CORPUSEN ETIKETATZE LINGUISTIKOA

I. Aldezabal, M. J. Aranzabe, A. Diaz de Ilarraza, A. Estarrona, N. Ezeiza, L. Uria
(IXA Taldea, Universidad del País Vasco/Euskal Herriko Unibertsitatea)

Abstract

In this article, we shall comment on the steps that have to be taken to give a linguistic label to a corpus and the difficulties that appear in this process. Our main objective was to highlight the importance of the labelling when preparing a corpus that is useful for linguistic research, and the need to establish criteria and to take the decisions that this entails. We also explain how semi-automatic methods are applied and how the manual revision that guarantees the quality of the corpus is carried out. Once the corpus has been revised and labelled, it will be useful both for carrying out linguistic analyses and for improving or assessing the linguistic tools and resources, and also for channelling automatic study.

1. Sarrera

Garai batean arrotz samarra zitzaigun *corpus* hitza, egun, ezagun bilakatu da orotariko hizkuntza-aztertzaileen artean. Areago, hizkuntza-aztertzaileon aldarrikapen orokorra da *erreferentziazko euskal corpus* baten beharra, hasi Euskaltzaindia (Oyharçabal 2004, Sagarna 2007) eta UZEItik (Urkia 2008) hizkuntza-teknologien alorrean gabiltzanera (Ixa taldea & Elhuyar Fundazioa 2007).

Esan dezagun, baina, laburki, zer den corpus bat eta zer den erreferentziazkoa izatea.

Leech-en (1997) hitzak hartu ohi dira corpusaren definizioaren aitzindari:

> a body of language material which exists in electronic form, and which may be processed by a computer for various purposes such as linguistic research and language engineering.

Alegia, corpus bat, izatez, testu-bilduma huts batek osa badezake ere, berez adierazten duena da euskarri elektronikoan dagoen hizkuntzaren zati handi bat, hizkuntzaren azterketara eta hizkuntza-teknologiak garatzera bideratuta.

Hizkuntzaren zati hau, gainera, ez da edonolakoa izatea komeni, baizik eta, esan dugun bezala, erreferentziazkoa. Urkiak (2008), EAGLEsen[1] estandarrek ematen duten definizioan oinarrituta, ederki adierazten du zer den erreferentziazko corpusa:

> Hizkuntza, bere osotasunean hartuta, erakusteko diseinatuta egon behar du corpusak: hizkuntzaren aldaera esanguratsuak adierazteko besteko tamaina eta kalitatea behar du.

[1] EAGLES (Expert Advisory Group on Language Engineering Standards): <http://ww.ilc.pi.cnr.it/EAGLES96/>.

Hizkuntzaren aldaera esanguratsuak, corpusaren tamaina eta corpusaren kalitatea aipatzen dira hitzotan. Baina zehatz dezagun apur bat zer esan nahi duen horrek, eta azpimarra dezagun benetan zerk ematen dion corpus bati benetako balioa. Aldaera esanguratsuak emango dizkigu delako hizkuntzan ekoitzi diren orotariko testuak, gaiak eta erregistroak dituen bilduma batek. Tamainari dagokionez, esan ohi da milioi bat hitzeko corpusa dela azterketetarako gutxienekoa. Eta kalitateari dagokionez, testu-bilduma jasoa izan den moduak du garrantzia (zuzenketa ortografikoak, adibidez), baina batez ere *corpusaren etiketatzeak*. Azken honek ematen dio, hain zuzen, corpus bati benetako balioa. Jo dezagun, berriz ere, Leech-en hitzetara:

> Corpus annotation is widely accepted as a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development.

Etiketatu beharrak berekin dakar aldez aurretik corpusa prestatu beharra. Testuak ondo kodetuta egon behar du: esanahi bat duten formatuzko adierazleak modu esplizitu batean ezarri behar dira —paragrafo hasiera-bukaerak, letra lodia/etzana…, akats ortografikoak identifikatuta…—. Honetarako lengoaia estandar aberats bat erabiltzeak berebiziko garrantzia du, alde batetik, ondorengo aukerak zabaltzen dituelako (kontsulta-sistema, datu-erauzketa…) eta bestetik, hizkuntzen arteko konparaketak eta azterketak ahalbidetzen dituelako. Eta, egun, egokien jotzen den estandarra XML-TEI lengoaia da (Artola et al. 2002). Honenbestez, testuaren aldez aurreko kodeketa eta etiketatze linguistikoa eskutik doaz, eta honi ere merezitako garrantzia eman behar zaio.

Behin testua kodetuta, gure aztergai izango diren elementu linguistikoak bereizi behar dira, eta horretarako, testua *token*etan banatu behar da; hots, zuriunetik zuriunera doan karaktere-segida oro banan-banan jarri behar da. Honi *tokenizazioa* deitzen zaio eta edozein etiketatze linguistikoren aurreko prozesua da (Ezeiza 2002); horrexegatik deitzen zaio aurreprozesua. Artikuluan, maila linguistikoak aipatzerakoan erabiliko dugun kontzeptua izango da.

Bestalde, etiketatze linguistikoaren prozesu osoan baliabide eta tresna konputazionalak erabiltzen dira (datu-base lexikalak, analizatzaile morfologikoak, analizatzaile sintaktikoak, etab.). Baliabide linguistikoak ezagutza linguistikoaren bidez eraikitzen dira eta tresnek, horiek baliatuz, prozesua automatizatzen dute, horrela etiketatze-lana erruz arinduz. Zenbaitetan, testuetatik ikasitako datu estatistiko hutsak ere erabiltzen dira (Ezeiza 2002), modu guztiz automatikoan. Etiketatze-lanaren zati bat eskuz eta bestea automatikoki egiteari *etiketatze erdi-automatikoa* esaten zaio. Ondoren, eta corpusak kalitatea izango badu, eskuzko orrazketa beti da beharrezkoa, egon litezkeen akats eta hutsuneak osatzeko eta, era berean, baliabide eta tresna horiek aberasteko eta ebaluatzeko. Honek agerian uzten du etiketatze linguistikoak dakarren lanaren tamaina, eta corpus bat garatzerakoan atazarik garrantzitsuenetakoa izanik, gutxik ematen diote behar duen garrantzia. Horretara dator, hain zuzen, hemen aurkezten dugun lana: corpus bat etiketatzeak duen garrantzia azpimarratzera, baina batez ere zeregin horrek ekartzen dituen zailtasunak, beharrezko irizpide/erabakiak eta etorkizunerako beharrak plazaratzera.

Ikus ditzagun, bukatzeko, euskaraz zein corpus ditugun eta zein ezaugarri dituzten. Horretarako Elhuyarren (Areta et al. 2008) zerrenda gurera ekarriko dugu eta zenbait daturekin eguneratu (Aranzabe 2008). Horretaz gain, hemen idatzizkoak bakarrik zerrendatuko ditugu:

Corpusa	Egilea	Data	Mota	Asmoa	Tamaina	Etiketatzea
Orotariko Euskal Hiztegiaren testu corpusa (OEHTC)	Euskaltzaindia	1984-2005	orokorra	deskriptiboa	6 M hitz	ez
XX. mendeko euskararen corpus estatistikoa (XXMECE)	Euskaltzaindia UZEI	2002	orokorra orekatua lagindua	deskriptiboa	4,6 M hitz	automatikoa eta eskuz: SGML
Ereduzko Prosa gaur (EPG)	EHU eta Donostiako Udala	2007	orokorra	ereduzkoa	25,1 M hitz	lema eta ezaugarri morfologiko batzuk automatikoki
Zientzia eta Teknologiaren Corpusa (ZTC)	IXA Taldea (EHU) eta Elhuyar Fundazioa	2002-2008	berezia (zientzia eta teknologia), orekatua, lagindua	deskriptiboa	bertsio berrienerako 8,5 M hitz; hortik 2 M inguru orekatuak	egiturazkoa eta linguistikoa: automatikoki eta eskuz. XML TEI-P4
Klasikoen Gordailua (KG)	Susa literatura argitaletxea		berezia: euskal literatura klasikoa	deskriptiboa	11,9 M hitz	lematizatua?? TEI
Ibinagabeitia Proiektua (IP)	Susa literatura-argitaletxea	2000-2004	berezia: literatura aldizkarien gordailua	deskriptiboa	451 aldizkari ale; 7.949 artikulu	linguistikoa: lematizazioa
Euskararen Prozesamendurako Erreferentziazko Corpusa (EPEC)	IXA Taldea (EHU)	Maila linguistikoaren arabera	orokorra XXMECEko eta *Egunkari*ko testu-zatiak	deskriptiboa	300.000 hitz	Morfologikoa Sintaktikoa Semantikoa: Izenak (Euskal Wordnet); rol tematikoak, oraintsu hasita Pragmatika (testu-zati batzuetan): anafora, diskurtso-markatzaileak

1. taula

Euskaraz dauden idatzizko corpusak eta euren ezaugarriak

Ikusten denez, hauetako corpus batzuk gordinak dira eta beste batzuk informazio linguistikoarekin hornituak. Horietatik Ixa taldea garatzen ari den *Euskararen Prozesamendurako Erreferentzia Corpusa* (EPEC) da maila morfologikotik harantz etiketatuta dagoen bakarra. Eta guk artikulu honetan, hain zuzen, berau hartuko dugu langai, edozein corpus etiketatzerakoan izan dezakegun problematikaren eredu. Aztertuko dugu nola dagoen osatua, zein maila bereizten ditugun, eta etiketatzerakoan topatu ditugun fenomeno batzuen aurrean nola jokatu dugun. Zehazki, eta lana mugatzearren, maila guztietan eragina izan duten hitz anitzeko adierazpideak nola etiketatzen joan garen erakutsiko dugu.

Artikuluko atalak, honenbestez, honelaxe antolatu ditugu: 2. atalean EPEC corpusa deskribatu dugu, 3. atalean corpus hau prozesatzeko jarraitzen ditugun urratsak azaldu ditugu, 4. atalean etiketatzerakoan izan ditugun arazoak, eta 5. atalean, ondorio gisa, dauden etorkizuneko lan eta beharrak.

2. EPEC corpusa

EPEC corpusa euskara estandarrean idatzitako 300.000 hitzek osatzen duten testu-bilduma da. Testu-bilduma honen zati bat *XX. mendeko euskararen corpus estatistiko*tik hartu da, eta beste bat *Euskaldunon Egunkaria*tik.

XX. mendeko euskararen corpus estatistikoa urtetan euskararen erreferentzia-corpustzat hartu da eta 4.658.036 testu-hitzek osatzen dute. UZEI Terminologia eta Lexikografia zentroak (http://www.uzei.com) egin du. Corpus honek XX. mendean argitaraturiko euskal testuak biltzen ditu, eta testuok garaia, euskalkia eta testu-mota irizpideen arabera sailkatuta daude. Hala, lau garai nagusitan banatzen dira (1900-1939, 1940-1968, 1969-1990, 1991-1999), sei euskalki desberdin (bizkaiera, gipuzkera, zuberera, lapurtera-nafarrera, euskara batua, eta sailkatu gabeak) eta hamalau testu-mota (saio-artikuluak, administrazio-idazkiak, ikasliburuak, saioliburuak, literatur prosa, poesia, antzerkia, bertsoak, ikerketa-lanak, haur- eta gazteliteratura, ahozko jardunen transkripzioak, liturgia, egunkariak eta aldizkariak). Liburu eta aldizkarietako artikulu bakoitzak egileari (edo egileei) buruzko informazioa du eta euren izenburua. Corpus honetatik 48.000 testu-hitz hartu dira EPEC osatzeko; zehazki, euskara batuan idatzitako azken garaiko testuak hartu dira, 1991-1999 bitartekoak, alegia.

EPEC corpusaren bigarren zatia *Euskaldunon Egunkaria*ren 1999ko urtarriletik 2000ko maiatza bitarteko ale guztiek osatzen dute. Hauek ere euskara batuan idatzitakoak dira.

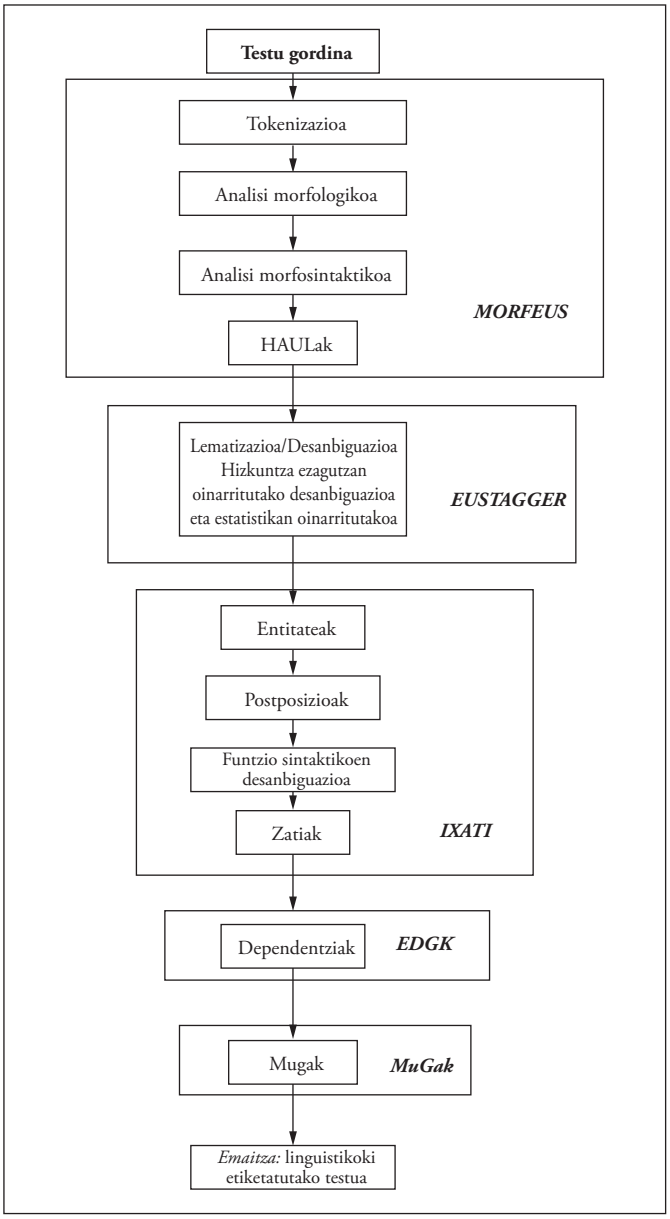
Urkiak (2008) esan bezala, erreferentzia-corpusak hizkuntza bati buruzko informaziorik osatuena eman behar du; gure kasuan, EPEC erreferentzia-corpusa dela diogunean, euskararen prozesamendurako hainbat tresna garatzeko eta hobetzeko erabiltzen den eta erreferentziatzat dugun corpusa dela ulertu behar da.

3. EPEC corpusaren kodeketa eta mailaketa

EPEC linguistikoki etiketatzen ari garen corpusa da; hots, corpuseko hitzak informazio linguistikoaz hornitzen ari gara. Horniketa horretarako Ixa taldean garatu diren tresnak erabiltzen ditugu: *Morfeus* (Alegria et al. 1997), *Eustagger* (Aldezabal et

al. 2007a), *Ixati* (Alegria et al. 2006, Aduriz et al. 2006a, Aduriz et al. 2008), *EDGK* (Aranzabe 2008) eta *MuGak* (Aduriz et al. 2006b).

1. irudian dago ikusgai corpusei ezartzen zaien prozesatze linguistikoaren eskema:



1. irudia

Etiketatze-prozesuaren eskema

Labur esanda, testu baten analisi-prozesuan honako urrats hauek egiten dira:

— Aurreprozesua edo tokenizazioa: sarrerako testu batean puntuazio-markak, zenbakiak, laburtzapenak eta antzeko beste edozein karaktere hitz ortografikoetatik bereizten dira; hots, analisi morfologikoan sarrera gisa erabiliko diren unitateak bereizten dira.
— Analisi edo segmentazio morfologikoa: hitz barruko osagaiak banatu eta bakoitzari dagokion informazio morfologikoa (eta hainbat kasutan sintaktikoa ere) eransten zaio; hau da, testuko hitza dagokion lema-, kategoria-, azpikategoria-, kasu- eta numero-informazioaz hornitzen da. Hitzen segmentazio-analisi hau *Euskararen Datu-Base Lexikaleko* (EDBL) informazioa (Aldezabal et al. 2001) baliatuta egiten da.

Hona, 2. irudian, segmentazioak ematen duen irteera:

ikuskizunekiko

ikuskizun [[Sarrera_ikuskizun--0][IZE][ARR][BIZ_-] + [ekin[DEK] [MUG] [P] [SOZ] [FS1_@ADLG]] + [ko[DEK] [MUG] [S] [GEL] [FS1_@IZLG>] [FS2_@<IZLG]]

2. irudia

ikuskizunekiko hitzaren analisia segmentazio mailan

— Analisi morfosintaktikoa: analisi morfologikotik hurrengo urratsetarako informazio esanguratsua *goratzen* edo hautatzen da (nahiz eta informazio guztia beste leku batean ere gordetzen den). Adibidez, 2. irudiko analisian bi deklinabide-kasu (SOZ eta GEL) eta mugatasun (P eta S) ditugu eta horietatik bakarra (azkena) *goratzen* da. Morfosintaxiaren helburua da, hain zuzen, eremu bakoitzeko balio bakar bat izatea.
— Hitz Anitzeko Unitate Lexikalen (HAUL) tratamendua: hitz-forma solteen analisitik haratago, elkarren mendeko diren hitz-konbinazio edo hitz anitzeko unitate lexikalak ezagutzen dira urrats honetan. Esate baterako, testuan *hala eta guztiz ere* aurkitzean, ezin ditugu hitzak independenteki interpretatu, elkarren ondoan daudenean beren funtzioa aldatzen baita.
— Lematizazioa/Desanbiguazioa: hitz batek dituen analisi aukeretatik zuzena markatzen da (okerrak baztertuz), testuinguruari erreparatuta. Desanbiguazioa, batetik, hizkuntza-ezagutzan oinarritutako murriztapen-gramatika (MG) baten bidez (Aduriz 2000) egiten da eta, bestetik, estatistikan oinarrituta (Ezeiza 2002).
— Entitateak: Hitz Anitzeko Unitate Lexikalen antzera, denbora-esapideak, zenbakizko esapideak eta izen bereziak (pertsona-, toki- eta erakunde-izenak) ezagutzen dira urrats honetan. Adibidez, *2004ko urtarrilaren 7an*, *berrehun eta hogeita zazpi*, *Juan Jose Ibarretxe*.
— Postposizio-lokuzioak: atzizki-postposizioak eta elementu beregainak osatzen duten multzo osoa unitate bakartzat hartu ondoren, horien ezagutza egiten da urrats honetan. Adibidez, *Lanez kanpoko harremanak dituzte* esaldiko *-z*

kanpoko postposizio-lokuzioa non hasten den eta non bukatzen den zehazten da.

— Funtzio sintaktikoen desanbiguazioa: anbiguoak diren etiketa sintaktikoen desanbiguazioa gauzatzen da MG gramatikako erregela sintaktikoen bitartez.

— Zatiak: elkarrekin sintaktikoki erlazionaturik dauden hitz multzoak atzematen dira. Esaterako, *Pentsamenduak iraungi egin ziren oinotsen aurrean* esaldian hiru zati hauek ezagutuko dira: i) *pentsamenduak*, ii) *iraungi egin ziren,* eta iii) *oinotsen aurrean*. Analisi hau *analisi sintaktiko partziala* izenarekin ere ezagutzen da.

— Dependentziak: esaldia osatzen duten hitzen arteko lotura gauzatzen da. Hitz hauek binaka lotuz, esaldiaren dependentzia-zuhaitza lortzen da. Halaber, zuhaitz hauek esaldi baten egitura, hau da, esaldiko hitzen arteko lotura, hierarkikoki adierazten dute. Honi, Hizkuntzalaritza Konputazionalean *analisi sintaktiko osoa* deritzo.

— Mugak: perpausen arteko amaiera-mugak esleitzen dira MG bitartez. Adibidez, *Gorka lanean dago eta Jon harantz doa* esaldian, bi perpaus ditugu. MuGak (Aduriz et al. 2006b) esaldi honetako bi perpausak bereizten ditu, *eta* bien arteko mugarria dela adieraziz.

Hemendik aurrerako mailak ez daude prozesamendu-kate honen barruan oraindik. Hala, aparteko modulu gisa tratatzen dira oraingoz.

Maila semantikoari dagokionez, batetik, corpuseko izenak *Euskal WordNet*-eko adierekin etiketatu dira (Pociello 2008), eta, bestetik, egun, rol tematikoen mailan etiketatzen ari gara (Aldezabal 2007b).

Diskurtsoaren tratamenduari dagokionez, esaldia gaindituz, paragrafoa eta paragrafoan agertzen diren elementuen arteko erlazioa aztertzen ari gara. Zehazki, anafora pronominalaren azterketari (Ceberio et al. 2008) eta diskurtsoko markatzaileak ezagutzeari ekin zaio (Iruskieta et al. 2008).

Esan dugun bezala, prozesu guztia aplikatu ondoren, eskuzko orrazketa beharrezkoa da kalitatea bermatzeko, baina orrazketa hau prozesuaren edozein puntutan lortzen den informazioaren gainean egin daiteke (Aduriz et al. 2004, Aduriz et al. 2006a). Maila bakoitzean nola aritu jakiteko, behar-beharrezkoa da etiketatze-eskuliburuak egitea, gida-lerro moduan. Horixe egin dugu, hain zuzen, orain arte landutako mailekin (Agirre et al. 2005, Aduriz et al. 2006b, Aldezabal et al. 2007a, Aldezabal et al. 2007c, Aduriz et al. 2008). Barne-txosten horietan ageri dira zehatzmehatz kasu orokor zein partikular bakoitzerako emandako irtenbideak. Eta tesi-lan ugariren muina ere izan dira (http://ixa.si.ehu.es/Ixa), ez bakarrik eskuzko orrazketarako, baita, eta batik bat, tratamendu automatikorako ere.

4. Etiketatze linguistikoan izandako arazoak

Etiketatze lanari ekiterakoan ez dira gutxi izan topatutako arazoak eta hartu beharreko erabakiak.

Hasteko, eta prozesamendu erdi-automatikoari dagokionez, urrats hauek guztiek zein *hurrenkera* hartu beharko luketen ez dago erabat finkaturik oraindik, ez guztietan behintzat. Esaterako, postposizio-lokuzioak etiketatzeko analisi morfosintaktikoaren emaitza hartzen da abiapuntutzat. Horrezaz gain, desanbiguazio morfo-

logikoa (Aduriz 2000) gauzatu ondoren aplikatzen da gramatika, hau da, esaldiko hitz bakoitzeko analisi bakarra (zuzena behar lukeena) lortutakoan.

Halaber, dependentzien bidez gauzatzen den analisi sintaktiko osoak ezinbestekoa du postposizio-lokuzioak egoki detektatuak izatea. Hala ez balitz, zenbait postposizio-lokuzioren osagaiak adabegi desberdinetan azalduko lirateke. Ostera, dependentzia-zuhaitza zuzena izan dadin, postposizio-lokuzio osoak adabegi berean egon behar du, eta hori bakarrik da posible postposizioak aurretiaz markatuz gero.

Edota perpausen amaiera-mugak ezagutzeko corpuseko esaldien dependentzia-zuhaitzak eginak baleude, hainbat errore edo hutsune konponduko lirateke. Baina egia da, era berean, esaldia osatzen duten hitzen arteko dependentzia bidezko loturak hobetzeko mugak zeintzuk diren jakitea garrantzitsua dela. Zein izan daiteke, beraz, analisi katearen hurrenkerarik egokiena? Posible al da gramatika hauek elkar elikatzea? Areago, modulu batzuk behin baino gehiagotan aplika daitezke (zenbait hitz, desanbiguatu aurretik eta beste zenbait, desanbiguatu ondoren, adibidez). Honako hau sakonki aztertu beharreko gaia da.

Hasiera-hasierako beste buruhauste bat izan da sarrera lexikal guztiak barne hartuko dituen *etiketa-sistema* bat aukeratzea, eta behin sistema aukeratuta, sarrerak modu koherente batean sailkatzea. Izan ere, jo dezagun gure etiketa-sisteman ez dugula onomatopeia (*danba*…) kategoriatzat hartu, eta ezta postposizio-lokuzioko hitz beregaina ere (*aurre, atze, at*…). Zein kategoria da, bada, egokiena horrelako elementu linguistikoak sailkatzeko?

Nola sailkatu hiztegietan kategoria argirik ez duten elementuak (*are*, *harik*)?

Edota hitz bat berez izen gisa erabiltzen bada oro har, baina tarteka adjektiboaren lekuan ere ager badaiteke (*teknikari* ***mekanikaria***), zer egin: bi sarrera eman lexikoan, edo sarrera bakarra eta gero sintaxiko erregelen bitartez nolabait konpondu?

Jakina denez, adjektiboak izenik gabe ere ager daitezke (*(ume) txikiek egiten dituzte horrelako gauzak*). Horrelakoetan zer egin? Bi sarrera eman (izen eta adjektiboa bezala); sintaxian erregelak egin adieraziz adjektiboek bakarrik ere izen sintagma bat osatzeko gaitasuna dutela nahiz eta izenik ez egon eta izen hori elidi tutzat jo?

Kategoriatik harantz, sintaxian gaudela, zein izango da gure azterketa-eremua? Esaldia? Eta zer da esaldia? Nola mugatu behar dugu zehatz-mehatz esaldia?

Behin esaldiaren mugak definitu ditugula, zein elementu linguistiko etiketatuko ditugu: bakarrik agerian daudenak ala elidi tutakoak ere bai? Hots, *Etorri da* moduko esaldi batean, *da*-k islatzen duenari jarraiki, HURA moduko zerbait agerian jarri eta etiketatuko dugu? Edo hori beste pauso eta maila baterako utziko dugu?

Puntuazio-markei dagokienez, esanguratsuak direnak (*eta* juntagailuaren funtzioa betetzen duen koma, esaterako) etiketatzera mugatuko gara ala guztiak markatuko ditugu? Nola?

Etiketatze sintaktikoa soilik gauzatuko da ala sintaxiarekin batera semantika eta diskurtso-egiturak ere etiketatuko dira?

Nola jokatu behar da corpusean agertzen diren esaldi okerrekin?

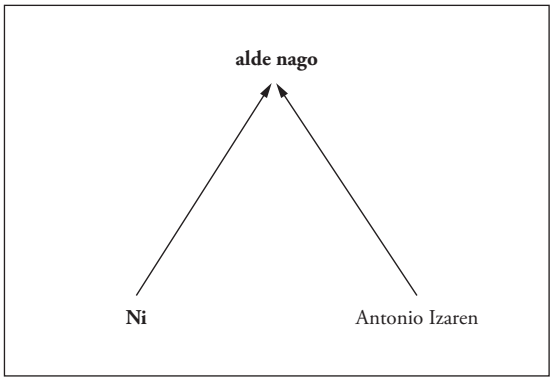
Baina gure lana mugatze aldera, esan bezala, jo dezagun kate osoan zehar eragina duten egitura adierazgarri batzuk aztertzera: hitz anitzeko adierazpideak (entitateak, postposizio-lokuzioak eta HAULak). Hauek, sintaxiari begira, unitatetzat hartu ohi dira eta horretarako eman ditugu elementu horiek analizatzeko urratsak. Beraz, hitz

anitzeko unitatetzat jotzen direnak sintaxiari begira ere osorik lantzea izan da gure irizpide nagusia.

Hala eta guztiz ere, arazoak sortzen dira, alde batetik, horrelakoak aldi berean suertatzen direnean, bestetik, aditzak direnean eta, azkenik, gure tresnek ezagutzen ez dituztenean:

(1) Ni Antonio Izaren alde nago

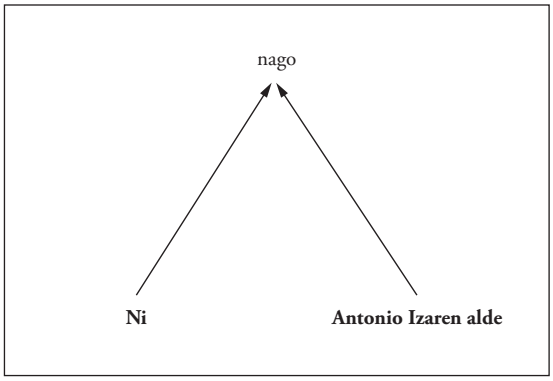
(1) adibidean dauzkagu: entitate bat (*Antonio Iza*), postposizio-lokuzio bat (*-en alde*) eta HAUL bat (*alde egon*). Horrelakoetan, nola jokatu behar da sintaxia aztertzeari begira: lehendabizi HAULa unitatetzat hartu eta gero zuhaitza eraiki (3. irudia)?



3. irudia

Ni Antonio Izaren alde nago esaldiaren dependentzia-zuhaitza I

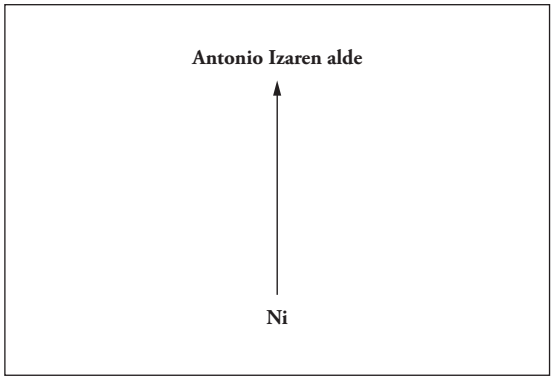
Edo lehendabizi postposizioa unitatetzat hartu eta zuhaitza eraiki (4. irudia)?



4. irudia

Ni Antonio Izaren alde nago esaldiaren dependentzia-zuhaitza II

Edo biak batera unitate konplexutzat hartu eta horrekin zuhaitza eraiki (5 irudia)?



5. irudia

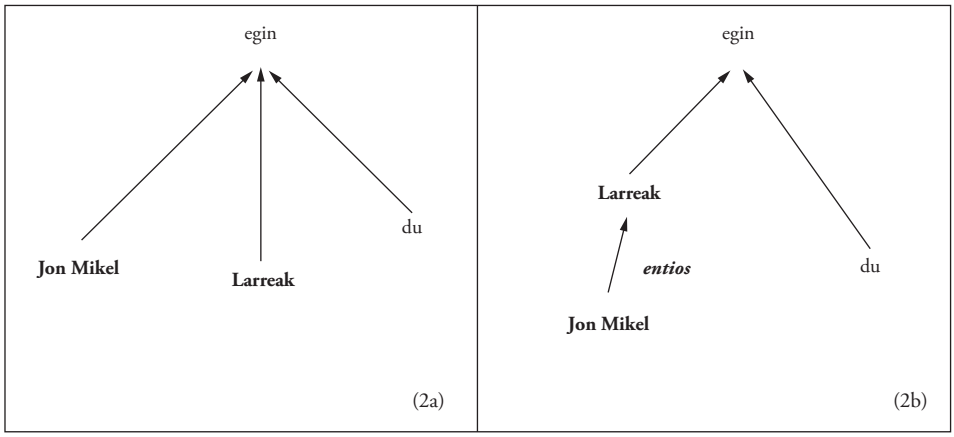
Ni Antonio Izaren alde nago esaldiaren dependentzia-zuhaitza III

Hartzen den erabakia hartzen dela, esaldiaren zuhaitza modu batera edo bestera eraiki eta interpretatzen dugu. Horrelakoen aurrean gure erabakia izan da lehendabizi HAULa kontsideratzea eta honek mendekotasunen bat badu unitatetik kanpo (***-en*** *alde egon*), genitiboari aditz konposatuaren modifikatzaile funtzioa jartzea. Hala, (1) adibideko esaldian, *alde egon* aditzaren mende dagoen *Izaren* hitza modifikatzailea da (3. irudia).

Aditz HAULen kasuan, osagarriren bat inplikatuta egon ohi da (***lo*** *egin*, ***suak*** *hartu*…). HAUL oso gisa analizatuz gero, aditzarekiko komunztadura galdu egiten da: laguntzaile iragankorra hartzen dutelarik, objektua/subjektua ez da inondik ageri. Kasu hauetan gure erabakia izan da guztiak unitate moduan analizatzea eta aditzaren barruan dagoen osagaiaren funtzioa aditzean bertan kodetuta agertzea.

Aztergai dugun (1) adibidean, 3., 4. eta 5. irudietako zuhaitzetan ageri den moduan, *Antonio Iza* osorik ezagutu da. Gerta daiteke, ordea, gure tresnek ez ezagutzea eta banaturik agertzea. Hots, nahi genukeen analisia ez izatea. Honek eskatzen du orrazketa-prozesuaren barruan aukera edukitzea guk nahi dugun analisi hori txertatzeko. Honek, aldi berean, hutsune horiek beteko duten etiketa lagungarriak sortzea eskatzen du.

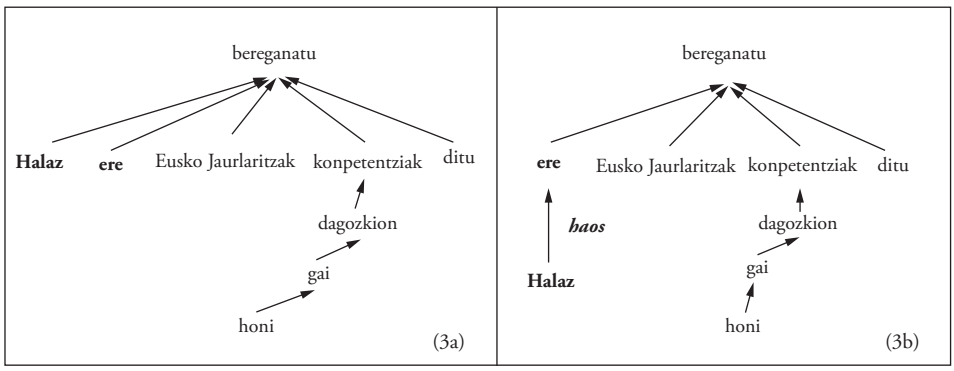
Esaterako, (2) *Jon Mikel Larreak egin du* adibidean, entitateen ezagutzaileak *Jon Mikel* entitate bat bezala analizatu du, eta *Larrea* beste entitate bat bezala. Emaitza hori abiapuntutzat hartuta, 6. irudiko (2a) zuhaitza izango genuke, eta hau ez litzateke bat etorriko erabakitakoarekin. Hori dela eta, *entios* izeneko etiketa sortu, *Jon Mikel*-ekin lotu eta orrazketaren ondoren, 6. irudiko (2b) zuhaitza eraikiko litzateke.



6. irudia

Jon Mikel Larreak egin du esaldiaren dependentzia-zuhaitza

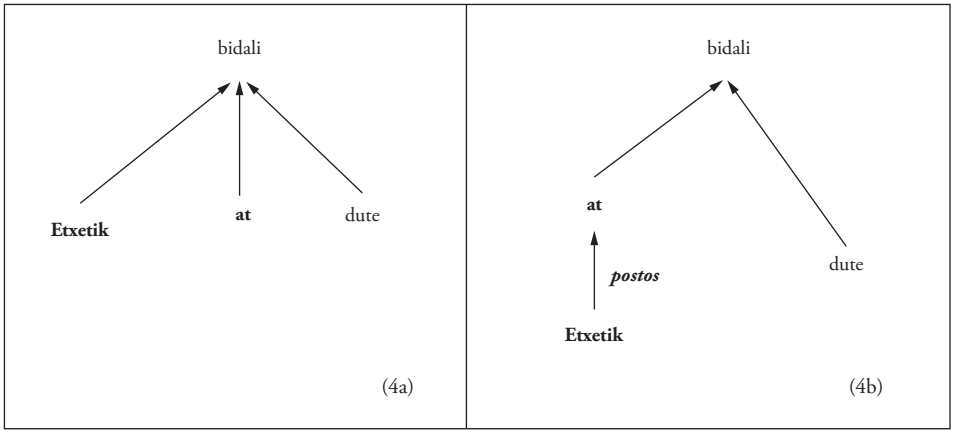
Ildo beretik, HAULa banatuta badator, adibidez, 7. irudiko (3a) analisiko *Halaz ere*, eta gure ustez (etiketatze egokiagoa lortzeko) lotuta egon behar badu, *haos* etiketaren bidez bateratu egingo da. Horren emaitza da 7. irudiko (3b) analisia.



7. irudia

Halaz ere Eusko Jaurlaritzak gai honi dagozkion konpetentziak bereganatu ditu esaldiaren dependentzia-zuhaitza

Postposizio-lokuzioen kasuan ere, *postos* etiketaren bidez solte datorren osagai hori postposizio-lokuzioari dagokiona dela adierazi nahi da. Horrelakoetan orain arteko egituren antzera jokatu da (4) *Etxetik at bidali dute* esaldiari dagokion 8. irudian ikus daitekeen moduan.



8. irudia

Etxetik at bidali dute esaldiaren dependentzia-zuhaitza

Edozein etiketatze-prozesutan horrelako kasuak aurreikusi eta irtenbide bat ematea behar-beharrezkoa da. Dena den, kontuan hartu behar da askotan erabakia ez dugula hartu soilik ikuspegi linguistikotik, baizik eta ondoren egingo den analisi automatikoa ere aintzat hartu dugula.

5. Ondorioak

Aipatuak ditugu corpusen zertarakoak: hizkuntzaren azterketa eta hizkuntza-teknologien garapena. Adibide konkretuetara joaz, corpus linguistikoki etiketatuak dagoeneko erabiliak izan dira perpaus erlatibo motak aztertzeko (egun garabidean), ikasleen akatsak (komunztadura, postposizioak…) detektatzeko (Oronoz 2009), egitura sintaktikoen maiztasunak ateratzeko (Agirre et al. 2009). Azken batean, aztertu nahi den fenomeno linguistikoari begira etiketatu behar da corpusa, eta hori aztertzea oso inportantea da, horrek baldintzatzen baitu corpusa diseinatzeko eta etiketatzeko modua.

Bestalde, ikuspegi konputazionaletik, ikasketa automatikorako ere erabili izan da, hala nola, kategoria-desanbiguaziorako (Ezeiza 2002), mugak esleitzeko (Alegria et al. 2008), analisi sintaktikoa egiteko (Bengoetxea & Gojenola 2007).

Linguistikoki etiketatutako corpusen beharra ezin ukatuzkoa da, eta baita honek berarekin dakarren lana ere: denbora, giza baliabide asko eta tresneria konputazionalen garapena. Horrelako proiektu bat aurrera aterako bada, babes ekonomikoa eta elkarlana ezinbestekoak dira, eta zalantzarik gabe, euren balioa areagotu egingo da guztion eskura jarrita, alegia, guztiok erabiltzeko moduan.

Benetako erreferentziazko euskal corpus batetik urrun samar bagaude ere, pauso handiak eman ditugu eta beste hizkuntza landuagoen bide beretik goaz, edo, bederen, saiatzen gara. Horretantxe ahalegindu beharko ginateke aurrerantzean ere.

Bibliografia

Aduriz, I., 2000, *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak*. Doktoretza-tesia, Filologia eta Geografia-Historia Fakultatea. UPV/EHU, Gasteiz.

—, Aranzabe, M. J., Arriola, J. M., Díaz de Ilarraza, A., Gojenola, K., Oronoz, M. & Uria, L., 2004, «A Cascaded Syntactic Analyser for Basque», *Computational Linguistics and Intelligent Text Processing,* Springer, Berlin, 124-135.

—, —, Arriola, J. M., Atutxa, A., Díaz de Illarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A. & Urizar, R., 2006a, «Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing», in A. Wilson, P. Rayson, D. Archer (arg.), *Corpus Linguistics Around the World* (Language and Computers 56), Netherlands: Rodopi, 1-15.

—, Arrieta, B., Arriola, J. M., Díaz de Ilarraza, A., Izagirre, E. & Ondarra, A., 2006b, *Muga Gramatikaren Optimizazioa*. Barne-txostena. UPV/EHU/LSI/TR 09-2006.

—, Aldasoro, E., Aldezabal, I., Aranzabe, M. J., Arriola, J. M., Ceberio, K., Estarrona, A., Iruskieta, M., Lersundi, M., Pociello, E., Uria, L. & Urizar, R., 2008, *Euskarazko postposizio-lokuzioen tratamendu konputazionala*. Barne-txostena. UPV/EHU/LSI/TR 07-2008.

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E. & Quintian, M., 2005, *EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editatze-, etiketatze- eta epaitze-lanak*. Barne-txostena. UPV/EHU/LSI/TR 23-2005.

—, Atutxa, A., Labaka, G., Lersundi, M., Mayor, A. & Sarasola, K., 2009, «Use of rich linguistic information to translate prepositions and grammar cases to Basque», *XIII Conference of the European Association for Machine Translation EAMT 2009,* 58-65, Barcelona.

Aldezabal, I., 2007, «Estudio preliminar para la creación de Euskal Propbank», in I. Castellón, A. Fernández (arg.), *Perspectivas de análisis de la unidad verbal,* SERES, Barcelona.

—, Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G. & Lersundi, M., 2001, «EDBL: a General Lexical Basis for the Automatic Processing of Basque», *Proceedings of the IRCS Workshop on Linguistic Databases,* Philadelphia (AEB).

—, Ceberio, K., Esparza, I., Estarrona, A., Etxeberria, J., Iruskieta, M., Izagirre, E. & Uria, L., 2007a, *EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua*. Barne-txostena. UPV/EHU/LSI/TR 11-2007, 01-45.

—, Aranzabe, M. J., Arriola, J. M., Díaz de Ilarraza, A., Estarrona, A., Fernández, K., Iruskieta, M. & Uria, L., 2007c, *EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua*. Barne-txostena. UPV/EHU / LSI / TR 12-2007.

Alegria, I., Artola, X. & Sarasola, K., 1997, «Improving a Robust Morphological Analyser using Lexical Transducers», in R. Mitkov, N. Nicolov (arg.), *Recent Advances in Natural Language Processing. Current Issues in Linguistic Theory (CILT) series* 136, John Benjamins, Amsterdam, 97-110.

—, Arregi, O., Ezeiza, N. & Fernandez, I., 2006, «Lessons from the Development of a Named Entity Recognizer», *Procesamiento del Lenguaje Natural* 36, 25-37.

—, Arrieta, B., Carreras, X., Díaz de Ilarraza, A. & Uria, L., 2008, «Chunk and Clause Identification for Basque by Filtering and Ranking with Perceptrons», *Revista del Procesamiento del Lenguaje Natural* 41, 5-12.

Aranzabe, M. J., 2008, *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitzbankua eta gramatika konputazionala*. Doktoretza-tesia. Euskal Filologia Saila, UPV/EHU, Leioa.

Areta, N., Gurrutxaga, A. & Leturia, I., 2008, «Begiratu bat corpus-baliabideei», *BAT Soziolinguistika aldizkaria* 62.

Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Hernández, G. & Soroa, A., 2002, «A Class Library for the Integration of NLP Tools: Definition and implementation of an Abstract Data Type Collection for the manipulation of SGML documents in a context of stand-off linguistic annotation», *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

Bengoetxea, K. & Gojenola, K., 2007, «Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera», *XXIII Congreso de la SEPLN,* Universidad de Sevilla, Sevilla.

Ceberio, K., Aduriz, I., Díaz de Ilarraza, A. & García, I., 2008, «Erreferentziakidetasunaren azterketa eta anotazioa euskarazko corpus batean», in X. Artiagoitia & J. A. Lakarra (arg.), *Gramatika Jaietan. P. Goenagaren omenez*, *ASJU*-ren Gehigarriak LI, UPV/EHU, Bilbo.

Ezeiza, N., 2002, *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua.* Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Iruskieta, M., Díaz de Ilarraza, A. & Lersundi, M., 2008, «Análisis de los marcadores del discurso para el euskera: denominación, clases, relaciones semánticas y tipos de ambigüedad», *XXVI Congreso Internacional de AESLA*, Almería.

Ixa taldea & Elhuyar Fundazioa, 2007, «Testu-corpusak: ezaugarriak, eraketa eta tresnak», in M. Jose Arrieta (koord.), *Hizkuntza, komunikazioaren eta teknologiaren garaian*. IVAP Herri Arduralaritzaren Euskal Erakundea, Vitoria-Gasteiz.

Leech, G., 1997, «Introducing Corpus Annotation», in R. Garside, G. Leech & T. McEnery (arg.), *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Longman, 1-18.

Oronoz, M., 2009, *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura.* Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Oyharçabal, B., 2004, «Euskaltzaindiaren corpusez», *Euskera* 49-1, 43-55.

Pociello, E., 2008, *Euskararen ezagutza-base lexikala: Euskal WordNet.* Doktoretza-tesia, Euskal Filologia Saila, UPV/EHU, Leioa.

Sagarna, A., 2007, «Euskara eta informazioaren teknologiak. Egungo egoeratik etorkizunera begira», *Euskera* 52-3, 843-858.

Urkia, M., 2008, «Euskararen erreferentzia-corpusaren beharraz» in *Euskalgintza XXI. mendeari buruz. XV. Biltzarra. Azkue eta Urkixoren omenez*, *Bilbon, 2001-9-17/19an*. Iker saila, 19, 307-312.