ERREFERENTZIAKIDETASUNAREN AZTERKETA ETA ANOTAZIOA EUSKARAZKO CORPUS BATEAN

Klara Ceberio
Ixa Taldea, UPV/EHU

Itziar Aduriz
Universitat de Barcelona

Arantza Diaz de Ilarraza
UPV/EHU

Inés M. Garcia Azkoaga
UPV/EHU

1. Sarrera

Ia hogeita hamar urte igaro dira Patxi Goenagak *Gramatika Bideetan* liburua argitara eman zuenetik, euskarak ofizialtasuna bereganatu zuen beste urte, eta denbora luze honetan euskarak aurrera egin du indartzearen, estandarizazioaren eta normalizazioaren bidean; poliki-poliki, lehen mugatuak zituen esparruetako ateak zabaldu zaizkio gizartean, besteak beste, hedabideetan, heziketan, unibertsitatean, eta zientzia eta teknikaren arlo askotan.

Era berean, zientziak eta teknikak sekulako aurrerapausoak eman ditu azken urteotan, eta horrekin batera erabilera eta azterketa esparru berriak sortu dira euskararentzat hizkuntzak eta teknologiak bat egiten duten eremuetan.

Hizkuntzaren erabiltzaileak gara, baina hizkuntzaz baliatzeko modu asko daude, eta informazio eta komunikazio teknologiari esker lehen pentsaezinak ziren bitartekoak ditugu euskara idatziari zein ahozkoari tarte berriak eskaintzeko eta gure hizkuntza mundu zabalean ibilarazteko.

Garai batean ahotsa grabatzeko erabiltzen genituen magnetofoi eta kasetek lekua utzi diete bitarteko digital eta informatikoei; telefonoz hitz egiteko dagoeneko ez dugu zertan egon hari bati lotuta, are gehiago, zenbaitetan telefonoz deitu eta gizaki baten antzean hitz egiten digun makina batek erantzuten digu; konputagailuen laguntzaz idazten dugu, eta konputagailuak ahotsaren bidez lan eginaraztea dugu amets; hizkuntzen arteko mugak apurtu nahi ditu gizakiak itzulpen automatikoaren bitartez, etab. Horren guztiaren ondorioz bide ugari hartu du hizkuntzalaritzak, eta gaur egun gauden jakintza eta informazioaren gizarte honetan, hizkuntzak modu naturalean egin du bat teknologiarekin. Azken batean, hizkuntza eta teknologia ez daude hain urrun, eta Lengoaia Naturalaren Prozesamendua (LNP) saiatzen da bi arloak uztartzen.

Hizkuntzalaritza eta informatikak bat egiten dute makina bati hitz egiten/hizkuntza ulertzen erakusteko orduan. Hizkuntzalariak tratatzen ari den hizkuntzako datuak

[1] Argitalpen hau EHUk babestutako 1/UPV 00033.130-H-15286/2003 ikerketa-proiektuaren barnean kokatzen da.

bildu, aztertu, xehatu eta programek erabiltzeko moduan deskribatu eta antolatzen ditu. Hau da, informazio linguistikoa automatikoki prozesatzeko moduan jartzen du. Informatikariak dira programen egileak, informazio linguistikoa erabiltzen dutenak testua prozesatzeko, testua ulertuko duten analizatzaileak lortzeko helburuarekin ala aplikazioak egiteko helburuarekin. Esate baterako, hizkuntzalariak erregela bidezko gramatika sintaktiko bat osatzen badu, informatikoak informazio hori prozesatuko lukeen analizatzaile sintaktikoa egingo luke, eta aplikazio bat eratu, zuzentzaile sintaktiko bat, adibidez.

Alegia, adituek aldez aurretik prestatu eta antolatutako informazioari esker makinak eta gizakiak hizkuntzaren bitartez elkarreragin dezakete. Elkarreragiketa horiek hizkuntzaren maila guztietan gerta daitezke, hala nola, morfologikoan, sintaktikoan, semantikoan ala testualean. Maila hauetan guztietan analisi osoa eta sendoa egin nahi bada, nahitaezko baldintza da elementu oro deskribatua egotea. Hala, morfologikoan postposizioen deskripzioa egina eta lexikoaren kategoria eta azpikategoriak emanak izatea, adibidez. Sintaktikoan, egitura sintaktikoak ezagutuak izatea, sintagmetatik hasita esaldietaraino, eta helburuaren arabera, mendekotasun-erlazioak azaleratuak. Semantikoan, adieren desanbiguazioarekin batera, paper tematikoen eta predikatuen berri izatea. Testuaren mailan sare erreferentziala aztertua izatea; izan ere, testu bat ekoizten dugunean erabiltzen dugun diskurtsoan ageri diren zenbait zantzuren bitartez gaiak harilkatzen dira eta informazioak aurrera egiten du, horri esker irakurleak edo entzuleak interpretatu egin dezake zeri buruz hitz egiten edo idazten ari garen. Gaien harilkatze hori anaforen eta sare erreferentzialen bitartez egiten da.

Euskarazko Lengoaia Naturalaren Prozesamenduaren arloan hamaika lan egin dira (eta egiten ari dira) batez ere maila morfologikoan eta sintaktikoan (Urkia 1997) (Gojenola 2000), maila semantikoko azterketa gaur egun pil-pilean dagoen gaia da eta euskaraz ere oraintxe lantzen ari den gaia da (Aldezabal 2004). Hutsune handiena arlo honetan testu mailako lanetan dago, ez baita euskaraz inongo azterketarik gai honi buruz, eta ondorioz, ez dago corpusik sare erreferentziala etiketatua duenik.

Hutsune hori betetzera dator lan hau. Orain dela bizpahiru urte EHUren barruan Euskal Filologia saila eta IXA taldearen artean hasitako lanen helburuak, ondorioak eta birplanteamenduen berri ematera gatozenak.

Artikulu honetan lehendabizi erreferentziakidetasunaren eta anaforaren arteko harremanaz arituko gara. Ondoren, hizkuntzaren prozesamendu automatikoaren esparruan anafora eta erreferentziakidetasuna markatzeko ikerketan zertan diren azalduko dugu labur. Jarraian etiketatzeko dauden tresnak eta egin ditugun aukerak azalduko ditugu. Gero, lan egingo dugun corpusaren ezaugarriez arituko gara eta lan honen ardatza izan den erreferentziakide eta anaforiko diren adierazpideak nola etiketatzen den azalduko dugu hurrengo puntuan. Azkenik, lan honen ondorioak eta etorkizuneko lanak aipatuko ditugu.

2. Erreferentziakidetasuna eta anafora

Erreferentziakidetasunaz hitz egiteak edo termino hori mugatu nahi izateak ezinbestean eramaten gaitu, besteak beste, anaforaren eta erreferentziakidetasunaren artean dagoen harremana tratatzera.

Garbi dago erreferente jakin bat izendapen ezberdinen bitartez berreskura daitekeela testuan zehar, adibidez:

(1) Abuztuaren 17an gorpu bat topatu zuten, deskonposizio egoeran Lodosako (Nafarroa) ubidean. 33 urteko emakume espainiar batena zela jakinarazi zuten ikerketa iturriek. *Emakume hori* Raquel Sanchez Suñen zen eta Calahorrakoa zen, Errioxa erkidegokoa. Gorpua topatu zutenean, deskonposizio egoeran zegoen, eta ADN frogak egin behar izan zizkioten. Hilabete lehenago desagertu zen *errioxarra bere* etxetik. *Raquel Sanchez Suñen*-en senideak eta polizia aspaldi ari ziren *haren* bila. [*Berria*-tik (2007-09-08) moldatua]

(2) ... Emakume hori → izena determinatzaile erakuslearekin
... Gorpua → izena artikuluarekin
... Errioxarra → jentilizioa
... bere → izenordaina
... Raquel Sanchez Suñen → izen berezia
... haren → erakuslea

Adierazpide horien artean erreferentziazko lotura dago erlazio semantikoa dagoelako, eta horretaz gain, harreman anaforikoa dago aurrekariarekiko, *emakume hori*, *gorpua*, *errioxarra*, *bere*, *Raquel Sanchez Suñen*, *haren* adierazpideen erreferentzia zentzua argitzeko sorburuan dagoen *gorpu bat* adierazpidera jo behar dugulako.

Aurrekari baten eta hura izendapen ezberdinarekin berreskuratzen duen adierazpide erreferentziakideren baten artean harreman anaforikoa egon daiteke, baina horrek ez du esan nahi anafora eta erreferentziakidetasuna beti parekideak direnik; hor daude, esate baterako, asoziazio bidezko anaforak, zero anaforak, izen berezien bitartez egiten diren berrartzeak, edo erreferente ebolutiboen kasuak. Halaber, erreferentziakidetasuna eta harreman anaforikoa bat datozen kasuetan ere, anafora uztartuen eta askeen (diskurtsiboen) artean bereizi beharko dugu.

Asoziazio bidezko anaforaren kasuan, adibidez, aurrekariaren eta adierazpide anaforikoaren arteko harremana ez da oinarritzen erreferentziakidetasunean, elementu biek baitute erreferente ezberdina:

(3) *Herri batera* iritsi ginen. *Eliza* tontor batean zegoen

Adibide klasiko horretan (3) ikusten den bezala, aurrekariak (*herri bat*) eta anaforak (*eliza*) erreferente ezberdina dute, adierazpide horietako bakoitzak errealitate ezberdin bat izendatzen baitu. Hala eta guztiz ere, harremanetan dauden bi adierazpide ditugu, bigarrenaren interpretazioa lehendabizikoaren bitartez egin behar dugulako, inferentziazko arrazonamenduaren bitartez; horregatik, bien artean dagoen harreman anaforikoa asoziazioaren bidezkoa dela esaten dugu.

Beste bi adibide hauetan ere nabaria da erreferentziakidetasun eza:

(4) *Istripu bat* egon zen... *Anbulantzia* iritsi zenean...
(5) Pablok *bost lehoi* hil ditu eta nik *hiru* (Kleiber 1994)

Istripuak erreferente bat du eta *anbulantziak* beste bat. Hurrengo adibidean (5) aipatzen diren lehoiak ere ez dira erreferentziakideak, lehoi ezberdinak baitira. Pablok hil dituen lehoiak eta nik hildakoak ez dira berberak. Identitate bakarra izendatzeko

moduan dago, kasu bietan izen berdina erabiltzen delako. Erreferentziakidetasun birtuala esaten dio Milnerrek (1982) horri. (3) eta (4) adibideetan asoziazio bidezko anafora baten aurrean egongo ginateke eta (5) adibidearen kasuan anafora lexikal edo nominal baten aurrean (Kleiber 1994).

*Zero anafora*ren kasuan ere zalantzazkoa izan daiteke erreferentziakidetasuna dagoen edo harremana anaforikoa den.

(6) *Mirenek sagarrak* bildu ditu. Bihar merkatura eramango ditu Ø Ø

Anafora berrartze erreferentzialean oinarrituta dagoela abiapuntu gisa hartuz gero, ezingo genuke kokatu elipsia fenomeno anaforikoen barruan, baina ikuspegi zabalago batetik, anafora baten aurrean egongo ginateke, elipsiaren interpretazioa hizkuntza-testu inguruan agertzen den beste adierazpide baten bitartez egin behar dugulako (Charolles 1991). Ikuspegi honetatik elipsiak ere harreman anaforikoaren adierazle izan daitezke.

(7) "Jakingo duzunez, *ura* ez da beti puru-purua izaten; hainbat hondakin eduki ditzake disoluzioan edo partikula solidoak eraman. Batzuetan Ø oso zikina izan daiteke (euri zaparrada baten ondorioz, putzu batean, etab.) eta beste batzuetan badirudi Ø garden edo garbi dagoela". (*Natur Zientziak*. "Ostadar" Proiektua, Elkar-G.I.E., DBH-1, 1996, 85. or.; García Azkoagan, 1999)

Era berean, litekeena da adierazpideak erreferentziakide izatea baina haien artean harreman anaforikorik ez egotea. Hori gertatzen da, hain zuzen**,** *izen berezien* kasuan:

(8) *Mikel* eta *Andoni* Gasteizko jaietara joan dira. *Mikel* goiz itzuli da etxera baina *Andoni* ez da agertu oraindik.

Bigarrenez aipatuta agertzen diren *Mikel* eta *Andoni* eta hasierakoak berdinak dira. Erreferentziakideak dira, baina lehendabiziko aipamenaren eta bigarrenaren artean ez dago harreman anaforikorik. Bigarren agerpeneko *Mikel* edo *Andoni* zuzenean interpretatzen ditugu, lehendabizikoak bezala; izan ere, izen berezia erreferentea zuzenean izendatzeko modua da.

Beste adibide honetan ere harreman anaforikoa zalantzazkoa litzateke:

(9) *Ibarretxek* bere agintaldiaren urte bukaerako lehendabiziko diskurtsoa irakurri zuen. *EEAko lehendakariak* esan zuenez..

Adierazpide biak erreferentziakideak dira erreferente berbera dutelako, baina era berean, *Ibarretxe* eta *EEAko lehendakariak* adierazpideak beregainak dira eta zuzenean interpreta ditzakegu. *Lehendakaria* hitzak pertsona jakin bati egiten dio erreferentzia eta dugun jakintza komunari esker ez dugu informazio gehiagorik behar *Ibarretxe* izenarekin lotzeko, hau da, lehendakaria zein den jakiteko ez da beharrezkoa aurrekarira jotzea. Hortaz, bigarren adierazpidea, EEAko lehendakariak, ez dugunez interpretatzen lehendabizikoaren bitartez, bien arteko harremana ez litzateke anaforikoa izango. Kleiber-en (1988) esanetan, anafora identifikatzeko orduan testuinguru linguistikoaren irizpidea hartzen badugu kontuan, ezin dira onartu anaforikoak ez diren erreferentziakidetasunak, alegia, berez interpreta daitezkeen adierazpidez osatuta dau-

denak; horrenbestez, adibide horretan autonomoak diren bi adierazpideren arteko erreferentziakidetasun erlazioa besterik ez genuke izango (Kleiber 1994). Corblin-ek (1983) anafora aurresuposizionalak deitzen die anaforikoak ez diren erreferentziakidetasun hauei.

Ez dira hor bukatzen erreferentziakidetasunaren inguruan sortzen diren arazoak. Erreferente ebolutiboen kasuan ere zalantzak dira nagusi:

(10) Har itzazu *lau sagar*. Zuritu eta zatitu. Eduki egosten ordu erdiz. Txiki-txiki egin. Hoztu ondoren, zerbitzatu *konpota hori* gailetatxoekin.

(11) *Ur garbitan* egosi, honako hauekin batera: *gatz pixka bat, tipula erdia, porrua, azenarioak, bi baratxuri-atal, txortenik gabeko tomatea, eta patata zuritua eta pusketan txikitua... Pureari* laguntzeko, ogi-azal batzuk friji daitezke... (Pedro Subijana (1994): *Denok Sukaldari*, ETB, S.A., 78 or.)

(12) *Arratoitxoa* begien itxi-ireki batean *adats ilegorridun eta begi distiratsudun neskatxa* bilakatu zen... *Neska gazteak* ibiltzeari ekin zion basotik irteteko asmoz... (Eguzkia baino ahaltsuagoa. *Ilargi Erditxoaren ipuinak*, 1993)

Denboran zehar gertatzen den itxuraldaketak, edo materiaren transformazioarekin (naturalak edo eragindakoak) gertatzen diren aldaketak adierazterakoan oso zaila da anaforaren mugak jartzea. Esate baterako, tartean prozesu bat azaltzen denean guztiz gal daiteke adierazpide baten identitate erreferentziala, lehen *sagarrak* zirenak *konpota* bihur daitezke prozesua bukatzerakoan, edo osagai sorta bat pure bilaka daiteke, etab., baina *sagarrak* eta *konpota* izaera ezberdina dute eta baita erreferente ezberdinak ere, beste horrenbeste gertatzen da lapikoan jartzen ditugun osagaiekin eta bukaeran lortzen dugun *pure*arekin.

Ipuinaren adibidera (12) jotzen badugu, hasieran *arratoitxoa* zena *neskatxa* bihurtzen da, baina *arratoiak* erreferente edo errealitate batera garamatza eta *neska ederrak* beste batera. Halere, ez da horrela gertatzen pertsona bati buruzko kontakizuna egiten denean; horrelakoetan bilakaera bat gerta daiteke: *umea* → *gaztea* → *gizona*, baina kasu honetan izaki bera da, pertsona berbera da kasu guztietan, jatorrizko erreferentea hiperonimo ezberdinen bitartez berreskuratzen da.

Kasu hauetan guztietan adierazpideen arteko harreman anaforikoa justifika daiteke, baina inola ere ez haien artean erreferentziakidetasuna dagoela.

Bestalde, erreferentziakidetasuna dagoen harremanetan jartzen badugu arreta, askotariko adierazpideak aurkituko ditugu eta bereizketa mota ezberdinak egin ahal izango ditugu, anafora pronominalak eta ez pronominalak, eta anafora uztartuak eta askeak kasu.

Anafora pronominalen artean izango genituzke pertsona izenordainak eta beste hizkuntza unitate batzuk ere: aditzondoak (han, hantxe, bertan...), elkar, X-en burua, eta izenordain genitiboak.

(13) *Gizon hark bere burua* hil nahi zuen

Adibide honetan *gizon* eta *bere burua* erreferentziakide dira eta gainera, bigarren adierazpidea uztartuta dago bere gobernu kategoriaren barruan, perpausean bertan baitu aurrekaria.

(14) *Andoni bere* etxera eraman nuen

Beste adibide horretan, Zabalak (1996) nabarmentzen duen bezala, *bere* izenordainaren erreferentzia zalantzazkoa izan daiteke Andoniri edo beste pertsona bati egin diezaiokeelako erreferentzia. Perpausaren barneko elementuari erreferentzia egiten badio, kasu honetan *Andoni*ri, izenordainaren erabilera bihurkariaren aurrean gaude.

Mota horretako izenordainen funtzionamendua gramatikaren bitartez azaldu behar da. Kasu horietan izenordainak anafora uztartuak dira. B elementuak aurrekari bat (A elementua) behar du txertaturik dagoen perpausean; A eta B erreferentziakideak dira eta uztartuak daude. Anaforaren funtzionamendua perpausaren barnera mugatzen da eta elementuen arteko harremana gramatikala da sintaxia eta semantikak baldintzatzen dutelako.

Hurrengo adibidean, ordea, izenordainaren funtzionamendua bestelakoa da, eraikuntza askea izango genukeelako; izenordaina eta aurrekaria erreferentziakideak dira baina izenordain anaforikoak ez du bere erreferentzia gauzatzen perpausa berean dagoen aurrekari baten bitartez.

(15) *Ozono geruza* oso garrantzitsua da lurraren bizitzarako. *Bera* / *hura* da erradiakzio kaltegarrietatik babesten gaituen filtro naturala.

Izenordain anaforikoa semantikoki osatu gabea den neurrian testuan agertu den elementuren bat behar du esanahia hartzeko, eta bi elementuren arteko harreman hori linguistikoa da. Hala eta guztiz ere, aurrekari egokia aukeratzeko orduan irizpide pragmatikoak hartzen dira kontuan sorburu esanguratsuena aukeratzeko.

Nolanahi ere, egile batzuen ustez uztartu/aske bereizketa horrek ez du balio praktiko handirik; Kempson-en (1986) arabera anafora aske baten izaera pragmatikoa agerian jartzen duen ezaugarria anafora uztartuetan ere agertzen da. (Kleiber 1998: 32).

Bestalde, izenordainen erabilera dela eta, euskarak aditzaren flexioan markatzen ditu subjektua, objektu zuzena eta zeharkako objektua, eta frantsesez edo ingelesez ez bezala, izenordainen erabilera ez da hain beharrezkoa hurrengo adibidean ikusten dugun moduan.

(16) *Ura* oso garrantzitsua da gure bizitzan. *Ø* Gure gorputzaren osagairik nagusiena da.

Eraikuntza askeetan anafora pronominalak ez ezik pronominalak ez direnak ere baditugu, eta hauen artean leialak direnak eta ez-leialak direnak.

Anafora leiala, zentzu hertsian, aurrekariaren berrartze lexiko-sintaktikoan datza. Kasu honetan adierazpide anaforikoak honako forma hauek har ditzake:

— Termino berberaren errepikapena (deklinabide-marka aldatu arren).

(17) ...*haurtzaindegiak* hazkuntzarako gune dira... *haurtzaindegietako* profesionalen lana ez da erraza...

— Izen-sintagma zehaztugabea (Izena + bat) lexema berdina duen izen-sintagma artikuludun baten bitartez berreskuratzen denean: 'Izena + (elem. atrib.) + bat' → 'Izena + (elem. atrib.) + *-a/-ak/-ok*'.

(18) ...*hiztegi berri bat* argitaratu dute... *hiztegiak*...

— Izen-sintagma zehaztugabea (Izena + *bat*), lexema berdina duen izen-sintagma erakusledun baten bitartez berreskuratzen denean: 'Izena + (elem. atrib.) + *bat*' → 'Izena + (elem. atrib.) + erakuslea'. Egile batzuek testu barneko erreferentziatze deiktikoa esaten diote aurrekaria berreskuratzeko modu honi.

(19) ...*hiztegi bat* argitaratu dute... *hiztegi berri honek*...

Anafora ez-leialen kasuan adierazpide anaforikoaren lexema eta aurrekariarena desberdinak dira. Hauen artean kokatzen dira anafora kontzeptualak; adierazpide hauek laburbildu egin dezakete aurrekariaren edukia:

(20) "Itsasoan *organismo mota desberdinak* aurki daitezke beren konplexutasun mailaren arabera eta bizi direneko sakontasunaren edo kostatik dagoen distantziaren arabera. Bizi direneko sakontasunari dagokionez, *itsasoko izakiak* honela sailkatzen dira: *planktona..., nektona..., bentosa...* (García Azkoaga, 1999: Natur Zientziak DBH-1, Ibaizabal 1996: 95)

(21) [...] Automobila, *alkilaturiko etxearen parean* gelditu zenean, anai-arrebak korrikan irten ziren maleten bila, irrika bizian baitzeuden *bizileku hura* ikusteko. [...] (García Azkoaga 2004: NE-DBH2-13)

(22) Bainatzen ari ginela, *izurde bat* agertu zen gugandik gertu. *Ugaztun hauek* normalean ez dira horrenbeste hurbiltzen hondartzetara.

Horretaz gain, zenbaitetan aurrekariari buruzko balorazioak adierazten dituzte:

(23) [...] Baina benetako damua gero etorri zitzaigun gogora, *bat-batean atea itxi zenean haize bolada handi bat* eragin zuen. *Egoera beldurgarri hura* artean sinestezina zen!! [...] (Garcia Azkoaga 2004: NE-DBH2-5)

Gainera, adibide horietan ikusten den bezala aurrekaria sintagma labur bat zein enuntziatu osoa izan daiteke.

Ikusten ari garenez, harreman anaforikoa erreferentziakidetasunean oinarritzeak arazo bat baino gehiago dakartza. Haatik, kasu horiek guztiek, hertsiki anaforiko izan ala ez izan, testuaren kohesioa eraikitzen eta elementu ezberdinen artean gerta daitezkeen erreferentzia sareak harilkatzen laguntzen dute. Lan honetan interesatzen zaiguna da testu batean elkarren artean lotura erreferentziala edo anaforikoa duten elementuak identifikatzea, eta horretarako kontuan hartu beharko genituzke erreferentziakidetasunezko harremanak zein erreferentziakidetasun gabeak.

Alabaina egun eskura ditugun tresna informatikoek ez digute oraindik ahalbidetzen erreferentziakide ez diren elementuak elkarren artean zalantzagarritasunik gabe lotzea. Hortaz, oraingoz erreferentziakidetasunezko harreman anaforikoen eta ez-anaforikoen esparrura mugatu beharko dugu azterketa eta, adibidez, horrelako adierazpide anaforikoak identifikatzen saiatuko gara: izen bereziak, izenordainak, erakusleak eta forma hauek hartzen dituzten anafora fidelak: lexema berberaren errepikapenak [izena + (elem. atrib.) + *-a/-ak/-ok*] forma duten adierazpideak, [izena + (elem. atrib.) + erakuslea] forma dutenak. Horietaz gain, orain arte aipatu ez dugun arren, aurrekariaren bitartez interpretatu behar diren zenbait aditzondo dugu, esate baterako lekua adierazten duten batzuk.

(24) *Koba-zuloan* sartu ginen. *Barruan* oso ilun zegoen dena eta hango isiltasuna beldurgarria zen.

3. Erreferentzialki zein anaforikoki etiketatutako corpusak

Badira zenbait urte Lengoaia Naturalaren Prozesamenduan anafora eta erreferentziakidetasunaren azterketarekin dihardutela. Erreferentziaren edo anaforaren ebazpen automatikorako tresna sendo bat garatzeko ezinbestekoa da corpus etiketatua izatea (Mitkov 2002) prozesuaren lehen urrats moduan. Alegia, corpusean eskuz markatu ohi dira erreferentziakidetasun-erlazioak. Horrela, oinarri horren gainean ebazpenerako garatutako aplikazioak modu automatikoan ikasiko ditu erlazio horiek gidatzen dituzten erregelak eta hala, markatuta ez dauden beste corpus batzuetan aplikatu ahalko dira.

Puntu honetan ikusiko dugun bezala, hainbat hizkuntzetan anaforak (mota batzuk), nahiz erreferentziakidetasunak markatuak dituzten corpusak badaude.

Lancaster-eko unibertsitatean dugu soilik anaforikoki etiketatua dagoen corpus bakarrenetakoa: Lancaster Anaphoric Treebank (UCREL) (Garside et al. 1997). Corpus honek 100.000 hitz inguru ditu eta Associated Press (AP)etik hartutako egunkarietako testuek osatzen dute eta UCRELeko markaketa sistema jarraitzen dute. Hasierako motibazioa, corpusa osatzerakoan anaforaren ebazpen automatikorako alde probabilistikoa trebatzea izan zen, hau da, eskuz etiketatutako corpusetik patroiak ateratzea aplikazio informatikoak ikasi ahal izateko. Corpus honetan kohesiozko erlazio ezberdinak etiketatzeko aukera dute, elipsia barne. Erreferentea non dagoen adierazi dute, alegia, elementu anaforikoa baino lehenago agertzen den edo ondoren, hau da, anafora edo katafora den esaten digu. Horretaz gain, corpus honetan etiketa bat ere darabilte, elementu anaforiko eta erreferentearen artean zein erlazio semantiko dagoen adierazteko.

Baditugu bestelako testuak korreferentzialki etiketatuak; MUC Coreference Task (Hirschman & Chincor 1997) corpusa egunkarietako testuek osatzen dute. Eskema hau jarraituta osatutako corpusean anafora eta bere aurrekaria, eta horien arteko erlazioa markatua dute. Halere izen-sintagma eta aurrekariaren arteko identitate-erlazioa adierazten dute bakarrik. 65.000 hitz inguru korreferentzialki markatutako corpus honen helburuetako bat, anaforaren ebazpenerako algoritmorako trebakuntza eta ebaluaziorako baliagarri izatea da. Honetaz gain informazio erauzketarako sistema automatikoetan ere erabili ahal izan dute. Sistema hauek eremu jakin bateko, zein produktu espezifiko baten inguruko hitzak erauzteko gai izaten dira, horregatik, garrantzitsua da testu bateko erreferentziakidetasunak ondo markatuak izatea.

Wolverhampton unibertsitatean, aurrekoaren antzeko eskema (MUC) jarraituz 60.000 hitz dituen corpusa osatu dute. Markatutako testuak zenbait tresna elektroniko arrunten eskuliburuetakoak dira. Corpusaren markaketarako ClinKa (Orasan 2000) izeneko tresna baten laguntza izan dute, unibertsitate horretan bertan garatutakoa.

Estatu Batuetan Brown-go unibertsitatean Penn Treebank-en zati batean (93.000 hitz inguru) agertzen diren 2.463 izenordain korreferentzialki markatu dituzte.

Ingelesez korreferentzialki markatuak dauden corpusen aipamena bukatzeko DRAMA *Scheme* corpusa aipatuko dugu. MUC-en eskema bera erabiltzen da anafora eta aurrekariak identifikatzeko eta euren arteko korreferentzia erlazioak markatzeko.

Honaino ekarri ditugun corpusen adibideak ingelesa dute aztergai baina badira beste hizkuntza askotan korreferentzialki markatutako corpusak. Alemanerako adibidez, TIGER proiektuan (Kunz & Hansen-Schirra 2003) markaketa morfologiko, sintaktiko eta semantikoaz gain korreferentzia erlazioak ere markatu dituzte. Antzera egin dute Pragako unibertsitatean (Hajič & Urešová 2004) corpusa pragmatikoki markatzeko saiakera egin dutenean.

Gaztelaniaren kasuan, Alacanteko Unibertsitatean arlo honetan egiten ari diren lana aipatu behar da. Anaforaren ebazpen automatikoan lortutako emaitzak hobetzeko asmoz anaforikoki etiketatutako corpusa osatzen ari dira unibertsitate horretan (Navarro et al. 2003).

Azkenik, ezin aipatu gabe utzi hemen aurkezten dugun lanaren aitzindari izan den corpusa (Aduriz et al. 2007). 50.000 hitz inguru dituen corpus honek egunkarietako testuak ditu oinarri, eta anaforaren fenomenoaren azterketa alde konputazionaletik egiteko lehen urrats moduan, anafora pronominalak ditu markatuak, bakoitza dagokion erreferentziarekin lotuta. Markaketa hau eskuz burutu da.

Erreferentziakidetasunaren eta anaforaren fenomenoak hainbat hizkuntzatako corpusetan markatu direla ikusi dugu orain arte. Corpus hauek guztiak baliagarri izango dira etorkizun batean, ez bakarrik erreferentziakidetasun edo anafora (erdi) automatikoki ebazteko, baizik beste helburu batzuetarako garatutako tresnetarako ere (laburpen automatikorako sistemak, termino erauzle automatikoak, galdera-erantzun sistemak, etab).

4. Etiketatzeko tresnak

Hastapenetan egindako lanetik abiatuta pauso bat haratago eman nahi izan dugu. Lehen pauso batean, anafora fenomenoaren alor espezifiko batean bakarrik zentratu ginen, hau da, anafora pronominalaren markaketa egin genuen, eta oraingoan urrutirago joan nahian, anafora multzo bakarra ez ezik, derrigorrez anafora ez diren erreferentziakidetasunak ere markatzeko asmoa dugu, honela testuaren ulermen zabalago batera iritsiko garelakoan gaude.

Corpusen etiketatze lan hauetan lagungarri izaten dira helburu horretarako garatutako tresnak. Honela guri ere etiketatze lana erraztuko digun tresna bat hautatu dugu zeregina bideratzeko, izan ere, aurreko lanean ikusi genuen interesgarria izango litzatekeela fenomeno hau etiketatzeko tresna aurreratua izatea.

Zenbait ezaugarri hartu ditugu kontuan tresna aukeratzeko orduan, besteak beste, honako hauek: zein mailatarainoko etiketatzea eskaintzen duten, gure tresnetara ondo egokitzen diren, formatu kontuak, etab.

Ikusi eta aztertutako guztien artean, aipagarrienak hurrengo hauek dira: Wolverhampton-go Unibertsitatean garatutako ClinkA tresna (Orasan 2000) sendoa izan

arren, batez ere egitura edo eskema aldetik ez digu gehiegi laguntzen, haiek MUC-7 Coreference Task Definition (Hirschman & Chinchor 1997) definitutako eskema jarraitzen baitute guk ez bezala. Guk erreferentziakidetasuna markatzeko ezarritako gidalerroak haiengandik urrun geratzen dira.

Aipatzeko beste tresnetako bat Alembic Workbench[1] etiketatzaile orokorra da. Beste maila batzuen artean erreferentziakidetasuna markatzeko aukera ere badu. Honetan ikusi dugun abantailetako bat etiketak zabaltzeko aukera ematen duela da. Hala ere, nahiko nahasgarria gertatzen da etiketa hauekin lan egiterakoan, testua bera eta etiketatzea nahasiz.

MATE Workbench-en (Dybkjær and Bernsen 2000) testuinguruan ahalmen handiko tresna garatu dute. Teorikoki gure etiketatzea gauzatzeko plataforma ideala izan zitekeen, baina testu handiekin zailtasunak izan dituztenen esperientziak irakurri ondoren, tresna hau ez erabiltzea erabaki genuen.

Tresnen zerrendatze honekin bukatzeko MMAX aplikazioa (Müller and Strube 2001) aipatuko dugu. Arina eta erabilerraza izateaz gain, norberak bere beharretara egokitzeko erraztasunak eskaintzen ditu. Bestetik, 'stand-off' deritzon informazioa metatzeko sistema erabiltzen du, hau da, oinarrizko datuak (testua bera) fitxategi batean eta bigarren mailakoak (informazio gramatikala zein testuala) beste fitxategi ezberdin batean gordetzen ditu. Horrez gain, etiketatze prozesuan pertsona batek baino gehiagok parte hartzeko aukera ere badu, eta etiketatzaileen arteko adostasun eta desadostasunak agertzeko aukera ematen du. Aztertutako tresnetatik gure eskakizunetara hoberena egokitzen denez, gure lana aurrera eramateko MMAX aplikazioa aukeratu dugu.

5. Corpusa

Lan honetarako oinarritzen garen corpusa EPEC izenekoa da (Aduriz et al. 2006). Corpus hau 3LB proiektuaren barruan sortu zen, gaztelaniazkoarekin eta katalanekoarekin batera (Palomar et al. 2004). Proiektuaren helburua corpusen etiketatze sintaktikoa eta semantikoa egitea zen. Euskarazkoari dagokionez, egunkarietako berrogeita hamar mila hitz anotatu ziren sintaktikoki, dependentzietako anotazio-sistema erabiliz (Aranzabe et al. 2003). Ondorengo lerroetan azalduko dugu corpusaren analisi-prozesu modularra (Aduriz et al. 2006), modulu bakoitzean zein informazio gehitzen zaion ikusteko, anaforaren eta aurrekarien markatzera iritsi baino lehen.

Lehen-lehenik corpusa morfologikoki analizatzen da Morfeus analizatzaile morfologikoa erabiliz (Aduriz et al. 1998). Morfeus-en lana da corpuseko hitz guztiak[2] analizatzea bere testuingurua kontuan hartu gabe. Lehen prozesu honen ondoren, analizatutako hitzek informazio morfosintaktikoa izango dute: kategoria gramatikala, azpikategoria, numero mugatasunari buruzko informazioa, kasu atzizkiena, eta gehienetan, funtzio sintaktikoari dagokion informazioa ere. Hona hemen morfosin-

[1] http://www.mitre.org/tech/alembic-workbench/

[2] Zuriunetik zuriunera doan hitza nahiz hitz anitzeko terminoa.

taktikoki etiketatutako esaldi baten adibidea: "*Udaberrian hegazti ugari pasatzen da gure mendien eta herrien gainetik.*"[3]

(25) /<Udaberrian>/<HAS_MAI>/
("udaberri" IZE ARR DEK NUMS MUGM DEK INE @ADLG)
/<hegazti>/
("hegazti" IZE ARR DEK ABS MG @OBJ @SUBJ @PRED)
("hegazti" IZE ARR @KM>)
/<ugari>/
("ugari" ADJ IZO DEK ABS MG @OBJ @SUBJ @PRED)
("ugari" ADJ IZO @<IA)
("ugari" DET DZG MG DEK ABS MG @SUBJ)
("ugari" DET DZG MG @ID>)
("ugaritu" ADI SIN AMM ADOIN @-JADNAG)
/<pasatzen>/
("pasatu" ADI SIN AMM ADIZE DEK INE @OBJ @-JADNAG_MP_OBJ)
("pasatu" ADI SIN AMM ADOIN ASP EZBU @-JADNAG)
/<da>/
("izan" ADL A1 NR_HU @+JADLAG)
("izan" ADT A1 NR_HU @+JADNAG)
/<gure>/
("gu" IOR PERARR NUMP GU DEK GEN DEK ABS MG @IZLG> @<IZLG @OBJ @SUBJ @PRED)
("gu" IOR PERARR NUMP GU DEK GEN @IZLG>)
("guretu" ADI SIN AMM ADOIN @-JADNAG)
/<mendien>/
("mendi" IZE ARR DEK GEN NUMP MUGM DEK ABS MG @IZLG> @<IZLG @OBJ @SUBJ @PRED)
("mendi" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<eta>/
("eta" LOT JNT EMEN @PJ)
("eta" LOT MEN KAUS @PJ)
/<herrien>/
("herri" IZE ARR DEK GEN NUMP MUGM DEK ABS MG @IZLG> @<IZLG @OBJ @SUBJ @PRED)
("herri" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<gainetik>/
("gain" IZE ARR DEK NUMS MUGM DEK ABL @ADLG)
/<.>/<PUNT_PUNT>/

Lehen urrats honen ondoren sortzen zaigun arazorik handienetakoa anbiguotasunarena da. Alegia, analisia testuingurutik at egiten denez, askotan gertatzen da hitzek anbiguotasuna izan dezaketela, lexikoak berak eraginda, kasuak eraginda ala funtzioak eraginda.

[3] Esaldiaren analisian agertzen diren laburduren azalpena 9. puntuan azaltzen da, glosarioan.

Desanbiguazio prozesua beste modulu batek eramaten du aurrera, EUSTAGGER izeneko analizatzailearen bitartez. Bere zeregin nagusia anbiguotasun morfosintaktikoa murriztea da, desanbiguatzea, alegia, testuinguru bakoitzean zuzena den aukera uzten du. Funtzio etiketa sintaktikoa ere esleitzen du. Ikus dezagun lehengo adibide bera morfosintaktikoki desanbiguatua.

(26) /<Udaberrian>/<HAS_MAI>/
("udaberri" IZE ARR DEK NUMS MUGM DEK INE @ADLG)
/<hegazti>/
("hegazti" IZE ARR @KM>)
/<ugari>/
("ugari" DET DZG MG DEK ABS MG @SUBJ)
/<pasatzen>/
("pasatu" ADI SIN AMM ADOIN ASP EZBU @-JADNAG)
/<da>/
("izan" ADL A1 NR_HU @+JADLAG)
/<gure>/
("gu" IOR PERARR NUMP GU DEK GEN @IZLG>)
/<mendien>/
("mendi" IZE ARR DEK GEN NUMP MUGM DEK ABS MG @IZLG>
@<IZLG @OBJ @SUBJ @PRED)
("mendi" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<eta>/
("eta" LOT JNT EMEN @PJ)
/<herrien>/
("herri" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<gainetik>/
("gain" IZE ARR DEK NUMS MUGM DEK ABL @ADLG)
/<.>/<PUNT_PUNT>/

Analisiaren puntu honetan aztergai izango dugun corpusa morfologikoki analizatua dago, funtzio sintaktiko nagusiak esleituta, eta ia erabat desanbiguatua. Prozesuaren azken urratsa burutzeko, zatitzailea ala "chunker"-a aplikatzen da. Modulu honek kateak ala sintagmak ezagutzen ditu. Ezagutuko diren egitura sintaktiko oinarrizkoak hauek izango dira: entitateak,[4] postposizio konplexuak[5] eta kateak[6] (sintagmak eta aditz kateak), ondoren dugun adibideak erakusten digun bezala.

(27) "<Udaberrian>"<HAS_MAI>"
"udaberri" IZE ARR DEK NUMS MUGM DEK INE @ADLG HAS_MAI
% SINT
"<hegazti>"
"hegazti" IZE ARR @KM> % SIH
"<ugari>"

---

[4] Entitateak: *Euskal Herria, Txomin*, etab.
[5] Postposizio konplexuak: *-ren ondoren, 0 gisa* …
[6] Izen-sintagma: *udaberrian* edo *etxe berri bat*; aditz-kateak berriz: *pasatzen da,* etab.

“ugari” DET DZG MG DEK ABS MG @SUBJ % SIB
“<pasatzen>”
“pasatu” ADI SIN AMM ADOIN ASP EZBU @-JADNAG NOTDEK % ADIKATHAS
“<da>”
“izan” ADL A1 NOR NR_HU @+JADLAG % ADIKATBU
“<gure>”
“gu” IOR PERARR NUMP GU DEK GEN @IZLG> % SIH
“<mendien>”
“mendi” IZE ARR DEK GEN NUMP MUGM @IZLG>
“<eta>”
“eta” LOT JNT EMEN @PJ AORG
“<herrien>”
“herri” IZE ARR DEK GEN NUMP MUGM @IZLG>
“<gainetik>”
“gain” IZE ARR DEK NUMS MUGM DEK ABL @ADLG % POS % SIB
“<$.>”<PUNT_PUNT>”
PUNT_PUNT

Maila morfologiko zein sintaktikoa automatikoki egin dugula ikusi dugu. Maila testualeko etiketatzea aldiz, eskuz burutuko dugu, lehen aipatu dugun tresnaren laguntzaz.

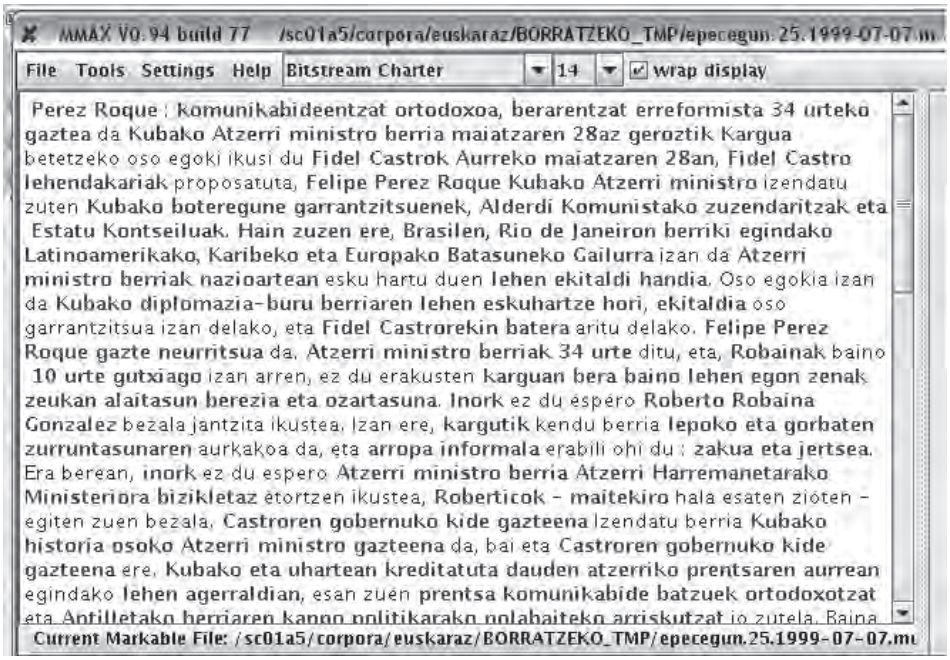
6. Etiketatze lana

Aurreko puntuan ikusi dugun bezala, corpus analizatutik abiatuko gara, ez testu hutsetik. Lehenagotik egindako analisiak testu egituratua emango digu. Honek etiketatzea errazteaz gain, gerorako baliagarri izango zaigun informazio linguistikoa emango digu.

Anafora pronominalak etiketatzeko egin genuen antzera (Aduriz et al. 2007), aditz kateak alde batera utzi eta izen kateetan jarriko dugu arreta. Halere, lan honetan gure aztergaia zabalduko dugu: izenordain funtzioa betetzen duten determinatzaile anaforikoez gain, erreferentziakidetasuna adieraz dezaketen izen kate guztiak markatuko ditugu, ondoren azalduko dugun bezala (ik. 6.1.1 atala). Horretarako, goian aipatutako MMAX tresnaren laguntza izan dugu.

6.1. MMAX tresna

MMAX izeneko tresnaren itxura eta funtzionamendua azaltzen saiatuko gara puntu honetan. Esan bezala, tresna honek etiketatzea erraztuko digu eta lehenagotik informazio linguistikoa automatikoki esleitu zaionez, testuak izen-sintagmak markatuak izango ditu ondorengo irudian (1. irudia.) ikus dezakegunez.



1. irudia

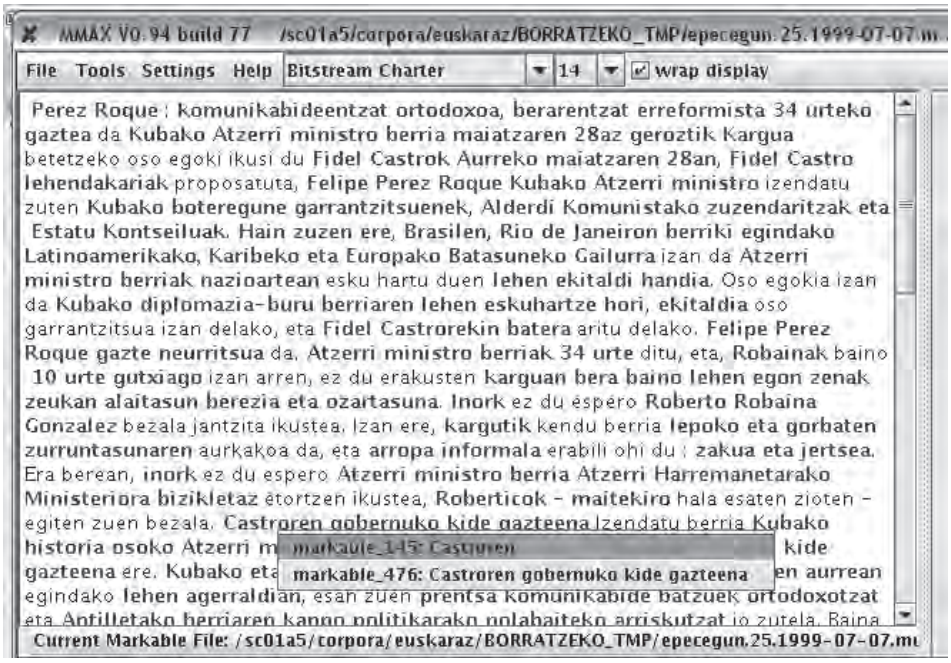
Erreferentziagai den izen-sintagma

Tresnak kolore kodea erabiltzen du testuaren ezaugarriak adierazteko (hemen ikusten ez bada ere), adibidez, kolore urdina erabiltzen da erreferentziakidetasunen bat izan dezaketen izen-sintagmak markatzeko. Kolore horia, berriz, izen-sintagma bereko parte denean.

6.1.1. 'Markables' kontzeptua

Etiketatzearen lehen fasean erreferentziakide edo anafora izan daitezkeen izen sintagma guztiak ondo markatuak dauden egiaztatu beharko dugu. 'Markables' kontzeptua diogunean, erreferentziakidegai diren izen sintagmez ari gara. Makinak berez markatuak etorriko dira, horrek ez du esan nahi guztiak ondo markatua izango dugunik. Baliteke zuzendu behar izatea, edota izen sintagma batzuk baztertzea, adibidez, 2. irudian ageri den *era berean* esamoldeak ez du testu honetan erreferente konkreturik izango, ez behintzat guk etiketatu nahi ditugun erreferentziakidetasunen artean.

Bestalde, gure tresnek sintagma osoak markatzeko gaitasuna dute, baina gerta daiteke izen sintagma horien azpian beste izen-sintagma motzagoak zein mendeko perpausak egotea. Izen-sintagma baten barruan beste osagai bat 'markable' bezala etiketatzea ahalbideratzen du tresnak, hurrengo irudiak erakusten digun moduan.



2. irudia

"Azpisintagma" posible bat

Izen-sintagmaren parte diren 'azpisintagmak' guk markatu beharko ditugu, gerora erreferentziak ezartzerakoan baliagarri izan baitaitezke, batzuetan erreferentziakidea 'azpisintagma' bakarrik izango delako.

(28) Castroren gobernuko kide gazteena da Perez Roque. Hark gobernuan dirauen artean Atzerri ministro lanetan arituko da.

6.2. Gidalerroak

Corpusa etiketatzeko erabili dugun tresnaren nondik norakoak eta beste kontzeptu batzuk argitu ondoren etiketatzaileek beharko dituzten gidalerroei buruz arituko gara. Hauek osatzeko aurretik egindako lanaren esperientzia erabili dugu (Aduriz et al. 2007), eta baita lan honetan egindako azterketa-lan teoriko-deskriptiboa.

Gidalerroak zehatzak eta argiak izatea ezinbestekoa da, kontuan izanda batzuetan, etiketatze lanaren kalitatea bermatze aldera, bi etiketatzaile baino gehiago aritzen direla. Hala ere, ezinezkoa izango da erabateko adostasuna gertatzea etiketatzaile guztien artean.

MMAX programa erabiliz etiketatzea bi fasetan egingo dugu, lehenengoan erreferentziakide izan daitezkeen izen-sintagmak identifikatuz eta bigarrenean erreferentziakideak direnak euren aurrekariekin markatuz.

Artikuluaren hasieran (ik. 2. puntua) aipatutako anafora fidelak ere markaketa honetan kontuan hartuko ditugu. Lehen fasean, erreferentziakide izateko aukera duten izensintagmek ondorengo ezaugarri gramatikalak izango dituzte eta honela sailka genitzake:

1. Pronominalak
    a. Izenordainak.
    b. Erakusleak, izenordain funtzioa betetzen dutenean eurak bakarrik izensintagma osatuz.
2. Termino berberaren errepikapenak
    a. Izena + (elem. atrib.) + artikulua *(-a/-ak/-ok)*
    b. Izena + (elem. atrib.) + erakuslea
    c. Izen berezia + (kasu marka)
3. Aditzondo zenbait (lekuzkoak)

Honako elementu hauek guztiak testuan lehenago aipatutako zerbaiti egin diezaiokete erreferentzia, honek hainbat erlazio mota ezarriko ditu testuko osagaien artean. Hurrengo ataletan erlazio mota hauen ezberdintasunak eta markatzeko irizpideak azalduko ditugu.

6.2.1. Erreferentzia sareak markatzen: erlazio motak

Behin balizko erreferentziakide osagaia eta bere aurrekaria markatuta daudela, bien arteko erlazio-mota zein den erabaki behar da. Aipatutako tresnak aukera ematen digu erreferentziakide osagaia eta bere aurrekaria modu eroso eta argian lotzeko. Hau guzti hau, 'xml' formatu estandarrean gordetzen du, honela, beste edozein aplikazio informatikoetan erabili ahal izango dugu.

Erreferentziakidetasun anaforikoen artean, adierazpide pronominalak eta nominalak markatuko ditugu, baina azken hauen artean, oraingoz, anafora leialak bakarrik; beste baterako utziko ditugu fidelak ez diren anaforen markatzea. Bestalde, balore anaforikoa duten lekuzko aditzondoak eta izen bereziak ere hartuko ditugu kontuan. Harreman anaforiko mota gehiago baldin badago ere, ez ditugu corpus honetan markatuko, izan ere, etorkizunera begira modu automatiko batean detektatzeko eragozpenak ikusten baititugu, giza-ulermen inferentzia beharrezkoa gertatzen delako.

Hona hemen laburki, etiketatzaileak ezarri beharko dituen erlazio motak:

— *Pronominalak:* izenordain funtzioa betetzen duten determinatzaile erakusleak, pertsona izenordainak, aditzondoak (jakin batzuk), *elkar*, *X-en burua*, eta izenordain genitiboak.

(29) *Ura* oso garrantzitsua da gure bizitzan. *Bera /hura* da gure gorputzaren osagairik nagusiena.

— *Anafora fidelak:* lexema bera duen hitza errepikatzen denean, nahiz eta honek atributuren bat edo deklinabide morfema ezberdinen bat izan.

(30) Igandean mendira igo zirenean, hango iturriko *ur freskoa* edan zuten. *Ur horrek* egingo zien kalte nonbait.

— *Lekuzko adberbioak:* testuan zehar aipatutako zerbaiti erreferentzia egiten diotenean.

(31) *Koba-zuloan* sartu ginen. Barruan oso ilun zegoen dena eta hango isiltasuna beldurgarria zen.

— *Izen bereziak:* aurretik edo atzetik erreferentziakide den elementuren bat agertzen denean.

(32) *Perez Roque*... Kubako diplomazia-buru berria... *Felipe Perez Roque*

7. Ondorioak eta etorkizuneko lana

Euskarazko corpusetan erreferentziakidetasun erlazioak nola gauzatzen diren aztertu dugu artikulu honetan, halaber, erlazio hauek modu emankorrago batean tratatzeko ahaleginean, EPEC izeneko corpusean erreferentziakidetasun horiek markatzen bigarren urratsari ekin diogu.[7]

Alderdi teorikotik egindako hausnarketa da lehen hurbilpenetik hona suma daitekeen ekarpenik handiena, gaia anafora pronominal soiletik beste erreferentziakidetasun kasuetara zabaldu dugulako.

Alderdi teorikotik gaia zabaltzeak, ezinbestean izan du alderdi praktikoan eragina. Honek eraman gaitu laneko gidalerroak aldatzera, eta era berean osatzera ere. Bestetik, corpusaren markaketari dagokionez, etiketatzeko tresnaren beharra eta honen hautaketa egin beharrak ere hamaika erabaki hartzera bultzatu gaitu. Puntu giltzarria da markatzeko tresnarena, izan ere, oinarrian erabiltzen dugun informaziora egokitu behar baitu modu erraz batean.

Oinarritzen garen corpusari dagokionez berriz, informazio morfologiko eta sintaktikoa bere baitan edukitzeak erraztuko du ondorengo etiketatze automatikoa, izan ere, soilik hasierako fasean egingo da eskuzko markaketa. Esan bezala, oinarrizko corpusak eskaintzen duen informazio morfologiko eta sintaktikoak batetik, eta markatzeko tresna egokiak bestetik, ahalbidetuko dute etiketatze automatikoa, azkartasuna emanez markatze-prozesuari. Izan ere, artikulu honetan azaltzen ari garena proiektu orokor baten barruan kokatzen da, alegia, EPEC corpusaren anotazio erabat semantikoa helburu duena.

Euskaraz horrelako corpus semantikoki anotatua izateak helburu bikoitza izango luke: ikuspegi konputazionaletik, erreferentziakidetasuna (erdi-) automatikoki ebatziko lukeen tresna baten garapena. Eta testuinguru horretan, corpusa tresna horren probaleku litzateke. Eta hizkuntzalaritzaren ikuspuntutik, euskarazko errefentziakidetasunaren berri izaten lagunduko digu, anaforaren erabilera zehazten eta horren erresoluziorako bideak jartzen.

Azkenik, eta epe luzeragorako, euskararako garatuko diren hainbat aplikazio informatikoetan, hala nola galdera-erantzun sistema automatikoetan, laburpen sistemetan edota itzulpen automatikoko aplikazioetan aztertutako hau guztia txertatzeko aukera ikusten dugu.

---

[7] Lehen urratsaren nondik norakoak (Aduriz et al. 2007) argitalpenean azalduta daude.

Bibliografia

Aduriz, I., Aranzabe, M., Arriola, J. M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A. & Urizar, R., 2006, "Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing". *Corpus Linguistics Around the World. Book series: Language and Computers*. Vol 56 (1-15. or.). Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands.

—, Ceberio, K., Díaz de Ilarraza, A., 2007, "Pronominal Anaphora in Basque: Annotation issues for later computational treatment", *6th Discourse Anaphora and Anaphor Resolution Colloquium. DAARC2007,* Lagos Portugal.

Aldezabal, I., 2004, *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz,* Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Aranzabe, M. J., Arriola, J. M., Atutxa, A., Balza, I., Uria, L., 2003, "Guía para la anotación sintáctica manual de Eus3LB (corpus del euskera anotado a nivel sintáctico, semántico y pragmático)", UPV/EHU/LSI/TR-13.

Charolles, M., 1991, "L'Anaphore. Definition et classification des formes anaphoriques", *Verbum* 14: 2-3-4, 203-216.

Corblin, F., 1983, "Défini et démonstratif dans la reprise immédiate", *Le Français moderne* 51, 118-133.

Dybkjær, L. and Bernsen, N. O., 2000, "The MATE Workbench", *Proceedings of the LREC'2000 workshop on Data Architectures and Software Support for Large Corpora,* Athens, 33-37 (a).

Garcia Azkoaga, I. M., 1999, "Elementu anaforikoak eskolako testuetan", *FLV* 82, 393-417.

—, 2004, *Kohesio anaforikoa hiru testu generotan. Adinaren araberako azterketa,* Bilbo, Euskal Herriko Unibertsitatea.

Garside, R., Leech, G. & McEnery, A. (arg.), 1997, *Corpus Annotation. Linguistic Information from Computer Text Corpora,* Longman, London.

Gojenola, K., 2000, *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan,* Donostia, Informatika Fakultatea, Euskal Herriko Unibertsitatea.

Hajič, J. & Urešová, Z., 2004, "The Prague Dependency Treebank", IXA taldeari egindako aurkezpena, Donostia.

Hirschman, L. and Chinchor, N., 1997, "MUC-7 coreference task definition", in *MUC-7 Proceedings,* Science Applications International Corporation.

Kempson, R., 1986, "Definite NPs and Context-Dependence: a Unified Theory of Anaphora". In T. Hyers et alii (arg.), *Reasoning and Discourse Processes*, Academic Press, London, 209-239.

Kleiber, G., 1988, "Peut-on définir une catégorie générale de l'anaphore?, *Vox Romanica* 47, 1-13.

—, 1994, *Anaphores et pronoms*, Louvain-la-Neuve, Duculot.

Kunz, K. & Hansen-Schirra, S., 2003, "Coreference Annotation of the TIGER Treebank". In *Proceedings of the Workshop Treebanks and Linguistic Theories*, Växjö, Sweden.

Milner, J. C., 1982, *Ordres et raisons de la langue*, Paris, Seuil.

Mitkov, R., 2002, *Anaphora resolution,* Longman, London.
Müller, C., Strube, M., 2001, "MMAX: A Tool for the Annotation of Multi-modal Corpora". In *Proc. of the 4th SIGDIAL*, Sapporo, Japan.
Navarro, B., Civit, M., Martí, M. A., Marcos, R., Fernández, B., 2003, "Syntactic, semantic and pragmatic annotation in Cast3LB", in *Proceedings of the Shallow Processing of Large Corpora. A Corpus Linguistics Workshop*, Lancaster, UK.
Orasan, C., 2000, "CLinkA a Coreferential Links Annotator", in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, Athens, Greece.
Urkia, M., 1997, *Euskal morfologiaren tratamendu informatikorantz***,** Filologia eta Geografia-Historia Fakultatea, Euskal Herriko Unibertsitatea.
Zabala, I., 1996, "Testu-lotura: lotura tematikoa eta erreferentzia-sareak testu teknikoetan", in Zabala I. (koord.). *Testu-loturarako baliabideak. Euskara Teknikoa*, Bilbo, Euskal Herriko Unibertsitatea, 15-44.

Glosarioa

% ADIKATBU: aditz-kate edo sintagma jarraitu bateko azken elementuak darama;
% ADIKATHAS: aditz-kate edo sintagma jarraitu bateko lehenengo elementuak darama;
% SIB: sintagmaren bukaera;
% SIH: sintagmaren hasiera;
% SINT: hitz bakarreko sintagma;
@+JADLAG: aditz laguntzailea;
@+JADLAG_MP: aditz laguntzaile mendekoa;
@+JADNAG: aditz nagusi jokatua;
@+JADNAG_MP: aditz nagusi jokatu mendekoa;
@<IA: eskuineko adjektiboa;
@<IZLG: eskuineko izenlaguna;
@ADLG: adizlaguna;
@ID>: ezkerreko determinatzailea;
@IZLG>: ezkerreko izenlaguna;
@-JADNAG: aditz nagusia ezjokatua;
@-JADNAG_MP_OBJ: objektu funtzioa duen aditz nagusi ezjokatu mendekoa;
@KM>: kasua daraman formaren modifikatzailea;
@OBJ: objektua;
@PJ: perpaus juntadura;
@PRED: predikatiboa;
@SUBJ: subjektua;
A1: DA / DU: izan / *edun-en orainaldia (indikatibozko orainaldia);
ABL: ablatiboa;
ABS: absolutiboa;
ABZ: hurbiltze-adlatiboa;
ADI: aditza;
ADIZE: aditz izena;
ADJ: adjektiboa;

ADL: aditz laguntzailea;
ADOIN: aditzoina;
ADT: aditz trinkoa;
AMM: aditz-mota morfema;
AORG: -a organikoa daukan hitzak darama;
ARR: (izen) arrunta;
ASP: aspektu-morfema;
DEK: deklinabidea;
DET: determinatzailea;
DZG: determinatzaile zehaztugabea;
EMEN: emendiozko juntagailua;
EZBU: aspektu-marka ezburutua;
GEN: genitibo edutezkoa;
HAS_MAI: letra larriz hasitako hitza;
INE: inesiboa;
IOR: izenordaina;
IZE: izena;
IZO: izenondoa;
JNT: juntagailua;
KAUS: kausala;
LOT: loturazko elementua: lokailu eta juntagailuak hartzen ditu bere baitan;
MEN: mendekoa;
MG: mugagabea;
MUGM = M: mugatua;
NOTDEK: deklinabidetik pasatzen ez den aditza;
NR(_)HU: NOR paradigmako HURA;
NUMP = PL = P: numero plurala;
NUMS = S: numero singularra;
PERARR: izenordain pertsonal arrunta;
PUNT_PUNT: puntua;
SIN: aditz sinplea.