eman ta zabal zazu

EUSKAL HERRIKO UNIBERTSITATEA

Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Doktoretza-tesia

Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala

Uxoa Iñurrieta Urmeneta

2019

EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala

Uxoa Iñurrieta Urmenetak Itziar Aduriz Agirreren eta Gorka Labaka Intxausperen zuzendaritzapean eginiko tesi-txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2019ko iraila.

Hizkuntza bakoitza da, osorik, esapide berezi bat (esapide-sareen palinpsesto bat, zehazkiago esanik, ezarian bezain etengabe aldatuz doana). Horregatik da hain zabala hizkuntzen ikasbidea, dela fraseologia hori itsatsirik daramaten bizipen askotarikoetan murgildurik zuzenean, dela gramatikaren pausoz pausoko zeharkabide analitikoan.

Horregatik da bereziki delikatua, hain zuzen, esapidezko sare zabal horietako batean emanik datorrena beste batean ematea. (...) Lan ekilibrista, teorian ezinezkoa lirudikeen horri esaten diogu itzulpen, hain zuzen.

Juan Garzia

Etxekoei

Esker ona

Eskerrik asko, bihotzez...

...umetatik hizkuntza-sena garatzen lagundu didazuenoi: **aitari**, hizkuntza-kontuekiko kezka nigan pizteagatik, eta irakaspenik baliagarrienak eskolaz kanpo jasotzeko aukera emateagatik; **amari**, lorpen txiki orotan aurkitzeagatik alabaz harro sentitzeko arrazoiak, eta nekaezin jarduteagatik nire lan ulergaitzak ulertzeko ahaleginean; eta **nebari**, munduari (eta, hortaz, azken urteotan nire munduaren parte handi izan den tesi-lanari) ikuspegi kritikoxeagoz begiratzen irakasteagatik.

...azken urteak nirekin *itzulpenetan kolokatuta* pasatu dituzuenoi: **Itziarri**, Qatarlunian egonik ere beti gertu egoteagatik hizkuntza-eztabaidetarako, tesiaren ilunak argitzeko eta, tarteka, garagardoa eskuetan hartu eta tesia alde batera uzteko; **Gorkari**, nire garunari falta zaion alderdi konputazionala betetzen laguntzeagatik, zeu izan baitzara lan honetan iparrorratz; **Arantzari**, halako lanek behar duten ikuspegi globala eta babesa emateagatik eta, behar izan denean, presio pixka bat ere egin izanagatik; eta **Kepari**, lan guztiei beti alderdi positiboa ikusteagatik eta, nola ez, urteotan inkondizionalki bete izanagatik taldeko bromazalearen papera.

...txosten hau bukatutzat(-edo) emateko giltzarri izan zareten adituoi: **Igoneri** eta **Rubeni**, zuen begi zoliez tesi honetako orrialderik zailenak hobetzen lagundu didazuelako; **Norari**, ingelesezko bertsioa fintzeko (eta, hala, ni lasaiago uzteko) hartu duzun lanagatik; eta **Anttoni**, beste hamaika zereginen artean beti bilatu duzulako tartea nire zalantzei eta laguntza-eskaerei taxuz erantzuteko.

...nola edo hala lan honi ekarpenen bat egin diozuen **Ixakide eta Ixakide ohi guztioi**, bai frikikeria linguistikoetan konplize izan zaituztedanoi (bereziki, **Ainarari**), eta bai ulertzen lagundu didazuenoi *if* eta *while* badirela zerbait lokailuez harago; halako talde baten parte izanik, arinxeagoak dira lanik nekezenak ere.

...**sailkideei**, aukera eman didazuelako probeten, xiringen eta engranajeen artean komunikazio-irakaskuntzako lehen urratsak egiteko; batez ere, **Maxuxi** eta **Izaskuni**, irakaskuntza eta tesigintza uztartze horretan adoreemaile jardun duzuelako azken txanpan.

...**EIZIEko lagunei**, hilean behin behintzat mundu konputazionaletik irtenarazten eta letren mundura itzularazten nauzuelako, horrek ere laguntzen baitu perspektiba pixka bat hartzen.

...pisukide ohiei: **Olatzi**, *Python*eko ateak ez ezik etxekoak ere sarri samar ireki behar izan dizkidazulako urteotan, eta **bi Maddiei**, Gasteizen elkartu ginenetik ezin hobetuzko bidaideak izan zaretelako niretzat, une onetan eta ez hain onetan.

...**Lizeoko zazpikoteari**, nerabezaroan elkarri lagundu geniolako zer bide aukeratu erabakitzen, eta oraindik ere elkarri laguntzen diogulako hartutako bideekiko gorabeherei aurre egiten.

...eta, ezin bestela, **Mikeli**, nire konpañero paregabeari, estres-momentuetako purrustadak ere lasaitasunez hartzeko duzun gaitasun ikaragarriagatik. Asko zor dizu tesi honek (eta, batez ere, tesigilearen osasunak!).

Lan hau Ekonomia eta Lehiakortasun Ministerioaren diru-laguntza bati esker (BES-2013-066372) egin da, SKATeR proiektuaren barruan (TIN2012-38584-C06-02).

Gaien aurkibidea

Gaien aurkibidea **ix**

Taulen zerrenda **xiii**

Irudien zerrenda **xvii**

1 Tesi-lanaren nondik norakoak **1**
1.1 Sarrera eta motibazioa 1
1.2 Lanaren kokapena 4
1.3 Helburuak eta hipotesiak 6
1.4 Tesi-txostenaren egitura eta argibideak 8
1.5 Argitalpenak 11

2 Unitate Fraseologikoak: oinarri teorikoak eta tratamendu konputazionala **15**
2.1 UFen oinarri teorikoak 15
2.1.1 UFen definizioa eta hizkuntza-ezaugarriak 16
2.1.2 Sailkapenak 22
2.1.3 Aditz-UFak eta haien ezaugarriak 31
2.1.4 Fraseologia eta itzulpengintza 37
2.2 UFak Hizkuntzaren Prozesamenduan 44
2.2.1 Erronkak 44
2.2.2 Erauzketa 47
2.2.3 Identifikazioa 51
2.2.4 Itzulpen automatikoa 57
Definizio laburren bilduma 69

3 Prestaketa-lana **71**
3.1 *Matxin*en erroreen analisia 72
3.2 *Elhuyar* hiztegiaren gaineko azterketa 75
3.2.1 Erauzitako hitz-konbinazioen ezaugarriak 77
3.2.2 Erauzitako ordainen ezaugarriak 82
Laburpena 97

4 Gaztelaniazko konbinazioen azterketa eta identifikazioa **99**
4.1 Eskuzko azterketa xehea 99
4.1.1 Ezaugarri lexiko-semantikoak: idiomatikotasunaren *continuuma* 100
4.1.2 Ezaugarri morfosintaktikoak 105
4.2 Lehen identifikazio-esperimentua 110
4.2.1 Erabilitako baliabideak eta alderatutako metodoak . . 110
4.2.2 Emaitzak 114
4.3 Azterketa xehearen ingeleserako aplikagarritasuna 117
4.3.1 Ingelesezko UFen azterketa 118
4.3.2 Ingelesezko identifikazio-esperimentua 123
4.4 Azterketa erdiautomatikoa 126
4.4.1 Informazio-bilketa corpus elebakarretatik (1. urratsa) . 127
4.4.2 Hautagaiak patroika sailkatzea (2. urratsa) 131
4.4.3 Sailkapena fintzea, corpus paraleloetan begiratuta (3. urratsa) 135
4.5 Bigarren identifikazio-esperimentua 139
4.5.1 Erabilitako baliabideak eta metodoak 140
4.5.2 Emaitzak 141
Laburpena 145

5 Euskarazko ordainen azterketa eta itzulpen automatikoa **147**
5.1 Eskuzko azterketa xehea 147
5.1.1 Ezaugarri lexikoak 148
5.1.2 Ezaugarri morfosintaktikoak 149
5.2 Lehen esperimentua *Matxin* itzultzailean 152
5.2.1 Informazio linguistikoa integratzeko proposamena . . . 152
5.2.2 Emaitzak 156
5.3 Ordainen azterketa erdiautomatikoa 161
5.3.1 Ordainen corpusetatiko erauzketa (4. urratsa) 162

5.3.2 Ordainen inguruko informazioaren erauzketa corpus elebakarretatik (5. urratsa) . . . 165
5.3.3 Ordainen patroikako multzokatzea (6. urratsa) . . . 168
5.4 Bigarren esperimentua *Matxin* itzultzailean . . . 170
5.4.1 Erabilitako metodologia . . . 170
5.4.2 Emaitzak . . . 173
Laburpena . . . 179

6 Konbitzul datu-basea 181
6.1 Kontsulta-tresnaren deskribapena . . . 182
6.2 Datu-basearen arkitektura . . . 184
6.3 Funtzionalitateak . . . 185
Laburpena . . . 187

7 Euskarazko aditz-UFak corpusean 189
7.1 Euskara PARSEMEren corpusean . . . 190
7.1.1 PARSEMEren gidalerroak . . . 190
7.1.2 Etiketatze-metodologia orokorra . . . 196
7.1.3 Corpus etiketatua eta handik ateratako zenbait ondorio 198
7.1.4 Euskarazko kasu nahasgarriak eta haiekiko erabakiak . 200
7.1.5 Gidalerroak hobetzeko proposamenak . . . 203
7.2 UFen agerpen literalak . . . 205
7.2.1 Kontzeptu nagusiak: UFen agerpen literalak eta kointzidentziazkoak . . . 206
7.2.2 Metodologia orokorra . . . 208
7.2.3 Etiketatze-lana eta gidalerroak . . . 210
7.2.4 Emaitza orokorrak . . . 212
7.2.5 Ondorio linguistikoak . . . 215
Laburpena . . . 219

8 Ondorioak, ekarpenak eta etorkizuneko lanak 221
8.1 Ondorioak . . . 221
8.2 Ekarpenak . . . 224
8.3 Etorkizuneko lanak . . . 226

Bibliografia 229

Eranskinak 255

A Datu multzoen argibideak **255**

B Fraseologia-baliabideak **259**
Hiztegi fraseologikoak 259
Corpus-bilatzaileak 265

C Itzulpen automatikoa ebaluatzeko gidalerroak **275**

Taulen zerrenda

3.1 Euskarazko konbinazioen kasu- eta postposizio-markak (*Elhuyar* hiztegia) 78
3.2 Gaztelaniazko konbinazioen egitura morfologikoak (*Elhuyar* hiztegia) 79
3.3 Euskarazko konbinazioetako aditzik ohikoenak (*Elhuyar* hiztegia) 80
3.4 Gaztelaniazko konbinazioetako aditzik ohikoenak (*Elhuyar* hiztegia) 81
3.5 Euskarazko konbinazioetako izenik ohikoenak (*Elhuyar* hiztegia) 82
3.6 Gaztelaniazko konbinazioetako izenik ohikoenak (*Elhuyar* hiztegia) 82
3.7 Gaztelaniazko ordain motak (*Elhuyar* hiztegia) 84
3.8 Gaztelaniazko ordain motarik ohikoenak euskarazko markaren arabera (*Elhuyar* hiztegia) 85
3.9 Euskarazko ordain motak (*Elhuyar* hiztegia) 87
3.10 Euskarazko ordain motarik ohikoenak gaztelaniazko egituren arabera (*Elhuyar* hiztegia) 89
3.11 Mugatasunaren eta numeroaren ezaugarriak gaztelaniaren eta euskararen artean parekatzeko proposamena 91
3.12 Mugatasuna eta numeroa euskaratik gaztelaniara (*Elhuyar* hiztegia) 92
3.13 Mugatasuna eta numeroa gaztelaniatik euskarara (*Elhuyar* hiztegia) 93
3.14 Izenen eta aditzen baliokidetza *Elhuyar* hiztegian 94

4.1 Anotatze lexiko-semantikoan lortutako adostasuna 104

4.2 Gaztelaniazko anotatze lexiko-semantikoan lortutako adostasun-matrizea . . . 105
4.3 Eskuzko sailkapen morfosintaktikorako erabaki-taula . . . 109
4.4 Gaztelaniazko anotatze morfosintaktikoan lortutako adostasun-matrizea . . . 109
4.5 Lehen identifikazio-esperimentuaren emaitzak . . . 115
4.6 PARSEMEren ataza partekatuko lehen edizioan, gaztelaniaz eta hizkuntza erromanikoen multzoan oro har izandako doitasun-markak . . . 116
4.7 PARSEMEren ataza partekatuko bigarren edizioan, gaztelaniaz eta hizkuntza guztietan oro har izandako doitasun-markak 117
4.8 Ingelesezko anotatze lexiko-semantikoan lortutako adostasuna 121
4.9 Ingelesezko anotatze lexiko-semantikoan lortutako adostasun-matrizea . . . 122
4.10 Ingelesezko anotatze morfosintaktikoan lortutako adostasun-matrizea . . . 123
4.11 Ingelesezko identifikazio-esperimentuaren emaitzak . . . 124
4.12 PARSEMEren ataza partekatuko bigarren edizioan, ingelesez eta hizkuntza guztietan oro har izandako doitasun-markak . . 126
4.13 *Elhuyar* hiztegiko hitz-konbinazioen patroi morfosintaktikoak . 132
4.14 *DiCE*ko hitz-konbinazioen patroi morfosintaktikoak . . . 133
4.15 Lehen sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κ-ren arabera . . . 134
4.16 Bigarren sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κren arabera . . . 136
4.17 Eskuz eta erdiautomatikoki landutako konbinazioen patroi morfosintaktikoak, konbinazio gehien dituztenetatik gutxien dituztenetara . . . 138
4.18 Bigarren identifikazio-esperimentuaren emaitzak . . . 142
4.19 PARSEMEren ataza partekatuko bigarren edizioan, gaztelaniaz eta hizkuntza guztietan oro har izandako emaitzak . . . 142
4.20 PARSEMEren ataza partekatuko bigarren edizioan gaztelaniaz izandako emaitzak, UF motaka . . . 143

5.1 *Matxin*en eginiko lehen esperimentuaren emaitzak, BLEU, NIST eta TER metriken arabera . . . 158
5.2 *Matxin*en eginiko lehen esperimentuaren emaitzak, ebaluatzailez ebaluatzaile . . . 159

5.3 *Matxin*en eginiko lehen esperimentuaren emaitzak, osotara, eskuzko ebaluazioaren arabera 160
5.4 Ordain-erauzketaren ebaluazioko emaitzak 165
5.5 Izena+aditza motako ordainen patroi morfosintaktikoak 168
5.6 Ordainen sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κ-ren arabera 169
5.7 *Matxin*en eginiko bigarren esperimentuaren emaitzak, BLEU, NIST eta TER metriken arabera 174
5.8 *Matxin*en eginiko bigarren esperimentuaren emaitzak, giza ebaluatzaileen arabera 176

7.1 Anotatze lexiko-semantikoan lortutako adostasuna 197
7.2 PARSEMEren euskarazko corpusaren datuak 198
7.3 Etiketen 100 esaldiko batezbestekoak, euskaraz, gaztelaniaz, frantsesez, ingelesez eta ataza partekatuko 20 hizkuntzetan oro har 199
7.4 Etiketatze-lanaren estatistika orokorrak, hizkuntza guztietan. Idiomatikotasun-tasa (*Idiomaticity rate*) honela kalkulatzen da: idiomatikoak/(literalak+idiomatikoak). 213
7.5 Idiomatikotasun-tasa zabaldua (*Extended Idiomaticity Rate*, EIR), kointzidentziazkotasun-tasa zabaldua (*Extended Coincidentality Rate*, ECR) eta literaltasun-tasa zabaldua (*Extended Literality Rate*, ELR). 214
7.6 Heuristikoen doitasuna, estaldura eta F neurria. 215

Irudien zerrenda

1.1 Tesi-lanaren egitura orokorra 8

2.1 Hitz-konbinazioen sailkapena, Howarthen (1996) arabera . . . 23
2.2 Hitz anitzeko unitateen sailkapena, Sag *et al.*en (2002) arabera 24
2.3 Unitate Fraseologikoen sailkapena, Corpas Pastorren (1996) eta Urizarren (2012) arabera 25
2.4 Izena+aditza konbinazioen sailkapena, Gurrutxagaren (2014) arabera 28
2.5 Aditz-UFen sailkapena, PARSEMEren corpusean (Savary *et al.*, 2018) 29
2.6 UFen prozesamenduaren eta bi aplikazioren arteko eragina . . 46
2.7 *Matxin* itzultzailearen arkitektura orokorra 63
2.8 *Matxin*en itzulpen-prozesuaren adibide bat: analisi-fasea . . . 64
2.9 *Matxin*en itzulpen-prozesuaren adibide bat: transferentzia-fasea 65
2.10 *Matxin*en itzulpen-prozesuaren adibide bat: sorkuntza-fasea . 66

3.1 *Elhuyar* hiztegitik konbinazio-zerrendak sortzeko prozesuaren adibide bat 76
3.2 Euskarazko konbinazioen kasu- eta postposizio-markak (*Elhuyar* hiztegia) 79
3.3 Gaztelaniazko konbinazioen egitura morfologikoak (*Elhuyar* hiztegia) 80
3.4 Gaztelaniazko ordain motak (*Elhuyar* hiztegia) 84
3.5 Gaztelaniazko ordain motarik ohikoenak euskarazko markaren arabera (*Elhuyar* hiztegia) 86
3.6 Euskarazko ordain motak (*Elhuyar* hiztegia) 88

3.7 Euskarazko ordain motarik ohikoenak gaztelaniazko egituraren arabera (*Elhuyar* hiztegia) 88
3.8 Mugatasuna eta numeroa euskaratik gaztelaniara (*Elhuyar* hiztegia) 92
3.9 Numeroa gaztelaniatik euskarara (*Elhuyar* hiztegia) 93
3.10 Izenen eta aditzen baliokidetza *Elhuyar* hiztegian 95

4.1 Idiomatikotasunaren continuuma, sailkapen-proposamenaren arabera 103
4.2 *Estar en forma* UFa duten hiru esaldiren dependentzia-analisia, Freeling 4.1en arabera. Hiru hitzen arteko dependentzia sintaktikoen zentzua biribilduta dago. 113
4.3 Lehen identifikazio-esperimentuaren emaitzak: identifikatutako konbinazioen ehunekoak, metodoaren arabera 114
4.4 Oxford Collocations Dictionary-ko sarrera baten adibidea: *connection* izenaren lehen adiera. 119
4.5 Ingelesezko identifikazio-esperimentuaren emaitzak: identifikatutako konbinazioen ehunekoak, metodoaren arabera 124
4.6 Azterketa linguistikoa erdiautomatizatzeko metodoa (identifikazioari dagokion zatia nabarmenduta) 128
4.7 Bigarren urratsean sortzen diren taulen adibide bat (zenbakiak, ehunekotan) 130
4.8 Lehen sailkapen-prozesuaren adibide bat 133
4.9 Bigarren sailkapen-prozesuaren adibide bat 135

5.1 *Matxin*en UFei buruzko informazioa integratzeko metodologia-proposamena 153
5.2 *Echar una siesta* UFaren itzulpen-prozesua: analisia eta transferentzia lexikoa 154
5.3 *Contraer matrimonio* UFaren itzulpen-prozesua: analisia eta transferentzia lexikoa 155
5.4 *Mantener el equilibrio* UFaren itzulpen-prozesua: analisia eta transferentzia osoa (lexikoa eta estrukturala) 156
5.5 *Mantener el equilibrio* UFaren itzulpen-prozesua: analisia eta transferentzia osoa (lexikoa eta estrukturala) 157
5.6 Azterketa linguistikoa erdiautomatizatzeko metodoa (itzulpenari dagokion zatia nabarmenduta) 161

5.7 Ordain-hautagaiak ateratzeko erabili dugun metodologiaren adibide bat . . . 163
5.8 Ordainen aukeraketaren adibide bat . . . 164
5.9 Bosgarren urratsean sortzen diren taulen adibide bat (zenbakiak, ehunekotan) . . . 167
5.10 Seigarren urratsean sortzen diren taulen adibide bat . . . 169

6.1 *Konbitzul* datu-basearen orri nagusia eta bilaketa-barra . . . 182
6.2 *Konbitzul*en gaztelaniazko *pelo* izena bilatuta erakusten diren UFak eta ordainak. . . . 183
6.3 *Konbitzul*en *poner en peligro* UFari eta *arriskuan jarri* ordainari dagokion informazio-taula. . . . 184

7.1 Heuristikoen arabera *arrastoa utzi* UFaren agerpen literalak izan litezkeen lau adibide. Dependentzia-erlazioen azalpenak: *acl*, adjektibo-perpausa; *dobj*, objektu zuzena; *nsubj*, izen-subjektua. Etiketa horiei buruzko informazio gehiago, Aranzabe *et al.*-en lanean (2019). . . . 209

B.1 Mokoroaren sareko hiztegian *arbola* bilatuta lortzen den emaitzaren zati bat . . . 261
B.2 Intza proiektuaren hiztegian *akatsa* kontzeptuan jasotako lokuzio-zerrenda . . . 262
B.3 Intza proiektuaren hiztegian *atzean zuloa ukan* lokuzioari dagokion fitxa . . . 262
B.4 Elhuyar web-corpusa: *amets*+aditza bilaketaren emaitza. . . . 267
B.5 Elhuyar web-corpusa: *amets egin* konbinazioaren adibideak. . 268
B.6 ETCn *amets* bilatuta Konbinatoria atalean lortzen den emaitza (aditzak bakarrik hautatuta). . . . 269
B.7 CORPESeko Concordancia atala: *inspirar+confianza* bilaketaren emaitza. . . . 270
B.8 CORPESeko Coapariciones atala: *admiración* bilaketaren emaitza. . . . 271
B.9 Elhuyar web-corpus paraleloa: *dar+confianza* bilaketaren emaitza. . . . 272
B.10 TextReference: *sin embargo* bilaketaren emaitza. . . . 273

1. KAPITULUA

Tesi-lanaren nondik norakoak

1.1 Sarrera eta motibazioa

Hiztunok txiki-txikitatik ikasten dugu, ia oharkabean, hitzak nola konbinatzen diren gure lehen hizkuntzan. Ikasten dugu zer *behar dugun* eta zer *nahi dugun* esaten, albokoari kontatzen nola *dugun izena* eta non *bizi garen*, eta norbaitek *eskerrak ematen* dizkigunean *ez horregatik* erantzuten zaiola, besteak beste. Konturatu ere ez gara egiten guretzat hain arruntak diren hitz batzuk ez ditugula edonola erabiltzen hizketan ari garenean, baizik eta beti beste hitz jakin batzuekin batera eta modu jakinetan.

Sarritan, beste hizkuntzaren bat ikasten hasten garenean jabetzen gara halako hitz-konbinazioen berezitasunez, hanka-sartzeren bat egin ondorenedo: norbaitek jakinarazten digunean gaztelaniaz ez dela naturala *sacar ruido* esatea eta, *zarata* euskaraz *atera* egiten bada ere, gaztelania-hiztunek *meter ruido* erabiltzen dutela normalean; edo, beharbada, ohartzen garenean ingeles-hiztunek *make mistakes* eta *do homework* erabiltzen dituztela, baina ez **do mistakes* eta **make homework*, euskaraz *egin* bakarrak laguntzen dien arren *akatsei* eta *etxeko lanei*, bi-biei.

Hitz-konbinazio horiek guztiak hizkuntzen fraseologiaren parte dira, eta Unitate Fraseologiko (UF) esaten zaie. Etengabe agertzen dira gure ahozko nahiz idatzizko jardunean, eta, hizkuntza bat zenbat eta hobeto hitz egin, orduan eta naturalago erabiltzen ditugu, orduan eta gehiago eta orduan eta konplikatuagoak. Ingelesa ikasi ahala, bereizten hasten gara zer den *take*

out ('atera'), zer *take off* ('aireratu'), zer *take up* ('ekin') eta zer *take after* ('antza izan'); lagunei idaztean, *keep in touch* jartzen diegu mezuen amaieran, 'jarrai dezagun harremanetan'; eta ahozko aurkezpen batean denboraz larri bagabiltza, *running out of time* ari garela esaten dugu, 'denborarik gabe geratzen' ari garela, alegia.

Hizkuntza bakoitzak bere UFak ditu, eta askotan ezin izaten dira hitzez hitz itzuli beste hizkuntza batzuetara, bereziki bi hizkuntzak oso desberdinak badira elkarren artean.

(1) EU: ***hala ere***
ES: ***sin embargo*** → eta ez *así también*

(2) EU: ***adarra jo***
ES: ***tomar el pelo*** → eta ez *tocar el cuerno*

Hiztunok pixkanaka ikasten ditugu halakoak, batzuetan entzunaren entzunez eta erabiliaren erabiliaz, eta beste batzuetan berariaz horretara jarrita, irakasleari adituz edo hiztegiak eta hizkuntza-liburuak eskuartean ditugula. Konturatu gabe ikasten dugu ez dugula lehen hizkuntzan hitzez hitz pentsatu behar beste hizkuntza batzuetan hitz egiteko. Hizkuntza-tresna aurreratuek, ordea, zailtasun gehiago izaten dituzte halakoak prozesatzeko.

UFek eragin handia izan dezakete hizkuntza-tresnen emaitzen kalitatean, konbinazioko osagai-hitzak batera tratatu ezean ezin baitira askotan testuak behar bezala interpretatu. Lan hori, ordea, erronka handia da Hizkuntzaren Prozesamendurako (HP), eta horren adibide dira itzultzaile automatikoen zenbait emaitza trakets, hala nola *Matxin* itzultzaile automatikoaren honako hauek[1]:

(3) ES: *No* ***correremos*** *ese* ***riesgo***.
Matxin: *Arrisku hori ez dugu korrika egingo.*
EU-zuz: *Ez dugu* ***arrisku*** *hori* ***hartuko***.

(4) ES: *No* ***plantó cara***. *No* ***plantó*** *nada de* ***cara***.
Matxin: *Ez zuen* ***aurpegi eman***. *Parez pare ezer ez zuen landatu.*

[1] Adibideetako markaketari buruzko azalpenak 1.4. atalean emango ditugu. Kasu hauetan, lehen lerroan jarri dugu itzulgaia (ES), bigarrenean itzulpen automatikoa (Matxin), eta hirugarrenean euskarazko itzulpen zuzen posible bat (EU-zuz); beltzez markatutakoak UFen barruko hitzak dira, eta azpimarratutakoak, UF ez direnak baina nabarmendu nahi direnak.

EU-zuz: *Ez zuen* ***aurpegi eman****. Ez zuen batere* ***aurpegirik eman****.*

(5) ES: ***Se buscan la vida*** *como pueden.*
Matxin: ***Bizimodua ateratzen*** *dira ahal duten bezala.*
EU-zuz: *Ahal duten bezala* ***ateratzen*** *dute* ***bizimodua****.*

Itzultzaile automatiko estatistikoek eta neuronalek corpusak dituzte oinarrian, eta, hortaz, zeharka bada ere, ikasten dute hitzak nola konbinatu ohi diren hizkuntza batean eta bestean. Erregeletan oinarritutako sistemek, aldiz, ez dute izaten arau linguistiko orokorretatik eta hiztegi elebidunetatik haragoko informaziorik normalean, eta hortik sortzen dira goiko adibideetako itzulpen okerrak (3–5) eta antzeko beste asko. Izan ere, halako sistemek bi ataza nagusiri egin behar diete aurre fraseologiari dagokionez: batetik, itzulgaian UFak identifikatzeari, eta bestetik, UF horiei ordaina emateari.

Identifikazio-lanerako, jakin behar da hitz-konbinazio bat noiz identifikatu behar den UF gisa eta noiz ez. Lehenik, baliabideak sortu behar dira sistemak zer identifikatu jakin dezan, UF jakin bat itzultzaile automatikoaren hiztegian ez badago ez baita inoiz UFtzat hartuko. Hori gertatzen da, adibidez, 3. adibidean: *Matxin* itzultzaile automatikoaren hiztegian ez dago *correr riesgo* sarrerarik, eta, ondorioz, itzulgaia hitzez hitz ekartzen da euskarara. Bigarrenik, UFen ahalik eta agerpen gehien identifikatu nahi badira, jakin egin behar da UF horiek testuetan nola agertzen diren ere. Izan ere, UF askok hainbat aldaki dituzte, eta identifikazio-metodo konplexuak behar dira aldaki horiek guztiak ezagutzeko. Horixe da, hain zuzen, 4. adibideko lehen esaldia ondo eta bigarrena gaizki itzultzearen arrazoia. Bigarren esaldi horretan, *plantar cara* UFaren osagai-hitzak ez daude bata bestearen jarraian, eta ez da UFaren agerpenik aurkitu; *de* eta *cara* hitzak bata bestearen alboan egonik, aldiz, *de cara* UF gisa identifikatu da, nahiz eta testuinguru horretan UFa ez izan.

Aditza buru sintaktikotzat duten UFak bereziki malguak izan ohi dira, eta, hitz-forma aldakorrak izateaz gain, maiz agertzen dira tartean beste hitz batzuk dituztela ere, edo hitz-hurrenkera aldatuta (6. adibidea). Horrek zaildu egiten du identifikazio-lana, eta horregatik da bereziki beharrezkoa kasu horietan informazio linguistikoa kontuan hartzen duten metodoak erabiltzea.

(6) ***Erabakia hartu*** *zuten.*
Erabaki *garrantzitsuak* ***hartu*** *zituzten.*
Hartutako erabakien *berri eman zuten.*

Bestetik, itzultzaile automatikoek xede-hizkuntzako ordaina eman behar diete identifikatutako UFei, eta hori ere ataza korapilatsua da erregeletan oinarritutako sistementzat. Osagaiak zein bere aldetik itzultzea ez da nahikoa eta, hortaz, UF osoari zer ordain eman jakin beharra dago. Hala ere, hiztegi elebidunetan jasotako UFen kopurua mugatua da (3. adibidea), eta, gainera, ez da inon jasotzen ordainak behar bezala erabiltzeko informazio linguistikorik. Gorago erakutsi dugun 5. adibidean, esaterako, *Matxin* itzultzaile automatikoak oker euskaratu du gaztelaniazko esaldia, *buscarse la vida* UFa hiztegian egon arren, ez baitaki *bizimodua atera* ordaina nola erabili zuzen esaldian.

Tesi-lan honetan, aditza+izena motako UFen azterketa linguistiko sakon bat egin dugu, ikusteko, batetik, zer ezaugarri lexiko-semantiko eta morfosintaktiko dituzten eta, bestetik, nola itzultzen diren gaztelaniaren eta euskararen artean. Ondoren, informazio linguistiko hori baliatu dugu fraseologiak HPko tresnetan sortzen dituen arazoak konpontzen laguntzeko. Erregeletan oinarritutako itzultzaile automatiko bat hartu dugu oinarritzat, eta, horren bidez, UFen identifikazioa eta itzulpena hobetzeko metodo bana garatu dugu.

Ikerketa-lan nagusi horrekin batera, bi baliabide fraseologiko ere sortu ditugu: aztertu ditugun UFak, ordainak eta informazio linguistikoa biltzen dituen datu-base bat, eta fraseologia mailan etiketatutako euskarazko corpus bat, beste hogei hizkuntzatako corpusen irizpide berberei jarraituz landua. Horrez gain, corpus etiketatu hori erabilita, UFen agerpen literalak ere etiketatu eta aztertu ditugu. Lan horien guztien inguruan jardungo dugu txosten honetan.

1.2 Lanaren kokapena

Tesi-lan hau Hizkuntzaren Prozesamenduan (HPn) kokatzen da, Ixa taldearen jardunean zehazki. Euskal Herriko Unibertsitateko ikerketa-taldea da Ixa, eta HPko hainbat azpiarlotan aritzen da duela hiru hamarkadatik, bai euskarazko hizkuntza-tresna aurreratuak sortzen, bai nazioarteko ikerketa-proiektuetan parte hartzen eta beste hizkuntza batzuetarako ere produktuak garatzen.

HPren alorrean, ikerketa-ildo nagusietako bat itzulpen automatikoa izan da azken urteotan, eta hala sortu dira, besteak beste, *Matxin* (Mayor *et al.*, 2009), *EUSMT* (Labaka, 2010) eta *MODELA* (Etchegoyhen *et al.*, 2018) itzultzaile automatikoak. Gure lanak lotura zuzena du sistema horietako

lehenarekin eta, era berean, fraseologia konputazionalarekin. Izan ere, tesi-lanaren muina UFen azterketa bat da: gaztelaniazko eta euskarazko aditza+izena motako UFak hartu ditugu oinarritzat, eta hizkuntza batetik bestera nola aldatzen diren aztertu dugu, ikusteko ea itzulpen-kalitatea hobetu daitekeen UFen informazio xehearen bidez.

Orain arte ere egin da lanik euskaraz UFen tratamendu konputazionalari dagokionez. Batetik, Urizarrek (2012) euskarazko lokuzioak aztertu zituen bere tesi-lanean, eta 2.207 lokuzioren zenbait ezaugarri deskribatu zituen Euskararen Datu Base Lexikalean (EDBLn; Aduriz *et al.*, 1998): zer murriztapen morfologiko dituzten, zer hurrenkera-aldaketa posible, eta murriztapen horiek betetzen dituztenean anbiguoak diren ala ez. Gainera, HABIL tresna ere sortu zen tesi-lan horren baitan, EDBLko datu horiek erabiliz lokuzioak corpusetan identifikatzeko.

Bestetik, Gurrutxagak metodologia bat proposatu zuen izena+aditza konbinazioak corpusetatik erauzteko eta sailkatzeko (Gurrutxaga, 2014). Neurri estatistikoen eta hainbat proba linguistikoren bidez, corpusetan bilatzen ditu elkarrekin agertzeko joera handia duten hitz-konbinazioak, eta hiru multzotan sailkatzen: lokuzioak, kolokazioak eta konbinazio libreak. Helburu lexikografikoetara bideratuta dago lan hori, batez ere, eta metodologia horixe da *Elhuyar Web Corpus*etan hitz-konbinazioak bilatzeko erabiltzen dena[2].

Itzulpen automatikoari dagokionez, ordea, ez da UFen azterketa sakonik egin. *Matxin* itzultzaileak, erregeletan oinarritutako sistema izanik, arau linguistikoak eta hiztegi elebidun bat darabiltza itzulpenak egiteko. Hiztegi elebidunak badauzka hitz batez baino gehiagoz osatutako sarrera batzuk, eta analizatzaile morfosintaktikoak ere badu haiek identifikatzeko modulu bat. Hala ere, datorren kapituluan azalduko dugunez, bai identifikazio-metodologia eta bai itzulpenerakoa oso mugatuak dira, eta muga horiek zabaltzera dator tesi-lan hau batez ere, bidean beste ekarpen batzuk ere egin ditugun arren.

Horrez gain, lan honek zerikusia du bi ikerketa-proiekturekin: batetik, SKATeR proiektuarekin, horren baitan egin baita tesi-lanaren zatirik handiena, Ekonomia eta Lehiakortasun Ministerioaren doktoretza aurreko dirulaguntza bati esker; bestetik, PARSEME proiektu europarrarekin ere lotura zuzena du, hainbat bilera, lantegi eta ikastaro antolatu baititu 2013tik 2017ra bitartean, fraseologia konputazionaleko ikertzaileak biltzeko asmoz. Jarduera horietako batzuetan parte hartu dugu, eta, geroago azalduko dugunez,

[2] `http://webcorpusak.elhuyar.eus/cgi-bin/kolokatuak.py`

tesi-lan honetako zenbait eduki egitasmo horretarako edo haren haritik egin dira.

1.3 Helburuak eta hipotesiak

Doktoretza-tesi honen helburu nagusia da aditza+izena motako UFen azterketa linguistikoa egitea eta, informazio horren bidez, era horretako UFen tratamendu konputazionala hobetzea. Gaztelaniatik euskararako itzulpenean jarri dugu arreta batez ere, baina azterketaren zati handi bat aplikagarria da sistema elebakarretan ere. Honako hipotesi nagusi hau hartu dugu oinarritzat: informazio linguistiko xehea, lexikoa eta morfosintaktikoa batez ere, baliagarria dela UFen prozesamendua hobetzeko.

Dena dela, txosten honetan azalduko ditugun lanetan, azpihelburu eta azpihipotesi gehiago ere izan ditugu gogoan, eta horiek zerrendatuko ditugu jarraian. Era berean, ideia horietako bakoitza zein kapitulutan landu dugun ere zehaztuko dugu, eta kapituluen amaieran laburbilduko dugu abiapuntu-hipotesiak zuzenak ote ziren eta helburuak zenbateraino bete ditugun.

Helburuak

Hizkuntzaren azterketari dagokionez,

[H1] Gaztelaniazko eta euskarazko aditza+izena motako UFen ezaugarri lexiko eta morfosintaktikoak aztertzea.

→ 3., 4., eta 7. kapituluak

[H2] Aditza+izena motako UFak gaztelaniaren eta euskararen artean nola itzultzen diren aztertzea.

→ 3. eta 5. kapituluak

[H3] Beste hizkuntza batzuetan ere erabilgarriak izan litezkeen azterketa-metodologiak sortzea eta erabiltzea.

→ 7. kapitulua

Hizkuntzaren prozesamenduari dagokionez,

[H4] Aditza+izena motako UFen identifikazio automatikoa hobetzea.

→ 4. kapitulua

[H5] Aditza+izena motako UFen informazio linguistikoak itzulpen automatikoan zer eragin duen aztertzea.

→ 3. eta 5. kapituluak

Hizkuntza-baliabideei dagokienez,

[H6] Aditza+izena motako UFak eta haien ordainak biltzea –euskaraz eta gaztelaniaz– eta eskuragarri jartzea, Hizkuntzaren Prozesamendurako aplikagarriak diren datu linguistikoekin batera.

→ 6. kapitulua

[H7] Gaztelaniazko eta –bereziki– euskarazko UFen corpus etiketatuak sortzea.

→ 7. kapitulua

Abiapuntu-hipotesiak

Fraseologiari dagokionez,

[A1] UFak, askotan, ez dira hitzez hitz itzultzen hizkuntza batetik bestera.

→ 3. eta 5. kapituluak

[A2] Aditza+izena motako UFak oso malguak izan ohi dira morfosintaxiari dagokionez, baina murriztapenak ere badituzte.

→ 4. eta 7. kapituluak

[A3] Fraseologia mailan aztertutako hizkuntza askoren aldean, euskaraz bereziki ohikoak dira aditz arinak barne hartzen dituzten UFak.

→ 3. eta 7. kapituluak

[A4] Hitz-konbinazio asko literalak zein idiomatikoak izan badaitezke ere, praktikan ia beti erabiltzen dira idiomatikoki, hau da, UF gisa.

→ 7. kapitulua

Hizkuntzaren Prozesamenduari dagokionez,

[A5] UFen inguruko informazio lexiko eta morfosintaktikoa kontuan hartzeak haien identifikazioa hobetu dezake.

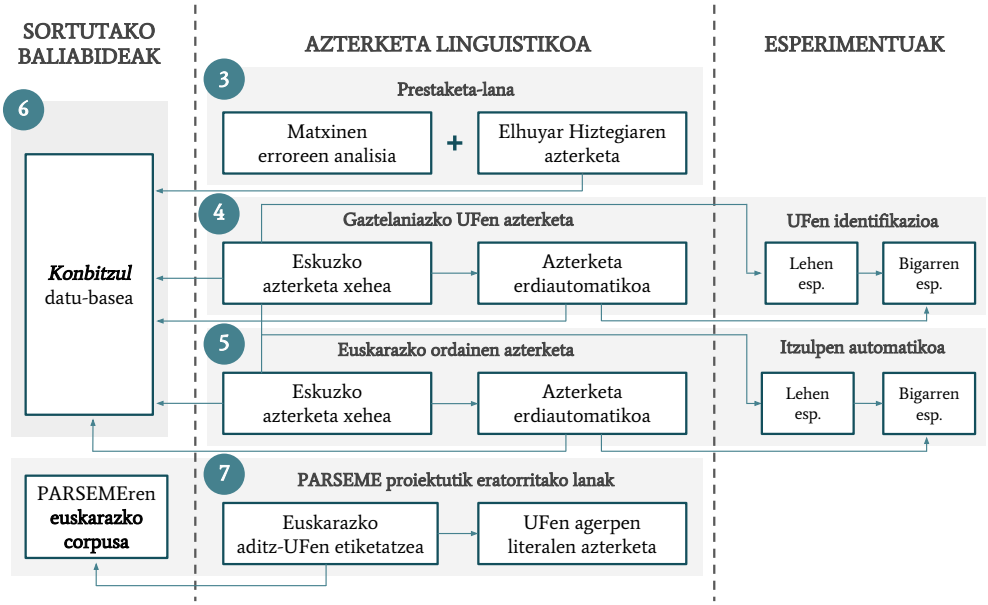
→ 4. kapitulua

[A6] UFei buruzko informazio morfosintaktikoa kontuan hartzea onuragarria izan daiteke itzultzaile automatikoentzat.

→ 3. eta 5. kapituluak

1.4 Tesi-txostenaren egitura eta argibideak

Tesi-txosten hau zortzi kapitulutan banatuta dago. Sarrerari, arloaren egoerari eta ondorioei dagozkien atalak alde batera utzita, gure lanaren egitura orokorra 1.1. irudian dago laburbilduta. Erdiko zutabean jarri ditugu azterketa linguistikoari dagozkion atalak, ezker-eskuin dituztelarik informazio horretatik sortutako baliabideak eta informazio hori erabiliz HPko tresnetan eginiko esperimentuak. Gezien bidez adierazi dugu atalak nola lotzen diren elkarren artean, eta zirkuluetan jarri dugu zer kapitulutan arituko garen zati bakoitzaren inguruan.



1.1 irudia – Tesi-lanaren egitura orokorra

Sarrera honen ostean, fraseologia konputazionalaren arloa gaur egun zertan den azalduko dugu (2. kapitulua): Hizkuntzalaritzaren esparruko fraseologia-

lanez jardungo dugu lehenik, eta Hizkuntzaren Prozesamenduan egin direnez ondoren.

Hurrengo bost kapituluetan aurkeztuko dugu gure ikerketa-lana bera. Hasteko, motibazioa indartze aldera eginiko bi azterketaren berri emango dugu 3. kapituluan: *Matxin* itzultzaile automatikoaren gainekoa bata, eta *Elhuyar* hiztegian oinarritua bestea.

Ondoren, 4. kapituluan, UFen ezaugarriez arituko gara, eskuz eta erdiautomatikoki egin ditugun azterketetatik abiatuta. Aztertutako informazioak UFen identifikazioan zer-nolako eragina duen ere erakutsiko dugu, bi esperimenturen berri emanez. Kapitulu horretan gaztelaniazko UFen azterketan eta identifikazioan jarriko dugu arreta batez ere, baina ingelesaren inguruko lan txiki bat ere aurkeztuko dugu, metodologia hori beste hizkuntza batzuetarako ere erabilgarria izan daitekeela erakusteko.

Hurrengo kapitulua, 5.a, UFen euskarazko ordainei eskainiko diegu. Gaztelaniazko UFekin egin bezala, eskuz eta erdiautomatikoki aztertu ditugu euskarazko ordainak ere, eta bi esperimentu egin ditugu informazio linguistiko hori guztia Hizkuntzaren Prozesamenduko tresnetan ebaluatzeko. Atal bana eskainiko diegu azterketei eta esperimentuei.

Aurreko lan horietatik ateratako informazio linguistiko guztia datu-base publiko batean gorde dugu, *Konbitzul*en, eta datu-base horren nondik norakoez jardungo dugu 6. kapituluan. Barruko egitura nolakoa den azaltzeaz gainera, interfazearen inguruko xehetasun batzuk ere emango ditugu.

Horrez gain, 7. kapituluan, PARSEME proiektu europarrarekin lotuta egin ditugun bi ekarpen aurkeztuko ditugu. Lehen azpiatalean, euskarazko corpus baten fraseologia mailako etiketatzeaz arituko gara, eta bigarrenean, berriz, corpus horretatik abiatuta UFen agerpen literalen gainean egin dugun lanaz.

Azkenik, tesi-lan osoaren ondorioei eta etorkizuneko lanei eskainiko diegu tartea, 8. kapituluan. Dena dela, atal bakoitzaren amaieran ere laburpen moduko bat egingo dugu, eta han azalduko dugu kapitulu bakoitzeko edukiak nola lotzen diren abiapuntuko hipotesiekin, bai eta helburuak betetzeko atalez atal zer pauso eman ditugun ere.

Lanaren garapen kronologikoa

Txostenari egitura hori eman badiogu ere, tesi-lanaren garapen kronologikoa ez da erabat halakoa izan, eta komeni da gogoan izatea laugarren eta bosgarren kapituluetan azaldutako lanak tartekatuta egin ditugula, hau da:

eskuzko azterketa xeheak eta haiei dagozkien esperimentuak egin ditugula lehenik, nola identifikaziokoa (4.1. eta 4.2. atalak) hala *Matxin*ekoa (5.1. eta 5.2. atalak), eta azterketa erdiautomatikoak eta haiekin batera egin ditugun esperimentuak geroago etorri direla, nola identifikaziokoa (4.4. eta 4.5. atalak) hala *Matxin*ekoa (5.3. eta 5.4. atalak). Ahalik eta garbien gera dadin datuak zer iturritatik eta zer hurrenkeratan bildu ditugun, datu multzoen eskema bat ere jarri dugu A. eranskinean.

Adibideen inguruko argibideak

Txosten honetako edukiei laguntzeko, adibide ugari txertatu ditugu datozen kapituluetan. Horietako asko tesi-lanean zehar erabili ditugun corpusetatik hartuak dira, zenbaiti moldaketaren bat egin diogun arren azalpenetara hobeto egokitzeko. Beste batzuk, berriz, geuk sortu ditugu, ez baitugu beti aurkitu eduki jakin batzuk argitzen laguntzeko adibide egokirik. Formatu berbera eman diegu zenbakitutako adibide guztiei, honela:

- Adibideak letra etzanez idatzi ditugu oro har, baina haien inguruko azalpenak, hizkuntza-kodeak eta abar, letra arruntez.

  (7) ES: *Ejemplo en castellano.*
  EU: *Euskarazko adibidea.* → Azalpena

- Letra lodiz markatu ditugu UFen parte diren hitzak. UF jakin bati buruzko azalpenak eman nahi izan ditugunean, adibidean UF horren osagaiak bakarrik markatu ditugu, nahiz eta esaldian UF gehiago ere egon.

  (8) *Txantxetan ari zen;* ***adarra jo*** *dizu.*
  (9) *Ederki* ***jo*** *dizu* ***adarra****!*

- Kasuan-kasuan markatu nahi genituen hitzak edo hitz-zatiak azpimarratu egin ditugu. Esate baterako, 10. adibidean determinatzailean jarri nahi da arreta, eta 11.ean partitibo-markan:

  (10) *Adar <u>bat</u> jo zuten.*
  (11) *Ez niri* ***adarr<u>ik</u> jo****!*

Ohar bedi idazkera hori hizkuntza bakoitzarekin lotuta eman dugula, hau da: hitz-konbinazio jakin bat UFa bada hizkuntza batean eta beste batean ez, adibide jakin horri dagokion hizkuntzari begiratu diogu nola markatu erabakitzeko. Txosten honetan UFen itzulpenaz jardungo dugunez, sarri samar agertuko dira adibide beraren barruan desberdin markatutako esaldiak, hizkuntza batean era batera eta beste batean beste era batera. Gisa horretakoa da, adibidez, ingelesezko UF bat Google Translate tresnak[3] nola itzultzen duen erakusten duen adibide hau:

(12) EN: *You **put** your **foot into** your **mouth**.*
Google: *Zure oina ahoan jarri duzu.*
EU-zuz: ***Hanka sartu** duzu.*

1.5 Argitalpenak

Tesi-lan honetako edukiak hainbat aldizkari, liburu eta kongresutan argitaratu ditugu, bai ingelesez eta bai euskaraz.

Aldizkari eta liburuetan argitaratutako artikuluak

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Learning about phraseology from corpora: a linguistically motivated method for Multiword Expression identification and translation**. Errebisio-prozesuan.

- Savary A., Cordeiro S.R., Lichte T., Ramisch C., Iñurrieta U., Giouli V. **Literal occurrences of multiword expressions: rare birds that cause a stir**. *The Prague Bulletin of Mathematical Linguistics*, 5–54. orr. 2019.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system**. *Multiword Units in Machine Translation and Translation Technologies*, 41–60. orr. John Benjamins Publishing Company. 2018.

[3] `https://translate.google.com/`

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Izen+aditz konbinazioen itzulpenaz eta tratamendu konputazionalaz**. *Senez* 47, 237–249. orr. 2016.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Izen+aditz konbinazioen azterketa elebiduna, hizkuntza-aplikazio aurreratuei begira**. *Linguamática* 6(2), 45–55. orr. 2014.

Kongresuetako argitalpenak

- Iñurrieta U. **Unitate Fraseologikoen agerpen literalak, *urre baina urri***. *IkerGazte: nazioarteko ikerketa euskaraz. Kongresuko artikulu-bilduma. Giza zientziak eta artea*, 139–147. orr. Baiona. 2019.

- Iñurrieta U., Aduriz I., Estarrona A., Gonzalez-Dios I., Gurrutxaga A., Urizar R., Alegria I. **Verbal Multiword Expressions in Basque corpora**. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, 86–95. orr. Santa Fe, New Mexico, AEB. 2018.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Konbitzul: an MWE-specific database for Spanish-Basque**. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC2018)*, 2500–2504. orr. Miyazaki, Japonia. 2018.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Aditza+izena konbinazioen itzulpen automatikoa, arau linguistikoen bidez**. *IkerGazte: nazioarteko ikerketa euskaraz. Kongresuko artikulu-bilduma. Giza zientziak eta artea*, 158–166. orr. Iruñea. 2017.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Rule-based translation of Spanish verb-noun combinations into Basque**. *Proceedings of the 13th Workshop on Multiword Expressions (at EACL2017)*, 149–154. orr. Valentzia, Espainia. 2017.

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K., Carroll J. **Using linguistic data for English and Spanish verb-noun combination identification**. *Proceedings of the 26th International*

Conference on Computational Linguistics (COLING2016): Technical Papers, 857–867. orr. Osaka, Japonia. 2016.

- Iñurrieta U. **Konbitzul: euskarazko eta gaztelaniazko izen+aditz konbinazioen datu-basea**. *IkerGazte: nazioarteko ikerketa euskaraz. Kongresuko artikulu-bilduma*, 32–38. orr. Durango. 2015.

Dibulgaziozko argitalpena

- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., Sarasola K. **Ez burua hautsi, Matxin!** *Elhuyar aldizkaria* 323, 49–51 orr. 2016.

2. KAPITULUA

Unitate Fraseologikoak: oinarri teorikoak eta tratamendu konputazionala

Kapitulu honetan, tesi-txosten honi dagokion marko teoriko-praktikoaz jardungo dugu, geroago azalduko ditugun edukiak testuinguru zabalagoan kokatzeko. Batetik, Unitate Fraseologikoei buruzko oinarri teorikoak aurkeztuko ditugu (2.1. atala) eta, bestetik, Hizkuntzaren Prozesamenduan haiek tratatzeko egiten diren lanak (2.2. atala). Azkenik, kapituluan zehar esandakoak sintetizatze eta argitze aldera, tesi-txosteneko gainerako kapituluetan sarri errepikatuko diren hainbat termino bilduko ditugu, eta definizio labur bana emango diegu.

2.1 Unitate Fraseologikoen oinarri teorikoak

Fraseologiaren alorra zaila da zedarritzen, eta horren erakusle dira, alorraren aztergaiari ematen zaizkion askotariko izendapenak ez ezik, kontzeptu nagusiak definitzean egile batetik bestera dauden aldeak ere. Fraseologiak hitz-konbinazio jakin batzuk aztertzen dituela jakinik, saia gaitezen hitz-konbinazio horien nondik norakoen berri ematen eta, hala, zedarriak pixka bat argitzen.

Lehenik, UFen definizioari eta hizkuntza-ezaugarriei buruz hitz egingo dugu (2.1.1), eta UFak nola sailkatu izan diren ere azalduko dugu (2.1.2).

Ondoren, izena+aditza motako UFen berezitasunez jardungo dugu bereziki, horiek baitira UFen barruan guri gehien interesatzen zaizkigunak (2.1.3), eta UFen itzulpenari buruzko zertzelada batzuk ere emango ditugu (2.1.4).

2.1.1 UFen definizioa eta hizkuntza-ezaugarriak

Unitate Fraseologikoek (UFek) hainbat izendapen jaso izan dituzte literaturan, eta ez dago adostasunik terminologiari dagokionez. Corpas Pastorrek (1996) hiru multzotan banatzen ditu **izendapen** horiek, eta dio multzoetako bakoitzak ezaugarri bat jartzen duela erdigunean:

- *Hitz Anitzeko Unitate* eta antzekoek, hitz batez baino gehiagoz osaturik daudela
- *Esapide finko* eta antzekoek, hitz-konbinazio egonkorrak direla
- *Unitate Fraseologiko* eta antzekoek, unitate semantikoa osatzen duten egitura sintaktikoak direla

Ingelesez, Hizkuntzaren Prozesamenduan behintzat, badirudi *Multiword Expression* terminoa zabaldu dela gehien, lehen multzoko termino bat alegia. Euskaraz, berriz, gehiago erabili izan da Unitate Fraseologiko (UF), azken urteotan bereziki (Urizar, 2012; Gurrutxaga, 2014; Sanz Villar, 2015b), eta termino horren alde egingo dugu guk ere txosten honetan. Izan ere, erabiliena izateaz gain, zehatzena ere horixe iruditzen zaigu, bi arrazoirengatik: batetik, Hitz Anitzeko Unitate askok –adibidez, izen bereziek eta termino polilexikoek– ez dutelako zerikusirik fraseologiarekin, gure aztergaiarekin; eta bestetik, landuko ditugun hitz-konbinazio asko oso malguak direlako, eta kontraesana litzatekeelako horiei *finko* deitzea. Beste hitz batzuetan esanda: UF terminoa erabiliko dugu, lehen multzoko izendapenak zabalegiak direlako guretzat, eta bigarrenekoak, murritzegiak.

Terminologiaz harago ere, alde nabarmenak daude fraseologiaren inguruko lanetan, alorra nola ulertzen den. Egile gehienek **bi ikuspegi** bereizten dituzte (Granger eta Paquot, 2008: 28–29. orr.; Evert, 2009: 1212–1213. orr.; Seretan, 2011: 11–13. orr.):

- Ikuspegi estatistikoa, Firthen (1957) eta Sinclairren (1991) ideietan oinarritua

- Ikuspegi linguistikoa, eskola errusiarrean sortua eta gerora jarraipen luzeagoa izan duena hizkuntzalaritzan batik bat (Corpas Pastor, 1996; Howarth, 1996; Cowie, 1998; Mel'čuk, 1998)

Gurrutxagak azaltzen duenez (2014: 15–16. orr.), bi ikuspegien arteko alde nagusia da lehenak hitzen agerkidetza hutsari ematen diola garrantzia, motibazio linguistikoa ia alde batera utzirik, eta bigarrenak, aldiz, hizkuntza-ezaugarrietan jartzen duela arreta. Horren adibide da ikuspegi estatistikoan hitzen arteko distantzia bakarrik erabiltzen dela konbinazioen testuingurua aztertzeko, eta eredu linguistikoaren aldekoek, ostera, osagaien arteko erlazio sintaktikoari ere erreparatzen diotela besteak beste. Gainera, Firthentzat eta Sinclairrentzat maiztasun handiz erabiltzen diren hitz-konbinazioak –eta, hala, esangura estatistikorik handienekoak– dira aztergairik garrantzitsuenak (*kolokazioak*), eta bigarren ikuspegiaren aldekoek esanahi aldetik konposizionalak ez diren hitz-konbinazioak jartzen dituzte erdigunean (*lokuzioak*). Geroxeago azalduko dugu hobeto zer desberdintasun dagoen bi UF mota horien artean (2.1.2. atala).

Gurea azterketa linguistikoa izanik, ikuspegi linguistikoari lotuko gatzaizkio hemen: UFen ezaugarri lexiko-semantikoei eta morfosintaktikoei begiratuko diegu. Badirudi ikuspegi hori gailentzen dela gaur egun fraseologia konputazionalean; izan ere, hizkuntzalaritzan ez ezik Hizkuntzaren Prozesamenduan ere, UFak prozesatzeko metodo gehienak –estatistikan oinarrituak barne– motibazio linguistikotik abiatzen baitira eta kontuan hartzen baitituzte UFen hizkuntza-ezaugarriak, batez ere sintaktikoak (Ramisch, 2015; Constant *et al.*, 2017).

Deskriba ditzagun, bada, **UFen hizkuntza-ezaugarri nagusiak**, eta begira diezaiegun, horretarako, gure gertueneko bi aurrekariei: Urizarren (2012) eta Gurrutxagaren (2014) doktoretza-tesiei. Ezaugarri desberdinei ematen diete garrantzia batak eta besteak, baina, oinarri-oinarrian, ez dago alde handirik bien ideien artean.

Urizarrek (2012: 56–68. orr.), Corpasek (1996) proposaturiko ereduari jarraituz, UFek honako ezaugarri hauek dituztela dio:

- **Polilexikalitatea**. Aintzat harturik hitza zuriuneen edo bestelako bereizleen arteko karaktere-kate bat dela (Martínez Linares, 2006; Savary, 2008), UFak hitz batez baino gehiagoz osaturiko unitateak dira.

- **Maiztasuna**. Bi eratara dira UFak usuak: agerkidetzari dagokionez eta erabilerari dagokionez. Agerkidetza-maiztasun altukoak direla

esatean, esan nahi da UFko osagai-hitzak askotan agertzen direla elkarrekin testuetan, ausaz pentsa litekeena baino sarriago. Erabileramaiztasunari dagokionez, berriz, ukaezina dirudi halakoek leku handia dutela hiztunon hiztegian[1]: Sinclairren arabera (1991), UF bat edo gehiago erabiltzen dugu batez beste esaldi bakoitzeko; Jackendoff-ek (1997) dio hiztunok hitz bakunak adina UF ditugula gure lexikoan; Sag *et al.*-en ustez (2002), hori baino are handiagoa da proportzioa; eta Monteiro Plantinen hitzetan ere (2011), UFek hiztunon lexikoaren % 50 baino gehiago osatzen dute.

- **Instituzionalizazioa**. Ale lexiko bat –kasu honetan, hitz-konbinazio jakin bat– hiztun-komunitate baten norman integratzean datza instituzionalizazioa. Lotura zuzena du maiztasunarekin –konbinazio jakin bat zenbat eta gehiago erabili are eta altuagoa baita haren instituzionalizazio maila–, bai eta egonkortasunarekin ere –konbinazio bat instituzionalizatzen doan heinean, forma aldetik ere egonkortzera jotzen baitu askotan–. Lipka *et al.*-en lana (2004) hartzen du Urizarrek oinarritzat ezaugarri honetaz hitz egitean.

- **Egonkortasuna eta aldakortasuna**. Hiru ezaugarri bereizten dira atal honetan: finkapena, espezializazio semantikoa eta aldakortasuna.

  - *Finkapena* deritzo UFek izan ohi dituzten murriztapen lexikoak eta sintaktikoak biltzen dituen ezaugarriari. Izan ere, UF askoren osagaiak ezin izaten dira sinonimoez ordezkatu edo ezabatu (*aldez edo moldez*, baina ez *aldez edo maneraz/moduz/gisaz*), haien hurrenkera aldatu (*argi eta garbi*, baina ez *garbi eta argi*), tartean beste elementurik sartu (**nahi handi dut*) eta abar.

  - Finkapen horrek aldaketak eragin ditzake hitz-konbinazio jakin baten jatorrizko interpretazioan, eta horri *espezializazio semantiko* deritzo, erabileraren ondorioz hitz-konbinazio baten interpretazioan gertatzen den aldaketari. Esate baterako, *muzin egin* hitz-konbinazioak aurpegiko keinu bat du esanahiaren oinarrian, zerbait fisikoa, baina UF horrek, gaur egun, askotan egiten dio erreferentzia ekintza psikiko eta abstraktu bati: erdeinatzeari edo arbuiatzeari.

[1] Estimazioen inguruko zerrendan, lehena eta azkena geuk gehitu ditugu.

  - Azkenik, *aldakortasuna* finkapenaren eskutik doan kontzeptua da. Nolabait kontrakoa esan nahi duen arren, UFen murriztapenak ez dira erabatekoak izaten normalean, eta, hala, UF gehienak ezaugarri jakin batzuekiko finkoak eta beste batzuekiko aldakorrak izaten dira: *beti parte hartzen du* edo *beti hartzen du parte*, baina ez, adibidez, **parte etengabea hartzen du*.

- **Konposizionaltasunik eza**. UF askoren esanahi globala ezin daiteke ondorioztatu osagai-hitzen esanahiak konbinatuz; esate baterako, *adarra jotzeak* ez du zerikusirik ez adarrekin eta ez jotzearekin. Urizarren hitzetan, espezializazio semantikoaren gradurik gorena da konposizionaltasunik eza, eta UFen ezaugarri nagusitzat hartu izan dute egile askok.[2]

- **Mailaketa**. Aurretik aipaturiko horiek guztiak UFen ezaugarriak dira, baina UF guztiek ez dituzte ezaugarriok beti maila berean izaten. Maila batetik besterako mugak ez dira beti argiak, eta etengabeko *continuum* bat osatzen dutela esaten da.

Gurrutxagaren ustez, eredu horretan, "instituzionalizazioa gertu dago gainerako ezaugarriak biltzen dituen ezaugarri orokor bat izatetik", lotura estua baitu beste ezaugarri gehienekin. Instituzionalizazio hori bere ereduko idiomatikotasunarekin parekatzen du nolabait, bere ustez UFak gainerako hitz-konbinazioetatik bereizten dituen ezaugarri nagusiarekin. Honela definitzen du Gurrutxagak **idiomatikotasuna** (2014: 25. orr.):

> «Idiomatikotasuna konbinazio bat UF izatea determinatzen duen propietatea da, muineko ezaugarritzat idiosinkrasia duena (hizkuntzaren ohiko portaeratik aldentzea, banako hizkuntza-elementuen konbinazio libreak aurreikusi edo esplika ezin dezakeena). Idiomatikotasuna konplexua eta graduala da, eta bere barnean zenbait propietate hartzen ditu: instituzionalizazioa, ez-konposizionaltasun semantikoa (osoa edo partziala) eta finkapena (morfosintaktikoa zein lexikala).»

[2]Konposizionaltasunik ezari *idiomatikotasun* ere esaten zaio, nahiz eta termino horrek bigarren adiera zabalago bat ere baduen, semantikatik haragokoa: partikulartasuna, hizkuntza batek berea eta berezia duenari dagokiona. Nolanahi ere, guk hemen *idiosinkrasia* esango diegu hizkuntza orokorrari dagozkion berezitasunei, eta *idiomatikotasuna*, berriz, hitz-konbinazioei dagokien idiosinkrasiari bakarrik Ikus kapitulu honen amaieran jarri ditugun definizioak (69. orrialdea).

Beraz, Corpas-Urizar ereduko ia ezaugarri guztiak agertzen dira Gurrutxagaren definizio horretan. Berariaz aipatzen ditu instituzionalizazioa, konposizionaltasunik eza eta finkapena, eta polilexikalitatea eta mailaketa ere kontuan hartzen ditu, zeharka izan arren: batetik, "hitz-konbinazioak" aipatzean aintzat hartzen ari baita UFak polilexikoak direla; eta bestetik, fenomeno "graduala" dela esatean mailaka daitekeela esaten ari delako. Maiztasuna ez da definizio horretan esplizituki agertzen, baina, geroxeago azalduko dugunez (2.2.2. atala), ezaugarri horrek ere garrantzi handia du Gurrutxagaren ereduan, corpusetatik UFak erauzteko lanean horixe baita idiomatikotasuna neurtzeko darabilen neurgailurik garrantzitsuenetako bat.

Hortaz, azaldu dugu UFen ezaugarri nagusiak zein diren, eta gatozen orain **definizio formaletara**. Hizkuntzaren Prozesamenduan gehien zabaldu diren bi definizioak Sag *et al.*-ena (2002) eta Baldwin eta Kimena (2010) izan dira.

- **Sag *et al.*-**en arabera, hitzen (edo zuriuneen) arteko mugetatik haragoko interpretazio idiosinkrasikoak dira UFak, eta hiru motatako idiosinkrasia izan dezakete: sintaktikoa, erabat malguak ez direlako; semantikoa, konposizionalak ez direlako; eta estatistikoa, maiztasun nabarmen handiz agertzen direlako.

- **Baldwin eta Kim**ek, berriz, definizio hori oinarritzat hartu eta proposamen zabaldu bat egiten dute. Haien ustez, UFak item lexikoak dira, eta honako bi ezaugarri hauek dituzte: (a) lexema batean baino gehiagotan deskonposa daitezkeela, eta (b) idiomatikotasun lexiko, sintaktiko, semantiko, pragmatiko eta/edo estatistikoa dutela.

Hortaz, Baldwin eta Kimek alderdi lexikoa eta pragmatikoa gehitzen dizkiote aurrekoen definizioari, eta *idiomatikotasun* terminoa darabilte *idiosinkrasia* beharrean. Lotura estua dute bi terminoek, baina guk ez ditugu berdin-berdin erabiliko, *idiosinkrasia* kontzeptu zabalagotzat hartzen dugulako, ezohikoa den edozeren nolakotasuntzat, eta *idiomatikotasuna*, berriz, hitz-konbinazioen ezaugarri espezifikotzat (ikus 2. oin-oharra eta 69. orrialdeko definizio laburrak). Bestalde, aipagarria iruditzen zaigu biek ere sintaxia bakarrik aipatzea, eta ez morfologia, gure ustez oso estuki lotuta egoten baitira bata eta bestea eta, hemendik aurrerako kapituluetan erakutsiko dugunez, morfologia ere gure lanen ardatzetako bat baita.

Azkenik, merezi du **PARSEME sare europarra**ren baitan (Savary *et al.*, 2015) eginiko lanei ere erreparatzea, proiektu hori baita azken urteo-

tan fraseologia konputazionalaren alorrean izan den egitasmorik handiena. Geroago ere emango dugu proiektuaren xehetasun gehiago, baina, oraingoz, ekar dezagun hona nola definitzen dituzten UFak (Savary *et al.*, 2018):

> «Multiword expressions (MWEs) are (continuous or discontinuous) sequences of words with the following compulsory properties:
>
> - They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language (...)
> - Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain is a syntactic (direct or indirect) dependence but it can also be e.g. a coordination. Depending on the category of the head word, the whole MWE can be nominal, adjectival, prepositional, verbal, sentential, etc.
> - At least two components of such a word sequence have to be lexicalized.»

Beraz, PARSEMEn, idiomatikotasun mailetan sartzen dituzte ortografia eta morfologia ere, eta definiziotik bertatik uzten dute argi UFaren osagaihitzak sintaktikoki erlazionaturik egoten direla, mendekotasunezko erlazioan normalean. Horrez gain, osagai *lexikalizatuez* hitz egiten dute UFan beti agertzen diren lexemei erreferentzia egiteko. Kontzeptu horrek lotura du, nolabait, lehen aipatu dugun finkapen lexikoarekin, eta garrantzitsua da hainbat atazatarako; testu-corpusetan UFak markatzeko, adibidez, oinarrioinarrizkoa da osagai lexikalizatuak (adibidean, beltzez) eta lexikalizatu gabeak bereiztea, hala bakarrik jakin baitaiteke non hasten den eta non amaitzen den UF bat zehazki.

(13) ***Ondorioak atera*** *zituen.*

(14) ***Ondorio*** *interesgarri bat* ***atera*** *zuen.*

(15) ***Atera*** *zeuk* ***ondorioak****.*

Instituzionalizazioa ez da giltzarritzat hartzen bere horretan, eta finkapen morfosintaktikoa ere ez da PARSEMEren definizioan aipatzen. Lan bereko beste atal batean, ordea, aditz-UFak tratatzeko erronkarik handienak zerrendatzean, aipatzen dute aditz-UFek hainbat aldaki morfosintaktiko izan

ditzaketela eta, are, izan ohi dituztela. Geroxeago hitz egingo dugu erronka horiez zabalago, 2.1.3. atalean.

Horren aurretik, baina, ikus dezagun nola sailkatu izan diren UFak, eta jarrai dezagun fraseologiaren inguruko kontzeptu gehiago argitzen.

2.1.2 Sailkapenak

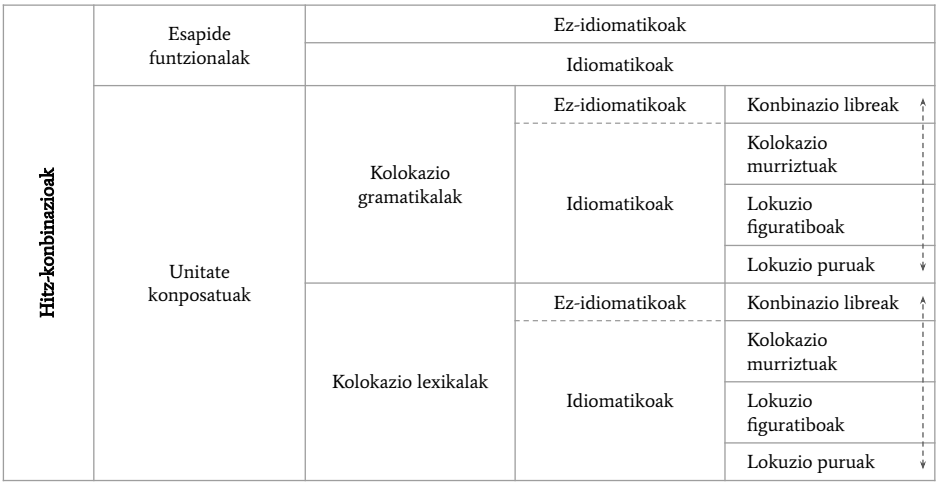
Fraseologiaren inguruko kontzeptu nagusiak definitzeko bezala, UFak sailkatzeko ere askotariko irizpideak erabili izan dira. Guk, atal honetan, arreta jarriko dugu gure lanean eragina izan duten sailkapenetan. Hasteko, era guztietako UFak kontuan hartzen dituzten hiru sailkapen orokor aurkeztuko ditugu: Howarthena (1996), Sag *et al.*-ena (2002), eta Corpasena (1996), Urizarrek (2012: 68–71. orr.) euskarara egokitua. Ondoren, berriz, aztergai espezifikoagoetara joko dugu, aditz-UFetara hain zuzen, eta Gurrutxagaren (2014) eta PARSEME proiektuaren (Savary *et al.*, 2018) sailkapenei begiratuko diegu. Sailkapen gehiagoren berri nahi duenak Gurrutxagaren tesira (2014: 33. orr.) jo dezake, beste hainbat egileren multzokatzeak ere jasotzen eta parekatzen baitira han.

2.1.2.1 UFen sailkapenak, oro har

Sailkapen orokorrez jarduteko, begira diezaiogun lehenik **Howarthen (1996) proposamena**ri (2.1. irudia). Hitz-konbinazio usuen lehen zatiketa *esapide funtzional*en eta *unitate konposatu*en artean egiten du. Lehen multzokoek zeregin jakin bat betetzen dute diskurtsoan, eta, askotan, hizketa-egintza edo enuntziatu osoak dira bere horretan[3] –atsotitzak (*You scratch my back and I'll scratch yours* lit. egidazu hazka bizkarrean eta nik ere egingo dizut, 'laguntzen badidazu, nik ere lagunduko dizut') eta errutinazko formulak (*What's up?* lit. zer da gora? 'zer berri?'), adibidez–. Bigarren multzokoek, aldiz, zeregin sintaktikoa dute esaldian edo perpausean.

Ondoren, unitate konposatuak sailkatzeko, hitz-konbinazioen osaera morfologikoa darabil irizpidetzat: kolokazio lexikoetan sartzen ditu klase irekiko bi osagai dituzten konbinazioak (izena+aditza, adjektiboa+aditza, etab.), eta kolokazio gramatikaletan, osagaietako bat klase itxikoa dutenak (izena+preposizioa, aditza+preposizioa, etab.). Horren ostean, lau multzoko

[3]Hizketa-egintza edo enuntziatu oso deitzen zaie bere horretan erabil daitezkeen hitz-konbinazioei, hau da, esaldi baten barruan txertatu gabe beregainak direnei.



2.1 irudia – Hitz-konbinazioen sailkapena, Howarthen (1996) arabera

banaketa bera egiten du batzuetan zein besteetan, idiomatikotasun txikienekoetatik handienekoetara: konbinazio libreak (*blow a trumpet* 'tronpeta jo'; *under the table* 'mahaiaren azpian'), kolokazio murriztuak (*blow a fuse* 'fusiblea erre'; *under attack* 'erasopean'), lokuzio figuratiboak (*blow your own trumpet* lit. nork bere tronpeta jo, 'harrokeriaz agertu'; *under the microscope* lit. mikroskopioaren azpian, 'xehetasun handiz'), eta lokuzio puruak (*blow the gaff* lit. etxeari putz egin, 'sekretuak agerian utzi', *under the weather* lit. eguraldiaren azpian, 'ondoezik'). Multzo batetik besterako mugak ez dira erabat zurrunak, eta idiomatikotasunak *continuum* kolokazional bat osatzen du –zeina guk, 2.1. irudian, gezi bikoitz eten baten bidez adierazi baitugu–.

Howarthek, sailkapen hori proposatzeko, lexikologiaren eragin handia duten hainbat egile (Arnold, 1986; Gläser, 1986; Cowie, 1988) hartzen ditu oinarritzat, eta esan liteke haien lanen arteko nahasketa bat duela berea. Aipagarria da *kolokazio* terminoa oso era orokorrean darabilela, unitate konposatuen barruko edozein hitz-konbinaziori deitzen baitio hala[4] eta lokuzioak ere kolokazioen barruan sartzen baititu. Beraz, Howarthen ereduan, elkarrekin maiz agertzen diren hitzen konbinazioak kolokazioak dira, bai idiomati-

[4]Sailkapenerako erabilitako terminologian, ez du *collocation* erabiltzen, baizik eta *composite*, baina testuan zehar sinonimotzat erabiltzen ditu bi terminoak.

koak eta bai idiomatikoak ez direnak ere.

Nolanahi ere, *kolokazio* terminoaren erabilera zabal hori ez da nahastu behar ikuspegi erabat estatistikoa duten egileen lanekin, Howarthen sailkapenaren benetako oinarria hitz-konbinazioen esanahian baitatza, idiomatiko ala ez-idiomatiko bereizketan. Gainera, estatistikaren garrantzia aintzat harturik ere, Howarth esplizituki kontrajartzen zaie maiztasuna irizpide bakartzat darabiltenei (Howarth, 1996: 27. orr.):

> «The approach followed here recognizes the enormous value of corpora large and small, but takes the view that phraseological significance means something more complex and possibly less tangible than what any computer algorithm can reveal.»

Sag *et al.*, 2002

Hitz anitzeko esapideak	Instituzionalizatuak	
	Lexikalizatuak	Finkoak
		Erdifinkoak
		Malguak

2.2 irudia – Hitz anitzeko unitateen sailkapena, Sag *et al.*en (2002) arabera

Hizkuntzaren Prozesamenduko nazioarteko lanetan, ziur asko, **Sag *et al.*-en (2002) sailkapena** izan da entzutetsuena (2.2. irudia), sailkapen horren egokitzapenak baitira gerora egin diren proposamen asko, Baldwin eta Kimena (2010) eta Ramischena (2015: 41–44. orr.) besteak beste. Bi multzo nagusi bereizten dituzte Sag *et al.*-ek: esapide *instituzionalizatuak* eta *lexikalizatuak*. Lehenengoetan sartzen dituzte elkarrekin agertzeko joera nabarmena duten hitz-konbinazioak baina konposizionalak direnak bai sintaktikoki eta bai semantikoki (*traffic light* lit. trafiko-argi, 'semaforo'), eta bigarrenetan, berriz, sintaxiari edo semantikari dagokionez idiosinkrasikoak direnak. Bigarren multzo hori, gainera, beste hiru multzotan banatzen dute, finkapen morfosintaktikoaren arabera: esapide finkoak (*by and large* lit. aurretik eta zabal, 'oro har'), erdi finkoak (*kick the bucket* lit. pertzari ostikoa jo, 'hil') eta malguak (*look up* lit. gora begiratu, 'bilatu').

Sailkapen horretan esplizituki agertzen ez bada ere, aipagarria da Sag *et al.*-ek ere Howarthek bezain era orokorrean darabiltela *kolokazio* terminoa, estatistikoki esanguratsua den edozein hitz-konbinaziori deitzen baitiote hala, bai UFak direnei eta bai hizkuntzaz kanpoko ezaugarriengatik elkarrekin agertzeko joera dutenei. Beste egile batzuek, ordea, jarraian hizpide izango ditugunek adibidez, ez diote esanahi hori ematen, eta hori, nolabait, kolokazioen inguruan literaturan dagoen adostasun-faltaren isla da.

Izan ere, *kolokazio* terminoa lehen aldiz aipatu zenetik (Firth, 1957)[5], termino hori era batera baino gehiagotara definitu izan da. Zenbaitek diote hiztun batzuek, aukeran dituzten hitz-konbinazio posible guztietatik, konbinazio jakin batzuk sortzera jotzen dutela, eta horiek direla kolokazioak (Haensch, 1982: 251. orr., Corpasen 2001eko lanetik); beste zenbaitek, ordea, erabileraren poderioz murriztapen lexikoak dituzten konbinaziotzat hartzen dituzte, eta *oinarri* eta *kolokatu* bereizketa egiten dute normalean, non lehen osagaia semantikoki autonomoa baita eta osagai horrek aukeratzen baitu bigarrena (Corpas Pastor, 1996: 66. orr.; Urizar, 2012: 89–99. orr.): *ardozale amorratu*, *zorra kitatu* (oinarria+kolokatua). Bigarren definizio horrekin du lotura hona ekarriko dugun hurrengo sailkapenak.

Corpas Pastor, 1996; Urizar, 2012

Unitate Fraseologikoak	Enuntziatuak (hizketa-egintzak)	Enuntziatu fraseologikoak
	Ez enuntziatuak (ez hizketa-egintzak)	Kolokazioak
		Lokuzioak

2.3 irudia – Unitate Fraseologikoen sailkapena, Corpas Pastorren (1996) eta Urizarren (2012) arabera

Corpasek (1996) eta Urizarrek (2012) hiru esferatan banatzen dituzte UFak, finkapen motaren arabera (2.3. irudia). Batetik, hizketa-egintza

[5] *Kolokazio* terminoa lehen aldiz Firthek erabili bazuen ere, ohar bedi ez zela bera izan kontzeptu horren inguruan hitz egiten lehena. Kennedyk (1998: 108. orr.) aipatzen duenez, Alexander Crudenek duela bi mende eta erdi aipatu zuen Biblian hitz batzuk sarri agertzen zirela elkarrekin. Gerora, XX. mendearen hasieran, Ballyk ere (1909) agerian jarri zuen hitz-konbinazio batzuetako hautapen lexikoa murritza zela, eta *groupements usuels* ('multzokatze usuak') deitu zien halakoei.

osoak direnak *enuntziatu fraseologiko*tzat jotzen dituzte, eta hizketan finkatuta daudela diote, hau da, bere horretan erabil daitezkeela, esaldi baten barruan txertatu beharrik izan gabe. Multzo horrek osatzen du lehen esfera, eta hor sartzen dira, esate baterako, atsotitzak (*zozoak beleari, ipurbeltz*) eta errutinazko formulak (*egun on*). Bigarren eta hirugarren esferetako UFek, berriz, ez dute hizketa-egintza osorik osatzen: bigarren esferan *kolokazioak* daude (*haserre bizi, zarata atera*), arauan finkatuak, sistemaren ikuspegitik sintagma libreak osatzen dituztenak; eta hirugarrenean, *lokuzioak* (*ziria sartu, hanka egin*), sisteman unitateak osatzen dituztenak eta, hortaz, sisteman finkatuak.

Beraz, konposizionaltasun semantikoa eta finkapen morfosintaktikoa eta lexikoa uztartzen dituzte UFak sailkatzeko. Ondoren, osaera morfosintaktikoaren araberako sailkapen xeheagoak ere egiten dituzte, eta UFen taxonomia zabalak eskaintzen, osagaien gramatika-kategoria eta osagaien arteko erlazio sintaktikoa oinarritzat harturik.

Urizarrek (2012: 109. orr.) lokuzioak lantzen ditu batez ere, eta bi multzotan banatzen ditu: lexikoak eta gramatikalak. Adibidez, lokuzio lexikoen barruan, honako bost multzo hauek bereizten ditu:

- Izen-lokuzioak (*a bildua, euskaldun berri*)
- Aditz-lokuzioak (*ahalak eta leherrak egin, ezagutzera eman*)
- Adjektibo-lokuzioak (*adin txikiko, batez besteko*)
- Adberbio-lokuzioak (*aldez edo moldez, behin eta berriro*)
- Interjekzio-lokuzioak (*hor konpon, ongi etorri*)

Aditz-lokuzioen barruan, ostera, honako beste multzokatze hau egiten du[6]:

- Izen- edo adjektibo-sintagma biluzia + aditza
  - Objektu zuzena + aditza (*min egin, beldur ukan*)
  - Objektuaren predikatzailea + aditza (*atsegin ukan, gogait egin*)

[6]Ohar bedi sailkapen honetako adibideak Urizarrenak berarenak direla. Horietako asko aditz arindunak dira, eta guk, 7.1.1.1. atalean argituko dugunez, ez ditugu halakoak lokuziotzat hartzen, baizik eta kolokazioen azpimultzotzat.

  - Subjektuaren predikatzailea + aditza (*falta izan, ados egon*)

- Izen- edo adjektibo-sintagma mugatua + aditza
  - Nominatibodun sintagma + aditza (*adarra jo*)
  - Ergatibodun sintagma + aditza (*suak hartu*)
  - Datibodun sintagma + aditza (*hitzari eutsi*)

- Adizlaguna edo aditzondoa + aditza
  - Argumentua + *egin* (*hegaz egin, haginka egin*)
  - Postposizio-sintagma + bestelako aditza (*aurrera eraman, gogora etorri*)

2.1.2.2 Aditz-UFen sailkapenak

Urizarren azpisailkapen hori aitzakiatzat harturik, gatozen orain buru sintaktikotzat aditza duten UFetara. **Gurrutxaga**k ere (2014: 53. orr), izena+aditza motako UFak sailkatzean, kolokazioak eta esapide idiomatikoak[7] banatzen ditu, Cowieren (1986) eta Howarthen (1996) lanetan oinarrituta batik bat. Ondoren, beste bina multzo bereizten ditu bi kategoria nagusi horietan, Ezeizak (2002: 96–97. orr.) bere tesi-lanean bereizitako berberak. Aurrekoekin egin bezala, 2.4. irudian jaso dugu Gurrutxagaren sailkapena.

Esapide idiomatikoen artean, opakoak eta figuratiboak banatzen ditu, eta kolokazioen artean, berriz, murriztuak eta irekiak. Nolanahi ere, Gurrutxagak aitortzen du kolokazio irekien multzoa nahiko eztabaidagarria dela, zalantzan jarri izan baita fraseologiaren alorreko aztergaia ote den benetan.

Tesi-txosten horretako diagrama batetik abiatuta, honela ezaugarritu daitezke sailkapen horretako lau multzoak (Gurrutxaga, 2014: 160. orr.):

- Esapide idiomatiko opakoak: hitz-konbinazioaren esanahia ezin da osagaien esanahietatik ondorioztatu edo ikasi.
  → *adarra jo, ziria sartu, hautsak harrotu*

- Esapide idiomatiko figuratiboak: esanahiak interpretazio metaforikoa edo figuratiboa onartzen du, edo izenak ez du bere adiera gordetzen.
  → *zubiak eraiki, burua hautsi, ateak zabaldu*

[7]Esapide idiomatikoak lokuzioen parekoak dira Gurrutxagaren ereduan, baina kategoria zabalxeagoa da berez, barne hartzen baititu hizketa-egintza osoak ere, atsotitzak eta halakoak.

2.4 irudia – Izena+aditza konbinazioen sailkapena, Gurrutxagaren (2014) arabera

- Kolokazio murriztuak: izenaren flexioa ez da erregularra, aditzak izenari dagokion ekintza bideratzeko funtzioa du (*arina* da) edo aditza ezin da sinonimo batekin ordezkatu.
  → *min eman, lan egin, beldur izan*

- Kolokazio irekiak: aditza sinonimoez ordezka daiteke, baina hala sortutako konbinazioak ez dira ohikoak.
  → *elkartasuna adierazi, legea urratu, konpromisoa berretsi*

Azkenik, esan bezala, konbinazio libreen multzoa ere (*liburua irakurri, armak saldu, partida jokatu*) idiomatikotasunaren continuumean sartzen du, lokuzio opakoen kontrako muturrean.

Sailkapenen atalarekin bukatzeko, komeni da PARSEMEren irizpideak ere aintzat hartzea, Hizkuntzaren Prozesamenduko ikertzaile-sare zabal baten onarpena izateaz gain tesi-txosten honen 7. kapituluan hizpide izango baititugu. Proiektu horretan, aditz-UFetan jarri dute arreta batez ere, eta edozein hizkuntzatarako aplikagarria den eredu bat sortu nahi izan dute. Hala, hogei hizkuntzatako testuak biltzen dituen corpus bat etiketatu dute fraseologia mailan, hizkuntza guztietan ere irizpide berberak eta sailkapen berbera kontuan harturik. Bi kategoria unibertsal, hiru ia unibertsal eta kategoria esperimental bat bereizi dituzte, 2.5. irudian ikus daitekeenez.

Kategoria unibertsalen artean, aditz-esapide idiomatikoak (*adarra jo, txoritxo batek esan*) eta aditz arindun konbinazioak (*negar egin, min hartu, eskubidea eman*) bereizten dituzte. Bi multzo horiek dira, hain zuzen, euskaraz

PARSEME proiektua (Savary *et al.*, 2018)

Aditz-UFak	Kategoria unibertsalak	Aditz-esapide idiomatikoak	
		Aditz arindun konbinazioak	Osoak
			Kausatiboak
	Kategoria ia unibertsalak	Erreflexiboa berezko duten aditzak	
		Aditza+partikula konbinazioak	
		Aditz anitzeko konbinazioak	
	Kategoria esperimentala	Adposizioa berezkoa duten aditzak	

2.5 irudia – Aditz-UFen sailkapena, PARSEMEren corpusean (Savary *et al.*, 2018)

aplikagarriak diren bakarrak, bai eta gure tesi-lanarekin zerikusi zuzena duten bakarrak ere, izena+aditza motako konbinazioak ezin baitira beste multzo batean ere sartu. Geroago hitz egingo dugu zabalago bi kategoria horien inguruan (193. orrialdetik aurrera), eta ez dugu xehetasun gehiago emango oraingoz.

Dena den, aipagarria da kolokazioak ez dituztela aintzat hartzen sailkapen horretan, fenomeno estatistiko hutsa baitira haien ustez, fraseologiatik kanpokoa. Gaztelaniazko eta euskarazko tradizio fraseologikoan, ostera, aditz arindun konbinazioak kolokazioen azpimultzotzat hartu ohi dira (Alonso Ramos, 2004; Bustos Plaza, 2005; Wanner *et al.*, 2006; Buckingham, 2009: 18. orr.; Gallego, 2010; Sanroman Vilas, 2017), nahiz eta gutxi batzuek kolokazioen eta lokuzioen artean kokatzen dituzten (Heine, 2006). Besteak beste, Urizarren eta Gurrutxagaren lanek agerian uzten dute mugako izaera hori: lehenak lokuziotzat hartzen ditu *min eman, lan egin* eta halakoak, eta bigarrenak, aldiz, kolokazio murriztuen balizko ezaugarrien artean sartzen du aditza arina izatea. Aditz arinak izenei laguntzen joan ohi direnez oro har, datorren atalean hitz egingo dugu gehiago horien inguruan.

Kategoria ia unibertsalei eta esperimentalari dagokionez, ingelesezko edota gaztelaniazko adibide batzuk emango ditugu jarraian. Azalpen gehiago nahi dituenak gidalerroetara jo dezake, Savary *et al.*-en artikulura (2018) nahiz sareko bertsiora[8].

[8] `http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=home`

- Erreflexiboa berezko duten aditzak (IReflV) → *abstenerse, suicidarse* (ES)
- Aditza+partikula konbinazioak (VPC) → *put off*, *blow up* (EN)
- Aditz anitzeko konbinazioak (MVC) → *make do* (EN)
- Adposizioa bereko duten aditzak (IAV) → *stand for, rely on* (EN); *entender de, contar con* (ES)

Teknikoki, adposizioa berezko duten aditzak baditugu euskaraz ere –Urizarrek lokuzio gramatikalen artean sartzen dituenak–, aditz-esapide idiomatikoez eta aditz arindun konbinazioez gain: *-tzat hartu, -(e)kin kontatu* eta gisa horretakoak. Hala ere, 7. kapituluan argituko dugunez, PARSEMEren ereduan hitz osoak bakarrik hartzen dira aintzat, eta, hortaz, hizkuntza ez-eranskarietan ez bezala, euskarazko konbinazio horiek eredutik kanpo geratzen dira.

2.1.2.3 *Kolokazio* eta *lokuzio* kontzeptuak lan honetan

Orain arte esandako guztia kontuan hartuta, bistan da ez dagoela fraseologia eta haren parte diren UF motak ulertzeko modu bakarra. Guri dagokigunez, esan dezakegu gertuen ditugun aurrekarien lanetan oinarrituko garela batez ere, Urizarrenean (2012) eta Gurrutxagarenean (2014). Dena dela, UFak sailkatzeko, egile horiek uztartu egiten dituzte alderdi lexiko-semantikoa eta morfosintaktikoa, eta guk, aldiz, kolokazio-lokuzio bereizketa maila lexiko-semantikoan bakarrik egingo dugu, ezaugarri morfosintaktikoak aparteko sailkapen baterako utzirik. Bigarren maila horretan, berriz, Sag *et al.*-en hirukoa erabiliko dugu, HPko tresnetan erabiltzeko asmoz: UF finkoak, erdifinkoak eta malguak.

Orain ez gara hasiko gure sailkapen-proposamenari buruzko xehetasunak ematen, laugarren kapituluan (103. orrialdetik aurrera) eskainiko baitiogu tartea proposamen horri eta proposamena egiteko arrazoiei. Nolanahi ere, kapitulu honen helburu nagusia kontzeptu gakoak kokatzea eta argitzea denez, ezin buka dezakegu atal hau aditz-UFen bi mota nagusiei definizio labur bana eman gabe, argi gera dadin zehazki zertaz ari garen hemendik aurrera termino horiek erabiltzen ditugunean.

Kontuan harturik landuko ditugun aditz-UFek ez dutela hizketa-egintzarik osatzen eta hitz-konbinazioko osagaiak sintaktikoki erlazionaturik egon ohi direla:

- **Kolokazio**tzat joko ditugu elkarrekin ausaz espero litekeena baino maizago agertzeko joera duten hitz-konbinazio usuak. Halakoetan, osagaietako bat (*oinarria*) buru semantikoa izaten da, eta horrek aukeratu ohi du beste osagaia (*kolokatua*). Aditz-UFetan, buru sintaktikoa aditza bada ere, oinarria izena izan ohi da, eta kolokatua, aditza. Askotan, aditz hori ezin izaten da sinonimoen bidez ordezkatu (*legea urratu, plazak bete*), edo ordezkatuz gero konbinazio arrotzak osatzen dira. Beste batzuetan, berriz, aditza *arina* izaten da: bere esanahia galtzen du nolabait, eta ezaugarri morfologikoak bakarrik gehitzen dizkio hitz-konbinazioari (*min hartu, lan egin*).

- **Lokuzio** deituko diegu semantikoki –behintzat– idiomatikoak diren hitz-konbinazioei. Halako UFen esanahi osoa ez da osagai-hitzen esanahien batura, baina batzuetan uler daitezke metafora bidez; horren arabera bereiziko ditugu lokuzio opakoak (*adarra jo, hanka egin*) eta metaforikoak (*zubiak eraiki, begi-bistatik galdu*).

2.1.3 Aditz-UFak eta haien ezaugarriak

Gure aztergaia aditz batez eta izen batez osaturiko konbinazioak dira zehazki, buru sintaktikotzat aditza duten UF motetako bat alegia. Argitu dezagun, atal honetan, zertan diren bereziak halakoak.

Urizarrek (2012: 117. orr.), aditz-lokuzioei buruz ari delarik, honako hau dio: "lokuzio-kategoria guztien artean ugarienetakoa izateaz gain, aditz-lokuzioek ager ditzaketen egitura motak ere askotarikoak dira, eta, halaber, aldakortasun maila handia erakusten dute gehienek." Hortaz, badira bi ezaugarri garrantzitsu izena+aditza motako UFak aztertzean kontuan hartu beharrekoak: haien maiztasuna batetik, eta aldakortasuna bestetik.

Lehen ezaugarriari dagokionez, esan beharra dago osaera horretako hitz-konbinazioak ez direla lokuzioetan bakarrik usuak, baizik eta UFetan oro har (Wotjak, 2018). PARSEMEren gaztelaniazko eta euskarazko corpusei begiratuta (Savary *et al.*, 2018; Ramisch *et al.*, 2018), adibidez, agerikoa da maiztasun hori: aditz arindun konbinazioak eta aditz-esapide idiomatikoak kontuan hartuta, gaztelaniazko corpuseko esaldien % 13k dute aditza+izena motako UFren bat etiketatuta batez beste, eta euskaraz, berriz, esaldien % 32k. Hortaz, maiz erabiltzen dira bietan, eta euskaraz bereziki –ia hiru aldiz gehiago–. Horietako gehien-gehienak aditz arindun konbinazioak dira, eta geroxeago eskainiko diegu tarte zabalagoa.

Bigarrenik, esana dugu aditz-UFak malgutasun handikoak izan ohi direla morfosintaxiari dagokionez eta hainbat aldaki izan ohi dituztela. Izena+aditza motakoak aditz-UFen azpimultzo bat direnez, ez dago esan beharrik aldakortasunak bete-betean eragiten diola gure aztergaiari. Morfosintaxia denez datozen kapituluetan azalduko dugun lanaren ardatza, sakondu dezagun beste pixka bat ezaugarri horretan.

2.1.3.1 Aditz-UFen malgutasun morfosintaktikoa

Fenomeno fraseologiko orok bezala, aditz-UFek ez dute talde homogeneorik osatzen. Batetik, aurreko atalean aipatu ditugun taldeetan bereiz daitezke, kolokazioetan eta lokuzioetan, oro har. Bestetik, morfosintaxiari dagokionez ere askotarikoak izan daitezke aditz-UFak, murriztapen gehiagokoak ala gutxiagokoak.

UFen malgutasun morfosintaktikoa aztertzeko erabili izan den metodo bat aldakortasun-testak egitea da. Parra Escartín *et al.*-ek (2018), adibidez, Nunberg *et al.*-en lanean (1994) oinarriturik, zortzi testen arabera sailkatzen dituzte gaztelaniazko UFak. Haien esanetan, gaztelaniazko UFak sailkatzeko eta aztertzeko egin diren proposamen gehienak ez dira baliagarriak HPko tresnak garatzeko, eta egokiagoa da Ramischen taxonomia (2015: 41–44. orr.), zeinak konbinazio malguak eta erdifinkoak bereizten baititu, Sag *et al.*-en lana (2002) oinarritzat harturik (ikus 24. orrialdeko 2.2. irudia). Hortaz, UFak sailkapen horretara ekartzeko asmoz proposatzen dituzte zortzi test hauek[9]:

- UFaren osagaiak flexionatu ote daitezkeen begiratzea.

  (16) *(ella)* ***corta el bacalao***; *(ellos)* ***cortan el bacalao***; *(ellos) cortan los bacalaos*

- Izen-sintagmetako determinatzaileak aldatu ote daitezkeen aztertzea.

  (17) ***hacer una foto; hacer varias fotos; hacer muchas fotos***

- UFaren zati bat pronominalizatzeko aukerari erreparatzea.

[9]Parra Escartín *et al.*-en lanean (2018), aditza+izena motakoak ez diren UFak ere kontuan hartzen dira. Adibideak gure aztergaira egokitze aldera, aditza+izena motakoak ekarri ditugu hona, eta, kasuren batean, geuk sortu edo egokitu ditugu.

(18) ***Dimos*** *un largo* ***paseo*** *por el campo y lo disfrutamos mucho.*

- Izen-sintagma edo sintagma osagarria topikalizatu daitekeen aztertzea.

(19) *¿Qué* ***trato*** *crees que* ***harán****?*

- Begiratzea ea UFaren osagaietako bat mendeko perpaus batean ager daitekeen edo mendeko perpausen bat ager daitekeen UFaren osagai bati loturik.

(20) *El* ***trato*** *que* ***hizo*** *consistía en...;* ***hizo*** *un* ***trato*** *que consistía en...*

- UFa egitura pasiboetan agertzen ote den aztertzea. Nolanahi ere, egileek aitortzen dute pasiboa ez dela ingelesez bezain erabilia gaztelaniaz, eta balitekeela test hau gaztelaniarako hain esanguratsua ez izatea.

(21) *La* ***decisión*** *fue* ***tomada*** *el lunes; la* ***decisión*** *se* ***tomará*** *el lunes.*

- UFko osagaien artean beste elementu batzuk (adjektiboak, adberbioak...) agertzeko aukerari erreparatzea.

(22) ***dar*** *un largo* ***paseo****;* ***echar*** *profundamente la* ***siesta***

- UFko elementuren bat eliditu ote daitekeen begiratzea.

(23) ***Tiene frío*** *y (tiene)* ***calor*** *al mismo tiempo.*

Buckingham-ek (2009: 32–40. orr.) ere test horietan agertzen diren ezaugarri gehienak aipatzen ditu, desberdin antolatzen eta karakterizatzen baditu ere. Esate baterako, koordinazioa eta erlazio anaforikoak aipatzen ditu berariaz, eta ez elipsia, baina haren adibideak (24 eta 25) goiko testetako azkenekoan sar litezke.

(24) *Juan le* ***dio*** *a María un* ***beso*** *y un* ***caramelo****.*

(25) **No pudimos* ***poner final*** *a la discusión entonces, pero al día siguiente lo pusimos.*

Dena dela, Buckinghamek mota jakin bateko aditz-UFak bakarrik hartzen ditu aztergaitzat: aditza *euskarria* dutenak edo, guk orain arte erabilitako terminologian, aditz arindun konbinazioak. Halakoei eskainiko diegu, hain zuzen, datorren azpiatala.

2.1.3.2 Aditz arindun konbinazioak

Hitz-konbinazio jakin batzuetan, gertatzen da aditzak esanahia galtzen duela nolabait. Gisa horretako aditzei erreferentzia egiteko, termino bat baino gehiago erabili da literaturan, *aditz arin* edo *aditz euskarri* batez ere. Bosquek dioenez (2001: 35. orr.), *aditz arin* terminoa erabiliz predikatuen zentzu abstraktua markatzen da, eta *aditz euskarri* erabiliz, aldiz, defektibotasun gramatikalari egiten zaio erreferentzia. Ingelesez lehena da erabiliagoa, *light verb* (Butt, 2010; Kearns, 1988; Oyharçabal, 2003; Stevenson *et al.*, 2004); gaztelaniaz, ostera, nahiko zabalduta dago bigarrena ere, *verbo de apoyo* (Alonso Ramos, 2004; García García, 2005). Guk, lan honen batasun terminologikoari euste aldera, *arin* adjektiboa erabiliko dugu hemen.

Lehenago ere azaldu dugunez, aditz arindun konbinazioak kolokazioen eta lokuzioen artean kokatu ohi dira, baina badirudi ohikoagoa dela kolokazioen azpimultzotzat hartzea (Alonso Ramos, 2004; Bustos Plaza, 2005), eta joera horrekin bat egingo dugu guk. Halako konbinazioak definitzean, esaten da aditzak bere eduki semantikoa galtzen duela, osorik edo ia osorik, eta beste hitzak, berriz, gorde egiten duela bere berezko esanahia (Tognini-Bonelli, 2001: 116. orr.). Buckinghamen arabera (2009: 19. orr.), aditzarekin batera doan hitza izena izaten da, adjektibo edo aditz batetik eratorria askotan, eta aditzak gehitzen duen bakarra balio aspektuala, kausatibotasuna edo diatesia eta gramatika-markak dira. Adposizioren bat ere ager daiteke izenari laguntzen.

PARSEMEren irizpideetan ere (Ramisch *et al.*, 2018; Savary *et al.*, 2018), ezaugarri oso antzekoak aipatzen dira, beste hitz batzuk erabiltzen badira ere: esan beharrean izena aditz edo adjektibo batetik eratorria izan ohi dela, esaten da izen horrek ekintza bat edo egoera bat adierazten duela. Ekintzak adierazteko kategoriarik ohikoena aditzena izanik (Altuna Díaz, 2018: 65–67. orr.) eta kontuan harturik adjektiboek maiz adierazten dituztela egoerak (2018: 74–75. orr.), bistan da badagoela lotura batak eta besteek esaten dutenaren artean. Gainera, PARSEMEren irizpideetan ere esplizituki aipatzen da aditz arinak bitarikoak izan daitezkeela: batetik, hitz-konbinazioari ezaugarri morfologikoak baino gehitzen ez dizkiotenak (*lan egin*), eta, bestetik, balio kausatiboa ere gehitzen diotenak (*min eman*). Bat datoz, hortaz, bigarren ezaugarri horri dagokionez ere.

Normalean, aditz arindun konbinazioez eta, oro har, kolokazioez hitz egiten denean, osagaiak sinonimoen bidez ordezkatzeko ezintasuna aipatzen da. Alonsok (2004: 52. orr.) dioenez, ordea, badira kasu batzuk non posible den

serie gisakoak osatzea, izen berbera aditz batekin baino gehiagorekin konbinatuz. Hala, antzeko esanahiak adieraztea lortzen da, baina aldea egon daiteke batzuen eta besteen ezaugarri aspektualetan (26. adibidea).

(26) ***tomar forma, tener forma, perder forma*** → hasierako fasea, jarraipen-fasea eta amaierako fasea

Bestalde, aditz bat edo beste bat aukeratzea dialektoaren araberakoa ere izan daiteke (Wierzbicka, 1982; Corpas, 2015). Gaztelaniaz, adibidez, hotzeria harrapatzeari ez zaio berdin esaten leku guztietan, eta Espainian ez da Txilen eta Mexikon erabiltzen den aditz berbera erabiltzen (Molero, 2003):

(27) ***coger*** *un* ***resfriado*** → Espainian

(28) ***pescar/pillar*** *un* ***resfriado*** → Txilen eta Mexikon

Aditz arindun konbinazio batzuek esanahi bereko aditz sinple homologo bat izaten dute, baina ez denek, beste askotan forma analitikoa bakarrik erabiltzen baita, eta ez sintetikorik (Rafel, 2004: 406. orr.). Gainera, aditz sinple homologo bat duten konbinazioen kasuan, forma sintetikoa eta forma analitikoko izena morfologikoki erlazionaturik egoten dira askotan (29a adibidea). Hala ere, beti-beti ere ez da hala (Piera eta Varela, 1999; 29b eta 29c adibideak), eta bi formak erlazionaturik daudenean ere ez dute beti esanahi berbera izaten, askotan aldaketak gertatzen baitira argumentu-egituran (Rafel, 2004: 408–410. orr.; Sanroman Vilas, 2017; 30. adibidea).

(29) a. ***dar*** *un* ***beso*** → *besar*

b. ***dar clase*** → *enseñar*

c. ***hacer huelga*** → **huelgar*

(30) *Nos mandaron* ***hacer silencio****.*

? *Nos mandaron* <u>*silenciar*</u>*.*

Aditz arindun konbinazioak oso ohikoak dira guk landuko ditugun bi hizkuntza nagusietan, gaztelaniaz eta euskaraz (Rafel, 2004; Zabala, 2004), bai eta beste hizkuntza askotan ere (Butt, 2010). Euskaraz, badirudi halakoak bereziki usuak direla –inguruko hizkuntzen aldean behintzat–, eta presentzia handia dute bai gramatiketan (Hualde *et al.*, 2003: 223–227, 235–246. orr.; Etxepare, 2003: 302–308. orr.; Zabala, 2003) eta bai ikerketa arloko beste hainbat lanetan ere (Oyharçabal, 2003; Zabala, 2004; Martinez, 2015).

Zabalak (2004) xehetasun handiz deskribatzen ditu euskarazko *predikatu konplexuak* –edo, gure hitzetara ekarrita, aditz-UFak–, hainbat irizpideren arabera. Batetik, halako konbinazioetako aditzak multzokatzen ditu, eta bereizi egiten ditu aditz arinez eta kopulatiboez osaturikoak (hurrenez hurren, *negar egin* eta *gose izan*, adibidez). Bestetik, aditzarekin batera doan sintagma nolakoa den, multzoak osatzen ditu:

- Aditz arindunen barruan, izen-sintagmadunak (*min eman, txaloak jo*), kasu-marka daramaten izen-sintagma determinatudunak (*loak hartu, bideari lotu*) eta postposizio-sintagmadunak (*kontuan hartu, harira etorri*)
- Aditz kopulatibodunen barruan, izen-sintagmadunak (*beldur izan, giro egon*) eta adjektibo-sintagmadunak (*posible izan, oker ibili*)

Gainera, lehen multzokoen artean, tarte berezia eskaintzen die izena+*egin* motako konbinazioei, ohikoenak izateaz gain aparteko berezitasunak dituztela iritzita. Oyharçabalek ere (2003) era horretako konbinazioei begiratzen die, eta haien egitura era batera baino gehiagotara uler daitekeela dio, hiztunaren eta dialektoaren arabera. Adibidez, *lan egin* hitz-konbinazioa hiru eratara erabil daitekeela dio: absolutibo-markadun izen-sintagma determinatua eta aditza (31. adibidea), absolutibo-marka espliziturik gabeko izen-sintagma eta aditza (32. adibidea) eta izen inkorporatua eta aditza (33. adibidea).

(31) ***Lan*** *ederra/ gutxi* ***egin*** *dugu.*

(32) *Ederki/Gutxi* ***egin*** *dugu* ***lan****.*

(33) *Ederki/Gutxi* ***lan egin*** *dugu.*

Martinezek ere, bere doktoretza-tesian (2015), izena+*egin* motako konbinazioetako izenek zer inkorporazio maila duten aztertzen du. Euskal hizkuntzalaritzako lanak eta nazioartekoak hartzen ditu abiapuntutzat, baina esan genezake aurrekarien artean pisu handia dutela Zabalaren lanak (2004) eta Rodríguez eta García Murga-renak (2003).

Martinezen ondorio nagusia da izena+*egin* motako konbinazioek –zeinak, bidenabar, aditz-lokuziotzat hartzen baititu– hainbat inkorporazio maila dituztela, bai morfosintaxian eta bai semantikan ere. Honela laburbil genitzake haren bi ideia nagusiak:

- Morfosintaxiari dagokionez, izenaren inkorporazioa hiru mailatakoa izan daiteke, baina *continuum* bat osatzen dute hirurek:

  - Lotura estua erakusten dutenak
  - Hurrenkeraren haustura erakusten dutenak
  - Joera bikoitza erakusten dutenak, osagaien arteko unitate banaezina nahiz haustura

- Semantikari dagokionez ere, askotarikoak dira izena+*egin* konbinazioak. Oro har, esanahi idiomatikoa dutenak inkorporazio sintaktikorik handienekoenak dira.

Hortaz, Oyharçabalen eta Martinezen inkorporazioa bateragarria da, nolabait, guk orain arte erabili izan dugun finkapen morfosintaktikoaren ideiarekin. Honako hau dio, hain zuzen, Martinezek (2015: 378. orr.): "erabilera-ren poderioz egitura batzuek ihartzera jo dute, eta zurruntasun sintaktikoa erakusten duten horiek dira inkorporazioaz azaldu ditugunak." Bestalde, testetan oinarritzen du bere azterketaren zati handi bat, eta horrek ere agerian uzten du lan hori bateragarria dela orain arte esandakoekin (ikus, bereziki, 2.1.3.1. atala). Honako ezaugarri hauek hartzen ditu bere testen ardatz: galdegaia, partitiboa, agintera, koordinazio-egituretako elipsia, mendeko perpaus osagarria eta lokailuak. Horietako gehienak aipatu ditugu Parra Escartín *et al.*-en (2018) eta Buckinghamen (2009) lanez hitz egin dugunean[10].

2.1.4 Fraseologia eta itzulpengintza

UFez eta itzulpengintzaz hitz egiten denean, orokortutako ideia da halako hitz-konbinazio askori ezin izaten zaiela hitzez hitzeko ordainik eman, eta ataza hori itzulpengintzaren erronkarik handienetakotzat jotzen dute askok (Cobeta Melchor, 2002; Richart Marset, 2008; Timofeeva, 2012). Hain zuzen ere, orain arte aipatu ditugun ezaugarri linguistiko guztien ondorioz gertatzen da hori, nolabait: konposizionaltasun ezak, finkapen lexikoak eta morfosintaktikoak, instituzionalizazioak eta halako ezaugarriek eragiten dute UFak itzultzeko lana hain korapilatsua izatea. Bada loturarik, beraz, bi diziplinen artean, eta lotura horien inguruan jardungo dugu atal honetan.

Hasteko, esan beharra dago zenbait egilek UFen definizioan bertan sartzen dutela hitz-konbinazioen itzulgarritasuna –edo, hobeto esanda, UFei hitzez hitzeko ordainak emateko ezintasuna–. Bar-Hillelek (1955: 50. orr., Richarten

[10]Lehenago zerrendatu ditugun test horiek eta Martinezenak alderatzeko, ohar bedi galdegaiak lotura duela topikalizazioarekin, eta partitiboak, ezezko egiturekin.

2008ko lanetik), adibidez, hitz-konbinazio baten idiomatikotasuna itzulpenaren mende definitzen du, eta dio hitz-konbinazio bat idiomatikoa dela baldin eta ez bada itzulpen ulergarririk edota egokirik lortzen sorburu-hizkuntzako konbinazio horri helburu-hizkuntzan ordain literal eta gramatikalki baliokide bat emanda. Hizkuntzaren barnean idiomatikoak diren hitz-konbinazioak ere badaudela aitortzen du, baina aparteko motatzat tratatzen ditu, eta *monolingual idiom* ('esapide idiomatiko elebakar') esaten die.

Beste egile batzuek (Morvay, 1996; Corpas Pastor, 2003), ostera, akatsa deritzote hitzez hitzeko itzulgarritasun eza UFen funtsezko ezaugarritzat hartzeari. Guretzat ere, aurreko ataletan argi utzi dugunez, fraseologia hizkuntzaren barneko fenomenoa da –edo, hobeto esanda, hizkuntzaren barnekoa *ere* bada–, eta askotxo iruditzen zaigu hitzez hitzeko itzulgarritasun eza UF guztien ezaugarria dela esatea. Egia da konparazioak egitea beharrezkoa dela hitz-konbinazio bat idiomatikoa den ala ez jakiteko (Greimas, 1960: 50. orr.), hala bakarrik jakin baitaiteke zerbait ohiz kanpokoa –edo idiosinkrasikoa– den benetan, baina konparazio horiek hizkuntza jakin baten barruko joerekikoak ere izan daitezke, eta ez nahitaez bi hizkuntzaren edo gehiagoren artekoak. Dena dela, ukaezina da fraseologia, oro har, asko aldatzen dela hizkuntza batetik bestera, eta kontuan hartu beharreko fenomenoa da, inondik ere, itzulpengintzan. Ildo beretik, itzulpenei begiratzea lagungarria gerta daiteke hitz-konbinazio bat UFa den ala ez ebazteko.

Itzulpen-teorian, **itzulgarritasuna**ren eta **hizkuntzen arteko baliokidetza**ren ideiak eztabaida luzea sortu du, eta UFak, askotan hitzez hitz itzulgarriak ez direnez, bete-betean sartzen dira auzi horretan. Jakobson-ek (1959) itzulpen deritze bi kode desberdinetan emaniko mezu baliokideei, eta dio desberdintasunean baliokidetza lortzea dela itzulpengintzaren –eta, oro har, hizkuntzalaritzaren– arazo nagusietako bat; *equivalence in difference* esaten dio. UFen esparrura etorrita, Corpasek (2003: 206–208. orr.), beste egile batzuek bezala (Snell-Hornby, 1986; Mellado Blanco, 2000), mailaka sailkatzen du baliokidetza fraseologikoa[11]:

- **Baliokidetza osoa**, hizkuntza bateko eta besteko UFek denotaziozko eta konnotaziozko esanahi berbera dutenean. Bestela esanda, hizkuntza bateko eta besteko UFak baliokideak direnean bai forman eta bai esanahian.

  (34) ***meter la pata → hanka sartu***

[11] Adibideak gureak dira. Ordainak *Elhuyar* hiztegitik hartu ditugu.

- **Baliokidetza partziala**, hizkuntza bateko eta besteko UFak, ordaintzat hartzen badira ere, desberdinak direnean denotaziozko edo konnotaziozko esanahiari dagokionez, edo hizkuntza bateko UFa unitate lexiko bakarraz itzultzen denean beste hizkuntzara.

  (35) ***estirar la pata*** → ***azken hatsa eman***

  (36) ***lan egin*** → *trabajar*

- **Baliokidetzarik eza**, hizkuntza bateko UF batek ez duenean ordainik beste hizkuntzan. Halakoetan, parafrasietara eta halako tekniketara jo beharra dago.

  (37) ***estar de Rodríguez*** → familia oporretara joandakoan lanagatik etxean geratzea

Dobrovol'skij-ek (2011) ere antzeko irizpideak darabiltza, eta bi multzo bereizten ditu: baliokidetza *osoa*, Corpasen lehen multzoaren parekoa, eta *funtzionala*, bigarren eta hirugarren multzoen bilketatzat har genezakeena. **Baliokidetza funtzionala** definitzean, Dobrovol'skijek dio hizkuntza bateko eta besteko adierazpideak, forma aldetik desberdinak izan arren, testuinguru berean erabil daitezkeela informazio-galerarik sortu gabe. Baliokidetza osoa gutxitan gertatzen dela uste du, baina funtzionala lor daitekeela eta ez zaiola zertan esapide idiomatiko bati beste esapide idiomatiko bat eman ordaintzat, askotan hitz bakar batek edo kolokazio batek hobeto adieraz baitezake sorburu-hizkuntzako esapidearen esanahia. Horrek erakusten du Corpasen hirugarren multzoa ere baliokidetza funtzionaltzat hartzen duela berak.

Zuluaga (1999) ere funtzionaltasunaren kontzeptuaren alde agertzen da, eta okertzat jotzen du eduki batzuk itzulezinak direla uste izatea, hain zuzen horrexetan baitatza itzulpena haren ustez, baliokidetzak bilatzean eta hautatzean. Arazoari ondo heltzeko, itzulpengintza kontzeptu erlatibo gisa ulertu behar dela dio, Albrechten (1990) bidetik: jatorrizkoak eta itzulpenak ezaugarri komunak eta desberdinak dituzte beti, zeri begiratzen zaion (komunitate jakin baten iruditeriari, estiloari, testuak hartzaileengan duen eraginari...), eta ez dago baliokidetza absoluturik, baliokidetza partzialak baizik.

Horrez gain, Zuluagak beste bi lanekin lotzen ditu aurretik aipatutako ideia horiek. Batetik, Lyonsen (1968) aplikazio-berdintasunaren kontzeptuarekin, zeinaren arabera hizkuntza desberdinetako bi adierazpidek aplikazio

berbera baitute egoera berean erabil badaitezke. Eta, bestetik, Komissaroven (1981) eta Coseriuren (1978) denotaziozko ereduarekin, eredu horrek defendatzen baitu jatorrizko testua eta itzulpena errealitate berberaren bi agerpen linguistiko direla.

Beraz, orain artekoak kontuan hartuta, esan dezakegu posible dela UFei ordain funtzionala ematea, hau da, UF jakin batek sorburu-hizkuntzako testuan duen esanahiaren baliokide bat ematea xede-hizkuntzako testuan. Hala ere, badirudi egile gehienak datozela bat esatean halako baliokidetza asko partzialak baino ez direla izaten, eta horrek esan nahi du hizkuntza bateko UFei sarri ematen zaizkiela forma aldetik parekoak ez diren hitzak edo hitz-konbinazioak ordain gisa.

UFei ordaina emateko prozesuari dagokionez, Bakerrek (1992: 65. orr.) **bi arazo nagusi** aipatzen ditu:

> «The main problems that idiomatic and fixed expressions pose in translation relate to two main areas: the ability to recognize and interpret an idiom correctly; and the difficulties involved in rendering the various aspects of meaning that an idiom or a fixed expression conveys into the target language.»

Hortaz, Bakerren arabera, sorburu-hizkuntzako UFa ezagutu behar da batetik, eta hari ordain egokia eman behar zaio bestetik, xede-hizkuntzaren ezaugarriak kontuan hartuta. Corpasek (2003), berriz, hiru fase bereizten ditu prozesu horretan: (1) identifikazioa, (2) interpretazioa eta (3) ordainen bilaketa. Faseak bata bestearen atzetik datozela uste du; izan ere, UF bat ez bada behar bezala identifikatzen, ezin da behar bezala interpretatu, eta ordainen bilaketa zaildu eta okertu egiten da. Dena den, lehen bi faseak estuki lotuta daudela dio, eta aldi berean egiten direla normalean, UF bat ondo identifikatzeko beharrezkoa baita hura ondo interpretatzea ere –eta alderantziz–. Beraz, funtsean, bat datoz Baker eta Corpas.

Identifikazioak hainbat zailtasun ditu. Izan ere, UFen finkapena eta idiomatikotasun lexiko-semantikoa gako baliagarriak dira halakoak identifikatzeko, baina ezaugarri idiosinkrasiko horiek ez dira beti begi-bistakoak hiztun –edo itzultzaile– guztientzat (Corpas Pastor, 2003). Ildo horretatik, Timofeevak (2012) uste du itzultzaileek sekuentzia diskurtsiboaren osagai arruntzat dituztela UF batzuk, eta ez dutela haiek identifikatzean pentsatu ere egiten. Hori gertatzen da, besteak beste, hainbat kolokaziorekin eta errutinazko formulekin, hitz bakar baten bidez itzultzen direnekin bereziki, itzultzaileen

hizkuntzekiko kontzepzioaren barruan sartzen baitira halakoak, beste hainbat eta hainbat ezaugarri bezala –tipologia sintaktikoa bezala, adibidez–.

Ordainen bilaketari dagokionez, bi mailatan egiten dela dio Corpasek (2003): maila lexiko-semantikoan eta maila diskurtsiboan. Maila lexiko-semantikoan, lehenago aipatu ditugun baliokidetza motak sartzen ditu (osoa, partziala eta baliokidetzarik eza); maila diskurtsiboa, aldiz, aurreko fase guztien bilketa dela dio, puntu horretan agertzen diren ahulguneak aurreko faseetatik ekarriak baitira. Geroago ikusiko dugunez, identifikazioa eta ordainen bilaketa bi pausotan bereizte hori berdin-berdin aplikatzen da itzulpengintza automatikoan, eta hala antolatu dugu guk ere gure lana.

Horrez gain, badirudi **UF motak** baduela eragina itzulpenaren zer-nolakoan. Izan ere, Richartek (2008), Bakerren (1992) lanari jarraituz, dio zailagoa dela esapide semantikoki idiomatikoei ordaina ematea kolokazioei eta halakoei ematea baino. Esapide idiomatiko horiek "itzulpenarekiko erresistentetzat" jotzen ditu, eta idazketarekin parekatzen du haien itzulpena, itzultzaileek berridazketa-lan bat egin behar izaten dutela argudiatuta.

Kolokazioei dagokienez, ostera, honako hau dio Corpasek (2015): "though bases are usually translated literally, collocates do not seem to follow this straightforward path: *to pay homage* cannot be translated into Spanish as **pagar homenaje*, but prototypically as *rendir homenaje*." Alegia, kolokazio jakin bateko oinarriari –gure kasuan, izenari– bere ohiko ordaina ematen zaio Corpasen arabera, baina kolokatua –gure lanean, aditza– sarri aldatzen da hizkuntza batetik bestera. Horrez gain, lan horretan, argudiatzen du dialektoen aukeraketak garrantzi handia duela itzulpenean, eta beharrezkoa dela ezaugarri horri ere begiratzea, xede-hizkuntzako testua naturala izango bada.

Horiek horrela, gatozen orain **euskarazko lanetara**. Esan beharra dago bilduma fraseologikoak aspalditik egin direla gurean eta ez direla gutxi euskarazko UFak beste hizkuntza batzuekin parez pare jarri dituzten egileak. Hor ditugu Garibai (atsotitz-bilduma Urquijok argitaratu zuen 1919an), Zamarripa (1913), Gilsou (1964), Izagirre (1981), Azkue (1989), Mokoroa (1990) eta Garate (2003), eta ezin ahaztu 1596ko *Refranes y sentencias*, euskal UFen lehen bilduma, egile ezagunik gabea (Lakarra, 1996). Horietako gehienei ez diegu tarte gehiagorik eskainiko hemen, baina bai Mokoroaren eta Izagirreren lanei, B. eranskinean, erreferentziatzat erabili ditugun fraseologia-baliabideen deskribapen labur bana egin baitugu. Gainerako bildumen inguruko xehetasun gehiago nahi dituenak Urizarren tesira (2012: 76–83. orr.) jo dezake, euskal fraseologiaren inguruko lanen ikuspegi zabal samarra ematen

baita han.

Bildumetatik harago, ordea, apenas egin da euskara kontuan hartzen duen itzulpen-ikerketarik fraseologiaren alorrean. Aierberena (2008) da euskarazko itzulpenetan fraseologia berariaz aztertu duen lan bakarretako bat –eta, guk dakigula, lehena–. Administrazio-hizkerari erreparatzen dio zehazki, eta, hala, fraseologia espezializatuaz dihardu, Cabréren (1999) eta Bevilacquaren (2001) lanetatik abiatuta.

UF espezializatuak aztergaitzat harturik, Aierbek hainbat ondorio interesgarri ateratzen ditu. Lehena, Euskaltzaindiak esandakoak onarpen zabala duela oro har, Euskaltzaindiak okertzat jotako forma linguistikoak gero eta gutxiago agertzen baitira administrazioko testu itzulietan. Bigarrena, euskara batuaren finkapen-faltak zuzenean eragiten diela erabilera espezializatuei eta, hortaz, fraseologia espezializatua ere finkatu gabea eta egonkortu gabea dela oraindik. Eta, azkenik, kontuan hartu behar dela euskaraz sortzen diren testu administratibo gehienak gaztelaniatiko itzulpenak direla; pixkanaka testuak euskaraz sortzeko joerak ere gora egin duela dio, eta, itzulpenen aldean, testu horiek ez dituztela hainbeste ahulgune fraseologia mailan.

Fraseologiaren finkapenari dagokionez, hizkuntza-baliabideek berebiziko garrantzia dute, eta, kontuan izanik testu espezializatuak biltzen dituzten euskarazko corpusak oso txikiak direla eta haietan kolokazioak erabiltzen direla batez ere, bada hutsune bat alor horretan. Izan ere, euskarazko bilduma fraseologikoek ez dituzte kolokazioak lantzen oro har (Gurrutxaga *et al.*, 2016), eta kontsulta fraseologikoak egiteko aukera ematen duten sareko corpusak, gaur-gaurkoz, ez dira espezializatuak (ñabardura eta zehaztapen gehiago, B. eranskinean).

Bestetik, Sanzek (2015b) UFen azterketa egiten du bere doktoretza-tesian, itzulpen literarioak oinarritzat harturik. Toury-ren (2012) hurbilpenetik abiatuta, itzulpen-portaera erregularrei erreparatzen die, eta arreta berezia eskaintzen die *estandarizazioaren legeari* –xede-hizkuntzako aukeren artean ohikoenen alde egiteko joerari– eta *inferentziaren legeari* –jatorrizko testuko egiturek xede-testuan duten eraginari–. Zehazki, UFak alemanetik euskarara nola itzuli izan diren deskribatzen du, alde batera utzirik UF bat beste UF baliokide batekin itzuli behar delako dogma (Farø, 2006).

Landutako bi lege horiei dagokienez, ondorio interesgarri bana ateratzen du. Batetik, estandarizazioaren legea ez dela erabat betetzen euskaratutako testuetan, zenbait itzultzaileren lanetan bai baita joera bat egunerokoan oso ohikoak ez diren UFak ere erabiltzeko, diskurtso literario jakin bat sortzeko ahaleginean. Eta, bestetik, aztertutako testuak alemanetik euskaratuak

izanik ere, maiz topatzen dela gaztelaniaren aztarna fraseologian, inferentzia gertatzen den seinale. Hala, lege-proposamen bat ere egiten du (Sanz Villar, 2015a: 226. orr.): "A hizkuntza batetik B hizkuntza gutxitu batera itzultzerakoan, B hizkuntza egoera diglosikoan baldin badago eta C hizkuntza nagusi batekin elkarrekin bizi bada, orduan C hizkuntzaren inferentzia egon daiteke B hizkuntzan."

Inferentziaren legeari loturik, bada euskarazko kalko fraseologikoen inguruko beste lan bat, Altzibarrek, Garcíak eta Alberdik (2011) egina. Lan horretan, kalko deritzote beste hizkuntza bateko elementu bat itzulpen literalaren bidez kopiatzeari, eta halakoek euskarazko komunikabideetako fraseologian zer-nolako eragina duten aztertzen dute. Haien ustez, kalkoak ez du konnotazio negatiborik berez, eta, are, euskaraz UF berriak sortzeko mekanismo baliagarritzat jotzen dute. Dena dela, gaztelaniaren eraginezko kalko okerrak ere usuak direla deritzote, eta halakoei eskaintzen diete arreta batik bat. Gure lana haienarekin lotze aldera, esan beharra dago berariaz aipatzen dutela kalko oker asko nominalizazioa gehiegi erabiltzetik datozela, hau da, behar ez denean ere izena+aditza motako hitz-konbinazioak erabiltzetik, gaztelaniaz hala erabili ohi direlako: *tener (buenas) sensaciones* → *sentsazio (onak) eduki* vs *ongi sentitu*. Hainbat hizkuntza-baliabide aztertu ondoren –besteak beste, *Zehazki* hiztegia (Sarasola, 2005) eta *Berria*ren estilo-liburua (Arrarats, 2006)–, euren iritzia ematen dute halako kalkoen onargarritasunari buruz, eta agerian uzten dute hizkuntza-baliabide horiek askotan zorrotzegi jokatzen dutela haien ustez, onargarriak diren kalko batzuk ere saihesgarritzat jotzen baitituzte.

Izan ere, badirudi kalko fraseologikoek kezka sorrarazi dutela euskal baliabideen sortzaileen artean. Altzibarren, Garcíaren eta Alberdiren (2011) arabera, *Berria*ren estilo-liburuak 30 kalko okerren berri ematen zuen 2006an, eta, kontuan harturik estilo-liburu hori oraindik ere eguneratzen ari direla etengabe, pentsatzekoa da kopuru horrek gora egingo zuela. Horrez gain, *Euskara batuaren ajeak* liburuan ere (Sarasola, 1997), ez dira gutxi kalko fraseologikoekin zerikusia duten sarrerak, ez eta EIMAren estilo-liburuko *Kalko okerrak* liburukian ere (Garzia, 2005). Bestalde, ezin ahaztu txosten honen hasieran (iii. orrialdea) eman dugun aipua jasotzen duen lana, *Joskera lantegi* (Garzia, 1997), zeinak sarreratik bertatik uzten baitu argi zer-nolako pisua duen fraseologiak itzulpengintzan eta, oro har, hizkuntza baten joskeran.

Horrenbestez, hementxe utziko ditugu UFen ikerketa linguistikoei dagozkienak, eta UFen tratamendu konputazionala zertan den azalduko dugu jarraian.

2.2 UFak Hizkuntzaren Prozesamenduan

Behin marko teorikoari dagozkionak argituta, egin diezaiogun begiratua marko praktikoari, UFen prozesamenduari. Lehenik, Hizkuntzaren Prozesamenduak (HPk) UFei dagokienez zer erronka dituen azalduko dugu (2.2.1. atala). Ondoren, atal bana eskainiko diegu UFen prozesamenduko bi ataza nagusiei, erauzketari (2.2.2. atala) eta identifikazioari (2.2.3. atala). Eta, azkenik, UFak itzulpen automatikoan nola aplikatu izan diren azalduko dugu (2.2.4. atala).

2.2.1 Erronkak

Aurreko kapituluan esan dugunez, UFen ezaugarri nagusia idiomatikotasuna da, esan nahi baita UFtzat hartzen ditugun hitz-konbinazioak ezohikoak direla hizkuntza-ezaugarriren bati –edo gehiagori– dagokionez. Ezohiko izaera horrek zaildu egiten du UFen prozesamendua, hizkuntza prozesatzeko metodo gehienek hitzez hitz edo morfemaz morfema egiten baitute lan. Hori hobeto azaltzeko, Constant *et al.*-ek (2017) PARSEME proiektuaren baitan eginiko lana hartuko dugu oinarritzat, eta han zerrendatzen diren ezaugarri arazotsuak ekarriko ditugu hona, HPn zer-nolako zailtasunak sortzen dituzten azaltzeko.

- **Agerkidetza arbitrarioa**. UFetako hautapen lexikoa arbitrarioa izateak arazoak sortzen ditu, adibidez, itzultzaile automatikoetan, hautapen hori hizkuntzaz hizkuntza egiten baita eta, horregatik, hitzez hitzeko itzulpen asko desegokiak edo lausoak izan ohi baitira (38. adibidea).

  (38) ES: ***meter ruido***
  EU: ***zarata atera*** (**zarata sartu*)

- **Konposizionaltasun eza**. Aurreko ezaugarria bezala, ezaugarri hau ere arazotsua da itzultzaile automatikoentzat, konposizionalak ez diren UFak ezin izaten baitira askotan hitzez hitz itzuli (39. adibidea).

  (39) ES: ***dormir a pierna suelta***
  EU: ***lo seko egon*** (**hanka soltera lo egin*)

- **Jarraitutasun eza**. UFko osagaien artean askotan kanpo-elementuak ager daitezkeenez, testuan hitz-segidak bilatzea ez da nahikoa UF asko identifikatzeko (40. adibidea).

  (40) ***Bete*** *gabe geratu diren* ***plazak***

- **Aldakortasuna**. Beti jarraituak ez izateaz gain, UF gehienak ez dira erabat finkoak formari dagokionez ere, eta malgutasun horrek zailtasunak sortzen ditu, besteak beste, analisi morfosintaktikoan. Izan ere, hitz-forma jakinak bakarrik bilatzea ez da nahikoa UFen agerpen asko ezagutzeko (41. adibidea), hitz-segidak bilatzea nahikoa ez den bezalaxe.

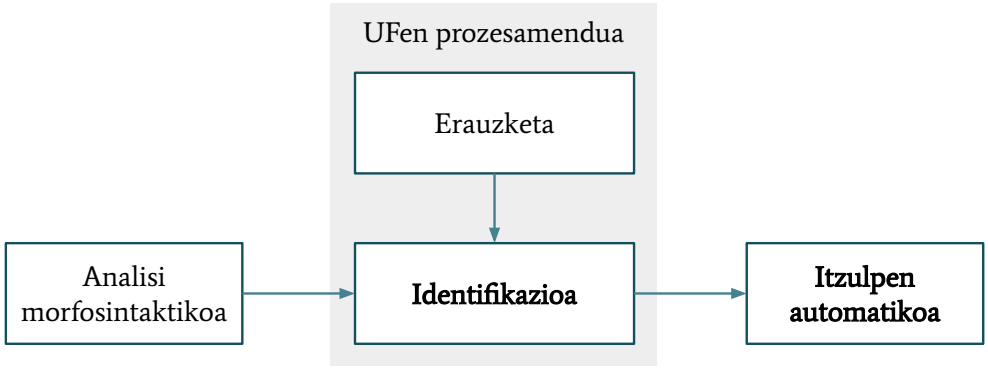
  (41) ***erabaki<u>a</u> hartu***
  erabaki *bat* ***hartu***
  erabaki<u>ak</u> hartu

- **Anbiguotasuna**. Hitz-konbinazio jakin bat esanahi idiomatikoaz ala literalaz erabilita dagoen bereizteak eragina izan dezake HPko hainbat atazatan, itzulpen automatikoan kasu (42. adibidea).

  (42) EU: ***ziria sartu***
  ES-lit.: *meter el palo*
  ES-idiom.: ***tomar el pelo***

Horiek guztiak kontuan harturik, UFen prozesamendua erronka handitzat hartzen da HPren alorrean, baina, aldi berean, aintzat ematen da haiek behar bezala tratatzea onuragarria izan daitekeela aplikazio askotarako. UFen prozesamenduak bi ataza nagusi hartzen ditu barnean, *erauzketa* eta *identifikazioa*, eta, Constant *et al.*-en arabera (2017), ataza horiek HPko aplikazioak hobetzeko aukeratzat hartu behar dira. Aplikazio horien artean bi aipatzen dituzte: analisi morfosintaktikoa eta itzulpen automatikoa. Hala, UFen prozesamenduko bi atazek eta aipaturiko bi aplikazioek elkarri eragiten diote[12] (2.6. irudia).

[12]Constant *et al.*-en (2017) irudian, gurean ez bezala, identifikaziotik aplikazioetarako geziak noranzko bikoitzekoak dira, eta aplikazioetatik erauzketara ere gezi bana ageri da. Guk irudi hori sinpletu dugu hemen, atazen eta aplikazioen arteko erlazioak zehazki gure lanean gertatzen diren bezala islatzeko. Gainera, beltzez jarri ditugu berariaz lantzen ditugun atalak: identifikazioa eta itzulpen automatikoa.



2.6 irudia – UFen prozesamenduaren eta bi aplikazioren arteko eragina

Defini ditzagun, beraz, bi ataza horiek, hobeto azaltzeko irudiko elkarreragina zertan datzan:

- **UFen erauzketa** deritzo corpusetatik UFak lortzeko lanari. Atazaren abiapuntutzat corpusak hartzen dira, eta UF-zerrenda bat lortzen da emaitza gisa.

- **UFen identifikazioa** esaten zaio aldez aurretik ezagutzen diren UFen agerpenak corpusetan bilatzeko lanari. UF-zerrenda bat behar izaten da atazari ekiteko, eta corpusaren gaineko etiketak dira emaitza.

Hori horrela, UFen erauzketa eta analisi morfosintaktikoa lagungarriak dira UFen identifikaziorako, eta UFen identifikazioa, itzulpen automatikorako (2.6. irudia). Identifikazio-lanerako beharrezkoa denez aldez aurretik UF-zerrenda bat izatea, erauzketa automatikoko sistemak baliagarriak izan litezke zerrenda hori osatzeko, fraseologia-baliabide osaturik ez dagoen kasuetan bereziki. Bestetik, aintzat harturik analizatzaile morfosintaktikoek corpusetako testuei etiketak gehitzen dizkietela (hitz bakoitzaren gramatika-kategoria, esaldi baten barruko hitzek zer dependentzia-erlazio duten elkarren artean, etab.), haien bidez lortutako informazioa ere baliagarria da UFen identifikazioa hobetzeko. Eta, azkenik, itzulpen automatikoan UFak ondo tratatu nahi badira, ezinbestekoa da UFak ondo indentifikatzea, ordaina ere behar bezala emateko.

Ikus dezagun, bada, jarraian, nola egin ohi den UFen erauzketa, nola identifikazioa, eta nola integratu izan den UFen inguruko informazioa itzultzaile automatikoetan.

2.2.2 Erauzketa

Esan dugunez, UFen erauzketa-lanaren helburua corpusetatik UFak lortzea da. Atal honetan, ataza horretarako zer-nolako lanak egin izan diren deskribatzen saiatuko gara, labur samar bada ere, ikuspegi orokor bat emate aldera. Xehetasun gehiago nahi dituenak eskura ditu ikuspegi orokor hori zabalago ematen duten beste lan batzuk, besteak beste, Baldwin eta Kimena (2010), Seretanena (2011), Ramischena (2015) eta Constant *et al.*-ena (2017).

Fraseologia konputazionalaren arloan, erauzketari garrantzi handia eman izan zaio 1980ko hamarkadaren amaieratik, hainbat egilek ohartarazi zutenetik ataza horrek zer-nolako garrantzia zuen HPren alorrean (Choueka, 1988; Church eta Hanks, 1990; Sag *et al.*, 2002; Bond *et al.*, 2003). Ordutik hona egin diren lan gehienak lexikoiak sortzera bideratuak izan dira, hiztegi orokorretan UFak gehitzeko edo UF-hiztegiak sortzeko asmoz eginak.

Aurreko atalean aipatu ditugun ezaugarrien artetik, biri arreta berezia jarri izan zaie ataza honetarako, ezaugarri gakoak direlakoan: agerkidetza arbitrarioari eta konposizionaltasunik ezari. Saia gaitezen hori hobeto azaltzen, UFak erauzteko erabili izan diren metodoak deskribatuz. Constant *et al.*-ek (2017) **lau metodo mota** bereizten dituzte:

- **Agerkidetza-neurrietan oinarritutakoak**. Corpuseko hitzen arteko agerkidetzaren garrantzia neurtzen dute, kontuan harturik (1) bi hitz –edo gehiago– zenbateko maiztasunez agertzen diren elkarrekin eta (2) zein den hitz horien banakako maiztasuna. Hala, hitzen banakako maiztasunetatik abiatuta, hitz horiek ausaz elkarrekin agertzeko probabilitatea kalkulatzen da, eta probabilitate hori alderatzen da corpusean benetan duten agerkidetza-maiztasunarekin. Agerkidetza-maiztasuna zenbat eta altuagoa izan ausazko probabilitatearen aldean, hitz-konbinazio jakin horrek orduan eta aukera gehiago ditu UFa izateko.

  Askotariko agerkidetza-neurriak erabiltzen dira lan hori egiteko: Pointwise Mutual Information (Church eta Hanks, 1990), Khi karratua (Dunning, 1993), Fisherren testa (Pedersen, 1996) eta beste hainbat. Hala ere, gaur arte behintzat, ez da lortu neurri horietan onena zein den

jakiterik, alderaketa-lan dezente egin bada ere (Pearce, 2002; Evert, 2005; Ramisch *et al.*, 2012; Garcia *et al.*, 2019). Aipagarria da, bestalde, halakoek emaitza onak lortzen dituztela bi hitzeko konbinazioei dagokienez, baina hitz gehiagoko hautagaiekin ezin dela hain ondorio orokorrik atera.

- **Ordezkapen-teknikak darabiltzatenak**. UFen murriztapen lexikoak eta morfosintaktikoak aintzat harturik, metodo hauek hitz-konbinazioen aldakiak sortzen dituzte automatikoki, eta aldaki artifizial horiek corpusean zenbatetan agertzen diren aztertzen dute. Pearcek (2001), esate baterako, hitz-konbinazioetako osagaiak WordNetetik lortutako sinonimoen bidez ordezkatzen ditu, eta sortutako konbinazio artifizialen maiztasunak alderatzen ditu jatorrizko hitz-konbinazioarenekin. Artifizialki sortutako konbinazioak oso gutxitan agertzen badira corpusetan, esan nahi du balitekeela jatorrizko hitz-konbinazioak murriztapenak edukitzea eta, hala, UFa izatea. Beste lan batzuetan, aldaki morfosintaktikoak sortzen dira artifizialki, aldaki lexikoak sortu beharrean (Villavicencio *et al.*, 2007; Ramisch *et al.*, 2008).

  Era horretako teknikak hainbat osaeratako UFak erauzteko erabili izan dira, besteak beste, aditza+partikula konbinazioak (McCarthy *et al.*, 2003), aditza+ izena esapide idiomatikoak (Fazly eta Stevenson, 2006; Cook *et al.*, 2007; Weller eta Heid, 2010) eta izen elkartuak (Farahmand eta Henderson, 2016). Duela gutxira arte, lan gehienek lexikoiak zerabiltzaten teknika horiek gauzatzeko (WordNet, VerbNet eta halakoak); azken urteotan, ordea, eredu distribuzionaletan oinarritutako lanak ugaritu egin dira, non hitzen arteko antzekotasuna automatikoki neurtzen baita corpuseko hitzen agerkidetzak kontuan hartuta (Riedl eta Biemann, 2015; Farahmand eta Henderson, 2016).

- **Antzekotasun semantikoari begiratzen diotenak**. Aintzat harturik UFen esanahia ez dela beti konposizionala, metodo hauek hitz-konbinazioen esanahia alderatzen dute konbinazioko osagai-hitzen esanahiekin. Gaur egun, esanahia konputazionalki errepresentatzeko erabiltzen diren metodo gehienak eredu distribuzionalen bidez eraikitzen dira: hitz jakin bat corpusean zer testuingurutan –zer beste hitzekin– agertzen den kontuan hartu, eta abstraktuki errepresentatzen da haren esanahia, bektoreen bidez normalean (Mikolov *et al.*, 2013). UFen erauzketa-lanerako, hitz-konbinazio usuei ere bektore bana sortzen zaie,

eta, ondoren, konbinazioaren bektorea osagai-hitzen bektoreekin alderatzen da.

Esate baterako, McCarthyk, Kellerrek eta Carrollek (2003) aditza+partikula motako hitz-konbinazioen eta haien barruko aditzen bektoreak alderatzen dituzte: *break up* eta *break*, *give up* eta *give*, eta abar. Hitz-konbinazioen bektorea urrun badago aditzaren bektoretik –*give up* eta *give*ren kasuan, adibidez–, esan nahi du aditzak ez duela bere ohiko esanahia gordetzen partikularekin batera doanean eta, hortaz, balitekeela UFa izatea. Beste egile batzuek (Baldwin *et al.*, 2003; Reddy *et al.*, 2011), berriz, hitz-konbinazioko osagai-hitz guztien banakako bektoreak batzen dituzte –adibidez, *adarra* hitzaren bektorea + *jo* hitzarena–, eta batura hori alderatzen dute hitz-konbinazioaren bektorearekin –adibidez, *adarra jo*renarekin–. Halako metodoek emaitza onak lortu dituzte lagin txikietan, aditza+partikula konbinazioei (Baldwin *et al.*, 2003; Bannard, 2005), aditza+izena esapide idiomatikoei (McCarthy *et al.*, 2007) nahiz izen elkartuei (Reddy *et al.*, 2011; Yazdani *et al.*, 2015; Cordeiro *et al.*, 2016) dagokienez.

- **Gainbegiratuak**. UF-zerrendak edo UFak etiketatuta dituzten corpusak oinarritzat hartu, eta haietatik ikasten dute hitz-konbinazio libreak eta UFak bereizten. Ikasketa automatikoko teknikak erabiltzen dira horretarako (Lapata eta Lascarides, 2003; Ramisch *et al.*, 2008; Rondon *et al.*, 2015), eta emaitzak oso onak izaten dira normalean. Hala ere, kontuan hartu behar da metodo hauek kanpoko datuak behar dituztela beti –UF-zerrendak edo corpus etiketatuak– eta datu horiek ez direla beti nahi bezain egokiak eta/edo osatuak.

Metodo horien bidez erauzitako hitz-konbinazioak **ebaluatu** egiten dira ondoren, erauzketa-metodoaren baliagarritasuna zenbatekoa den ikusteko. Erauzitako konbinazioen zerrenda eskuz aztertzea da ebaluazio-metodorik zehatzena, erauzitakoak benetan UFak diren ala ez jakiteko (Ferreira Da Silva *et al.*, 1999; Seretan, 2011). Askotan, ordea, eskuzko ebaluazioek denbora eta lan gehiago eskatzen dutenez, ebaluazio hori automatikoki egiten da, aurretik sortutako UF-zerrendaren bat oinarritzat harturik (Evert, 2009; Pecina, 2008; Yazdani *et al.*, 2015) edo emaitzak zuzenean hiztegi-sarrerekin alderatuta (Ramisch *et al.*, 2012; Riedl eta Biemann, 2015). Hala ere, kontuan hartu behar da bi aukera horiek ez direla erabat fidagarriak, emaitzak UF-zerrendekin edo hiztegi-sarrerekin bakarrik alderatuz gero, zerrenda ho-

rietatik kanpoko edozein hitz-konbinazio okertzat jotzen baita zuzenean, eta gerta baitaiteke jaso gabeko konbinazio batzuk zuzenak izatea.

Bada beste aukera bat ere, aipatu ditugun bi horiez gain: erauzitako zerrendak beste aplikazio edo ataza batzuetan integratzea eta horien emaitzak ebaluatzea zuzenean, identifikazio-sistemenak (Riedl eta Biemann, 2016), analizatzaile morfosintaktikoenak (Villavicencio *et al.*, 2007) edo itzultzaile automatikoenak (Ruiz Costa-Jussà *et al.*, 2010). Hala, ebaluazioa ziur asko hain zehatza ez bada ere, erauzketa-metodoek aplikazio errealetan zer eragin duten ikus daiteke.

Aipagarria da erauzketa-metodoen kalitatea asko aldatzen dela erabiltzen diren baliabideen arabera. Izan ere, entrenamendurako corpus txiki bat bakarrik erabiltzen bada, litekeena da informazio faltagatik ez lortzea oso emaitza onik, baina, era berean, eskala handiko esperimentuak ebaluatzea ere zaila da. Gainera, erauzi nahi den konbinazio motak ere eragin zuzena du emaitzetan, askoz ere errazagoa baita oso finkoak diren UFak automatikoki ezagutzea libreagoak direnak ezagutzea baino. Aditz-UFak dira erauzteko UF motarik zailenetakoak, eta horren adierazgarri da, besteak beste, Ramischek aditza+izena motako UFei *difficulty class* 'zailtasun-klase' izendapena eman izana bere sailkapenean (Ramisch, 2015: 41–44. orr.).

Hain zuzen ere, izena+aditza motako UFen gainekoa da euskaraz egin den erauzketa-lanik aipagarriena, Gurrutxagarena (2014). Bere doktoretzatesian, bi ataza nagusi izan zituen Gurrutxagak: UFak erauztea batetik, eta karakterizatzea bestetik. Erauzketarako, komunikabideetako testuz osatutako corpus bat hartu, automatikoki analizatu, eta aipatu ditugun tekniken konbinazio bat erabili zuen. Honako ezaugarri hauei erreparatu zien bereziki: agerkidetzari, antzekotasun distribuzionalari, malgutasun morfosintaktikoari eta malgutasun lexikoari. Ondoren, ezaugarri horien arabera, ikasketa automatikoko teknikak aplikatu, eta bi eratara karakterizatu zituen erauzitako UF-hautagaiak: ranking baten bidez, UF izateko aukera gehien dutenetatik aukera txikien dutenetara, eta sailkapen baten bidez, erauzitako konbinazioak lokuzioak, kolokazioak ala konbinazio libreak ziren ebazteko.

Oso emaitza onak lortu zituen bi atazetan ere, eta ondorioztatu zuen teknika semantikoek eta malgutasun morfosintaktikoaren neurketak bereziki laguntzen dutela, euskaraz behintzat, erauzketa–lanean. Eta, era berean, ikusi zuen UF-kategoriaren eta konbinazioaren aditzaren artean badagoela korrelazioa, eta korrelazio hori lagungarria dela UFak kategorizatzeko. Bestalde, espero zuenaren kontra, finkapen lexikoaren neurketak ez omen zion hainbeste lagundu, ez batean eta ez bestean.

2.2.3 Identifikazioa

UFen prozesamenduko bi atazen artetik, identifikazioa da, ziur asko, HPko aplikazioetan eraginik handiena duena. Esan dugunez, UF jakinen agerpenak testuetan hautemateari esaten zaio identifikazioa. Analizatzaile morfosintaktikoetan, anbiguotasuna txikiagoa izan ohi da UFak kontuan hartzen direnean (43. adibidea), eta, itzultzaile automatikoetan ere, hitzez hitz itzuli ezin diren UFak ezagutzeak itzulpen trakets asko ekidin ditzake (44. adibidea).

(43) a. *Istripua izan zuen;* ***hori dela eta****, ez du lanera etortzerik izango.*
→ Ondoriozko lokailua

b. *Esan didate atzerapenaren arrazoia hori dela eta pixka batean zain egon beharko dugula oraindik.*
→ Mendeko perpausa eta juntagailua

(44) ES: *Debemos* ***tomar las riendas*** *en este asunto.*
Matxin: *Aho-uhalak hartu behar ditugu gai honetan.*
EU-zuz: ***Agintea hartu*** *behar dugu gai honetan.*

Bi aplikazio horiez gain, beste batzuk ere hobetu daitezke UFen identifikazioaren bidez. Esate baterako, rol semantikoak etiketatzeko lanean, UFen identifikazioak eragina izan dezake, aditz arindun konbinazio eta lokuzio askotan argumentu-egitura ez baita aditzaren araberakoa. Estarronak, bere doktoretza-tesian (2014: 161–165. orr.), UFen barruan agertzeko joera duten zenbait aditz aipatzen ditu, eta azaltzen du zer-nolako zalantzak sortu dizkieten halakoek EPEC-RolSem corpusa etiketatzerakoan[13]. Kasu zalantzagarri horietako bat da, adibidez, *bat etorri* hitz-konbinazioa, *etorri* aditzak, UF horren barruan doanean, ez baititu onartzen normalean hain ohikoak dituen ablatiboa (abiapuntua, nondik) eta adlatiboa (helburua, nora).

Erronkez jardun dugunean aipatu dugunez (2.2.1. atala), UFek badituzte euren identifikazioa zailtzen duten hainbat ezaugarri, bereziki **jarraitutasun eza**, **aldakortasuna** eta **anbiguotasuna**. Ezaugarri horiek kontuan hartuta, mutur batean leudeke (1) hitz-segidak bere horretan bilatzen dituzten identifikazio-metodoak, hitz-segida finko aldaezinak balira bezala, eta beste muturrean, berriz, (2) hitz-konbinazioen forma flexionatu guztiak balizkotzat jotzen dituztenak. Bai mutur batekoek eta bai bestekoek dituzte mugak:

[13]Euskarazko aditzen argumentu-egiturak e-ROLda datu-basean kontsultatu daitezke: `http://ixa2.si.ehu.es/e-rolda/`.

1. Hitz-segida finkoak bilatzen dituzten metodoak UF mota jakin batzuentzat bakarrik dira aproposak, batere aldaketarik onartzen ez duten UFentzat bakarrik (adib.: *hala eta guztiz ere, horrez gain*). Izan ere, inongo flexio-aldaketarik kontuan hartzen ez dutenez, finkoak ez diren UFen agerpen gutxi batzuk bakarrik identifikatzeko gai dira (adib.: *ondorioak atera* bai, baina *ondorioa/ondoriorik/ondorio bat atera* ez). Aditz-UFen identifikaziorako murritzegiak dira, lehenago esan dugunez (2.1.3. atala), halakoak malgutasun morfosintaktiko handikoak izaten baitira.

2. UFko osagai-hitzen forma flexionatu guztiak aintzat hartzen dituzten metodoei, berriz, kontrakoa gertatzen zaie. Alde batera uzten dituzte UFen murriztapen morfosintaktikoak, eta, hala, agerpen gehiegi identifikatzen dituzte, bai UFenak eta bai UFenak ez direnak (Carpuat eta Diab 2010; Ghoneim eta Diab 2013). Esate baterako, *ziria sartu* UFaren agerpenak identifikatu nahi bagenitu eta ez bagenu kontuan hartuko *ziri* izena beti mugatu singularrean agertzen dela UFaren barruan, berdin-berdin identifikatuko lituzke *ziriak/ziriek/zirietan/ziriekin sartu* eta beste hainbeste, horiek UFak ez izan arren.

Bistan denez, tarteko metodoak behar dira UFen identifikazioa behar bezalakoa izan dadin, bereziki morfologia aberatseko hizkuntzetan (Alegria *et al.*, 1996; Urizar, 2012: 59. orr.). Hitz egin dezagun, jarraian, tarteko metodo horietako batzuei buruz.

Zenbait egilek **erregeletan oinarritutako teknikak** proposatu dituzte, metodorik sinpleenetik abiatuta, aldakortasun morfologikoa kontuan hartzeko. Freeling analizatzaile morfosintaktikoan (Padró eta Stanilovsky, 2012) eta Apertium itzultzaile automatikoan (Forcada *et al.*, 2011), adibidez, UFetako zati aldagarriak zehazten dira, eta zati horrek ager ditzakeen forma flexionatu guztiak zerrendatzen dira automatikoki, flexio-arauak erabiliz. Hala, esate baterako, gaztelaniazko *echar de menos* UFaren kasuan, *echar* aditza hartzen da zati aldagarritzat, eta UFaren agerpenak identifikatzen dira aditza edozein forma flexionatutan dagoela ere (*echamos, echa, echaron, echabais...*). Zehaztasun handiko metodoa da, baina agerpen dezente uzten dira alde batera, ez baita onartzen osagai-hitzen artean beste inongo hitzik agertzerik. Hortaz, hein batean aurre egiten zaio aldakortasun morfologikoaren arazoari, baina ez jarraitutasun ezak sortzen dituen zailtasunei.

Beste lan batzuetan, analizatzaile morfosintaktikoen emaitzak eta UFen

identifikaziorako berariaz sortutako arauak konbinatzen dira, bai arau orokorrak (Oflazer *et al.* 2004), bai UF motaka prestatutakoak (Copestake *et al.* 2002), eta bai UFz UF aplikatutakoak, lexikoi batean gordetako ezaugarriei eta murriztapenei begiratuta (Hashimoto *et al.* 2006). **Euskararen prozesamendua**n, azken multzoko metodo bat erabiltzen da gehien, Urizarren tesi-lanean (2012) garatua. Lan eskerga horren ekarpen garrantzitsuenetako bat da euskarazko 2.207 lokuzio EDBLn (Euskararen Datu Base Lexikalean, Aduriz *et al.*, 1998) deskribatu izana; datu-base horretan jasotzen dira, lokuzioetako osagai-hitzen gramatika-kategoriez eta buru sintaktikoaz gain, lokuzio bakoitzaren *gauzatze-eskemak* ere. Eskema horietan, lokuzio bakoitza testuan zer forma desberdinetan ager daitekeen zehazten da, hiru ezaugarriren bidez:

- Hurrenkera: osagai-hitzak zein hurrenkeratan agertzen diren, eta tartean beste hitzik sartu ote daitekeen –edo sartu behar ote den–.
- Flexio-murriztapenak: osagai-hitzetako bakoitzak zer flexio onartzen dituen.
- Ziurtasuna: osagai-hitzek aurreko baldintzak betetzen dituztenean beti ote diren UFaren agerpenak ala badagoen anbiguotasunik.

Lokuzio bakoitzaren deskribapena kontuan hartuta, identifikazio-lana Murriztapen Gramatika formalismoa (Karlsson *et al.*, 1995) erabiliz egiten da, berariaz horretarako sortutako HABIL tresnaren bidez. Hala, testuan UF bateko osagai-hitzak aurkitzen direnean, EDBLra jotzen da, eta HABIL tresnak aztertzen du UFaren murriztapen guztiak betetzen ote diren ala ez; betetzen badira, hitz-konbinazioa UFtzat hartzen da, eta bestela, ez.

Euskarazko lan horrez gain, aipagarria da MWEtoolkit tresna ere (Ramisch *et al.*, 2010), zeinak aukera ematen baitu bilaketa-heuristikoak moldatzeko, UFko osagai-hitzen arteko hitz-kopurua zehazteko, inguruko hitzen gramatika-kategoriak murrizteko eta abar. Metodologia orokor eta moldagarria du, edozein hizkuntzatarako da aplikagarria, eta erauzketarako nahiz identifikaziorako erabil daiteke.

Erregeletatik harago, izan dira lan batzuk UFen identifikazioa **adieradesanbiguazioko teknikak** erabiliz egin dutenak (Uchiyama *et al.*, 2005; Katz eta Giesbrecht, 2006; Cook *et al.*, 2007; Hashimoto eta Kawahara, 2008; Boukobza eta Rappoport, 2009; Sporleder eta Li, 2009; Tu, 2012).

Lan horietan, hitz-konbinazioen inguruko hitzei begiratzen zaie –eta, batzuetan, hitz horien ezaugarri morfosintaktikoei–, eta semantika distribuzionala erabiltzen da. Erauzketa-metodoez hitz egin dugunean azaldu dugun bezala (48. orrialdea), semantika distribuzionalean, hitzak eta hitz-konbinazioak abstraktuki errepresentatzen dira, bektoreen bidez, eta bektore horien arteko aldeei eta antzekotasunei begiratzen zaie hitz-konbinazioak sailkatzeko. Bada, identifikazio-lanean ere gauza bera egiten da: lehenik corpusean maiz agertzen diren hitz-konbinazioak erauzten dira, eta, jarraian, adieradesanbiguazioko tekniken bidez erabakitzen da hitz-konbinazio bakoitza benetako UFa den ala ez.

Ikasketa automatikoan oinarritutako metodoak ere erabiltzen dira. Corpus etiketatuak oinarritzat harturik, hitz-konbinazio jakin baten eta inguruko hitzen ezaugarriei begiratzen zaie, eta, horietatik abiatuta, ikasketa automatikoko hainbat algoritmo aplikatzen dira hitz-konbinazioak UFtzat edo konbinazio libretzat sailkatzeko. Teknika horietako gehienek hitz-segida jarraituak bakarrik identifikatzen dituzte (Blunsom eta Baldwin, 2006; Constant eta Sigogne, 2011; Shigeto *et al.*, 2013), nahiz eta egileren batek garatu duen eredu konplexuagorik, osagai-hitzak bereiz agertzen direneko agerpenak ere ezagutu ahal izateko (Schneider *et al.*, 2014).

Halako lanetan, oinarritzat hartzen den corpus etiketatuaren tamainak eta kalitateak zuzenean eragiten diote ikasketa-prozesuari: etiketatze-lana txikiegia edo irizpide garbirik gabea izan bada, oso litekeena da sistemak ondorio traketsak ateratzea entrenamendutik eta, hortaz, hautagaiak oker sailkatzea. Gainera, UFak berariaz etiketatuta dituzten corpusak urri samarrak dira, hizkuntza batzuetan bereziki (Losnegaard *et al.*, 2016).

Hain zuzen ere, baliabide-falta horrek bultzaturik, fraseologia mailan etiketatutako corpus eleaniztun bat sortzea izan da PARSEME proiektuaren ekarpen handietako bat. Corpus hori oinarritzat harturik, aditz-UFen identifikaziora bideratutako bi ataza partekatu antolatu dituzte, eta horietaz hitz egingo dugu jarraian.

2.2.3.1 PARSEMEren identifikazio-ataza partekatuak

PARSEMEren ataza partekatuak **hainbat pausotan** antolatu dira. Lehenik, corpusak sortu eta etiketatu dira, hizkuntza guztietan gidalerro berberei jarraituz[14]. Ondoren, corpus horiek bi zatitan bereizi dira: entrenamendu-

[14]Geroago emango dugu etiketatze-lanaren berri (7. kapitulua), eta ez dugu xehetasun gehiagorik emango orain. Gidalerroak interneteko lotura honetan daude eskuragarri:

corpusa eta testeko corpusa. Entrenamendu-corpusa etiketa eta guzti argitaratu da, parte-hartzaileek haren gainean presta zitzaten euren sistemak, eta testekoa, berriz, *itsu* –UF-etiketarik gabe–, ebaluazioa ahalik eta objektiboena izan zedin. Parte-hartzaileek testeko corpusak etiketatu dituzte euren sistemen bidez, eta antolatzaileek corpus horiek ebaluatu dituzte, eskuzko etiketekin alderatuz.

Erabilitako identifikazio-metodoen artean, denetarik egon da. Hona hemen laburpen orokor bat.

- **Lehen edizioa**n (Savary *et al.*, 2017), 7 sistemak hartu dute parte, eta gehien-gehienek analisi morfosintaktikoko teknikak erabili dituzte:
  - Foufi, Nerima eta Wehrli (2017) erregeletan oinarritutako analizatzaile eleaniztun batez baliatu dira.
  - Klyueva, Vernerová eta Qasemizadeh-ek (2017) eredu neuronaletan[15] oinarritu dute euren sistema.
  - Gainerako sistema guztiek ikasketa automatikoko teknikak erabili dituzte: Boros *et al.*-ek (2017) eta AlSaied *et al.*-ek (2017) trantsizioen bidezko dependentzia-analisia egin dute; Simkó, Kováks eta Vincze-k (2017) analizatzaile morfosintaktiko baten bi modulu erabili dituzte gramatika-kategoriak eta dependentziak analizatzeko; Maldonado *et al.*-ek (2017) eta Buljan eta Snajder-ek (2017) sekuentzia-etiketatzeak egin dituzte.
- **Bigarren edizioa**n (Ramisch *et al.*, 2018), berriz, 17 sistemak hartu dute parte:
  - Erdiak baino gehiagok neurona-sareak erabili dituzte (Berk *et al.*, 2018; Boros eta Burtica, 2018; Ehren *et al.*, 2018; Stodden *et al.*, 2018).
  - Bi sistema metodo estatistikoen eta agerkidetza-neurrien gainean eraiki dira.

`https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/`

[15]Eredu neuronaletan oinarrituak dira, adibidez, lehenago aipatu ditugun semantika distribuzionaleko teknikak. Halako ereduek giza garunaren egitura imitatu nahi dute, konputazionalki neurona-sare modukoak osatuz, eta gero eta gehiago erabiltzen dira, hizkuntza prozesatzeko ez ezik, irudiak prozesatzeko eta inteligentzia artifizialeko beste hainbat atazatarako ere.

– Gainerako sistemek ikasketa automatikoko teknikak erabili dituzte: horietako hiru[16] (Waszczuk, 2018) zuhaitz sintaktikoetan eta analizatzaile morfosintaktikoetan oinarritu dira, eta besteek beste algoritmo batzuk baliatu dituzte (Pasquer *et al.*, 2018; Moreau *et al.*, 2018).

Sistema horiek **ebaluatzeko**, doitasuna, estaldura eta F neurria erabili dira, HPren alorrean ohikoa den moduan. Esan bezala, PARSEMEren corpusaren testeko zatia hartu da oinarritzat, eta sistema bakoitzaren erantzunak eskuzko etiketekin alderatu dira. Labur azaltzeko, honela definitu litezke hiru neurri horiek, UFen identifikazioari dagokionez:

- **Doitasuna** (P) deritzo identifikatu diren UFetatik zuzenak zenbat diren adierazten duen neurriari, edo, bestela esanda, sistemak UF etiketa eman dien hitz-konbinazioetatik benetan UFak zenbat diren adierazten duenari.
- **Estaldura** (R) da sistemak identifikatu behar zituen UF guztietatik zenbat identifikatu dituen adierazten duen neurria.
- **F neurria**k aurreko biak biltzen ditu: $(P + R)/2$.

Emaitzak aztertuta, esan daiteke denetarik egon dela edizio batean zein bestean, baina bi datu azpimarragarriak iruditzen zaizkigu. Batetik, argi ikusten dela corpus etiketatuak eragin zuzena duela identifikazio-lanaren kalitatean, corpusik txikienak dituzten hizkuntzetan emaitza nabarmen okerragoak lortu baitira. Eta, bestetik, agerian geratzen dela 2.2.1. atalean aipatu ditugun ezaugarriek asko zailtzen dutela ataza: emaitzak 17 puntu jaisten dira batez beste UF jarraituetatik ez jarraituetara, eta 20 puntu batez beste entrenamendu-corpusekoen berdinak diren UF-agerpenetatik beste aldaki morfosintaktiko batzuetara.

Horrez gain, espero izatekoa denez, sistema guztiek izan dituzte nahiko emaitza kaskarrak entrenamendu-corpusean agertu ez diren UFak identifikatzen. Era horretakoak ere ezagutu nahi badira, ez da nahikoa UFak etiketatuta dituzten corpus txikiak bakarrik erabiltzea, UF gehienak oso gutxitan errepikatu baitira ia hizkuntza guztietako corpusetan (Savary *et al.*, 2017). Emaitzak alderdi horretatik hobetzeko, UF-lexikoi handiak edo erauzketarako erabiltzen diren beste teknika batzuk baliatu beharra dago.

[16]Ohar bedi sistema guztiek ez dutela argitalpenik.

Gure identifikazio-lanen berri ematean (2.2.3. atala), ataza partekatuko emaitzekin alderatuko ditugu guk lortutakoak. Beraz, momentuz, ez dugu xehetasun gehiagorik emango emaitza horien inguruan, eta HPren alorreko aplikaziorik konplexuenetako batera egingo dugu jauzi: itzulpen automatikora.

2.2.4 Itzulpen automatikoa

Itzultzaile automatikoen helburua da sorburu-hizkuntzako testu bat helburu-hizkuntzan ematea automatikoki, esanahia gordez eta naturaltasunari eutsiz. Behin baino gehiagotan esan dugunez, ordea, UF asko ez dira hitzez hitz itzultzen, eta hortaz, ez da harritzekoa itzultzaile automatikoek arazoak izatea hizkuntza batetik bestera UFen esanahia gordetzeko eta naturaltasunari eusteko lanean. Izan ere, 2.2.1. atalean aipatu ditugun ezaugarri arazotsuetatik den-denek eragiten diote aplikazio horri: agerkidetza arbitrarioki usuak, konposizionaltasunik ezak, jarraitutasun ezak, aldakortasunak eta anbiguotasunak.

Atal honetan, hasteko, gaur egun gehien erabiltzen diren itzultzaile automatiko motez jardungo dugu, eta, labur samar bada ere, argitzen saiatuko gara nola burutzen duten itzulpen-prozesua eta nola egiten dioten aurre UFen prozesamenduari. Horren ostean, gehixeago sakonduko dugu *Matxin* itzultzaile automatikoan (2.2.4.1. atala), huraxe erabili baitugu tesi-lan honetako zenbait esperimentutan.

UFen prozesamenduak itzultzaile automatikoetan zer leku duen azaltzen hasi aurretik, deskriba ditzagun, oro har, gaur egun gehien erabiltzen diren hiru **itzulpen-sistema motak**[17].

- **Erregeletan oinarritutako itzulpen automatikoa**. Lexikoi elebidun zabalak eta erregela linguistikoak erabiltzen dituzte itzulpen-prozesurako. Erregela horietan deskribatzen da hizkuntza batetik besterako transferentzia nola egiten den, nola aukeratzen den hitz bakoitzarentzako ordain egokia, eta zer gramatika-ezaugarri eta murriztapen dituzten bai sorburu- eta bai helburu-hizkuntzak.

[17] Adibideetan oinarritutako itzulpen automatikoa apartekotzat hartzen dute zenbaitek (Somers, 1999), baina guk ez diogu tarterik eskainiko metodo horri, euskaraz ez baita halako sistemarik garatu gaur arte –guk dakigula–, eta beste egile askok ere hiru mota bakarrik desberdintzen baitituzte (Aranberri eta Labaka, 2017)

- **Itzulpen automatiko estatistikoa**. Aurrekoek ez bezala, sistema estatistikoek corpus paraleloak eta elebakarrak dituzte oinarrian, eta ikasketa automatikoko teknikak erabiltzen dituzte itzultzen ikasteko. Ikasketa automatikoko edozein atazak bezala, itzulpen-prozesuak bi fase nagusi izaten ditu: entrenamendu-fasea eta deskodetze-fasea. Entrenamendu-fasean, sistemak hizkuntza-baliabideetatik ikasten du, eta bi eredu sortzen ditu: itzulpen-eredua –corpus paraleloetatik abiatuta–, eta hizkuntza-eredua –helburu-hizkuntzako corpus elebakarretatik abiatuta–. Metodo probabilistikoak erabiltzen dira horretarako, testuak hitz-segidaka zatituz[18]. Deskodetze-fasean, berriz, aurreko fasean ikasitakoa praktikan jarri, eta itzulpena ematen zaio itzulgai berri bati. Sorburu-hizkuntzako esaldi bakoitzarentzat itzulpen hipotetiko batzuk sortzen dira, eta probabilitaterik altuenekoa aukeratzen da. Halako sistema gehienek hitz-segidetan bakarrik jartzen dute arreta (Koehn *et al.*, 2003), baina beste batzuek egitura sintaktikoak eta bestelako datu linguistikoak ere hartzen dituzte kontuan (Chiang, 2007; Koehn eta Hoang, 2007; Koehn, 2010).

- **Itzulpen automatiko neuronala**. Hauek ere itzulpen-probabilitateak kalkulatzen eta erabiltzen dituzte, baina beste metodo batzuen bitartez. Neurona-sareetan oinarritzen dira, eta kodetze- eta deskodetze-arkitektura bat izaten dute, hau da: sorburu-hizkuntzako itzulgaia abstraktuki errepresentatzen –edo *kodetzen*– dute lehenik, bektoreen bidez, eta bektore horiek deskodetzen dituzte jarraian, helburu-hizkuntzako itzulpena lortzeko (Kalchbrenner eta Blunsom, 2013; Cho *et al.*, 2014). Hitzei zein ordain eman erabakitzeko, esaldi berean aurretik jarritako ordainak ere erabiltzen dira probabilitaterik altueneko aukera zein den kalkulatzeko. Halako sistema gehienek ez dute hitz-formez haragoko informazio linguistikorik erabiltzen, sistema bakanen batek salbu (Sennrich eta Haddow, 2016).

Sistema estatistikoak eta neuronalak, corpusetan oinarritzen direnez eta hitz-segidei erreparatzen dietenez, gai izaten dira UF jarraituak –edota tartean hitz gutxi dituztenak– ondo itzultzeko. Azaldu dugu, ordea, UF asko ez direla beti jarraituak izaten, eta ezin daiteke esan fraseologiak halako

[18]Bosgarren kapituluan, 163. orrialdean zehazki, prozesu horren adibide bat emango dugu, gure metodologiaren zati bat deskribatzeko. Beharbada argigarri gerta daiteke adibide hori, puntu honetako edukiak hobeto ulertzeko.

itzultzaileei inongo arazorik sortzen ez dienik. Bistan da, dena den, erregeletan oinarritutako sistemak direla zailtasun gehien dituztenak UFei ordaina emateko orduan, arau linguistikoak baitituzte oinarrian eta, hain zuzen ere, "arauz kanpoko" izaera baita –idiomatikotasuna– UFak definitzen dituena.

Itzulpen-estrategia edozein dela ere, UFen tratamendua bi pausotan egin beharra dago itzultzaile automatikoetan: identifikazioa batetik, eta ordain-ematea bestetik. Sistemaren arabera, zein metodo erabili nahi den, itzulpen-prozesuaren fase batean edo beste batean egin beharko da lan hori.

Erregeletan oinarritutako itzultzaileek, normalean, analisi-fasearen ondoren identifikatzen dituzte UFak, analizatzaile morfosintaktikoaren informazioa ere erabiltzen baitute hala. Sistema horien oinarria lexikoiak eta arau linguistikoak direnez, ezinbestekoa da lexikoiak UFak edukitzea, eta, lexikoi hori nahikoa zabala bada, UF finko samarrei nahiko ondo ematen zaie ordaina metodologia konplexurik gabe ere (Barreiro, 2008). Halakoetan, beraz, lexikoia eguneratuta izatea da garrantzitsuena, eta erauzketa-teknikak lagungarri gertatzen dira askotan, esate baterako, izen bereziak (Tan eta Pal, 2014) edo terminoak (Bouamor *et al.*, 2012) lexikoian biltzeko. Finkoak ez diren UFentzat, berriz, beste estrategia batzuk dira beharrezkoak, eta erregela gehigarrien bidez egin ohi dira bai identifikazio-lana eta bai ordain-ematea (Anastasiou, 2008; Forcada *et al.*, 2011; Monti *et al.*, 2011).

Wehrli *et al.*-ek (2009), adibidez, analisi-fasearen ondotik patroi linguistikoak aplikatuz identifikatzen dituzte UFak, eta, haiek helburu-hizkuntzara transferitzeko, zuhaitz sintaktikoen araberako errepresentazio formal bat ematen diete. Beste egile batzuek, berriz, identifikatzen dituztenetik transferitzen dituztenera bitartean, beste era batera kodetzen dituzte UFak, *interlingua* edo bitarteko hizkuntza gisara (Oepen *et al.*, 2004; Monti *et al.*, 2011: 2013). Bitarteko errepresentazio horietako bat funtzio lexikoen bidezkoa da (Mel'čuk, 1998), Zentzu-Testu Teoriaren barnekoa. Funtzio lexikoek hitz-konbinazio jakin bateko osagaien artean zer erlazio semantiko dagoen zehazten dute; Heylen, Maxwell eta Verhagen-en arabera (1994), *interlingua* moduan erabiltzen badira, kolokazio bateko oinarriaren ordaina edukitzea nahikoa da kolokazio osoa itzultzeko, oinarriaren eta funtzio lexikoaren bidez jakin baitaiteke zer kolokatu behar den helburu-hizkuntzan (45. adibidea). Horretarako, noski, informazio hori jasotzen duten lexikoi aberastuak behar dira, gaztelaniazko DiCE hiztegia (Alonso Ramos, 2017) eta halakoak.

(45) EN: **Magn**(smoker) = heavy
FR: **Magn**(fumeur) = grand

Itzultzaile estatistikoei dagokienez, berriz, badira lan batzuk entrenamendu-fasearen aurretik identifikatzen eta itzultzen dituztenak UFak. Lan horietan, UFak hitz bakarra balira bezala tratatzen dira gehienetan, eta hala ematen zaie ordaina, hitz bakarrekoa, hitz-konbinazioa edo, kasu batzuetan, parafrasia (Ullman eta Nivre, 2014). Carpuat eta Diabek (2010), lexikoi bat oinarritzat harturik, erregelen bidez identifikatzen dituzte UFak, eta osagaihitzak marra baten bidez lotzen dituzte, elkarrekin tratatu beharrekoak direla adierazteko. Beste lan batzuetan, berriz, erregela espezifikoak aplikatzen dira jarraitutasun eza eta aldaki morfosintaktikoak ere kontuan hartzeko, esate baterako, aditz arinei marka berezi bat jarriz (Cap *et al.*, 2015).

Lan gehienetan, ordea, UFen prozesamendua itzulpen-prozesuaren baitan egiten da. Bi aukera daude horretarako: entrenamenduko datuak egokitzea, edo entrenamendutik sortutako itzulpen-hautagaien aukeraketan eragitea. Lehen aukera erabiltzen dutenen artean, egile askok hiztegi elebidunetako UFak erauzten dituzte, eta entrenamendu-corpusean sartzen dituzte ondoren, gainerako esaldi-pareekin batera (Babych eta Hartley, 2010; Tan eta Pal, 2014). Bigarren aukeraren alde egiten dutenek, berriz, hainbat teknika erabiltzen dituzte: batzuek UFen erauzketarako metodoak integratzen dituzte sorburu- eta helburu-hizkuntzako hitz-segidak lerrokatzeko, eta hala lortzen dute itzulpen-hautagaien zerrendan UFak berariaz tratatzea (Bouamor *et al.*, 2012; Kordoni eta Simova, 2014; Pal *et al.*, 2013; Lambert eta Banchs, 2005); beste zenbaitek pauso bat gehiago ematen dute, eta UFen itzulpenei dagokien probabilitatea igoarazten dute zuzenean (Ren *et al.*, 2009); eta morfosintaxiari ere erreparatzen dioten sistema estatistikoetan, berriz, gramatikak ere erabiltzen dira prozesu horietan eragiteko (Na *et al.*, 2010).

Euskarazko itzultzaile automatikoetara etorrita, askotariko sistemak garatu dira ikerketarako[19]:

- *Matxin*, euskararako lehen itzultzaile automatikoa, erregeletan oinarritua (Mayor *et al.*, 2011)
- *EUSMT*, estatistikan oinarritua, zeinak kontuan hartzen baititu, hitz-segida hutsez gain, euskarazko hitzen osagai diren morfemak (Labaka, 2010)

[19]Zerrendan bildutakoez gain, ikerketa-helburuetara bideratuta ez dauden Eusko Jaurlaritzaren itzultzaileak (`http://www.itzultzailea.euskadi.eus/traductor/welcome.do`) eta Google Translatek ere (`https://translate.google.es/?hl=eu`) itzultzen dute euskarara/euskaratik.

- Sistema hibrido bat, *Matxin* eta *EUSMT* biltzen dituena (Labaka *et al.*, 2014)
- *MODELA*, eredu neuronaletan oinarritua (Etchegoyhen *et al.*, 2018)

Horien artetik, *Matxin*ek bakarrik du UFen tratamendurako estrategia espezifiko bat, orokor samarra bada ere. Geroxeago azalduko dugu hobeto nolakoa den tratamendu hori, non kokatzen den sistemaren itzulpen-prozesuan, eta zer indargune eta ahulgune dituen (2.2.4.1. azpiatala eta 3. kapitulua). Gainerako sistemek, aldiz, ez dute tratamendu berezirik, baina, lehen esan dugunez, gai dira askotan UF jarraituei ordain txukun samarrak emateko.

Ebaluazioari dagokionez, itzulpen-kalitatearen neurketak bi eratara egin ohi dira: eskuz eta metrika automatikoak erabiliz. Eskuzko ebaluazioetan, egileek zehaztu ohi dituzte irizpideak: batzuetan erroreen analisia egiten da, eta beste batzuetan, besterik gabe, itzulpenak ulergarriak ote diren aztertzen da. Metrika automatikoek, berriz, erreferentziazko itzulpenak hartzen dituzte oinarritzat, eta haiekin alderatzen dute sistemak egindako itzulpena (ikus azalpena beherago). MODELAk aurrerapen ikaragarria ekarri du euskarazko itzulpen automatikoaren esparruan, askogatik hobetu baititu aurreko sistemen itzulpenak, bai metrika automatikoen eta bai giza ebaluatzaileen arabera. Aurreko hiruren artean, berriz, ebaluazio-metodo batzuen eta besteen emaitzak ez datoz bat: metrika automatikoek sistema hibridoa jotzen dute kalitaterik handienekotzat, baina giza ebaluatzaileek, *Matxin* (Labaka *et al.*, 2014).

Izan ere, BLEU, NIST, TER eta halako metrikei hein batean bakarrik egin behar zaie jaramon, hain ebaluazio-irizpide itxia eta orokorra izanik, itzulpen-kalitatearen zantzu bat besterik ez baitute ematen. Azal dezagun, oro har, zertan datzan aipatutako hiru metriketako bakoitza:

- **BLEU** da itzulpen automatikoaren alorreko ebaluazio-metrikarik ezagunena. Sistemaren erantzunak esaldika banatu, eta ereduzko itzulpenekin –corpus paralelo batekoekin– alderatzen ditu. Sistemak sortutako esaldiak zatikatzen joaten da –hitz bakarra lehenik, gero bi hitzeko multzoak, hiru hitzekoak ondoren, eta laukoak azkenik–, eta erreferentziazko corpuseko itzulpenekin alderatzen ditu, aztertzeko zenbateraino diren antzekoak sistemaren itzulpenak eta eskuzkoak.
- **NIST**ek ere oso era antzekoan egiten du lan, BLEUn oinarrituta baitago, baina ezaugarri gehigarri bat hartzen du kontuan: ebaluatzen

dituen hitz multzoen maiztasuna. Hitz multzo jakin batek zenbat eta maiztasun txikiagoa izan ereduzko corpusean, sistemaren itzulpena zuzena bada, orduan eta pisu handiagoa ematen dio, eta alderantziz –zenbat eta maiztasun handiagoa izan, orduan eta pisu txikiagoa–.

- **TER**, berriz, desberdin samarra da. Metrika horrek ere ereduzko itzulpenekin alderatzen ditu sistemaren emaitzak, baina beste era batera egiten ditu kalkuluak: itzulpen automatikotik eskuzkora iristeko egin beharreko moldaketa-kopuruaren arabera. Hortaz, BLEUn eta NISTen ez bezala, zenbat eta txikiagoa izan TERen balioa, esan nahi du orduan eta hobea dela itzulpen automatikoaren kalitatea.

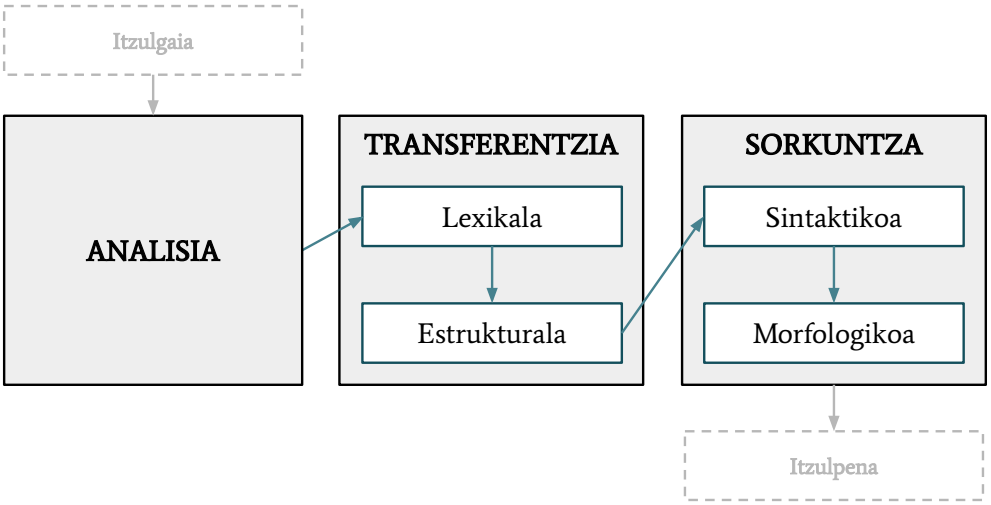
Bistan denez, ebaluazio-irizpide horiek orokorregiak dira gure alorrerako, eta ezin dira, bere horretan, UFen itzulpenaren kalitatearen adierazgarritzat hartu (Constant *et al.*, 2017). Dena den, oraindik ez da sortu UFak berariaz ebaluatzeko erabat aproposa den metodologiarik edo gidalerro bateraturik (Monti *et al.*, 2012; Ramisch *et al.*, 2013; Barreiro *et al.*, 2014). Gainera, UFak etiketatuta dituzten corpus paraleloak ere garrantzitsuak lirateke, baina halako gutxi samar sortu dira gaur-gaurkoz, eta UF mota eta hizkuntza gutxi batzuetarako bakarrik (Monti *et al.*, 2013; Schottmüller eta Nivre, 2014).

Txosten honen 5. kapituluan, gure emaitzen berri ematen dugunean, emango dugu ebaluazio-lanaren inguruko xehetasun gehiago. Oraingoz, baina, hortxe utziko ditugu itzulpen-kalitateari dagozkionak, eta *Matxin* itzultzailearen metodologia azalduko dugu sakonxeago.

2.2.4.1 *Matxin* itzultzaile automatikoa

Erregeletan oinarritutako itzultzaile automatiko orok bezala, *Matxin*ek hiru fase nagusitan egiten du lan testuak gaztelaniatik euskaratzeko: analisia, transferentzia eta sorkuntza. Horrez gain, transferentzia eta sorkuntza ere bina azpifasetan banatzen dira: transferentzia lexikoa eta estrukturala, eta sorkuntza sintaktikoa eta morfologikoa. Sistemaren arkitektura 2.7. irudian jaso dugu.

Azal dezagun orain, pausoz pauso, zer egiten den fase bakoitzean, eta erakuts dezagun prozesu hori guztia hobeto adibide baten bidez. Honako esaldi hau itzularazi diogu *Matxin*i: *Sinceramente, pienso que el informe da en el clavo.*



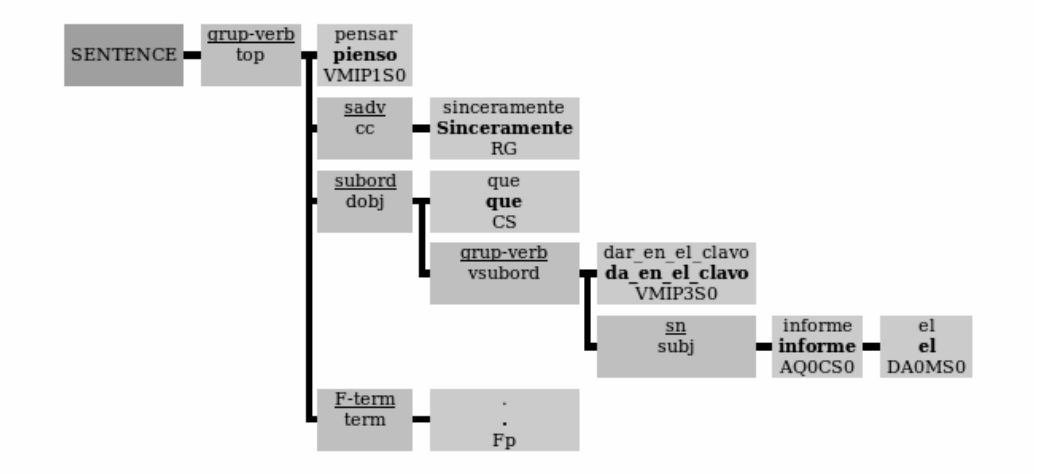
2.7 irudia – *Matxin* itzultzailearen arkitektura orokorra

1. **Analisia**. Sorburu-hizkuntzako (gaztelaniazko) esaldia aztertzen da, *Freeling* analizatzaile morfosintaktikoa (Padró eta Stanilovsky, 2012) erabiliz. Hala lortzen dira, besteak beste, hitz-forma bakoitzari dagokion lema, informazio morfologikoa (adib.: numeroa, aditz-denbora, pertsona...) eta *chunk*ak, sintagma gisako hitz multzokatzeak[20]. Horrez gain, chunkek elkarren artean zer dependentzia-erlazio duten (adib.: subjektua, objektua, modifikatzailea...) eta chunken barruko hitzak nola erlazionatuta dauden ere zehazten da. Prozesu honen amaieran, itzulgaiko esaldia abstraktuki errepresentatuta gelditzen da, 2.8. irudian ikusten den bezala.

2. **Transferentzia**. Transferentzia-fasea bi azpifasetan banatzen da: lexikoa transferitzen da batetik, eta egitura bestetik (2.9. irudia).

    - **Transferentzia lexikoa**. Lexikoa transferitzean, analisi-faseko nodo bakoitzari bere ordain lexikoa ematen zaio euskaraz, halakorik baldin badauka beti ere. Izan ere, preposizioen, artikulu mugatuen, aditz laguntzaileen eta halakoen kasuan, ez dago euskarazko

[20] *Chunk*ei *zati* izena ere eman izan zaie euskaraz, baina guk, hemen, *chunk* erabiliko dugu, *zati* terminoa oso zabala delako eta iruditzen zaigulako zenbait kasutan nahasteren bat sor dezakeela.



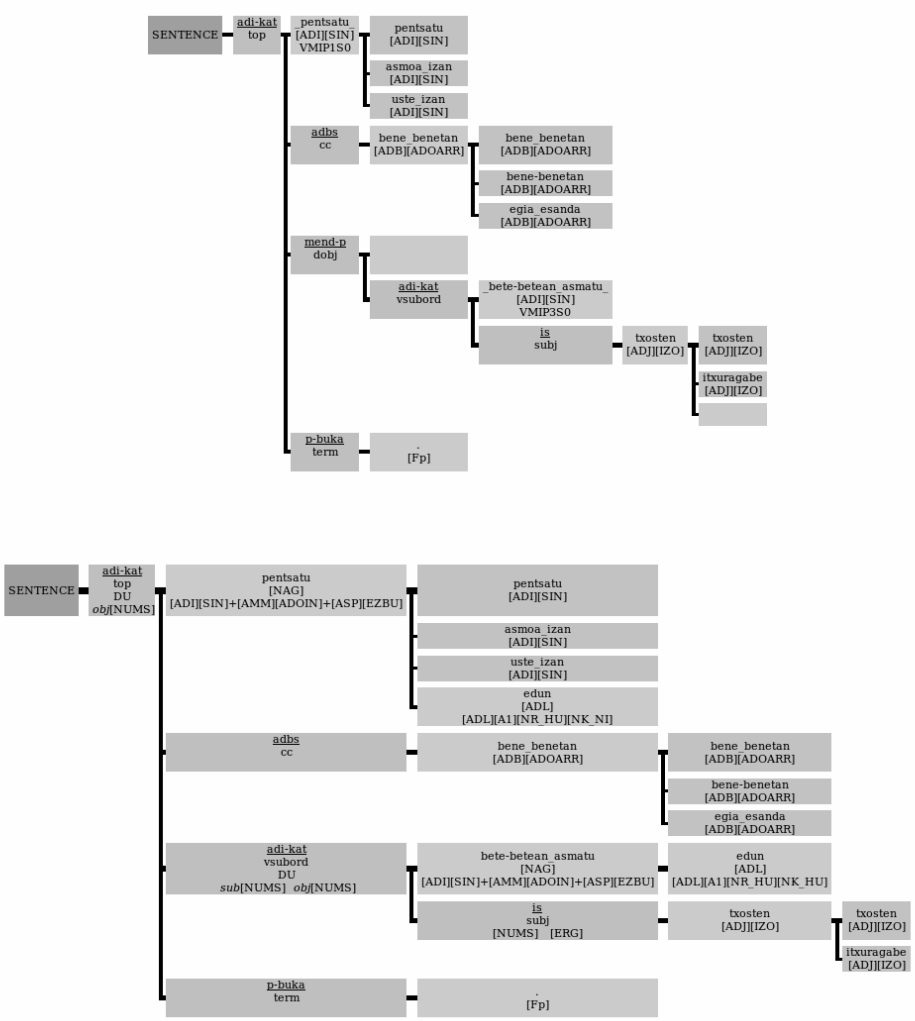
2.8 irudia – *Matxin*en itzulpen-prozesuaren adibide bat: analisi-fasea

ordain lexikorik, eta era horretako informazioa etiketa morfologiko gisa gordetzen da, geroago erabiltzeko. Ordaintza lexikoak bilatzeko, *Matxin*en lexikoi elebidunera jotzen da, eta, ordain bat baino gehiago dituzten lemen kasuan, desanbiguazio-estrategia sinple bat erabiltzen da: lehen adierako lehen ordaina aukeratzen da normalean, ordainik erabiliena hura delakoan. Beste zenbaitetan, berriz, ordain jakin bat adiera batean baino gehiagotan errepikatzen bada, ordain errepikatu hori aukeratzen da.

- **Transferentzia estrukturala**. Egitura mailako transferentzia egiten denean, nodoen barruko informazio morfologikoa euskarazko nodoetara pasatzen da, eta chunken arteko erlazio-markak ere bai. Hori, baina, dirudiena baino prozesu konplexuagoa da, gaztelania eta euskara tipologikoki hain desberdinak direnez, erregela gehigarriak behar baitira. Esate baterako, preposizioak postposizio-informazio gisa errepresentatzeko, baliokidetzak jasotzen dituen erregela sorta espezifiko bat dauka sistemak.

3. **Sorkuntza**. Azken fasea ere beste bi azpifasetan banatzen da: sintaxi mailako sorkuntza egiten da lehenik, eta morfologia mailakoa ondoren. Prozesu hori eta azken emaitza 2.10. irudian jaso dugu.

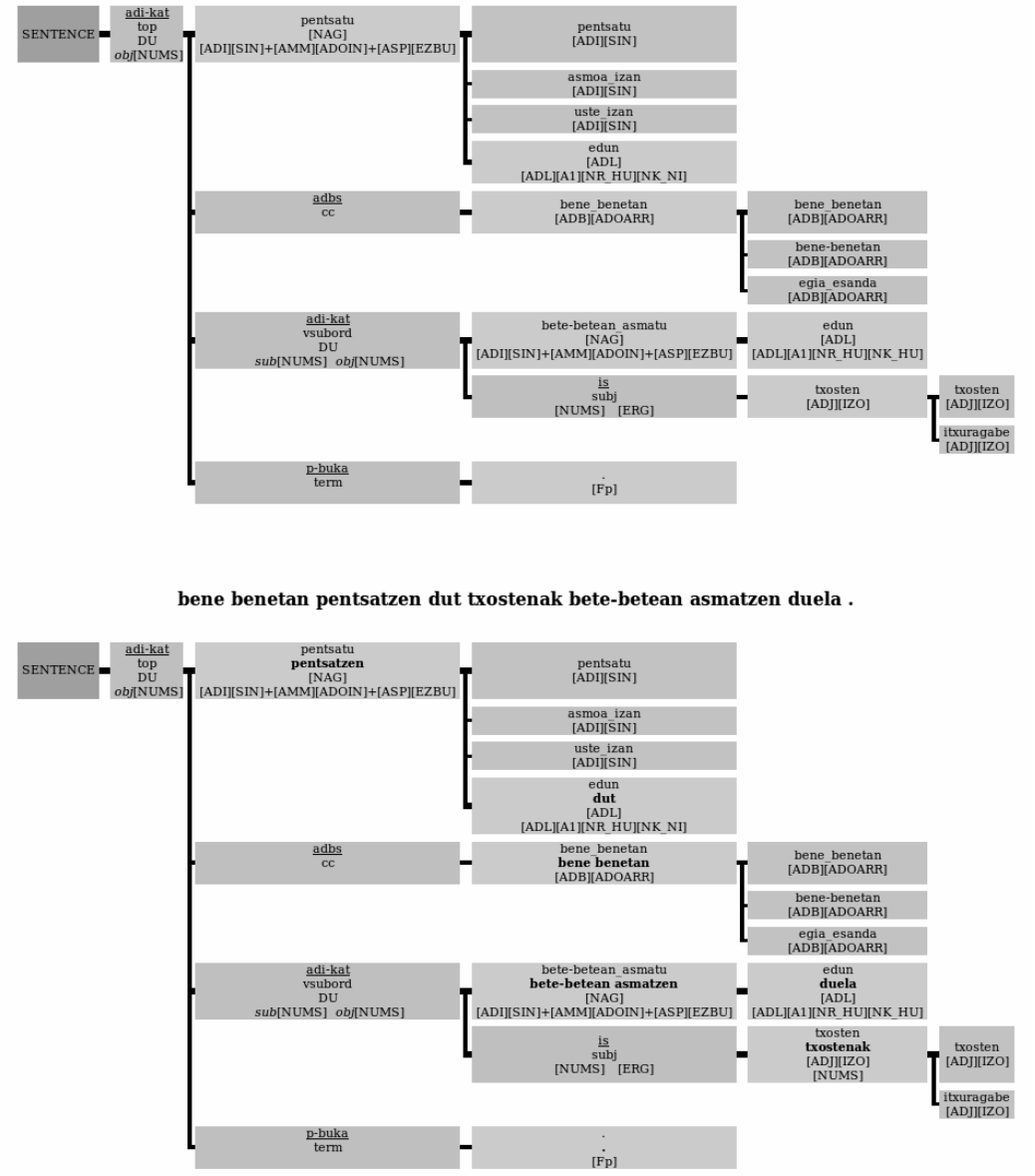
   - **Sorkuntza sintaktikoa**. Zuhaitzeko chunken eta nodoen antolamendua oinarritzat harturik, chunkek euskaraz behar duten orde-



2.9 irudia – *Matxin*en itzulpen-prozesuaren adibide bat: transferentzia-fasea

na eta chunken barruko nodoena zehazten dira. Beste egokitzapen batzuk ere egiten dira, esate baterako, postposizio-informazioa chunken amaierako nodora eramatea.

- **Sorkuntza morfologikoa**. Azken pauso honetan, lemak eta haiekin batera zehaztutako informazio morfologikoa kontuan hartu, eta hitz-formak sortzen dira.



2.10 irudia – *Matxin*en itzulpen-prozesuaren adibide bat: sorkuntza-fasea

Prozesu horretan, **UFen tratamendua** estrategia sinple baten bidez egiten da:

- Identifikaziorako, analisi-fasean, lexikoiko UFak hitz-segida gisa bila-

tzen dira itzulgaian, eta, baten bat topatzen bada, hitzak azpimarra baten bidez lotzen dira, batera tratatu behar direla adierazteko (ikus *da_ en_ el_ clavo* 2.8. irudian). Hala, UFak hitz bakarra balira bezala tratatzen dira hortik aurrera, eta gramatika-kategoria ere bakarra ematen zaie, lexikoian adierazitakoa. Aditza burutzat duten UFen kasuan, hitz-segidak bilatzean, aditzaren flexioa ere kontuan hartzen da; gainerako hitz guztiak, berriz, hiztegi-sarreran duten forma berberean.

- Transferentzian, berriz, lexikoian jasotako ordaina ematen zaio UFari, eta hori ere hitz bakar gisa tratatzen da (ikus *bete-betean_ asmatu* 2.9. irudian). Aditza burutzat duten UF-ordainen kasuan, sorkuntza egitean, behar besteko aldaketak egiten zaizkio aditzari (adib.: behar duen laguntzailea gehitzea), baina gainerako hitzak bata bestearen jarraian eta forma berean ematen dira.

Hori horrela izanik, adibide gisa erabili dugun esaldiko *dar en el clavo* UFa ondo identifikatzen da itzulgaian (2.8. irudia), eta ordaina ere ondo ematen zaio euskaraz (2.9. irudia). Estrategia horrek, sinplea izanik ere, oso emaitza onak lortzen ditu lexikoian dauden UFak era jarraituan eta lexikoiko aldaki berean agertzen direnean itzulgaian. Lexikoitik kanpoko UFak edo lexikoiko UFen beste aldaki morfosintaktiko batzuk agertzen direnean, berriz, hanka sartzen du. Datorren kapituluan hitz egingo dugu sakonago estrategia horren mugez, eta emango ditugu hanka-sartze horien adibide erakusgarri batzuk.

Definizio laburren bilduma

Hemendik aurrerako kapituluen amaieran, eginiko lana lotuko dugu tesiaren hasieran aipatu ditugun abiapuntu-hipotesiekin eta helburuekin. Oraingo honetan, ordea, gure lana testuinguruan kokatzen eta oinarrizko kontzeptuak argitzen jardun dugunez, definizio laburren bilduma bat egin dugu, aurrerantzean sarri samar agertuko diren terminoak hautatuta.

- **Idiosinkrasia**. Hizkuntza jakin baten ohiko arauetatik kanpo geratzen diren hizkuntza-ezaugarrien eta -fenomenoen nolakotasuna. Hizkuntzaren alderdi guztiak hartzen dira kontuan zerbait *idiosinkrasikoa* dela esateko: fonologia, ortografia, morfologia, sintaxia, lexikoa, semantika eta pragmatika.

- **Idiomatikotasuna**. Hitz-konbinazio jakin bat Unitate Fraseologikoa izatea eragiten duen ezaugarria. Idiosinkrasia du muinean, baina termino zehatzagoa da, hitzak konbinatzeko moduari dagokion idiosinkrasia bakarrik hartzen baitu kontuan.

- **Unitate Fraseologikoak (UF)**. Bi hitzen edo gehiagoren konbinazio ohikoak, morfosintaxiari edo/eta lexiko-semantikari dagokionez idiomatikoak. Osagai-hitzak sintaktikoki erlazionaturik egoten dira, dependentzia-erlazioan normalean.

- **Aditz-esapide idiomatikoak**. Burutzat aditza duten hitz-konbinazio idiomatikoak. Kasu honetan, *idiomatiko* adjektiboak semantikari egiten dio erreferentzia (bestela, oro har, *aditz-UF* erabiliko dugu), esan nahi baita konbinazioko osagai-hitzen esanahiak batuta ez dela esapide osoaren esanahia lortzen. Aditz-esapide idiomatikoetan sar daitezke, aditz-lokuzioak ez ezik, burutzat aditza duten atsotitzak eta halakoak ere. Euskarazko adibideak: *ikusi eta ikasi, katuak mingaina jan, ziria sartu.*

- **Lokuzioak**. Semantikoki (behintzat) idiomatikoak diren hitz-konbinazioak, enuntziatu osoak ez direnak. Lokuzioen esanahi osoa ez da osagai-hitzen esanahien batura, baina batzuetan uler daitezke metafora bidez; horren arabera bereizten dira lokuzio **opakoak** eta **metaforikoak**. Adibideak: *adarra jo, hanka egin* (opakoak), *zubiak eraiki* (metaforikoa).

- **Kolokazioak**. Hitz-konbinazio usuak, elkarrekin ausaz espero litekeena baino maizago agertzeko joera dutenak. Osagaietako bat buru semantikoa izaten da, eta horrek aukeratu ohi du beste osagaia, *kolokatua*. Askotan, kolokatu hori ezin izaten da sinonimoen bidez ordezkatu, edo ordezkatuz gero konbinazio arrotzak osatzen dira. Adibideak: *legea urratu, plazak bete, zarata atera*.

- **Aditz arindun konbinazioak**. Aditz arin batez eta beste osagai batez (normalean, izen batez) osaturiko kolokazioak. Aditzak ez beste osagaiak ekintza edo egoera bat adierazi ohi du, eta aditzak esanahi horri aditz-izaera ematen laguntzen du nolabait, normalean ezaugarri morfologikoak bakarrik gehituz. Horregatik deitzen zaie aditz *arinak*, ez dutelako ia esanahirik hitz-konbinazioaren barruan. Adibideak: *lo egin, min hartu, musu eman*.

- **Konbinazio libreak**. Hizkuntzaren ohiko bideei jarraituz osaturiko konbinazio ez-idiomatikoak, fraseologiaren alorretik kanpokoak. Adibideak: *liburua irakurri, oparia erosi, zapatak jantzi*.

- **Konbinazio (erdi)finkoak**. Morfosintaxiari dagokionez murriztapenak dituzten hitz-konbinazioak. Erabat zurrunak direnak *finko*tzat hartzen ditugu, eta murriztapen batzuk izan arren aldaketa morfosintaktikoren bat ere onartzen dutenak, berriz, *erdifinko*tzat. Adibideak: *hala eta guztiz ere*, *eskerrik asko* (finkoak), *lo egin, txantxetan aritu* (erdifinkoak).

- **Konbinazio malguak**. Morfosintaxiari dagokionez ohiko hizkuntza-arauak jarraitzen dituzten hitz-konbinazioak. Ez dute murriztapen morfosintaktikorik, baina baliteke lexiko-semantikari dagokionez idiomatikoak izatea eta, hortaz, UFtzat hartzea. Adibideak: *dirua egin, erabakia hartu, zailtasuna izan*.

- **UFen erauzketa**. Corpusetatik UFak lortzeko lana. Emaitzatzat UF-zerrenda bat lortzen da.

- **UFen identifikazioa**. Testuetan UFen agerpenak identifikatzera bideratutako ataza. Aldez aurretik jakin behar da zer UF bilatu nahi diren, eta, UFa osatzen duten hitzak testuan agertzen direnean, UFaren agerpen bat den ala ez desberdintzean datza.

3. KAPITULUA

Prestaketa-lana: *Matxin* itzultzaile automatikoaren eta *Elhuyar Hiztegia*ren azterketa

Doktoretza-tesi honen hipotesiak zerrendatzean (1.3. atala), esan dugu UFak ez direla beti hitzez hitz itzultzen, eta geroago (2.2.4. atala) aipatu dugu horrek, sarritan, zailtasunak ematen dizkiela itzultzaile automatikoei. Abiapuntu hori nolabait indartze aldera, eta HPrako aplikagarriak diren proposamenak egiten hasi aurretik, itzultzaile automatiko bat eta hiztegi elebidun bat aztertu nahi izan ditugu. Hain zuzen ere, ondorengo lanetan ere oinarritzat hartuko ditugun tresna bati eta baliabide bati buruz jardungo dugu kapitulu honetan: *Matxin* itzultzaile automatikoari eta *Elhuyar* hiztegiari buruz.

Hasteko, 3.1. atalean, *Matxin* itzultzaile automatikoak UFak itzultzerakoan zer-nolako zailtasunak izaten dituen erakutsiko dugu. Gero, 3.2. atalean, *Elhuyar* gaztelania-euskara eta euskara-gaztelania hiztegian jasotako aditza+izena konbinazioei buruz jardungo dugu. Eta azkenik, hemendik aurrerako kapitulu guztietan bezala, kapituluan zehar azaldutakoak laburtu, eta abiapuntu-hipotesiekin eta tesiaren helburuekin lotuko ditugu.

3.1 *Matxin*en erroreen analisia

Lehenago azaldu dugunez (2.2.4. atala), *Matxin* itzultzaileak erregela linguistikoak ditu oinarrian, eta itzulpen-prozesua hiru fasetan banatzen du: analisia, transferentzia eta sorkuntza (Mayor *et al.*, 2009). Sistema gai da UF batzuk zuzen itzultzeko, baina beste askotan oso itzulpen traketsak sortzen ditu.

Oro har, esan daiteke *Matxin*en akatsek lau arazo-iturri nagusi dituztela, UFen itzulpenari dagokionez:

- Lexikoi elebidunaren mugak
- UFak identifikatzeko metodoaren gabeziak
- UFak itzultzeko metodoaren gabeziak
- Testuinguruaren tratamendurik eza

Tesi-txosten honetan ez dugu testuinguruarekin zerikusia duen akatsik aztertuko. Izan ere, badirudi UFak gutxitan erabiltzen direla literalki corpusetan (7. kapitulua), eta, beraz, pentsatzekoa da morfosintaxian sakontzeak ekarpen handiagoa egingo duela. Hala, alde batera utziko ditugu, momentuz, semantikari begiratuta baino ebatzi ezin diren kasuak, 46. eta 47. adibideetakoak bezalakoak.

(46) ES: *Qusieron curarlo, pero estaba muy enfermo y* ***estiró la pata****.*
Matxin: *Sendatu nahi izan zuten, baina oso gaixoa zegoen eta* ***azken hatsa eman*** *zuen.*

(47) ES: *Estiró la pata y el brazo.*
Matxin: *Eta besoa* ***azken hatsa eman*** *zuen.*
EU-zuz: *Hanka eta besoa luzatu zituen.*

Bi esaldi horietan, beltzez markatutako hitz-segidari itzulpen berdinberdina eman zaio, batean eta bestean oso esanahi desberdina badu ere: 46. adibidean idiomatikoa, eta 47.ean, literala. Bigarren esaldiari eman behar litzaiokeen itzulpena ez dator bat *Matxin*enarekin, baina sistemak, gaur egun, ez du testuingurua kontuan hartzen, eta ez du halakoak desberdintzeko modurik.

Adibideen laguntzaz, zerrendako beste hiru arazo-iturriak zertan dautzan azalduko dugu orain, argiago gera dadin nondik abiatu garen datozen kapituluetan azalduko dugun lanean.

Lexikoi elebidunaren mugak

Esan dugunez, *Matxin*ek lexikoi elebidun bat du oinarrian, hainbat hiztegitatik sortua. Lexikoi horrek baditu hitz batez baino gehiagoz osatutako sarrera batzuk, baina, batetik, ez dira asko, eta bestetik, tipologia bakarrekoak –lokuzioak– dira oro har, hiztegi orokorrek halakoak bakarrik jaso ohi baitituzte. Horrek esan nahi du, salbuespenak salbuespen, *Matxin*ek ez duela kolokaziorik bere lexikoian, eta dauzkan lokuzioen kopurua ere mugatua dela.

(48) ES: *La pareja* ***contrajo matrimonio****.*
Matxin: *Bikotea ezkontza uzkurtu zen.*
EU-zuz: *Bikotea ezkondu (egin) zen.*

Muga horiek sistemari akatsak sorrarazten dizkiote askotan. Esate baterako, *Matxin*en lexikoian ez dagoenez *contraer matrimonio* sarrerarik, *contraer* eta *matrimonio* hitzei zeini bere aldetik ematen zaie ordaina 48. adibideko esaldian. Bi hitz horiek, ordea, UF bat osatzen dute, eta okerreko itzulpena sortzen da haiek hitzez hitz euskaratuta: *ezkontza uzkurtu*, *ezkondu* beharrean.

Halako mugek zer-nolako garrantzia duten irudikatzeko, har dezagun PARSEME proiektuak sortutako gaztelaniazko corpusa. Corpusak egunkarietatik eta testu orokorretatik bildutako 5.515 esaldi ditu, eta aditza+izena motako 662 UF markatuta. *Matxin*en lexikoiko UFak corpuseko etiketa horiekin alderatuz gero, lexikoiaren mugak agerian geratzen dira, 53 agerraldi bakarrik baitatoz bat lexikoiko UFekin. Horrek esan nahi du *Matxin* ez dela gai corpuseko UFen % 92 ezagutzeko, lexikoian agertzen ez diren UFen 609 agerraldi baitira.

UFak identifikatzeko metodoaren gabeziak

Bestalde, UF jakin bat lexikoian badago, *Matxin*ek hitz-segida hori itzulgaian bilatzen du, eta, aurkituz gero, lexikoian daukan ordaina ematen dio. Horregatik itzultzen da zuzen 49. adibideko esaldia, lexikoiko *gastar una broma* sarrera eta itzulgaian agertzen den hitz-segida bat datozelako.

(49) ES: *Ayer le* ***gastaron una broma****.*
Matxin: *Atzo* ***broma egin*** *zioten.*

Baina UF bat lexikoian egoteak ez du esan nahi sistemak UF hori beti ondo itzuliko duenik. Aurreko esaldia ondo itzultzen da, *gastar una broma* hitz-konbinazioa era jarraituan eta –aditzaren flexioa gorabehera– forma berean agertzen delako. Aldiz, beste hitzetako bati aldaketaren bat egiten badiogu edo hiru hitzen artean beste elementuren bat sartzen badugu, emaitza bestelakoa da.

(50) ES: *La* ***broma*** *que le* ***gastaron*** *fue cruel.*
Matxin: *Gastatu zioten broma krudela izan zen.*
EU-zuz: ***Egin*** *zioten* ***broma*** *krudela izan zen.*

Ikusten denez, 50. adibidean *broma* eta *gastar* ez daude bata bestearen ondoan, izen-sintagmak ez darama *una* determinatzailerik, eta hitz-hurrenkera ere desberdina da. *Matxin*ek, UFen aldaki morfosintaktikorik kontuan hartzen ez duenez, ez du esaldi horretan gaztelaniazko UFrik topatzen, eta okerreko ordaina ematen dio euskarazko aditzari: *gastatu*, *egin*en ordez.

Gainera, aintzat hartu behar da *Matxin*ek ez diela UFen barruko hitzei kategoriarik esleitzen, eta hortik ere interpretazio okerrak etortzen dira sarri. Har dezagun, adibidez, aditza+izena motako *hacer mal* UFa.

(51) ES: *El colesterol* ***hace mal*** *al corazón.*
Matxin: *Kolesterolak bihotzari* ***kalte egiten*** *dio.*

(52) ES: *Ha hecho mal el examen.*
Matxin: *Azterketa* ***kalte egin*** *du.*
EU-zuz: *Azterketa gaizki egin du.*

Sistemak ez daki *mal* zer kategoriatakoa den, eta *hacer* aditzaren atzetik datorren guztietan UF bat identifikatzen du, bai benetan UFa dagoenean (51. adibidean) eta bai ez dagoenean ere (52. adibidean). Lehen esaldian ez dago arazorik, baina bai bigarrenean, beharrezkoa baitzen sistemak *hacer mal* UFtzat ez hartzea, hitz bakoitza bere aldetik itzultzeko eta, hala, *kalte*ren ordez *gaizki* jarri ahal izateko.

UFak itzultzeko metodoaren gabeziak

Identifikazio-lanean ez ezik, itzulpena sortzerakoan gertatzen diren nahasteek ere lotura zuzena dute morfosintaxiarekin. Esan dugu *Matxin*ek ez diola gaztelaniazko UFko hitzen kategoriari begiratzen, eta aditzaren flexioa dela

kontuan hartzen duen aldaketa morfosintaktiko bakarra. Transferentzia- eta sorkuntza-faseetan ere antzeko zerbait gertatzen da:

(53) ES: *Miren* ***tiene frío***.
Matxin: *Miren* ***hotz da***.

(54) ES: *Miren tiene frío el plato.*
Matxin: *Mirenek platera* ***hotz da***.
EU-zuz: *Mirenek platera hotz du*.

Bi adibide horiek erakusten dutenez, ordainen ezaugarri morfologikoak ez dira kontuan hartzen, eta, UF bat tartean dagoenean, sistema ez da gai subjektuaren eta aditzaren arteko komunztadurarik egiteko. Hori dela-eta, 53. adibideko itzulpena egokia bada ere, 54. adibidekoa okerra da: aditzak *ukan* iragankorra izan beharko luke, *izan* iragangaitza beharrean.

Hortaz, azken azpiatal honetan azaldutakoak argiago erakusten du *Matxin*ek hutsuneak dituela UFei dagokienez. Datozen bi kapituluetan azalduko dugu zein den gure proposamena informazio hori tratatzen laguntzeko, baina, horren aurretik, ikus dezagun zer-nolakoak diren *Elhuyar* hiztegian jasotako konbinazioak eta ordainak, hori ere lagungarria izango baita ikusteko zenbateraino den konplexua UFen itzulpena.

3.2 *Elhuyar* gaztelania-euskara hiztegiaren gaineko azterketa

Prestaketa-lanaren bigarren urratsa hiztegien gainean egin dugu, zer hitz-konbinazio eta zer ordain jaso ohi dituzten aztertzeko. Badakigu hiztegietan biltzen diren UFak lokuzioak izaten direla oro har eta beste mota batzuetako hitz-konbinazioak ez direla hainbeste landu izan hiztegigintzan, kolokazioak bereziki (Gurrutxaga, 2014; Urizar, 2012). Aurrerago joko dugu beste baliabide batzuetara konbinazio gehiagoren bila, baina, hasteko, hiztegi orokor hauetan jasotakoaren argazki orokor bat egingo dugu, ondorengo pausoetarako prestaketa gisa.

Oinarritzat hartu ditugun konbinazio-zerrendak *Elhuyar*ren gaztelania-euskara eta euskara-gaztelania hiztegitik erauzi ditugu. Horretarako, Eustagger (Alegria *et al.*, 1996) eta Freeling (Padró eta Stanilovsky, 2012) tresnen bidez analizatu ditugu euskarazko eta gaztelaniazko hiztegi-sarrera eta

ordain guztiak. Analizatzaile horiek hainbat etiketa esleitu dizkiete sarreren barruko hitzei (gramatika-kategoria, kasu- eta postposizio-markak, numeroa, etab.), eta etiketa horiez baliatu gara aditza+izena motako konbinazioak hautatzeko eta zerrenda bat osatzeko.

Prozesu automatiko gehienetan bezala, emaitza oker batzuk ere lortu ditugu benetan gure aztergai diren konbinazioekin batera, eta zerrenda gainbegiratu behar izan dugu akatsak baztertzeko. Esate baterako, *argi eduki* konbinazioa izena+aditza motakotzat identifikatu du Eustaggerrek, testuingururik gabe zaila baita jakitea *argi* adjektiboa ala izena den. Gaztelaniazko *tener claro*ren ordaina izanik, ordea, adjektibo kategoria zen aukera zuzena eta, hortaz, eskuz baztertu behar izan dugu konbinazio hori.

Hiztegi-sarrera diren izena+aditza konbinazioez gain, beste egitura batzuetako sarreren ordaintzat ageri diren izena+aditza konbinazioak ere erauzi ditugu, pentsatuz, hiztegi-sarreren ordaintzat jasota badaude, balitekeela konbinazio horiek ere hein batean idiomatikoak izatea. Adibidez, gaztelaniazko *poner barreras* hiztegi-sarrera eta haren *oztopatu* ordainaz gainera, euskarazko *abadetu* sarreraren ordaintzat ageri den *ordenarse sacerdote* ere erauzi dugu, ordaina izena+aditza motakoa delako (3.1. irudia).

3.1 irudia – *Elhuyar* hiztegitik konbinazio-zerrendak sortzeko prozesuaren adibide bat

Gaztelaniazko 2.343 konbinazio landu ditugu guztira: gaztelania-euskara hiztegitik erauzitako 1.205 hiztegi-sarrera, eta euskara-gaztelania hiztegitik erauzitako 1.138 ordain. Hiztegi-sarrera gehienek ordain bat baino gehiago dutenez, ez da harritzekoa euskarazko baliokideak askoz ere gehiago izatea: 6.587.

Euskarazko konbinazioei dagokienez, berriz, 2.954ko zerrenda osatu dugu, euskara-gaztelania hiztegiko 1.576 sarrera eta gaztelania-euskara hiztegiko 1.378 ordain batuta. Horien gaztelaniazko baliokideak 6.390 dira guztira.

Datozen azpiataletan bildu ditugu horien guztien azterketatik ateratako emaitzak. Lehenik, 3.2.1. atalean, erauzitako konbinazioez arituko gara: zer ezaugarri morfologiko dituzten eta nolako aditzez eta izenez osatuta dauden. Ondoren, 3.2.2. atalean, ordainak nolakoak diren aztertuko dugu, bai eta haien ezaugarriak beste hizkuntzako hitz-konbinazioenekin bat ote datozen ere.

3.2.1 Erauzitako hitz-konbinazioen ezaugarriak

Esana dugu gure aztergai nagusia aditza+izena motako UFak direla eta, prestaketa-lan honetarako, era horretako 5.297 konbinazio erauzi ditugula *Elhuyar* hiztegitik: gaztelaniazko 2.343 eta euskarazko 2.954.

Gogoan izan behar da, dena den, aditza+izena diogunean, multzo horretan hitz-konbinazio gehiago sartzen direla aditz batez eta izen batez bakarrik osatutakoak baino. Izan ere, arrazoi tipologikoak direla-eta, euskarazko izena+aditza konbinazioen barruan hainbat morfema ager daitezke izenari itsatsita: artikuluak, kasu-markak eta postposizioak. Gaztelaniaz pareko egiturako konbinazioak bildu nahi genituenez, aditzaren eta izenaren artean preposizioren bat edota determinatzaileren bat duten konbinazioak ere onartu ditugu.

Datozen azpiataletan emango ditugu ezaugarri morfologikoei buruzko datu xeheagoak, bai eta konbinazioetako izenei eta aditzei buruzko hainbat jakingarri ere.

Ezaugarri morfologikoak

Atal honen hasieran esan dugunez, **euskara**zko konbinazioak aztertzeko, Eustagger analizatzailearen etiketak hartu ditugu oinarritzat. Etiketa automatikoak eskuz zuzendu ondoren, izenei lotutako kasu- eta postposiziomarkei begiratu diegu lehenik. Kontaketak egitean, berehala konturatu gara

alde nabarmena zegoela etiketen araberako multzoen artean.

Urizarren (2012) eta Zabalaren (2004) lanetan aipatzen denez, izen batez eta aditz batez osatzen diren konbinazio gehien-gehienetan, izenek absolutibo-marka izan ohi dute[1]. Horixe bera ondorioztatu dugu guk ere, konbinazio guztien hiru laurden baino gehiago baitira multzo horretakoak gure aztergaien artean: *denbora galdu, dei egin, gobada jo, itzal egin, hitza bete...*

Inesiboa daramaten izenak ere nahiko sarri ageri dira besteen aldean (*sutan egon, jokoan jarri*), eta adlatibodunek (*aurrera egin, eskura ordaindu*), instrumentaldunek (*eskuz jo, aurrez prestatu*) eta ablatibodunek (*burutik egon, hutsetik hasi*) jarraitzen diete kopurutan. Gehien errepikatu diren bost markak 3.1. taulan jaso ditugu agerraldien eta ehunekoaren arabera, eta 3.2. irudian ikus daiteke marka bakoitza zein proportziotan erabili den.

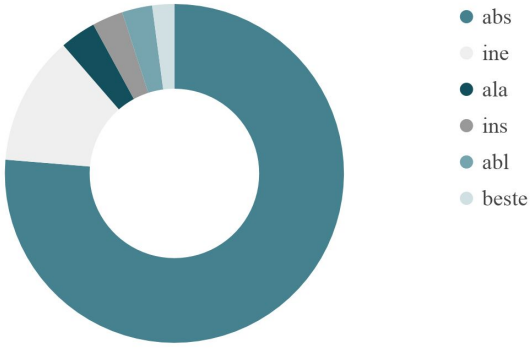
EU marka	**Agerr.**	**Ehunekoa**
absolutiboa (abs)	2248	% 76,10
inesiboa (ine)	366	% 12,39
adlatiboa (ala)	101	% 3,42
instrumentala (ins)	87	% 2,94
ablatiboa (abl)	85	% 2,88
beste batzuk	67	% 2,27

3.1 taula – Euskarazko konbinazioen kasu- eta postposizio-markak (*Elhuyar* hiztegia)

Gainerako markak nahiko bakanak dira, eta, banaka hartuta, ez dira konbinazio guztien % 1era ere heltzen: ergatiboa (*deabruak hartu*), datiboa (*amuari lotu*), adlatiboa+abutiboa (*leporaino egon*), soziatiboa (*suarekin jolastu*), genitiboa+absolutiboa (*gorrarena egin*), lekuzko genitiboa+absolutiboa (*hitzekoa izan*), adlatiboa+lekuzko genitiboa (*bururako gorde*) eta lekuzko genitiboa (*zorioneko egon*).

Gaztelaniari dagokionez, berriz, lau egituratako konbinazioak bildu ditugu. Batzuk izen eta aditzez bakarrik osatuak dira, eta beste batzuek determinatzaile edota preposizioren bat dute tartean. Gehien errepikatzen den egitura aditza+determinatzailea+izena da (*dar un toque, ser una pena, hacer*

[1]Lan honetan, aintzat hartu dugu batere markarik gabe agertzen diren izen-sintagmak (*lan egin*eko *lan*, *min eman*eko *min*, eta halakoak) absolutiboan daudela, gure helburu aplikaturako horixe baita modurik errazen eta eraginkorrena. Dena dela, Oyharçabalek (2003) ohartarazten duenez, ez dago argi halakoak benetan absolutiboan dauden ala ez, ez baitute berez inongo marka espliziturik.



3.2 irudia – Euskarazko konbinazioen kasu- eta postposizio-markak (*Elhuyar* hiztegia)

un favor), eta beste hiru multzoen artean ez dago alde handirik: aditza+izena (*meter baza, tener afecto*), aditza+preposizioa+izena (*saber de memoria, tener a favor*), aditza+preposizioa+determinatzailea+izena (*dejar a un lado, caerse por su peso*). Beraz, egitura morfologikoen proportzioak orekatuagoak dira gaztelaniaz euskaraz baino, 3.2. taulako datuek eta 3.3. irudiak agerian uzten dutenez.

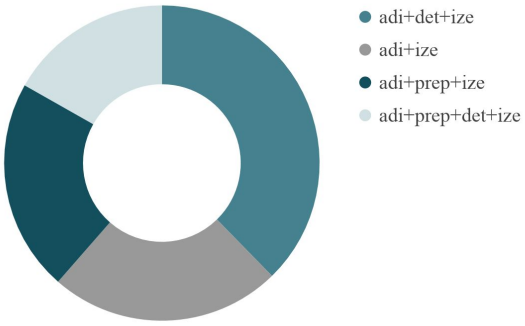
ES egitura	**Agerr.**	**Ehunekoa**
adi+det+ize	877	% 37,43
adi+ize	579	% 24,71
adi+prep+ize	499	% 21,30
adi+prep+det+ize	388	% 16,56

3.2 taula – Gaztelaniazko konbinazioen egitura morfologikoak (*Elhuyar* hiztegia)

Ezaugarri lexikoak

Behin ezaugarri morfologikoak ezagututa, hizkuntza bateko eta besteko konbinazioak zein aditzez eta izenez osatuta dauden begiratu dugu, bi hizkuntzak parez pare jarri eta euren artean antzekotasunik ba ote dagoen ikusteko.

Fraseologiari buruzko lanetan aipatu izan denez (Oyharçabal, 2003; Butt, 2010; Buckingham, 2009), UFetan erabiltzen diren aditzak *arinak* izaten dira



3.3 irudia – Gaztelaniazko konbinazioen egitura morfologikoak (*Elhuyar* hiztegia)

maiz, UFen barruan esanahia nolabait galtzen dutenak. Oso aditz arruntak izaten dira, eta nahiko talde mugatua osatu ohi dute. Ez da harritzekoa, beraz, zer aurkitu dugun: hizkuntza bateko eta besteko hitz-konbinazioetako aditzik usuenen artean, gehienak oso arruntak dira eta, gainera, asko baliokideak dira euren artean. Hain zuzen ere, euskaraz gehien errepikatzen diren sei aditzen ordainak gaztelaniazko hitz-konbinazioetako zortzi aditzik errepikatuenen artean aurkitu ditugu: *egin – hacer, izan – ser/estar/tener, eman – dar, hartu – tomar, egon – estar, jarri – poner* (ikus 3.3. eta 3.4. taulak).

Aditza	**Agerr.**	**Ehunekoa**
egin	614	% 20,79
izan	296	% 10,02
eman	217	% 7,35
hartu	147	% 4,98
egon	133	% 4,50
jarri	92	% 3,11
Guztira	**1.499**	**% 50,74**

3.3 taula – Euskarazko konbinazioetako aditzik ohikoenak (*Elhuyar* hiztegia)

Hala ere, agerraldiei begiratuz gero, bat aise konturatzen da euskaraz askoz ere alde handiagoa dagoela gehien errepikatzen diren aditzen eta gainerakoen artean. Izan ere, euskarazko konbinazioetan 310 aditz desberdin agertu dira guztira, baina konbinazio guztien bi heren 3.3. taulan bildutako

Aditza	**Agerr.**	**Ehunekoa**
hacer	215	% 9,18
dar	178	% 7,60
estar	103	% 4,40
tener	101	% 4,31
poner	95	% 4,05
echar	64	% 2,73
ser	56	% 2,39
tomar	54	% 2,30
Guztira	**866**	**% 36,96**

3.4 taula – Gaztelaniazko konbinazioetako aditzik ohikoenak (*Elhuyar* hiztegia)

seiekin bakarrik osatzen dira.

Gaztelaniaz ere badago aldea, baina ez da inondik inora ere euskarazkoa bezain handia; zortzi aditzik errepikatuenek konbinazioen % 36,96 osatzen dute, euskarazko sei aditzek osatzen duten proportzioaren ia erdia. Dena den, datu hori ez da hain harritzekoa ere, euskaraz oso ohikoak baitira hitz batez baino gehiagoz osatutako aditz "konplexuak" (Zabala, 2004), hain zuzen ere aditz arinekin osatzen direnak, *egin, eman, hartu* eta halakoekin. Txosten honetako kapitulu gehiagotan ere, 7.ean bereziki, emango ditugu datu gehiago erakusteko halakoak bereziki ohikoak direla euskaraz.

Izenen agerraldiak ere kontatu ditugu, antzeko ondoriorik atera ote genezakeen ikusteko. Susmoa genuen izen desberdinen kopurua aditzena baino askoz ere altuagoa izango zela baina, era berean, izen batzuek beste batzuek baino joera handiagoa izango zutela konbinazioetan agertzeko. Susmoa ez zen okerra, 3.5. eta 3.6. taulek agerian uzten dutenez.

Izenetan ere badago antzekotasunik bi hizkuntzen artean, gehien errepikatzen direnetako batzuk bat baitatoz euskaraz eta gaztelaniaz: *buru – cabeza, begi – ojo, esku – mano, kontu – cuenta.* Gainera, zerrenda horietan argi ikusten da lokuzioetan zeinen sarri agertu ohi diren gorputz-atalak izendatzen dituzten izenak. Hortik aurrera, ordea, ez dago aditzena bezain datu deigarririk.

Izena	Agerr.	Ehunekoa
buru	65	% 2,20
begi	54	% 1,83
aurre	43	% 1,46
kontu	33	% 1,12
atze	24	% 0,81
bide	23	% 0,78
esku	21	% 0,71
Guztira	**263**	**% 8,91**

3.5 taula – Euskarazko konbinazioetako izenik ohikoenak (*Elhuyar* hiztegia)

Izena	Agerr.	Ehunekoa
mano	28	% 1,19
cabeza	27	% 1,15
ojo	21	% 0,90
vista	21	% 0,90
oído	18	% 0,77
cuenta	18	% 0,77
Guztira	**133**	**% 5,68**

3.6 taula – Gaztelaniazko konbinazioetako izenik ohikoenak (*Elhuyar* hiztegia)

3.2.2 Erauzitako ordainen ezaugarriak

Behin konbinazioak nolakoak diren aztertuta, haien ordainetan jarri dugu arreta, hizkuntza batetik bestera ordainak ematean gertatzen diren aldaketetan bereziki. Aurreko azpiatalean egin dugun bezala, hemen ere, ezaugarri morfologikoei begiratu diegu lehenik, eta ezaugarri lexikoei ondoren.

Gaztelaniazko ordainen ezaugarri morfologikoak

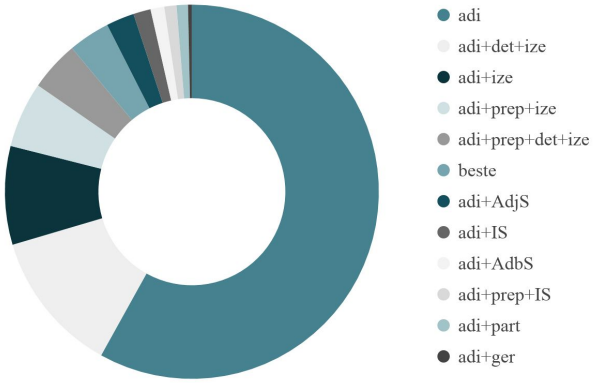
Bi hizkuntzen ezaugarri morfologikoak alderatzen hasteko, euskarazko konbinazioen gaztelaniazko ordain guztiak analizatu, eta hainbat multzotan sailkatu ditugu. Honako sailkapen hau sortu dugu:

- Aditz soila (**adi**): *descentrar, lanzar, descansar.*

- Aditza, determinatzailea eta izena (**adi**+**det**+**ize**): *matar el tiempo, hacer la colada, echar la llave.*
- Aditza eta izena (**adi**+**ize**): *ganar tiempo, tener noticia, hacer caso.*
- Aditza, preposizioa eta izena (**adi**+**prep**+**ize**): *arder en deseos, llevar por acompañante, comer con apetito.*
- Aditza, preposizioa, determinatzailea eta izena (**adi**+**prep**+**det**+**ize**): *estar en los huesos, faltar a su palabra, hacerse a la mar.*
- Aditza eta adberbioa edo adberbio-sintagma (**adi**+**AdbS**): *salir adelante, sentar bien, llegar lejos.*
- Aditza eta adjektiboa edo adjektibo-sintagma (**adi**+**AdjS**): *estar claro, volverse loco/a, resultar complicado.*
- Aditza, eta izen soila ez den izen-sintagma (**adi**+**IS**): *poner mala cara, levantar la tapa de los sesos, hacer buen tiempo.*
- Aditza, preposizioa, eta izen soila ez den izen-sintagma (**adi**+**prep**+**IS**): *dejar a medio camino, entrar en uso de razón, pagar en dinero contante.*
- Aditza eta partizipioa (**adi**+**part**): *pillar inadvertido, dejar frito, estar preocupado.*
- Aditza eta gerundioa (**adi**+**ger**): *salir corriendo, estar ardiendo, andar endredando.*
- Aurreko multzoetan sartzen ez direnak (**beste**): *estar por suceder, dar mucho que decir, entablar amistad con alguien.*

Kasu- eta postposizio-marka guztiak kontuan hartuta, gehien ageri den ordain mota aditz soila da nabarmen (3.4. irudia, 3.7. taula): ordain guztien % 58,07. Datu hori gutxi-gorabehera espero izatekoa zen, lehenago ere aipatu dugunez, euskaraz oso ohikoak baitira aditz arinekin osatzen diren UFak, beste hizkuntza askotan baino gehiago ziur asko, eta halakoen ordainak maiz izaten dira aditz soilak beste hizkuntza batzuetan.

Interesgarria da, hala ere, beste ordainen egiturei ere begiratu bat egitea. Aditz soilak alde batera utzita, lau egitura morfologiko nabarmentzen dira



3.4 irudia – Gaztelaniazko ordain motak (*Elhuyar* hiztegia)

Egitura	**Agerr.**	**Ehunekoa**
adi	3711	% 58,08
adi + det + ize	787	% 12,32
adi + ize	545	% 8,53
adi + prep + ize	364	% 5,70
adi + prep + det + ize	274	% 4,29
adi + AdjS	155	% 2,43
adi + IS	96	% 1,50
adi + AdbS	76	% 1,19
adi + prep + IS	67	% 1,05
adi + part	63	% 0,99
adi + ger	21	% 0,33
beste	231	% 3,62

3.7 taula – Gaztelaniazko ordain motak (*Elhuyar* hiztegia)

besteen gainetik: *adi+det+ize, adi+ize, adi+prep+ize* eta *adi+prep+det+ize*, hurrenez hurren. Egitura horiexek dira, hain zuzen ere, gaztelaniazko konbinazioetan bildu ditugunak (3.2.1. atala), eta, agerraldien arabera antolatuta, ordena ere berbera da.

Bestalde, markaz markako azterketa eginez gero, absolutiboan dauden konbinazioen ordainek atentzioa ematen dute, denen ia bi heren hartzen baitituzte aditz soilek bakarrik: *uzta bildu – cosechar; oreka galdu – desequilibrar...* Gainerako ordainetan, berriz, agerpen gehien dituztenak preposiziorik

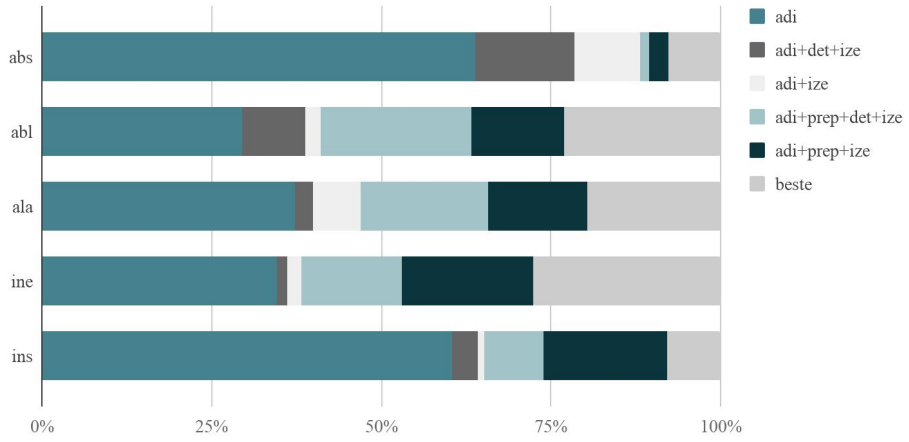
gabekoak dira, determinatzailedunak eta determinatzailerik gabeak, hurrenez hurren: *itxura egin – hacer el paripé; hotz izan – tener frío*.

EU marka	**ES egitura**	**Ehunekoa**
abs	adi	% 63,90
	adi + det + ize	% 14,49
	adi + ize	% 9,81
abl	adi	% 29,50
	adi + prep + det + ize	% 22,30
	adi + prep + ize	% 13,67
ala	adi	% 37,28
	adi + prep + det + ize	% 18,86
	adi + prep + ize	% 14,47
ine	adi	% 34,64
	adi + prep + ize	% 19,35
	adi + prep + det + ize	% 14,87
ins	adi	% 60,32
	adi + prep + ize	% 18,25
	adi + prep + det + ize	% 8,73

3.8 taula – Gaztelaniazko ordain motarik ohikoenak euskarazko markaren arabera (*Elhuyar* hiztegia)

Horrez gain, izenek postposizioren bat izateak ere badu eraginik ordainen egituretan (3.8. taula, 3.7. irudia). Izan ere, sarrien agertzen diren postposizioei –ablatiboari, adlatiboari, inesiboari eta instrumentalari– begiratuz gero, aditz soilen ondoren preposiziodun egiturak nagusitzen direla ikusten da, absolutibodun konbinazioen ordainetan ez bezala: *armairutik atera – salir del armario; eskura ordaindu – pagar en mano; bistan egon – saltar a la vista; barrez ito – morirse de risa*.

Hori horrela, eta euskarazko postposizioak gehienetan preposizioen bidez itzultzen direla aintzat hartuta, batek pentsa lezake konbinazioen itzulpena ez dela hain irregularra zentzu horretan –eta, hein batean, zuzen legoke–. Izan ere, kasuan kasuko azterketa eginda, ikusi dugu postposizio batzuen baliokidetzat ageri diren preposizioak nahiko sarri direla edozein hiztunek espero litzakeenak. Ablatiboaren pare, esaterako, *de* eta *por* ageri dira ordainen % 86,89tan (*ahotik kendu – quitar de la boca; zeharretik irten – salirse por la tangente*), eta adlatiboaren ordez, berriz, *a* erabili da % 76,47tan (*belarrira esan – decir al oído*).



3.5 irudia – Gaztelaniazko ordain motarik ohikoenak euskarazko markaren arabera (*Elhuyar* hiztegia)

Dena den, baliokidetza horiek hein batean baino ez dira erregularrak, postposizio guztiekin ez baita gauza bera gertatzen. Inesibodun konbinazioen ordainetan, adibidez, % 57,7tan baino ez da agertzen postposizio hori itzultzeko erabili ohi den preposizioren bat (*en*, *por* edo *sobre*); gainerako kasuetan, beste preposizioren bat ageri da: *baxoerditan ibili – ir de poteo, txantxetan hartu – tomar a broma.*

Euskarazko ordainen ezaugarri morfologikoak

Gaztelaniazko ordainekin egin dugun bezala, euskarazkoak ere egitura morfologikoaren arabera multzokatu ditugu hasteko:

- Aditz soila (**adi**): *geldotu, neskazahartu, txunditu.*
- Absolutibodun izena eta aditza (**ize.abs**+**adi**): *bizia arriskatu, ahotsa goratu, eskua jaso.*
- Postposizio-markadun izena eta aditza (**ize.pos**+**adi**): *buruan sartu, berriketan ibili, negarrari eman.*

- Adberbioa edo adberbio-sintagma eta aditza (**AdbS+adi**): *azkarrago ibili, alferrik galdu.*
- Adjektiboa edo adjektibo-sintagma eta aditza (**AdjS+adi**): *nabaria izan, izugarria izan, argal mantendu.*
- Izen soila ez den izen-sintagma (absolutiboan) eta aditza (**IS.abs+adi**): *kontu ezaguna izan, hitzaren jabe egon.*
- Izen soila ez den izen-sintagma (absolutiboa ez beste kasu- edo postposizio-markaren batekin) eta aditza (**IS.pos+adi**): *bere onetik atera, bere kabuz utzi.*
- Aurreko multzoetan sartzen ez direnak (**beste**): *amaitutzat jo, tentuz ibiltzeko esan, hortzen artean hitz egin.*

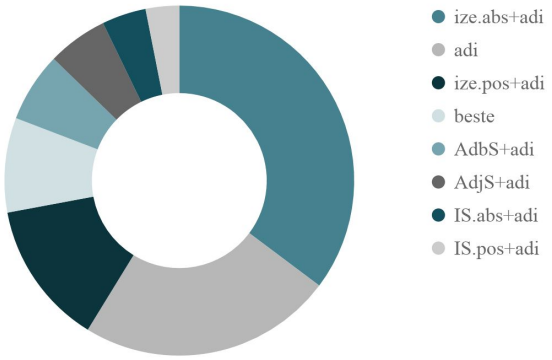
Multzo bakoitzaren agerraldiak 3.9. taulan eta 3.6. irudian bildu ditugu.

Egitura	**Agerr.**	**Ehunekoa**
ize.abs+adi	2316	% 35,22
adi	1546	% 23,51
ize.pos+adi	878	% 13,35
AdbS+adi	422	% 6,42
AdjS+adi	364	% 5,54
IS.abs+adi	269	% 4,09
IS.pos+adi	204	% 3,10
beste	576	% 8,74

3.9 taula – Euskarazko ordain motak (*Elhuyar* hiztegia)

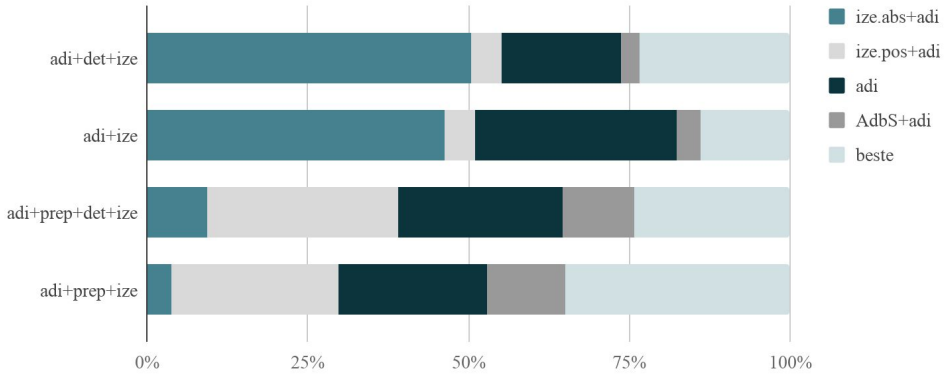
Euskarazko ordain guztiak kontuan hartuz gero, garbi ikusten da bi egitura beste guztiak baino askoz ere gehiago errepikatzen direla: *ize.abs+adi* motako konbinazioak lehenik, eta aditz soilak ondoren. Nolanahi ere, gaztelaniazko konbinazioak egituraren arabera bereizita, ikusten da determinatzaileen eta preposizioen agerpenak baduela zerikusia ordain motarekin, kontrako zentzuan gertatzen den bezala.

Har ditzagun, esaterako, aditzez eta izenez osaturiko konbinazioak batetik, eta aditzez, determinatzailez eta izenez osaturikoak bestetik. Agerpen gehien dituzten ordainak *ize.abs+adi* multzokoak eta aditz soilak dira bi kasuetan, eta gainerako ordain motak oso gutxi agertzen dira bi horien aldean.



3.6 irudia – Euskarazko ordain motak (*Elhuyar* hiztegia)

Hala ere, gaztelaniaz determinatzailerik ez duten konbinazioetan, *ize.abs+adi* multzoko ordainen eta aditz soilen arteko aldea ez da hain nabarmena ageRraldiak kontuan hartuz gero; determinatzailea duten konbinazioetan, aldiz, *ize.abs+adi* egiturako ordainak aditz soilak baino ia hiru bider gehiago dira (ikus aldea 3.10. taulan).



3.7 irudia – Euskarazko ordain motarik ohikoenak gaztelaniazko egituraren arabera (*Elhuyar* hiztegia)

Horrez gain, hemen ere badago loturarik preposizioen eta postposiziomarken artean. Izan ere, aurreko bi kasuetan absolutibo-markadun izenak

ES	EU egitura	Ehunekoa
adi+det+ize	ize.abs+adi	% 50,36
	adi	% 18,55
	adi.pos+adi	% 4,83
	AdbS+adi	2,96
adi+ize	ize.abs+adi	% 46,34
	adi	% 31,38
	ize.pos+adi	4,62
	AdbS+adi	3,75
adi+prep+det+ize	ize.pos+adi	% 30,07
	adi	% 25,51
	AdbS+adi	% 11,25
	ize.abs+adi	% 9,02
adi+prep+ize	ize.pos+adi	% 26,18
	adi	% 23,01
	ize.abs+adi	% 14,16
	AdbS+adi	% 12,10

3.10 taula – Euskarazko ordain motarik ohikoenak gaztelaniazko egituren arabera (*Elhuyar* hiztegia)

nagusi baziren ere, preposiziodun konbinazioen ordainetan ohikoagoak dira postposizio-markaren bat daramaten izenak. Hain zuzen ere, gaztelaniazko konbinazioak *adi+prep+ize* edo *adi+prep+det+ize* multzoetakoak direnean, sarrien agertzen den ordain mota *ize.pos+adi* da.

Aurreko atalean bezala, hemen ere, preposizioen baliokidetzat agertzen diren postposizioak aurreikusteko modukoak dira batzuetan: *en* preposizioa duten konbinazioen ordainetatik % 87,68k inesibodun izen bat daramate (*estar en la inopia – ametsetan egon*), *con* preposizioaren ordez instrumentala eta soziatiboa agertzen da ordainen % 87,5etan (*andar con cuidado – kontuz ibili, ir con el cuento – koplarekin etorri*), eta *por* preposizioa ere ablatiboaren eta inesiboaren bidez ordezkatu da % 88,56 kasutan (*pasar por el tamiz – galbahetik pasatu, pasar por las armas – armetan iragan*).

Hemen ere, ordea, hein batean baino ez da erregulartasun hori gordetzen. Esate baterako, *a* preposizioa % 56 kasutan bakarrik itzuli da adlatiboaren edo datiboaren bidez (*traer a la memoria – gogora ekarri*), eta gainerako ordain guztiek beste marka batzuk dituzte: *andar a la greña – istilutan ibili, saltar a la vista – begi-bistakoa izan.*

Azkenik, interesgarria da nabarmentzea adberbioak ere nahiko maiz agertzen direla preposiziodun konbinazioen ordainetan; 3.8. taulan ikus daitekeenez, *adi+prep+det+ize* multzoan, absolutibodun izenekin sortzen diren egiturak baino are sarriago agertzen dira adberbioekin sortzen direnak: *caer por su peso – argi egon, dar en el clavo – bete-betean asmatu,* eta abar.

Mugatasuna, definitutasuna eta numeroa

Ezaugarri morfosintaktikoak aztertzen eta alderatzen jarraitzeko, mugatasunari, definitutasunari eta numeroari begiratu diegu, hizkuntza batetik bestera gordetzen ote diren jakiteko. Dena dela, kontuan hartu behar da ezaugarri horiek, berez, oso zailak direla hizkuntza baten eta bestearen artean alderatzeko, gaztelaniazko determinatzaileen definitutasuna eta euskarazko mugatasuna ez baitira baliokideak.

Bat-etortze falta horren erakusgarri da, adibidez, *Elhuyar* hiztegian gaztelaniazko artikuluak bilatuta aurkitzen dena: *el/la* sarreran, euskarazko *-a* artikulua agertzen da ordain gisa, baina *un/una* sarreran ere, *bat* determinatzaileaz gain, agertzen da *-a* artikulua. Izan ere, euskarazko artikulu mugatuaren erabilera oso zabala da, zabalagoa gaztelaniazko eta beste hizkuntza askotako artikulu zehaztuena baino[2] (Trask, 2003: 96. orr.). Hortaz, kontuan hartu behar da guk hemen azalduko duguna parekatze-proposamen bat baino ez dela eta, ziur asko, kasuren batean zurrunegia eta besteren batean orokorregia izango dela.

Euskaraz, lau ezaugarri desberdindu ditugu: mugagabea (mg), singularra (s), plurala (pl), eta zalantzazkoa (*), singularra zein mugagabea izan litezkeen kasuetarako. Gaztelaniaz, berriz, singularra (s) eta plurala (pl) bakarrik. Parekatze-proposamena 3.11. taulan laburbildu dugu. Taulako informazioa behar bezala antolatzeko, gaztelaniazko numero singularra beste bi azpimultzotan banatu dugu[3]: artikulu zehaztudun izen-sintagmak (zeh) eta determinatzailerik gabeak (-).

[2]Traskek (2003: 96. orr.) honako hau dio, euskarazko artikulu mugatuaren askotariko erabilerak zerrendatu aurretik: “The label ‘definite article’ is misleading, since this article is of much broader use than the English definite article”.

[3]Azterketa honetan, aintzat hartu dugu artikulu zehaztuak ez diren determinatzaileek ez zutela parekorik izango euskarazko konbinazioetan. Izan ere, guk, euskaraz, izena+aditza osaerako konbinazioak bakarrik landu ditugu, eta ez dugu kontuan hartu izen-sintagman sar litekeen beste determinatzailerik, *bat, asko* eta halakorik, gaztelaniazko konbinazioetan askotariko determinatzaileak ager daitezkeen arren. Horregatik daude 3.11. taulan artikulu zehaztudun izen-sintagmak eta determinatzailerik gabeak bakarrik.

			EU			
			MG	*****	**S**	**PL**
ES	**S**	**zeh**		X	X	
		-	X	X		
	PL					X

3.11 taula – Mugatasunaren eta numeroaren ezaugarriak gaztelaniaren eta euskararen artean parekatzeko proposamena

Hau da: gaztelaniatik euskararako zentzuan, parekatze hau egin dugu:

- Izen-sintagma singularra (*dar una paliza, volver la espalda, hacer carrera*)
  - → Artikulu mugatu singulardun izena (*egurra eman*)
  - → Izen mugagabea (*bizkar eman*)
  - → Singularra zein mugagabea izan litekeen izena (*arrakasta izan*)
- Izen-sintagma plurala (*dar recuerdos*)
  - → Artikulu mugatu pluraldun izena (*goraintziak eman*)

Euskaratik gaztelaniarako zentzuan, berriz, honela:

- Artikulu mugatu singulardun izena (*adarra jo*)
  - → Artikulu zehaztu singulardun izen-sintagma (*tomar el pelo*)
- Artikulu mugatu pluraldun izena (*belarriak zabaldu*)
  - → Izen-sintagma plurala (*aguzar las orejas*)
- Izen mugagabea (*su egin*)
  - → Determinatzailerik gabeko izen-sintagma singularra (*abrir fuego*)
- Singularrean zein mugagabean egon litekeen izena (*hanka egin, zarata egin*)
  - → Artikulu zehaztu singulardun izen-sintagma (*ahuecar el ala*)

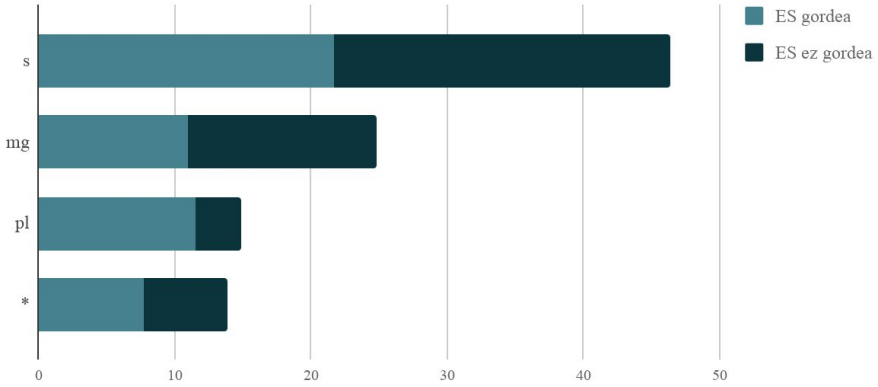
→ Determinatzailerik gabeko izen-sintagma singularra (*armar bulla*)

Parekatzeak behar bezala egiteko, konbinazio-zerrenda osoaren zati bat bakarrik hartu dugu, hain zuzen ere, ordain gisa ere konbinazioak dituzten konbinazioez osatua.

Euskarazko konbinazioetan, mugatu singularrean daude izenen ia erdia (*umea izan, ahoa garbitu*), eta mugagabearen (*ohar egin, zerraldo utzi*) eta mugatu pluralaren (*goraintziak eman, erroak bota*) agerraldien artean ez dago desberdintasun handirik. Bestalde, konbinazioen hamarren bat baino gehiago zalantzazko kasuak dira (*denbora egin, lotsa izan*), ezin baita zehatz jakin mugatu singularrean ala mugagabean dauden.

	EU konbinazioak	**ES gordea**
s	% 46,35	% 53,23
mg	% 24,87	% 56,12
pl	% 14,92	% 22,45
*****	% 13,86	% 44,69

3.12 taula – Mugatasuna eta numeroa euskaratik gaztelaniara (*Elhuyar* hiztegia)



3.8 irudia – Mugatasuna eta numeroa euskaratik gaztelaniara (*Elhuyar* hiztegia)

Eskuineko zutabean (3.12. taula), euskarazko ezaugarria gaztelaniaz zenbatetan gorde den ikus daiteke. Deigarriena mugatu pluralaren kasua da

beharbada, baliokideen laurdenak baino gutxiagok baitu artikulu zehaztu plurala gaztelaniaz. Zalantzazko kasuen ordainetan, ia hiru laurdenek dute pareko mugatasuna eta numeroa, eta mugagabea eta singularra erdia baino gehixeagotan gordetzen dira.

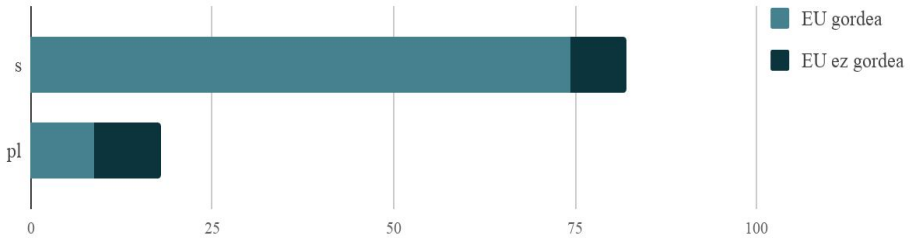
Gaztelaniazko konbinazioak abiapuntutzat hartuz gero, ordea, emaitzak aldatu egiten dira.

	ES konbinazioak	**EU gordea**
s	% 82,09	% 90,66
pl	% 17,91	% 48,78

3.13 taula – Mugatasuna eta numeroa gaztelaniatik euskarara (*Elhuyar* hiztegia)

Batetik, konbinazio gehienak singularrean daude (*sacar provecho, estar en auge, perder el juicio, quitar de la cabeza*), eta pluralean, berriz, oso gutxi (*parar los pies, subirse por las paredes*).

Bestetik, ordainetan mugatasuna eta numeroa zenbatetan gordetzen den begiratuz gero, alde handia nabari da bien artean. Singularrean dauden konbinazioen ordainak singularrean daude gehien-gehienetan, baina ez da gauza bera gertatzen pluralarekin, erdirak baino gehiagok ez baitute ordain pluralik.



3.9 irudia – Numeroa gaztelaniatik euskarara (*Elhuyar* hiztegia)

Hortaz, badirudi gaztelaniatik euskarara numero singularra nahiko erregularki gordetzen dela baina, gainerakoan, irregulartasun handia dagoela mugatasunari eta numeroari dagokienez: bai gaztelaniatik euskarara izen-sintagma pluraletan, bai eta euskaratik gaztelaniara, ezaugarria edozein dela ere.

Ezaugarri lexikoak: aditzen eta izenen baliokidetza

Aurreko atalean (3.2.1), hizkuntza bateko eta besteko konbinazioen osagai nagusiak aztertu ditugu: aditzak eta izenak. Oraingoan, berriz, konbinazioen osagaiak euren ordainenekin alderatuko ditugu. Mugatasuna eta numeroa aztertzeko egin dugun bezala, izenez eta aditzez osatutako ordainak bakarrik hartu ditugu, dagozkien hitz-konbinazioekin batera. Ondoren, bi hizkuntzetako aditzak eta izenak parekatu, eta hiztegian bilatu dugu ea bata bestearen ordain gisa ageri diren.

Adibidez, *deabruetara bidali – mandar al infierno* bikotea hartuta, *deabru – infierno* eta *bidali – mandar* hiztegian baliokidetzat jasota ote dauden begiratu dugu. Kasu horretan, aditzak bakarrik agertzen dira ordaintzat, *deabru* sarreraren ordainetan ez baita *infierno* agertzen.

Hasi aurretik, gure irudipena zen nahiko gutxitan agertuko zirela ordaintzat bai izena eta bai aditza, eta, oro har, gure ustea bete da, nahiz eta proportzioa desberdina izan euskaratik gaztelaniarako eta gaztelaniatik euskararako bikoteetan. Gure kontaketetatik atera ditugun ehunekoak 3.14. taulan jaso ditugu.

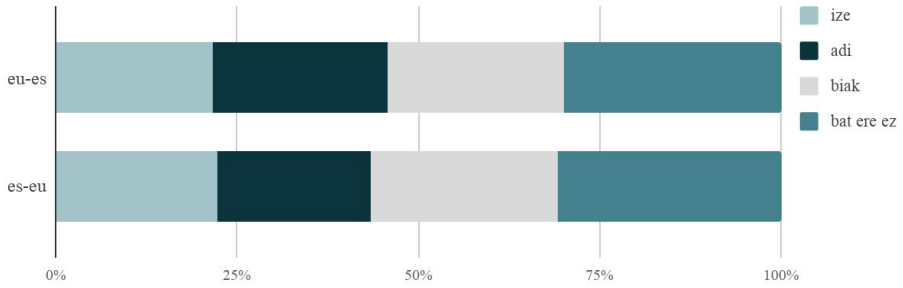
	ize	**adi**	**biak**	**bat ere ez**
eu–es	% 21,62	% 24,01	% 24,47	% 29,90
es–eu	% 22,29	% 21,16	% 25,74	% 30,81

3.14 taula – Izenen eta aditzen baliokidetza *Elhuyar* hiztegian

Taulako azken zutabeak atentzioa ematen du, izan ere, zentzu bateko zein besteko konbinazio-pareen artean, ia herena dira ez izenik eta ez aditzik baliokidetzat ez dutenak. Bai izena eta bai aditza ordaintzat dituztenak, ostera, laurdena baino ez dira; beraz, nahiko argia da izena+aditza UFen itzulpena oso gutxitan egin daitekeela hitzez hitz, hiztegiaren arabera behintzat.

Horrez gain, kontuan izan behar da guk hitzak hiztegian baliokidetzat agertzen diren ala ez bakarrik begiratu dugula baina, baiezko kasuetan, ez dugula zehaztu ordaina lehena den ala ez. Esaterako, *bidea urratu – abrir camino* parean, aditzak baliokidetzat hartu ditugu, *urratu* sarreraren barruan *abrir* agertzen delako ordainen artean; euskaraz eta gaztelaniaz egiten dugunok, ordea, nekez erabiliko genituzke ordain gisa, *bide*ren alboan ez bada.

Laburbilduz, atal honek guztiak agerian uzten du hitz-konbinazioen itzulpena zenbateraino den irregularra eta, hori kontuan hartuta, zenbateraino



3.10 irudia – Izenen eta aditzen baliokidetza *Elhuyar* hiztegian

den beharrezkoa haiek tratatzeko estrategia landuak erabiltzea itzultzaile automatikoetan, xede-hizkuntzan testu txukun bat sortu nahi bada.

Laburpena

Kapitulu honetan, bi azterketa deskriptibo egin ditugu: batetik, *Elhuyar* hiztegian jasotako aditza+izena UFena, eta bestetik, *Matxin* itzultzaileak halako UFei ematen dien itzulpenena. Lan horiek burutzeko, hiru hipotesi izan ditugu gogoan. Hona hemen hiru hipotesiak eta kapitulu honetako edukiek zer dioten haiei buruz.

[A1] UFak, askotan, ez dira hitzez hitz itzultzen hizkuntza batetik bestera.
Elhuyar hiztegiaren arabera, hala da. Landutako aditza+izena konbinazio guztietatik, herenak baino gutxiagok du hitzez hitzeko ordaina, gaztelaniatik euskarara nahiz euskaratik gaztelaniara. Gainera, aldaketak ez dira lexiko mailakoak bakarrik, gaztelaniazko % 49 konbinaziok eta euskarazko % 30ek bakarrik baitituzte izen batez eta aditz batez osatutako ordainak; gainerako ordainek bestelako osaera morfologiko bat dute.

[A3] Fraseologia mailan aztertutako hizkuntza askoren aldean, euskaraz bereziki ohikoak dira aditz arinak barne hartzen dituzten UFak.
Egindako azterketak baietz iradokitzen du, bi arrazoirengatik. Batetik, hiztegiko izena+aditza konbinazioen artean euskaraz gehien errepikatzen diren sei aditzek bakarrik konbinazio guztien ia % 51 osatzen dutelako, eta gaztelaniaz, berriz, konbinazio guztien % 36 dira zortzi aditzik errepikatuenak dituztenak. Bestetik, euskarazko konbinazioen gaztelaniazko ordainen artean, aditz soilak dira ohikoenak alde handiz (% 58). Bigarren datu hori esanguratsua da, aditz arindun UFek sarri izaten baitituzte aditz soilak ordaintzat beste hizkuntza batzuetan eta, gainera, horrek argiago erakusten baitu hitz-konbinazio horiek esanahi-unitate bakarra adierazten dutela. Beste hizkuntza batzuk ere hartuko ditugu kontuan geroago, 7. kapituluan.

[A6] UFei buruzko informazio morfosintaktikoa kontuan hartzea onuragarria izan daiteke itzultzaile automatikoentzat.
Matxin itzultzaileari erreparatuta, badirudi hala dela, informazio morfosintaktikoaren falta baita UFak itzultzean egiten dituen akats askoren iturria, bai identifikazioari dagokionez eta bai itzulpenari berari dagokionez ere.

Bestalde, bi azterketa horien bidez, 1.3. atalean zerrendatutako hiru helburu betetzeko pausoak eman ditugu.

[H1] Gaztelaniazko eta euskarazko aditza+izena motako UFen ezaugarri morfosintaktikoak aztertzea
Elhuyar hiztegiko konbinazioak aztertu ditugu, lexikoari eta ezaugarri morfologikoei erreparatuz. Dena dela, kontuan hartu behar da konbinazioen forma kanonikoa bakarrik hartu dugula kontuan, alegia, hiztegian agertzen dena bakarrik, haren aldaki morfosintaktikoak alde batera utzita. Datozen kapituluetan sakonduko dugu gehiago bigarren alderdi horretan.

[H2] Aditza+izena motako UFak gaztelaniaren eta euskararen artean nola itzultzen diren aztertzea
Hiztegietan jasotako UFek zer-nolako ordainak dituzten aztertu dugu, hasierako urrats gisa. Hemen ere lexikoan eta ezaugarri morfologikoetan jarri dugu arreta. Aurrerago joko dugu corpus paraleloetara, hiztegietatik ateratako ondorioak corpusetatik ateratakoekin bat ote datozen ikusteko.

[H5] Aditza+izena motako UFen informazio linguistikoak itzulpen automatikoan zer eragin duen aztertzea.
Matxin itzultzaileak UFak itzultzean egiten dituen akatsak nolakoak diren aztertu dugu. Testuinguruak sortzen dituen arazoak alde batera utzita, hiru multzotan banatu ditugu sistemaren zailtasunak: lexikoi elebidunaren mugak, identifikazio-metodoaren gabeziak eta itzultze-metodoaren hutsuneak. Datozen bi kapituluetan azalduko dugu zer egin dugun arazo horiek konpontzen laguntzeko.

4. KAPITULUA

Gaztelaniazko konbinazioen azterketa eta identifikazioa

Aurreko kapituluan erakutsi dugunez, *Matxin* itzultzaileak egiten dituen akatsetako batzuk UFak ondo ez identifikatzeagatik sortzen dira. Akats horiek nolakoak diren ikusita, badirudi informazio morfosintaktikoa giltzarria dela arazo horri aurre egin ahal izateko, eta ideia horretatik abiatu gara gu kapitulu honetan azalduko ditugun lanetan. *Matxin*ek gaztelaniatik euskarara itzultzen duenez, identifikazio-arazoak sortzen dizkioten UFak gaztelaniazkoak dira, eta haien inguruan jardungo dugu hemen.

Hasteko, aditza+izena motako UFen ezaugarriak xeheki aztertu ditugu (4.1. atala), eta identifikazio-esperimentu bat egin dugu azterketatik ateratako informazioa lagungarria ote den jakiteko (4.2. eta 4.3. atalak). Ondoren, behin esperimentu horren emaitzak ikusita, aurreko azterketa erdiautomatizatzeko metodologia bat sortu dugu (4.4. atala), eta identifikazio-esperimentua errepikatu dugu (4.5. atala), eskuzko informazioaren ordez erdiautomatikoki lortutakoa erabilita.

Datozen ataletan azalduko ditugu lau pauso horien nondik norakoak, eta, kapituluaren amaieran, edukiak laburbilduko ditugu.

4.1 Eskuzko azterketa xehea

Ezaugarri morfosintaktikoetan jarri dugu arreta batez ere, informazio hori baita, gure ustez, UFen identifikazioari ekarpen handiena egingo diona.

Horren aurretik, ordea, ezaugarri lexiko-semantikoei ere begiratu bat egin nahi izan diegu. Izan ere, ikerketa-lan batek baino gehiagok aipatzen du lexiko-semantika eta morfosintaxia lotuta daudela fraseologian (Corpas Pastor, 1996; Urizar, 2012; Markantonatou *et al.*, 2018), eta mota lexiko-semantiko bereko UFek antzeko ezaugarri morfosintaktikoak izan ohi dituztela.

Gainera, azterketa honetan, *Elhuyar* hiztegitik erauzitako hitz-konbinazioak hartuko ditugu oinarritzat, prestaketa-lanetan erabilitako berberak, eta gogoan izan behar da, hiztegi-sarrerak ez ezik, ordaintzat jasotako aditza+izena konbinazioak ere badaudela sorta horretan. Horrek esan nahi du balitekeela konbinazio horietako batzuk UFak ez izatea eta, beraz, ez behar izatea tratamendu berezirik.

Ezaugarri lexiko-semantikoez arituko gara lehenik (4.1.1. atala), eta geroago sakonduko dugu morfosintaxian (4.1.2. atala).

4.1.1 Ezaugarri lexiko-semantikoak: idiomatikotasunaren *continuuma*

Esan bezala, atal honetan azalduko duguna ere prestaketa-lanean aztertu ditugun hitz-konbinazioen gainean egin dugu. Bigarren azterketa hau, ordea, aurrekoa baino xeheagoa izan da, eta konbinazio guztietatik usuenak bakarrik aukeratu ditugu, 150 guztira. Hainbat generotako testuetatik bildutako corpus bat hartu dugu, 491.853 esaldikoa, eta Freeling analizatzailearen identifikazio-metodoa erabili dugu maiztasunak kalkulatzeko, *Matxin*ek erabiltzen duen identifikazio-metodo berbera. Hau da: konbinazioko hitzak bilatu ditugu, bata bestearen jarraian eta, aditzaren flexioa gorabehera, forma berean. Gehien agertu diren 150 konbinazioak hautatu ditugu azkenean.

Hurrengo pausoak honakoak izan dira: lehenik, aztergai genituen konbinazioei begiratuta, sailkapen-proposamen bat egin dugu, eta bigarrenik, anotatze-lan bat egin dugu sailkapen hori egokia ote den ikusteko. Datozen lerroetan azalduko dugu gure sailkapena zertan datzan eta zer ondorio atera ditugun anotatze-lanetik.

Sailkapen-proposamena

Txosten honen 2. kapituluan azaldu dugunez, UFak sailkatzeko, gure lanaren bi aurrekaririk gertukoenek aldi berean hartzen dituzte kontuan ezaugarri lexiko-semantikoak eta morfosintaktikoak (Urizar, 2012; Gurrutxaga, 2014). Guk, ordea, morfosintaxian berariaz sakontzeko asmoa dugunez, bi alorrak

bereiz lantzea erabaki dugu: semantikari eta lexikoari dagozkion ezaugarriak alde batetik, eta morfosintaxiari dagozkionak beste batetik.

Gainera, gogoan izan behar da lan horiek hizkuntza bakarreko UFak sailkatzen dituztela, eta ikerketa-lan honetan, berriz, pisu handia duela itzulpenak. Hori dela-eta, kolokazio-lokuzio bereizketa egiteaz gain, lokuzio opakoak eta metaforikoak ere bereizi egin ditugu, Gurrutxagak (2014) bezala.

Gure lanaren mamia morfosintaxian dagoenez, ez gara askorik luzatuko multzo lexiko-semantikoen inguruko azalpenetan. Nolanahi ere, ildo interesgarria da etorkizunerako, bereziki jakingarria litzatekeelako ikustea zer portaera duen benetan multzoetako bakoitzak itzulpenetan. Hemen, momentuz, multzo bakoitzaren ezaugarri nagusiak aipatuko ditugu, eta gure ustez haien itzulpenek zer berezitasun izan ohi duten ere esango dugu, etorkizuneko lanetarako hipotesi gisa.

- **Lokuzio opakoak**. Konbinazio osoaren esanahia ez da osagai-hitzen esanahien batura, eta ezin da ulertu baldin eta konbinazio osoa nolabait ezagutzen ez bada. Oso litekeena da hizkuntza batetik bestera hitzez hitzeko itzulpenik ez izatea edo, are gehiago, ez izatea UFrik ordaintzat.

  (55) ES: *Dice lo que piensa sin* ***cortarse un pelo****.*
  EU: *Batere lotsatu gabe esaten du zer pentsatzen duen.*

  (56) EU: *Beti ari da mundu guztiari* ***adarra jotzen****.*
  ES: *Siempre está* ***tomando el pelo*** *a todo el mundo.*

- **Lokuzio metaforikoak**. Konbinazio osoaren esanahia ez da osagai-hitzen esanahien batura, baina metafora bidez uler liteke. Lokuzio opakoak ez bezala, lokuzio metaforikoak hizkuntza batetik bestera hitzez hitz itzulita ere, baliteke helburu-hizkuntzan ulergarriak izatea.

  (57) ES: *Ese tipo de experiencias* ***dejan huella****.*
  EU: *Halako bizipenek* ***aztarna uzten*** *dute.*

  (58) EU: *Bere intereseko gaiez ari garenean,* ***belarriak luzatzen*** *ditu berehala.*
  ES: *Cuando hablamos de temas que le interesan,* ***aguza las orejas*** *enseguida.*

- **Kolokazioak (aditz arindun konbinazioak barne)**. Konbinazioko hitzek ausaz espero litekeena baino joera handiagoa dute elkarrekin agertzeko eta, normalean, esanahiaren pisurik handiena izenak izaten du. Hautapen lexikoa murriztua izaten da, alegia, ezin izaten dira hitzak sinonimoen bidez ordezkatu, berez esanahi berbera adierazten badute ere (59. adibidea). Aditz arinak barne hartzen dituzten konbinazioen kasuan, aditzak esanahia "galtzen" duela esan ohi da, izenak adierazten duen esanahiari aditz-izaera emateko balio duela nolabait (60. adibidea). Itzulpenetan, litekeena da izenari bere ohiko ordaina ematea baina aditzak ohikoa ez den ordainen bat behar izatea.

(59) ES: *No hacen más que gritar y* ***meter ruido****.*
EU: *Oihuka aritu eta* ***zarata atera*** *besterik ez dute egiten.*

(60) EU: *Haurrak bere lehen* ***urratsak egin*** *zituen.*
ES: *La niña* ***dio*** *sus primeros* ***pasos****.*

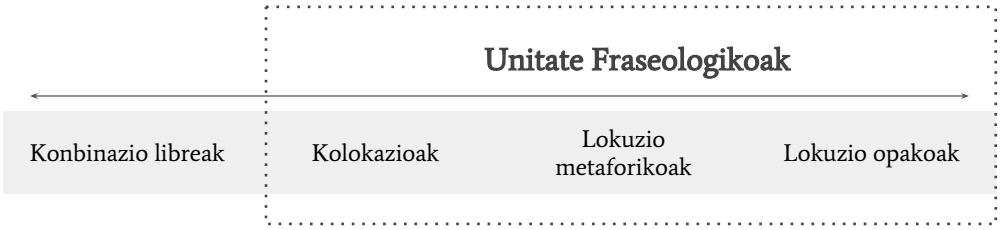
UFetatik kanpo, berriz, **konbinazio libreak** dauzkagu, non hitzak era askean konbinatzen baitira, hizkuntzen ohiko lexiko- eta gramatika-arauei jarraituz. Hizkuntzen ezaugarri tipologikoak gorabehera, hitzez hitzeko itzulpenek ez dute arazorik ematen gehienetan.

(61) ES: *Este año iremos a un lugar diferente.*
EU: *Aurten beste leku batera joango gara.*

(62) EU: *Beti denda berean erosten ditu liburuak.*
ES: *Siempre compra sus libros en la misma tienda.*

Dena dela, lehenago ere aipatu dugu (2.1.2. atala) idiomatikotasunak nolabaiteko *continuum* bat osatzen duela eta oso zaila dela multzoen arteko mugak ondo definitzea. Gure lau multzoak bata bestearen alboan jarrita, continuumak 4.1. irudiko forma hartzen du.

Behin sailkapena sortuta, gaztelaniazko konbinazioen anotatze-lanari ekin diogu, sailkapena egokia ote den eta zer ondorio atera daitekeen ikusteko. Bestalde, esan bezala, interesgarria litzateke aztertzea multzo bakoitzeko konbinazioek zer portaera duten egiaz corpus paraleloetan eta, hala, begiratzea ea emaitzak bat datozen guk sailkapena proposatzeko gogoan izan ditugun itzulpen-hipotesiekin. Lehen lanari buruz datorren atalean jardungo dugu, eta etorkizunerako utziko dugu bigarrena.



4.1 irudia – Idiomatikotasunaren continuuma, sailkapen-proposamenaren arabera

Anotatze-lana

Hiru hizkuntzalarik jardun dute anotatze-lanetan. Kategoria bakoitzari buruzko azalpen labur batzuk –aurreko atalekoen gisakoak– jaso ostean, hiru anotatzaileetako batek, anotatzaile nagusiak[1], 150 konbinazioko sorta osoa etiketatu du, eta beste biek, berriz, erdibana, 75na konbinazio. Hala, jasotako zerrendako elementuak lehen aipaturiko lau multzoetan sailkatu behar izan dituzte, inongo testuingururik gabe. Kasuren bat multzo batean baino gehiagotan sar zitekeela uste bazuten, erabilerarik ohikoenaren taldea esleitzeko eskatu diegu.

Jarraian, anotatzaile nagusiaren emaitzak eta beste bi anotatzaileenak alderatu ditugu. Bi eratara kalkulatu dugu adostasuna: ehunekotan, eta Cohen kappa κ (Cohen, 1960) erabiliz. Bigarren neurri horrek, adostasuna kalkulatzeko, kontuan hartzen du anotatzaileak ausaz zenbatetan bat etorriko liratekeen ere, aukeran dauzkaten kategorien arabera. Ausaz asmatzea erraza bada, bat-etortzeen pisua jaitsi egiten da, eta alderantziz. Lortu ditugun emaitzak 4.1. taulan daude jasota.

Ehunekotan dagoen emaitzak ez du interpretazio-arazorik sortzen: anotatzaileek kategoria berbera eman diote konbinazio guztien hiru laurdeni baino gehiagori. Kontuan hartuta fraseologian kategorien arteko mugak ez direla inoiz behar bezain argiak, esan daiteke emaitza txukun samarra dela. Cohen κ ona den ala ez kalkulatzeko, aldiz, hainbat alderdiri begiratu behar zaio.

Landis eta Kochen (1977) arabera, gure emaitzek dezenteko adostasu-

[1]Lan honetan, *anotatzaile nagusi* deituko diogu anotatze-lanaren zatirik handiena egiten duenari eta, horrez gain, atazaren baten amaieran erabakiak bateratu beharra egonez gero, erabakiak hartzeko ardura duenari.

	4 kat.	**3 kat.**	**2 kat.**
Bat-etortzea	% 76,00	% 76,00	% 83,34
Cohen κ	0,63	0,62	0,61

4.1 taula – Anotatze lexiko-semantikoan lortutako adostasuna

na erakusten dute, *substantial agreement* multzoan sartzen baitituzte 0,61 eta 0,80 arteko neurriak. Hala ere, ikertzaile guztiek ez dituzte haien irizpideak erabat onartzen; batzuek, uste dutelako ez direla nahikoa objektiboak (Gwet, 2012), eta beste batzuek, emaitzak beste era batera interpretatzen dituztelako (Krippendorff, 1980).

Lehenago ere aipatu dugunez, azken urteotan UFen alorrean egin den anotatze-lanik garrantzitsuena PARSEME proiektuan eginikoa da. Aditz-UFak identifikatzeko proposatu zuten ataza partekaturako, 18 hizkuntzatako corpus anotatuak argitaratu zituzten 2017an (Savary *et al.*, 2017), eta 20 hizkuntzatakoak 2018an (Ramisch *et al.*, 2018). Lan zabal hori corpusetan oinarritua bada ere, jakin daiteke hizkuntza bakoitzeko anotatzaileek zenbateko adostasuna izan duten UF motei dagokienez, eta emaitza horiek alderatu ditugu gureekin.

Hizkuntza guztiak kontuan hartuta, gure emaitzak 2018ko batezbestekoa (0,67) baino lau puntu beherago leudeke, baina gaztelaniazkoa (0,57) baino sei puntu gorago. Gainera, aipagarria da gaztelaniaz lortutako adostasuna nabarmen txikia dela hizkuntza gehienen aldean, PARSEMEren gidalerroak denentzat berberak izan arren. Nolanahi ere, gure lana eta PARSEMEn eginikoa ez dira erabat alderagarriak, multzo desberdinak erabili baititugu, bai izaeran eta bai kopurutan.

Guretik gertuen dagoen lana, ataza honetan, Gurrutxagarena (2014: 157–163 orr.) da. Lokuzioen, kolokazioen eta konbinazio libreen arteko bereizketa eginda, 0,51 eta 0,62 arteko Cohen κ lortu dute Gurrutxagarren laneko anotatzaileek, euskarazko corpusetatik automatikoki erauzitako izena+aditza konbinazioak multzokatzerakoan. UFak eta konbinazio libreak bakarrik bereizita, berriz, 0,53tik 0,66rakoak. Hau da, hiru multzo erabilita lortu duten Cohen κ neurririk onena gurearen berdin-berdina da; bi multzo bakarrik erabilita, berriz, haien neurririk onena gurea baino hobea da, baina batez bestekoa (0,58), apalxeagoa.

Bestalde, desadostasunei pixka bat gehiago erreparatuta, bada ondorio

interesgarriak ateratzeko moduko daturik. Begira, bestela, continuumaren arabera sortu dugun adostasun-matrizean (4.2. taula) non kokatzen diren anotatzaileak bat etorri ez direneko kasuak.

		Beste anotatzaileak			
		Lok.Op.	**Lok.Met.**	**Kolokazioa**	**Librea**
Anotatzaile nagusia	**Lok.Op.**	**1**	0	1	0
	Lok.Met.	0	**20**	2	1
	Kolokazioa	0	8	**69**	15
	Librea	0	1	8	**24**

4.2 taula – Gaztelaniazko anotatze lexiko-semantikoan lortutako adostasun-matrizea

Kontuan hartuta ezkerretik eskuinera eta goitik behera doan lerro diagonalak bat-etortzeak erakusten dituela, deigarria da desadostasun gehiengehienak lerro horren alboan daudela. Alegia, hiru kasutan izan ezik, desadostasuna sortu duten beste 33 konbinazioak gure continuumean ondoz ondo dauden kategorietan sailkatu dira: lokuzio metaforikoetan eta kolokazioetan, edo kolokazioetan eta konbinazio libreetan. Eta, era berean, erabat kontrako etiketarik ez da inoiz eman: anotatzaile batek lokuzio opakotzat etiketatutakoa ez du besteak inoiz sailkatu konbinazio libretzat. Hortaz, lan honen emaitzak fraseologian sarri aipatu izan den continuumaren erakusgarri garbia dira (ikus, besteak beste, 2.1.1. orrialdean esaten dena).

Azkenik, esanguratsua da sorta osoan konbinazio bakarra etiketatu dela lokuzio opakotzat, eta kolokazioen etiketa izan dela erabiliena. Geroago ere, 4.4. atalean eta 7. kapituluan, emango dugu datu gehiago hori hala izan ohi dela erakusteko.

4.1.2 Ezaugarri morfosintaktikoak

Morfosintaxia aztertzen hasteko, alde batera utzi ditugu sailkapen lexiko-semantikoan konbinazio libretzat etiketatutakoak, eta UFtzat sailkatutako 117ekin egin dugu aurrera. Azal dezagun orain zer aztertu dugun zehazki eta zer sailkapen egin dugun.

Sailkapen-proposamena

Sag *et al.*-i (2002) eta HPko beste egile batzuei jarraituz, hiru multzo morfosintaktiko bereizi ditugu: konbinazio finkoak, erdifinkoak eta malguak. Horretarako, Buckinghamek (2009), Parrak (2018) eta beste batzuek egin bezala, test gisako galderetan oinarritu dugu gure azterketa. Konbinazio bakoitzeko, honako galdera hauek egin ditugu:

- Izen-sintagmak badarama determinatzailerik? Beti (63), batzuetan (64) ala inoiz ez (65)?

  (63) *Te* ***toma el pelo***.
  **Te toma pelo.*

  (64) ***Hace frío***.
  Hace *un* ***frío*** *insoportable.*

  (65) *El moderador* ***dio paso*** *a la siguiente pregunta.*
  **El moderador dio el paso a la siguiente pregunta.*

- Determinatzailea badarama, nolakoa da? Beti zehaztua[2] (66), beti zehaztugabea (67) ala bata zein bestea (68)?

  (66) *Te* ***toma el pelo***.
  **Te toma un pelo.*

  (67) *Ese niño* ***es un amor***.
  **Ese niño es el amor.*

  (68) ***Hazme*** *un* ***favor***.
  Hazme *el* ***favor*** *que te pido.*

- Zer numero du izen-sintagmak? Beti singularra (69), beti plurala (70) ala bata zein bestea (71)?

[2] Azterketa honetan, *determinatzaile zehaztu* diogunean, hiru determinatzaile motari buruz ari gara, erreferente ezagunak izendatzeko balio duten hirurez: artikulu zehaztuez, determinatzaile posesiboez eta erakusleez. Multzokatze hori lan gehiagotan ere egiten da, hala nola RAEren gramatikan (RAE, 2009) eta determinatzaileen inguruan Gutierrezek eginiko azterketa sakonean (2008). Gainerako determinatzaileak zehaztugabetzat markatu ditugu.

(69) *Han **llevado a cabo** el proyecto.*
**Han llevado a cabo<u>s</u> el proyecto.*

(70) *Se enfadaron tanto que **llegaron a <u>las</u> manos**.*
**Se enfadaron tanto que llegaron a <u>la</u> mano.*

(71) ***Cumpliremos con** <u>el</u> **plazo** acordado.*
***Cumpliremos** con <u>los</u> **plazo<u>s</u>** acordados.*

- Egon al daiteke modifikatzailerik izen-sintagmaren barruan? Adjektiborik, adibidez. Bai (72) ala ez (73)?

(72) *Hemos **sacado fotos**.*
*Hemos **sacado** <u>bonitas</u> **fotos**.*

(73) *Una decisión equivocada puede **poner en juego** la empresa.*
Una decisión equivocada puede **poner en <u>importante</u> **juego** la empresa.*

- Egon al daiteke elementurik aditzaren eta izen-sintagmaren artean? Adberbiorik, adibidez. Bai (74) ala ez (75)?

(74) *Se reunirán para **tomar** una **decisión**.*
Se reunirán para **tomar** <u>al fin</u> una **decisión**.

(75) *Estaba enfermo y **estiró la pata**.*
*?Estaba enfermo y **estiró** <u>desgraciadamente</u> **la pata**.*[3]

- Alda al daiteke konbinazioko osagaien hurrenkera? Bai (76) ala ez (77)?

(76) *La invitada **dio** una **charla** de hora y media.*
*La **charla** que **dio** la invitada duró hora y media.*

(77) ***Estamos en contacto** continuamente.*
**El contacto en que estamos es continuo.*

[3]Adibide hau oso arrotza izanik ere, ez gara ausartu * ikurra eman eta erabat okertzat jotzen. Izan ere, ez dugu inon topatu *estirar desgraciadamente la pata* hitz-segidarik, baina bai oso antzekoak diren beste batzuk, esate baterako, *estirar repentinamente la pata*. Geroago ere hitz egingo dugu ezaugarri honetaz, 4.1.2. atalean –eta 10. oin-oharrean bereziki–.

Hala, anotatzaileek etiketak jarri dizkiete berriro 117 konbinazioei, eta lehen aipatutako hiru taldeetan sailkatu dituzte (anotatze-metodologiari buruzko xehetasun gehiago, datorren azpiatalean). Hona hemen taldeetako bakoitzaren ezaugarri nagusiak eta, gure ustez, zer informazio behar den mota bakoitzeko konbinazioak identifikatzeko.

- **Konbinazio finkoak**[4]. Ez dute batere aldaketa morfosintaktikorik onartzen. Haiek identifikatzeko, nahikoa da hitzak bata bestearen segidan bilatzea, beti hurrenkera berean eta, aditzak salbu, forma berean.

- **Konbinazio erdifinkoak**. Ez dira guztiz finkoak, baina ezta guztiz malguak ere (78. adibidea). Konbinazioak hitz-segida finkotzat bilatuta, agerraldi asko kanpoan utziko lirateke, baina bi hitzak erlazio sintaktiko jakin batean bilatzea bakarrik ere ez da nahikoa. Kasuan-kasuan, informazio linguistiko gehigarria behar da identifikazioa ondo egin ahal izateko.

  (78) *Siempre* ***presta atención***.
  La compañía mejorará la ***atención*** *que* ***presta*** *a sus clientes.*
  Se ***prestará*** *una especial* ***atención*** *a los pacientes más graves.*
  **Siempre prestan atenciones.*

- **Konbinazio malguak**. Beti erlazio sintaktiko jakin batean agertzen dira, baina, horrez gain, ez dute bestelako murriztapen morfosintaktikorik. Beraz, multzo honetako konbinazioak identifika daitezke hitzen lemak erlazio sintaktiko jakin batean bilatuz.

  (79) ***Tiene dificultad*** *para moverse a causa del dolor por la artritis.*
  Varios miembros señalaron las ***dificultades*** *que* ***tienen***.
  Los jóvenes ***tienen*** *una gran* ***dificultad*** *para encontrar trabajo.*

Jarraian emango ditugu anotatze-lanari buruzko datu gehiago.

[4]Multzo honetan ez dugu adibiderik jarri, tesi-lan honetan guztian aztertu ditugun UFetatik bakar bat ere ez zaigulako erabat finkoa iruditu; goiko galderak oinarritzat hartuta, den-denek izan dute hein handiagoan edo txikiagoan aldagarria den ezaugarriren bat.

Anotatze-lana

Hiru kategoriak kontuan hartuta, anotatzaileei konbinazioak sailkatzeko eskatu diegu. Lehen egin bezala, anotatzaile nagusiak 117ko sorta osoa etiketatu du, eta beste biek erdibana egin dute lana: 58 konbinazio batak, eta 59 besteak. Hala, goian zerrendatutako galderak gogoan izanik, 4.3. taula hau sortu dugu, konbinazioak nola sailkatu erabakitzeko.

	Finkoa	**Malgua**
Determinatzailerik?	beti inoiz ez	batzuetan
Determinatzaile mota	zehaztua zehaztugabea	biak
Numeroa	singularra plurala	biak
Modifikatzailerik?	ez	bai
Adi-IS bereizgarria?	ez	bai
Hurrenkera-aldaketarik?	ez	bai

4.3 taula – Eskuzko sailkapen morfosintaktikorako erabaki-taula

Honako argibideok eman dizkiegu anotatzaileei: konbinazio bati dagozkion erantzun guztiak ezkerreko zutabean badaude, konbinazioa guztiz finkotzat sailkatu beharrekoa da; denak eskuineko zutabean badaude, berriz, malgutzat sailkatu beharrekoa; eta zutabe batean zein bestean badaude, erdifinkoen multzoan sartzekoa. Galderei erantzuteko, anotatzaileak euren jakintzaz baliatu dira batez ere, baina sareko corpusetan ere egin dituzte bilaketak. Etiketa guztiak 4.4. taulan jaso ditugu.

		Beste anotatzaileak		
		Finkoa	**Erdifinkoa**	**Malgua**
Anotatzaile nagusia	**Finkoa**	**0**	0	0
	Erdifinkoa	4	**38**	9
	Malgua	0	12	**54**

4.4 taula – Gaztelaniazko anotatze morfosintaktikoan lortutako adostasun-matrizea

Anotatzaileek % 81,34ko adostasuna izan dute ataza honetan, eta 0,61ekoa Cohen κri dagokionez. Bi datu horiei eta 4.4. taulako matrizeari begiratuta, bi ondorio nagusi atera daitezke. Batetik, aditza+izena motako konbinazioak nahiko malguak izan ohi direla morfosintaktikoki, konbinazio bakar bat ere ez baitu anotatzaile batek baino gehiagok finkotzat etiketatu. Bestetik, hiztun batetik bestera intuizioak aldatu egiten direla alderdi morfosintaktikoari dagokionez ere, konbinazioen ia % 19k desadostasunak sortu baitituzte.

Behin ataza hori bukatuta, datuak prestatzen hasi gara, ikusteko ea aztertutako informazioa erabilgarria ote den UFak corpusetan identifikatzeko (4.2. atala). Horretarako, anotatzaile nagusiak erabakiak hartu ditu desadostasuna sortu duten kasuen gainean, eta, ondoren, erdifinko gisa markatutako konbinazio bakoitzeko, 4.3. taulako galderei erantzun die banan-banan. Behin datu horiek bilduta, martxan jarri dugu lehen identifikazio-esperimentua; eman dezagun orain esperimentu horri buruzko xehetasun gehiago.

4.2 Lehen identifikazio-esperimentua, eskuz aztertutako datuak erabiliz

Eskuzko azterketatik ateratako datuek UFen identifikazioa hobetu ote lezaketen ikusteko, esperimentu txiki bat egin dugu. Hiru identifikazio-metodo alderatu ditugu: Freeling analizatzaileak –eta, hortaz, *Matxin*ek– darabilena, eta beste bi, analizatzaile sintaktikoaren emaitzak eskuzko informazioarekin bateratuta. Erabilitako baliabideen eta alderatutako metodoen berri 4.2.1. azpiatalean emango dugu, eta 4.2.2.ean erakutsiko ditugu lortutako emaitzak.

4.2.1 Erabilitako baliabideak eta alderatutako metodoak

Lehen esperimenturako, gaztelaniazko 15.182.385 esaldiko corpus bat erabili dugu: 2013ko WMT itzulpen automatiko estatistikoari buruzko lantegian argitaratutako ingelesa-gaztelania corpus paraleloaren zati bat, zeinak hainbat generotako testuak biltzen baititu. Freeling 3.0 (Padró eta Stanilovsky, 2012) analizatzaileaz baliatu gara, eta UFak identifikatzeko hiru metodo alderatu ditugu: A metodoa, Freeling analizatzaileak –eta *Matxin* itzultzaileak– darabiltena; B metodoa, eskuz aztertutako informazio linguistikoa eta *Freeling*

analizatzaileak sortutako *chunk*[5] automatikoak konbinatzen dituena; C metodoa, eskuz aztertutako informazio linguistikoa eta *Freeling* analizatzaileak sortutako dependentzia sintaktikoak konbinatzen dituena.

A metodoak UFak nola identifikatzen dituen azaldu dugu lehenago ere (3.1. atala): UFak hitz-segida finkoak balira bezala bilatzen ditu, osagai guztiak beti jarraian, hurrenkera berean eta, aditzaren flexioa salbu, forma berean. Azal dezagun orain, oro har, zer egiten duten B eta C metodoek. UF bat morfosintaktikoki malgua bada...

- **B metodoa**k UFko osagaien lemak bilatzen ditu, edozein hurrenkeratan baina tartean gehienez ere chunk bat dutelarik. Tartean chunk bakarra onartzea erabaki dugu, hori baino zurrunago jokatuta agerraldi asko kanpoan utziko genituzkeelako –A metodoarekin gertatzen den bezala– baina, bestetik, zenbat eta hitz gehiago sartu osagaien artean, orduan eta arrisku gehiago dagoelako aditza eta beste osagaiak benetan erlazionatuta ez egoteko.

- **C metodoa**k osagaien lemak bilatzen ditu, dependentzia sintaktikoen zentzu jakin batean: aditzak beti izan behar du izenaren gobernatzailea eta, preposizioak daudenean salbu, ezin da beste osagairik egon bi osagai horien artean dependentzia-zuhaitzean. Hortaz, dependentzien zentzua bai, baina erlazio mota ez dugu kontuan hartzen, bi arrazoirengatik: batetik, analizatzaile sintaktiko automatikoek akats asko egiten dituztelako erlazio horiek ezartzean, eta ez genuelako nahi akats horiek gure atazan eragina izaterik; eta bestetik, UF askotako osagaiak ager daitezkeelako erlazio mota bakarrean baino gehiagotan, bereziki esaldiak egitura kanonikoetatik ez-kanonikoetan aldatzen direnean (pasiboan, inpertsonalean edo izen-sintagmak erlatibozko perpausetan sartzen direnean, adibidez).

Konbinazio erdifinkoak identifikatzeko, berriz, goiko informazio hori erabiltzeaz gain, kasuan kasuko murriztapenak ere gehitzen dituzte. Adibidez, demagun *estar en forma* UFaren hiru agerraldi hauek ditugula.

(80) *Según los médicos, ahora* ***está en forma*** *de nuevo.*

(81) *Según los médicos, ahora* ***está en*** *buena* ***forma*** *de nuevo.*

[5]Gogoratu, 2.2.4.1. orrialdean esan dugunez, HPren alorrean *chunk* esaten zaiela sintagma gisako hitz multzoei. Informazio gehiago, Aranzaberen doktoretza-tesian (2008).

(82) ***Está*** *ahora, según los médicos, de nuevo* ***en*** *buena* ***forma****.*

Eskuz aztertutako informazioaren arabera, UF hori erdifinkoa da: izen-sintagmak ez du determinatzailerik izaten eta beti singularrean egoten da, baina modifikatzaileak onartzen ditu; aditza eta izen-sintagma bereiz daitezke, baina hitz-hurrenkera ez da inoiz aldatzen. Hori gogoan izanda, alderatu ditugun metodoek honela jokatuko lukete hiru adibideekin.

- A metodoak, hitz-segida finkoak bilatzen dituenez, 80. adibidean baino ez luke UFa topatuko, bigarren eta hirugarrenean UFko osagaiak beste hitz batzuez bereizita agertzen direlako.

- B metodoak, 80. adibidekoaz gain, 81.eko agerraldia ere identifikatuko luke, UFaren izen-sintagman modifikatzaileak ager daitezkeenez, onartzen duelako *forma* izena barne hartzen duen chunkean beste hitz batzuk ere egotea. Hirugarren esaldian (82. adibidea), ordea, ez luke UFrik identifikatuko, aditzaren eta gainerako hitzen artean chunk bateko muga daukalako ezarrita eta, kasu honetan, hiru ageri direlako:

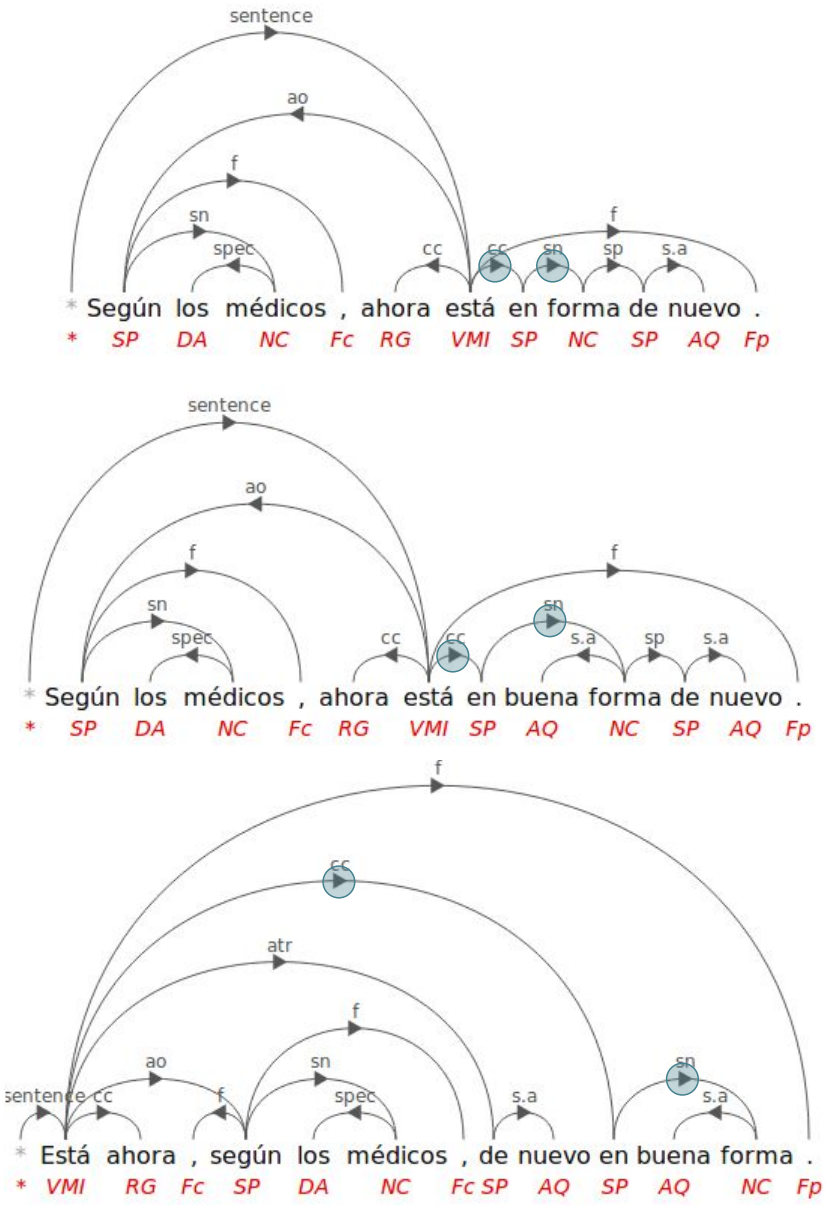
  [**está**] **[**ahora**]** [,] **[**según los médicos**]** [,] **[**de nuevo**]** [**en** buena **forma**]

- C metodoa gai izango litzateke hiru agerraldiak identifikatzeko, hirurek betetzen baitituzte beharrezko baldintzak: *forma* izena singularrean eta determinatzailerik gabe ageri da; hitz-hurrenkera kanonikoa da; *forma* izenaren gobernatzailea *en* preposizioa da, eta preposizioarena, *estar* aditza (4.2. irudia).

Bestalde, UF bateko hitzak esaldi batean agertzen direnean baina ez direnean benetan UF baten parte, nahiz eta hitz horien tartean bi chunk edo gutxiago izan eta dependentzia-zentzu berberean egon, proposatutako bi metodoak gai dira UFrik ez identifikatzeko, baldin eta morfosintaxian badatza UFen eta UF ez direnen arteko aldea. Esate baterako, 83. adibidean, izen-sintagman *una* determinatzailea agertzen denez, B eta C metodoek ez dute *estar en forma* identifikatzen, eskuzko murriztapenetan zehazten baita UF horrek ez duela determinatzailerik izaten.

(83) *La información está escrita en una forma fácilmente comprensible.*

Jakina, kasu guztiak ez dira halakoak, eta, batzuetan (84. adibidea), morfosintaxia baino gehiago behar da hitz-konbinazio jakin bat UFa den ala ez bereizteko.



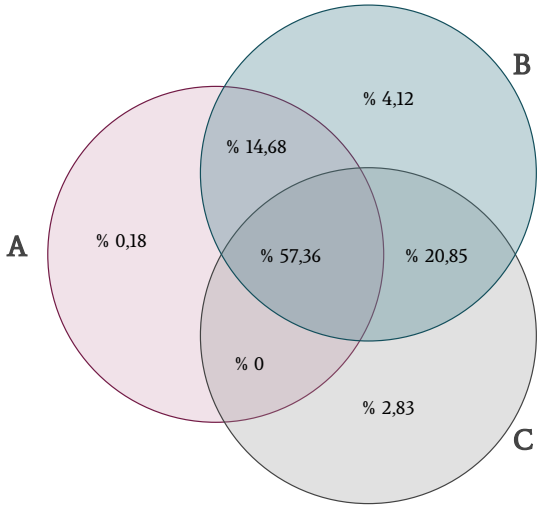
4.2 irudia – *Estar en forma* UFa duten hiru esaldiren dependentzia-analisia, Freeling 4.1en arabera. Hiru hitzen arteko dependentzia sintaktikoen zentzua biribilduta dago.

(84) *Los informes pueden ser de papel o estar en forma electrónica.*

Lehenago ere aipatu dugunez, 80. adibidetik 83.era arteko kasuak eta antzekoak ebaztea da gure asmoa, eta ez 84. adibidearen gisakoak. Hala, proposatutako bi metodoek ezaugarri morfosintaktikoak dituzte oinarritzat, eta ez semantikoak. Ikus dezagun orain kontuan hartu ditugun ezaugarri morfosintaktikoek zer-nolako ekarpena egiten duten UFen identifikazioan.

4.2.2 Emaitzak

Hiru metodoen artean, 433.092 konbinazio identifikatu dira UFtzat guztira, eskuzko azterketan landutako 117 konbinazioetatik abiatuta. Horietatik, ia % 28 izan dira A metodoak hauteman gabeak, hau da, gure metodo-proposamenei esker identifikatuak (4.3. irudia).



4.3 irudia – Lehen identifikazio-esperimentuaren emaitzak: identifikatutako konbinazioen ehunekoak, metodoaren arabera

Doitasuna kalkulatzeko, multzo bakoitzeko lagin adierazgarri bana ebaluatzeko eskatu diegu bi hizkuntzalariri. 4.5. taulan, erdiko zutabeak erakusten du identifikatutako konbinazio guztietatik zer ehuneko identifikatu duen metodo bakoitzak, eta eskuinekoak, berriz, zer doitasunez identifikatu diren identifikatutako konbinazioak.

	Identifikatutako UFak	**Doitasuna**
A metodoa (guztira)	% 72,20	0,99
B+C metodoak (A gabe)	% 20,85	0,97
B bakarrik	% 4,12	0,93
C bakarrik	% 2,83	0,83

4.5 taula – Lehen identifikazio-esperimentuaren emaitzak

Taula horretan ikusten denez, proposatutako metodoen doitasuna ez da Freelingen doitasuna bezain altua, hura ia erabatekoa baita. Dena dela, bi metodo berrien emaitzak ere oso onak dira: identifikatutako konbinazio guztiak kontuan hartuta, 0,98 inguruko doitasuna dute[6], eta A metodoak identifikatzen dituenak alde batera utzita ere, 0,95ekoa.

Bestalde, 4.3. irudian ikus daiteke A metodoa gai izan dela B eta C metodoek identifikatu gabeko konbinazio gutxi batzuk identifikatzeko. Horiek ere banan-banan aztertu ditugu, eta denak dira analisi sintaktiko okerren ondorioz sorturikoak. Hala ere, gure asmoa hiru metodoak batera erabiltzea denez, B eta C metodoek agerraldiren bat kanpoan uzteak ez luke arazorik ekarriko.

Esperimentu hau egin dugunean, oraindik ez dugu gaztelaniazko corpus etiketaturik izan eskura, eta ezin izan dugu estaldurarik eta F neurririk kalkulatu. Nolanahi ere, lan hau proba moduko bat izan da, eta bete du bere egitekoa: egiaztatu dugu aztertutako informazioa baliagarria izan daitekeela UFak identifikatzeko, oso doitasun ona lortzen baitu eta, estaldura zehazki zenbatekoa den jakin gabe ere, bai baitakigu lehen geneukana baino hobea dela.

Gainera, PARSEMEren ataza partekatuko gaztelaniazko corpus etiketatuari begiratuta (Savary *et al.*, 2017: 6. taula), bada gure estaldurarekin lotzeko moduko datu interesgarri bat. Etiketatutako aditz-UF guztietatik, % 70 agertzen dira era jarraituan corpusean, eta gainerako % 30ak beste hitzen bat du UFko osagaien artean. Datu hori pareka liteke nolabait gure estaldurarekin, B eta C metodoek UFen % 28 agerraldi gehiago identifikatu baitituzte eta ehuneko horretan sartzen diren agerraldi gehienak ez-jarraituak baitira hain zuzen. Dena dela, aintzat hartu behar da PARSEMEren corpu-

[6]Emaitza orokor hori lortzeko, metodo bakoitzaren doitasuna identifikatutako UFen ehunekoekin biderkatu, eta dena batu dugu.

sean izena+aditza motakoak ez diren beste UF batzuk ere badaudela etiketatuta eta, hortaz, ez-jarraituak diren % 30 horietan den-denak ez datozela bat gure aztergaiarekin.

Momentuz, ezin dugu gehiagorik esan estaldurari buruz; geroago zabalduko eta hobetuko dugu hasierako proposamen hau (4.4. atala), eta lortuko dugu neurri hori ere. Doitasuna, ordea, alderatu daiteke nolabait beste lan batzuekin, eta PARSEMEren ataza partekatuko emaitzetara joko dugu horretarako. Kontuan hartu behar da emaitza horiek eta gureak ez direla erabat alderagarriak, baina laguntzen dute gure emaitzak beste batzuen artean kokatzen.

Lehen edizioan (Savary *et al.*, 2017), bost sistemak hartu dute parte gaztelaniazko atalean; haien doitasunak 4.6. taulan laburbildu ditugu, hizkuntza erromanikoen multzo osoan izandakoekin batera. Bai gaztelaniaz eta bai hizkuntza erromanikoen multzo osoan, TRANSITION sistemak (Al Saied *et al.*, 2017) lortu ditu emaitzarik onenak oro har, hau da, F neurriari dagokionez. Esan beharra dago, dena den, doitasun-markarik altuena ez dela sistema horrena izan eta, aldiz, doitasunik oneneko sistemak, RACAIk (Boros *et al.*, 2017), emaitzarik okerrenak lortu dituela F neurrian bost parte-hartzaileetatik. Horrez gain, aipagarria da gaztelaniazko batez besteko doitasun-marka –bai eta estaldura eta F neurria ere– 8 puntu beherago dagoela hizkuntza erromanikoen batezbestekotik.

ES	0,50 (0,26-0,64)
H. erromanikoak	0,58 (0,06-0,87)

4.6 taula – PARSEMEren ataza partekatuko lehen edizioan, gaztelaniaz eta hizkuntza erromanikoen multzoan oro har izandako doitasun-markak

Bigarren edizioan (Ramisch *et al.*, 2018), berriz, doitasun-markarik onena erdietsi duten sistemak izan dira F neurriari dagokionez ere mailarik altuenekoak: TRAVERSAL sistema hizkuntza guztiak kontuan hartuta (Waszczuk, 2018), eta TRAPACC_S gaztelaniazko atalean (Stodden *et al.*, 2018). Horrez gain, bigarren edizioan lehenengoan baino are nabarmenagoa da gaztelaniazko emaitzen eta batez bestekoen arteko aldea: 17 puntukoa (4.7. taula).

Laburbilduz, bi ondorio atera daitezke esperimentu honetatik. Batetik, gure metodoaren doitasuna oso altua dela, aipatutako marka guztietatik bakar bat ere ez baita guk lortutakoa bezain ona. Eta, bestetik, badirudiela

ES	0,19 (0,00-0,32)
H. guztiak	0,36 (0,00-0,68)

4.7 taula – PARSEMEren ataza partekatuko bigarren edizioan, gaztelaniaz eta hizkuntza guztietan oro har izandako doitasun-markak

gaztelaniazko UFak bereziki zailak izan direla identifikatzeko PARSEMEren ataza partekatuetan, emaitza nabarmen kaskarragoak lortu baitira beste hizkuntzen aldean. Horren arrazoia ez dago argi: baliteke zailtasun batzuk gaztelaniaren ezaugarriei zor izatea, baina, ziur asko, corpuseko etiketen kalitateak ere eragin handia izango zuen.

Jakina, gure proposamenaren estaldura zehazki kalkulatu ez badugu ere, badakigu alderdi horretan oso motz geratuko ginatekeela orain arte egindakoarekin bakarrik: hiztegietako konbinazioak hartzen ditugunez oinarritzat, orain arteko bideak bakarrik erabilita ezingo genuke hiztegian jaso gabeko UFrik inola ere identifikatu, eta halakoak ez dira gutxi izaten testuetan. Esperimentu honetan, corpusa berariaz prestatu dugu guk landutako UFentzat, baina ekar dezagun gogora lehenago ere eman dugun datu erakusgarri bat: PARSEMEren ataza partekatuko corpusean, *Elhuyar* hiztegiko konbinazioen 53 agerraldi agertzen dira guztira, eta hiztegian jaso gabeko UFen beste 609. Hau da, UF gehiago landu ezean, corpus orokor horretan lor genezakeen estaldurarik onena 0,08koa litzateke.

Ahulgune hori kontuan harturik baina, era berean, metodoa baliagarria izan daitekeela baieztatu ondoren, erronka berri bat dugu orain: nola zabal dezakegu egindako lana, eskala handiagoan erabili ahal izan dadin? Gai horri ere heldu diogu, eta 4.4. atalean azalduko dugu nola. Horren aurretik, ordea, ikus dezagun orain arteko azterketa nola aplika daitekeen ingelesera, gure proposamena beste hizkuntza batzuetan ere erabilgarria izan litekeela erakusteko.

4.3 Azterketa xehearen ingeleserako aplikagarritasuna

Ingelesez ere, gaztelaniaz eginikoa errepikatu dugu: batetik, hiztegi batetik erauzitako hitz-konbinaziorik usuenak sailkatu ditugu ezaugarri lexiko-

semantikoen eta morfosintaktikoen arabera, eta, bestetik, informazio linguistikoa baliatu dugu UFak corpusean identifikatzen laguntzeko. Atal honetan azalduko dugu nola egin ditugun anotatze-lana eta identifikazio-esperimentua, eta emaitzak gaztelaniazkoekin alderatuko ditugu.

4.3.1 Ingelesezko UFen azterketa

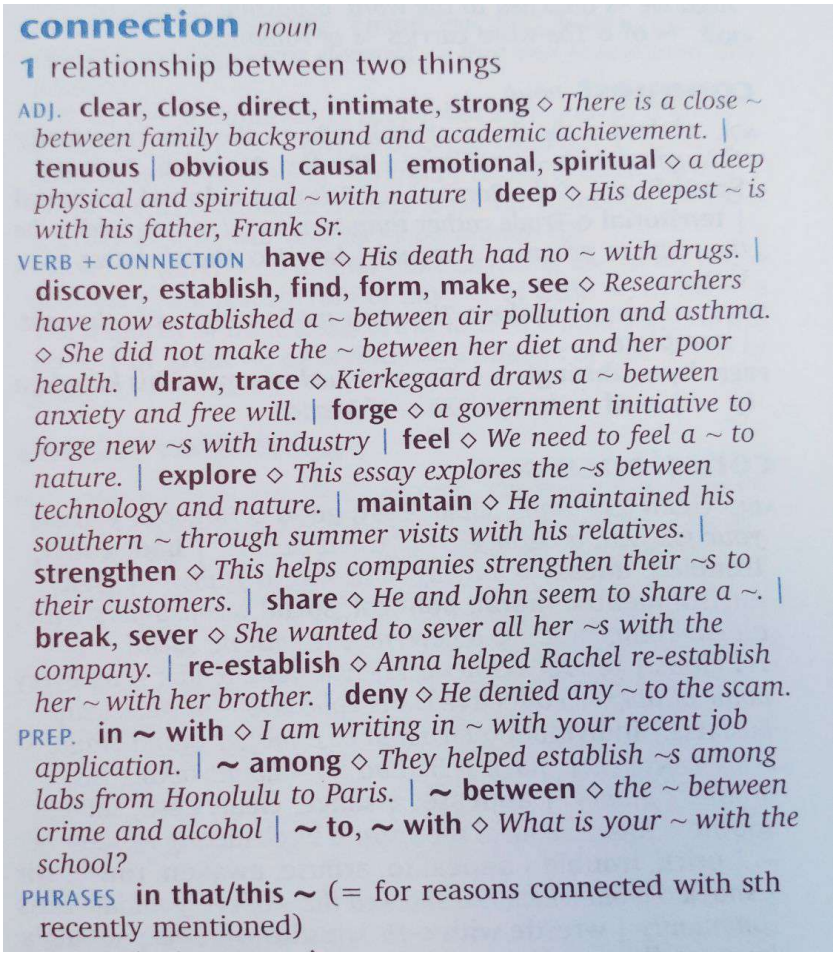
Azterketari ekin ahal izateko, hitz-konbinazioen zerrenda osatu dugu lehenik. Gure hasierako asmoa *Elhuyar*ren ingelesa-euskara hiztegia erabiltzea zen, gaztelaniazko azterketatik ahalik eta gutxien urruntzeko, baina beste hiztegi batera jotzea erabaki dugu azkenean, ikusirik *Elhuyar*renean hitz batez baino gehiagoz osatutako oso sarrera gutxi zeudela. *Oxford Collocations Dictionary* (McIntosh, 2009) erabili dugu, ingelesezko hiztegi fraseologiko bat. Anotatze-lanari buruzko xehetasunak eman aurretik, deskriba dezagun hiztegi hori labur, eta azal dezagun nola aukeratu ditugun ondorengo lanerako hitz-konbinazioak.

Hitz-konbinazioen aukeraketa

Oxford Collocations Dictionary, izenburuak agerian uzten duenez, hiztegi fraseologikoa da, eta kolokazioak biltzen ditu oro har. Hain zuzen, hiztegiaren hitzaurrean, esaten da konbinazio libreak alde batera uzten direla, eta lokuziorik ere ez sartzea izan dela irizpide orokorra. Dena dela, guretzat garrantzitsuak diren bi ohar ere egiten dira: (1) praktikan, lokuzio batzuk ere onartu direla hiztegian, hein batean baino ez baziren idiomatikoak edo, bestela esanda, erabat opakoak ez baziren –alegia, gure lokuzio metaforikoak eta gisa horretakoak–; eta (2) askotariko kolokazioak hartu direla kontuan, maila idiomatiko txikikoetatik hasi eta oso idiomatikoak direnetaraino. Hiztegia pixka bat aztertuta, ikusi dugu "maila idiomatiko txikikotzat" jotzen dituzten horietako asko guk ez genituzkeela kolokaziotzat sailkatuko, baizik eta konbinazio libretzat. Hortaz, desberdintasunak desberdintasun, iturri egokia iruditu zaigu gure azterketarako, baita sailkapen lexiko-semantikorako ere, izenburuak bestela iradoki lezakeen arren.

Hori horrela izanik, sarrerak izen, aditz edo adjektibo hutsak dira, eta sarreren azpian biltzen dira hitz nagusi horrekin batera erabili ohi diren *kolokatuak*, hau da, sarrera-hitzarekin batera hitz-konbinazioa osatzen duten beste osagaiak. Ondoren, kolokatu horiek gramatika-kategoriaren arabera

multzokatzen dira sarrera bakoitzean. Ikus sarrera baten adibidea 4.4. irudian.



connection *noun*

1 relationship between two things

ADJ. **clear, close, direct, intimate, strong** ◇ *There is a close ~ between family background and academic achievement.* | **tenuous** | **obvious** | **causal** | **emotional, spiritual** ◇ *a deep physical and spiritual ~ with nature* | **deep** ◇ *His deepest ~ is with his father, Frank Sr.*

VERB + CONNECTION **have** ◇ *His death had no ~ with drugs.* | **discover, establish, find, form, make, see** ◇ *Researchers have now established a ~ between air pollution and asthma.* ◇ *She did not make the ~ between her diet and her poor health.* | **draw, trace** ◇ *Kierkegaard draws a ~ between anxiety and free will.* | **forge** ◇ *a government initiative to forge new ~s with industry* | **feel** ◇ *We need to feel a ~ to nature.* | **explore** ◇ *This essay explores the ~s between technology and nature.* | **maintain** ◇ *He maintained his southern ~ through summer visits with his relatives.* | **strengthen** ◇ *This helps companies strengthen their ~s to their customers.* | **share** ◇ *He and John seem to share a ~.* | **break, sever** ◇ *She wanted to sever all her ~s with the company.* | **re-establish** ◇ *Anna helped Rachel re-establish her ~ with her brother.* | **deny** ◇ *He denied any ~ to the scam.*

PREP. **in ~ with** ◇ *I am writing in ~ with your recent job application.* | **~ among** ◇ *They helped establish ~s among labs from Honolulu to Paris.* | **~ between** ◇ *the ~ between crime and alcohol* | **~ to, ~ with** ◇ *What is your ~ with the school?*

PHRASES **in that/this ~** (= for reasons connected with sth recently mentioned)

4.4 irudia – Oxford Collocations Dictionary-ko sarrera baten adibidea: *connection* izenaren lehen adiera.

Gure kasuan, izen motako sarreretara jo, eta bi multzori begiratu diogu, VERB+*gakoa* edo *gakoa*+VERB motakoei, hala baitaude jasota gure aztergaiarekin bat datozen hitz-konbinazioak, bai aditza+izena motakoak eta bai aditzaren eta izenaren artean preposizioa edota determinatzailea dutenak ere. Halako guztiak erauzi ditugu lehenik, eta hitz-konbinazio bakoitzaren maiztasuna kalkulatu dugu ondoren, British National Corpus-en (Burnard, 2007) oinarrituta: ondoz ondoko chunketan bilatu ditugu aditza eta izena

(eta, zegokionean, preposizioa), eta agerraldiak kontatu ditugu. Azkenik, gaztelaniazkoaren antzeko tamainako zerrenda sortu nahi genuenez, 500 agerralditik gorako hitz-konbinazioak hautatu ditugu, eta 173ko sortarekin egin dugu aurrera.

Behin zerrenda osatuta, anotatze-lanari ekin diogu, lexiko-semantika kontuan harturik lehenik, eta morfosintaxiari begiratuz ondoren.

Anotatze lexiko-semantikoa

Erabili dugun sailkapen lexiko-semantikoa gaztelaniazko berbera denez, ez ditugu multzoen inguruko azalpenak berriro errepikatuko, 4.1.1. atalean baitaude. Dena dela, ekar dezagun gogora zer lau multzo erabili ditugun, eta eman dezagun ingelesezko adibide bana.

- **Lokuzio opakoak**

  (85) EN: *Do not believe her, she is just **pulling** your **leg**.*
  EU: *Ez sinetsi, **adarra jotzen** baino ez zaizu ari.*

- **Lokuzio metaforikoak**

  (86) EN: *If you need a good teacher, she **is** your **woman**.*
  EU: *Irakasle on bat behar baduzu, bera da egokiena.*

- **Kolokazioak**

  (87) EN: *Volunteers are needed to **give support** to the organization.*
  EU: *Boluntarioak behar dira erakundeari **laguntza ematekо**.*

- **Konbinazio libreak**

  (88) EN: *They are using a new technique now.*
  EU: *Teknika berri bat erabiltzen ari dira orain.*

Gaztelaniazkoan bezala, ataza honetan ere hiru anotatzailek hartu dute parte, ingelesaren ezagutza maila altua duten hiru hizkuntzalarik. Lehengo

bideari jarraituz, anotatzaile nagusiak 173 konbinazioak sailkatu ditu, eta beste biek erdibana egin dute lana: batek 86 konbinazio, eta besteak 85.

Adostasuna gaztelaniaz baino txikiagoa izan da, bai ehunekotan eta bai, bereziki, Cohen κ-ri begiratuta (4.8. taula). Hain zuzen ere, azken neurri horri dagokionez, gaztelaniazko emaitzak baino 8 puntu baxuagoak dira ingelesezkoak, PARSEMEren 2018ko batezbestekoa baino 12 puntu baxuagoak eta, ataza horretako ingelesezko corpusari bakarrik begiratuta, 23 puntu baxuagoak; Landis eta Kochen arabera (1977), *moderate agreement* multzoan leudeke. Bat-etortzeari ehunekotan begiratuta, aldiz, ez dago hainbesteko alderik gaztelaniazko emaitzekin (4-5 puntu), eta esan liteke adostasun onargarria lortu dela.

Dena dela, berriz ere, gogora ekarri behar da PARSEMEren corpusean askotariko osaerako aditz-UFak daudela etiketatuta, ez aditza+izena motakoak bakarrik, eta euren sailkapena ez datorrela guztiz bat gurearekin, ez izaeran eta ez kopurutan; sei multzo bereizten dituzte, baina horietako hirutan bakarrik sartzen dira aditza+izena motakoak. Hortaz, bistan da emaitzak ez direla erabat alderagarriak. Bestetik, PARSEMEren corpusaren anotatzeprozesuan ama-hizkuntzatzat ingelesa duten adituek parte hartu dute, eta ezaugarri horri ere zor izango zaio, ziur asko, halako desberdintasuna izatea bai gaztelaniazko emaitzen eta bai PARSEMEren corpusekoen aldean.

	4 kat.	**3 kat.**	**2 kat.**
Bat-etortzea	% 0,71	% 0,71	% 0,78
Cohen κ	0,55	0,56	0,50

4.8 taula – Ingelesezko anotatze lexiko-semantikoan lortutako adostasuna

Adostasun-matrizea (4.9. taula), berriz, gaztelaniazkoaren oso antzekoa da distribuzioari dagokionez. Kolokazioen multzoa da nabarmen handiena, konbinazio libreak eta lokuzio metaforioak dezente gutxiago dira, eta anotatzaile bakarrak erabili du lokuzio opakoen etiketa, behin bakarrik. Gainera, desadostasun-kopurua altuagoa bada ere, agerian geratzen da beste behin ere idiomatikotasunak *continuum* bat osatzen duela, ondoz ondoko multzoetan sailkatu baitira, hiru kasutan izan ezik, desadostasun-iturri izan diren hitz-konbinazio guztiak.

		Beste anotatzaileak			
		Lok.Op.	**Lok.Met.**	**Kolokazioa**	**Librea**
Anotatzaile nagusia	**Lok.Op.**	**0**	0	0	0
	Lok.Met.	1	**24**	0	1
	Kolokazioa	0	12	**73**	22
	Librea	0	2	13	**25**

4.9 taula – Ingelesezko anotatze lexiko-semantikoan lortutako adostasun-matrizea

Anotatze morfosintaktikoa

Morfosintaxiari dagokionez ere, 4.1.2. atalean azaldutako hiru multzoei eutsi diegu. Konbinazioak multzo horietan sailkatzeko eskatu diegu anotatzaileei, multzoen inguruko azalpenez eta 109. orrialdeko 4.3. erabaki-taulaz lagunduta. Hona hemen hiru multzoak eta, dagokionean, adibide bana[7].

- **Konbinazio finkoak**
- **Konbinazio erdifinkoak**

  (89) *She **is in love**.*
  *She **is** always **in love**.*
  **She is in the love.*

- **Konbinazio malguak**

  (90) *They are **making money** with their new business.*
  *They are **making** lots of **money**.*
  *Do not waste the **money** you are **making**!*

Kasu honetan ere, adostasuna gaztelaniazkoa baino baxuagoa izan da Cohen κ-ri dagokionez, 0,56koa (0,61en aldean). Bat-etortzea ehunekotan kalkulatuta, aldiz, gaztelaniazkoa baino hobea da emaitza, % 85ekoa (% 81en aldean). Aldeak alde, gaztelaniazko atazatik ateratako ondorioak berresten dira hemen ere: konbinazioak morfosintaktikoki sailkatzeko atazak objektiboagoa dirudien arren, erabilerarekiko intuizioa aldakorra dela hiztun batetik bestera.

[7]Gaztelaniaz gertatu zaigun bezala, ingelesezko multzoan ere ez dugu erabat finkotzat sailkatzeko konbinaziorik topatu.

		Beste anotatzaileak		
		Finkoa	**Erdifinkoa**	**Malgua**
	Finkoa	**0**	0	0
Anotatzaile	**Erdifinkoa**	2	**24**	10
nagusia	**Malgua**	0	14	**123**

4.10 taula – Ingelesezko anotatze morfosintaktikoan lortutako adostasun-matrizea

Adostasun-matrizeari erreparatuta, ikus daiteke konbinazio malguen kopurua oso-oso altua dela, gaztelaniazkoen artean baino askoz ere altuagoa.

4.3.2 Ingelesezko identifikazio-esperimentua

Identifikazio-esperimenturako, gaztelaniaz bezala, sailkapen lexiko-semantikoan UFtzat hartutako hitz-konbinazioak erabili ditugu, 133 UF guztira. Anotatzaile nagusiak banan-banan zehaztu dizkie aldakortasun morfosintaktikoari buruzko ezaugarriak, 4.1.2. ataleko galderei erantzunez (105. orrialdea), eta datu horiez lagunduta egin dugu esperimentua. Azal dezagun orain zer metodologia erabili dugun identifikazio-lanerako, eta eman dezagun emaitzen berri.

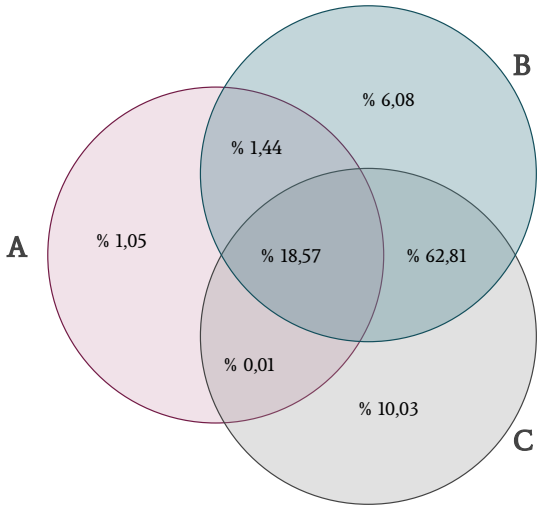
4.3.2.1 Erabilitako baliabideak eta alderatutako metodoak

Esperimenturako erabili dugun corpusak British National Corpus du izena (Burnard, 2007), eta domeinu orokorreko testuak biltzen ditu, 100 milioi hitz guztira. Corpus hori analizatzeko, berriz, Stanford CoreNLP analizatzailea erabili dugu (Manning *et al.*, 2014), eta hala lortu dugu chunkei eta dependentzia-zuhaitzei buruzko informazioa.

Alderatutako metodoak gaztelaniaz alderatutako berberak izan dira: (A) hitzak karaktere-segidatzat hartzen dituen oinarrizko metodo bat, aditzaren flexioaz gain inolako aldaketarik onartzen ez duena; (B) UFei eskuz zehaztutako datu morfosintaktikoak eta chunken inguruko informazioa uztartzen dituen metodo bat; eta (C) UFei eskuz zehaztutako datu morfosintaktikoak eta automatikoki analizatutako dependentzia sintaktikoak darabiltzan metodo bat. Haien funtzionamenduari buruzko xehetasun gehiago 4.2.1. atalean daude jasota, gaztelaniazko esperimentuaren inguruko azalpenetan.

4.3.2.2 Emaitzak

Landutako 133 UFak oinarritzat harturik, 152.051 agerpen identifikatu dira hiru metodoak konbinatuta, eta horietako % 79 ez ditu A metodoak ezagutu, hau da, informazio linguistiko xehea darabilten bi metodoei esker identifikatu dira. Metodo bakoitzak identifikatutako agerpenen ehunekoak 4.5. irudian eta 4.11. taulan jaso ditugu, eta doitasuna ere zehaztu dugu taulan.



4.5 irudia – Ingelesezko identifikazio-esperimentuaren emaitzak: identifikatutako konbinazioen ehunekoak, metodoaren arabera

	Identifikatutako UFak	**Doitasuna**
A metodoa (guztira)	% 21,08	0,99
B+C metodoak (A gabe)	% 62,81	0,96
B bakarrik	% 6,08	0,70
C bakarrik	% 10,03	0,79

4.11 taula – Ingelesezko identifikazio-esperimentuaren emaitzak

Emaitzek argi erakusten dute informazio linguistiko xehea baliagarria dela identifikazio-lanerako, oinarrizko metodoak agerpen guztien % 21 baino ez baititu identifikatzen. Beraz, hobekuntza gaztelaniazkoa baino askoz ere

nabarmenagoa da, B eta C metodoek identifikatutako agerpen gehigarriak % 79 baitira ingelesez, eta gaztelaniaz % 28 bakarrik, ingelesezkoen herena pasatxo.

Bi arrazoi nagusirengatik dago halako aldea, ziur asko. Batetik –eta batez ere–, hitz-konbinazioak biltzeko erabili ditugun iturriek eta metodologiak eragina izango zuten ziur asko. Izan ere, lehen ere aipatu dugu gaztelaniazko eta ingelesezko hiztegiak oso desberdinak zirela, helburu desberdinetara bideratuak, eta konbinazioen maiztasuna kalkulatzeko modua ere ez da berdina izan batean eta bestean. Gaztelaniazko hiztegiko sarrerak forma jakin batean zeudenez –adibidez, *dar pasos* pluralean, eta *dar un repaso* artikulu zehaztugabearekin–, agerraldiak kontatzean konbinazioko osagaiak bata bestearen jarraian bilatu ditugu beti, eta aditza ez beste osagai guztiak forma berean. Ingelesezko hiztegian, berriz, sarrerak eta azpisarrerak lemen arabera daudenez antolatuta, edozein formatan bilatu ditugu izen eta aditz horien lemak, eta ez beti elkarren jarraian baizik eta ondoz ondoko chunken barruan. Horrek esan nahi du, maiztasunak kalkulatzetik bertatik, malgutasun gehiagoz jokatu dugula ingelesez, eta baliteke horregatik izatea malguagoak hizkuntza horretako konbinazioak gure datuen arabera.

Bestetik, hizkuntza baten eta bestearen ezaugarriek ere izango zuten eragina beharbada. Esate baterako, adjektiboek izen-sintagman izan ohi duten kokalekua ez da berdina gaztelaniaz eta ingelesez: gaztelaniaz izenaren alde batean zein bestean joan daitezke (91. adibidea), baina ingelesez aurretik bakarrik (92. adibidea). Hori horrela izanik, baliteke A metodoa eraginkorragoa izatea gaztelaniaz, eta horregatik egotea alde txikiagoa beste metodoekiko.

(91) ***dar*** *importantes* ***pasos***
dar pasos *importantes*

(92) ***take*** *important* ***steps***
****take steps*** *important*

Doitasun-emaitzarik okerrena B metodoak lortu du. Erroreen azaleko analisi bat eginda, ikusi dugu errore asko eta asko datozela aditz arindun konbinazioetatik, bereziki *have* aditza dutenetatik, aditz hori laguntzaile ere izan baitaiteke. Honako esaldi honetan, adibidez, *have influences* UFa identifikatu da, nahiz eta *have* egiaz *have been likened* aditz konposatuaren parte izan testuinguru horretan.

(93) *These influences have also been likened to the forces effected by a millenarian journey to a new faith...*

Emaitzak beste lan batzuekin alderatzeko, PARSEMEren ataza partekatura joko dugu berriro ere (Ramisch *et al.*, 2018), eta, estaldura zehazki kalkulatzerik ez dugunez, doitasunari begiratuko diogu, lehen bezala (4.12. taula).

EN	0,31 (0,01–0,59)
H. guztiak	0,36 (0,00–0,68)

4.12 taula – PARSEMEren ataza partekatuko bigarren edizioan, ingelesez eta hizkuntza guztietan oro har izandako doitasun-markak

Aurkeztu diren 12 sistemen doitasun-marketatik altuena 0,59koa izan da (Pasquer *et al.*, 2018), eta batezbestekoa 0,31koa. Hortaz, ingelesez ere marka baxu samarrak lortu dira hizkuntza guztien batezbestekoaren aldean, gaztelaniaz bezala, eta bat ere ez da geure doitasun-marka baino hobea. Agerian geratzen da, beraz, geure metodoak ekarpena egiten duela UFen identifikazioan, oso doitasun altua lortzen baitu eta, estaldurari buruzko datu zehatzik ez badugu ere, agerpen gehigarri asko ezagutzea lortzen baitu.

4.4 Identifikaziorako azterketa erdiautomatizatzeko proposamena

Orain arte eginiko azterketa erdiautomatizatzeko, lehenik eta behin, UF-zerrenda bat sortu dugu. Bi iturri erabili ditugu horretarako: batetik, *Elhuyar* hiztegia, orain arte erabili dugun berbera, eta bestetik, gaztelaniazko kolokazioen *DiCE* hiztegia (Vincze *et al.*, 2011). *Elhuyar*retik, 3. kapituluko azterketarako erauzi ditugun 1.205 hiztegi-sarrerez baliatu gara. *DiCE*tik, berriz, aditza+izena motako beste 4.504 konbinazio ere erauzi ditugu, eta maiztasunik altueneko 500ak aukeratu ditugu azterketa honetarako.

Kontuan izan behar da, dena den, bi iturri horiek oso desberdinak direla eta, hortaz, behar duten tratamendua ere ez dela beti berdin-berdina. Izan ere, batetik, *DiCE*k gaztelaniazko kolokazioak bakarrik biltzen ditu, eta *Elhuyar* hiztegiak, oro har, lokuzioak. Eta, bestetik, *Elhuyar* hiztegian ez bezala, *DiCE*n hiztegi-sarrerak izen hutsak dira; aditzak –eta gainerako kolokatuak– sarrera bakoitzaren barruan daude gordeta, eta ez da zehazten

izenaren eta aditzaren artean zer beste elementu egon ohi den[8] –hots, zer preposizio edota determinatzaile–.

Beraz, hainbat neurri hartu ditugu ziurtatzeko iturri bata zein bestea zirela bateragarriak azterketa honetarako. Lehenik eta behin, *DiCE*ko 500 kolokazioen artean lehen 22ak aukeratu ditugu[9], eta eskuz aztertu ditugu, 4.1.2. atalean bezalaxe, egiaztatzeko zehaztutako irizpideak egokiak ote ziren *DiCE*ko konbinazioentzat ere. Baietz ikusi dugunez, aurrera egin dugu, eta moldaketa gehiago ere egin ditugu azterketa erdiautomatikoan zehar, iturriaren arabera. Eskuz aztertutako 22 konbinazioak eta *Elhuyar*ren ere jasotakoak baztertuta, *DiCE*tik erauzitako konbinazioekin 437ko zerrenda osatu dugu azkenean.

Oro har, behin UF-hautagaiak lortuta, sei urratsetan banatu dugu azterketa linguistikoa erdiautomatizatzeko prozesua (4.6. irudia). Lehenengotik hirugarrenera arteko urratsak –irudian nabarmendutako hirurak– dira atal honi dagozkionak, identifikazioa hobetzeko informazioa erauztera bideratuak. Beste hiruren inguruan, berriz, datorren kapituluan hitz egingo dugu, 5.3. atalean.

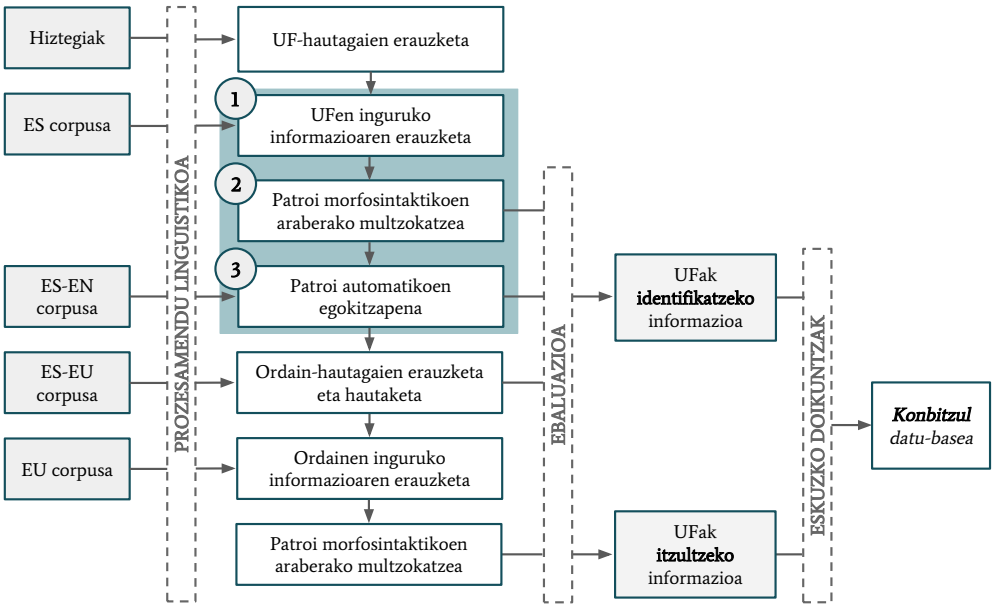
Azpiatal bana eskainiko diegu orain 4.6. irudian markatutako hiru urratsei. Azalpenak eta adibideak emango ditugu, eta, hala dagokionean, lortutako emaitzak ere erakutsiko ditugu.

4.4.1 Informazio-bilketa corpus elebakarretatik (1. urratsa)

Behin UF-hautagaien zerrenda osatuta, corpus elebakarretan nola erabiltzen diren begiratu dugu. Gaztelaniazko 15 milioi esaldiko corpus bat hartu dugu, hainbat generotako testuak biltzen dituena, eta, oraingoan ere, Freeling 3.0 (Padró eta Stanilovsky, 2012) erabili dugu testuak analizatzeko. Jarraian, hautagai bakoitzeko aditzen eta izenen lemak bilatu ditugu, aditza izenaren buru sintaktiko deneko kasuetan.

[8]Berez, *DiCE*ren sareko kontsultetan badago modua osagaien arteko preposizioak ikusteko, baina guk, lan hau egin dugunean, ez dugu informazio hori eskura izan.

[9]*DiCE*ko konbinazioak maiztasunen arabera ordenatuta, iruditu zaigu 8.000 agerralditik gorakoak hartzea zela aproposena ataza honetarako, atalasea hortik behera jarriz gero nabarmen igotzen baitzen aztergaien kopurua. Nahikoa iruditu zaigu 22 konbinazioko lagina osatzea, jakin nahi baikenuen, besterik gabe, posible ote zen orain arteko metodologia beste iturri bateko konbinazioei ere aplikatzea; gainerakoak azterketa erdiautomatikorako utzi ditugu.



4.6 irudia – Azterketa linguistikoa erdiautomatizatzeko metodoa (identifikazioari dagokion zatia nabarmenduta)

Urrats honetan, hautagaien inguruko informazio morfosintaktiko xehea lortu nahi izan dugu; hain zuzen ere, aurreko atalean azaldutako ezaugarrien ingurukoa[10]:

- ISaren numeroa: singularra (Sing.) ala plurala (Pl.)
- ISko determinatzaileak (Det.)
- ISaren definitutasuna, determinatzaileren bat dagoen kasuetan: zehaztua (Def.) ala zehaztugabea (Ind.)
- ISaren barruko modifikatzaileak (Mod.)
- Aldaketak osagai hitzen arteko hurrenkeran (Hurr.)

[10]Eskuzko azterketa xehean (4.1.2. atala), aditzaren eta izen-sintagmaren artean beste hitzik ager litekeen ere aztertu dugu. Ondoren, ordea, iruditu zaigu informazio hori ez zela hain esanguratsua, landutako UF guztiak baitziren bereizgarriak, hein batean behintzat (ikus 75. adibidea eta harekin batera dagoen oin-oharra). Hortaz, hemendik aurrera alde batera utziko dugu ezaugarri hori.

UF-hautagai bakoitzaren agerraldi bakoitzeko, informazioa gordetzen joan gara (geroago erakutsiko dugu adibide bat). Batzuetan, ordea, aditz eta izen berbera izan daitezke UF baten baino gehiagoren parte (94. eta 95. adibideak), edo UF baten eta UF ez den konbinazio baten parte (96. adibidea), nahiz eta aditza izenaren buru sintaktikoa izan kasu guztietan.

(94) *No es posible **dar paso** a muchas preguntas hoy.*

(95) *Estamos a punto de **dar** un **paso** transcendental.*

(96) *Los pasos dieron media vuelta y se marcharon.*

Aurreko azterketan ikusitakoen arabera, gure hipotesia da ezaugarri jakin batzuk bereziki lagungarriak direla halakoak bereizteko: erlazio sintaktikoa (97. adibidea), izen-sintagmaren barruan determinatzaileak egoteko aukera (98. adibidea), eta aditzaren erabilera pronominala (99. adibidea). Beraz, urrats honetan, informazioa bereiz gorde dugu hiru alderdi horietan desberdinak diren agerraldietarako.

(97) a. *Pueden **tomar parte** en los debates.*
→ Objektua
b. *Cada parte tomará todas las medidas necesarias.*
→ Subjektua

(98) a. *Estas cuestiones pueden **ser de interés** para los participantes.*
→ Determinatzailerik gabea
b. *Esto debería **ser del interés** del cliente.*
→ Determinatzaileduna

(99) a. ***Nos damos** perfecta **cuenta** de lo ocurrido.*
→ Aditza forma pronominalean
b. *Las autoridades deben **dar cuenta** de lo ocurrido.*
→ Aditza forma ez-pronominalean

Gerora, bosgarren urratsean (4.4.3. atala), corpus paraleloak erabili ditugu bereizitako hautagaiak benetan bi hitz-konbinazio desberdin diren ala bakarrean bildu behar diren erabakitzeko. Esate baterako, 97. eta 99. adibideetan bina UF desberdin daude, baina 98. adibideko hitz-konbinazioak UF bakarraren bi aldaki morfosintaktiko dira –eta, hortaz, hautagai bakarrean bildu beharrekoak–.

Bestetik, lehentxeago aipatu dugu *DiCE* hiztegiko hitz-konbinazioak preposiziorik gabe gordetzen direla, nahiz eta horietako asko preposizio eta guzti agertu testuetan. Halakoetan ere, preposizioka desberdindu ditugu hautagaiak (100. adibidea).

(100) a. *No lo* ***dejaremos a*** *un* ***lado***.
b. *No lo* ***dejaremos de lado***.

		IS numeroa			Det. mota			
	Pron.	Sing.	Pl.	Det.	Def.	Ind.	Mod.	Ord.
TOMAR - subj - - **PARTE**	0	100	0	0	0	0	3,55	89,94
TOMAR - obj - - **PARTE**	0	100	0	0	0	0	4,85	78,97
SER - ccomp **DE** - **INTERÉS**	0	100	0	0	0	0	47,21	6,61
SER - ccomp **DE** * **INTERÉS**	0	83,98	16,02	95,39	86,65	8,74	43,69	5,83
DAR pron obj - - **CUENTA**	100	100	0	0	0	0	85,71	2,38
DAR - obj - - **CUENTA**	0	100	0	0	0	0	97,37	25,66
DEJAR - ccomp **A** * **LADO**	0	92,86	7,14	100	16,67	83,33	9,52	2,38
DEJAR - ccomp **DE** * **LADO**	0	100	0	100	30	70	20	10

4.7 irudia – Bigarren urratsean sortzen diren taulen adibide bat (zenbakiak, ehunekotan)

Hautagai bakoitzeko, ehunekotan gorde dugu atal honen hasieran zerrendatutako sei ezaugarrien inguruko informazioa. Orain arte emandako hainbat adibideren datuak, adibidez, 4.7. irudiko taulan ikus daitezke: ezaugarri bakoitzari dagozkion ehunekoak zutabeka, eta izen eta aditz berbera duten hautagaiak bereizteko arrazoiak nabarmenduta.

Hiztegietako hautagai guzti-guztiak ez dira corpusean agertu, eta beste batzuk –oso agerraldi gutxi izan dituztenak[11]– alde batera utzi ditugu. Hala,

[11]Hamar agerraldi baino gutxiago izan dituzten hautagaiak ez ditugu kontuan hartu. Adibideei begiratuta erabaki dugu atalasea hamar agerralditan jartzea, iruditu baitzaigu erauzitako informazioaren kalitatea jaitsi egiten zela atalase horretatik behera eta datu horiek ez zirela nahikoa onak orokortzeak ganoraz egiteko.

guztira 979 UF-hautagairekin egin dugu aurrera: *Elhuyar* hiztegiko 435ekin eta *DiCE*ko 544rekin.

Zenbaki horiei erreparatuta, aipagarria da *Elhuyar*reko 1.205 konbinaziok 435 hautagai bakarrik sortu izana, batez ere ikusirik *DiCE*ko 437k 544 sortu dituztela. Gogoan izan behar da, dena den, iturriak oso desberdinak direla elkarren artean, eta kolokazioak izan ohi direla UF moten artean maizen erabiltzen direnak, hau da, *DiCE*n jasotzen diren konbinazioen gisakoak.

Datorren atalean (4.4.2) azalduko dugu xeheago nola sailkatu ditugun hautagai horiek patroi morfosintaktikoen arabera.

4.4.2 Hautagaiak patroika sailkatzea (2. urratsa)

Sailkapena ehunekotan oinarritu dugunean, honako ideia hau izan dugu gogoan: hautagai baten agerraldi gehien-gehienak era batera agertzen badira testuetan, gainerako agerraldiak ziur asko ez direla esanguratsuak hautagai horri dagokionez, analisi-erroreetatik datozelako edo, zenbaitetan, beste esanahi bati dagozkiolako. Hala, ehunekoen araberako atalaseak ezarri ditugu, mugatzeko zein puntutatik aurrera hartuko dugun ezaugarri bakoitza erabakigarritzat.

Atalase horien arabera[12], ezaugarriak hiru multzotan sailkatu ditugu: Y (bai, ezaugarri hori beti agertzen da era jakin batean hautagaiaren agerraldietan), O (aukerakoa da; ezaugarri hori era batera edo bestera ager daiteke hautagaiaren agerraldietan), edo N (ez, ezaugarri hori ez da inoiz agertzen era jakin horretan hautagaiaren agerraldietan).

Erlazio sintaktikoa salbu, kontuan hartu ditugun ezaugarri guztiak morfologikoak izan dira urrats honetan. Izan ere, aurreko azterketan ikusi dugu UF gehienak nahiko malguak zirela izen-sintagmaren barruko modifikatzaileei eta hitz-hurrenkerari zegokienez eta, hain zuzen, halakoetan murriztapenak zituztenak osaera morfologiko jakinetakoak izaten zirela ia beti. Beraz, lotura hori aintzat harturik, bi ezaugarri horiei buruzko informazioa geroago gehitzea erabaki dugu, patroiz patroi.

Proba-saioetarako, 490 konbinazio erabili ditugu, iturri batetik eta bestetik lortutako hautagaien erdiak: *Elhuyar* hiztegiko 218 eta *DiCE* hiztegiko 272. Ondoren, ezaugarri antzekoak dituzten hautagaiak taldekatu ditugu.

[12]Proba batzuk egin ondoren, 90 eta 10 inguruko atalaseekin lortu ditugu emaitzarik onenak. Alegia: UF bat beti singularrean erabiltzen dela erabakitzeko, adibidez, agerpenen % 90 edo gehiago izan behar izan ditu singularrean, eta inoiz singularrean erabiltzen ez dela zehazteko, aldiz, agerpenen % 10 edo gutxiago.

Oso hautagai gutxiko taldeak orokortu ondoren, hamabi patroi morfosintaktiko ezarri ditugu *Elhuyar* hiztegiko hautagaientzat. Patroi bakoitzaren ezaugarriak 4.13. taulan daude jasota.

	Pron.	**IS numeroa**		**Det.**	**Det.**	
		Sing.	**Pl.**		**Def.**	**Ind.**
FREE	N	O/N	O/N	Y/O/N	Y/O/N	Y/O/N
PL_NO-DET	N	N	Y	N	O/N	O/N
PL_DET_DEF	N	N	Y	Y	Y	N
PL	N	N	Y	Y	Y/O/N	Y/O/N
SING_NO-DET	N	Y	N	N	N	N
SING_DET_DEF	N	Y	N	Y	Y	N
SING_DET_IND	N	Y	N	Y	N	Y
SING	N	Y/O	N	Y/O	O/N	O/N
P_PL	Y	N	Y	Y/O	Y/O/N	Y/O/N
P_SING_NO-DET	Y	Y	N	N	N	N
P_SING_DET_DEF	Y	Y	N	Y	Y	N
P_SING	Y	Y	N	Y/O	Y/O/N	Y/O/N

4.13 taula – *Elhuyar* hiztegiko hitz-konbinazioen patroi morfosintaktikoak

Jarraian, patroi berberak erabili ditugu *DiCE*ko hautagaiak sailkatzeko ere, baina, espero genuen bezala, ez dute oso emaitza onik eman eta egokitu egin behar izan ditugu. Izan ere, *DiCE*n jasotako hitz-konbinazioek murriztapen morfosintaktiko gutxiago dituzte oro har, eta patroi orokorragoak egokiagoak dira halakoentzat.

Egokitzapenak egin ostean, *Elhuyar*rerako sortutako hamabi patroiak bostera murriztea erabaki dugu (4.14. taula). Zazpi patroi baztertu ditugu guztira, *DiCE*ko hautagaien artean agertu ez direnez erabilgarriak ez zirela iritzita. Hain zuzen ere, forma pronominalean bakarrik erabiltzen diren konbinazioak eta pluralean bakarrik erabiltzen direnak utzi ditugu alde batera.

Urrats honetan lortzen den informazioak 4.8. irudiko taularen itxura du. Irudi horretan, 97-100. adibideetako hautagaiak nola sailkatu diren erakusten da, eta nabarmenduta dago zer ezaugarri izan diren erabakigarriak hautagai bakoitza sailkatzeko.

	Pron.	IS numeroa		Det.	Det.	
		Sing.	Pl.		Def.	Ind.
FREE	N	O/N	Y/O/N	Y/O/N	Y/O/N	Y/O/N
SING_NO-DET	N	Y	N	N	N	N
SING_DET_DEF	N	Y	N	Y	Y	N
SING_DET_IND	N	Y	N	Y	N	Y
SING	N	Y/O	N	Y/O	O/N	O/N

4.14 taula – *DiCE*ko hitz-konbinazioen patroi morfosintaktikoak

	IS numeroa			Det. mota		
	Pron.	Sing.	Pl.	Det.	Def.	Ind.	
TOMAR - subj - - **PARTE**	N	Y	N	N	N	N	DISCARD
TOMAR - obj - - **PARTE**	N	Y	N	N	N	N	SING_NO-DET
SER - ccomp **DE** - **INTERÉS**	N	Y	N	N	N	N	SING_NO-DET
SER - ccomp **DE** * **INTERÉS**	N	O	O	O	O	O	FREE
DAR pron obj - - **CUENTA**	Y	Y	N	N	N	N	P_SING_NO-DET
DAR - obj - - **CUENTA**	N	Y	N	N	N	N	SING_NO-DET
DEJAR - ccomp **A** * **LADO**	N	Y	N	Y	O	O	SING
DEJAR - ccomp **DE** * **LADO**	N	Y	N	Y	O	O	SING

4.8 irudia – Lehen sailkapen-prozesuaren adibide bat

Aurreko atalean azaldu dugunez, konbinazio batzuk hautagai batean baino gehiagotan bereizi ditugu, eta informazio morfosintaktikoa bereiz jaso dugu haietako bakoitzarentzat. Hala, bikoiztutako hautagai gehienei patroi desberdina eman zaie. Hautagai horiek benetan bereiz tratatu beharrekoak diren ala konbinazio bakarraren aldakiak diren erabakitzeko, corpus paraleloetara jo dugu hurrengo urratsean, eta horri buruz jardungo dugu 4.4.3. atalean.

Lehenago, ordea, ikus dezagun zer-nolako emaitzak lortu ditugun lehen sailkapenaren ondoren.

Lehen sailkapenaren ebaluazioa

Patroiak sortzeko probak hautagai guztien erdiekin egin ditugu –490ekin guztira– eta hala erabaki dugu zer atalase erabili eta zer ezaugarriren arabera sortu patroiak. Ebaluaziorako, berriz, gainerako hautagaiak erabili ditugu: *Elhuyar* hiztegiko 217 eta *DiCE*ko 272.

Hautagaiei eskuzko patroiak egokitu dizkiegu lehenik, eta automatikoki esleitutako patroiak eskuzkoekin alderatu ditugu ondoren. Iturri bateko eta besteko patroi morfosintaktikoak desberdinak direnez, ebaluazioa ere bereiz egin dugu. Bi eratara kalkulatu dugu bat-etortzea (4.15. taula): ehunekotan eta Cohen κ erabilita (Cohen, 1960).

	Elhuyar		**DiCE**	
	Konbinazioak	**%**	**Konbinazioak**	**%**
Zuzen	118	54,38	148	54,42
Oker	99	45,62	124	45,59
Cohen κ	0,45		0,39	

4.15 taula – Lehen sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κ-ren arabera

Taulan ikusten denez, corpus elebakarrak bakarrik erabilita ere, hautagaien erdia baino gehiago ondo sailkatzen dira automatikoki. Ehunekotan, oso emaitza antzekoak lortu dira bi iturrietako hautagaiekin, baina, Cohen κri begiratuta, aldea igartzen da batzuen eta besteen artean. Dena den, ez da harritzekoa *DiCE*ko hautagaiek Cohen κ baxuagoa lortzea, gogoan izan behar baita neurri horrek kontuan hartzen duela ausazkotasuna ere. Hau da: *DiCE*ko hautagaientzat ezarritako patroiak bost bakarrik izan direnez –*Elhuyar* hiztegirako ezarritakoak baino zazpi gutxiago–, patroiak ausaz asmatzea errazagoa izango litzateke programarentzat, eta horrek jaitsi egiten du Cohen κ bidez kalkulatutako adostasuna.

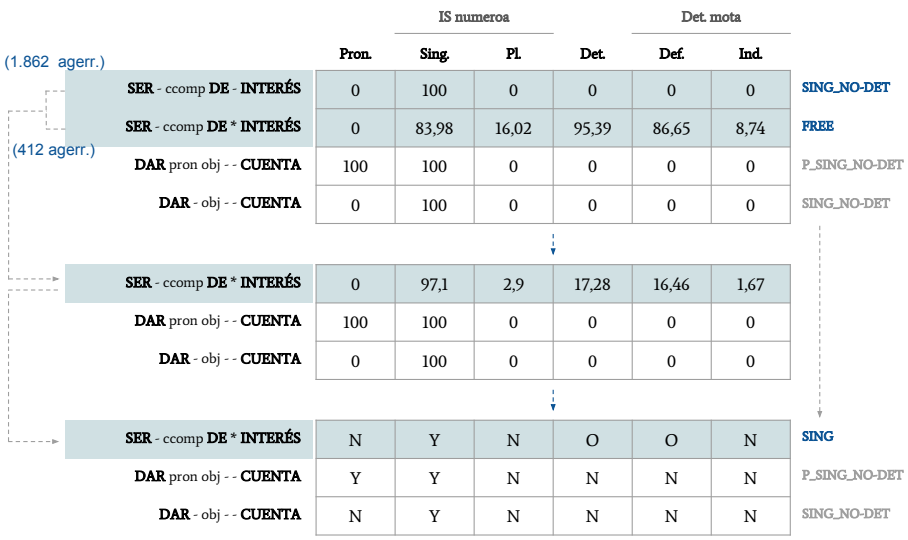
Hurrengo urratsean, emaitza horiek hobetzen saiatu gara corpus paraleloen laguntzaz; azal dezagun nola.

4.4.3 Sailkapena fintzea, corpus paraleloetan begiratuta (3. urratsa)

Corpus elebakarrak erabiliz lortutako emaitzak fintzeko, corpus paraleloak erabili ditugu. Gaztelania-euskarazko 7,5 milioi esaldiko corpus batekin eta gaztelania-ingelesezko 15 milioi esaldiko batekin. egin dugu proba, eta bigarrena erabiltzea erabaki dugu azkenean. Izan ere, besteak, tamaina txikiagokoa izanik, ez du hain eragin handirik izan.

Urrats honetan, honako ideia hau hartu dugu oinarritzat: hitz berberez osatutako bi hautagaik oso antzeko itzulpenak izan badituzte corpusetan, UF berberaren aldaki morfosintaktikoak dira; oso itzulpen desberdinak eman bazaizkie, aldiz, bi UF desberdin dira.

Beraz, honela erabaki dugu lema bereko hautagaiekin zer egin. Lehenik, hautagai bakoitzaren itzulpenak erauzi ditugu corpusetik, mGiza tresnaren bidez. Ondoren, hautagaiek zenbat itzulpen berdin zituzten kontatu dugu, eta, itzulpen guztien % 35 edo gehiago berdinak izan badituzte, bi hautagaiak bakarrean bildu ditugu. Bataren eta bestearen ezaugarri guztiak batu ditugu, eta berriro sailkatu dugu hitz-konbinazioa (4.9. irudia).



4.9 irudia – Bigarren sailkapen-prozesuaren adibide bat

Prozesu hori hobeto azaltzeko, ikus dezagun zer gertatu den 98. eta 99. adibideetako hitz-konbinazioekin: *SER - ccomp DE - INTERÉS* eta *SER -*

*ccomp DE * INTERÉS*, eta *DAR - obj - - CUENTA* eta *DAR pron obj - - CUENTA*. Lehen sailkapenean, hautagai bakoitzari patroi bana esleitu zaio. Corpus paraleloan begiratuta, ordea, ikusi dugu hautagai-bikoteek honako bat-etortzea zeukatela itzulpenei zegokienez:

- ***SER*** - *ccomp* ***DE*** _ ***INTERÉS***
  SER - *ccomp* ***DE*** _ * ***INTERÉS***
  → % 54,94

- ***DAR*** _ *obj - -* ***CUENTA***
  DAR *pron obj - -* ***CUENTA***
  → % 15,78

Bigarren bikoteak ez bezala, lehenak gainditu du % 35eko atalasea; beraz, hautagaien informazioa bakarrean bildu da[13], eta berriro sailkatu.

Bigarren sailkapenaren ebaluazioa

Lehenago erakutsi dugunez, corpus elebakarreko informazioa bakarrik erabilita, hautagaien erdia ondo sailkatu da automatikoki. Beste urrats honetan, ordea, corpus paraleloen bidez emaitza horiek hobetzen saiatu gara. Bigarren sailkapenaren emaitzak 4.16. taulan daude jasota.

	Elhuyar		**DiCE**	
	Konbinazioak	**%**	**Konbinazioak**	**%**
Zuzen	127	58,53	161	59,19
Oker	90	41,47	111	40,81
Cohen κ	0,50		0,45	

4.16 taula – Bigarren sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κren arabera

Ehunekotan, sailkapena ia 5 puntu hobetzen da bi iturrietako hautagaientzat, eta Cohen κri dagokionez, berriz, 0,05 eta 0,06 puntu –*Elhuyar* hiztegiko

[13]Hautagaien informazioa batzeko, bataren eta bestearen pisua corpuseko agerraldien arabera normalizatu dugu: ezaugarri bakoitzaren balioa kasuan-kasuan dagokion agerraldiekin biderkatu, eta bildu beharreko hautagaien agerraldi guztiekin zatitu dugu.

eta *DiCE*ko hautagaientzat, hurrenez hurren–. Hortaz, argi dago corpus paraleloak lagungarriak direla bikoiztutako hautagaiekin zer egin erabakitzeko, beti ere corpusa nahikoa handia bada.

Egia da emaitza horiek, horrela ikusita, ez diruditela oso altuak, baina gogoan izan behar da metodologia honen helburua eskuzko lanerako lagungarri izatea dela. Kontuan harturik 10 hautagaitik ia 6 ondo sailkatzen direla, esan dezakegu helburu hori bete dugula eta metodologia baliagarria dela UFen eskuzko azterketa errazteko. Gainera, 4.5. atalean eta 5.4. atalean erakutsiko dugunez, metodologia erabat automatikoki erabilita ere informazio horrek badu eragina identifikazio-atazan eta itzulpen automatikoan.

Bestetik, garrantzitsua da aipatzea emaitzak jaitsarazten dituzten hautagai asko eta asko baztertzekoak direla, ez arrazoi morfosintaktikoengatik, baizik eta lexiko-semantikoengatik, hain zuzen ere guk lantzen ez ditugun ezaugarri horiengatik. Etorkizunean, beraz, interesgarria litzateke aztertzea nola txerta daitekeen informazio mota hori prozesu honetan guztian.

Azken datu multzoari buruzko oharrak

Behin azterketa osorik bukatutakoan, datu guztiak bildu, eta eskuzko azterketan landutakoekin (4.1. atala) batu ditugu. Horretarako, lehenik, eskuzko azterketako konbinazioei patroiak esleitu dizkiegu, bai *Elhuyar*reko 117ei eta bai *DiCE*ko 22ei. Ondoren, erdiautomatikoki landutakoekin elkartu, errepikaturen bat bazegoen bakarra utzi, eta, errepaso-lan pixka baten ostean, azken zerrenda osatu dugu: 668 UF guztira. Haien patroi morfosintaktikoei dagozkien datuak 4.17. taulan daude jasota.

Ikusten denez, multzorik handiena (% 35,63) konbinazio erabat malguena da, eta singularrean baino erabiltzen ez direnak ere asko dira (lau patroiak batuta, % 57,04). Gainera, badirudi pluralean baino erabiltzen ez diren UFak ez direla hain usuak (hiru patroiak batuta, % 4,19) eta pronominalean bakarrik erabiltzen direnak are gutxiago direla oraindik (osotara, % 3,14). Dena dela, ez da harritzekoa azken multzo horiek txikiagoak izatea, *DiCE*ko konbinazioetatik bakar bat ere ez baitugu horietan sailkatu.

Eskuzko azterketan (4.1.2. atala), landutako 117 UFetako % 56 erabat malgutzat jo ditugu, eta gainerako % 44ak erdifinkotzat. Sailkapen orokor horretara 4.17. taula osorik ekarrita, % 35 dira erabat malguak, eta gainerako guztiak erdifinkoak dira. Beraz, lehen baino nabarmenagoa da orain murriztapen morfosintaktikoen garrantzia.

Horrez gain, aipagarria da 27 UFk dutela hitz berberez osatutako beste

	Kopurua	**%**
FREE	238	% 35,63
SING	189	% 28,29
SING_NO-DET	110	% 16,47
SING_DET_DEF	62	% 9,28
SING_DET_IND	20	% 2,99
PL_NO-DET	10	% 1,50
PL_DET_DEF	9	% 1,35
PL	9	% 1,35
P_SING_DET_DEF	6	% 0,90
P_PL	6	% 0,90
P_SING_NO-DET	5	% 0,75
P_SING	4	% 0,60

4.17 taula – Eskuz eta erdiautomatikoki landutako konbinazioen patroi morfosintaktikoak, konbinazio gehien dituztenetatik gutxien dituztenetara

UF bat 668 UF horien artean, hitzen arteko erlazio sintaktiko berberekoa. UF-pare horien artean daude, esate baterako, honako hauek:

(101) a. ***DAR*** - *obj* - - ***VOZ***
dar voces ‘oihuka aritu’
→ PL_NO-DET
b. ***DAR*** - *obj* - *?* ***VOZ***
dar voz ‘ahotsa eman’
→ SING

(102) a. ***HACER*** *pron obj* - *?* ***ILUSIÓN***
hacerse ilusiones ‘itxaropenak egin/izan’
→ P_PL
b. ***HACER*** - *obj* - *?* ***ILUSIÓN***
hacer ilusión ‘ilusioa egin, poz eman’
→ SING

Datuak biltzean, halako UFei marka bat jarri diegu patroi morfosintaktikoarekin batera: lehentasun-zenbaki bat. Izan ere, zenbait UF-pare nahasgarriak izan litezke identifikazio-lanerako, eta garrantzitsua da lehentasuna

zeini ematen zaion zehaztea. Adibidez, lehenago ere esan dugunez, *dar* eta *paso* hitzak bi UFren parte izan daitezke.

(103) a. ***DAR*** - *obj* - - ***PASO***
dar paso 'bide eman'
→ SING_SING_NO-DET
b. ***DAR*** - *obj* - *?* ***PASO***
dar paso(s) 'pauso(ak) eman'
→ FREE

Patroi bat bestea baino orokorragoa denez, bigarren patroiak ere identifika litzake lehen patroiaren agerraldiak, eta okerreko UFtzat identifikatuko lituzke. Hortaz, lehentasun-zenbakia zehaztean, patroirik orokorrena amaierarako utzi dugu, eta murriztapen gehien dituena jarri dugu lehenengo, zalantza-kasuak ebazte aldera. Horiek horrela, bigarren identifikazio-esperimentuari ekin diogu; datorren atalean azalduko dugu nola.

4.5 Bigarren identifikazio-esperimentua, automatikoki aztertutako datuak erabiliz

Aurreko esperimentuan lortutako emaitzetatik (4.2. atala), ondorioztatu dugu informazio morfosintaktikoa baliagarria dela UFen identifikazio-lanerako. Hala ere, oso UF gutxi (117) genituen aztertuta esperimentu horren aurretik eta, gainera, ezin izan ditugu nahi beste emaitza kalkulatu, ez baitugu erreferentziazko corpus etiketaturik izan eskura.

Bigarren esperimentu honetan, ordea, badugu aditz-UFak etiketatuta dituen corpus bat, PARSEMErena[14], eta huraxe erabiliko dugu gure proposamenaren doitasuna ez ezik estaldura eta F neurria ere kalkulatzeko. Nolanahi ere, PARSEMEren corpusa irizpide jakin batzuei jarraituz etiketatu da, eta irizpide horiek ez dira guk orain arte erabili ditugun berberak (7. kapituluan emango ditugu irizpide horien inguruko xehetasun gehiago). Besteak beste, haiek ez dute kolokaziorik etiketatzen, eta, kontuan harturik guk UF asko *DiCE*tik erauzi ditugula, kolokazioen hiztegi batetik, gure orain arteko datu multzoak eta corpus hori ez dira erabat bateragarriak. Gainera, corpusean

[14]Gaztelaniazko corpusa helbide honetatik eskuratu daiteke: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842`

askotariko aditz-UFak biltzen dira, aditza+izena motakoak eta beste osaera askotakoak (aditza+adjektiboa, aditza+adberbioa, etab.).

Esperimentua prestatzeko lehen lana, hortaz, datu multzoa eta corpusa atazara egokitzea izan da. Azal dezagun nola egin dugun egokitzapen hori, eta erakuts ditzagun, ondoren, lortutako emaitzak.

4.5.1 Erabilitako baliabideak eta metodoak

Baliabideak atazarako prestatzeko, batetik corpusa eta bestetik identifikatzeko UFen zerrenda moldatu behar izan ditugu. PARSEMEren corpusak, testuaz gain, bi eratako informazioa dauka zehaztuta: testuko zer hitz-konbinazio diren UFak, eta analisi morfosintaktikotik ateratako informazioa (gramatika-kategoriak eta dependentzia-erlazioak, adibidez). Hortaz, corpus hori gure atazara egokitzeko, UF gisa etiketatutako hitz-konbinazioetan begiratu dugu ea osagaien gramatika-kategoriak bat ote zetozen gure aztergaiarekin –hau da, ea *aditza+(preposizioa)+(determinatzailea)+izena* motakoak ziren–, eta, besterik gabe, osaera hori ez zuten UFen etiketak ezabatu egin ditugu.

Azkenean, 5.515 esaldiko corpusean 662 UF gelditu dira etiketatuta guztira. Honela daude banatuta etiketa horiek corpusaren hiru zatietan: 355 entrenamendu-corpusean, 136 garapenekoan, eta 171 testekoan.

Bestetik, identifikatu beharreko UFak eta haien inguruko datuak prestatzeko, corpus horretan bertan agertzen diren 662 etiketak erauzi, eta automatikoki aztertu ditugu, 4.4. atalean azaldu dugun metodologiaren bidez[15]. Azkenean, azterketa-prozesuaren ondoren, 156 UFren patroi morfosintaktikoak lortu ditugu.

Jakina, corpuseko etiketa guztiak ez dira 156 UF horien agerpenak, esana baitugu azterketa-prozesu automatikoan ez dela informaziorik lortzen zenbait hitz-konbinaziorentzat, corpusean gutxiegi agertzen diren –edo behin ere agertzen ez diren– hitz-konbinazioentzat, hain zuzen. Hala ere, lortu ditugun 156 UF horiek etiketen % 59 biltzen dituzte, 389 etiketa, eta gainerako 273ak dira patroi morfosintaktikorik gabe gelditu zaizkigunak. Hala, aintzat harturik UF batzuen patroi morfosintaktikoak bagenituela baina beste batzuenak ezetz, identifikazio-metodoa ere UFen bitariko izaera horretara egokitu behar izan dugu.

[15] Azterketa automatikorako erabili dugun corpusa 4.4.1. atalean erabili dugun berbera izan da, eta patroi sorta, berriz, *DiCE*ko konbinazioak sailkatzeko erabili duguna.

Honako hau izan da, oro har, erabili dugun identifikazio-metodoa:

- Corpuseko testu hutsa hartzen da oinarritzat, eta Freeling analizatzailearen bidez lortzen dira gramatika-kategoriei, chunkei eta dependentzia sintaktikoei buruzko datuak.
- Datu horien gainean, UFak bilatzen dira:
  - Patroi morfosintaktikoa zehaztuta duten UFen osagaiak esaldi berean agertzen direnean, UF horien patroiei begiratzen zaie, eta, patroi bakoitzari dagozkion murriztapen morfosintaktiko guztiak betetzen badira, identifikazioa gauzatzen da.
  - Patroi morfosintaktikorik gabeko UFak, berriz, bi eratara tratatzen dira:
    A. Hitz-segida jarraitu gisa, aditzaren flexioa kontuan hartuz baina gainerako osagaiak corpuseko etiketaren forma berberean bakarrik bilatuz
    B. Osagaien lemak bilatuz, elkarren artean gehienez ere bi chunk dituztelarik eta dependentzia-zuhaitzean erlazio zuzena dutelarik

Emaitzak PARSEMEren ataza partekatukoekin ahalik eta alderagarrienak izan zitezen, haien testeko corpusa erabili dugu esperimentu honetan, baina, lehen esan bezala, aditza+izena motako etiketak bakarrik utzi ditugu. Datu multzoa, berriz, bi eratara integratu dugu: batetik, entrenamendu- eta garapen-corpuseko UFen inguruko informazioa bakarrik erabiliz (*train+dev*), ataza partekatuan benetan egiten den bezala; eta bestetik, corpuseko UF guztien inguruko informazioa erabiliz (*denak*), testeko UFena barne. Hala, metodoaren estaldura errealaren berri izango dugu batetik, lehen datu multzoaren bidez, baina, horrez gain, jakin ahal izango dugu gure metodoak zer emaitza lor litzakeen identifikatu beharreko UF guzti-guztiak aldez aurretik ezagutuko bagenitu.

4.5.2 Emaitzak

Aurreko atalean azaldutako corpusak eta datu multzoak erabilita, 4.18. taulako emaitzak lortu ditugu. Ikusten denez, emaitzak txukunak dira, entrenamendu- eta garapen-corpuseko UFak bakarrik erabiliz 0,51ko F neurria lortzen baita, eta datu multzo osoa erabiliz, berriz, 0,71koa. Bestetik, aipagarria

da patroirik gabeko UFak hitz-segida jarraitu gisa tratatzea hobe dela erabat malgutzat tratatzea baino, estaldura pixka bat baxuagoa izan arren, doitasun dezente altuagoa lortzen delako.

	P	**R**	**F**
A-train+dev	0,74	0,29	**0,51**
B-train+dev	0,60	0,31	0,46
A-denak	0,84	0,60	**0,72**
B-denak	0,76	0,67	0,71

4.18 taula – Bigarren identifikazio-esperimentuaren emaitzak

Hortaz, patroi morfosintaktikoen bidez lortzen diren emaitzak onak dira, nahiz eta, datu multzo mugatua bakarrik erabiltzen denean –hau da, testeko UFak kontuan hartu gabe–, estaldura apal samarra izan. Nolanahi ere, joera hori nabarmena da ataza partekatuan parte hartu duten sistema guztien artean (Ramisch *et al.*, 2018), eta, gurea baino estaldura hobeak egon diren arren (Boros eta Burtica, 2018; Pasquer *et al.*, 2018; Stodden *et al.*, 2018), oro har, estaldura ona lortzen duten sistemek doitasun-marka txarrak lortu dituzte. Hala, emaitza globaletan askogatik gainditzen ditugu gaztelaniazko atalean parte hartu duten sistemen emaitzak, F neurririk onena 0,38koa izan baita (Boros eta Burtica, 2018), gure metodoarena baino 13 puntu baxuagoa. Ikus PARSEMEren bigarren ataza partekatuko emaitzen laburpena, 4.19. taulan[16].

	P	**R**	**F**
ES	0,19 (0,00-0,32)	0,33 (0,00-0,49)	0,23 (0,00-0,38)
Guztiak	0,36 (0,00-0,68)	0,29 (0,01-0,53)	0,31 (0,00-0,54)

4.19 taula – PARSEMEren ataza partekatuko bigarren edizioan, gaztelaniaz eta hizkuntza guztietan oro har izandako emaitzak

Jakina, kontuan hartu behar da emaitza batzuk eta besteak ez direla erabat alderagarriak, guk corpuseko etiketa batzuk bakarrik erabiltzen baiti-

[16]Emaitza guztiak eskuragarri daude webgune honetan: `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_50_Shared_task_results#lang-ES`

tugu, aditza+izena osaera morfologikoa dutenak. Hala ere, parte-hartzaileen emaitzak UF motaka ere argitaratu dira, eta badirudi guk lantzen ditugun UFak direla, hain zuzen, identifikatzen zailenak. Izan ere, gaztelaniazko sei UF motetatik (ikus mota guztiak 2.1.2. ataleko 29. orrialdean), hirutan bakarrik sar daitezke aditza+izena osaerakoak[17] –LVC.cause, LVC.full eta VID–, eta horiexetan lortu dira emaitzarik kaskarrenak (mota bakoitzaren inguruko azalpen gehiago, 193. orrialdean).

	P	**R**	**F**
IAV	0,12 (0-0,28)	0,34 (0-0,84)	0,17 (0-0,37)
IRV	0,20 (0-0,35)	0,49 (0-0,79)	0,28 (0-0,46)
LVC.cause	0,13 (0-100)	0,02 (0-0,21)	**0,03 (0-0,30)**
LVC.full	0,17 (0-0,48)	0,14 (0-0,41)	**0,13 (0-0,33)**
MVC	0,15 (0-0,25)	0,33 (0-0,73)	0,20 (0-0,35)
VID	0,14 (0-0,46)	0,08 (0-0,23)	**0,10 (0-0,31)**

4.20 taula – PARSEMEren ataza partekatuko bigarren edizioan gaztelaniaz izandako emaitzak, UF motaka

Laburbilduz, esperimentuaren emaitzek erakusten dute UFen inguruko patroi morfosintaktikoak lagungarriak direla identifikazio-lanerako eta, horrez gain, halako daturik eskura ez dagoenean, hobe dela UFak hitz-segida finkoak balira bezala tratatzea, hitz-konbinazio erabat malguak balira bezala tratatzea baino.

[17]LVC.cause eta LVC.full etiketa daramaten UF guztiak dira aditza+izena motakoak; VID etiketa daramatenen artean, berriz, batzuk bakarrik.

Laburpena

Kapitulu honetan, gaztelaniazko aditza+izena konbinazioak aztertu ditugu, maila lexiko-semantikoan eta, batez ere, morfosintaktikoan. Azterketa hori erabat eskuz egin dugu lehenik, eta automatizatzeko metodologia bat proposatu dugu ondoren. Bateko zein besteko datuak oinarritzat hartuta, bi identifikazio-esperimentu ere egin ditugu, datu horiek erabilgarriak ote ziren ikusteko. Hona hemen lan horietan gogoan izan ditugun abiapuntu-hipotesiak (1.3. atalean zerrendatutakoak) eta gure ondorioak.

[A2] Aditza+izena motako UFak oso malguak izan ohi dira morfosintaxiari dagokionez, baina murriztapenak ere badituzte.

Azterketa osotik ateratako datuen arabera, landu ditugun UFetatik % 35,63 erabat malguak dira, eta gainerako guztiak erdifinkoak, hau da, murriztapen morfosintaktikodunak. Hortaz, eskuz edo erdiautomatikoki aztertutako 668 UFek, behintzat, betetzen dute hipotesi hori. Gainera, aipagarria da murriztapenen bat duten konbinazioen artean asko direla singularrean baino erabiltzen ez direnak (% 57).

[A5] Informazio morfosintaktikoa kontuan hartzeak UFen identifikazioa hobetu dezake.

Bi esperimenturen bidez erakutsi dugu hala dela. Lehenik, eskuz aztertutako datu morfosintaktikoak erabili ditugu gaztelaniazko 117 eta ingelesezko 133 UFren agerpenak corpusetan identifikatzeko. Hitzsegidak bakarrik bilatzen dituen oinarrizko metodo batekin alderatuta, ikusi dugu askoz ere agerpen gehiago identifikatzen direla informazio xehearen bidez (gaztelaniaz % 28 eta ingelesez % 79 gehiago), eta oso doitasun altuz. Bigarrenik, berriz, PARSEMEren ataza partekatuko corpusa erabili dugu, eta, entrenamendu-corpuseko UFei patroi morfosintaktikoak automatikoki esleituta, azken edizioko parte-hartzaileen emaitzak nabarmen hobetzea lortu dugu: 0,51ko F neurria erdietsi dugu corpus horretako UFak zeintzuk ziren aldez aurretik jakin gabe, eta UFak aldez aurretik ezagututa, aldiz, 0,71koa.

Helburuei dagokienez, berriz, honako hauek landu ditugu.

[H1] Gaztelaniazko eta euskarazko aditza+izena motako UFen ezaugarri morfosintaktikoak aztertzea.

Kapitulu honetan gaztelaniazko UFei begiratu diegu batez ere, eta azterketa morfosintaktiko xehea egin dugu, hasieran eskuz, eta ondoren erdiautomatikoki. UF gutxi batzuen kasuan, erabat automatikoki ere egin dugu azterketa hori, baina kalitatea bermatzeko iragazki batzuk jarrita. Hainbat ezaugarri hartu ditugu kontuan (izen-sintagmaren numeroa, determinatzaileak eta modifikatzaileak, eta UFko osagaien hurrenkera), eta horien araberako patroi morfosintaktikoak sortu ditugu ondoren, identifikazio-lanerako multzoak sortzeko. Azkenean, eskuz eta erdiautomatikoki aztertutako 668 UF lortu ditugu patroi eta guzti, eta automatikoki landutako beste 156 UF.

[H4] Aditza+izena motako UFen identifikazio automatikoa hobetzea.

Bi identifikazio-esperimentu egin ditugu, eta lortu dugu orain arteko identifikazio-metodo askoren emaitzak hobetzea. Izan ere, PARSEMEren ataza partekatuan parte hartu duten sistemen aldean, gure metodoak oso emaitza onak lortzen ditu F neurriari dagokionez: gaztelaniazko emaitzekin alderatuta, zehazki, sistemarik onenak baino 13 puntu gehiago erdiesten ditu gureak.

5. KAPITULUA

Euskarazko ordainen azterketa eta itzulpen automatikoa

Tesi-lana kokatzean azaldu dugunez (1. kapitulua), gure helburuen artean dago UFen identifikazioa hobetzea ez ezik aztertutako informazio linguistikoa itzultzaile automatikoetan ere integratzea. Horretarako, *Matxin* itzultzaile automatikoa hartu dugu oinarritzat.

Ataza honetan ere, identifikazioa hobetzeko baliatu dugun estrategia berbera erabili dugu, oro har. Hasteko, eskuzko azterketa xehe bat egin dugu UF gutxi batzuen gainean (5.1. atala), eta aztertutako informazioa itzultzaile automatikoetan nola aplika daitekeen ikusi dugu, esperimentu baten bidez (5.2. atala). Ondoren, azterketa linguistiko hori erdiautomatizatzeko metodo bat sortu dugu (5.3. atala), eta esperimentua errepikatu dugu datu berriak baliagarriak ote diren jakiteko (5.4. atala).

Jarraian emango ditugu prozesu horri guztiari buruzko xehetasun gehiago, eta, atalaren amaieran, tesiaren hipotesiekin eta helburuekin lotuko dugu egindako lana.

5.1 Eskuzko azterketa xehea

UFen euskarazko ordainak aztertzeko, ordain-zerrenda sortzeari ekin diogu lehenik. Batetik, eskuz landutako gaztelaniazko 117 UFak hartu ditugu berriro, *Elhuyar* hiztegitik erauzitakoak, eta bestetik, *DiCE* hiztegiko 22 kolokaziorik usuenak, 8.000 agerpenetik gorakoak. Ondoren, jakinik *Matxin*ek

lexikoiko sarrera bakoitzaren ordain bakarra erabiltzen duela, zerrendako UF bakoitzari ordain bana eman diogu eskuz.

*Elhuyar*reko konbinazioak hiztegi elebidun batetik atereak direnez, ordaina emateko, hiztegira bertara jo dugu. Ordain bakarra zuten UFei zuzenean esleitu diegu hiztegian jasotakoa, eta gainerakoentzat aukeren arteko bat hautatu dugu, maizen erabili dena edo gure ustez egokiena dena ahalik eta testuinguru gehienetarako erabilgarria izan dadin. Lan horretan ari ginela, ohartu gara hainbat UF anbiguoak zirela baina horietako batzuk, 14 zehazki, bereiz zitezkeela ezaugarri morfosintaktikoen bidez (104. adibidea).

(104) **a.** ES: ***dar paso*** *(a algo)*
EU: *(zerbaiti)* ***bide eman***
b. ES: ***dar*** *un* ***paso***
EU: ***pauso*** *bat* ***eman***

Hala, UF horiek bikoiztu, identifikaziorako ezaugarriak egokitu, eta bikoiztapenei ere ordaina eman diegu. *DiCE*ko UFei, berriz, erreferentziazko ordainik ez genuenez, geuk eman dizkiegu ordainak eskuz, erreferentzia-corpusen eta beste hainbat kontsulta-tresnaren laguntzaz (ikus B. eranskina). Azkenean, 153ko sorta osatu dugu.

Behin ordain sorta prestaturik, haien ezaugarri linguistikoak aztertzeari ekin diogu, gogoan hartuz lortutako datuak *Matxin*en erabiliko ditugula geroago. Bi multzotan bereizi ditugu aztertutako ezaugarriak: lexikoari dagozkionak batetik, eta morfosintaktikoak bestetik. Azal dezagun zer ezaugarriri erreparatu diogun zehazki.

5.1.1 Ezaugarri lexikoak

Aztergai ditugun UFen osagai nagusiak aditz bat eta izen bat direnez, gaztelaniazko bi osagai horiei euskarazko zer ordain ematen zaien zehaztu dugu lehenik. Jakin badakigun arren UFei osorik ematen zaiela ordaina eta ez osagai bakoitzari bere aldetik, informazio hori *Matxin*en errazago tratzeko asmoz, euskarazko ordaina ere bitan zatitu dugu. Esate baterako, 105. adibidean, *interés* izenari *interes* eman zaio euskarazko ordaintzat, eta *mostrar* aditzari, *agertu*. Izena izen batez eta aditza aditz batez ordeztu dira.

(105) ES: ***mostrar interés***
EU: ***interesa agertu***

Beste zenbait kasutan, aldiz, gaztelaniazko izenari zegokion euskarazko ordaina ez da izena izan, esana baitugu aditza+izena motako UFak ez direla beti aditz batez eta izen batez itzultzen. Hona hemen bi adibide:

(106) ES: ***dar voces***
EU: *oihuka aritu*

(107) ES: ***contraer matrimonio***
EU: *ezkondu*

Lehenengo adibidean (106), izenari adberbio bat eman zaio ordaintzat: *voces* → *oihuka*. Bigarrenean (107), berriz, ez diogu ordainik esleitu izenari, UF osoaren euskarazko ordaina aditz soil bat delako: *ezkondu*.

Beraz, lexikoari dagokionez, horixe egin dugu, ordainak aztertzen hasteko: UFen ordainek barruan zer lema duten esatea. Horren ondotik zehaztu dugu ordain bakoitza zer kategoriatako hitzez osatuta dagoen eta nola erabiltzen den, hurrengo urratsean.

5.1.2 Ezaugarri morfosintaktikoak

Morfosintaxiari dagokionez, gaztelaniazko UFak lantzean eginiko azterketaren antzekoa egin dugu ordainekin ere. Zer ezaugarri aztertu ditugun azaltzeko, gogoan izan ditugun galderak zerrendatuko ditugu jarraian, banan-banan.

Hasteko, ordaina zer motatako hitzek osatzen duten aztertu dugu.

- Zer kategoriatako hitzek osatzen dute ordaina? Izen batek eta aditz batek (108), adberbio batek eta aditz batek (109), adjektibo batek eta aditz batek (110), aditz batek bakarrik (111)...?

  (108) ES: ***rendir cuentas***
  EU: ***kontu eman*** → ize+adi

  (109) ES: ***estar de acuerdo***
  EU: ***ados egon*** → adb+adi

  (110) ES: ***ser un caso***
  EU: *berezia izan* → adj+adi

  (111) ES: ***llevar a cabo***
  EU: *burutu* → adi

Ondoren, izena+aditza motako ordainak bakarrik aintzat hartuta, beste hiru galdera egin ditugu.

- Zer kasu-marka edo postposizio lotzen zaio izenari? Absolutiboa (112), datiboa (113), inesiboa (114)...?

(112) ES: ***dar miedo***
EU: ***beldur eman*** → abs

(113) ES: ***mantener*** *el* ***equilibrio***
EU: ***oreka<u>ri</u> eutsi*** → dat

(114) ES: ***estar en contacto***
EU: ***harreman<u>etan</u> egon*** → ine

- Izen-sintagmak har dezake –artikulu mugatuez gain– determinatzailerik? Bai (115) ala ez (116)?

(115) ES: ***cumplir*** *un* ***plazo***
EU: ***epe*** <u>*bat*</u> ***bete*** → determinatzailea

(116) ES: ***ser una pena***
EU: ***pena izan*** → determinatzailerik ez

- Nolakoa da izen-sintagma, mugatasunari eta numeroari dagokionez? Mugagabea (117), mugatu singularra (118) ala mugatu plurala (119)?

(117) ES: ***dar paso***
EU: ***bide eman*** → mg

(118) ES: ***conocer mundo***
EU: ***munduan ibili*** → sing

(119) ES: ***tomar las aguas***
EU: ***urak hartu*** → pl

Azkenik, gaztelaniazko UFak aztertzean zehaztu ez ditugun hiru ezaugarriri ere erreparatu diegu, ordaina ondo emateko garrantzitsuak zirelakoan.

- Ordainak ba al du kanpo elementu irekirik? Alegia, ba al da, lexikoari dagokionez aldakorra izan arren, ordainarekin batera beti agertzen den sintagmarik? Egonez gero, gaztelaniazko zer elementu ordezkatu behar da euskarazko zer elementurekin, eta zer marka jarri behar zaio? Gaztelaniazko objektu zuzena euskarazko modifikatzailearekin, genitibo-marka jarrita (120); gaztelaniazko modifikatzailea euskarazko beste modifikatzaile batekin, prolatibo-marka jarrita (121)...?

  (120) ES: ***echar en falta*** *(X)*
  EU: *(X)ren* ***falta sumatu*** → obj:mod-gen

  (121) ES: ***tener consideración*** *(de X)*
  EU: *(X)tzat hartu* → mod:mod-pro

- Gaztelaniazko UFa forma pronominalean dagoenean, euskarazko ordaina intrantsitibo bihurtzea komeni da? Bai (122), *Matxin*en gehienetan egiten den bezala, ala ez (123)?

  (122) ES: ***se abrió paso***
  EU: ***aurrera egin*** **zen* → ez

  (123) ES: *se* ***sacó a la luz***
  EU: ***argitara atera*** *zen* → bai

- Gaztelaniazko UFa ezezkoan dagoenean, euskarazko ordainak partitiboa izan dezake? Bai (124) ala ez (125)?

  (124) ES: *no le* ***hizo caso***
  EU: *ez zion* ***kasurik egin*** → bai

  (125) ES: *no* ***merece la pena***
  EU: *ez* ***du *merezirik*** → ez

Behin informazio hori guztia aztertutakoan, datuak tauletan jaso ditugu, eta *Matxin*en sisteman sartu ditugu, zer hobekuntza ekar lezaketen ikusteko. Datorren atalean azalduko dugu nola egin dugun hori eta zer emaitza lortu ditugun.

5.2 Lehen esperimentua *Matxin*en, eskuz aztertutako datuak erabiliz

Esan dugun bezala, *Matxin* itzultzaile automatikoak eginkizun bikoitza du UFei dagokienez: sorburu-hizkuntzako UFak identifikatzea batetik, eta haiei xede-hizkuntzako ordaina ematea bestetik. Beraz, bi ataza horietarako informazioa eman diogu guk, atal honetan azalduko dugun esperimentuan.

Hain zuzen ere, aurreko atalean aztergai izan ditugun 153 ordainak eta haiei dagozkien UFak erabili ditugu, landutako informazio linguistikoarekin batera. Datu horiei buruz arituko gara atal honetan *informazio linguistiko xehe* esatean. Azal dezagun, bada, nola sartu dugun informazio hori guztia itzultzaile automatikoan.

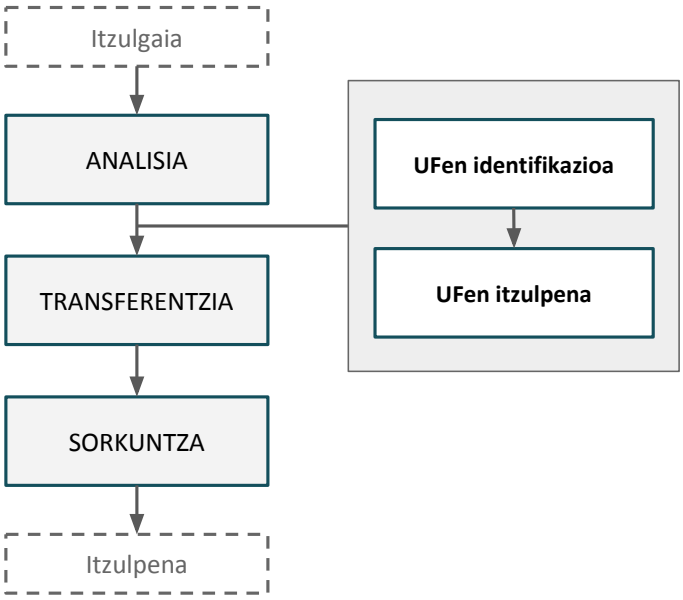
5.2.1 Informazio linguistikoa integratzeko proposamena

Matxin deskribatu dugunean (2.2.4.1. atala), azaldu dugu itzulpen-prozesuak hiru fase nagusi dituela: analisia, transferentzia eta sorkuntza. Guk, esperimentu honetan, lehen bien artean tratatu ditugu UFak (5.1. irudia), eta hala sortu dugu *Matxin-UF*, UFen inguruko informazioa ere barne hartzen duen sistema moldatua.

Analisi-fasea

Analisi-fasean, itzulgaia analizatzaile automatikoaren bidez aztertzen da, eta etiketa morfosintaktikoak jartzen zaizkio, jatorrizko sisteman bezalaxe: hitz bakoitza zer kategoriatakoa den, zer hitz multzo edo *chunk* dauden esaldian, eta nola erlazionatzen diren hitzak euren artean. Ondoren, analizatzailearen emaitzen gainean identifikatzen dira UFak, 4.2. atalean azaldu dugun metodologia berbera erabiliz. Hau da, hiru modutara bilatzen dira UFak itzulgaian:

- Hitz-segida finkoak balira bezala, aldaketa morfosintaktiko bakartzat aditzaren flexioa hartuz
- Eskuz landutako informazio linguistikoa eta analsi automatikotik ateratako *chunk*ak erabiliz
- Eskuz landutako informazio linguistikoa eta analisi automatikotik ateratako dependentzia sintaktikoak erabiliz

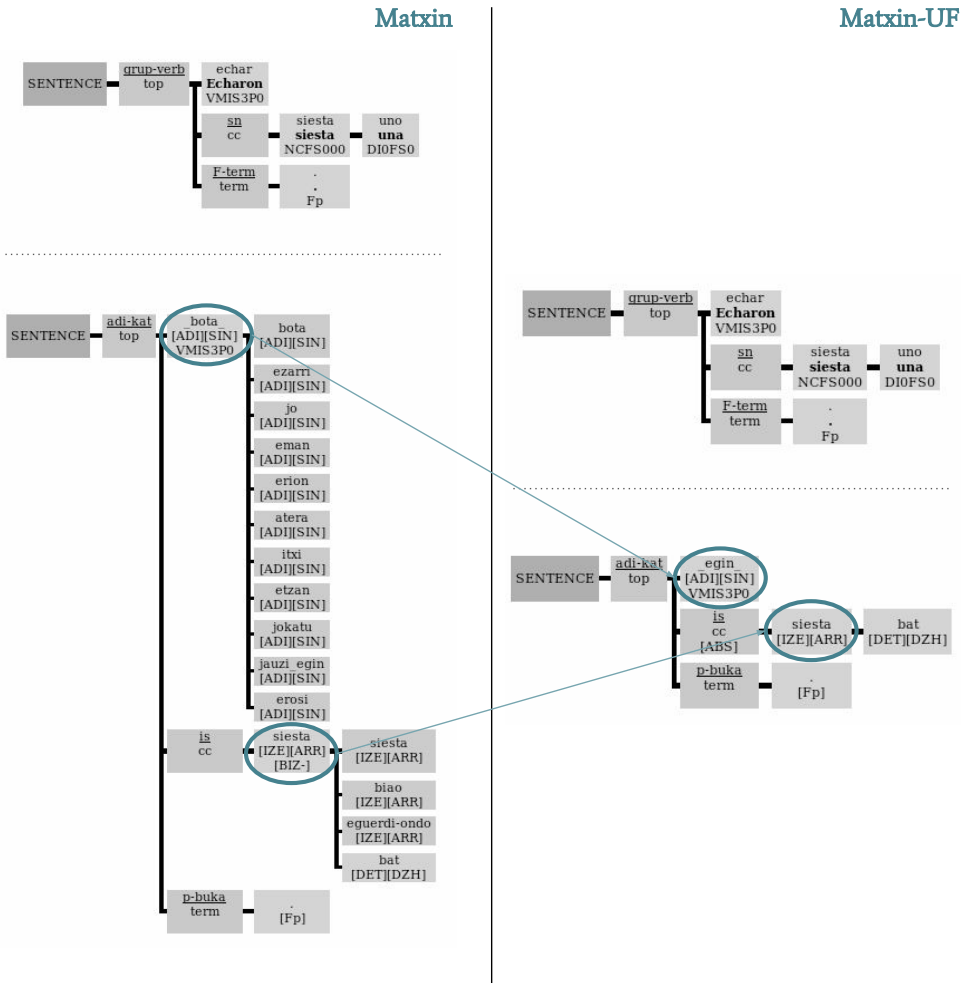


5.1 irudia – *Matxin*en UFei buruzko informazioa integratzeko metodologia-proposamena

Transferentzia-fasea

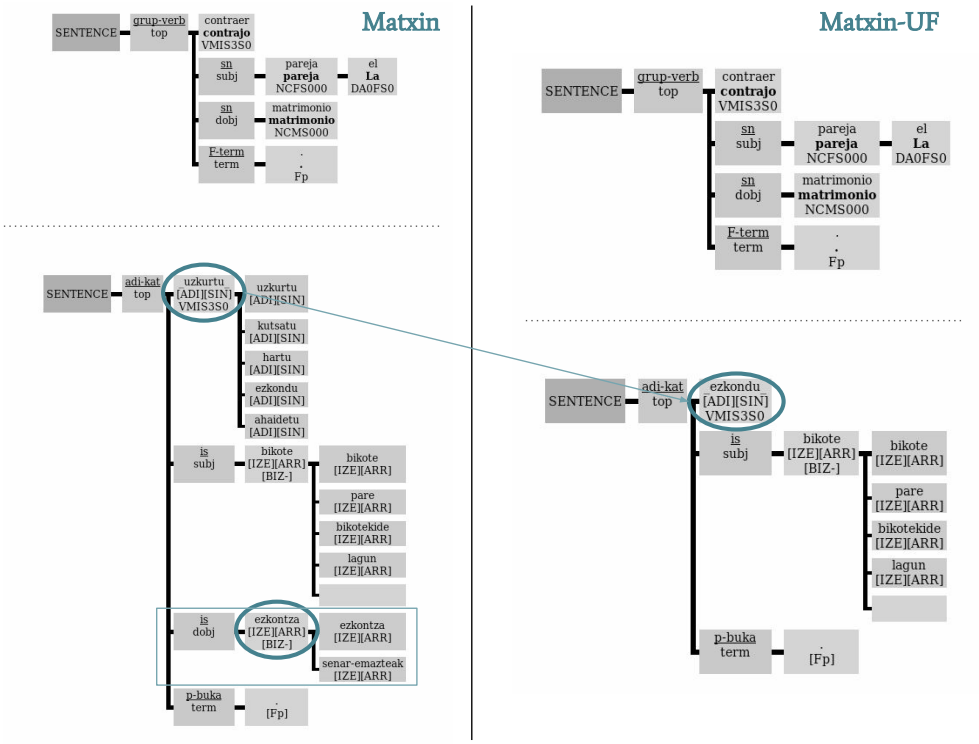
Behin UFak identifikatutakoan, haien transferentzia zuzenean egiten da, itzulgai osoaren transferentzia-fasearen aurretik. Hortaz, transferentzia-fase nagusiaren bi azpifaseetara heltzerako, transferentzia lexikora zein estrukturalera, sistemak aldez aurretik du zehaztua zein den UFaren ordain lexikoa, bai eta zer datu morfosintaktiko moldatu behar den ere, daturen bat moldatu beharra dagoen kasuetan.

Hala, *Matxin*en itzulpen-prozesua erakusten duten adibide-irudietan ikus daitekeenez, analisi-fasearen emaitza jatorrizko sistemarenaren berdin-berdina bada ere, transferentzia lexikoari dagokiona desberdina da (5.2. irudia), UFei buruzko informazio linguistikoa lehenago hartu delako kontuan. Informazio linguistiko xehearen arabera, *echar (una) siesta* UFaren ordaina *siesta (bat) egin* da. Horregatik ematen zaio aditzari jatorrizko sistemarena ez beste ordain bat *Matxin-UF*n: *bota*ren ordez, *egin*. Izena, berriz, berdin itzultzen da batean zein bestean, kasu horretan aditza bakarrik aldatzen delako.



5.2 irudia – *Echar una siesta* UFaren itzulpen-prozesua: analisia eta transferentzia lexikoa

UF-ordaina hitz batez baino gehiagoz osatuta badago, aditzari aditza ematen zaio baliokidetzat, eta izenari, gainerako zatia. Aurreko atalean esan dugu, ordea, UFaren ordaina izan daitekeela aditz bat bakarrik ere. Halakoetan, aditzari bakarrik ematen zaio ordaina, eta izenari dagokion lekua hutsik uzten da. Horixe gertatzen da, adibidez, *Matxin-UF*ri *contraer matrimonio* hitz-konbinazioa itzularazten badiogu: transferentzian, izenari dagokion zatia desagertu egiten da (5.3. irudia).

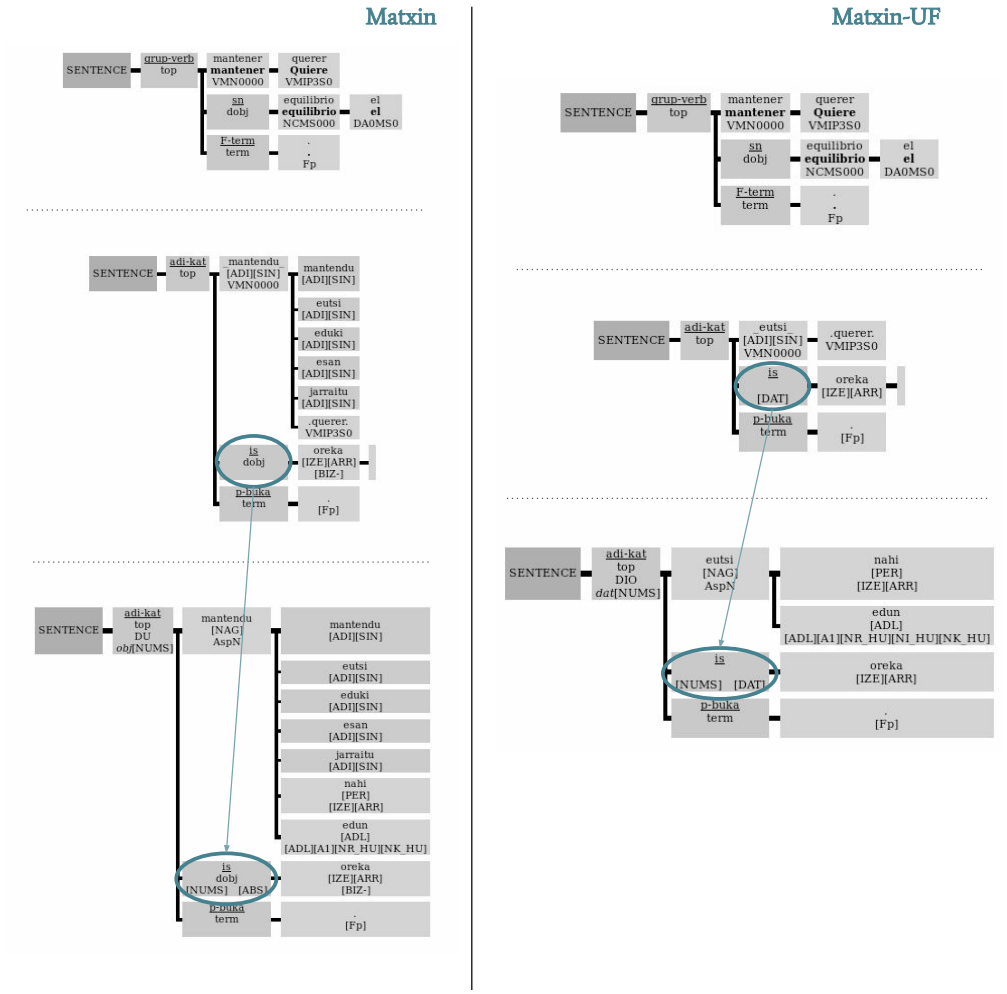


5.3 irudia – *Contraer matrimonio* UFaren itzulpen-prozesua: analisia eta transferentzia lexikoa

Esan bezala, transferentzia estrukturalean ere bere horretan uzten da UF-ordainari buruzko informazioa, eta gainerako osagaiei dagokien informazioa bakarrik aldatzen da. Esate baterako, *mantener el equilibrio* UFa euskararara ekartzeko, informazio linguistiko xeheak dio lexikoaz gain kasu-marka ere aldatu beharra dagoela. Gaztelaniazko UFa erregularki itzuliko balitz, absolutiboa jarriko litzaioke euskarazko ordainari (*oreka mantendu*), baina datiboa behar du (*orekari eutsi*). Datu hori transferentzia-fasearen aurretik zehaztu denez, transferentzia estrukturalean ez da aldatzen (5.4. irudia).

Sorkuntza-fasea

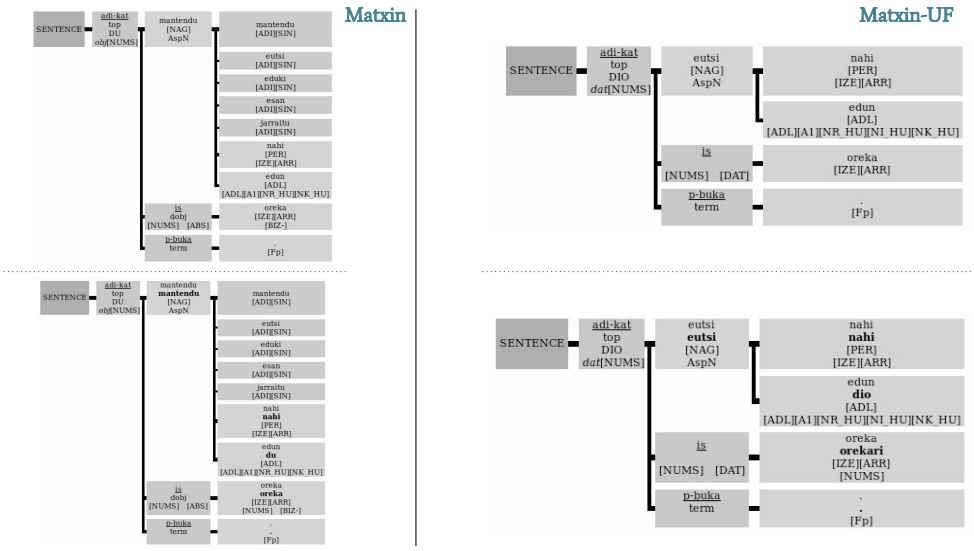
Azkenik, sorkuntza-faseak bukatzen du itzulpen-prozesua, ohiko moduan: sintaxiari eta morfologiari dagozkion moldaketak egiten ditu, UFari arreta berezirik jarri gabe (5.5. irudia).



5.4 irudia – *Mantener el equilibrio* UFaren itzulpen-prozesua: analisia eta transferentzia osoa (lexikoa eta estrukturala)

5.2.2 Emaitzak

Proposatu dugun metodoa *Matxin*entzat lagungarria den ala ez jakiteko, aztertutako UFei berariaz egokitutako corpus bat sortu dugu. Gaztelaniaren eta euskararen artean itzulitako testuak biltzen dituen corpus paralelo ba-



5.5 irudia – *Mantener el equilibrio* UFaren itzulpen-prozesua: analisia eta transferentzia osoa (lexikoa eta estrukturala)

tetik[1], guk aztertutako 153 UFen izenak eta aditzak barne hartzen dituzten esaldiak hautatu ditugu, eta 1.990 esaldi-pareko azpicorpus bat osatu dugu. Ondoren, bi eratara neurtu ditugu emaitzak: ebaluazio-metrika automatikoen bidez eta eskuz.

Ebaluazio-metrika automatikoen arabera

Metrika bat baino gehiago dago itzultzaile automatikoen kalitatea automatikoki neurtzeko, eta guk horietako hiru aukeratu ditugu: BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002) eta TER (Snover *et al.*, 2006). Horien funtzionamenduaren azalpen labur bana 2.2.4. atalean eman dugu, 61. orrialdean.

UFei buruzko informazioa sisteman gehitu ondoren, hiru ebaluazio-metriken emaitzak hobetu dira (5.1. taula), eta, hiruretatik, BLEUn lortu da alderik handiena bi sistemen artean, % 3koa –0,22 puntukoa–.

Nolanahi ere, esan beharra dago hobekuntza hori oso apala dela, bereziki kontuan izanik erabili dugun corpusa esperimentu honetarako sortu dugula

[1]Corpus horretan, berriak, administrazio-testuak eta saretik automatikoki erauzitako testuak biltzen dira.

Sistema	**BLEU**	**NIST**	**TER**
Matxin	7,28	3,88	84,36
Matxin-UF	7,50	3,90	84,27

5.1 taula – *Matxin*en eginiko lehen esperimentuaren emaitzak, BLEU, NIST eta TER metriken arabera

espreski, landutako konbinazioen itzulpena aztertzeko. Corpus orokor handiago bat erabiliko bagenu, 153 UFk oso agerpen gutxi izango lituzkete ziur asko, eta metrika automatikoetan lortutako hobekuntza ia hautemanezina izango litzateke.

Hortaz, lehen identifikazio-esperimentutik (4.2. atala) atera dugun ondorio berbera atera daiteke itzulpen automatikoari dagokionez ere: metodo-proposamena baliagarria izan daitekeela frogatu badugu ere, eskala handiagora eraman beharra dagoela.

Bestalde, eskuzko ebaluazioan erroreei ere begiratu diegu, proposamena hobetzeko ideiak biltze aldera.

Eskuzko ebaluazioaren arabera

Eskuzko ebaluazioak eman dizkigu etorkizunerako arrastorik interesgarrienak. Ebaluazio kualitatibo hori egiteko, bi sistemek desberdin itzulitako esaldi sorta erakusgarri bat eman diegu hiru ebaluatzaileri: (A) lehen hizkuntza euskara duen hizkuntzalari bati, (B) gaztelaniatik euskarara itzultzen duen itzultzaile bati eta (C) euskaraz eta gaztelaniaz arazorik gabe egin arren hizkuntza-ikasketa bereziturik ez duen hiztun bati. Gaztelaniazko esaldi bakoitzarekin batera, sistema baten eta bestearen itzulpenak erakutsi dizkiegu ausazko hurrenkeran, eta hiru aukera hauen arteko bat hautatzeko eskatu diegu:

- Lehen emaitza bigarrena baino hobea da.
- Bigarren emaitza lehena baino hobea da.
- Bi itzulpenak dira maila berekoak.

Ebaluatzaileen erantzunei begiratuta, nabaria da hobekuntza egon dela jatorrizko sistematik berrira (5.3. taula). Hala ere, beste ondorio interesgarri

bat ere atera liteke erantzunak ebaluatzailez ebaluatzaile aztertuta (5.2. taula): hizkuntzetan adituak direnentzat begi-bistakoak diren hobekuntza batzuk ez direla hain agerikoak hizkuntza-ikasketa bereziturik gabeko hiztunentzat.

Sistema onena	**A**	**B**	**C**
Matxin-UF	% 77,50	% 77,50	% 46,50
Matxin	% 8	% 6,50	% 40,50
Maila berekoak	% 14,50	% 16	% 13

5.2 taula – *Matxin*en eginiko lehen esperimentuaren emaitzak, ebaluatzailez ebaluatzaile

Izan ere, esaldi guztien % 43,52k kontraesanak sortu dituzte hiru ebaluatzaileen artean, baina kontraesan horietako % 78,57tan C anotatzailea izan da esaldia desberdin ebaluatu duena –esaldi guztien % 33tan, alegia–.

Euskara normalizazio-prozesuan egonik, ez da harritzekoa UF batzuen aurrean hiztun guztiek sentipen berbera ez izatea. Esate baterako, kontraesan gehien sortu dituen konbinazioa *dar pasos* izan da: *Matxin*en jatorrizko sistemak *urratsak eman* itzultzen zuen, eta UFen inguruko informazio gehigarria darabilenak, berriz, *pausoak eman*. Euskarazko tradizioan *urratsak egin* zein *pausoak eman* erabili izan dira batez ere, *urratsak eman*en agerpen bakanen bat ere badagoen arren –Orotariko Euskal Hiztegiaren arabera (Michelena, 1987), hegoaldeko autore moderno batzuen testuetan–. Gaur egungo testuetan begiratuta ere, lehen biak dira usuen agertzen direnak, baina hirugarrena ere gero eta sarriago erabiltzen da[2]. Beraz, ez da harritzekoa hizkuntzan adituek *pausoak eman* aukeratu izana, baina ezta beste hiztun askori *urratsak eman* normal-normala iruditzea ere.

Ebaluazioan egon dira era horretako kasu batzuk, non A eta B ebaluatzaileek bi aukeretako bat jo duten hobetzat eta C ebaluatzaileak bestea. Halakoetan ere hobekuntza badagoela uste dugu guk, nahiz eta hiztun guztientzat begi-bistakoa ez izan. Hortaz, oro har, sistema hobetu dela ulertu dugu honako kasuotan: hiru ebaluatzaileek *Matxin-UF*ren alde egin dutenean, bik berriaren alde egin eta hirugarrenak bi sistemak berdintzat jo dituenean, eta

[2]Ibon Sarasolak horren inguruko artikulu bat sareratu zuen 2004an, azalduz, hain zuzen, urratsak *egin* eta pausoak *eman* egiten direla, edo hala erabili izan dela euskarazko tradizioan behintzat. Duela 15 urte argitaratutako artikulu hartan ere aipatu zuen gero eta gehiago erabiltzen ari zela *urratsak eman* konbinazioa.

Sistema	**Hobe**	**Berdin**	**Desados**	**Okerrago**
Matxin-UF	% 76	% 10	% 1	% 13

5.3 taula – *Matxin*en eginiko lehen esperimentuaren emaitzak, osotara, eskuzko ebaluazioaren arabera

A eta B ebaluatzaileek sistema berriaren alde egin eta kontraesana C ebaluatzaileak sortu duenean. Horiek guztiak batuta, hobekuntza osoa % 76koa dela ondorioztatu daiteke eskuzko ebaluaziotik (5.3. taula)[3].

Bestetik, gaizki itzulitako esaldiei begiratzean, zerbaitez ohartu gara: zenbat eta gramatika-aldaketa gehiago egin UFa itzultzeko, orduan eta zailagoa da itzulpen txukunak lortzea. Hona hemen bi adibide.

(126) ES: *El barco ha* ***atracado en*** *un nuevo* ***puerto****.*

Matxin-UF: *Itsasontzia berri batean porturatu da.*

(127) ES: *No* ***echo en falta*** *el mar.*

Matxin-UF: ***Faltarik*** *ez dut* ***sumatzen*** *itsasoaren.*

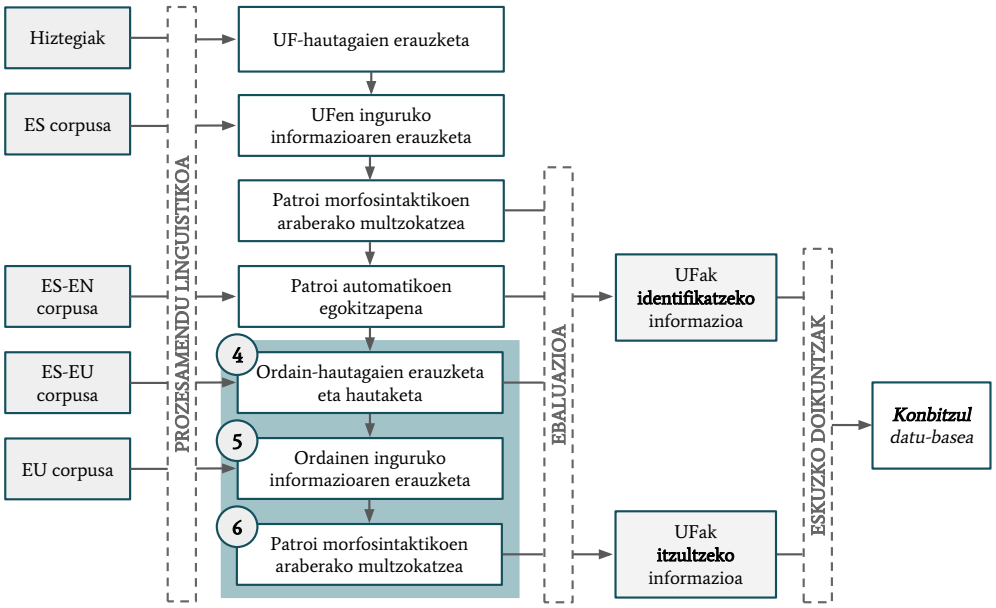
Esate baterako, bi adibide horietako lehenengoan (126), *atracar en puerto* UFari *porturatu* ordaina eman zaio, aditz bakarra, hitz-konbinazio bat eman beharrean. Gaztelaniazko izen-sintagman adjektibo bat ere badago, ordea, eta itzultzaileak, euskarazko esaldian *puerto*ren baliokiderik ez zegoenez, ez du jakin adjektibo hori zeri lotu. Bigarrenean (127), aldiz, informazio morfologikoa ondo aldatu da (*echar de menos (algo)* → *(zerbait)en falta sumatu*), baina, hitz-hurrenkera bere horretan gorde denez, ez da espero zen emaitza lortu.

Oro har, beraz, ondorio nagusi bat atera dugu itzulpen okerrak aztertzetik: gramatika-aldaketek akatsak sorrarazten dituztela sarri eta, halakoak saihesteko, hobe dela, posible denean, UFaren ezaugarri morfosintaktikoak gordetzea ordainean ere. Geroago erakutsiko dugunez (5.4. atala), hurrengo esperimentuan kontuan izan dugu ondorio hori.

[3]Taula honetako *Desados* zutabean, ez ditugu C ebaluatzaileak sortutako kontraesanak kontatu. Berez, horiek ere kontuan hartuta, % 43,52 esaldik sortu dute kontraesana.

5.3 Ordainak erdiautomatikoki erauzteko eta aztertzeko proposamena

Proposatu dugun metodoa baliagarria izan litekeela ikusi dugun arren, identifikazioa lantzean gertatu zaigun bezala, badakigu azterketa linguistikoa erabat eskuz egiteak eskalabilitate-arazoa sortzen digula itzulpen automatikoari dagokionez ere: informazio linguistikoak benetan eragina izan dezan, askoz ere UF gehiago landu beharra dago, baina dena eskuz aztertzeak denbora asko eskatzen du. Hortaz, oraingoan ere, azterketa linguistikoa erdiautomatizatzeko metodo bat sortu dugu.



5.6 irudia – Azterketa linguistikoa erdiautomatizatzeko metodoa (itzulpenari dagokion zatia nabarmenduta)

Aurreko kapituluan erakutsi dugu hiru urrats nagusitan egin dugula identifikaziorako azterketa erdiautomatikoa, eta 546 UF lortu ditugu prozesu horren amaieran: *Elhuyar* hiztegiko 282 eta *DiCE*ko 264. Itzulpenari dagokion azterketa ere beste hiru urratsetan banatu dugu, 5.6. irudian erakusten denez: ordain-hautagaiak erauzi eta hautatu ditugu lehenik (5.3.1. atala); ondoren, hautatutako ordainei buruzko informazioa erauzi dugu corpusetik

(5.3.2. atala); eta, azkenik, patroi morfosintaktikoetan sailkatu ditugu ordainak, erauzitako informazioaren arabera (5.3.3. atala). Azal dezagun hori guztia pausoz pauso.

5.3.1 Ordainen corpusetatiko erauzketa (4. urratsa)

UFen euskarazko ordainak lortzeko, 7 milioi esaldiko corpus paralelo batera jo dugu. Lehenik, UF bakoitzeko ordain posibleen sorta bat erauzi dugu *mGIZA* lerrokatze-tresna erabiliz (Gao eta Vogel, 2008), eta, jarraian, aukera guztien artetik bakarra hautatu dugu UFko. Ordain posible horiei *ordain-hautagai* deituko diegu, eta amaieran aukeratutakoari, berriz, *ordain hautatu*.

Lehenik, ordain-hautagaiak erauzteko, lerrokatze-tresna gure beharretara egokitu dugu. Azal dezagun, oro har, nola egin dugun lerrokatzea, kontuan izanik lehen pausoa orokorra dela –mGIZAk defektuz egiten duena, alegia– eta beste biak, aldiz, geuk erantsitakoak. Azalpenei laguntzeko, adibide bat ere gehitu dugu 5.7. irudian.

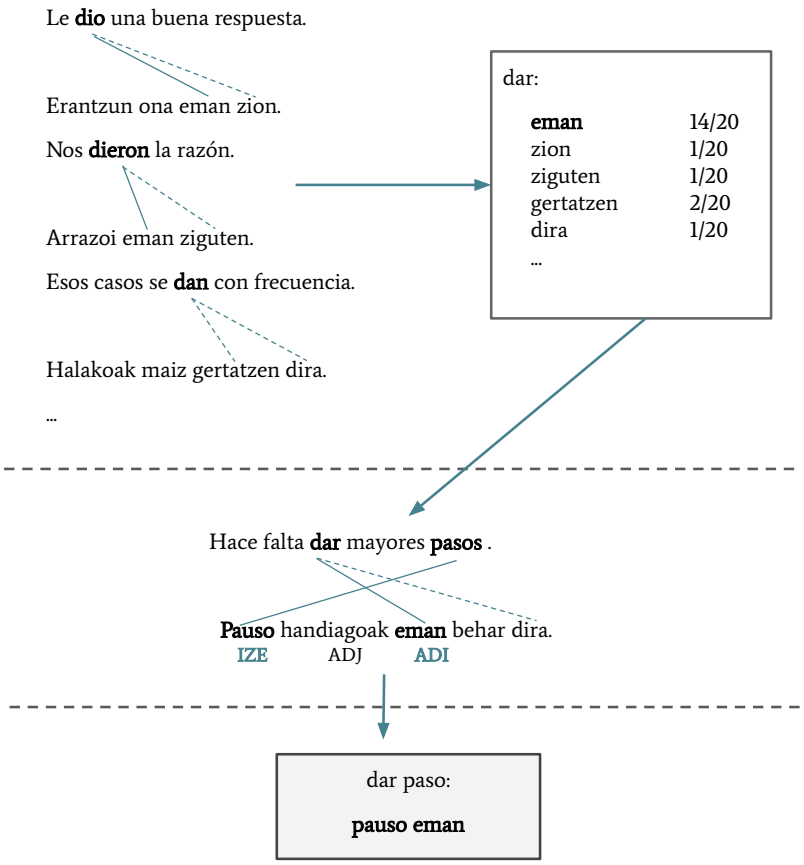
- Gaztelaniaren eta euskararen arteko corpus paralelo bat oinarritzat izanik, hizkuntza bateko eta besteko esaldiak parez pare hartu, eta gaztelaniazko hitzei euskarazko baliokide posibleak esleitzen joaten da *mGIZA*, baliokide posible bakoitzari dagokion probabilitatearekin batera. Prozesu hori hainbat bider errepikatzen da entrenamendu-corpus baten gainean, *Expectation Maximization* ikasketa-algoritmoa erabiliz, eta probabilitateak fintzen joaten dira hala.

- Entrenamendu-fasean ikasitakoa erabiliz, landutako UFak barne hartzen dituzten esaldi paraleloak lerrokatzen dira, *mGIZA*k lortutako hitz mailako lerrokatzeetan oinarrituz. Hala, esaldi bakoitzeko ordain-hautagai bat ematen zaio UF bakoitzari.

- Hirugarrenik, euskarazko hitzen kategoriei begiratu, eta iragazkiak aplikatu ditugu, batetik ordain-hautagaiak garbitzeko, eta bestetik bat etor daitezen, ahal denean, guk lehenetsi nahi ditugun egitura morfosintaktikoekin –hau da, *Matxin*ek errazago prozesatzen dituenekin–. Hau izan da erabili dugun leheneste-zerrenda:

  1. izena+aditza konbinazioak
  2. adjektiboa+aditza edo aditza+adberbioa konbinazioak

3. aditza bakarrik edo aditza beste edozein kategoriatako hitzekin
4. beste egitura morfosintaktiko batzuk



5.7 irudia – Ordain-hautagaiak ateratzeko erabili dugun metodologiaren adibide bat

Prozesu horren guztiaren ondoren, UFen agerpen bakoitzari ordain bana esleitzen zaio automatikoki, eta, hala, UF bakoitzeko hainbat ordain-hautagai biltzen dira. Hautagai horiekin batera, agerraldiei dagozkien ehunekoak ere jartzen dira.

Demagun, adibidez, *generar confianza* konbinazioa 255 aldiz agertu dela corpus paraleloan eta agerpen horien arteko 202ri *konfiantza sortu* ordaina esleitu zaiela lerrokatze-tresnaren bidez. Bada, UFaren ordain-hautagaiak biltzean, *konfiantza sortu* % 79,21etan agertu dela jasotzen da (5.8. irudia).

Behin UFen ordain-hautagaiak bilduta, haien artetik *Matxin*erako egokiena zein den hautatzen da. Horretarako, garbiketa pixka bat egin behar izan dugu lehenik, batere aditzik gabeko hautagaiak baztertzeko eta, hitz edo hitz-konbinazio berberaren aldaki bat baino gehiago hautagai desberdinetan jaso badira, denak ordain-hautagai berean biltzeko. UF bakoitzarentzat agerraldi gehien zituen ordaina hautatu dugu azkenean, eta horrekin egin dugu aurrera.

GENERAR - *obj* - - **CONFIANZA**

konfiantza (abs) sortu	% 79,21
konfiantza (abs) sorrarazi	% 6,67
konfiantza (abs) galdetu	% 3,53
konfiantza (abs) areagotu	% 3,53
konfiantza (abs) eman	% 3,53
konfiantza	% 3,53

5.8 irudia – Ordainen aukeraketaren adibide bat

Ordain-erauzketaren ebaluazioa

Atal honen hasieran esan dugun bezala, automatikoki erauzitako ordainen kalitatea neurtu dugu hurrena. UF batzuei ez zaie ordainik eman (*Elhuyar*reko 70i eta *DiCE*ko 52ri), corpus paraleloan agertu ez direlako gehienetan, eta, beste zenbait kasutan, erauzitako ordainak baztergarriak zirelako, batere aditzik gabeak. Izan ere, corpusaren tamainak garrantzi handia du ordain-erauzketaren kalitatean, eta, corpusa txikia denean, arriskua dago oso maiztasun handikoak ez diren UF batzuk batere ordainik gabe geratzeko (Maniez, 2001).

Ordain automatikoak eskuz zuzendu ditugu, eta, aldaketarik egin badugu, aldaketa hori nolakoa izan den ere zehaztu dugu: lexiko mailakoa ala gramatika-ezaugarriei zegokiena[4]. Ordainik gabe geratu diren UFak alde batera utzita, 5.4. taulan jaso ditugu emaitzak.

	Elhuyar		**DiCE**	
	Ordainak	%	**Ordainak**	%
Zuzen	98	46,23	102	48,11
Aldaketa lexikoan	100	47,17	101	47,64
Aldaketa gramatikan	14	6,60	9	4,25

5.4 taula – Ordain-erauzketaren ebaluazioko emaitzak

Taulan ikusten denez, automatikoki lortutako ordainen ia erdia izan dira zuzenak, eta, eskuz zuzendu ditugunen artean, gehien-gehienek lexiko mailako aldaketak behar izan dituzte. Emaitzak ez dira bereziki onak, baina esan beharra dago erauzketaren kalitatea corpus paraleloaren araberakoa dela erabat. UF jakin batzuentzat ordainik lortu ez izana corpusaren tamainari zor zaio, eta eskuz egin beharreko zuzenketa kopurua ere askoz ere handiagoa izan da agerraldi gutxi izan dituzten UFen kasuan. Hortaz, oro har, pentsatzekoa da zenbat eta corpus handiagoa izan hainbat eta hobea izango dela ordain-erauzketaren kalitatea ere.

Etorkizunean, interesgarria litzateke urrats hau errepikatzea, batetik, beste lerrokatze-tresna batzuk erabilita, emaitzak antzekoak ote diren ikusteko, eta bestetik, corpus paralelo handiagoren bat erabilita. Momentuz, baina, eskuz hobetutako konbinazioekin nahikoa dugu ezaugarri morfosintaktikoak aztertzeari ekiteko.

5.3.2 Ordainen inguruko informazioaren erauzketa corpus elebakarretatik (5. urratsa)

Euskarazko ordainei buruzko informazio morfosintaktikoa corpusetik erauzteko, lehenik, bi multzotan bereizi ditugu ordainak: izena+aditza motakoak

[4]Batzuetan, gertatzen da lexikoa ondo egon arren informazio gramatikala zuzena ez izatea, eta halakoei dagokie etiketa hau. Esate baterako, *tratar con respeto – errespetu tratatu* lerrokatzen da, baina euskarazko izenak postposizio-marka instrumentala behar du.

eta bestelakoak. Izena+aditza motako konbinazioekin egin dugu aurrera urrats honetan, eta gainerakoak alde batera utzi ditugu oraingoz, gramatika-kategoria zehaztuta. Esate baterako:

(128) ES: ***ser*** *un* ***consuelo***
EU: *kontsolagarri izan* → adj+adi

(129) ES: ***dar alivio***
EU: *lasaitu* → adi

Ondoren, izena+aditza motako ordainen inguruko datuak erauzi ditugu corpusetik. Gaztelaniazko UFen informazioa erauzteko erabili dugun metodologia bera (4.4.1. atala) erabili dugunez, aztertu ditugun ezaugarriak ere ia berdinak izan dira:

- ISaren numeroa: singularra (Sing.) ala plurala (Pl.)
- ISko determinatzaileak (Det.)
- ISaren mugatasuna: mugagabea (Ind.) ala ez mugagabea (Def.)
- ISaren barruko modifikatzaileak (Mod.)
- Aldaketak osagai hitzen arteko hurrenkeran (Ord.)

Dena dela, hizkuntza batetik besterako jauzian, zenbait xehetasun moldatu behar izan ditugu programan. Izan ere, gaztelaniaz, definitutasunari dagozkion ehunekoak determinatzailearen arabera gorde ditugu, hau da: analizatzaileak emandako informazio linguistikoan bilatu dugu determinatzailea zehaztua ala zehaztugabea den. Euskaraz, ordea, mugatasunaren inguruko datuak dira itzultzaile automatikoan sartu nahi ditugunak, eta informazio hori bestela bilatu beharra dago: izen-sintagma izen batez bakarrik osatua denean, izenak eraman ohi du numeroari eta mugatasunari buruzko informazioa; osagai gehiagoko izen-sintagmetan, aldiz, beste elementuren batek izan lezake informazio hori (Hualde *et al.*, 2003: 135. orr). Hortaz, honela jokatu dugu: numeroari eta mugatasunari buruzko informazioa izenean bilatu dugu zuzenean, eta, izenak halakorik izan ez duenean, dependentzia-zuhaitzean beherantz jarraitu dugu bila, izen-sintagmaren barruan erantzuna aurkitu arte.

Bestalde, aipatu beharra dago euskarazko determinatzaile mota guztietatik (Hualde *et al.*, 2003: 92. orr.) hitz beregainak direnak bakarrik kontatu ditugula bigarren ezaugarrian, eta alde batera utzi ditugula, praktikotasunagatik, izen-sintagmako osagaiei itsatsita agertu ohi diren artikulu mugatuak (*-a(k)*). Horrez gain, mugatu-mugagabe bereizketa egiteko ere, moldaketa bat egin dugu: mugagabeak bildu ditugu batetik (*pauso eman* eta halakoak), eta mugagabe ez diren guztiak bestetik (*pausoak eman, pauso bat/asko eman* eta halakoak). Determinatzaileei buruzko informazioarekin eta mugatasunari buruzkoarekin, ondoriozta daiteke IS jakin batek noiz duen benetan artikulua izenari itsatsita: ez bada mugagabea eta ez badu hitz beregaina den determinatzailerik, mugatua da, eta izenari itsatsita darama artikulua.

Horretaz landa, ez dugu beste aldaketarik egin, eta informazioa tauletan gorde dugu, gaztelaniaz bezalaxe. Taula horien adibide bat 5.9. irudian dago jasota.

	IS numeroa			Mugatasuna			
	Sing.	**Pl.**	**Det.**	**Def.**	**Ind.**	**Mod.**	**Ord.**
AIRE * **ine** ccomp **EGON**	100	0	0	100	0	17,78	4,44
AUKERA * **abs** obj **APROBETXATU**	42,23	44,44	52,22	86,67	13,33	32,22	28,89
ESKU * **abs** obj **GARBITU**	0	100	0	100	0	33,33	44,44
HITZ * **abs** obj **EMAN**	39,77	1,14	3,41	40,91	59,09	19,32	12,50
PIKU - **ala** ccomp **BIDALI**	0	0	0	0	100	0	0

5.9 irudia – Bosgarren urratsean sortzen diren taulen adibide bat (zenbakiak, ehunekotan)

Esan beharra dago, dena den, eskuzko azterketan landutako ezaugarri guzti-guztiak ez direla agertzen hemen zerrendatu ditugun ezaugarrietan. Hain zuzen ere, 5.1.2. atalaren amaiera aldera aipatutako hirurak falta dira: kanpo elementu irekiei dagokiena, forma pronominalaren itzulpenari dagokiona eta ezezkoetako partitiboari dagokiona. Hiru ezaugarri horiek azterketa erdiautomatikoan ez sartzea erabaki dugu, itzultzaile automatikoetan haiek orokortzeko beste modu bat pentsatu dugulako. Geroago argituko dugu zer-

tan datzan orokortze hori (5.4 atala), baina, horren aurretik, azal dezagun zer egin dugun urrats honetatik ateratako informazioa sailkatzeko.

5.3.3 Ordainen patroikako multzokatzea (6. urratsa)

Gaztelaniazko UF-hautagaiekin egin dugun bezala (4.4.2. atala), euskarazko ordainak ere bi multzotan banatu ditugu sailkapen-prozesurako[5]: batetik, 268 ordaineko sorta bat, probak egiteko, zer patroi sortu erabakitzeko eta metodologia fintzeko; eta bestetik, 267 ordaineko beste bat, ebaluaziorako.

Lehen bezala, bi fasetan egin dugu sailkapena. Hasteko, aurreko urratseko ehunekoei atalaseen araberako balioak eman dizkiegu: Y (bai, ezaugarri hori beti agertzen da era horretan ordainaren agerraldietan), O (aukerakoa da; ezaugarri hori era batera edo bestera ager daiteke ordainaren agerraldietan), edo N (ez, ezaugarri hori ez da inoiz agertzen era jakin horretan ordainaren agerraldietan)[6].

Ondoren, balio horiek multzokatu, eta sei patroi sortu ditugu (5.5. taulan daude patroi bakoitzari dagozkion balioak), denak ere ordainen ezaugarri morfologikoetan oinarrituak. Aurreko urratseko adibide berberak hizpide hartuta, 5.10. irudian ikus daitezke sailkapen-prozesuaren emaitza batzuk.

	IS numeroa		**Det.**	**Mugatasuna**	
	Sing.	**Pl.**		**Def.**	**Ind.**
FREE	O/N	O/N	Y/O/N	O/N	Y/O/N
IND	N	N	N	N	Y
SING_DEF	Y	N	N	Y	N
SING	Y	N	Y/O	Y/O/N	N
PL_DEF	N	Y	N	Y	N
PL	N	Y	Y/O	Y/O/N	N

5.5 taula – Izena+aditza motako ordainen patroi morfosintaktikoak

[5]Zenbaki hauetan, izena+aditza motakoak ez diren ordainak ere –aurreko urratsean behin-behinean alde batera utzitakoak– kontuan hartu ditugu. Izan ere, aurreko urratsean izena+aditza motako ordainen informazioa bakarrik erauzi dugu corpusetik, baina hemendik aurrerakoa ordain guztiei dagokie.

[6]Identifikazioa lantzean egin dugun bezala, hainbat atalaserekin egin ditugu probak, eta hemen ere 90 eta 10 inguruko atalaseek eman dituzte emaitzarik onenak.

	IS numeroa			Mugatasuna		
	Sing.	Pl.	Det.	Def.	Ind.	
AIRE * **ine** ccomp **EGON**	Y	N	N	Y	N	SING_DEF
AUKERA * **abs** obj **APROBETXATU**	O	O	O	O	O	FREE
ESKU * **abs** obj **GARBITU**	N	Y	N	Y	N	PL_DEF
HITZ * **abs** obj **EMAN**	O	N	N	O	O	FREE
PIKU - **ala** ccomp **BIDALI**	N	N	N	N	Y	IND

5.10 irudia – Seigarren urratsean sortzen diren taulen adibide bat

Patroi horiez gainera, aurreko urratsean behin-behinean alde batera utzitako ordainak ere berreskuratu ditugu, hau da, izena+aditza motakoak ez zirenak. Haiei osaera morfologikoa jarri diegu zuzenean patroitzat, ezaugarri hori izango baita itzultzaile automatikoan sartu beharrekoa.

Sailkapenaren ebaluazioa

Ebaluaziorako, esan bezala, 267 ordain erabili ditugu. Hala ere, horietako 61 ez dira euskarazko corpusean 10 aldiz baino gehiagotan agertu, eta ezin izan dugu haien inguruko informaziorik erauzi. Hortaz, horiek alde batera utzita egin dugu ebaluazioa, eta 5.6. taulako emaitzak lortu ditugu, ehunekotan eta Cohen κ (Cohen, 1960) erabilita.

	Sorta osoa		**Izena+aditza motakoak**	
	Ordainak	%	**Ordainak**	%
Zuzen	150	72,82	110	66,27
Oker	56	27,18	56	33,73
Cohen κ	0,62		0,53	

5.6 taula – Ordainen sailkapen automatikoaren emaitzak, ehunekotan eta Cohen κ-ren arabera

Emaitzak bi eratara kalkulatu ditugu: sorta osoa kontuan hartuta ba-

tetik, eta izena+aditza motako ordainak bakarrik kontuan hartuta bestetik. Espero izatekoa denez, sorta osoa aintzat hartuta lorturiko emaitzak hobeak dira, izena+aditza motakoak ez diren 40ei osaera morfologikoa zuzen jarriz gero automatikoki esleitzen baitzaie patroi zuzena. Dena dela, izena+aditza motakoei bakarrik begiratuta ere bi heren sailkatzen dira ondo, eta Cohen κ ere nahiko onargarria da. Ezin esan emaitzak bere horretan eta inongo iragazkirik gabe *Matxin*en sartzeko modukoak direnik, baina badira erabilgarriak hiztegietatik haragoko informazioa berrerabiltzeko eta eskuzko azterketa arintzeko. Gainera, datorren atalean erakutsiko dugunez, informazio erabat automatikoa ere oso lagungarria gerta daiteke, baldin eta kalitatea bermatzeko maiztasun-iragazkiak jartzen badira.

5.4 Bigarren esperimentua *Matxin*en, erdiautomatikoki lortutako datuak erabiliz

Prozesu horren guztiaren ondotik, bigarren esperimentu bati ekin diogu, eskuzko datuak ez ezik erdiautomatikoki bildutakoak ere erabiliz. Horiez gain, bigarren identifikazio-esperimentuan bezala (4.5. atala), erabat automatikoki lortutako datu multzo bat ere erabili dugu, itzulpen-kalitateari nola eragiten dion ikusteko. Azal dezagun, atal honetan, zer metodologia erabili dugun zehazki (5.4.1. azpiatala) eta zer emaitza lortu ditugun (5.4.2. azpiatala).

5.4.1 Erabilitako metodologia

Datu linguistikoen tratamendua pixka bat orokortze aldera, egokitzapen batzuk egin ditugu aurreko esperimentutik (5.2. atala) hona. Batetik, identifikaziorako eta itzulpenerako datuak orokortu ditugu, azterketa-prozesu automatikorako zehaztutako patroi morfosintaktikoen bidez. Hala, bai UFen agerpenak identifikatzeko eta bai identifikatutako UFei ordaina emateko, *Matxin*ek patroi bakoitzari dagozkion murriztapenei bakarrik begiratu beharko die orain, ezaugarriz ezaugarri joan beharrean.

Bestetik, ordainen patroietan kontuan hartzen ez diren hiru ezaugarriri ere erregela orokorragoak aplikatu dizkiegu. Izan ere, euskarazko ordainetan ezaugarri askotxo zehaztu ditugu aurreko esperimentuan (5.2. atala), eta ikusi dugu, zenbat eta gramatika-informazio gehiago aldatzen saiatu UF bakoitzeko, *Matxin*ek itzulpen zuzenak sortzeko aukerak jaitsi egiten zirela. Horrexegatik erabaki dugu UFetako osagai irekiei zegozkien erregelak,

partitiboari zegozkionak eta gaztelaniazko aditzen erabilera pronominalari zegozkionak orokortzea.

- **Osagai irekiek** sortzen dituzten arazoetako batzuk konpontzeko, honako erregela hau sortu dugu: gaztelaniazko osagaien artean objektuerlaziorik ez badago baina euskarazko osagaien artean bai, eta gaztelaniazko UFarekin objektu zuzen bat ageri bada, euskaraz zehar-objektu bihurtzen da osagai hori.

  (130) ***CASTIGAR*** - *ccomp* ***CON*** * ***PENA*** *(a alguien)*
  → ***ZIGOR*** * *abs obj* ***EZARRI*** *(norbaiti)*

  (131) ***TENER*** - *ccomp* ***EN*** * ***ESTIMA*** *(a alguien)*
  → ***ESTIMU*** * *abs obj* ***UKAN*** *(norbaiti)*

  Erregela hau noiz aplikatu erabakitzeko UFaren osagaien arteko erlazio sintaktikoa hartzen denez kontuan, kasu batzuetan, baliteke erregela hori aplikatzea gaztelaniazko UFak osagai irekirik eduki ez arren. Hala izanik ere, erregelak ez du kalterik eragiten, ez duelako benetan ezer aldatzen.

  (132) ***REQUERIR*** - *ccomp* ***DE*** * ***ATENCIÓN***
  → ***ARRETA*** * *abs obj* ***ESKATU***

- **Partitiboaren erabilera**ri dagokionez, ezezkoetan izen-sintagmari partitiboa jartzen zaio oro har, bai UFetan eta bai bestela. Hala ere, erabaki dugu salbuespentzat tratatzea euskaraz mugagabean bakarrik erabiltzen diren eta aditza *izan/ukan* duten UFak.

  (133) *Ez* ***dut nahi*** *(*nahirik)*

  (134) *Ez* ***naiz bizi*** *(*bizirik)*

  (135) *Ez* ***dut merezi*** *(*merezirik)*

  Hemen ere, salbuespena era horretan orokortuz gero, testuinguru batzuetan partitiboa onartzen duen UFren bat ere jarriko da partitiborik gabe ezezkoetan, baina, gure azterketaren arabera, oso kasu bakanetan lekarke horrek kalterik.

  (136) *Ez* ***dut behar/beharrik***

- Hirugarren araua, gaztelaniazko **aditz pronominalekin** lotura duena, denetan orokorrena da, ez baitie UFei bakarrik eragiten, baizik eta itzulgai guzti-guztiei. Freeling analizatzaileak halakoei etiketa bat edo beste ematen die, testuinguruaren arabera, estatistika kontuan harturik. *Matxin*ek, etiketa horiei erreparatzen dienez, ez ditu aditz pronominalak beti berdin itzultzen:

  (137) ***Se gana la vida*** → ***Bizimodua ateratzen*** *da*.

  (138) ***Se juega el tipo*** → ***Bizia arriskatzen*** *da*.

  (139) *Se busca trabajo* → *Lana bilatzen du*.

  Gehiegizko orokortzeek sistemari kalte egin liezaiokeela iruditzen zaigu, eta, hortaz, kasu gutxi –baina ziur– batzuk bakarrik konpontzeko erregela bat gehitu dugu *Matxin*en: aditz pronominalak bai subjektua eta bai objektua agerian baditu itzulgaian, esaldi horretan aditza trantsitiboa izango da euskaraz. Agerian ez baditu, berriz, ez dugu ezer egingo, inpertsonalak diren zenbait adibidetan hanka-sartzeak eragin litzake eta.

  (140) *(Alguien)* ***se toma*** *su* ***tiempo*** → *(Norbaitek) bere* ***denbora hartzen*** *du*.

  (141) *(Alguien)* ***se juega el tipo*** → *(Norbaitek)* ***bizia arriskatzen*** *du*.

  (142) *Se* ***da prioridad*** *a los casos más graves* → *Lehentasunezko* ***arreta ematen*** *zaie kasurik larrienei*.

Erregela horiek transferentzia-fasearen amaieran aplikatzen dira, sorkuntzaren aurretik. Hori kenduta, erabili dugun metodologia 5.2. ataleko berbera izan da, hau da: UFen identifikazioa eta itzulpena analisi-fasearen ondoren egin ditugu, transferentzia-fasearen aurretik.

Esperimentu honetarako erabili dugun datu multzoari dagokionez, eskuz landu ditugun UFez eta ordainez gain (4.1. eta 5.2. atalak), prozesu erdiautomatikoan aztertutako UFak eta ordainak ere erabili ditugu oraingoan. Horiez gain, erabat automatikoki ere aztertu ditugu PARSEMEren corpuseko eta *DiCE* hiztegiko zenbait hitz-konbinazio, eta horiek ere sartu ditugu *Matxin*en, zer-nolako eragina zuten ikusteko. Bitan banatu ditugu azken horiek: corpusean hamar agerraldi baino gehiago izan dituzten hitz-konbinazioak eta

ordainak batetik, eta agerraldi gutxiagokoak bestetik. Honela geratu da, azkenean, 1.108 UFko eta beste hainbeste ordaineko datu multzo osoa:

- Eskuz aztertutakoak: 133
  (5.2. atalean landutakoak, baina errepaso-lana egin ondoren gutxi batzuk baztertuta, esanahi aldetik oso anbiguoak izanik sistemari on baino kalte handiagoa ekarri diotelakoan)
- Erdiautomatikoki aztertutakoak: 535
  (5.3. atalean landutakoak, errepikatuak baztertuta)
- Automatikoki aztertutakoak, baina corpus elebidunean hamar aldiz baino gehiagotan agertu direnak: 226
  (PARSEMEren corpusetik eta *DiCE*tik lortutako konbinazioak, automatikoki lortutako ordainekin batera)
- Automatikoki aztertutakoak, corpusean hamar aldiz baino gutxiagotan agertu direnak: 214

Ikus dezagun, bada, zer emaitza lortu ditugun proposamen findu honekin, eta ba ote dagoen alderik datuen azterketa-moduaren arabera.

5.4.2 Emaitzak

*Matxin*en egin dugun lehen esperimentuan bezala, oraingoan ere bi eratara kalkulatu ditugu emaitzak: BLEU, NIST eta TER metrika automatikoak erabiliz[7], eta eskuz, hiru adituri iritzia eskatuta.

Metrika automatikoen arabera

Metrika automatikoen emaitzak bost eratara kalkulatu ditugu, datu multzo bakoitzak sisteman zer eragin daukan ikusteko. Hona hemen nola deitu diogun sistemaren bertsio bakoitzari:

- ***Matxin***: jatorrizko sistema
- ***Matxin+***: jatorrizko sistema, 5.4.1. atalean azaldu ditugun erregela orokorrak gehituta

[7]Metrika bakoitzari buruzko azalpenak, 61. orrialdean.

- ***Matxin-UFeskuz***: eskuz edo erdiautomatikoki landutako 668 UFen eta ordainen datuak erabilita (173. orrialdeko puntuetatik, lehen eta bigarren multzokoak)
- ***Matxin-UFauto1***: aurrekoez gain, automatikoki baina maiztasun-iragazkia jarrita landu ditugun 226 UFen eta ordainen datuak erabilita (173. orrialdeko puntuetatik, lehen hiru multzoetakoak)
- ***Matxin-UFdenak***: datu guzti-guztiak erabilita, aurretik aipatutako guztienak eta maiztasun-iragazkirik gabe automatikoki landutakoenak (173. orrialdeko puntu guztiak)

Erreferentziatzat erabili dugun corpusak 21.786 esaldi-pare biltzen ditu (gaztelaniazko 746.370 hitz eta euskarazko 517.921), eta denek ere datu multzoko UFren bateko izena eta aditza dituzte barnean –nahiz eta agerpen horiek ez izan beti UFak–. Emaitzak 5.7. taulan jaso ditugu.

Sistema	**BLEU**	**NIST**	**TER**
Matxin	7,08	4,04	85,90
Matxin+	7,17	4,05	85,44
Matxin-UFeskuz	7,23	4,07	85,34
Matxin-UFauto1	7,24	4,07	85,31
Matxin-UFdenak	7,24	4,08	85,30

5.7 taula – *Matxin*en eginiko bigarren esperimentuaren emaitzak, BLEU, NIST eta TER metriken arabera

Ikusten denez, hiru metriken arabera hobetzen dira emaitzak, pausoz pauso: jatorrizko sistematik erregela orokorrak dituenera, bigarren horretatik eskuzko datuak darabiltzanera, eta hirugarren horretatik informazio automatikoa integratuta duen bertsiora. BLEUren arabera, hobekuntza osoa % 2,25ekoa da, eta hobekuntzarik txikiena automatikoki landutako azken datu multzoak dakar, maiztasun-iragazkirik gabe lortutako datuek, alegia.

Hortaz, badago hobekuntza jatorrizkotik gure proposamena kontuan hartzen duen bertsiora, baina, lehen esperimentuan bezala, hobekuntza hori oso txikia da. Nolanahi ere, komeni da bi ohar kontuan hartzea, emaitzak behar bezala interpretatzeko. Batetik, jatorrizko sistemaren emaitzak hain apalak izanik, zaila dela fraseologia bezain aztergai konplexu baten bidez emaitza horiek hobetzea. Izan ere, *Matxin*en BLEU marka oso baxua da eredu neuronalak darabiltzaten sistema berriagoen aldean; MODELA gaztelania-euskara

itzultzailearena, adibidez, 30etik gorakoa da (Etchegoyhen *et al.*, 2018), *Matxin*ena halako laukoa baino altuagoa. Eta bestetik, metrika automatikoak ez direla oso egokiak fraseologia bezain fenomeno konplexua ebaluatzeko (Constant *et al.*, 2017), sistemaren emaitzak erreferentzia-corpus bakarrarekin eta estatistika hutsez alderatzeak ezer gutxi esan baitezake itzulpenetako fraseologiaren kalitateaz.

Ikus dezagun, beraz, zer gehiago esan dezakegun eskuzko ebaluazio-lanetik abiatuta.

Eskuzko ebaluazioaren arabera

Hiru ebaluatzailek hartu dute parte ataza honetan: lehen hizkuntza euskara duten bi hizkuntzalarik eta itzultzaile batek. Metrika automatikoekin egin bezala, ebaluazio hau ere datu multzoen arabera antolatu dugu, multzo bakoitzak sistemari zer hobekuntza dakarkion ikusteko. Horretarako, gorago aipatutako sistemek desberdin itzulitako esaldiak bildu[8], eta honela osatu dugu ebaluaziorako seta:

- *Matxin+* sistemak era batera eta *Matxin-UFeskuz* sistemak beste era batera itzuli dituzten 150 esaldi-pare
- *Matxin-UFeskuz* sistemak era batera eta *Matxin-UFauto1* sistemak beste era batera itzuli dituzten 150 esaldi-pare
- *Matxin-UFauto1* sistemak eta *Matxin-UFdenak* sistemak desberdin itzuli dituzten 50 esaldi-pare

Hala, ebaluatzaileei binaka eman zaizkie sistema baten eta bestearen emaitzak, ausazko hurrenkeran, eta hiru aukeraren arteko bat markatu behar izan dute: lehen sistema hobea den, bigarren sistema hobea den, ala bi sistemen emaitzek duten kalitate berbera. Oraingo honetan, gainera, aurreko esperimentuko (5.2.2. atala) ebaluatzaileen gomendioei jarraituz, gidalerro moduko batzuk sortu ditugu ebaluatzaileentzat, irizpide orokor batzuk izan zitzaten; C. eranskinean bildu ditugu gidalerro horiek.

Ebaluazioaren emaitzak 5.8. taulan daude jasota. Ikusten denez, eskuzko datuak edo maiztasun-iragazkidun datu automatikoak erabilita, hobekuntza

[8]Guztira, corpuseko 21.786 esaldietatik, 6.093 aldatu dira *Matxin+* sistematik UFen informazioa kontuan hartzen duten beste hiruretara: 4.527 *Matxin-UFeskuz* sistemara, 1.472 *Matxin-UFauto1* sistemara, eta 94 *Matxin-UFdenak* sistemara.

nabaria da, eta oso gutxitan sortzen dira aurreko sisteman baino emaitza okerragoak. Maiztasun-iragazkirik gabeko datu automatikoek, ostera, onura baino askoz ere galera gehiago dakarte.

Sistema	**Hobe**	**Berdin**	**Desados**	**Okerrago**
Matxin-UFeskuz	% 62	% 11	% 19	% 8
Matxin-UFauto1	% 65	% 12	% 16	% 7
Matxin-UFdenak	% 14	% 14	% 16	% 56

5.8 taula – *Matxin*en eginiko bigarren esperimentuaren emaitzak, giza ebaluatzaileen arabera

Hortaz, ikusten da landu ditugun datu linguistikoak lagungarriak direla itzultzaile automatikoek UFak hobeto prozesa ditzaten. Dena dela, kontuan hartu behar da ebaluatzaileek UFei bakarrik begiratu diotela ataza honetan eta, askotan, UFari ordain hobea emanda ere, sistemaren itzulpenek, oro har, lehen bezain traketsak izaten jarraitzen dutela, bereziki itzulgaia konplexua denean (143. adibidea).

(143) ES: *Esta tarea de dos semanas* ***despertó*** *en él la* ***fascinación*** *por el tema circense, que finalmente dio como resultado su Circo Calder (Cirque Calder), una performance en la que intervenían figuras creadas con alambre.*

Matxin+: *Bi astez lan honek hartan esnatu zuen zirku gaiagatik lilura, azkenean eman zuen emaitza haren Circo Calder (Cirque Calder), figura sortuak esku hartzen zituzten alanbrearekin performance bat.*

Matxin-UFauto1: *Bi astez lan honek hartan* ***eragin*** *zuen zirku gaiagatik* ***zirrara****, azkenean eman zuen emaitza haren Circo Calder (Cirque Calder), figura sortuak esku hartzen zituzten alanbrearekin performance bat.*

Izan ere, ebaluazio-metrika automatikoek agerian uzten duten bezala, UFen inguruko datuek corpus osoan sortzen duten aldaketa ez da estatistikoki oso esanguratsua, eta hori, jakina, halaxe islatzen da esaldien ulergarritasun osoari dagokionez ere. Gainera, fraseologia, askotan, zuzentasunarekin baino gehiago, hizkuntza baten egokitasunarekin dago lotuta, estiloarekin kasik, eta halakoek benetako eragina izan dezaten, beharrezkoa da oinarrian duten testuak gutxieneko maila bat izatea. Oinarrizko gramatika-kontuetan

ere akatsak sarri samar egiten dituen sistema batez ari garelarik, ez da harritzekoa tarteka agertzen diren hitz-konbinazio batzuek dakarten onura oso handia ez izatea.

Laburpena

Aurreko kapituluan gaztelaniazko UFei eta haien identifikazioari buruz jardun ostean, UF horien euskarazko ordainei eskaini diegu bosgarren kapitulu hau. Eskuzko azterketa xehe bat eta azterketa erdiautomatiko bat egin ditugu, eta horietatik ateratako informazio linguistikoa integratu dugu *Matxin* itzultzailean, bi esperimenturen bidez. Laburpen gisa, azal dezagun nola lotzen den lan hori 1.3. ataleko hipotesiekin.

[A1] UFak, askotan, ez dira hitzez hitz itzultzen hizkuntza batetik bestera. Kapitulu honetan azterketa kuantitatiborik egin ez badugu ere, euskarazko ordainen gainean egin dugun azterketak argi erakusten du hori, hainbat arrazoirengatik: batetik, hiztegien gaineko azterketan ikusi dugun bezala (3. kapitulua), aditza+izena motako UFen ordain guztiak ez direlako izena+aditza motakoak euskaraz; eta bestetik, izena+aditza motako ordainen artean ere, informazio lexikoa edota morfosintaktikoa esplizituki zehaztean, ordainaren ezaugarriak askotan urrundu direlako gaztelaniazko UFaren ezaugarrietatik. Gainera, ordainak corpusetik automatikoki erauzi ditugunean ere, erdia inguru bakarrik erauzi dira erabat zuzen, UFen parte diren osagaiak parekatzea ataza konplexua den seinale.

[A6] UFei buruzko informazio morfosintaktikoa kontuan hartzea onuragarria izan daiteke itzultzaile automatikoentzat.

Egin ditugun bi esperimentuetan ikusi dugunez, hala da, aztertu dugun informazio linguistikoa *Matxin* itzultzaile automatikoan integratzean hobetu egiten baita itzulpenen kalitatea, bai metrika automatikoen eta bai giza ebaluatzaileen arabera. Nolanahi ere, hobekuntza hori txikia da estatistikoki (% 2,25 BLEU neurriaren arabera), eta, jatorrizko sistemaren itzulpena traketsa denean, UFen ordainak zuzenduta ere ez da itzulpen osoaren kalitatea askorik aldatzen. Emaitzak benetan txukunak izan daitezen, sistemaren oinarrizko gramatika-akatsak konpondu beharko lirateke lehenik, bestela, UFei ordain egokia emanda ere, askotan sortzen baitira itzulpen ulergaitzak, bereziki gaztelaniazko itzulgaia konplexua bada.

Hipotesi horiekin loturik, kapitulu honetako edukiek tesi-lanaren bi helburu betetzen lagundu digute.

[H2] Aditza+izena motako UFak gaztelaniaren eta euskararen artean nola itzultzen diren aztertzea. Azterketa horixe izan da kapitulu honen ardatza, eta, egin ditugun lanen ostean, 894 UFri buruzko informazio xehea lortu dugu, erabat eskuz (133), erdiautomatikoki (535), edo automatikoki baina kalitatea bermatzeko iragazkiak aplikatuta (226). Bildu ditugun datuen artean daude, UF horietako bakoitzari dagokion ordaina ez ezik, ordainen murriztapen morfosintaktikoak ere.

[H5] Aditza+izena motako UFen informazio linguistikoak itzulpen automatikoan zer eragin duen aztertzea. Oro har, uste dugu lortu dugula erakustea UFen inguruko informazio linguistikoak eragin positiboa duela itzulpengintza automatikoan. Metodologia erdiautomatiko bat proposatu dugu corpusetatik informazio fraseologikoa lortzeko, bai UFen aldakortasunaren eta bai UFen ordainen ingurukoa, eta, azterketa horretan bildutako datuak *Matxin* itzultzaile automatikoan sartuta, itzulpenen kalitatea hobetu egiten da.

6. KAPITULUA

UFen eta haien datu linguistikoen gordailua: Konbitzul datu-basea

Aurreko kapituluetan azaldu ditugun lanetan askotariko informazio linguistikoa bildu dugunez, datu horiek guztiak eskuragarri jarri nahi izan ditugu, eta datu-base publiko bat sortu dugu horretarako. *Konbitzul* du izena, eta interneteko helbide honetan dago: `http://ixa2.si.ehu.es/konbitzul/`.

Datu-baseak, guztira, gaztelaniazko 1.927 UF eta euskarazko 2.074 biltzen ditu gaur egun, euskarazko 4.043 eta gaztelaniazko 3.022 ordainekin batera. Horietatik, gaztelaniazko 894 UFk eta haien ordain banak patroi morfosintaktikoa dute zehaztuta, gure lanetan zehar bildu dugun informazioa alegia, UFen identifikaziorako eta itzulpen automatikorako erabili duguna.

*Konbitzul*i bilatzaile gisako interfazea sortu diogu batetik, nahi duen orok kontsultak egin ahal izan ditzan, baina, bestetik, osorik deskargatzeko aukera ere eskaintzen dugu, informazio hori hizkuntza-tresna aurreratuetan erabili nahi duenari lana errazte aldera: `http://ixa.eus/node/4484`. CSV formatuko bost fitxategitan deskargatzen da informazioa, taulak ulertzeko argibideak jasotzen dituen beste dokumentu batekin batera.

Kapitulu honetan, kontsulta-tresna deskribatuko dugu lehenik (6.1. atala), datu-basearen barruko arkitektura nolakoa den erakutsiko dugu jarraian (6.2. atala), eta, amaitzeko, funtzionalitateez jardungo dugu (6.3. atala).

6.1 Kontsulta-tresnaren deskribapena

Esan bezala, bilatzaile gisako interfazea eman diogu datu-baseari, interesa duen edonork erraz kontsultatu ahal izan dezan barruko informazioa. Webgunearen orri nagusian, hiru ataleko menua agertzen da goialdean:

- **Sarrera**. Datu-basearen aurkezpena, bilaketak egiteko argibide gutxi batzuk, hainbat datu orokor (tamaina, zer informazio jasotzen den eta nola dagoen informazio hori antolatuta) eta erreferentziak.
- **Bilatzailea**. Datu-basean bilaketak egiteko barra, hainbat iragazki jartzeko aukerarekin (informazio gehiago, jarraian).
- **Nor garen**. Datu-basearen sortzaileen izenak eta harremanetarako informazioa.

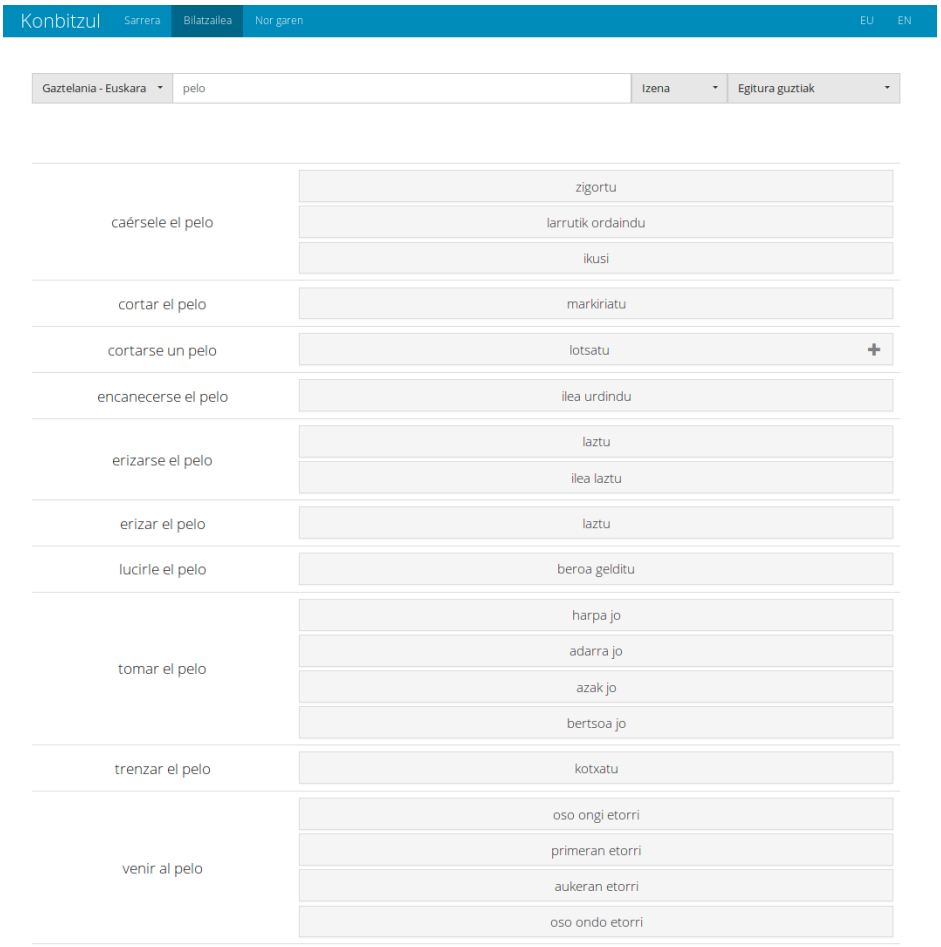
Webgunera sartzean, bilatzailea da defektuz agertzen den orri nagusia. Hala, 6.1. irudiko bilaketa-barra agertzen da, non aukera ematen baita, bilatu nahi d(ir)en hitza(k) idazteko ez ezik, bilaketei hainbat iragazki aplikatzeko ere.

6.1 irudia – *Konbitzul* datu-basearen orri nagusia eta bilaketa-barra

Lehenik, hizkuntza-noranzkoa aukeratu beharra dago: gaztelania-euskara ala euskara-gaztelania. Bilaketak noranzko horren lehen hizkuntzan egiten dira beti, eta bilaketa zeren arabera egin nahi den ere zehaztu behar da, hau da: hitz-konbinazio oso bat bilatu nahi den, ala izen edo aditz bat bakarrik, hura barne hartzen duten UFak ikusteko. Ondoren, nahi bada, egitura morfologikoari dagokion iragazkia ere aplikatu daiteke: euskarazko UFen kasuan, izenari itsatsiriko 12 markaren artean aukeratu daiteke (absolutiboa, ergatiboa, ablatiboa...), eta gaztelaniazko UFen kasuan, berriz, osaera morfologikoa (aditza+izena, aditza+preposizioa+izena...).

Bilaketa-gakoa idazteko, hitz-konbinazio osoaren araberako bilaketa egin nahi bada, forma osoa idatzi beharra dago; esate baterako, *caérsele el pelo*

idatzita, UF hori eta hiru ordain erakusten dira, baina, *caer pelo* bakarrik idatziz gero, ez da emaitzarik aurkitzen. Izena edo aditza bakarrik bilatzen bada, berriz, lema da idatzi beharrekoa, hau da, adibide berberarekin jarraituz, *caer* edo *pelo* jarri beharko genituzke, eta ez *caérsele*, *pelos* eta halakorik (6.2. irudia).



6.2 irudia – *Konbitzul*en gaztelaniazko *pelo* izena bilatuta erakusten diren UFak eta ordainak.

Nolanahi ere, bada hirugarren aukera bat, tartekoa-edo: % ikurra erabiltzea. Ikur hori jartzerakoan, ikurraren lekuan edozein karaktere-segida onartzen da, eta, hala, esate baterako, *pel%* idatzi eta izena bilatzen ari garela zehazten badugu, *pel-* hizkiez hasten diren izenak barne hartzen di-

tuzten UFak erakusten dira: *poner en peligro, hacer pellas, jugarse el pellejo* eta beste batzuk.

Emaitzak erakusten direnean, bilaketarekin bat datozen UFen zerrenda eta UF bakoitzari dagozkion ordainak bistaratzen dira. Baina, horrez gain, linguistikoki aztertu ditugun UFei eta ordainei [+] ikurtxo bat agertzen zaie aldamenean, eta haren gainean klikatuz ikus daiteke informazio gehigarria: osagaien gramatika-kategoriak, osagaien arteko erlazio sintaktikoa, eta UFari edo ordainari dagokion patroi morfosintaktikoa. Kurtsorea patroiaren gainean jarriz gero, patroi horri dagokion informazio xehea erakusten da (6.3. irudia).

poner en peligro

arriskuan jarri	
Identifikaziorako informazioa	
Aditza	poner
Izena	peligro
Preposizioa	en
Erlazio sintaktikoa	PREP-CC
Patroi morfosintaktikoa	SING_NO-DET
Transferentziarako informazioa	
Aditza	jarri
Izena	arrisku
Kasua/Postposizioa	INE
Erlazio sintaktikoa	PREP-CC
Patroi morfosintaktikoa	SING_DEF
Aztertze-metodoa	
Eskuzkoa	

6.3 irudia – *Konbitzul*en *poner en peligro* UFari eta *arriskuan jarri* ordainari dagokion informazio-taula.

6.2 Datu-basearen arkitektura

Datu-basean gordetako informazioa bost taulatan dago antolatuta, honela:

- Gaztelaniazko UFak eta euskarazko hitz-konbinazioen gaztelaniazko ordainak. UFen azterketa linguistikorik egin badugu, osagaien gramatika-kategoriak, haien arteko erlazio sintaktikoa eta UFari dagokion patroi morfosintaktikoa ere zehazten dira; halako informaziorik ez badugu, hitz-konbinazioa eta haren osagaien gramatika-kategoriak bakarrik.

- Euskarazko UFak eta gaztelaniazko hitz-konbinazioen euskarazko ordainak. Gure azterketa linguistikoetan landu ditugun ordainen kasuan, osagai(ar)en gramatika-kategoria(k) eta ordain bakoitzari dagokion patroi morfosintaktikoa zehazten da, eta, ordaina hitz batez baino gehiagoz osatua bada, baita osagaien arteko erlazio sintaktikoa ere.

- Gaztelaniazko eta euskarazko UFak eta ordainak nola lotzen diren elkarren artean, edo, bestela esanda, gaztelaniazko UF bakoitzari euskarazko zer ordain dago(z)kion –eta alderantziz–.

- Gaztelaniazko patroi morfosintaktiko bakoitzak zer ezaugarri xeheri egiten dien erreferentzia.

- Euskarazko patroi morfosintaktiko bakoitzak zer ezaugarri xeheri egiten dien erreferentzia.

6.3 Funtzionalitateak

Orain arte azaldutakoak kontuan harturik, datu-basea eta kontsulta-tresna hainbat helburutarako izan daitezke baliagarriak. Hona hemen zer-nolako informazioa lor daitekeen *Konbitzul*etik, besteak beste:

- Hitz jakin bat zer beste hitzekin konbinatu ohi den

- UF jakin batek zer ordain dituen

- UFrik usuenek zer ezaugarri morfosintaktiko dituzten

- UF horien ordainik usuenek zer ezaugarri morfosintaktiko dituzten

Bestalde, datu-basea osorik ere deskargatu daitekeenez, erabiltzaileek aukera dute informazio hori nahieran erabiltzeko, hizkuntza-tresnen garapenean nahiz azterketa linguistikoetan. Deskargagatzen den karpeta konprimatuak cvs formatuko bost fitxategi dauzka, taulak ulertzeko argibideak jasotzen dituen beste dokumentu batekin batera. Honako lotura honetatik eskuratu daiteke informazio guztia: `http://ixa2.si.ehu.eus/konbitzul/deskargak/Konbitzul.zip`.

Laburpena

Kapitulu honetan, *Konbitzul* datu-basea deskribatzen jardun dugu: nolakoa den kontsulta-tresna, nola antolatuta dauden datuak, eta zer funtzionalitate dituen datu-baseak, oro har. Eduki horiek ez dutenez ikerketa-lanen berri ematen, baizik eta gure azterketetatiko datuak biltzen dituen tresna baten berri, kapitulu honetan ez dugu abiapuntu-hipotesirik landu. Bai, ordea, tesi-lanaren helburuetako bat.

[H6] Aditza+izena motako UFak eta haien ordainak biltzea –euskaraz eta gaztelaniaz– eta eskuragarri jartzea, Hizkuntzaren Prozesamendurako aplikagarriak diren datu linguistikoekin batera.

Landu ditugun UFak, ordainak eta batzuen zein besteen inguruko datu linguistiko guztiak publiko egin ditugu, *Konbitzul* datu-basearen bidez. Datu-basea publikoa da, eta, bilaketak egiteko interfaze erabilerraza edukitzeaz gain, posible da informazio guztia .csv formatuan deskargatzea ere. Guztira, gaztelaniazko 1.927 UF (euskarazko 4.043 ordainekin) eta euskarazko 2.074 (gaztelaniazko 3.022 ordainekin) jasotzen ditu, eta, horietatik, gaztelaniazko 894 UFk eta haien ordain banak patroi morfosintaktikoa dute zehaztuta, identifikazio-lanerako eta itzulpen automatikorako erabilgarri.

7. KAPITULUA

Euskarazko aditz-UFak corpusean: etiketatze-lana eta agerpen literalen azterketa

Behin baino gehiagotan aipatu dugu PARSEME proiektu europarra txosten honetan zehar. Fraseologia konputazionalaren inguruko ikertzaileak batzea izan du helburutzat, eta, gu ere haren parte izan garenez, kapitulu hau harako –edo haren haritik– egin ditugun ekarpenei eskainiko diegu.

Askotariko jarduerak antolatu dira PARSEMEren baitan, baina, ziur asko, proiektuaren amaiera aldera eginiko ataza partekatuak izan du oihartzunik handiena. Bi edizio egin dira gaur arte, eta, ataza gauzatu ahal izateko, hainbat hizkuntzatako corpusak etiketatu dira fraseologia mailan. Bitan hartu dugu parte guk. Batetik, gaztelaniazko lantaldean jardun dugu lehen ediziorako etiketatze-lanean (Savary *et al.*, 2017), eta, bestetik, euskarazko corpusa sortu dugu bigarren ediziorako (Ramisch *et al.*, 2018). Kapitulu honetako 7.1. atalean, euskarazko corpusaren etiketatze-lanaz jardungo dugu.

Horrez gain, PARSEMEk lankidetzarako ateak ireki ditu proiektua bera amaitu eta gerora ere, eta proiektukide ohi batzuen artean eginiko lan batez hitz egingo dugu, hain zuzen, 7.2. atalean. Lankidetza horretan, UFen agerpen literalak aztertu ditugu, familia desberdineko bost hizkuntza oinarritzat harturik: alemana, euskara, greziera, poloniera eta portugesa. Lan horren berri emango dugu hemen, euskarazko zatiari arreta berezia eskainiz.

7.1 Euskara PARSEMEren corpusean

Esan dugunez, PARSEMEren ataza partekatuak bi edizio izan ditu, eta bigarrenerako sortu dugu euskarazko corpusa, zeina beste hemeretzi hizkuntzatako corpusekin batera argitaratu baita.

Euskarazko etiketatze-lanaren berri emateko, PARSEMEren gidalerroak zertan dautzan azalduko dugu lehenik, eta zertzelada batzuk emango ditugu gidalerroak euskaraz erabili ahal izateko eginiko moldaketez (7.1.1. atala). Ondoren, corpusa etiketatzeko metodologia orokorra azaldu (7.1.2. atala), eta corpus etiketatuaren ezaugarri nagusiez jardungo dugu (7.1.3. atala). Etiketatze-lanean nahastea sortu duten hainbat kasu aipatuko ditugu jarraian, eta haien aurrean hartutako erabakien berri emango dugu (7.1.4. atala). Eta, azkenik, gidalerroak hobetzeko proposamen gutxi batzuk plazaratuko ditugu (7.1.5. atala).

7.1.1 PARSEMEren gidalerroak

PARSEMEren gidalerroak gidalerro unibertsalak izateko asmoz sortu dira, hizkuntzaz hizkuntzako tradizio fraseologikoak nolabait uztartu eta proposamen bateratu bat egon dadin. Hortaz, garrantzitsua da nabarmentzea gidalerroetako edukiak ez direla beti bateragarriak hizkuntzaz hizkuntzako literaturarekin. Guri dagokigunez zehazki, haien sailkapena ez dator guztiz bat 4. kapituluan jaso dugunarekin, multzoak oro har antzeko samarrak diren arren. Atal honetan eta ondorengoetan, aipamen bat baino gehiago egingo diegu desberdintasun horiei.

Euskarazko corpusa sortzen hasi ginenerako gidalerroen lehen bertsioa argitaratuta bazegoen ere, aldaketa dezente egin ziren lehen bertsio hartatik bigarrenera. Eztabaida mamitsuak izan ziren aldatu beharrekoen inguruan[1], eta, gure ekarpenen artean, azpimarratzekoa da gidalerroak hizkuntza ez-prepositiboentzat ere aplikagarri egin izana[2].

Behin bigarren bertsio egonkor samar bat lortuta, hizkuntza bakoitzeko arduradunek adibideak gehitu zizkieten azalpenei, eta gidalerroak honako webgune honetan jarri ziren eskuragarri: `http://parsemefr.lif.univ-mrs.`

[1]Eztabaidak GitLabeko orri honetan daude ikusgai: `https://gitlab.com/parseme/sharedtask-guidelines/issues`

[2]Ikus eztabaida (`https://gitlab.com/parseme/sharedtask-guidelines/issues/34`) eta gidalerroetako 7.1.2 puntuan esaten dena.

`fr/parseme-st-guidelines/1.1/`. Ez ditugu gidalerro mardul horiek osorik ekarriko hona, baina bai kontzeptu nagusiak eta etiketa bakoitzaren ezaugarriak, ondorengo edukietarako argibide gisa.

7.1.1.1 Kontzeptu nagusiak

Hasteko, gogora dezagun PARSEMEren corpusean **aditz-UFak** daudela etiketatuta, hau da, forma kanonikoan buru sintaktikotzat aditza dutenak –ikus forma kanonikoei buruzko azalpena beherago–. Horrek esan nahi du, 144. adibidea etiketatzekoa bada ere, 145.eko izen elkartua ez dela kontuan hartzen[3], izena duelako buru sintaktikotzat eta ez aditza.

(144) *Ikastaroan* ***izen eman*** *zuen.*

(145) *Izen-ematea atzo amaitu zen.*

Aditz nagusiarekin batera agertzen diren hitzek edozein gramatika-kategoria izan dezakete, eta, hortaz, etiketatze-lana ez da, gure orain arteko lanetan bezala, izena+aditza motako UFetara bakarrik mugatzen. Esate baterako, 146. adibidetik 149. adibidera arteko UFak etiketatu beharrekoak dira, izena+aditza motakoa lehena bakarrik bada ere.

(146) ***Erabakia hartu*** *zuen.*

(147) ***Nabari da*** *jendea etorri dela.*

(148) ***Ikusi eta ikasi***!

(149) ***Aho hertsitik ez da eulirik sartzen****.*

UFen barruko **osagai lexikalizatuak** etiketatzen dira, hau da, UFan beti agertzen diren lemak. Nolanahi ere, euskararen izaera aglutinatiboa dela-eta, lexikalizatu gabeko morfema batzuk ere etiketatu behar izaten ditugu guk, lexikalizatutako lemaren bati lotuta egon direnean. Adibidez, 150–153. adibideetan *pauso* eta *eman* dira UFko osagai lexikalizatuak, baina, etiketatzea hitz mailan egiten denez, izenari itsatsitako markak ere etiketaren barruan sartu behar dira nahitaez: 150. adibidean artikulua (*-ak*), eta 153.ean artikulua eta postposizio instrumentala (*-ez*).

(150) ***Pausoak ematen*** *ari da.*

[3]PARSEMEren gidalerroekin lotura zuzenik ez badu ere, merezi du gogora ekartzea Azkarateren doktoretza-tesian (1987: 407–415. orr) badela eranskin oso bat hitz elkartuen eta UFen –haren hitzetan *aditz-esapideen*– arteko desberdintasunen inguruan.

(151) ***Pauso*** *bat* ***eman*** *zuen.*

(152) *Hainbat* ***pauso*** *oker* ***eman*** *zituen.*

(153) ***Emandako pausoez*** *damutu zen.*

UFen aldaki morfosintaktikoak kontuan hartzen dira PARSEMEren corpusean, eta horregatik etiketatzen da *pauso eman* UFa lau adibide horietan guztietan, esaldi batetik bestera desberdin erabilita egon arren. Lehen hiru adibideetan (150–152) aditza da buru sintaktikoa, eta azkenekoan, ordea, ez. Hala ere, esana dugu UFen **forma kanonikoa**n pentsatu behar dela hitz-konbinazio jakin bat UFa den ala ez erabakitzeko, zenbaitetan posible baita UFetako osagaien erlazio sintaktikoa aldatzea –erlatibozko perpausetan eta halakoetan–. Hau diote forma kanonikoei buruz PARSEMEren gidalerroetan[4]:

> Aditz-UF baten forma kanonikoa aditz-sintagma bat da, zeina ahots aktiboan baitago, buru sintaktikotzat aditza baitu eta gainerako osagai lexikalizatuak aditzaren mende edo beste osagai lexikalizaturen baten mende baitaude.

Hori aintzat harturik, 153. adibidea forma kanonikora ekarrita, *pausoak eman* edo halakoren bat litzateke, eta beteko luke burutzat aditza izateko baldintza. Ez da gauza bera gertatzen, ordea, 145.eko hitz elkartuarekin, adibide horretan hitz-konbinazioa forma kanonikoan baitago dagoeneko, izena buru duelarik: ez da *izen eman*, baizik eta *izen-emate*.

Beraz, 150. eta 153. adibideen bidez erakutsi dugunez, lexikalizatu gabeko morfema batzuk markatu egin behar izaten ditugu guk. Hala ere, kontrakoa ere gertatzen da zenbaitetan: morfema lexikalizatu bat ezin izatea etiketatu, lexikalizatuta ez dagoen lema bati itsatsita joateagatik. Esate baterako, 154. adibidean, *falta* eta *sumatu* bakarrik etiketatuko genituzke, *haren* gabe, nahiz eta *falta sumatu* UFak derrigorrezkoa izan genitibodun osagai bat.

(154) *Haren* ***falta sumatzen*** *dut.*

Etiketak hitz mailan ematen direnez, ezin dugu *haren* ere markatu, horrek esan nahi bailuke *hura* lexikalizatutzat hartzen dugula, eta ez da hala, lema

[4] Aipu gisa baina komatxorik gabe jartzen ditugun testu-zatiak jatorrizko testutik itzuliak dira.

hori ordezka baitaiteke beste askorekin: *norbaiten falta sumatu*, *zure falta sumatu* eta abar.

Azkenik, gidalerroek arreta berezia eskaintzen diete **kolokazioak** eta **metaforak** aditz-UFetatik kanpo uzteko argibideei. Hitz gutxitan, honako hau jasotzen da gidalerroen 1.6 eta 1.7 ataletan:

- Kolokazioak ez dira UFtzat hartzen PARSEMEren atazan, haien idiomatikotasuna erabat estatistikotzat jotzen delako. Bestela esanda: kolokazioko hitzek ausaz aurreikus litekeena baino joera handiagoa dute elkarrekin agertzeko, baina ez dute bestelako idiosinkrasia ortografiko, morfologiko, sintaktiko edo semantikorik. Hortaz, 155. eta 156. adibideen gisakoak albo batera uztekoak dira.

  (155) *interesa agertu*

  (156) *autobusa hartu*

- Metaforak ere UFetatik kanpo uztekoak dira oro har, esapide idiomatikoekin lotura estua badute ere. UF asko metaforetan oinarritutakoak diren arren (157. adibidea), metafora guztiak ez dira UFak, baizik eta, askotan, une jakinetan esanahi bat adierazteko sortutako konparazio modukoak (158. adibidea). Hizkuntza bateko hiztegian nahikoa egonkortuta dauden metaforak UFtzat hartuko ditugu, baina ez unean uneko behar estilistikoei erantzuteko sortutakoak.

  (157) *Ez zaitez kezkatu, ez ezazu* ***burua hautsi****.*

  (158) *Laino beltz batean egon gara orain arte.*

Gure orain arteko lan-ildoak argi uzten duenez, kolokazioez diotenarekin ez gatoz guztiz bat, eta eskainiko diogu horri beste tarte bat 7.1.5. atalean.

7.1.1.2 UFen sailkapena

PARSEMEren sailkapenak sei aditz-UF mota bereizten ditu guztira. Hala ere, sei mota horietako bi baino ez dira unibertsalak, eta bi horiek bakarrik dauzkagu, hain zuzen, euskaraz: aditz-esapide idiomatikoak (*Verbal Idioms*, VID) eta aditz arindun konbinazioak (*Light Verb Constructions*, LVC). Beste laurak ez dira guretzat aplikagarriak: erreflexiboa berezko duten aditzak (*Inherently Revlexive Verbs*, IRV), aditza+partikula konbinazioak

(*Verb Particle Constructions*, VPC), aditz anitzeko konbinazioak (*Multi-verb Constructions*, MVC) eta adposizioa berezko duten aditzak (*Inherently Adpositional Verbs*, IAV)[5]. Guk lehen biez bakarrik hitz egingo dugu hemen; gainerakoen inguruko xehetasunak gidalerroetako bosgarren atalean daude irakurgai, eta txosten honetako 2.1. atalean ere jaso ditugu adibide batzuk.

Aipagarria da gidalerroak erabaki-zuhaitz itxurako testez hornituta daudela, bai hitz-konbinazio jakin bat UFtzat etiketatu ala ez erabakitzen laguntzeko, eta bai UF bakoitzari zer etiketa eman erabakitzen laguntzeko ere. Guk ez ditugu testak bere horretan ekarriko hona (horiek ere gidalerroetako 5. atalean daude jasota), baina landu ditugun moten ezaugarri nagusiak aipatuko ditugu.

Aditz arindun konbinazioak (LVC)

Aditz arindun konbinazioen artean bi azpimota bereizten dira aditzaren ezaugarri semantikoen arabera: konbinazio osoak (LVC.full) eta kausalak (LVC.cause). Hona hemen LVCen ezaugarri nagusiak eta zer desberdintasun duten azpimota batek eta besteak:

1. Aditz batez eta izen (bakar edo elkartu) batez osatutako konbinazioak dira. Izena aditzaren mendekoa da beti, eta zenbaitetan artikulua, kasu-marka edo postposizioren bat izan dezake.

   (159) ***lan egin***

   (160) ***aurrera egin***

2. Izena predikatiboa da, eta ekintza, gertaera edo egoera bat adierazten du. Izena predikatiboa dela esatean, esan nahi da argumentu semantikoak behar dituela bere esanahia osatzeko. Honako bi adibide hauetan, UFek adierazten duten ekintzek derrigorrez behar dute subjektu semantikoa:

   (161) ***negar egin*** → izenak ekintza bat adierazten du

   (162) ***lo egin*** → izenak egoera bat adierazten du

[5] Adposizioa berezko duten aditzak, berez, baditugu euskaraz ere: *-tzat hartu*, adibidez. Nolanahi ere, PARSEMEn hitz mailako etiketak bakarrik jartzen direnez (ikus 154. adibidea eta azalpena), euskaraz ezin izan dugu halakorik etiketatu.

3. Aditzaren arabera, LVC.full eta LVC.cause bereizten dira.

    - UFa **LVC.full** motakoa da aditza erabat *arina* denean, hau da, ezaugarri morfologikoak bakarrik gehitzen dizkionean LVC osoaren esanahiari: pertsona, numeroa, denbora edota aspektua. Bestela esanda, aditzaren subjektu semantikoa izenaren argumentu semantikoa da.

        (163) ***eskubideak ukan*** → Xk eskubideak baditu, Xren eskubideez ari gara

        (164) ***min hartu*** → Xk min hartu badu, Xren minaz ari gara

    - UFa **LVC.cause** motakoa da aditza kausatiboa denean, hau da, aditzaren subjektua izenak adierazten duen ekintza, gertaera edo egoeraren kausa denean. Bestela esanda, izenak baditu subjektua ez beste elementuen bidez adierazten diren argumentu semantikoak, eta aditzaren subjektuak informazioa gehitzen du.

        (165) ***min eman*** → Xk min ematen badio Yri, Yren minaz ari gara eta ez Xrenaz

        (166) ***eskubideak eman*** → Xk eskubideak ematen badizkio Yri, Yren eskubideez ari gara eta ez Xrenez

Ezaugarri horiek gure sailkapenarekin alderatuz, esan beharra dago multzo hau gure kolokazioen multzoa baino dezente mugatuagoa dela. Batetik, aditz arindun konbinazioak kolokazioen barruko zati bat baino ez direlako, eta, bestetik, guk aditz arintzat hartzen ditugun guztiak ere ez datozelako bat etiketa honen ezaugarriekin. Geroago hitz egingo dugu gehixeago horretaz, 7.1.4. eta 7.1.5. ataletan.

Aditz-esapide idiomatikoak (VID)

Aditz-esapide idiomatikoen multzoa, VIDena, LVCena baino zabalagoa da, aditz-lokuzioez gain barne hartzen baititu enuntziatu fraseologikoak ere[6]. VIDek bi osagai lexikalizatu dituzte gutxienez, aditz bat eta haren mendeko osagai bat, eta mendeko osagai hori askotarikoa izan daiteke:

[6]Enuntziatu fraseologikoen inguruko informazio gehiago, Urizarren doktoretza-tesian (2012: 84–87. orr).

(167) ***txoritxo batek esan*** → subjektua

(168) ***adarra jo*** → objektua

(169) ***begi-bistatik galdu*** → osagarri zirkunstantziala

Horrez gain, etiketatzean sor litezkeen zalantzak argitzeko, badira bi gako erabakigarri:

- LVCetan aditza+izena motako konbinazioak bakarrik sartzen direla eta, hortaz, aditzarekin batera izena ez beste osagairen bat badago, VID gisa markatzekoa dela –beti ere UFtzat markatzekoa bada–.

  (170) ***nabari izan*** → adjektiboa+aditza

  (171) ***adarka egin*** → adberbioa+aditza

- LVCetan aditzaz gain beste osagai bakarra onartzen denez, mendeko osagai bat baino gehiago daudenean ere UFa VID gisa markatzekoa dela.

  (172) ***katuak mingaina jan*** → subjektua eta objektua

  (173) ***gizakia ez da ogiz bakarrik bizi*** → subjektua, atributua eta osagarri zirkunstantziala

Batzuetan, 173. adibidean esaterako, gerta liteke VID baten barruan beste UFren bat egotea (kasu honetan *bizi izan*, LVC.full), eta halakoetan biak markatu beharra dago, bai UF nagusia eta bai barruan daramana.

7.1.2 Etiketatze-metodologia orokorra

Sei etiketatzailek hartu dute parte ataza honetan: bost hizkuntzalarik eta lexikografo batek. Horietako bostek eskarmentua zuten lehenagotik fraseologia-kontuetan, eta besteak, fraseologiaren alorrean ez baina bai bestelako etiketatze linguistikoak egiten.

Etiketatze-lana hainbat fasetan egin dugu. Lehenik, trebakuntza-saio gutxi batzuk antolatu ditugu, bi helbururekin: batetik, parte-hartzaileei argibideak emateko gidalerroen eta etiketatze-plataformaren inguruan, eta bestetik, aditz-UFen etiketatze-lanak euskaraz bereziki sor litzakeen kasu nahasgarriak identifikatu eta ebazpideak emateko. Horren ostean egin dugu benetako

etiketatze-lana, eta berrikuspen orokor bat ere egin dugu lan osoaren amaieran, desberdin etiketatutako konbinazioren bat edo beste bateratzeko.

Atal honetan, trebakuntza-saioen inguruko azalpenak emango ditugu lehenik (7.1.2.1), eta etiketatze-lanaren eta etiketatzaileen arteko adostasunaren ingurukoak ondoren (7.1.2.2).

7.1.2.1 Trebakuntza-saioak

Gidalerroen eta etiketatze-plataformaren inguruko argibideak eman ostean, parte-hartzaile guztiei 500 esaldiko lagin berbera eman diegu lehen proba egiteko. Hasieran, desadostasun ugari egon dira etiketatzaile batzuen eta besteen artean: testu berberari 85etik 170 etiketara bitarte eman dizkiote.

Desadostasun-iturri izan diren adibide guztiak bildu, eta gaiaren arabera antolatu ditugu, beste hiru saiotan haien inguruan jarduteko. Desadostasun asko erraz samar konpondu ditugu, gidalerroak ondo ez ulertu izanagatik eginiko akatsak izan baitira. Gai arazotsuagoen inguruan, berriz, erabakiak hartu behar izan ditugu corpus koherente bat sortze aldera, eta barne-txosten bat osatu dugu, etiketatzaileek gidalerroekin batera erabil zezaten. Txosten horren eduki nagusiak 7.1.4. atalean daude laburbilduta.

7.1.2.2 Etiketatze-lana

Trebakuntza-saioen ondotik, etiketatzaile bakoitzari hainbat testu banatu dizkiogu, eta corpusaren zati txiki bat (871 esaldi) birritan etiketatu dugu, adostasuna kalkulatzeko. Adostasunerako lagin hori osorik etiketatu du etiketatzaile batek, eta beste bik erdibana. Lortutako adostasun-emaitzak 7.1. taulan daude jasota, hizkuntza guztien batezbestekoekin batera.

	F	κ_{span}	κ_{cat}
EU	0,859	0,820	0,859
Batez beste	0,691	0,644	0,844

7.1 taula – Anotatze lexiko-semantikoan lortutako adostasuna

Adostasuna hiru eratara kalkulatu da: F neurriaren bidez (etiketatzaile batek bestearen etiketak asmatu nahi izan balitu bezala), κ_{span} neurriaren bidez (corpuseko aditz guztietatik, zenbatetan etorri diren bat bi etiketatzaileak aditzak UFen partetzat etiketatzean) eta κ_{cat} neurriaren bidez (bi

etiketatzaileek etiketatutako hitz-konbinazioak kontuan hartuta, zenbatetan etorri diren bat UF mota zehaztean).

Taulak erakusten duenez, emaitzak oso altuak dira, eta hiru neurrien bidez lortu dira batezbestekoa baino emaitza hobeak. Ziur asko, gidalerroen zehaztasuna izan da emaitza horien arrazoietako bat, eta trebakuntza-saioek ere lagunduko zuten etiketatzaileen arteko koherentzia lortzen.

Gainera, behin etiketatze-lana amaituta, azken ikuskapen bat ere egin dugu PARSEMEk horretarako prestatutako tresnen bidez[7], eta aukera izan dugu testu batean eta bestean desberdin etiketatutako hitz-konbinazio batzuk topatzeko eta zuzentzeko. Hortaz, kontuan izanik ikuskapenaren aurretik ere adostasun handia genuela, pentsatzekoa da azken corpusaren kalitatea are maila altuagokoa dela.

7.1.3 Corpus etiketatua eta handik ateratako zenbait ondorio

Euskarazko corpus etiketatuak bi iturritako testuak biltzen ditu: *Dependentzia Unibertsalen corpus*eko 6.621 esaldi, hau da, corpus osoa (Aranzabe *et al.*, 2019), eta *Elhuyar Web Corpus*eko 4.537 esaldi[8]. Hortaz, 11.158 esaldi ditu guztira, 157.807 hitz. Corpusari buruzko xehetasunak 7.2. taulan daude jasota, jarritako etiketei dagozkienak barne.

Esaldiak	**Hitzak**	**UFak**	**LVC.cause**	**LVC.full**	**VID**
11.158	157.807	3.823	183	2.866	774

7.2 taula – PARSEMEren euskarazko corpusaren datuak

Ikusten denez, LVC.full gisa markatutako UFen kopurua beste biena baino nabarmen handiagoa da: etiketa guztien % 75. Horiei LVC.cause marka dutenak ere batuta, aditz arindun konbinazioen multzo osoa are handiagoa da, etiketa guztien % 80 hartzen baitu. Nolanahi ere, oso UF gutxi sailkatu dira LVC.cause kategorian, eta badirudi joera hori nahiko orokorra dela hizkuntza gehienetan. Gaztelaniaz, ingelesez eta frantsesez, adibidez, honako

[7]PARSEMEk corpusen prozesamendurako sortutako tresnak, lotura honetan: `https://gitlab.com/parseme/utilities/tree/master/1.1`

[8]*Dependentzia Unibertsalen corpus*ak albisteak biltzen ditu, eta *Elhuyar Web Corpus*ak, berriz, saretik lortutako askotariko testuak.

ehuneko hauek dagozkio etiketa horri: % 10, % 10 eta % 2, hurrenez hurren.

Bestalde, interesgarria da corpuseko gainerako hizkuntzen aldean UFen maiztasuna zenbatekoa den ikustea, alde nabarmena baitago hizkuntza batetik bestera (7.3. taula).

	UFak 100 esaldiko	**LVCak 100 esaldiko**
Euskara	34	27
Batez beste	18	11
Frantsesa	20	9
Gaztelania	15	9
Ingelesa	6	4

7.3 taula – Etiketen 100 esaldiko batezbestekoak, euskaraz, gaztelaniaz, frantsesez, ingelesez eta ataza partekatuko 20 hizkuntzetan oro har

Corpus osoaren batezbestekoaren aldean, euskarazko UFen kopurua ia bikoitza da, eta LVCena, berriz, bikoitza baino handiagoa. Izan ere, corpuseko 20 hizkuntzetatik bik baino ez dute euskarak baino LVC etiketa gehiago jarri esaldi bakoitzeko: persierak eta hindiak. Gainera, euskal hiztunek gehien hitz egiten dituzten hiru erdarak kontuan hartuta, aldea are nabarmenagoa da[9], LVCen kopurua hirukoiztu egiten baita frantsesaren eta gaztelaniaren aldean, eta ia zazpi bider handitzen ingelesarenean.

Azkenik, morfologiari begiratu bat eginez gero, aipatzekoa da corpusean etiketatutako aditz-UFen artean izena+aditza motakoak izan direla gehiengehienak (% 94). Datu hori aurresangarria zen, kontuan izanik LVCen multzoan osaera horretako konbinazioak bakarrik onartzen direla PARSEMEren gidalerroetan. Izen horietatik % 85ek absolutibo-marka dute[10], eta beste guztien artean inesiboa da markarik errepikatuena; hortaz, corpuseko datuak bat datoz 3. kapituluan eman ditugun datuekin.

[9]Alderaketa egiteko, kategoria unibertsalak bakarrik hartu ditugu kontuan: LVC.full, LVC.cause eta VID.

[10]Gogoan izan batere markarik gabeko izenak ere, *lan egin*eko *lan* eta gisa horretakoak, absolutibotzat hartzen ditugula.

7.1.4 Euskarazko kasu nahasgarriak eta haiekiko erabakiak

Esan dugunez, trebakuntza-saioetan hartutako erabakiak barne-txosten batean bildu ditugu. Txosten horretako gai orokorrak ekarriko ditugu orain hona: LVCetako izenen aldakortasun morfologikoa, *izan* aditza duten LVCen geroaldia, UFetako izen eta adjektibo batzuen arteko muga lausoa, eta LVCetako (itxurazko) *cranberry* hitzak.

7.1.4.1 LVCetako izenen aldakortasun morfologikoa

Euskaraz, izen-sintagmak ia beti eraman ohi du determinatzailea, eta oso kasu gutxitan erabiltzen da batere determinatzaile gabeko izenik (Trask, 2003: 92. orr; Laka, 1996: 6. atala). Salbuespen gehienek lotura dute fraseologiarekin, ezaugarri hori nahiko ohikoa baita UFetan, aditz arindun konbinazioetan batez ere. Izen soilak bereziki maiz agertzen dira oso ekintza ohikoak adierazten dituzten UFetan (174–175. adibideak), zeinak inguruko hizkuntza askotan aditz bakarrez adierazten baitira.

(174) ***lo egin*** → (ES) dormir; (FR) dormir; (EN) to sleep

(175) ***hitz egin*** → (ES) hablar; (FR) parler; (EN) to speak

Halako UF batzuek ez dute ia aldaketa morfologikorik onartzen, baina beste batzuek bai. Adibidez, 176. adibideko UFa erabil daiteke forma mugatuan.

(176) a. ***lan egin*** → izen soila

b. ***lan<u>a</u> egin*** → izen mugatua, artikuluduna

Anotatzaile batzuek zalantza izan dute halakoekin, ziur asko euskarazko analizatzaile automatikoek era horretako konbinazioen artikulurik gabeko aldakiak bakarrik analizatzen dituztelako UFtzat gaur egun (Alegria *et al.*, 2004). Zenbaitetan izen mugatuak UFen aldakitzat ez hartzea proposatu izan bada ere, etiketatze-lan honetan 176b. adibidekoa eta gisakoak ere kontuan hartzea erabaki dugu, PARSEMEko gidalerroen test guztiak gainditzen baitituzte.

7.1.4.2 *Izan* aditza duten LVCen geroaldia

Izan da LVCetan gehien agertzen den aditzetako bat euskaraz, eta haren erabilera ezohiko samarra da morfosintaxiari dagokionez. Halako aditz batzuei *erdilaguntzaile* ere deitu izan zaie literaturan (Ortiz de Urbina, 2003b), eta haien berezitasuneko bat LVCaren geroaldiko formetan datza. Izan ere, ekintza bat etorkizunean gertatuko dela adierazteko, aditzaren partizipioari *-ko/go* morfema erantsi ohi zaio euskaraz. *Izan* aditza barne hartzen duten LVCetan, berriz, morfema hori ez zaio beti aditzari eransten; izenari ere eransten zaio batzuetan (177. adibidea).

(177) a. ***behar dut***
b. ***behar izango*** *dut*
c. *beharko dut*

Adibide horretan, *behar izan* UFaren hiru aldaki daude: bata orainaldian (a), eta beste biak geroaldian (b eta c). Lehenengoan eta bigarrenean, UFko bi osagai lexikalizatuak daude agerian, *behar* eta *izan*; hirugarrenean, ordea, bakarra, hor agertzen den *izan* aditza laguntzailea baita, eta ez LVCaren parte dena.

Izan aditza laguntzailea ere izateak pentsaraz lezake 177c. adibidean eta halakoetan aditz laguntzailea ere UFaren parte dela. Ez da hala, ordea, eta etiketatzaileek gogoan izan behar dute desberdintasun hori. Halakoak argitzeko, Morfeus analizatzaile morfologikoa lagungarria izan liteke (Alegria *et al.*, 1996), adibidez, *beharko* hitzari lema ematean *behar_izan* esleitzen baitio zuzenean, eta ez *behar* izena soilik.

7.1.4.3 UFetako izen eta adjektibo batzuen arteko muga lausoa

Euskarazko zenbait UFtan, ez dago erabat argi aditza ez beste osagaia izena ala adjektiboa den, eta analizatzaile morfosintaktiko batzuek zailtasunak izaten dituzte halakoei kategoria esleitzean. Esate baterako, *gose* bata ala bestea izan daiteke testuinguruaren arabera, gaur egun adjektibo gisa oso gutxi erabiltzen bada ere. Hitz hori eta *izan* aditza barne hartzen dituen bi UF ageri dira beheko adibideetan. Zenbait egilek diote *gose* adjektiboa dela 178.ean eta izena 179.ean, baina analizatzaile askok izentzat hartzen dituzte biak.

(178) ***Gose naiz.***

(179) ***Gosea dut.***

Berez, izena+aditza motakoak ez diren UFak ezin dira LVCen multzoan sartu, eta, horren arabera, 178. adibidea VIDen multzoan sartu beharko genuke adjektiboa balitz. Nolanahi ere, gidalerroetan ohar hau egiten da LVCek izenak baino onartzen ez dituztela esan ondoren:

> Hindiaren kasuan, ager daiteke adjektibo bat izenaren ordez, adjektibo hori izen predikatibo baten berdin-berdina bada morfologikoki.

Aipamen hori euskararako ere aplikagarria dela uste dugu, goiko adibideetatik bi-biek betetzen baitituzte LVCen baldintza eta test guztiak. Gainera, badira beste kasu berezi batzuk non ematen duen beti kategoria batekoa den hitz batek beste kategoria bat duela UF jakin baten barruan:

(180) ***Nahi dut.***

(181) ***Nahiago dut.***

Esate baterako, 180. eta 181. adibideetan ikusten da *nahi* izenak *-ago* konparazio-atzizkia har dezakeela. Atzizki hori adjektiboak eta adberbioak graduatzeko erabiltzen da normalean, eta horrek adieraz lezake *nahiago izan* konbinazioa adjektiboa+aditza motakoa dela.

Koherentziari euste aldera, halako kasuak ere LVC gisa markatzea erabaki dugu, adjektiboek LVCetako izenen ezaugarriak betetzen zituztela iruditu bazaigu.

7.1.4.4 LVCetako (itxurazko) *cranberry* hitzak

Cranberry hitz –edo, zentzu zabalagoan, *cranberry* morfema– deritze esapide jakin batetik kanpo esanahirik ez duten hitzei (Aronoff, 1976; Richter eta Sailer, 2003). Honela daude definituta PARSEMEren gidalerroetan:

> Cranberry hitzak hitz beregainak ez diren tokenak dira. Ez dute esanahi beregainik, baina kategoria sintaktikoa eta inflexio-paradigma izan ditzakete. Esapide jakin batean –edo esapide-zerrenda itxi batean– bakarrik erabiltzen dira, eta ez dira testuinguru batean baino gehiagotan agertzen.

Halako hitzen bat barne hartzen duten UFak zuzenean VIDetan etiketatzekoak dira gidalerroen arabera. Esate baterako, VIDetan sartuko genuke 182. adibidea, *ospa* hitza ez delako erabiltzen ez bada interjekzio gisa edo UF baten barnean. Eta 183. adibidearekin ere beste hainbeste egingo genuke, *txint* hitza oso testuinguru mugatuan erabiltzen baita.

(182) ***ospa egin***

(183) ***txintik*** *ere* ***ez esan***

Badira beste hitz batzuk ere gaur egun UFen barnean bakarrik erabiltzen direnak ia. Lehen begiratuan, hiztun askok pentsa lezakete, adibidez, *merezi* izena ez dela erabiltzen *merezi izan* UFtik kanpo, hori baita gaur egun duen erabilera nagusia. Hala ere, hiztegietan definizio eta guzti jasotzen da *merezi* izena, eta hura UFtik kanpo ageri den adibideak ere badaude, adibidez, Orotariko Euskal Hiztegian (Michelena, 1987):

(184) *Nik emango dizut zure merezia.*

Hasieratik aintzat hartu beharrean halakoak *cranberry* hitzak direla, bilaketa batzuk egitea komeni da etiketatze-lana ondo egiteko. Izan ere, *merezi* izena dela jakinik, *merezi izan* UFak erraz gainditzen ditu LVCen testak eta, hortaz, multzo horretan sartu beharrekoa da, eta ez VIDetan.

Antzekoa da *ari izan*en kasua ere. Ortiz de Urbinaren arabera (2003a: 223–227. orr), lehenago aipatu ditugun aditz erdilaguntzaileen multzoan sar liteke UF hori, eta badu aditz modalekin zerikusirik, *behar izan* eta gisakoen ezaugarri asko betetzen baititu. Lan horretan esaten da aukerarik egokiena *ari* izentzat hartzea dirudiela, eta horixe uste dugu guk ere. Gainera, izena bera Euskaltzaindiaren Hiztegian (Euskaltzaindia, 2012) gutxi erabilitzat markatu bada ere, testu idatzi batzuetan badira oraindik 185. adibidea bezalakoak.

(185) *Ez zuen utzi bere aria.*

Hortaz, *ari* ere *cranberry* hitzetatik kanpo uztea eta *ari izan* LVCtzat markatzea erabaki dugu, *behar izan*, *nahi izan*, *merezi izan* eta antzekoekin batera.

7.1.5 Gidalerroak hobetzeko proposamenak

Aurreko atalak argi uzten du desadostasun eta zalantza gehien sortu duten UFak LVCen multzokoak izan direla. Arrazoietako bat LVCek euskaraz

duten ugaritasuna izango zen, noski. Dena dela, esan dugu PARSEMEren gidalerroak ez datozela beti bat hizkuntzaz hizkuntzako lan fraseologikoekin, eta euskara ere sartzen da multzo horretan.

Gidalerro unibertsalak egitea lan nekeza da edozein atazatarako, baina are gehiago fraseologia konputazionalaz ari bagara, hitz-konbinazio moten arteko mugak ez baitira inoiz Hizkuntzaren Prozesamenduko tresnek behar bezain argiak. PARSEMEren gidalerroak aurrerapauso handia dira bateko eta besteko ikusmoldeak bateratzeko bidean, eta testen eta definizioen zehaztasunak izugarri errazten du etiketatze-lana. Nolanahi ere, badira gure ustez hobetu litezkeen bi puntu, eta horietaz jardungo dugu atal honetan: batetik, kolokazioak fenomeno estatistikotzat soilik hartzeaz (7.1.5.1. azpiatala), eta, bestetik, LVCak izena+aditza konbinazioetara bakarrik mugatzeaz (7.1.5.2. azpiatala).

7.1.5.1 Kolokazioak, UF motatzat

Euskarari eta gaztelaniari buruzko azterketa fraseologikoetan, aditz arindun konbinazioak kolokazioen azpimultzotzat hartu izan dira. PARSEMEren gidalerroetan, ordea, kolokazioak bazter uzten dira, ez LVCen multzotik bakarrik, baizik eta UF guztietatik (7.1.1.1. atala). Kasu batzuetan, ez zaigu erabat argia iruditu zergatik uzten diren kolokazio batzuk UFetatik kanpo, gure ustez etiketatzen diren beste LVC batzuen oso antzekoak baitira.

(186) ***deia egin*** → LVC motako UFa

(187) # *deia jaso* → ez UFa

Esate baterako, *dei* izenak beti aukeratzen du *jaso* aditza 187. adibideko hitz-konbinazioaren esanahia adierazteko, eta oso kasu bakanetan agertzen da antzeko beste aditz batzuekin, *eduki*rekin edo *izan*ekin adibidez. LVC.full multzoak aditz erabat arinak hartzen ditu, baina ez LVC.cause motakoek, horietan aditzak argumentu predikatibo bat gehitzen baitio ekintzari edo egoerari, eragilea. Ildo beretik, *jaso* aditzak ere argumentu semantiko bat gehitzen dio deitzeko ekintzari, deiaren hartzailea.

Erabaki-test xeheei esker, erraz samar baztertu ditugu 187. adibidea eta halakoak etiketatze-lanetik, baina UFetan sartzeko modukoak direla uste dugu hala ere. LVC.cause multzoan zer konbinazio gutxi sailkatu diren ikusita (7.1.3. atala), bai euskaraz eta bai gainerako hizkuntzetan, beharbada multzo hori zabaltzea eta kausalak ez diren aditz batzuk ere kontuan hartzea izan liteke halakoak atazan sartzeko bide bat.

7.1.5.2 LVCak, izena+aditza konbinazioetatik harago

Definizioz, aditz arindun konbinazio deritze pisu semantikoa aditzak ez beste osagaiak hartzen dueneko konbinazioei. PARSEMEren gidalerroetan, osagai hori izena da beti –hindiaren kasuan izan ezik–, eta ekintza, gertaera edo egoera bat adierazten du. Lehen esan dugunez (7.1.4.3. atala), euskarazko etiketatze-lanean adjektiboa+aditza konbinazio batzuk ere LVCen multzoan sartzea erabaki dugu guk, baldin eta aditza ez beste osagaia adjektiboa edo izena izan badaiteke (*bizi, gose*, etab.).

Hala ere, adjektiboak ez dira LVCetako izenekin parekatu litezkeen bakarrak, euskal adberbio askok ere aise betetzen baitituzte LVCen parte izateko ezaugarriak (188–189. adibideak).

(188) ***korrika egin***

(189) ***hazka egin***

Bi adibide horietan, *korrika* eta *hazka* adberbioak predikatiboak dira: ekintza bat adierazten dute, eta argumentu semantikoak behar dituzte esanahia osatzeko. Aditzak ez du egiten adberbioari aditz-izaera ematea besterik, eta ondo uztartzen da LVC.full multzoko beste UFekin.

Hortaz, LVCen multzoa murritzegia dela iruditzen zaigu, eta, euskarari dagokionez batik bat, uste dugu adjektibo eta adberbio predikatiboak ere LVCen partetzat hartzea komeni dela, koherentziari euste aldera.

7.2 UFen agerpen literalen azterketa

Aipatu dugunez, PARSEME proiektuak lankidetzarako bidea eman digu proiektua bukatu eta gerora ere, eta, atal honetan, proiektukide ohi batzuen artean eginiko lan batez jardungo dugu. UFen agerpen literalak izango ditugu hizpide, hain zuzen.

Izan ere, UF asko literalki nahiz idiomatikoki uler litezke, eta testuinguruaren arabera argitu ohi da hitz-konbinazio jakin batek esanahi bat ala bestea duen. Esate baterako, *ziri* eta *sartu* hitzak idiomatikoki erabilita daude 190. adibidean, eta literalki 191.ean.

(190) *Ez zen benetan ari.* ***Ziria sartu*** *zizun!*

(191) *Mutikoak egurrezko ziri bat sartu zuen zuloan.*

Esanahi idiomatikoen eta literalen arteko bereizketa Hizkuntzaren Prozesamenduko erronkarik handienekotzat hartzen da, 2. kapituluan esan dugunez. Atal honetan, ordea, erakutsiko dugu UFen agerpen literalak oso urriak direla praktikan, eta, gainera, ebidentzia gehiago emango dugu frogatzeko bereizketa hori egiteko garrantzi handia dutela, semantikak ez ezik, morfologiak eta sintaxiak ere. Horretarako, bost familia desberdinetako hizkuntza bana aztertu dugu, PARSEMEren corpusetik abiatuta. Lan osoa (Savary *et al.*, 2019) *Prague Bulletin of Mathematical Linguistics* aldizkarian argitaratu dugu, baina haren moldaketa bat egingo dugu hemen, euskarazko zatiari arreta berezia jarriz.

Hasteko, lan honen kontzeptu nagusiak azalduko ditugu (7.2.1. atala). Ondoren, erabilitako metodologia orokorrari buruz jardungo dugu (7.2.2. atala), etiketatze-lana bera nola egin dugun azaltzeko (7.2.3. atala). Eta, azkenik, emaitza orokorrak erakutsiko eta aztertuko ditugu, euskarazkoak batez ere (7.2.4. atala).

7.2.1 Kontzeptu nagusiak: UFen agerpen literalak eta kointzidentziazkoak

Lan honek PARSEMEren corpus etiketatua du oinarrian, eta, hortaz, kontzeptu nagusi gehienak aurreko atalean (7.1) azaldu ditugun berberak dira. Ekar dezagun gogora PARSEMEren corpusean **aditz-UFak** daudela etiketatuta, hau da: hitz lexikalizatu batez baino gehiagoz osatutako konbinazioak, forma kanonikoan aditza dutenak buru sintaktikotzat eta esanahi idiomatikoa dutenak.

Etiketa horietatik abiatuta, corpusetik automatikoki erauzi ditugu etiketatutako hitz-konbinazioen agerpen literalak izan litezkeenak, honako ideia hau oinarritzat harturik: UF baten barruko lemak corpusean elkarrekin agertu badira baina ez badira UF gisa etiketatu, baliteke hitz-konbinazio hori jatorrizko UFaren agerpen literal bat izatea. Metodologiari buruzko xehetasunak datorren atalean emango ditugu (7.2.2), eta azalduko dugu hautagaien erauzketa nola egiten den. Momentuz, baina, azal ditzagun lan honen ardatz diren hiru kontzeptuak: agerpen idiomatikoak, agerpen literalak eta kointzidentziazko agerpenak.

Aintzat harturik UF baten barruko lemak elkarrekin agertzen direla corpusean, hitz-konbinazio hori **agerpen idiomatiko**tzat (AI) jotzen dugu, baldin eta honako baldintza hauek betetzen baditu:

- Egitura sintaktikoa bat dator jatorrizko UFak forma kanonikoan duen egitura sintaktikoarekin, edo haren baliokidea da[11].

- Esanahi idiomatikoa du.

Bestalde, UF baten **agerpen literal**tzat (AL) jotzen ditugu honako ezaugarri hauek betetzen dituztenak:

- Egitura sintaktikoa bat dator jatorrizko UFak forma kanonikoan duen egitura sintaktikoarekin, edo haren baliokidea da.

- Ez du esanahi idiomatikorik.

Azkenik, **kointzidentziazko agerpen** (KA) deitu diegu honako ezaugarri hau betetzen duten hitz-konbinazioei:

- Egitura sintaktikoa ez dator bat etiketatutako UFak forma kanonikoan duen egitura sintaktikoarekin, eta ez da haren baliokidea.

Esate baterako, *adarra jo* UFa aintzat hartuta, 192. adibideko hautagaia agerpen idiomatikoa litzateke, 193.ekoa literala, eta 194.ekoa, berriz, kointzidentziazkoa.

(192) *Ez egin jaramonik,* ***adarra jotzen*** *ari zaizu eta.*
→ egitura sintaktiko berdina, esanahi idiomatikoa

(193) *Pinaburuak adarra jo zuen erortzean.*
→ egitura sintaktiko berdina, esanahi literala

(194) *Zuhaitzaren adar bat hautsi zuen baloia jo nahian.*
→ egitura sintaktiko desberdina

Azal dezagun orain zer metodologia erabili dugun halakoak corpusean aztertzeko.

[11] Aurreko atalean azaldu bezala (7.1.1.1), UFek hainbat aldaki morfosintaktiko izan ditzakete, eta forma kanonikoa hartzen da kontuan zalantzak argitzeko. UFen agerpenek egitura sintaktiko baliokideak dituztela diogu, forma kanonikora ekarrita egitura berbera badute.

7.2.2 Metodologia orokorra

Hasteko, lan honi zegozkion bost hizkuntzetako zatiak bildu ditugu PARSEMEren corpusetik: alemanezkoa, grezierazkoa, euskarazkoa, polonierazkoa eta portugesezkoa. Hortik abiatuta, etiketatutako UFen agerpen literal izan litezkeenak –hemendik aurrera, *hautagaiak*– erauzi, eta etiketak jarri dizkiegu. Hautagaiak nola erauzi ditugun azalduko dugu jarraian (7.2.2.1. atala), eta etiketatze-lanean erabilitako sailkapenaren berri emango dugu ondoren (7.2.3. atala).

7.2.2.1 Hautagaiak erauzteko heuristikoak

Esan bezala, corpus osoa hutsetik etiketatu beharrean, markatutako UF-etiketetatik abiatu gara haien agerpen literalak izan litezkeen hautagaiak erauzteko. Horretarako, UFtzat etiketatutako hitz-konbinazioen lemak erabili ditugu corpusean bilaketak egiteko.

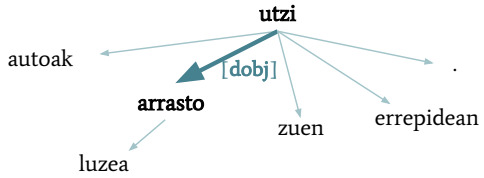
UF baten lemak corpusean agertu diren aldiro, lau heuristikok erabaki dute hautagaietan sartzekoak ziren ala ez. Azal dezagun zer egiten duten heuristiko horiek, 7.1. irudiko adibideen laguntzaz. Beheko zerrendan argituko dugu hitz-konbinazio baten lemek corpusean zer ezaugarri bete behar dituzten heuristikoek hautagaitzat erauz ditzaten. Eta, horrez gain, aintzat harturik corpusean *arrastoa utzi* UFa etiketatuta dagoela eta etiketa horri dagokion dependentzia-erlazioa objektu zuzena dela, heuristikoek 7.1. irudiko adibideekin zer egingo luketen ere esango dugu. Honako hemen lau heuristikoak:

- **WindowGap**: lemek hitz-leiho baten barruan agertu behar dute testuan, tartean gehienez ere bi hitz dituztelarik.
  → Hautagai bana erauziko luke lau esaldietatik, lauretan agertzen baitira *arrasto* eta *utzi* lemak tartean gehienez ere bi hitz dituztela.

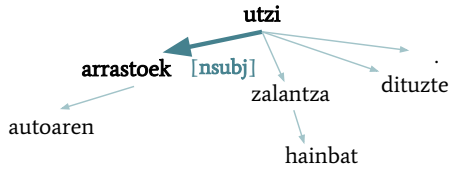
- **BagOfDeps**: lemek elkarri lotuta agertu behar dute dependentzia-zuhaitzean, baina berdin du zein hurrenkeratan dauden eta zer erlazio mota duten elkarren artean.
  → Hautagai bana erauziko luke (a), (b) eta (c) esaldietatik, baina ez (d) esaldidik, azken horretan *utzi* eta *arrasto* ez baitaude elkarri lotuta dependentzia-zuhaitzean.

- **UnlabeledDeps**: lemek elkarri lotuta agertu behar dute dependentzia-zuhaitzean, UF etiketatuaren osagaien noranzko berean.

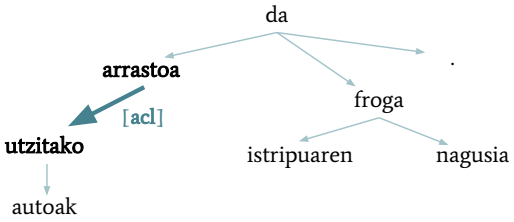
(a) Autoak arrasto luzea utzi zuen errepidean.



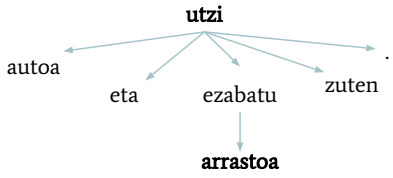
(b) Autoaren arrastoek hainbat zalantza utzi dituzte.



(c) Autoak utzitako arrastoa da istripuaren froga nagusia.



(d) Autoa utzi eta arrastoa ezabatu zuten.



7.1 irudia – Heuristikoen arabera *arrastoa utzi* UFaren agerpen literalak izan litezkeen lau adibide. Dependentzia-erlazioen azalpenak: *acl*, adjektibo-perpausa; *dobj*, objektu zuzena; *nsubj*, izen-subjektua. Etiketa horiei buruzko informazio gehiago, Aranzabe *et al.*-en lanean (2019).

→ Lehen bi esaldietatik hautagaiak erauziko lituzke, baina ez beste bietatik. Izan ere, (a) eta (b) esaldietan bakarrik agertzen dira bi lemak elkarri lotuta eta aditza izenaren gobernatzaile delarik.

- **LabeledDeps**: lemek elkarri lotuta agertu behar dute dependentziazuhaitzean, UF etiketatuaren osagaien noranzko berean eta erlazio mota berberarekin.
  → Lau esaldietatik hautagai bakarra erauziko luke, (a) esaldian bakarrik agertzen baita *arrasto* lema *utzi*ren objektu zuzentzat.

Behin hautagaiak erauzirik, etiketatze-lanari ekin diogu. Datorren atalean azalduko dugu hautagaiak nola sailkatu ditugun.

7.2.3 Etiketatze-lana eta gidalerroak

Hautagai-zerrendak osatu ondoren, hizkuntza bakoitzeko etiketatzaileei taula banatan gorde zaizkie honako datuak:

- Zein UFtatik abiatuta erauzi duten heuristikoek hautagai bakoitza
- Jatorrizko UFak zer etiketa zuen corpusean
- Hautagai bakoitzari dagokion esaldia osorik, hautagaia bera markatuta
- Analizatzaileak zer etiketa eman dien hautagaiaren barruko lemei
- Zer heuristikok erauzi du(t)en hautagai bakoitza

Informazio hori guztia eta kontzeptu nagusietan azaldutakoak (7.2.1. atala) kontuan harturik, hautagaiak sailkatzeko eskatu zaie etiketatzaileei. Hizkuntza bakoitzeko hiztun aditu banak hartu du parte atazan, portugesezko zatian salbu, horretan bi etiketatzailek jardun baitute.

Bederatzi multzoko sailkapena erabili dugu: lehen bost multzoek erroreekin dute zerikusia, bai PARSEMEren corpusekoekin eta bai hautagaiak erauzteko heuristikoekin ere; beste laurak, berriz, zuzenean daude lotuta gure aztergaiarekin. Azal dezagun zer sartu dugun multzo bakoitzean, eta eman dezagun adibide bana.

Erroreei dagozkien etiketak

1. ERR-FALSE-IDIOMATIC: hitz-konbinazio jakin bat UFtzat etiketatuta dago jatorrizko corpusean, baina ez da UFa benetan; corpuseko positibo faltsu bat da. Esate baterako, era horretakoa da 195. adibidea, *beharko liguke* ez baita UFa, 7.1.4.2. atalean argitu dugunez.

   (195) *Hausnarketa eragin beharko liguke horrek.*

2. ERR-SKIPPED-IDIOMATIC: hitz-konbinazio jakin bat ez dago UFtzat etiketatuta jatorrizko corpusean, baina egon beharko luke; corpuseko negatibo faltsu bat da. Beheko esaldian, adibidez, *atzera bota* UFa etiketatu gabe zegoen.

   (196) *Udalbatzak aho batez erabaki du alegazioa* ***atzera botatzea****.*

3. NONVERBAL-IDIOMATIC: Idiomatikoak izan arren aditz-UFak ez diren hitz-konbinazioak, azterketa honetatik kanpokoak. Esate baterako, *parte-hartze* izen elkartua da 197. adibidean eta, aditz-UF batetik eratorritakoa izan arren, ez da gure aztergaian sartzekoa (azalpen gehiago, 7.1.1.1. atalean).

   (197) *Nahiko nahasiak izaten dira parlamentarioen parte hartzeak.*

4. MISSING-CONTEXT: testuinguru gehigarririk gabe anbiguoak diren hitz-konbinazioak. Beheko esaldian, adibidez, ondoko beste esaldi batzuk irakurri gabe ezin da jakin *ateak zabaldu* metaforikoki ala literalki erabilita dagoen.

   (198) *Ateak zabaldu zizkion.*

5. WRONG-LEXEME: lema edo gramatika-kategoria gaizki analizatzetik sortutako erroreak. *Eragin* hitza, adibidez, aditza da 199. adibidean, eta ez izena *eragina izan* UFan bezala.

   (199) *Presa baten matxurak hondamendia eragin du.*

Agerpen literalei eta kointzidentziazkoei dagozkien etiketak

6. COINCIDENTAL: lemak eta gramatika-kategoriak zuzenak dira, baina dependentzia-erlazioa ez da UFarenaren berdina. Adibidez, *eragina izan* UFaren kointzidentziazko agerpen bat dago esaldi honetan:

   (200) *Egunero sufritzen dugu obrako zarataren eragina.*

7. LITERAL-MORPH: agerpen literal bat da, eta UFtik bereiz daiteke murriztapen morfologikoak kontuan hartuz. Esate baterako, 201. adibidekoa *aurrera egin* UFaren era horretako agerpen literal bat da:

   (201) *Larrialdien aurrean egin beharrekoa jasotzen du txostenak.*

8. LITERAL-SYNT: agerpen literal bat da, eta UFtik bereiz daiteke murriztapen sintaktikoak kontuan hartuz. *Gauza izan* UFtik 202. adibideko hitz-konbinazioa bereizteko, adibidez, nahikoa da izenarekin batera izen-sintagma horretan dauden determinatzaileari eta modifikatzaileari begiratzea, UFak ez baitu halakorik onartzen.[12]

   (202) *Bi gauza desberdin dira.*

9. LITERAL-OTH: agerpen literal bat da, eta murriztapen morfosintaktikoak ez dira nahikoa UFtik bereizteko; testuinguruari, semantikari edo hizkuntzaz kanpoko ezaugarriei begiratu beharra dago. Horixe gertatzen da, adibidez, esaldi honetan, ezin baita horko *izena jarri* literala idiomatikotik bereizi ez bada esanahia kontuan hartuta.

   (203) *Agiriak bete beharko dituzte, deialdiaren izena jarriz.*

7.2.4 Emaitza orokorrak

Etiketatze-lanaren emaitzetik ateratako estatistika nagusiak 7.4. taulan jaso ditugu. Oro har, idiomatikotasun-tasa oso altua da, eta emaitzak oso antzekoak dira hizkuntza guztietan: % 96tik % 98ra bitartekoak. Hortaz, lehen ondorio nagusia da gure hipotesia betetzen dela: UFen agerpen idiomatikoak oso urriak dira testu errealetan.

[12] Halako kasuak morfologiaren eta sintaxiaren arteko mugakoak diren arren, ataza honetan zera erabaki dugu: fenomeno morfologikotzat hartzea hitz baten barruan gertatzen diren aldaketa guztiak (*adarra, adarretik, adarrek...*), eta sintaktikotzat hitzetik kanpo gertatzen direnak (*adar bat, adar luzeak...*).

	DE	EL	EU	PL	PT
UF etiketatuak	3,823	2,405	3,823	4,843	5,536
Hautagaiak	926	451	2,618	332	1,997
ERR-FALSE-IDIOM	21,5%(199)	12,0% (54)	9,4% (246)	0,0% (0)	3,8% (76)
ERR-SKIPPED-IDIOM	27,0%(250)	47,5%(214)	17,3% (453)	5,4% (18)	10,7%(213)
NONVERBAL-IDIOM	0,0% (0)	0,0% (0)	0,2% (6)	0,0% (0)	0,5% (9)
MISSING-CONTEXT	0,3% (3)	0,2% (1)	0,5% (12)	2,1% (7)	0,7% (13)
WRONG-LEXEMES	40,1%(371)	0,9% (4)	26,7% (700)	1,8% (6)	38,1%(760)
COINCIDENTAL	**2,6%** (24)	**27,9%**(126)	**42,4%**(1110)	**61,1%**(203)	**33,5%**(668)
LITERAL	**8,5%** (79)	**11,5%** (52)	**3,5%** (91)	**29,5%** (98)	**12,9%**(258)
↪ literal-morph	0,8% (7)	5,5% (25)	1,9% (51)	1,2% (4)	3,7% (73)
↪ literal-synt	1,5% (14)	2,0% (9)	0,7% (19)	8,1% (27)	2,2% (44)
↪ literal-other	6,3% (58)	4,0% (18)	0,8% (21)	20,2% (67)	7,1%(141)
Idiomatikotasun-tasa	**98%**	**98%**	**98%**	**98%**	**96%**

7.4 taula – Etiketatze-lanaren estatistika orokorrak, hizkuntza guztietan. Idiomatikotasun-tasa (*Idiomaticity rate*) honela kalkulatzen da: idiomatikoak/(literalak+idiomatikoak).

Arrazoi tipologikoak direla medio, hautagaien kopuruan alde nabarmena dago hizkuntza batetik bestera: mutur batean poloniera dago, heuristikoek 384 hautagai erauzi baitituzte 5.152 UF etiketatatik abiatuta, eta beste muturrean, berriz, euskara, 2.618 hautagai izan baitira 3.823 UF etiketatatari esleitutakoak. Aurreraxeago azalduko dugu alde horren zergatia, 7.2.5. atalean.

Bestalde, taula horrek erakusten du morfosintaxiaren bidez ebatzi ezin diren kasuak oso gutxi direla, eta euskaraz bereziki. Hain zuzen, erroredun hautagaiak alde batera utzita, gainerakoen artean % 2 baino gutxiago dira multzo horretakoak.

Hizkuntzaren Prozesamenduari begira, interesgarria da kointzidentziazko agerpenak ere kontuan hartzea. Izan ere, eskuzko etiketatze-lanean erraz samar bereizten dira UFak kointzidentziazko agerpenetatik, baina UFen identifikazio-lan automatikoan garrantzi handia dute agerpen horiek ere, 4. kapituluan zehar erakutsi nahi izan dugunez. Hortaz, 7.4. taulan jasotako idiomatikotasun-tasez gain, idiomatikotasun-, kointzidentziazkotasun- eta

	DE			EL			EU			PL			PT		
	LVC	**VID**	**All**	**LVC**	**VID**	**All**	**LVC**	**VID**	**All**	**LVC**	**VID**	**All**	**LVC**	**VID**	**All**
Id-tasa	100	99	**98**	99	95	**98**	99	93	**98**	99	96	**98**	99	88	**96**
EIR	100	97	**98**	94	92	**94**	86	58	**78**	94	90	**94**	92	73	**86**
ECR	0,3	1	**0,6**	5	3	**5**	14	37	**20**	5	7	**4**	7	18	**10**
ELR	0	1	**2**	1	5	**2**	1	5	**2**	1	3	**2**	1	10	**4**

7.5 taula – Idiomatikotasun-tasa zabaldua (*Extended Idiomaticity Rate*, EIR), kointzidentziazkotasun-tasa zabaldua (*Extended Coincidentality Rate*, ECR) eta literaltasun-tasa zabaldua (*Extended Literality Rate*, ELR).

literaltasun-tasa *zabaldu*ak ere bildu ditugu 7.5. taulan[13], hau da, kointzidentziazko agerpenak ere kontuan hartuz kalkulatutakoak[14].

Oro har, idiomatikotasun-tasa zabaldua ere altua da hizkuntza guztietan, baina, idiomatikotasun-tasa soilean ez bezala, alde handia dago hizkuntza batetik bestera. Kasu honetan ere, euskara dago mutur batean, gurea izan baita tasarik apalena, % 78koa. Alde hori euskarazko kointzidentziazko agerpenen ugaritasunak eragin du batez ere, gure kointzidentziazkotasun-tasa % 20koa baita, hurrengo tasarik altuena duen hizkuntzarena halako bikoa.

Hautagai kopuru altuaz gain, kointzidentziazko agerpenen ugaritasuna ere zuzenki lotuta dago euskararen ezaugarri tipologikoekin, eta 7.2.5. atalean hitz egingo dugu horretaz. Lehenago, ordea, heuristikoen emaitzak aztertuko ditugu.

7.2.4.1 Heuristikoen emaitzen azterketa

Heuristikoek agerpen literalak hautematerakoan lortutako doitasuna (P), estaldura (R) eta F neurria (F) 7.6. taulan jaso ditugu. Hizkuntzaz hizkuntza eta heuristikoz heuristiko bildu ditugu datuak, bai eta heuristiko guztiak batuta ere.

Doitasunak eskuz literaltzat etiketatutako hautagaiei egiten die erreferentzia, hau da, heuristikoek erauzitako hautagaietatik zenbat izan diren benetan literalak. Espero izatekoa zen bezala, LabeledDeps heuristikoak lor-

[13]Kapitulu honetan bereziki euskarazko UFez ari garenez, euskararako aplikagarriak diren bi etiketak bakarrik sartu ditugu taulan. Nolanahi ere, gogoan hartu behar da gainerako hizkuntzek beste kategoria bat edo bi ere badituztela; datu osatuagoak lan honi buruzko artikuluan daude jasota (Savary *et al.*, 2019).

[14]Idiomatikotasun-tasa zabaldua: idiomatikoak/(lit+idiom+kointz)

	WindowGap			**BagOfDeps**			**UnlabeledDeps**			**LabeledDeps**			**Denak**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EU	0.05	**0.94**	0.05	0.07	0.72	0.06	0.06	0.50	0.06	0.07	0.18	0.05	0.05	1.00	0.05
DE	0.08	0.78	0.07	0.12	**0.90**	0.11	0.13	0.90	0.11	0.14	0.77	0.12	0.09	1.00	0.08
EL	0.11	0.86	0.10	0.15	**0.88**	0.13	0.15	0.80	0.13	0.16	0.51	0.12	0.11	1.00	0.10
PL	0.30	**0.96**	0.23	0.43	0.75	0.27	0.49	0.69	0.28	0.52	0.22	0.15	0.27	1.00	0.21
PT	0.14	**0.98**	0.13	0.17	0.62	0.14	0.20	0.59	0.15	0.34	0.37	0.18	0.13	1.00	0.11

7.6 taula – Heuristikoen doitasuna, estaldura eta F neurria.

tu du doitasunik altuena hizkuntza guztietan, eta WindowGap-ek, berriz, baxuena, horiek baitira, hurrenez hurren, erauzketarako murriztapen gehien eta gutxien jartzen dituzten heuristikoak.

Lan honetan, ordea, estaldura ahalik eta handiena lortzeari eman diogu garrantzia, hautagaiak eskuz etiketatu behar genituelako. Ataza honetan, zehazki, estaldurak honako hau erakusten du: heuristiko guztiek erauzitako hautagaietatik heuristiko bakoitzak zenbat erauzi dituen. Hortaz, heuristiko guztiak batuta lortzen den estaldura erabatekoa izateak ez du esan nahi corpusean egon litezkeen agerpen literal guztiak erauzi direnik, kontuan hartzen ditugunak heuristiko batek gutxienez erauzitakoak bakarrik baitira. Dena dela, Savary eta Cordeiro-ren lanean (2018) erakusten denez, hautagaiak erauzteko hurbilpen hori oso zehatza da polonieraz behintzat, lehen 1.000 esaldiak aztertuta ez baita agerpen literal bakar bat ere kanpoan utzi.

WindowGap heuristikoaren estaldura denetan altuena izan da, alemanierarako eta grezierarako izan ezik. Alemanierazko emaitza bereziki baxua da, baina bat dator PARSEMEren corpusak erakusten duen ezaugarri batekin (Savary *et al.*, 2017): UFko osagaien artean zenbat hitz dauden kalkulatuta, alemanieraz gainerako hizkuntzetan baino askoz ere gehiago daude, ia hiru hitz batez beste.

Oro har, esan liteke espero zena bete dela: WindowGap-ek estaldura handiagoa du BagOfDeps-ek baino, BagOfDeps-ek hobea UnlabeledDeps-ek baino, eta UnlabeledDeps-ek hobea LabeledDeps-ek baino. Eta doitasunari begiratuta, kontrako noranzkoan gertatzen da hori. Alegia, zenbat eta murriztapen gehiago hartu kontuan, orduan eta hobea da doitasuna, baina estaldura okerragoa.

7.2.5 Ondorio linguistikoak

Behin emaitzak ikusita, azal dezagun oro har nolakoak diren multzo bakoitzean sailkatutako hautagaiak. Atal honetan, agerpen literalen berri emango

dugu lehenik (7.2.5.1. atala), kointzidentziazkoen berri ondoren (7.2.5.2. atala), eta, azkenik, ohar batzuk egingo ditugu erroredun agerpenen inguruan ere (7.2.5.3. atala).

7.2.5.1 Agerpen literalen ezaugarriak

Agerpen literalen ezaugarriak desberdin samarrak dira UF mota batetik bestera. Esate baterako, 7.5. taulan ikus daiteke idiomatikotasun-tasa askoz ere altuagoa dela LVCetan VIDetan baino, edozein hizkuntzaz ari garela ere. Euskaraz eta portugesez, bereziki alde nabarmena dago batzuen eta besteen artean: % 58tik % 85erakoa eta % 73tik % 92rakoa, hurrenez hurren.

Izan ere, LVCak nahiko konposizionalak dira semantikari dagokionez, izenak bere ohiko esanahia gordetzen baitu eta aditzak ezaugarri morfologikoak baino ez baitizkio gehitzen normalean (7.1.1.2. atala). Intuizioz ere, ez da hain erraza multzo horretako UFek agerpen literalak dituztela pentsatzea. Dena dela, badaude halako kasu batzuk, non LVC barruko bi lemak elkarrekin agertzen baitira baina ez baitituzte LVCaren ezaugarriak betetzen. Hori gertatu ohi da, adibidez, izen batek esanahi predikatiboa eta ez-predikatiboa izan dezakeenean, alegia, batzuetan bakarrik egiten dionean erreferentzia ekintza edo egoera bati (204–205. adibideak).

(204) *Sekulako **laguntza eman** dit kirolean eta kiroletik kanpo.*

(205) *Enpresa berriak sustatzeko laguntzak emango ditu Udalak.*

Goiko adibideetako lehenengoan, *laguntza eman* idiomatikoki dago erabilita, baina ez bigarrenean, diru-laguntza bati buruz ari baita eta ez laguntzeko ekintzari buruz. Esanahi idiomatikoan *laguntza* beti singularrean erabiltzen denez eta 205. adibidean pluralean dagoenez, bigarren adibide horri eta halakoei LITERAL-MORPH etiketa eman diegu.

Bestalde, VID motako UF asko metaforetatik datoz, eta intuizioz errazagoa da halakoek esanahi figuratiboa eta literala dutela pentsatzea. Era horretakoa da, adibidez, *atzera bota* UFa; 206. esaldian idiomatikoki erabilita dago, eta 207.ean, aldiz, literalki, zerbait fisikoki atzerantz botatzea adierazten baitu. Esaldi horri LITERAL-OTH etiketa dagokio.

(206) *Irakasleen eskaerak **atzera bota** ditu Hezkuntzak.*

(207) *Pase aparta eman, eta baloia atzera bota dio taldekideari.*

VIDetako asko bereiz daitezke morfosintaxiari begiratuta, bereziki euskaraz, grezieraz eta portugesez. Esate baterako, *gai izan* UFan izena ez da inoiz adjektibo batez lagunduta egoten, eta ezaugarri horrexeren bidez jakin liteke 209. adibideko agerpena literala dela, LITERAL-SYNT motakoa zehazki.

(208) *Lau langiletik bat **gai da** euskaraz aritzeko.*

(209) *Horixe da gaurko gai nagusia.*

7.2.5.2 Kointzidentziazko agerpenen ezaugarriak

Euskara da, alde handiz, kointzidentziazko agerpen gehien dituen hizkuntza (7.4. taula). Kointzidentziazkotasun-tasa zabaldurik altuena ere badu (7.5. taula), bereziki VIDetan.

Kointzidentziazko agerpen askok eta askok postposiziodun izenak dituzte barnean, jatorrizko UFan izenak halakorik ez bazuen ere. Izan ere, heuristikoek lemari begiratuta bakarrik erauzten dituzte hautagaiak, eta, postposizioak eta kasu-markak lematik kanpo geratzen direnez, izenari eransten zaizkion markak ez dira kontuan hartzen.

Hori gertatzen da, esate baterako, ondorengo adibideetan. *Aurre egin* UFa kontuan hartuta, *aurre* eta *egin* lemak bilatu dira, eta hautagaitzat erauzi dira 211. eta 212. adibideetako hitz-konbinazioak.

(210) *Arazoei **aurre egin** zien.*

(211) *Donostiako udaletxearen aurrean egin dute elkarretaratzea.*

(212) *Hitz egiten hasi aurretik egin beharrekoak.*

Era horretako konbinazioetan, postposizioa gehitzeak izenaren eta aditzaren arteko erlazio sintaktikoa aldatzen du, eta, hortaz, hautagaia kointzidentziazko agerpentzat sailkatzen da. Etiketa bestelakoa litzateke jatorrizko UFak postposizioren bat barne hartuko balu (esate baterako: *atzera egin*), litekeena baita halakoetan erlazio sintaktikoa berbera izatea hautagairen batean ere (adibidez: *atzean egin*), eta LITERAL-MORPH etiketa beharko bailuke kasu horietan.

Bistan denez, heuristikoak ez dira hizkuntza aglutinatiboetan pentsatuz sortu hasiera batean. Ahalik eta heuristikorik orokorrenak sortu nahi izan direnez, lemari bakarrik begiratu zaio, baina horrek alferreko hautagai asko eta asko erauzarazi ditu euskaraz. Proposamen gisa, hobe litzateke postposiziomarkak ere lexikalizatutzat hartzea beti, gainerako hizkuntzetan preposizio lexikalizatuak –hitz beregainak izanik– kontuan hartzen diren modu berean.

7.2.5.3 Erroredun agerpenen ezaugarriak

Erroredun agerpenetan, WRONG-LEXEME multzoan sailkatutako hautagaietan jarriko dugu arreta, horiek baitira heuristikoekin zuzenean loturikoak. Gainerakoak corpuseko etiketatze-lanean eginiko akatsak dira, etiketatzaileei ihes egindako UFak edo etiketatu behar ez zirenak.

Hasieran esan dugunez, heuristikoek hitzen lemari begiratzen diote hautagaiak erauzteko, baina ez gramatika-kategoriari; hortik etorri dira errore dezente, bereziki alemanari, euskarari eta portugesari dagokienez. Akats gehienak hitz homografoek sortuak izan dira. Adibidez, euskaraz, 7.1.4.3. atalean esan dugunez, badaude izenen berdin-berdinak diren hainbat adjektibo, eta ezaugarri horrexeren eraginez erauzten dira 214. adibideko hitz-konbinazioa bezalakoak.

(213) *Planaren **berri eman** ziguten.*

(214) *Plan berria eman ziguten.*

(215) *Plana berriz eman ziguten.*

Aldiz, 215. adibideko *berriz* adberbioa *berri* lemari postposizio instrumentala erantsiz sortua da. Analisi morfosintaktikoa eskuz egina izan balitz, lematzat *berriz* esleituko zitzaion zuzenean, baina corpusaren zati bat automatikoki etiketatua denez, behin baino gehiagotan etiketatu da postposiziodun izentzat.

Laburpena

Kapitulu honetan, PARSEME proiektu europarrean eginiko ekarpenen berri eman dugu. Lehenik, azaldu dugu euskarazko corpus bat nola etiketatu dugun fraseologia mailan, eta bigarrenik, UFen agerpen literalen inguruan aritu gara, corpus horretatik bertatik abiatuta. Lan horiek 1.3. ataleko lau abiapuntu-hipotesi berresteko balio izan digute.

[A2] Aditz-UFak oso malguak izan ohi dira morfosintaxiari dagokionez, baina murriztapenak ere badituzte.
Aditz-UFen agerpen literalak aztertu ditugunean, ikusi dugu literaltzat etiketatutako hitz-konbinazioetako gehienak (% 80) ezaugarri morfosintaktikoen bidez bereiz daitezkeela agerpen idiomatikoetatik. Hortaz, corpusak ere agerian uzten du UF batzuek badituztela murriztapen morfosintaktikoak.

[A3] Fraseologia mailan aztertutako hizkuntza askoren aldean, euskaraz bereziki ohikoak dira aditz arinak barne hartzen dituzten UFak.
Hala erakusten du PARSEMEren corpusak: batez beste, 100 esalditik 27k aditz arindun konbinazioren bat dute barnean. Hogei hizkuntzatako testuak etiketatu dira gidalerro berberei jarraituz, eta bik bakarrik dauzkate euskarak baino aditz arindun konbinazio gehiago esaldiko: persierak eta hindiak. Gainera, euskal hiztunek gehien hitz egiten dituzten erdarei begiratuta, aldea are nabarmenagoa da: euskarazko aditz arindun konbinazioen maiztasuna hiru aldiz handiagoa da frantsesezkoa eta gaztelaniazkoa baino, eta ingelesarena halako seikoa.

[A4] UF asko literalak zein idiomatikoak izan badaitezke ere, praktikan gutxitan erabiltzen dira literalki testu errealetan.
Familia desberdineko bost hizkuntzatako corpusak aztertu ditugu, eta erakutsi dugu hala dela, aditz-UFak oso gutxi erabiltzen direla literalki praktikan: agerpen idiomatikoak eta literalak kontuan hartuta, idiomatikotasun-tasa % 96–98 artekoa da hizkuntza guztietan, % 98koa euskaraz.

[A5] UFen inguruko informazio morfosintaktikoa kontuan hartzeak haien identifikazioa hobetu dezake.

Corpus etiketatu bateko UFen lemak oinarritzat hartu, eta corpus beretik haien agerpen literalak izan litezkeen hautagaiak erauzi ditugu, hitz-leihoak eta dependentzia sintaktikoak erabiliz. Hizkuntza ia guztietan, erauzitako hautagai gehienek ez dute UFaren egitura sintaktiko berbera, eta gainerakoen artean ere oso gutxi dira morfosintaxiaren bidez ebatzi ezin direnak; euskaraz, zehazki, % 2 baino ez.

Bestalde, kapitulu honetako bi azterketen bidez, 1.3. atalean zerrendatutako hiru helburu betetzeko pausoak eman ditugu.

[H1] Gaztelaniazko eta euskarazko aditza+izena motako UFen ezaugarri morfosintaktikoak aztertzea.
Kasu honetan, euskarazko aditz-UFei begiratu diegu zehazki. Corpus etiketatuari begiratuta, emaitzetako batzuk bat datoz lehenagoko lanekin; esate baterako, corpusean ikusten da aditza+izena motako UFen artean absolutiboan daudela gehien-gehienak, hiztegietan bezalaxe.

[H3] Beste hizkuntza batzuetan ere erabilgarriak izan litezkeen azterketa-metodologiak erabiltzea.
PARSEME proiektuaren baitan sortutako gidalerro unibertsalak geure egin, eta euskarazko corpusa etiketatu dugu, beste hogei hizkuntzatakoarekin batera argitara zedin. Gainera, gidalerroak hobetzeko prozesuan parte hartu dugu, eta amaitu ondoren ere proposamenak egin ditugu etorkizunerako. Bestalde, agerpen literalak aztertzean ere corpusa ustiatzeko eta etiketatzeko metodologia berbera erabili dugu bost hizkuntzatan.

[H7] Gaztelaniazko eta –bereziki– euskarazko UFen corpus etiketatuak sortzea.
PARSEMEren ataza partekatuetarako, hogei hizkuntzatako corpus etiketatuak sortu dira. Guk gaztelaniazko corpusa etiketatzen lagundu dugu lehen ediziorako, eta euskarazkoa sortu dugu bigarrenerako. Gaztelaniazkoak 5.515 esaldi ditu (bi edizioetarako eginiko lana kontuan hartuta), eta 2.739 aditz-UF guztira; euskarazkoak, berriz, 11.158 esaldi ditu, eta 3.823 aditz-UF.

8. KAPITULUA

Ondorioak, ekarpenak eta etorkizuneko lanak

Tesi-txosten honen hasierako kapituluan (1.3. atala), lanerako abiapuntutzat hartu ditugun hipotesiak eta helburuak zerrendatu ditugu. Kapitulu hau azkena izanik, zerrenda horri helduko diogu berriro, eta hala emango diogu amaiera txostenari: hipotesietatik abiatuta laburbilduko ditugu lan osotik atera ditugun ondorio nagusiak (8.1. atala), eta helburuetatik abiatuta, egin ditugun ekarpenak (8.2. atala). Azkenik, etorkizuneko ikerketa-ildoei ere eskainiko diegu tartea (8.3. atala).

8.1 Ondorioak

Tesi-lan hau egiteko, gure hipotesi nagusia izan da informazio linguistikoa, lexikoa eta morfosintaktikoa bereziki, lagungarria dela UFen tratamendu konputazionala hobetzeko. Eginiko lanen bidez, hori hala dela erakutsi dugu, informazio linguistikoa erabiliz lortu baitugu UFen kalitatezko identifikazio-metodo bat garatzea, bai eta itzultzaile automatiko baten emaitzak hobetzea ere, neurri txikiagoan.

Hipotesi nagusi horrez gain, beste sei azpihipotesi ere izan ditugu gogoan: lehen laurek fraseologiarekin eta hizkuntzaren azterketarekin dute lotura, eta beste biek, ostera, hizkuntzaren prozesamenduarekin. Egin ditugun lanek azpihipotesi guztiak berresteko balio izan digute, eta, hortaz, hipotesiok ondorio ere badira, nolabait. Zerrenda ditzagun, beraz, sei ondorioak hemen, eta laburbil dezagun zer ebidentzia lortu dugun gure lanetan ondorio horietara

heltzeko.

[1] UFak, askotan, ez dira hitzez hitz itzultzen hizkuntza batetik bestera. Gure lanen arabera, hala gertatzen da gaztelaniaren eta euskararen artean. Batetik, *Elhuyar* hiztegi elebidunean jasotako aditza+izena konbinazioak eta haien ordainak aztertu ditugunean, ikusi dugu gaztelaniazko sarreren erdiak eta euskarazkoen herenak bakarrik dituztela aditz batez eta izen batez osaturiko konbinazioak ordain gisa eta, gainera, horien laurdena bakarrik direla hitzez hitzeko itzulpenak; hau da, osotara, gaztelaniazko sarreren % 11ri eta euskarazko sarreren % 7ri bakarrik ematen zaie hitzez hitzeko ordaina beste hizkuntzan. Bestetik, geroko lanetan ere argi ikusi dugu ezaugarri horrek arazoak sortzen dituela Hizkuntzaren Prozesamenduko ataza batean baino gehiagotan, bereziki, UFen ordainak corpus paraleloetatik automatikoki erauzteko lanean –guk hala erauzitako ordainen erdia inguru okerrak izan baitira– eta erregeletan oinarritutako itzulpen automatikoan, non itzulpenak hitzez hitz egiten baitira oro har.

[2] Aditz-UFak oso malguak izan ohi dira morfosintaxiari dagokionez, baina murriztapenak ere badituzte. Identifikazio-lanerako aztertu ditugun gaztelaniazko aditza+izena UFetatik, bat ere ez da erabat finkoa: aztertutako UFen azken multzoan, % 36 UF erabat malguak dira, eta gainerako guztiak erdifinkoak. Horrez gain, UFen agerpen literalak aztertu ditugunean, ikusi dugu agerpen literal horietako gehienak morfosintaxiaren bidez bereiz daitezkeela agerpen idiomatikoetatik. Hain zuzen ere, euskarazko etiketatze-lanean, % 1era ere ez da iritsi morfosintaxiaren bidez ebatzi ezin diren agerpen literalen kopurua, eta gainerako hizkuntza gehienetan ere –polonieraz izan ezik– oso apala izan da ehuneko hori.

[3] Fraseologia mailan aztertutako hizkuntza askoren aldean, euskaraz bereziki ohikoak dira aditz arinak barne hartzen dituzten UFak. *Elhuyar* hiztegitik erauzi ditugun izena+aditza konbinazioetatik, erdia baino gehiago sei aditzik usuenekin osatzen da, eta aditz horiek (*egin, izan, eman, hartu, egon* eta *jarri*), hain zuzen, arinak izan ohi dira UFen barruan. Horrez gain, PARSEMEren euskarazko corpusa beste hemeretzi hizkuntzaren irizpide berberei jarraituz etiketatu ostean, etiketak beste hizkuntzekin alderatu ditugu, eta ikusi dugu, batez beste, bi hizkuntzak baino ez dutela erabiltzen euskarak

baino aditz arindun konbinazio gehiago esaldi bakoitzeko: persierak eta hindiak. Maiztasunari dagokion aldea bereziki nabarmena da erkaketa gaztelaniarekin, frantsesarekin eta ingelesarekin eginez gero: euskarazko aditz arindun konbinazioen maiztasuna (100 esalditik 27) hiru aldiz handiagoa da frantsesezkoa eta gaztelaniazkoa baino, eta ingelesarena halako seikoa.

[4] Hitz-konbinazio asko literalak zein idiomatikoak izan badaitezke ere, praktikan ia beti erabiltzen dira idiomatikoki, hau da, UF gisa. PARSEMEren corpus etiketatua oinarritzat harturik, UF izan ohi diren hitz-konbinazioen agerpen literalak aztertu ditugu, familia desberdineko bost hizkuntzatan: alemanez, grezieraz, euskaraz, polonieraz eta portugesez. Idiomatikotasun-tasa oso altua da guztietan ere, agerpen guztien % 2 inguru bakarrik baitira literalak.

[5] UFen inguruko informazio morfosintaktikoa kontuan hartzeak haien identifikazioa hobetzen du. Analizatzaile morfosintaktiko baten emaitzei UFen inguruko informazio morfosintaktikoa gehituta, identifikazioaren kalitatea nabarmen hobetzea lortu dugu. Lehenik, datuak banan-banan eta eskuz aztertu ditugu UF gutxi batzuentzat, eta, ondoren, azterketa-prozesu hori automatizatzeko metodo bat proposatu dugu. Gure esperimentuek erakutsi dutenez, azterketa erabat automatikoki eginda ere identifikazio-lanaren emaitzak nabarmen hobetzen dira, azterketa-metodo horrek oinarrian dituen irizpide linguistikoak baliagarriak diren seinale. Hain zuzen ere, identifikatu nahi diren UF guztien zerrenda aldez aurretik ezagutuz gero, 0,72ko F neurria lortzen dugu, eta emaitza hori nabarmen hobea da PARSEMEren ataza partekatuan parte hartu duten beste sistemena baino.

[6] UFei buruzko informazio morfosintaktikoa kontuan hartzea onuragarria da itzultzaile automatikoentzat. *Matxin* itzultzaile automatikoan UFen inguruko informazio linguistikoa gehituta, UF-itzulpenen % 63 inguru hobetzen dira gure eskuzko ebaluazioaren arabera, eta oso gutxi okertzen dira, % 8 inguru. Metrika automatikoek ere erakusten dute sistemaren itzulpen-kalitatea hobetu egiten dela informazio horren bidez, baina kontuan hartu behar da hobekuntza hori oso txikia dela estatistikoki –% 2,25ekoa BLEU metrikaren arabera–, eta eragina ez dela hain nabarmena UF-itzulpenei soilik begiratu beharrean sistemaren kalitate osoari begiratuz gero.

8.2 Ekarpenak

Hipotesiak berresteaz gain, tesi-lanaren helburuak ere bete ditugu, eta helburu horietatik abiatuta laburbilduko ditugu gure ekarpenak. Oro har, esan dezakegu helburu nagusia bete dugula, hau da, aditza+izena motako UFen azterketa linguistikoa egin dugula eta, informazio horren bidez, era horretako UFen tratamendu konputazionala hobetzea lortu dugula. Dena dela, bide horretan, ekarpen gehiago ere egin ditugu, eta horiez jardungo dugu jarraian. Hona hemen gure ekarpenak:

[1] Gaztelaniazko eta euskarazko aditza+izena motako UFen ezaugarri morfosintaktikoak HPrako era aplikagarrian aztertu izana. Beste egile batzuek eginiko azterketen aldean, gure ekarpena izan da, batetik, aztertu ditugun datu gehienak kuantifikatu egin ditugula, aditza+izena motako UFen ezaugarri bereizgarrien proportzioak zenbatekoak diren argiago erakusteko asmoz, eta, bestetik, landutako ezaugarriak HPren alorrerako era aplikagarrian aztertu ditugula. Horrez gain, 6. ekarpenean ere azalduko dugunez, datu horiek guztiak publiko egin ditugu.

[2] Aditza+izena motako UFak gaztelaniaren eta euskararen artean nola itzultzen diren aztertu izana. Gaztelaniaren eta euskararen arteko UFen itzulpena ia ikertu gabeko esparrua izan da orain arte, eta urrats batzuk egin ditugu bide horretan. Hasteko, *Elhuyar* hiztegiko aditza+izena motako sarrerak eta haien ordainak aztertu ditugu, eta, horrez gain, beste baliabide batzuetako UFei ere automatikoki lortu dizkiegu ordainak corpusetatik. Horietan guztietan, hizkuntza bateko UFak beste hizkuntzara ekartzean gertatzen diren aldaketa lexiko eta morfosintaktikoei erreparatu diegu.

[3] Beste hizkuntza batzuetan ere erabilgarriak izan litezkeen azterketa-metodologiak sortu eta erabili izana. Lehenik, identifikazioa hobetzeko egin dugun lana beste hizkuntza batzuetan erabilgarria ote den ikusteko, gure lehen esperimentua ingelesez ere egin dugu, eta ikusi dugu hizkuntza horretan ere emaitza onak lortu ditugula. Bigarrenik, gaztelaniazko UFen azterketa automatizatzeko egin dugun metodoproposamena ere erraz molda daiteke beste hizkuntza batzuetara; izan ere, guk metodo horixe egokitu eta erabili dugu euskarazko ordainen informazioa corpusetatik eraucteko ere. Hirugarrenik, euskarazko corpus

bat fraseologia mailan etiketatzeko, PARSEME proiektuaren gidalerro unibertsalen irizpideak jarraitu ditugu, beste hemeretzi hizkuntzak egin duten bezala. Laugarrenik, UFen agerpen literalak aztertu ditugunean ere, bost familiatako hizkuntza bana hartu dugu oinarritzat azterketa-irizpideak ezartzeko. Eta, azkenik, aipatzekoa da guk landutako UFak eta azterketa-metodologia gaztelania-katalana hizkuntza-bikoterako ere berrerabili dela berriki, Bartzelonako Unibertsitateko Filologia Fakultatean ikasle batek aurkezturiko proiektuan.

[4] Aditza+izena motako UFen identifikazio automatikoa hobetu izana. Gure identifikazio-proposamenak nabarmen hobetzen ditu beste metodo batzuen emaitzak. Besteak beste, PARSEMEren ataza partekatuko baldintzetara ahalik eta gehien gerturatuta egin dugun esperimentu baten bidez, erakutsi dugu proposamen horrek emaitza hobeak lortzen dituela gaztelaniazko corpusean eginiko beste identifikazio-lanek baino: F neurriari dagokionez, 0,51ko marka lortu dugu, batez bestekoa baino 28 puntu hobea eta emaitzarik onena baino 13 puntu hobea. Garrantzitsua da, batez ere, agerian jarri dugula zehazki zer ezaugarri morfosintaktikok laguntzen duten ataza horretan.

[5] Aditza+izena motako UFen informazio linguistikoak itzulpen automatikoan zer eragin duen aztertu izana. Erregeletan oinarritutako *Matxin* itzultzailea oinarritzat harturik, UF sorta baten eta haien ordainen inguruko informazio lexikoa eta morfosintaktikoa erabili dugu, eta *Matxin*ek UF horiek hobeto itzultzea lortu dugu, bai metrika automatikoen eta bai eskuzko ebaluazio baten arabera: BLEU metrikan % 2,25eko hobekuntza lortu dugu, eta, eskuzko ebaluazioak dioenez, UF-itzulpenen % 62-65 hobetzen dira informazio morfosintaktikoa erabiliz; okertu, berriz, % 7-8 bakarrik. Nolanahi ere, 6. ondorioan esan dugun bezala, hobekuntza horrek sistemaren itzulpen-kalitate osoan duen eragina txiki samarra da.

[6] Aditza+izena motako UFak eta haien ordainak bildu izana –euskaraz eta gaztelaniaz– eta eskuragarri jarri izana, Hizkuntzaren Prozesamendurako aplikagarriak diren datu linguistikoekin batera. Helburu horrekin sortu dugu, hain zuzen, *Konbitzul* datu-basea[1]. Edonork kontsultak erraz egin ahal izan ditzan, bilatzai-

[1] *Konbitzul* datu-basean kontsultak egiteko: `http://ixa2.si.ehu.es/konbitzul/`

le gisako interfaze bat sortu diogu, zeinak aukera ematen baitu UFak hainbat irizpideren eta iragazkiren arabera egiteko. Horrez gain, datubaseko informazioa osorik deskargatzeko aukera ere eskaintzen dugu[2], aztertutako informazioa hizkuntza-tresnetan integratu nahi duenak ere eskura izan dezan. Guztira, gaztelaniazko 1.927 UF (euskarazko 4.043 ordainekin) eta euskarazko 2.074 (gaztelaniazko 3.022 ordainekin) jaso ditugu, eta, horietatik, gaztelaniazko 894 UFk eta haien ordain banak patroi morfosintaktikoa dute zehaztuta, identifikazio-lanerako eta itzulpen automatikorako erabilgarri.

[7] Gaztelaniazko eta –bereziki– euskarazko UFen corpus etiketatuak sortu izana. PARSEME proiektuaren baitan, hogei hizkuntzatako corpusak etiketatu dira fraseologia mailan, gidalerro berberei jarraituz. Bi fasetan egin da lan hori, eta bietan parte hartu dugu guk, bigarrenean bereziki: lehenengoan, gaztelaniazko corpusaren etiketatzaile-lantaldean parte hartu dugu, eta bigarrenean, gidalerroak hobetzeko zenbat proposamen egiteaz gain, euskarazko corpus etiketatua ere sortu dugu. Bata zein bestea eskuragarri daude sarean, gainerako hizkuntza guztiekin batera[3]. Bestalde, beste etiketatze-lan bat ere egin dugu euskarazko corpusaren gainean, agerpen literalena, eta hori ere eskuragarri jarri dugu[4].

8.3 Etorkizuneko lanak

Hainbat alderditatik zabaldu daiteke tesi honetan eginiko lana. Hona hemen ildo horietako batzuk:

- ***Konbitzul* datu-basea zabaltzen jarraitzea**. Datu-basea elikatzen joatea da gure asmoa, bai UF eta ordain berriekin, eta bai datubaseko UFen eta ordainen inguruko informazio gehiagorekin. Bereziki, euskaratik gaztelaniarako UF-itzulpenak aztertu nahi genituzke, orain arte eginikoa kontrako zentzurako egin baitugu ia osorik. Horrez gain, aurrera begira, interesgarria litzateke beste osaera batzueta-

[2]Datu-basea osorik deskargatzeko: `http://ixa.eus/node/4484`

[3]PARSEMEren corpus etiketatua: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842`

[4]Agerpen literalen corpusa: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2966`

ko UFak ere kontuan hartzea, hala nola, aditza+adjektiboa motakoak, aditza+adberbioa motakoak eta izena+adjektiboa motakoak. Eta, azkenik, 3. ekarpenean aipatu dugunaren haritik, landutako UFen katalanerako itzulpenak ere datu-basean sar litezke.

- **UFen inguruko informazioa itzultzaile automatiko estatistiko eta neuronaletan integratzea**. Orain arte, landu dugun informazioa erregeletan oinarritutako itzultzaile batean bakarrik erabili dugu. Hala ere, uste dugu egin dugun azterketaren zati handi bat erabilgarria dela beste itzultzaile mota batzuetan ere, eta saiakera egin nahi dugu ikusteko zer eragin duen UFen informazio gehigarriak halakoetan.

- **Anbiguotasun semantikoa ebazteko bideak ikertzea**. Tesi honetan, HPko tresnetarako erabilgarrien iruditu zaizkigun ezaugarriei begiratu diegu, eta morfosintaxian jarri dugu arreta batez ere. Hortaz, alde batera utzi ditugu morfosintaxiaren bidez ebatzi ezin diren UF anbiguoak, baina bide hori ikertzea ere beharrezkoa dela uste dugu, eredu distribuzionalak eta antzeko teknikak erabiliz.

- **Corpus orokorretatik espezializatuetarako jauzia egitea**. Fraseologia oso aldakorra da testu mota batetik bestera, eta corpus orokorretan egin duguna corpus espezializatuetara eraman nahi genuke. Izan ere, corpus mota batean eta besteetan erabiltzen diren UFak oso desberdinak izan ohi dira, eta azterketa horrek aukera emango liguke bateko eta besteko UFen ezaugarriak alderatzeko. Horrez gain, fraseologia espezializatua ia ikertu gabeko arloa da euskaraz, eta lan horren bidez urratsak egin litezke, besteak beste, testu espezializatuetako UFak eta ordainak biltzeko.

- **UFen itzulpen-portaerak sakonago aztertzea**. UFen sailkapen lexiko-semantikoa proposatu dugunean, haien itzulpen-portaeren inguruko hipotesiak egin ditugu, baina ez dugu ikerketa-ildo horretatik jarraitu, beste lan batzuei lehentasuna eman diegulako. Etorkizunean, corpus paraleloak oinarritzat hartu, eta aztertzen jarraitu nahi genuke ea UF mota bakoitzak joera duen benetan itzulpen-portaera jakin batzuk izateko (adibidez, ea kolokazioetan izenak ohiko ordaina jasotzen duen eta aditzarena izaten den irregularra, ea lokuzio opakoek itzulpen erabat irregularra izaten duten, etab.).

Bibliografia

Abel A. Towards a systematic classification framework for dictionaries and call. *Proceedings of ELex 2009, eLexicography in the 21st Century: New Challenges, New Applications*, 15–18. Louvain-la-Neuve, Belgika, 2010.

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., eta Insausti J. Edbl: a multi-purposed lexical support for the treatment of basque. *Proceedings of LREC 1998, the 1st International Conference on Language Resources and Evaluation*, 2 lib., 821–826. Granada, Espainia, 1998.

Aierbe A. La traducción a la lengua vasca de las unidades fraseológicas especializadas del lenguaje administrativo. *A Multilingual Focus on Contrastive Phraseology and Techniques for Translation.*, 27–44, 2008.

Al Saied H., Candito M., eta Matthieu C. The ATILF-LLF System for Parseme Shared Tak: a Transition-based Verbal Multiword Expressions Tagger. *Proceedings of MWE 2017, the 13th Workshop on Multiword Expressions (at EACL 2017)*, 127–132. Valentzia, Espainia, 2017.

Albrecht J. Invarianz, äquivalenz, adäquatheit. In Von Arntz H. eta Thone G., editors, *Übersetzungswissenschaft: Ergebnisse und Perspektiven*, 71–81. Tübingen, 1990.

Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., eta Urizar R. Representation and treatment of multiword expressions in basque. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (at ACL 2004)*, 48–55. Bartzelona, 2004.

Alegria I., Artola X., Sarasola K., eta Urkia M. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203, 1996.

Alonso E. Linguee y las nuevas formas de traducir. *Skopos. Revista internacional de Traducción e Interpretación*, 2:5–28, 2013.

Alonso Ramos M. *Las construcciones con verbo de apoyo*. Visor Libros, 2004.

Alonso Ramos M. Learning resources for spanish collocations: From a dictionary towards a writing assistant. In Sanromán Vilas B., editor, *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, 65–95. Société Néophilologique de Helsinki, 2016.

Alonso Ramos M. Diccionarios combinatorios. *Estudios de lingüística del español*, 38:173–201, 2017.

Alonso Ramos M., Nishikawa A., eta Vincze O. DiCE in the web: An online Spanish collocation dictionary. *Proceedings of eLex 2009, eLexicography in the 21st Century: New Challenges, New Applications*, 369–374. Louvain-la-Neuve, Belgika, 2010.

Altuna Díaz B. *Euskarazko denbora-egituren azterketa eta corpusaren sorrera*. Doktoretza-tesia, UPV/EHU, 2018.

Altzibar X., García J., eta Alberdi X. Calcos fraseológicos en el euskera de los medios de comunicación. In Luque L., Pamies A., eta Pazos J.M., editors, *Multi-Lingual Phraseography: Second Language Learning and Translation Applications*, 215–224. Baltmannsweiler, Schneider, 2011.

Anastasiou D. Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. *Proceedings of EAMT 2008, the 12th Conference of the European Association of Machine Translation*, 12–20. Hanburgo, Alemania, 2008.

Aranberri N. eta Labaka G. Euskarazko itzulpen automatikoa. *Senez*, 48:18, 2017.

Aranzabe M.J., Atutxa A., Bengoetxea K., de Ilarraza A.D., Goenaga I., Gojenola K., eta Uria L. Dependentzia unibertsalen eredura egokitutako euskarazko zuhaitz-bankua. *EKAIA, EHUko Zientzia eta Teknologia aldizkaria*, 2019.

Aranzabe Urruzola M.J. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Doktoretza-tesia, UPV/EHU, 2008.

Arnold I.V. *The English Word*. Vysšaja Škola, 2nd edition, 1986.

Aronoff M. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge*, (1):1–134, 1976.

Arrarats I. *Berriaren estilo-liburua*. Berria, 2006. URL `https://www.berria.eus/estiloliburua/`.

Azkarate M. *Hitz elkartuak euskaraz*. Doktoretza-tesia, Deustuko Unibertsitatea, 1987.

Azkue R.M. *Euskalerriaren yakintza: Literatura popular del País Vasco. III. liburukia*. Euskaltzaindia eta Espasa-Calpe, Bilbo/Madril, 1989.

Babych B. eta Hartley A. Automated error analysis for multiword expressions: using bleu-type scores for automatic discovery of potential translation errors. In Daelemans W. eta Hoste V., editors, *Evaluation of Translation Technology*, 8 lib., 81–104. Department Vertalers & Tolken Artesis Hogeschool Antwerpen, 2010.

Baker M. *In other words: A coursebook on translation*. Routledge, London, 1992.

Baldwin T., Bannard C., Tanaka T., eta Widdows D. An empirical model of multiword expression decomposability. *Proceedings of the ACL workshop on Multiword Expressions: analysis, acquisition and treatment (at ACL 2003)*, 89–96. Sapporo, Japonia, 2003.

Baldwin T. eta Kim S.N. Multiword expressions. *Handbook of Natural Language Processing*, 2:267–292, 2010.

Bally C. *Traité de stylistique française*, 1 lib. Librairie Klincksieck, 1909.

Bannard C. Learning about the meaning of verb–particle constructions from corpora. *Computer Speech & Language*, 19(4):467–478, 2005.

Bar-Hillel Y. Idioms. *Language and information: selected essays on their theory and application*, 47–55. Addison-Wisley Publishing Company, 1955.

Barreiro A. *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation*. Doktoretza-tesia, Universidade do Porto, 2008.

Barreiro A., Monti J., Orliac B., Preuß S., Arrieta K., Ling W., Batista F., eta Trancoso I. Linguistic evaluation of support verb constructions by openlogos and google translate. *Proceedings of LREC 2014, the 9th Language Resources and Evaluation Conference*, 35–40. Reykjavik, Islandia, 2014.

Benson M., Benson E., eta Ilson R. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins Publishing Company, Amsterdam, 1986.

Berk G., Erden B., eta Güngör T. Deep-BGT at parseme shared task 2018: Bidirectional lstm-crf model for verbal multiword expression identification. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 248–253. Santa Fe, AEB, 2018.

Bevilacqua C. Unidades fraseológicas especializadas (UFE): elementos para su identificación y descripción. *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*, 113–141. Universitat Pompeu Fabra, Bartzelona, 2001.

Blunsom P. eta Baldwin T. Multilingual deep lexical acquisition for HPSGs via supertagging. *Proceedings of the 2006 conference on Empirical Methods in Natural Language Processing*, 164–171. Sydney, Australia, 2006.

Bond F., Korhonen A., McCarthy D., eta Villavicencio A. *Proceedings of the ACL Workshop on Multiword Expressions (at ACL 2003)*. Sapporo, Japonia, 2003.

Boros T. eta Burtica R. GBD-NER at PARSEME Shared Task 2018: Multiword Expression Detection Using Bidirectional Long-Short-Term Memory Networks and Graph-Based Decoding. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 254–260. Santa Fe, AEB, 2018.

Boros T., Pipa S., Barbu V., eta Dan Tufis M. A data-driven approach to Verbal Multiword Expressions detection. PARSEME Shared Task system description paper. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 121–126. Valentzia, Espainia, 2017.

Bosque I. On the weight of light predicates. In Herschensohn J., Mallén E., eta Zagona K., editors, *Features and Interfaces in Romance: Essays in honor of Heles Contreras*, 23–38. John Benjamins Publishing Company, 2001.

Bosque I. *Redes: diccionario combinatorio del español contemporáneo*. Ediciones SM, 2004.

Bosque I. *Diccionario combinatorio práctico del español contemporáneo: las palabras en su contexto*. Ediciones SM, 2006.

Bouamor D., Semmar N., eta Zweigenbaum P. Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (at COLING 2012)*, 95–108. Mumbai, India, 2012.

Boukobza R. eta Rappoport A. Multi-word expression identification using sentence surface features. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: volume 2*, 468–477. Singapur, Singapur, 2009.

Buckingham L. *Las construcciones con verbo soporte en un corpus de especialidad*. Peter Lang, 2009.

Buljan M. eta Šnajder J. Combining linguistic features for the detection of croatian multiword expressions. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 194–199. Valentzia, Espainia, 2017.

Burnard L. Reference Guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services, 2007.

Bustos Plaza A. *Combinaciones verbonominales y lexicalización*. Peter Lang, 2005.

Butt M. The light verb jungle: still hacking away. *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge University Press, 2010.

Buyse K. eta Verlinde S. Possible effects of free on line data driven lexicographic instruments on foreign language learning: The case of linguee and the interactive language toolbox. *Procedia: Social and Behavioral Sciences*, 95:507–512, 2013.

Cabré M.T. Hacia una teoría comunicativa de la terminología: aspectos metodológicos. *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos.*, 129–150. Universitat Pompeu Fabra, 1999.

Cap F., Nirmal M., Weller M., eta Im Walde S.S. How to account for idiomatic german support verb constructions in statistical machine translation. *Proceedings of the 11th Workshop on Multiword Expressions (at ACL 2015)*, 19–28. Denver, AEB, 2015.

Carpuat M. eta Diab M. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 242–245. Los Angeles, AEB, 2010.

Chiang D. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., eta Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar, 2014.

Choueka Y. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *Proceedings of RIAO 88: Recherche d'Information Assistée par Ordinateur*, 609–623. Cambridge, AEB, 1988.

Church K.W. eta Hanks P. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Cobeta Melchor M.d.M. Paremiología y traducción. *Actas de las II^{a} Jornadas de Jóvenes Traductores: diciembre 1998*, 107–118. Universidad de las Palmas de Gran Canaria, 2002.

Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Constant M., Eryiğit G., Monti J., Van Der Plas L., Ramisch C., Rosner M., eta Todirascu A. Multiword Expression processing: a survey. *Computational Linguistics*, 43(4):837–892, 2017.

Constant M. eta Sigogne A. MWU-aware part-of-speech tagging with a CRF model and lexical resources. *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world*, 49–56. Portland, AEB, 2011.

Cook P., Fazly A., eta Stevenson S. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. *Proceedings of the workshop on a broader perspective on Multiword Expressions*, 41–48. Association for Computational Linguistics, 2007.

Copestake A., Lambeau F., Villavicencio A., Bond F., Baldwin T., Sag I., eta Flickinger D. Multiword expressions: Linguistic precision and reusability. *Proceedings of the Language Resources and Evaluation Conference*, 1941–1947. Kanariar Irlak, Espainia, 2002.

Cordeiro S., Ramisch C., Idiart M., eta Villavicencio A. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1 lib., 1986–1997. Berlin, Alemania, 2016.

Corpas G. Translating english verbal collocations into spanish: On distribution and other relevant differences related to diatopic variation. *Lingvisticæ Investigationes*, 38(2):229–262, 2015.

Corpas Pastor G. *Manual de fraseología española*. Editorial Gredos, 1996.

Corpas Pastor G. En torno al concepto de colocación. *Euskera*, 1(XLVI), 2001.

Corpas Pastor G. *Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos*. Iberoamericana, 2003.

Coseriu E. *Falsche und richtige Fragestellungen in der Ubersetzungstheorie*. Peter Lang, 1978.

Cowie A.P. Stable and creative aspects of vocabulary. *Vocabulary and language teaching*, 126–139. Longman, 1988.

Cowie A.P. Collocational dictionaries - a comparative view. *Proceedings of the Fourth Joint Anglo-Soviet Seminar on English Studies*, 61–69. The British Council, 1986.

Cowie A.P. *Phraseology: theory, analysis, and applications*. Oxford University Publishing, 1998.

Dobrovol'skij D. Cross-linguistic equivalence of idioms: does it really exist. In Pamies A. eta Dobrovol'skij D., editors, *Linguo-cultural competence and phraseological motivation*, 7–24. Schneider Verlag Hohengehren, 2011.

Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research*, 138–145. Morgan Kaufmann Publishers Inc., 2002.

Dunning T. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

Ehren R., Lichte T., eta Samih Y. Mumpitz at PARSEME Shared Task 2018: A Bidirectional LSTM for the Identification of Verbal Multiword Expressions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 261–267. Santa Fe, AEB, 2018.

Esnal P. Ortik eta emendik: euskal lokuzioak eta fraseologia baino ere haratago. *Euskera: Euskaltzaindiaren lan eta agiriak*, 46(1):137–144, 2001.

Estarrona Ibarloza A. *EPEC corpusa predikatu-mailan etiketatzeko oinarriak: EPEC-RolSem, BVI eta e-ROLda*. Doktoretza-tesia, UPV/EHU, 2014.

Etchegoyhen T., Martinez Garcia E., Azpeitia A., Labaka G., Alegria I., Cortes Etxabe I., Jauregi Carrera A., Ellakuria Santos I., Martin M., eta Calonge E. Neural machine translation of basque. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, 139–148. Alicante, Espainia, 2018.

Etxepare R. Valency and argument structure in the basque verb. In Hualde J.I. eta Ortiz de Urbina J., editors, *A grammar of Basque*, 282–323. De Gruyter, 2003.

Euskaltzaindia. *Euskaltzaindiaren Hiztegia*. Elkar, 2012.

Evert S. *The statistics of word cooccurrences: word pairs and collocations*. Doktoretza-tesia, University of Stuttgart, 2005.

Evert S. Corpora and collocations. *Corpus linguistics. An international handbook*, 2 lib., 1212–1248. De Gruyter, 2009.

Ezeiza N. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua*. Doktoretza-tesia, Informatika Fakultatea, UPV/EHU, 2002.

Farahmand M. eta Henderson J. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. *Proceedings of the 12th workshop on Multiword Expressions*, 61–66. Berlin, Alemania, 2016.

Farø K. Dogmatismus, skeptizismus, nihilismus und pragmatismus bei der idiomübersetzung: Grundfragen zu einer idiomtranslatorischen theorie. In Häcki A. eta Burger H., editors, *Phraseology in Motion I. Methoden und Kritik*, 189–202. Baltmannsweiler, Schneider, 2006.

Fazly A. eta Stevenson S. Automatically constructing a lexicon of verb phrase idiomatic combinations. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 337–344. Trento, Italia, 2006.

Ferreira Da Silva J., Dias G., Guilloré S., eta Pereira Lopes J.G. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Progress in Artificial Intelligence: 9th Portuguese Conference on Artificial Intelligence*, 113–132. Springer, 1999.

Firth J.R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1–32. Basil Blackwell, 1957.

Forcada M.L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J.A., Sánchez-Martínez F., Ramírez-Sánchez G., eta Tyers

F.M. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.

Foufi V., Nerima L., eta Wehrli E. Parsing and MWE detection: Fips at the PARSEME shared task. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valentzia, Espainia, 2017.

Gallego Á.J. Predicados ligeros y valoración de rasgos. *Dicenda*, 28:27–55, 2010.

Gao Q. eta Vogel S. Parallel implementations of word alignment tool. *Software engineering, testing, and quality assurance for Natural Language Processing*, 49–57. Columbus, AEB, 2008.

Garate G. *Atsotitzak*. Bilbao Bizkaia Kutxa Fundazioa, Bilbo, 2003.

Garcia M., García Salido M., eta Alonso Ramos M. A comparison of statistical association measures for identifying dependency-based collocations in various languages. *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 49–59, 2019.

García García M. Construcciones con verbo soporte y otras construcciones afines. *Estudios filológicos alemanes: revista del Grupo de Investigación Filología Alemana*, 7:75–96, 2005.

García Salido M. eta Alonso Ramos M. Asignación de niveles de aprendizaje a las colocaciones del diccionario de colocaciones del español. *Revista Signos*, 51(97):153–174, 2018.

Garzia J. *Joskera lantegi*. 1997.

Garzia J. *Kalko okerrak*. 2005.

Ghoneim M. eta Diab M. Multiword expressions in the context of statistical machine translation. *Proceedings of the sixth International Joint Conference on Natural Language Processing*, 1181–1187. Nagoya, Japonia, 2013.

Gilsou P. *Errantegia*. Argitaratu gabea. Parisen 1964ko ekainean burutua, 1964.

Gläser R. *Phraseologie der englischen Sprache*. Max Niemeyer, 1986.

Granger S. eta Paquot M. Disentangling the phraseological web. *Phraseology: an interdisciplinary perspective*, 28:27–49, 2008.

Granger S. eta Paquot M. Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1):118–141, 2015.

Greimas A.J. Idiotismes, proverbes, dictons. *Cahiers de lexicologie*, 2:41–61, 1960.

Gurrutxaga A. *Idiomatikotasunaren katakterizazio automatikoa: izena+aditza konbinazioak*. Doktoretza-tesia, UPV/EHU, 2014.

Gurrutxaga A. eta Alegria I. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. *Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world (at ACL 2011)*, 2–7. Portland, AEB, 2011.

Gurrutxaga A., Alegria I., eta Artola X. Idiomatikotasunaren karakterizazio automatikoa: izena+ aditza. *Ekaia: Euskal Herriko Unibertsitateko zientzia eta teknologia aldizkaria*, 29:47–68, 2016.

Gutiérrez-Rodríguez E. Rasgos categoriales de los determinantes. *Actas del XXXVII Simposio Internacional de la Sociedad EspaÃśola de Lingüística (SEL)*. Iruñea, 2008.

Gwet K.L. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters*. Advanced Analytics, LLC, 2012.

Haensch G. La lengua española y la lexicografía actual. *LEA: Lingüística española actual*, 4(2):239–252, 1982.

Hashimoto C. eta Kawahara D. Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features. *Proceedings of the conference on empirical methods in natural language processing*, 992–1001. Honolulu, Hawaii, 2008.

Hashimoto C., Sato S., eta Utsuro T. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. *Proceedings of the COLING/ACL on Main conference poster sessions*, 353–360. Sydney, Australia, 2006.

Heine A. *Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-) Lexikographie*. Peter Lang, 2006.

Heylen D., Maxwell K.G., eta Verhagen M. Lexical functions and machine translation. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japonia, 1994.

Howarth P.A. *Phraseology in English academic writing: some implications for language learning and dictionary making*. Walter de Gruyter, 1996.

Hualde J.I., Oyharçabal B., eta Ortiz de Urbina J. Verbs. *A grammar of Basque*, 155–198. De Gruyter, 2003.

Izagirre K. *Euskal lokuzioak. Espainolezko eta frantsesezko gida-zerrendarekin*. Hordago, Donostia, 1981.

Jackendoff R. *The architecture of the language faculty*. MIT Press, 1997.

Jakobson R. On linguistic aspects of translation. *On translation*, 3:30–39, 1959.

Kalchbrenner N. eta Blunsom P. Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. Seattle, AEB, 2013.

Karlsson F., Voutilainen A., Heikkilae J., eta Anttila A. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Walter de Gruyter, 1995.

Katz G. eta Giesbrecht E. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12–19. Sydney, Australia, 2006.

Kearns K. *Light verbs in English*. MIT Press, 1988.

Kennedy G. *An introduction to corpus linguistics*. Longman, 1998.

Klyueva N., Vernerová A., eta QasemiZadeh B. Querying Multiword Expressions Annotation with CQL. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 73–79. Praga, Txekiar Errepublika, 2017.

Koehn P. *Statistical Machine Translation*. Cambridge University Press, 2010.

Koehn P. eta Hoang H. Factored translation models. *Proceedings of the 2007 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Praga, Txekiar Errepublika, 2007.

Koehn P., Och F.J., eta Marcu D. Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48–54. Edmonton, Kanada, 2003.

Kordoni V. eta Simova I. Multiword Expressions in Machine Translation. *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*, 1208–1211. Reykjavik, Islandia, 2014.

Krippendorff K. *Content analysis: an introduction to its methodology*. Sage Publications, 1980.

Labaka G. *EUSMT: incorporating linguistic information to SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation*. Doktoretza-tesia, UPV/EHU, 2010.

Labaka G., España-Bonet C., Màrquez L., eta Sarasola K. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125, 2014.

Laka I. *A brief grammar of Euskara, the Basque language*. University of the Basque Country, UPV/EHU, 1996.

Lakarra J.A. *Refranes y Sentencias (1596): Ikerketak eta edizioa*. Euskaltzaindia, 1996.

Lambert P. eta Banchs R. Data inferred multiword expressions for statistical machine translation. *Proceedings of Machine Translation Summit X*, 396–403. Phuket, Thailandia, 2005.

Landis J.R. eta Koch G.G. The measurement of observer agreement for categorical data. *Biometrics*, 159–174, 1977.

Lapata M. eta Lascarides A. Detecting novel compounds: The role of distributional evidence. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 235–242. Budapest, Hungaria, 2003.

Leturia I. Evaluating different methods for automatically collecting large general corpora for basque from the web. *Proceedings of COLING 2012, the International Conference on Computational Linguistics*, 1553–1570, 2012.

Lipka L., Handl S., eta Falkner W. Lexicalization and institutionalization. *The encyclopedia of language and linguistics*, 4:2164–2167, 2004.

Losnegaard G.S., Sangati F., Escartín C.P., Savary A., Bargmann S., eta Monti J. PARSEME survey on MWE resources. *9th International Conference on Language Resources and Evaluation*, 2299–2306. Portoroz, Eslovenia, 2016.

Lyons J. *Introduction to theoretical linguistics*. Cambridge University Press, 1968.

Maldonado A., Han L., Moreau E., Alsulaimani A., Chowdhury K., Vogel C., eta Liu Q. Detection of verbal multiword expressions via conditional random fields with syntactic dependency features and semantic re-ranking. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 114–120. Valentzia, Espainia, 2017.

Maniez F. Extraction d'une phraséologie bilingue en langue de spécialité: corpus parallèles et corpus comparables. *Meta: Journal des traducteurs/-Meta: Translators' Journal*, 46(3):552–563, 2001.

Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., eta McClosky D. The Stanford CoreNLP Natural Language Processing toolkit. *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations*, 55–60. Baltimore, AEB, 2014.

Markantonatou S., Kouris P., eta Maistros Y. Fixed Similes: measuring aspects of the relation between MWE idiomatic semantics and syntactic flexibility. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 51–61. Santa Fe, AEB, 2018.

Martinez A. *[Izen+egin] aditz-lokuzioak: inkorporazio-mailak.* Doktoretzatesia, UPV/EHU, 2015.

Martínez Linares M.A. *Palabra y lexía.* Liceus, Servicios de Gestión, 2006.

Mayor A., Alegria I., De Ilarraza A.D., Labaka G., Lersundi M., eta Sarasola K. Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, 25(1):53–82, 2011.

Mayor A., Alegria I., Díaz de Ilarraza A.D., Labaka G., Lersundi M., eta Sarasola K. Matxin, euskararako lehenengo itzultzaile automatikoa. *Senez*, (37):197–220, 2009.

McCarthy D., Keller B., eta Carroll J. Detecting a continuum of compositionality in phrasal verbs. *Proceedings of the ACL 2003 workshop on Multiword Expressions: analysis, acquisition and treatment*, 73–80. Sapporo, Japonia, 2003.

McCarthy D., Venkatapathy S., eta Joshi A. Detecting compositionality of verb-object combinations using selectional preferences. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 369–379. Praga, Txekiar Errepublika, 2007.

McIntosh C. *Oxford collocations dictionary for student of English.* Oxford, 2009.

Mel'čuk I. Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102, 1996.

Mel'čuk I. Collocations and lexical functions. *Phraseology. Theory, analysis, and applications*, 23–53. Clarendon Press, Oxford, 1998.

Mel'čuk I. eta Polguere A. A formal lexicon in the meaning-text theory:(or how to do lexica with words). *Computational linguistics*, 13(3-4):261–275, 1987.

Mellado Blanco C. Formas estereotipadas de realización no verbal en alemán y español: los cinegramas desde un enfoque contrastivo-histórico. In Corpas Pastor G., editor, *Las lenguas de Europa: estudios de fraseología, fraseografía y traducción*, 389–410. Editorial Comares, 2000.

Michelena L. *Orotariko euskal hiztegia*. Euskaltzaindia, 1987.

Mikolov T., Sutskever I., Chen K., Corrado G.S., eta Dean J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119. MIT Press, 2013.

Mokoroa Mugica J.M. *Ortik eta emendik: Repertorio de locuciones del habla popular vasca, oral y escrita, en sus diversas variedades*. Labayru eta Eusko Jaurlaritza, Bilbo, 1990.

Molero A. *El español de España y el español de América: vocabulario comparado*. Ediciones SM, Madrid, 2003.

Monteiro Plantin R. Foreword. In Pamies A., Luque-Nadal L., eta Pazos-Bretaña J.M., editors, *Multi-lingual phraseography: second language learning and translation applications*, page 1. Schneider-Verlag Hohengehren, 2011.

Monti J., Barreiro A., Elia A., Marano F., eta Napoli A. Taking on new challenges in multi-word unit processing for machine translation. *Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, 11–19. Bartzelona, 2011.

Monti J., Barreiro A., Oroliac B., eta Batista F. When Multiwords go bad in Machine Translation. *Machine Translation Summit XIV*, 26–33. Niza, Frantzia, 2013.

Monti J., Elia A., Postiglione A., Monteleone M., eta Marano F. In search of knowledge: text mining dedicated to technical translation. *Proceedings of ASLIB 2011 Translating and the Computer Conference*. Londres, Erresuma Batua, 2012.

Moreau E., Alsulaimani A., Maldonado A., eta Vogel C. Crf-seq and crf-deptree at parseme shared task 2018: Detecting verbal mwes using sequential and dependency-based approaches. *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) at the 27th International Conference on Computational Linguistics (COLING 2018)*, 241–247. Santa Fe, AEB, 2018.

Morvay K. Harri batez bi kolpe. cuestiones de fraseología comparada. *Euskera, XLI*, 3:719–767, 1996.

N Von Vilen K. Linguistische modelle des übersetzungprozesses. *Übersetzungswissenschaft*, 535:171, 1981.

Na H., Li J.J., Lee Y., eta Lee J.H. A synchronous context-free grammar using dependency sequence for syntax-base statistical machine translation. *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*. Denver, AEB, 2010.

Nunberg G., Sag I.A., eta Wasow T. Idioms. *Language*, 70(3):491–538, 1994.

Oepen S., Dyvik H., Lønning J.T., Velldal E., Beermann D., Carroll J., Flickinger D., Hellan L., Johannessen J.B., eta Meurer P. Som å kappete med trollet?-towards mrs-based norwegian-english machine translation. *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Alicante, Espainia, 2004.

Oflazer K., Say B., *et al.*. Integrating morphology with multi-word expression processing in turkish. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 64–71. Bartzelona, 2004.

Ortiz de Urbina J. Periphrastic constructions. In Hualde J.I. eta Ortiz de Urbina J., editors, *A grammar of Basque*, 223–234. De Gruyter, 2003a.

Ortiz de Urbina J. Semiauxiliary verbs. In Hualde J.I. eta Ortiz de Urbina J., editors, *A grammar of Basque*, 235–346. De Gruyter, 2003b.

Oyharçabal B. Basque light verb constructions, 2003.

Padró L. eta Stanilovsky E. Freeling 3.0: towards wider multilinguality. *Proceedings of the Language Resources and Evaluation Conference, LREC2012*, 2473–2479. Istanbul, Turkia, 2012.

Pal S., Naskar S., eta Bandyopadhyay S. A hybrid word alignment model for phrase-based statistical machine translation. *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 94–101. Sofia, Bulgaria, 2013.

Papineni K., Roukos S., Ward T., eta Zhu W.J. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, AEB, 2002.

Parra Escartín C., Nevado Llopis A., eta Sánchez Martínez E. Spanish multiword expressions: Looking for a taxonomy. In Sailer M. eta Markantonatou S., editors, *Multiword expressions, insights from a multilingual perspective*, 271–323. Language Science Press, 2018.

Pasquer C., Ramisch C., Savary A., eta Antoine J.Y. VarIDE at PARSEME Shared Task 2018: Are Variants Really as Alike as Two Peas in a Pod? *Proceedings of the COLING Workshop on Linguistic Annotation, Multiword Expressions and Constructions*. Santa Fe, AEB, 2018.

Pearce D. Synonymy in collocation extraction. *Proceedings of the workshop on WordNet and other lexical resources, second meeting of the North American chapter of the Association for Computational Linguistics*, 41–46. Pittsburgh, AEB, 2001.

Pearce D. A comparative evaluation of collocation extraction techniques. *Language Resources and Evaluation Conference*, 1530–1536. Kanariar Irlak, Espainia, 2002.

Pecina P. A machine learning approach to multiword expression extraction. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, 54–61. Marrakech, Maroko, 2008.

Pedersen T. Fishing for exactness. 188–200. Austin, AEB, 1996.

Piera C. eta Varela S. Relaciones entre morfología y sintaxis. *Gramática descriptiva de la lengua española*, 3:4367–4422, 1999.

RAE. *Nueva gramática de la lengua española*. Espasa, 2009.

Rafel J. Los predicados complejos en español. In Zabala I., Pérez Gaztelu E., eta Gràcia Sole L., editors, *Las fronteras de la composición en lenguas románicas y en vasco*, 393–443. Deustuko Unibertsitatea, 2004.

Ramisch C. *Multiword expressions acquisition: a generic and open framework*. Springer, 2015.

Ramisch C., Besacier L., eta Kobzar A. How hard is it to automatically translate phrasal verbs from english to french? *MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*, 53–61. Niza, Frantzia, 2013.

Ramisch C., Cordeiro S.R., Savary A., Vincze V., Mititelu V.B., Bhatia A., Buljan M., Candito M., Gantar P., Giouli V., Güngür T., Hawwari A., Iñurrieta U., Kovalevskaite J., Krek S., Lichte T., Liebskind C., Monti J., Parra C., QasemiZadeh B., Ramisch R., Schneider N., Stoyanova I., Vaidya A., eta Walsh A. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, 222–240. Santa Fe, AEB, 2018.

Ramisch C., De Araujo V., eta Villavicencio A. A broad evaluation of techniques for automatic acquisition of Multiword Expressions. *Proceedings of ACL 2012 Student Research Workshop*, 1–6. Jejudo, Hego Korea, 2012.

Ramisch C., Villavicencio A., eta Boitet C. MWEtoolkit: A framework for multiword expression identification. *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, 662–669. Valletta, Malta, 2010.

Ramisch C., Villavicencio A., Moura L., eta Idiart M. Picking them up and figuring them out: verb-particle constructions, noise and idiomaticity. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 49–56. Manchester, Erresuma Batua, 2008.

Reddy S., McCarthy D., eta Manandhar S. An empirical study on compositionality in compound nouns. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 210–218. Chiang Mai, Thailandia, 2011.

Ren Z., Lü Y., Cao J., Liu Q., eta Huang Y. Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the ACL Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, 47–54. Singapur, 2009.

Richart Marset M. Las unidades fraseológicas y su resistencia a la traducción. *Foro de profesores de E/LE*, 4:1–10, 2008.

Richter F. eta Sailer M. Cranberry words in formal grammar. *Empirical issues in formal syntax and semantics*, 4:155–171, 2003.

Riedl M. eta Biemann C. A single word is not enough: Ranking multiword expressions using distributional semantics. *Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing*, 2430–2440. Lisboa, Portugal, 2015.

Riedl M. eta Biemann C. Impact of MWE resources on multiword recognition. *Proceedings of the 12th Workshop on Multiword Expressions*, 107–111. Berlin, Alemania, 2016.

Rodríguez S. eta García Murga F. Izen+egin predikatuak euskaraz. *Euskal gramatikari eta literaturari buruzko ikerketak XX1. mendearen atarian, Gramatika gaiak, Iker-14*, 417–436, 2003.

Rondon A., Caseli H., eta Ramisch C. Never-ending Multiword Expressions learning. *Proceedings of the 11th Workshop on Multiword Expressions*, 45–53. Denver, AEB, 2015.

Ruiz Costa-Jussà M., Daudaravicius V., eta Banchs R.E. Integration of statistical collocation segmentations in a phrase-based statistical machine translation system. *EAMT 2010: proceedings of the 14th annual conference of the European Association for Machine Translation*. Saint-Raphael, Frantzia, 2010.

Sag I.A., Baldwin T., Bond F., Copestake A., eta Flickinger D. Multiword expressions: a pain in the neck for nlp. *International Conference on Intelligent Text Processing and Computational Linguistics*, 1–15. Springer, 2002.

Sanroman Vilas M.B. *¿Es posible definir un verbo ligero?* Lingua Americana, 2017.

Sanz Villar Z. Alemanetik euskaratutako unitate fraseologikoen itzulpen-azterketa: tesiaren nondik norakoak. *Senez*, 46:211–230, 2015a.

Sanz Villar Z. *Unitate Fraseologikoen itzulpena: alemana-euskara*. Doktoretza-tesia, Letren Fakultatea, UPV/EHU, 2015b.

Sarasola I. *Euskara batuaren ajeak*. Alberdania, 1997.

Sarasola I. *Zehazki: gaztelania-euskara hiztegia*. Alberdania eta UPV/EHU, 2005. URL `http://www.ehu.eus/ehg/zehazki/sarrera.htm`.

Savary A. Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, (1-2):1–53, 2008.

Savary A., Candito M., Barbu Mititelu V., Bejček E., Cap F., eta Gompel M.v. PARSEME multilingual corpus of Verbal Multiword Expressions. *Multiword Expressions at length and in-depth: extended papers from the MWE 2017 workshop*, 87–147. Language Science Press, 2018.

Savary A., Cordeiro S., Lichte T., Ramisch C., Inurrieta U., eta Giouli V. Literal occurrences of multiword expressions: rare birds that cause a stir. *Prague Bulletin of Mathematical Linguistics*, 112:1–44, 2019.

Savary A. eta Cordeiro S.R. Literal readings of multiword expressions: as scarce as hen's teeth. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 64–72. Praga, Txekiar Errepublika, 2018.

Savary A., Ramisch C., Cordeiro S., Sangati F., Vincze V., QasemiZadeh B., Candito M., Cap F., Giouli V., eta Stoyanova I. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. *Proceedings of the 13th Workshop on Multiword Expressions (at EACL 2017)*, 121–126. Valentzia, Espainia, 2017.

Savary A., Sailer M., Parmentier Y., Rosner M., Rosén V., Przepiórkowski A., Krstev C., Vincze V., Wójtowicz B., Losnegaard G.S., *et al.*. PARSEME–PARSing and Multiword Expressions within a European multilingual network. *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznan, Polonia, 2015.

Schneider N., Danchik E., Dyer C., eta Smith N.A. Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, 2014.

Schottmüller N. eta Nivre J. Issues in translating verb-particle constructions from German to English. *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*, 124–131. Gűteborg, Suedia, 2014.

Sennrich R. eta Haddow B. Linguistic input features improve neural machine translation. *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, 83–91. Berlin, Alemania, 2016.

Seretan V. *Syntax-based collocation extraction.* Springer Science Business Media, 2011.

Shigeto Y., Azuma A., Hisamoto S., Kondo S., Kouse T., Sakaguchi K., Yoshimoto A., Yung F., eta Matsumoto Y. Construction of English MWE dictionary and its application to POS tagging. *Proceedings of the 9th Workshop on Multiword Expressions*, 139–144. Atlanta, AEB, 2013.

Simkó K.I., Kovács V., eta Vincze V. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 48–53. Valentzia, Espainia, 2017.

Sinclair J. *Corpus, concordance, collocation.* Oxford University Press, 1991.

Snell-Hornby M. *Übersetzungswissenschaft. Eine Neuorientierung.* Tübingen, 1986.

Snover M., Dorr B., Schwartz R., Micciulla L., eta Makhoul J. A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 223–231. Cambridge, AEB, 2006.

Somers H. Example-based machine translation. *Machine Translation*, 14(2): 113–157, 1999.

Sporleder C. eta Li L. Unsupervised recognition of literal and non-literal use of idiomatic expressions. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 754–762. Atenas, Grezia, 2009.

Stevenson S., Fazly A., eta North R. Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 1–8. Bartzelona, 2004.

Stodden R., QasemiZadeh B., eta Kallmeyer L. Trapacc and trapacc-s at parseme shared task 2018: Neural transition tagging of verbal multiword

expressions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 268–274. Santa Fe, AEB, 2018.

Tan L. eta Pal S. Manawi: Using multi-word expressions and named entities to improve machine translation. *Proceedings of the 9th Workshop on Statistical Machine Translation (at ACL 2014)*, 201–206. Baltimore, AEB, 2014.

Timofeeva L. Sobre la traducción fraseológica. *Estudios de lingüística (ELUA)*, 26:405–432, 2012.

Tognini-Bonelli E. *Corpus linguistics at work*. John Benjamins Publishing Company, 2001.

Toury G. *Descriptive translation studies and beyond*. John Benjamins Publishing Company, 2012.

Trask R. The noun phrase: nouns, determiners and modifiers; pronouns and names. In Hualde J.I. eta Ortiz de Urbina J., editors, *A grammar of Basque*, 92–134. De Gruyter, 2003.

Tu Y. *English complex verb constructions: identification and inference*. Doktoretza-tesia, University of Illinois at Urbana-Champaign, 2012.

Uchiyama K., Baldwin T., eta Ishizaki S. Disambiguating japanese compound verbs. *Computer Speech & Language*, 19(4):497–512, 2005.

Ullman E. eta Nivre J. Paraphrasing swedish compound nouns in machine translation. *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 99–103. Gűteborg, Suedia, 2014.

Urizar R. *Euskal lokuzioen tratamendu konputazionala*. Doktoretza-tesia, Informatika Fakultatea, UPV/EHU, 2012.

Urquijo J. *El refranero vasco: I. Los refranes de Garibay*. Imprenta Martín, Mena y Compª, 2. edizioa, Bilbo, 1919.

Villavicencio A., Kordoni V., Zhang Y., Idiart M., eta Ramisch C. Validation and evaluation of automatically acquired multiword expressions for

grammar engineering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1034–1043. Praga, Txekiar Errepublika, 2007.

Vincze O. eta Alonso Ramos M. Incorporating frequency information in a collocation dictionary: establishing a methodology. *Procedia, Social and Behavioral Sciences*, 95:241–248, 2013.

Vincze O., Mosqueira E., eta Alonso Ramos M. An online collocation dictionary of Spanish. *Proceedings of the 5th International Conference on Meaning-Text Theory*, 275–286. Bartzelona, 2011.

Wanner L., Bohnet B., eta Giereth M. Making sense of collocations. *Computer Speech & Language*, 20(4):609–624, 2006.

Wanner L., Verlinde S., eta Alonso Ramos M. Writing assistants and automatic lexical error correction: word combinatorics. *Proceedings of eLex 2013. Electronic lexicography in the 21st century: Thinking outside the paper*, 17–19. Tallin, Estonia, 2013.

Waszczuk J. TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 275–282. Santa Fe, AEB, 2018.

Wehrli E. Traduction, traduction de mots, traduction de phrases. *Proceedings of TALN XI*, 483–491. Fes, Maroko, 2004.

Wehrli E., Seretan V., Nerima L., eta Russo L. Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy. *13th Annual Conference of the European Association for Machine Translation*, 128–135. Bartzelona, 2009.

Weller M. eta Heid U. Extraction of german multiword expressions from parsed corpora using context features. *Language Resources and Evaluation Conference*. Valleta, Malta, 2010.

Wierzbicka A. Why can you have a drink when you can't* have an eat? *Language*, 58(4):753–799, 1982.

Wotjak G. Reflexiones acerca de construcciones verbo-nominales. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, 1:3–32, 2018.

Yazdani M., Farahmand M., eta Henderson J. Learning semantic composition to detect non-compositionality of multiword expressions. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1733–1742. Lisboa, Praga, 2015.

Zabala I. Nominal predication: copulative sentences and secondary predication. 428–448. De Gruyter, 2003.

Zabala I. Los predicados complejos en vasco. In Zabala I., Pérez Gaztelu E., eta Gràcia Sole L., editors, *Las fronteras de la composición en lenguas románicas y en vasco*, 445–534. Deustuko Unibertsitatea, 2004.

Zamarripa P. *Manual del vascófilo*. Imp. y Enc. de Jose A. de Lerchundi, Bilbo, 1913.

Zuluaga A. Traductología y fraseología. *Paremia*, 8:537–549, 1999.

A. ERANSKINA

Datu multzoen argibideak

Tesi-txosten honetan datu multzo ugari agertzen denez, eranskin honetan datu horien inguruko argibideak jarri ditugu eskematikoki. Atalez atal bildu ditugu datuak, eta datu multzoak zein atali dagozkion ere zehaztu dugu.

Hiztegien gaineko azterketa

(3.2. atala)

- *Elhuyar* gaztelania-euskara hiztegitik:
  - Gaztelaniazko 2.343 aditza+izena konbinazio, hiztegi-sarrerak direnak edota euskarazko hiztegi-sarreren ordainetan agertzen direnak.
  - Euskarazko 6.587 ordain.
- *Elhuyar* euskara-gaztelania hiztegitik:
  - Euskarazko 2.954 izena+aditza konbinazio, hiztegi-sarrerak direnak edota euskarazko hiztegi-sarreren ordainetan agertzen direnak.
  - Gaztelaniazko 6.390 ordain.

Azterketa xehea

- Identifikazioari begira (4.1. atala):
  - Gaztelaniaz, hasiera batean, hiztegiko 150 konbinaziorik usuenak. Etiketatze-lan lexiko-semantikoa egin ondoren, UFtzat jotako 117ak.
  - Ingelesezko zatirako, hasieran, 173 konbinazio usu. Etiketatze-lan lexiko-semantikoa egin ondoren, UFtzat jotako 133ak.
- Itzulpen automatikoari begira (5.1. atala):
  - Identifikaziorako landutako 117 gaztelaniazko konbinazioak, hiztegiko ordain banarekin batera.
  - *DiCE* hiztegitik hartutako 22 konbinaziorik usuenak, eskuz emandako ordain banarekin batera.
  - Aurreko konbinazioetako 14 UF anbiguoren bikoizketak, hau da, zerrendako UFen osagai lexiko berberak izan arren ezaugarri morfosintaktiko desberdinak dituzten UFak.

Azterketa erdiautomatikoa

(4.4. eta 5.3. atalak)

- Oinarritzat hartu dira:
  - *Elhuyar*reko gaztelaniazko 1.205 hiztegi-sarrerak.
  - *DiCE*ko 500 kolokaziorik usuenei aurreko atalean aztertutako 22ak eta *Elhuyar*ren ere badaudenak kenduta, geratzen diren 437 kolokazioak.
- Azterketa-metodoa aplikatu ondoren, emaitza automatikoak eskuz zuzenduta, 668 UFren inguruko informazioa lortzen da.

Azterketa automatikoa

(4.4. eta 5.3. atalak)

- Oinarritzat hartu dira:

  - PARSEMEren gaztelaniazko corpusean etiketatutako 662 aditza+izena konbinazioak.
  - *DiCE* hiztegitik aztertu gabeko aditza+izena konbinazioak.

- Azterketa-metodoa aplikatu ondoren 266 konbinazio landu dira maiztasun-iragazkia jarrita, eta beste 214 iragazkirik jarri gabe. Azken horiek baztertu egin dira azkenean.

B. ERANSKINA

Fraseologia-baliabideak

Tesi-lanean zehar egin ditugun azterketa askotan, eskuz zehaztu behar izan dugu hitz-konbinazio jakin bat zuzena den ala ez, zer aldagarritasun morfosintaktiko daukan, eta abar. Lan horietan sortu zaizkigun zalantzak argitzeko, hainbat baliabide erabili ditugu erreferentzia gisa, eta horiek bilduko eta deskribatuko ditugu hemen. Hala, bide batez, erakutsiko dugu zer-nolako fraseologia-baliabideak ditugun eskuragarri gaztelaniaz eta euskaraz, sarean batez ere.

Hiztegi fraseologikoak

Gaztelaniaren eta euskararen arteko UF-ordaintzak jasotzen dituzten baliabideak murritz samarrak dira, esapide idiomatikoak bakarrik hartzen baitituzte kontuan, lokuzioak edota atsotitzak batez ere, kolokazioak alde batera utzita. Hortaz, gaur arte, ez da sortu hiztegirik argitzen duenik nola itzuli kolokazioak gaztelaniaren eta euskararen artean, nahiz eta halakoek ere berebiziko garrantzia duten itzulpenetan eta, oro har, edozein idazketa-lanetan.

Eranskin honetarako, gure lanerako esanguratsuak iruditu zaizkigun hiztegi batzuk hautatu ditugu: Mokoroaren *Ortik eta emendik* hiztegia, Intza proiektuaren sareko lokuzio-bilduma, gaztelaniazko *Redes* eta *Práctico* konbinazio-hiztegiak eta gaztelaniazko kolokazioen *DiCE* hiztegia. Sarean eskuragarri dauden baliabideak dira ia denak, *Redes* eta *Práctico* salbuetsita. Hala ere, interesgarria iruditu zaigu bi horiei ere tartea eskaintzea, hain zuzen

ere euskaraz falta den baliabide motakoak baitira.

Ortik eta emendik (Mokoroa Mugica, 1990)

Euskarazko fraseologia-baliabideen artean, Justo Maria Mokoroarena da gaur arte egin den bildumarik handiena: *Ortik eta emendik, repertorio de locuciones del habla popular vasca, oral y escrita, en sus diversas variedades*. Bi liburutan argitaratu zen 1990an, eta 92.000 esaeratik gora jasotzen ditu. Gaur egun, sarean ere eskuragarri dago, `hiru.eus` webgunean, eta bilaketak hainbat irizpideren arabera egin daitezke: euskarazko gako-hitzak idatziz, gaztelaniazko gako-hitzak idatziz, iturria adieraziz edo euskalkia zehaztuz.

Izenburuan bertan esaten denez, bildumako informazioa idatzizko eta ahozko jardunetik jasoa da, eta egileak 1.300 lekuko inguru eta 385 bibliografia-iturri aipatzen ditu. Dena dela, Esnalek (2001) eta Urizarrek (2012: 80–81. orr.) ohartarazten dutenez, bildumak ez ditu lokuzioak bakarrik jasotzen; lokuzioez gain, atsotitzak eta kolokazioren bat edo beste ere jasotzen ditu, bai eta UFtzat hartuko ez genituzkeen beste hainbat hitz-konbinazio ere. Hain zuzen, horregatik erabili dugu aurreko paragrafoan *esaera* terminoa, eta ez UF, Antonio Zavalak bildumaren hitzaurrean egin bezala.

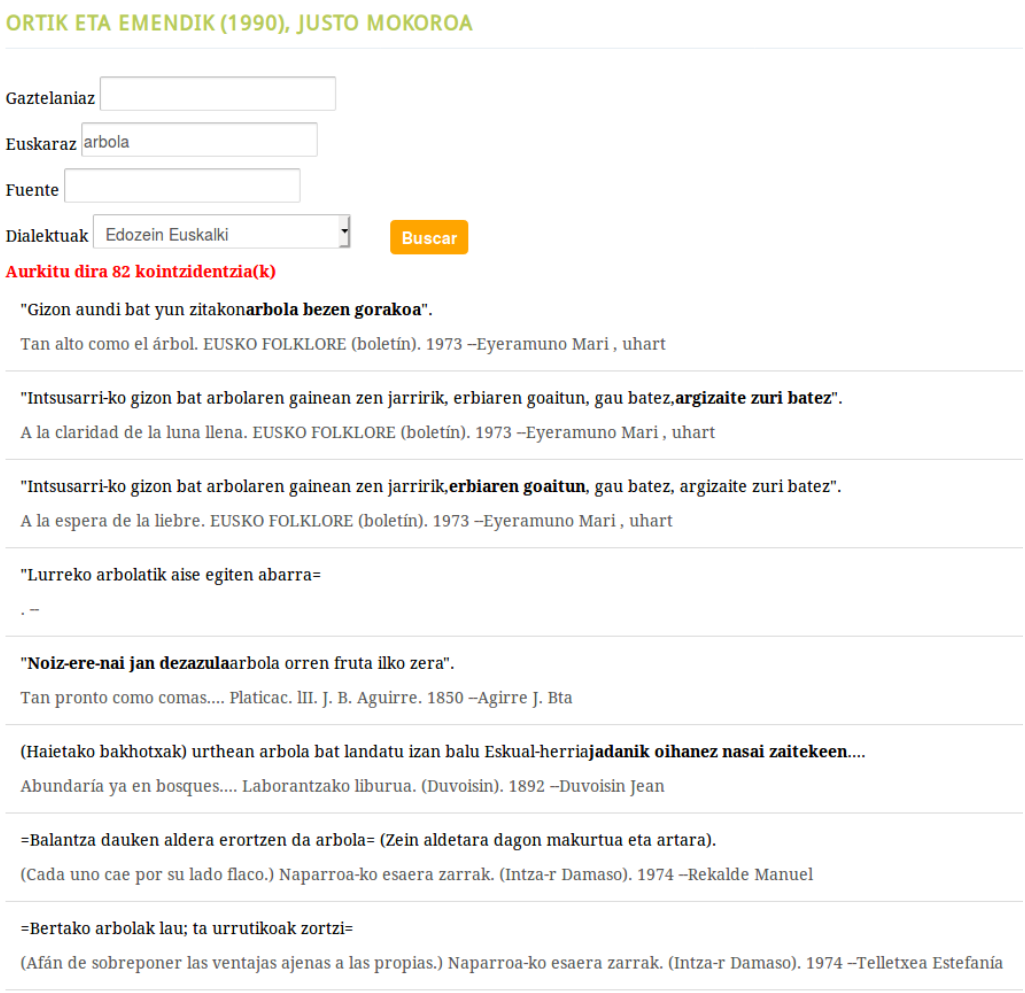
Sarrera bakoitzean, gaztelaniazko azalpena –ez beti ordaina– eta iturria ere zehazten dira (B.1). Hiztegia digitalizaturik egoteak asko errazten du haren erabilera, baina aipagarria da ez dagoela lematizatuta eta, hortaz, emaitzak ez direla beti nahi bezain zehatzak.

Euskal lokuzioak sarean (Intza Proiektua)

Intza proiektua Koldo Izagirreren *Euskal lokuzioak* liburuan (Izagirre, 1981) oinarritua da. Izagirrek, lan horretan, *Auspoa* saileko liburuetako lokuzioak jasotzen ditu batez ere, ahozko hizkeran halako gehiago erabiltzen direlakoan, baina kontuan hartzen ditu beste lokuzio-bilduma batzuk eta zenbait idazleren lanak ere. Guztira 7.000 lokuzio inguru jasotzen ditu, azalpenekin eta adibideekin batera, eta gaztelaniazko eta frantsesezko ordainak ere ematen dira. Sarreren erdia inguru aditz-lokuzioak dira.

Webgunean[1] aipatzen denez, sareko bertsioa "osagarri garrantzitsuz hornitua" dago, batetik berrikusketa- eta zuzenketa-lanak egin zaizkiolako jatorrizko lanari, eta bestetik sarrerak gehitzen ari zaizkiolako. Hala, *Euskal*

[1] `https://intza.armiarma.eus`

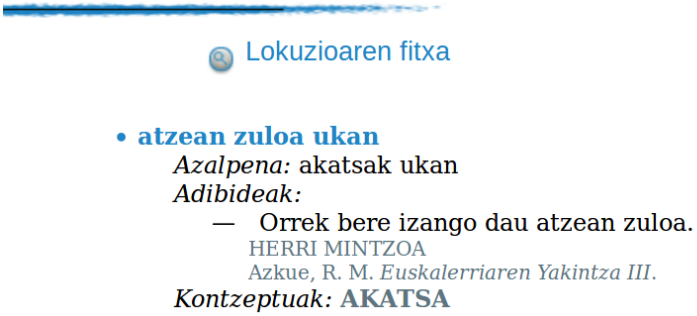


B.1 irudia – Mokoroaren sareko hiztegian *arbola* bilatuta lortzen den emaitzaren zati bat

lokuzioak sarean bildumaren helburua da "ondare linguistiko bat unibertsalizatzea" eta edonoren eskura jartzea euskarazko lokuzioak ezagutzeko aukera.

Gako-hitzak lokuzioetan, azalpenetan nahiz ordainetan bilatu daitezke. Gainera, lokuzioetako asko kontzeptuka antolatuta daude, eta kontzeptu horien aurkibidea ere eskaintzen da webgunean. Adibidez, B.2. irudian ikusten den emaitza *akatsa* kontzeptuan klik eginez lortzen da. Gero, lokuzio bakoitzak bere fitxa du, eta han ematen da informazio xeheagoa (B.3. irudia).

B.2 irudia – Intza proiektuaren hiztegian *akatsa* kontzeptuan jasotako lokuzio-zerrenda



B.3 irudia – Intza proiektuaren hiztegian *atzean zuloa ukan* lokuzioari dagokion fitxa

Redes eta *Práctico* (Bosque, 2004: 2006)

Gaztelaniaz, beste hizkuntza batzuetan bezala, ahaleginak egin dira azken urteotan konbinazio-hiztegiak osatzeko, hizkuntza-ikasleentzako laguntza-tres-

na gisara batez ere, hitz jakin bat zer beste hitzekin batera erabili ohi den jasotzeko. Ingelesezko adibide batzuk ematearren, multzo horretakoak dira *BBI Combinatory Dictionary of English* (Benson *et al.*, 1986), ingelesezko lehen konbinazio-hiztegitzat hartzen dena, eta *Oxford Collocations Dictionary* (McIntosh, 2009), tesi-txosten honetan erabili eta deskribatu duguna (4.3. atala).

Gaztelaniazko hiru hiztegi sartzen dira konbinazio-hiztegien multzoan, eta horietako bi RAEren baitan eginak dira: *Redes* eta *Práctico*. Bigarrena lehenengoan oinarritua denez, oso eredu antzekoa darabilte, baina lehena zabalagoa da bigarrena baino. Hona hemen bataren eta bestearen ezaugarri nagusiak:

- ***Redes: Diccionario combinatorio del español contemporáneo*** (Bosque, 2004). Konbinazio-hiztegi mardula da, eta bi eratako hiztegi-sarrerak jasotzen ditu: analitikoak –edo luzeak– eta laburtuak.

  – Sarrera analitiko gehienak aditzak, adjektiboak edo adberbioak dira, eta haien azpian jasotzen dira, sarreraren adierei buruzko azalpenak ez ezik, adiera bakoitzarekin konbinatu ohi diren hitzak, ezaugarri semantikoen arabera antolatuta. Esate baterako, *barajar* aditzaren barruan, multzo batean sartzen dira aukerei eta alternatibei dagozkien izenak (*posibilidad, hipótesis, opción, expectativa*), beste batean pentsamendu-unitateei dagozkien izenak (*propuesta, idea, teoría, plan, proyecto, tesis*), hirugarren batean datuei edo emaitzei dagozkien izenak (*cifra, dato, resultado*) eta abar. Horietako bakoitzari adibide bana ere ematen zaio.

  – Sarrera laburtuak, berriz, analitikoetatik automatikoki sortuak dira. Sarrerekin konbinatu ohi diren hitzak zerrenda moduan ematen dira, kategoria gramatikalaren arabera antolatuta, baina kategoria esplizituki zehaztu gabe. Adibidez, *hipótesis* hiztegi-sarreraren azpian, multzo batean zerrendatzen dira adjektiboak (*acertado, arriesgado, atinado, atrevido*), bigarren multzo batean aditzak (*afianzar(se), airear, alimentar*), eta abar.

  Kazetaritza-corpus batean oinarritua bada ere, corpusean agertzen ez diren konbinazio batzuk ere jasotzen ditu hiztegiak, lexikografoak naturaltzat hartu baditu.

- ***Práctico: Diccionario combinatorio Práctico del español contemporáneo*** (Bosque, 2006). *Redes*en bertsio sinpletua da. Sarreren azpian jasotzen da sarrera bakoitza zer hitzekin konbinatzen den, kategoria gramatikalaren arabera, eta azpimultzoak ere bereizten dira, semantikoki antzekoak diren lemak bilduta. Esate baterako, *hipótesis* sarreraren azpian, adjektiboak agertzen dira lehenik, eta honako azpimultzo hauek egiten dira, besteak beste: *acertada, atinada, certera, válida, cierta, fuerte, falsable*; *errónea, desacertada; descabellada, disparatada...*

DiCE: Diccionario de Colocaciones del Español (Alonso Ramos *et al.*, 2010)

Aurreko biak ez bezala, hasieratik sarean argitaratutako hiztegia da *DiCE*, eta sentimenduak adierazten dituzten izenekin osatzen diren kolokazioak biltzen ditu. Hortaz, oinarri-oinarrian badu aurrekoetatik bereizten duen ezaugarri garrantzitsu bat: *Redes*ek eta *Práctico*k konbinazio usuak lantzen dituzte, oro har; honek, berriz, kolokazioak. Alonsoren (2017) arabera, aurrekoek ere kolokazioak jasotzen dituzte funtsean, baina, ziur asko, "arrisku kontzeptualagatik" saihesten dute termino hori, hau da, kolokazio kontzeptuak literaturan sortu izan duen eztabaidagatik.

Bestalde, antolamenduari dagokionez, *Redes* eta *Práctico* hiztegiek murriztapen semantikoei begiratzen diete batez ere, eta *DiCE*k, aldiz, murriztapen lexikoei. Alonsoren (2017) ustetan, murriztapen lexikoak dira interesgarrienak gaztelaniazko testuak sortu behar dituen edonorentzat, eta *DiCE* hiztegiak erabilgarria izan nahi du, gaztelania-ikasleentzat ez ezik, itzultzaileentzat eta irakasleentzat ere.

*DiCE*ren oinarri teorikoa Zentzu-Testu Teorian datza (*Meaning-Text Theory*, Mel'čuk eta Polguere, 1987). Hiztegi-sarrerak izenak dira, kolokazioetako oinarriak, eta haien azpian jasotzen dira izenek izan ditzaketen adierak eta adiera bakoitzari dagozkion kolokatuak. Bi eratako informazioa jasotzen da sarrera bakoitzaren azpian:

- Informazio zentrala: semantikoa eta konbinazioei dagokiena. Adibidez, *admiración* sarreraren barruan, kolokatuak morfosintaxiaren arabera multzokatzen dira lehenik (*admiración+adjetivo, admiración+verbo, verbo+admiración*, etab.), eta multzo horietako bakoitza, esanahika (sentir → *deber, dispensar, profesar...*; ser objeto de → *gozar, tener*;

continuar sintiendo → *conservar*, etab.). Esanahi horiek funtzio lexikalen araberakoak dira (Mel'čuk, 1996), baina parafrasian ere ematen dira, erabiltzaileen erraztasunaren mesedetan.

- Maiztasunari eta ikasketa-mailari dagokion informazioa. Maiztasuna izen-adiera bakoitzari esleitzen zaio (Vincze eta Alonso Ramos, 2013), eta kolokazioen zailtasun-maila maiztasun horren arabera kalkulatzen da (García Salido eta Alonso Ramos, 2018), Hizkuntzen Europako Erreferentzia Marko Bateratua oinarritzat harturik.

Horrez gain, *HARenEs* tresna ere sortu dute (Alonso Ramos, 2016), *DiCE*ko edukietan oinarritua, ikasleen kolokazio-akatsak automatikoki identifikatzera eta hobetzeko proposamenak egitera bideratua. Izan ere, Alonsoren arabera (2017), beharrezkoa da hiztegi independenteetatik harago joatea, baliabide horiek corpusekin eta beste hiztegi batzuekin lotuz edo haien ezagutza beste tresna batzuetan integratuz, lagungarri izan daitezen, besteak beste, idazketa-lanetarako edo hizkuntzen irakaskuntzarako (Wehrli, 2004; Abel, 2010; Wanner *et al.*, 2013; Granger eta Paquot, 2015).

Corpus-bilatzaileak

Esan dugunez, euskaraz ez dugu *Redes* eta *DiCE*ren gisako konbinazio-hiztegirik. Hala ere, azken urteotan asko ugaritu dira sarean kontsultagarri dauden euskarazko corpusak, eta horietako batzuk hasiak dira hitz-konbinazioen bilaketak egiteko aukerak eskaintzen. Hortaz, hiztegietan dagoen hutsunea corpusen bidez betetzen ari da, nolabait, pixkanaka. Dena dela, gaztelaniazko eta euskarazko konbinazioak berariaz lantzen dituzten corpus kontsultagarriak elebakarrak dira, eta bi hizkuntzen arteko ordaintzak topatzeko ez dago fraseologiaren alderdia hain landuta duen bilaketa-tresnarik oraingoz.

Atal honetan, UFen informazioa eskuratzeko baliagarriak diren corpus-bilatzaile batzuez jardungo dugu. Euskarazkoen artean, *Elhuyarren hitz-konbinazioen corpusa*ri eta *Egungo Testuen Corpusa*ri begiratuko diegu; gaztelaniazkoen artean, berriz, *CORPES*i. Azkenik, bi corpus paralelo ere deskribatuko ditugu, fraseologia berariaz lantzen ez badute ere: *Elhuyar web-corpus paraleloa* eta *TextReference* testuingurudun hiztegia.

Ez dago esan beharrik corpus horiek baino gehiago ere badaudela sarean erabilgarri, baina, hiztegiekin egin bezala, corpusen artean ere hautaketa bat

egin dugu eranskin honetarako. Tesi-lanerako interesgarrientzat jo ditugunak ekarri ditugu, erabilgarrienak iruditu zaizkigunak bilaketak egiteko moduagatik, tamainagatik edo bestelako berezitasunen batengatik. Azalpenak emateko, hizkuntzaren arabera multzokatuko ditugu baliabideak.

Elhuyarren hitz-konbinazioen corpusa eta *Egungo Testuen Corpusa*

Elhuyar web-corpus elebakarrak bilaketa-tresna espezifiko bat dauka hitz-konbinazioentzat[2]. Corpusa internetetik automatikoki bildutako testuz osatuta dago (Leturia, 2012), eta ia 125 milioi hitz ditu guztira. Hitz-konbinazioen bilaketak egiteko tresna hizkuntza-teknologia aurreratuen bidezkoa da, Elhuyar Fundazioko I+G taldearen eta, zehazki, Gurrutxagaren tesilanaren baitan inplementatua (Gurrutxaga eta Alegria, 2011; Gurrutxaga *et al.*, 2016).

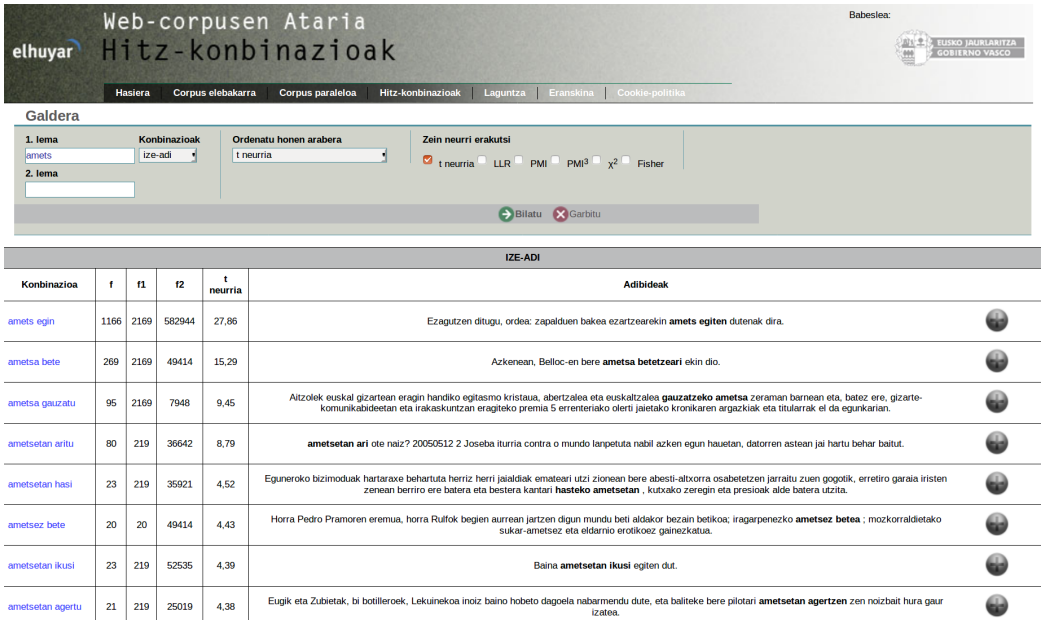
Hiru osaera morfologikotako konbinazioak bila daitezke: izena+aditza, izena+izena eta izena+adjektiboa. Bilaketan, erabiltzaileak nahi duen lema edo lema-parea idatz dezake, eta hitz-konbinazioko osagai-hitzen kategoriak ere zehaztu ditzake. Gero, emaitzak taula batean biltzen dira, zutabeka: (1) lortutako hitz-konbinazioak, (2) hainbat neurri estatistiko eta (3) konbinazio bakoitzeko adibide bana. Ordenatze-irizpidea aukerakoa da, eta zer neurri estatistiko ikusi nahi diren ere egokitu daiteke.

Emaitzetan jasotzen diren konbinazioak hainbat idiomatikotasun-mailatakoak dira: lokuzioak, kolokazioak eta, tarteka, konbinazio libreren bat, neurketa estatistikoen ondorioz lortua. Bestetik, egileek eurek ohartarazten dutenez, testuen zuzentasuna ez da erabatekoa, testu horien bilketa ere automatikoki egina delako. Hona hemen, B.4. irudian, *amets* idatzita eta izena+aditza kategoria zehaztuta lortzen den emaitzaren zati bat.

Gero, konbinazioaren gainean klik eginez gero, adibideak bistaratzeko aukera dago, eta bilatutako lemak adibideen erdialdean eta nabarmenduta agertzen dira, *KWIC* eran (*Key Words in Context*). Esate baterako, B.5. irudian ikus daitezke *amets egin* konbinazioaren adibideak.

Beste corpus batzuk ere hasiak dira konbinazioei arreta eskaintzen, eta, fraseologiari *Elhuyar web-corpusa*k bezainbesteko garrantzia ematen ez badiote ere, zenbaitek aukera ematen dute hitz jakin bat zer beste hitzekin konbinatu ohi den ikusteko. Halakoa da, besteak beste, *Egungo Testuen*

[2] `http://webcorpusak.elhuyar.eus/cgi-bin/kolokatuak.py`

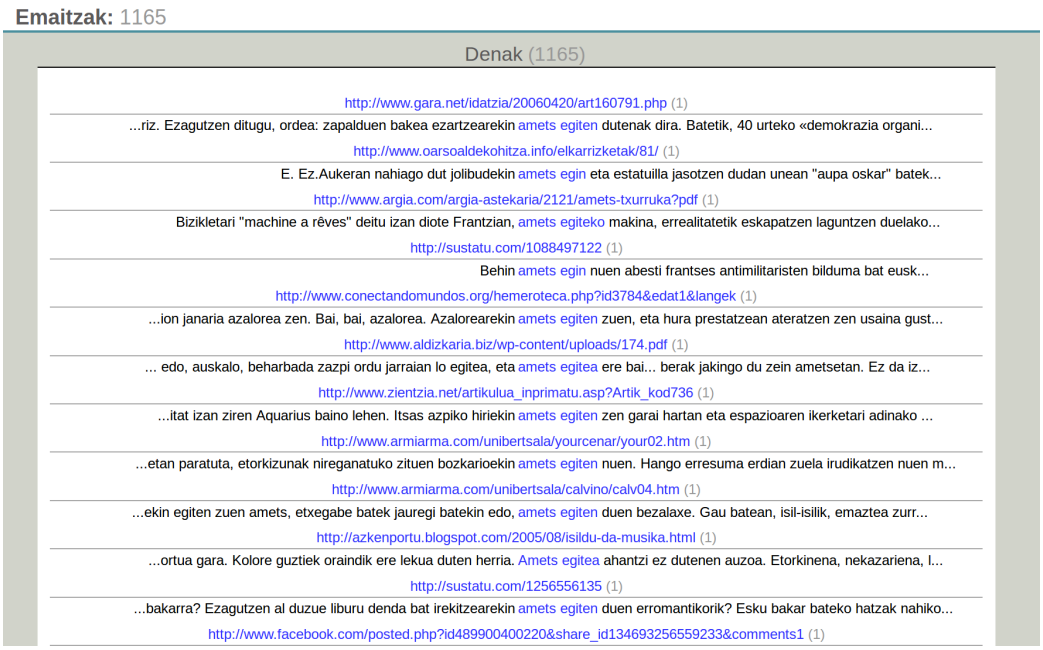


B.4 irudia – Elhuyar web-corpusa: *amets*+aditza bilaketaren emaitza.

Corpusa[3]. Hainbat generotako testuak biltzen ditu, 2001etik 2015era bitartean argitaratuak, prentsakoak eta literatur-liburuetakoak gehienak, baina baita lan zientifiko-teknikoetako eta *Wikipedia*ko batzuk ere. Guztira 269 milioi hitz baino gehiago dauzka.

Bilaketak lemen arabera egiten dira, eta askotariko informazioa ematen da emaitzetan: zer formatan erabili ohi den lema hori, urtetik urtera zer bilakaera izan duen, jatorrizko testuetan ala itzulpenetan agertu den gehiago, zer iturritan agertu den eta, azkenik, zer beste hitzekin konbinatzen den. Azken atal horretan, konbinazioak kategoria gramatikalaren arabera iragazteko aukera dago, eta posible da emaitzetan lortutako adibide guztiak fitxategi batean jaistea. *Elhuyar web-corpus*ean eginiko bilaketa berbera eginda, B.6. irudiko emaitza lortzen da Konbinatoria atalean. Klik eginez gero, hemen ere *KWIC* eran bistaratzen dira corpuseko adibideak.

[3] `https://www.ehu.eus/etc/`



B.5 irudia – Elhuyar web-corpusa: *amets egin* konbinazioaren adibideak.

CORPES, Corpus del Español del Siglo XXI

CORPES corpusa[4] RAEk sortua da, eta 285.000 dokumentuz osatuta dago –286 milioi hitzez–. Dokumentu gehienak liburuetakoak dira, baina prentsako testuek ere corpusaren zati handi samar bat osatzen dute; gainerakoak internetetik eta bestelako iturrietatik bildutako lanak dira. Corpusaren % 90 idatzizko testuei dagokie, eta Espainiako nahiz Ameriketako dialektoetako lanak jasotzen dira (% 30 eta % 70, hurrenez hurren).

Bilaketak egitean, bi aukera daude hitz-konbinazioen inguruko informazioa lortzeko.

- *Concordancia* atalean, bilatu nahi den lema idaztea eta, *+Proximidad* sakatuta, bigarren lema bat gehitzea. Aukera dago, gainera, bi lemen artean gehienez ere hitz-kopuru jakin bateko distantzia nahi dela zehaztekoa, bai eta hitzen hurrenkera zehaztekoa ere.

- *Coapariciones* atalean, bilatu nahi den lema idaztea, besterik gabe. Hala, lema horrekin batera agertzen diren beste lemen zerrenda bat

[4] `http://web.frl.es/CORPES/view/inicioExterno.view`


konbinazioak beste lemekin

	pisua	ager.
amets bete	8,74	969
amets zapuztu	8,23	130
amets egin	8,21	2.679
amets gauzatu	8,12	281
amets prefabrikatu	7,73	16
amets ero	6,85	75
amets esnatu	6,80	88
amets ilusitu	5,71	2
amets iratzarri	5,54	21
amets doitu	5,30	18
amets gorpuztu	5,17	17
amets kontatu	5,16	81

B.6 irudia – ETCn *amets* bilatuta Konbinatoria atalean lortzen den emaitza (aditzak bakarrik hautatuta).

ematen da, neurri estatistikoak erabiliz lortua, eta, haietako baten gainean klik eginez gero, bi lemak barne hartzen dituzten esaldien adibidezerrenda bat bistaratzen da.

Aurreko bietan bezala, emaitzak KWIC eran erakusten dira; B.7. eta B.8. irudietan jaso dugu emaitzen adibide bana.

Elhuyar web-corpus paraleloa eta *TextReference*

Gaztelania-euskara corpus elebidunei dagokienez, esana dugu gaur egun eskuragarri dauden bilatzaileek ez dutela aukerarik ematen fraseologia-bilaketa espezifikoak egiteko. Hala ere, haietako batzuek badute gure helburuetarako interesgarria den aukeraren bat.

Elhuyar web-corpus paraleloan, adibidez, bi lema bilatzeko aukera ematen da. Interfazea, defektuz, lema bat hizkuntza batean eta bestea beste hizkuntzan bilatzeko konfiguratuta agertzen da, baina aukera ematen da hizkuntza aldatzeko. Hala, hizkuntza bereko bi lema zehazten baditugu, posible

B.7 irudia – CORPESeko Concordancia atala: *inspirar+confianza* bilaketaren emaitza.

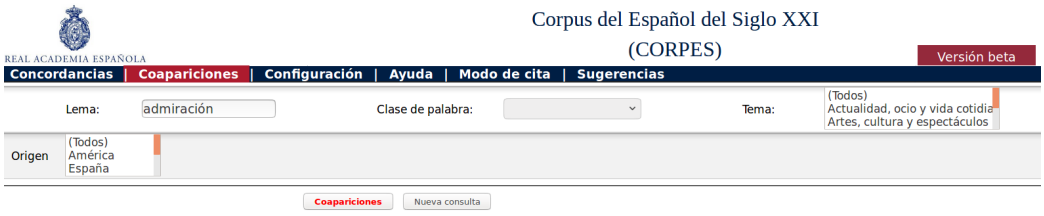
da bi lema horien konbinazioa nola itzuli izan den ikustea. Esan beharra dago, dena den, bilaketa-tresna ez dagoenez berez hitz-konbinazioak bilatzeko prestatuta, topatzen diren esaldi guztiek ez dutela hitz-konbinazioa nahitaez topatzen, baizik eta bi lemak barne hartzen dituzten esaldiak, oro har (B.9. irudia).

Bestetik, corpusen erabilera handitzearekin batera, badirudi gero eta egile gehiagok ikusi dutela hiztegiak eta corpusak nolabait uztartzeko beharra, batzuen eta besteen arteko mugak gero eta lausoagoak direla uste izanik (Alonso Ramos, 2017). Hala sortu dira *testuingurudun hiztegi* deritzenak, non bilaketak corpusetan egiten baitira baina ordain-zerrenda bat ere proposatzen baita. Zenbaiten ustez, halakoek aukera berriak eskaintzen dituzte hainbat erabilerentzat, hizkuntza-ikasleentzat eta itzultzaileentzat besteak beste (Alonso, 2013; Buyse eta Verlinde, 2013). Ingelesezkoen artean, *Linguee*[5] da, ziur asko, azken urteotan ezagunen egin denetako bat.

Euskarara etorrita, bada sortu berria den testuingurudun hiztegi bat: *TextReference*[6]. Bilaketak sareko hainbat corpus libretan egiten dira, 700.000

[5] `https://www.linguee.es`

[6] `http://www.textreference.com/eu/`. Mikel Artetxe da tresnaren egilea, eta *Opendata* lehiaketako saria jaso du proiektu horrengatik. Sortu berria denez, ez dago oraindik tresnaren inguruko dokumentazio zabalik, baina bai erabilera-argibideak ematen di-



	Clase	Freq	MI	LL SIMPLE	T-SCORE
profesar	verbo	58	11,47	351,81	7,61
granjear	verbo	15	11,26	88,96	3,87
rendido	adjetivo	24	11,1	140,09	4,89
gratitud	sustantivo	39	10,65	216,89	6,24
envidia	sustantivo	81	10,39	437,94	9,00
respeto	sustantivo	406	10,32	2.189,56	20,14
exclamación	sustantivo	15	10,3	80,17	3,87
silbido	sustantivo	21	9,87	106,75	4,58
aprecio	sustantivo	19	9,83	96,12	4,35
cariño	sustantivo	107	9,6	527,70	10,34
suscitar	verbo	50	9,58	245,54	7,07
mutuo	adjetivo	58	9,55	284,01	7,61
afecto	sustantivo	55	9,27	259,97	7,41
asombro	sustantivo	43	9,2	201,34	6,55
incredulidad	sustantivo	10	9,14	46,44	3,16
digno	adjetivo	77	9,11	356,48	8,77
simpatía	sustantivo	32	9,04	146,67	5,65
agradecimiento	sustantivo	27	8,97	122,69	5,19
despertar	verbo	194	8,76	859,40	13,92
signo	sustantivo	83	8,68	363,13	9,11

1 - 20 de 393 página: 1 2 3 4 5 6 7 8 9 ... 20

B.8 irudia – CORPESeko Coapariciones atala: *admiración* bilaketaren emaitza.

itzulpen-adibide ingurutan oraingoz. Bilaketa-barran nahi den lema idatzita, lema hori barnean duten esaldiak eta haien itzulpenak bilatzen dira, eta, esaldi horiek oinarritzat hartuta, neurri estatistikoak erabiltzen dira lemen ordainak automatikoki lortzeko. Hala, emaitzak erakustean, ordain-zerrenda bat bistaratzen da lehenik, ordain bakoitzaren agerpen-ehuneko eta guzti, eta azpian jasotzen dira ordain bakoitzari dagozkion bina adibide. Ondoren, aukera ematen da corpuseko adibide gehiago ere ikusteko.

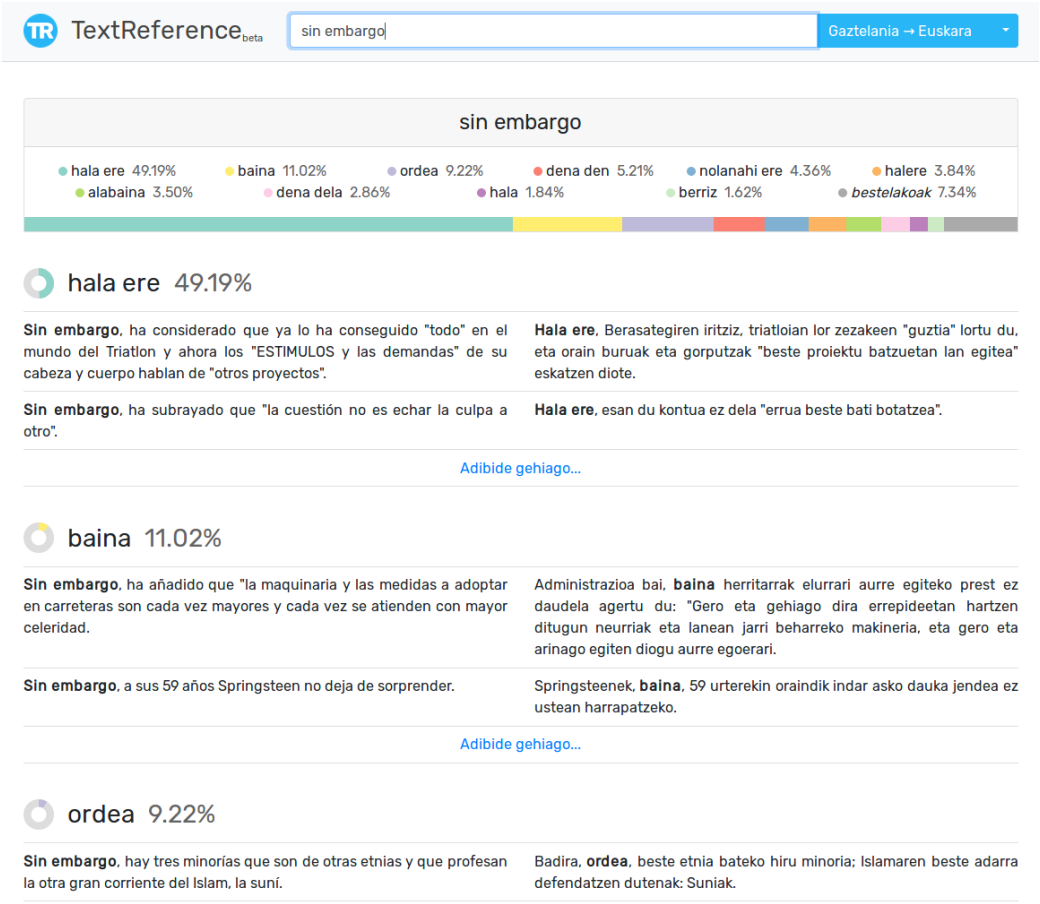
Hitz-konbinazioei dagokienez, askotan errepikatzen diren hitz-segida jarraituak ere bila daitezke, eta, hala, posible da UF batzuen ordainak ere lortzea (B.10. irudia).

Sarean dagoena tresnaren *beta* bertsioa baino ez da, eta esan beharra dago emaitzak ez direla beti nahi bezain finak gaur-gaurkoz, batez ere ez delako inongo tratamendu morfologikorik egiten oraindik eta, oinarrian duen corpusa txiki samarra denez, bilaketa batzuentzat ez delako kalitatezko emaitzarik

tuen txosten bat, sariketaren webgunean: `http://opendata.euskadi.eus/contenidos/informacion/concurso_apps_candidaturas_18/es_def/adjuntos/15.pdf`.

B.9 irudia – Elhuyar web-corpus paraleloa: *dar+confianza* bilaketaren emaitza.

lortzen. Hala ere, proiektua hasi baino ez da egin, eta etorkizun interesgarriko lana da, inondik ere.



B.10 irudia – TextReference: *sin embargo* bilaketaren emaitza.

C. ERANSKINA

Itzulpen automatikoa ebaluatzeko gidalerroak

Ebaluazioa egiteko urratsak

1. Begiratu taulako hirugarren zutabean zein den ebaluatu beharreko hitz-konbinazioa. Aditza eta izena agertuko zaizkizue, marra batez elkartuta. Adib.: *tener_ repercusión*

2. Irakurri gaztelaniazko esaldia, eta bilatu hitz-konbinazio hori esaldi barruan. Gero, irakurri euskarazko itzulpenak, eta bilatu zer ordain eman zaion batean eta bestean hitz-konbinazioari. Normalean, gaztelaniazko esaldiaren barruan dagoen leku berean egongo da euskarazkoan ere, gutxi-gorabehera.

3. Irakurri gaztelaniazko esaldia, eta bilatu hitz-konbinazio hori esaldi barruan. Gero, irakurri euskarazko itzulpenak, eta bilatu zer ordain eman zaion batean eta bestean hitz-konbinazioari. Normalean, gaztelaniazko esaldiaren barruan dagoen leku berean egongo da euskarazkoan ere, gutxi-gorabehera. Hitz-konbinazio horri emandako ordainak kontuan hartuta, dokumentu honetako irizpide nagusiak gogoan harturik, erabaki:

    - Lehen itzulpena hobea den (A sistema)
    - Bi itzulpenak diren berdinak (biak berdin)
    - Bigarren itzulpena hobea den (B sistema)

Oharra: A eta B sistemek ez diote beti sistema berari erreferentzia egiten; ausaz jartzen dira itzulpenak hurrenkera batean edo bestean.

Ebaluaziorako irizpideak

Ebaluatzaile bakoitzak bere senaren arabera erabaki beharko du itzulpen bat ala bestea hobetsi. Dena dela, honako irizpide hauek lagungarriak izan daitezke zer baloratu erabakitzeko orduan.

- Hitz-konbinazioari emandako itzulpenetan lexikoa desberdina bada:
  - Bai izena eta bai aditza badira desberdinak, aditzak argitzen du normalean itzulpenetako bat ala bestea den egokiagoa. Hortaz, zalantza kasuetan, eman lehentasuna aditzari. Adib.: *llamar la atención* > *arreta deitu* ala *atentzioa eman*?
  - Aditzak alderatzean, lagungarria izan liteke pentsatzea ea bietako bat gaztelaniazko kalkoa ote den. Itzulpen bat gaztelaniazkotik nabarmen kalkatua bada, baliteke bestea txukunagoa izatea. Dena dela, kontuan hartu kalko guztiak ere ez direla okerrak eta norberak erabaki beharko duela kalko jakin bat zenbateraino den onargarria. Adib.: *poner fin* > *amaiera jarri* ala *eman*?
  - Izenei begiratzean, gutxi batzuetan nabarmena da erantzun bat bestea baino hobea dela. Beste askotan, ordea, itzulpen bateko eta besteko izenak sinonimoak-edo izaten dira. Bi izenak onargarriak direla uste baduzue (eta beste ezer aldatzen ez bada), jo itzazue itzulpenak berdintzat. Adib.: *arreta* eta *atentzio*.
- Ezaugarri morfosintaktikoak desberdinak badira:
  - Lexikoa ere desberdina bada itzulpen batean eta bestean:
    - Bi itzulpenak badira onargarriak lexikoari dagokionez, baina ezaugarri morfosintaktikoak egokiagoak badira bietako batean, itzulpen hori da hobetzat jotzekoa.
    - Lexikoan itzulpen bat bestea baino nabarmen hobea bada, baina morfosintaxiari dagokionez okerragoa, bi itzulpenak berdintzat jotzekoak dira.

– Lexikoa ez bada aldatzen batetik bestera, begiratu, besterik gabe, zein den hobea gramatika aldetik. Akats gehienak komunztaduran egiten dira; beraz, jarri arreta ezaugarri horri.