Laburpena

Lan honen helburua kazetaritza-corpusean esaldia osatzen duten elementuak zein hurrenkeratan agertzen diren aztertzea izan da. Euskarari dagokionez, automatikoki burutu den lehen azterketa izango da hau ziurrenik ere. Betebehar horiei buru egiteko, esaldi bakunak analizatzeko egokia den eta izen-sintagmak eta adizlagunak tratatzeko estaldura ia osoa duen euskarazko analizatzaile sintaktiko konputazionalaz baliatu gara, PATRII-z (Shieber M., 1986), alegia. Ondoren, euskararen hitz-hurrenkera zein den zehazteko ezagutzen diren eta eskuz burutuak izan diren lehen azterketa estatistikoekin (de Rijk, 1969; Hidalgo, 1995) alderatu dugu.
Hala, baliatu ditugun tresna eta errekurtso horiei esker kazetaritza-corpusean nagusitzen den hurrenkera SOA dela egiaztatu dugu.

Abstract

The goal of this research has been the study of the Basque word-order in a corpus of journalism. As far as the Basque language is concerned, this is the first time this task is automatically pursuit. By means of the unification-formalism named PATR-II (Shieber M., 1986), we have produced a computational grammar which describes the structure of the Basque noun phrase and that of the matrix simple sentences. After specifying the main word-order in the mentioned corpus, we compared it with the first statistic studies manually carried out by de Rijk (1969) and Hidalgo (1985) in order to analyze the usual Basque word-order.
This way, and thanks to the means and resources used in this work, we conclude the main word-order in the analyzed corpus of journalism is SOV.

AURKIBIDEA:

1. SARRERA

1.1 Motibazioa eta helburua

J.H. Greenberg-ek (1961/63) proposatutako hizkuntzen tipologia kontuan izanik, euskara SOA[1] motako hizkuntza dela onartu izan da urteetan. Ideia hau hartu dute abiapuntutzat esaldia osatzen duten elementuen hurrenkera aztergai izan duten ondorengo zenbait lanek.

Bestelako iritzirik ere izan da, ordea; Mitxelenak, esaterako, de Rijk-en lanari bedeinkazioa ematen zion artikulu berean eginiko oharretan (1978; 1987)[2] euskal tradizio moldeak jarraitzen dituzten hurrenkerak sumatzen direla zioen. Berrikiago B. Hidalgok *Hitzen ordena euskaraz* (1995) izeneko bere tesia osatzerakoan, zalantzan jarri du euskara SOA motako hizkuntza dela. Honek, euskarazko hitzen ordenaren inguruan topiko ugari oker zabaldu dela dio. Hori egiaztatzeko garai desberdinetako idazleen testuetara jo du eta ondorioztatu du berak egindako azterketak ez datozela bat, esaterako, de Rijk-ek (1969)[3] egindakoekin eta aurrerantzean eztabaida honek irekia beharko lukeela izan.

Horixe da, hain zuzen ere, lan honen **motibazioa**, Hidalgok zabalik utzi duen atetik sartuko gara eta zabalagoa den corpusean, kazetaritza-corpusean zehazki, esaldia osatzen duten elementuak zein hurrenkeratan agertzen diren aztertuko dugu. Ondoren, de Rijk-en (1969) eta Hidalgoren (1995) azterketetako datuekin alderatuko ditugu.

Ukaezina da hizkuntzaren azterketan corpusek duten garrantzia. Testu-corpusak testu-masa handiak dira eta euren helburu nagusietako bat da ebidentzia linguistikoa adieraztea. Hizkuntzalariak ere, haren teoriak babestuko dituen eta hizkuntzaren joera nagusiak erakutsiko dizkion erreferentzia-elementuak (datu enpirikoak) behar ditu, eta corpusek erreferentzia hori sistematizatzen duten testu edo hizketa multzoa osatzen dute.

Analisiaren abiapuntua izango den eta aztergaia osatuko duen enuntziatu multzoa, corpusa alegia, tamaina handiagokoa da ildo honetan burutu diren azterketa estatistikoak baino. Kontuan izan behar dugu corpusaren tamainak duen garrantzia, hain zuzen ere, corpusak ugaltzen diren neurrian, berauetan oinarritutako ikerketek sakontasunean eta zehaztasunean irabazten dutelako.

---

[1] SOA : subjektu, objektu, aditza, alegia.

[2] *"En estas notas se examina un cierto tipo de ordenaciones, a título de muestra, como posible señal de una manera de contar que, por los textos en que lo hallamos, puede considerarse con alguna confianza como continuador de moldes tradicionales. Se alude, en resumen, al hecho de que, tanto en cantares épicos como en refranes, se encuentran ciertas ordenaciones que, aun cuando no sean mayoritarias -y acaso precisamente por eso mismo-, tienen una frecuencia estadística suficiente para que merezcan ser tenidas en cuenta, al menos a título de ensayo, en un estudio de nuestra estilística"* FLV, 1978.

[3] R.P.G. de Rijk (1969): *"Is Basque an S.O.V. Language?"* FLV I-3 (1969), 319-351

Ikerlan honetan erabiliko dugun corpuseko informazioa erauzteko IXA lantaldean[4] garatu dugun analizatzaile sintaktiko partzialaz (PATR II)[5] baliatuko gara. Partziala da, perpausaren ulerkuntza osoa lortu gabe emaitza erabilgarriak lor daitezkeelako, hau da, mota honetako analizatzaileek ohiko analisi sintaktikoaren informazio zati bat, ez guztia lortzen dute. Beraz, mekanismo honen bitartez, esaldi batean ageri diren aditza eta berari dagozkion sintagmak lor ditzakegu. Esku artean dugun lanaren **helburua** da, ordea, esaldia osatzen duten aditza eta honekin komunztadura egiten duten *absolutibo* eta *ergatibo* kasuak zein hurrenkeratan agertzen diren aztertzea.

Helburu aplikatuei dagokienean, Lengoaia Naturalaren Prozesamendurako (LNP) lanetan ezinbestekoa da euskara bezalako hizkuntza batean esaldia osatzen duten hitzen hurrenkera eta aldi berean esaldiena bera ere aztertzea. Azterketa hau baliagarria izan dugu esaterako, euskararen *treebank*-ari buruzko ikerlanean[6] jarraitu beharreko formalismoa aukeratzeko garaian.

1.2 Ikerlanaren eskema

Ikerlanaren motibazioa eta helburua azaldu ondoren, 2. atalean euskararen hitz-hurrenkeraren inguruko zenbait lanen berrikustapena egingo dugu, bereziki estatistika baliatu dutenena: de Rijk eta Hidalgoren azterketa estatistikoak aipatuko ditugu. 3.ean ikerlan hau burutzeko beharrezkoak izan ditugun baliabideak deskribatuko ditugu, corpusa eta analizatzaile sintaktiko partziala alegia. 4. atalean, euskarazko kazetaritza-corpusaren azterketa eta bertatik ateratako emaitzen berri emango dugu. 5. atalean, aurreko atalean lortutako datu horietako batzuk orain arte egindako azterketa estatistikoekin erkatuko ditugu. Bukatzeko, 6. atalean ikerlan honetatik atera ditugun ondorioak aipatuko ditugu.

2. AURREKARIAK

Hogeigarren mendearen hasieran S. Altubek zenbait arau eman zituen euskararen hitz hurrenkera jatorra biltzen zutelakoan. Ikus daitekeenez, gaur eguneko testu idatziek, oraindik ere, betetzen dituzte bi lege nagusi horiek, bata *galdegaiarena* (Azkuek 1894rako arautu zuena)[7] eta bestea, berriz, aditza atzeratzearena esaldietan.

Horren harira, euskararen esaldiaren hurrenkera oinarrizkoa edo neutroa[8] SOA dela onartu izan da. Euskararekin ukipenean dauden latinetiko hizkuntzak eta ingelesa, aldiz, SAO motakoak dira.

---

[4] Ikus http://ixa.si.ehu.es

[5] Gojenola, K. (2000) *Euskararen sintaxi konputazionalerantz*. Doktoretza-tesia.

[6] "*Corpusaren etiketatze sintaktikoa analizatzailea eraikitzeko"* M.J. Aranzaberen doktoretzako ikerlana.

[7] Azkue, R.M. (1894) *Ensayo Práctico* eskuizkribu argitarabean dio: *"Como se pregunta se contesta.../Esto quiere decir que si por ejemplo se pregunta ¿Non dago Txomin? se debe contestar: Txomin etxean dago"*.(33. or.)

[8] Osa E. (1990) "*Euskararen hizordena komunikazio zereginaren arauera* doktoretza-tesian hala dio: *"Beraz, ordena neutroaz hitz egiten dugunean zera adierazi nahi dugu, testu(inguru)arekiko menpekotasunik txikiena adierazten duena... Esan dugun bezala, neutraltasun hau ipintzen dugu guk euskarazko ordena kanonikoaren oinarrian, maiztasun estatistikoaren gainetik"* .

Greenberg-ek agerpen-maiztasunarekin lotzen du ordena neutroa eta badira euskararen hitz-hurrenkera metodo estatistikoen bidez aztertu duten hizkuntzalariak: de Rijk eta Hidalgo.

Atal honetan Altuberen eskolaren ondorioez hitz egingo dugu lehenik eta ondoren, azterketa estatistikoen ondorioez.

2.1 Altuberen eskola

Altubek esaldi barruko elementuak zein hurrenkeratan ezarri behar diren, edota esaldi konposatuetan bata bestearen ondoan nola kateatu behar diren jakiteko ematen dituen erregelek eragin handia izan dute bere garaiko eta ondorengo idazleen idatzietan.

Bere garaian hitzen arloan nabarmentzen zen purismoaren edo garbizalekeriaren kontra borrokatu bazuen ere, aldi berean euskararen joskera jator garbia zein zen aztertzearen alde azaldu zen. Horrela, bere aurrekoak ziren Cardaveraz eta Azkueren antzera, gaztelaniaren egituratik urruntzen zen era proposatu zuen euskaraz idazteko, hiztunek egiten zutena kontuan hartu gabe, edo hobeto esan, hiztunek egiten zutena zuzentzearren, erdal moldeen arabera egindakoa zelakoan.

Galdegaia, aditza atzeratzea eta esaldiaren luzera oinarri zuten lege horiek, esan bezala, ildo nabaria utzi dute urteetan euskal idazle, irakasle eta hizlarien artean. Aipagarria da zenbait idazleren lanetan nabarmentzen den estilo aldaketa; hau da, teoria hau aldarrikatu aurretik tradiziozko ordena batean idatzirik zeuden lanen argitalpen berria egitean, eskola honen eraginez hitz-ordena berria erakusten dute; honen adierazgarri ditugu Azkueren beraren lanak[9].

Dena dela, Altuberen hitz-ordenari buruzko arau hauek jaso izan dute kritikaren bat edo beste; horien artean Mitxelenak (1953)[10] egiten diona daukagu. Honek, Altube ametsezko euskal joskera ideal baten bila dabilela dio eta ez duela kontuan hartzen, behar den neurrian ez behintzat, euskara idatziaren tradizioa.

2.2 Azterketa estatistikoak

2.2.1 de Rijk

Bereak dira, euskarazko hitzen ordenamenduaren inguruan egin diren lehen estatistika ezagunak.

Azterketa hau egiteko hiru multzo desberdinetan banatzen ditu erabilitako testu-corpusak:

---

[9] 1888ko "Grankanton arrantsaleak" (*Euskera*, 1988, 379-389), adibidez.

[10] *"Eztiot nik ezertxo ere kenduko Altube jaunaren lanari, bere gaiean alderatzeko gauzarik ez baitu bestek egin. Zerbait esan diteke orratik: bi gauza, berezirik egon bear luketenak, elkarrekin naasten dituela, dena eta bear litzakeana...-Labur esateko, Altube jaunak finkatutako legeak, euskerak gorde bear lituzkeanak izango dira agian, baiña ez euskerak gorde edo gordetzen dituenak. Eztira, geienez ere, euskerarenak, zenbait dialekturenak baizik .."* L. Mitxelena in "Arnaut Oihenart", *Boletín de Amigos del País*, 1953, 460.

- I testu-corpusa, Aita J.M. Barandiaranen 1920-1936 bitarteko ipuin-bilduma folklorikoek osatzen dute.

- II corpusean Nemesio Etxanizek idatzitako zenbait antzerki-lan ditugu, *Euskal-Antzerkiak* liburuan (Kuliska Sorta 27-28, Itxaropena, Zarauz 1958: 7-132 or.) argitaratuak 1958an.

- IIIkoak, N. Etxanizek Mérimée-renetik eginiko kontakizun laburren itzulpenak dira; aipaturiko *Euskal Antzerkiak* liburuan daudenak (135-159 or.)

Corpus horietatik ateratzen dituen emaitzak honako hauek dira:

	I	II	III	Guztira
Esaldi kopurua	209	183	67	459
SOA	138	80	41	259
SAO	48	67	21	136
OAS	11	17	3	31
OSA	5	13	1	19
ASO	6	4	1	11
AOS	1	2	0	3

Ehunekotan:

	I	II	III	batez bestekoa
SOA	66	44	61	57
SAO	23	37	31	30
OAS	5	9	5	6
OSA	2.5	7	1.5	4
ASO	3	2	1.5	2.5
AOS	0.5	1	0	0.5

Guztietan, ikus daitekeen bezala, S/O/A hurrenkera da nagusitzen dena; beraz, esaldiko hiru elementu nagusien (subjektu, objektu eta aditza) ordena erlatiboa kontuan hartuz, euskara estatistikoki S/O/A saileko hizkuntza dela ondorioztatzen du.

2.2.2 B. Hidalgo

Honek, dio aipatu berri ditugun azterketa estatistiko horiek, de Rijk-enak alegia, ez direla fidagarriak, ondoren datozen bi arrazoi hauengatik:

a) estatistika horiek egin ahal izateko de Rijk-ek erabili dituen baiezko esaldien kopurua, batez ere aztergai diren hiru osagaiak batera azaltzen direnekoa, oso txikia delako.
b) Aukeratu duen corpusa ez delako egokia. Altuberen legearen eragina duten testuez baliatu delako azterketa horiek egiteko.

Gauzak horrela, euskararen esaldien egitura zein den ezagutzeko, garai desberdinetako, gaur egungo, nola aurreko mendeetako idazleen eta hiztunen testuak oinarri dituzten corpusak aztertzera jo du Hidalgok. Hona hemen azterketa horietako batzuk[11]:

- XVII. mendeko bi prosa nagusi: Axular eta Tartas

	Axular		Tartas	
	Kopur.	%	Kopur.	%
SAO	98	48.0	108	44.3
SOA	52	25.5	87	35.7
OAS	25	12.2	26	10.7
ASO	20	9.8	9	3.7
OSA	6	2.9	12	4.9
AOS	3	1.5	2	0.8
Denera	204	%100	244	%100

Ikus daitekeen bezala, bi autore hauetan SAO ordena da nagusi, nahiz bietan oso altua den SOA esaldien portzentaia.

- XIX. mendeko lau autore aukeratu

	Moguel		J.B. Aguirre		Duvoisin		Pach. Cherren.	
	Kopur.	%	Kopur.	%	Kopur.	%	Kopur.	%
SAO	92	69.7	111	55.5	150	59.3	55	63.2
SOA	5	3.8	17	8.5	93	36.8	10	11.5
OAS	10	7.6	31	15.5	3	1.2	18	20.7
ASO	15	11.4	33	16.5	1	0.4	2	2.3
OSA	3	2.3	0	0.0	5	2.0	0	0.0
AOS	7	5.3	8	4.0	1	0.4	2	2.3
Denera	132	%100	200	%100	253	%100	87	%100

[11] B. Hidalgoren "Ohar estatistiko garrantzitsuak euskararen hitz ordenaren inguru. Euskara, S.V.O.? *FLV* (1995) lanetik hartuak.

Erabiltzen duten hizkuntza-ereduagatik kritika ona jaso duten lau autore hauen idatzietan ere, esaldi erdiak baino gehiago daude SAO hurrenkeran emanak. Duvoisinengan bakarrik suma liteke proportzio sendo bat SOA esaldiena (%36.8), Axular edo Tartasen antzekoa.

Aztertu dituen beste corpusetan ere antzekoak dira lortzen dituen emaitzak. Horien artean, Azkue aipatzea merezi du, bere lanak erkatu dituenean Altubetar iraultzaren eragina dela medio, nonbait, joskera estilo desberdinak aurkitu baititu.

Beraz, azterketa horietatik guztietatik atera dituen datuei esker, euskaran maiztasun handiena azaltzen duen hurrenkera S/A/O dela egiaztatu ahal izan du Hidalgok.

Horrela, zabaldutako joskera eredu azkue-altubetarra juzkatu nahian egiten dituen azterketa estatistiko hauetako emaitzak de Rijk-ek lortzen dituenekin bat ez datozela ondorioztatu du batetik, eta bestetik eztabaidatzea merezi duen gaitzat jo du hitz-hurrenkera hau.

2.2.3 Esku artean dugun lana

Gure aurrekariak diren azterketa hauek guztiak kontuan izanik, gure lanean corpus zabalago[12] bat, kazetaritza-corpusa, erabiliz esaldia osatzen duten hiru elementu horiek zein hurrenkeratan emanak datozen aztertuko dugu. Bertatik ateratzen ditugun emaitzak azaldu ditugun beste bi azterketa estatistiko horietako emaitzekin erkatuko ditugu.

Bideratu dugun azterketa honek eta hasieran aipaturiko tresna informatikoari esker, azterketa gehiago egiteko aukera emango digu; adibidez, aditzez aditz zenbait ondorio atera ahal izango ditugu eta aldi berean ondorio orokorrekin bat datozen ala alde nabaria dagoen ikusi, etab.

Dena dela, azterketa honi ekin baino lehen, hurrengo atalean erabiliko dugun corpusaren eta analizatzaile sintaktiko partzialaren deskribapena egingo dugu.

3. BALIABIDEAK

3.1 Corpusa

Esaldia osatzen duten elementuak zein hurrenkeratan ematen diren aztertzea du helburu nagusi ikerlan honek. Horretarako corpus zabala erabiliko dugu; beraz, ezer baino lehen corpusak duen garrantziaz jabetzea komeni da.

Testuek edo corpusek, benetan erabiltzen den hizkuntza idatziaren neurria ematen dute. Arrazoi horregatik, besteak beste, areagotu egin da egun beren erabilpena.

[12] zabalagoa da esaldi kopuruari dagokionez.

Esan bezala, testu-corpusak testu-masa handiak dira, eta gramatikarekin batera informazio linguistikoaren iturri direla aitortu behar da batetik, eta bestetik aplikazio eta tresna informatikoetarako probaleku ezinbestekoak, sistemen zehaztasuna neurtzeko.

Izan ere, gogoan izan behar dugu gure lana diziplina artekoa dela, bertan neurri batean behintzat, hizkuntzalaritza eta informatika uztartzen direlako. Hala, ondoren azalduko dugun analizatzaileari esker, eskatutako datuen erauzketa egin ahal izango dugu.

Azterketa honetan kazetaritza-corpusa erabili dugu; euskarri elektronikoan ditugun 1999ko urtarriletik 2000ko maiatza bitarteko *Euskaldunon Egunkaria*ren ale guztiak, zehazki. *Estilo Liburuan (2001)* zehazten den bezala EGUNKARIAn erabiltzen den hizkuntza ereduaren ezaugarriak hauek dira:

a) estandarra, euskara batua da hizkuntza eredua.
b) Argia. Horretarako hiru puntu hauetan oinarritzen dena:
   - Antolamendua. Berria ondo egituratu behar da, osotasuna eta koherentzia zainduz.
   - Laburtasuna eta sintaxi egokia. Testua modu arin eta ulergarrian idatzi behar da.
   - Lexiko zehatza eta argia. Zehaztasunik gabeko hitzetatik bezainbat urrundu behar da teknizismo ulergaitzetatik.

3.2 Analizatzaile sintaktiko partziala (PATR II)

Hizkuntzalaritza orokorrean bezalaxe, hizkuntzalaritza konputazionalean ere ikertzaileek hitzen arteko harremana eta perpausen egiturak konputazionalki lortzen saiatu dira. Hasieran espero ez bezain zaila suertatu da, ordea, erronka hori. Kasu guztietan aurre egin beharreko arazoa anbiguotasunarena baita.

Dena dela, testu errealetako perpaus edo esaldi oso-osoen analisia lortzeko asmorik gabe, hizkuntza teknologiako aplikazioak garatzeko (hizketa-sorkuntzan edo informazio bilaketan) baliagarriak izango diren tresnak garatu dira.

Euskarari dagokionean, hor ditugu, esaterako, sintaxi partziala[13] lantzeko taldean garatu diren Murrizpen Gramatika (Karlsson *et al.*, 1995) eta PATR II formalismoak (Shieber, 1986). Sintaxia aztertzeko bi bide eskaintzen dituzte mekanismo hauek, baina bigarrenak anbiguotasunari aurre egin ahal izateko lehenengotik zerbait jasotzen du.

Azterketa honetan bigarren mekanismo hori erabiliko dugu. Analizatzaile honen bitartez testu errealetako ia izen-sintagma eta adizlagun guztiak analiza daitezke; hau da, puntuazio markarik gabe honelako elementuek osatzen dituzten sekuentziak:

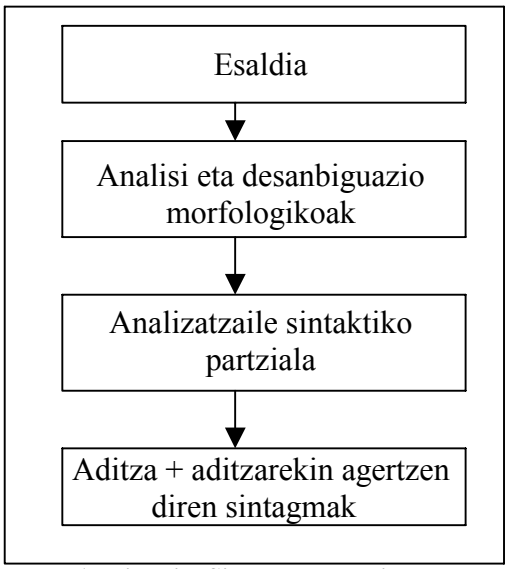
- aditza
- aditzarekin komunztadura egiten duten kasuak: ergatiboa, absolutiboa eta datiboa.
- postposizio-sintagmak

[13] Sintaxi partziala: azaleko egituretatik abiatuta zenbait erlazio sintaktiko adierazten ditu.

• adberbioak
• nominalizazioak (adibidez, *ekartze*)
• mendeko perpaus erlatibozkoak (adibidez, *gizonak ekarri duen*)
• mendeko perpaus konpletiboak (adibidez, *umeak ekarri duela*)
• mendeko perpaus moduzkoak (adibidez, *oinez nenbilela*)
• mendeko perpaus denborazkoak (adibidez, *kalean nengoenean*)
• zehar-galderak (adibidez, *nor etorri den*)

3.2.1 Prozesuaren deskribapena

1. irudian sistemaren arkitektura ikus daiteke, hainbat moduluren konbinazioa dena:



1. irudia. Sistemaren arkitektura

• Analisi eta desanbiguazio morfologikoak[14]

Aztergai den esaldia analizatzaile morfologikotik pasa ondoren, esaldi horretako hitz-forma bakoitzari bere interpretazio posibleak ematen zaizkio, ezaugarri morfosintaktikoen zerrenda baten bidez.
Ondoren, anbiguotasunaren ebazpena dator. Honen helburua da hasieran aukera posible guztiak ematen dituen prozesuko interpretazioak kentzeko erregelen bidez aukera zuzenarekin geratzea.

• Analisi sintaktiko partziala. Honek oinarrizko unitate sintaktiko guztiak (izen-sintagmak, adizlagunak eta zenbait mendeko perpaus) hartzen ditu.

[14] Aduriz, I. (2000) *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz*. Doktoretza-tesia. EHU/UPV.

Horrela, tresna informatiko honen bitartez, aditz batekin doazen osagaien kasua, lema eta numeroa atera daitezke izen-sintagma eta adizlagun bakoitzeko, eta mendekotasun mota mendeko esaldiekin, (1) adibidean agertzen den moduan.

Sarrera:	*Berezitasun gehiago ere baditu operak, zeluloidezko irudietan agertuko baitira une batez protagonistak*
Irteera:	*ine (irudi, m, p)* *agertu* *ins (une, s)* *abs (protagonista, m, p)*
	irudietan *agertuko baitira* *une batez* *protagonistak*

(1) adibidea. Esaldi batetik ateratako informazioa (*agertu* aditzaren azterketan)

Azterketarako baliatuko dugun kazetaritza-corpusa pasa dugu analizatzaile sintaktiko honetatik, eta bertan 1.400 aditzi dagozkion 111.000 esaldi (milioi eta erdi hitz) aztertu dira. Horrekin batera, gure analizatzaile automatikoa ebaluatu da.

4. HITZ-HURRENKERAREN AZTERKETA KAZETARITZA-CORPUSEAN

4.1 Zer aztertu

Aurreko atalean deskribatu dugun euskarazko analizatzaile sintaktiko partzialak, kazetaritza-corpus honetan jasotzen den informazioa eskuratzeko aukera ezin hobea eskaintzen digu. Orain arte egindako gramatika hau erabili dugu taldean, esaterako, aditzen azpikategorizazioari buruzko informazioa automatikoki lortzeko, testu idatzietako informazioa analizatuz, eta lortu diren emaitzak onak dira (Aldezabal *et al.*, 2001), nahiz eta oraindik jorratzeko bide luzea dugun.

Geure azterketa, ordea, corpus horretako 14.557 esaldietara mugatuko dugu. Esaldi diogunean, puntutik punturako tarteaz ari gara, eta tarte horren mugetan aditzari dagokion perpaus-eremua lortzeko prozedurak jarriko dira martxan (azpi-esaldiak, gure hitzetan). Horrela esaldi bateko aditzari dagozkion osagaiak analizatzerakoan, puntuazio markaz harantzago daudenak patroi horretatik kanpo geratuko dira. Honek, dena dela, ez dio baliotasunik kenduko sistemari.

14.557 esaldi hauetan, aditzarekin komunztadura egiten duten kasuei (ergatibo, absolutibo, datibo) dagokienez, hauek gauzatuta azaltzen dituzten esaldiak bakarrik hartu ditugu kontuan, beste osagaiak badituzten edo ez aintzat hartu gabe. Horretaz gain, baiezko esaldi nagusiak bakarrik izango ditugu aztergai.

Aldi berean zenbait aditz eskuz aztertuko ditugu analizatzaile sintaktikoa ebaluatzeko helburuarekin.

4.2 Emaitzak

4.2.1 Oro har

Corpusean ageri diren aditz guztiak eta hauekin komunztadura egiten duten deklinabideko hiru kasuak (absolutiboa, ergatiboa eta datiboa) kontuan hartuz, emaitza hauek lortu ditugu:

4.2.1.1 Aditzarekin komunztadura egiten duten hiru kasuen eta aditzaren, jokatu zein jokatugabearen arteko hurrenkera, maiztasun handienetik txikienera:

Hurrenkera	Esaldi kopurua	%
absolutiboa / aditza	10447	71.7
ergatiboa / aditza	1680	11.5
aditza / absolutiboa	796	5.4
ergatiboa / absolutiboa / aditza	412	2.8
datiboa / aditza	380	2.6
aditza / ergatiboa	299	2
datiboa / absolutiboa / aditza	110	0.7
...	...	...

Oro har aditzarekin komunztadura egiten dutenetako kasu bakarra eta aditza azaltzen dituzten esaldiak nagusitzen dira; hain zuzen, maiztasun handieneko hurrenkera 'abs./A'[15] modukoa da.

[15] Irakur bedi *absolutiboa / aditza.*

4.2.1.2 Osagaien arabera

	Hurrenkera[16]	**Kopuruak**	**%**
osagai bat / aditza	abs. / A	10447	71.7
	erg. / A	1680	11.5
	A / abs.	796	5.4
	dat. / A	380	2.6
	A / erg.	299	2
bi osagai / aditza	erg. / abs. / A	412	2.8
	dat. / abs. / A	110	0.75
	erg. / A / abs.	82	0.5
	abs. / erg. / A	80	0.5
	abs. / A / erg.	73	0.5
hiru osagai / aditza	erg. / dat. / abs. / A	5	0.03
	abs. / A / erg. / dat.	5	0.03
	erg. / dat. / A / abs.	2	0.01
	erg. / abs. / A / dat.	1	0.006
	erg. / A / abs. / dat.	1	0.006

Ikus daitekeenez, ez dira asko aditzarekin komunztadura egiten duten hiru kasuak eta aditza batera azaltzen dituzten esaldiak. Hortaz ez dirudi osagai hauei bakarrik begiratzea egokia denik hizkuntza baten hurrenkerari buruz ondorio argigarriak lortzeko.

4.2.1.3 Aditz laguntzailearen arabera

Aztertu ditugun 14.557 esaldi horietatik 5.823 dira aditz laguntzailea daramatenak. Horietan 'da' eta 'du' motako laguntzaileak dira nagusitzen direnak, eta bertan 'abs./A' eta 'erg./A' hurrenkerak. Honek badu bere arrazoia, era honetako testuetan gertaera edo berri bat adierazi behar denean, izen-abizenik ematen ez bada ere, iturria beti aipatu behar delako ahalik eta zehatzen, *Egunkariaren Estilo Liburuan* (2001) jasotzen den bezala.

[16] Zutabe honetan eta lanean zehar agertuko diren laburdura hauen irakurketa: A : aditza; abs.: absolutiboa; erg.: ergatiboa; eta dat.: datiboa.

Aditz laguntzailea	Hurrenkera	Kopurua	%
DA	abs. / A	3128	53
	A / abs.	473	8
ZAIO	dat. / A	75	1.2
	dat. / abs. / A	17	0.2
	abs. / A / dat.	8	0.1
DU	erg. / A	1145	19
	erg. / abs. / A	291	4.9
	A / erg.	287	4.9
DIO	erg. / dat. / A	21	0.3
	A / erg. / dat.	13	0.2
	abs. / A / erg. / dat.	5	0.08

4.2.1.4 Aditz jokatugabeak

8.734 esaldi dira aditz laguntzailerik gabe agertzen direnak. Hona zenbait hurrenkera osagai-kopurua kontuan izanik:

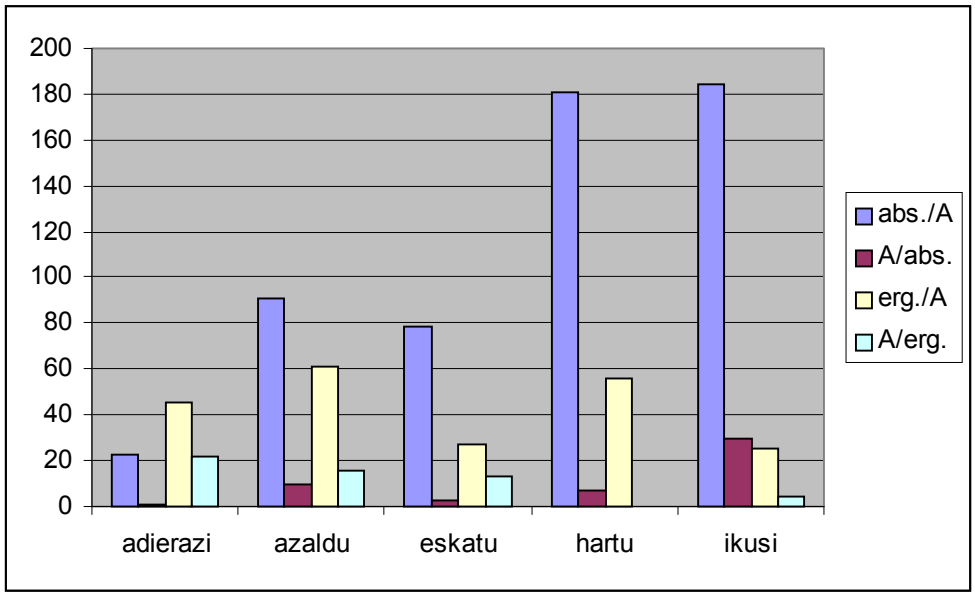
	Hurrenkera	Kopuruak	%
osagai bat / aditza	abs. / A	7259	83
	erg. / A	532	6
	dat. / A	304	3.4
bi osagai / aditza	erg. / abs. / A	119	1.3
	abs. / erg. / A	29	0.3
	abs. / dat. / A	27	0.3
hiru osagai / aditza	erg. / dat. / abs. / A	3	0.03

Aditz jokatugabeetan ateratzen diren emaitzak, aditz jokatutakoetan ateratakoekin erkatuz gero, azpimarra daiteke alde handirik ez dagoela datuen artean.

4.2.2 Aditzez aditz

Grafiko honetan, batetik corpusean maiztasun handienarekin agertzen diren lehen bost aditzak[17] (adierazi, azaldu, eskatu, hartu eta ikusi) erakusten dira eta bestetik, bostetan nagusitzen diren kasuak eta aipaturiko aditz hauekin batera gertatzen den hurrenkera.

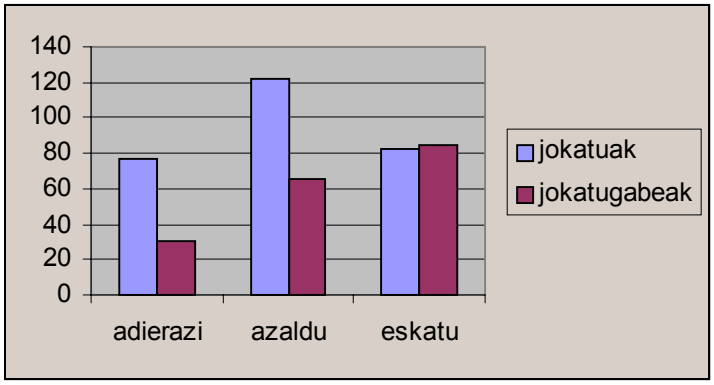
[17] aditz jokatuak nahiz jokatugabeak



Aditzez aditz egindako azterketa honetan ere 'abs./A' hurrenkera da nagusitzen dena, nabarmenki gainera. Salbuespena 'adierazi' aditza dugu, 'erg./A' hurrenkerarekin. Aditzaren kokalekuari erreparatuz gero, gehienetan kasuaren atzetik agertzen da, beraz eta emaitza orokorrekin erkatuz, ez da alde handiegirik nabari.

4.2.2.1 'adierazi', 'azaldu' eta 'eskatu' aditzen azterketa zehatza

Aztergai ditugun hiru aditz hauetan gailentzen diren osagaiak zein diren eta esaldiaren barruan nola antolatuta agertzen diren emango dugu jakitera. Azterketa hori bitan banatuko dugu, batetik aditza jokatuta duten esaldiak hartuko ditugu eta bestetik, jokatugabekoak. Ondoren datorren irudiak esaldien banaketa hori isladatzen du.



Ikus daitekeen bezala, 'adierazi' eta 'azaldu' aditzetan, aditza jokatuta daramaten esaldiak dira nagusi; eta 'eskatu' aditzean, aldiz, jokatugabeak, gutxigatik bada ere.

	adierazi		**azaldu**		**eskatu**	
	Kop.	%	Kop.	%	Kop.	%
Jokatuak	77	71.9	122	64.8	82	49.1
jokatugabeak	30	28.03	66	35.1	85	50.8

Jarraian esaldi horietan nabarmentzen diren hurrenkerak ditugu aditzez aditz:

	Adierazi			**azaldu**			**eskatu**		
Jokat.	hurrenkera	Kop.	%	hurrenkera	kop.	%	hurrenkera	kop.	%
DA	-	-	-	abs./A	36	29.5	abs./A	17	20.7
DU	erg./A	43	55.8	erg./A	50	40.9	erg./A	19	23.1
ZAIO	-	-	-	-	-	-	dat./A	2	2.4
DIO	erg./dat./A	2	2.5	A/dat./erg.	1	0.8	A/erg./dat.	4	4.8
jokag	abs./A	22	73	abs./A	50	75	abs./A	62	72.9

4.3. Ebaluazioa

Ebaluazioak, nola ez, garrantzi handia du LNPko aplikazioen alorrean.

Corpusean maiztasun handienarekin agertzen diren lehen hiru aditzak /adierazi/ /azaldu/ eta /eskatu/ eskuz aztertu ditugu analizatzailea ebaluatzeko helburuarekin. Eskuz eta automatikoki lortutako emaitzak alderatu ondoren lortzen den doitasuna %81.7koa da.

Sistemaren fidagarritasuna neurtu nahian ateratako emaitza hauek adierazgarriak dira, errore kopuruak txikiak direlako. Erroreen proportzioak ez dio kentzen sinesgarritasuna sistemari; beraz fidagarria da.

5. AZTERKETA ESTATISTIKOEN ERKAKETA

5.1 Euskarak SOA hurrenkera du?

5.1.1 Hidalgoren azterketa estatistikoak *versus* de Rijk-enak

Hidalgok euskararen SOA hitz-hurrenkera hori benetan hala den egiaztatze aldera zenbait azterketa estatistiko egin ondoren, eginda zeudenekin erkatu ditu eta bere azterketatik atera dituen ondorioak adierazgarriak dira, orain arteko uste hori zalantzan jartzeraino iritsi baita.

Horrela, lehenik, de Rijk-ek aztertutako corpus bera aztertu du[18] eta Greenberg eta de Rijk-ek proposatutako zenbaketa-sistema bera erabiliz atera dituen ondorioak de Rijk-ek ateratakoen antzekoak dira, nahiz eta esaldien kopurua zertxobait aldatzen den.

	de Rijk	**B. Hidalgo**
SOA	138	115
SAO	48	53
OAS	11	17
OSA	5	5
ASO	6	8
AOS	1	0
Guztira:	209	198

Ondoren, aurreko azterketa hori egiteko erabilitako testu-corpusak fidagarriak ez direlako, garaian indarrean zegoen Altuberen legeen eragina zutelako-edo, beste testu-corpus batzuk aukeratu eta beste azterketa bat egin du. Hemen[19] Greenberg eta de Rijk-ek erabilitako irizpide berberaz baliatuko da, jarraian emaitzak konparatu ahal izateko. Horrela, berak lortutako emaitzak de Rij-ek lortutakoekin bat ez datozela aditzera eman du:

	de Rijk	**Hidalgo**
SOA	%57	%23.4
SAO	%30	%55.3

Beraz, corpusa aldatzeak garrantzi handia izan du.

5.1.2 Hidalgoren azterketa estatistikoak *versus* kazetaritza-corpusa

Gure lanean, Hidalgok erabili dituen irizpide berberei jarraituz, corpus zabalago bat aztertzeari ekingo diogu, ondoren bi lan horiek erkatzeko.

Horrela, hasierako 14.557 esaldi horietatik 5.639 esaldi baino ez dugu hartu, Hidalgoren lanarekiko alderaketa osoa egiteko datiboa eta aditz jokatugabeak alde batera utziz. Hala, euskararen hitz hurrenkera SOA den ala ez frogatze aldera, ergatiboa, absolutiboa eta aditza elementuak zein hurrenkeratan emanak datozen bost mila eta koska esaldi horietan aztertu ditugu.

[18] *FLV*, 1995, 401-420 or.
[19] B. Hidalgo : "Hitz ordenaren estatistikak euskaraz" *ASJU*, XXXIII-2, 1999.

Emaitzak:

Kazetaritza-corpusean aditza amaieran agertzen deneko esaldiak dira nagusitzen direnak: 4671 esaldi (% 82.8); eta horietatik gauzatzen diren hiru elementu hauek kontuan hartuz, SOA hurrenkera da nagusi.

	Kazetaritza-corpusa	
	kopurua	**%**
SOA	291	56.8
SAO	76	14.8
OAS	71	13.8
OSA	51	9.9
AOS	17	3.3
ASO	6	1.1
Guztira:	512	100

Erreferentzia modura dauzkagun azterketa estatistikoekin erkatuz gero:

	de Rijk	**Hidalgo**	**Kazetaritza-corpusa**
SOA	%57	%23.4	%56.8
SAO	%30	%55.3	%14.8

Datu hauek guztiak erkatu ondoren, kazetaritza-corpusean ateratzen diren emaitzak Hidalgok ematen dituenarekin bat ez datozela egiazta daiteke batetik, eta bestetik SOA hurrenkera dela nagusitzen dena. Horrek esan nahi du Altuberen eragina nabaria dela oraindik ere.

5.2 Aditzaren kokalekua esaldian.

Hidalgok erabilgarri dituen testu-corpusaz baliatuz, aditzaren gunea esaldian zein den aztertu du, euskal aditzak esaldia amaitzeko joera duen ala ez egiaztatzeko. Azterketa horretatik kanpo geratu dira aditzik gabeko esaldiak, aditz jokaturik gabekoak, eta aditz jokatua izanagatik, osagarririk gabekoak. Baita ere, galdera, agintera eta ezezko esaldiak.

Corpusak dituen 4.681 esaldietatik, aurrizki baieztatzailea duten 193 esaldi aditz trinkodunak alde batera utzi eta 4.488 esaldi baliatu ditu.

Emaitzak:

	Hidalgo	
	Kopurua	**%**
Baiezko esaldi nagusiak	4488	100
Aditza hasieran	640	14.3
Aditza tartean	2586	57.6
Aditza amaieran	1262	28.1

Guk, une honetan, tresna informatikoak dituen mugak direla-eta, absolutiboa, ergatiboa eta aditz jokatua (konbinazio guztiak direlarik posible) gauzatuta azaltzen dituzten esaldiak bakarrik kontuan hartuta, honako emaitza hauek atera ditugu:

	Kazetaritza-corpusa	
	Kopurua	**%**
Baiezko esaldi nagusiak	5639	100
Aditza hasieran	804	14.2
Aditza tartean	165	2.9
Aditza amaieran	4670	82.8

Aditza amaieran agertzeak ez du esan nahi esaldia aditzarekin amaitzen denik, honen ondoren ergatiboa eta absolutiboa ez den beste osagairen bat edo batzuk ager baitaitezke. Beraz, honekin adierazi nahi da esaldia osatzen duten eta aztergai ditugun ergatibo, absolutibo eta aditza, hiru elementu hauen hurrenkera aztertzerakoan, aditza bi horien atzetik joango dela.

6. ONDORIOAK

Ikerlan honetan euskararen hitz-hurrenkera zein den zehazteko ezagutzen diren eta eskuz burutuak izan diren lehen azterketa estatistikoez jardun dugu. Eskuz egindako lan hauen mugak (aztertutako esaldi kopurua txikiagoa, antzeko fenomenoak modu desberdinean tratatzeko arriskua, ...) aintzat harturik, corpus zabalago batean gertatzen diren egiturak eta hurrenkerak automatikoki aztertzeari ekin diogu. Era honetan burutu den lehen azterketa izango da hau ziurrenik ere.

Betebehar horiei buru egiteko, esaldi bakunak analizatzeko egokia den eta izen-sintagmak eta adizlagunak tratatzeko estaldura ia osoa duen euskarazko analizatzaile sintaktiko konputazionalaz baliatu gara eta testu libretik, kazetaritza-corpus batetik zehazki, era askotako informazioa erauzteko aukera izan dugu. Egindako azterketa apal honen bitartez eskuzko emaitzak eta automatikoak erkatu ditugu eta horrek aukera ematen digu esateko, taldean garatutako analizatzaile sintaktiko hau hizkuntzalarientzat

baliagarria izango dela etorkizuneko euskararen gramatika oso baten hazi gisa, edo beste azterketetarako corpusak eta adibideak biltzeko.

Bestalde, lan honetan, beste askotan bezalaxe, corpusak duen garrantzia nabarmenduko genuke. Hizkuntzaren baitan gertatzen denaren ikuspegi gero eta sakonago eta zuzenagoa izateko, eta ezagutza hori beste hainbat aplikazio-eremuetara eraman ahal izateko, ezinbestekoak baitira ondo eraikitako corpusak, behar bezala kodetuak eta etiketatuak.

Lan honen helburu nagusiari helduz, berriz, baliatu ditugun tresna eta errekurtso horiei esker kazetaritza-corpusean nagusitzen den hurrenkera SOA dela egiaztatu dugu eta lortutako emaitza hauek Hidalgoren aurreikuspenarekin bat datozela ondorioztatu. Hau da, Azkue/Altubek XX. mende hasieran asmatutako edo ezarritako legeek duten eragina oso nabarmena da oraindik ere. Beraz, badirudi Hidalgok dioen moduan "gure gaurko estandarrean" beste inongo, inoizko eta inolako euskaldunek inoiz hitz egin edo idatzi ez duten bezala idazten dela[20].

[20] Hidalgo, B. (1998) "Baina, zer da euskal joskera?" *Administrazioa euskaraz*, 21. IVAP/HAEE

ERREFERENTZIAK

Atal honetan, ikerlanean zehar erreferentziatu diren obrak aurkituko dira.

Aduriz, I. (2000) *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz.* Doktoretza-tesia, Euskal Filologia Saila, EHU/UPV.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K. & Goenaga P. (2001) *Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus.* Actas del XVII Congreso de la SEPLN Universidad de Jaén, septiembre de 2001.

Altube, S. (1929) *Erderismos*. Euskera. X urtea I-IV zenbakia.

Azkue, R.M. (1888) "Grankanton arrantsaleak". *Euskera*, XXXIII, 1988, 379-89.

_____, (1894) *Ensayo Práctico.*(1896ko, *Método Práctico*-aren aurreko eskuizkribu argitaragabea, Euskaltzaindiako, Azkue bibliotekan).

Egunkaria (2001) *Estilo liburua*

Gojenola, K. (2000) *Euskararen sintaxi konputazionalerantz*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, EHU/UPV.

Greenberg, J.H. (1963) "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements". In: J.H. Greenberg (ed.), *Universals of Language*, 58-90, Cambridge (Mass), MIT Press. (2nd. ed., 1966, 73-113).

Hidalgo, V. (1995) *Hitzen ordena euskaraz*. Doktoretza-tesia, EHU, Euskal Filologia, Gasteiz.

_____, (1995) '*Ohar estatistiko garrantzitsuak euskararen hitz ordenaren inguru. Euskara, SVO?'*. FLV, 70, 1995, 401-420.

_____, (1998) "Baina, zer da euskal joskera?" in *Administrazioa euskaraz*, 21 IVAP/HAEE.

_____, (1999) *'Hitz ordenaren estatistikak euskaraz'*. Anuario del Seminario Julio de Urquijo, 1999-2, 393-451.

Karlsson F., Voutilainen A., Heikkila J. & Anttila A. (1995) *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Mitxelena, L. (1953) *Arnaut Oihenart*. BRSVAP, 1953, 445-463. (Orain in *Mitxelenaren Euskal Idazlan Guztiak* -MEIG-, V, 35-47).

_____, (1978) "Miscelánea filológica vasca I". *FLV*, 1978, 205-228.

_____, (1987) *Palabras y textos*, EHU, 1987, 363-385.

Osa, E. (1990) *Euskararen hitzordena komunikazio zereginaren arabera*. Doktoretza-tesia. EHUko Argitalpen Zerbitzua.

Rijk, R.P.G. de, (1969) “Is basque an SOV language?”. *FLV*, 3, 1969, 319-351.

Shieber S.M. (1986) *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, 4 zenbakia, Stanford.