Euskararako aipamen-detektatzailea: korreferentzia-ebazpenerako sistema baten lehen urratsak

Egilea: Ander Soraluze Irureta
Tutoreak: Xabier Arregi Iparragirre
Olatz Arregi Uriarte

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2012ko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Master bukaerako proiektu honetan, euskarazko korreferentzia-ebazpena gauzatzeko sortu nahi den sistema baten garapenean egindako lehen urratsak azaltzen dira. Lehen pauso bezala, aipamenen azterketa linguistiko batean oinarritutako aipamen-detektatzailea aurkezten dugu. Sistema euskaraz idatzitako testuetan korreferentzia-kateetako partaide izan daitezkeen aipamenak detektatzeko gai da. Aipamen-detektatzailea erregelatan oinarritutakoa da, eta egoera finituko teknologia erabiliz inplementatu da. % 77,58ko F-measure balioa lortzen du*Exact Matching* ebaluazio-irizpidea erabiliz, eta % 82,81koa, berriz, *Lenient Matching* erabiltzean. Aipamen-detektatzaileari aplikazio erreal bat eman nahi izan zaio. Horretarako EPEC corpusa automatikoki etiketatu da eta detektatuko aipamenak MMAX2 etiketatze tresnan erabiltzeko prestatu dira.

Abstract

This paper presents the first steps in the development of a Basque coreference resolution system. We propose a mention detector system based on a linguistic study of the nature of mentions. The system identifies mentions that are potential candidates to be part of coreference chains in Basque written texts. The mention detector is rule-based and has been implemented using finite state technology. It achieves a F-measure of 77.58% under the Exact Matching protocol and of 82.81% under Lenient Matching. The mention detector has been used to automatically tag an EPEC corpus with mentions, for later use them in MMAX2 annotation tool.

Gaien aurkibidea

Glosategia

A

Aipamen (mention) Mundu errealeko objektu bati erreferentzia egiten dion espresio testuala.

E

Eihera Euskarazko entitate izendunak ezagutzen eta sailkatzen dituen tresna.

Entitate (entity) Mundu errealeko objektua edo objektu multzoa.

Eustagger IXA taldean sortutako euskarako analizatzaile/desanbiguatzaile morfosintaktiko automatikoa.

F

Foma Helburu anitzetarako erabil daitezkeen egoera finituko automata eta transduktoreak sortzeko konpiladorea, programazio lengoaia eta Cko liburutegia da.

I

IXATI Euskararako analizatzaile sintaktikoa (azalekoa).

K

Kate (chunk) Katea sintagma kategoriako zatia da eta, sintaktikoki erlazionaturiko hitzez osatua dago. Gainjartzen ez diren eta elkarrekin sintaktikoki erlazionaturik dauden hitz multzoak dira kateak. Hitz multzo horiek, gainera, ez-errekurtsiboak izango dira; hau da, ezin dute beren baitan beste hitz multzorik edota katerik izan.

Korreferentzia Objektu edo entitate bera adierazten duten bi elementu testualen arteko erlazioa.

Korreferentzia-ebazpen Testu batean dauden korreferentziak detektatzen eta multzokatzen dituen prozesua nahiz emaitza.

M

ML-IXATI IXA taldean sortutako kate eta perpaus identifikatzaile automatikoa.

Morfeus IXA taldean sortutako euskarako analizatzaile morfologiko automatikoa.

HAP masterra

Murriztapen-gramatika (Constraint Grammar) Patroiak identifikatzeko eta etiketak jarri, kendu edo aldatzeko aukera ematen duen formalismoa.

P

Perpaus Aditz baten inguruan osatzen den hitzen multzoa da perpausa. Hortaz, aditzak eta aditzari dagozkion elementuek osatzen dute perpausa. Aditza, dena dela, ez da beti testuan esplizituki agertuko; hots, aditza perpausaren ardatza izanagatik ere, aditzaren beraren elipsia egon daiteke, eta horrek ez dio perpaus-izaera kentzen. Bestalde, bi perpaus mota definitzen ditugu: markatuak (menderatuak) eta markatu gabeak. Menderagailua daramatenak izango dira markatuak edo mendeko perpausak, eta markarik ez dutenak perpaus bakunak izango dira.

U

Urre-patroi (gold standard) Automatikoki eskuratutako emaitzak ebaluatu ahal izateko, eskuz sortzen diren emaitza prototipikoak.

X

Xerox Finite State Tool (XFST) Adierazpen erregularrak jaso, eta hauek transduktore bihurtzen dituen tresna. Egoera finituko kalkulua ahalbidetzen duen eragiketa multzo aberatsa du.

Laburtzapenak

ADB Adberbioa

ADIZE Aditz-izena

AKB Aditz-kate bukaera

AKH Aditz-kate hasiera

AK Aditz-katea

EAM Erlatibozko egitura duen aipamena

EA Erlatibozko atzizkia duen aditza

ELI Elidítutako izena

EM *Exact Matching* parekatze-metodoa

ERL Erlatibozko atzizkia duen aditza

H Hitza

JUNT Juntagailua

LM *Lenient Matching* parekatze-metodoa

MP Mendeko perpausa

PB Perpaus bukaera

PH Perpaus hasiera

POSB Postposizio-lokuzio amaiera

POSH Postposizio-lokuzio hasiera

SIB Izen-kate bukaera

SIH Izen-kate hasiera

SINT Izen-katea

1 Proiektuaren definizioa

Testu bateko bi espresio testualek objektu berbera adierazi edo erreferentziatzen dutenean, bi espresio horien artean korreferentzia-erlazio bat dagoela esan ohi da. Testu batean ager daitezkeen espresio testual horien arteko korreferentzia-erlazioak ebaztea helburu duen atazari korreferentzia-ebazpena deritzo.

Ataza honetan sarritan erabiltzen diren bi termino *entitatea* eta *aipamena* dira. Entitate bat mundu errealeko objektua edo objektu multzoa dela esaten da, aipamena aldiz, entitate bati erreferentzia egiten dion espresio testuala da (Doddington et al., 2004).

Azaldutako terminoak modu argiagoan ulertzeko ikus dezagun adibide bat.

> [Miguel Indurain] erretiratu zenean, [[hura] ordezkatuko zuen pertsona bat] bilatu nahian zebiltzan.

Goiko adibidean, kortxete artean hiru aipamen ikus ditzakegu, [Miguel Indurain], [hura] eta [hura ordezkatuko duen pertsona bat]. Garbi ikusten da [Miguel Indurain] eta [hura] aipamenek mundu errealeko objektu berbera erreferentziatzen dutela, beraz, korreferenteak direla esan dezakegu.

Gaur egun, korreferentzia-ebazpena gakotzat har dezakegu testuak ulertu ahal izateko; ondorioz, beharrezkoa da diskurtsoaren ulerkuntza sakonagoa eskatzen duten Lengoaia Naturalaren Prozesamenduko (NLP) hainbat atazatan; adibidez, informazioaren erauzketan (McCarthy and Lehnert, 1995), testuen laburpenean (Steinberger et al., 2007), galdera-erantzuteko sistemetan (Vicedo and Ferrández, 2006), itzulpen automatikoan (Peral et al., 1999), sentimentuen analisian (Nicolov et al., 2008) edota irakurketa automatikoan (Poon et al., 2010).

Ohikoa da korreferentzia-ebazpena bi azpi-ataza nagusitan banatzea: aipamenen detekzioa, batetik, eta erreferentzien ebazpenean, bestetik (Pradhan et al., 2011). Aipamenen detekzioa testu batean entitate baten aipamenak topatzean datza; erreferentzien ebazpena, berriz, entitate berdinari erreferentzia egiten dioten aipamenak identifikatzean.

Lan honen helburua aipamen-detektatzaile eraginkor bat garatzea da, ondoren euskararako korreferentzia-ebazpenerako sistema batean erabili ahal izateko. Aipamenen azterketa linguistiko batean oinarritu gara lan honetan, eta azterketa horretatik eratorritako deskribapenak, ondoren, egoera finituko teknologia erabiliz espresio erregularren bidez kodetu ditugu.

Ezaguna da euskara bezalako hizkuntza gutxitu batean gertatu ohi den baliabide linguistikoen urritasuna. Hori dela eta, aipamenen detekzioa moduko atazetarako tresna eraginkorrak garatzea erronka izan ohi da.

Memoria hau honela egituratzen da. 2. kapituluan aurrekariak aztertu ondoren, 3. kapituluan euskarazko aipamenen azterketa linguistikoa azalduko da. 4. kapituluan garatu dugun sistemaren ikuspegi orokorra eskainiko da, aurreprozesaketarako erabilitako tresnak deskribatuz eta sistema garatzeko erabili diren teknologiak eta garatutako erregelak azalduz. Ebaluazioan lortutako emaitzak 5. kapituluan aztertuko dira. 6.ean, berriz, aipamen-detektatzaileari eman zaion erabilera azalduko da, aipamen-detektatzailearen irteera MMAX2 tresnan erabili ahal izateko nola prestatu den azalduz. Azkenik, 7. kapi-

tuluan proiektu honetatik ateratako ondorio nagusiak eta etorkizunean gauzatzeko gelditu diren lanei tartetxo bat eskainiko diegu.

2 Aurrekariak

Korreferentzia-ebazpenak presentzia nabarmena izan du informazioaren erauzketarekin erlazionatutako atazetan. Lehen aldiz, seigarren eta zazpigarren *Message Understanding Conference* (MUC-6, 1995; MUC-7, 1998) barnean antolatu zen korreferentzia-ebazpenari zegokion ataza. Gero, 2000 eta 2001 urteetan, *Automatic Content Extraction* (ACE) programaren esfortzua *Entity Detection and Tracking*-ean (EDT) zentratu zen. EDT atazan, entitate berberari erreferentzia egiten dioten aipamen guztiak topatu behar ziren ondoren baliokidetza-klaseetan biltzeko.

Hala ere, MUC eta ACE programak informazioaren erauzketarako diseinatuak izan ziren, ondorioz, korreferentzia-elementuei buruz hartutako erabakiak atazaren beharretara egokituak izan ziren, zehaztasun linguistikoaren kaltetan (van Deemter and Kibble, 1995; Recasens, 2010).

Duela gutxi, zehazki korreferentzia-ebazpenera mugatutako hainbat ataza antolatu dira. SemEval-2010en korreferentziaren ebazpena gauzatu behar zen hainbat hizkuntzatan (Recasens, 2010). Urtebete beranduago, CoNLL-2011n (Pradhan et al., 2011), parte hartzaileek OntoNotes corpusean (Pradhan et al., 2007) korreferentzia-ebazpena gauzatu behar izan zuten.

Korreferentzia ebazteko, lehenik korreferentzia-kateak osatuko dituzten aipamenak topatu beharra dago. Ikertzaile ugarik aipatu duten moduan (Stoyanov et al., 2009; Hacioglu et al., 2005; Zhekova and Kübler, 2010), aipamenen detekzioa garrantzi berezikoa da korreferentzia-ebazpenerako sistema batean. Ataza honetan egindako erroreek, hedatu eta hurrengo pausoetako eraginkortasuna murrizten dute. Hori dela eta, korreferentzia-ebazpenerako sistemetan erabiltzen diren aipamen-detektatzaileen hobekuntzak artearen egoera nabarmen hobetuko luke. Ideia hau defendatzen duten azterketa ugari argitaratu dira, eta hauetan aipamenen detekzioak korreferentzia-ebazpenean duen eragina kuantifikatu da. Uryupina (2008) artikuluan diote haiek garatutako korreferentzia-ebazpenerako sistemak egindako estaldura erroreen % 35 detekzio-fasean galdutako aipamenen ondorioz sortuak direla. Horrez gain, Uryupina (2010) artikuluan, doitasun erroreen % 20 aipamenen detekzio desegokiak sortuak direla gehitzen dute. Chang et al. (2011) lanean bi sistema konparatzen dituzte, batetik, aipamenak automatikoki detektatzen dituen sistema, bestetik, urre-patroia (ingelesez, *gold standard*) erabiltzen duen sistema. Azken honek emaitzak % 15 eta % 18 bitartean hobetzen ditu. Domeinu espezifikoetan ere aipamen-detekzio egokia egitea funtsezkoa dela defendatu da. Biomedikuntza bezalako arlo espezifikoetarako prestatuko korreferentzia-ebazpenerako sistemetan ere aipamen-detektatzaile eraginkor bat izatea oso garrantzitsua da (Kim et al., 2011). Hauek behatu duten arabera, automatikoki detektatutako aipamenak edo urre-patroiak erabiliz, MUC ebaluazio-irizpidearen arabera, sistemak lortzen dituen emaitzak % 49,69 izatetik % 87,32 izatera pasatzen dira, hobekuntza nabarmena da, zalantzarik gabe.

Aipamenen detekzioa erronkatzat hartzen da ataza honetan detektatu behar diren espresioak bai sintaktikoki bai semantikoki egitura konplexuak izan ohi direlako.

Aipamenen detektatzailea garatzeko erabili diren teknologiak bi multzo nagusitan bana ditzakegu. Batetik, erregelatan oinarritutako teknikak ditugu eta bestetik, ikasketa

automatikoan oinarritutako ereduak.

SemEval-2010 kongresuan sistema gehienek erregelatan oinarritutako teknikak erabili zituzten aipamenen detekzioa gauzatzeko (Stoyanov et al., 2009; Hacioglu et al., 2005; Zhekova and Kübler, 2010); halatsu gertatu zen CoNLL-2011n ere, non lau sistemak soilik erabili zituzten ikasketa automatikoan oinarritutako ereduak. Lehiaketa honetan emaitzarik onena lortu zuen sistema guztiz erregelatan oinarritutakoa izan zen (Lee et al., 2011). Ikasketa automatikoan oinarritutako ereduek doitasun eta estaldura balio orekatuagoak eta F-measure balio altuagoak lortzeko joera erakusten duten arren, lortzen duten estaldura erregelatan oinarritutako teknikek lortzen dutena baino baxuagoa izan ohi da. Aipamenen detekzio-fasean estaldura balio baxuak lortzeak albo-ondorio negatiboak sortzen ditu korreferentzia-ebazpenean. Aurreko fasean galdu diren aipamenak ezingo dira berreskuratu eta, horrek korreferentzia-ebazpenerako sistemaren eraginkortasuna guztiz baldintzatuko du.

3 Aipamenen azterketa linguistikoa

Kapitulu honetan, aipamen-detektatzailea garatzeko erabili dugun aipamenen azterketa linguistikoa azaldu nahi da. Azterketa linguistiko hau garrantzitsua izan da garapen-faserako, bertan zehaztu baitira zein egitura kontsideratuko diren aipamentzat eta zein ez, baita bakoitzaren ezaugarri linguistikoak ere.

Recasens-en (2010) aipatzen den moduan, korreferentziak etiketatzea eta horrek lehenagotik eskatzen duen aipamen-etiketatzea ez da ataza hutsala. Urteetan zehar anotazio-eskema proposamen ugari egon dira eta ohikoa da corpus bakoitzak bere beharretara egokitutako eskema propioa erabiltzea. Beraz, esan dezakegu ez dagoela estandartzat har daitekeen proposamenik, eta sarritan kontraesanean (van Deemter and Kibble, 1995) dauden eskemak aurkitzen direla.

Gainera, korreferentzia edo erreferentziakidetasuna arlo pragmatikoan kokatzen bada ere, hizkuntza bakoitzaren berezitasunak ere kontuan hartu behar dira. Honek guztiak gure anotazio-eskema proposatzera bultzatu gaitu.

Bestalde, korreferentzia-ebazpenerako sistema garatu eta ebaluatzerakoan aukeratzen den corpusak garrantzi handia du, eta corpusa etiketatu den moduak eragina du sistema bat diseinatu eta eraikitzeko orduan. Hori dela eta, aipamen-azterketa linguistikoa egitean, kontuan izan dugu lan honen helburua aplikazio automatiko bat sortzea dela.

Hona hemen guretzat korreferentzia-katea osa dezaketen elementuak, aipamenak, zein diren:

1. **Izen-kate arruntak**

   Orokorrean izen-kate guztiak aipamentzat hartuko ditugu, baina multzo honetan izen-kate arruntak hartuko ditugu. Horien artean honako azpi-sailkapena egin dugu:

   1.1 Artikulu mugatuarekin bukatzen den izen-katea. Artikuluak ez du berez inolako informazio erreferentzialik, baina mugatzen duten izena identifikagarria dela adierazten dute:

   (a) Lotinak [lasaitasun*a*] eskatu du.

   1.2 Izen-katearen azken osagaia determinatzaile erakuslea da:

   (b) [Hitzaldi *horietan*] oso ondo adierazten zituen bezeroen gogoak.

   1.3 Izen-katea mugagabea da:

   (c) Ikerketatik ateratzen dituen datuetatik [prozesu *bat*] irekiko dela espero da.

2. **Izen bereziak**

   Izen-katearen burua izen berezia da. Izen bereziak erreferentziatzen duen objektua bakarra eta ez anbiguoa da.

   (d) [*Argentinan*] egindako krimenak ikertzen hasiko dira.

3. **Izenordainak**

   Euskarazko izenordainen sailkapenera jotzen badugu, euskal gramatiketan benetazko izenordaintzat jotzen direnak, lehen (*ni, gu*) eta bigarren (*zu, zuek*) pertsonako izenordainak dira. Hirugarren pertsonako izenordainik ez da berariaz aipatzen. Hala ere, honela aipatzen du Euskaltzaindiaren gramatikak:

   > [...] zenbait gramatikatan *bera* eta gainerako erakusleak (*hau, hori, hura*) hartu ohi dira hirugarren pertsonako izenordaintzat. (Euskaltzaindia, 1985).

   Etiketatze honetan, lehen eta bigarren pertsonako izenordainak ez ditugu kontuan hartuko, kasu gehienetan deiktikoak izango direlako (erreferentzia ez da testuan bertan egongo, testuinguruan baizik).

   Guk aipamen gisa, berez, gramatiketan determinatzaile erakusle bezala izendatzen direnak hartuko ditugu. Izen-katearen osagai bakarra erakusle horietako bat (*hau, hori, hura*) denean, izenordain kontsideratuko dugu hurrengo adibidean ikusiko dugun bezala:

   (e) LDPko buruek Mori hautatu zuten apirilean Keizo Obuchi orduko lehen ministroa ordezkatzeko, [*hark*] tronbosia izan ostean.

4. **Edutezkoak**

   Euskarazko posesiboa edo edutezkoa izenordainari genitibozko kasua gehituz osatzen da. Orokorrean, izen-kate osoak hartuko baditugu ere, genitibozkoen kasuan, katea zatitu eta parte bat hartuko dugu, genitiboan doan determinatzaile erakuslea aipamentzat hartuz.

   (f) Epitieren kasuan [[*bere*] helburua] lortu dezakela dirudi eta baliteke denboraldia Lehen Mailan hastea.

   Edutezko determinatzaile hauek, batzuetan, eurek bakarrik osatzen dute sintagma, izen-katearen buru bihurtzen dira, posesiboaren eta kasu-markaren artean izenaren elipsia gertatzen delako. Kasu horietan erreferentzia bikoitza izan dezakete, objektuarena edo erreferentearena batetik, eta norena den bestetik. Horrelakoetan ere, aipamen gisa joko ditugu.

   (g) Escuderok euskal musika tradizionala eraberritu eta indartu zuen. [*Harenak*] dira, esate baterako, Illeta, Pinceladas Vascas eta Eusko Salmoa obrak.

5. **Aditz-izenak**

   Atal honen hasieran aipatu dugu orokorrean izen-kateak kontsideratuko ditugula aipamen gisa, baina badira kasu berezi batzuk, non aditzek izenaren funtzioa betetzen duten, aditz-izenak alegia. Aditz-izenen artean, batzuk izena ordezkatzeko gaitasuna izango dute.

(h) Garrantzi handia du [tokietako hizkeren berezitasunen *jasotze* honek].

Adibideko *jasotze* aditz-izena beste izen arrunt batengatik ordezka dezakegu esaldian. Horrelako aditz-izenek, orokorrean, izen arrunt batek har ditzakeen modifikatzaileak onartzen dituzte (izenlagunak, determinatzaileak, etab.). Guk, aditz-izen horiek osatzen dituzten izen-kateak aipamen gisa etiketatuko ditugu.

Badira beste aditz batzuk, esaldi nagusiko aditzak hala eskatuta, aditz-izen forma hartuko dutenak:

(i) [Instalazio militarrak *ixtea*] eskatuko dute

Adibide honetan *ixtea* aditz-izena mendeko perpausaren parte da (*Instalazio militarrak ixtea*) eta ezingo litzateke aditz-izena bakarrik izen batekin ordezkatu; horregatik, kasu hauetan ez dugu aipamentzat hartuko, ez duelako berak bakarrik izenaren funtzioa betetzen.

6. **Postposizio-lokuzioetako aipamenak**

Egungo ikuspegitik, postposizioen artean bi mota nagusi bereiz daitezke: alde batetik, lemari erantsirik ageri diren atzizkiak (argi-*rik*), eta bestetik, osagai nagusia elementu beregaina dutenak (argi-*rik gabe*) (Aduriz et al., 2008).

Lehenengo multzoko postposizioei (argi-*rik*) atzizki-postposizioak deituko diegu (Zabala and Odriozola, 2004). Bigarren multzokoak (argi-*rik gabe*), ostera, postposizio-lokuzioak direla esango dugu. Hemen postposizio beregainaren aurreko izena hartuko dugu kontuan aipamentzat, izenaren lemari erantsirik ageri zaizkion atzizkiak barne, hurrengo adibidean (*argindarrik*) izango litzateekena:

(j) Ondorioz, [*argindarrik*] *gabe* geratu ziren hiriko zenbait auzo.

7. **Mendeko perpausa duten izen-kateak**

Izen-kate hauetan burua izen arrunta izango da eta izen horrek modifikatzaile gisa jokatuko duen mendeko perpaus bat izango du. Hurrengo adibidean ikusten dugun bezala, mendeko perpaus osagarriak izenari informazioa gehitzen dio, eta horregatik, mendeko perpausa ere aipamenean sartu dugu.

(k) 1. [[Gaintzako udaletxe berria] *egiteko* lanak] hasi dira asteburu honetan.

Izen-katearen burua *lanak* litzateke eta azpian duen perpausa *Gaintzako udaletxe berria egiteko*. Guk kate guztia markatuko dugu, eta aldi berean, azpian izan ditzakeen izen-kate guztiak ere markatuko ditugu.

Honela, goiko adibidean ondorengo aipamenak izango genituzke *lanak* izenaren inguruan.

(k) 2. [Gaintzako udaletxe berria]

3. [Gaintzako udaletxe berria egiteko lanak]

Izenak mendeko perpaus gisa izan dezakeen beste perpaus-mota erlatibozkoa da. Horrelakoetan ere erlatibozko perpausa eta izena hartuko ditugu kontuan aipamen osoa markatzerakoan, baita barruko izen-kateak ere:

(l) [[Igandeko partiduak] *duen* garrantzia] dela eta, lasai egotea beharrezkoa dutela esan zuen Lotinak.

8. **Koordinazioa**

Izen-kate koordinatuak juntagailu (*eta, edo, edota, nahiz...*) baten bitartez lotzen diren izen-kateak dira. Egitura koordinatu horietan ez da beti erraza izango erabakitzen zein kasutan osagaiak bere horretan aipamen gisa hartu behar diren eta zeinetan ez, kasuistika zabala baita.

Orokorrean koordinazio-egiturak banatzea komeni da baldin eta koordinazioaren ezker aldeko osagaia eta eskuin aldekoa argi eta garbi bereiz baditzakegu; hau da, "A eta B" koordinazio-egitura batean "A" osagaia eta "B" osagaia bereiz baditzakegu.

Azter ditzagun ondorengo adibide hauek:

(m) [[Palacios] eta [Alex]] partidutxoa amaitu aurretik erretiratu ziren

(n) Egunkarietan [Juan Epitie eta Nacho Zaragozaren kasua] agertu da.

1. adibidean ikus dezakegu juntagailuaren (*eta*) ezker aldeko eta eskuin aldeko osagaiak bereizgarriak direla. Ezker aldean *Palacios* markatuko genuke, eskuin aldean berriz *Alex*, bi pertsona ezberdinei buruz ari delako. Ezker aldean nahiz eskuin aldean lortzen diren egiturek zentzu osoa dute eurek bakarrik, ondorioz, koordinazio-egitura hori zatigarria da.

Bigarren adibideak, aldiz, beste koordinazio-mota konplexuagoa erakusten digu. Ezker aldean dagoen *Juan Epitie* izen bereziak, berez, egitura koordinatuaren bukaeran agertzen den *kasua* izenarekin batera osatuko du izen-katea. Honela, ezker eta eskuin aldeko egiturak bereiziko bagenitu *Juan Epitie* eta *Nacho Zaragozaren kasua* lortuko genituzke, eta ez genituzke behar bezala interpretatuko. Horregatik, kasu hauetan elementu koordinatu osoa izango da guretzat aipamena eta ez ditugu banatuko, koordinazio-egitura banatuz gero ez baita jatorrizko zentzua mantentzen.

9. **Elipsia**

Euskaraz elipsia hainbat mailatan gertatzen da. Esaldi-mailan adibidez, subjektua, objektua eta zehar-objektua momentuko behar komunikatiboen arabera isiltzeko aukera izaten da. Ondorengo esaldia bost modutara idatz daiteke eta guztiak egokiak dira, baldin eta isildutako elementua testuinguruaren arabera argi badago zer edo zein den.

(o) 1. [Pellok] [hari] eman dio [liburua]

2. [Pellok] Ø eman dio [liburua]
3. Ø [Hari] eman dio [liburua]
4. Ø Ø Eman dio [liburua]
5. Ø Ø Eman dio Ø

Etiketatze honetan agerian dauden elementuak hartuko ditugu kontuan aipamentzat, isilduak bazter utziz.

Bada beste elipsi-mota bat maila morfosintaktikoan ematen dena, erlatibozko perpauseko aditzaren eta ondorengo kasu-marken artean. Aditzari zuzenean lotzen zaio kasu-marka (*zen-ø-ak*) eta izena isilduz elipsia gertatzen da. Honela, berez etiketatuko ez genituzkeen aditzek, izen izaera hartu eta izen-kate gisa jokatzen dute. Horregatik aipamen gisa kontuan hartuko ditugu:

(p) [Bigarren sailkatu *zenak*], segundo bakarra kendu zion.

10. **Adberbioak**

Orain arte esan dugu korreferentzia eskuarki izen-kateen artean gertatzen dela, zehatzago esanda, izen-kateen artean ematen den erlazioa dela. Hala ere, zenbait kasutan, adizlagun edota perpaus eta esaldien artean ere gertatzen da fenomeno hau.

Corpuseko testuetan ugariak dira funtzio kohesibo hori duten lekuzko adberbioak. Multzo honetan sailkatuko ditugu aurretik aipatutako leku bati erreferentzia egiten dioten adberbioak. Euskaraz, benetazko lekuzko adberbioak hiru dira: *hemen, hor, han*. Lotura estua dute determinatzaile erakusleen hiru mailekin. Badago gainera, erakusle indartutik (*bera*) sortzen den adberbioa ere: *bertan*.

(q) Etxeko atezainarenean giltza bat utziko ziola, eta apartamentura igo eta [*han*] itxaroteko.

(r) Segurtasun neurri handien artean iritsi ziren bi epaileak Urruñako etxera eta [*bertan*] izan ziren eguerdira arte.

Horregatik, lekuzko adberbioak aipamentzat hartuko ditugu.

11. **Aipamentzat hartuko ez direnak**

Aipamentzat kontsideratu behar ditugunak azaldu ostean, zein elementu aipamentzat ez ditugun kontsideratuko zehaztuko dugu. Aurreko puntuan aipatu ditugun arren puntu honetan laburpen modura zerrendatu ditugu.

- **Lehen eta bigarren pertsonako izenordainak** ez ditugu kontuan hartuko lan honetan, testuan zehar erreferentziakide ez direlako izango, deiktikoak baizik.

  (s) *Nik* ez dut etxera joateko gogorik, hala ere *zuek* erabaki beharko duzue, zer esan behar diozuen jendeari, galdezka hasiko da-eta.

- **Adjektibo soilez osaturiko sintagmak** ere ez ditugu aipamentzat kontsideratuko. Orokorrean adjektiboek ez diote objektu konkretu bati erreferentzia egingo, objektu horien ezaugarriren bat deskribatuko dute.

  (t) Etxeak *politak* dira.

- **Aditz-izenen** kasuan esan dugunez, bera bakarrik izen batez ordezkatu ezin badugu ez dugu aipamentzat hartuko.

  (u) Kirola *egitea* ona da.

- **Elipsia** subjektu-, objektu- edo zeharobjektu-funtzioan gertatzen denean ez dugu etiketatuko. Beheko adibidean subjektu eta zeharobjektua isiltzen dira eta horiek ez genituzke markatuko.

  (v) Arraunlariak berandu iritsi arren $\emptyset$ $\emptyset$ ez zioten inongo azalpenik eman.

- **Adberbioei** dagokienez, lehen aipaturiko lekuzko adberbioak izan ezik, besteak ez ditugu kontuan edukiko. Horrelako adberbioek ez diote objektu konkretu bati erreferentziarik egingo, objektuaren nolakotasunari edota moduari baizik.

  (w) *Atzo* eman ziguten azterketaren emaitza.

  (x) Zentzugabea eta ulertezina da gure aurka *horrela* jotzea.

4 Sistemaren ikuspegi orokorra

Kapitulu honetan, aipamen-detektatzailearen garapenenari dagozkion nondik norakoak azalduko ditugu. Lehenik, egoera finituko teknologiaren inguruko xehetasunei buruzko laburpena egingo da, teknologia hau baita gure aipamen-detektzailea garatzeko erabili duguna. Ondoren, aipamen-detektatzaileak sarrera moduan jasotzen duen testua aurreprozesatzeko erabili den analisi-katea azalduko da. Kapitulua amaitzeko, egoera finituko teknologia erabiliz 3. kapituluan zehaztutako aipamenen definizioak nola kodetu ditugun azalduko da.

4.1 Egoera Finituko Teknologia

Egoera Finituko Makinak (*Finite State Machines*, FSM) eta beraien propietateak ondo ezagutzen diren objektu matematikoak dira. Hala ere, hauen erabilera Lengoaia Naturalaren Prozesamenduaren arloan 80ko hamarkadaren erdi aldetik aurrera aztertu da. FSMak erabiliz sor daitezkeen aplikazioen artean honakoak aurki ditzakegu: zuzentzaile ortografikoak, kategoria gramatikalen desanbiguatzaileak, tokenizatzaileak, azaleko analizatzaileak eta bereziki analizatzaile morfologikoak (Beesley and Karttunen, 2003).

Hizkuntzaren prozesamendurako erabili nahi diren tresna linguistikoak sortzeko, beharrezkoa da Egoera Finituko Makinek oinarrian nola lan egiten duten ezagutzea.

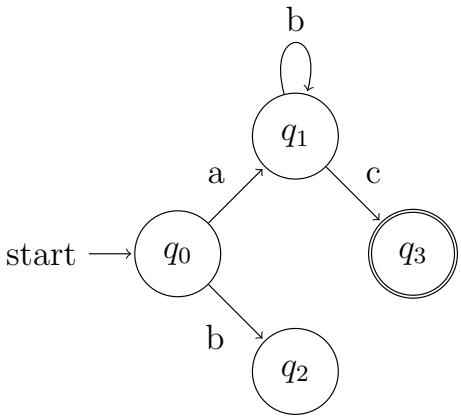
Egoera finituko teknologiaren oinarrian Egoera Finituko Makinak aurkitzen dira. Egoera Finituko Makina baten ezaugarriei dagokionez honakoak aipa daitezke.

- *Egoeraz* osatutako sarea da, hasierako egoera bat eta amaierako egoera bat edo gehiago dituelarik.
- *Trantsizioa* egoera batetik beste egoera batera egiten den aldaketa da.
- *Bidea*, egoera konkretu batera iristeko *arkuen* artean egin daitezkeen *trantsizioen* sekuentzia da.
- Finituak dira, hau da, egoera kopuru finitu bat modela dezakete eta ez egoera kopuru infinitu bat.
- Konbentzioz, egoerak zirkulu moduan adierazten dira eta berain arteko trantsizioak etiketatutako arku moduan. Gainera, hasierako egoera gezi batez markatzen da eta amaierako egoera zirkulu bikoitzen bidez adierazten da (Maleki et al., 2009).

Oraintxe aipatutako kontzeptuak argiago ikusteko 1. irudian agertzen den egoera finituko makinan zentratuko gara. Bertan agertzen zaizkigun egoerak q_0, q_1, q_2 eta q_3 dira. q_0 hasierako egoera da gezi batez markatua baitago, q_3 berriz, amaierako egoera da zirkulu bikoitzez adierazia baitago. Egin daitekeen trantsizio posible bat q_0 egoeratik q_1 egoerara igarotzea da, sarreran a karakterera ezagutu bada. Bide posible bat, berriz, hasierako egoeran hasi (q_0) eta amaierarako egoerara (q_3) iristea da. Horretarako q_0, q_1 egoeren arteko trantsizioa lehenik, eta ondoren, q_1,q_3 arteko trantsizioa eginez.

Egoera Finituko Makinen terminologiari buruz hitz egitean ohikoa da honako hiru termino hauek definitzea:

- Alfabetoa: Makina batek onartzen dituen balizko sinboloen multzoa.
- Hitzak: Sinboloen sekuentziak dira.
- Lengoaia: Hitzen multzo osoa da.



1. irudia: Egoera Finituko Makina (FSM) baten adibidea.

1. irudian agertzen den FSMan, alfabetoa {a, b, c} izango litzateke. Hitz posible bat, *abbc* eta egoera finituko makinak onartzen duen lengoaia $ab^{\star}c$ espresio erregularraren bidez adierazten dena da.

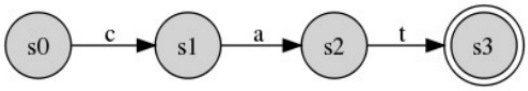
Egoera Finituko Makinak bi multzo nagusitan bana daitezke:

1. Egoera Finituko Automata (*Finite State Automata*, FSA):

   Emandako string multzo bat (lengoaia) soilik onartzen duen FSMa da. Lengoaiak deskribatzen dituzte. *Lookup* (analisia) egiteko erabil daitezke.

2. Egoera Finituko Transduktorea (*Finite State Transducer*, FST):

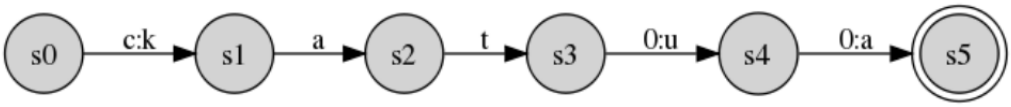
   Egoera Finituko Automatek sarrerako hitzak onartu edo baztertzen dituzte, hala ere, ezin dute sarrera hori aldatu. Murriztapen hau gainditzeko FSTak erabil daitezke. FST bat FSA baten oso antzekoa da, hala ere, badago bien arteko desberdintasun nabarmen bat. Desberdintasun hori, FST batean, arkuetan aurkitzen diren etiketak sinbolo-bikote baten bidez adieraz daitezkeela da. Sinbolo-bikote hauek adierazteko a:b formatua erabiltzen da, non ezker aldeko karaktereak sarrera moduan espero den karakterea adierazten duen, eta eskuin aldekoak irteera bezala itzuliko dena. *Lookup* (analisia) edo *lookdown* (sorkuntza) egiteko erabil daitezke.



2. irudia: FSA baten adibidea.

2. irudian *cat* hitza soilik onartzen duen FSA bat ikus dezakegu.

3. irudian berriz, sarrera gisa *cat* hitza jasotzen duenean emaitza bezala *katua* itzultzen duen transduktorea ikus dezakegu. Hau da, ingeleseko *cat* eta euskarazko *katua* hitzen arteko trasliterazioa egiten du.



3. irudia: FST baten adibidea.

4.2 Analisi-katea

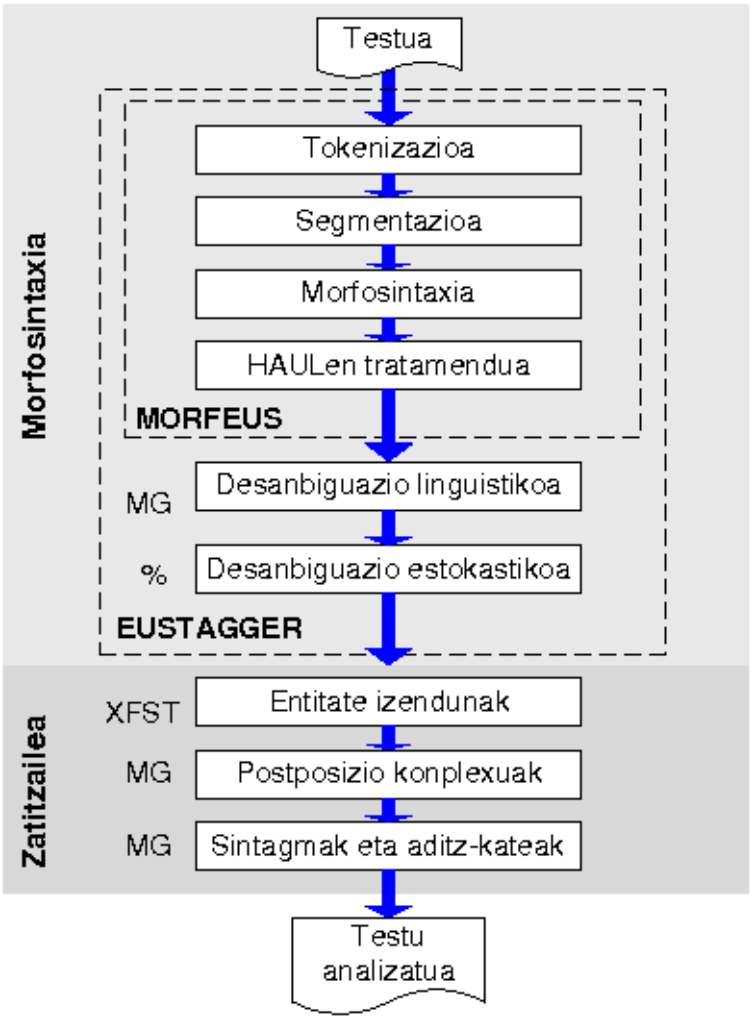
Lan honetan IXA taldean sortutako analisi-katea erabili da. Kate honetan, hainbat tresna erabiltzen dira, bakoitzak bere zeregin zehatza duelarik. Tresna horiek, testuen analisi sakona egiten dute urrats desberdinetan banatuta (tokenizazioa, analisi morfologikoa...). Analizatzailearen ezaugarriak Oronoz-en (2008) tesi-txostenean azalduta datoz, zehatz-mehatz. Horregatik, laburpen bat soilik egingo dugu hemen.

4.2.1 Deskribapena

IXA taldean mendekotasun-egituretan oinarritzen den sintaxi-analizatzaile sendoa erabiltzen da. Sintaxi-analizatzaileak geruzaka egiten du analisia eta geruza bakoitzean hizkuntza-ezagutza maila desberdina erabiltzen du. Analisirako geruzak katean erabiltzen dira modu sekuentzial batean. Geruza bakoitzak aurreko geruzak eskaintzen dion informazioa erabiltzen du sarrera moduan, eta jasotako analisia informazio linguistiko berriarekin aberasten du.

Sintaxi-analisia urratsez-urrats egiten da eta erabiltzailearen esku gelditzen da erabili nahi duen hizkuntza-ezagutza maila aukeratzea.

Geruzetako bakoitzean bereizketa argia egiten da gramatika eta hauek erabiliko dituzten programen artean. Gramatikak, Murriztapen-gramatika (MG) eta XFST tresnek definitutako arauen arabera kodetzen dira. Bi formalismo hauek ordena askeko elementuekin lan egiteko metodologia eta tresna egokiak eskaintzen dituztelako aukeratu dira. 4. irudian ikus ditzakegu analisi-kateko moduluak eta haien analisi-geruzak.

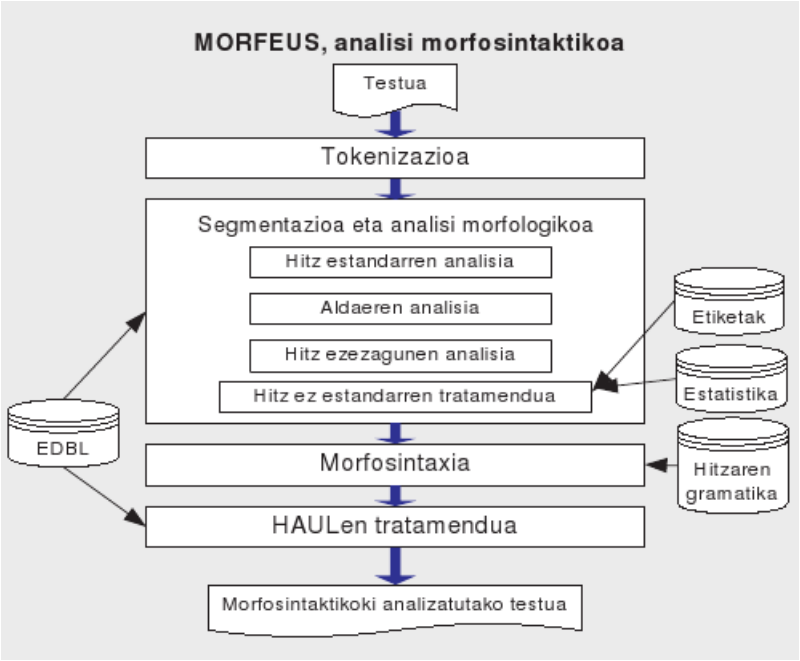


4. irudia: Analisi-katea.

4.2.2 MORFEUS, analizatzaile morfosintaktikoa

Analisi prozesuari hasiera MORFEUS (Alegria et al., 1996) analizatzaile morfosintaktikoarekin ematen zaio. Honek, sarrerako testua jaso eta tokenetan banatu ondoren, horietako bakoitzarentzat lema eta morfema konbinazio posible guztiak ematen ditu, dagokien informazio morfologikoarekin batera. Euskaraz, hizkuntza-unitateak propietate morfologikoak eta sintaktikoak edukitzen dituenez *morfosintaxia* terminoa erabiltzen da morfologia erabili ordez. Ezeiza-ren (2002) tesi-lanean informazio zabalagoa eskaintzen da MORFEUS-i buruz.

Modulu hau osatzen duten analisi-geruzak deskribatuko ditugu jarraian.



5. irudia: MORFEUS, analizatzaile morfosintaktikoa.

1. Tokenizazioa:

   Tokenizatzaileak testua unitate edo tokenetan, eta gero esaldietan, banatzen du. Testuan honakoak identifikatzen ditu: hitzak, zenbakiak, arruntak eta erromatarrak, deklinatu gabeak eta deklinatuak; laburdurak eta siglak; zuriuneak eta puntuazio-markak... Tokenei beharrezko informazio tipografikoa gehitzen zaie etiketen bidez (hasiera maiuskulaz, hitz osoa maiuskulaz; siglak eta zenbaki deklinatuak). Tokenizatzaileak sarrera gisa testu gordina eta XML formatuak onartzen ditu.

2. Segmentazioa edo analisi morfosintaktikoa:

   Segmentatzaileak, testu-hitz bakoitza esanahia duten unitate txikienetan, lema eta morfemetan banatzen du. Osagai hauei buruz, morfotaktika (hitz bakoitzaren ondo-

ren ager daitezkeen morfema segida posibleak eta beraien ordenaren murriztapenak) eta informazio morfologikoa (hitzaren segmentu bakoitzari dagozkion analisi morfologikoak) ematen dira. Tokenizatzailetik jasotako testu-hitz bakoitzeko interpretazio posible guztiak identifikatzen ditu segmentatzaileak. Interpretazio bakoitzean, izen eta adjektiboen kasuan hitz-formaren kategoria, azpikategoria, deklinabide atzizkia, numeroa eta mugatasuna ezaugarriak ematen dira; modu/denbora eta aspektua aditzen kasuan. Hizkuntza-informazio hau guztia *Euskararen Datu-Base Lexikaletik* (EDBL) (Aldezabal et al., 2001) jasotzen du.

3. Morfosintaxia:

   Segmentatzailean lortutako morfemei buruzko informazio morfologikoa (eta hainbat kasutan baita sintaktikoa ere) *biltzea* eta *optimizatzea* da analizatzaile morfosintaktikoaren helburua.

   Analisi morfosintaktikoa ahalbidetzen duen *gramatika*, hitzaren egitura deskribatzen duen testuingururik gabeko gramatika bat da. Gramatika honek morfemetatik lortutako informazioa konbinatu egiten du hitz-formaren interpretazio bakoitzeko ezaugarri-egitura bat emanez.

4. Hitz anitzeko unitate lexikalen (HAUL) tratamendua:

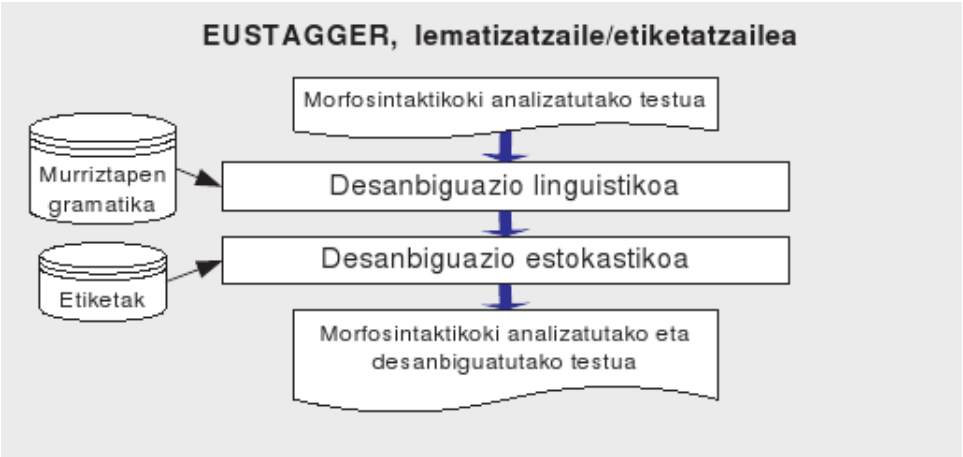
   HAUL kontsideratzen dira hitz elkartuak, lokuzioak eta kolokazio murriztuak (Alegria et al., 2004). Zehazkiago, EDBLn landu diren elementuak, inolako morfosintaxi aldaketarik behar ez duten adierazpide finkoak (in situ, a posteriori, e.a), esamolde deskonposagarri eta ez-deskonposagarriak eta lexikalizatutako termino konposatuak izan dira. Atsotitzak, kolokazioak, makulu-hitzak eta alderaketak ez dira kontuan hartu. Entitate izendunentzako, datentzako eta zenbakientzako aparteko tratamendua egin da (ikus 4.2.4).

4.2.3 EUSTAGGER, lematizatzaile/etiketatzailea

EUSTAGGER euskarako lematizatzaile/etiketatzaileak MORFEUS-ek emandako analisia jasotzen du sarrera gisa, eta sarrera horretako hitz-forma bakoitzerako testuinguru horretan hitz-forma horri dagozkion lema eta etiketak ematen ditu. MORFEUS-ek analisi morfosintaktiko guztiak ematen dituenez, EUSTAGGER arduratzen da testuinguruaren arabera analisi morfosintaktiko onartezinak baztertu eta ezabatzeaz. Hiru anbiguotasun morfosintaktiko mota bereizten dira (Ezeiza, 2002): kategoriari, morfema ez-askeei eta sintaxiari dagozkionak. Kategoriari dagokion anbiguotasuna, adibidez aditz/adjektibo/adberbio, izen/aditz... moduko anbiguotasuna, omen da zailtasunik handiena duena.

Gainera, euskararen kasuan, kategoria mailako anbiguotasunari, atzizkien anbiguotasun morfosintaktikoa gehitzen zaio.

Anbiguotasun hau ebazteko, hainbat teknika erabiltzen ditu EUSTAGGER-ek (ikus 6. irudia): hala nola, hizkuntzan oinarria duten teknikak (sinbolikoak) eta teknika estatistikoak (enpirikoak).



6. irudia: EUSTAGGER, lematizatzaile/etiketatzailea.

1. Hizkuntza-ezagutza:

   Desanbiguazio morfosintaktikorako erabiltzen den gramatikak 1.113 erregela ditu. Hauek definitzeko, IXA taldean etiketa sistema bat (Aldezabal, 2007) definitu da. Etiketa sistema honetan lau maila zehazten dira. Lehen mailan hitzaren gramatika-kategoria erabiltzen da (guztira 20 etiketa), bigarrenean, aurrekoaz gain gramatika-azpikategoria erabiltzen da (45 etiketa), hirugarrenean, aurrekoari informazio interesgarria gehitzen zaio, deklinabide-kasua adibidez, eta azkenik, analizatzaile morfologikoak emandako informazioa gehitzen da. Desanbiguazio-gramatika multzoetan banatu da aipatu berri dugun etiketa-sistemaren arabera.

2. Estatistika:

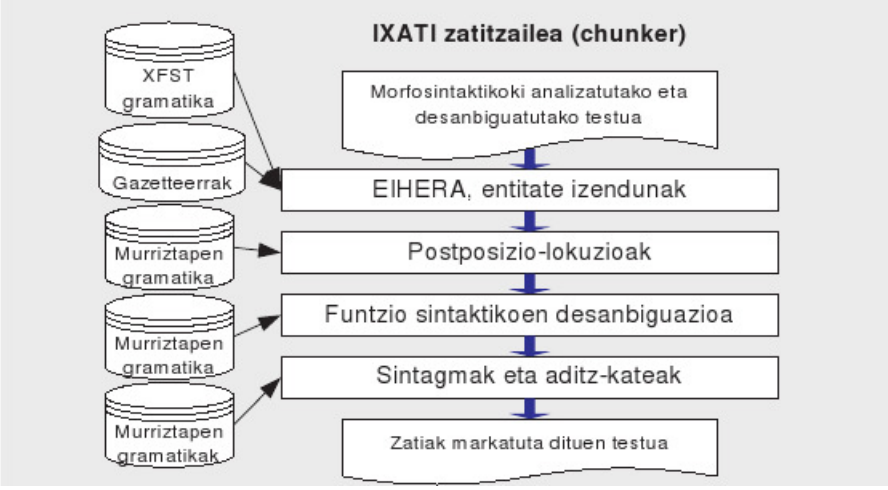
   Corpusetan oinarritutako teknika erabiltzen duen desanbiguaziorako moduluak lehen mailako Markov-en eredu ezkutuak (HMM, Hidden Markov Model) ditu oinarri. Kasu honetan, corpus handi batetik erauzitako estatistikak erabiltzen dira desanbiguazioa burutzeko.

4.2.4 IXATI zatitzailea

IXATI zatitzaileak testua *kateetan* banatzea du helburu, ingelesez *chunk* deritzonetan. Katea sintagma kategoriako zatia da eta sintaktikoki erlazionatutako hitzez osatzen da. Beraz, testua kateetan zatitzea gainjartzen ez diren eta sintaktikoki elkarrekin erlazionaturik dauden multzoak detektatzean datza. Analisi-katean sintaktikoki erlazionaturiko hitz multzo gisa kontsideratu ditugunak sintagmak eta aditz-kateak, postposizio-lokuzio eta etiketa izendunak dira. Hauen markaketa egiteko, zatitzaileak morfosintaktikoki analizatutako eta desanbiguatutako testua jasotzen du.

7. irudian ikus dezakegunez IXATI zatitzailean lau geruza-maila bereizten dira, honakoa egiten da maila bakoitzean:

HAP masterra



7. irudia: IXATI zatitzailea.

- Entitate izendunak (EIHERA):

  Entitate izendunak ezagutu eta sailkatzeko bi urratsetako metodoa konbinatzen da euskararako EIHERA tresnan (Alegria et al., 2003). Lehen pausoan, informazioa morfosintaktikoa erabiliz entitate izendunak erauzten dira helburu honetarako sortutako XFST tresnako formatua jarraitzen duen gramatika baten bidez. Bigarrenez, jadanik ezagutu diren entitate izendunak pertsona, erakundea eta tokia multzoetan sailkatzen dira testuinguruko informazioa eta *gazetteer*rak erabiliz. Entitate hauek elementu bakarrekoak (Egipto, tokia) edo hitz anitzekoak (Euskal Herriko Unibertsitatea, erakundea) izan daitezke.

- Postposizio-lokuzioak. Postposizio-lokuzioak perpausaren "sintagmen arteko erlazio gramatikalak adierazten dituzten forma askeak direla" adierazten da EGLU-I (Euskaltzaindia, 1985) gramatikan (adibidez, etxea-ren barruan, neska honi buruz...). IXATIn islapen-erregelekin osatutako gramatikak erabiltzen dira egitura hauek ezagutzeko.

- Funtzio sintaktikoen (FS) desanbiguazioa (Aduriz, 2000). Kateko analisietako bakoitzak, gramatikak hala eskatzen duelako, funtzio sintaktiko bat (edo gehiago) izaten du esleituta. Funtzio sintaktiko hauek, esaldiaren azaleko sintaxiaren berri ematen dute, batzuetan funtzio tradizionalekin (@subj, @obj...), eta besteetan sintagmak eta aditz-kateak lotzeko erabiltzen diren funtzioekin (@km>, @-jadnag_mp_subj...). Funtzio sintaktiko hauetako batzuk EDBLtik datoz, beste batzuk, ordea, morfosintaxia lantzen denean esleitzen dira. Horrela, forma bati izan ditzakeen funtzio sintaktiko guztiak esleitzen zaizkio; beraz, batzuetan, hitzak FS bat baino gehiago baditu, hitza sintaktikoki anbiguo izango da. Adibidez, oso ohikoa da absolutibo kasuetan @subj/@obj/@pred anbiguotasuna. Urrats honetan Murriztapen Gramatika erabiltzen da funtzio sintaktikoen desanbiguazioa egiteko.

- Izen-kateak eta aditz-kateak. Zati horiek hainbat murriztapen-gramatika erabiliz ezagutzen dira, horretarako funtzio sintaktikoei buruzko informazioa baliatuz. Horrela, aditz-kateak ezagutzeko, aditz-erlazio etiketak (@+jnag, @-jnag, @+jlag...) eta partikula batzuk (ezeztapenerako partikula, partikula modala...) erabiltzen dira. Sintagmak azalarazteko, berriz, honakoa onartzen da: funtzio modifikatzailea duen edozein hitz, sintaxi-funtzio etiketa nagusi bat duen (@subj, @obj, @zobj) hitz edo hitz multzoren batera lotuko da (adibidez, [Nire (@km¿) txostena (@subj)] luzea da). Are gehiago, sintaxi-funtzio etiketa nagusi bat duen hitz batek, berak bakarrik osa dezake sintagma (adibidez, [Nik (@subj)] jaso dut). Beraz, hitz bakarreko edo hainbat hitzeko sintagmak topa ditzakegu. Era berean, aditz-kateak ere elementu bakarrekoak edo hainbat elementukoak izan daitezke. Azken kate hauetan, gainera, aditz-kate jarraituak eta ez jarraituak (adibidez, [Ez da] euririk [espero]) topa ditzakegu. Gehienez ere, hiru osagai dituzten aditz-kate ez-jarraituak ezagutu dira.

4.2.5 ML-IXATI

ML-IXATI tresna euskarako kateen eta perpausen identifikatzailea da. Kate eta perpausak identifikatzeko erregelak eta FR-Perceptron ikasketa-algoritmoa erabiltzen ditu (Arrieta, 2010). Gure lanerako ML-IXATIk eskaintzen dizkigun perpaus-mugak soilik erabiliko ditugu.

4.3 Definitutako Erregelak

Aipamenen detekzioa egin aurretik, komenigarria da tratatu nahi diren testuen aurreprozesaketa bat egitea. Aurreprozesu hau egiteko Lengoaia Naturalaren Prozesamendurako tresna orokorrak erabiltzea lagungarria bada ere, tresna hauek zehazki ez dira aipamenen detekzioa egiteko sortuak izan. Horrek, tresna hauen emaitzetatik lortzen diren aipamenen mugak zehazki zuzenak ez izatea dakar, beraz, tresna hauek doitu egin behar dira aipamenen detekzioa modu egokian egin ahal izateko. Hori dela eta, sortu dugu aipamen-detektatzailea. Hau, eskuz definitutako erregelen bidez inplementatua izan da eta erregela horiek konpilatuz Egoera Finituko Transduktoreak (Finite State Transducers, FST) lortu dira.

Egoera Finituko Teknologiak datu multzo handiak azkar eta memoria erabilera txikia eginez prozesatzeko aukera eskaintzen digu. Foma (Hulden, 2009), automata eta transduktoreekin lan egiteko aukera eskaintzen duen kode irekiko aplikazioa, erabili dugu erregelak definitu eta transduktoreak lortzeko. Guztira 24 erregela definitu ditugu eta behin konpilatu ostean 8 FST lortu dira.

Gure FSTek aurreprozesaketan IXATIk itzulitako izen-sintagmak eta ML-IXATIk lortutako perpaus mugak erabiltzen dituzte aipamenak eta beraien mugak identifikatzeko. Lortutako FST bakoitza aurreko kapituluan aipatutako azterketa linguistikoko aipamen mota bat identifikatzeaz arduratzen da.

Jarraian garatu ditugun FSTak azalduko ditugu. Azalpenak gauzatzeko adibide errealez baliatuko gara, FST bakoitzaren funtzionamendua modu errazagoan ulertu ahal izateko

asmoarekin.

1. **Izenordainak**

   IXATI zatitzaileak, izenordain gehienak itzultzen baditu ere, badira batzuk ongi detektatzen ez dituenak. Gabezia horri aurre egiteko sortu dugu FST hau eta aurreprozesaketan detektatu gabe utzi dituen izenordainak identifikatzeaz arduratzen da.

   8. irudian IXATI zatitzaileak eskaintzen dizkigun etiketak[1] ikus ditzakegu. Irudi horretan agertzen den adibidean, *hark* ez da izen-kate moduan etiketatzen. Ondorioz ez da aipamen moduan kontsideratzen hasiera batean. Hala ere, definitu dugun FSTak hau detektatu eta aipamentzat kontsideratuko du.

	[Mugabe presidentea]	presionatzeko	[asmoa]	dute	hark	[okupazioarekin]	buka	dezan.	
IXATI:	SIH	SIB	AK	SINT	AK	SINT	AKH	AKB	
FST:	[Mugabe presidentea]	presionatzeko	[asmoa]	dute	[hark]	[okupazioarekin]	buka	dezan.	

8. irudia: IXATIk detektatze ez duen izenordain baten adibidea.

2. **Edutezkoak**

   Izen-kate baten barruan egon daitezkeen edutezko egiturak detektatzeaz arduratzen da. Garatu den FSTak, izen-kate baten barnean berariazko edutezko bat topatzen badu, edutezko hori aipamentzat kontsideratzen du. 9. irudian ikus dezakegunez [*bere jokalariak*] izen-kateak barnean genitiboan dagoen edutezko bat du (*bere*). Aipamentzat kontsideratzeko irizpideak betetzen direnez [*bere*] aipamen berria detektatu eta markatuko du FSTak.

	[Lotinak]	[Oronozera]	eraman	zituen	atzo	[bere	jokalariak.]	
IXATI:	SINT	SINT	AKH	AKB	SINT	SIH	SIB	
FST:	[Lotinak]	[Oronozera]	eraman	zituen	atzo	[[bere]	jokalariak.]	

9. irudia: Edutezko aipamen baten adibidea.

3. **Aditz-izenak**

   FSTak lehenik aditz-izenak (ADIZE) detektatzen ditu analisi sintaktikoa erabiliz, eta ondoren aipamenaren eskuin-muga ezartzen du aditz-izen horren ondorengo izen-kate edo koordinazioaz lotutako izen-kateen amaieran. Ezker-muga ezartzeko ML-IXATI itzultzen duen lehen ezker-muga erabiltzen da, aipamenaren hasiera bertan finkatuz. Azterketa linguistikoan aipatu den moduan, aditz-izen egitura baten barnean sinpleagoak izan daitezkeen izen-kateak aurki ditzakegu. Izen-kate horiek IXATIk itzuliko dizkigu eta aipamentzat kontsideratuko dira.

[1] Adibideetan agertzen diren laburtzapenen esanahia *Laburtzapenak* atalean dago

10. irudian dagoen adibideari begiratzen badiogu, lehenik FSTak *zabaltze* aditz-izena identifikatzen du, ondoren aditz-izen honen atzetik datorren izen-kate (*honek*) edo koordinazioz lotutako izen-kateak bilatzen ditu eta hemen aipamenaren amaiera ezartzen du. FSTak aipamenaren hasiera topatzeko, ML-IXATIk aditz-izenaren aurretik eskaintzen duen gertueneko perpaus-muga (PH) zein den begiratzen du (*EBren*) eta bertan aipamenaren hasiera ezartzen du, [*EBren zabaltze honek*] aipamen berria lortuz.

	[EBren	zabaltze	[honek]	arazo	asko	konpontzera	behartuko	du	
IXATI:	SIH	ADIZE	SINT	SIH	SIB	AK		AKH	AKB
ML-IXATI:	PH			PH		PB			PB
FST:	[[EBren	zabaltze	[honek]]	arazo	asko	konpontzera	behartuko	du	

10. irudia: Aditz-izen bat duen aipamen baten adibidea.

4. **Postposizio-lokuzioak**

Definitu den FSTa, postposizio-lokuzioetatik egituraren parte den izen-katea lortzeko gai da. Lortutako izen-kate hori aipamen bat izango da.

11. irudian ikus dezakegunez, *Montenegroko jarraitzaile sutsuen ondoan* izen-kateak postposizo-lokuzio bat (*sutsuen ondoan*) dauka bere barnean. Definitu den FSTak potsposizio-lokuzioa identifikatu eta izen-katearen amaiera, potposizio-lokuzioaren hasieran (POSH) ezartzen du, [*Montenegroko jarraitzaile sutsuen*] aipamen berria lortuz.

	[Norvegiarrak]	[ardi	otzanak]	dira	[Montenegroko	jarraitzaile	sutsuen	ondoan.]	
IXATI:	SINT	SIH	SIB	AK	SIH		POSH	POSB	
ML-IXATI:	PH							PB	
FST:	[Norvegiarrak]	[ardi	otzanak]	dira	[Montenegroko	jarraitzaile	sutsuen]	ondoan.	

11. irudia: Postposizio-lokuzio bat duen aipamen baten adibidea.

5. **Mendeko perpausa duten izen-kateak**

FST hau, mendeko perpausa duten izen-kateak tratatzeko sortu da. FSTak izen-kate bat eta berau modifikatzen duen mendeko perpaus osagarri bat topatzen duenean, aipamenaren eskuin-muga izen-katearen amaieran ezartzen du. Ezker-muga, berriz, ML-IXATIk eskaintzen dion ezker aldeko gertueneko perpaus-muga erabiltzen du, hau da, izen-katetik gertuen dagoen perpaus-muga, hain zuzen ere.

12. irudian dagoen adibidearen kasuan, mendeko perpausa (MP) non hasten den topatzen du lehenik FSTak. Ondoren, aipamenaren amaiera MP etiketaren ondoren datorren izen-katean (*konponbideak*) ezartzen du. Aipamenaren hasiera berriz, MP etiketaren ezker aldera kokatzen den lehen perpaus-mugan (*erregaien*) jartzen

du. Pauso hauek jarraituz [*erregaien krisia saihesteko konponbideak*] aipamen berria lortzen du FSTak.

	[Helburua]	[erregaien	krisia]	saihesteko	[konponbideak]	bilatzea	da.
IXATI:	SINT	SIH	SIB	MP	SINT	AKH	AKB
ML-IXATI:	PH	PH		PB			PB
FST:	[Helburua]	[[erregaien	krisia]	saihesteko	konponbideak]	bilatzea	da.

12. irudia: Perpaus osagarria duen aipamen baten adibidea.

Erlatibozko atzizkia duten aditza agertzen den egiturak modu bertsuan tratatzen dira: Aipamenaren hasiera, erlatibozko atzizkiaren ezker aldean dagoen perpaus-muga gertuenekoan ezartzen da, eta amaiera, berriz, erlatibozko atzizkia duen aditzaren ondoren dagoen izen-katean.

13. irudian ikus dezakegun adibidean, erlatibozko atzizkia duen aditza (*duen*) da. Hortik abiatuz, ezker aldera gertuen dagoen perpaus-muga *Morik* hitzean aurkitzen da eta aipamenaren hasiera hemen finkatzen da. Aipamenaren amaiera, berriz, *duen* aditzaren ondoren datorren izen-katean (*hirugarren kabinetea*). Pauso hauek jarraituz [*Morik apiriletik egingo duen hirugarren kabinetea*] aipamen berria lortzen da.

	[Morik]	[apiriletik]	egingo	duen	[hirugarren	kabinetea]	izango	da
IXATI:	SINT	SINT	AKH	ERL	SIH	SIB	AKH	AKB
ML-IXATI:	PH	PH		PB				PB
FST:	[[Morik]	[apiriletik]	egingo	duen	hirugarren	kabinetea]	izango	da

13. irudia: Erlatibozko atzizkia duen aipamen baten adibidea.

6. **Elipsia**

   Definitu den FSTak analisi sintaktikoa erabiltzen du elidítutako izen bat topatzeko. Eskuin-muga elidítutako izenaren amaieran ezartzen du eta ezker-muga, elidítutako izenaren ezker aldean gertuen dagoen perpaus-mugan.

   14. irudian ikus dezakegun adibidean FSTak elidítutako izena (*gertatutakoa*) topatzen du IXATIk eskaintzen dion ELI etiketa erabiliz, eta aipamenaren amaiera bertan ezartzen du. Ondoren, elidítutako izen horren ezkerrean dagoen perpaus-muga gertuenean zein den begiratzen du (*Azken*) bertan aipamenaren hasiera ezarriz. Horrela, [*Azken estropadan gertatutakoa*] aipamen berria lortzen du.

7. **Koordinazioa**

   Koordinazio-egituren parte diren aipamenen detekziorako FSTa garatzeak, eskatu digu lan gehien. Ez da hain ebidentea zein kasutan lortu behar diren aipamen berriak eta zein kasutan ez.

	[Azken bi estropadetan]	gertatutakoa	ikusita	[garbi]	dago	[Zumaia]	[gorantz]	doala	
IXATI:	SIH	SIB	ELI	AK	SINT	AK	SINT	SINT	AK
ML-IXATI:	PHPH			PB			PH		PB
FST:	[[Azken bi estropadetan] gertatutakoa]			ikusita	[garbi]	dago	[Zumaia]	[gorantz]	doala

14. irudia: Elipsia gertatzen den aipamen baten adibidea.

FST honek beste gramatikek lortu dituzten aipamenak jasotzen ditu sarrera moduan eta koordinazio-elementuren bat badute (eta, edo, edota...) eta banatuak izateko baldintzak betetzen badituzte, koordinazioaren ezker eta eskuin aldeetako hitzekin aipamen berriak lortzen ditu.

15. irudian agertzen den adibidean, koordinazioa duen izen-kate bat ikus dezakegu, [*Meatzari eta kobre altzairutegietako langileek*], hain zuzen. Izen-kate hau aipamentzat kontsideratuko da, 3. kapituluan aipatu dugun moduan izen-kate guztiak aipamentzat kontsideratuko baititugu. FSTak aipamen hau topatzean eta barnean koordinazio-elementu bat duela ikusirik, egitura zatigarria den kontsideratuko du. Kasu honetan hala denez, aipamen horretan koordinazioaren ezker eta eskuin aldeko testuekin bi aipamen berri lortuko ditu, [*Meatzari*] eta [*kobre altzairutegietako langileek*], hain zuzen ere.

	[Meatzari	eta	kobre altzairutegietako langileek]	bat	egin	zuten	[protestekin.]	
IXATI:	SIH	JUNT	SIB	SINT	AKH	AKB	SINT	
ML-IXATI:	PH						PB	
FST:	[[Meatzari]	eta	[kobre altzairutegietako langileek]]	bat	egin	zuten	[protestekin.]	

15. irudia: Zatigarria den koordinazio-egitura baten adibidea.

8. **Adberbioak**

FST honek lekuzko adberbioak identifikatzeko gaitasuna dauka. Euskarazko lekuzko adberbioen lema guztiak lortzeko EDBL erabili da. Lortutako lemak espresio erregular batean bildu eta konpilatu egin dira FST hau lortuz.

16. irudian dagoen adibidean, FSTak [*han*] adberbioa topatuko luke.

	[Indarrarekin]	banago	han	izango	naiz.
IXATI:	SINT	AK	ADB	AKH	AKB
FST:	[Indarrarekin]	banago	[han]	izango	naiz.

16. irudia: Lekuzko adberbio baten adibidea.

Orain arte aipamen mota bakoitzaren adibideetan oinarritu bagara ere, definitu diren erregelaren bat azaltzea komenigarria dela iruditzen zaigu. Erregelaren portaera hobekiago ulertzeko, 17. irudian agertzen den adibidea erabiliko dugu. Adibide horretan

aipamen-detektatzaileak sarrera moduan jasoko duen adibide bat, *Armada Britainiarrak Ipar Irlandan dituen bi kuartel eta beste begiratoki eraitsi dituzte*, ikus dezakegu. Horretaz gain, IXATIk itzultzen dizkigun kateen mugak (SIH, SIB, AKH, AKB) eta ML-IXATIk itzulitako perpaus-mugak (PH eta PB) ikus ditzakegu.

Sarrera horretatik, aipamentzat kontsideratuko diren izen-kateez gain, [*Armada Britaniarrak Ipar Irlandan dituen bi kuartel eta beste bi begiratoki*] aipamen berria lortzea nahiko genuke, erlatibozko aditz bat duen egitura denez, aipamentzat kontsideratu behar baita.

Jasotzen dugun sarreran oinarrituz, 18. irudian ikus daitekeen erregela definitu dugu, erlatibozko egiturak detektatu eta aipamen berria lortzeko. Lehenik, erlatibozko atzizkia duen aditza (EA) definitzen da, analisi sintaktikoan ERL etiketa duen aditz bat moduan. EA, erlatibozko aipamenaren (EAM) definizioan erabiltzen da ondoren.

Erregelak erlatibozko atzizkia duen aditz bat topatzen du lehenik (*dituen*). Ondoren, eskuin-muga ezartzen da erlatibozko atzizkia duen aditzaren ondoren dagoen izen-kate edo koordinatutako izen-kateen amaieran (*bi kuartel eta beste begiratoki*). Azkenik, ezker-muga aditzaren ezker aldetik gertuen dagoen perpaus-mugan (PH) ezartzen da (*Armada*).

Pauso hauek jarraituz, sistemak aipamen berri zuzen bat lortzen du eta egitura osoa, <AIPAMENA> eta </AIPAMENA> etiketen artean mugatzen du ondoren tratatu ahal izateko.

	[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	[bi	kuartel]	eta	[beste	bi	begiratoki]	eraitsi	dituzte
IXATI	SIH	SIB	SIH	SIB	ERL	SIH	SIB	JUNT	SIH		SIB	AKH	AKB
ML-IXATI	PH	PH			PB								PB
FST:	[[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	[bi	kuartel]	eta	[beste	bi	begiratoki]]	eraitsi	dituzte

17. irudia: Aipamen-detektatzailearen sarrera posible baten adibidea.

```
define EA Aditza & $['ERL'];
define EAM [PH H+ EA SINT [JUNT SINT]*] @-> "<AIPAMENA>"... "</AIPAMENA>";
```

18. irudia: Erlatibozko atzizkia duten aditzen egituretatik aipamenak identifikatzeko erregela.

5 Ebaluazioa

5.1 Aipamenen detekzioa ebaluatzearen garrantzia

Autore ugarik argudiatu dute aipamenen detekzioari eta korreferentzia-ebazpenari dagozkien ebaluazioak bakoitza bere aldetik egin behar direla. Recasens-en (2010) aipatzen denez, aipamenen detekzioa erabat desberdina da korreferentzia-ebazpenarekin alderatuz, ondorioz, ataza osotasunean (aipamenen detekzioa+korreferentzia-ebazpena) ebaluatzeak bi atazak bananduta ebaluatzeak baino informazioa gutxiago eskaintzen digu. Sistema batek korreferentzia-ebazpena oso ondo egin dezake baina gabeziak eduki ditzake aipamenen detekzioan, edota alderantziz gerta daiteke. Ataza bakoitzaren ebaluazioa bestearekiko banatuta egiteak gure sistemek gabeziak zein atazatan dituzten argitzeko balio du. Popescu-Belis et al. (2004) artikuluan ere aipamenen detekzioa bere baitan kontsideratu behar den ataza dela diote autoreek, eta bere ebaluazioa korreferentzia-ebazpenaren ebaluaziotik banatu behar dela. Aipamenen detekzioaren ebaluazioak sistema batek hurrengo pausoetan korreferentzia-kateetan bildu beharko diren aipamenak identifikatzeko duen gaitasuna neurtzen du.

5.2 Ebaluaziorako irizpide eta metrikak

Aipamenen detekzioa ebaluatzeko orokorrean erabili ohi diren neurriak doitasuna (P), estaldura (R) eta F-measure balioak dira. Balio hauek kalkulatzeko eskuz etiketatutako aipamenak (GOLD) eta aipamen-detektatzaileak identifikatuak (SYS) konparatzen dira. Bi multzo hauetan komunak diren aipamenak, aipamen zuzentzat hartzen dira (COR). Hau honela izanik, aipatu diren balio horiek kalkulatzeko honako formulak erabiltzen dira.

$$P = \frac{COR}{SYS} \qquad R = \frac{COR}{GOLD} \qquad F_{measure} = \frac{2.P.R}{P+R}$$

Aipamen bat zuzena dela kontsideratzen da, automatikoki detektatu den aipamenaren mugak urre-patroiaren mugen barnean badaude eta burua (head word) ere aipamenaren barnean kokatzen bada (Kummerfeld et al., 2011). Parekatze mota hau *Lenient Matching* edo *Partial Matching* bezala ezagutzen da. Hala ere, parekatze-metodo zorrotzagoak aplikatu izan dira. Adibidez, CoNLL-2011 Shared Task-en (Pradhan et al., 2011), *Strict Matching* metodoa erabili zen. Metodo honen arabera, aipamen bat zuzena dela kontsideratzen da, baldin eta soilik baldin urrezko aipamenaren berdina bada.

Gure aipamen-detektatzailea ebaluatzeko bai *Lenient Matching* bai *String Matching* parekatze-metodoak erabili ditugu.

5.3 Erabilitako corpusa

Sistema garatu eta ebaluatzeko erabili den corpusa, EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) (Aduriz et al., 2006) corpusaren zati bat izan da. EPEC corpusa Euskaldunon Egunkarian 2000. urtean argitaratu ziren berriez osatua dago. Erabili

ditugun fitxategiak bi multzo nagusitan banatu ditugu: lehenengo zatia sistemaren garapenerako erabili da eta guztira 278 aipamenez osatua dago, bigarren zatia berriz sistema ebaluatzeko erabili da eta zati hau 394 aipamenez osatua dago. Bi zatiak hizkuntzalari batek eskuz etiketatuak izan dira.

5.4 Emaitzak

Oinarri-lerroa finkatzeko, aipamen moduan IXATIk itzulitako kateak (aditz-kateak izan ezik) kontuan hartu ditugu eta hauek urre-patroiarekin konparatu. 1. taulan gure sistemak *Exact Matching* eta *Lenient Matching* parekatze-metodoak erabiliz lortutako emaitzak azaltzen dira, oinarri-lerroko balioekin konparatuz.

	Oinarri-lerroa			Aipamen-detektatzailea		
	P	R	F_1	P	R	F_1
EM	63.37	70.33	66.65	76.85	78.59	**77.58**
LM	72.01	79.75	75.65	81.96	83.97	**82.81**

1. taula: Oinarri-lerroa eta sistemak lortutako balioak.

5.5 Erroreen azterketa kuantitatibo nahiz kualitatiboa

Exact Matching parekatze-metodoa erabiltzean sistemak % 77,58ko F-measure balioa lortzen du; *Lenient Matching* erabiltzean emaitza hobea lortzen da, % 82,81 hain zuzen ere. Bi parekatze-metodoen arteko balioen desberdintasunaren arrazoia da, lehen aipatu bezala, *Lenient Matching* parekatze-metodoa ez dela *Exact Matching* bezain zorrotza.

Oinarri-lerroa eta aipamen-detektatzaileak lortzen dituen emaitzen arteko desberdintasunei begiratzen badiegu, gure sistemak lortzen duen hobekuntza nabarmena dela ikus dezakegu. *Exact Matching* erabiltzean, oinarri-lerroa 11 puntuan hobetzen da, eta 7 puntuan, berriz, *Lenient Matching* erabiltzean. Ziur gaude aipamenen detekzioan lortutako hobekuntza honek hurrengo pausoetarako onura nabarmena ekarriko duela.

Aipamen-detektatzailearen doitasun eta estaldura balioak kasu gehienetan oso antzekoak dira. Sistemak egindako erroreen ebaluazio kualitatiboa egin da, eta honakoa ondorioztatu dugu: gure sistemak aipamen hautagai bat proposatzen duen askotan, aipamen hori urre-patroiko aipamenaren berdina da. Hau kontuan edukiz argudia dezakegu gure sistemak aipamen bat lortzen badu, aipamen hori kalitate altukoa dela, hau da, aipamen zuzena izateko aukera handia duela.

Ebaluazio kualitatiboak, erroreen kausa argitzeko aukera ere eskaini digu. Puntu honetan garrantzitsua da gogoraraztea gure sistemak guztiz automatikoak diren tresnak erabiltzen dituela. Sistemak bere sarrera moduan ez du urrezko analisi sintaktikorik edota izen-kateen urrezko mugarik jasotzen. Jakina da tresna eta baliabide automatikoen erabilerak prozesuaren hasieratik errore-tasa batekin lan egitea suposatzen duela.

Egindako azterketa kualitatiboan behatu ditugun errore gehienak aurreprozesaketan erabilitako tresnen eraginez sortuak izan dira. Eskuarki, tresna hauek itzulitako izen-sintagmen mugak, urrezko aipamenen mugak gainditzen dituzte. Bistakoa da aipamen hauek okertzat kontsideratuko direla bi parekatze-metodoak erabiltzean. *Exact Matching* protokoloak okertzat kontsideratuko du, lortutako aipamena urrezko aipamenaren berdina ez baita izango eta *Lenient Matching* protokoloak, berriz, urrezko aipamenaren mugak gainditzen dituelako.

6 Integrazioa

Kapitulu honetan garatu dugun aipamen-detektatzailea erabiliz corpus baten etiketazio automatikoa nola egin dugun eta zerk bultzatu gaituen hau egitera azaldu nahi ditugu.

6.1 Motibazioa

Behin aipamen-detektatzailea garatu eta inplementatu ostean, hurrengo pausoa korreferentzia-ebazpena gauzatzea da. Korreferentzia-ebazpen automatikorako, ordea, komenigarria da eskuz etiketatutako corpus bat izatea, non testu batean entitate berberari erreferentzia egiten dioten aipamenak korreferentzia-kateetan bilduta egongo diren.

Euskarari dagokionez orain arte korreferentzia eskuz etiketatuta duen corpusa oso txikia da. Hori dela eta, corpus handiago baten beharra nabarmena da korreferentzia-ebazpen automatikoa modu eraginkor batean egin ahal izateko.

Eskuz etiketatutako corpusak lortzea, ordea, lan neketsua izan ohi da, denbora asko eskatu ohi du eta garestia izan ohi da, etiketatzaile profesionalen lan intelektual eta eskuzkoa eskatzen baitu.

Hizkuntzalarien etiketatze-lana erraztu ahal izateko eta aipamenak testu hutsetik etiketatzen hasi beharrik ez izateko, garatu dugun aipamen-detektatzaile automatikoa erabiliz EPEC corpusaren zati bat etiketatu dugu.

Modu honetan, corpusen etiketatzeak hizkuntzalari bati suposatzen dion denbora murriztu nahi izan da. Lortu den corpusa ezingo da urre-patroitzat kontsideratu, hizkuntzalari batek eskuz errebisatzen ez badu; hala ere, lan hau erraztu egin daitekeela uste dugu. Aipamen-detektatzaileak itzultzen dituen aipamen gehienak zuzenak dira eta oker daudenen artean, mugak okerrak dituztenak dira gehienak. Urre-patroia lortzeko bidean, hizkuntzalariak egin beharko lukeen lanik handiena aipamen-detektatzaileak itzultzen dituen aipamenak errebisatu eta muga okerrak dituzten kasuan muga horiek zuzentzea da.

6.2 Etiketatze-eredua

Aipamenen eta korreferentzien anotazioa formatu estandar batean egitea da egokiena. Gaur egun, XML lengoaian kodetutako MMAX2 tresnak oinarrian erabiltzen duen formatua oso erabilia da eta hizkuntzaren prozesamenduan estandartzat har daiteke.

Formatu estandarrak erabiltzeaz gain, corpus baten etiketatze prozesuak mailakatua eta hedagarria izan beharko luke eta honek ez luke arazorik sortu behar jadanik existitzen diren etiketatzeekin. Hau da, etiketatze berriek ez lukete azpian dagoen corpusa aldatu beharko. Idealena, etiketatzeak jatorrizko corpusarekiko fisikoki banatuak egon beharko lukete eta etiketatze horiek corpusa erreferentziatu bakarrik egin beharko lukete. Etiketatze modu hau *stand-off* etiketatze moduan ezagutzen da (Ide, 1998; Thompson and McKelvie, 1997).

Stand-off etiketatzea gaur egun corpusak errepresentatzeko ia estandartzat kontsidera dezakegu, corpus batean guztiz desberdinak diren fenomeno linguistikoen etiketatzeak

banatuta edukitzeko aukera eskaintzen baitigu. Etiketatze maila desberdin hauek fisikoki banatuta badaude ere, posible da maila desberdinetako fenomenoen arteko erlazioak adieraztea. Eta hau maila anitzeko etiketatze (*multi-level annotation*) moduan ezagutzen da.

Lehen aipatutako abantailez gain, etiketatze desberdinak fisikoki banatuta egiteak anotazio-lanak talde espezializatu desberdinen artean banatzeko aukera eskaintzen digu. Behin anotazio bakoitza amaitzean, anotazio guztiak maila anitzeko anotazio-metodoa erabiliz konbina daitezke.

6.3 MMAX2 anotazio tresna

Anotazio-lanetarako, anotazio software espezializatuak erabili ohi dira, etiketzaileei euskarri egokia eskaini ahal izateko. Etiketaziorako tresna espezializatu horietako bat MMAX2[2] tresna da. Tresna honek oinarrian MMAX2 formatu estandarra erabiltzen du aipamenak eta korreferentziak anotatu ahal izateko. Berau Hizkuntzalaritza Konputazionaleko proiektu errealen testuinguruan erabiltzeko sortua izan da (Müller and Strube, 2006). MMAX2 tresnak maila anitzetako etiketatze linguistikoak sortu, erakutsi eta bilaketak egiteko aukera eskaintzen digu.

MMAX2 anotazio-tresnak maila anitzeko etiketatze-sistema erabiltzen duela eta anotazioak formatu estandar bat erabiliz gordetzen dituela jakinik, MMAX2 tresna aukeratu dugu gure lanerako. Alde batetik, formatu estandarren erabilera ziurtatuta daukagu. Bestalde, orain momentuan aipamenen etiketatzea soilik egin bada ere, etorkizun hurbilean korreferentziaren etiketatzearekin jarraituko dugu. Maila anitzeko etiketatzeari esker, bi fenomeno horien etiketatzea paraleloan egin ahal izango da, lan bakoitza pertsona desberdinek egiteko aukera izanik. Gainera, bi fenomeno horietaz gain beste fenomeno linguistikoren bat etiketatu nahi badugu, ez dugu inolako arazorik izango orain arte egin den etiketatze-lanarekin.

6.4 Aipamenak MMAX2 formatuan

Aipamenen anotaziorako erabiliko den formatu estandarra zehaztuta, gure aipamenen detektzailearen irteera MMAX2 tresnak erabiltzen duen XML formatura pasa dugu.

Egindako lana aztertu aurretik ordea, komenigarria da MMAX2 formatuan aipamenak adierazteko erabiltzen diren fitxategi garrantzitsuenen nondik norakoak azaltzea.

Honakoak dira fitxategi garrantzitsuenak:

- *fitx.words.mmx.xml* fitxategia

  Fitxategi batean dauden hitz guztiak errepresentatzen dira bertan, hitz bakoitzak unibokoa eta bakarra den identifikadore (id) bat du.

  19. irudian hitzak nola errepresentatzen diren ikus dezakegu.

[2] http://mmax2.sourceforge.net/

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE words SYSTEM "words.dtd">
<words>
<word id="word_1">Miel</word>
<word id="word_2">Saralegi</word>
<word id="word_3">Harri-jasotzailea</word>
<word id="word_4">"</word>
<word id="word_5">Belauna</word>
<word id="word_6">ez</word>
<word id="word_7">daukat</word>
<word id="word_8">orain</word>
<word id="word_9">10</word>
<word id="word_10">urte</word>
<word id="word_11">bezala</word>
<word id="word_12">;</word>
</words>
```

19. irudia: MMAX2 tresnan hitzak errepresentatzeko modua.

- *words.dtd* fitxategia

  Hitz batek behar dituen eta izan ditzakeen atributuak definitzen dira DTD (*Document Type Definition*) fitxategi honetan. **.words.mmx.xml* motako XML fitxategi bat ondo eratuta dagoen egiaztatzeko balio digu, fitxategia DTDaren kontra balidatuz.

  20. irudian ikus daiteke *words.dtd* fitxategia nola osatuta dagoen.

```
<!ELEMENT words (word*)>
<!ELEMENT word (#PCDATA)>
<!ATTLIST word id ID #REQUIRED>
<!ATTLIST word starttime CDATA #IMPLIED>
<!ATTLIST word endtime CDATA #IMPLIED>
```

20. irudia: words.dtd fitxategia.

- *fitx_coref_level.xml* fitxategia

  Fitxategi honetan aipamenak (*markables*) errepresentatzen dira. Aipamen bakoitzak unibokoa eta bakarra den identifikadore bat, aipamena zein hitzetan hasi eta amaitzen den (*span*) eta zenbait kasutan aipamenaren mota(*np_form*) atributuak ditu. Zenbait kasutan, bere numeroa (*number*) eta funtzio sintaktikoa (*grammatical_role*) atributuak izan ditzake. Aipatu beharra dago *span* atributuak hartzen

dituen balioak, fitx.words.mmx.xml hitzen identifikadoreei dagozkiela. Adibidez, *span="word_61..word_62"* balioak adierazten digu aipamena fitx.words.mmx.xml fitxategian 61 identifikadorea duen hitzean hasi eta 62 identifikadorea duenean amaitzen dela.

21. irudian aipamenak nola errepresentatzen diren ikus dezakegu.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/coref">
<markable id="markable_1" mmax_level="coref" span="word_1..word_1"> </markable>
<markable id="markable_4" mmax_level="coref" span="word_23..word_23"> </markable>
<markable id="markable_5" mmax_level="coref" span="word_58..word_58"> </markable>
<markable id="markable_7" mmax_level="coref" span="word_61..word_62"> </markable>
<markable id="markable_9" mmax_level="coref" span="word_92..word_92"> </markable>
<markable id="markable_12" mmax_level="coref" span="word_637..word_638"grammatical_role="obj"> </markable>
<markable id="markable_80" mmax_level="coref" span="word_220..word_221"np_form="defnp"number="P">
</markable>
<markable id="markable_81" mmax_level="coref" span="word_224..word_224"number="S"> </markable>
</markables>
```

21. irudia: MMAX2 tresnan aipamenak errepresentatzeko modua.

- *markables.dtd* fitxategia

  Aipamen bakoitzak izan behar dituen eta izan ditzakeen atributuak zehazten diren DTDa adierazten da fitxategi honetan. fitx_coref_level.xml fitxategiaren balidazioa egiteko erabiltzen da.

  22. irudian markables.dtd fitxategia ikus dezakegu.

```
<!ELEMENT markables (markable*)>
<!ATTLIST markable id ID #REQUIRED>
<!ATTLIST markable span CDATA #REQUIRED>
<!ATTLIST markable type CDATA #REQUIRED>
<!ATTLIST markable member CDATA #IMPLIED>
<!ATTLIST markable pointer IDREF #IMPLIED>
```

22. irudia: markables.dtd fitxategia.

Behin MMAX2 tresnak oinarrian dituen fitxategi garrantzitsuenak aztertuta, gure aipamen-detektatzailearen irteera formatu honetara nola egokitu den azalduko dugu.

4.2 azpi-atalean aztertu dugun analisi-kateak, testu bateko hitz bakoitzari identifikadore uniboko bat esleitzen dio. MMAX2k erabiltzen duen formatuan ere hitz

bakoitzak identifikadore uniboko bat behar duela aipatu dugu oraintxe. Hori horrela izanik, fitx.words.mmma.xml fitxategia sortzeko analisi-kateko jatorrizko fitxategia erabili da, eta analisi-kateko hitzaren identifikadorea fitx.words.mmax.xml fitxategiko identifikadorearen berdina izango da.

Aipamen edo markagaien fitxategia sortzeko, berriz, gure aipamen-detektatzailearen irteera erabili dugu. Irteera horretan, lerro bakoitzean aipamen bat lortuko dugu eta aipamen bakoitzean, bera osatzen duten hitzak izango ditugu. Hitz bakoitzak berekin doan identifikadorea eta analisi sintaktikoa izango ditu. Beraz, erabat tribiala da fitx_coref_level.xml fitxategia sortzea: aipamen bakoitzaren lehen hitzaren eta azken hitzaren identifikadoareak hartu eta span atributuan ezarri.

Aipamen-detektatzailea EPEC corpuseko fitxategi guztietatik pasa eta bere emaitzak MMAX2 tresnan erabiltzeko prest utzi dira.

6.5 Aipamenen sailkatze automatikoa

Aipamen-detektatzailearen emaitzak MMAX2 tresnarako prestatu ditugu. Gainera, aipamen bakoitza zein motatakoa den (izen berezia, izenordain-pertsonala, lekuzko adberbioa, izen-kate mugatua...), zein numero duen (singularra, plurala edo mugagabea) eta zein funtzio sintaktiko (subjektua, objektua, adizlaguna...) betetzen duen adierazi dugu. Informazio hau oso baliagarria da ondoren korreferentziaren ebazpena egin ahal izateko, izan ere, aipamenaren ezaugarrien arabera korreferentzia-kate baten parte den edo ez jakiteko erabakigarria izango baita.

Anali sintaktikotik datorkigun informazioa erabiliz eta 3. kapituluan azaldu dugun sailkapena uztartuz, aipamenak 2. taulan agertzen diren moduan sailkatu ditugu. Horretarako gramatika txiki bat prestatu *foma* bidez eta FST batean konpilatu dugu. FST honek modu automatikoan, aipamen bakoitzaren burua topatu ondoren, aipamena-mota zehaztuko du eta dagokion etiketa jarriko dio.

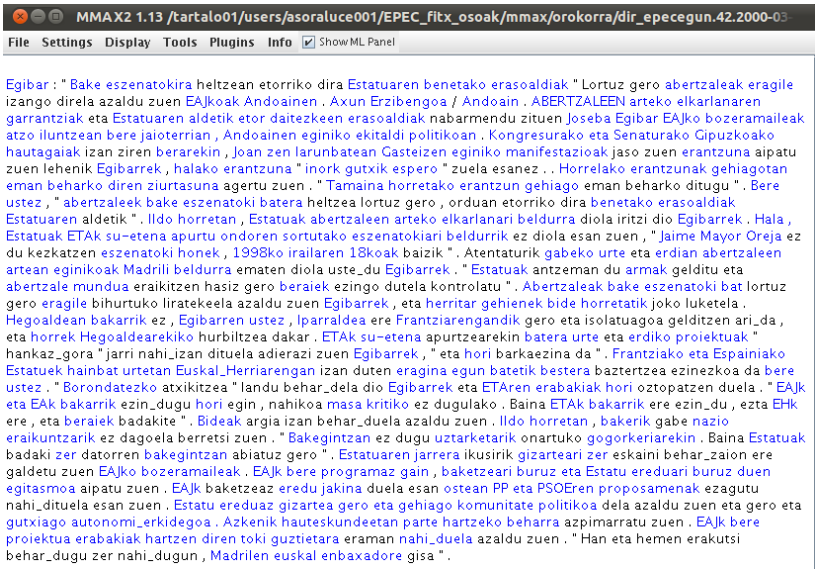
Egindako lana argiago ikusi ahal izateko 23. irudian, aipamenak automatikoki etiketatu zaizkion fitxategi baten adibidea ikus dezakegu, MMAX2 tresnan ageri den bezala. Irudi horretan urdinez agertzen den testu zati bakoitza aipamen bat da.

24. irudian, berriz, *EAJkoak* aipamen konkretuaren informazioa ikus dezakegu. *ne_o* etiketa ezarri zaio, hau da, erakunde bat adierazten duen izen berezi bat da, numeroari dagokionez singularrean dago eta subjektu funtzioa betetzen du.

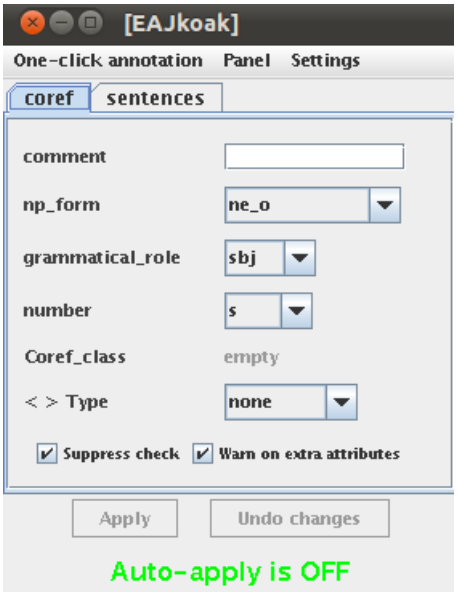
6.6 Integrazioaren emaitzak

EPEC corpusaren zati batean aipamenak automatikoki etiketatu ditugu. Corpus hau, 240779 hitzez osatua dago eta guztira 65525 aipamen automatikoki etiketatzea lortu da. Corpusean aipamenak automatikoki etiketatzeaz gain, hauen sailkapen automatikoa ere egin da. 3. taulan ikus dezakegu aipamen-mota bakoitzaren corpuseko agerpen kopurua.

Etiketazio prozesua estandarrak erabiliz gauzatu da eta IXAren analisi-katearekin ondo uztartzen den lana egin da.



23. irudia: Aipamenak automatikoki etiketatuak dituen fitxategi baten adibidea.



24. irudia: Aipamen zehatz baten informazioa erakusteko leihoa.

Etiketa	Azalpena	Adibidea
ne_p	Izen bereziak: Person	Txillardegi
ne_l	Izen bereziak: Location	Euskal Herria
ne_o	Izen bereziak: Organization	Eusko Jaurlaritza
adv_l	Lekuzko adberbioak	hemen / hor / han
post_pre	Postposizioan aurretik doazen izen-kateak	etxearen (ondoan)
defnp	Izen-kate mugatuak, artikuludunak	etxea
defnp_dem	Izen-kate mugatuak, erakusledunak	etxe hau / hori / hura...
defnp_pos	Izen-kate mugatuak, aurretik posesiboa dutenak	honen / horren / haren etxea
defnp_pos_enf	Izen-kate mugatuak, posesibo indartuak	bere
indefnp	Izen-kate mugagababeak	hainbat etxe
p_per	Izenordain pertsonalak	berau
p_pos	Izenordain posesiboak	harena, berea,...
p_ds	Izenordain gisa jokatzen duten erakusleak	hau / hori / hura
p_enf	Izenordain gisa jokatzen duten erakusle indartuak	bera
p_rec	Izenordain elkarkariak	elkar

2. taula: Aipamen moten sailkapena.

Aipamen-mota	Agerpen-kopurua
Izen Berezia: Person	3907
Izen Berezia: Location	3305
Izen bereziak: Organization	3015
Lekuzko adberbioak	501
Postposizioan aurretik doazen izen-kateak	2240
Izen-kate mugatuak, artikuludunak	34862
Izen-kate mugatuak, erakusledunak	1735
Izen-kate mugatuak, aurretik posesiboa dutenak	271
Izen-kate mugatuak, posesibo indartuak	1360
Izen-kate mugagababeak	9804
Izenordain pertsonalak	514
Izenordain posesiboak	1631
Izenordain gisa jokatzen duten erakusleak	1446
Izenordain gisa jokatzen duten erakusle indartuak	281
Izenordain elkarkariak	74

3. taula: Aipamen-mota bakoitzaren agerpen-kopurua.

7 Ondorioak eta etorkizuneko lana

Egoera finituko teknologia erabiliz garatu dugun euskararako aipamen-detektatzaile bat aurkeztu dugu. Aipamen-detektatzailea euskarazko testuetako aipamenen azterketa linguistiko sakon batean oinarrituta garatu da.

Autore askok argudiatu duten moduan, aipamenen detekzioa ataza erabakigarria da korreferentzia-ebazpenerako sistema batean. Hala ere, lan batzuetan aipamena zer den edo zer ez den azaletik tratatu izan da. Gure hipotesiaren arabera, aipamenen inguruko azterketa linguistiko bat egin eta aipamen hauen ezaugarri linguistikoak ondo definitzeak emaitzak hobetzen lagundu dezake eta lan honi esker hori frogatu ahal izan dugu.

Aipamen-detektatzailea garatzeaz gain, honi aplikazio erreal bat eman zaio. Corpus oso bateko aipamen guztiak (65525 guztira) automatikoki etiketatu dira eta corpus hori etiketatzaileentzako oso lagungarria den etiketatze-tresna batean, MMAX2 tresnan, erabiltzeko prest utzi da. Aipamenak automatikoki etiketatzeaz gain, gainera, hauen sailkapen automatikoa egin da. Argi dago aplikazio erreal honek hizkuntzalarien etiketatze-lana asko erraztu duela.

Sistemak lortu dituen emaitzak artearen egoerarekin bat datozenak dira, kontuan izanik euskararako korreferentzia-ebazpenerako egin den lehen saiakera izan dela. Sistemak lortutako F-measure balioa % 77,58koa da *Exact Matching* ebaluazio protokoloa erabiltzean eta % 82,81koa *Lenient Matching* erabiltzean.

Etorkizunean, aipamen-detektatzailearen ebaluazioa sakondu eta eraginkortasuna hobetu nahi dugu, aipamenen muga hobeak lortzeko hainbat erregela espezifiko definituz.

Etorkizuneko beste lanetako bat, aipamenen sailkatze automatikoan gehiago sakontzea da. Garatu den FSTa oso sinplea da eta datorkion analisi-sintaktikoa erabiliz egiten du aipamen bakoitzaren sailkapena. Sailkapena hobetzeko asmotan, erabili ditzakegun estrategia berriak aztertu nahi dira.

Azkenik, aipamenen etiketatze automatikoa hizkuntzalari batek gainbegiratu eta dauden akatsak zuzentzeko asmoa daukagu, honela, aipamenen urre-patroizko corpusa lortuz. Corpus hau, hurrengo pausoan, korreferentzia-ebazpenean, baliabide erabilgarria izango da. Hala ere, aipamen-detektatzaile automatikoak lortzen dituen emaitzekin korreferentzia-ebazpenari ekiteko moduan gaude urre-patroizko corpusa eduki aurretik.

Erreferentziak

Aduriz, I. (2000). *EUSMG: Morfologiatik sintaxira murriztapen gramatika erabiliz. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak*. PhD thesis, Filologia eta Historia-Geografia Fakultatea, UPV-EHU.

Aduriz, I., Aldezabal, I., Aranzabe, M., Arriola, J. M., Ceberio, K., Estarrona, A., Iruskieta, M., Lersundi, M., Pociello, E., Uria, L., Urizar, R., and Aldasoro, E. (2008). Euskarazko postposizio-lokuzioen tratamendu konputazionala. Technical report.

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., Diaz-De-Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and 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. pages 1–15. Rodopi. Book series: Language and Computers.

Aldezabal, I. (2007). Ixa taldeko etiketen eskuliburua. Technical report.

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernandez, G., and Lersundi, M. (2001). EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *IRCS Workshop on linguistic databases. Philadelphia (USA).*

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and Treatment of Multiword Expressions in Basque. In *ACL workshop on Multiword Expressions*, pages 48–55.

Alegria, I., Artola, X., Sarasola, K., and Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203.

Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información*, (JOTRI 2003), pages 198–203, Madrid, Spain.

Arrieta, B. (2010). *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* PhD thesis, University of the Basque Country.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology.* CSLI Publications.

Chang, K.-W., Samdani, R., Rozovskaya, A., Rizzolo, N., Sammons, M., and Roth, D. (2011). Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 40–44, Stroudsburg, PA, USA.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program–Tasks, Data, and

Evaluation. In *Proceedings of Language Resources and Evaluation Conference*, (LREC 2004), pages 837–840.

Euskaltzaindia (1985). *Euskal Gramatika. Lehen Urratsak-I.* Euskaltzaindia, Burlata.

Ezeiza, N. (2002). *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* PhD thesis, University of the Basque Country.

Hacioglu, K., Douglas, B., and Chen, Y. (2005). Detection of entity mentions occurring in english and chinese text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (HLT 2005), pages 379–386, Stroudsburg, PA, USA.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (EACL 2009), pages 29–32, Stroudsburg, PA, USA.

Ide, N. (1998). Corpus encoding standard: Sgml guidelines for encoding linguistic corpora. In *In Proceedings of the First International Language Resources and Evaluation Conference*, pages 463–70.

Kim, Y., Riloff, E., and Gilbert, N. (2011). The Taming of Reconcile as a Biomedical Coreference Resolver. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 89–93, Portland, Oregon, USA.

Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. (2011). Mention Detection: Heuristics for the OntoNotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 102–106, Stroudsburg, PA, USA.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 28–34, Stroudsburg, PA, USA.

Maleki, J., Yaesoubi, M., and Ahrenberg, L. (2009). Applying Finite State Morphology to Conversion Between Roman and Perso-Arabic Writing Systems. In *Proceeding of the 2009 conference on Finite-State Methods and Natural Language Processing*, pages 215–223, Amsterdam, The Netherlands, The Netherlands. IOS Press.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for conference resolution. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, (IJCAI 1995), pages 1050–1055, San Francisco, CA, USA.

MUC-6 (1995). Coreference Task Definition (v2.3, 8 Sep 95). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 335–344, Columbia, Maryland, USA.

MUC-7 (1998). Coreference Task Definition (v3.0, 13 Jul 97). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment Analysis: Does Coreference Matter? In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, pages 37–40.

Oronoz, M. (2008). *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura.* PhD thesis, University of the Basque Country.

Peral, J., Palomar, M., and Ferrández, A. (1999). Coreference-oriented interlingual slot structure & machine translation. In *Proceedings of the Workshop on Coreference and its Applications*, (CorefApp 1999), pages 69–76, Stroudsburg, PA, USA.

Poon, H., Christensen, J., Domingos, P., Etzioni, O., Hoffmann, R., Kiddon, C., Lin, T., Ling, X., Mausam, Ritter, A., Schoenmackers, S., Soderland, S., Weld, D., Wu, F., and Zhang, C. (2010). Machine reading at the University of Washington. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, (FAM-LbR 2010), pages 87–95, Stroudsburg, PA, USA.

Popescu-Belis, A., Rigouste, L., Salmon-Alt, S., and Romary, L. (2004). Online Evaluation of Coreference Resolution. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1507–1510, Lisbon, Portugal.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 1–27, Stroudsburg, PA, USA.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, (ICSC 2007), pages 517–526, Washington, DC, USA. IEEE Computer Society.

Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation.* PhD thesis, University of Barcelona.

Steinberger, J., Poesio, M., Kabadjov, M. A., and Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore.

Thompson, H. S. and McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97: The next decade – Pushing the Envelope*, page 227–229.

Uryupina, O. (2008). Error Analysis for Learning-based Coreference Resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Uryupina, O. (2010). Corry: A system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 100–103, Uppsala, Sweden. Association for Computational Linguistics.

van Deemter, K. and Kibble, R. (1995). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26:629–637.

Vicedo, J. and Ferrández, A. (2006). Coreference In Q&A. In *Advances in Open Domain Question Answering*, volume 32 of *Text, Speech and Language Technology*, pages 71–96. Springer.

Zabala, I. and Odriozola, J. C. (2004). Los complejos posposicionales en vasco. In *Las fronteras de la composición en lenguas románicas y en vasco*, pages 281–315. E.Pérez Gaztelu and I.Zabala and L.Gràcia (eds.).

Zhekova, D. and Kübler, S. (2010). UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 96–99, Stroudsburg, PA, USA.