ARTIKULUAK

Euskarazko anafora pronominala: ikuspuntu konputazionala eta corpus baten garapena

ITZIAR ADURIZ*, KLARA CEBERIO** ETA ARANTZA DÍAZ DE ILARRAZA**

* Department of General Linguistics
University of Barcelona
** IXA Group (http://ixa.si.ehu.es)
Department of Computer Languages and Systems, UPV-EHU

(Pronominal Anaphora in Basque: computational point of view and the development of a corpus)

Abstract

This paper describes the process of annotating pronominal anaphor in a corpus of Basque which consists of 54.000 words. Our aim is to use this annotation as a basis for later computational processing. The linguistic study carried out and the criteria defined for the tagging process are also presented in the paper.

Keywords: *Pronominal anaphora, anaphoric annotation of a corpus.*

1. Sarrera

Gure eguneroko bizitzan gero eta arruntagoa gertatzen da ordenagailuaren erabilera eta honek bere baitan hartzen du testuaren prozesamendua. Bertan erabiltzen diren programa ohikoenak testua prozesatzen dute eta horretarako Lengoaia Naturalaren Prozesamenduaren arloan garatzen diren aplikazio eta teknologiez baliatzen gara.

Arlo honetan azken urteetan eman diren aurrerapausoak nabarmenak izan dira hizkuntza guztietan, eta euskara ez da atzean geratu. Lan horretan dihardu, hain zuzen ere IXA taldeak 15 urte inguru.

Testua prozesatzearen helburuetako bat, ahal den neurrian, gizakiak duen hizkuntzaren ezagutza-maila bera ordenagailuari ematea da. Horretara-

ko corpusaren etiketatze morfologikoa, sintaktikoa eta semantikoa egitearekin batera, diskurtsoan ematen diren zenbait elementu etiketatu nahi dira, beste batzuen artean anaforaren fenomenoa.

Lehenengo pauso batean, hitza (morfologia) eta hitzen arteko erlazioak (sintaxia eta semantika) hartu dira kontuan. Testuaren tratamentuan unitatea zabaltzen da, esaldia gaindituz paragrafoa eta paragrafoan agertzen diren elementuen arteko erlazioa da aztertu beharrekoa, eta interpretazio arazoak agertzen dira.

Maila honetan guk aztertu eta etiketatu nahi dugun fenomenoa anafora da eta adibideak erakusten digun moduan, hartzaileak berak duen ezagutzari esker jakiten du kohesiorako balio duten osagai laburtu horiek zein elementuri egiten dion erreferentzia:

> *Mitterrand$_i$ libre uzteko presazko helegitea aurkeztu zuen bere defentsak hura$_i$ espetxeratu ostean, (…)*

Hemen ikusten da *hura* determinatzaileak *Mitterrand* pertsonari egiten diola erreferentzia eta ez perpauseko beste elementuren bati. Hizkuntza ezagutu eta ulertzeaz gain (informazio morfologiko, sintaktiko eta semantikoaren ezagutza) munduaren ezagutzak (pragmatikak) esaten digu pertsona izango dela espetxeratuko dutena eta ez perpauseko beste elementuren bat. Lan honetan aurkeztuko dugun aztergaia hauxe izango da, diskurtsoan ematen den anafora fenomenoaren ingurukoa eta honek ematen dituen interpretazio arazoak.

Anafora orokorrean definitzea zaila gertatzen da, alorraren arabera kontzeptu oso desberdinak hartzen baititu bere baitan. Lan hau hizkuntzalaritza konputazionalaren arloan kokatzen denez, hizkuntzalaritzan anafora kontsideratzen dena izango dugu kontuan. Definizio labur bat ematearren, anafora, diskurtsoan lehenago aipatutako zerbaiti erreferentzia egiten dion elementua da, kasu batzutan testua errepikakorra gerta ez dadin erabiltzen dena.

Lan hau Lengoaia Naturalaren Prozesamenduan (LNP) barruan kokatzen da, eta aurrerago ikusiko ditugun hainbat arrazoirengatik, anaforaren azpimultzo baten azterketa egitera mugatuko gara.

Ondorengo atalean, anafora eta diskurtsoan ematen diren beste fenomenoen arteko ezberdintasunak argitzen saiatuaz batera, ikuspegi konputazionaletik abiatuz anafora linguistikoaren sailkapen bat azalduko dugu. Jarraian anaforikoki markatutako corpusa osatzeko oinarriak finkatuko ditugu, horretarako hainbat iturritatik jasotako informazioaz baliatuz. Azkenik corpusaren markaketaren aurrekariak eta guk osatu dugun corpusaren nondik norakoak azalduko ditugu, hortik ateratako ondorioak etorkizun batean garatuko ditugun tresna informatikoetarako baliagarri izango zaizkigulakoan.

2. Anafora eta korreferentzia

Lanaren aztergaia zehazten hasi aurretik kontzeptu orokor batzuk argitzen saiatuko gara, anaforaren fenomenoa hobeto ulertu ahal izateko.

Testu bat irakurtzean, bere osotasunean ondo ulertzeko nolabaiteko lotura egon ohi da esaldien artean. Ez dira esaldi solteak izaten, esaldi bakoitzak aurretiaz esan denarekin zerikusia izango du. Esaldi arteko lotura eta koherentzia horri kohesioa deitu ohi zaio eta beharrezkoa izaten da hartzaileak diskurtsoa inongo arazorik gabe uler dezan.

Kohesio honen barruan koka dezakegun fenomeno linguistiko bat korreferentzia da. Diskurtsoko bi elementuk mundu errealeko erreferente bera izatean ematen da, hau da, aurrerago argiago azalduko dugun bezala anaforamota batzuetan ematen dena. Ikus dezagun adibide bat:

> *[Ben Amor]*[1]$_i$ *ere ez da Mundiala amaitu arte etorriko Irunera, honek*$_i$ *ere Tunisiarekin parte hartu baitu Mundialean.*

Ben Amor izen bereziak eta *honek* erakusleak pertsona berari egiten diote erreferentzia, errealitateko erreferentea bera da eta korreferentzia ematen dela esan dezakegu. Aldi berean, anafora ere ematen da, *honek* determinatzaile erakusleak testuan lehenago azaldu den entitate bati egiten baitio erreferentzia, *Ben Amor*-i hain zuzen, baina izena ez errepikatzearren forma laburtu bat erabili du testua idatzi duenak.

Segidan dakargun adibidean aldiz, desberdina da erreferentzia mota:

> *Gizon*$_i$ *bakoitzak bere*$_i$ *patua du.*

Irakurleak badaki *bere patua*-k erreferentzia *gizon bakoitzak*-i egiten diola, baina horrek ez du esan nahi erreferente bera dutenik. Honela bada, kasu honetan anafora dugu, baina ez dago korreferentziarik. Anafora mota batzuetan (aurrerago ikusiko dugu sailkapena) beraz, korreferentzia ematen da, baina ez beti.

Alderantziz ere gertatu ohi da, korreferentzia ematen den kasuan eta aldiz, anaforarik gertatzen ez denean. Kasu hori dokumentu ezberdinetan pertsona edo elementu berari erreferentzia egiten zaionean ematen da, korreferentzia egon badago elementu berdinari buruz ari garelako, baina erlazio anaforikorik ez.

---

[1] Egunkaria, 2001; adibideetan '$_i$' ikurrarekin elementu anaforikoa eta bere erreferentea markatuta daude. Erreferentea bi elementuk osatzen badute kortxete artean azalduko da. Adib.: [*Ben Amor*]$_i$.

Anaforaren fenomenoak alderdi bat baino gehiago hartzen ditu bere baitan. Batetik fenomeno semantiko bat dela esan dezakegu, perpausen eta hitzen arteko semantikarekin zerikusia duelako. Bestetik, diskurtso mailako fenomenotzat har dezakegu, diskurtsoa lotzeko erabiltzen diren elementuekin harreman estua baitu. Halere, autoreen arteko desadostasuna nabarmena da anaforari dagokionez, Kleiberrek [14] dioen moduan: «*Désaccord sur la définition même du phénomène anaphorique, désaccord encore sur la façon de concevoir les processus d'interprétation référentielle et sur le statut des mécanismes d'interprétation (…)*» Desadostasun hauek anaforaren definiziotik hasiz interpretazio prozesu eta mekanismoetaraino heltzen dira. Definizioak aztertzerakoan arlo bakoitzak duen ikuspegia erabakigarria izan ohi da.

Lehenago azaldu dugunez, gure lana Lengoaia Naturalaren Prozesamenduaren arloan kokatzen da eta guretzat baliagarri izango diren definizioak, gaia ikuspegi konputazionaletik landu dutenak izango dira. Esate baterako Hirst-ek [11] testu edo diskurtsoan entitate bati edo gehiagori erreferentzia laburtua egiten uzten digun mekanismoa dela dio, diskurtso edo testuaren hartzaileak, zeri edo zeini erreferentzia egiten dion jakingo duen ustetan edo konfiantzan.

Ricoren ustez [24], bere tesian aipatutako gisan, anafora diskurtsoan edo testuan modu inplizitu edo esplizitu batean aipatua izan den objektu, pertsona edo egoera eta forma linguistiko baten artean ezartzen den erreferentziazko erlazioa da.

Anafora, beraz, edozein kategoria gramatikaleko unitate lexikoa izan daiteke, hau da, izena, adjektiboa, izenordaina, aditza, etab. Erreferentzia egiten dion elementuari erreferente edo aurrekari esaten zaio eta edozein unitate lexiko edo sintagma izan daiteke hau ere. Erreferentea kasu batzuetan elementu anaforikoaren aurretik agertuko da eta beste batzuetan ondoren. Horrelakoetan kataforaz hitz egingo dugu, erreferentea ondoren izango duelako.

Hurrengo puntuan ikusiko dugun bezala, lan honetan anafora fenomenoaren azterketa egingo dugu, mota ezberdinak deskribatu eta gure aztergaia mugatuko dugu.

3. Anaforaren sailkapena

3.1. Anaforaren testuinguruak

Anafora testuinguru desberdinetan eman daiteke. Pertsona batek hitz egiterakoan edo testuren bat idazterakoan, inkontzienteki anafora darabil, hitz jakin batzuk erabiltzen ditu ordura arte hitz egindakoari edo idatzita-

koari erreferentzia egiteko. Honela bada anafora testuinguru ezberdinetan eman daiteke:

— **Testuinguru situazionala**: erreferentzia egiten dion aurrekaria ez da esplizituki agertzen testuinguruan. Horiei erreferentzia keinu bidez egiten zaie, idazkeraren kasuan izan ezik.

Kontuz horrekin!

Erakusleak erreferentzia egiten dion elementua egoeraren arabera alda daiteke.

— **Testuinguru konbentzionala**: inongo elementuri erreferentziarik egiten ez dioten anaforak (exofora ere deituak). Esaldi egituratuak edo eginak izaten dira, aurrekaririk ez dutenak.

Hor konpon!

— **Testuinguru linguistikoa**: Lan honetan aztertuko dena da. Anafora testuan dagokion aurrekariarekin agertzen deneko kasua dugu hau. Aurrekaria izen sintagma bat, perpaus bat edota testu zati bat izan daiteke. Aurrekaria testuan esplizituki agertzen da:

Ane$_i$ mendira joan zen kirol pixka bat egiteko gogoa zuelako. Berak$_i$ ez zuen etxean geratu nahi.

Testuinguru honetan agertzen diren anaforak ere mota bat baino gehiagokoak izan daitezke eta zenbait faktoreren arabera sailka ditzakegu.

3.2. **Anafora linguistikoaren sailkapena**

Bada arlo honetan euskarara egokitutako lana, non sailkapena ikuspuntu diskurtsibo batetik egiten den [6], baina hasieran aipatu dugun moduan gure lana Lengoaia Naturalaren arloan kokatu dugunez, gazteleraz eta ingelesez, helburu konputazionaletarako erabili diren sailkapenak hartu ditugu eredutzat.

Horrela, puntu honetan Ferrández-ek [5] bere tesian gaztelerarako aurkezten duen sailkapenaz gain, ingelesez Mitkov-ek [18] proposatzen duen sailkapena oinarri bezala hartuko ditugu ikuspegi konputazionalari jarraituz.

Euskarara egokitutako anafora-moten deskripaben honetan, anafora osatzen duen elementua izango dugu kontuan, baita aurrekariak izango dituen ezaugarri ezberdinak ere.

3.2.1. *Anaforaren kategoria gramatikalaren araberako sailkapena*

Sailkapen honetan anafora beraren kategoria gramatikala bakarrik izango da kontuan eta ez aurrekariarena.

Pronominala. Elementu anaforikoa determinatzailea edo izenordaina izan daitezke. Hala ere, anafora bezala jokatzen duen determinatzaileak izenordain funtzioa hartuko du esaldiko beste elementuren bati erreferentzia eginez. Izenordain funtzioa betetzen duten determinatzaile hauek ematen diguten informazio lexikoa urria denez, oso murriztailea dela esan genezake eta horregatik, aurrekariak erraz detektatzeko moduan agertu behar du esaldian. Aurrekaria eta anaforaren arteko tartea ez da oso zabala izango, entzuleak bestela ez baitu jakingo esaldiko zein elementuri egiten ari zaion erreferentzia eta ez du izango izenordaina egoki interpretatzeko informazio nahikoa [2]. Adibide gisa:

[Tourrerako apostu gogorra]$_i$ egitea ona da, baina hura$_i$ ondo irteten ez bada (...)

Lengoaia Naturalaren Prozesamenduaren arloan anafora-mota hau gehien aztertu dena izan da, automatikoki detektatzeko eman dezakeen erraztasunagatik, kasu batzutan nahikoa baita informazio morfologiko eta sintaktikoa bere aurrekaria automatikoki ebazteko.

Erreferentea elementu pronominalaren ondoren etor daiteke, horrelakoetan kataforaz mintzatuko gara, hona adibidea:

Pellok hau$_i$ esan zuen, [ez zitzaiola axola etxera joatea]$_i$.

— **Deskribapen zehaztuak**[2]. Honako anafora hau izen sintagma batek osatzen du. Autore batzuen ustetan artikulu mugatu edo erakuslearekin hasiz osatzen den edozein izen-sintagma izango da. Kasu honetan aurrekaria anaforatik urrutiago egon liteke, anafora betetzen duen elementuak informazio lexikoa baitu bere baitan.

Mirenek Bilbora joan zen azkenengoan [bi aterki]$_i$ erosi zituen. [Aterki berdea]$_i$ asko gustatzen zait.

Detektatze automatiko posibleari begira, anafora honek arazoak sor ditzake elementu anaforikoak batzuetan aurrekari osoari edo eta beste batzuetan aurrekariaren zati bati bakarrik egin baitiezaioke erreferentzia.

— **Anafora adjektiboa.** Anaforatzat hartzen den izen-sintagmaren osagai nagusia adjektiboa da. Elipsi kasu bat ere izan zitekeen, adjektiboak izen funtzioa betetzen baitu, aurretiaz aipatutako izena errepikatu behar ez izateko.

---

[2] Ingelesez 'definite description' bezala ezagutzen dena.

Mirenek [gona gorri bat]$_i$ eta [berde bat]$_i$ erosi zuen.

— **Aditz-anafora**. Beste kasu honetan anaforak aditz bati edo aditz sintagma bati egiten dio erreferentzia. Askotan izenordain edo erakusle batek eta aditz laguntzaileak osatzen dute anafora.

Ixonek [gozokia lurrera bota zuen]i, Anderrek ere [hori egin]i zuen.

— **Adberbio eta osagarri zirkunstantzialak**. Lekuzko edo denborazko erreferentzia egiten dute anafora hauek. Aurrekaria halaber lekuzko edo denborazko osagarri bat izan ohi da.

Aurreko asteburuan Gasteiza$_i$ joan ginen kontzertu bat entzutera. Han$_i$ geundela aprobetxatuz, Artium ere bisitatu genuen.

— **Elipsia edo zero-anafora.** Izenak adierazten digun moduan anaforamota hau elipsiak osatzen duena da. Testuan esplizituki agertzen ez diren edozein sintagmak osa dezake eta adibideetan ikur honekin (Ø) agertuko zaigu.

Japonieraz$_i$ hitz egitea erraza da baina Ø$_i$ idaztea ez.

Orain arte azaldutako sailkapenean anafora osatzen duten elementuak izan dira aztergai. Ondorengoak aurrekariaren ezaugarri desberdinak hartuko ditu aintzat.

3.2.2. *Aurrekariaren posizioaren araberako sailkapena*

Aurrekaria eta anaforaren arteko distantziak baldintzatzen du sailkapen hau. **Esaldi barnekoa**, anafora eta aurrekaria esaldi berean daudenean izango da:

Zuzendariaren $_i$ izena ez da garrantzitsuena, hura $_i$ kontratatzeko sistema jarri baitu (…)

Esaldi artekoa aldiz, anafora eta aurrekaria esaldi ezberdinetan agertzen bazaizkigu:

Antzinako euskal dantzak taularatzen esperientzia du Urbeltzek$_i$. Harenak$_i$ dira (…)

3.2.3. *Aurrekariaren eskuragarritasunaren araberako sailkapena*

Kasu honetan, aurrekaria detektatzeko erraztasuna oinarri izanda egiten da sailkapena

— **Anafora morfosintaktikoa.** Izenak berak argitzen digun bezala, analisi morfologiko eta sintaktikoa egin ondoren ateratako emaitzen ondorioz erabakitzen da zein izango den aurrekaria. Bien arteko kasu eta numeroan ematen diren erlazioak aztertzen dira. Denetan errazen detektatzen den anafora dugu hau eta guk tratatu duguna ere hemen sartuko genuke.

Joni$_i$ ez zaizkio sagarrak gustatzen. Berari$_i$ gehien gustatzen zaio fruta marrubia da.

— **Anafora semantikoa.** Analisi semantikoan datza, hitz eta perpausen esanahiak izaten ditu kontuan. Deskribapen zehatzeko anaforak sartzen dira hemen, anaforak bere aurrekariekin zein erlazio duten jakiteko informazio semantiko gehiago behar da. Erlazio semantiko ezberdinetatik ateratako anaforak izango dira. Makina edo aplikazio batek detektatzeko zaila gertatzen da hau.

Scolak$_i$ ezingo du datorren igandeko partidua jokatu. Saskibaloi-jokalariak$_i$ aurreko partidan min hartu zuen (…)

— **Anafora pragmatikoa.** Diskurtsoan agertzen den informazio pragmatikoa behar da anafora hau ulertzeko. Testuan esplizituki agertzen ez diren informazioak jakintzat ematen ditu, testuingurutik at. Munduko ezagutza aurresuposatzen du (adibidean bezala, bi titulu horiek liburuenak direla jakin behar dugu), eta horregatik, aplikazio bati begira anafora guztietatik harrapatzen zailena da.

Titulu ezberdinetan ikusi dezakegu hau$_i$:[«Dr. Jekyll eta Mr. Hyde», «Altxorraren irla»]$_i$ (…)

Hala ere, ez da beti erraza izango anafora semantiko eta pragmatikoaren artean bereiztea. Bi arlo hauen artean beste hainbat kasuetan gertatzen dena baitugu hemen, bata bestearen arteko mugen ezarpen zaila.

3.2.4. *Aurrekari motaren araberako sailkapena*

Sailkapen hau egiteko aurrekaria nolakoa den aztertuko da eta honela bi motatakoak izango dira.

— **Anafora konkretua**. Erreferentea objektu edo pertsona konkretu bat denean, normalean izen-sintagma bat izaten da.

Toshaki$_i$ asko gustatzen zaio berak$_i$ kontrol handia izatea, eta orain (…)

— **Anafora abstraktua**. Aurrekari bezala entitate abstraktu bat duenean. Sintaktikoki ikusita, aditz sintagma ala perpaus bat edo gehiago izan daitezke aurrekari bezala jokatuko dutenak.

[Lantokian gaixo ez jartzeko eskubidea]$_i$. Hau$_i$ adibide bat besterik ez da (…)

3.2.5. *Erreferentzia motaren araberako sailkapena*

Anafora eta aurrekariaren arteko erlazioak izaten dira kontuan sailkapen honetan.

— **Anafora sakona**. Lehenago aipatu den objektu bati erreferentzia osoa egiten dio anaforak. Korreferentziazko erlazio bat eratzen da aurrekari eta anaforaren artean.

Muskarditzek$_i$ adierazi zuenez, berak$_i$ ere ez zuen espero hasieratik nagusitzea.

— **Azaleko anafora**. Honek aurrekariaren zati bati egiten dio aipu, hau da, erreferentzia partziala egiten dio aurrekariari. Anaforak diskurtso edo testuan objektu berri bat sartzen du, lehenago aipatu den objektu horrekin zerikusia duena.

Amonaren baserrian [animalia asko]$_i$ daude. Niri gehien gustatzen zaidana astoa$_i$ da.

4. Anafora pronominala euskaraz

Aurreko puntuan ikusi dugun sailkapena zehazterakoan aipatu dugun bezala, osatu nahi dugun corpusean, hasiera batean markatuko dugun anafora-mota pronominala izango da. Anafora pronominala euskaraz zer kontsideratuko dugun zehazten laguntzeko Alacanteko Unibertsitatean Cast3LB [19] corpusaren inguruan egiten ari diren lana interesgarria iruditu zaigu segidan azalduko ditugun zenbait arrazoiengatik.

Gaztelera, katalana eta euskara morfologikoki, sintaktikoki eta semantikoki markatzea helburu zuen 3LB corpus [22] orokorraren barruan garatu da Cast3LB corpusa. Azken honetan, gaztelerazko corpusaren markaketa izan dute helburu, eta pauso bat gehiago emanez anafora-mota batzuk etiketatu dituzte.

Honela, erlazio anaforikoak eta korreferentziazkoak etiketatu dituzte, elementu anaforikoak eta horien aurrekariak markatzea adostuz (kataforak

oraingoz alde batera utzita). Elementu anaforiko hauek elipsi anaforikoa eta anafora pronominala dira (aurrerago ikusiko ditugun gaztelerazko izenordain tonikoak eta atonikoak, baita izenordain erakusleak ere). Unitate hauek korreferentziazko kateak osatzen dituzte eta markatuak egon behar dute testuaren kohesioa eta koherentzia erakusteko.

Markaketarako erabili duten sailkapena oinarritzat hartuz zerrendatutako adibideetara joko dugu. Adibide hauek lagunduko digute euskarazko corpusean zer markatu behar genukeen erabakitzen:

1. Izenordain pertsonal klitikoak: gaztelerazko *lo, la, le* eta *los, las, les*.

   Ana abre [la verja]$_i$ y [la]$_i$ cierra tras de sí.

   Anak hesia$_i$ irekitzen du eta bere ondoren ixten duØ $_i$.

   Adibide hau euskarara itzultzerakoan ez dugu pareko egiturarik aurkitzen, izenordain pertsonal klitikorik ez baitago. Euskaraz klitikoen pareko egitura, aditz jokatuaren barruan artizki gisa agertzen da [16], adibidean ikusten dugun bezala. Bestela, elipsia erabil daiteke, objektu lexikoaren agerpena ez baita nahitaezkoa euskaraz. Corpusa etiketatzeko lehen fase honetan ez dugu horrelakorik markatuko.

2. Izenordain bihurkariak.

   Ana$_i$ abre la verja y la cierra tras de sí$_i$.

   Anak$_i$ hesia irekitzen du eta bere$_i$ ondoren ixten du.

   Adibide honen kasuan, gazteleraz izenordain bihurkariak agertzen dira eta euskaraz badiren arren, ez dira modu berean erabiltzen. Esaldi hori *bere* izenordainarekin postposizio konplexu bat osatuz itzuli dugu. Honelakoak aurrerago ikusiko dugun bezala, corpus honetan bertan baina hurrengo pausu batean markatuko ditugu, ez oraingoan.

3. Isileko izenordainak (eliptikoak).

   Ana$_i$ abre la verja y ø$_i$ la cierra tras de sí.

   Anak$_i$ hesia irekitzen du eta ø$_i$ bere ondoren ixten du.

   Isileko izenordainari dagokionez, gazteleraz bezalaxe euskarazko itzulpenean ere izenordaina isildu egiten dugu, ez da esplizituki agertzen esaldian. Hauek ez ditugu etiketatuko lehen pauso honetan. Kategoria lexikoak hartuko ditugu kontuan, elipsia oraingoz alde batera utzita.

4. Preposizio-sintagma (preposizio + IS) bat osatzen EZ duten izenordain pertsonalak: *él, ella, ello,* eta *ellas, ellos*.

   Andrés$_i$ es mi vecino. Él $_i$ vive en el segundo piso.

   Andrés$_i$ nire bizilaguna da. Bera$_i$ bigarren pisuan bizi da.

5. Preposizio-sintagma baten barnean dauden izenordainak, preposizio + *él, ella, ello* izango liratekenak.

   Juan$_i$ debe asistir pero Pedro lo hará por él$_i$.

   Juanek$_i$ joan behar du baina Pedro joango da bere$_i$ ordez

   4. eta 5. puntuei dagokienez, itzulpena egiterakoan euskaraz ere hirugarren pertsonako izenordaina erabiltzen dugu. Genero ezberdintasunik ez dagoenez, bai femeninorako, bai maskulinorako *bera* izenordaina hartuko genuke.
   Oinarritzen ari garen sailkapenean izenordain pertsonalak preposizio sintagma osatzen duten ala ez hartzen dute kontuan multzoak osatzeko garaian. Lehenengo multzoa sintagma berak bakarrik osatzen duten izenordain pertsonalak izango dira eta bigarren multzoa preposizio-sintagma bat osatzen duten izenordainak dira. Guk oraingoz, ez dugu antzeko ezberdintasunik egingo, kasu gramatikalen eta postposizioen arteko bereizketarik ez dugu egingo lehen pauso honetan.

6. Izenordain erakusleak, preposizio-sintagma bat osatzen EZ dutenak:

   El Ferrari$_i$ ganó al Ford. Éste$_i$ es el mejor.

   Ferrariak$_i$ Ford-ari irabazi zion. Hura$_i$ zen hoberena

7. Izenordain erakusleak, preposizio-sintagma batean daudenak.

   Ana vive con Paco$_i$ y cocina para éste$_i$ diariamente.

   Ana Pacorekin$_i$ bizi da eta harentzat$_i$ egiten du jana egunero.

   Esaldi hauen euskarazko baliokideak determinatzaile erakusleen (*hau, hori, hura)* bitartez itzuli ditugu.
   Izenordain erakusleen kasuan ere bi multzo egiten dituzte eta gaztelerazko sintagma preposizionalarekin guztiz baliokideak izan ez arren adibide hauekin guztiekin burura datorkigun beste azterketa posible bat hau da: determinatzailea kasuaren arabera sailkatzea, erreferentearen agerpenean ezberdintasunik izango litzatekeen aztertuz.

Erkaketa honen ondorioz gure corpusean lehenengo fase batean markatuko duguna zehazten joan gintezke: batetik *hau, hori, hura* determinatzaile erakusleak; eta horien pluraleko formak, *hauek, horiek, haiek* hartzen dituzten kasu guztietan. Hauek guztiak noski, izenordain funtzioa betetzen ari direnean, markatzen dugun elementua bakarrik doanean eta aurreko zerbaiti erreferentzia egiten dionean.

Itzulpenetan agertu zaizkigun 3. pertsonako izenordain pertsonalak *bera* singularreko forma eta pluraleko *beraiek (berak, eurak*), aztertu ditugu baina

hurrengo fase batean etiketatuko ditugu, elementua anaforikoaz gain aurrekariaren azterketa egin nahi dugulako.

Corpus honen markaketan etiketatuko ditugun elementuak zehaztu ondoren, elementu hauei euskal gramatiketan eman zaien deskribapen laburtua ekarriko dugu hurrengo puntuan.

4.1. **Anafora Euskal Gramatiketan**

Ondorengo lerroetan, euskal gramatika ezberdinetan gure aztergaia izango denak zein tratamendu duen ikusiko dugu. Batetik *hau, hori, hura/ hauek, horiek, haiek* determinatzaile erakusleak eta bestetik *bera/ beraiek* izenordain indartuei erreparatu diegu.

4.1.1. *Euskal Gramatika Laburra [3][4]*

Euskaltzaindiak *hau, hori, hura*, singularrean eta *hauek, horiek, haiek*, determinatzaile kontsideratzen ditu eta hiru gradu desberdinak aipatzen ditu. Haien funtzioa erakustea (hitzak berak esaten duen bezala) edo seinalatzea denez, hiru gradu bereizten dira, bai denboran baita espazioan ere. Honetaz gain, gradu hauek balio nozionala ere badutela dio, leku eta denboratik at egonda kontaketan eta elkarrizketan erabiltzen dena.

Gramatika honetan erakusle indartuak *hauxe, horixe, huraxe* kontsideratzen dituzte bere forma guztietan, erakusle arruntei *-xe* atzizkia gehituz osatzen direnak. Erakusle arruntei *«ber-»* aurrizkia gehituz gero maiz izenordain funtzioa hartuko duten erakusle indartuak izango ditugu: *bera, berau, berori*.

4.1.2. *Euskal Gramatika Osoa [25]*

Determinatzaileen atalean erakusle arruntak ditugu, hiru hurbiltasun gradu erakusten dutenak eta hona gertuenetik urrunenera, *hau, hori, hura*, singularrean eta *hauek, horiek, haiek*, pluralean.

Erakusle indartuen artean, batetik erakusle arruntei *-xe* atzizkia gehituta osatzen diren *hauxe, horixe, huraxe* ditugu deklinabide forma guztiekin eta bestetik maiz izenordain gisa erabiltzen diren *berau, berori, bera* erakusle arruntei *ber-* aurrizkia gehituz. Anafora esplizituki aipatzen ez badute ere, hona hemen zer esaten duten, bere ezaugarrietan bai aipatzen dute aurrekoari erreferentzia egiten diotela esanez:

> *»Forma hauek erakusle gisa erabil badaitezke ere, gehienetan izenordainaren funtzioa betetzen dute eta aldez aurretik aipatutako pertsona edo objektu bati erreferentzia egiteko erabili ohi dira.»*

4.1.3. *Grammar of Basque [16]*

Interneten ikusgai dagoen gramatika honek, beste gramatiketan dagoenaren antzerakoa dio *hau, hori, hura-* ren gradu ezberdintasunari buruz. Halaber puntu bat litzateke interesgarria:

> *«There are no distinct forms for third person pronouns in Euskara, and demonstratives are used as third person pronominals»*

Beste gramatiketan izenordain indartutzat hartzen dituztenei buruz, Lakak hurrengo hau dio:

> *«They are used anaphorically, that is, when the entity they referred to is already known in the discourse (...) The third one in the series, bera, is very frequently used as third person pronoun, alternating with the third demonstrative hura».*

4.1.4. *A Grammar of Basque [12]*

Liburu honetan, erakusle arruntak *hau, hori, hura* direla esaten digu. Hurbiltasunaren hiru graduak igorlea eta jasotzailearen gertutasunaren araberakoa dela kontsideratzen ditu.

> *»In general, the first indicates proximity to the speaker, the second proximity to the addressee, and the third remoteness from both, though on occasion hori and hura merely indicate differing degrees of remoteness from the speaker.»*

Izenordainak deskribatzen dituen puntuan zera dio:

> *»In general, Basque lacks true third-person pronouns, and demonstratives are used when third-person pronouns are required for thematic purposes».*

Lakak [16] bere gramatikan esaten duenaren antzera, izenordain pertsonal indartuak sortzen dira *hau, hori* eta *–a*-ri *ber-* aurrizkia gehituz gero: *berau, berori* eta *bera*.

Ezaugarri nagusiak laburbilduz, gramatika guztiek aipatzen dute determinatzaile erakusleak euskaraz izenordain funtzioa bete dezaketela. Horrez gain, anaforak duen berrartzearen funtzioaren berri ematen zaigu esku artean izan ditugun gramatiketan. Honela, aurreko puntuan adibideetan itzultzeko hartutako bidea egokia izan daitekeela berretsi digute.

Ondorengo puntuan, gramatiketan izan ezik, euskal hizkuntza aztergai izan duten gai honen inguruko lanei erreparatu diegu.

4.2. **Anafora beste lanetan**

Erlazio anaforikoak izan ditzaketen determinatzaile eta izenordainen deskribapenak izan ditugu hizpide aurreko puntuan eta hau osatzeko euska-

raz anaforaren inguruan aztertu dena begiratu dugu García Azkoagaren [6] eta Garzia Garmendiaren [8] artikuluak aztertuz.

4.2.1. *Anafora argudiozko testuetan [6]*

Argudiozko testuetatik abiatzen da anaforaren azterketa egiteko. Testua linguistikoki egituratutako mezu bat duen ekoizpena den heinean bere destinatarioarengan eragin koherentea izatea bilatzen du. Helburu hau lortzeko garrantzitsua da testuak barne-antolakuntza bat izatea, horretarako testu ekoizlea testualizazio mekanismoez baliatzen da. Mekanismo horietako bati 'izen-kohesioa' deritzo eta honen barruan gertatzen den fenomenoetako bat anafora da.

Anaforaren fenomenoa ikuspuntu zabalago batetik hartzen du, definizio klasiko batetik abiatuz: *«anafora aipatutako segmentu bati erreferentzia egiten dion adierazpen bat izango da, eta segmentu hori aurretik (anafora) edo atzetik (katafora) ager daiteke»* [17]. Anafora, beraz, diskurtsoan bi elementuren artean sortzen den erlazioa da.

Erlazio anaforikoa menpekotasun interpretatibo eta diskurtsibo bezala abiapuntutzat hartuz, idazle trebatu zein ikasleek idatzitako lau adibideetan oinarrituta adierazpide anaforiko nabarienen azterketa egiten du. Ondorio orokor batzuk ateratzearekin batera, sailkapen bat egin du. Bi taldetan banatzen ditu testu horietan aurkitzen dituen anaforak: *«uztartuak eta ez-uztartuak; lehenengokoak sintaxiari lotuta daude eta euren funtzionamendua esaldi mailakoa da. Anafora ez-uztartuak edo askeen funtzionamenduak, aldiz, esaldiaren esparrua gainditzen du eta orduan esaldi kanpokoak diren fenomeno anaforikoen aurrean aurkituko gara.»*

Diskurtsoan anafora askeak betetzen duen papera garrantzitsua denez hauen sailkapen bat egiten du. Honenbestez, gure aztergaitik gehien urruntzen diren anafora-moten azterketa egiten du, esaldiaren esparrua gaindituz, diskurtsoan ematen diren elementu anaforiko orokorragoen azterketa eginez.

4.2.2. *Hura, bera eta abarren adar gehiago [8]*

Determinatzaile eta izenordainen erabilera egokiari buruz hitz egiten du, itzultzaileek beste hizkuntzetan agertzen diren izenordainak itzultzerakoan aurkitzen dituzten zailtasunen aurrean laguntzeko asmoz. Deskribapen honetatik, euskaraz anaforaren fenomenoarekin lotuta dauden determinatzaile eta izenordainak nola erabiltzen diren aztertu (literaturako zenbait autoreren adibideen bitartez) eta nola erabili behar liratekeen gomendatzen du.

> Autorearen hitzetan, *hau, hori, hura,* erakusleen arteko ezberdintasuna: *«Oinarrizko oposaketa hiztunaren kokalekutiko hurbil-urruntasunari dagokio, eta korrelazio aski estuan dago, beraz, pertsona gramatikalekin (nahiz hirurak egokitu daitezkeen, jakina, hirugarren pertsonekin): ni>hau, hi>hori, hura>hura. Denboran, berriz, lehen gradua (hau) presenteari dagokio eta hirugarrena (hura) iragan nahiz etorkizunari (bigarren graduak (hori) eginkizun gutxiago du oinarrizko kokapen horretan)».*

Bere esanetan, diskurtsoan erabili den erreferente bat errepikatzen dela adierazteko (hura osorik ez errepikatzeko, alegia, hura *ordezteko)*: funtzio horri deritza anafora, eta erakusleek diskurtsoan duten funtzio guztiz garrantzizkoa da anaforikoa.

Labur-labur esanda hiru funtzio elkargaintzen dira erakusleetan: «*deixi arrunta, (kanpoko) anafora, eta diskurtso (edo testu) barneko deixia (anafora diskurtsibo-testuala), bada beste funtzio diskurtsibo-testual bat, kataforiko deritzana eta erakuslearen erreferentea diskurtsoan lehenago agertu ordez ondoren etorriko da*».

Funtzio anaforiko-kataforikoetan ere badute beren espezializazioa erakusleen hiru graduek eta honela erabili beharko lirateke:

1. *HAU:* lehen gradua, kataforetarako, bi puntuen ondoko azalpena aurrez ordezten duenean. Situazioan ere hurbiltzat jotzen den zerbaitekin, lehen pertsonarekin, orainaldiarekin lotua dagoenean.
2. *HORI:* bigarren gradua, diskurtsoan aurreko zati bat seinalatu eta ordezteko (diskurtso barneko anafora), ordeztu nahi dena erreferentzia argia bada.
3. *HURA:* hirugarren gradua, diskurtsoak adierazten duen munduko erreferente baten ordezkapenerako (kanpoko anafora, arrunta).

Bi artikulu hauetatik ateratzen ditugun ondorioak desberdinak dira. Lehenengoan, García Azkoagak [6] ematen digun sailkapena interesgarria den arren, esaldiaz gaindiko sailkapena den heinean oraingoz ez dugu gurearekin bateratuko, lehenik eskuz eta etorkizun batean konputazionalki markatzeko kontuan hartzeko ezaugarriak urruti geratzen baitira. Garzia Garmendiak [8] bere artikuluan guk egin nahi dugun korpusaren osaketan lagunduko digu, batez ere anaforaren aurrekaria errazago detektatzeko orduan, erabiltzen den erakuslearen arabera, lehenago edo ondoren egon baitaiteke erreferentea.

Azterketa hau guztia egin ondoren, oinarri batzuk finkatu nahi izan ditugu hurrengo atalean ikusiko dugun corpusa osatzeko.

5. Corpusaren osaketa

Euskaraz anaforikoki markatutako corpusa osatzen hasteko, euskal gramatiketan eta fenomeno honen inguruko beste lanak aztertzeaz gain, corpusaren osaketan beste hizkuntza batzuetan zein lan egin diren ikusiko dugu.

5.1. **Corpus etiketatuen aurrekariak**

Badira zenbait urte Lengoaia Naturalaren Prozesamenduan anafora eta korrereferentziaren azterketarekin dihardutela. Hala ere ezin esan corpus asko daudenik anaforikoki edo korreferentzialki etiketatuak. Anaforaren ebazpenerako corpusaren beharra Mitkov-ek (2002) ondo adierazten du.

> *«The annotation of corpora is an indispensable, albeit time-consuming, preliminary to anaphora resolution (and to most NLP tasks or applications), since the data they provide are critical to the development optimisation and evaluation of new approaches.»*

Liburuan corpusa osatzerakoan jarraituko beharreko irizpide batzuk azpimarratzen ditu:

1. Etiketatze eskema espezifiko baten beharra.
2. Etiketatzen lagunduko duen tresna bat.
3. Etiketatzaileen arteko adostasunaren garrantzia.

Erlazio anaforikoak edo korreferentziazkoak markatuta dituzten zenbait corpus ezagutzen ditugu nahiz eta guztien arazoa bera izan.

Lancaster-eko unibertsitatean dugu anaforikoki etiketatua dagoen corpus bakarrenetakoa: Lancaster Anaphoric Treebank (UCREL) [7]. 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 probabilistikoa trebatzea izan zen. Kohesiozko erlazio ezberdinak etiketatzeko aukera ematen du, elipsia kontuan izanaz. Erreferentea non dagoen adierazten digu, alegia, elementu anaforikoa baino lehenago agertzen den edo ondoren, anafora edo katafora den esanez. Horretaz gain, etiketa bat ere badute, elementu anaforiko eta erreferentearen artean zein erlazio semantiko ematen den adierazteko.

Baditugu bestelako testuak korreferentzialki etiketatuak; MUC Coreference Task [9] corpusa egunkarietako testuek osatzen dute. Anafora eta bere aurrekaria, eta horien arteko erlazioa markatzen dute. Halere izen-sintagmak bere aurrekariarekin identitatezko erlazioa bakarrik duenean adierazten dute. 65.000 hitz inguru korreferentzialki markatutako corpus honen helburuetako bat, anaforaren ebazpenerako algoritmorako trebakuntza eta ebaluaziorako baliagarri izatea da. Honetaz gain, informazio erauzketarako sistemetan ere erabili ahal izan dute.

Wolverhampton unibertsitatean, aurrekoaren antzeko eskema (MUC) jarraituz 60.000 hitzeko corpusa osatu dute. Markatutako testuak zenbait tresnen eskuliburuetakoak dira. Corpusaren markaketarako ClinKa [21] 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 hartuz anafora eta aurrekariak identifikatzen ditu eta euren arteko korreferentzia erlazioak markatzen.

Hona ekarri ditugun orain arteko corpusen adibideak hizkuntza ingelesa dute aztergai baina badira beste hizkuntza askotan korreferentzialki markatutako corpusak. Alemanerako adibidez, TIGER proiektuan [15] markaketa morfologiko, sintaktiko eta semantikoaz gain korreferentzia erlazioak ere markatu dituzte. Antzera egin dute Pragako unibertsitatean [10] corpusa pragmatikoki markatzeko saiakera eginez.

Gaztelerarako berriz, Alacanteko Unibertsitatean arlo honetan egiten ari diren lana ere aipatu behar da. Hauek anaforaren ebazpen automatikoan lortutako emaitzak hobetzeko asmoz osatzen ari dira anaforikoki etiketatutako corpusa [19], [20].

Aurreko puntuetan anaforikoki etiketatutako corpusa osatzen joateko beharrezkoa izango zaigun informazioa biltzearekin batera aztergaia mugatu eta finkatu dugu. Atal honetan anaforaren fenomeno corpusean nola markatuko dugun eta zein baliabide izan ditugun azalduko dugu. Gerora tresnek anaforaren fenomenoa automatikoki errazago detekta dezaten markatutako corpus hau ezinbestekoa izango da.

5.2. Euskarazko corpusaren osaketa

Lehen esan bezala, sailkapenean azaldutako **anafora pronominala** markatuko dugu. Ikuspegi konputazionaletik, detektatzen errazena gerta daitekeelako eta etorkizun batean etekin handiena beroni atera diezaiokegulako, lehenengo pauso honetan anafora-mota hau markatzeko erabakia hartu dugu.

Corpusa osatzen hasteko IXA taldean garatutako zenbait tresna informatikoz baliatu gara. Eskura dugun corpusa, euskal egunkari bateko 2001. urteko testuez osatzen da eta IXA taldeko analizatzaile morfologiko eta sintagma zatikatzaile automatikoaren bidez analizatua egongo da. Guk horren gainean lan egingo dugu, eta gure eskuzko markaketa erraztearren, aztergaia mugatuko duen tresna informatiko batez baliatu gara. Ondorengo adibidean dakusagun moduan eskura izango ditugun testuek sintagmak markatuak izango dituzte, horrela errazago izango baitzaigu gure lana aurrera eramatea. Testu gordinak analizatzaile morfologikotik pasa ondoren sintagma-zatikatzaileaz baliatuz, izen-kateak eta aditz-kateak markatuak izango ditugu:

Ben Amor ere ez da Mundiala amaitu arte etorriko Irunera, honek ere Tunisiarekin parte hartuko baitu Mundialean.

```
/<Ben>/<HAS_MAI>/
(«» /Ben/ IZE IZB DEK DEK ABS MG ENTI_HAS_PER @IZLG> %SIH)
/<Amor>/<HAS_MAI>/
(«» /Amor/ IZE IZB PLU- DEK ABS MG ENTI_BUK_PER @SUBJ %SIB)
/<ere>/
(«ere» LOT LOK EMEN @LOK)
/<ez>/
(«ez» PRT EGI @PRT %ADIKATHAS)
/<da>/
(«izan» ADT A1 NR_HU @+JADNAG %ADIKATBU)
/<Mundiala>/<HAS_MAI>/
(«Mundiala» IZE IZB PLU- DEK ABS MG ENTI_PER @OBJ @PRED %SINT)
/<amaitu>/
(«amaitu» ADI SIN AMM PART @-JADNAG %ADIKAT)
/<arte>/
(«arte» IZE ARR DEK ABS MG @OBJ @PRED %SINT)
/<etorriko>/
(«etorri» ADI SIN AMM PART DEK NUMS MUGM DEK GEL DEK ABS MG @PRED
%SINT)
/<Irunera>/<HAS_MAI>/
(«irun» ADI SIN AMM PART DEK NUMS MUGM DEK ALA @ADLG %SINT)
/<,>/<PUNT_KOMA>/
/<honek>/
(«hau» DET ERKARR NUMS DEK ERG NUMS MUGM @SUBJ %SINT)
/<ere>/
(«ere» LOT LOK EMEN @LOK)
/<Tunisiarekin>/<HAS_MAI>/
(«Tunisia» IZE LIB PLU- DEK SOZ MG ENTI_LOC @ADLG %SINT)
/<parte>/
(«parte» IZE ARR DEK ABS MG @OBJ @PRED %SINT)
/<hartuko>/
(«hartu» ADI SIN AMM PART ASP GERO @-JADNAG %ADIKATHAS)
/<baitu>/
(«*edun» ERL MEN KAUS ADL A1 NR_HU NK_HU @+JADLAG_MP %ADIKATBU)
/<Mundialean>/<HAS_MAI>/
(«mundial» ADJ IZO DEK NUMS MUGM DEK INE @ADLG %SINT)
/<.>/<PUNT_+PUNT>/
```

Izen-kateei dagozkien etiketak hauxe adieraziko digute:

%SINT → Hitz bakar batek sintagma osatzen duenean.
%SIH → Sintagmak osagai bat baino gehiago duenean, sintagma-katearen hasiera adierazten diguna.
%SIB → Osagai bat baino gehiago duten sintagmen bukaera markatzeko.

Adibidean ikus dezakegun moduan, *hau* determinatzailea izenordain funtzioa ari da betetzen eta berak bakarrik osatzen du izen-sintagma osoa, horregatik %*SINT* batez markatua dago. *[Ben Amor]* izen berezian ikus dezakegu, sintagma hasiera markatzeko *%SIH* darabilela eta sintagmaren bukaera adierazteko *%SIB*.

Gure lanari helduz eta orain arte egindako azterketaren ondorio gisa, goran aipatu bezala corpus honetan lehenengo fase batean markatu beharreko elementuak hurrengoak izatea erabaki dugu:

— Determinatzaile erakusle arruntak *(DET ERKARR)*[3]:

Singularrean: *hau, hori, hura*
Pluralean: *hauek, horiek, haiek*

Ondorengo fase batean berriz, beste hauek hartuko genituzke kontuan:

— Determinatzaile erakusle indartuak (*DET ERKIND*):

Singularrean: *berori, bera*
Pluralean: *berauek, beroriek*

— Izenordain pertsonal indartuak (*IOR PERARR*):

Singularrean: *berau*
Pluralean: *beraiek*

Orokorrean, izenordainei anafora marka gehitzerakoan ez genuke arazo handirik izango, haiek bakarrik sintagma bat osatzen baitute *(%SINT)*.

Determinatzaileekin aldiz gauzak bestelakoak izango dira, testu batean determinatzaileak ez dira beti anafora izango, honela, izen baten determinatzaile bezala agertzen zaigunean *(gizon hau)* sintagma baten bukaera markatuko du *(%SIB)* eta ez da anafora izango. Baina determinatzaileak berak bakarrik sintagma osatzen duenean *(%SINT),* hau da, izenordain funtzioa betetzen ari denean, anafora kontsideratuko dugu. Kasu guztietan aurrekaria bilatzen saiatuko gara eta izenordaina baino lehenagoko esaldietan aurkitzen ez badugu, izenordainaren ondoren datozen esaldietan begiratuko dugu, erreferentea ondoren ere etor daitekeelako.

5.3. **Markatze-sistema eta erreferentearen identifikatzea**

Guzti honen eskuzko etiketatzea errazteko aipatutako euskal egunkari baten corpusean oinarrituta eta baliabide informatiko baten laguntzaz, ele-

[3] IXA Taldeko kategoria sistemaren arabera

mentu bakoitza markatuko dugu. Anafora edo katafora ager dakigukeenez, kontuan hartu beharko dugu, erreferentea elementu anaforikoa baino lehen zein ondoren etor daitekeela. Hori dela eta, erreferentea detektatzeko aurreko hiru esaldi eta ondorengo bi esaldi hartuko ditugu kontuan.

Anafora pronominala elementu bakarrez osatutako sintagma izango da, eta [ANAznb] etiketa ipiniko diogu, zenbakia (znb) agerpen ordenaren araberakoa jarriaz. Honi dagokion aurrekaria edo erreferentea hitz bakar batek osatzen badu [REFznb] jarriko zaio, anaforaren zenbaki berdinarekin. Aurrekari edo erreferente hau hitz bat baino gehiagoz osatua badago, [REF-Bznb] marka erantsiko diogu eta bukaeran berriz [REF-Eznb]. Ikus dezagun hau guztia adibide batekin:

Beste gauza bat da PSOEk_i ahal izatea, edo nahi izatea, edo PPk honi_i uztea.

```
/<PSOEk>/<DEN_MAI_DEK>/[REF11]
(«PSOE» SIG DEK ERG MG ENTI_ORG @SUBJ %SINT)
(…)
/<honi>/[ANA11]
(«hau» DET ERKARR NUMS DEK DAT NUMS MUGM @ZOBJ %SINT)
```

Analisiak erakusten digun moduan *hau* determinatzaileak berak bakarrik osatzen du sintagma eta bere aurrekaria osagai bakar batekoa denez [REFznb] marka du.

[Ben Amor]$_i$ ere ez da Mundiala amaitu arte etorriko Irunera, honek_i ere Tunisiarekin parte hartuko baitu Mundialean.

```
/<Ben>/
/<HAS_MAI>/[REF-B5]
(«» /Ben/ IZE IZB DEK DEK ABS MG ENTI_HAS_PER @IZLG> %SIH)
/<Amor>/<HAS_MAI>/[REF-E5]
(«» /Amor/ IZE IZB PLU- DEK ABS MG ENTI_BUK_PER @SUBJ %SIB)
(…)
/<honek>/[ANA5]
(«hau» DET ERKARR NUMS DEK ERG NUMS MUGM @SUBJ %SINT)
```

Oraingoan *hau* determinatzaile erakusleak sintagma bat osatzen du eta erreferentzia egiten dion entitatea, hau da, aurrekaria *Ben Amor* izen berezia izango da. Erreferenteak bi osagai dituenez, hasieran [REF-Bznb] jarrioko diogu eta bukaeran [REF-Eznb].

Bada beste kasu bat ere, gutxiagotan agertu arren, beste etiketa bat jartzea erabaki duguna, kataforarena hain zuzen. [CATznb] marka erabiliko dugu

kataforaren papera betetzen ari den elementua markatzeko eta ondoren bilatu beharko dugun errefenteak [CAT-REFznb] etiketa izango du. Erlazio kataforiko hauetan ere ondoren datorren erreferenteak osagai bat baino gehiago izan dezake eta orduan hasieran [CAT-REF-Bznb] marka gehituko diogu eta erreferentziak amaiera non duen jakiteko [CAT-REF-Eznb]. Adibidez:

Gainera, eta hau$_i$ da beharbada garrantzitsuena, [zehatz-mehatz azaldu beharrekoa beraien asmoa da]$_i$.

```
/<hau>/ [CAT63]
(«hau» DET ERKARR NUMS DEK ABS @SUBJ)
(…)
/<zehatz-mehatz>/ [CAT-REF-B63]
(«zehatz-mehatz» ADB ADOARR @ADLG %SINT)
(…)
/<beraien>/
(«beraiek» IOR PERARR NUMP HAIEK DEK GEN @IZLG> @<IZLG %SIH)
/<asmoa>/
(«asmo» IZE ARR DEK ABS NUMS MUGM @SUBJ %SIB)
/<da>/ [CAT-REF-E63]
(«izan» ADT A1 NR_HU @+JADNAG %ADIKAT)
/<.>/<PUNT_PUNT>//<.>/<PUNT_PUNT>/
```

Sailkapenean ikusi dugunez, anafora eta kataforetan, badira kasuak, non aurrekaria gertueneko esaldietan agertzen ez den. Esku artean dugun tartean ez badugu erreferente garbirik ikusten [?ANAznb] galdera ikurra gehitu diogu etiketari, gertueneko esaldietan aurrekaririk ez dugula aurkitu argi gera dadin.

50.000 hitz inguruko corpusa osatu dugu eta markaketarekin hasteko, *hau, hori* eta *hura* determinatzaile erakusleekin hasi gara. Lehen urrats honetan honako eragozpen eta zalantzazko puntu hauek topatu ditugu:

— Kohesiozko elementuak: *hau da, harekin eta honekin, honetaz gain, hau honela izanik, horren ondorioz,* etab. Batzuetan badute erreferente argi bat testuan baina badira kasuak, diskurtsoan kohesio funtzioa betetzen ari direnak, aurretik esandako guztiari erreferentzia eginez.
— Esaldi kopulatiboetan, hau da, perpaus nagusian *izan* aditza dagoenean, benetako anaforen aurrean ote gauden zalantzan jartzen dugu. Adibidez:

Jolasa$_i$ baita hau$_i$

Argi dago *hau* izenordainak aurreko *jolasa* izenari egiten diola erreferentzia baina izen horren garrantzia indartzeko bakarrik agertzen ote den ere pentsa daiteke.

— Generoa duten beste hizkuntza batzuekin alderatuz gero (gaztelera, frantzesa, alemana, …), aurrekaria aukeratzeko garaian batzuetan erabakigarria izan daitekeen generoaren laguntzarik ez dugu izango euskaraz [14].

Hasiera batean, kohesiozko elementu hauek nahiz esaldi kopulatiboak kasu berezitzat hartuko ditugu eta aparteko tratamendua emango diegu.

5.4. **Zenbait ondorio**

Corpus honetako anaforen eta hauen erreferenteen hainbat berezitasun aipatuko ditugu oinarrizko ezaugarri hauek kontuan hartuz:

— Anafora eta bere erreferentearen arteko tartea.
— Aurrekariaren osagai kopurua.
— Erreferentearen egitura sintaktikoa (izena, perpausa…)
— Izen-sintagmen kasuan, anaforaren deklinabide kasuarekiko eta numeroarekiko bateragarritasuna.

Ezaugarri hauei erreparatuz, bakoitzaren berezitasunak aipatuko ditugu segidan:

Hau, izenordain gisa 177 aldiz agertu da 50.000 hitzeko corpusean. Erreferentearen posizioari dagokionez, % 86an erreferentzia egiten ari zaion elementua anaforaren aurretik agertzen da testuan, eta gainerantzean, % 14an, kataforaren fenomenoa dugu, erreferentea elementu anaforizatzailearen ondoren agertuz[4]. Erreferentziako elementu hori gainera, kasu askotan pertsona edo izaki konkretu bat da Anafora eta aurrekariaren arteko distantziari dagokionez % 59an esaldi berean agertzen da. Bestela, aurreko esaldian etorriko da eta oso kasu gutxietan bi edo hiru esaldi lehenago. Aurrekariaren egitura aintzat hartuz, izen-sintagma ala perpausa den aztertzen badugu, % 73an izen-sintagma osatuz agertzen da eta geratzen den % 27an perpausak osatuz agertzen da. Determinatzaile honek maizen betetzen dituen ezaugarriak dituen adibidea ekarri dugu hona:

> *[Ben Amor]$_i$ ere ez da Mundiala amaitu arte etorriko Irunera, honek$_i$ ere Tunisiarekin parte hartuko baitu Mundialean.*

Hori determinatzailea 50.000 hitzeko corpusean izenordain anaforiko gisa 251 aldiz agertu da. Ia kasu guztietan, % 99an, anafora da eta determinatzaile honetan aurrekaria esaldi berean, % 50ean, zein lehenagoko esaldian kokatzen da, % 47an. Aurrekariaren egiturari begiratuz gero, aurrekari guztiak perpausak dira eta nahiko adierazgarri gertatzen da, pertsona edo elementu

[4] *hau* izenordaina da kataforan gehien agertzen dena.

konkretu bati baino, orokorragoa den zerbaiti egiten diola erreferentzia, *hau* determinatzailearen kasuan ez bezala. Ikus dezagun hau adibide batekin:

> *[Errepide beltzenak ohikoak izan dira 2000. urtean ere]$_i$: (…). Horrek$_i$ esan nahi du (…)*

Hura, berriz, esku artean dugun corpusean 321 aldiz agertu da. Nabarmena da determinatzaile hau ia kasu guztietan izenordain anaforiko gisa agertzen dela, % 98an, hau da, erreferentea beti elementu anaforikoa baino lehen agertzen da eta ez da kataforarik ematen. Aurrekaria esaldi berean agertzen da usuen, % 64an, eta aurreko esaldian agertzea ere nahiko normala gertatzen da, % 33an. Aurrekarien agerpen guztiak izen-sintagma egiturakoak dira, eta beste determinatzaileetan baino gehiagotan ikusi dugu erreferentea izen berezi batek osatzen duela.

> *Eta kasu horretan jeneralari$_i$ desmen-egitea egotziko litzaioke eta Guzman epaileak hura$_i$ prozesatu eta atxilotu ahal izango luke.*

Orokorrean, esan genezake, kasu gehienetan erreferentea elementu anaforikoaren aurretik agertu ohi dela. Beste hizkuntzetan gertatzen denaren antzera kataforaren kasua, hau da, erreferentea ondoren etortzea, oso gutxitan ematen den fenomenoa da. Euskaraz, fenomeno hau ia beti *hau* determinatzailearekin ematen da. Distantziari dagokionez, hiru kasuek antzeko jokaera erakusten dute, aurrekaria esaldi berean agertzearena, alegia. Aurrekariaren egitura sintaktikoari begira, berriz, diferentziak agertzen dira. *Hau* izenordainaren kasuan, aurrekari gehienak izen-sintagmak dira eta pertsona edo elementu konkreturen bati egiten diote erreferentzia, berau hiztunarengandik gertu dagoen elementu bat izaten da. *Hura* izenordainak aurrekari bezala guztiak ditu izen-sintagma eta oro har, erreferentziatzat duena pertsona bat izaten da. Guztiz alderantzizkoa gertatzen da *hori* izenordainarekin, izan ere, kasu guztietan erreferentziaren egitura sintaktikoa perpausa da, eta pertsona edo gauza ez den zerbaiti egiten dio erreferentzia, askotan, diskurtsoan lehenago aipatu den ekintza edo ideiaren bati.

Aurrekaria izen-sintagma denean, hiru kasuetan elementu anaforikoaren eta erreferentearen kasu-konkordantzia aztertu dugu. Esku artean dugun laginean, elementu anaforikoa absolutibo kasuan doanean bakarrik etortzen dira bat kasuan bi elementuok. Anafora beste kasuren batean agertzen denean ere, aurrekaria gehienetan absolutiboa izaten da. Adibidez:

> *(…) gainera, ez du hartu nahi izan [erakundea bera]$_i$, nahiz eta honek$_i$ abenduaren 11ko 372/2000 Foru Dekretua onartu (…)*

Halaber, anafora zein erreferentearen numeroari erreparatu diogu eta ia kasu guztietan biak bat datoz, %99an.

Aipatutako ondorioak, ondorengo taulan ikus ditzakegu laburtuta:

	HAU	HORI	HURA
Agerpenak	177	251	321
Anafora	% 86	% 99	% 98
Katafora	% 14	% 1	% 2
Aurrekaria esaldi berean	% 59	% 50	% 64
Aurrekaria esaldi bat lehenago	% 32	% 47	% 33
Aurrekaria izen-sintagma	% 73	% 33	% 100
Aurrekaria perpausa	% 27	% 67	% 0

Azkenik aipatu, etorkizunari begira, aurrekari egokia aukeratzeko beharrezkoa den informazioa gehitzeko asmoa badugula. Lehen esan bezala generoaren informazioa erabili ezin dugunez, beste datu batzuetaz baliatu beharko dugu, adibidez, erreferentearen biziduntasuna, azpikategorizazioa, etab. Posible dugun informazio gehiena erabiltzea interesatzen zaigu, anafora pronominalaren ebazpen automatikoan pentsatuz, erreferentea errazago detektatzeko lagungarri izan baitaiteke.

6. Ondorioak eta etorkizuneko lana

Lan honetan zehar aztertutakoarekin corpusa osatzen hasteko oinarri batzuk finkatu ditugu. Anaforaren sailkapena euskarara egokitu eta konkretuki anafora pronominalaren arloan sakondu ahal izan dugu.

Aztergaia mugatu ostean, markatzea erabaki dugun determinatzaile eta izenordainen portaera egiaztatzeko atera ditugun laginak oso handiak ez izan arren, lanean zehar aipatu dugun bibliografian ikusitako azalpen teorikoak betetzen direla ikusi dugu. Halaber, kontuan izanik fenomeno honen inguruan euskaraz oraindik badagoela aztertzeko, lagin handiagoetan egiaztatzeko geratu zaizkigun zenbait ondorio ikusi ahal izan ditugu.

Etorkizunera begira, corpusetik adibide gehiago atera ahal izango ditugu eta lehenengo lagin honetan atera ditugun ondorioak egiaztatu eta berretsi ahal izango ditugu. Honekin batera, aztergaia zabaltzeko aukera izango dugu, hurrengo pauso batean izenordaina eta agian elkarkari eta bihurkariak kontuan hartzeko.

Lehenago azaldu dugun moduan, azterketa hau IXA taldearen lanen barruan kokatua dago eta etorkizunean ildo honetatik jarraitzeko asmoa badago. Horregatik, hemen aztertutakoak gerora corpusaren markaketa zabalago bat egiterakoan baliagarri izango zaizkigu. Izan ere, honako lan hau lehen hurbilpen bat izan baita.

Ildo honetatik jarraituz epe erdira zenbait aplikazio aurreikusten ditugu, horietako bat anafora erdi automatikoki ebazteko tresna litzateke. Tresnak anafora ebazteko erabiliko duen algoritmoa trebatzeko corpusaren beharra argi dago, erabiliko duen ezagutza zenbat eta zabalagoa izan emaitzak hobetzeko aukera handiagoa izango baitugu.

Epe luzeragorako berriz, euskararako garatuko diren aplikazio ezberdinetan, laburpen sistemetan edota itzulpen automatikoko aplikazioetan aztertutako guzti hau txertatzeko aukera ikusten dugu.

Eskertzak

Lan hau Euskal Herriko Unibertsitatearen (9/UPV00141.226-14601/2002), Eusko Jaurlaritzako Industria Sailaren (HIZKING21 ETORTEK2002) eta Zientzia eta Hezkuntza Ministerioaren (TIN2004-07918-C04-01) diru-laguntzarekin burutu da.

7. Bibliografia

[1] AZKARATE M. & ALTUNA P. (2001). *Euskal morfologiaren historia*. Donostia: Elkarlanean.

[2] CORNISH F. (1999). *Anaphora, discourse and understanding: evidence from English and French*. Oxford: Oxford University Press.

[3] EUSKALTZAINDIA (1985). *Euskal Gramatika: Lehen Urratsak I*. Nafarroako Foru Gobernua: Euskaltzaindia.

[4] EUSKALTZAINDIA (1993). *Euskal Gramatika Laburra: Perpaus bakuna*. Bilbo: Euskaltzaindia.

[5] FERRÁNDEZ RODRÍGUEZ A. (1998). *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos*. Doktoretza-tesia. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

[6] GARCÍA AZKOAGA I. (1998). *Erlazio anaforikoak argudiozko testuetan*. In Koherentzia, kohesioa eta konexioa: testuratze baliabideak. Hizkuntzaren azterketa eta irakaskuntza, Larringan & Idiazabal (ed.). Gasteiz: EHU-UPV & Arabako Foru Aldundia.

[7] GARSIDE R., LEECH G. & MCENERY A. (eds.) (1997). *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London: Longman.

[8] GARZIA GARMENDIA J. (1996). *Hura, bera, eta abarren adar gehiago*. Senez 18. zenbakia. EIZIE Euskal Itzultzaile, Zuzentzaile eta Interpreteen Elkartea.

[9] GRISHMAN R. & SUNDHEIM B. (1996). *Message Understanding Conference - 6: A Brief History*. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark.
[10] HAJIČ J. & UREŠOVÁ Z. (2004). *The Prague Dependency Treebank*. IXA taldeari egindako aurkezpena. Donostia.
[11] HIRST G. (1981). *Anaphora in Natural Language Understanding*. Berlin: Springer-Verlag.
[12] HUALDE J.I. & ORTIZ DE URBINA J. (2003). *A grammar of basque*. Berlin: Mouton de Gruyter.
[13] KAMP H. & REYLE U. (1993). *From discourse to logic: introduction to modeltheoretic semantics of natural language, formal language, formal logic and discourse representation theory*. Dordrecht: Kluwer Academic Publishers.
[14] KLEIBER G. (1994). *Anaphores et pronoms*. Louvain-la-Neuve: Duculot.
[15] Kunz K. & Hansen-Schirra S. (2003). *Coreference Annotation of the TIGER Treebank*. In Proceedings of the Workshop Treebanks and Linguistic Theories, Växjö, Sweden.
[16] LAKA I. (1998). *A Brief Grammar of Euskara, the Basque Language*. HTML-ko dokumentua. Euskarako errektoreordetza, Euskal Herriko Unibertsitatea. http://www.ehu.es/grammar
[17] MAILLARD M. (1974). *Essai de typologie des substituts diaphoriques*. Langue Française, 21. Paris: Larousse.
[18] MITKOV R. (2002). *Anaphora resolution*. London: Longman.
[19] 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.
[20] NAVARRO B., IZQUIERDO R., SAIZ-NOEDA M. (2004). *Exploiting semantic information for manual annotation in Cast3LB corpus*. In Proceedings 42nd Annual Meeting of the ACL. Barcelona.
[21] ORASAN C. (2000). *CLinkA a Coreferential Links Annotator*. In Proceedings of LREC'2000, Athens, Greece.
[22] PALOMAR M., CIVIT M., DÍAZ A., MORENO L., BISBAL E., ARANZABE M., AGENO A., MARTÍ M.A. Y NAVARRO B. (2004). *3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y español*. XX. Congreso SEPLN, Barcelona.
[23] PALOMAR M., FERRÁNDEZ A., MORENO L., MARTÍNEZ-BARCO P., PERAL J., SAIZ-NOEDA M. and MUÑOZ R. (2001). *An Algorithm for Anaphora Resolution in Spanish Texts*. Computational Linguistics, Vol. 27, Number 4.
[24] RICO C. (1994). *Aproximación estadístico-algebraica al problema de la resolución de la anáfora en el discurs*. Doktorego-tesia. Universidad de Alicante.
[25] ZUBIRI I., & ZUBIRI E. (1995). *Euskal Gramatika osoa*. Bilbo: DIDAKTIKER.