S5S0Q0soc

Buscar Búsqueda avanzada Comprar Vender Mi eBay Comunidad Ayuda

Bienvenido Identificarse o registrarse. Contactar con eBay | Mapa del sitio

Categorías de eBay

Demuestra que eres el que más sabe de música

¿Y ahora qué hago?
Compra con total seguridad en eBay

Portada > Comprar > Instrumentos Musicales > Guitarras

Todos los artículos Subasta ¡Cómpralo ya! Identificate para ver tus opciones de búsqueda personalizada

Buscar en el título y en la descripción

Búsquedas relacionadas: fender, gibson, ibanez, epiphone, telecaster

Categorías

Guitarras

- Accesorios (113)
- Acústica (48)
- Bajos (57)
- Eléctrica (269)
- Electro-acústica (39)
- Española (39)
- Otros (4)

Opciones de búsqueda

Ubicación:

☐ España ☐ Artículos a 200

597 artículos encontrados en Guitarras Guardar esta búsqueda

En lista | Imagen de la Galería Ordenados por: Hora: primero en finalizar Personalizar pantalla

Título del artículo	PayPal	Precio*	Pujas	Envío a ESP	Tiempo restante
Las mejores Oportunidades					
 Guitarra eléctrica RP-A600 Serie STRATOCASTER - NUEVA		30,00 EUR	9	30,00 EUR	3h 33m
 ALBA TELECASTER EDITION - GUITARRA ELECTRICA - NUEVA		51,50 EUR	20	25,00 EUR De Alemania	9h 27m

Eginda

Aplikazioak Places Desktop

az., mai 16, 12:29

demo - OpenTrad - Mozilla Firefox

Enxategia Editatu Ikusi Joan Lastar-markak Tresnak Laguntza

http://www.opentrad.org/demo/libs/habigatzailea.php?language=eu

Opentrad Orri hau automatikoki itzuli da, hemen duzu jatorrizko orria

Buscar Aitzindari bilaketa Erostea Saltzea Nire ebay Komunitasuna Laguntza

Edo miatu Bienvenido Identificarse. Ebay-ekin | Iekuren Mapa harremanetan jartzea

Ebay-en kategoriak

Participa en nuestra encuesta ¡Y gana 1500€!

¿ETA orain zerk egiten du? ¿Ebay-en erabateko segurtasunarekin Erosketa?

Portada > Erostea > Musika-tresnak > Gitarra

Gai guztiak Enkantea Jadanik eros ezazu! Bilaketa izena aipatuaren zure aukerak ikusteko Identificate

Buscar Bilaketa aurreratu

✓ Izenburuan eta deskribapenean bilatzea

Búsquedas relacionadas: fender, gibson, ibanez, epiphone, telecaster

Kategoriak

Gitarak

- Akzesorioak (124)
- Akustika (48)
- Baxuak (57)
- Elektronika (292)
- Electro-acústica (43)
- Espaniara (39)
- Batzuk besteak (4)

Bilaketaren aukerak

Aurkako 597 gaiGitarak Guardar esta búsqueda

Imagen De La Galería azkarrean Ordenatuak: Ordua: bukatu lehen Izena aipatu pantaila

Gaiaren izenburua	Paypal	Prezioa*	Saioak	Bidalketa ESP	Tiempo restante
Oportunidades hobea					
 RP-BERRIA A600 SERIE STRATOCASTER - electrica gitara		30,00 EUR	9	30,00 EUR	3h 33m

Eginda

Aplikazioak Places Desktop

az., mai 16, 12:29

Irudia A.2: Interneteko web orrien itzulpena. Salmentak.

A cuenta de prometer el reino de los cielos, algunos vivillos lo que están haciendo es su propio cielo particular en la Tierra. Compre un pedazo de cielo, pagando su cuota mensual. ¡Salve Regina! Mater Misericordia. Hay que estar majareto para hablar de amor de Dios y al mismo tiempo en sus escuelas preparar los cuadros de mando de la represión fascista. ¡Cómo se puede ser tan fariseo! ¡Salve Regina! Mater Misericordia. Control económico es control del poder, control mental, control sexual. Realmente tíos, nunca he visto religión que pretenda tanto salvarnos a hostias. ¡Salve! Opus Dei ¡Oh no! La polla records	Zeruko erreinua agindu kontura, Egiten ari diren vivillo batzuk bere zeru partikularra da Lurrengan. Zeruaren puska bat eros beza, haren hileko kuota ordaindu. Salve Regina!Mater Misericordia. Jainkoaren maitasunaren majareto-a egon behar da mintzatzeko eta denbora berari haren eskoletan prestatzea errepresio faxistaren agintearen koadroak. Nola izan ahal da honenbeste fariseua! Salve Regina!Mater Misericordia. Kontrol ekonomikoa boterearen kontrola da, Kontrol mentala, kontrol sexuala. Osabak, Egiaz inoiz ez nahi izan dezan erlijioa ikusi natzaie honenbeste salbatu ostiei. Agur ! Oh ez Opus Dei! Records-ak oilanda
---	---

Taula A.7: Itzulpen-adibideak. Kanta bat

El ingenioso hidalgo Don Quijote de la Mancha En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lantejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo para las fiestas, con sus pantuflos de lo mismo, y los días de entresemana se honraba con su vellorí de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años; era de complexión recia, seco de carnes, enjuto de rostro, gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada, o Quesada, que en esto hay alguna diferencia en los autores que deste caso escriben; aunque, por conjeturas verosímiles, se deja entender que se llamaba Quejana. Pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.	Kapare argi Kixote Mantxako Orbanaren leku batean, ez dudanez erabaki nahi zeinen izenen, bizi zen denbora asko ez dago lantzak kapare bat ontzian, adarga zaharra, lasterkaria zamari gozozale eta argala. Zerbaiten vaco gehiago aharia, eltze bat gau gehiagotan, haragia dueluak eta gallerak larunbatak, Ostiralean lanteja-ak, gehigarriaren igandeak, haitz-usakumeren bat haren haziendaren 3 parte suntsitzen zuten. Gainerako della-k kota amaitzen zuten esna egon zure burua izatera, jaientzat bilotsuaren galtzak, mesmo-aren haren txinalekin, eta entresemana-aren egunak mehe gehiagoren haren vellorí-arekin ohoratzen zen. Haren etxean zuen pasatzen zen 40tako etxeakoandre bat, eta iloba bati ez zen iristen 20ri, eta mendiaren ezkongai bat eta plaza, honela inauskaia hartzen edun zamaria zelatzen zuela. Gure kaparearen adina frisar-tzen zuen 50 urterekin; gorpuzkera marduletik zen, haragien lehorra, aurpegiaren argala, goiztiar handia eta ehizaren adiskidea. Esan nahi dute zuen Masailezurraren ezizena, edo Quesada, honetan diferentziaren bat dagoela egileengan eman dezazun kasua idazten dutela; , egiantzeko uste, ulertzea uzten da Kexaa deitzen zuela. Baina honek gutxi axola da gure ipuinera; nahikoa da dél kontaketan egiaren puntu bat ez dadila atera.
---	--

Taula A.8: Itzulpen-adibideak. Literatura

B. ERANSKINA

Parole kategorija-sistema

C. ERANSKINA

EDBL kategorija-sistema

Tesi hau
idazten bukatu zen
2007ko uztailaren 23an.

*eta hala bazan, ez bazan,
sar dadila kalabazan
eta geldi dadila betirako bertan.*