EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

ASKHi: Analisi sintaktiko konputazional hibridoa paradigma desberdinen konbinazioan oinarrituta

Iakes Goenaga Azkarate

2017

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

ASKHi: Analisi sintaktiko konputazional hibridoa paradigma desberdinen konbinazioan oinarrituta

Iakes Goenaga Azkaratek Koldo Gojenola Galletebeitia eta Nerea Ezeiza Ramosen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2017ko otsaila.

Eskerrak

Tesi-lan honekin gora eta behera bost urte luzez jardun eta gero, urte horietan guztietan nire ondoan egon zaretenei eskerrak emateko ordua iritsi zait. Benetan diotsuet, zuek uste baino gehiago lagundu didazue tesia aurrera eramateko bide honetan.

Lehenik eta behin, nire bi zuzendariak eskertu nahi ditut urte hauetan egin duten lanagatik. Koldo eta Nerea, Nerea eta Koldo, tesi-lan hau ez zen berdina izango zuen laguntza izugarririk gabe. Oso ondo zuzendu duzue tesia eta beti egon zarete hor laguntza behar izan dudan bakoitzean. Ez naiz inoiz bakarrik sentitu bide luze honetan eta galduta egon naizen momentuetan ondo zuzendu nauzue berriro bide onera. Eskerrik asko, benetan.

Esker bereziak eman nahi dizkiet IXA taldeko kideei eta kide ohiei. IXA taldean ikasi dut ikerketa mundua zer den eta zuek bezalako lankideak izanda dena izan da askoz errazagoa. Profesional bikainak zarete, baina ez da ezaugarri hori zuen ezaugarririk hoberena, ezer bueltan eskatu gabe laguntzeko duzuen erraztasuna baizik. Oso giro ona izan dugu taldean hasiera-hasieratik eta hori ezin da diruarekin ordaindu. Kafe orduan, bulegoan, bazkarietan, afarietan eta hainbat aisialditan sekulako barreak bota ditut zuekin, eta horrek pila bat lagundu dit. Zuei guztioi ere, eskerrik asko, benetan.

Nire lagunei ere eskerrak eman nahi dizkiet. Zuekin egindako afariek eta zuekin hartutako zerbezek asko lagundu didate momentu txarretan deskonektatzeko eta indarberritzeko. Hori gutxi balitz, ideia on bat baino gehiago eman didazue. Zuei ere, eskerrik asko, benetan.

Etxekoei, amari, aitari, arrebari, ilobei eta amonari. Zuek jakin gabe

zuekin pasatutako momentu bakoitzak izugarri lagundu didalako. Zuek nire ondoan izateak dena erraztu didalako eta estres handiena izan dudan momentuetan tesia nola doan ez galdetzeagatik. Aita, orain dela 7 urte utzi gintuzun, oraindik informatikan ingeniaria ez nintzela. Dena ondo badoa hemendik gutxira informatikan doktorea izango naiz. Zauden lekuan zaudela, espero dut ni zutaz sentitu naizen bezain harro sentitzea nitaz. Egunen batean azalduko dizkizut liburuxka honek gordetzen dituen sekretuak. Ama, aita, amona, Legen, Aimar eta Markel, eskerrik asko familia!

Bukatzeko, urte hauetan gehien lagundu didan pertsonari eskerrak eman nahi dizkiot. Amane, zu izan zara momentu txarretan beti lagundu didana, gauzak beste era batera ikusten lagundu didana, ahoan irribarre batekin aurrera egiten lagundu didana, nire itsasargia izan zara, nire desertuko oasia. Egun luze eta gogor baten ostean etxera bueltatu eta zu etxean aurkitzeak bizia eman dit, zu nire ondoan izateak edozertarako indarrak eman dizkit. Amane, batez ere zuri, eskerrik asko bihotz bihotzez!

Laburpena

Hizkuntzaren Prozesamenduan sintaxiak berebiziko garrantzia du. Hainbat atazatan erabiltzen da sintaxitik eratorritako informazioa, esaterako itzulpen automatikoan, rol semantikoen etiketatzean eta sentimenduen analisian. Tesi-lan honetan sintaxi konputazionala landu da, zehazki dependentzietan oinarritutako sintaxia jorratu da analizatzaile automatikoen bidez. Dependentzien analisi sintaktiko automatikoa hobetzeko bide desberdinak aztertu dira: izaera desberdinetako analizatzaileen hibridazioa, ezaugarrien ingeniaritzako tekniken erabilpena, multzokatze mota desberdinen esperimentazioa eta automatikoki analizatutako zuhaitz-bankuetatik eratorritako ezaugarrien erabilpena.

Bide horiek guztiak jorratzearen arrazoi nagusia morfologikoki aberatsak diren hizkuntzen dependentzien analisia hobetzen lagundu dezaketen era desberdinak aztertzea da. Hori dela eta, egindako esperimentu gehienak bost hizkuntza desberdinetan probatu dira (euskara, frantsesa, alemana, hungariera eta suediera), eta hizkuntza horietan guztietan probatu ezin izan diren bideak euskararekin probatu dira, euskararen dependentzien analisia baita bereziki hobetu nahi dena.

Tesi-lan honen beste atal garrantzitsua euskararako baliabideak sortzea da, sintaxiari hertsiki lotutako baliabideak hain zuzen ere. Tesiak iraun dituen urteetan baliabide desberdinak sortu dira, baina bi dira nabarmentzeko modukoak. Alde batetik, 150 milioi hitzeko zuhaitz-bankua etiketatu da sintaktikoki era automatikoan; beste aldetik, euskarazko jatorrizko zuhaitz-

bankua nazioarteko Dependentzia Unibertsalak proiektuan proposatzen den formatura bihurtu da. Lehenengo corpusa tesi-lan honetan erabili da automatikoki analizatutako zuhaitz-bankuetatik eratorritako ezaugarriak sortzeko, eta bigarrena edozeinek erabil dezake, publikoki atzigarri baitago.

Gaien aurkibidea

1 Tesi-lanaren aurkezpen orokorra

1.1 Sarrera

Esaldi bat ondo ulertzeko prozesuan hizkuntzaren maila desberdinak hartu behar dira kontuan. Morfologiak, esaterako, hitzen barneko loturak ongi eratuta daudela bermatzen du. Har ditzagun adibide bezala *lagunero* eta *laguntzaile* hitzak. Lehenengo hitzaren kasuan, *lagun* hitza eta maiztasuna adierazten duen *-ero* atzizkia lotu dira. Lotu diren bi elementuak zuzenak diren arren, sortu den hitz berria ez da morfologikoki zuzena. *laguntzaile* hitzaren kasua, aldiz, desberdina da. Kasu horretan, *lagun* hitza eta ekintza burutzen duena adierazten duen *-tzaile* atzizkia lotu dira morfologikoki zuzena den hitz bat sortzeko.

Bestalde, sintaxiak esaldiko hitzak modu egokian erabili direla bermatzen du, hitzen arteko lotura egokiak ziurtatuz. Adibidez, *daiteke ikus etxea* esaldiko hitzak zuzenak diren arren (existitzen dira), esaldia ez dago ongi eratuta, esaldia sintaktikoki zuzena izateko hitzen ordena *etxea ikus daiteke* edo *ikus daiteke etxea* izan behar delarik.

Semantikaren kasuan, berriz, esaldiaren esanahiak berebiziko garrantzia du. Esaldi bat semantikoki zuzena izateko ez da nahikoa hitzen barneko lotura eta hitzen arteko lotura egokiak erabiltzea, esaldiak zentzua izan behar du. Adibidez, *aulkia berdea da* esaldia semantikoki zuzena da, baina *gorria berdea da* esaldia ez.

Esaldi baten ulermenerako garrantzitsuak diren mailekin bukatzeko pragmatika geratzen zaigu. Pragmatikak, esaldi bat testuinguru batean egokia dela bermatzen du. Adibidez, taxilari bati *zenbat egunero egiten duzu he-*

gan? galdetzea pragmatikaren ikuspuntutik ez dirudi zuzena, baina galdera berdina hegazkineko pilotu bati egitea bai.

Esanak esan, tesi-lan hau sintaxian kokatzen da, zehazki, sintaxi konputazionalean. Sintaxi konputazionalean, esaldien egitura sintaktikoak zehazten dira ordenagailuen bidez. Sintaxia modu horretan lantzearen abantaila nagusia azkartasuna da, milaka esaldi aztertu baitaitezke minutuetan. Desabantaila nagusia, ostera, lortutako analisi sintaktikoaren kalitatea eskuz aztertutakoarena baino eskasagoa dela da. Hori dela eta, tesi-lan honen helburu nagusia analisi sintaktiko automatikoa hobetzen lagunduko duten teknikak aztertzea da, batez ere euskararen analisi sintaktikoa hobetzeko, baina ikuspegi eleaniztuna ahaztu gabe.

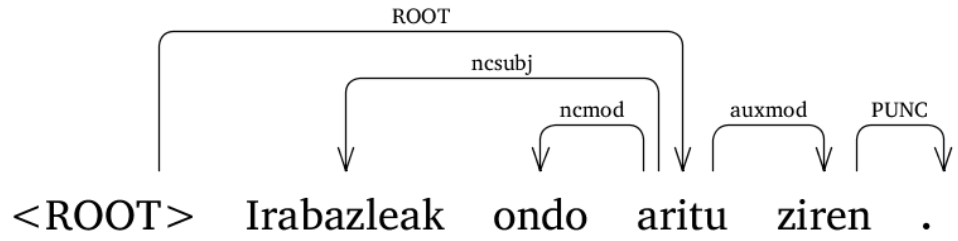
Sarrera bezala erabiliko dugun kapitulu hau modu honetara antolatu dugu: 1.2 puntuan, sintaxi konputazionala zer den aztertuko dugu sakonago. 1.3 puntuan, tesi-lan hau IXA taldeko lanen artean kokatuko dugu. 1.4 puntuan, berriz, definitu ditugun helburuak eta helburu horiek lortzeko planteatu diren hipotesiak azalduko ditugu. Bukatzeko, 1.5 puntuan, tesia nola antolatu dugun azalduko dugu, atal bakoitzarekin lotuta dauden argitalpenekin eta sortutako baliabideekin batera.

1.2 Sintaxi konputazionala

Sintaxi konputazionala hizkuntzalaritza konputazionalaren azpialor bat da, eta esaldien analisi sintaktikoa, azalekoa nahiz osoa, ordenagailuen bidez automatikoki egitean datza. Sintaxi konputazionalean, esaldien analisi sintaktikoa automatikoki burutzeko bide bat baino gehiago aurki daiteke. Bide horien artean erabilienak ezagutza linguistikoan oinarritutakoak (erregeletan oinarritutakoak) eta estatistiketan oinarritutakoak dira.

Ezagutza linguistikoan oinarritutako analizatzaile bat erabiltzeko ezinbestekoa da aztertu nahi diren hizkuntz fenomenoak deskribatzen dituzten erregelak eta prozedurak definituta egotea (gramatika). Kasu gehienetan, gramatika deskribatzen duena pertsona bat izaten da eta gramatika horrekin hizkuntza bakar bat landu daiteke.

Estatistiketan oinarritutako analizatzaile bat sortzeko, aldiz, analizatzaileak ikasteko erabiliko duen corpusa (zuhaitz-bankua) beharrezkoa da. Oso garrantzitsua da corpus horren kalitatea ona izatea, analizatzaileak erdietsiko duen emaitza ikasiko duen corpus horri hertsiki lotuta baitago. Mota honetako analizatzaileen abantaila hizkuntzarekiko independenteak direla da,



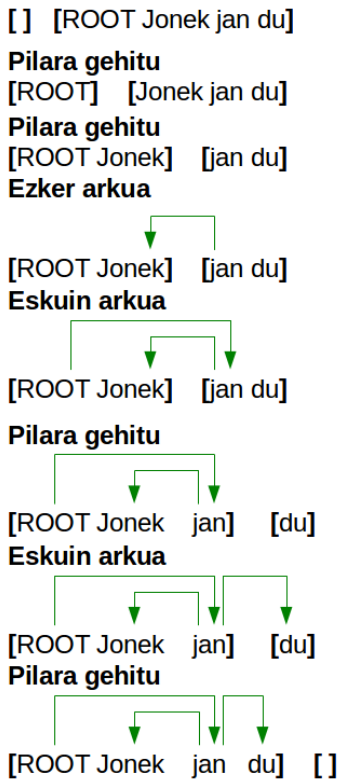
1.1 irudia – *Irabazleak ondo aritu ziren* esaldiaren dependentzia egitura.

aztertu nahi den hizkuntzaren corpus bat izatea nahikoa delarik hizkuntza horretako esaldiak sintaktikoki aztertzen ikasteko.

Tesi-lan honetan, aipatutako analizatzaile sintaktikoen bi motak erabili diren arren, oinarri bezala beti hartu dugu estatistiketan oinarritutakoa. Landu ditugun esperimentuetan beti saiatu gara sintaxi osoa lantzen duten (zuhaitz osoa eraikitzen duten) estatistiketan oinarritutako analizatzaileak edo *parser*rak ezaugarri berrien edo teknika desberdinen bidez aberasten, oinarrizko emaitzak hobetzeko xedearekin.

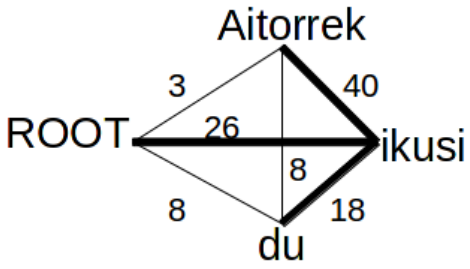
Egun, parserrek bi egitura nagusirekin egiten dute lan. Batzuek osagai egiturarekin egiten dute lan (*constituency-based*) eta beste batzuek dependentzia egiturarekin (*dependency-based*). Osagai egituren bidezko analisian lortutako osagaiak eta beren kategoriak definituz ematen da emaitza (izensintagmak, esaldiak, ...). Dependentzia egituren kasuan, aldiz, erlazionatuta dauden osagaien arteko erlazioak deskribatzen dira. Esan beharra dago tesi-lan honen ardatza dependentzietan oinarritutako analisia dela (ikus 1.1 irudia). Izan ere, gure asmoa analisi sintaktikoa hobetzeko bide desberdinak aztertzea baita eta dependentzietan oinarritutako analisi sintaktikoa lantzeko baliabide gehiago daude eskuragarri (hizkuntza desberdinetako zuhaitzbankuak, parserrak, ...).

Ildo beretik, dependentzietan oinarritutako analisi sintaktikoa lantzen duten parserren artean bi mota nagusi bereizten dira: trantsizioetan oinarritutakoak (Yamada eta Matsumoto, 2003; Titov eta Henderson, 2007; Nivre *et al.*, 2007b) eta grafoetan oinarritutakoak (Eisner, 1996; McDonald *et al.*, 2005, 2006; Bohnet, 2010). Trantsizioetan oinarritutako analizatzaileek normalean pila bat eta buffer bat erabiltzen dute trantsizio batetik bestera iragateko, eta egin beharreko trantsizioa ikasketa automatikoko algoritmo batek erabakitzen du ikasketa fasean ikasitakoaren arabera (ikus 1.2 irudia).



1.2 irudia – Trantsizioetan oinarritutako analizatzaile batekin *Jonek jan du* esaldiari egindako analisia. Ezkerraldeko kortxeteek pila datu-egituraren papera betetzen dute eta eskuinaldekoek bufferrarena. Goitik behera letra beltzez adierazitako ekintzaren ondoren eragindako aldaketak adierazten dira.

Grafoetan oinarritutako parserretan, berriz, grafo-egitura bat erabiltzen da esaldia osatzen duten hitzen arteko loturak erabakitzeko. Grafoan, hasieran hitz guztiak (adabegiak) lotzen dira arkuen bidez eta ikasketa algoritmoaren bidez lotura bakoitzari balio bat ematen zaio, azken emaitza baliorik altuena duten hitzen arteko loturek sortuko duten zuhaitza izango delarik (ikus 1.3 irudia).



1.3 irudia – Grafoetan oinarritutako analizatzaile batek *Aitorrek ikusi du* esalditik sortutako zuhaitza.

Gure esperimentuetarako trantsizioetan eta grafoetan oinarritutako parserrak erabili ditugu, dependentzia egituren analisiari zukua ateratzeko ikuspuntu desberdinak jorratzen dituzten analizatzaileak erabiltzea lagungarria dela uste dugulako.

1.3 Lanaren kokapena

IXA taldean hiru hamarkadatan zehar aritu dira Hizkuntzaren Prozesamenduaren (HP) alor desberdinetan lanean. Urte horietan guztietan informatikarien eta hizkuntzalarien elkarlanari esker hainbat tesi irakurri dituzte, baita hainbat produktu garatu ere. Taldearen helburu nagusia euskararen azterketan eta prozesamenduan lagundu dezaketen gaietan sakontzea da, ikerkuntzan nahiz bizitza errealean lagungarriak izan daitezkeen tresnak ere garatuz. Egindako lan guztia lerro hauetan biltzea ezinezkoa den arren, uste dugu merezi duela ekarpen garrantzitsuenak aipatzeak.

Artola (1993) eta Arregi (1995) tesi-lanetan ikus daitekeen bezala, taldean aspalditik landu izan da lexikografia. Lan horri esker sortu zen IXA taldeko baliabideen artean garrantzitsuenetakoa den euskararen datu-base lexikala (EDBL) (Aldezabal *et al.*, 2001). Euskara bezalako hizkuntza aberats bat landu nahi bada, ezinbestekoa da morfologiaren azterketa egitea. Alegria (1995) tesi-lanean euskararen morfologiari buruzko deskribapena eta inplementazioa egin zen. Maritxalar (1999) tesian, berriz, bigarren hizkuntza baten irakaskuntzarako sistema baten garapena egiten da alde morfologikoa landuz. Morfologian egin diren lanei hertsiki lotuta dagoen tresna da MORFEUS analizatzaile morfosintaktikoa (Aduriz *et al.*, 1998), IXA taldeko

analisi katearen ardatza delarik.

Azken urteetan semantikarekin erlazionatutako hainbat tesi irakurri dira (Agirre, 1999; Lopez de Lacalle, 2009; Aldabe, 2011; Otegi, 2012) taldean, baita pragmatikarekin erlazionatutako bat ere (Iruskieta, 2014). Sintaxiaren eremu zabala dela eta, sintaxiarekin erlazionatutako tesi ugari irakurri dira. Landu diren bideen artean, sintaxi partziala euskararen murriztapen-gramatika (ingeleseko *Constraint Grammar*) formalismoa erabiliz (Aduriz, 2000; Arriola, 2000), euskararen sintaxia lantzeko oinarrizko baliabideen garapena (Gojenola, 2000a; Ezeiza, 2002), aditzaren azpi-kategorizazioaren azterketa (Aldezabal, 2004), murriztapen gramatika formalismoa jarraituz garatutako dependentzien analizatzaile sintaktiko osoa (Aranzabe, 2008), euskararako *chunker* estatistiko baten garapena (Arrieta, 2010a) eta estaldura zabaleko euskararako analizatzaile sintaktiko estatistiko baten planteamendua (Bengoetxea, 2014). Lan horietan aztertutakoarekin orain arte euskararako landu gabe zeuden ikerlerro berriak jorratzeko aukera zabaldu da. Ikerlerro horietako bat erregeletan oinarritutako analizatzaile sintaktikoen eta estatistiketan oinarritutakoen arteko hibridazioa edo konbinaketa da eta ikerlerro hori tesi-lan honetako ardatzetako bat da.

Aipatu diren sintaxiarekin erlazionatutako lan guztiek euskararekin lan egiten dutela dute ezaugarri komuna. Egia da Bengoetxea (2014) tesian momenturen batean ingelesa ere lantzen dela, baina orokorrean lanaren helburua euskararako estaldura zabaleko analizatzaile sintaktikoa sortzea da. Eleaniztasun falta hori kontuan hartuta, tesi-lan hau zulo hori betetzera dator, ahal den neurrian, egingo diren esperimentuak hizkuntza desberdinetan probatuko dira, planteatutako hobekuntzak hizkuntzekiko independenteak diren ala ez zehazteko asmoarekin.

Tesi-lan hau 2012an hasi zen fruituak ematen, zehazki SPMRL 2012 (*Statistical Parsing of Morphologically Rich Languages*) *workshop*ean. Workshop horretan, morfologikoki aberatsak diren hizkuntzetan sintaxi konputazionalaren inguruan egindako lanak aurkezten dira. Urte horretatik aurrera workshop horretara ia urtero aurkeztu garen arren, tesi honetan berebiziko garrantzia dute 2013 eta 2014 urteko edizioek. Izan ere, urte horietan, bakoitzak garatutako sistema hizkuntza desberdinetan probatzea ahalbidetzen zuen ataza-partekatua antolatu baitzuten. Gure helburu nagusia, sortutako ezaugarri eta teknika berriekin dependentzia sintaktikoen ikuspegi eleaniztuna lantzea izanik, ataza partekatu horietan landutakoa tesiaren ardatz bihurtu da. Alde batetik, zuhaitz-bankuetan etiketatutako kategoria, azpikategoria eta ezaugarri morfologikoei etekina ateratzen saiatu gara ezaugarrien

ingeniaritza erabiliz, hau da, zuhaitz-bankuetan eskuragarri dagoen informazioaz baliatu gara, baina ohiko ordena eta erabilitako informazioa aldatuz. Beste aldetik, ez-gainbegiratutako multzokatze (*clustering*) tekniken bidez erauzitako ezaugarriek analisi sintaktikoan duten eragina neurtu da.

Ez dugu ahaztu nahi tesiak euskararekin duen lotura, egindako ikerlan osoan euskara izan baita behin eta berriro erabili dugun hizkuntza bakarra. IXA taldean aztertzen den hizkuntza nagusia euskara izanda, euskararekin erlazionatutako zenbait baliabide garatu ditugu urte hauetan, aipagarriena Dependentzia Unibertsalak[1] (*Universal Dependencies*) proiekturako sortu dugun euskararako zuhaitz-bankua. Aurrerago azalduko dugun bezala, nazioarteko proiektu horretan, hainbat hizkuntzatako zuhaitz-bankuak bihurtu dira etiketatze estandarrera. Modu horretara, gidalerro berdinak erabilita etiketatutako kategoriak, ezaugarriak eta dependentziak biltzen dituzten zuhaitz-bankuak sortu dira partu hartu duten hizkuntza guztietarako.

1.4 Helburuak

Aurretik aipatu den bezala, tesi-lan honen helburu nagusia dependentzia sintaktikoen ikuspegi eleaniztuna lantzea da eta helburu hori betetzeko bidean ezaugarri berriak eta teknika berriak probatuko ditugu. Ahal izan den neurrian, esperimentu guztiak hizkuntza desberdinetan probatzen saiatuko gara, baina hori ezinezkoa gertatzen zaigun kasuetan euskaran zentratuko gara, hizkuntza hori baita gure lehentasuna. Hori dela eta, euskararako baliagarriak izan daitezkeen baliabideak sortzea ere gure helburuen artean sartu dugu.

Deskribatu diren helburuak betetzeko hainbat eginkizun eramango dira aurrera eta tesiaren ondorioen atalean erantzungo diren eginkizun horietako bakoitzarekin erlazionatuta dauden zenbait ikerkuntza galdera planteatu dira:

Izaera desberdinetako analizatzaileen hibridazioarekin lortu daitezkeen emaitzak neurtu (*4. kapitulua*)

- Erregeletan oinarritutako analizatzaileak analizatzaile estatistikoekin konbinatzen direnean sintaxi osoa lantzen duten erregeletan oinarritutako analizatzaileen ekarpena azaleko sintaxia lantzen dutenena baino handiagoa izango al da?

[1] `http://universaldependencies.org/`

– Azaleko sintaxia lantzen duten analizatzaileen artean erregeletan oinarritutakoen ekarpena estatistiketan oinarritutakoena baino handiagoa izango al da?

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarri morfologiko bakoitzak analisi sintaktikoan duen eragina neurtu (*5. kapitulua*)

– Morfologia aberatsagoa duten hizkuntzetan, neurtutako ezaugarri morfologikoek sintaxian duten ekarpena handiagoa izango al da?

– Ondorioztatu al daiteke ezaugarri mota batzuek analisi sintaktikoan duten pisua beste batzuena baino handiagoa dela orokorrean?

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarrien ingeniaritza aplikatzeak analisi sintaktikoan duen eragina neurtu (*5. kapitulua*)

– Analisi sintaktikoan pisu gehien duten ezaugarriak erabiltzean, ezaugarri guztiak erabiltzean lortzen direnekin konparagarriak diren emaitzak erdietsiko al dira?

– Analizatzaileari ezaugarri morfologikoak pasatzeko moduak eraginik al du analisi sintaktikoan?

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarrien ingeniaritza aplikatzean erdietsitako analisi desberdinen konbinazioak duen eragina neurtu (*5. kapitulua*)

– Ondorioztatu al daiteke zenbat eta morfologia aberatsagoa izan hizkuntza batek orduan eta emaitza hobeak lortzen dituela ezaugarrien ingeniaritza aplikatzean erdietsitako analisi desberdinen konbinazioaren bidez?

Morfologikoki aberatsak diren zenbait hizkuntzatan multzokatze mota (*clustering*) desberdinek analisi sintaktikoan duten eragina neurtu (*5. kapitulua*)

– Ondorioztatu al daiteke multzokatze mota bat bestea baino lagungarriagoa dela analisi sintaktikorako?

Morfologikoki aberatsak diren zenbait hizkuntzatan multzokatze mota desberdinak aplikatzean erdietsitako analisien konbinazioak duen eragina neurtu (*5. kapitulua*)

- Ondorioztatu al daiteke zenbat eta morfologia aberatsagoa izan hizkuntza batek orduan eta emaitza hobeak lortzen dituela multzokatze mota desberdinak aplikatzean erdietsitako analisi desberdinen konbinazioaren bidez?

Erdi-gainbegiratutako zuhaitz-bankuetatik[2] erauzitako ezaugarri mota desberdinek morfologikoki aberatsak diren zenbait hizkuntzatan duten eragina neurtu (*5. kapitulua*)

- Ondorioztatu al daiteke erdi-gainbegiratutako zuhaitz-bankuetatik erauzitako ezaugarri mota batzuk besteak baino hobeak direla?
- Ondorioztatu al daiteke erdi-gainbegiratutako zuhaitz-bankuetatik erauzitako ezaugarriak erabiltzen direnean zenbat eta morfologia aberatsagoa izan hizkuntza batek orduan eta emaitza hobeak lortzen direla?

Ezaugarrien ingeniaritzarekin, multzokatze mota desberdinekin eta erdi-gainbegiratutako zuhaitz-bankuetatik erauzitako ezaugarriekin erdietsitako analisien konbinazioak morfologikoki aberatsak diren zenbait hizkuntzatan duen eragina neurtu (*5. kapitulua*)

- Konbinazioan erabilitako iturri desberdinetatik eratorritako analisiak osagarriak izango al dira?

Euskarazko zuhaitz-bankua bihurtu Dependentzia Unibertsalak proiektuan zehazten diren gidalerroak jarraituz (*6. kapitulua*)

- Zehaztutako gidalerroen arabera euskarazko zuhaitz-banku osoa bihurtu ahal izango da?
- Bihurtutako zuhaitz-bankuarekin oinarrizkoarekin erdietsitakoak baino emaitza hobeak lortuko al dira?

Behin tesi-lan honetako helburu, eginkizun nagusi eta horiei lotutako ikerketa galderak planteatu ditugula, hurrengo kapituluetan zehar galdera horiei erantzunak emango dizkieten emaitzak ikusten joango gara, ondorioen atalean banan-banan erantzungo ditugularik.

[2]Erdi-gainbegiratutako teknikak erabiliz sortutako zuhaitz-bankuak.

1.5 Tesi-lanaren eskema eta argitalpenak

Tesi-lan hau 4 atal nagusitan banatu da, eta tesia osatzen duten 7 kapituluak 4 atal horietan bildu dira. Banaketa hori azaltzeko 1.1 taula erabiliko dugu; taularen ezkerraldean aipatutako atalak zehaztuko dira eta eskuinaldean kapituluak. Ondorengo lerroetan, kapituluak modu horretara sailkatzearen arrazoiak eta kapituluetan azalduko diren gaiak deskribatuko dira labur-labur.

Atalak	**Kapituluak**
Motibazioa eta gaiaren kokapena	1. Tesi-lanaren aurkezpen orokorra
	2. Lanaren kokapena
Esperimentazio-ingurunea	3. Esperimentazio-ingurunea
Teknika elebakarrak	4. Hibridazioa
Teknika eleaniztunak	5. Analisi sintaktiko eleaniztuna
Euskarazko baliabideen sorkuntza	6. Dependentzia Unibertsalak
Ondorioak eta zabaldutako ikerlerroak	7. Ondorioak eta etorkizuneko lanak

1.1 taula – Tesi-lanaren antolaketa.

Lehen atala esku artean dugun kapituluak eta *Lanaren kokapena* deituriko kapituluak osatzen dute. Bertan, tesiaren helburu nagusiak deskribatzen dira, ikerkuntza galderak planteatzen dira, tesia IXA taldean egindako lanen artean eta artearen egoeran egindako lanen artean kokatzen da, eta hurrengo kapituluetan azalduko dena ulertzeko beharrezkoak diren kontzeptuak zehazten dira.

Bigarren atalean, *Esperimentazio-ingurunea* deituriko atalean, izen bera duen kapitulua dago. Atal honetan, tesian egin diren esperimentuak aurrera eramateko beharrezkoak diren tresnak eta baliabideak zehaztuko dira.

Hirugarren atalean, *Teknika elebakarrak* deiturikoan, *Hibridazioa* deituriko kapitulua bildu da. Kapitulu horretan, izaera desberdineko analizatzaileak konbinatzeak analisi sintaktikoan duen eragina neurtuko da. Horretarako, erregeletan oinarritutako analizatzaileen eta azaleko sintaxia lantzen duten analizatzaileen irteera erabiliko dute sarrera gisa analizatzaile estatistiko desberdinek.

Laugarren atalean, *Teknika eleaniztunak* deiturikoan, *Sistema eleaniztunak* deituriko kapitulua bildu da. Kapitulu horretan, jarraian aipatzen diren hiru ikerlerroek morfologikoki aberatsak diren bost hizkuntzatan duten eragina neurtuko da. Alde batetik, ezaugarrien ingeniaritza erabiliz analisi

sintaktikoa hobetzeko tartea dagoen ala ez aztertuko da; beste aldetik, multzokatzearen bidez sortutako ezaugarri berriek analisi sintaktikoan duten eragina neurtuko da; eta bukatzeko, erdi-gainbegiratutako zuhaitz-bankuetatik eratorritako ezaugarri berriek analisi sintaktikoan duten eragina zehaztuko da.

Bosgarren atala, *Euskarazko baliabideen sorkuntza* deiturikoa, *Dependentzia Unibertsalak* deituriko kapituluak osatzen du. Kapitulu honetan, euskara Dependentzia Unibertsalak (*Universal Dependencies*) proiektuaren kide izateko sortutako euskarazko zuhaitz-bankuaren bihurketa azalduko da.

Tesi-lanari amaiera emateko, urte hauetan ateratako zenbait ondorio eta ikerkuntzarako zabaldu diren bideen berri emango da zazpigarren kapituluan. Kapitulu hori erabiliko da esku artean dugun sarrerako kapitulu honetan planteatu diren ikerkuntza galderak erantzuteko ere.

Argitalpenak

Tesi-lan honetan ikertutako gaiek hainbat artikulu argitaratzeko bidea eman digute. Ondorengo lerroetan zehaztuko da argitalpen horien zerrenda, kapituluen arabera sailkatuta.[3]

- 4. kapitulua – Hibridazioa:
  - Aranzabe Maria Jesus, Bengoetxea Kepa, Díaz de Ilarraza Arantza, Ezeiza Nerea, Goenaga Iakes, Gojenola Koldo **Combining Rule-Based and Statistical Syntactic Analyzers** *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012).* 48-54, Association for Computational Linguistics (ACL), USA, ISBN: 978-1-937284-30-5, July 12, 2012, Jeju Island, Republic of Korea
- 5. kapitulua – Analisi sintaktiko eleaniztuna:
  - Goenaga Iakes, Gojenola Koldo, Ezeiza Nerea **Exploiting the Contribution of Morphological Information to Parsing: the BASQUE TEAM system in the SPRML 2013 Shared**

[3] Argitalpen hauen egileen ordena alfabetikoa da, 5. kapituluko lehenengo biei dagokienean izan ezik.

Task *Workshop on Statistical Parsing of Morphologically Rich Languages.*Pages 71-77. SPRML 2013 Shared Task, Seattle, EMNLP Workshop. ISBN 978-1-937284-97-8

- Goenaga Iakes, Gojenola Koldo, Ezeiza Nerea **Combining Clustering Approaches for Semi-Supervised Parsing: the BASQUE TEAM system in the SPRML 2014 Shared Task** *Workshop on Statistical Parsing of Morphologically Rich Languages* SPRML 2014 Shared Task, Dublin, COLING Workshop.
- Atutxa Aitziber, Ezeiza Nerea, Goenaga Iakes, Gojenola Koldo **Experiments on Semi-supervised Dependency Parsing of a Morphologically Rich Language** *6th Workshop on Statistical Parsing of Morphologically Rich Languages* (SPMRL 2015), Bilbao, IWPT Workshop.

- 6. kapitulua − Dependentzia Unibertsalak:
  - Aranzabe Maria Jesus, Atutxa Aitziber, Bengoetxea Kepa, Díaz de Ilarraza Arantza, Goenaga Iakes, Gojenola Koldo, Uria Larraitz. **Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies** *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)*, 2015. 233-241. Institute of Computer Science of the Polish Academy of Sciences, Warszawa, Poland. ISBN: 978-83-63159-18-4

- Beste hauek ere, kapitulu zehatz batekin erlaziorik ez badute ere, tesiarekin erlazionatuta daude:
  - Arriola Jose Mari, Aranzabe Maria Jesus, Goenaga Iakes **Reutilizacion del Treebank de Dependencias del Euskera para la Construccion del Gold Standard de la Sintaxis Superficial** *Procesamiento del Lenguaje Natural 2013*, 51, 83-90. (ISSN edicion impresa: 1135-5948) (ISSN edicion digital: 1989-7553)
  - Otegi Arantxa, Ezeiza Nerea, Goenaga Iakes, Labaka Gorka **A Modular Chain of NLP Tools for Basque** *Proceedings of the 19th International Conference on Text, Speech and Dialogue - TSD 2016*, Brno, Czech Republic, volume 9924 of Lecture Notes in Artificial Intelligence, pp. 93-100. ISBN 978-3-319-45509-9

Sortutako baliabideak

Tesi-lan honetan zehaztu diren helburuak betetzeko hainbat esperimentu eraman ditugu aurrera, eta esperimentu horiek jorratzeko baliabide desberdinak erabili ditugu. Ondorengo lerroetan guk sortutako baliabide garrantzitsuenak bildu ditugu, baliabide horiek erabili diren kapituluekin batera.

- 5. kapitulua – Analisi sintaktiko eleaniztuna:
  - Sintaktikoki analizatutako euskarazko 150 milioi hitzeko zuhaitz-bankua
- 6. kapitulua – Dependentzia Unibertsalak:
  - Euskarazko UD zuhaitz-bankua (1.1 bertsioa) (Aranzabe *et al.*, 2015)

2

Lanaren kokapena

Kapitulu honetan, tesi-lanarekin erlazionatuta dauden lan esanguratsuak aztertuko dira. Hala ere, lan horiekin hasi baino lehen, sarrerako kapituluan gainetik azaldu dugun sintaxi konputazionalaren atal oso garrantzitsua den dependentzien analisia aurrera eramateko erabiltzen diren bi ikuspuntu nagusiak azalduko dira sakonki: trantsizioetan oinarritutako dependentzien analisia eta grafoetan oinarritutako dependentzien analisia. Bi ikuspuntu horiek azaltzea ezinbestekoa dela uste dugu tesi-lan honen ardatz diren parserrak bi multzo horietan banatzen direlako, trantsizioetan oinarritutako parserrak eta grafoetan oinarritutako parserrak, hain zuzen ere.

Dependentzien analisian nagusi diren bi ikuspuntuak azaldu ondoren, tesi-lanean nolabaiteko garrantzia duten lanak aipatuko dira. Azalpen hori bi zatitan banatu dugu. Alde batetik, ikuspegi elebakarra lantzeko erabili ditugun teknikekin lotura duten lanen azalpena; bestetik, ikuspegi eleaniztuna lantzeko erabili ditugun teknikekin erlazioa duten ataza partekatu eta lanen deskribapena.

2.1 Dependentzien analisia

Dependentzien analisia aurrera eramateko bide bat baino gehiago daude. Erregeletan oinarritutako sistemak izan ziren dependentzien analisi automatikoa burutzen lehenengotarikoak. Sistema horiek, aurretik definitutako gramatiketan oinarritzen dira eta aztertzen duten hizkuntzari oso lotuta daude. Izan ere, erregeletan oinarritutako sistema horiekin beste hizkuntza bat az-

tertu nahi izango balitz, erregela berriak idatzi beharko bailirateke.

Dena den, azken urteetan estatistiketan oinarritutako sistemen erabilpena hazten joan da, neurri batean erregeletan oinarritutako sistemak atzean utziz. Gertaera horren arrazoiak bat baino gehiago dira. Ikasketarako corpus bat dugun bitartean, ikasketa automatikoa erabiltzen duten sistemekin hizkuntza bakoitzerako analizatzaile bat izan dezakegu. Egia da ere, erregeletan oinarritutako sistemek baino emaitza hobeak lortzen dituztela orokorrean eta behin sistema inplementatuta dagoela oso mantentze lan txikia eskatzen dutela. Estatistiketan edo ikasketa automatikoan oinarritutako dependentzia analizatzaileak bi multzo nagusitan banatzen dira: trantsizioetan oinarritutakoak eta grafoetan oinarritutakoak. Ondorengo azpi-puntuetan, aipatutako bi hurbilpen horiek azalduko dira esaldi baten analisiaren adibide batekin batera.

2.1.1 Trantsizioetan oinarritutako dependentzien analisia

Mota honetako sistemei (Nivre *et al.*, 2007b; Attardi *et al.*, 2007; Duan *et al.*, 2007) trantsizioetan oinarritutako sistemak (ingelesezko *transition-based*) deitzearen arrazoia beraien izaera da. Esaldi baten dependentzien analisia, trantsizio sistema abstraktu edo egoera makinen bidez bide optimoa aurkitzera mugatzen baita. Sistema hauek, hurrengo trantsizioa zein izango den iragartzen dute. Horretarako, aurretik egindako urratsak, hitzen ezaugarriak eta sarrerako esaldia hartzen dituzte kontuan. Labur esateko, dependentzia analizatzailea hasierako egoera batetik hasten da eta sortutako ereduaren iragarpenak aintzat hartuta teknika edo algoritmo jaleak erabilita beste egoeretara pasatzen da bukaerako egoerara iritsi arte. Mota honetako sistemek betetzen dituzten oinarriak ondorengoak dira:

- Gehienetan bi datu-egitura erabiltzen dira: pila bat eta buffer bat.
- Analizatzaileak esaldia ezkerretik eskuinera aztertzen du modu deterministan.
- Ikasketa automatikoan oinarritutako sailkatzailea da hurrengo urratsa erabakitzen duena.
- Sistemak eman behar duen hurrengo urratsa iragartzeko, ikasketa automatikoko algoritmoak ondorengoen artean aukeratu behar du: ezker

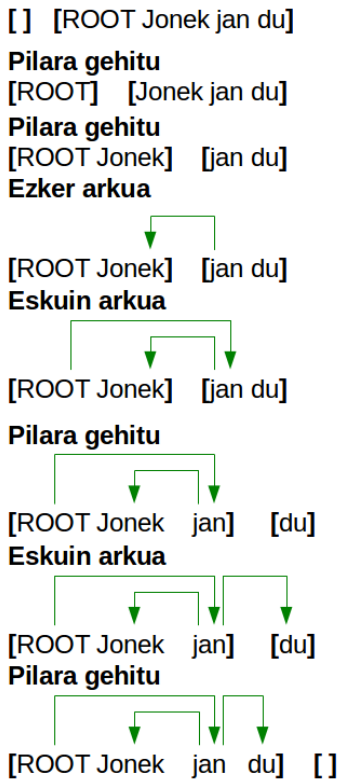
arkua, eskuin arkua, pilara gehitu edo trukatu.

– Ikasketa automatikoko sailkatzaileak hitzen ezaugarriak (forma, lema, kategoria, azpikategoria) erabiltzen ditu urratsik hoberena zein izango den erabakitzeko.

Mota honetako sistemen funtzionamendua hobeto azaltzeko, 2.1 taulan trantsizioetan oinarritutako sistema batek esaldi sinple bat aztertzeko ematen dituen urratsak erakusten dira. 2.1 irudian 2.1 taulan ematen diren urratsek zuhaitzaren egituran duten eragina irudikatu da. Taularen lehenengo lerroan pilaren eta bufferraren hasierako egoerak bildu dira. Bufferrean, aztertu beharreko esaldia eta esaldiaren erroa definitzeko erabiltzen den elementua (ROOT) daude. Pila hutsik dago oraindik. Hurrengo lerroetan trantsizioetan oinarritutako analizatzaileak bufferreko esaldia analizatzeko burututako ekintza bakoitzaren eragina ikusten da. Pilara gehitu (*shift*), Ezker arkua (*left-arc*) eta Eskuin arkua (*right-arc*) ekintzen bidez bukaerako lerroan *Jonek jan du* esaldiaren zuhaitz sintaktikoa eraiki da.

Pila	**Bufferra**	**Ekintza**
-	ROOT Jonek jan du	Pilara gehitu
ROOT	Jonek jan du	Pilara gehitu
ROOT Jonek	jan du	Ezker arkua
ROOT Jonek	jan du	Eskuin arkua
ROOT Jonek	jan du	Pilara gehitu
ROOT Jonek jan	du	Eskuin arkua
ROOT Jonek jan	du	Pilara gehitu
ROOT Jonek jan du	-	-

2.1 taula – Trantsizioetan oinarritutako analizatzaile batekin *Jonek jan du* esaldiari egindako analisiaren taula.



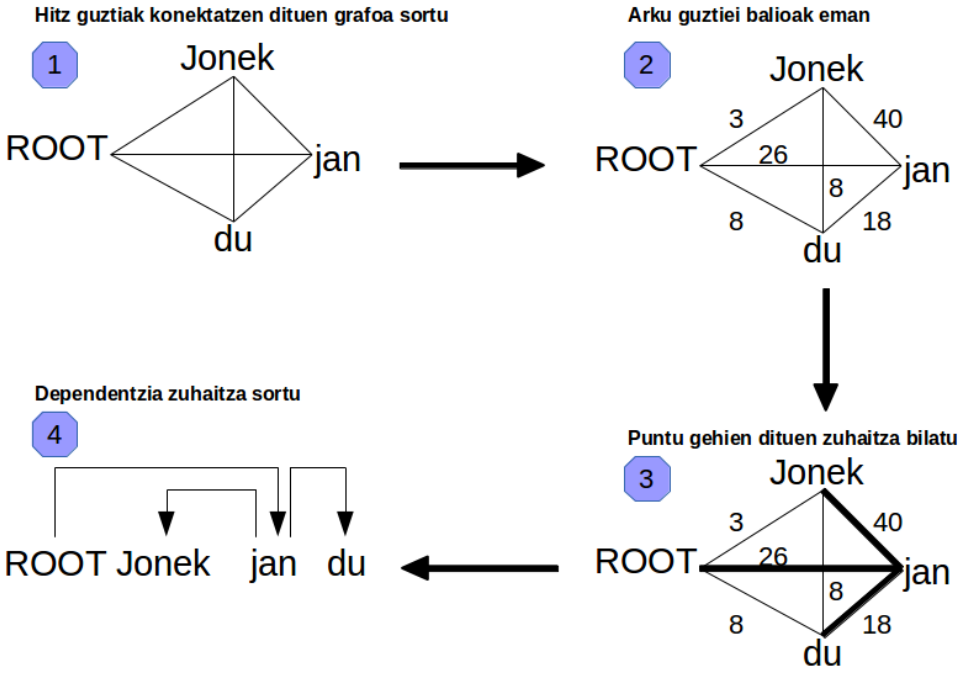
2.1 irudia – Trantsizioetan oinarritutako analizatzaile batekin *Jonek jan du* esaldiari egindako analisia. Ezkerraldeko datu-egitura pila da eta eskuinaldekoa bufferra.

Trantsizioetan oinarritutako sistemen abantailetako bat, esaldi proiektiboen analisia (dependentzia zuhaitzaren arkuak gurutzatzen ez direnean) denbora linealean egiteko gai direla da, kasu gehienetan ez-proiektiboak (dependentzia zuhaitzaren arkuak gurutzatzen direnean) denbora koadratikoan egiteko gai diren bitartean. Esandakoa betetzearen arrazoi nagusia algoritmo jaleak behin erabakia hartzen duenean ez duela erabaki hori berrikusiko da. Bilaketa espazioa oso txikia izaten da eta trantsizio sistemak hurrengo trantsizioa zein izango den aukeratu behar duenean aukerak normalean lau baino gutxiago izaten dira.

2.1.2 Grafoetan oinarritutako dependentzien analisia

Trantsizioetan oinarritutako sistemen eta grafoetan oinarritutako (ingelesezko *graph-based*) arteko desberdintasuna ataza bera ebazteko jarraitzen duten bidea da. Grafoetan oinarritutako dependentzia analizatzaileek (McDonald *et al.*, 2005; Spranger, 2007; Nguyen *et al.*, 2007), aztertu nahi den esaldiarekin sortutako grafoaren hitzen arteko arku guztiak aztertzen dituzte eta sailkatzaile baten bidez aukeratzen dituzte probabilitate altuena duten arkuak. Gauzak horrela, grafo horretatik erauzten dira esaldia osatzen duten dependentzia egiturak.

Trantsizioen sistemarekin egin den bezala, grafoetan oinarritutako dependentzien analizatzaile batek esaldi sinple bat aztertzeko jarraitzen dituen urratsak bildu dira 2.2 irudian.



2.2 irudia – Grafoetan oinarritutako analizatzaile batekin *Jonek jan du* esaldiari egindako analisia.

Grafoetan oinarritutako dependentzia analizatzaileek orokorrean denbora gehiago behar izaten dute emaitzak bueltatzeko, baita memoria gehiago ere. Esandakoa ulertzeko 2.2 irudia aztertuko dugu. Irudiko esaldia hiru hitzez

soilik osatuta dagoen arren, sei arku posible sortu dira grafoan. Egoera horren aurrean, esaldi luzeetan sortzen diren arku posibleak asko dira eta bilaketa espazioa handia izaten da. Kasurik okerrenean exekuzio denbora kubikoa izaten da eta kasurik hoberenean koadratikoa.

2.2 Ikuspegi elebakarra lantzeko teknikak

Puntu honetan, ikuspegi elebakarraren aldetik landu diren teknikekin erlazionatutako lanak azalduko dira. Tesi-lan honen helburuetako bat ahal den guztietan ikuspegi eleaniztuna lantzea izanik, ikuspegi elebakarrarekin erlazionatutako lanetan azpi-puntu bakarra sortu dugu, sistemen hibridazioa guk nahi bezala lantzeko oso zaila gertatu baitzaigu euskararako eskuragarri ditugun baliabideekin konparagarriak diren baliabideak aurkitzea.

2.2.1 Sistemen hibridazioa

4. kapituluan, analizatzaile sintaktiko desberdinak konbinatzeak zabaltzen dituen bide desberdinak landuko dira. Hori dela eta, hibridazioarekin erlazionatuta dauden eta gure tesi-lanean eragina izan duten antzeko lanak aztertzea beharrezkoa dela uste dugu.

Azken urteetan, asko izan dira analizatzaile sintaktiko desberdinak konbinatu dituztenak (Surdeanu eta Manning 2010), horiek bakarka lortzen dituzten emaitzak hobetuz. Erdietsitako emaitzen arabera, konbinaketa edo hibridazioa aurrera eramateko hurbilpen arrakastatsuenak pilaketa (*stacking*) (Martins *et al.* 2008) eta botaketa (*voting*) (Sagae eta Lavie 2006; Nivre eta McDonald 2008; McDonald eta Nivre 2011) izan dira.

Bestalde, morfologikoki aberatsak diren hizkuntzak aztertzea betidanik izan da erronka bat analizatzaile sintaktiko estatistikoetarako, hizkuntza horietarako oso zaila gertatzen delarik ingelesa bezalako hizkuntzetan erdiesten den eraginkortasunaren parekorik lortzea. Hala ere, kategoria gramatikalaren etiketatzean (*POS tagging*) egin den modura, erregeletan oinarritutako sistemak eta estatistiketan oinarritutako sistemak konbinatuz (Ezeiza *et al.* 1998; Tapanainen eta Voutilainen 1994), uste dugu hibridazioak bide luzea duela aurretik morfologikoki aberatsak diren hizkuntzen analisi sintaktikoan, eta oso interesgarria dela ikerketa-lerro horrek eskaintzen dituen aukerak ikuspuntu desberdinetatik lantzea.

2.3 Ikuspegi eleaniztuna lantzeko teknikak

Ikuspegi eleaniztuna lantzeko teknikekin erlazionatutako lanen azalpena era honetara emango da: hasteko, tesi-lan honetan berebiziko garrantzia duten SPMRL 2013 eta SPMRL 2014 ataza partekatuak azalduko dira, ataza partekatu horietan landutakoa baita *Sistema Eleaniztunak* kapituluan (5. kapitulua) azalduko denaren zatirik handiena; ondoren, kapitulu horretan azalduko diren teknikekin erlazionatutako gainerako lanak aztertuko dira.

2.3.1 SPMRL 2013 eta SPMRL 2014 ataza partekatuak

2013an, *Statistical Parsing of Morphologically Rich Languages* kongresuan (Seddah *et al.*, 2013) morfologikoki aberatsak diren hizkuntzetan zentratu zen ataza partekatua antolatu zuten. Bertan, dependentzietan oinarritutako edo osagaietan oinarritutako zuhaitz-bankuetan, hizkuntza desberdinen analisi sintaktikoan emaitzarik hoberenak erdiesteko hainbat partaide lehiatu ziren. Partaideen esku utzi zen bakoitzak aztertu nahi zituen hizkuntzak eta zuhaitz-bankuen egitura mota aukeratzea (dependentzietan oinarritutakoa edo osagaietan oinarritutakoa).

Hurrengo urtean, berriro antolatu zen kongresua (Seddah *et al.*, 2014) eta urte horretako helburuetako bat analisi sintaktikoan erdi-gainbegiratutako iturrietatik erauzitako ezagutza erabiltzea izan zen. Hori dela eta, hizkuntza bakoitzerako oinarrizko zuhaitz-bankuez gain, dependentziak automatikoki etiketatuko zuhaitz-banku erraldoiak utzi ziren eskuragarri.

SPMRL 2013an, dependentzietan oinarritutako 9 corpus desberdin zeuden eskuragarri eta beste horrenbeste osagaietan oinarritutakoak. SPMRL 2014an ere corpus kopuru bera zegoen eskuragarri mota bakoitzerako. Seddah *et al.* (2013) eta Seddah *et al.* (2014) lanetan deskribatzen den bezala, SPMRL 2013 eta SPMRL 2014 ataza partekatuetako oinarrizko corpusak ondorengoak dira:

- **Arabiera:** Arabierarako, LDC Penn Arabic Treebanks (PATB) (Maamouri *et al.*, 2004) zuhaitz-bankutik eratorritako bi corpus utzi ziren eskuragarri. Dependentzietan oinarritutako Columbia Arabic Treebank-a (CATiB) (Habash eta Roth, 2009) eta PATB zuhaitz-bankuaren osagaietan oinarritutako bertsioa den Stanford Arabic Phrase Structure Treebank-a (Green eta Manning, 2010).

- **Euskara:** Euskararako, EPEC-DEP zuhaitz-bankuaren (Aduriz *et al.*,

2003) bigarren bertsioaren (Aranzabe, 2008; Aldezabal *et al.*, 2009) zati bat utzi zen eskuragarri dependentzien analisi automatikoa lantzeko. Osagaien analisirako, berriz, aurretik aipatutako dependentzien zuhaitz-bankua bihurtu zen osagaien formatura.

- **Frantsesa:** Frantseserako eskuragarri zeuden bi corpusak French Treebank (Abeillé *et al.*, 2003) zuhaitz-bankutik eratorriak izan ziren. Lehenengo, osagaietan oinarrituko bertsioa sortu zen automatikoki eta ondoren, azken horretan oinarrituta, dependentzietan oinarritutako bertsioa sortu zen (hau ere automatikoki).

- **Alemana:** Alemanerako corpusak TiGer zuhaitz-bankuaren 2.2 bertsioaren (Brants *et al.*, 2002) aldaerak dira. Osagaietan oinarritutako bertsioan jatorrizko bertsioarekiko aldaketa txikiak egiten diren arren, dependentzietan oinarritutako bertsioa lortzeko bihurketa automatikoa egin behar izan da (Seeker eta Kuhn, 2012).

- **Hebreera:** Hebreerako corpusak Hebrew Treebank V2 corpusaren (Sima'an *et al.*, 2001; Guthmann *et al.*, 2008) aldaerak dira. Bat dependentzietan oinarritutakoa, eta bestea osagaietan oinarritutakoa.

- **Hungariera:** Hungarierarako zeuden bi corpusak, Szeged Treebankean oinarritutako osagaietarako bertsio bat (Csendes *et al.*, 2004) eta dependentzietarako bertsio bat (Vincze *et al.*, 2010) dira.

- **Koreera:** Koreerarako osagaietarako corpusa, KAIST Treebank-etik (Choi *et al.*, 1994) osagaietan oinarritutako zuhaitzak jasota sortu da. Dependentzietan oinarritutako bertsioa sortzeko, osagaietan oinarritutako bertsioa bihurtu da.

- **Poloniera:** Polonierazko osagaietan oinarritutako zuhaitz-bankua *Składnica* deiturikoa da (Świdziński eta Woliński, 2010; Wolinski *et al.*, 2011). Dependentzietan oinarritutako bertsioa (Wróblewska, 2012), eskuz desanbiguatutako osagaietako zuhaitzen bihurketa automatikoaren bidez sortu zen.

- **Suediera:** Suedierarako osagaietan oinarritutako zuhaitz-bankua Swedish Treebank-aren (Nivre eta Megyesi, 2007) Talbanken ataletik (Nivre *et al.*, 2006) erauzita dago. Talbanken ataleko etiketatze sintaktikoa

osagaien formatura bihurtu zen eta, ondoren, azken hori bihurtu zen dependentzietan oinarritutako formatura.

Azaldu diren corpus horiek guztiak ez dira erabili tesi-lan honetan, erabili direnei buruzko informazio gehigarria 3. kapituluan emango da.

2.3.2 Ezaugarrien ingeniaritza

Ezaugarrien ingeniaritzaren bidez analisi sintaktikoa hobetzeko zenbait bide landuko dira. Horrenbestez, aurretik bide horretan egin diren lanak eta gure lanarekin lotuta dauden hurbilpenak aztertzea beharrezkoa dela deritzogu.

Analisi sintaktikoa egiterakoan analizatzaile sintaktikoari pasatzen zaizkion ezaugarri bakunek azken emaitzan duten ekarpena zein den jakitea ez da erraza. Izan ere, egileek gehienetan beraien eskura dituzten ezaugarri morfologiko guztiak[1] erabiltzen baitituzte, hauen aniztasunaz baliatzeko helburuarekin emaitza sendoak eskuratzeko.

Badaude ezaugarri morfologikoen benetako eragina zein den neurtu duten egileak ere. Ambati *et al.* (2010) lanean hindi hizkuntzako ezaugarri morfosintaktikoak dependentzien analisi sintaktikoan integratzeko erarik hoberena zein den aztertzen dute. Ezaugarri multzo desberdinekin hainbat proba egin dituzte grafoetan oinarritutako eta trantsizioetan oinarritutako analizatzaile sintaktikoekin. Alde batetik, hitz-formaz gain hitz bakoitzaren kategoria bakarrik erabilita probatzen dute, jasotako emaitza oinarri bezala finkatzeko. Beste alde batetik, oinarrizko konfigurazio horri ezaugarri morfologikoak gehitu dizkiote hauen eragina neurtzeko. Ezaugarri horiek zehazki, hitzaren erroa, azpikategoria, generoa, kasua, numeroa, pertsona eta hitzaren atzizkia dira eta ezaugarri multzo hauek gehituta emaitza hobeak lortzen direla frogatu dute.

Ildo berari jarraituz, Seeker eta Kuhn (2013) lanean, hitzaren kasua bereziki esanguratsua dela ondorioztatu dute txekierarako, alemanerako eta hungarierarako dependentzietan oinarritutako analisi sintaktiko zuzen bat lortu nahi bada. Izan ere, analizatzaile sintaktikoan erabilitako ezaugarri morfologikoen murriztapenak (adibidez, emaitzarik hoberenak lortu dituztenak bakarrik erabiltzea) analizatzaileari bilaketa espazioa txikitzen laguntzen baitio. Era honetara, ezaugarri morfologiko guztiak erabilita lortzen den

[1]Hau da, ezaugarriak zehazteko garaian ezaugarri morfologiko guztiak tratatzen dituzte era berean, eta ikasketa algoritmoari uzten diote horietako bakoitzari ematen zaion pisua erabakitzeko ardura.

emaitza gainditzen da.

Bestalde, Çetinoğlu eta Kuhn (2013) lanean analizatzaile sintaktiko batzuek emaitza hobeak lortzeko joera dutela frogatu da hitzaren kategoriaren ordez hitzaren ezaugarri morfologiko jakin batzuk erabiltzen badira turkierarako. Analizatzaileari normalean kategoria eta azpikategoria pasatzen zaizkion zutabeetan ezaugarri morfologikoak pasatzen dizkiote banan-banan ikusteko ea inolako aldaketarik gertatzen den oinarrizko konfigurazioarekin alderatuta (ezaugarri bakoitza bere zutabean eta eskuragarri dauden ezaugarri guztiak erabilita). Esan bezala, ezaugarriak modu horretara erabilita hobekuntzak lortzen dituzte, erdietsitako emaitzen artean hoberena, hitzaren kasua kategoriaren ordez jarrita lortzen dutelarik. Lan horretako esperimentuetatik lortutako emaitzek iradokitzen dute, analizatzaile sintaktiko batzuek ez dutela modu egokienean erabiltzen eskura duten informazio morfosintaktiko guztia eta analizatzailearen ikasketa algoritmoak pisu handiagoa ematen diola hitzaren kategoria eta azpikategoriaren lekuan pasatzen zaion informazioari. Ondorioz, erabilitako informazioaren arteko desoreka gerta daiteke ezaugarri batzuei estatus altuagoa ematen baitiete analizatzaileek.

2.3.3 Multzokatzea

Multzokatze teknika desberdinek dependentzien analisi sintaktikoan duten eragina ere neurtu nahi da. Esperimentu horiek aurrera eramateko bi lan hartu ditugu nagusiki oinarritzat. Koo *et al.* (2008) lanean 3. kapituluan sakonduko dugun Brown algoritmoarekin lortutako multzoak erabiltzen dituzte dependentzietan oinarritutako analisi sintaktikoa hobetzeko. Brown multzoak, analizatu behar den hitz bakoitzaren informazioaren (hitz-forma, lema, kategoria, azpikategoria...) gehigarri bezala erabiltzen dituzte txekierarako eta ingeleserako eta bi hizkuntzetarako lortzen dituzte hobekuntzak.

Ingelesa eta txekiera hain hizkuntza desberdinak izanik eta txekiera morfologikoki aberatsa kontsideratzen den hizkuntza izanik, morfologikoki aberatsak diren hizkuntzetan, multzokatze teknika desberdinetatik erauzitako multzoak informazio gehigarri bezala erabiltzea erabaki da, emaitza konparagarriak lortzeko asmoarekin.

Bestalde, multzoak egiteko bi kasu bereizi ditugu: hitz-formak morfemetan banatuta eta banatu gabe. Ohikoena ez banatzea da. Hala ere, Kim *et al.* (2014) lanean, koreerarako emaitza interesgarriak lortzen dituzte hitzak morfemetan banatu eta banatutako informazio hori rol semantikoak etiketatzeko sistema batean erabilita. Emaitza horiek ikusita, multzoak sortzeko garaian

hitzak banatuta erabiltzeak emaitza positiboak erdietsiko dituela uste dugu.

2.3.4 Meta-ezaugarriak

Meta-ezaugarrien inguruan egindako lana azaltzen dugunean, dependentziak automatikoki etiketatuta dituen corpus erraldoi batetik jasotako informazioa erabiliko da morfologikoki aberatsak diren hizkuntzen analisi sintaktikoa hobetzeko. Dependentzia sintaktikoak hobetzeko erdi-gainbegiratutako teknikak erabili dituzten autoreak ugari diren arren (McClosky *et al.*, 2006; Sagae eta Tsujii, 2007; Koo *et al.*, 2008; Chen *et al.*, 2009), tesi-lan honetan egindako esperimentuak Chen *et al.* (2013) lanean oinarritu dira.

Lan horretan, oinarri bezala analizatzaile sintaktiko estatistiko batek automatikoki etiketatu duen milioika hitz dituen zuhaitz-banku bat hartzen da. Ondoren, zuhaitz-banku horretako hitzen dependentzia loturen arteko maiztasunetatik ezaugarri berriak (meta-ezaugarriak deiturikoak) sortzen dira. Esan beharra dago, ezaugarri berri horien erabilpena ez dela ohiko eran egiten, hots, ez zaio hitz bakoitzari ezaugarri berri bat esleitzen. Izan ere, ezaugarriak hitz bat baino gehiagoren arteko loturetatik erauzitako informazioa direnez, horiek erabiltzeko modurik egokiena analizatzaile estatistikoak hitzen arteko loturak egiten dituenean erabiltzea baita.

Esanak esan, guk Chen *et al.* (2013) lanean egindakoa zabaldu egingo dugu ezaugarri mota gehiago sortuz eta morfologikoki aberatsak diren hizkuntzekin lan eginez.

Laburpena

Kapitulu honetan, tesi-lanarekin erlazioa duten hurbilpenak azaldu ditugu, irakurleari atzera egitea eta erreferentzia bila kapitulu hau kontsultatzea lagungarria izango zaiolakoan. Landu den ikerlerro bakoitza aurrera eramateko beste egile batzuen lanetan oinarritu gara. Hibridazioarekin lotutako esperimentuak planifikatzeko iturri batetik baino gehiagotik edan dugu. Hala ere, aipagarrienak Ezeiza *et al.* (1998) eta Tapanainen eta Voutilainen (1994) lanak direla deritzogu. Bi lan horietan, erregeletan oinarritutako sistemak eta estatistiketan oinarritutako sistemak konbinatzen dituzte emaitza interesgarriekin. Emaitza horietan oinarrituta, dependentzien analisi sintaktikoa hobetzeko, sistemen hibridazioak eskaintzen dituen bide desberdinak landu dira.

Ezaugarrien ingeniaritzaren inguruan egindako esperimentuekin lotutako lan nagusiak hiru izan dira. Ambati *et al.* (2010) eta Seeker eta Kuhn (2013) lanak aztertuz, ezaugarri morfologikoak erabiltzeko moduaren garrantziaz jabetu gara. Izan ere, ezaugarri bakoitzak ekarpen desberdina baitu analisi sintaktikoan, kasu batzuetan ekarpen negatiboa dutelarik. Çetinoğlu eta Kuhn (2013) lana aztertuz, aldiz, analizatzaile sintaktiko estatistikoek sartutako informazioari lehentasun desberdina ematen diotela konturatu gara, informazio bera leku desberdinetan erabilita emaitza desberdinak erdiesten direlarik.

Dependentzien analisia hobetzeko multzokatze teknikak erabiltzeko ideia, batez ere, Koo *et al.* (2008) lanetik atera dugu. Bertan, Brown multzoak erabiltzen dituzte ingeleserako eta txekierarako dependentzien analisia hobetzeko. Tesi-lan honetan, bertan landutako ideiak hartu ditugu eta horietan oinarritutako aldaera berriak aplikatu ditugu hainbat hizkuntzatan.

Erdi-gainbegiratutako tekniketan oinarritutako esperimentuei lotutako lanen artean aipagarriena Chen *et al.* (2013) da. Lan horretan, dependentziak automatikoki etiketatuta dituen corpus erraldoi batetik jasotako informazioa erabiltzen dute ezaugarri berezi batzuk sortzeko. Ezaugarri berri horiek oso baliagarriak dira dependentzia zuhaitzean guraso eta semeen arteko loturak zeintzuk izango diren ondorioztatzeko. Guk, lan horretan egin diren esperimentuak zabaldu egin ditugu iturri desberdinetatik sortutako meta-ezaugarriak hizkuntza desberdinetan aplikatuz.

3

Esperimentazio-ingurunea

Kapitulu honetan, segidan datozen kapituluetan aipatuko diren formatu berezi, baliabide eta ebaluaketarako erabili diren neurriei buruzko informazioa azalduko da. Are gehiago, baliabideren bat hobeto ulertzeko behar diren kontzeptuen azalpena ere emango da. Esanak esan, kapitulu hau hiru atal nagusitan banatu da: 3.1 puntuan, esaldiak sintaktikoki adierazteko erabiltzen diren eredu nagusiak aipatuko dira, baita eredu horiek errepresentatzeko erabili diren formatuak ere. 3.2 puntuan, aldiz, esku artean dugun lan hau aurrera eramateko erabili diren aurreprozesaketarako tresna, analizatzaile, corpus eta algoritmoei buruzko informazioa emango da. Bukatzeko, 3.3 puntuan, tesi-lan honetan aipatzen diren emaitza desberdinak adierazteko erabili diren neurriak aipatuko dira.

3.1 Etiketatze sintaktikoak eta formatu erabilienak

Esan bezala, esaldiak sintaktikoki etiketatzeko ereduen artean, bi dira erabilienak: dependentzietan oinarritutako eredua eta osagaietan oinarritutako eredua. Dependentzietan oinarritutako ereduan (ikus 3.1.1), esaldi bat osatzen duten hitzen arteko dependentzia irudikatzen da, egitura sintaktikoa errepresentatzen duen zuhaitzean hitz bat beste baten mende dagoela adieraziz. Osagaietan oinarritutako ereduan (ikus 3.1.2), ostera, esaldia osagaietan banatzen da. Era berean, osagai horiek osagaietan banatu daitezke, guztien artean esaldiaren egitura sintaktikoa irudikatzen dutelarik.

Esaldien egitura sintaktiko osoa adierazteko aipatutako ereduak nagusitzen diren arren, badaude azaleko sintaxia irudikatzeko ereduak ere. Hori dela eta, 3.1.3 puntuan, tesi lan honetan erabili direnak azalduko dira.

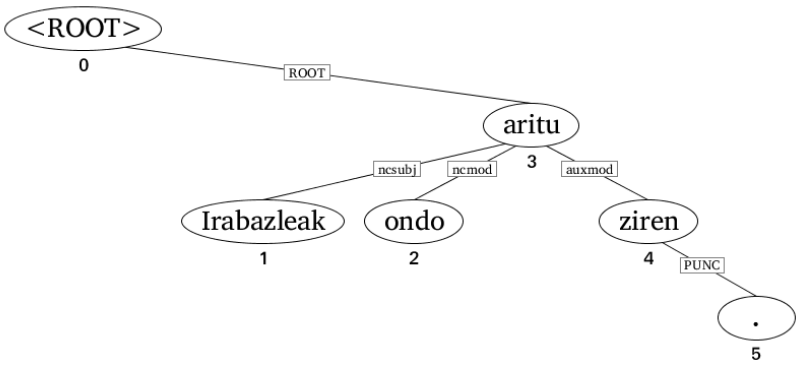
3.1.1 Dependentzietan oinarritutako eredua

Dependentzietan oinarritutako etiketatze sintaktikoan, esaldia osatzen duten elementu lexikalen artean binakako erlazioak ematen dira. Binakako dependentzia erlazio hori gobernatzailearen eta mendekoaren artean gertatzen da. Gauzak horrela, i j-ren gobernatzailea, burua edo gurasoa dela esaten da eta j i-ren mendekoa edo umea dela. Dependentzietan oinarritutako eredua jarraitzen duen sintaktikoki etiketatutako esaldi batean, guraso eta semearen arteko erlazio mota zein den ere zehaztu egiten da. Hori dela eta, esaldi baten egitura sintaktikoa grafikoki zuhaitz (ikus 3.1 irudia) edo grafo (ikus 3.2 irudia) gisa irudikatzen denean, adabegiek eta nodoek guraso eta umeak irudikatzen dituzte, eta arku zuzenduek, ostera, guraso eta umeen arteko erlazio motak.

Esandakoa formalki zehazteko har ditzagun n hitz dituen $E = w_1, ..., w_n$ esaldia eta arkuak etiketatzeko erabiliko den A etiketa multzoa. P esaldiaren grafo zuzendu edo zuhaitz etiketatuaren definizio formala *G = (I,R,L)* hirukotearen bidez egingo da:

- *I = 1,...,n*, *E* esaldiaren hitz-formen identifikadoreen multzoa.
- R $\subseteq$ I x I, E esaldiaren dependentzia-arkuen multzoa.
- L : I $\rightarrow$ A, E esaldiaren dependentzia-arkuak etiketatzeko dependentzia-etiketen multzoa.

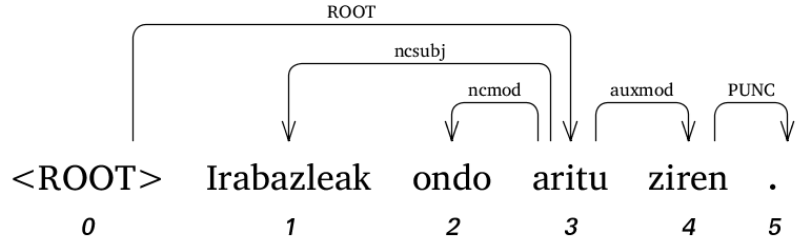
R multzoko dependentzia-arkuak guraso edo gobernatzailetik seme edo mendekora doaz eta arku horiek L dependentzia-etiketen bidez etiketatuko dira.



3.1 irudia – Dependentzietan oinarritutako eredua jarraituz etiketatutako esaldi baten zuhaitzaren irudikapena.

Esanak esan, ondo eraikitako zuhaitz edo grafo batek, adibidez 3.1 irudiko zuhaitzak, ondorengo irizpideak bete behar ditu:

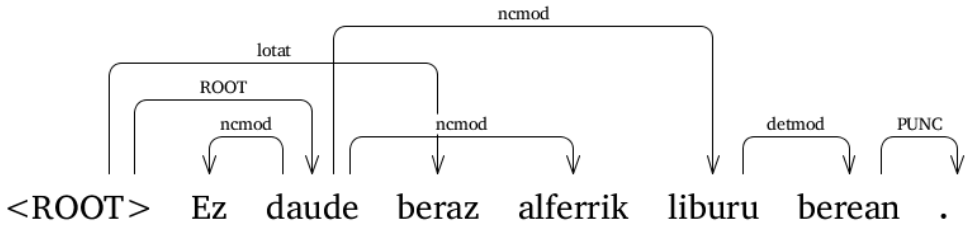
- Guraso eta umeen arteko loturak aziklikoak izan behar dira, hots, umeetatik ezin da gurasora itzuli arku zuzenduak zeharkatuz.
- Ume batek guraso bakarra izan dezake, baina guraso batek ume bat baino gehiago izan ditzake.



3.2 irudia – Dependentzietan oinarritutako eredua jarraituz etiketatutako esaldi baten grafo zuzenduaren irudikapena.

Aurretik aipatu den bezala, dependentzia-egitura modu batean baino gehiagotan irudika daiteke. Hala ere, gauzak erraztearren hemendik aurrera zuhaitza edo dependentzia-zuhaitza deituko diogu dependentzietan oinarrituta etiketatua izan den egitura sintaktikoari. Ildo beretik jarraituz,

dependentzia-zuhaitzean arkuak gurutzatzen ez direnean, zuhaitza proiektiboa dela esaten da. Era berean, arkuak gurutzatu egiten direnean zuhaitza ez-proiektiboa dela esaten da. Esandakoaren adibide dira 3.2 (proiektiboa) eta 3.3 (ez-proiektiboa) irudietako zuhaitzak. Esan beharra dago 3.2.1.1 atalean azalduko den dependentzien euskal zuhaitz-bankuan agertzen diren zuhaitz gehienak proiektiboak direla, ez-proiektiboen kopurua % 3tik beherakoa delarik.



3.3 irudia – Zuhaitz ez-proiektibo baten irudikapena.

Irakurlea konturatuko zen jada, azaldu diren irudietan *ROOT* izeneko elementua gehitu dela. Elementu horren helburua zuhaitzaren erroaren (gorengo adabegia) funtzioa betetzea da, esaldiko hitz guztiek buru edo gobernatzaile bat izan dezaten.

Dependentzietan oinarritutako etiketatze sintaktikoaren azalpenarekin bukatzeko, dependentziak errepresentatzeko erabiltzen diren formatu nagusiak azalduko dira ondorengo puntuetan. Formatu horietan ez da zuhaitza gurasо, seme eta dependentzia erlazio motekin soilik errepresentatzen; hitz-forma bakoitzaren ezaugarriak ere biltzen dira (lema, kategoria, azpikategoria eta ezaugarri morfologikoak).

Lehendabizi, dependentziak errepresentatzeko estandar bezala hartzen den formatua azalduko da 3.1.1.1 puntuan. Ondoren, azken horretan oinarritutako bi aldaera aurkeztuko dira 3.1.1.2 eta 3.1.1.3 puntuetan. Bukatzeko, 3.1.1.4 puntuan, aurrekoetatik aldentzen den dependentziak etiketatzeko formatu bat aurkeztuko da.

3.1.1.1 CoNLL-X

Computational Natural Language Learning (*CoNLL*) konferentzian (Buchholz eta Marsi, 2006), 1997. urtetik aurrera, hizkuntzaren prozesamenduarekin lotutako gaietan ataza partekatuak antolatu dira urtero. Ataza partekatu

edo lehiaketa horietan, partaide bakoitzak datu multzo beraren gainean egiten ditu esperimentuak bere sistemarekin. Ondoren, partaide guztien sistemek erdietsitako emaitzak konparatzen eta komentatzen dira. 2006. urtean, *CoNLL* konferentziaren 10. lehiaketa antolatu zen eta urte horretako ataza partekatuaren gaia dependentzietan oinarritutako analisi sintaktiko automatikoa izan zen. Partaideek eskuragarri zituzten zuhaitz-bankuak formatu konkretu batean zeuden, *CoNLL-X* formatuan hain zuzen ere. Formatu horretan, esaldiak lerro huts baten bidez daude banatuta eta esaldiko hitz bakoitzaren informazioa lerro batean jartzen da (ikus 3.1 taula). Hitz-forma bakoitzarentzat tabuladore baten bidez banatutako ondorengo informazioa gordetzen da:

1. ID: hitzak esaldian duen posizioaren zenbakia
2. WORD: hitz-forma
3. LEM: hitzaren lema
4. CPOS: hitzaren kategoria (coarse-grained POS)
5. POS: hitzaren kategoria eta azpikategoria (fine-grained POS)
6. FEATS: hitzaren ezaugarri morfosintaktikoak
7. HEAD: hitzaren burua
8. DEP: buruarekiko duen dependentzia erlazioa

Informazioa 10 zutabetan ematen den arren, azken bi zutabeetan ematen den informazioa hautazkoa da eta kasu gehienetan ez dira erabiltzen zutabe horiek. Hori dela eta, bi zutabe horiek ez ditugu aipatu azaldu berri dugun zerrendan eta 3.1 taulan ez ditugu sartu.

Esan bezala, aipatu berri dugun taulan (3.1 taula), CoNLL-X formatuan etiketatutako euskarazko esaldi bat aurkezten da. Bertan, adibidez, *horretan* hitza determinatzaile erakusle arrunta dela adierazten da, baita bere kasua inesiboa, numeroa singularra eta mugatua dela ere. Dependentzia sintaktikoari buruz, berriz, bere burua *Eraso* hitza dela esaten da (HEAD = 1) eta beraien arteko dependentzia erlazioa *detmod* motakoa dela. Era berean, *Eraso* hitza izen arrunta dela esaten da eta bere buruarekin (*hil*) *ncmod* dependentzia erlazioa duela.

ID	WORD	LEM	CPOS	POS	FEATS	H	DEP
1	Eraso	eraso	IZE	IZE_ARR	_	6	ncmod
2	horretan	hori	DET	DET_ERKARR	KAS:INE|NUM:S|MUG:M	1	detmod
3	ez	ez	PRT	PRT	MOD:EGI	6	ncmod
4	zen	izan	ADL	ADL	MDN:B1|NOR:HURA	6	auxmod
5	inor	inor	IOR	IOR_IZGMGB	KAS:ABS|MUG:MG	6	ncsubj
6	hil	hil	ADI	ADI_SIN	ADM:PART|ASP:BURU	0	ROOT
7	.	.	PM	PUNT_PUNT	_	6	PUNC

3.1 taula – Euskarazko zuhaitz-bankuaren esaldi bat CoNLL-X formatuan. H = HEAD, PM=PUNTU_MARKA.

3.1.1.2 CoNLL 2009

2009. urtean, *CoNLL* konferentzian antolatu zen ataza partekatuan landu zen gaia rol semantikoen etiketatze automatikoa izan zen (*Semantic Role Labeling*). Rol semantikoak etiketatzeko sistema askok hitzen arteko dependentzia sintaktikoak erabiltzen dituztenez, 2009ko atazan eskuragarri utzi zen informazioa dependentziak etiketatzeko *CoNLL-X* formatuan oinarritzen den formatuan banatu zen. Horretarako, *CoNLL-X* formatuan zeuden zutabeei rol semantikoen etiketatzerako beharrezkoak diren beste zutabe batzuk gehitu zitzaizkien. Esanak esan, hitz-forma bakoitzarentzat tabuladore baten bidez banatutako ondorengo informazioa gordetzen da:

1. ID: hitzak esaldian duen posizioaren zenbakia
2. FORM: hitz-forma
3. LEMMA: hitzaren lema
4. PLEMMA: hitzaren lema (automatikoa)
5. POS: hitzaren kategoria
6. PPOS: hitzaren kategoria (automatikoa)
7. FEATS: hitzaren ezaugarri morfosintaktikoak
8. PFEATS: hitzaren ezaugarri morfosintaktikoak (automatikoa)
9. HEAD: hitzaren burua
10. PHEAD: hitzaren burua (automatikoa)
11. DEPREL: buruarekiko dependentzia erlazioa
12. PDEPREL: buruarekiko dependentzia erlazioa (automatikoa)
13. FILLPRED: PRED zutabea bete behar den ala ez
14. PRED: predikatu bati esleitu zaion PropBank adiera
15. APREDs: predikatuaren adierak hartzen dituen argumentuak

Aurreko zerrendan, parentesi artean *automatikoa* hitza duten zutabeetan automatikoki erauzitako informazioa erabiltzen dela esan nahi da. Adibidez, PLEMMAk ingelesezko *predicted lemma* esan nahi du. Aipatzekoa da ere predikatu bati esleitu zaion adierak dituen argumentu kopurua haina zutabe gehitzen direla APREDs atalean. Hala ere, gure esperimentuetarako lehenengo 12 zutabeak soilik erabili ditugu, eta gainontzeko zutabeei azpimarra esleitu zaie. Bestalde, tresna automatikoen informazioa behar den zutabeetan urre-patroi informazioa erabili da. Horrenbestez, adibidez, LEMMA eta PLEMMA zutabeetan informazio bera erabili da, POS eta PPOS, FEATS eta PFEATS, HEAD eta PHEAD, eta DEPREL eta PDEPREL zutabeetan egin den bezala.

3.1.1.3 CoNLL-U

6. kapituluan sakonago azalduko den Universal Dependencies (UD) edo Dependentzia Unibertsalak proiektuaren helburuetako bat hainbat hizkuntzatarako era berean etiketatutako zuhaitz-bankuak sortzea da. Helburu hori bete ahal izateko gidalerro batzuk jarraitu behar dira, bakoitzak bere hizkuntzan duen zuhaitz-bankua bihurtu dezan. Gidalerro horietan, kategoriak, ezaugarri morfologikoak eta dependentzia erlazioak bihurtzeko azalpenak ematen dira, baita zuhaitz-bankua zein formatutan sortu behar den ere, zehazki *CoNLL-U* formatuan.

Aurretik esan den bezala, Dependentzia Unibertsalak proiektu barnean erabiltzen den formatua *CoNLL-X* formatuan oinarritzen da. Hori dela eta, azken hori hartu da eta proiektuaren beharrizanetara egokitu da. Gauzak horrela, hitz-forma bakoitzarentzat tabuladore baten bidez banatutako ondorengo informazioa gordetzen da bertan:

1. ID: hitzak esaldian duen posizioaren zenbakia
2. FORM: hitz-forma
3. LEMMA: hitzaren lema
4. UPOSTAG: hitzaren UD kategoria
5. XPOSTAG: hitzaren jatorrizko kategoria (bihurtu gabea)
6. FEATS: hitzaren UD ezaugarri morfosintaktikoak
7. HEAD: hitzaren burua
8. DEPREL: buruarekiko duen dependentzia erlazioa
9. DEPS: bigarren mailako dependentzien zerrenda
10. MISC: beste edozein ohar edo ezaugarri

ID	W	LEM	UP	XP	FEATS	H	DEP
1	Has	hasi	VERB	_	VerbForm=Inf	0	root
2	dadila	*edin	AUX	_	Number[abs]=Sing|Person[abs]=3	1	aux
3	liga	liga	NOUN	_	Case=Abs|Definite=Def|Number=Sing	1	nsubj
4	.	.	PUNCT	_	_	3	punct

3.2 taula – Euskarazko zuhaitz-bankutik UD formatura bihurtutako esaldi baten adibidea. W = WORD, LEM = LEMMA, UP = UPOSTAG, XP = XPOSTAG, H = HEAD eta DEP = DEPREL.

Aipatzekoa da euskarazko UD zuhaitz-bankuan azken bi zutabeak ez direla erabili; 5. zutabeari ere azpimarra esleitu zaio. Hala ere, etorkizunean zutabe horretako informazioa behar izanez gero, modu errazean berreskura daiteke. 3.2 taulan euskarazko UD zuhaitz-bankutik ateratako esaldi bat bildu dugu. Bertan *liga* hitzari erreparatzen badiogu, bere kategoria izena dela (*NOUN*), bere kasua absolutiboa dela, mugatua dela eta singularrean dagoela adierazten da. Are gehiago, *liga* hitza, esaldian erroaren papera betetzen duen *Has* hitzaren subjektua dela ere esaten da.

3.1.1.4 Murriztapen Gramatika formatua (*VISL CG Format*)

Esaldiak edo hitzak Murriztapen Gramatika formalismoa (Karlsson *et al.* 1995) erabilita etiketatzeko modu bat baino gehiago dagoen arren, tesi honetan VISL CG formatua (hemendik aurrera MG formatua) soilik erabili da. MG formatuan, esaldia osatzen duten hitzak goitik behera erakusten dira eta hitz bakoitzaren atzetik hurrengo lerroan berari buruzko informazioa aurkezten da. Informazio hori mota askotakoa izan daiteke: lema, kategoria, azpikategoria, ezaugarri morfologikoak, dependentzia sintaktikoa... 3.4 irudian, MG formatuan etiketatuta dagoen esaldi bat ikus daiteke. Bertan hitz bakoitzaren lema, kategoria, azpikategoria, ezaugarri morfologikoak, funtzio sintaktikoa, hitz segida mota eta dependentzia etiketa zehazten dira hurrenez hurren. Gainera, hitz bakoitzaren dependentzia etiketarekin batera etiketa hori duen hitzaren burua esaldian zein noranzkotan dagoen ere adierazten da '<' (aurretik) eta '>' (ondoren) ikurren bitartez.

```
"<Hiru>"
        "hiru" DET DZH NMGP ZERO @ID> %SIH &DETMOD>
"<puntuak>"
        "puntu" IZE ARR BIZ- ABS NUMP MUGM @OBJ %SIB &NCOBJ>
"<lortu>"
        "lortu" ADI SIN PART BURU NOTDEK @-JADNAG %ADIKATHAS &ADITZ_NAGUSI
"<ditugu>"
        "*edun" ADL A1 NOR_NORK NR_HAIEK NK_GUK @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"
       PUNT_PUNT
```

3.4 irudia – MG formatuan etiketatutako esaldi baten adibidea.

3.1.2 Osagaietan oinarritutako eredua

Izenak dioen bezala, osagaietan oinarritutako etiketatze sintaktikoaren oinarria osagaia da eta elementu hori unitate edo sintagma bakarra osatzen duen hitz multzo gisa definitu daiteke. Esanak esan, esaldia izen-sintagma, aditz-sintagma eta antzerako egituren papera hartzen duten osagaietan banatzen da. Era berean, elementu horiek osagai txikiagoetan banatu daitezke, denen artean zuhaitz bat eratzen dutelarik.

Esaldiaren osagai bakoitza nola deskonposatu azaltzen duten formalismo gehienak Chomsky-ren Testuingururik Gabeko Gramatikan (hemendik aurrera TGG) oinarritzen dira (Chomsky 1956, 1965, 1981). Esan daiteke, TGG bat elementu bakoitza zein osagaitan banatu behar den edo zein osagairekin ordezkatu behar den adierazten duen erregela multzoa dela. Erregela horietan bi motatako sinboloak aurki ditzakegu: bukaerakoak eta ez-bukaerakoak. Bukaerakoak, hitz arruntak dira (kotxea, euria, etorri...); ez-bukaerakoak, ostera, guztiz deskonposatu gabe dauden eta oraindik ordezkatu daitezkeen elementuak dira (izen-sintagma, aditz-sintagma, determinatzailea...).

TGGren deskribapen formala egiteko, G = (N, Σ, P, S) laukotea erabiliko da:

- N: ez-bukaerako ikurren multzoa
- Σ: bukaerako ikurren multzoa
- P: $A \rightarrow \alpha$ bezalako produkzioen edo erregelen multzoa, non $A \in N$ eta $\alpha \in (\Sigma \bigcup N)^*$ multzoaren ikurren katea den
- S: hasierako ikurra

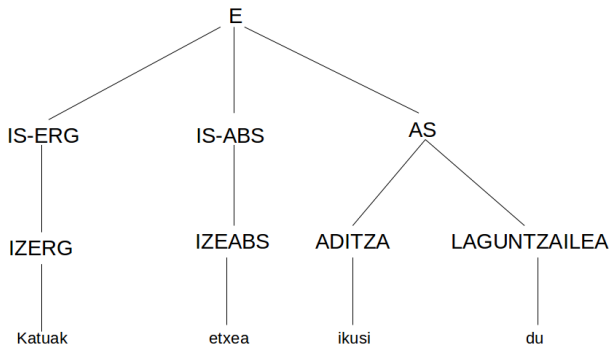
Esandakoa hobeto ulertzeko, 3.3 taulan euskarazko TGG bat aurkezten da. Taula aztertzen bada, gramatika horren arabera esaldia izen-sintagma

1.	E → IS-ERG IS-ABS AS
2.	IS-ERG → IZERG
3.	IS-ABS → IZEABS
4.	AS → ADITZA LAGUNTZAILEA
5.	IZERG → **gizonak** | **katuak**
6.	IZEABS → **etxea** | **baloiak**
7.	ADITZA → **ikusi** | **edan** | **erabili**
8.	LAGUNTZAILEA → **du** | **da** | **ditu**

3.3 taula – Euskarazko esaldi baten erregela sorta. Bukaerako ikurrak beltzez.

ergatibo, izen-sintagma absolutibo eta aditz-sintagma batez osatzen da. Era berean, esaterako, aditz-sintagma aditz batez eta laguntzaile batez osatzen da.

Bestalde, erregela multzo hori erabilita eraiki den *Katuak etxea ikusi du* esaldiaren zuhaitza aurkezten da 3.5 irudian. Zuhaitzean hiru motatako adabegiak aurki ditzakegu: erroa (gorengo adabegia), barne-adabegiak (gurasoa eta umeak dituztenak) eta hostoak (umerik ez dutenak). 3.5 irudian erroa *E* elementua da (esaldia) eta barne-adabegiak *IS-ERG* (izen-sintagma ergatiboan), *IZERG* (izena ergatiboan), *IS-ABS* (izen-sintagma absolutiboan), *IZEABS* (izena absolutiboan), *AS* (aditz-sintagma), *ADITZA* eta *LAGUNTZAILEA* dira. Hostoak, berriz, esaldian erabili diren hitzak dira: *Katuak*, *etxea*, *ikusi* eta *du*.



3.5 irudia – Euskarazko esaldi baten osagaietan oinarritutako zuhaitz-sintaktikoa.

3.1.3 Azaleko sintaxia etiketatzeko formatuak

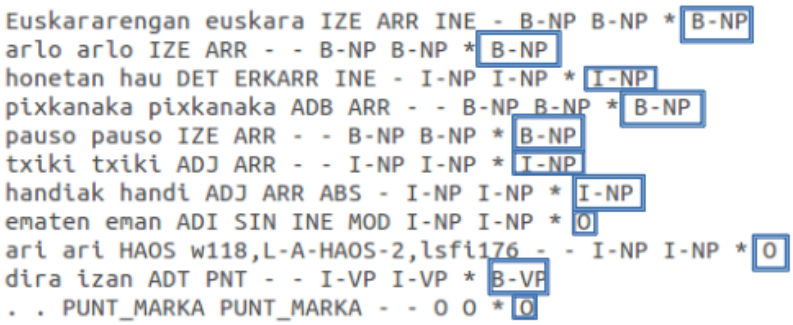
Esaldiak sintaktikoki etiketatzeko erabili ditugun eredu nagusiak dependentzietan oinarritutakoa eta osagaietan oinarritutakoa diren arren, azaleko sintaxiarekin lan egiten duten analizatzaileak ere erabili dira tesi honetan (aurrerago azalduko diren IXATI eta ML IXATI analizatzaileak). Analizatzaile horiek, *IOB* formatuan oinarritzen dira lan egiteko. Horrenbestez, 3.1.3.1 puntuan etiketatzeko estilo edo formatu hori aztertuko dugu.

3.1.3.1 *IOB* formatua (*Inside/Outside/Beginning*)

IOB formatua, *chunken* hasierak eta bukaerak markatzeko erabiltzen den formatua da. Chunkak (Abney, 1991) sintagma kategoriako zatiak dira, eta sintaktikoki erlazionaturiko hitzez osatuta daude. Chunken barneko egitura aztertuz gero, burua eta bere modifikatzaileak bereiziko genituzke. Horrela bada, testua chunketan zatitzea gainjartzen ez diren eta elkarrekin sintaktikoki erlazionaturik dauden hitz segidak identifikatzean datza. Era horretara, esaldi bateko izen-sintagmen eta aditz-kateen hasierak eta bukaerak markatu daitezke. Formatu honetan erabiltzen diren etiketak ondorengoak dira:

- **I:** Hitza chunkaren barruan dagoela esateko erabiltzen da
- **O:** Hitza ez dela chunkaren atal bat esateko erabiltzen da
- **B:** Hitza chunkaren hasierako hitza dela esateko erabiltzen da

Esan beharra dago formatu honen aldaera asko erabiltzen direla eta bakoitzak bere beharrizanetara moldatzen duela formatua. Esaterako, guk erabili ditugun chunkerrek formatu hau erabiltzen dute emaitza bueltatzeko, baina bakoitzak bere erara.



3.6 irudia – IOB formatuan etiketatutako esaldi baten adibidea. Kasu honetan, hitz bakoitza chunkaren atal den ala ez adierazteaz gain chunk mota ere adierazten da (urdinez).

3.2 Oinarrizko baliabideak

Gure tesi lanean erabili diren oinarrizko baliabideak azaltzeko aukeratu dugun puntu hau hiru zati nagusitan antolatu dugu: lehenengo zatian, esperimentuetan baliatu ditugun corpusak aztertuko dira; bigarren zatian, ostera, hitzen aurreprozesaketarako erabili ditugun tresnak aztertuko dira. Bukatzeko, analisi sintaktikoa egiteko baliatu ditugun analizatzaile desberdinak aztertuko dira.

3.2.1 Corpusak

Esperimentuak aurrera eramateko corpus desberdinez baliatu gara. Oinarri bezala euskarazko zuhaitz-bankua hartu dugun arren (ikus 3.2.1.1), hizkuntza desberdinetako zuhaitz-bankuekin lan egiteko aukera ere izan dugu, diseinatutako esperimentuak hizkuntza desberdinetan probatzeko (ikus 3.2.1.2). Bestalde, zenbait kasutan ezaugarri berriak gehitzeko etiketatu gabeko corpusak ere erabili dira, 3.4 taulako erdi-gainbegiratutako corpusetatik ateratako testua, hain zuzen ere.

Bukatzeko, euskarazko fenomeno linguistiko konkretuekin erlazionatutako probak egin direnez, fenomeno horietaz osatutako zuhaitz-bankuak ere erabili dira (ikus 3.2.1.3).

3.2.1.1 Dependentzien euskal zuhaitz-bankua

EPEC corpusa euskarazko 300.000 hitz biltzen dituen corpusa da. Corpusean bildutako informazioa EEBS proiektutik[1] eta Euskaldunon Egunkaria egunkaritik lortu da eta domeinu desberdinak jorratzen dira bertan: kirola, politika, aisialdia, kultura, iritzia eta abar. EPEC corpusa Hizkuntzaren Prozesamenduaren hainbat atazatan erabili da eta bere tamaina oso handia ez den arren, euskara bezalako hizkuntza baterako oso baliabide garrantzitsua da.

EPEC corpusean dependentzia sintaktikoak etiketatu zirenean EPEC-DEP izeneko corpusa sortu zen (Aranzabe, 2008; Aldezabal *et al.*, 2009). Esku artean dugun dependentzien euskal zuhaitz-bankua EPEC-DEP corpusa *CoNLL-X* formatura egokitu ondoren sortu zen. Egokitzapen horren ondoren, dependentzien euskal zuhaitz-bankuak 150.000 hitz ditu 11.225 esaldietan banatuta. Maila desberdinetako urre-patroi analisiak biltzen ditu: lema, kategoria, azpikategoria, ezaugarri morfologikoak eta dependentzia sintaktikoak. Zuhaitz-bankuan 22 kategoria desberdin bildu dira, 27 azpikategoria, 17 ezaugarri morfologiko eta 31 dependentzia sintaktiko mota. Arku ez-proiektiboen kopurua ere aipatzekoa da, % 1,3a hain zuzen. Zuhaitz-bankuan erabilitako dependentzia etiketen azalpenak A eranskinean bildu dira eta analisi kateko etiketen azalpenak, aldiz, `http://ixa2.si.ehu.es/iakesgoenaga/AnalisiKatekoLaburtzapenak.ods` dokumentuan.

3.2.1.2 SPMRL 2013-2014 corpusak

2013an *Statistical Parsing of Morphologically Rich Languages* kongresuan (Seddah *et al.*, 2013) morfologikoki aberatsak diren hizkuntzetan zentratu zen ataza partekatua antolatu zuten. Bertan, dependentzietan oinarritutako edo osagaietan oinarritutako zuhaitz-bankuak ardatz gisa hartuz, hizkuntza desberdinen analisi sintaktikoan emaitzarik hoberenak erdiesteko hainbat partaide lehiatu ziren. Partaideen esku utzi zen bakoitzak aztertu nahi zituen hizkuntzak eta zuhaitz-bankuen egitura mota aukeratzea (dependentzietan oinarritutakoa edo osagaietan oinarritutakoa).

Hurrengo urtean, berriro antolatu zen kongresua (Seddah *et al.*, 2014) eta urte horretako helburuetako bat analisi sintaktikoan erdi-gainbegiratutako iturrietatik erauzitako ezagutza erabiltzea izan zen. Hori dela eta, hizkuntza

[1] `www.euskaracorpusa.net`

bakoitzerako oinarrizko zuhaitz-bankuez gain, dependentziak automatikoki etiketatutako zuhaitz-banku erraldoiak utzi ziren eskuragarri.

Guk bi kongresuetan hizkuntza berdinak aukeratu ditugu: euskara, frantsesa, alemana, hungariera eta suediera. Horrenbestez, 3.4 taulan bildu ditugu hizkuntza horietarako oinarrizko zuhaitz-bankuen eta automatikoki etiketatutako zuhaitz-bankuen ezaugarriak. Esan beharra dago, oinarrizko zuhaitz-bankuen lema, kategoria, azpikategoria eta ezaugarri morfologikoak automatikoki eratorritakoak direla. Euskarazko zuhaitz-bankuen tamaina dependentzien euskal zuhaitz-bankuaren tamaina baino txikiago da. Izan ere, ataza partekatuetan dependentzietan oinarritutako zuhaitz-bankuetan eta osagaietan oinarritutako zuhaitz-bankuetan esaldi berdinak ordena berdinean joan behar baitira, eta parekatze hori % 100ean ezin izan baita burutu antolatzaileek eskatzen duten datarako.

Hizkuntza	**Train**	**Dev**	**Test**	**Erdi-gainbegiratutakoa**	
Euskara (Aduriz *et al.*, 2003)	96.368	13.851	11.457	150M	Orekatua
Frantsesa (Abeillé *et al.*, 2003)	443.113	38.820	75.216	120M	Berrien zerbitzua
Alemana (Seeker eta Kuhn, 2012)	719.532	76.704	92.004	205M	Wikipedia
Hungariera (Vincze *et al.*, 2010)	170.141	29.989	19.908	100M	Berrien zerbitzua
Suediera (Nivre *et al.*, 2006)	76.357	9.341	10.690	24M	Orekatua

3.4 taula – Esperimentuetan erabili diren hizkuntza desberdinetako zuhaitz-bankuen ezaugarriak.

3.2.1.3 Fenomeno linguistiko konkretuen zuhaitz-bankuak

4. kapituluan hibridazio mota desberdinak landuko dira. Horietako batean estatistiketan oinarritutako analizatzaile sintaktikoak eta erregeletan oinarritutako analizatzaile sintaktikoak konbinatzen dira. Konbinazio horrekin, alde batetik, euskararen analisi sintaktiko orokorrean lortzen diren emaitzak neurtu nahi dira; beste aldetik, hibridazio mota horrek euskararen fenomeno linguistiko konkretuetan erdiesten dituen emaitzak aztertu nahi dira. Hori dela eta, azken esperimentu horiek aurrera eramateko, hiru fenomeno linguistiko aukeratu dira eta fenomeno horietaz osatutako zuhaitz-bankuak bildu nahi dira.

Aukeratutako fenomenoak, galderak, koordinazioa eta esaldi konpletiboak dira. Hala ere, oraingoz galderazkoen zuhaitz-bankua soilik osatu da, gainontzekoak osatzeko lehen urratsak eman direlarik. Galderazko esaldiak, galdera

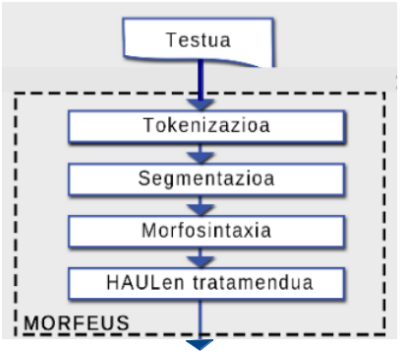
erantzun (Question answering) corpus batetik hartu dira eta lema, kategoria, azpikategoria eta ezaugarri morfologikoak automatikoki esleitu ondoren, dependentzia sintaktikoak eskuz etiketatu dira (*CoNLL-X* formatua). Gauzak horrela, 102 esalditan banatutako 803 hitzeko zuhaitz-bankua osatu da. Esaldien luzerari dagokionez, esaldiak 5-15 hitz bitartekoak dira.

3.2.2 Aurreprozesaketarako tresnak

Puntu honetan, analisi sintaktikorako lagungarria den informazioa sortzen duten tresnak aztertuko dira. Maila desberdinetako informazioa sortzen dute, tokenizaziotik hasita multzokatze mota desberdinetako informaziora arte. Esan beharra dago, ondorengo lerroetan tesi-lan honetan erabili diren aurreprozesaketarako tresnak soilik azalduko direla. Horrenbestez, euskararekin soilik lan egiten duten tresnak (ikus 3.2.2.1 eta 3.2.2.2 puntuak) eta hizkuntza desberdinekin lan egiten duten tresnak aztertuko dira (ikus 3.2.2.3 eta 3.2.2.4 puntuak).

3.2.2.1 MORFEUS

MORFEUS (Aduriz *et al.*, 1998) analizatzaile morfosintaktikoaren lana 4 atal edo modulutan banatzen da: tokenizazioa, segmentazioa, morfosintaxia eta Hitz Anitzeko Unitate Lexikalen (HAUL) tratamendua.



3.7 irudia – MORFEUS tresnaren moduluen irudia.

Tokenizazioa Tokenizatzaileak testu gordina jaso eta elementu bakunetan banatzen du. Elementu bakun edo token horietako bakoitza, hitza, zenbaki

arrunta, zenbaki erromatarra, hitz deklinatu gabea, deklinatua, sigla, zuriunea, puntuazio marka edo laburdura den identifikatzen da eta kasu bakoitzean gehitu behar zaizkion ezaugarriak gehitzen zaizkio.

Segmentazioa edo analisi morfologikoa Segmentatzailearen lana, token bakoitza lema eta morfemetan zatitzea da, posible diren interpretazio guztiak bilduta; horretarako, euskarri bezala Euskararen Datu Base Lexikala (EDBL) (Aldezabal *et al.*, 2001) hartzen du. Esaterako, tokena izena edo izenondoa denean, bere kategoria, azpikategoria, kasua, numeroa, etab... identifikatzen dira. Tokena aditza denean, berriz, modua, denbora, edo aspektua bezalako ezaugarriak detektatzen dira.

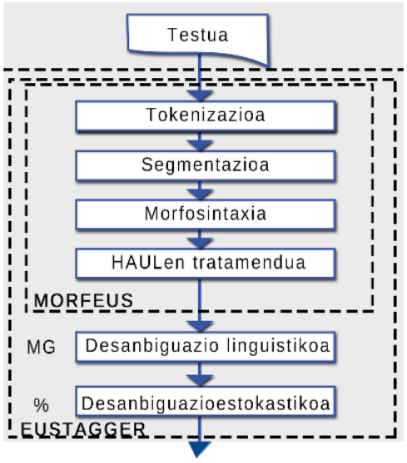
Morfosintaxia Morfosintaxiaren bidez hitzen barne-egitura eta egitura horiek beste forma batzuekin sintagma eta perpausak osatzeko konbinatzen direnean finkatzen diren erlazioak aztertzen dira. Hitzen egitura zehazteko, testuingururik gabeko gramatika bat erabili da (Gojenola, 2000b; Aduriz *et al.*, 2000) eta PATR formalismoaren inplementazio baten bidez (Shieber, 1986) hitzen barruko analisi morfosintaktikoa egin da. Beste era batera esanda, aurreko fasean sortutako morfemetatik abiatuz hitz-formaren interpretazio posible bakoitzarentzako ezaugarri-egitura bat lortu da.

Hitz Anitzeko Unitate Lexikalen (HAUL) tratamendua HAULak tratatzeko HABIL (Urizar, 2012) tresna erabiltzen da. Tresna horrek EDBLn gordeta dauden HAULen deskribapenak aztertzen ditu, HAUL hautagaia benetan HAULa den identifikatzeko. Ondoren, HAULaren interpretazio guztien analisi morfosintaktikoak sortuko ditu osagaien interpretazioetan oinarrituta katean integratzeko.

3.2.2.2 EUSTAGGER lematizatzailea/etiketatzailea

EUSTAGGERen (Ezeiza, 2002) lana MORFEUSen irteeran sortutako analisi morfosintaktiko guztien artetik testuinguruan egokia den analisia aukeratzea da. EUSTAGGERrek bi anbiguotasun morfosintaktiko motarekin egiten du lan: kategoriarena eta morfema ez-askearena. Anbiguotasun hori ebazteko, desanbiguazio morfologikoaren modulua (Ezeiza *et al.*, 1998) bi urratsetan konbinatu da:

- Lehenengo urratsean, Murriztapen Gramatika formalismoaren bitartez, hizkuntza-ezagutzan oinarritutako desanbiguazioa egiten da.
- Bigarren urratsean, aldiz, corpusetan oinarritutako teknika estatistikoa (Markov-en eredu ezkutua) erabiltzen da. Teknika horren funtsa, hitzak dituen interpretazio guztien artetik egokiena, corpus handi batetik ateratako estatistikak kontuan hartuta, aukeratzean datza.



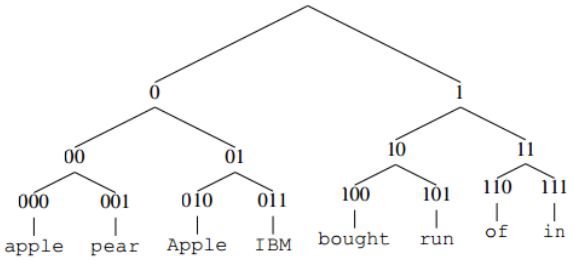
3.8 irudia – MORFEUS eta EUSTAGGER tresnen moduluen irudia.

3.2.2.3 Brown multzoak sortzeko tresna

Brown multzoak sortzeko Liang (2005) lanean sortutako Brown algoritmoaren (Brown *et al.*, 1992) inplementazioa erabili dugu. Tresna hori C++ lengoaia erabilita inplementatu da eta sarrera gisa testu gordina hartu eta hitz bakoitzarentzat bere Brown multzoa bueltatzen du. Tresnak, hitzak nahi diren haina multzotan banatzeko aukera ematen du. Hala ere, oso kontuan hartzekoa da tresnak multzoak bueltatzeko denbora hertsiki lotuta dagoela pasatako corpusean dauden hitz desberdinen kopuruarekin eta sortu nahi ditugun multzo kopuruarekin, hau da, bere konplexutasun funtzioa $O(NC^2)$ da, non N corpuseko hitz kopurua eta C sortu nahi diren multzoen kopurua diren.

Esanak esan, Brown multzoak zer nolakoak diren azaltzeko, ondorengo lerroetan Brown algoritmoaren nondik norakoak aztertuko dira.

Brown algoritmoa Brown multzokatze algoritmoak sarrera bezala N tokeneko corpus bat jasotzen du, bertan m hitz desberdin daudelarik (m hitzeko hiztegia). Algoritmoak hasieran hitz bakoitza multzo desberdinean banatzen du. Ondoren, behin eta berriro Markoven eredu ezkutuaren arabera probabilitatearen beherakada txikiena eragiten duen multzo bikotea batzen du, hitz guztiak definitutako multzoetan banatu arte.



3.9 irudia – Brown algoritmoaren bidez sortutako zuhaitz bitarra.

Algoritmoak bukatzen duenean, zuhaitz bitar bat bezala irudikatu daitezkeen multzo desberdinen hierarkia bat lortzen da (ikus 3.9 irudia). Zuhaitz bitar horren barruan, hitz bakoitza modu argian dago identifikatuta zuhaitz bitarraren errotik hitzerainoko bidearen bidez. Are gehiago, aipatutako bide hori bit-kate bezala errepresentatu daiteke. Aurkeztutako irudiari erreparatzen badiogu, sortutako hitz multzoak identifikatu nahi badira, adibidez biko sakoneran dauden multzoak, biko sakoneran dauden lau nodoak hartu behar dira: [*apple,pear*], [*Apple,IBM*], [*bought,run*] eta [*of,in*]. Aipatzekoa da ere, multzo berdinak lortu daitezkeela hitz bakoitzaren bit-katearen hasierako bi zenbakiak hartuta. Bestalde, bit-kateen luzera desberdinak erabiliz orokortze maila desberdinetako multzoak sortzen dira. Esaterako, *000* eta *001* multzoek sagarra eta udarea biltzen dituzte, hurrenez hurren; baina maila bat gorago joanez gero aurkitzen dugun multzoak (*00* multzoa) frutak errepresentatzen dituela esan daiteke. Berdina gertatzen da beste kasuetan ere: *01* multzoak, ordenagailuekin erlazionatutako enpresak biltzen ditu; *10* multzoak, aditzak biltzen ditu eta *11* multzoak preposizioak.

3.2.2.4 Word2vec tresna

Word2vec tresna (Mikolov *et al.*, 2013c, b, a) bi eredutan oinarritzen da hitz-bektoreak sortzeko. Hitz-bektoreak, hizkuntzaren prozesamendu espa-

rruan erabiltzen diren hizkuntzen ereduak erauzteko eta ezaugarriak ikasteko teknikak erabiliz lortzen diren hitzen errepresentazio bektorialak dira. Hitz bakoitzari bektore bat esleitzen zaio, non bektore bakoitza dimentsio bakoitza errepresentatzen duen koordenatuez osatua dagoen. Era horretara, hitz bakoitza errepresentatzen duen bektoreak, esleitutako dimentsio kopurua hainako luzera izango du.

Word2vec tresna, bi geruza dituen sare neuronalekin eratutako bi eredutan oinarritzen da: *Skip-gram* eta *CBOW* (continuous bag-of-words). *CBOW* ereduan, uneko hitza zein den asmatzen saiatzen da bere testuinguruaren arabera. Demagun *Ander etxera joan da* esaldia dugula eta bilaketa leihoa batekoa dela. Uneko hitza *etxera* dela kontsideratzen bada, eredua uneko hitz probableena lortzen saiatzen da *Ander* eta *joan* hitzak (testuingurua) erabilita. *Skip-gram* ereduan kontrakoa egiten da; *etxera* hitza hartuta, bateko leiho barruan testuinguru probableena zein den asmatzen saiatzen da.

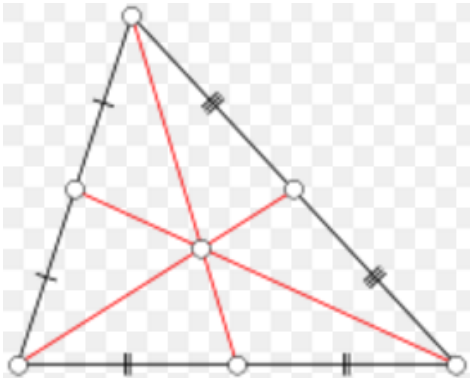
Esanak esan, Word2vec tresnak testu gordina jasotzen du sarrera bezala eta hitz bakoitzarentzako bektore bakarra bueltatzen du irteeran. Hitzbektoreek daukaten ezaugarri interesgarrietako bat, beraien artean eragiketa aljebraikoak egin daitezkeela da. Adibidez, *errege* hitzaren bektoreari *gizon* hitzaren bektorea kentzen bazaio eta *emakume* hitzaren bektorea gehitzen bazaio, *erregina* bektoretik oso hurbil dagoen bektorea ateratzen da. Hurbiltasuna, 3.3 puntuan azalduko dugun kosinu distantziarekin kalkulatzen da.

Dena den, tesi-lanean ez dira hitz-bektoreak bere horretan erabili, horietan oinarritutako multzoak baizik. Hitz-bektoreetan oinarritutako multzoak sortzeko ere Word2vec tresna erabili da, zeinak *K-means* multzokatze algoritmoa inplementatzen duen.

K-means multzokatze algoritmoa K-means multzokatze algoritmoa (Hartigan eta Wong, 1979), gehienbat datu meatzaritzan erabiltzen den algoritmoa da. Algoritmoaren funtsa, n elementu k multzotan banatzea da, n elementuetako bakoitza zentroide hurbilena duen multzoan banatzen delarik. Algoritmoak bi urrats hauek ematen ditu gelditzeko baldintza bete arte:

- **Esleipen urratsa:** Elementu bakoitzari batezbesteko hurbilena duen multzoa esleitzen zaio.
- **Eguneraketa urratsa:** Multzoetako zentroide berriak kalkulatzen dira. Zentroidea n dimentsio dituen espazio batekoa den elementu bat

hiperplanoarekiko n-bolumena duten bi zati berdinetan banatzen duten hiperplano guztien ebakidura da (ikus 3.10 irudia).

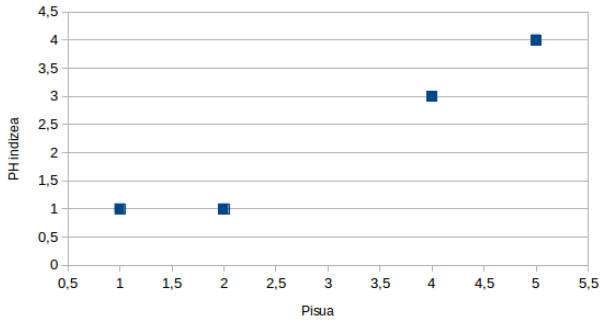


3.10 irudia – Triangeluaren zentroidea.

Algoritmoa esleipen urratsean aldaketarik ez dagoenean gelditzen da. Esandakoa hobeto ulertzeko adibide bat erabiliko dugu. Demagun 3.5 taulako elementuak (A, B, C eta D) multzokatu nahi ditugula K-means algoritmoarekin.

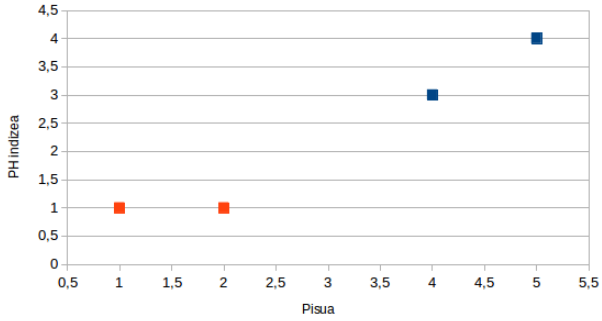
Elementua	**Pisua**	**Altuera**
A	1	1
B	2	1
C	4	3
D	5	4

3.5 taula – A, B, C eta D elementuen pisu eta altueren balioak.



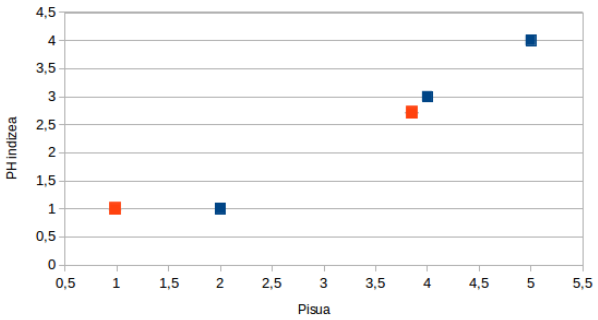
3.11 irudia – 3.5 taulako elementuen irudikapen grafikoa.

Taulako elementuak grafikoki 3.11 irudian irudikatu dira. K-means algoritmoaren lehenengo urratsa k zentroide hasieratzea da. Kasu honetan $k = 2$ izatea eta hasierako zentroideak A eta B elementuak izatea aukeratu da (ikus 3.12 irudia, zentroideak gorriz).



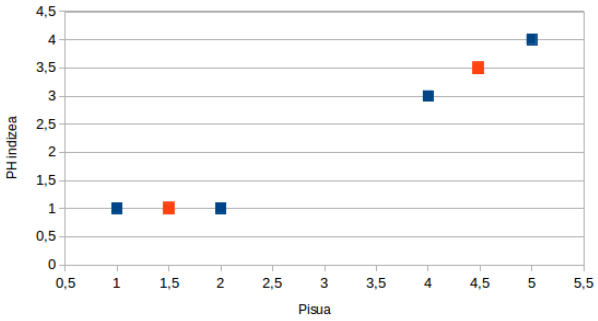
3.12 irudia – Zentroideen hasieraketa.

Esleipen urratsean A elementua lehenengo multzoari esleitu zaio eta B, C eta D elementuak bigarren multzoari esleitu zaizkio, hurbilen duten zentroidea bigarrena baita. Eguneraketa urratsean zentroide berriak kalkulatzen dira (ikus 3.13 irudia).



3.13 irudia – Zentroideen eguneraketa.

Hurrengo esleipen urratsean A eta B elementuak lehenengo multzoari esleitu zaizkio eta C eta D elementuak bigarren multzoari. Hurrengo eguneraketa urratsean zentroide berriak kalkulatzen dira berriro (ikus 3.14 irudia). Hurrengo esleipen urratseko esleipenak ez dira aldatzen eta algoritmoak ez du aurrera jarraitzen. Horrenbestez, A eta B elementuak lehenengo multzoari esleitu zaizkio eta C eta D elementuak bigarren multzoari.



3.14 irudia – Zentroideen eguneraketa berria.

3.2.3 Analizatzaile sintaktikoak

Puntu honetan, tesi-lanean erabili diren analizatzaile sintaktikoak aztertuko dira. Mota desberdinetako analizatzaile sintaktikoak erabili dira. Horrenbestez, azalpena hiru zatitan banatu da: 3.2.3.1 puntuan estatistiketan oinarritutako analizatzaile sintaktikoei buruzko informazioa emango da; 3.2.3.2

puntuan, berriz, erregeletan oinarritutako analizatzaileak azalduko dira. Bukatzeko, 3.2.3.3 puntuan, analisi sintaktikoak bozketa bidez konbinatzen dituen tresna estatistikoa azalduko da.

3.2.3.1 Estatistiketan oinarritutako analizatzaile sintaktikoak

Estatistiketan oinarritutako analizatzaile sintaktiko desberdinez baliatu gara esperimentuak gauzatzeko. Alde batetik, dependentzien analisi sintaktiko osoa egiten dutenak (*parser*) erabili ditugu; beste aldetik, chunker estatistiko bat erabili dugu euskarazko testuetan aplikatzeko.

Emango diren azalpenetan, lehenengo, dependentzien analisi sintaktiko osoa egiten duten analizatzaileei buruzko informazioa emango da bi zatitan banatuta: trantsizioetan oinarritutako analizatzaileak (3.2.3.1.1) eta grafoetan oinarritutako analizatzaileak (3.2.3.1.2). Ondoren, chunker estatistikoa azalduko da 3.2.3.1.2 puntuan.

3.2.3.1.1 Trantsizioetan oinarritutako parserrak Trantsizioetan oinarritutako analizatzaile sintaktikoen funtzionamendua azaldu da 2.1.1 puntuan. Horrenbestez, puntu honetan, erabili diren mota horretako analizatzaileak aztertuko dira. Trantsizioetan oinarritutako MaltParser (Nivre, 2006; Nivre *et al.*, 2007b) parserra eta hori optimizatzeko MaltOptimizer tresna (Ballesteros eta Nivre, 2012b) erabili dira tesi-lan honetako zenbait esperimentu burutzeko. Ondorengo lerroetan tresna horiei buruzko informazioa emango da, esperimentuetarako horiek aukeratu izanaren arrazoiekin batera.

MaltParser MaltParser[2] analizatzaileak sarrera gisa *CoNLL-X* formatuan dagoen zuhaitz-bankua jasotzen du ikasketarako eta testerako, eta formatu horretan bertan bueltatzen ditu esaldien dependentzia zuhaitzak. Parserraren funtzionamenduan hiru atal nagusi bereizi daitezke: analisi sintaktikorako algoritmoa, ezaugarri-eredua eta ikasketa automatikorako aukeratzen den sailkatzailea.

MaltParserrek trantsizioetan oinarritutako hainbat algoritmo eskaintzen ditu analisi sintaktikorako: Nivre algoritmoak, Pila algoritmoak, Covington algoritmoak eta Multiplanar algoritmoak. Guk erabili duguna Pila algoritmoen familiako *Stacklazy* algoritmo ez-proiektiboa (Nivre, 2009; Nivre *et al.*,

[2] `http://www.maltparser.org/`

2009) izan da, egindako probetan algoritmo hori erabilita lortu ditugulako emaitzarik hoberenak.

Hurrengo trantsizioa zein izango den erabakitzeko ezaugarri-eredua sortzen da; hori definitzeko, analizatzaileak emandako urratsen historia eta sarreran dauden uneko elementuen ezaugarriak erabiltzen dira. Ezaugarri-eredua, ezaugarri sinplez osatutako dimentsio handiko bektore bat da, eta ezaugarri sinple horiek definitzeko bi funtzio erabiltzen dira:

- **Helbide funtzioa:** Funtzio honen bidez, sortu diren egitura partzialetatik edozein hitzetara heltzeko erabakiak jaso daitezke.

- **Egozpen funtzioa:** Egozpen funtzioaren bidez, hitzen ezaugarri morfosintaktikoak lortzen dira (hitz-forma, lema, kategoria, azpikategoria etab...).

Esperimentuetan ezaugarri-ereduaren balio lehenetsiak erabili ditugu: pila eta bufferrean dauden hitzen kategoria gramatikalak, hitz-formak eta gurasoekin dituzten dependentzia erlazioak. Ezaugarrien arteko konbinaketak egiteko, aldiz, kategorien arteko n-gramak eta kategoria eta hitz-formen arteko bikoteak erabili dira.

Sarrerako datuetatik abiatuz, ikasketa automatikorako sailkatzaileak datu horiei dagokien klasea iragarriko du. MaltParserren kasuan, sailkatzailea ezaugarri-ereduan oinarrituta hurrengo trantsizioa zein izango den erabakitzeko erabiltzen da, kasu batzuetan analisi sintaktikorako algoritmoak informazio hori eskatuko baitio sailkatzaileari. MaltParser, ikasketa automatikorako sostengu bektore makinez baliatzen da (*Support Vector Machine*, *SVM*) (Cortes eta Vapnik, 1995). Are gehiago, bi pakete daude hautagai *SVM* sailkatzaileen artean: LibSVM (Chang eta Lin, 2001) eta LibLINEAR[3] (Fan *et al.*, 2008) paketeak, guk azken hori baliatu dugularik.

Behin MaltParser analizatzailea azalduta, hori erabiltzeko arrazoi nagusiak emango dira hurrengo lerroetan. Arrazoi nagusiak hiru izan dira:

– **Ugaritasuna**: Metodologia desberdinak aplikatzen dituzten analizatzaile sintaktikoak erabiltzea pentsatu dugu emaitza aberatsagoak lortzeko asmoarekin. Gauzak honela, MaltParser, trantsizioetan oinarritzen diren parserren artean ezagunena dela esan daiteke.

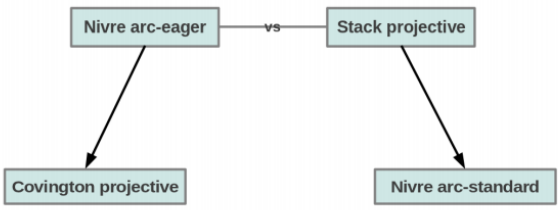
[3] `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`

– **Eraginkortasuna**[4]: Analizatzaile sintaktiko ugari eskura izan arren, MaltParser erabiltzea pentsatu dugu topologikoki desberdinak diren hainbat hizkuntzatan aplikatu ondoren emaitza onak lortu dituztelako era guztietako esperimentuetan urteetan zehar.

– **Azkartasuna**: Argi dago parser bat eraginkorra izatea komeni dela emaitza onak lortzeko, baina azkarra izatea ere oso garrantzitsua da ahalik eta esperimentu gehien egin ahal izateko ahalik eta denbora laburrenean. Erreferentzia bezala, 100.000 hitzeko zuhaitz banku batetik ikasten, MaltParser gai da ordu bete eta hamar minututan emaitzak bueltatzeko 4Gb-eko Intel Core i5 batean (ikasketa + testa).

MaltOptimizer MaltOptimizer[5], ikasketarako corpusean oinarrituta MaltParserrek eman beharreko urratsak optimizatzen dituen tresna da. MaltOptimizer tresnak burutzen duen optimizazioa hiru fasetan banatzen da:

– **Datuen baliozkotzea eta azterketa:** Datuen baliozkotzean, sartutako informazioa formatu egokian dagoen (*CoNLL-X*) begiratzen da, baita dependentzia zuhaitzak ongi eraikita dauden ere. Datuen azterketan, ostera, ikasketarako zuhaitz-bankuko informazioa biltzen da. Adibidez, hitz/esaldi kopurua eta ez-proiektiboak diren arku/zuhaitzen portzentajea.



3.15 irudia – Algoritmo proiektibo hoberena aukeratzeko erabaki zuhaitza.

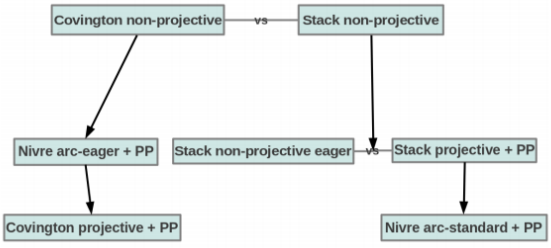
[4]Analizatzaile bat aukeratzeko arrazoiak ematean eraginkortasuna aipatzen dugunean, emaitzei lotutako eraginkortasuna adierazi nahi da eta ez algoritmoaren exekuzio denborarekin lotutako eraginkortasuna (eraginkortasun algoritmikoa).

[5]`http://nil.fdi.ucm.es/maltoptimizer/install.html`

– **Analisi sintaktikorako algoritmoa:** Bigarren fase honetan, analisi sintaktikorako algoritmo egokiena aukeratzen da lehenengo fasean bildutako informazioan oinarrituta. Ikasketarako zuhaitz-bankuan zuhaitz ez-proiektiborik ez badago, algoritmo proiektiboen artean aukeratzen da (ikus 3.15 irudia). Zuhaitz-bankuan zuhaitz ez-proiektibo asko badaude, berriz, zuhaitz ez-proiektiboak tratatzeko algoritmoak aztertzen dira (ikus 3.16 irudia). Bestalde, zuhaitz ez-proiektibo eta proiektiboen kopurua antzerakoa bada, aipatutako bi algoritmo motak aztertzen dira.

– **Ezaugarrien aukeraketa:** Hirugarren fasean, ezaugarri-eredua optimizatzen da. Horretarako, lehendabizi, oinarrizko ezaugarri-ereduan sartu diren ezaugarri guztiek ekarpena duten ala ez aztertzen da. Ondoren, ahalmen handiena duten ezaugarriekin esperimentuak egiten dira horiek banan-banan eta konbinaketen bidez probatuz.

MaltOptimizer aukeratu izanaren arrazoiak MaltParser aukeratu izanaren oso antzekoak dira, bata bestearen optimizazioa baita. Hala ere, badaude zenbait desberdintasun. MaltOptimizerrek eraginkortasun hobea dauka, kasu batzuetan 3 puntuko hobekuntza erdiesten delarik. Ezaugarri negatibo bezala, aldiz, MaltOptimizerrek denbora gehiago behar du emaitza bueltatzeko. Erreferentzia bezala, 100.000 hitzeko zuhaitz-banku batetik ikasten, MaltParserrek baino hogeita hamabost minutu gehiago behar ditu 4Gb-eko Intel Core i5 batean (ikasketa + testa).



3.16 irudia – Algoritmo ez-proiektibo hoberena aukeratzeko erabaki zuhaitza, non PP = erdi-proiektiboa den.

3.2.3.1.2 Grafoetan oinarritutako parserrak Grafoetan oinarritutako analizatzaile sintaktiko estatistikoen eredua 2.1.2 puntuan azaldu da. Hori

dela eta, tesi lan honetan erabili diren mota horretako parserretan jarriko dugu arreta. Grafoetan oinarritutako bi parserretaz baliatu gara: MSTParser (McDonald *et al.*, 2005, 2006) eta Mate (Bohnet, 2010) parserrak.

MSTParser MSTParser[6] analizatzaile estatistikoak *CoNLL-X* formatuan dagoen zuhaitz-bankua jasotzen du eta emaitzak formatu horretan bueltatzen ditu. Bi urrats nagusi ematen dira emaitza bueltatzeko: lehendabizi, esaldiko elementu bakoitzaren burua aukeratzen da; ondoren, bi elementuen arteko dependentzia erlazioa esleitzen da. MaltParserren kasuan ikusi den bezala, MSTParserren funtzionamendua ere hiru atal nagusitan bereiz daiteke: ikasketarako algoritmoa, ezaugarri-eredua eta ikasketa automatikorako sailkatzailea.

Ikasketarako algoritmoen artean aukera bat baino gehiago daude eskuragarri MSTParser analizatzailearekin erabiltzeko. Bueltatuko den azken zuhaitza zein den erabakitzeko, esaldia osatzen duten zuhaitz-arkuen puntuazioaren batuketa maximizatzen duen *Maximun Spanning Tree* (*MST*) algoritmo globala erabiltzen da. Hala ere, bilaketarako bi algoritmo nagusitzen dira. Analisi proiektiborako Eisner (1996) algoritmoa erabiltzen da, zeinak zuhaitz-proiektibo guztien gaineko bilaketa egiten duen lehen edo bigarren mailako ezaugarriak erabiliz. Analisi ez-proiektiborako, berriz, Chu-Liu-Edmonds (Chu eta Liu, 1965; Edmonds, 1967) deituriko algoritmoaz baliatzen da. Algoritmo horrek ere zuhaitz guztien gaineko bilaketa gauzatzen du lehen edo bigarren mailako ezaugarriez baliatuz. Guk erabilitako zuhaitzen artean gehiengoa ez-proiektiboa denez, guk gehienbat Chu-Liu-Edmondsen algoritmoa erabili dugu.

Aipatutako algoritmoek bigarren mailako ezaugarriak erabiltzen dituztenean, bigarren mailako algoritmo bezala kontsideratzen dira. Bigarren mailako algoritmoek, zuhaitz probableena aukeratzeko garaian, guraso bereko bi elkarren ondoko umeen arkuak ere kontuan hartzen dituzte, lehenengo mailakoek guraso eta umearen arteko arkuak soilik hartzen dituzten bitartean.

Ezaugarri-eredua sortzeko, ezaugarri desberdinak kontsideratzen dira erabiliko den algoritmoaren mailaren arabera. Lehen mailako algoritmoetan, guraso eta umearen hitz-forma, kategoria, azpikategoria eta ezaugarri morfologikoak kontsideratzen dira; baita guraso eta umearen artean dauden hitz guztien kategoria eta azpikategoria ere, guraso eta umearen eskuineko eta ezkerreko hitzen kategoria eta azpikategoriarekin batera. Bigarren maila-

[6] `http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html`

ko algoritmoetan, aldiz, lehenengo mailako ezaugarriez gain, gurasoaren eta guraso hori duten elkarren ondoko umeen hitza eta azpikategoria, beraien artean dagoen distantziarekin konbinatzen dira.

MSTParserrek, ezaugarri-ereduan erabilitako ezaugarrien pisua ikasteko *Margin Infused Relaxed Algorithm* (*MIRA*) (Crammer eta Singer, 2003; McDonald *et al.*, 2005) sailkatzailea erabiltzen du. Sailkatzaile honek egiten duen iterazio bakoitzean, entrenamendurako jaso dituen instantzia (esaldi) guztiak tratatzen ditu eta pisuen bektorea eguneratzen du. Prozesuak iterazio gehiago egiten ditu instantzia guztiak linealki banatu arte ala iterazio kopuru maximora iritsi arte.

Esanak esan, gure esperimentuetarako MSTParser aukeratzeko arrazoiak ondorengoak izan dira:

- **Kodea moldatzeko erraztasuna**: Zenbait kasutan, esperimentuak aurrera eramateko analizatzailearen kodean aldaketak egin behar izan ditugu. Parser honen kodean murgiltzea nahiko erraza da beste batzuekin konparatzen badugu. IXA taldean parser hau moldatu duen jendea egoteak ikasketarako lana aurreztu digu.

- **Eraginkortasuna**: Analizatzaile sintaktiko ugari eskura izan arren, MSTParser erabiltzea pentsatu dugu, urteetan zehar hainbat atazatan eta hizkuntzatan modu arrakastatsuan erabili delako, artearen egoerako emaitzak lortuz.

- **Azkartasuna**: Erreferentzia bezala, 100.000 hitzeko zuhaitz banku batetik ikasten, parser hau gai da ordu bete eta berrogeita bost minututan emaitzak bueltatzeko 4Gb-eko Intel Core i5 batean (ikasketa + testa).

Mate Mate[7] parserraren bertsio desberdinak dauden arren, grafoetan oinarritutako bertsioaz baliatu gara tesi lan honetan. Mate analizatzaile sintaktiko estatistikoak *CoNLL 2009* formatuan jasotzen du zuhaitz-bankua eta formatu horretan bueltatzen du emaitza. Aurrekoetan bezala, bere funtzionamendua hiru ataletan banatu daiteke: ikasketarako algoritmoa, ezaugarri-eredua eta sailkatzailea.

Ikasketarako algoritmo bezala, MSTParser analizatzailean erabiltzen den bigarren mailako algoritmoaren hedapen bat (Carreras, 2007) erabiltzen da.

[7] `https://code.google.com/archive/p/mate-tools/downloads`

Hedapen horretan, probabilitateak lortzeko kasu gehiago hartzen dira kontuan. Adibidez, esaldian gurasoaren eta umearen artean dagoen uneko umearen umea ere kontuan hartzen da.

Ezaugarri-ereduari dagokionez, egiten diren konbinaketa guztiak 3.17 irudian bildu ditugu.

#	Standard Features	#	Linear Features		Linear G. Features		Sibling Features
1	l,h_f,h_p,d(h,d)	14	l,h_p,h+1_p,d_p,d(h,d)	44	l,g_p,d_p,d+1_p,d(h,d)	99	l,s_l,h_p,d(h,d)⊕r(h,d)
2	l,h_f,d(h,d)	15	l,h_p,d-1_p,d_p,d(h,d)	45	l,g_p,d_p,d-1_p,d(h,d)	100	l,s_l,d_p,d(h,d)⊕r(h,d)
3	l,h_p,d(h,d)	16	l,h_p,d_p,d+1_p,d(h,d)	46	l,g_p,g+1_p,d-1_p,d_p,d(h,d)	101	l,h_l,d_p,d(h,d)⊕r(h,d)
4	l,d_f,d_p,d(h,d)	17	l,h_p,h+1_p,d-1_p,d_p,d(h,d)	47	l,g-1_p,g_p,d-1_p,d_p,d(h,d)	102	l,d_l,s_p,d(h,d)⊕r(h,d)
5	l,h_p,d(h,d)	18	l,h-1_p,h+1_p,d-1_p,d_p,d(h,d)	48	l,g_p,g+1_p,d_p,d+1_p,d(h,d)	75	l,∀d_m,∀s_m,d(h,d)
6	l,d_p,d(h,d)	19	l,h_p,h+1_p,d_p,d+1_p,d(h,d)	49	l,g-1_p,g_p,d_p,d+1_p,d(h,d)	76	l,∀h_m,∀s_m,d(h,s)
7	l,h_f,h_p,d_f,d_p,d(h,d)	20	l,h-1_p,h_p,d_p,d-1_p,d(h,d)	50	l,g_p,g+1_p,h_p,d(h,d)		**Linear S. Features**
8	l,h_p,d_f,d_p,d(h,d)		**Grandchild Features**	51	l,g_p,g-1_p,h_p,d(h,d)	58	l,s_p,s+1_p,h_p,d(h,d)
9	l,h_f,d_f,d_p,d(h,d)	21	l,h_p,d_p,g_p,d(h,d,g)	52	l,g_p,h_p,h+1_p,d(h,d)	59	l,s_p,s-1_p,h_p,d(h,d)
10	l,h_f,h_p,d_f,d(h,d)	22	l,h_p,g_p,d(h,d,g)	53	l,g_p,h_p,h-1_p,d(h,d)	60	l,s_p,h_p,h+1_p,d(h,d)
11	l,h_f,d_f,h_p,d(h,d)	23	l,d_p,g_p,d(h,d,g)	54	l,g_p,g+1_p,h-1_p,h_p,d(h,d)	61	l,s_p,h_p,h-1_p,d(h,d)
12	l,h_f,d_f,d(h,d)	24	l,h_f,g_f,d(h,d,g)	55	l,g-1_p,g_p,h-1_p,h_p,d(h,d)	62	l,s_p,s+1_p,h-1_p,d(h,d)
13	l,h_p,d_p,d(h,d)	25	l,d_f,g_f,d(h,d,g)	56	l,g_p,g+1_p,h_p,h+1_p,d(h,d)	63	l,s-1_p,s_p,h-1_p,d(h,d)
77	l,h_l,h_p,d(h,d)	26	l,g_f,h_p,d(h,d,g)	57	l,g-1_p,g_p,h_p,h+1_p,d(h,d)	64	l,s_p,s+1_p,h_p,d(h,d)
78	l,h_l,d(h,d)	27	l,g_f,d_p,d(h,d,g)		**Sibling Features**	65	l,s-1_p,s_p,h_p,h+1_p,d(h,d)
79	l,h_p,d(h,d)	28	l,h_f,g_p,d(h,d,g)	30	l,h_p,d_p,s_p,d(h,d) ⊕r(h,d)	66	l,s_p,s+1_p,d_p,d(h,d)
80	l,d_l,d_p,d(h,d)	29	l,d_f,g_p,d(h,d,g)	31	l,h_p,s_p,d(h,d)⊕r(h,d)	67	l,s_p,s-1_p,d_p,d(h,d)
81	l,d_l,d(h,d)	91	l,h_l,g_l,d(h,d,g)	32	l,d_p,s_p,d(h,d)⊕r(h,d)	68	s_p,d_p,d+1_p,d(h,d)
82	l,d_p,d(h,d)	92	l,d_p,g_p,d(h,d,g)	33	l,p$_f$,s_f,d(h,d)⊕r(h,d)	69	s_p,d_p,d-1_p,d(h,d)
83	l,d_l,h_p,d_p,h_l,d(h,d)	93	l,g_l,h_p,d(h,d,g)	34	l,p_p,s_f,d(h,d)⊕r(h,d)	70	s_p,s+1_p,d-1_p,d_p,d(h,d)
84	l,d_l,h_p,d_p,d(h,d)	94	l,g_l,d_p,d(h,d,g)	35	l,s_f,p_p,d(h,d)⊕r(h,d)	71	s-1_p,s_p,d-1_p,d_p,d(h,d)
85	l,h_l,d_l,d_p,d(h,d)	95	l,h_l,g_p,d(h,d,g)	36	l,s_f,d_p,d(h,d)⊕r(h,d)	72	s_p,s+1_p,d_p,d+1_p,d(h,d)
86	l,h_l,h_p,d_p,d(h,d)	96	l,d_l,g_p,d(h,d,g)	37	l,s_f,d_p,d(h,d)⊕r(h,d)	73	s-1_p,s_p,d_p,d+1_p,d(h,d)
87	l,h_l,d_l,h_p,d(h,d)	74	l,∀d_m,∀g_m,d(h,d)	38	l,d_f,s_p,d(h,d)⊕r(h,d)		**Special Feature**
88	l,h_l,d_l,d(h,d)		**Linear G. Features**	97	l,h_l,s_l,d(h,d)⊕r(h,d)	39	∀l,h_p,d_p,$x_p$between h,d
89	l,h_p,d_p,d(h,d)	42	l,g_p,g+1_p,d_p,d(h,d)	98	l,d_l,s_l,d(h,d)⊕r(h,d)		
41	l,∀h_m,∀d_m,d(h,d)	43	l,g_p,g-1_p,d_p,d(h,d)				

3.17 irudia – Mate parserrak ezaugarri-ereduan biltzen dituen konbinaketa guztiak, non l dependentzia etiketa, h gurasoa (burua), d semea, s semearen anaia, g biloba, **d**(x,y,[,z]) hitzen ordena eta **r**(x,y) distantzia diren.

Ikasketarako algoritmoari laguntzen dion sailkatzaile bezala, erabilitako datu-egituretan hiztegi (*hash*) funtzioa aplikatzen duen Hash Kernel-a erabiltzen da. Funtzio horren bitartez bilatzen dena, iragartzen den zuhaitz sintaktikoan gertatzen diren errore kopurua minimizatzea da. Horretarako, ezaugarrietan oinarritutako pisu bektoreak funtsezkoak dira. Sailkatzailearen algoritmoak, lehendabizi, ikasketako instantzia bakoitzean ezaugarriak erauzten ditu; ondoren, hash funtzioaren bidez ezaugarriak pisu bektorearentzako indizeetan bihurtzen dira eta pisu bektoreak kalkulatzen dira. Pisu bektoreekin esaldi bakoitzarentzako dependentzia zuhaitz probableena sortzen da.

Hurrengo urratsean *MIRA* algoritmoaren eguneraketa funtzioaren hedapena den eguneraketa funtzioa aplikatzen da. Funtzioak bi pisu bektore, une-

ko esaldia, esaldiaren urre-patroiko dependentzia egitura, esaldiaren dependentzia egitura automatikoa eta pisuen eguneraketa erabiltzen ditu. Gaizki etiketatutako arku kopurua gutxienez batekoa bada, pisu bektoreak eguneratu egiten dira eta prozesua errepikatzen da. Beste era batera esanda, algoritmoa ikasten doan heinean errore tasa gutxitzen doa gehiago ikasi ezin duen arte. Prozesuak, iterazio maximoa egin arte ala errorerik ez izan arte jarraitzen du.

Aipatzekoa da ere paralelizazioan burutzen duen lana. Parser baten algoritmoak paralelizatu daitezkeen atazak dauzka. Esaterako, ezaugarrien erauzketa eta pisu bektoreen kalkulua. Ataza horiek eta beste batzuk paralelizatuz kalkuluen lan karga banatzea lortzen da. Hori dela eta, prozesu osoaren denbora asko murrizten da.

Esanak esan, tesi-laneko esperimentuetarako Mate parserra aukeratzeko arrazoiak ondorengoak dira:

- **Ugaritasuna**: Metodologia desberdinak aplikatzen dituzten analizatzaile sintaktikoak erabiltzea pentsatu dugu emaitza aberatsagoak lortzeko asmoarekin. Gauzak honela, Mate, MSTParserrekin batera, grafoetan oinarritzen diren parserren artean ezagunena dela esan daiteke.

- **Eraginkortasuna**: Erabili ditugun parserren artean eraginkorrena da eta horrela dela erakutsi da urteetan zehar hizkuntza desberdinetan egindako lanetan.

- **Azkartasuna**: Zalantzarik gabe, probatu dugun parserrik azkarrena da. Erreferentzia bezala, 100.000 hitzeko zuhaitz banku batetik ikasita, Mate gai da hamabi minututan emaitzak bueltatzeko 4Gb-eko Intel Core i5 batean (ikasketa + testa).

ML IXATI Erabilitako analizatzaile sintaktiko estatistikoekin bukatzeko, ML IXATI (Arrieta, 2010b) chunkerra azaltzea besterik ez da falta. Chunkerrak erabiltzen duen sailkatzailearen algoritmoa Carreras (2005) lanean azaldutakoa da. Bertan, Pertzeptroien algoritmo tradizionalaren (Rosenblatt, 1958) orokortze bat egiten da. Algoritmoaren oinarrizko funtzionamendua ondorengoa da: lehendabizi, esaldiko hitz guztiak markatzen dira hitz-multzo edo kate bezala. Ondoren, kate guzti horiek jasotzen dituen funtzio batek (*score* funtzioa delakoa) iragarpena zuzena den ala ez aztertzen du urre-patroiarekin alderatuz. Zuzena ez bada, beharrezkoak diren zuzenketak egiten dira erregela sinple batzuen bidez hurrengo iteraziorako. Algoritmoak

iterazio kopuru maximo haina aldiz ala egindako akatsa minimoa izan arte iteratzen du.

Carreras (2005) lanean sortutako algoritmoak euskaraz ahalik eta eraginkortasun handiena eduki zezan, bi egokitzapen egin behar izan zituzten: alde batetik, atributu gehiago sartu zizkieten (lema, azpikategoria, deklinabide kasua, mendeko perpaus mota, IXATI zatitzailearen irteera etab...); bestetik, algoritmoan esplizituki agertzen ziren ingeleseko zenbait termino euskarara ekarri zituzten.

ML IXATIk, sarrera bezala 3.1.1.4 puntuan aipatu dugun *MG* formatuan dagoen informazioa erabiltzen du. Irteera bezala, berriz, hitz bakoitza lerro batean bueltatzen du bere ezaugarriekin eta iragarritako etiketarekin. Iragarritako etiketa 3.1.3.1 puntuan azaldutako *IOB* formatuan dago eta etiketa horien bitartez markatzen dira hitz-kateen mugak. Etiketak eta beraien esanahiak ondorengoak dira:

- **B-NP:** izen-sintagmaren hasiera
- **I-NP:** izen-sintagmaren atala (ez hasierakoa)
- **B-VP:** aditz-katearen hasiera
- **I-VP:** aditz-katearen atala (ez hasierakoa)
- **O:** ez da izen-sintagma edo aditz-katea

Chunker hau erabiltzeko arrazoi nagusiak ondorengoak izan dira:

- **Espezializazioa:** ML IXATI, euskara ardatz duen chunker estatistiko bakarra da.
- **Eraginkortasuna:** Egindako azken esperimentuetan % 90eko F-measure balioa lortu da.

3.2.3.2 Erregeletan oinarritutako analizatzaile sintaktikoak

Jarraian erregeletan oinarritzen diren erabilitako bi analizatzaile sintaktiko azalduko ditugu. EDGK analizatzaileak (Aranzabe, 2008), esaldiko hitzetan dependentzia etiketak esleitzen ditu; IXATI chunkerrak (Aduriz *et al.*, 2004), berriz, esaldian aurkitzen dituen chunkak markatzen ditu.

EDGK EDGK analizatzaileak 3.1.1.4 puntuan azaldu dugun *MG* formatuan jasotzen du informazioa (hitza, lema, kategoria, azpikategoria, funtzio sintaktikoa...) eta jasotakoari iragarritako emaitza esleitzen dio *Constraint Grammar Parser CG-2* analizatzailearen bidez (Tapanainen, 1996). Adibide bezala 3.18 irudia erabiliko dugu. Irudia aztertuta, hitz bakoitzaren analisiaren bukaeran hitzaren dependentzia sintaktikoa adierazten dela konturatuko gara.

```
"<Hiru>"
        "hiru" DET DZH NMGP ZERO @ID> %SIH &DETMOD>
"<puntuak>"
        "puntu" IZE ARR BIZ- ABS NUMP MUGM @OBJ %SIB &NCOBJ>
"<lortu>"
        "lortu" ADI SIN PART BURU NOTDEK @-JADNAG %ADIKATHAS &ADITZ_NAGUSI
"<ditugu>"
        "*edun" ADL A1 NOR_NORK NR_HAIEK NK_GUK @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"
        PUNT_PUNT
```

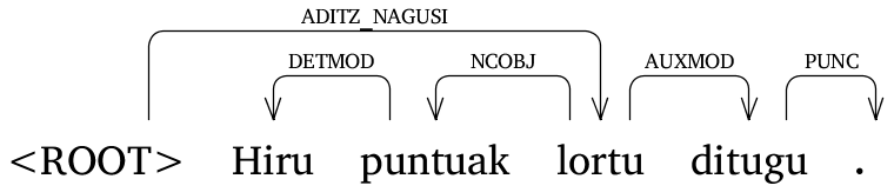
3.18 irudia – MG formatuan etiketatutako esaldi baten adibidea. Urdinez EDGKren emaitza.

Dependentziak etiketatzeko era hau desberdina da aurretik azaldu direnekin alderatuta. Kasu horretan ez da adierazten zehazki burua edo gobernatzailea zein hitz den; horren ordez, dependentzia erlazioaren izena jartzeaz gain, '<' eta '>' ikurrak erabiltzen dira gobernatzailea zein noranzkotan dagoen adierazteko. Dena den, hitz bakoitzak bere dependentzia etiketa duenean, beste kasuetan egin den bezala, dependentzia zuhaitza eraiki daiteke.

Hitza	Hiru	puntuak	lortu	ditugu
Dependentzia etiketa	DETMOD>	NCOBJ>	ADITZ_NAGUSI	<AUXMOD

3.6 taula – Euskarazko esaldi bati esleitutako dependentzia etiketak.

Aurkeztu dugun *Hiru puntuak lortu ditugu.* adibidean hitzei esleitu zaizkien dependentzia etiketak 3.6 taulan bildu ditugu. Ildo beretik jarraituz, taulako dependentzietatik abiatuta 3.19 irudiko zuhaitz sintaktikoa eraikitzen da, aurretik erakutsi diren dependentzietan oinarritutako zuhaitzen baliokidea delarik.



3.19 irudia – *Hiru puntuak lortu ditugu* esaldiaren dependentzia zuhaitza.

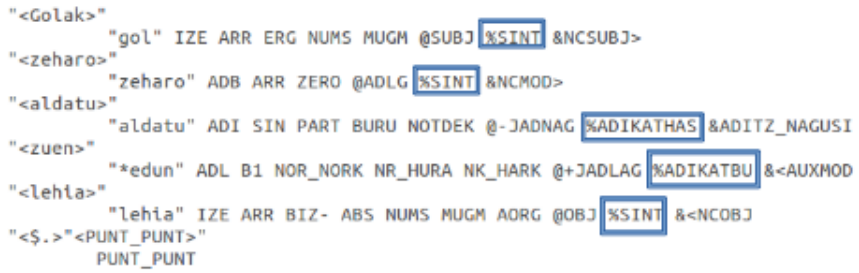
Aipatzekoa da EDGK erabiltzearen arrazoi nagusia, euskara ardatz duen erregeletan oinarritutako analizatzaile sintaktiko bakarra dela da, hau da, euskarazko esaldien dependentzietan oinarritutako analisi sintaktiko osoa egin daiteke berarekin, eta hori oso ondo datorkigu hibridazioarekin lotutako esperimentuak aurrera eramateko.

IXATI IXATI zatitzaileak, sintagmak, aditz-kateak, postposizio-lokuzioak[8] eta entitate izendunak[9] markatzen ditu esaldian. Hala ere, tesi-lan honetan sintagma eta aditz-kateetan zentratu gara, chunk horiek IXATI zatitzailearekin modu honetara markatzen direlarik:

- **%SIH:** sintagmaren hasiera
- **%SIB:** sintagmaren bukaera
- **%SINT:** sintagma (hitz bakarrekoa denean)
- **%ADIKATHAS:** aditz-katearen hasiera
- **%ADIKATBU:** aditz-katearen bukaera
- **%ADIKAT:** aditz-katea (hitz bakarrekoa denean)

3.20 irudian, IXATI zatitzaileak *Golak zeharo aldatu zuen lehia* esaldian aurkitu dituen chunkak azaltzen dira. Irudian, *Golak*, *zeharo* eta *lehia* hitzak sintagma bezala markatu dira. *aldatu*, aldiz, aditz-katearen hasiera bezala definitu da eta *zuen* aditz-katearen bukaera bezala.

[8]Perpausaren sintagmen arteko erlazio gramatikalak definitzen dituzten forma askeak.
[9]Entitate-izenak (pertsonak, lekuak eta erakundeak), denbora-adierazpenak eta zenbakizko espresioak biltzen dira entitate izendunetan.



3.20 irudia – MG formatuan etiketatutako esaldi baten adibidea. Bertan, chunkak markatuta agertzen dira urdinez.

IXATI erabiltzearen arrazoiak EDGK erabiltzearen arrazoien antzekoak dira. Euskarazko esaldien chunkak mugatzen dituen erregeletan oinarritutako zatitzaile bakarra da IXATI. Horrenbestez, honek irteeran ematen duen emaitza hibridazioa lantzen den esperimentuetarako baliagarria dela deritzogu.

3.2.3.3 Konbinaketarako tresnak

Tesi-lan honetako zenbait esperimentutan, analizatzaile desberdinek sortzen dituzten analisiak konbinatzea erabaki da analisi bakoitzak eskaini dezakeen ekarpena azken emaitzan islatu dadin. Ataza hori aurrera eramateko MaltBlender (Hall *et al.* 2010) tresna erabiltzea pentsatu dugu.

MaltBlender MaltBlender tresnak funtsean *CoNLL-X* formatuan jasotako analisi guztien arteko konbinaziorik onenarekin erdietsitako emaitza bueltatzen du. Demagun, zuhaitz-banku bera bost parser desberdinekin aztertu dugula eta ondorioz zuhaitz-banku beraren bost analisi desberdin ditugula. MaltBlenderrek bost analisi horiek jasoko ditu eta konbinaketak egiten hasiko da bozketa sistema baten bidez. Bozketa sistema horren bitartez analisietan ikusitako portaeren arteko konbinaketarik hoberena (emaitzarik hoberena) aukeratzen da eta konbinazio horren emaitza bueltatzen da azken analisi batean.

3.3 Erabilitako neurriak

Tesi-lan honetan erabili diren neurriak ondorengo lerroetan azalduko ditugunak dira. Sistema konkretu baten eraginkortasun orokorra neurtzeko erabili dugun neurria *Labeled Attachment Score* (LAS) neurria izan da. Dependentzia etiketa zehatz batean sistemak duen eraginkortasuna neurtzeko, aldiz, doitasuna, estaldura eta F-measure neurriak erabili dira. Bukatzeko, hitzbektoreen arteko antzekotasuna neurtzeko kosinu distantziaz baliatu gara eta hobekuntzak estatistikoki esanguratsuan diren ala ez zehazteko McNemar testa erabili dugu.

– **Labeled Attachment Score:** Gurasoa eta dependentzia etiketa ongi esleitu zaien hitzen portzentajea.

$$LAS = \frac{Gurasoarekin\_dependentzi\_zuzenarekin\_lotutako\_hitzak}{Hitz\_kopurua} \tag{3.1}$$

– **Doitasuna:** Sistemak itzulitako emaitza zuzenen kopuruak, itzulitako guztien artean osatzen duen portzentajea.

$$Doitasuna = \frac{Emaitza\_zuzenak}{Itzulitako\_emaitzak} \tag{3.2}$$

– **Estaldura:** Sistemak itzulitako emaitza zuzenen kopuruak, urre patroiko emaitzen artean osatzen duen portzentajea.

$$Estaldura = \frac{Emaitza\_zuzenak}{Urre - patroiko\_kasuak} \tag{3.3}$$

– **F-measure/F1:** Doitasuna eta estalduraren arteko batezbesteko harmonikoa da.

$$F - measure = 2 * \frac{Doitasuna * Estaldura}{Doitasuna + Estaldura} \tag{3.4}$$

– **Kosinu distantzia:** Bektoreen arteko angeluaren kosinua bueltatzen du

$$Kosinu\_distantzia = Cos\frac{V1 + V2}{\sqrt{V1 + V1} * \sqrt{V2 + V2}} \tag{3.5}$$

– **McNemar testa:** McNemar testa bi sailkatzaileen arteko aldea esanguratsua den ala ez erabakitzeko erabiltzen da. Horretarako, corpusa bi zatitan banatu behar da: ikasketa corpusa eta test corpusa. Bi sailkatzaileak (A, B) ikasketa corpus bera erabiliz ikasi ondoren, test corpus beraren gainean ebaluatu behar dira. Hipotesi nuluaren arabera, A sailkatzaileak ondo eta B sailkatzaileak gaizki sailkatutako adibideen kopuruak A sailkatzaileak gaizki eta B sailkatzaileak ondo sailkatutako adibideen kopuruaren berdina izan behar du. Datu hauen arabera, χ^2 testean oinarritzen da McNemar testa, hipotesi nulu hau uka daitekeen edo ez erabakitzeko. Hipotesia errefusatu baldin badaiteke, bi sailkatzaileen arteko aldea esanguratsua dela esaten da.

Laburbilduz, 3. kapitulu honetan hurrengo kapituluetan erabiliko diren eta tesi-lan honen ulermenerako beharrezkoak diren formatu berezi, baliabide eta neurriak azaldu dira. Hurrengo hiru kapituluetan, sarrera gisa erabili den kapituluan zehaztu diren helburuak betetzeko aurrera eraman diren esperimentuak azalduko dira, eta azkenengo kapituluan, esperimentu horietatik eratorritako ondorioak eta horiei lotutako etorkizuneko lanak zehaztuko dira.

4 Hibridazioa

Hibridazioa naturan, teknologian eta hainbat zientziatan ematen den prozesua da. Bertan, izaera desberdinetako elementuak modu batera edo bestera konbinatzen dira, erabilitako elementuen ezaugarriak biltzen dituen elementu berri bat sortuz. Kasu gehienetan, hibridazioa modu batean baino gehiagotan landu daiteke, beti, lortu nahi den emaitzaren edo hibridoaren arabera. Tesi-lan honen testuinguruan, beraien jatorrian ezberdinak diren analisi sintaktiko automatikoa lantzen duten sistemak konbinatzea izango litzateke hibridazioa. Horrela, gure emaitzak hobetzea espero dugu Hall *et al.* (2007) lanean adierazten den modura:

> "The evaluation shows that hybrid representations can be produced with only a marginal loss in accuracy for dependency and constituency considered separately. With better tuning we believe it will be possible to eliminate this loss and perhaps even achieve better accuracy than for separate constituency and dependency parsing." (Hall *et al.* 2007)

4.1 Sarrera

2.2.1 puntuan, tesi-lan honetan hibridazioaren inguruan egin ditugun esperimentuei lotutako lanak aztertu dira. Emandako informazioa zabaltzeko asmoarekin, esan beharra dago analisi sintaktikoarekin lotuta dauden hibridazioaren inguruko artearen egoerako lan gehienetan, estatistiketan oinarritutako sistema desberdinak beraien artean konbinatzen direla. Kapitulu ho-

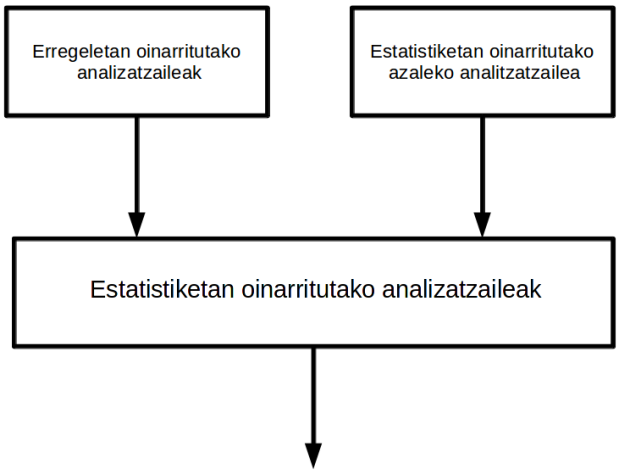
netan gehiago zentratu gara erregeletan eta estatistiketan oinarritutako sistemen hibridazioan, gutxiago landuta dagoen ikuspuntua izateaz gain, beraien erroan guztiz desberdinak diren sistemen konbinazioak analisi sintaktikoari onura handiagoa ekar diezaiokeela uste dugulako.

Kapitulu honen helburu nagusia morfologikoki aberatsak diren hizkuntzen analisi sintaktikoa hibridazioaren bitartez hobetzea da. Kasu honetan, aukeratu dugun hizkuntza euskara izan da, Ixa taldean gehien landu den hizkuntza izanik, baliabide ugari ditugu eskuragarri. Gainera orokorrean analizatzaile sintaktiko estatistikoek ez dituzte oso emaitza onak ematen euskararako beste hizkuntza batzuekin alderatzen badugu. Are gehiago, euskararako analisi sintaktikoan hobekuntzak lortzen badira, sintaxia erabiltzen duten beste tresna batzuen eraginkortasunak ere gora egingo du, horrek euskalgintzarako suposatzen duen onurarekin.

Erabilitako sistemen konbinazioak euskararen analisi sintaktikoan duen eragina sakonki aztertuko dugu atal honetan, ez bakarrik erdietsitako emaitza orokorraren analisia eginda, baita dependentzia esanguratsuenek bakarka erdiesten dituzten emaitzen analisia burututa ere. Arlo honetan aurrera eraman ditugun lanekin bukatzeko, hibridazioak euskararen fenomeno linguistiko konkretu batzuetan duen eragina ere azalduko dugu, era honetara, sistema hibridoaren bertuteak eta ahuleziak fenomenoka identifikatzeko gai izango gara. Ondorioz, etorkizunean, landutako fenomeno konkretu bat hobetzea interesatzen bada, ikuspegi hibridoaren bitartez egiteak merezi duen ala ez erabakitzeko oinarria izango dugu.

4.2 Metodologia

Kapitulu honetan, tresna desberdinen hibridazioa aztertuko dugu euskararako analisi sintaktiko sendoagoa lortzeko asmoarekin. Horretarako, IXA taldean eskuragarri ditugun tresnak, eta sarean atzigarri dauden tresnak konbinatuko dira egingo diren esperimentu desberdinetan. Zehazkiago, dependentzia analizatzaile estatistikoak oinarri izanda, izaera desberdinetako tresnen irteera gehituko zaie pilaketa bidez. Gehitutako informazio hori erregeletan oinarritutako analizatzaile desberdinen emaitza izango da gehienbat, dependentzia analizatzaileak eta azaleko analizatzaileak, baina estatistiketan oinarritutako azaleko analizatzaileen informazioa ere gehituko da (ikus 4.1 irudia).



4.1 irudia – Konbinaketetan erabili diren analisi motak.

Aipatutako konbinaketek euskararen analisi orokorrean duten eragina neurtuko dugu, baita dependentzia erlazio konkretuek duten portaera ere. Gure emaitzen analisia ez da analizatzaileek bakarka erdietsitako emaitzen azterketara soilik mugatuko, analizatzaileen arteko adostasun eta desadostasunak ere aztertuko dira euskararen analisi osoan erdietsitako emaitzetatik abiatuta eta aukeratutako dependentzia erlazioetan ikusitakoarekin jarraituz.

Ondoren, aukeratutako zenbait fenomeno linguistikotan prozedura bera aplikatuko da. Hasiera batean, hibridazioaren eragina galderazko esaldietan aztertuko da eta etorkizunean berdina egingo da koordinazioan eta esaldi konpletiboetan.

4.3 Esperimentazio-ingurunea

Atal honetako esperimentuekin euskararen analisi sintaktikoa hobetzeko erabili daitezkeen tresnen hibridazioa aztertu dugu. Hori dela eta, espresuki euskararako inplementatuta dauden tresnak eta hizkuntza-independenteak diren tresnak erabili dira. Hurrengo lerroetan gainetik azalduko ditugu:

- **EDGK:** 3.2.3.2 puntuan aurkeztu dugun erregeletan oinarritutako dependentzia analizatzaile honek, aldez aurretik definituta dauden dependentzia erlazioei dagozkien etiketak asignatzen dizkie hitzei, non etiketa bakoitza dependentzia erlazioaren izenarekin eta bere buruaren

(gurasoa) norantzarekin (ezker edo eskuin) osatuta dagoen.

- **MaltParser (MaltOptimizer), MST eta Mate:** 3.2.3.1 puntuan sakondu diren analizatzaile estatistiko hauek, gaur egun dependentzietan oinarritutako analisi sintaktikoa burutzeko artearen egoeran aurki daitezkeen aukeren artean hoberenetarikoak direla esan daiteke. Azken urteetan hainbat ataza aurrera eramateko erabili dira modu arrakastatsuan, beraien azkartasuna eta asmatze-tasa altua direlarik hizkuntza gehienetarako izan duten arrakasta horren giltzetariko bat.

- **Ixati**: 3.2.3.2 puntuan azaldu dugun azaleko analizatzaile honek, izensintagma, aditz-kate eta postposizio-sintagmen hasiera eta bukaera markatzen ditu.

- **ML Ixati:** 3.2.3.1.2 puntuan aurkeztu dugun chunker hibrido honek, ikasketa automatikoak eta erregelek eskaintzen dituzten onurak aplikatzen ditu kateen identifikazio automatikoa egiteko, identifikatutako kateak IOB estiloaren arabera etiketatzen dituelarik.

Arlo honetan landu diren esperimentu gehienak SPMRL 2013 (Seddah *et al.* 2013) eta SPMRL 2014 (Seddah *et al.* 2014) ataza partekatuetan eskuragarri izan ditugun eta 3.2.1.1 puntuan azaldu ditugun euskarazko zuhaitz-bankuen gainean jorratu dira. Erabilitako zuhaitz-bankuekin bukatzeko, hibridazioaren eragina neurtzeko aukeratu diren fenomeno linguistikoak biltzen dituzten zuhaitz-bankuak eskuz sortuko ditugu. Aurretik aipatu dugun bezala, orain arte galderazkoak landu dira (ikus 3.2.1.3), eta koordinaziozkoak eta konpletiboak etorkizunerako utzi dira.

Azkenik, analizatzaile sintaktiko hibrido desberdinek bueltatzen dituzten analisiak ebaluatzeko 3.3 atalean azaldu ditugun *Labeled Attachment Score* (LAS) eta F-measure neurriak erabili ditugu, lehenengoa analisi osoaren kalitatea neurtzeko eta bigarrena dependentzia konkretuen emaitzak neurtzeko.

4.4 Erregeletan eta estatistiketan oinarritutako sistemen hibridazioa

Puntu honetan, erregeletan oinarritutako sistemen eta estatistiketan oinarritutakoen hibridazioan zentratu gara. Hibridazioa modu horretan aurrera

eramatea erarik egokiena dela iruditzen zaigu, konbinatutako sistemek informazioa era desberdinean erabiltzen dutelako, eta bakoitzak eskaintzen dituen ezaugarriak beraien artean osagarriak izan daitezkeela uste dugulako. Hori kontuan hartuta, kapitulu honetako esperimentu gehienetan, erregeletan eta estatistiketan oinarritutako sistemen hibridazioa landu dugu. Analisi sintaktikoari dagokionean, berriz, hibridazioak testu orokorretan eta fenomeno zehatzetan duen eragina aztertu eta konparatu nahi da.

Atal hau honela dago antolaturik: hasteko, 4.4.1 puntuan, erregeletan eta estatistiketan oinarritutako sistemen konbinazioaren inguruan aplikatu dugun gure hurbilpenaren nondik-norakoak zehaztuko ditugu, eta 4.4.2 puntuan, jorratu diren esperimentuak sakon azalduko dira, bai orokorrean duen eragina neurtzen duten esperimentuak ikusiz, baita galderazko esaldietan aplikatu ondoren erdietsitako emaitzak komentatuz ere.

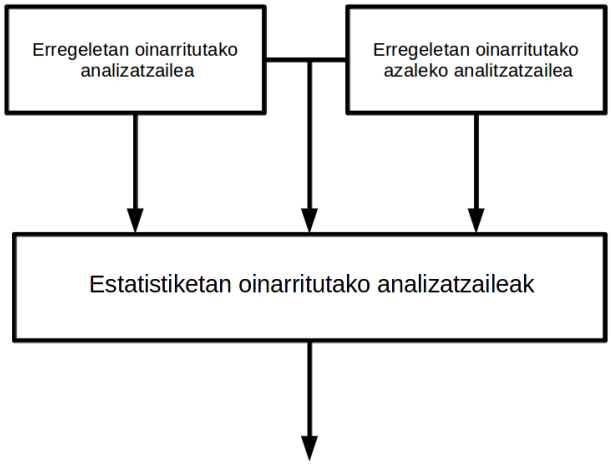
4.4.1 Gure hurbilpena

Atal honetan, erregeletan eta estatistiketan oinarritutako sistemen hibridazioak analisi sintaktikoa hobetzeko eskaintzen dituen aukera desberdinak aztertu ditugu. Alde batetik, landutako hibridazio motek euskararen analisi sintaktiko orokorrean duten eragina neurtu dugu. Bestetik, euskararen fenomeno linguistiko jakin batzuen analisi sintaktikoan duen eragina aztertu dugu, baita dependentzia sintaktiko esanguratsuenek duten portaera ere. Helburu horrekin ondorengo konbinaketak landu dira pilaketaren bidez (ikus 4.2 irudia):

- Erregeletan oinarritutako dependentzia analizatzailea eta estatistiketan oinarritutako analizatzaile desberdinak.
- Erregeletan oinarritutako azaleko analizatzailea eta estatistiketan oinarritutako analizatzaile desberdinak.
- Erregeletan oinarritutako dependentzia analizatzailea, erregeletan oinarritutako azaleko analizatzailea eta estatistiketan oinarritutako analizatzaile desberdinak.

Behin badakigunean zeintzuk esperimentu burutuko ditugun, komeni da jakitea analizatzaile mota bakoitzari, honek emaitza bueltatzeko, pasatu zaion informazioa zein den. Erregeletan oinarritzen diren sistemetatik hasten bagara, erabilitako biek behar dute IXA taldeko analisi kateak bueltatzen duen

informazioa. Lehen urratsa, hitzen analisi morfologikoa egitea da. Euskarazko hitz bakoitzak hainbat atzizki izan ditzake eta hitz-forma bera sortzeko hainbat morfema konbinaketa erabil daitezke. Ondorioz, euskarazko analizatzaile morfologikoa (ikus 3.2.2.1 atala) anbiguotasun handiarekin topatzen da. Honen aurrean, desanbiguazio morfologikoa ezinbestekoa da. Desanbiguatzaile morfologikoaren lana, hitz-forma bakoitzeko interpretazio bakarra bueltatzea da, testuinguruari dagokiona, ahalik eta akats kopuru txikienarekin. Esan beharra dago, ez dela erraza beti interpretazio zuzena bueltatzea, kasu askotan testuinguru lokalen azterketa sakon bat beharrezkoa delako horretarako. Dena den, desanbiguatzailearen lanak berebiziko garrantzia dauka, kategoriaren edo ezaugarri morfologikoen esleipen okerrak analizatzaile sintaktikoetan eragin negatiboa baitauka.



4.2 irudia – Estatistiketan eta erregeletan oinarritutako hibridazioan erabili diren analisi motak.

Hitz-formen informazio guztia desanbiguatuta dugunean, Ixatik informazio hori erabiliko du izen-sintagma, postposizio eta aditz-kateak markatzeko. EDGKk, aurretik jasotako informazio guztia erabiltzen du hitz bakoitzari dependentzia etiketa bat esleitzeko, baita Ixatiren emaitza ere. Bukatzeko, analizatzaile estatistikoek jasoko duten informazioa, berriz, analizatzaile morfologikoak sortu eta desanbiguatzaile morfologikoak aukeratzen dituen kategoria eta ezaugarri morfosintaktikoak izango dira, landu nahi den esperimentu bakoitzean gehituko zaien informazioarekin batera.

4.4.2 Esperimentuak eta emaitzak

Aurretik esan bezala, atal honetan erregeletan eta estatistiketan oinarritutako tresna edo sistemen konbinaketa landu da, 4.4.2.1 puntuan hibridazio mota horrek euskararen analisi sintaktikoan duen eragina aztertuko delarik, 4.4.2.2 puntuan hibridazioak galderazko esaldietan duen eragina neurtuko den bitartean.

4.4.2.1 Fenomeno orokorren tratamendua

Izaera desberdinak dituzten analizatzaileen konbinazioaren bidez euskara bere orokortasunean tratatzeko aurrera eraman diren esperimentuak eta erdietsitako emaitzak azalduko dira atal honetan. Alde batetik, analizatzaile estatistikoak ditugu. Bestetik, erregeletan oinarritutako dependentzia analizatzailea eta azaleko analizatzailea. Gure helburua mota bakoitzeko analizatzaileak beste motako analizatzaileekin konbinatzea da. Esanak esan, 4 dependentzia analizatzaile estatistiko erabili dira: MaltParser, MaltOptimizer (MaltParserren hobekuntza bat), MST eta Mate. Erregeletan oinarritutako analizatzaileak, aldiz, 2 erabili dira: EDGK dependentzia analizatzailea eta Ixati azaleko analizatzailea. Aipatzekoa da hasiera batean eskuragarri izan dugun euskarazko zuhaitz-bankuaren gainean egin direla esperimentuak momentuko analizatzaile estatistiko eraginkorrenak erabilita, hots, MaltParser eta MST erabilita. Hala ere, denbora aurrera joan ahala, euskarazko zuhaitz-bankuaren tamaina handitzea lortu da eta analizatzaile estatistikoak hobetzen joan dira. Horrenbestez, MaltParser eta MST erabilita egindako esperimentuak zuhaitz-banku txikiagoaren gainean daude landuta eta erdietsitako emaitzak aurkezteko arrazoi nagusia hibridazioak tamaina desberdinetako zuhaitz-bankuen analisi sintaktikoan duen eragina aztertzea da.

Ildo beretik jarraituz, egindako esperimentuetan, aipatutako 4 analizatzaile estatistikoetako bakoitzari ondorengo informazioa gehitu zaio pilaketa bidez: EDGKren irteera, Ixatiren irteera, eta EDGK eta Ixatiren irteera batera. Kasu bakoitzean erdietsitako emaitzak aurkeztu baino lehen, EDGKren eta Ixatiren irteerak zeintzuk diren gogoraraztea komenigarria dela uste dugu. EDGKk, esaldi bateko hitzei dependentzia etiketak esleitzen dizkie zuhaitz sintaktikoan beraien gurasoa zein noranzkotan dagoen adierazten duen ikurrarekin batera ($<$ ezkerretara adierazteko eta $>$ eskuinetara adierazteko). Bestalde, Ixatik, esaldi bateko kateak identifikatzen ditu: izensintagma hasiera, bukaera, aditz-kate hasiera, bukaera... Aurrerago kontu-

ratuko garen bezala, EDGKk eta Ixatik ez dituzte hitz guztiak etiketatzen. Azaldutakoa biribiltzeko, 4.3 irudian bi analizatzaileek esaldi bat nola etiketatu duten ikusten da eta 4.4 irudian, berriz, erregeletan oinarritutakoen analizatzaileen irteerak parserrei nola pasatu zaizkien.

```
"<Noiz>"<HAS_MAI>"
        "noiz" ADB GAL ZERO HAS_MAI w1,L-A-ADB-GAL-8,lsfi1 @ADLG %SINT &NCMOD>
"<aipatu>"
        "aipatu" ADI SIN PART BURU NOTDEK w2,L-A-ADI-SIN-228,lsfi2 @-JADNAG %ADIKATHAS &ADITZ_NAGUSI
"<zituen>"
        "*edun" ADL ZHG B1 NOR_NORK NR_HAIEK NK_HARK w3,L-A-ADL-58,lsfi5 @+JADLAG_MP_OBJ %ADIKATBU
"<Andre>"<HAS_MAI>"
        "andre" IZE ARR BIZ+ ENTI_PER HAS_MAI w4,L-A-IZE-ARR-765,lsfi7 @KM> %SIH
"<Thevet>"<HAS_MAI>"
        EZEZAG "Thevet" IZE IZB PLU- ENTI_PER HAS_MAI w5,L-G-IZE-IZB-89,lsfi8 @KM> &<NCSUBJ
"<kosmografoak>"
        EZEZAG "kosmografo" ADJ ARR ERG NUMS MUGM w6,L-G-ADJ-ARR-248,lsfi9 @SUBJ %SIB &<NCMOD
"<pinguinoak>"
        "pinguino" IZE ARR ABS NUMP MUGM w7,L-A-IZE-ARR-771,lsfi10 @OBJ %SINT &<NCOBJ
"<edo>"
        "edo" LOT JNT HAUT w8,L-A-LOT-JNT-5,lsfi13 @PJ
"<hegazti>"
        "hegazti" IZE ARR BIZ+ ZERO w9,L-A-IZE-ARR-773,lsfi14 @KM> %SIH &NCMOD>
"<zuri-beltzak>"
        "zuri-beltz" IZE ARR ABS NUMP MUGM w10,L-A-IZE-ARR-775,lsfi15 @OBJ %SIB &<NCOBJ
"<$?>"<PUNT_GALD>"
        PUNT_GALD
```

4.3 irudia – Esaldi bat EDGK eta Ixatiren bidez etiketatuta. Urdinez eta % hasierarekin Ixatiren etiketak eta berdez eta & hasierarekin EDGKrenak.

```
1   Angolan  Angola  IZE   IZE_LIB     KAS=INE|NUM=S|IXATI=SINT|EDGK=&NCMOD>                    4  ncmod
2   48       48      DET   DET_DZH     KAS=ZERO|IXATI=SIH|EDGK=&DETMOD>                        3  detmod
3   lagun    lagun   IZE   IZE_ARR     KAS=ABS|IXATI=SIB|EDGK=NULL                             4  ncsubj
4   zendu    zendu   ADI   ADI_SIN     ADM=PART|ASP=BURU|IXATI=ADIKATHAS|EDGK=&ADITZ_NAGUSI    0  ROOT
5   dira     izan    ADL   ADL         MDN=A1|DADUDIO=NOR|NOR=HAIEK|IXATI=ADIKATBU|EDGK=&<AUXMOD  4  auxmod
6   hegazkin hegazkin       IZE IZE_ARR KAS=ZERO|IXATI=SIH|EDGK=NULL                            7  ncmod
7   istripu  istripu IZE IZE_ARR        KAS=ZERO|IXATI=NULL|EDGK=NULL                           9  ncmod
8   baten    bat     DET   DET_DZH     KAS=GEN|IXATI=SIB|EDGK=NULL                             7  detmod
9   eraginez eragin  ADI   ADI_SIN     KAS=INS|ADM=PART|NUM=P|IXATI=ADIKAT|EDGK=NULL           4  ncmod
10  .        .       PUNT  PUNT_PUNT   IXATI=NULL|EDGK=NULL                                    9  PUNC
```

4.4 irudia – Analizatzaile estatistikoek jasoko duten esaldi baten adibidea EDGK eta Ixatiren irteeren informazioa gehituta (urdinez).

Egindako esperimentuetan jasotako emaitzak 4.1 taulan bildu ditugu, oinarrizko konfigurazioekin[1] jasotako emaitzekin batera. Emaitza horiek aztertuta ez da erraza ondorio garbirik ateratzea. Ematen duenez, trantsizioetan oinarritutako parserrek (MaltParser eta Maltoptimizer) hobeto aprobetxatzen dituzte jasotako ezaugarri berriak. Bestalde, esan daiteke, ikasketarako

[1]Balio lehenetsiekin, ezer gehitu edo kendu gabe.

	MaltParser	MST	MaltOptimizer	Mate
	train 84.000 hitz test 12.625 hitz	**train 84.000 hitz test 12.625 hitz**	**train 96.368 test 13.851**	**train 96.368 test 13.851**
Oinarrizkoak	76,77	77,96	80,04	83,00
+Ixati	77,10 (+ 0,33)	77,99 (+ 0,03)	79,99 (- 0,05)	82,26 (- 0,74)
+EDGK	77,15 (+ 0,38)	78,03 (+ 0,07)	80,15 (+ 0,11)	82,34 (- 0,66)
+Ixati+EDGK	77,25 (+ 0,48)	78,00 (+ 0,04)	80,11 (+ 0,07)	81,45 (- 1,55)

4.1 taula – Tresna desberdinen hibridazioaren emaitza *Labeled Attachment Score* (LAS) neurrien arabera.

corpusa handia denean, parserrak ez duela behar sartu diogun informazioa, beste informazioa erabilita gai baita pareko emaitzak edo hobeak erdiesteko. Ondorioz, uste dugu zenbat eta ikasketa corpus txikiagoa erabili orduan eta probetxu gehiago aterako zaiola sartutako informazioari. Hala ere, esandakoa hipotesi bat besterik ez da eta aztertzeke gelditzen da hibridazioa tamaina txikiagoko ikasketa corpus desberdinetan aplikatzeak duen eragina.

Ildo beretik jarraituz, erabilitako tresna desberdinen konbinaketarekin lortutako hobekuntzak ez dira espero diren bezain onak. Hala ere, emaitza orokor hauek modu zehatzagoan aztertzeko asmoarekin eta hibridazioak dependentzia erlazio konkretuetan duen eragina zein den jakiteko, dependentzia erlazio esanguratsuenek izandako emaitzak aztertu ditugu 4.2 taulan. Ikasketa corpus txikiagoa erabili den esperimentuak aztertzen badira, MaltParserrekin lortutako dependentzia konkretuen emaitzek hobekuntza handiagoak jasan dituzte MSTrekin erdietsitakoek baino. Esan daiteke, aukeratu ditugun dependentzia erlazioetan eragin handia duela hibridazioak, kasu gehienetan hobekuntzak gertatu baitira. Are gehiago, kasu askotan gertatzen dira 1,5 eta 2 puntuko hobekuntzak.

Emaitzetan sakonduz, MaltParser eta EDGKren arteko konbinaketa objektua (+2,33), subjektua (+2,04) eta zehar objektua (+1,40) bezalako dependentzia erlazioak ebazteko oso onuragarria dela nabaritzen da. Predikatiboarengan hasiera batean ez du ematen hibridazioak eragin positibo nabarmenik duenik, baina Ixati, EDGK eta MaltParser konbinatzen direnean 2,13 puntuko igoera gertatzen da. MaltParser erabilita, *ncmod* dependentzia erlazioa da hobekuntza txikienak jasan dituena (+1,11), igoera hori, MaltParser, EDGK eta Ixati konbinatuta erdietsi da. Maiztasun handieneko erlazioetako bat da *ncmod* eta hura ahalik eta gehien hobetzea komeni da, emaitza orokorrean eragin handia baitu dependentzia honek.

	MaltParser				MST Parser			
Dep erlazioa	**Oinarria**	**+Ixati**	**+EDGK**	**+Ixati +EDGK**	**Oinarria**	**+Ixati**	**+EDGK**	**+Ixati +EDGK**
ncmod	75,29	75,90	76,08	**76,40**	77,15	**77,44**	76,39	76,92
ncobj	67,34	68,49	**69,67**	69,54	64,85	64,86	65,56	**66,18**
ncpred	61,37	61,92	61,26	**63,50**	**60,37**	57,55	58,44	59,27
ncsubj	61,92	61,90	**63,96**	63,91	59,19	59,26	**62,23**	61,61
nczobj	75,76	76,53	**77,16**	76,29	74,23	**74,47**	72,16	69,08
	MaltOptimizer				**Mate**			
Dep erlazioa	**Oinarria**	**+Ixati**	**+EDGK**	**+Ixati +EDGK**	**Oinarria**	**+Ixati**	**+EDGK**	**+Ixati +EDGK**
ncmod	76,62	76,85	**77,21**	76,63	**80,61**	80,08	79,54	79,06
ncobj	74,49	**74,55**	73,90	73,91	**76,53**	76,13	75,24	74,66
ncpred	58,10	58,28	58,29	**59,70**	**63,42**	58,19	61,01	59,87
ncsubj	66,30	65,67	**67,47**	**67,47**	69,36	67,93	**69,83**	65,15
nczobj	76,03	**77,42**	76,53	76,33	79,72	**79,86**	77,69	74,82

4.2 taula – Tresna desberdinen hibridazioaren emaitza dependentzia erlazio bakoitzerako F-measure neurriaren arabera. Beltzez dependentzia bakoitzerako analizatzaile bakoitzak lorturiko emaitzarik onena.

MSTrekin lortutako emaitzak aztertzen badira, aurretik esan bezala, pilaketa bidez egindako hibridazioak eragin gutxiago du analizatzaile sintaktiko honekin erdietsitako dependentzia erlazioetan. Hala ere, esaldiko subjektuarengan (*ncsubj*) eta objektuarengan (*ncobj*) bai ikusten da eragin positiboa duela MST, Ixati eta EDGK konbinatzen direnean, +1,33 eta +2,42eko igoera, hurrenez hurren. MSTren kasuan ere, MaltParserren kasuan ikusi den bezala, hibridaziorako onuragarriago da EDGK erabiltzea Ixati baino. Ematen du, EDGKn landu diren erregelen bitartez, analizatzaile estatistikoek identifikatu ezin dituzten fenomeno linguistiko batzuk errazago identifikatzen direla. Hala ere, ez da ahaztu behar EDGKrekin *ncmod* erlazioan erdietsitako emaitzak oinarrizko sistemarekin lortutakoak baino okerragoak direla eta erlazio horren identifikazioa okertzea ez dela komenigarria.

Corpus handiagoarekin ikasi duten bi analizatzaileek erdietsitako emaitzak aztertzen badira, MaltOptimizerrek hibridazioari nolabaiteko etekina ateratzen diola ikusten da aukeratutako dependentzia erlazioetarako. Izan ere, beti baitago hibridazioarekin erdietsitako emaitzen bat oinarrizko emaitzaren gainetik. Mate analizatzailearekin, ostera, oso zaila da oinarrizko emaitzak gainditzea eta gainditzen direnean, hobekuntza oso txikia da.

Hori ikusita, emaitza orokorrak gehiago ez hobetzearen arrazoia bilatzeko beste esperimentu bat burutu dugu. Esperimentu horretan, EDGKko urre-patroi etiketak gehitu zaizkio pilaketa bidez MaltParser analizatzaileari,

EDGK eta MaltParserren hibridazioaren bidez noraino iritsi gaitezkeen jakiteko, goi-muga finkatzeko. Espero bezala, hobekuntza nabarmena lortzen da, % 95eko *Labeled Attachment Score* (LAS). Dena den, dependentzia erlazio desberdinek jaso dituzten emaitzak aztertu ditugunean, EDGKko urrepatroiko etiketak dependentzia erlazio batzuetarako onuragarriak direla ikusi dugu, baina kaltegarriak beste batzuetarako. Adibidez, *ncmod* erlazioak 3,25 puntuko gorakada izan du F-measurean, *ccomp_zobj* erlazioak 0,46 puntuko galera izan duen bitartean. Gertaera hori oso bitxia da, emaitzak okertu baitira. Hortik etiketa automatikoak erabiltzean lortutako emaitza orokorraren hobekuntza txikia, nahiz eta maiztasun handiko erlazioetan, ikusi bezala, igoera nabarmena izan. Hori jakinda, ez da harritzekoa erlazio askotan emaitzak okertzea EDGKren etiketa automatikoak eta analizatzaile estatistikoak konbinatzen direnean.

Arrazoi horiengatik guztiengatik, ematen du analizatzaile estatistikoei pasatako dependentzia etiketen eta erdietsitako emaitzen arteko erlazioa konplexua dela, eta uste dugu sakonago aztertu behar direla beraien izaera zehazteko. Erabilitako informazio sintaktiko mota bakoitzak dependentzia erlazio konkretuetan eragin handia izan dezake eta beraien azterketa analizatzaile sintaktikoen konbinaziorako garrantzitsua izan daiteke.

Atal honekin bukatzeko, erregeletan oinarritutako analizatzailetik eratorritako emaitzak parserrari zein kasutan lagundu diezaioken ikusteko, EDGKren eta bi parser berrienen arteko konparaketa egin dugu ikasketarako zuhaitz-banku handiena erabiliz (96.368 hitz). Esperimentu horretan, 13.851 hitzeko zuhaitz-bankuan EDGKk, MaltOptimizerrek eta Matek lortu dituzten emaitzetan oinarritu gara. Azterketarako 7 kasu definitu ditugu: 1) analizatzaile bakoitzak etiketatu dituen dependentzia kopurua, 2) zenbat kasutan etiketatu duten ondo uneko parserrak eta EDGKk, 3) zenbat kasutan etiketatu duten gaizki, 4) uneko parserrak zenbat kasutan huts egin duen EDGKk ondo etiketatu duenean, 5) uneko parserrak zenbat kasutan asmatu duen EDGKk gaizki etiketatu duenean, 6) zenbat kasutan asmatu duen parserrak EDGKk emaitzik ez duenean bueltatu eta 7) zenbat kasutan huts duen parserrak EDGKk emaitzik ez duenean bueltatu. Konparaketaren emaitzak 4.3 taulan bildu ditugu. Taula horretan ikusten denaren arabera, EDGKk, 13.851 hitzetik 7.086 kasutan soilik bueltatzen du dependentzia etiketa, hots, kasuen % 51a etiketatzen du, parserrek % 100 etiketatzen duten bitartean. Taulako gainontzeko kasuekin jarraitzen badugu, parserrek % 80ko gutxieneko asmatze-tasa dutela ikusten da EDGKk % 14a soilik duenean (986 kasu ondo 7.086tik).

		MaltOptimizer	Mate
Etiketatze kopurua		13.851 (% 100)	13.851 (% 100)
EDGK ondo	**Parserra ondo**	868	893
	Parserra gaizki	118	93
EDGK gaizki	**Parserra ondo**	5.101	5.189
	Parserra gaizki	999	911
EDGK emaitza hutsa	**Parserra ondo**	5.675	5.833
	Parserra gaizki	1.090	932

4.3 taula – Parserren eta EDGKren arteko konparaketa.

Emaitza horien aurrean, zenbait puntu kontuan hartzekoak direla uste dugu. Hasteko, erregeletan oinarritutako analizatzaileak ez du estaldura osoa eta, egoera honetan, oso zaila da analizatzaile estatistikoei laguntzea. EDGKk eta uneko parserrak ondo egiten duten kasuetan, 868 eta 893 hurrenez hurren, ezin zaio lagundu parserrari, berak jada ondo egiten baitu. Biek gaizki egiten duten kasuetan ere ezin dio EDGKk lagundu, honek ere gaizki egiten duelako. Erregeletan oinarritutako analizatzaileak lagundu dezakeen kasu bakarra, berak ondo eta uneko parserrak gaizki egiten duenean da. Zoritxarrez, kasu guztien % 1ean baino gutxiagotan eta EDGKk emaitza bueltatzen duen kasuen % 1ean gertatzen da hori, gehienez 118 alditan bakarrik.

Hori dela eta, EDGKren estaldura hobetzea ezinbestekoa dela deritzogu, analizatzaile estatistikoari gaizki ematen zaizkion fenomenoak lantzea edo biek gaizki etiketatzen dituzten kasuak hobeto lantzea.

4.4.2.2 Fenomeno konkretuen tratamendua

Erregeletan oinarritutako analizatzaileen eta estatistiketan oinarritutako analizatzaileen hibridazioak euskararen zenbait fenomeno linguistikotan duen eragina neurtzeko asmoarekin, aurreko atalean bezala, erregeletan oinarritutako dependentzia analizatzailearen emaitza pilaketa bidez gehitu zaie parserrei. Hiru fenomeno aztertzea pentsatu dugu: galderazko esaldiak, koordinazioa eta esaldi konpletiboak. Tesi-lan honetan, bere osotasunean lehenengoa bakarrik landu ahal izan dugu eta beste biak bideratu egin ditugu.

Galderazko esaldien tratamendua

Hibridazioak galderazko esaldietan duen eragina zehazteko bi zuhaitz-banku erabili dira. Ikasketarako 96.368 hitzeko zuhaitz-bankua (ikus 3.2.1.1) eta ebaluaziorako 803 hitzeko (102 galderazko esaldi) zuhaitz-bankua (ikus 3.2.1.3). Bi corpus horiek EDGKrekin analizatu dira eta hitz bakoitzerako lortutako emaitza MaltOptimizer eta Mate parserrei gehitu zaie orain arte egin den bezala. Hibridazioarekin galderazko esaldietan erdietsitako emaitzak 4.4 taulan aurkeztu dira. Emaitzetan ikusten den bezala, MaltOptimizerrek, fenomeno orokorren tratamenduan gertatu den bezala, probetxu handiagoa ateratzen dio hibridazioari Mate analizatzaileak baino (+ 0,52). Hala ere, ez da ahaztu behar Maten oinarrizko emaitza hobea dela (+ 1,74) eta hobekuntza tarte txikiagoa duela.

Kasua	**Emaitza**
MaltOptimizer	81,82
MaltOptimizer + EDGK	83,56 (+ 1,74)
Mate	83,56
Mate + EDGK	84,68 (+ 1,12)

4.4 taula – EDGK, galderazko esaldien gainean MaltOptimizer eta Materekin konbinatu ondoren lortutako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

Bestalde, ikus daiteke sistema hibridoek gutxienez 1,12 puntuko aldea ateratzen dietela analizatzaile estatistiko puruei, alde hori 1,74raino doalarik MaltOptimizer eta EDGKk lortutako emaitza MaltOptimizerrek erdietsitakoarekin konparatzen dugunean. Lortutako hobekuntzak McNemar testaren arabera ($p < 0,05$) esanguratsuak ez diren arren, uste dugu emaitzak eskalagarriak izan daitezkeela eta zuhaitz-banku handiagoetan emaitzak mantendu daitezkeela galderen egitura zurrunagoa dela medio.

Hibridazioak galderazko esaldietan duen eragina sakonago aztertzeko, dependentzia erlazio esanguratsuenetan zer gertatzen den aztertuko dugu (ikus 4.5 taula). Bertako emaitzen arabera, EDGKk esleitutako etiketak lagungarriak dira *ncmod* eta *ncobj* dependentziak asmatzeko, erabilitako bi analizatzaile estatistikoetan ikusten baitira emaitzetan hobekuntzak. Aztertu ditugun beste hiru dependentzietan, berriz, sistemaren portaera guztiz desberdina da erabili den analizatzaile estatistikoaren arabera. MaltOptimizer eta EDGK konbinatzen ditugunean, *ncpred* dependentzia erlazioan 21,50

		MaltOptimizer		Mate	
Dep erlazioa	**Kasu kopurua**	**Oinarria**	**+EDGK**	**Oinarria**	**+EDGK**
ncmod	227	75,43	**77,09 (+ 1,66)**	79,91	**80,87 (+ 0,96)**
ncobj	70	74,84	**82,19 (+ 7,35)**	76,31	**77,63 (+ 1,32)**
ncpred	14	**66,66**	45,16 (- 21,50)	56,25	**68,29 (+ 12,04)**
ncsubj	91	67,48	**72,72 (+ 5,24)**	**70,73**	67,49 (- 3,24)
nczobj	5	60,00	**72,72 (+ 12,72)**	**72,72**	45,00 (- 27,72)

4.5 taula – Sistema desberdinen hibridazioaren emaitzak dependentzia erlazio bakoitzerako F-measure neurriaren arabera.

puntuko galera dagoela ikusten da. Hala ere, mota horretako 14 kasu bakarrik egonda, galera ez da inolaz ere hasiera batean uste bezain handia. Gauza bera gertatzen da *nczobj* erlazioarekin, hots, mota horretako 5 kasu soilik izanda, 12,72 puntuko hobekuntza ez da guk nahi bezain handia. *ncsubj* erlazioan irabazitako 5,24 puntuak garrantzitsuagoak direla deritzogu, kasu horretako erlazio kopurua 91 baita.

Bestalde, Mate eta EDGK konbinatzen ditugunean, aurreko kasuan gertatzen denaren kontrakoa gertatzen da *ncpred* erlazioan hobekuntza dago eta *ncsubj* eta *nczobj* erlazioetan galera. Dena den, Mate eta EDGK konbinatzen ditugunean hiru erlazio horiekin gertatzen dena okerragoa da MaltOptimizerren gertatzen den baino. Izan ere, *ncsubj* erlazioko kasuen okertzeak eragin negatibo handiagoa baitu emaitza orokorrean beste bienak baino.

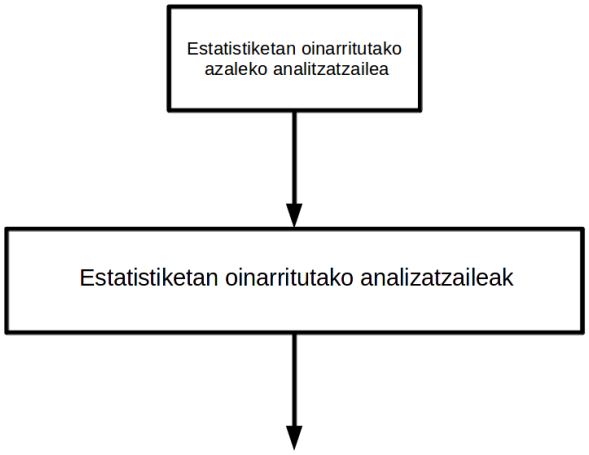
Atal honetan aurkeztu ditugun emaitza guztiek, etorkizunerako pista ona ematen digute hibridazioak galderazko esaldien analisi sintaktikoa hobetu dezakeen ala ez zehazteko. Ebaluatutako laginean hobekuntzak ikusten dira. Halarik ere, kontuan hartu behar da lagin hori nahiko txikia dela (102 esaldi eta 803 hitz) eta emaitza horien balioa behin-betiko finkatzeko, esperimentuak lagin handiago baten gainean errepikatzea ezinbestekoa dela, ataza hori etorkizunerako utzi dugularik.

Koordinazio eta esaldi konpletiboen tratamendua

Koordinazioa eta esaldi konpletiboen tratamendua aurrera eramateko lehen urratsak eman dira. Zuhaitz-bankutik mota bakoitzerako zoriz aukeratutako laginak erauzi dira eta hizkuntzalarien eskuetan utzi da fenomeno horiek landu edo hobetzeko EDGKko erregelak lantzea. Ondoren, galderazkoetan erabilitako metodologia berdina aplikatuko da beraiekin.

4.5 Estatistiketan oinarritutako sistemen hibridazioa

Aurreko puntuan estatistiketan eta erregeletan oinarritutako sistemak konbinatu ditugu euskararen analisi sintaktikoa hobetzeko asmoarekin. Hala ere, sistema desberdinen hibridazioa modu batera baino gehiagotara landu daiteke. Hori kontuan hartuta, atal honetan estatistiketan oinarritzen diren analizatzaile sintaktikoak konbinatzea erabaki dugu (ikus 4.5 irudia). Hori dela eta, atal honetan estatistiketan oinarritutako sistemen hibridazioak euskarako testu orokorren prozesamenduan duen eragina neurtu dugu, fenomeno linguistiko konkretuetan izan dezakeen eragina alde batera utzita.



4.5 irudia – Estatistiketan oinarritutako analizatzaileen hibridazioan erabili diren analizatzaileak.

Euskararen analisi sintaktikoa hobetzeko, estatistiketan oinarritutako azaleko analizatzaile sintaktiko bat (chunker bat) eta estatistiketan oinarritutako analizatzaile desberdinak pilaketaren bidez konbinatuko ditugu eta, aurreko puntuan egin den bezala, dependentzia sintaktiko esanguratsuenek duten portaera aztertuko dugu. Ondorengo atalean burututako esperimentuen berri emango dugu zehaztasun osoarekin.

4.5.1 Esperimentuak eta emaitzak

Puntu honetan, trantsizioetan oinarritzen den analizatzaile estatistiko bati eta grafoetan oinarritzen den beste analizatzaile estatistiko bati MLIxati chunkerraren emaitza gehitu nahi izan diegu pilaketa eskema jarraituz. Egin diren esperimentuetan, zehazki MaltOptimizer eta Mate analizatzaileei gehitu zaie informazio hori.

Komeni da gogoratzea analisi sintaktikoa aurrera eramateko aukeratu ditugun bi parserrek zein informazio erabiltzen duten, chunkerraren etiketez gain. Biek analisi katetik lortutako informazioa jasotzen dute formatu egokian (CoNLL 2007 eta CoNLL 2009), zehazkiago hitz-forma, lema, kategoria, azpikategoria eta ezaugarri morfologikoak (kasua, numeroa, aditzaren aspektua,...).

Bestalde, MLIxati chunkerrak (ikus 3.2.3.1.2 puntua), hitz-forma, lema, kategoria, azpikategoria, ezaugarri morfologikoak, funtzio sintaktikoa (batzuetan guztiz desanbiguatu gabe *istiluak* hitzaren kasua) eta Ixati azaleko analizatzaile sintaktikoaren irteera (sintagmen hasiera eta bukaeren informazioa) erabiltzen ditu (ikus 4.6 irudia).

```
"<Joan>"<HAS_MAI>"
        "joan" ADI SIN PART BURU NOTDEK HAS_MAI w1,L-A-ADI-SIN-2920,lsfi1 @-JADNAG %ADIKATHAS
"<den>"
        "izan" ADL ZHG A1 NOR NR_HURA w2,L-A-ADL-480,lsfi3 @+JADLAG_MP_OBJ %ADIKATBU
"<irailaren>"
        "irail" IZE ARR BIZ- GEN NUMS MUGM ZERO w3,L-A-IZE-ARR-7091,lsfi6 @IZLG> %SIH
"<28an>"<ZEN_DEK>"
        "28" IZE ZKI INE NUMS MUGM ZEN_DEK w4,L-A-IZE-ZKI-342,lsfi7 @ADLG %SIB
"<hasi>"
        "hasi" ADI SIN PART BURU NOTDEK w5,L-A-ADI-SIN-2924,lsfi8 @-JADNAG %ADIKATHAS
"<ziren>"
        "izan" ADL B1 NOR NR_HAIEK w6,L-A-ADL-482,lsfi9 @+JADLAG %ADIKATBU
        "izan" ADL MOS B1 NOR NR_HAIEK w6,L-A-ADL-485,lsfi12 @+JADLAG_MP_ADLG %ADIKATBU
"<istiluak>"
        "istilu" IZE ARR BIZ- ABS NUMP MUGM w7,L-A-IZE-ARR-7097,lsfi13 @OBJ %SINT
        "istilu" IZE ARR BIZ- ABS NUMP MUGM w7,L-A-IZE-ARR-7097,lsfi15 @SUBJ %SINT
```

4.6 irudia – MLIxatik ikasketarako jasotzen duen informazio linguistikoaren adibidea MG formatuan. Urdinez Ixati analizatzaile sintaktikoaren irteera.

MLIxatiren emaitza beste bi analizatzaileei gehitu diegu ezaugarri morfologikoen zutabean, 4.7 irudian ikus daitekeen bezala. Informazio hori erabilita egindako esperimentuen emaitzak 4.6 taulan agertzen dira oinarrizko analizatzaile estatistikoen emaitzekin batera, konparaketa egin ahal izateko.

```
1   Angolan   Angola   IZE   IZE_LIB   KAS=INE|NUM=S|CHUNK=B-NP                  4   ncmod
2   48        48       DET   DET_DZH   KAS=ZERO|CHUNK=B-NP                      3   detmod
3   lagun     lagun    IZE   IZE_ARR   KAS=ABS|CHUNK=I-NP                       4   ncsubj
4   zendu     zendu    ADI   ADI_SIN   ADM=PART|ASP=BURU|CHUNK=B-VP             0   ROOT
5   dira      izan     ADL   ADL       MDN=A1|DADUDIO=NOR|NOR=HAIEN|CHUNK=I-VP  4   auxmod
6   hegazkin  hegazkin IZE   IZE_ARR   KAS=ZERO|CHUNK=B-NP                      7   ncmod
7   istripu   istripu  IZE   IZE_ARR   KAS=ZERO|CHUNK=I-NP                      9   ncmod
8   baten     bat      DET   DET_DZH   KAS=GEN|CHUNK=I-NP                       7   detmod
9   eraginez  eragin   ADI   ADI_SIN   KAS=INS|ADM=PART|NUM=P|CHUNK=I-NP       4   ncmod
10  .         .        PUNT  PUNT_PUNT CHUNK=O                                  9   PUNC
```

4.7 irudia – MLIxatiren irteera ezaugarri morfologikoen zutabean gehituta parserrek erabil dezaten.

Kasua	Emaitza
MaltOptimizer	80,04
MaltOptimizer + MLIxati	79,50 (- 0,50)
Mate	83,00
Mate + MLIxati	83,15 (+ 0,15)

4.6 taula – MLIxati, MaltOptimizer eta Materekin konbinatu ondoren lortutako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

Taulako emaitzak aztertzen baditugu, chunkerrak analizatzaile sintaktiko estatistikoei ez diela laguntzen konturatzen gara, MaltOptimizerrekin konbinatu den kasuan bere ekarpena kaltegarria delarik analisi sintaktikorako. Egoera honen aurrean, aurreko kasuetan egin den antzera, hibridazio mota honek dependentzia erlazio esanguratsuenetan duen eragina ere neurtu nahi izan dugu, posible delako orokorrean hobekuntzarik ez ekartzea, baina zenbait dependentzia erlazio ebazteko lagungarria izatea. Horrenbestez, chunkerraren eta parserren konbinaketek dependentzia konkretuetan erdietsitako emaitzak 4.7 taulan bildu ditugu, modu errazean alderatu ahal izateko emaitza guztiak.

	MaltParser (MaltOptimizer)		Mate	
Dep erlazioa	**Oinarria**	**+MLIxati**	**Oinarria**	**+MLIxati**
ncmod	**76,62**	76,45 (- 0,17)	80,61	**81,36** (+ 0,75)
ncobj	**74,49**	73,73 (- 0,76)	**76,53**	76,48 (- 0,05)
ncpred	**58,10**	57,80 (- 0,30)	**63,42**	63,00 (- 0,42)
ncsubj	**66,30**	65,51 (- 0,79)	**69,36**	69,29 (- 0,07)
nczobj	**76,03**	75,79 (- 0,24)	**79,72**	79,46 (- 0,26)

4.7 taula – Sistema desberdinen konbinaketen emaitzak dependentzia erlazio bakoitzerako F-measure neurriaren arabera. Beltzez kasu bakoitzeko emaitzarik onena.

Chunker estatistikoa eta analizatzaile sintaktiko estatistikoak konbinatuta euskararen analisi sintaktikoan lortutako emaitzak aztertuta ez da harritzekoa konbinaketa mota horrek dependentzia erlazio konkretuetan ere hobekuntza handirik ez lortzea. Emaitzek behera egiten dute kasu guztietan MaltOptimizer eta MLIxati konbinatzen direnean, eta ia gauza bera gertatzen da chunkerra Materekin konbinatzen dugunean. Igoera positibo bakarra *ncmod* dependentziarekin bakarrik ikusten den arren, kontuan hartzekoa izan daiteke, dependentzia hori oso garrantzitsua delako, gehien agertzen den dependentzia baita. Hori dela eta, *ncmod* dependentzian lortzen diren hobekuntzak beti dira garrantzitsuak, analisi sintaktiko orokorrean duen eraginagatik. Esandakoa argi ikusten da 4.6 eta 4.7 tauletan, dependentzia erlazio gehienetan emaitza negatiboak lortu arren, *ncmod* dependentzia erlazioaren hobekuntzarekin oinarrizko analisi sintaktiko orokorra hobetzeko (+ 0,15) gai izan baita sistema hibridoa.

4.6 Ondorioak

Kapitulu honetan ikuspegi hibridoari heldu nahi izan diogu, alde batetik, erregeletan oinarritutako sistemak eta estatistiketan oinarritutakoak konbinatu ditugu eta, bestetik, estatistiketan oinarritutako tresna desberdinen hibridazioa landu dugu, beti ere euskararen analisi sintaktikoa hobetzeko helburuarekin. Horretarako, eskuragarri izan ditugun baliabideak erabili ditugu, gure esperimentuen emaitzak errepikagarriak izan daitezen eta antzeko esperimentuak modu errazean burutu ahal izateko. Ikuspegi hibridoa euskararen analisi sintaktikoa hobetzeko jorratu dugun arren, egindako esperimentuetatik jasotako ezagutza euskararekin konparagarriak diren hizkuntza

morfologikoki aberatsetan aplikagarria izatea izan da gure helburuetako bat, hizkuntza horietan ere guk erdietsitako emaitzetatik oso urrun ez dauden emaitzak jasotzea espero delarik.

Landu ditugun hurbilpenak bi izan dira: erregeletan oinarritutako sistemen eta estatistiketan oinarritutako sistemen arteko hibridazioa, eta estatistiketan oinarritutako sistemen arteko hibridazioa. Horietako bakoitzean ikusitako emaitzak aztertuta hurrengo ondorioak atera ditugu hurbilpen bakoitzerako:

- **Erregeletan eta estatistiketan oinarritutako sistemen hibridazioa:** Hibridazio mota honek euskararen analisi sintaktikoan duen eragina zehaztu dugu egindako esperimentuetan oinarrituta. Erdietsitako emaitzak ikusita, orokorrean esan dezakegu erregeletan oinarritutako analizatzaileek egiten duten ekarpena positiboa dela parserrekin konbinatzen dituguneanean. Kontuan hartu behar da erabili diren erregeletan oinarritutako analizatzaileak fenomeno linguistiko konkretu batzuen estalduran eta egitura sintaktiko batzuen tratamenduan pentsatuta diseinatu direla. Horrenbestez, ez da harritzekoa euskararen analisi sintaktiko orokorrean gehiegi ez laguntzea estaldura eta doitasuna bezalako neurrien arabera. Dependentzia erlazio konkretuetan jasotako emaitzak aztertzen baditugu, aipatu dugun hipotesia bete egiten dela ikusten da, hau da, erabilitako erregeletan oinarritutako sistemen diseinua dela eta, dependentzia erlazio batzuetan oso hobekuntza onak erdiesten diren bitartean, beste batzuetan emaitza negatiboak jasotzen dira. Ondorioz, esan daiteke ideia ona dela hibridazio mota hau erabiltzea erregeletan oinarritutako analizatzaileak ondoen egiteko diseinatu diren dependentzietan hobekuntzak erdietsi nahi badira. Gure helburua euskararen analisi sintaktikoa modu orokorrean hobetzea bada, ikusi da hobekuntzak lortzen direla, baina ez da ahaztu behar hibridazio mota honek bere mugak dituela. Bestalde, erregelak hobetuz gero, emaitzak are gehiago hobetu daitezkeela uste dugu.

  Ildo beretik jarraituz, hibridazio mota honek galderazko esaldietan duen eragina ere aztertu dugu. Kasu honetan, erabilitako bi analizatzaile estatistikoetan erdietsitako emaitzak itxaropentsuak izan dira. Bi kasuetan gainditu dira analizatzaile estatistikoen oinarrizko emaitzak. Hala ere, erabilitako lagina txikia izanik, ezinbestekoa da esperimentuak lagin handiago batekin errepikatzea, hobekuntzak mantentzen diren ala ez zehazteko. Dependentzia konkretuetan gertatutakoari bu-

ruz antzeko ondorioak atera daitezke.

- **Estatistiketan oinarritutako sistemen hibridazioa:** Hurbilpen honetan, chunker estatistiko bat eta analizatzaile sintaktiko estatistikoak konbinatzeak duen eragina zehaztu dugu euskararen analisi sintaktikoan. Kasu honetan ere garrantzia dauka konbinazioan erabili den parserrak, aukeratutako analizatzaile estatistikoaren arabera emaitzak asko aldatzen baitira. Trantsizioetan oinarritutako parserra (MaltOptimizer) erabiltzen bada, emaitzek behera egiten dute orokorrean eta aztertutako dependentzia erlazio guztietan. Grafoetan oinarritutakoa aukeratzen bada (Mate), ostera, hobekuntza txikia nabaritzen da analisi sintaktiko orokorrean, baina ia dependentzia erlazio gehienetan emaitzek okerrera egiten dute. Aztertutako dependentzietan gorantz egiten duen bakarra *ncmod* erlazioa da eta bere eragin positiboa dela medio, parserra gai da analisi orokorraren oinarrizko emaitza gainditzeko beste dependentzien eragin negatiboari aurre eginez, esaldietan *ncmod* erlazioen kopurua oso altua dela aprobetxatuta.

Oro har esanda, euskararen analisi sintaktiko orokorraren emaitzek nabarmen gora egin ez duten arren, emaitzetan sakontzen bada, argi ikusten da dependentzia erlazio konkretuetan hobekuntzak egon direla erabilitako hibridazio mota desberdinetan. Horretaz gain, erregeletan oinarritutako analizatzaileen eta parserren arteko hibridazioa onuragarria dela ikusi da galderazko esaldien analisi sintaktikoan. Bide horri jarraituz, alde batetik, galderazko esaldien lagin handiago baten gainean errepikatu nahi ditugu esperimentuak, erdietsitako hobekuntzan mantentzen diren ala ez zehazteko; beste aldetik, koordinazioaren eta esaldi konpletiboen gainean aplikatu nahi dira galderazko esaldietan egin diren esperimentuak, horretarako, fenomeno horiekin lotutako erregelak landu behar dira. Lehenengo urratsak eman dira bide horretan; fenomeno bakoitzerako laginak atera dira eta hizkuntzalarien esku utzi da erregelen garapena.

Bukatzeko, estatistiketan oinarritutako sistemen konbinaketak emaitza interesgarriak lortu ditu analisi sintaktiko orokorrarentzako oso garrantzitsua den *ncmod* dependentzia erlazioan. Hala ere, gainontzeko dependentzia erlazioetan ez dugu lortu hobekuntzarik. Emaitza horien aurrean, zergatik erdietsi diren hobekuntzak *ncmod* erlazioan eta ez besteetan aztertuko da. Gertaera horren zergatia jakinda, agian posible izango da *ncmod* erlazioan emaitza hobeak lortzea eta gainontzeko erlazioetan aurrerakuntzak ikustea.

Hala ere, erregeletan zein estatistiketan oinarritutako bi chunkerren informazioaren erabilerak ez dirudi ekarpen handia egiten dutenik. Esan bezala, azterketa sakonagoa merezi du gai honek, batez ere maiztasun handieneko etiketen esleipen egokirako lagungarria izan badaiteke.

5 Analisi sintaktiko eleaniztuna

5.1 Sarrera

Esku artean dugun tesiaren helburu nagusietako bat hizkuntza desberdinetan analisi sintaktiko automatikoarekin lortzen diren emaitzak hobetzen dituzten teknikak jorratzea eta gure sisteman aplikatzea da. Lortu nahi den xedea ez da erraza, hizkuntza ezberdinek beraien artean dituzten desberdintasunak askotan handiak eta ugariak direlako. Esandakoa kontuan hartuta, kapitulu honetan erabili diren hizkuntza ezberdinek badituzte beraien artean desberdintasun ugari, baina badute, bestalde, ezaugarri komun bat: denak dira morfologikoki aberatsak, eta justu aberastasun horretaz baliatu gara lan honetan eskuratu nahi ditugun emaitzak erdiesteko.

Aipatu ditugun morfologikoki aberatsak diren hizkuntza horiek euskara, alemana, frantsesa, hungariera eta suediera dira. Ondorengo lerroetan kapitulu hau nola antolatu den aurkeztuko da. 5.2 atalean, ikuspegi eleaniztuna jorratzeko jarraitu dugun metodologia aurkeztuko dugu, 5.3 atalean, egindako esperimentuetan erabilitako baliabide orokorrak, eta ondorengo hiru ataletan, 5.4, 5.5, eta 5.6, ikuspegi eleaniztuna lantzeko aplikatu ditugun hiru teknika azalduko ditugu: ezaugarrien ingeniaritza, multzokatzea eta meta-ezaugarriak, hurrenez hurren. 5.7 atalean aurreko hiru teknikekin sortuko ditugun analisiak konbinatuko ditugu, beraien artean osagarriak diren aztertzeko. Bukatzeko, 5.8 atalean, egindako esperimentu guztietatik atera ditugun ondorioak azalduko ditugu zenbait etorkizuneko lanekin batera.

5.2 Metodologia

Kapitulu honetan ikuspegi eleaniztuna landu nahi izan da, euskararekin konparagarriak diren hizkuntzak aztertuz. Horretarako, Seddah *et al.* (2013) eta Seddah *et al.* (2014) ataza partekatuetan eskuragarri egon diren baliabideak erabili dira. Ataza horietan, morfologikoki aberatsak diren hainbat hizkuntzetako zuhaitz-bankuak eta corpus baliabideak izan ditugu atzigarri. Hala ere, zuhaitz-banku guztiak ez daude era berean etiketatuta, denak era berean landu ahal izateak suposatzen duen arazoarekin. Ondorioz, atal honetan aplikatu nahi ditugun hurbilpenak ahalik eta konparagarrienak izateko, ahalik eta aldaketa gutxien eskatzen duten hizkuntzak aukeratu dira.

Atal honetako helburuetako bat zuhaitz-banku, corpus eta tresna eskuragarriak erabiltzea da, horiek ematen digutenari ahalik eta etekin handiena ateratzeko, ondorioztatutako teknika eta ezaugarriak, ahalik eta hizkuntza gehienetan aplikatzeko. Horretarako, hizkuntza bakoitzaren zehaztasun berezietatik aldendu egingo gara orokortze maila altuagoa lortzeko asmoarekin, analizatzaile sintaktikoek, ikuspegi ez-gainbegiratuak eta erdi-gainbegiratuak ematen dizkiguten aukera desberdinak aztertuz, eta, lortutako emaitzak, ahal den heinean, morfologia xumeagoa duten hizkuntzekin erdietsitako emaitzekin alderagarriak izatea espero dugu.

5.3 Esperimentazio-ingurunea

Atal honetako esperimentuekin, hizkuntza desberdinetan modu arrakastatsuan aplikatu ahal diren hiru hurbilpen aztertu nahi dira. Hori dela eta, gure esperimentuak hizkuntza desberdinetako zuhaitz-bankuen gainean burutu ditugu, hizkuntza horietako bakoitzerako hiru hurbilpenak aplikatuz.

Erabilitako hizkuntzak euskara, frantsesa, alemana, hungariera eta suediera dira eta horiei dagozkien zuhaitz-bankuak 3.2.1 atalean azaldu ditugunak dira.

Erabili ditugun analizatzaile sintaktikoak 3.2.3.1 atalean aurkeztu ditugun MaltOptimizer, Mate eta MST dira. Lehenengo biak 5.4 eta 5.5 ataletan erabiliko dira eta horiek aukeratzearen arrazoi nagusiak emaitzak bueltatzeko duten azkartasuna eta oinarri desberdinak dituztela dira, lehenengoa trantsizioetan oinarrituta eta bigarrena grafoetan oinarrituta daudelarik. Atal horietako bakoitzean lortutako analisiak MaltBlender tresnaren bidez (ikus 3.2.3.3 atala) konbinatu egingo direnez, iturri desberdinetatik lortutako ana-

lisiak edukitzea komeni zaigu, ugaritasunak konbinaketetan asko eragiten baitu. MST, berriz, 5.6 atalean erabili dugu eta hura aukeratzearen arrazoi nagusia, batez ere bere kodean aldaketak egiteko ematen duen erraztasuna da. 5.7 atalean aurreko hiru ataletan erdietsitako analisiak konbinatuko dira MaltBlender tresnaren bidez.

5.4 Ezaugarrien ingeniaritza

Ezaugarrien ingeniaritza ditugun datuak gure atazarako ikasketa automatikoko ereduetarako erabilgarriagoak izango diren modu batera bihurtzeko prozesua da, era honetara orain arte tratatu gabeko datuetan emaitza hobeak lortuz. Esan daiteke dauzkagun datuetatik ateratako ezaugarriek eragin zuzena dutela ikasketa automatikoko ereduetan, hots, zenbat eta esanguratsuagoak izan ezaugarri horiek, orduan eta eraginkorragoa izango da datu eta ezaugarri horien gainean sortuko den eredua. Domingosek dioen bezala:

> "At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used." (Domingos 2012)

Ondorioz, ezaugarrien ingeniaritzaren helburuetako bat gure datuen azpian dauden egiturak azaleratzea da ikasketa automatikoko algoritmoak gure datuak hobeto uler ditzan.

5.4.1 Sarrera

Morfologikoki aberatsak diren hizkuntzak erronka bat dira egungo analizatzaile sintaktiko automatikoen eraikuntzarako, batez ere hizkuntza hauetan lortzen dituzten emaitzak ingelesa bezalako hizkuntza ez-eranskariekin alderatzen baditugu. Nivre *et al.* (2007) lanean argi geratzen dira analizatzaile sintaktikoek morfologikoki aberatsak diren hizkuntzekin dituzten zailtasunak, non euskarak, grekoak eta arabierak ez duten ingelesak eta italierak lortzen dituzten emaitzak erdiesten. Ildo horretatik jarraituz, morfologikoki aberatsak diren hizkuntzak tratatzeko nahiko ohikoa da horien berezko ezaugarria den informazio morfologiko aberatsa erabiltzea analizatzaile sintaktikoan (Tsarfaty *et al.* 2010), baina batzuetan posible da guk nahi ditugun emaitzak ez lortzea, hau da, posible da ezaugarri morfologiko batzuek ekarpen negatiboa izatea. Adibidez, euskararako emaitza hobeak lortzen dira

Mate analizatzaile sintaktikoarekin ezaugarri guztiak erabili beharrean kasua, numeroa eta mendeko esaldi mota soilik erabilita, aurrerago 5.2 taulan ikusiko den bezala.

Bengoetxea eta Gojenola (2010) lanean hainbat ezaugarri morfologiko probatzen dituzte banan-banan MaltParser analizatzailean, ezaugarri horietako bakoitzak euskarako dependentzia zuhaitzen azterketan duen eragina aztertzeko. Ateratako ondorioen arabera, euskararako mendekotasun zuhaitzak aztertzeko garaian, hitzen kasuak eta zer-nolako mendeko esaldia den adierazten duten ezaugarriek garrantzi handia dute. Kasuak hitzaren funtzio sintaktikoa ondorioztatzen laguntzen du eta mendeko motak, berriz, esaldi nagusi eta mendekoen arteko mugak markatzen laguntzen du, eta horrek berebiziko garrantzia du aditz nagusiaren eta mendeko esaldien arteko dependentzia erlazioa zehazteko garaian. Behin euskararen ezaugarri morfologiko bakoitzaren eragin zehatza jakinda, analizatzaile sintaktikoari ezaugarri guztiak pasatzen dizkiote, baina ongi berezitutako bi multzotan: hitzaren kasua eta mendeko esaldi mota alde batetik, eta gainontzeko ezaugarriak beste aldetik (ikus euskarazko ezaugarri morfologikoen azalpena A eranskinean). Informazioa modu horretara erabiliz hobekuntza esanguratsuak lortzen dituzte ezaugarri guztiak batera erabiliz lortutakoarekin alderatuta.

Marton *et al.* (2010) lanean ere antzeko saiakerak aurkeztu dira arabierarako, dependentzietan oinarritutako zuhaitz sintaktikoak aztertzeko garaian ezaugarri morfologiko lagungarrienak zeintzuk diren jakiteko. Analisi sintaktikoan ezaugarri morfologiko bakoitza bakarka erabili ondoren lortutako emaitzak alderatu dira ondorio hauetara iritsiz: zuhaitz-bankuan erabilitako ezaugarri morfologikoak zuzenak direnean (urre-patroia), gehien laguntzen duen ezaugarria hitzaren kasua da, baina erabilitako ezaugarriak automatikoki sortutakoak badira, aldiz, kontrakoa gertatzen da, hots, hitzaren kasua erabiltzen den egoera gehienetan emaitzek okerrera egiten dute.

Bestalde, 2.3.2 puntuan aipatu den bezala, analizatzaile sintaktiko batzuek eskuragarri duten informazioa erarik eraginkorrenean ez dutela erabiltzen ikusten da Çetinoğlu eta Kuhn (2013) lanean, zutabe batzuetan sartutako informazioari garrantzi handiagoa ematen diotelarik. Analizatzaileari normalean kategoria eta azpikategoria pasatzen zaizkion zutabeetan ezaugarri morfologikoak pasatzen dizkiote banan-banan ikusteko ea inolako aldaketarik gertatzen den ezaugarri bakoitza bere zutabean eta eskuragarri dauden ezaugarri guztiak erabilita lortutakoarekin alderatuta. Baldintza horietan turkierarako analizatzaile sintaktiko batzuek emaitza hobeak lortzeko joera dutela frogatu da hitzaren kategoriaren ordez hitzaren ezaugarri morfologiko

jakin batzuk erabiltzen badira.

5.4.2 Gure hurbilpena

Ezaugarrien ingeniaritzarekin lotutako atal honetan, bide horrek eskaintzen dituen aukera desberdinak aztertuko ditugu morfologikoki aberatsak diren hizkuntzetako analisi sintaktikoa hobetzeko, eta helburu hori erdiesteko, ondorengo atazak landu dira hautatutako hizkuntza bakoitzerako:

- Ezaugarri morfologiko bakoitzaren ekarpena neurtu. Horretarako ezaugarriak banan-banan erabilita lortutako emaitza ezaugarririk gabeko konfigurazioarekin alderatuko da.
- Emaitzetan hobekuntza handiena eragiten duen ezaugarri morfologikoen azpimultzo egokia zehaztu.
- Analizatzaile sintaktikoek pasatako ezaugarriei ematen dieten garrantzia jakinda, estandartzat jotzen den ordena aldatu, bestelako ezaugarri morfologikoei lehentasuna emateak duen ekarpena neurtzeko. Horretarako ez-estandarrak diren ordenekin erdietsitako emaitzak estandarrekin lortutakoekin konparatuko dira.
- Ezaugarrien ingeniaritza aplikatu ondoren lortutako analisi sintaktiko desberdinak konbinatu, modu bakoitzean lortutako analisiak beraien artean osagarriak diren zehazteko.

5.4.3 Esperimentuak eta emaitzak

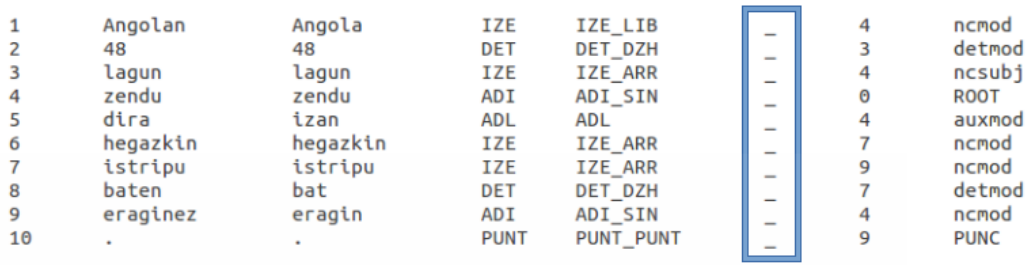
Puntu honetan arestian aipatutako atazekin lotutako esperimentuak deskribatuko dira, bakoitzaren ebaluazioarekin batera.

5.4.3.1 Ezaugarrien eragina neurtzeko esperimentuak

Analisi sintaktikorako esanguratsuenak diren ezaugarri morfologikoak zeintzuk diren jakiteko, analizatzaileek ezaugarri horiek zehazteko erabiltzen duten posizioan (ezaugarri morfologikoen zutabean) pasako dira banan-banan.

	Eus	**Fra**	**Ale**	**Hun**	**Sue**
Oinarrizkoak					
Denak	83,0	84,2	91,0	82,8	76,7
Hutsa	76,5	84,2	90,9	75,3	76,9
Izenarekin Lotutakoak					
Kasua	**82,2***	84,2	**91,0**	**80,9***	76,8
Numeroa	**77,7***	**84,3**	**90,9**	75,7	76,3
Generoa		**84,3**	**91,0**		76,0
Pertsona		**84,3**	90,9	75,7	
Egilearen Pertsona				**76,3***	
Egilearen Numeroa				**76,3***	
Jasotzailearen Numeroa				75,4	
Aditzarekin Lotutakoak					
Aditz-forma					**77,0**
Erlazioa	**77,1**				
Nork	76,7				
Nori	76,2				
Nor	76,6				
Dadudio	76,8				
Aspektua	76,4				
Modua	76,6	84,2	90,8	75,5	76,2
Denbora		84,2	90,9	75,3	
Aditz-mota	76,5				
Azpikategoria		84,2		75,9	
Bestelakoak					
Gradua			90,8	75,0	**76,8**
Koordinazio-mota				75,3	
Forma				75,3	
Mota				75,3	
Mugatasuna				75,7	**76,8**
Laburdura					76,3
Burutua					76,3
Izenordainaren funtzioa					76,1

5.1 taula – Hizkuntza bakoitzeko ezaugarri bakoitzak analisi sintaktikoan duen eragina *Labeled Attachment Score* (LAS) neurriaren arabera. Eus=euskara, Fra=frantsesa, Ale=alemana, Hun=hungariera eta Sue=suediera. Letra lodiz hizkuntza bakoitzean erdietsitako hiru emaitza hoberenak. *= Estatistikoki esanguratsua McNemar testaren arabera, $p < 0{,}05$.

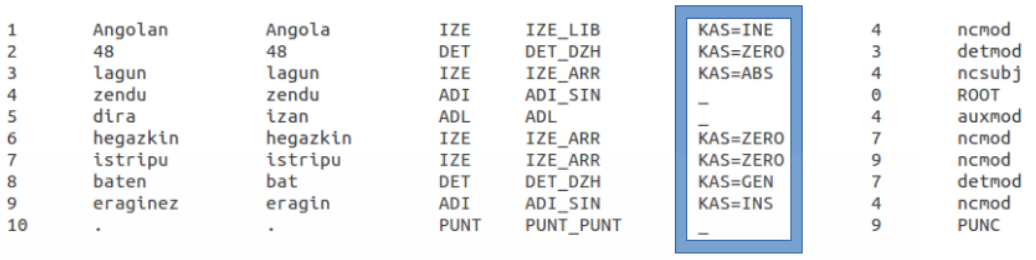


1	Angolan	Angola	IZE	IZE_LIB	_	4	ncmod
2	48	48	DET	DET_DZH	_	3	detmod
3	lagun	lagun	IZE	IZE_ARR	_	4	ncsubj
4	zendu	zendu	ADI	ADI_SIN	_	0	ROOT
5	dira	izan	ADL	ADL	_	4	auxmod
6	hegazkin	hegazkin	IZE	IZE_ARR	_	7	ncmod
7	istripu	istripu	IZE	IZE_ARR	_	9	ncmod
8	baten	bat	DET	DET_DZH	_	7	detmod
9	eraginez	eragin	ADI	ADI_SIN	_	4	ncmod
10	.	.	PUNT	PUNT_PUNT	_	9	PUNC

5.1 irudia – Euskarazko zuhaitz-bankutik hartutako esaldia, ezaugarrien zutabea hutsik duela.

Ataza hau aurrera eramateko Mate analizatzaile sintaktikoa erabili dugu, orain arte ez delako ezaugarrien bakarkako ekarpena neurtzeko erabili, MaltOptimizer eta MST erabili diren bezala.

Lehenik eta behin, hizkuntza bakoitzeko ezaugarrien zutabea hutsik utzi dugu (ikus 5.1 irudia), analizatzaile sintaktikoak lortzen duen emaitza erreferentzia bezala erabiltzeko. Hurrengo pausoa gainontzeko ezaugarriak ezaugarrien zutabean (ikus 5.2 irudia) banaka sartzea da analisi sintaktikoa gauzatzean lortutako erreferentziarekin konparatzeko; zenbat eta emaitza altuagoa lortu erreferentziarekiko, erabilitako ezaugarri morfologikoa orduan eta esanguratsuagoa izango da. Bukatzeko, ezaugarrien zutabean ezaugarri guztiak erabilita (ikus 5.3 irudia) ere lortu ditugu emaitzak, ikusteko ea ezaugarri konbinazio batzuek eragin negatiboa duten analisi sintaktikoan. Izan ere, printzipioz, zenbat eta ezaugarri gehiago erabili orduan eta emaitza hobeak espero baitira.



1	Angolan	Angola	IZE	IZE_LIB	KAS=INE	4	ncmod
2	48	48	DET	DET_DZH	KAS=ZERO	3	detmod
3	lagun	lagun	IZE	IZE_ARR	KAS=ABS	4	ncsubj
4	zendu	zendu	ADI	ADI_SIN	_	0	ROOT
5	dira	izan	ADL	ADL	_	4	auxmod
6	hegazkin	hegazkin	IZE	IZE_ARR	KAS=ZERO	7	ncmod
7	istripu	istripu	IZE	IZE_ARR	KAS=ZERO	9	ncmod
8	baten	bat	DET	DET_DZH	KAS=GEN	7	detmod
9	eraginez	eragin	ADI	ADI_SIN	KAS=INS	4	ncmod
10	.	.	PUNT	PUNT_PUNT	_	9	PUNC

5.2 irudia – Euskarazko zuhaitz-bankutik hartutako esaldia, ezaugarrien zutabean kasua bakarrik utzita.

5.1 taulako emaitzak aztertzen baditugu, ikus daiteke euskara dela ezaugarri morfosintaktikoak gehituta aldaketa nabarmenenak jasaten dituen hiz-

kuntza. Kasua bakarrik erabilita, % 5,7ko hobekuntza lortzen da ezaugarrien zutabea hutsa (5.1 taulan Hutsa) mantenduz lortzen denarekin konparatuta. Argi geratzen da kasua oso garrantzitsua dela dependentzia zuhaitzak analizatzaile sintaktikoekin eraikitzeko garaian. Izan ere, ezaugarri guztiak erabilita lortzen den LASaren eta kasua bakarrik erabilita lortzen denaren arteko aldea oso txikia baita (+% 0,8). Hurrengo bi ezaugarri esanguratsuenak euskararako, alde nabarmenarekin, numeroa eta mendeko esaldi mota (erlazioa) dira, hauek soilik erabilita % 1,2 eta % 0,6ko hobekuntza lortzen delarik ezaugarririk ez erabiltzearekin alderatuta, hurrenez hurren.

1	Angolan	Angola	IZE	IZE_LIB	KAS=INE|NUM=S	4	ncmod
2	48	48	DET	DET_DZH	KAS=ZERO	3	detmod
3	lagun	lagun	IZE	IZE_ARR	KAS=ABS	4	ncsubj
4	zendu	zendu	ADI	ADI_SIN	ADM=PART|ASP=BURU	0	ROOT
5	dira	izan	ADL	ADL	MDN=A1|DADUDIO=NOR|NOR=HAIEK	4	auxmod
6	hegazkin	hegazkin	IZE	IZE_ARR	KAS=ZERO	7	ncmod
7	istripu	istripu	IZE	IZE_ARR	KAS=ZERO	9	ncmod
8	baten	bat	DET	DET_DZH	KAS=GEN	7	detmod
9	eraginez	eragin	ADI	ADI_SIN	KAS=INS|ADM=PART|NUM=P	4	ncmod
10	.	.	PUNT	PUNT_PUNT		9	PUNC

5.3 irudia – Euskarazko zuhaitz-bankutik hartutako esaldia, hitz bakoitzerako ezaugarrien zutabean ezaugarri guztiak erabilita.

Frantseserako lortutako emaitzak ikusita esan daiteke euskararekin gertatzen denaren kontrakoa gertatzen dela, hots, ezaugarri morfologikoek ez dute ia eraginik. Ezaugarririk esanguratsuenak generoa, numeroa eta pertsona dira. Horien eragina oso txikia da analisi sintaktikoan eta ia ez dago alderik ezaugarri morfologikorik ez erabiltzearen eta horietakotako bat erabiltzearen artean (+% 0,1). Gainera, hizkuntza honetan erabili diren ezaugarri guztiak beraien artean osagarriak ez direla ikusten da emaitzetan. Izan ere, ezaugarri bakarra erabiltzen den kasu batzuetan ezaugarri guztiak erabilita lortzen den emaitza gainditzen baita.

Frantsesarekin gertatzen denaren antzera, alemaneko ezaugarri morfosintaktikoek oso eragin txikia dute analisi sintaktikoan. Hiru ezaugarri esanguratsuenak kasua, generoa eta numeroa dira eta hauek banaka erabilita lortzen diren hobekuntzak +% 0,1, +% 0,1 eta +% 0,0ekoak dira, hurrenez hurren. Interesgarria da ikustea nola kasua bakarrik erabilita edo generoa bakarrik erabilita ezaugarri guztiak erabilita lortzen den emaitza bera erdiesten dela. Fenomeno hau gerta daiteke ezaugarri batzuen eragin ahularen eta beste batzuen eragin negatiboaren ondorioz. Modu honetara, ezaugarri batzuen eragin positiboa beste batzuen eragin negatiboarekin baliogabetzen da. Ale-

manean eta frantsesean ezaugarri morfologikoek duten eragin ahula ikusita, esan daiteke bi hizkuntza horietan analizatzaileak erdiesten dituen emaitzak beste ezaugarri batzuen eraginez lortzen direla, esate baterako, lemaren, kategoriaren eta azpikategoriaren eraginez.

Hungariera da ezaugarri morfosintaktiko gehien dituen hizkuntza aztertu ditugunen artean, 14 guztira. Hizkuntza hau da, euskararekin batera, aldaketa gehien jasaten dituena ezaugarri morfosintaktikoak gehitzen zaizkionean. Hiru ezaugarri esanguratsuenak kontuan hartuta, kasuarekin % 5,6ko hobekuntza erdiesten da, egilearen pertsonarekin % 1eko hobekuntza eta egilearen numeroarekin % 1koa. Bestalde, azpikategoriak +% 0,6an laguntzen du eta numeroak, pertsonak eta mugatasunak +% 0,4ko hobekuntza lortzen dute. Gainontzeko ezaugarriek ez dute modu esanguratsuan laguntzen eta gradua soilik erabiltzen bada analisi sintaktikoan emaitzek okerrera egiten dute (-% 0,3). Ezaugarri guztiak erabiltzen baditugu analisi sintaktikoan, hungarieran % 82,8ko LASa lortzen dugu. Emaitza hau ikusita, hungarierako ezaugarrien ekarpena analisi sintaktikoan % 7,5ekoa dela esan daiteke, aztertu ditugun bost hizkuntzetatik handiena, adibidez, euskaraz % 6,5ekoa delarik.

Aztertutako hizkuntzekin amaitzeko, frantsesarekin eta alemanarekin gertatzen den antzera, suedierako ezaugarri morfosintaktikoek ez diote analizatzaile sintaktikoari asko laguntzen. Gainera, hizkuntza honetako ezaugarriekin fenomeno arraroak ikus daitezke. Euskaran, alemanean eta hungarieran kasuak ezaugarririk ez erabiltzearekin alderatuta eragin positiboa duen arren (frantsesean ez du eraginik), suedieran emaitza negatiboa lortzen da berarekin. Orokorrean, aditz-forma kontuan hartu gabe, ezaugarri guztiekin emaitza negatiboak lortzen dira. Egoera honen aurrean, ez da harritzekoa ezaugarri guztiak erabiltzen direnean analizatzaile sintaktikoan emaitzek okerrera egitea (-% 0,2). Hau esanda, suedierarako emaitzarik hoberenak lortu dituzten ezaugarriak aditz-forma (+% 0,1), mugatasuna (-% 0,1), gradua (-% 0,1) eta kasua (-% 0,1) dira. Hurrengo azpiatalean hizkuntza bakoitzeko hiru ezaugarririk eraginkorrenak erabiliko direnez esperimentu desberdinak egiteko, suedierarako kasua ez dugu aukeratuko ehunekoak kontuan hartuta mugatasunak eta graduak baino emaitza baxuagoa lortzen duelako analisi sintaktikoan.

Behin hizkuntza guztietako emaitzak agerian daudela, argi geratzen da euskararen eta hungarieraren ezaugarriek eragin handiagoa dutela analisi sintaktikoan gainontzeko hiru hizkuntzetakoek baino. Gertaera horri azalpen bat emateko okurritzen zaigun arrazoi nagusia euskarazko eta hungarierako tresna automatikoek kalitate altuagoko ezaugarriak lortzen dituztela da. Izan

ere, hizkuntza horietarako espresuki garatutako tresnekin erdietsitako ezaugarriak erabili baitira hizkuntza horietako zuhaitz-bankuak sortzeko, batez ere kategoria, azpikategoria eta ezaugarri morfologikoak sortzeko. Frantseserako eta alemanerako eta suedierarako analizatzaile eleaniztunak erabilita sortu dira ezaugarri horiek.

5.4.3.2 Ezaugarrien ingeniaritzako esperimentuak

Aurreko atalean, analisi sintaktikorako ezaugarri morfosintaktiko esanguratsuenak zeintzuk diren aztertu dugu aukeratutako bost hizkuntzetarako. Horretan oinarrituta, informazio hori erabiliko dugu atal honetan esperimentu desberdinak egiteko. Hizkuntza bakoitzerako eta analizatzaile sintaktiko bakoitzerako (MaltOptimizer eta Mate) 6 esperimentu egin ditugu ezaugarri guztiak erabilita lortutako emaitzekin konparatzeko:

1) $3{-}best_{Malt}$: Hiru ezaugarri hoberenak erabili dira ezaugarrien zutabean eta MaltOptimizer analizatzailearekin analizatu da.

2) $CPOS - best_{Malt}$: Ezaugarririk esanguratsuena ipini da hitzaren kategoriaren lekuan eta MaltOptimizer analizatzailearekin analizatu da. Kategoria ezaugarrien zutabean gehitzen da beste ezaugarri bat izango balitz bezala.

3) $POS - best_{Malt}$: Ezaugarririk esanguratsuena ipini da hitzaren azpikategoriaren lekuan eta MaltOptimizer analizatzailearekin analizatu da. Azpikategoria ezaugarrien zutabean gehitzen da beste ezaugarri bat izango balitz bezala.

4) $3{-}best_{Mate}$: Hiru ezaugarri hoberenak erabili dira ezaugarrien zutabean eta Mate analizatzailearekin analizatu da.

5) $CPOS - best_{Mate}$: Ezaugarririk esanguratsuena ipini da hitzaren kategoriaren lekuan eta Mate analizatzailearekin analizatu da. Kategoria ezaugarrien zutabean gehitzen da beste ezaugarri bat izango balitz bezala.

6) $POS - best_{Mate}$: Ezaugarririk esanguratsuena ipini da hitzaren azpikategoriaren lekuan eta Mate analizatzailearekin analizatu da. Azpikategoria ezaugarrien zutabean gehitzen da beste ezaugarri bat izango balitz bezala.

Esperimentu horiek guztiak zuhaitz-bankuko hitzen ordena ezkerretik eskuinera (hemendik aurrera ordena arrunta) eta eskuinetik ezkerrera (hemendik aurrera alderantzizko ordena) erabiliz egin dira hizkuntza guztietarako. Ezaugarrien zutabean hiru ezaugarri hoberenak erabiltzea erabaki da, konbinazio desberdinekin (3-5 ezaugarri) probak egin ondoren lortutako emaitzaren eta erabilitako ezaugarri kopuruaren arteko erlaziorik hoberena lortu delako modu horretara.

	Eus	**Fra**	**Ale**	**Hun**	**Sue**
Oinarriak					
MaltOptimizer	80,0	79,9	87,6	77,3	73,4
Mate	83,0	84,2	91,0	82,8	76,7
Ezkerretik Eskuinera					
$3 - best_{Malt}$	79,9	79,9	87,6	75,9	73,4
$CPOS - best_{Malt}$	80,3	79,7	87,5	76,6	72,9
$POS - best_{Malt}$	78,7	78,7	86,6	77,2	72,8
$3 - best_{Mate}$	**83,4**	84,3	90,8	82,4	76,6
$CPOS - best_{Mate}$	82,7	84,3	91,0	**82,7**	76,8
$POS - best_{Mate}$	82,2	83,4	90,5	82,5	76,5
Eskuinetik Ezkerrera					
$3 - best_{Malt}$	80,1	78,9	86,9	75,3	69,3
$CPOS - best_{Malt}$	80,0	79,0	86,7	76,6	69,3
$POS - best_{Malt}$	**81,0***	77,8	85,4	74,9	70,2
$3 - best_{Mate}$	83,3	84,3	90,9	82,1	76,5
$CPOS - best_{Mate}$	83,1	**84,6**	**91,0**	82,6	**77,0**
$POS - best_{Mate}$	81,6	83,5	90,6	82,4	76,4

5.2 taula – Hizkuntza bakoitzerako ezaugarrien ingeniaritza teknikekin egin diren esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz hizkuntza bakoitzean analizatzaile bakoitzak erdietsitako emaitzak. *= Estatistikoki esanguratsua McNemar testaren arabera, $p < 0{,}05$.

Esperimentu horien guztien emaitzak 5.2 taulan bildu ditugu. Euskararako emandako emaitzak aztertzen baditugu, erabilitako bi parserren arteko ezberdintasuna nabaria da. Izan ere, Mate erabilita, batezbestekoan 3 puntuko hobekuntza lortzen baita MaltOptimizerrekin alderatzen badugu. Esan beharra dago hau ez dela euskaraz soilik gertatzen, probatu ditugun

hizkuntza guztietan antzeko aldea ikus baitaiteke bi analizatzaile sintaktiko horien artean. Euskaraz emaitzarik hoberena lortu duen konfigurazioa Mate analizatzailearekin ezaugarrien zutabean hiru ezaugarririk esanguratsuenak erabiltzen dituena izan da (% 83,4 LAS). Emaitza hori oinarrizko konfigurazioarekin lortutakoarekin (% 83,0) alderatzen bada, % 0,4ko hobekuntza dagoela ikus daiteke, nahiz eta hobekuntza hori McNemar testaren arabera ($p < 0{,}05$) estatistikoki esanguratsua ez izan. MaltOptimizerrekin lortutako emaitzak aztertzen badira, hobekuntza bakarra euskaraz lortu da. Zehazki, ezaugarririk hoberena kategoriaren ordez jarrita ($POS - best_{Malt}$) hitzen alderantzizko ordenarekin % 1eko hobekuntza estatistikoki esanguratsua ($p < 0{,}05$) lortu da (% 81,0 vs % 80,0).

Bestalde, frantsesarekin, Mate parserrarekin soilik lortzen da oinarrizko konfigurazioa gainditzea. Analizatzaile horrekin emaitzarik hoberena azpikategoriaren ordez ezaugarri esanguratsuena erabilita ($CPOS - best_{Mate}$) alderantzizko ordenarekin lortzen da (+ 0,4), hobekuntza hori ez delarik estatistikoki esanguratsua McNemarren testaren arabera ($p < 0{,}05$). MaltOptimizerren kasuan, ematen du erabili ditugun konfigurazio ezberdinek ez dutela eragin positiborik izan analisi sintaktikoan, ez baitira gainditzen oinarrizko konfigurazioekin lortutako emaitzak. Hala ere, ikusten da emaitza konparagarriak erdiesten direla 3 ezaugarri esanguratsuenak ($3 - best_{Malt}$) ordena arruntean erabilita. Esan beharra dago, azken esperimentu honetan, analizatzaile sintaktikoari lan-karga kendu zaiola bere kalkuluak ezaugarri guztien gainean egin beharrean hiru ezaugarriren gainean bakarrik egin behar dituelako. Horregatik interesgarria izan daiteke ezaugarri guztiak erabili beharrean 3 ezaugarri esanguratsuenak erabiltzea.

Alemanean zentratzen bagara, ez da lortu oinarrizko konfigurazioekin lortutako emaitzak gainditzea, baina bai pareko emaitzak erdiestea ezaugarri gutxiagorekin ($3 - best_{Malt}$). Aurreko atalean zehaztu dugun bezala, alemana da erabilitako ezaugarriekiko aldaketa gutxien jasan duen hizkuntza. Informazio hori jakinda, ez da harritzekoa atal honetan erabili diren konfigurazio desberdinek berarengan eragin gutxien izan duten hizkuntza alemana izatea. Hala ere, frantseserako esan bezala, analizatzaileari kalkulu-karga kentzeko baliagarriak dira esperimentu hauek.

Hungarieran probatu ditugun esperimentuek ere ez dute lortu oinarrizko konfigurazioarekin erdietsitako emaitza gainditzea. Hala ere, kasu askotan oinarrizko konfigurazioarekin lortutako emaitzen parekoak dira. Kontuan hartuta emaitza horiek ezaugarriak modu desberdinetan erabiliz atera direla, hurrengo puntuan analisi desberdinak konbinatzen direnean ugaritasun horri

probetxua ateratzea espero da.

Suedierari dagokionean, 5.2 taulari erreparatzen badiogu, konturatuko gara esperimentu guztietatik bitan hobetu direla oinarrizko emaitzak. Hobekuntza horiek Mate analizatzaile sintaktikoa erabilita erdietsi dira eta ez dira estatistikoki esanguratsuak McNemarren testaren arabera ($p < 0{,}05$). Zehazki, aldaketarik nabarmenena (+% 0,3) hitzaren azpikategoria ezaugarri esanguratsuenarekin ordezkatuz eman da ($CPOS - best_{Mate}$) alderantzizko ordenarekin. Beste igoera, berriz, konfigurazio berdinarekin lortu da, baina hitzen ordena arruntarekin (+% 0,1).

Aplikatutako konfigurazio ezberdinek hizkuntza ezberdinetan izan duten eragina ikusita, euskararako eta alemanerako espero zitezkeen emaitzak direla esan beharra dugu. Izan ere, bakarkako ezaugarri morfologikoek eragin handiena izan duten hizkuntzetan espero baitira emaitzarik hoberenak eta bakarkako ezaugarri morfologikoek eragin ahulena izan duten hizkuntzetan espero baitira emaitzarik okerrenak. Ondorio horretara iristeko arrazoi nagusia egin ditugun esperimentuak ezaugarri horietan oinarrituta daudela da (hiru hoberenak erabili, kategoriarekin trukatu, ...). Hori dela eta, hungarieran bakarkako ezaugarriek izan duten eragina kontuan hartuta emaitza hobeak espero dira, euskararen parekoak. Frantsesean eta suedieran, aldiz, kontrakoa gertatzen da, hots, emaitza okerragoak espero dira, alemanaren antzera.

Hala ere, hurrengo atalean aplikatuko den konbinaketarekin hizkuntza guztietan emaitzek gora egitea espero dugu, teknika honek, lortutako analisien ugaritasuna aprobetxatzen baitu, xede horrekin, besteak beste, sortu direlarik atal honetako analisi desberdinak.

5.4.3.3 Konbinaketa Esperimentuak

Aurreko azpiatalean erabilitako konfigurazio batzuek guk nahi genuen eragin positiboa izan ez duten arren, lortutako analisien ugaritasuna konbinaketaren bidez analisi sintaktiko sendoago bat erdiesteko erabil dezakegula uste dugu. Hizkuntza guztietarako konbinaketaren emaitzak 5.4 irudian aurkeztu ditugu oinarrizko konfigurazioarekin[1] lortutako emaitzekin batera, horiekin alderatu ahal izateko.

5.4 irudia aztertzen bada, konbinaketari probetxu gehien atera dion hizkuntza euskara dela konturatuko gara. Modu honetara, % 3,2ko hobekuntza

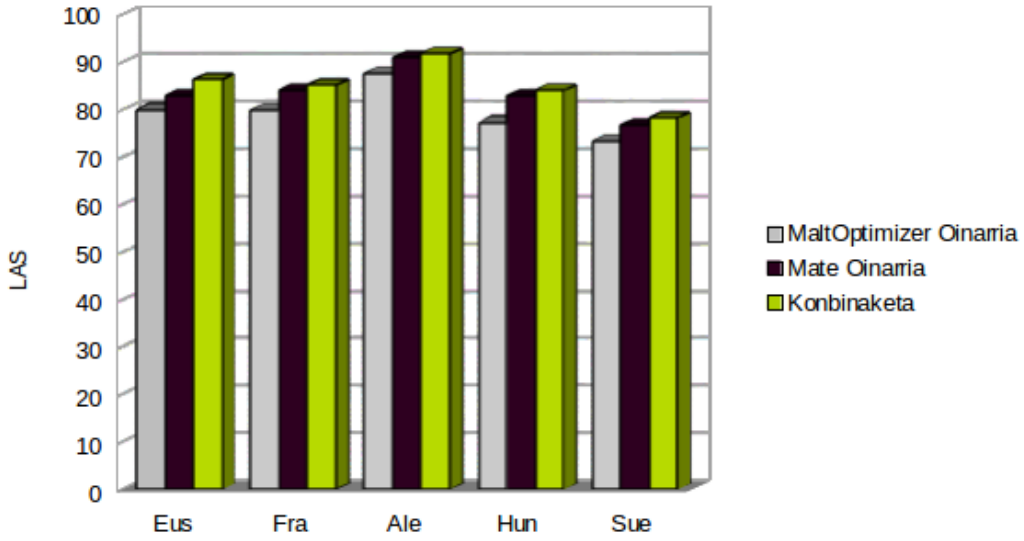
[1]Ezaugarri guztiak dituena, aldaketarik jasan ez duena.

lortu da Mate analizatzaileak emandako emaitzarekin konparatzen bada eta, % 6,2koa MaltOptimizer analizatzailearekin alderatzen bada. Frantsesarekin, berriz, % 0,9 puntuko gorakada erdiesten da Mate analizatzailearen oinarrizko emaitzarekiko eta % 5,2ko hobekuntza MaltOptimizerren oinarrizko emaitzarekiko.

Alemanean, aldiz, % 0,8ko gorakada dagoela erakusten dute zenbakiek Matekin lortutako oinarrizko emaitzarekiko eta % 4,2koa MaltOptimizerren emaitzarekiko. Hungarieran ematen da gorakadarik nabarmenena oinarrizko emaitzekiko: % 6,8 puntuko hobekuntza MaltOptimizerrekin jasotako oinarrizko emaitzarekin alderatuta. % 1,8koa, berriz, Matenarekin konparatuta.

Suedieran konbinaketaren bidez % 4,8ko hobekuntza lortu da MaltOptimizerren oinarrizko emaitzarekin konparatuta eta % 1,3koa Mate analizatzaile sintaktikoarekin jaso dugun oinarrizko emaitzarekin alderatuta.

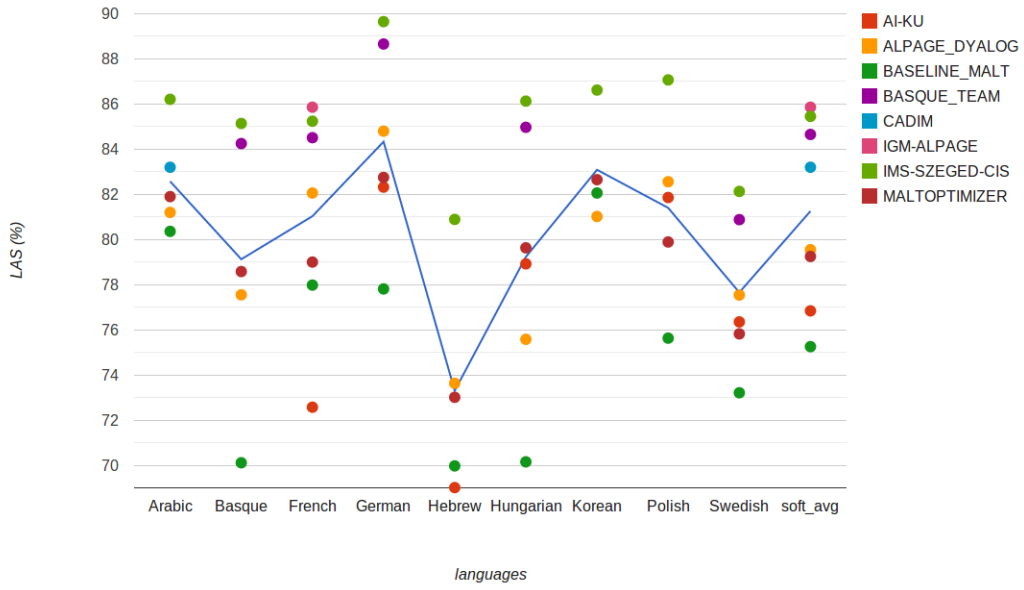
Aipatu ditugun hobekuntza guztiak dira estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$), bai MaltOptimizerren oinarrizko konfigurazioarekiko eta bai Mateen oinarrizko konfigurazioarekiko.



5.4 irudia – Konbinaketa erabiliz egindako esperimentuen emaitzen irudikapena *Labeled Attachment Score* (LAS) neurriaren arabera.

Ezaugarrien ingeniaritza landu den atal honetan ikusitakoaren balorazio bezala, atal honetan egindako esperimentuetan oinarrizko konfigurazioarekin

lortutako emaitzak hizkuntza guztietan hobetzen ez diren arren, esperimentu horietan sortutako analisiak konbinatzeak emaitza guztietan hobekuntza esanguratsuak lortzeko balio izan digu, baita parte hartu dugun ataza partekatuan emaitza onak erdiesteko ere (ikus 5.5 irudia). Horrenbestez, esan daiteke eskura izan ditugun ezaugarriak analizatzaileari era desberdinetan pasatuz lortutako analisiak beraien artean osagarriak direla, hots, analizatzaileak analisi bakoitzean jasotako informazioari probetxua atera dio emaitzak hobetzeko. Azkenik, aipatu beharra dago hurbilpen hori konputazionalki garestiagoa dela oinarrizko konfigurazioan erabiltzen dena baino.



5.5 irudia – Ataza partekatuan parte hartu duten sistemen emaitzak, gure sistema BASQUE_TEAM da (kolore morea).

5.5 Multzokatzea

Multzokatzea (*clustering*) elementu edo objektu ezberdinen sailkapen teknika bat da zeinetan elementuak beraien ezaugarrien arabera multzo ezberdinetara biltzen diren. Multzo horietan sartzen diren elementuak erabilitako algoritmoaren arabera doaz aldatzen, eta oso arrunta da bi algoritmo desberdinek sortutako multzoek inolako zerikusirik ez edukitzea beraien artean. Bestal-

de, multzo berean dauden elementuek beraien artean antzekotasun handiagoa izaten dute beste multzoetako elementuekiko baino. Gehienetan bi elementuen arteko antzekotasuna distantzia formula baten bitartez neurtzen da, elementuen ezaugarri edo datuetan oinarrituz.

Euskara bezala morfologikoki aberatsak diren hizkuntzetan lema bakoitzak hainbat hitz-forma izan ditzake (*sparseness*, hemendik aurrera urritasuna) eta hori arazo bat da analizatzaile sintaktikoentzako, hitz bakoitzari dependentzia etiketa bat esleitzeko garaian beraien ataza zailagoa bihurtzen baita. Candito eta Seddah-ren arabera multzokatzeak arazo horri aurre egiten laguntzen du.

> "Clustering words seems useful as a way of addressing the lexical data sparseness problem, since counts on clusters are more reliable and lead to better probability estimates." (Candito eta Seddah 2010)

5.5.1 Sarrera

Azken urteetan gero eta gehiago erabiltzen ari dira multzokatze teknikak Hizkuntzaren Prozesamenduko (HP) hainbat atazatan, lan askotan emaitza interesgarriak lortu direlako teknika horien bitartez. Jarraian gure lanarekin erlazioa duten zenbait lanen berri emango dugu gure hurbilpena azaldu aurretik.

Miller *et al.* (2004) lanean ikasketarako etiketatutako datu kopurua handitzeko etiketatu gabeko corpus handi batetik eratorritako hitzen multzokatzea erabiltzen da. Teknika hori ebaluatu ondoren, izenak automatikoki etiketatzeko sistema baten emaitza berdinak lortzen zirela frogatu zen, baina azken honek behar duen etiketatutako informazioaren % 13a soilik erabiliz.

Koo *et al.* (2008) lanean hitz multzoen edo *cluster*-en erabilera dependentzietan oinarritutako analisi sintaktikorako nahiko lagungarria dela frogatzen dute. Bertan, dependentzietan oinarritzen diren analizatzaileak trebatzeko metodo erdi-gainbegiratua aurkezten dute, analisian etiketatu gabeko corpus handi batetik eratorritako hitz multzoak ezaugarri bezala erabiliz. Beraien esperimentuak Penn Treebank eta Prague Dependency Treebank-ean landu dituzte, eta multzoetan oinarritutako ezaugarriak gehituz hobekuntza esanguratsuak erdiesten dituzte multzoak erabili gabeko konfigurazioarekiko. Ingelesaren kasuan % 1,14ko gorakada lortzen dute eta % 1ekoa txekieraren kasuan.

Kim *et al.* (2014) lanean, aldiz, koreeran, morfologikoki aberatsa den hizkuntza eranskarian, hitzak morfemetan banatzen dituzte rol semantikoak automatikoki etiketatzen dituen sistema batean erabiltzeko. Modu honetara, rol semantikoen etiketatzean (*Semantic Role Labeling*) hizkuntza eranskari baterako lortu den emaitzarik hoberena (% 81,07 F1) erdiesten dute rol semantikoak automatikoki etiketatzen dituen beraien sistemarekin.

Nahiz eta sintaxian hitz bektoreak erabiltzea nahiko berria den, azken urteetan hainbat autorek oso ideia interesgarriak azaldu dituzte ataza desberdinetan aplikatu ahal izateko. Hitz bektoreekin asko esperimentatu duen egile bat Mikolov da (Mikolov *et al.* 2013c, b, a). Denboran zehar, hitzen irudikapen jarraitua egiten duten hainbat eredu azaldu dira. Hala ere, Mikolovek eta bere lankideek neurona sareekin ikasitako hitzen irudikapena landu dute, hitzen arteko erregulartasun linealak hobeto mantentzen direla ikusi baitzuten aurretik. Hitzen irudikapenak ikasteko konplexutasun konputazionala minimizatzeko bi arkitektura eredu berri aurkeztu dituzte: *Continuous Bag-of-Words* (CBOW) eredua eta *Skip-gram* eredua. CBOW ereduan, leiho zehatz bat kontuan hartuta uneko hitza asmatzen da bere inguruan dauden hitzetan oinarrituz, eta *Skip-gram* ereduan kontrakoa, hau da, uneko hitza jakinda bere testuingurua zein den asmatzea. Bi eredu horiek erabiltzen dituen neurona sarea gai da minutu gutxitan milioika hitz dituen corpus batean azaltzen diren hitz bakoitzarentzako bektore bat sortzeko.

Azkenik, Andreas eta Klein (2014) lanean, hitz-bektoreek osagaietan oinarritutako analisi sintaktiko automatikoan laguntzeko zenbait bide aztertzen dituzte. Hitz-bektoreak osagaietan oinarritutako analisi sintaktikoan oso lagungarriak izan ez diren arren, dependentzietan oinarritutako analisi sintaktiko automatikoan lagun dezaketela uste dute egileek, osagaietan oinarritutako corpusetan esplizituki ematen den abstrakzio sintaktikoa lortu daitekeelako bektoreen bidez.

Gure hurbilpenean bi multzokatze teknika erabili ditugu: Brown multzokatzea eta hitz-bektoreetan oinarritzen den K-means multzokatzea. Brown multzoak aukeratzeko arrazoia aurretik aipatu dugun urritasunaren arazoari aurre egiteko baliagarria dela frogatu delako da (Koo *et al.*, 2008). Hitz-bektoreetan oinarritzen diren K-means multzoak erabiltzearen arrazoia, ostera, dependentzien analisian oso gutxitan aplikatu direla da, hots, beraien ekarpena zehaztu nahi da aukeratu ditugun bost hizkuntzetan. Bestalde, konbinaketak egiteko garaian oinarri desberdineko analisiak izatea garrantzitsua denez, izaera desberdineko multzoak aplikatuta erdietsitako analisiak sortzea baliagarria izango zaigula uste dugu.

5.5.2 Gure hurbilpena

Atal honetan, multzokatze teknika desberdinek eskaintzen dituzten aukerak aztertu ditugu morfologikoki aberatsak diren hizkuntzetako analisi sintaktikoa hobetzeko. Horretarako ondorengo atazak landu dira hautatutako hizkuntza bakoitzerako:

- Aukeratutako multzokatze teknika bakoitza hitzen gainean aplikatu ondoren sortutako multzoek analisi sintaktikoan duten eragina neurtu.
- Multzokatze teknika desberdinetatik eratorritako ezaugarriak aplikatu ondoren lortutako analisi sintaktiko desberdinak konbinatu, modu bakoitzean lortutako analisiak beraien artean osagarriak diren zehazteko.

5.5.3 Esperimentuak eta emaitzak

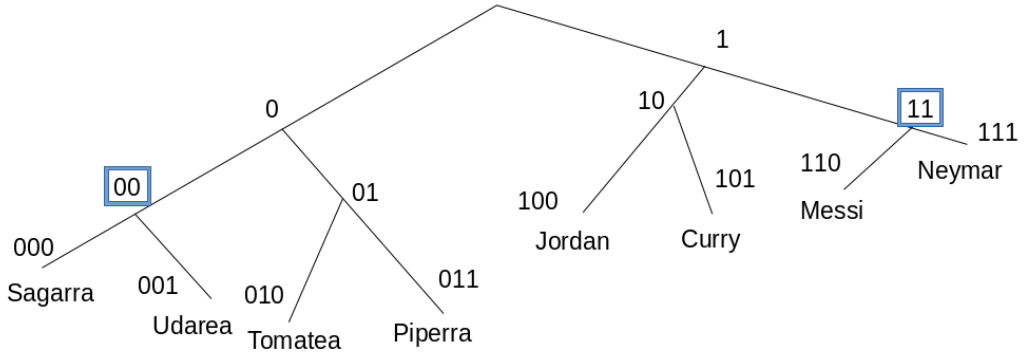
Lan honetan, 3.2.1.2 atalean azaldutako oinarrizko zuhaitz-bankuak erabili dira, baita erdi-gainbegiratutako zuhaitz-bankuetatik ateratako testu hutsa ere. Izan ere, Brown multzoak eta hitz-bektoreen gaineko K-means multzoak sortzeko beharrezkoak baitira etiketatu gabeko corpusak. Multzo mota bakoitza sortzeko, kasu bakoitzean eskuragarri izan dugun corpusik handiena erabili dugu.

Atal honetan, egin ditugun esperimentuak modu zehatzagoan aurkeztuko ditugu erdietsitako emaitzekin batera. Hizkuntza bakoitzerako multzokatze mota bakoitzak izan duen eragina ikusiko dugu lehenik eta azkenik aurreko emaitzak konbinatuz lortu ditugun emaitzak aurkeztuko ditugu.

5.5.3.1 Brown Multzokatze Morfologikoa

Brown multzokatzea modu arrakastatsuan erabili da dependentzietan oinarritutako analisi sintaktikoan (Haffari *et al.* 2011; Koo *et al.* 2008), entitate izendunen ezagutzan (Turian *et al.* 2010) eta galdera erantzun sistemetan (Momtazi eta Klakow 2009). Algoritmo honen emaitza zuhaitz bitar bat da, non hitz bakoitzari errotik hasita berarengana iristeko bidea zehazten duen bit kate bat dagokion (ikus 3.2.2.3).

Demagun testu baten gainean aplikatu dugula Brown multzokatzea eta honek emaitza bezala 5.6 irudian agertzen den zuhaitz bitarra sortu duela. Zuhaitz horretako adabegi batek (00 bit kateari dagokiona) bi ume ditu (000 eta 001) eta hauek *sagarra* eta *udarea* hitzei esleituta daude. Bestalde,



5.6 irudia – Brown multzokatzearen adibidea.

beste adabegi batek (11 bit kateari dagokiona) bi ume ditu (110 eta 111) eta hauek *Messi* eta *Neymar* hitzei lotuta daude. Orduan, esan dezakegu 00 bit kateari dagokion multzoak frutak dituela barnean eta 11 bit kateari dagokionak futboleko jokalariak. Esperimentuetan mota honetako jakintza gehitu nahi diogu analizatzaile sintaktikoari emaitzak hobetzeko, orokortze gaitasun horrek ikasketa corpusaren tamainak dituen mugak gainditzen lagunduko duelakoan.

Bestalde, Brown multzoen emaitzak modu desberdinetan erabiltzeko aukerak daude. Hitz bakoitzari dagokion bit katearekin jokatu daiteke orokortze maila desberdinak lortu nahi badira. Bit guztiak erabiliz gero, hitzaren errepresentazio zehatzena lortzen da, eta zenbat eta gutxiago erabili (bitak eskuinetik ezkerrera kentzen dira), orduan eta orokortze altuago lortzen da. Koo *et al.* (2008) lanean bit kateen luzera desberdinak erabiltzen dituzte orokortze maila bakoitzak eskaintzen dituen onurak neurtzeko. Beraien esanetan, oso garrantzitsua da bit katearen luzera egokia aurkitzea emaitza onak erdietsi nahi badira analisi sintaktikoan. Hainbat probaren ostean, ondorioztatu dute bit kate motzak erabiltzea dela (4-6 bit) aukerarik egokiena hitzaren kategoria ordezkatzeko, eta bit kate osoa hitz-forma ordezkatzeko. Hau jakinda, guk ere probak egin ditugu 3, 4, 5 eta 6 biteko luzerako kateekin, eta emaitzarik hoberena 4 biteko kateekin lortu dugu, orokortze maila egokiena eskaintzen digun luzera, hain zuzen ere. Sortuko diren multzoen kopurua zehazteko ere metodologia bera erabili dugu. Multzo kopuru desberdinekin egindako hainbat probaren ostean (200, 400, 500, 600, 700, 800, 1.000, 2.000, ...) emaitzarik hoberena 800 multzorekin lortu da. Aipatzekoa da atal honetako esperimentuetan erabili diren Brown multzo guztiak

sortzeko Liang (2005) inplementazioa erabili dugula.

Ondorengo lerroetan, adibide bezala euskarazko hitzak erabiliz, analizatzaile sintaktikoari ezaugarrien zutabean gehituko dizkiogun ezaugarri berriak nola sortu ditugun deskribatuko dugu:

– **Brown multzoak sortu**: Brown algoritmoa aplikatu dugu hizkuntza bakoitzerako 800 multzo sortzeko:

Etxera[10100] *noa*[1000000] *gaur*[1010101] *gauean*[111111]

– **Hitz bakoitzeko bi ezaugarri berri lortu**: Ikasketarako zuhaitzbankuko eta ebaluatuko den zuhaitz-bankuko hitz bakoitzari bi ezaugarri berri gehituko zaizkio ezaugarrien zutabean, adibide bezala *etxera* hitza erabiliko dugu:

  – **Hitzaren bit kateko lehenengo 4 bitak:** 1010
  – **Hitzaren bit kate osoa:** 10100

Gehitutako ezaugarriek hizkuntza bakoitzean izan duten eragina islatzen duten emaitzak 5.3 taulan aurkeztu ditugu. Emaitzak aztertzen baditugu, orokorrean sartutako ezaugarriek eragin positiboa dutela ikusten da. Frantsesa da ezaugarri berriekin hobekuntzarik txikiena erakutsi duen hizkuntza, hobekuntza horiek ez direlarik estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$). Hungariera, ostera, hobekuntzarik handienak lortu dituen hizkuntza da eta kasu honetan hobekuntza horiek estatistikoki esanguratsuak dira. Euskarak ere hobekuntza estatistikoki esanguratsuak erdietsi ditu oinarrizko emaitzekin alderatzen bada.

Alemanean lortutako emaitzak bereziki interesgarriak dira, oso zaila baita hain oinarrizko emaitza altuak dituen hizkuntza batean horiek gainditzea eta hobekuntza estatistikoki esanguratsuak lortzea. Suedieran, berriz, Mate analizatzailearekin bakarrik lortu da oinarrizko emaitza gainditzea, nahiz eta hobekuntza hori ez den estatistikoki esanguratsua, hobekuntza nahiko positiboa dela esan behar da kontuan hartzen bada beste hizkuntzekin konparatuta multzokatze teknikak aplikatzeko corpus txikia duela.

Bestalde, emaitza hauek ezaugarrien ingeniaritzako esperimentuetan aurkeztu ditugunekin alderatzen baditugu, esan daiteke orokorrean Brown multzokatzearekin erdietsitako emaitzak (ikus 5.2 taula) hobeak direla (frantseserako izan ezik). Dena den, konparaketa osoa egin ahal izateko, konbinaketaren atalean lortutako emaitzak aztertu behar dira.

	Eus	Fra	Ale	Hun	Sue
Oinarriak					
MaltOptimizer	80,0	79,9	87,6	77,2	73,4
Mate	83,0	84,2	91,0	82,8	76,7
Brown Multzoak					
MaltOptimizer	80,9 (+ 0,9)	80,1 (+ 0,2)	88,0 (+ 0,4)	78,4 (+ 1,2)	73,4 (+ 0,0)
Mate	83,6 (+ 0,6)	84,4 (+ 0,2)	91,8 (+ 0,8)	83,7 (+ 0,9)	77,2 (+ 0,5)

5.3 taula – Brown multzoekin egindako esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

Azaldutako esperimentuetan Brown multzoak lortzeko hitz-formak bere horretan erabili dira. Hala ere, Brown multzoak lortzeko garaian hitz-formak lema eta atzizkietan banatuta probatzea ere pentsatu dugu, morfologikoki aberatsak diren hizkuntzetan analizatzaile sintaktikoei oso zaila egiten baitzaie lema bera duten forma desberdinen (etxeari, etxera, etxetik, etxearentzat...) kopuru handiak kudeatzea. Esan beharra dago multzokatze morfologikoa euskara, alemana eta hungarieran bakarrik aplikatu dela, frantsesean eta suedieran oso zaila edo ezinezkoa delako lema eta atzizkia banatzea eskuragarri ditugun tresnekin.

Ondorengo lerroetan, adibide bezala euskarazko hitzak erabiliz, analizatzaile sintaktikoari gehituko dizkiogun ezaugarri berriak nola sortu ditugun deskribatuko dugu:

- **Hitzak banatu**: Hitzak lema eta atzizkietan banatu ditugu:

  Etxera noa gaur gauean => Etxe ra noa gaur gau ean

- **Brown multzoak sortu**: Brown algoritmoa aplikatu dugu hizkuntza bakoitzerako 800 multzo sortzeko:

  Etxe[10100] *ra*[1000] *noa*[1000000] *gaur*[1010101] *gau*[111111] *ean*[1000]

- **Hitz bakoitzeko hiru ezaugarri berri lortu**: Analizatzaile sintaktikoak erabiliko duen ikasketa zuhaitz-bankuko hitz bakoitzeko eta analizatuko den zuhaitz-bankuko hitz bakoitzeko hiru ezaugarri berri gehituko dira ezaugarrien zutabean, adibide bezala *etxera* hitza erabiliko dugu:

  - **Lemaren bit kateko lehenengo 4 bitak:** 1010

– **Lemaren bit kate osoa:** 10100

– **Hitzaren atzizkia:** ra

Hiru ezaugarri horiekin batez ere lema orokortu nahi da analizatzailearen lana errazteko. Gehitutako ezaugarriek hizkuntza bakoitzean izan duten eragina islatzen duten emaitzak 5.4 taulan aurkeztu ditugu. Erabilitako hizkuntzen artean euskara da hobekuntzarik handienak lortu dituena. Alemanean, berriz, ez da gorakada handirik erdietsi oinarrizko konfigurazioarekin lortutakoarekin konparatuta, baina oinarrizko emaitzarik hoberena duen hizkuntza dela kontuan hartuta, hobetze-tarte txikiena duen hizkuntza da alemana. Hungarieran, Brown multzokatze morfologikoarekin sortu ditugun ezaugarri berriekin hobekuntza txikiak ikusten dira bi analizatzaile sintaktikoetan. Aipatzekoa da hizkuntza desberdinetan erdietsitako hobekuntzak ez direla estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$). Bestalde, emaitzak ikusita, Brown multzokatzea egiteko garaian hitz-formak ez banatzea hobe dela esan daiteke analisi sintaktikoa hobetu nahi bada.

Ezaugarrien ingeniaritzako esperimentuen atalean egindako esperimentuekin alderatuta, banakako esperimentuetan orokorrean emaitza hobeak lortzen dira Brown multzoak erabilita. Hala ere, multzoak erabiliz erdietsitako konbinaketa esperimentuen emaitzak aztertu behar dira ezaugarrien ingeniaritzako konbinaketa esperimentuetan lortutako emaitzekin konparaketa osoa egiteko.

	Eus	**Ale**	**Hun**
Oinarriak			
MaltOptimizer	80,0	87,6	77,2
Mate	83,0	91,0	82,8
Brown Multzoak			
MaltOptimizer	80,5 (+ 0,5)	87,7 (+ 0,1)	77,5 (+ 0,3)
Mate	83,4 (+ 0,4)	91,1 (+ 0,1)	82,7 (- 0,1)

5.4 taula – Brown multzoekin morfemetan banatutako hitzen gainean egindako esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

5.5.3.2 Hitz-bektoreen gaineko K-means multzokatze morfologikoa

Mikolov *et al.* (2013b) lanean azaltzen den bezala, hitz-bektoreak erabiliz hitzen arteko antzekotasun semantiko nahiz sintaktikoa modu errazean aurki daitezke. Hori frogatzeko 150 milioi hitzeko euskarazko corpus batetik eratorritako hitz-bektoreen gainean zenbait esperimentu egin ditugu, hitz-bektoreak eta hitz-multzoak sortzeko *word2vec* (Mikolov *et al.* 2013c) tresna erabili delarik. Adibidez, kosinu distantzia erabiliz (ikus 3.3 atala) *Bilbo* hitzaren hitz hurbilenak zeintzuk diren kalkulatu dugunean, 3 hitz hurbilenak euskal herriko hiriak izan dira. Hori gutxi balitz, 40 hitz hurbilenak munduko hiriak izan dira. Horrek hitz-bektoreek informazio semantikoa barneratuta daukatela frogatzen du eta guk informazio hori erabili nahi dugu sintaxian lagunduko duelakoan.

Ildo beretik jarraituz, esperimentu berdina errepikatu dugu, oraingoan *aitaren* hitza erabili da honen hitz hurbilenak aurkitzeko. Lortu ditugun emaitzak nahiko itxaropentsuak izan dira: hitz hurbilenen artean gehienak bilatutako hitzarekin semantikoki erlazionatuta daude (*amaren*, *arrebaren*, *anaiaren*, *amonaren*, *aitonaren*...) eta guztiek daramate *-ren* atzizkia.

Hitz-bektoreekin egindako esperimentuen helburua, hitz-bektoreetatik eratorritako ezaugarriak erabiltzea da analizatzaile sintaktikoek ikuspegi semantiko nahiz sintaktiko zabalagoa izan dezaten. Erabili beharreko multzo kopurua zehazteko balio desberdinekin hainbat proba egin dira (50, 100, 200, 300, 400, 500, 800, 1.000, 2.000, ...) eta 500 multzorekin lortu dira emaitzarik hoberenak.

Ondorengo lerroetan, adibide bezala euskarazko hitzak erabiliz, analizatzaile sintaktikoari gehituko dizkiogun hitz-bektoreetatik eratorritako ezaugarri berriak nola sortu ditugun deskribatuko dugu:

- **K-means multzoak sortu**: Hizkuntza bakoitzerako dugun etiketatu gabeko corpus erraldoietatik (50-250 milioi hitz hizkuntzaren arabera) hitz-bektoreak eratorri ditugu. Ondoren, 500 multzo sortu dira K-means algoritmoa aplikatuz hitz-bektoreetan:

  Etxera[134] *noa*[78] *gaur*[87] *gauean*[87]

- **Hitz bakoitzeko ezaugarri berria lortu**: Ikasketa zuhaitz-bankuko hitz bakoitzeko eta analizatuko den zuhaitz-bankuko hitz bakoitzeko

ezaugarri berri bat gehituko da. Hitza corpusean ez badago ez zaio ezer gehitzen. Adibide bezala *etxera* hitza erabiliko dugu:

– **Hitzaren multzoa:** 134

Sortu diren ezaugarri berriek analisi sintaktikoan duten eragina 5.5 taulan aurkeztu dugu. Emaitzei erreparatzen badiegu, orokorrean sartu ditugun ezaugarri berriek eragin positiborik ez dutela konturatuko gara eta erdietsi diren hobekuntza gehienak ez dira estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$). Suedieran lortu da hobekuntza handiena Mate analizatzaileari ezaugarri berriak gehituta lortzen den emaitza oinarrizko konfigurazioarekin erdietsitakoarekin alderatzen bada. Gertaera hori suedierako ikasketarako zuhaitz-bankuaren tamaina txikiaren ondorio bat izan daiteke. Izan ere, ikasketarako corpusetik jasotzen ez duen informazioa ezaugarri berrietatik jaso baitezake. Hala ere, hobekuntza analizatzaile bakarrarekin gertatu denez, esandakoa hipotesi bat besterik ez da eta etorkizunean aztertu beharko litzateke benetako arrazoia zein den.

Multzokatzearen atalean erabili ditugu bi multzokatze desberdinak konparatzen baditugu, esan daiteke Brown multzokatzearekin emaitza hobeak lortu direla hitz-bektoreen gaineko K-means multzokatzearekin baino. Bestalde, hungarieran eta suedieran K-means multzokatzearekin ezaugarrien ingeniaritzako bakarkako esperimentuetan erdietsi diren emaitzak gainditzen dira, baina kontrakoa gertatzen da euskara, frantsesa eta alemanean, hau da, ezaugarrien ingeniaritzako bakarkako esperimentu (ikus 5.2) hoberenek emaitza hobeak lortzen dituzte. Hori dela eta, ez da erraza orokorrean zein ikuspuntu den eraginkorragoa esatea, hizkuntza batzuentzat eraginkorragoa baita ezaugarrien ingeniaritza aplikatzea eta beste batzuentzat K-means multzokatzea erabiltzea.

	Eus	**Fra**	**Ale**	**Hun**	**Sue**
Oinarriak					
MaltOptimizer	80,0	79,9	87,6	77,2	73,4
Mate	83,0	84,2	91,0	82,8	76,7
K-means Multzoak					
MaltOptimizer	79,9 (- 0,1)	79,9 (+ 0,0)	87,5 (- 0,1)	76,9 (- 0,3)	73,4 (+ 0,0)
Mate	83,0 (+ 0,0)	84,3 (+ 0,1)	91,1 (+ 0,1)	83,0 (+ 0,2)	77,2 (+ 0,5)

5.5 taula – K-means multzoekin egindako esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

Brown multzoekin egin den antzera, hitz-bektoreetan oinarritutako K-means multzokatze morfologikoa aplikatzea pentsatu dugu euskara, alemana eta hungarieran. Horretarako, hitzak lema eta atzizkietan banatu dira eta hitz-bektoreetan oinarritutako K-means multzokatzea aplikatu zaie. Multzokatze morfologikoaren bidez sortu diren hitz-bektoreetan oinarritutako ezaugarri berriekin erdietsitako emaitzak 5.6 taulan aurkeztu ditugu. Euskararako eta alemanerako jaso ditugun emaitzak aztertuta, esan daiteke sartutako ezaugarriek ez dutela eragin positibo nabarmenik izan analisi sintaktikoan, gehienez % 0,1eko hobekuntza lortu baita. Bestalde, hungarieran % 0,2ko gorakada lortu da MaltOptimizer analizatzaile sintaktikoarekin, baina % 0,2ko beherakada Mate analizatzailearekin. Esan beharra dago hobekuntza horiek ez direla estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$).

	Eus	**Ale**	**Hun**
Oinarriak			
MaltOptimizer	80,0	87,6	77,2
Mate	83,0	91,0	82,8
K-means Multzoak			
MaltOptimizer	80,1 (+ 0,1)	87,7 (+ 0,1)	77,4 (+ 0,2)
Mate	82,9 (- 0,1)	91,1 (+ 0,1)	82,6 (- 0,2)

5.6 taula – K-means multzoekin banatutako hitzen gainean egindako esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

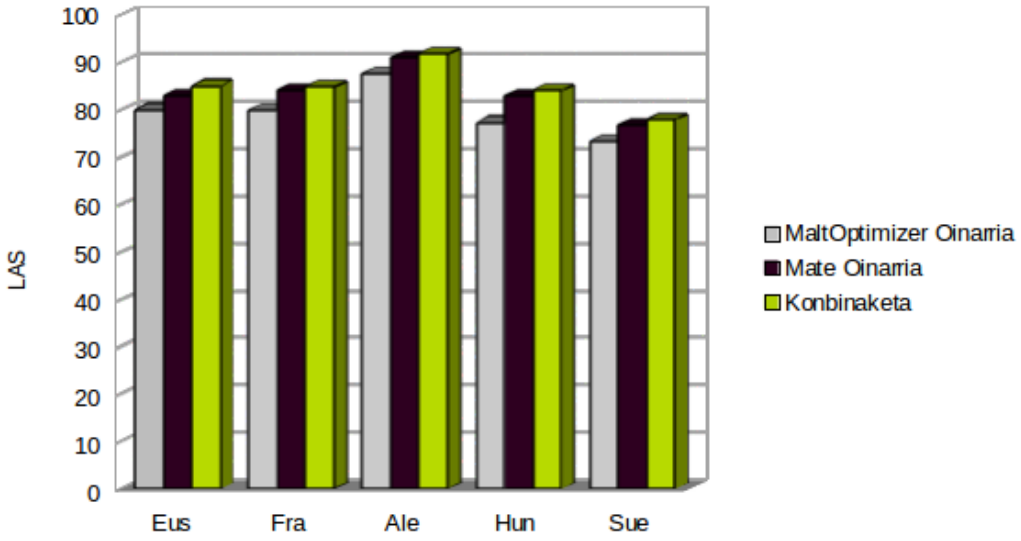
Ezaugarrien ingeniaritzako esperimentuetan egin den bezala, hainbat multzokatze teknika aplikatuta, analisi sintaktikoa hobetu nahi izan da, bai teknika horiek bakarka aplikatuta, bai denak batera konbinatuta. 5.5 eta 5.6 tauletako emaitzek, teknika horien bakarkako aplikazioak eragin positibo nabarmenik izan ez duela erakusten dute. Hala ere, hurrengo azpiatalean azalduko diren konbinaketa esperimentuek multzokatze teknika desberdinek eskaintzen duten abstrakzio semantikoa eta sintaktikoa aprobetxatzea espero da analisi sintaktikoa hobetzeko, ezaugarrien ingeniaritzako esperimentuetan gertatu den antzera.

5.5.3.3 Konbinaketa

Aurreko puntuan ikusi da hitz-bektoreen gaineko K-means multzoetatik eratorritako ezaugarri berriek gehienetan ez dutela lortzen analisi sintaktikoan

eragin positiboa izatea. Hala ere, posible da hasiera batean ezaugarri batzuek analisi sintaktikoan hobekuntzarik ez suposatzea, baina beste ezaugarri batzuekin konbinatzean aurreikusi ez diren emaitzak lortzea. Ondorioz, hasieratik esan den bezala, lortutako analisi sintaktikoak konbinatu egingo ditugu erabili ditugun bi multzokatze moten onurak jasotzeko asmoarekin. Aipatzekoa da konbinatuko diren analisiak banatu gabeko hitzetatik eratorritako multzoekin lortutakoak direla, bakarkako emaitza hobeak erdiesten dituztela ikusi baitugu aurreko puntuetan.

Konbinaketa horrekin erdietsi diren emaitzak 5.7 irudian aurkeztu dira oinarrizko konfigurazioekin egindako esperimentuen emaitzekin batera. Behin baino gehiagotan gertatu den bezala, euskara, sartutako aldaketei probetxu gehien atera dien bi hizkuntzen artean dago; % 5eko hobekuntza MaltOptimizerrekin lortutako oinarrizko emaitzarekiko, eta % 2koa Mateekin jasotakoarekiko. Frantsesean ere % 5eko hobekuntza ikusi da MaltOptimizerrekin erdietsitako oinarrizko emaitzarekiko, eta % 0,7ko gorakada Mateekin lortutakoarekiko. Alemana da hizkuntza guztien artean hobekuntza txikiena erakutsi duena. Hala ere, ez dago aldaketa nabarmenik beste hizkuntzek jasotako emaitzekiko, batez ere, oinarrizko konfigurazioarekin jasotako emaitzarik altuenak dituela kontuan hartzen bada. 4,3 puntuko gorakada dago MaltOptimizerrekin erdietsitako oinarrizko emaitzarekiko, eta % 0,9ko hobekuntza Mateekin lortutakoarekiko. Hungariera, berriz, alemanaren beste muturrean kokatzen da, emaitzarik hoberenak lortu baitira bertan; % 7ko eragin positiboa ikusi da konbinaketa erabiliz MaltOptimizerrekin erdietsitako oinarrizko emaitzarekiko, eta % 1,4eko gorakada Mateekin lortutakoarekiko. Emaitzekin bukatzeko, suedierarekin jaso direnak soilik falta dira. Hizkuntza honetan ere, erabili diren ezaugarri berriek eragin positibo nabarmena izan dute; % 4,5 puntuko gorakada MaltOptimizerrekin lortutako oinarrizko emaitzarekiko, eta % 1,2koa Mateekin erdietsitakoarekiko. Aipatu diren hobekuntza guztiak estatistikoki esanguratsuak dira McNemar testaren arabera ($p < 0{,}05$) Mateen oinarrizko emaitzekiko eta MaltOptimizerren oinarrizko emaitzekiko. Gainera, Brown multzoen eta hitz-bektoreetatik eratorritako K-means multzoen konbinaketarekin Brown multzoekin soilik erdietsitako emaitza guztiak gainditu dira.



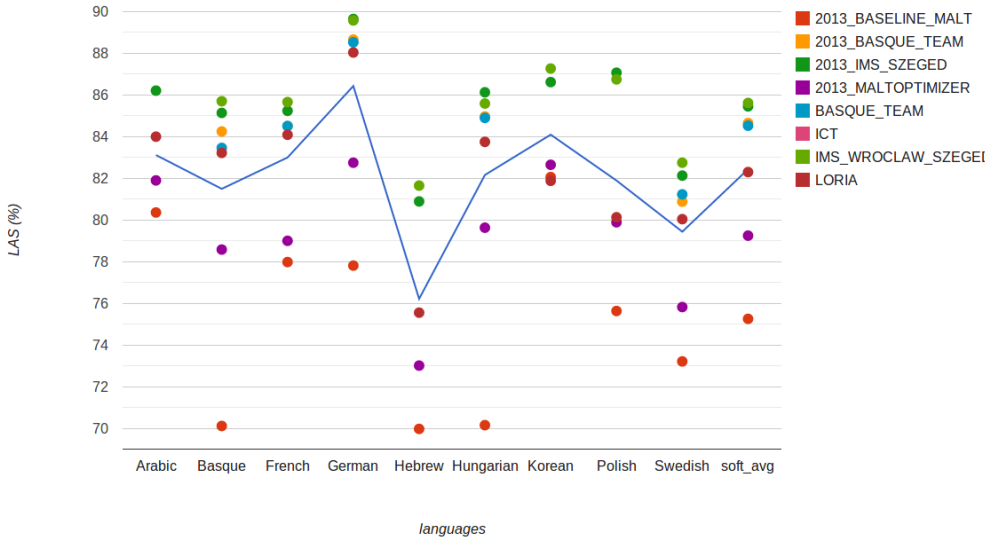
5.7 irudia – Konbinaketa erabiliz egindako esperimentuen emaitzen irudikapena *Labeled Attachment Score* (LAS) neurriaren arabera.

Puntu honetan, konbinaketarekin hizkuntza guztietan emaitzak asko hobetzen direla ikusita, eta multzokatze desberdinetatik eratorritako ezaugarriek bakarkako analisi sintaktikoan lortzen dituzten emaitzak ez direla hain esanguratsuak jakinda, multzokatze teknika desberdinek konbinaketan duten eragina neurtu nahi izan da. Horretarako, konparaketa bat egin da oinarrizko konfigurazioarekin jasotako analisien konbinaketaren (MaltOptimizer eta Mate oinarrien konbinaketa) eta analisi guztien (oinarrizkoak eta multzokatzeetakoak) konbinaketaren artean. Konparaketa horren emaitza 5.7 taulan bildu da. Emaitzak aztertuta, agerikoa da sortu ditugun ezaugarri berriek konbinaketan izan duten eragin positiboa. Euskaran % 1,6ko gorakada ikusten da oinarrizko konbinaketarekin alderatuta. Frantsesean, aldiz, % 0,6eko hobekuntza eta alemanean % 0,9koa. Hungarieran % 1,3ko hobekuntza lortzen da eta suedieran puntu bateko hobekuntza. Gainera, erakutsitako hobekuntza horiek guztiak estatistikoki esanguratsuak dira McNemar testaren arabera ($p < 0{,}05$).

	Eus	Fra	Ale	Hun	Sue
Oinarrizkoak	83,4	84,3	91,0	82,9	76,9
Guztiak	85,0	84,9	91,9	84,2	77,9

5.7 taula – Oinarrizkoen konbinaketaren emaitza eta analisi guztien konbinaketaren emaitza *Labeled Attachment Score* (LAS) neurriaren arabera.

Aurkeztu berri diren emaitzek, sartu diren aldaketek analisi sintaktikoan eragin positiboa dutela iradokitzen dute. Gainera, beraien eragina ez da mugatzen hizkuntza jakin batzuetara bakarrik, hizkuntza guztietan ikusi baitira hobekuntzak. Esandakoak ataza partekatuan emaitza onak lortzea ahalbidetu du (ikus 5.8 irudia).



5.8 irudia – Ataza partekatuan parte hartu duten sistemen emaitzak aurreko urtean lortutako emaitzekin batera, gure sistema BASQUE_TEAM da (kolore urdina).

Multzokatzearen atal honekin bukatzeko, 5.8 taulan multzokatzearen konbinaketa esperimentuetan erdietsi diren emaitzak eta ezaugarrien ingeniaritzako konbinaketa esperimentuetan lortutako emaitzak bildu ditugu. Aipatutako bi ikuspegiek ezaugarri morfologikoak era desberdinetan erabiltzen dituztenez, konparaketa bat egin nahi da beraien arteko antzekotasunak eta

desberdintasunak argitzeko eta bi ikuspegi horien artean eraginkorrena zein den zehazteko.

	Eus	**Fra**	**Ale**	**Hun**	**Sue**
Ezaugarrien ingeniaritza	86,2	85,1	91,8	84,1	78,1
Multzokatzea	85,0	84,9	91,9	84,2	77,9

5.8 taula – Ezaugarrien ingeniaritzaren eta multzokatzearen konbinaketa esperimentuen emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera.

5.8 taulako emaitzak aztertzen badira, hizkuntza gehienetan bi emaitzen artean desberdintasun nabarmenik ez dagoela ikusten da. Euskara da bi emaitzen artean alde nabarmena erakusten duen hizkuntza bakarra. Izan ere, ezaugarrien ingeniaritzako konbinaketa esperimentuan % 1,2 puntu gehiago lortzen baitira multzokatzearen konbinaketa esperimentuan baino. Gainera, alde hori estatistiko esanguratsua da McNemar testaren arabera ($p < 0{,}05$).

Emaitzei soilik begiratuta, esan daiteke ezaugarrien ingeniaritzako esperimentuetan erabilitako ezaugarriak orokorrean eraginkorragoak direla multzokatzearen esperimentuetan erabilitako ezaugarriak baino, antzeko emaitzak lortzen direlako hizkuntza gehienetan eta euskaraz emaitza hobeak lortzen direlako. Hala ere, ez da ahaztu behar multzokatzearen konbinaketa esperimentuan 6 elementu soilik konbinatu direla eta ezaugarrien ingeniaritzaren konbinaketa esperimentuan 14. Hori dela eta, esan daiteke multzokatzearen konbinaketaren emaitza erdiestea baliabide eta denbora aldetik merkeago dela. Ondorioz, emaitzak lortzeko denbora garrantzitsua bada, multzokatzearen konbinaketa egitea gomendatzen da.

5.6 Meta-Ezaugarriak

Ikasketa gainbegiratua aplikatzen den ataza askotan erabilitako sailkatzaileek hainbat ezaugarri jasotzen dituzte sarrera gisa, ahalik eta informazio gehien izateko sailkapenerako. Sailkatzaileek aukeraketa bat egiten dute ezaugarri guztien artean, sailkapena ahalik eta zuzenena izateko eta konplexutasun konputazionala murrizteko. Dena den, modu horretan hautatutako ezaugarriak besteetatik bereizten dituzten propietateak zeintzuk diren jakitea oso zaila da, eta sailkatzaileak eraginkortasuna galtzen du orain arte agertu ez

diren ezaugarriak jasotzen dituenean. Informazioa aberasteko hainbat modu aurkeztuko dira atal honetan, horien artean meta-ezaugarriak.

Meta-ezaugarriak oinarrizko ezaugarrietatik eratortzen diren eta horien arteko erlazio edo propietateak finkatzen dituzten ezaugarri bereziak dira. Ezaugarri berezi horiekin, bestela lortuko ez litzatekeen orokortze maila lortzea espero da, eta ezaugarri ezezagunen aurrean, sailkatzaileek aparteko baliabidea erabiltzeko aukera izan dezakete. Chen *et al.* (2013) lanean aukezten den bezala, meta-ezaugarriak oso egokiak dira ezaugarri ezezagunak kudeatzeko:

> "Further analysis indicate that the meta features are very effective in processing the unknown features." (Chen *et al.* 2013)

5.6.1 Sarrera

Analisi sintaktiko gainbegiratuak, Hizkuntzaren Prozesamenduko hainbat atazatan gertatzen den bezala, datuen botila-lepoaren arazoa pairatzen du. Ikasketa algoritmoek sortzen dituzten ereduetan orain arte ikusi ez diren hitzen parametrizazioa ez da oso zehatza izaten orokorrean, eta honek errendimendu galera suposatzen du. Nahiz eta hainbat lanetan zuhaitz-bankuko informazioa aberasteko iturri desberdinak erabili izan diren, oraindik ez dago argi informazio gehigarri hori modu eraginkorrean nola gehitu behar den analizatzaile sintaktikoaren emaitzak hobetzeko.

Ohikoa da zuhaitz-bankutik kanpoko informazioa erabiltzeko algoritmo ez-gainbegiratuen bidez hitzen errepresentazio desberdinak sortzea. Gehien erabiltzen direnak hitz multzoak (Koo *et al.* 2008; Candito eta Seddah 2010; Haffari *et al.* 2011; Täckström *et al.* 2012) eta hitz-bektoreak (Andreas eta Klein 2014; Bansal *et al.* 2014) dira. Bestalde, lan askotan WordNet bezalako datu-base lexiko-semantikoetatik ateratako informazioa erabiltzen dute (Agirre *et al.* 2008; Bengoetxea *et al.* 2014).

Ildo berari jarraituz, Bengoetxea *et al.* (2014) lanean, informazio semantiko ez-gainbegiratuak artearen egoerako zenbait analizatzaile sintaktikoko emaitzetan duen eragina neurtzen da. Lan horretan, WordNet2.1etik jasotako *synset*a eta fitxategi semantikoak erabiltzen dira Brown multzoekin batera analisi sintaktikoan emaitzak hobetzeko. Aurkeztutako emaitzek hobekuntza txikiak erakusten dituzte orokorrean, eta emaitzarik hoberenak Brown multzoekin lortzen dira MST analizatzaile sintaktikoarekin (McDonald *et al.* 2005, 2006), % 1,12ko gorakada hain zuzen ere.

Azken urteotan, HPko hainbat atazatan aplikatu dira hitz-bektoreak, eta analisi sintaktikoa ez da izan salbuespena. Andreas eta Klein (2014) lanean bide hori aztertzen dute; etiketatu gabeko corpus erraldoietan aurki daitekeen informazioa aprobetxatzeko, hitz-bektoreak sortzen dituzte, osagaietan oinarritutako analizatzaile batek (Petrov eta Klein 2007) erabil ditzan gainontzeko ezaugarriekin batera. Egileen esanetan, hitz-bektoreak erabiltzeko arrazoi nagusia hitz ezezagunei probabilitateak esleitzeko ematen duten laguntza da. Ondorioz, ebaluatu nahi den testuaren zuhaitz-bankuan hitz ezezagun bat aurkitzen duten bakoitzean, hitz horren hitz-bektoretik hurbilen dagoen hitz-bektoreaz ordezkatzen dute. Modu honetara lortutako emaitzetan, hobekuntza txikiak ikusten dira ikasketarako erabili den zuhaitz-bankua txikia denean, eta hobekuntza horiek galdu egiten dira zuhaitz-bankuaren tamaina hazten doan heinean.

Chen *et al.* (2013) lanean ere etiketatu gabeko corpus erraldoitik ateratako informazioa erabiltzen da, baina beraien asmoa morfologikoki aberatsak diren hizkuntza askotan gertatzen den arazo bati aurre egitea da. Beraien proposamena, oinarrizko ezaugarrietatik meta-ezaugarriak deituriko maila altuagoko ezaugarri bereziak sortzea da. Meta-ezaugarrien ideia oinarrizko ezaugarriak beraien maiztasunaren arabera multzokatzea da. Horretarako, tamaina handiko corpus bat analizatzen da sintaktikoki, eta oinarrizko ezaugarrien maiztasunak biltzen dira ondoren. Lau meta-ezaugarri mota sortzen dituzte maiztasunetan oinarrituta: A, T, G eta B (Asko, Tartean, Gutxi eta Bestelakoak). Oinarrizko ezaugarriekin egiten den bezala, meta-ezaugarriak sortzeko ere ordena desberdinak aztertzen dituzte: aita-ume (lehen ordena), aitona-aita-ume (bigarren ordena) eta abar. Meta-ezaugarriak erabilita erdiesten dituzten emaitzen arabera, beraien sistemak erdi-gainbegiratuak diren hainbat sistema gainditzen ditu, beraien artean Koo *et al.* (2008) lanean aurkeztutako sistema.

Hori jakinda, lan honetan meta-ezaugarriak erabiltzea erabaki da morfologikoki aberatsak diren hizkuntzetan aplikatzeko. Erdi-gainbegiratutako zuhaitz-bankuetatik abiatuta iturri desberdinetatik eratorritako (hitz-formak eta multzoak) meta-ezaugarriak sortuko dira. Horretarako zuhaitz-bankuetan ematen diren lotura sintaktikoek berebiziko garrantzia dute, lotura horien maiztasunetatik sortuko baititugu meta-ezaugarriak. Adibidez, meta-ezaugarri mota bat guraso eta umeen hitzen kategoriaren maiztasunen kontaketatik erator daiteke. Erdi-gainbegiratutako zuhaitz-bankuan izen arruntak eta aditzak lotuta askotan agertzen badira, informazio horretatik eratorritako meta-ezaugarriak izen arrunt bat eta aditz bat lotuta egoteko probabilitatea

kalkulatzen lagunduko dio analizatzaileari. Meta-ezaugarriak mota desberdinetako kontaketetatik eratorri daitezke: guraso eta umeen kategorien kontaketetatik, guraso eta umeen hitz-formen kontaketetatik, guraso eta umeen multzoen (cluster) kontaketetatik eta abar.

Meta-ezaugarriekin, beraien artean lotura sendorik ez duten ezaugarrien arteko nolabaiteko loturak sortu nahi dira analizatzaile sintaktikoak ezaugarri guztiak hobeto maneia ditzan. Horretaz gain, meta-ezaugarrien bidez analizatzailea gai izango da kalitate altuagoko ezaugarriak aukeratzeko, honek suposatu dezakeen errendimenduaren igoerarekin.

5.6.2 Gure hurbilpena

Morfologikoki aberatsak diren hizkuntzetarako meta-ezaugarriak analisi sintaktikoan aplikatzeak eskaintzen dituen aukerak aztertzeko ondorengo atazak landu dira hautatutako hizkuntza bakoitzerako:

- Iturri desberdinetako lotura sintaktikoen maiztasunen kontaketetatik eratorritako meta-ezaugarriek analisi sintaktikoan duten eragina neurtu (hitz-forma eta multzoen kontaketa desberdinak).

- Emaitzarik hoberenak lortu dituzten meta-ezaugarrien konbinaketak analisi sintaktikoan duen eragina neurtu.

5.6.3 Esperimentuak eta emaitzak

Lan honetarako 3.2.1.2 atalean azaldutako oinarrizko zuhaitz-bankuetan aldaketak egin ditugu. Ikasketarako zuhaitz-bankuen tamaina 100.000 hitzera mugatu dugu erabiliko den MST analizatzailearen baliabide kontsumoa dela eta. Bestalde, Brown multzoak eta hitz-bektoreen gaineko K-means multzoak sortzeko 3.2.1.2 atalean azaldutako erdi-gainbegiratutako zuhaitz-bankuetatik ateratako testua erabili dugu, kasu bakoitzean eskuragarri izan dugun hitz kopuru handiena erabiliz. Meta-ezaugarrietan zentratzen bagara, 3.2.1.2 atalean azaldutako erdi-gainbegiratutako zuhaitz-bankuak erabili ditugu bere horretan, lotura sintaktiko desberdinen maiztasunen araberako multzoak sortzeko.

Meta-ezaugarriak sortzeko hiru iturri desberdinetan oinarritu gara: hitz-formak, Brown multzoak eta hitz-bektoreetan oinarritutako K-means multzoak. Hori dela eta, erdietsitako emaitzak ere erabili den iturri motaren

arabera banatu dira. Horietako bakoitzean erdi-gainbegiratutako zuhaitzbankuan gertatzen diren lotura desberdinen maiztasunak erabili ditugu meta-ezaugarriak sortzeko ardatz bezala. Hitz-formetan oinarritzen diren esperimentuetan hitz-forma eta hitzen kategorien kontaketak egin dira meta-ezaugarriak sortzeko, Brown multzoetan oinarritutakoetan Brown multzo eta hitzen kategorien kontaketak, eta K-means multzoetan oinarritutako esperimentuetan K-means multzo eta kategorien kontaketak.

Kontaketak egiteko garaian antzeko patroiak bilatu dira erabilitako iturri guztiekin, desberdintasun bakarra erabilitako iturria delarik. 5.9 taulan hitz-formetan oinarritutako meta-ezaugarriak sortzeko bilatu diren patroiak bildu dira. Brown multzoetan eta K-means multzoetan oinarritutako meta-ezaugarriak sortzeko patroiak zeintzuk diren jakiteko hitz-formak Brown multzo eta K-means multzootaz ordezkatzea besterik ez dago.

Meta-ezaugarriak erabiltzeko arrazoi nagusia analizatzaileari exekuzio garaian lotura bakoitzari buruzko informazioa pasatzea da. Esaterako, Brown multzoetan oinarritutako meta-ezaugarri bat sortzeko zuhaitzean guraso diren hitzen multzoen eta ume diren hitzen multzoen maiztasunak kontatu dira, hau da, (guraso multzo)-(ume multzo) patroia bilatu da kontaketak egiteko. Modu horretara, bi multzok aita-ume papera askotan betetzen duten ala ez jakingo dugu eta analizatzaileak aita-ume lotura horri probabilitatea emateko garaian meta-ezaugarri hori kontuan izango du.

Esanak esan, atal honetan egin ditugun esperimentuak modu zehatzagoan aurkeztuko ditugu erdietsitako emaitzekin batera. Hizkuntza bakoitzerako iturri desberdinetatik eratorritako meta-ezaugarri mota bakoitzak analisi sintaktikoan izan duen eragina ikusiko dugu ondorengo puntuetan.

5.6.3.1 Hitz-formetatik eratorritako meta-ezaugarriak

Hizkuntza bakoitzeko corpusean hitz-formetatik eratorritako meta-ezaugarriak sortzeko zenbait kontaketa egin dira bilaketa patroi[2] bakoitzaren maiztasuna zein den jakiteko. Hitz-formekin erlazionatuta dauden bilaketa patroiak 5.9 taulan bildu ditugu, hitzen kategoriekin soilik erlazionatutako patroiekin

[2]Bilaketa patroien izenak modu honetara definitu dira: K letra dutenak kategoriarekin soilik daude erlazionatuta eta H letra dutenak hitz-formarekin nahiz kategoriarekin egon daitezke erlazionatuta. Bukaeran H bat dutenak beste bi bilaketa patroiren bildura dira, adibidez N1H = N1H1 + N1H2. N letraren atzetik datorren zenbakiak bilatzen diren elementu kopurua adierazten du, adibidez N2H1 bilaketa patroian bi elementu hartzen dira kontuan, gurasoaren hitz-forma eta umearen hitz-forma.

(N1K, N2K, N3K eta N4K) batera. Aipatutako kontaketa horiek hitzekin erlazionatutako kontaketekin batera jartzea erabaki dugu elkarrekin oso lotuta daudelako eta kasu horretan banaketa egiteak merezi ez duela uste dugulako. Meta-ezaugarriak sortzeko, bilatutako patroi bakoitzarentzat lortu

Izena	**Bilaketa patroia**
N1K	g_k, n(g, u)
N1H	g_h, n(g, u); g_h, g_k, n(g, u)
N1H1	g_h, n(g, u)
N1H2	g_h, g_k, n(g, u)
N2K	g_k, u_k, n(g, u)
N2H	g_h, u_h, n(g, u); g_h, s_k, n(g, u)
N2H1	g_h, u_h, n(g, u)
N2H2	g_h, u_k, n(g, u)
N3K	g_k, u_k, a_k, n(g, u, a)
N3H	g_h, u_h, a_h, n(g, u, a); u_h, u_{+1k}, a_k, n(g, u, a)
N3H1	g_h, u_h, a_h, n(g, u, a)
N3H2	u_h, u_{+1k}, a_k, n(g, u, a)
N4K	g_k, g_{+1k}, a_k, a_{1+k}, n(g, u, a)
N4H	g_h, g_{+1h}, a_h, a_{1+h}, n(g, u, a); g_h, u_{+1k}, a_k, a_{1+k}, n(g, u, a)
N4H1	g_h, g_{+1h}, a_h, a_{1+h}, n(g, u, a)
N4H2	g_h, u_{+1k}, a_k, a_{1+k}, n(g, u, a)

5.9 taula – Hitz-formetan oinarritutako meta-ezaugarriak sortzeko bilatu diren patroi desberdinak. g=gurasoa, u=umea, a=anaia, n(g, u)=guraso eta umearen arteko loturaren noranzkoa, n(g, u, a)= guraso, ume eta anaiaren loturaren arteko noranzkoa, h=hitza, k=kategoria, +1=unekoaren ondorengoa. Horrela, adibidez, g_{+1k}= gurasoaren ondorengo hitzaren kategoria da.

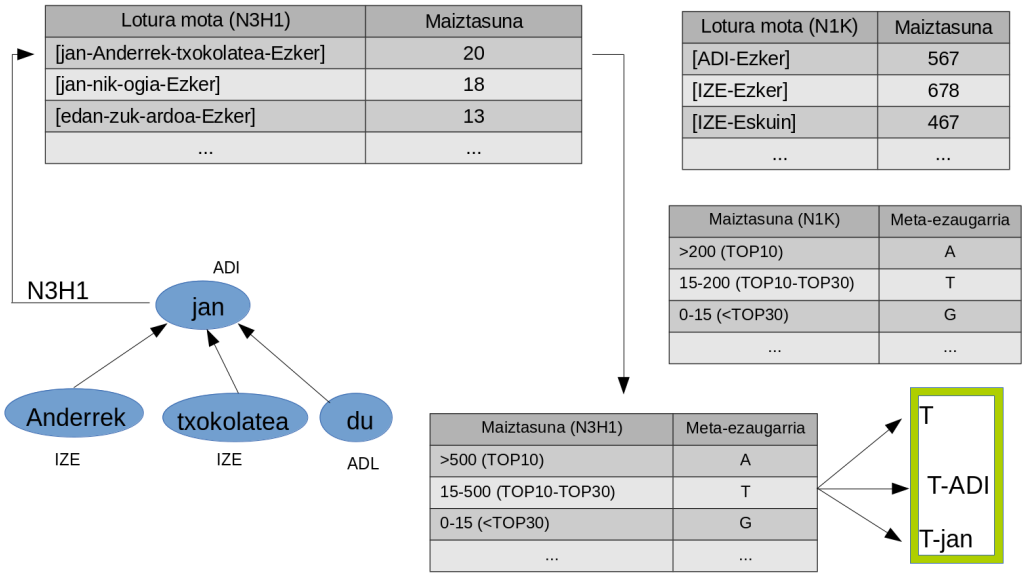
ditugun maiztasun horiek handitik txikira ordenatuko ditugu, eta 5.1 mapaketa funtzioaren arabera maiztasun bakoitzari etiketa bat esleituko zaio, A(asko), T(tartekoa) edo G(gutxi). Funtzio horren arabera, maiztasunaren posizioa (funtzioan P) ordenatutako zerrendan lehenengo % 10en artean badago (TOP10) eta gertaera hori 50 alditan edo gehiagotan gertatu bada (funtzioan Kop), A etiketa esleituko zaio. Lehenengo % 30en (TOP30) eta % 10en (TOP10) artean badago eta 10 alditan edo gehiagotan gertatu ba-

da, T etiketa jasoko du, eta G etiketa bestela. Murriztapen horien esleipena ez da zorizkoa izan eta beraien erabilpena murriztapen maila desberdinekin probak egin ondoren lortutako emaitzen ondorioa da. Esan genezake murriztapen horiekin lortu dugula emaitzen eta sortzen den informazio tamainaren erlaziorik egokiena, zenbat eta murriztapen zorrotzagoa orduan eta txikiagoa baita gorde behar den informazioaren tamaina, eta ondorioz analizatzaile sintaktikoaren lana errazagoa da.

Hitz-formekin erlazionatutako meta-ezaugarriak lortzeko jarraitzen den prozedura argiago ikusteko, adibide batekin azalduko dugu (ikus 5.9 irudia). Suposatu *Anderrek txokolatea jan du* esaldia dugula eta *Anderrek*, *txokolatea* eta *jan* hitz-formen arteko meta-ezaugarriak sortu nahi ditugula. Esaldi horretan *jan* hitza lotura sintaktikoaren gurasoa da, *txokolatea* umea eta *Anderrek txokolatea*-ren ezker anaia hurbilena. Aurretik esan bezala, meta-ezaugarriak lortzeko bilaketa patroien maiztasunak erabiliko ditugu ardatz gisa.

$$\Phi(m) = \begin{cases} A & \text{baldin } P(m) \leq \text{TOP10 eta } Kop(m) \geq 50 \\ T & \text{baldin TOP10} < P(m) \leq \text{TOP30 eta } Kop(m) \geq 10 \\ G & \text{bestela} \end{cases} \tag{5.1}$$

Behin bilaketa patroi bakoitzaren maiztasunari dagokion etiketa dakigula, hiru meta-ezaugarri sortuko ditugu patroi bakoitzarentzat: $\Phi(m)$, $\Phi(m) - g_k$ eta $\Phi(m) - g_h$, non, $\Phi(m)$, 5.1 funtzioaren bidez lortutako etiketa, g_k, lotura sintaktikoko gurasoaren kategoria, eta, g_h, lotura sintaktikoko gurasoaren hitz-forma diren. Hau da, meta-ezaugarri berri horiek guztiak $\Phi(m)$ tarte berdinean daudela adieraziko da. Bilaketa patroi bakoitzari dagokion esperimentuan hiru meta-ezaugarri horiek pasatuko zaizkio analizatzaileari eta horiek duten eragina neurtuko da.



5.9 irudia – Hitz-formetatik meta-ezaugarriak lortzeko prozesuaren irudikapena.

Kasu honetan 5.9 taulan N3H1 izena duen bilaketa patroia erabiliko dugu azalpenerako, g_h, u_h, a_h, n(g, u, a), non gurasoaren, umearen eta anaiaren hitz-forma erabiltzen diren, horien arteko dependentziaren norabidearekin batera. Demagun automatikoki etiketatutako gure corpusean P_s:[jan, txokolatea, Anderrek, Ezker] patroia 20 alditan gertatzen dela eta sailkapenean TOP10 eta TOP30 artean dagoela. 5.1 mapaketa funtzioaren arabera T balioa jasoko luke P_s patroiak. Modu honetara hiru meta-ezaugarri hauek sortuko lirateke: N3H1-T, N3H1-T-ADI eta N3H1-T-jan.

Izena	Eus	Fra	Ale	Hun	Sue
Oinarria	82,69	76,78	85,25	77,65	73,27
N1K	82,74 (+ 0,05)	76,53 (- 0,25)	**85,42 (+ 0,17)**	77,69 (+ 0,04)	73,19 (- 0,08)
N1H	**82,90 (+0,21)**	76,43 (- 0,35)	85,29 (+ 0,04)	77,62 (- 0,03)	73,20 (- 0,07)
N1H1	82,79 (+0,10)	76,53 (- 0,25)	85,36 (+ 0,11)	77,64 (- 0,01)	73,16 (- 0,11)
N1H2	82,77 (+ 0,08)	76,38 (- 0,40)	**85,42 (+ 0,17)**	77,66 (+ 0,01)	**73,56 (+ 0,29)**
N2K	82,41 (- 0,28)	76,35 (- 0,43)	85,34 (+ 0,09)	77,61 (- 0,04)	73,38 (+ 0,11)
N2H	82,66 (- 0,03)	**76,70 (- 0,08)**	**85,43 (+ 0,18)**	**77,84 (+ 0,19)**	73,44 (+ 0,17)
N2H1	82,62 (-0,07)	**76,77 (- 0,01)**	85,40 (+ 0,15)	**78,12 (+ 0,47)**	**73,76 (+ 0,49)**
N2H2	**82,96 (+ 0,27)**	76,51 (- 0,27)	85,33 (+ 0,08)	77,59 (- 0,06)	73,27 (+ 0,00)
N3K	82,69 (+ 0,00)	76,39 (- 0,39)	85,32 (+ 0,07)	77,51 (- 0,14)	73,34 (+ 0,07)
N3H	82,54 (- 0,15)	**76,79 (+ 0,01)**	85,40 (+ 0,15)	**77,71 (+ 0,06)**	73,24 (- 0,03)
N3H1	82,65 (- 0,04)	76,53 (- 0,25)	85,28 (+ 0,03)	77,59 (- 0,06)	73,36 (+ 0,09)
N3H2	82,41 (- 0,28)	76,66 (- 0,12)	85,37 (+ 0,12)	77,54 (- 0,11)	73,35 (+ 0,08)
N4K	**82,95 (+ 0,26)**	76,42 (- 0,36)	85,30 (+ 0,05)	77,60 (- 0,05)	73,38 (+ 0,11)
N4H	82,61 (- 0,08)	76,38 (- 0,40)	85,33 (+ 0,08)	77,51 (- 0,14)	**73,59 (+ 0,32)**
N4H1	82,62 (- 0,07)	76,59 (- 0,19)	85,38 (+ 0,13)	77,70 (+ 0,05)	73,38 (+ 0,11)
N4H2	82,69 (+ 0,0)	76,48 (- 0,30)	85,39 (+ 0,14)	77,69 (+ 0,04)	73,40 (+ 0,13)

5.10 taula – Hitz-formetan oinarritutako bilaketa patroi bakoitzetik eratorritako meta-ezaugarriak erabiliz hizkuntza bakoitzerako lortutako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz hiru emaitza hoberenak.

5.10 taulan, bilaketa patroi bakoitzetik eratorritako meta-ezaugarriak erabiliz erdietsitako emaitzak bildu dira. Euskararako lortu diren emaitzak aztertzen badira, gora-behera handiak ikusten dira. Kasu batzuetan emaitza negatiboak erdiesten dira eta beste batzuetan hobekuntzak. Hala ere, lortzen diren hobekuntzak ez dira estatistikoki esanguratsuak McNemar testaren arabera ($p < 0{,}05$). Hiru emaitzarik hoberenak N1H, N2H2 eta N4K bilaketa patroietatik eratorritako meta-ezaugarriekin lortu dira.

Frantseserako lortu dira emaitzarik baxuenak, hobekuntza ikusten den emaitza bakarra N3H bilaketa patroitik eratorritako meta-ezaugarriekin lortu delarik. Euskararen kasuan gertatu den bezala, hobekuntza hori ez da estatistikoki esanguratsua. Emaitza horiek jasotzeko arrazoia zuhaitz-bankuan bildu diren hitzen kategoriak automatikoki sortzeko erabilitako tresnaren eraginkortasuna izan daiteke. Kategoriak sortzeko tresna hori ez baita euskararen kasuan bezala hizkuntza bakarrean zentratzen, eta horrek askotan emaitzak okertzea dakar. Alemanean frantsesean gertatzen denaren kontrakoa ikusten da, kasu guztietan lortzen baitira hobekuntzak. Hiru emaitzarik hoberenak N1K, N1H2 eta N2H patroietatik eratorritako meta-ezaugarriekin erdietsi dira. Dena den, hobekuntza horiek ez dira estatistikoki esangura-

tsuak eta multzokatze desberdinekin lortutako emaitzak aztertu beharko dira meta-ezaugarriek alemanean duten ekarpena esanguratsua den ala ez ondorioztatzeko.

Hungarierarako taulan bildu ditugun emaitzetan gora-beherak ikusten dira, emaitza positiboak eta negatiboak tartekatzen direlarik. Hiru emaitzarik hoberenak N2H, N2H1 eta N3H bilaketa patroietatik eratorritako meta-ezaugarriekin erdietsi dira. N2H1 patroiaren kasuan ia puntu erdiko hobekuntza ikusten da emaitzetan (+ 0,47), baina gainontzeko emaitza guztietan bezala, hobekuntza hori ez da estatistikoki esanguratsua.

Suedierarako, aldiz, orain arteko hobekuntzarik handiena (+ 0,49) erdietsi da N2H1 patroitik eratorritako meta-ezaugarriak erabilita. Emaitza guztiak aztertuta, nahiz eta hobekuntza estatistikoki esanguratsurik lortu ez den, esan daiteke suediera dela hitz-formetan oinarritutako meta-ezaugarriei probetxu gehien atera dien hizkuntza. Izan ere, emaitza negatibo gutxi eta hobekuntza altuenak baititu.

5.10 taulan bildutakoa ikuspegi eleaniztunetik aztertzen bada, bilaketa patroi batzuk hizkuntza desberdinetan hiru emaitzarik hoberenen artean errepikatzen direla konturatuko gara. Adibidez, N2H patroia frantsesa, alemana eta hungarieran dago hiru emaitzarik hoberenak lortu dituzten patroien artean. N2H1 patroia, berriz, frantsesa, hungariera eta suedieran dago hiru emaitzarik hoberen artean.

Puntu honekin bukatzeko, patroi bat baino gehiagotik eratorritako meta-ezaugarriak erabiltzen dituzten kasuak (N1H= $N1H1$+$N1H2$, N2H= $N2H1$+$N2H2$, N3H= $N3H1$+$N3H2$ eta N4H= $N4H1$+$N4H2$) aztertuta, batutako bi patroiek bakarka lortutako emaitzak baturarekin lortutako emaitzarekin erlaziorik ez duela ikusten da. Izan ere, kasu askotan patroiek bakarka lortu dituzten emaitzen batura patroiak batuta lortutako emaitza baino altuagoa baita.

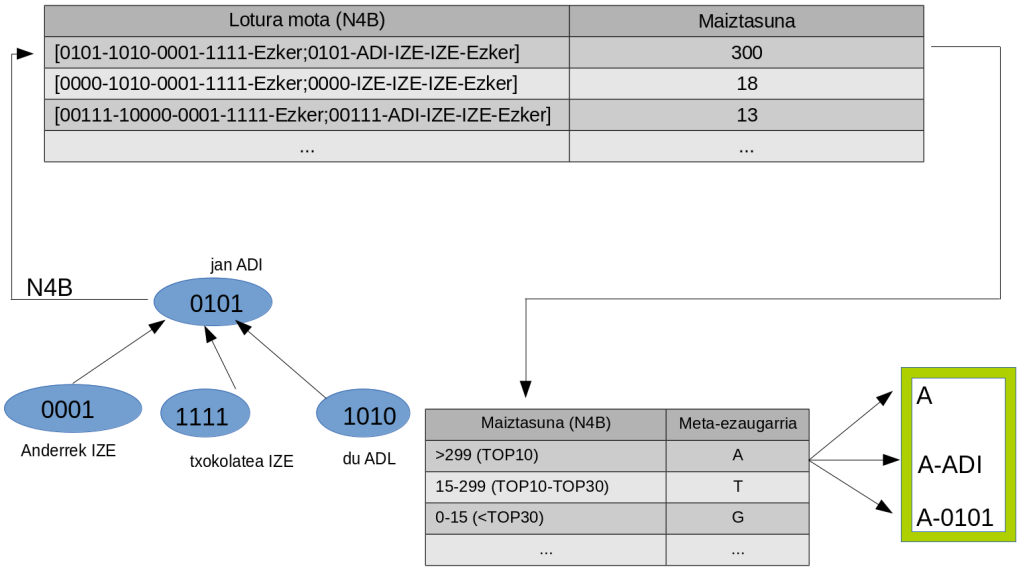
5.6.3.2 Brown multzoetatik eratorritako meta-ezaugarriak

Hitzetatik eratorritako meta-ezaugarriak sortzeko erabili ditugun corpus berak erabili ditugu hizkuntza bakoitzerako. Hala ere, meta-ezaugarriak sortzeko hitzetan oinarritu beharrean Brown multzoetan oinarritu garenez, kasu honetan hitz-forma bakoitzari dagokion Brown multzoa zein den ere jakin beharra dugu, horretarako 3.2.1.2 atalean aipatu ditugun automatikoki sintaktikoki etiketatutako zuhaitz-bankuetatik (erdi-gainbegiratutako zuhaitz-bankuak) ateratako testu hutsa erabili dugularik hizkuntza bakoitzerako. Ka-

su honetan Brown algoritmoak hitz bakoitzerako sortzen duen bit kate osoa erabili da. Brown multzoak erabiltzearen helburu nagusia automatikoki etiketatutako corpusean espresuki agertzen ez diren loturen meta-ezaugarriak lortzea da.

Kasu honetan egin ditugun kontaketak zehazteko 5.9 taulan hitz-formei dagozkien kontaketak Brown multzoei dagozkienekin ordezkatzea besterik ez dugu. Adibidez, N2H txantiloiaren Brown multzoetan oinarritutako bertsioa hau litzateke: N2B = g_b, u_b, n(g, u); g_b, u_k, n(g, u) non, g_b, lotura sintaktikoko gurasoaren hitz-formaren Brown multzoa eta, u_b, umearen hitz-formaren Brown multzoa diren.

Behin gertaera guztien maiztasunak dakizkigunean, aurreko puntuan egin dugun bezala, maiztasun horiek handitik txikira ordenatuko ditugu, eta 5.1 mapaketa funtzioaren arabera gertaera bakoitzaren maiztasun bakoitzari etiketa bat esleituko zaio. Brown multzoetatik eratorritako meta-ezaugarriak sortzeko jarraitu dugun prozedura azaltzeko aurreko puntuan erabili dugun esaldi bera erabiliko dugu, *Anderrek txokolatea jan du*, hain zuzen ere (ikus 5.10 irudia). Orain demagun *Anderrek*, *txokolatea*, *jan* eta *du* hitz-formei dagozkien Brown multzoak 0001, 1111, 0101 eta 1010 direla, hurrenez hurren. Hori esanda, lotura sintaktikoaren gurasoa 0101 multzoa litzateke (*jan* hitz-formari dagokion multzoa), 1111 (*txokolatea*) uneko umea eta 0001 (*Anderrek*) 1111en ezker anaia hurbilena. Puntu honetan, 5.9 taulako N4H txantiloiaren Brown multzoen bertsioa erabiliko dugu azalpenekin jarraitzeko: N4B = g_b, g_{+1b}, a_b, a_{1+b}, n(g, u, a); g_b, u_{+1k}, a_k, a_{1+k}, n(g, u, a), non, g_b, lotura sintaktikoko gurasoaren hitz-formaren Brown multzoa, g_{+1b}, lotura sintaktikoko gurasoaren hurrengo hitzaren Brown multzoa, a_b, uneko umearen anaiaren hitz-formaren Brown multzoa, a_{1+b}, uneko umearen anaiaren hurrengo hitz-formaren Brown multzoa, s_{+1k}, uneko umearen hurrengo hitzaren kategoria, a_k, uneko umearen anaiaren kategoria, a_{1+k}, uneko umearen anaiaren hurrengo hitzaren kategoria eta, n(g, u, a), gurasoaren, umearen eta anaiaren arteko dependentziaren norabidea diren.



5.10 irudia – Brown multzoetatik meta-ezaugarriak lortzeko prozesuaren irudikapena.

Kontatu behar ditugun gertaera sintaktikoak azalduta, N4B txantiloian agertzen diren elementuak gure adibideko elementuekin lotu ditzakegu: g_b=0101, g_{+1b}=1010, a_b=0001, a_{1+b}=1111, n(g, s, a)=Ezker; g_b=0101, u_{+1k}=ADI, a_k=IZE, a_{1+k}=IZE, n(g, u, a)=Ezker. Puntu honetan, automatikoki etiketatutako corpusean puntu komaz banatutako bi gertaerak aldi berean zenbat alditan gertatu diren kontatu behar dugu. Demagun G_s:[0101, 1010, 0001, 1111, Ezker;0101, ADI, IZE, IZE, Ezker] gertaera sintaktikoaren maiztasuna 300 dela eta ranking-ean TOP10 barruan dagoela. Orduan, 5.1 mapaketa funtzioaren arabera A balioa jasoko luke eta hiru meta-ezaugarri hauek sortuko genituzke analizatzaile sintaktikoak probabilitateen kalkuluan kontuan har ditzan: $\Phi(m) \Rightarrow N4B - A$, $\Phi(m) - g_k \Rightarrow N4B - A - ADI$ eta $\Phi(m) - g_b \Rightarrow N4B - A - 0101$.

5.11 taulan, bilaketa patroi bakoitzetik eratorritako meta-ezaugarriak erabiliz erdietsitako emaitzak bildu dira. Euskararako lortu diren emaitzak aztertzen badira, kasu batzuetan emaitza negatiboak erdiesten dira eta beste batzuetan hobekuntzak. Hala ere, lortzen diren hobekuntzak ez dira estatistikoki esanguratsuak eta hitz-formetan oinarritutako patroiak erabiliz erdietsitako emaitzak baino baxuagoak dira. Hiru emaitzarik hoberenak N3B2,

Izena	Eus	Fra	Ale	Hun	Sue
Oinarria	82,69	76,78	85,25	77,65	73,27
N1B	82,64 (- 0,05)	**76,67 (- 0,11)**	**85,38 (+ 0,13)**	77,90 (+ 0,25)	73,35 (+ 0,08)
N1B1	82,71 (+ 0,02)	76,51 (- 0,27)	85,30 (+ 0,05)	77,77 (+ 0,12)	**73,84 (+ 0,57)**
N1B2	82,45 (- 0,24)	76,51 (- 0,27)	85,30 (+ 0,05)	77,77 (+ 0,12)	**73,84 (+ 0,57)**
N2B	82,61 (- 0,08)	**76,67 (- 0,11)**	**85,38 (+ 0,13)**	77,90 (+ 0,25)	73,35 (+ 0,08)
N2B1	82,54 (- 0,15)	76,51 (-0,27)	85,30 (+ 0,05)	77,77 (+ 0,12)	**73,84 (+ 0,57)**
N2B2	**82,73 (+ 0,04)**	76,51 (- 0,27)	85,30 (+ 0,05)	77,77 (+ 0,12)	**73,84 (+ 0,57)**
N3B	82,64 (- 0,05)	**76,74 (- 0,04)**	**85,46 (+ 0,21)**	77,91 (+ 0,26)	73,39 (+ 0,12)
N3B1	82,54 (- 0,15)	76,51 (- 0,27)	85,30 (+ 0,05)	77,77 (+ 0,12)	**73,84 (+ 0,57)**
N3B2	**82,85 (+ 0,16)**	76,49 (- 0,29)	85,30 (+ 0,05)	77,95 (+ 0,30)	73,25 (- 0,02)
N4B	**82,74 (+ 0,05)**	76,53 (- 0,25)	85,31 (+ 0,06)	**78,02 (+ 0,37)**	73,21 (- 0,06)
N4B1	82,59 (- 0,10)	76,64 (- 0,14)	85,33 (+ 0,08)	**78,06 (+ 0,41)**	73,16 (- 0,11)
N4B2	**82,73 (+ 0,04)**	76,64 (- 0,14)	85,33 (+ 0,08)	**78,06 (+ 0,41)**	73,16 (- 0,11)

5.11 taula – Brown multzoetan oinarritutako bilaketa patroi bakoitzetik eratorritako meta-ezaugarriak erabiliz hizkuntza bakoitzerako lortutako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz hiru emaitza hoberenak.

N2B2 eta N4B bilaketa patroietatik eratorritako meta-ezaugarriekin lortu dira.

Frantseserako lortu dira berriro emaitzarik baxuenak, hobekuntza batik ikusten ez delarik. Hitz-formekin gertatu den bezala, emaitza horiek jasotzeko arrazoia zuhaitz-bankuan bildu diren hitzen kategoriak automatikoki sortzeko erabilitako tresna dela uste dugu.

Alemanean, hitz-formekin gertatu den bezala, frantsesean gertatzen denaren kontrakoa ikusten da, kasu guztietan lortzen baitira hobekuntzak. Hiru emaitzarik hoberenak N1B, N2B eta N3B patroietatik eratorritako meta-ezaugarriekin erdietsi dira, baina hobekuntza horiek ez dira estatistikoki esanguratsuak.

Hungarierarako taulan bildu ditugun emaitza guztietan ikusten dira hobekuntzak. Hiru emaitzarik hoberenak N4B, N4B1 eta N4B2 bilaketa patroietatik eratorritako meta-ezaugarriekin erdietsi dira. Kasu guztietan emaitza positiboak erdietsi direla eta orokorrean emaitza hobeak lortu direla ikusita, esan daiteke Brown multzoetan oinarritutako patroiak erabiliz eratorritako meta-ezaugarriek eragin positiboagoa dutela hitz-formetan oinarritutako patroiak erabiliz eratorritako meta-ezaugarriek baino.

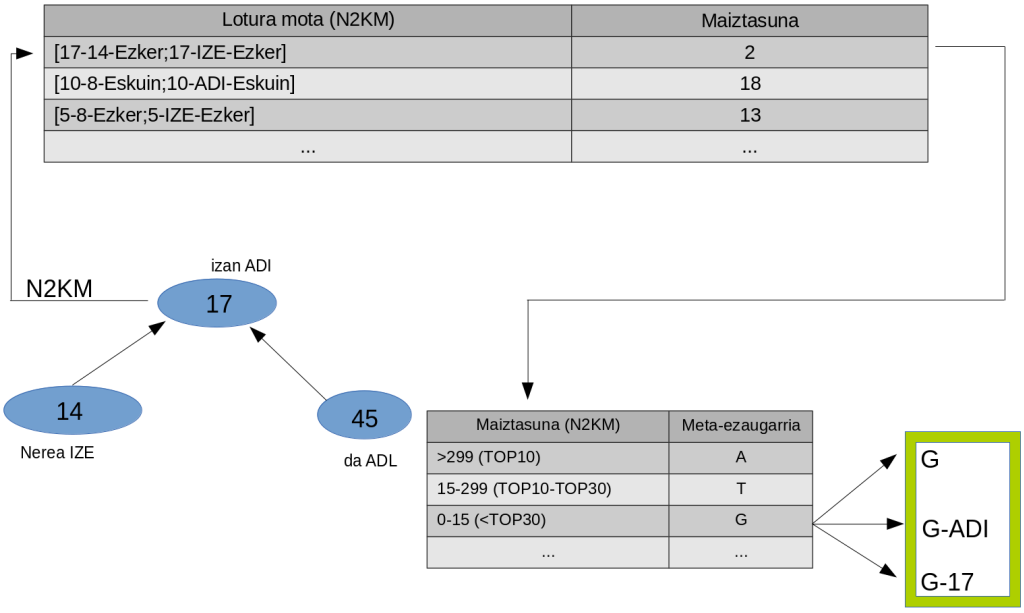
Suedierarako, aurreko puntuan gertatu den bezala, orain arteko hobekuntzarik handiena (+ 0,57) erdietsi da bost bilaketa patroi desberdinekin (N1B1, N1B2, N2B1, N2B2 eta N3B1). Emaitza guztiak aztertuta, hitz-

formetan ikusi den bezala, esan daiteke suediera dela Brown multzoetan oinarritutako meta-ezaugarriei probetxu gehien atera dien hizkuntza. Izan ere, bost emaitzarik hoberenak suedierarako erdietsi baitira.

5.11 taulan bildutakoa ikuspegi eleaniztunetik aztertzen bada, ez da erraza hizkuntza guztietarako Brown multzoetan oinarritutako patroietatik eratorritako meta-ezaugarriak hitz-formetan oinarritutako patroietatik eratorritako meta-ezaugarriak baino hobeak direla baieztatzea dependentzien analisirako. Hungarierarako eta, batez ere, suedierarako Brown multzoetan oinarritutako esperimentuen emaitzak orokorrean hitz-formetan oinarritutako esperimentuen emaitzak baino hobeak dira. Hala ere, gainontzeko hiru hizkuntzetan, euskaran, frantsesean eta alemanean, ez da ikusten alde handirik bi iturri motekin lortutako emaitzen artean.

5.6.3.3 Hitz-bektoreen K-means multzoetatik eratorritako meta-ezaugarriak

Hitz-bektoreen gaineko K-means multzoetatik eratorritako meta-ezaugarriak sortzeko prozesua, Brown multzoetatik eratorritako meta-ezaugarriak sortzeko prozesuaren oso antzekoa da, desberdintasun bakarra erabili den multzokatze mota izanik. Orduan, atal honetan egin ditugun K-means multzoen kontaketak zehazteko 5.9 taulan hitz-formei dagozkien kontaketak K-means multzoei dagozkienekin ordezkatzea besterik ez dugu. Adibidez, N2H txantiloiaren K-means multzoetan oinarritutako bertsioa hau litzateke: N2KM = g_{km}, u_{km}, n(g, u); g_{km}, u_k, n(g, u) non, g_{km}, lotura sintaktikoko gurasoaren hitz-formaren K-means multzoa eta, u_{km}, umearen hitz-formaren K-means multzoa diren.



5.11 irudia – K-means multzoetatik meta-ezaugarriak lortzeko prozesuaren irudikapena.

Maiztasunak aurreko puntuetan bezala handitik txikira ordenatuko ditugu, eta 5.1 mapaketa funtzioaren arabera gertaera bakoitzaren maiztasun bakoitzari etiketa bat esleituko zaio. K-means multzoetatik eratorritako meta-ezaugarriak sortzeko jarraitu dugun prozedura azaltzeko *Nerea izan da* esaldia eta aurretik aipatu dugun N2KM txantiloia erabiliko dugu oraingoan (ikus 5.11 irudia). Esaldi horretan *izan* hitza gurasoa da eta *Nerea* eta *da* umeak. Suposatuko dugu uneko umea kasu honetan *Nerea* hitza dela. Demagun ere 14, 17 eta 45 direla *Nerea*, *izan* eta *da* hitz-formen K-means multzoak, hurrenez hurren. Hori esanda, gure esaldiko elementuak txantiloiko elementuekin lotu ditzakegu modu honetara: g_{km}=17, u_{km}=14, n(g, u)=Ezker; g_{km}=17, u_k=IZE, n(g, u)=Ezker. Puntu honetan, automatikoki etiketatutako corpusean puntu komaz banatutako bi gertaerak aldi berean zenbat alditan gertatu diren kontatu behar dugu. Demagun G_s:[17, 14, Ezker; 17, IZE, Ezker] gertaera sintaktikoaren maiztasuna 2 dela. Orduan, 5.1 mapaketa funtzioaren arabera G balioa jasoko luke eta hiru meta-ezaugarri hauek sortuko genituzke analizatzaile sintaktikoak probabilitateen kalkuluan kontuan har ditzan: $\Phi(m) \Rightarrow N2KM-G$, $\Phi(m)-g_k \Rightarrow N2KM-G-ADI$

eta $\Phi(m) - g_{km} \Rightarrow N2KM - G - 17$.

Izena	Eus	Fra	Ale	Hun	Sue
Oinarria	82,69	76,78	85,25	77,65	73,27
N1KM	82,64 (- 0,05)	76,51 (- 0,27)	**85,33 (+ 0,08)**	**77,92 (+ 0,27)**	**73,20 (- 0,07)**
N1KM1	82,75 (+ 0,06)	**76,55 (- 0,23)**	**85,32 (+ 0,07)**	77,53 (- 0,12)	72,95 (- 0,32)
N1KM2	**82,79 (+ 0,10)**	**76,55 (- 0,23)**	**85,32 (+ 0,07)**	77,53 (- 0,12)	72,95 (- 0,32)
N2KM	82,65 (- 0,04)	76,51 (- 0,27)	**85,33 (+ 0,08)**	**77,92 (+ 0,27)**	**73,20 (- 0,07)**
N2KM1	82,64 (- 0,05)	**76,55 (- 0,23)**	**85,32 (+ 0,07)**	77,53 (- 0,12)	72,95 (- 0,32)
N2KM2	82,64 (- 0,05)	**76,55 (- 0,23)**	**85,32 (+ 0,07)**	77,53 (- 0,12)	72,95 (- 0,32)
N3KM	**82,76 (+ 0,07)**	**76,62 (- 0,16)**	**85,33 (+ 0,08)**	77,71 (+ 0,06)	72,92 (- 0,35)
N3KM1	82,67 (- 0,02)	**76,55 (- 0,23)**	**85,32 (+ 0,07)**	77,53 (- 0,12)	73,14 (- 0,13)
N3KM2	**82,76 (+ 0,07)**	**76,56 (- 0,22)**	**85,37 (+ 0,12)**	77,67 (+ 0,02)	**73,27 (+ 0,00)**
N4KM	**82,83 (+ 0,14)**	76,52 (- 0,26)	85,27 (+ 0,02)	77,70 (+ 0,05)	73,01 (- 0,26)
N4KM1	82,74 (+ 0,05)	76,32 (- 0,46)	85,30 (+ 0,05)	**77,72 (+ 0,07)**	73,16 (- 0,11)
N4KM2	82,56 (- 0,13)	76,32 (- 0,46)	85,30 (+ 0,05)	**77,72 (+ 0,07)**	73,16 (- 0,11)

5.12 taula – Hitz-bektoreen K-means multzoetatik eratorritako meta-ezaugarriak erabiliz kategoria bakoitzean hizkuntza bakoitzerako lortutako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz hiru emaitza hoberenak.

5.12 taulan, bilaketa patroi bakoitzetik eratorritako meta-ezaugarriak erabiliz erdietsitako emaitzak bildu dira. Euskararako lortu diren emaitzak aztertzen badira, lortzen diren hobekuntzak ez dira estatistikoki esanguratsuak. Emaitza horiek hitz-formetan oinarritutako patroiak erabiliz erdietsitako emaitzak baino baxuagoak dira orokorrean, eta Brown multzoetan oinarritutako patroiak erabiliz lortu diren emaitzen parekoak. Hiru emaitzarik hoberenak N1KM2, N3KM eta N4KM bilaketa patroietatik eratorritako meta-ezaugarriekin lortu dira.

Frantseserako errepikatzen dira berriro emaitzarik baxuenak, hobekuntza bat bera ere erdietsi ez delarik. Hori gutxi balitz, hitz-bektoreetan oinarritutako K-means multzoak meta-ezaugarriak sortzeko iturri bezala erabiltzean hitz-formekin eta Brown multzoekin erdietsi diren emaitzak baino emaitza okerragoak lortu dira frantseserako.

Alemanean, hitz-formekin eta Brown multzoekin gertatu den bezala, kasu guztietan lortzen dira hobekuntzak, baina hobekuntza horiek ez dira estatistikoki esanguratsuak. Orokorrean esan daiteke erabili ditugun hiru iturrietatik alemanerako emaitzarik okerrenak lortu direla K-means multzoekin.

Hungarierarako taulan bildu ditugun emaitzetan gora-beherak ikusten dira. Kasu batzuetan analizatzaileari gehitu zaion informazioa baliagarria izan da, baina beste kasu batzuetan kaltegarria izan da. Hiru emaitzarik hobe-

renak N1KM, N2KM eta N4KM1 bilaketa patroietatik eratorritako meta-ezaugarriekin erdietsi dira. Hungarieraren kasuan esan daiteke erabili diren hiru iturrien artean hitz-bektoreetan oinarritutako K-means multzoak hoberena ez dela, Brown multzoekin emaitza hobeak lortu baitira. Hala ere, hitz-formekin lortu diren emaitzen parekoak erdietsi dira iturri honekin.

Suedierarako orain arteko emaitzarik okerrenak erdietsi dira iturri honetako patroietatik eratorritako meta-ezaugarriekin. Are gehiago, ez da hobekuntza behin ere lortu. Horrenbestez, emaitza guztiak aztertuta, esan daiteke suedierarako iturririk hoberena Brown multzoak direla. K-means multzoak, ostera, aurkako aldean kokatzen dira, eta hitz-formak erdian.

Aurreko bi puntuetan eta hemen ikusitako emaitzak aztertzen badira, orokorrean hizkuntza guztietarako lagungarriena den iturria Brown multzoak direla esan daiteke. Dena den, ez da ikusten hitz-formekin alde handirik dagoenik. Bestalde, gutxien laguntzen duen iturria K-means multzoak dira.

Meta-ezaugarriekin erlazionatutako esperimentuetan hainbat hobekuntza erdietsi diren arren, kasu gutxitan izan dira estatistikoki esanguratsuak. Hori dela eta, hurrengo puntuan, hizkuntza bakoitzerako emaitzarik hoberenak lortu dituzten hiru patroiak bilduko dira azken esperimentu batean. Modu honetara, patroi desberdinetatik eratorritako meta-ezaugarriak beraien artean osagarriak diren neurtuko da, hots, patroien bildurarekin patroi bakoitzarekin lortutako emaitza gainditzen den neurtuko da.

5.6.3.4 Hiru patroi hoberenen meta-ezaugarrien batura

Aurreko hiru puntuetan, erabilitako iturriaren arabera banakako patroietatik eratorritako meta-ezaugarrien eraginkortasuna neurtu da. Puntu honetan, aldiz, hizkuntza bakoitzean emaitzarik hoberenak erdietsi dituzten hiru patroietatik eratorritako meta-ezaugarri guztiek duten eragina neurtuko da. Hiru hoberenak erabiltzea erabaki da, lortu nahi den aniztasunaren eta gehituko diren meta-ezaugarri kopuruaren arteko erlazioa egokia iruditzen zaigulako. Izan ere, gehituko diren meta-ezaugarriak osagarriak diren ala ez neurtuko baita, horretarako gutxienez hiru erabiltzea komenigarria delarik, eta analizatzaileari ezaugarri berri gehiegi gehitzea ez baita aukerarik hoberena oinarrizko ezaugarrien eta gehitutako ezaugarrien arteko loturak ondorioztatzeko.

Hizkuntza bakoitzerako emaitzarik hoberenak lortu dituzten hiru patroiak ondorengoak dira, letra lodiz hizkuntza bakoitzean emaitzarik hoberena lortu duen patroia:

- Euskara: N1H, **N2H2** eta N4K
- Frantsesa: N2H1, N3B eta **N3H**
- Alemana: N1K, N2H eta **N3B**
- Hungariera: **N2H1**, N4B1 eta N4B2
- Suediera: **N1B1**, **N2B1** eta **N3B1**

5.13 taulan hizkuntza bakoitzean hiru patroi hoberenen baturatik eratorritako meta-ezaugarriek erdietsitako emaitzak bildu dira, oinarrizko emaitza eta bakarkako patroiekin orain arte lortu diren emaitzarik hoberenekin batera.

Hizkuntza	**Oinarria**	**Hoberena**	**Hiru hoberenen batura**
Eus	82,69	82,96	**82,98 (+ 0,29)(+ 0,02)**
Fra	76,78	76,79	76,40 (- 0,28)(- 0,29)
Ale	85,25	85,46	85,37 (+ 0,12)(- 0,09)
Hun	77,65	78,12	77,93 (+ 0,28)(- 0,19)
Sue	73,25	73,84	73,08 (- 0,17)(- 0,76)

5.13 taula – Hiru patroi hoberenen baturarekin lortutako emaitzen, oinarrizko emaitzen eta bakarkako emaitzarik hoberenen arteko konparaketa *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz bakarkako emaitzarik hoberena gainditu duten emaitzak.

5.13 taula aztertuta orain arteko emaitzarik hoberena euskarak soilik gainditzen duela ikusten da. Gainontzeko hizkuntzetan patroien baturarekin emaitzek okerrera egin dute. Egia da alemanaren eta hungarieraren kasuan oinarrizko emaitzak gainditzen direla, baina, aurreko puntuetan patroi bat baino gehiago erabiltzen duten esperimentu askotan ikusi den bezala, batu diren patroietatik eratorritako meta-ezaugarriak beraien artean ez direla osagarriak ikusi da.

Patroi desberdinetatik eratorritako meta-ezaugarriek emaitza orokorrean hobekuntza esanguratsurik erdiesten ez dituzten arren, meta-ezaugarri horiek dependentzia sintaktiko desberdinetan duten eragina aztertuko da. Horretarako, hizkuntza guztietan parekoak diren dependentzia etiketetan erdietsitako emaitzak bildu dira 5.14 taulan. Era horretara, hizkuntza bakoitzean

meta-ezaugarriek zein dependentziari eragiten dioten zehaztuko da eta hizkuntza desberdinetan antzeko portaera duten ikusiko da. Hizkuntza desberdinetako dependentzia etiketak ez dira inoiz izango guztiz konparagarriak, hizkuntza bakoitzak bere arau sintaktikoak dituelako. Hala ere, esan daiteke 5.14 taulan bildu ditugun dependentziak beraien artean baliokideak direla nahiz eta baliokidetza hori erabatekoa ez izan. Aukeratu ditugun dependentziak esaldiaren egitura definitzeko garaian pisu gehien duten dependentziak izan dira.

		ncmod	**ncobj**	**root**	**ncsubj**
Eus	**Oinarria**	81,65	74,93	88,42	70,20
	Batura	81,69	74,66	**89,46 (+ 1,04)**	69,98
Fra	**Oinarria**	70,35	88,26	88,10	83,56
	Batura	69,64	**88,32 (+ 0,06)**	87,29	82,34
Ale	**Oinarria**	96,63	76,31	88,02	85,35
	Batura	96,59	76,24	87,98	**85,65 (+ 0,30)**
Hun	**Oinarria**	84,08	87,20	84,14	77,83
	Batura	84,27	87,01	84,19	**78,66 (+ 0,83)**
Sue	**Oinarria**	61,83	74,13	86,95	85,84
	Batura	60,90	**74,97 (+ 0,84)**	86,34	85,66

5.14 taula – Hiru bilaketa patroi hoberenen baturatik eratorritako meta-ezaugarriek hizkuntza desberdinetako dependentzia etiketetan jasotako emaitzak F-measure neurriaren arabera. Letra lodiz hizkuntza bakoitzerako erdietsitako hobekuntzarik handiena.

5.14 taulan bildutako emaitzen arabera, meta-ezaugarriek ez dute izan ia eraginik aukeratu ditugun dependentzietan frantsesean eta alemanean. Frantsesa izan da aukeratutako dependentzietan eragin positibo gutxien jasan duen hizkuntza. Gertaera hori ez da harritzekoa emaitza orokorrean kasu gehienetan eragin negatiboa izan baitute meta-ezaugarriek. Alemanean, subjektuan lortu da emaitzarik hoberena (taulan *ncsubj*), oinarrizko emaitzan dependentzia horretan lortzen den emaitzan baino 0,30 puntu gehiago lortuz.

Euskarazko esaldien erroa definitzen duen dependentzia etiketan (taulan *root*) lortu da hobekuntzarik handiena, oinarrizko emaitzan erdietsi dena baino puntu bat gehiago (+ 1,04) lortuz. Hobekuntza hori oso positiboa da eta oso kontutan hartzekoa, esaldiaren egitura zehazteko esaldiaren erroa identifikatzea oso garrantzitsua baita. Euskararen kasuan gainontzeko dependentzietan pareko emaitzak edo emaitza negatiboak jaso ditugu. Hala

ere, esperotako emaitzak dira, erroaren kasuan jasotako emaitza gehiagorekin 5.13 taulan erakutsitako emaitza orokorrak altuagoak izango baitziren.

Hungarieraren kasuan meta-ezaugarrien eragin positiboa gehien jasan duen dependentzia subjektua (taulan ncsubj) izan da 0,83 puntuko hobekuntza erdietsiz. Zoritxarrez, dependentzia horretan jasandako hobekuntza ez da 5.13 taulako emaitzetan islatzen 5.13 taulan bildu ez diren hainbat dependentzietan izandako emaitza negatiboen ondorioz. Horrenbestez, euskararen kasuan esaldiaren erroaren dependentziarekin gertatzen den bezala, hungarierako subjektua hobetzeko erabili daitezke meta-ezaugarriak.

Suedieran, aldiz, objektu zuzenaren dependentzia (taulan *ncobj*) izan da nabarmen hobetu den dependentzia bakarra (+ 0,84), baina ez da hobekuntza hori 5.13 taulako emaitzetan islatzen. 5.13 taulan ikusi denaren arabera, hiru patroi hoberenetatik eratorritako meta-ezaugarriekin oinarrizko emaitzak okertu egiten dira. Horrenbestez, esan daiteke okerrera egin duten dependentziek eragin handiagoa izan dutela emaitzan objektu zuzenaren dependentziak baino. Okerrera egin duten dependentzien artean esaldiaren erroa definitzen duen dependentzia etiketa (taulan *root*) nabarmentzen da, oinarrizko konfigurazioarekin erdietsitako emaitza baino 0,61 puntu gutxiago erdietsi direlarik bertan. Aurretik esan den bezala, dependentzia hori oso garrantzitsua da eta horren emaitza negatiboa izan daiteke hiru patroi hoberenetatik eratorritako meta-ezaugarriekin emaitza orokorra ez gainditzearen arrazoi nagusia.

5.7 Erabilitako tekniken konbinaketa

Kapitulu honetako aurreko hiru atal nagusietan analizatzaileei teknika desberdinetan oinarritutako informazioa pasatu zaie, informazio horrek dependentzien analisi sintaktikoan duen eragina neurtu ahal izateko. Erabilitako informazioaren arabera eta aztertutako hizkuntzaren arabera, analizatzaileek oso emaitza desberdinak bueltatu dituzte. Nahiz eta orokorrean lortu nahi denaren arabera (eraginkortasuna, azkartasuna, etab.) informazio motaren bat erabiltzea besteak erabiltzea baino komenigarriagoa gertatu daitekeen, ezin dugu esan informazio mota bat hobea dela hizkuntza guztietarako aspektu guztietan. Hori dela eta, atal honetan orain arte erabilitako hiru tekniketatik eratorritako informazioa konbinatuko dugu hiru informazio moten konbinazioak analisi sintaktikoan duten eragina neurtzeko.

5.7.1 Gure hurbilpena

Morfologikoki aberatsak diren hizkuntzetarako erabilitako hiru tekniken konbinazioa aplikatzeak eskaintzen dituen aukerak aztertzeko ondorengo atazak landu dira hautatutako hizkuntza bakoitzerako:

- Ezaugarrien ingeniaritzaren bitartez sortutako informazioaren, multzokatzearen bitartez sortutako informazioaren eta meta-ezaugarrien informazioaren konbinaketak analisi sintaktikoan duen eragina neurtu.
- Erabilitako hiru informazio motak osagarriak diren ala ez aztertu.

5.7.2 Esperimentuak eta emaitzak

Erabilitako hiru informazio motak konbinatzeko, ezaugarrien ingeniaritzaren bitartez sortutako analisiak, multzokatzearen bitartez sortutako analisiak eta meta-ezaugarriekin sortutako analisiak MaltBlender programari pasatu zaizkio. Meta-ezaugarrien esperimentuak azaldu ditugunean hainbat esperimentu ikusi ditugu, bilaketa patroiaren eta iturri motaren araberako esperimentuak. Esperimentu horietako guztietako analisiak konbinaketan erabiltzea gehiegizkoa denez, meta-ezaugarriak konbinaketan bi eratara erabiliko ditugu: hiru bilaketa patroi hoberenen analisiak konbinaketan bakarka gehituz eta hiru bilaketa-patroi hoberenen baturatik eratorritako meta-ezaugarriekin erdietsitako analisiak gehituz (5.6.3.4 puntuan ikusitakoa).

Informazio bakoitzak analisian duen eragina neurtzeko, informazio motak banaka joango gara gehitzen. Erdietsi ditugun emaitzak 5.15 taulan bildu ditugu.

	Eus	**Fra**	**Ale**	**Hun**	**Sue**
EI	86,2	85,1	91,8	84,1	78,1
EI+MU	86,2	85,1	91,8	84,1	78,1
EI+MU+MEBAK	86,2	85,1	91,8	84,1	78,1
EI+MU+MEBAT	**86,6 (+ 0,4)**	85,1	**91,9 (+ 0,1)**	**84,4 (+ 0,3)**	**78,3 (+ 0,2)**

5.15 taula – Konbinaketan informazio mota bakoitzak duen eragina neurtzeko erdietsitako emaitzak *Labeled Attachment Score* (LAS) neurriaren arabera. Letra lodiz bakarkako emaitzarik hoberena gainditu duten emaitzak. EI=Ezaugarrien ingeniaritza, MU=Multzokatzea, MEBAK=hiru bilaketa-patroi hoberenen meta-ezaugarriak bakarka, MEBAT=hiru bilaketa-patroi hoberenen meta-ezaugarrien batura.

5.15 taulako emaitzen arabera, oso zaila gertatu da ezaugarrien ingeniaritzako esperimentuetan egindako konbinaketen emaitzak gainditzea. Are gehiago, frantsesean ez da lortu emaitza hori gainditzea.

Gainontzeko lau hizkuntzetan konbinaketan multzokatzearekin sortutako analisiak gehitu dituguneanean, ez da lortu ezaugarrien ingeniaritzako analisien konbinaketarekin erdietsitako emaitza gainditzea. Gauza bera gertatu da hiru bilaketa-patroi hoberenekin eratorritako meta-ezaugarrien analisiak bakarka gehitu dituguneanean ere.

Hiru bilaketa-patroi hoberenetatik eratorritako meta-ezaugarrien batura gehitu duguneanean konbinaketan lortu dira hobekuntzak. Ikusitako hobekuntzak McNemar testaren arabera estatistikoki esanguratsuak ez diren arren, kontuan hartzekoa da konbinaketara analisi bakarra gehituta lortu direla hobekuntzak. Gainera, ez da ahaztu behar gainditu beharreko emaitzak altuak direla. Adibidez, euskararen kasuan gainditu beharreko emaitza oinarrizko emaitzarik hoberena baino 3,2 puntu altuagoa da, eta alemanaren kasuan % 91,8ko emaitza da gainditu beharrekoa. Zenbat eta altuagoa izan gainditu beharreko emaitza, orduan eta hobekuntza tarte txikiagoa dago, eta oso zaila da % 90etik gora dauden emaitzak gainditzea.

5.8 Ondorioak

Kapitulu honetan ikuspegi eleaniztunari heldu nahi izan diogu egindako aurrerakuntzak ahalik eta hizkuntza gehienetan modu errazean aplikatu daitezkeen teknikak erabiliz. Horretarako, eskuragarri izan ditugun baliabideak erabili ditugu, sortutako ezagutza arazorik gabe aplikatzeko hizkuntza desberdinetan. Ikuspegi eleaniztuna jorratu dugun arren, euskararekin konparagarriak diren hizkuntza morfologikoki aberatsak landu dira, lortutako emaitzetatik ezaugarri morfologiko konplexuagoak dituzten hizkuntzekin lotutako ahalik eta ondorio orokorrenak ateratzeko. Hala ere, kapitulu honetan landutako hurbilpenak ia edozein hizkuntzatan landu daitezke aldaketa minimo batzuk soilik aplikatuta.

Landu ditugun hurbilpenak lau izan dira: ezaugarrien ingeniaritza, multzokatzea, meta-ezaugarriak eta hiru horien konbinaketa. Horietako bakoitzean ikusitako emaitzak aztertuta hurrengo ondorioak atera ditugu hurbilpen bakoitzerako:

- **Ezaugarrien ingeniaritza:** Aztertutako hizkuntza bakoitzerako, analisi sintaktikorako ezaugarri morfologikoen eragina zehaztu dugu. Er-

dietsitako emaitzak aztertzen badira, orokorrean ezaugarririk esanguratsuena kasua dela esan genezake, berebiziko garrantzia duelarik euskaran eta hungarieran. Ezaugarri morfologiko bakoitzaren pisua jakinda landu ditugun esperimentuekin zenbait hobekuntza lortu ditugu, batez ere euskaran, frantsesean eta suedieran. Hungarieran eta alemanean ez dugu lortu inolako hobekuntzarik ezaugarri esanguratsuenak bakarrik erabilita, ezta ezaugarri garrantzitsuena analizatzaileak haztapen altuagoa ematen dion lekuan jarrita.

Azken emaitza horiek ikusita, bi hizkuntza horietan zergatik ez den inolako hobekuntzarik nabaritzen galdetu diogu geure buruari. Alemanaren kasuan ezaugarri morfologikoek analisi sintaktikoan bakarka eraginik ez dutela ikusi da eta hori izan daiteke emaitza horien arrazoia, aplikatutako esperimentuak ezaugarri morfologikoetan oinarrituta baitaude. Hungarieraren kasuan, bakarkako ezaugarri gehienek analisi sintaktikoan eragin nabarmena dutela ikusi da, baina ezaugarri horietan oinarritutako esperimentuetan ez da lortu oinarrizko konfigurazioarekin lortutako emaitza gainditzea. Emaitza horiek jaso izanaren arrazoi nagusia hungarierako ezaugarri morfologiko gehienak beraien artean osagarriak direla izan daiteke. Hori dela eta, ezaugarri guztiak erabili beharrean hiru hoberenak erabiltzean informazioa galdu da. Bestalde, ezaugarririk hoberena kategoriaren eta azpikategoriaren ordez jarri dugunean analizatzaileak ez dio informazio horri haztapen altuagoa eman, ezaugarri morfologikoen zutabean jaso dituen ezaugarriek informazioa oso zabala eta osagarria eskaintzen dutelako.

Dena den, hasierako esperimentuetan aplikatutako tekniken eragina asko nabaritzen da konbinaketa esperimentuak lantzen ditugunean, hizkuntza guztietarako hobekuntza sendoak erdiesten direlarik. Ondorioz, uste dugu aplikatutako teknikekin lortutako analisi bakoitzak baduela berezko ezaugarriren bat beste analisiekin osagarria dena eta analisi sintaktiko sendoagoa erdiesteko lagungarria dena.

- **Multzokatzea:** Hurbilpen honetan, multzokatzeak eskaintzen dituen aukera desberdinak aztertu ditugu analisi sintaktikoa hobetzeko. Alde batetik, euskara, alemana eta hungarierarako multzokatze algoritmo desberdinak aplikatu baino lehen, hitz-forma bakoitza zatitu egin dugu lema eta atzizkian, multzokatze morfologikoa deitu diogunaren eragina neurtzeko hiru hizkuntza horietan. Beste alde batetik, multzokatze

arruntaren eragina neurtu dugu aukeratutako bost hizkuntzetan. Multzokatze morfologikoaren eta multzokatze arruntaren emaitzak ikusita esan daiteke Brown multzoak erabiltzen direnean eraginkorragoa dela multzokatze arrunta aplikatzea, K-means multzoak erabiltzen direnean, berriz, ez dago ia alderik bi ikuspegien artean.

Erabilitako multzokatze algoritmoetatik eratorritako ezaugarrien eraginkortasuna aztertzen bada, alde handirik ez badago ere, Brown multzokatze algoritmoan oinarritutako ezaugarrien eragina positiboagoa da analisi sintaktikoan, nahiz eta Brown multzoak sortzeko corpus txikiagoa erabili den. Bukatzeko, multzokatze mota bakoitzetik eratorritako ezaugarri berriekin lortutako analisiak konbinatu ditugunean, hobekuntza nabarmenak ikusi dira aztertutako hizkuntza guztietan.

Gertaera horretatik, erabili ditugun multzokatze mota biak osagarriak izan daitezkeela ondoriozta daiteke. Hala ere, kasu honetan, esperimentu bat burutu dugu gure ondorioak ziurtatzeko. Multzokatzetik eratorritako ezaugarririk gabe lortutako analisiak konbinatu dira, multzokatze mota desberdinetatik eratorritako ezaugarriekin lortutako analisien konbinaketarekin erdietsitako emaitzekin alderatzeko. Konparaketa horrekin lortu diren hobekuntzen zati handiena multzokatze teknika desberdinen konbinaziotik datorrela ikusten da bertan.

Horrenbestez, esan daiteke multzokatze mota desberdinek informazio baliagarria eskaintzen diotela analizatzaileari, bakarka erdietsitako emaitzetan eta oinarrizko konfigurazioarekin erdietsitako emaitzetan hobekuntza estatistikoki esanguratsuak lortuz ia hizkuntza guztietarako (4 hizkuntzatarako).

- **Meta-ezaugarriak:** Kasu honetan, iturri desberdinetatik eratorritako meta-ezaugarriek hizkuntza desberdinen analisi sintaktikoan duten eragina neurtu dugu. Iturri horiek hitz-formak, Brown multzoak eta hitz-bektoreetan oinarritutako K-means multzoak dira. Iturri horietako bakoitzarekin hainbat bilaketa-patroi definitu dira, iturri desberdinen bilaketa-patroien arteko desberdintasun bakarra corpusean maiztasunak kontatzeko garaian kontatu beharreko elementu mota izan delarik, hau da, hitz-formak, Brown multzoak eta hitz-bektoreetatik eratorritako K-means multzoak.

  Bilaketa-patroi bakoitzeko esperimentu bat gauzatu dugu hizkuntza bakoitzeko, bilaketa-patroi horri buruzko kontaketak egin ondoren sor-

tutako meta-ezaugarriak erabiliz. Era horretan jaso ditugun emaitzak aztertuta, hizkuntza eta patroi desberdinekin hainbat hobekuntza erdietsi dira. Hobekuntza horiek kasu gehienetan estatistikoki esanguratsuak ez diren arren meta-ezaugarriak baliagarriak direla esan daiteke eta beraiek aplikatzeko modu desberdinak erabil daitezke, horrek zabaltzen dituen aukera desberdinekin. Guk iturri eta bilaketa-patroi zehatz batzuk erabili ditugu, baina iturri desberdinak eta patroi desberdinak erabil daitezke.

Hizkuntza bakoitzari era desberdinean eragin die iturri bakoitzetik eratorritako meta-ezaugarriek. Euskaran eta alemanean eragin positiboagoa izan dute hitz-formetatik eratorritako meta-ezaugarriek. Hungarieran eta suedieran, berriz, Brown multzoetatik eratorritako meta-ezaugarriek eragin handiagoa izan dute, bereziki suedieran. Frantsesa suedieraren kontrako aldean kokatu da, meta-ezaugarriekin ez baitugu lortu oinarrizko emaitza hobetzea. Hitz-bektoreetatik eratorritako K-means multzoetan oinarritutako meta-ezaugarriekin kasu batzuetan oinarrizko emaitzak gainditu diren arren, egin diren esperimentuetan beste iturri biei probetxu gehiago atera diete analizatzaileek.

Hizkuntza bakoitzean emaitzarik hoberenak lortu dituzten hiru patroiak batu direnean erdietsi diren emaitzen arabera, ez da lortu bakarkako emaitzarik hoberenak gainditzea kasu gehienetan. Euskararen kasuan soilik gertatu da bakarkako emaitzarik hoberena gainditzea, baina aldea oso txikia izan da. Hori dela eta, batu ditugun bilaketa-patroietatik eratorritako meta-ezaugarriak beraien artean osagarriak ez direla esan daiteke. Suedieraren eta frantsesaren kasuak bereziki negatiboak izan dira, ez baita lortu oinarrizko emaitzak eta emaitzarik hoberenak gainditzea.

Dena den, bilaketa-patroi desberdinetatik eratorritako meta-ezaugarrien baturak dependentzia sintaktiko zehatzetan eragina baduela ikusi da. Euskararen kasuan, esaldiaren erroa definitzen duen dependentzia nabarmen hobetu da; hungarieraren eta alemanaren kasuan, subjektua definitzen duen dependentzia izan da hobetu dena; eta suedieraren kasuan, objektu zuzena definitzen duen dependentzia hobetu da. Frantsesean, ostera, objektu zuzena definitzen duen dependentzia hobetu den arren, hobekuntza oso txikia izan da. Emaitza hori ez da harritzekoa izan, egindako esperimentuetan ez baitugu lortu meta-ezaugarriekin frantseserako oinarrizko emaitzak gainditzea.

- **Guztien konbinaketa:** Atal honetan aurretik landutako hiru hurbilpenak konbinatu dira. Alde batetik, konbinaketarekin erdietsitako emaitzak neurtu nahi izan dira. Beste aldetik, erabilitako hiru hurbilpenak beraien artean osagarriak diren ala ez neurtu nahi izan da.

  Ikusitako emaitzen arabera, hiru hurbilpenak konbinatu direnean frantsesean izan ezik banakako konbinaketetan jasotako emaitzak gainditu dira. Hobekuntzak estatistikoki esanguratsuak ez diren arren, kontuan hartzekoa da gainditu beharreko emaitzak jada emaitza onak direla.

  Erabilitako hiru hurbilpenak kasu gehienetan beraien artean osagarriak ez direla ikusi da. Ezaugarrien ingeniaritza bidezko konbinaketarekin erdietsitako emaitzak ezin izan dira gainditu multzokatzea ere konbinaketan erabili dugunean. Horrenbestez, esan daiteke ezaugarrien ingeniaritzako analisiak eta multzokatzeko analisiak ez direla osagarriak. Lortu nahi dena ahalik eta sistemarik eraginkorrena izatea bada, hobe da ezaugarrien ingeniaritzako analisiak erabiltzea. Bestalde, lortu nahi dena sistema azkarragoa bada, multzokatzearen analisiak erabiltzea komenigarria da. Izan ere, konbinaketan erabiltzen den analisi kopurua askoz txikiagoa baita eta emaitzak antzekoak.

  Meta-ezaugarriekin lortutako analisiak bi eratara gehitu dira konbinaketan: emaitzarik hoberenak lortu dituzten hiru bilaketa-patroietatik eratorritako meta-ezaugarriek lortutako analisiak bakarka, eta hiru bilaketa-patroi horietatik eratorritako meta-ezaugarrien baturarekin lortutako analisia. Lehenengo aukerarekin ez da lortu ezaugarrien ingeniaritzaren konbinaketarekin erdietsitako emaitzak gainditzea. Bigarren aukerarekin, berriz, hobekuntza txikiak ikusi dira ia hizkuntza guztietan. Kontuan hartzen bada hobekuntzak konbinaketara analisi bakarra gehituta lortzen direla, hobekuntza horiek oso positiboak direla deritzogu.

6 Dependentzia Unibertsalak

6.1 Sarrera

Universal Dependencies edo Dependentzia Unibertsalak, hizkuntza desberdinetarako sortuta dauden zuhaitz-bankuak etiketatze estandar batera bihurtzea helburu duen proiektua[1] da, horretarako zenbait gidalerro eskaintzen dituelarik. Modu honetara, etiketatze sistema bera jarraituta etiketatuak izan diren hainbat zuhaitz-banku egongo dira eskuragarri bakoitza bere hizkuntzarako. Modu berean etiketatuta dauden zuhaitz-banku desberdinak edukitzeak hainbat esperimentazio-bide zabaltzen ditu. Hizkuntza askotan erabil daitezkeen analizatzaile sintaktiko estatistikoen garapenerako oso lagungarria izan daiteke, parserrek etiketatze sistema bakarrarekin lan egin behar dutelako esaldien egitura sintaktikoak ikasten dituzten bitartean. Topologikoki antzekoak diren hizkuntzetan ere esperimentu interesgarriak eraman daitezke aurrera. Adibidez, hizkuntza jakin baten egitura sintaktikoekin entrenatu daiteke parserra eta ikasitakoa beste hizkuntza bati aplikatu, bi hizkuntzen arteko antzekotasun sintaktikoa zehazteko.

Etiketatze eskemari dagokionez, dependentzia sintaktikoak Stanfordeko dependentzien eboluzioan (De Marneffe *et al.* 2006; De Marneffe eta Manning 2008; De Marneffe *et al.* 2014) daude oinarrituta, kategoriak, aldiz, Googleeko kategorietan (Petrov *et al.* 2012), eta ezaugarri morfosintaktikoak Interset Interlinguan (Zeman 2008). Stanfordeko dependentziak 2005ean garatu ziren antzeko testuak identifikatzen zituen sistema bati laguntzeko, baina hainbat

[1] `http://universaldependencies.org/introduction.html`

hizkuntzatarako analisi sintaktikoa burutzeko dependentzia etiketa sistema bezala erabiltzen dira gaur egun. Googleeko kategorietarako etiketa unibertsalen multzoa, berriz, hasiera batean ikasketa ez-gainbegiratua erabiliz kategoriaren esleipen automatikorako sortu zen, eta orduz geroztik, ohiko estandar bezala erabili da.

Ezaugarri morfosintaktikoetarako aukeratutako etiketatze estandarra, ostera, hizkuntza desberdinetako etiketa morfosintaktikoen arteko bihurketak egiteko sortu zen eta lehen aldiz hizkuntzen arteko analizatzaile sintaktiko estatistiko baten egokitzapenean erabili zen (Zeman eta Resnik 2008). Urte batzuk beranduago, HamleDT proiektuan ere erabili zen geruza morfologiko gisa (Zeman *et al.* 2014), proiektuaren helburua hizkuntza desberdinetako zuhaitz-bankuak etiketatze eskema berarekin etiketatuta izatea zelarik.

Stanfordeko dependentzia etiketak eta Googleeko kategoria etiketa unibertsalak biltzeko egin zen lehen ahalegina *Universal Dependency Treebank* (UDT) izeneko proiektuarekin hasi zen (McDonald *et al.* 2013). 2013. urtean etiketatze estandarra zuten 6 hizkuntzatarako zuhaitz-bankuak bildu ziren eta 11 hizkuntzatarako zuhaitz-bankuak hurrengo urtean. Ezaugarri morfologikoak ere estandarizatuta sartzeko lehen proposamena 2013an egin zen (Tsarfaty 2013).

Bestalde, HamleDT proiektuaren bigarren bertsioan (Rosa *et al.* 2014), 30 hizkuntzatarako Stanfordeko dependentziak eta Googleeko kategoria etiketak erabiltzen zituzten zuhaitz-bankuak biltzea lortu zen, honen ostean Stanfordeko dependentzia unibertsalen garapena bukatu zelarik (De Marneffe *et al.* 2014).

Esandakoa laburtzeko, kapitulu honetan landutakoa, aurretik aipatu diren ahalegin guztien bildura dela esan daiteke, zeinetan Stanfordeko dependentzia etiketa unibertsaletan oinarritutako etiketak, Googleeko kategoria etiketa unibertsalen bertsio hedatua, Interset ezaugarrien azpimultzo gainbegiratua eta CoNLL-X formatuaren bertsio gainbegiratua erabiltzen diren; proiektu horretarako gure ekarpena, euskarazko zuhaitz-bankua Dependentzia Unibertsalen formatu estandarrera bihurtzea izan da (CoNLL-U edo UD formatua deiturikoa).

Dependentzia Unibertsalak proiektuko partaideek hizkuntza desberdinetan egin duten lana hobeto kokatzeko euskararen, hungarieraren eta suedieraren kategoriak, ezaugarri morfologikoak eta dependentzia erlazioak UD formatuan eskatzen direnekin kuantitatiboki konparatuko dira. Euskarazko jatorrizko zuhaitz-bankuak 16 kategoria desberdin ditu, ezaugarri morfologikoen 69 bikote (esaterako $KAS = ABS$ bikotea) eta 30 dependentzia sin-

Hizkuntza	Hitz kopurua	Hizkuntza	Hitz kopurua	Hizkuntza	Hitz kopurua
Greko zaharra	244.000 206.000	Alemana	293.000	Poloniera	83.000
Arabiera	282.000	Gotikoa	56.000	Portugalera	226.000
Portugalera Brasiliera	298.000	Bulgariera	156.000	Hebreera	115.000
Errumaniera	12.000	Katalana	530.000	Indiera	351.000
Errusiera	99.000	Txinera	123.000	Hungariera	26.000
Esloveniera	140.000	Kroaziera	87.000	Indonesiera	121.000
Espainiera	423.000 547.000	Txekiera	1.503.000	Irlandera	23.000
Suediera	96.000	Daniera	100.000	Italiera	252.000
Tamilera	9.000	Nederlandera	209.000	Japoniera	267.000
Estoniera	9.000	**Euskara**	121.000	Grekoa	59.000
Ingelesa	254.000	Latina	47.000 259.000 165.000	Norvegiera	311.000
Finlandiera	181.000 159.000	Eliza Zaharreko Eslovakiera	57.000	Frantsesa	389.000
Persiera	151.000				

6.1 taula – Dependentzia Unibertsalak proiektuan parte hartu duten hizkuntzen bilduma. Laster hizkuntza gehiago espero dira.

taktiko desberdin. Hungarierako jatorrizko zuhaitz-bankuan ere 16 kategoria desberdin daude, ezaugarri morfologikoen 70 bikote eta 32 dependentzia etiketa desberdin. Suedieraren kasuan, aldiz, 15 dira kategoriak, 27 ezaugarri morfologiko bikoteak eta 35 dependentzia sintaktiko desberdinak. UD zuhaitz-bankuek 17 kategoria desberdin dituzte, 17 ezaugarri morfologiko mota[2] eta 37 dependentzia etiketa desberdin.

Kapitulu honetako sarrerarekin bukatzeko, Dependentzia Unibertsalak proiektuan parte hartu duten hizkuntza guztiei buruzko informazioa bildu dugu 6.1 taulan, kasu bakoitzean sortu den zuhaitz-bankuaren tamaina adieraziz. Zenbait hizkuntzatan zuhaitz-banku bat baino gehiago daudela ikus daiteke. Horren arrazoia erakunde desberdinek sortutako zuhaitz-bankuak direla da, tamaina eta eduki desberdinekin.

Kapitulu honetan azaltzeko gelditzen diren atalak modu honetara antolatu dira: 6.2 atalean, UD formatuaren eta euskarazko zuhaitz-bankuaren ezaugarriak eta desberdintasunak azalduko dira, baita bihurketa egiteko jarraituko diren irizpide nagusiak ere. 6.3 atalean, zuhaitz-bankuko etiketa mota bakoitza bihurtzeko jarraitu dugun prozesua zehaztuko da (kategoriak,

[2]Hizkuntza bakoitzaren arabera ezaugarri morfologiko bikoteak desberdinak dira

ezaugarriak eta dependentzia erlazioak), eta 6.4 atalean, egindako bihurketa prozesua aztertu ondoren ateratako ondorioak azalduko ditugu zenbait etorkizuneko lanekin batera.

6.2 Euskarazko Dependentzia Unibertsaletarako irizpideak

Aurretik esan bezala, kapitulu honetan 3.2.1.1 puntuan azaldu dugun euskarazko zuhaitz-bankua Dependentzia Unibertsalak proiektuan ematen diren gidalerroak[3] jarraituta UD formatura pasatu dugu. Formatu horretako kategoriak Googleeko kategoria unibertsaletan oinarrituta daude, baina birdefinitu egin dira UD espezifikazioak betetzen dituen formatu berrira egokitzeko. Berdina gertatzen da ezaugarri morfologiko eta dependentziekin ere; ezaugarri morfologikoak Interset ezaugarrietan oinarrituta dauden UD espezifikazioak jarraitzen dituzten ezaugarrietara bihurtu dira, eta euskarazko dependentziak, Stanfordeko dependentzia etiketa unibertsaletan oinarritutako etiketetara bihurtu dira. Euskarazko zuhaitz-bankuaren formatua, aldiz, 3.1.1.3 atalean azaldu dugun ConLL-U formatura bihurtu da. Zuhaitz-bankuaren bihurketa egiteko jarraitu ditugun irizpide nagusiak bi izan dira:

- **Automatikoa:** Bihurtu nahi den hitz kopurua milaka hitzekoa izanik, ezinbestekoa da bihurketa prozesua automatikoa izatea ahal den denbora eta baliabide (pertsonak) gutxien inplikatzeko.

- **Zuzena:** Nahiz eta bihurketa automatikoki egin den, bihurtu diren esaldiak guztiz zuzenak izatea da gure helburua. Horretarako, zalantzazko kasuak baztertu egin dira, ziurtasun handiarekin ondo dauden esaldiak bakarrik bihurtuz.

Euskarazko zuhaitz-bankua, sortu zen momentutik izan da CoNLL-X formatura bihurgarria den dependentzietan oinarritutako zuhaitz-bankua. Maila desberdinetan etiketatuta dagoenez, esaldian agertzen den hitz-forma bakoitzari buruzko informazio ugari biltzen da bertan: lema, kategoria, azpikategoria, ezaugarri morfosintaktikoak eta dependentzia erlazioa esaldiko zein hitzekin duen. Kasua, numeroa, mugatasuna edo mendeko esaldi mota bezalako ezaugarri morfosintaktikoak hitzen erroei lotutako aurrizki edo

[3] `http://universaldependencies.org/`

atzizkiei dagozkie. Jatorrizko euskarazko zuhaitz-bankuaren tamaina gutxi gora-behera 150.000 hitzekoa da (11.225 esaldi), dependentzia erlazioek osatzen dituzten arkuen % 1,3a arku ez-proiektiboak izanik. Zuhaitz-bankuan aurki daitezkeen dependentzien mota kopurua 28 da, kategorien kopurua 16 eta ezaugarri morfosintaktikoen mota kopurua 10. Ezaugarri morfosintaktikoen mota bakoitzak har ditzakeen balio kopurua desberdina izan daiteke, adibidez kasu morfologikoak 14 balio har ditzake (genitiboa, ergatiboa, absolutiboa...) eta numeroak 2 (singularra eta plurala).

UD zuhaitz-bankuarekin zenbait ezaugarri ditu komunean (ikus 6.1 adibidea): biek jarraitzen dute sintaxiaren hurbilpen lexikalista, hau da, dependentzia erlazioak zatitu gabeko hitz-formen artean gertatzen dira, eta izen-sintagmen eta aditz-kateen buruak zeintzuk diren ere partekatzen dute. Bestalde, jatorrizko zuhaitz-bankuko kategoriak eta ezaugarri morfosintaktikoak bihurtzea nahiko prozesu gardena eta zailtasun handirik gabekoa izan da. Esanak esan, UD formatuan lortutako zuhaitz-bankuaren tamaina 121.443 hitzekoa da, 8.993 esaldi. Datuak aztertu ondoren, bidean ia 30.000 hitz gelditu direla esan daiteke, baina ez dira 30.000 hitz utzi ezin izan direlako hitz horiek UD formatura bihurtu. Irizpide nagusietan aipatu bezala, esaldi bateko hitz bat ezin bada bihurtu, esaldi osoa baztertu behar da. Beste modu batera esanda, bihurtu gabe geratu diren hitzak 6.500 hitz baino gutxiago izan dira, gainontzekoak horien ondorioz baztertu direlarik.

```
1       Protestak       protesta        IZE     IZE_ARR BIZ:-|KAS:ABS|NUM:P|MUG:M       2       ncsubj
2       ugaritu         ugaritu         ADI     SIN     ADM:PART|ASP:BURU               0       ROOT
3       dira            izan            ADL     ADL     MDN:A1|NOR:HAIEK                2       auxmod
4       ekialdean       ekialde         IZE     ARR     KAS:INE|NUM:S|MUG:M             2       ncmod
5       energia         energia         IZE     ARR     BIZ:-                           6       ncmod
6       mozketak        mozketa         IZE     ARR     KAS:ABS|NUM:P|MUG:M             7       ncobj
7       salatzeko       salatu          ADI     SIN     ADM:ADIZE|ERL:HELB|KAS:ABS      2       xmod
8       .               .               PUNT    PUNT_PUNT       _                       7       PUNC

1       Protestak       protesta        NOUN    Animacy=Inan|Case=Abs|Definite=Def|Number=Plur  2       nsubj
2       ugaritu         ugaritu         VERB    Aspect=Perf|VerbForm=Part                       0       root
3       dira            izan            AUX     Mood=Ind|Number[abs]=Plur|Person[abs]=3         2       aux
4       ekialdean       ekialde         NOUN    Case=Ine|Definite=Def|Number=Sing               2       nmod
5       energia         energia         NOUN                    _                               6       nmod
6       mozketak        mozketa         NOUN    Case=Abs|Definite=Def|Number=Plur               7       dobj
7       salatzeko       salatu          VERB    Case=Abs|Definite=Ind                           2       advcl
8       .               .               PUNCT                   _                               2       punct
```

6.1 irudia – Goian, euskarazko zuhaitz-bankuko adibide bat. Behean, esaldi bera UD formatuan.

Dependentzia erlazioak bihurtzeko garaian, aldiz, desberdintasunak aurkitu ditugu zenbait fenomeno linguistiko etiketatzeko eran eta moldaketak egin behar izan ditugu bihurketa ahalik eta egokiena izateko; bihurtu gabe gelditu diren hitzak fenomeno horiekin erlazionatuta dauden hitzak dira

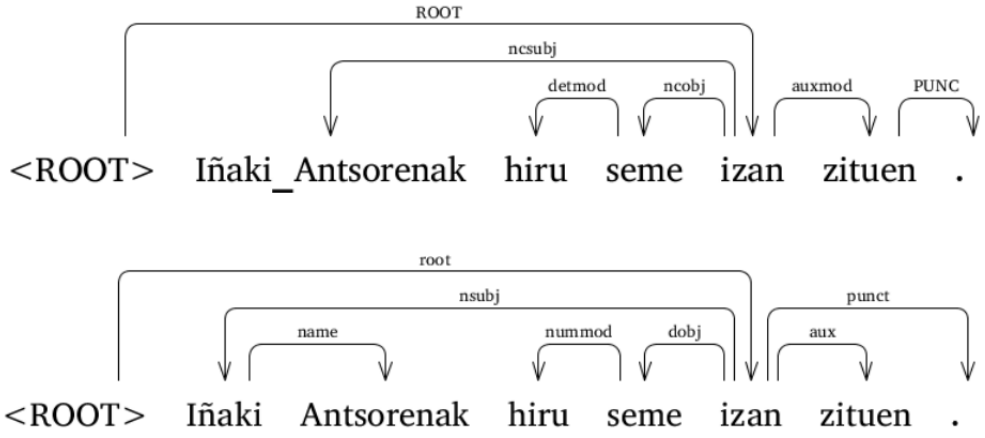
gehienbat. Hori dela eta, fenomeno horiek zeintzuk diren azalduko dugu dependentzia erlazioen bihurketa azaltzen dugun atalean (6.3.4 atala), fenomeno horiek dependentzia erlazioei hertsiki lotuta baitaude.

6.3 Euskarazko zuhaitz-bankuaren bihurketa

Aurretik aipatu den bezala, euskarazko zuhaitz-bankuaren bihurketa prozesuan hainbat atal daude. Atal batzuetan euskarazko etiketen bihurketa zuzena egin daiteke, baina beste batzuetan zuhaitzaren egitura aldatu behar da etiketen bihurketa aplikatu baino lehen. Argi utzi nahi da bihurketa prozesua ez dela izan euskarazko etiketak hartzea eta beraiei dagozkien UD etiketetara bihurtzea bakarrik. Prozesua konplexua eta luzea da, eta hainbat atal ditu: hitz anitzeko esapideen elementuen analisiak lortu, puntuazioa moldatu, koordinazioa moldatu, etiketa bakoitzerako kasuak aztertu eta bere UD parea identifikatu, ordena zuzena erabaki... Gauzak horrela, ondorengo lerroetan euskarazko jatorrizko zuhaitz-bankua UD formatura bihurtzean emandako urratsak azalduko ditugu zehaztasunez. Azalpena lau ataletan banatu dugu: hitz anitzeko esapideak bihurtzeko jarraitu ditugun urratsak azaltzeko atala, kategorien bihurketa azaltzekoa, ezaugarri morfologikoen bihurketa azaltzekoa eta dependentzien bihurketa azaltzekoa.

6.3.1 Hitz anitzeko esapideak

Euskarazko zuhaitz-bankuak hitz anitzeko esapideak osatzeko hitzak multzokatzea ahalbidetzen duen arren, UD gidalerroen arabera hitzak biltzeko joera hori saihestu egin behar da (ikus 6.2 irudia). Hori dela eta, euskarazko zuhaitz-bankuko hitz anitzeko esapideak hitz bakunetan banatu behar izan dira. Era desberdinetan sorturiko hitz anitzeko esapideak daude, baina oraingoz kategoria eta azpikategoria konbinazio erabilienak dituztenak bihurtu dira UD formatura.



6.2 irudia – Goian, euskarazko zuhaitz-bankuko adibide bat, hitz anitzeko esapide bat (*Iñaki Antsorenak*) bilduta azaltzen delarik bertan. Behean, esaldi bera UD formatuan, baina hitz anitzeko esapidea banatuta.

Bihurketa prozesuaren funtsa, dependentzia erlazioa esleitzen den bitartean hitz anitzeko esapidea osatzen duten hitz bakoitzaren lema, kategoria, azpikategoria eta ezaugarri morfosintaktikoak berreskuratzean datza. Ildo beretik jarraituz, hitz anitzeko esapidearen burua eta semea(k) zeintzuk diren zehazteak arreta berezia merezi du. Izan ere, kasu batzuetan hitzak deklinatu egin baitaitezke burua edo semea izateko aukera desberdinak emanez. Har dezagun adibide bezala *mendiaren gainean* postposizio konplexua. Kasu honetan, postposizioa osatzen duten bi hitzak daude deklinatuta, kasu genitiboarekin eta inesiboarekin, hurrenez hurren. Hasiera batean hitz anitzeko esapidearen burua *mendiaren* hitzak dirudien arren, genitiboak osagarri bezala jokatzen du eta burua *gainean* dela iradokitzen du.

Hitz anitzeko esapideen bihurketa hiru atazatan banatu dugu:

- **Hitz bilduen bihurketa:** *behar da*, *ezin dugu*, *uste dut*, *nahi dute...*
- **Izen berezien bihurketa:** *Patxi Lopez*, *Gaizka Garitanoren*, *Xarm el Xeikhen*, *Ariel Sharonek...*
- **Postposizio konplexuen bihurketa:** *mendiaren gainean*, *egun hartaz gero*, *adiskide bezala*, *Afrikan barrena...*

Sortuko den UD zuhaitz-bankuaren kalitatea ahalik eta altuena izateko, ataza horietako bakoitzerako azterketa linguistiko bat egin da. Zoritxarrez,

jatorrizko zuhaitz-bankuko 4.922 hitz anitzeko esapide desberdinetik 1.904 bihurtu ahal izan dira. Denak ez bihurtzeko arrazoia, kasu gehienetan hitz anitzeko esapideak osatzen dituzten hitz bakunen analisia falta dela da.

Esanak esan, ikus dezagun zehaztasunez kasu bakoitzean jarraitu den prozesua.

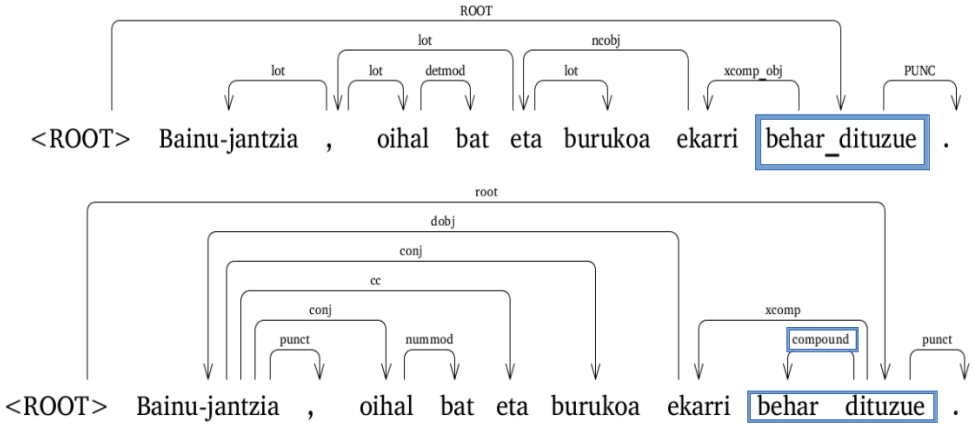
6.3.1.1 Hitz bilduak

Hitz bilduak banatzeko IXA taldeko amaraunean oinarritu gara (Artola *et al.* 2005). Bertan, milaka hitz bildu osatzen dituzten hitzen eskuz errebisatutako analisi bakunak gordetzen dira. Modu honetara, jatorrizko zuhaitz-bankuko hitz anitzeko esapide bat banatu behar den bakoitzean, termino hori osatzen duten hitz bakunen analisia amaraunean gordeta dagoen aztertzen da. Analisia ez badago gordeta, esapidea baztertu egiten da esaldi osoarekin batera. Bestela, hitz bakunei amarauneko analisia esleitzen zaie eta zein dependentzia erlazioren bidez lotuko diren erabakitzen da. Erabaki hori hartzeko, elementu bakoitzaren hitz-forma, kategoria, azpikategoria eta ezaugarri morfologikoetan oinarritu gara, kasu bakoitzean behar ditugunak erabiliz (ikus 6.2 taula). Jatorrizko zuhaitz-bankuan dauden 1.282 hitz bilduetatik 536 bihurtu dira.

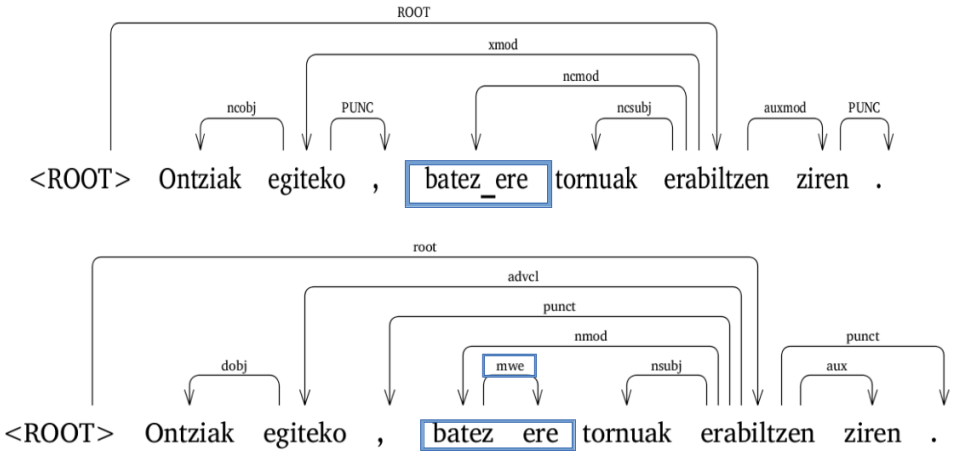
1. hitza	2. hitza	3. hitza	UD erlazioa	Adibidea
Izen arrunta	Aditz sinplea		compound:lvc	*ahal* ***izateko***
Izen arrunta	Aditz trinkoa		compound:lvc	*behar* ***da***
Izenondoa	Aditz trinkoa		compound:lvc	*bizi* ***dugun***
Izenondoa	Aditz sinplea		compound:lvc	*bizi* ***izan***
Genitiboan			nmod	*bien* ***bitartean***
Hitz-anitzeko osagaia	Aditz trinkoa		compound	*ari* ***dira***
Leku izen berezia	Izen arrunta		compound	*Europako* ***Batasunaz***
Izen arrunta	Izen arrunta		compound	*gisa* ***eskubideen***
Izen arrunta	Aditzondoa		compound	*behar* ***bezala***
Aditzondoa	Izen arrunta		compound	*bertan* ***behera***
Izen arrunta	Leku izen berezia		compound	*Hego* ***Amerika***
Izan	*ere*		compound	-
Det zenbatzaile zehaztugabea	Aditzondoa		compound	*besterik* ***gabe***
Loturazko juntagailua	Izen arrunta		compound	*eta* ***abar***
Izen arrunta	Izenondoa		amod	***Estatu*** *Batuetako*
Pertsona izen berezia	Izenondoa		amod	***Maria*** *ahalguztiduna*
Leku izen berezia	Izenondoa		amod	***Britainia*** *Handia*
Aditzondoa	Aditzondoa		mwe	***Gaur*** *egun*
Aditzondoa	Loturazko lokailua		mwe	***Hala*** *ere*
Det zenbatzaile zehaztua	Loturazko lokailua		mwe	***batez*** *ere*
Loturazko lokailua	Loturazko juntagailua		mwe	***nahiz*** *eta*
besteak	*beste*		mwe	-
Izen arrunta	Det erakusle indartua		mwe	***aldi*** *berean*
Det zenbatzaile zehaztua	Det zenbatzaile zehaztua		mwe	***batik*** *bat*
Izen arrunta	Loturazko juntagailua		mwe	***ahalik*** *eta*
izan	*ezik*		mwe	-
Aditzondoa	Izenondoa		mwe	***hain*** *zuzen*
Det erakusle arrunta	Izen arrunta		mwe	***horrez*** *gain*
Aditzondoa	Aditzondo galdetzailea		mwe	***hala*** *nola*
Det zenbatzaile zehaztu ord	Izen arrunta		amod	*azken* ***aldian***
Izen arrunta	Det zenbatzaile zehaztua		nummod	***hein*** *bat*
gutxi	*batzuk*		det	-
Aditzondoa	Aditz sinplea		advmod:lvc	*gainezka* ***egin***
Izen arrunta ez genitiboa	Izen arrunta ez genitiboa	Izen arrunta	compound compound	*Euskal jai* ***taldea***
Izen arrunta genitiboa	Izen arrunta genitiboa	Izen arrunta	nmod nmod	*Gipuzkoako kanpuseko* ***ikasleak***
Izen arrunta genitiboa	Izen arrunta ez genitiboa	Izen arrunta	nmod compound	*Gipuzkoako herri* ***asanblada***
Izen arrunta ez genitiboa	Izen arrunta genitiboa	Izen arrunta	compound nmod	*Euskal Herriko* ***Unibertsitatea***

6.2 taula – UD formatura bihurtzean hitz bilduei kasu bakoitzean esleitu zaizkien dependentzia erlazioak eta adibideak.

Hitz bilduak UD formatura bihurtzeko prozesuari buruzko azalpenekin bukatzeko, horrelakoak dituzten pare bat esaldiren egitura sintaktikoak erakusten dira, 6.3 eta 6.4 irudiak, UD formatura pasatu aurretik eta ondoren.



6.3 irudia – Goian, jatorrizko zuhaitz-bankuko esaldi baten egitura sintaktikoa. Behean, goikoaren parekoa den UD zuhaitz-bankuko esaldi baten egitura sintaktikoa. *behar_dituzue* hitz bildua banaturik ageri da bertan eta bere osagaiak *compound* dependentzia erlazioarekin lotuta.



6.4 irudia – Goian, jatorrizko zuhaitz-bankuko esaldi baten egitura sintaktikoa. Behean, goikoaren parekoa den UD zuhaitz-bankuko esaldi baten egitura sintaktikoa. *batez_ere* hitz bildua banaturik ageri da bertan eta bere osagaiak *mwe* dependentzia erlazioarekin lotuta.

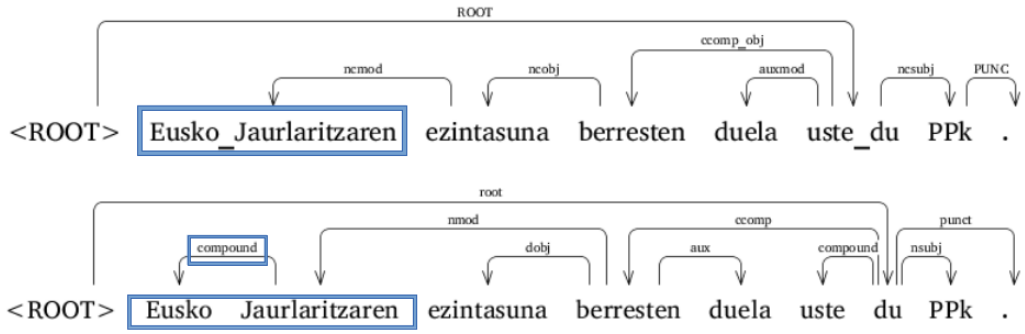
6.3.1.2 Izen bereziak

Izen bereziak banatzeko ere Ixa taldeko amaraunean oinarritu gara. Bertan, milaka izen berezi osatzen dituzten hitzen analisi bakunak gordetzen dira. Hitz bilduekin gertatzen den bezala, behar dugun analisia ez badago gordeta, terminoa baztertu egiten da esaldi osoarekin batera. Bestela, hitz bakunei amarauneko analisia esleitzen zaie eta zein dependentzia erlazioren bidez lotuko diren erabakitzen da. Hori erabakitzeko, elementu bakoitzaren hitz-forma, kategoria, azpikategoria eta ezaugarri morfologikoetan oinarritu gara, kasu bakoitzean behar ditugunak erabiliz (ikus 6.3 taula). Taula aztertzen badugu, kasu batzuetan hitz bilduetan hartu diren erabaki berdinak hartu direla ikusiko dugu. Izan ere, bi hitz anitzeko esapide mota horien arteko muga ez baitago beti argi. Jatorrizko zuhaitz-bankuan dauden 1.943 izen berezietatik 856 bihurtu dira.

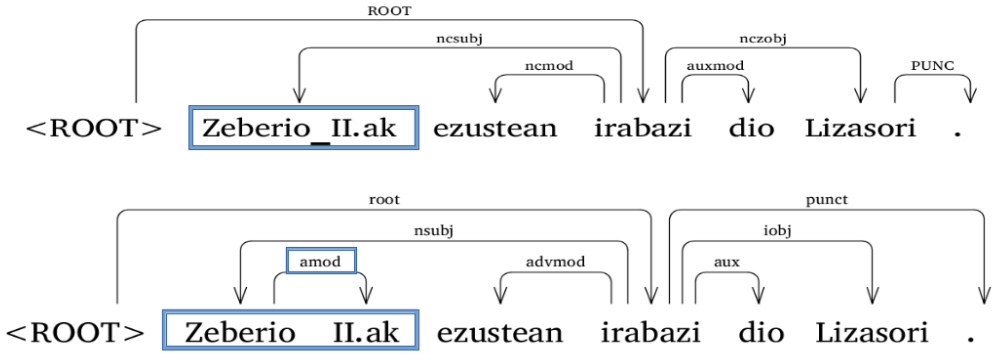
1. hitza	**2. hitza**	**3. hitza**	**UD erlazioa**	**Adibidea**
Genitiboan			nmod	*Heriotzaren* ***karabana***
Izen arrunta	Ez izenondoa Ez aditz sinplea		compound	*Eusko* ***Basquek***
Ez genitiboan	Izen arrunta		compound	*Frantziako* ***Ligan***
	Det zenbatzaile zehaztu ordinala		amod	***Egiguren*** *II.a*
Izen arrunta	Izenondoa		amod	***Lege*** *Gorenak*
Izen arrunta	Aditz sinplea Partitiboan		amod	***Indar*** *Armatu*
Pertsona izen berezia	Izenondoa		amod	***Arturo*** *Morearen*
Leku izen berezia	Pertsona izen berezia		name	*Donostiako* ***Martin***
Det zenbatzaile zehaztu ordinala	Izenondoa		amod	*Bigarren* ***Gorria***
Izen arrunta	Aditz sinplea aditz izena konpletiboa		amod	*Euskal* ***Herriratzea***
Pertsona izen berezia	Pertsona izen berezia		name	***Hugo*** *Gutierrez*
Pertsona izen berezia	Pertsona izen berezia	Pertsona izen berezia	name name	***Miguel*** *Angel Lotina*
Genitiboan	Genitiboan		nmod nmod	*Aznarren* ***Gobernuaren***
Izen arrunta	Izen arrunta genitiboan	Izen arrunta	compound nmod	*Gizarte Segurantzaren* ***kontua***
Izen arrunta	Leku izen berezia genitiboan	Izen arrunta	compound nmod	*Espetxe Madrildarraren* ***atean***
Izen arrunta	Izenondoa genitiboan	Izen arrunta	amod nmod	*Gurutze Gorriaren* ***autoa***
Pertsona izen berezia	Pertsona izen berezia genitiboan	Izen arrunta	name nmod	*Luis Fernandezen* ***alaba***
Pertsona izen berezia	Leku izen berezia genitiboan	Izen arrunta	name nmod	*Juan Donostiaren* ***liburua***
Leku izen berezia	Pertsona izen berezia genitiboan	Izen arrunta	name nmod	*Donostia Maxuxen* ***heriotza***
Leku izen berezia	Leku izen berezia genitiboan	Izen arrunta	name nmod	*Deutsche Banken* ***zigorra***

6.3 taula – UD formatura bihurtzean izen bereziei kasu bakoitzean esleitu zaizkien dependentzia erlazioak eta adibideak.

Hitz bilduen atalean egin den bezala, egindako lana hobeto ulertzeko, izen bereziak dituzten pare bat esaldiren egitura sintaktikoak erakutsiko dira, 6.5 eta 6.6 irudiak, UD formatura pasatu aurretik eta ondoren.



6.5 irudia – Goian, jatorrizko zuhaitz-bankuko esaldi baten egitura sintaktikoa. Behean, goikoaren parekoa den UD zuhaitz-bankuko esaldi baten egitura sintaktikoa. *Eusko_ Jaurlaritzaren* izen berezia banaturik ageri da bertan eta bere osagaiak *compound* dependentzia erlazioarekin lotuta.



6.6 irudia – Goian, jatorrizko zuhaitz-bankuko esaldi baten egitura sintaktikoa. Behean, goikoaren parekoa den UD zuhaitz-bankuko esaldi baten egitura sintaktikoa. *Zeberio_ II.ak* izen berezia banaturik ageri da bertan eta bere osagaiak *amod* dependentzia erlazioarekin lotuta.

6.3.1.3 Postposizio konplexuak

Bihurtu beharreko hitz-anitzeko terminoa postposizio konplexua denean, ez dugu jarraitu aurreko mota bietan jarraitu dugun bide bera. Izan ere, tamalez, kasu honetan ezin izan baikara Ixa taldeko amaraunean oinarritu, postposizio konplexuei buruzko informazioa ez dagoelako biltegiratuta. Ondorioz, postposizio konplexuak bihurtzeko beste prozedura bat jarraitu dugu

eta jatorrizko zuhaitz-bankuan dauden 1.697 postposizio konplexuetatik 512 bihurtu dira. Jatorrizko zuhaitz-bankuan postposizioak etiketa berezi baten bidez markatuta daudenez, etiketa hori jasotzen dugun bakoitzean ondorengo prozesua jarraitzen dugu postposizioa banatzeko:

1) Bilduta dagoen postposizio konplexuaren lema banatu. Atal bakoitza postposizioaren osagai bakoitzaren lema izango da. Demagun *zalantzarik_gabe* postposizioa bihurtu nahi dugula. Bere lema *zalantza_gabe* da, orduan banatutako postposizioaren lehen osagaiaren (*zalantzarik*) lema *zalantza* izango da eta bigarren osagaiarena *gabe*.

2) Bildutako postposizioaren kategoria eta azpikategoria lehen osagaiari esleitu. Adibidez, *ate_ondoan* postposizioaren kategoria izen motakoa da eta azpikategoria izen arrunt motakoa. Orduan, *ate* hitzari esleitzen zaizkio kategoria eta azpikategoria horiek.

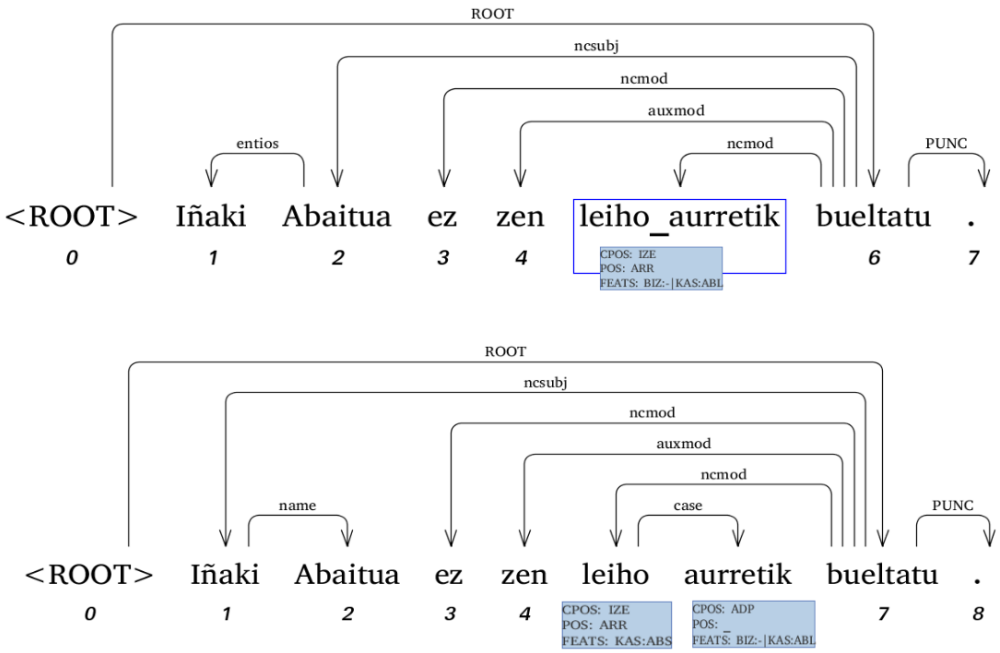
3) Postposizio konplexuaren osagaien kasuak erauzi hitz-formen atzizkietan oinarrituta (azken osagaiarena izan ezik). 12 kasu mota landu dira. Adibidez, *mendiaren_gainean* postposizioaren lehen elementuaren kasua genitiboa da, *mendiaren* hitz-forma -en atzizkiarekin bukatzen baita.

4) Postposizioaren azken osagaiari, bildutako postposizioaren ezaugarri morfologikoak esleitu, kasu gehienetan bildutako postposizioaren ezaugarriak azken hitzari dagozkion ezaugarriak baitira. Adibidez, *epaitegiaren_aurrean* postposizio konplexuaren ezaugarri morfologikoak dira inesiboan dagoela, ez dela biziduna eta numero singularra duela, *aurrean* hitzari dagozkionak, hain zuzen ere.

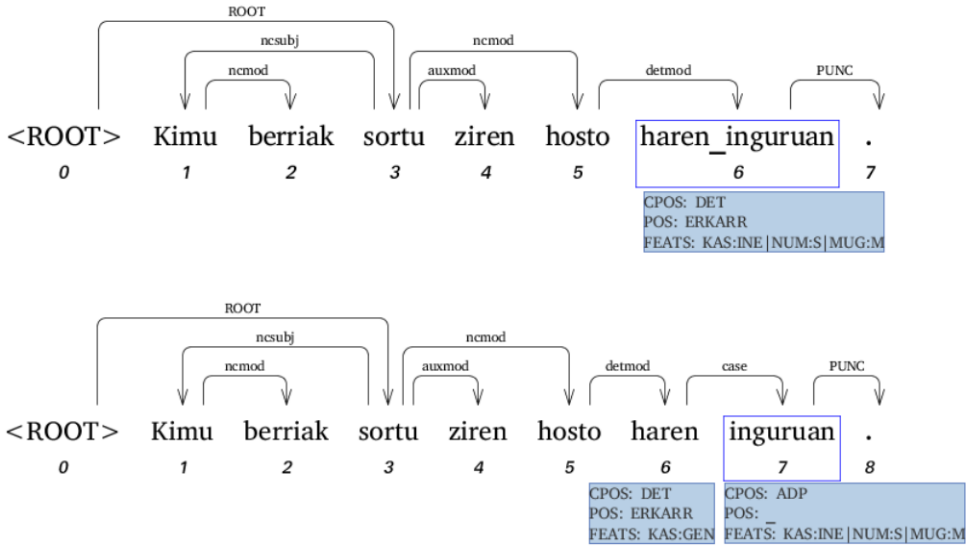
5) Postposizioaren azken osagaiari *ADP* kategoria esleitu (*adposition*) UD gidalerroek postposizioentzako esleitzea gomendatzen duten bezala.

6) Postposizioaren elementuak *case* dependentzia erlazioaren bidez lotu beraien artean.

Azaldutakoa hobeto ulertzeko, 6.7 eta 6.8 irudietan bildu ditugun bi adibideetan zentratuko gara. Adibide horietan, *leiho_aurretik* eta *haren_inguruan* postposizio konplexuen bihurketa prozesuaren hasierako egoera eta bukaerako egoera irudikatzen dira, goian azaldutako urratsak erraz identifikatzen direlarik.



6.7 irudia – *Leiho_ aurretik* postposizio konplexuaren bihurketaren hasierako eta bukaerako egoerak.



6.8 irudia – *haren_inguruan* postposizio konplexuaren bihurketaren hasierako eta bukaerako egoerak.

6.3.2 Kategoriak

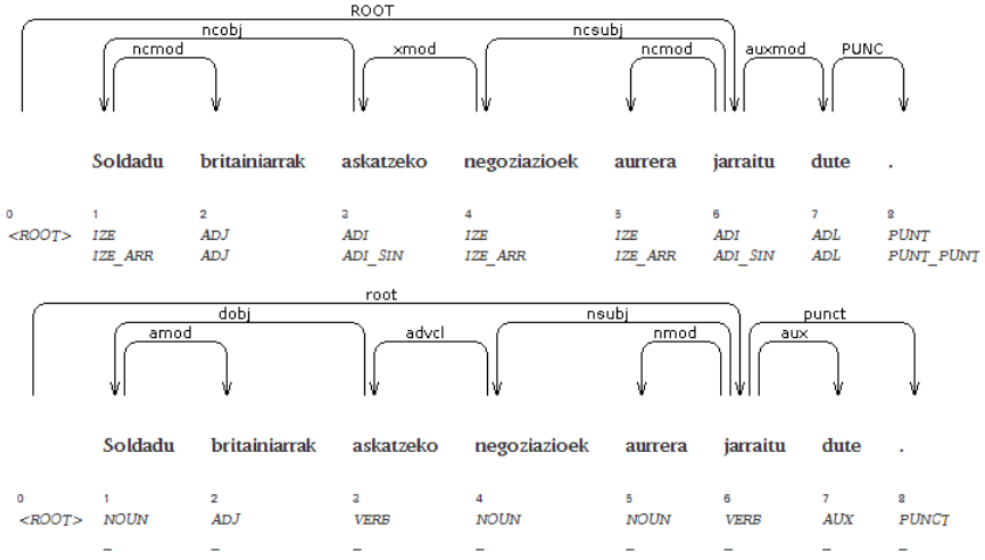
Jatorrizko zuhaitz-bankuko kategorien bihurketa prozesua azalduko dugu atal honetan. Esan beharra dago, jatorrizko kategoria bakoitzari dagokion UD kategoria zein den jakiteko jatorrizko kategoriak eta UD gidalerroak sakon aztertu direla. Modu honetara, kategoria bakoitzari dagokion UD etiketa esleitu zaio. Izan ere, kasu askotan bihurketa zuzena egin daitekeen arren, beste kasu batzuetan bihurketa ez baita nabaria.

Jatorrizkoa		**UD**	**Adibidea**
Adjektiboa		ADJ	*gorria*
Adizlaguna		AUX	*zuen*
Interjekzioa		INTJ	*to*
Aditz perifrastikoa		ADP	*gainean*
Puntuazio marka		PUNCT	.
Aditza		VERB	*izan*
Determinatzailea	Banatzailea	DET	*gehiago*
	Zenbatzaile zehaztua	NUM	*batekin*
	Zenbatzaile zehaztu ordinala	NUM	*lehena*
Aditzondoa		ADV	*bezala*
Loturazkoa		CONJ	*eta*
Partikula		PART	*ez*
Izenordaina		PRON	*gure*
Bereiz		PUNCT	
Aditz trinkoa		VERB	*zekien*
Izena	Berezia	PROPN	*Lazkao*
	Zenbakia	NUM	*1995eko*
	Sinboloa	SYM	*kg*
	Arrunta	NOUN	*txartela*
Bestelakoak	Laburtzapena	NOUN	*LABeko*
	Bestela	X	*de*

6.4 taula – UD formatura bihurtzean jatorrizko kategoriei kasu bakoitzean esleitu zaizkien UD kategoriak eta adibideak.

Esanak esan, bihurketa egiteko jarraitu dugun prozesua ondorengoa izan da: jatorrizko zuhaitz-bankutik hitz bat irakurtzen den bakoitzean, 6.4 taulan bere kategoriari dagokion UD kategoria esleitzen zaio. Taula aztertuta kasu batzuk ez direla zuzenean esleitzen konturatuko gara. Esaterako, jatorrizko kategoria izena denean. Kasu horietan, hitzaren azpikategoria edo hitz-forma ere kontuan hartzen dira jatorrizko kategoriari dagokion UD etiketa egokia zein den erabakitzeko.

Kategorien bihurketa prozesuaren ikuspegi zabalagoa izateko, jatorrizko esaldien bihurketa irudikatzen duen adibide bat aukeratu dugu. 6.9 irudian, bihurketa hasi baino lehen eta bihurketa amaitu ondoren esaldi berdinak dituen kategoria eta dependentzia etiketak irudikatu dira.



6.9 irudia – Goian, jatorrizko zuhaitz-bankuko esaldi baten kategoria, azpikategoria eta dependentzia zuhaitza. Behean, esaldi berdinaren UD bertsioa, kategoriak eta dependentzia erlazioak UD formatuan.

6.3.3 Ezaugarri morfologikoak

Kategoriekin egin den antzera, jatorrizko ezaugarri morfologikoei dagozkien UD ezaugarriak zeintzuk diren azalduko dugu ondorengo lerroetan. Jatorrizko etiketak eta UD gidalerroak aztertu ondoren jatorrizko ezaugarri batzuk zuzenean bihurtu daitezkeela ikusi dugu, baina beste etiketa batzuk bihurtzeko, hitzaren beste ezaugarri batzuetan oinarritu behar izan gara. 6.5 eta 6.6 tauletan, bihurketa zuzena aplikatu ahal izan zaien jatorrizko etiketei esleitu zaizkien etiketa berriak bildu ditugu. Taulak modu honetara antolatu dira: lehenengo zutabean jatorrizko ezaugarriak bildu dira, bigarrenean, aldiz, ezaugarri horien UD bertsioak. Hirugarrenean, ezaugarri bakoitzak har ditzakeen balioak daude eta laugarrenean, beraien UD bertsioak. Bosgarren zutabean kasu bakoitzaren adibide bat sartu dugu, etiketa bakoitza hobeto ulertzeko.

Esan bezala, zuzeneko UD ordainik ez duten ezaugarriak ere badaude jatorrizko zuhaitz-bankuan. Are gehiago, kasu batzuetan, hitzaren kategoria eta azpikategorian oinarrituta sortu behar dira UD ezaugarri berriak. Esaterako, aztertzen ari garen hitza izenordaina denean, *PronType* izeneko ezau-

Ezaugarria	UD Ezaugarria	Balioa	UD Balioa	Adibidea
Numeroa	Number	Singularra	Sing	*etxea*
		Plurala	Plur	*etxeak*
Kasua	Case	Absolutiboa	Abs	*lagun*
		Ergatiboa	Erg	*Nik*
		Datiboa	Dat	*Niri*
		Lekuzko genitiboa	Loc	*beheko*
		Genitiboa	Gen	*zure*
		Instrumentala	Ins	*iritziz*
		Soziatiboa	Com	*zurekin*
		Inesiboa	Ine	*etxean*
		Adlatiboa	All	*autora*
		Ablatiboa	Abl	*basotik*
		Motibatiboa	Cau	*nigatik*
		Destinatiboa	Ben	*hiretzat*
		Partitiboa	Par	*jotzerik*
		Adlatibo bide zuzenezkoa	Lat	*gorantz*
		Prolatiboa	Ess	*hildatzat*
Mugatasuna	Definite	Mugagabea	Ind	*sutsu*
		Mugatua	Def	*bilera*
Maila	Degree	Konparatiboa	Cmp	*sendoagoa*
		Superlatiboa	Sup	*ahulenak*
		Gehiegizkoa	Abs	*altuegia*

6.5 taula – UD formatura bihurtzean jatorrizko ezaugarriei kasu bakoitzean esleitu zaizkien UD ezaugarriak eta adibideak.

garria gehituko da ezaugarri morfologikoen zutabean. Izenordaina arrunta denean, Prontype ezaugarriak *Prs* balioa jasoko du (pertsona izenordaina, pertsona izenordain posesiboa edo determinatzailea), eta *Int* balioa bestela (izenordain galdetzailea, determinatzailea, numerala edo adberbioa). Aztertzen ari garen hitza aditz faktitiboa denean, *Voice* UD ezaugarria sortzen da *Cau* balioarekin (boz kausatiboa).

Zenbakiekin ere antzeko prozesua jarraitzen da. Aztertzen ari garen hitza zenbakia denean, *NumType* UD ezaugarria gehitzen da ezaugarrien zutabean. Zenbakia kardinala denean, UD ezaugarriak *Card* balioa jasoko du (zenbaki kardinala), eta ordinala denean, *Ord* balioa (zenbaki ordinala).

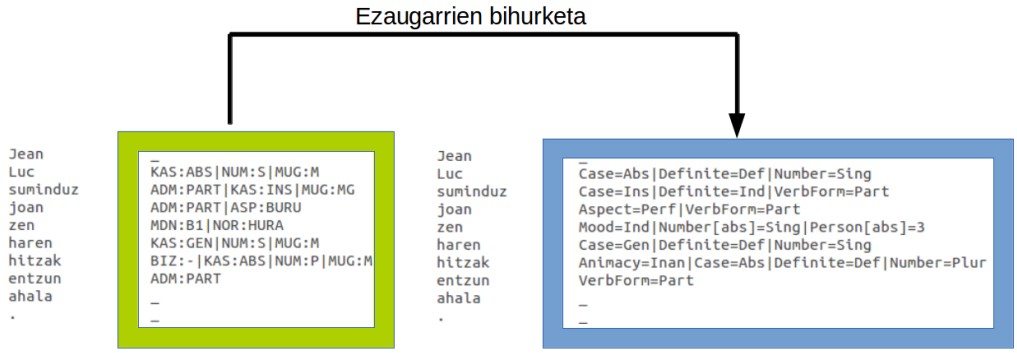
Bestalde, UD ezaugarri bat baino gehiago dagozkien jatorrizko ezaugarriak daude: NOR, NORI, NORK eta HIT (hitanoa). Ezaugarri horiei *Number* eta *Person* UD ezaugarriak dagozkie. Hala ere, hitza hitanoan dagoenean *Polite* ezaugarria gehitzen da eta batzuetan generoa adierazten duen *Gender* ezaugarria ere bai. Ildo beretik jarraituz, NOR ezaugarria *abs* (absolutive) osagaiari lotuta doa, NORK *erg* (ergative) osagaiari, eta NORI *dat* (dative) osagaiari.

Zoritxarrez, ezin izan dira jatorrizko ezaugarri guztiak bihurtu. Kasu

Ezaugarria	UD Ezaugarria	Balioa	UD Balioa	Adibidea
Aditz-mota	Verbform	Aditz-oina	Inf	*erakusten*
		Partizipioa	Part	*izan*
Modua/Denbora	Mood	Indikatibozko orainaldia	Ind	*du*
		Indikatibozko lehenaldia	Ind	*zuten*
		Inperatibozko orainaldia	Imp	*itzazu*
		Subjuntibozko baldintza	Cnd	*badakit*
		Indikatibozko baldintza (orain-gero)	Cnd	*litzateke*
		Indikatibozko baldintza (lehen)	Cnd	*zatekeen*
		Indikatibozko baldintza (aurrekoa)	Cnd	*bagenitu*
		Subjuntibozko baldintza (lehenaldia)	Cnd	*balego*
		Ahalezko orainaldia	Pot	*daiteke*
		Ahalezko lehenaldia	Pot	*genezake*
		Ahalezko lehenaldi arrunta	Pot	*zitakeela*
		Subjuntibozko orainaldia	Sub	*dezala*
		Subjuntibozko alegiazkoa	Sub	*lezan*
		Subjuntibozko lehenaldia	Sub	*zezaten*
		Subjuntibozko lehenaldia (nendin)	Sub	*zintezen*
Biziduna	Animacy	Ez biziduna	Inan	*uztailean*
		Biziduna	Anim	*bertsolaria*
Aspektua	Aspect	Ez burutua	Imp	*erabiltzen*
		Partizipioa	Perf	*egin*
		Etorkizuna	Pro	*engainatuko*
		Puntukaria	Prog	*zituela*
Pertsona	Person	Ni	1	*naiz*
		Hi	2	*haiz*
		Hura	3	*dator*
		Gu	1	*guk*
		Zuek	2	*zuena*
		Zu	2	*zure*
		Haiek	3	*beraien*

6.6 taula – UD formatura bihurtzean jatorrizko ezaugarriei kasu bakoitzean esleitu zaizkien UD ezaugarriak eta adibideak.

batzuetan ez da izan beharrezkoa bihurketa egitea, jatorrizko ezaugarriak adierazten dituen propietateak jadanik bildu direlako. Esaterako, hitz anitzeko esapide bat postposizioa dela adierazten duen POS ezaugarriaren UD ordaina ez da beharrezkoa, UD gidalerroen arabera postposizioa *case* UD dependentzia erlazioaren bidez adierazten baita. Dena den, bihurtu ez diren ezaugarri gehienak ez bihurtzearen arrazoi nagusia UD ordain egokia ez aurkitu izana da. Mota horietako ezaugarrien artean aipagarrienak entitate mota adierazten duen ezaugarria, izenondo bat izenaren aurretik joan daitekeen adierazten duen ezaugarria, metakategoria mota adierazten duen ezaugarria, eta aditza eta osagaien arteko mendekotasun mota adierazten duen ezaugarria dira. Azken ezaugarri hori kanpoan utzi beharra izan da aurkitu dugun hutsunerik handienetakoa. Dena den, UD gidalerroek hizkuntza ba-



6.10 irudia – Jatorrizko zuhaitz-bankuko esaldi bateko ezaugarrien bihurketa adibidea.

koitzaren berezitasunak islatzen dituzten ezaugarriak uzteko aukera ematen dutenez, etorkizunerako utzi dugu euskarazko jatorrizko ezaugarri horiekin datu bereziak definitzea eta UD zuhaitz-bankura gehitzea.

Bukatzeko, jatorrizko ezaugarrien bihurketa prozesuaren ulermenerako, dependentzia zuhaitzak erabili beharrean, esaldi bateko ezaugarri guztien bihurketa biltzen duen adibide bat irudikatzea pentsatu dugu. 6.10 irudia aztertuta, irakurleak, emandako azalpenak esaldian aurki daitezkeen hitzen ezaugarrien bihurketa adibideen laguntzarekin argiago izatea espero da.

6.3.4 Dependentzia erlazioak

Etiketen bihurketekin bukatzeko, dependentzia erlazioei dagozkienak azaltzea soilik geratzen zaigu. Azalpena hiru ataletan banatuko dugu: 6.7 taulan, modu zuzenean bihurtu diren erlazioak bildu ditugu, 6.8 taulan, baldintza sinpleren baten menpe dauden erlazioak, eta 6.3.4.1, 6.3.4.2, 6.3.4.3 eta 6.3.4.4 ataletan, bihurketa konplexuagoa eskatzen duten erlazioak edo fenomenoak. Azken horietan jatorrizko zuhaitzaren egitura moldatu behar da.

Jatorrizkoa	**UD**	**Adibidea**
Aditz laguntzailea	aux	*Ezagutu ZUEN*
Mendeko perpaus osagarri jokatua, objektua	ccomp	*EGINGO duela uste dut*
Mendeko perpaus osagarri jokatua, subjektua	csubj	*litekeena da akordioa IZATEA*
Graduatzailea	advmod	kaleetan HAIN txukun
Aditzaren indartzailea	aux	*galdu IZAN zituzten*
Objektua (ez-perpausa)	dobj	*BRONTZEA lortu zuen*
Subjektua (ez-perpausa)	nsubj	*Gimnastika erritmikoko NESKA arabarrek*
Zehar-objektua (ez-perpausa)	iobj	*HORRI emandako erantzuna*
Mendeko perpaus osagarri ez jokatua, objektua	xcomp	*akordioa LORTU nahi badu*
Mendeko perpaus osagarri ez jokatua, subjektua	csubj	*Ematen du Egibarrek PSOEri JAKINARAZI diola*
Mendeko perpaus osagarri ez jokatua, zehar-objektua	advcl	*ezin da lanaldiaren murrizketa orokor bat PROPOSATU*
Aditz nagusia	root	*Etxera ETORRI zen*
Entitate-osagaia	name	*GIBRALTARKO zelaia*
Loturazko elementua	conj	*Ez ote dira senar eta EMAZTE mundu beraren bi parte?*

6.7 taula – UD formatura bihurtzean jatorrizko dependentziei zuzeneko bihurketaren bidez esleitu zaizkien UD dependentziak eta adibideak.

Jatorrizkoa	**Baldintza**	**UD**	**Adibidea**
Determinatzailea	Numerala	nummod	*Ur pixka BAT*
	Bestela	det	*Ur PIXKA bat*
Mendeko perpaus ez jokatua; adizlaguna edo izenlaguna	Erlatibozkoa	acl	*DATORREN igandean*
	Bestela	advcl	*baimen berezi bat ESKATZEKO*
Aposizioan dagoen mendeko perpaus jokatua	Izena	appos	*121. kilometroan (URBASAN)*
	Bestela	parataxis	*urduri samar (bost aldiz TRABATU zen)*
Aposizioa (ez-perpausa)	Izena	appos	*Busturian (BIZKAIA)*
	Bestela	parataxis	*Irujok irabazi zuen bere kontra (**22-18**)*

6.8 taula – UD formatura bihurtzean jatorrizko dependentziek betetzen duten baldintzaren arabera esleitu zaizkien UD dependentziak eta adibideak.

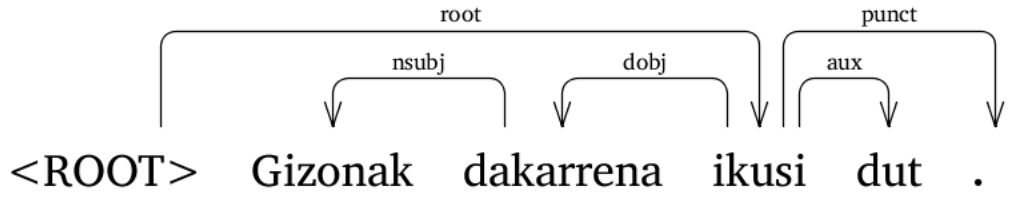
6.3.4.1 Elipsia

Euskarak, mendeko erlatibozko sintagmen bidez edo genitiboen bidez hitz-formaren barruan elipsia sortzea ahalbidetzen du. Har dezagun adibide bezala ondorengo hitza:

$$dakarrena = dakarren + \text{-}a$$

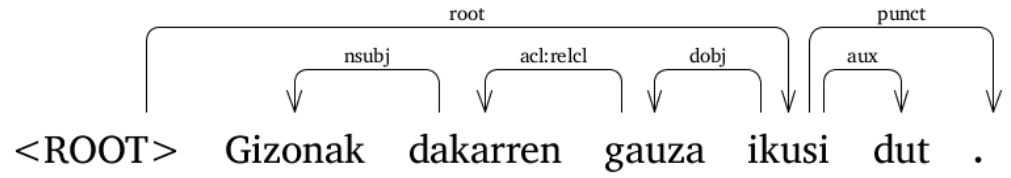
Hitz hori erlatibozko sintagma baten adibidea da, non artikulu mugatu batekin konbinatzen denean elipsia sortzen den. Hitz-forma bakoitzak kategoria

bakarra izan dezakeenez, kasu honetan erabaki egin behar da hitzaren erroa kontuan hartuta aditza izango den, ala hitz-forma osoak esaldian betetzen duen objektu funtzioa kontuan hartuta izena izango den. Are gehiago, *dakarrena* hitzak, izenek dauzkaten ezaugarri morfosintaktikoak ditu, esaterako kasua.



6.11 irudia – Izenaren barruan elipsia duen erlatibozko sintagmaren adibidea.

6.11 irudiak aipatutako fenomenoa irudikatzen du. Irudiaren arabera, aipatutako hitza aditz nagusira lotuta dago objektu zuzena (dobj) erlazioaren bidez. Gertaera hori kontraesankorra dela esan daiteke, hitza aditz bezala etiketatuta dagoelako. 6.12 irudian aurreko irudian agertzen den esaldiaren parekoa den esaldia irudikatzen da, baina elipsirik gabe. Adibide horretan, *Gizonak* hitz-formak mendeko esaldiaren subjektu funtzioa hartzen du, *dakarrena* hitza *gauza* hitzarekin erlazionatuta dago, eta azken honek objektu funtzioa betetzen du.

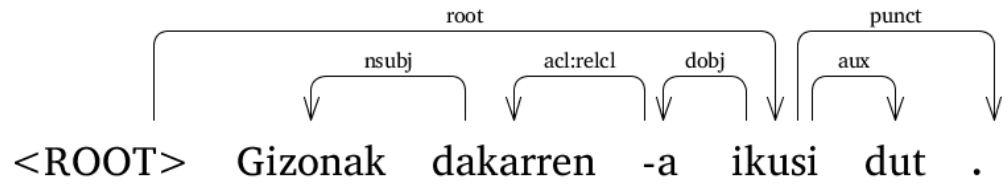


6.12 irudia – 6.11 irudian azaldutakoaren parekoa den elipsirik gabeko sintagmaren adibidea.

Aurretik esan bezala, Dependentzia Unibertsalak etiketatzea sintaxiaren hurbilpen lexikalistan oinarritzen da, dependentzia erlazioak hitzen artean eman behar direlarik. Hala ere, badira gaztelerazko klitikoak bezalako salbuespenak. Turkieran ere saiatu dira antzerako ideia martxan jartzen multzo malgukarietan, baina ez dira adostasunera iritsi hitzak moztearen inguruan.

6.13 irudian 6.11 irudian aurkeztu den arazoaren konponbide posible bat aurkezten da, bertan *dakarrena* hitz-forma osatzen duten aditzaren eta izenaren informazioa banatzen delarik. Modu honetara, irudian aurkezten den analisia eta 6.12 irudian aurkeztu dena simetrikoak dira, eta 6.11 irudian aurkeztu den aditz/izen dikotomia ebazten da. Hitz-forma banatzea automatikoki egin daiteke, elipsiak ez baitu interpretazio asko izateko aukerarik ahalbidetzen.

Dena den, argi utzi nahi da 6.13 irudian aurkeztutakoa proposamen posible bat besterik ez dela eta oraindik ez dela aplikatu euskarazko UD zuhaitz-bankuan, batez ere UD gidalerroetan gomendatzen denaren aurka doalako. Hori dela eta, euskarazko elipsia ez da tratatu eta 6.11 irudian aurkeztutakoa aplikatu zaio euskarazko UD zuhaitz-bankuari.



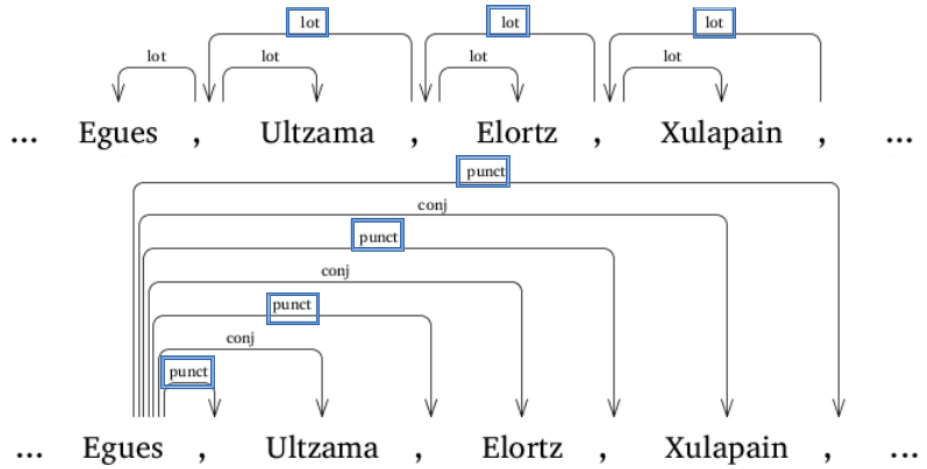
6.13 irudia – 6.11 irudian aurkeztutakoaren analisi alternatiboa.

6.3.4.2 Puntuazioa

Jatorrizko zuhaitz-bankuan normalean puntuazioa adierazteko erabiltzen den dependentzia erlazioak UD ordain zuzena (*punct*) duen arren, moldaketa behar duten erlazioen multzoan sartu dugu, zuhaitzaren egitura aldatzea eskatzen duen erlazioa baita.

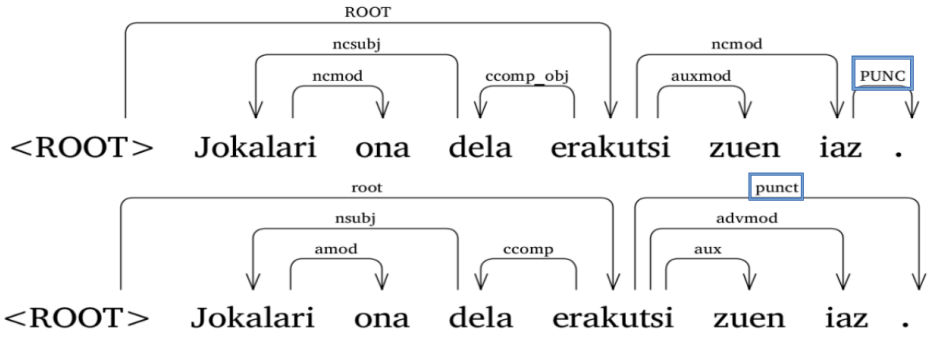
Jatorrizko zuhaitz-bankuan, puntuazio ikurrak normalean beraien aurreko hitzera lotuta doaz, baina UD gidalerroen arabera, UD formatuko zuhaitz-bankuan puntuazio ikurrak eduki hitzetara (*content words*) doaz lotuta. Hori dela eta, puntuazioaren egituraketa berri bat egin behar da bihurketa modu egokian egiteko. Ataza hori aurrera eramateko, kasu bakoitzean puntuazio marka zein elementuri lotu behar zaion jakiteko ondorengo irizpideak jarraitu dira:

1) Koordinazioaren atalak banatzen dituen puntuazio marka, lehenengo elementuari lotu behar zaio.



6.14 irudia – Puntuazioa zein elementuri lotu behar zaion erabakitzeko erabili dugun lehen irizpidearen adibidea. Goian, jatorrizko zuhaitz-bankuko esaldi bat. Behean, irizpidea aplikatu ondoren, esaldi berdina UD formatura bihurtuta.

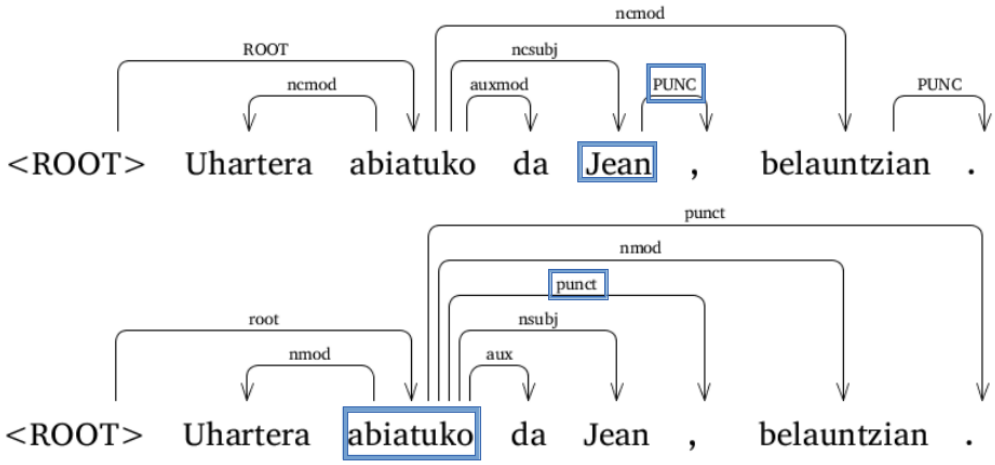
2) Mendeko atalaren aurretik edo atzetik doan puntuazio marka, perpausaren buruari lotu behar zaio.



6.15 irudia – Puntuazioa zein elementuri lotu behar zaion erabakitzeko erabili dugun bigarren irizpidearen adibidea. Goian, jatorrizko zuhaitz-bankuko esaldi bat. Behean, irizpidea aplikatu ondoren, esaldi berdina UD formatura bihurtuta.

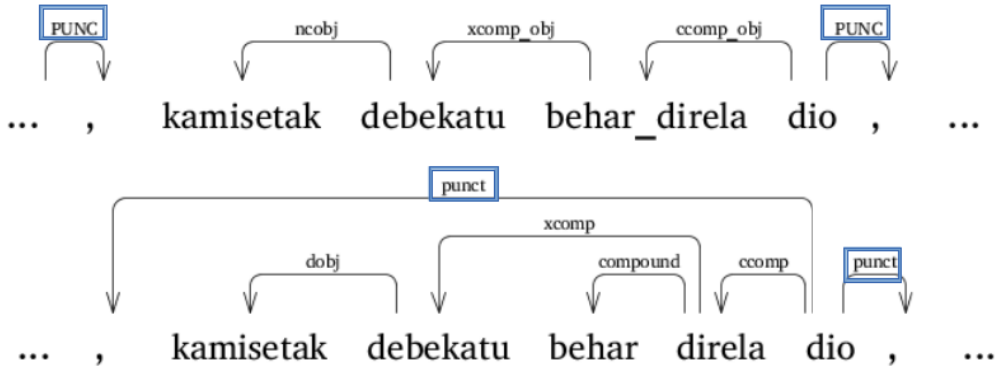
3) Atal baten barruko puntuazio marka, atal horren barruan proiektibita-

tea mantentzen duen gorengo nodoari lotu behar zaio.



6.16 irudia – Puntuazioa zein elementuri lotu behar zaion erabakitzeko erabili dugun hirugarren irizpidearen adibidea. Goian, jatorrizko zuhaitzbankuko esaldi bat. Behean, irizpidea aplikatu ondoren, esaldi berdina UD formatura bihurtuta.

4) Binaka datozen puntuazio markak (parentesiak, komatxoak, kortxeteak, komak...), proiektibitatea mantentzen den bitartean, hitz berdinera lotuta egon behar dira. Hitz hori, normalean, binakako puntuazio marken barruan aurkitzen den atalaren burua izaten da.



6.17 irudia – Puntuazioa zein elementuri lotu behar zaion erabakitzeko erabili dugun laugarren irizpidearen adibidea. Goian, jatorrizko zuhaitz-bankuko esaldi bat. Behean, irizpidea aplikatu ondoren, esaldi berdina UD formatura bihurtuta.

6.3.4.3 Koordinazioa

Koordinazio egiturak etiketatzeko era asko daude. Izan ere, koordinazioaren burua izateko aukeratu den hitzaren arabera egitura aldatu egiten baita. Euskarazko jatorrizko zuhaitz-bankuan koordinazioaren burua juntagailua da (eta, edo, ala, baina...), baina UD zuhaitz-bankuan koordinazioaren lehen argumentua da egituraren buruaren papera hartzen duena. UD zuhaitz-bankuan, koordinaziorako loturak irizpide hauen arabera erabakitzen dira:

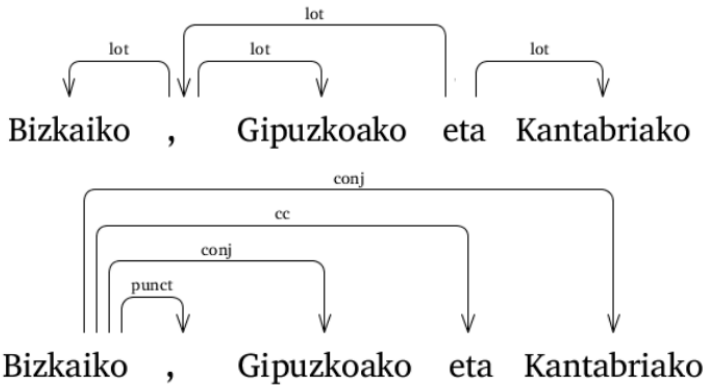
1) Sortuko den erlazio berriaren burua, koordinazioaren lehenengo atalaren burua izango da, eta gainontzeko atalak horri lotuta joango dira *conj* UD dependentzia erlazioaren bidez.

2) Koordinaziorako erabiltzen diren hitzak lehenengo atalaren buruarekin lotuko dira *cc* (coordinating conjunction) UD dependentzia erlazioaren bidez.

Koordinazioa etiketatzeko era horrek, alegia, elementu guztiak koordinazioaren lehen argumentuari lotzeak, jatorrizko zuhaitz-bankuan inplizitu dagoen informazioaren galera ekar dezake. Esaterako, jatorrizko zuhaitz-bankuan posible da modifikatzaile bat koordinazio egituraren lehen argu-

mentuaren umea edo koordinazio egitura guztiaren umea izatea. UD zuhaitz-bankuan, ostera, aipatutakoa ezinezkoa da.

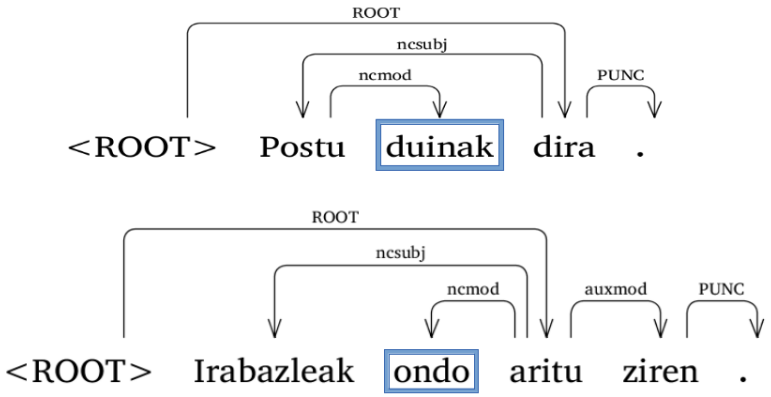
Koordinazioa nola bihurtu dugun hobeto azaltzeko, 6.18 irudian esaldi berdinaren jatorrizko egitura eta bihurtutako egitura irudikatu dira (jatorrizkoa goian eta bihurtutakoa behean), modu honetara egiturak konparatu ahal izango ditugu eta hobeto ulertu bihurketa prozesua.



6.18 irudia – Jatorrizko zuhaitz-bankuan eta UD zuhaitz-bankuan koordinazioa irudikatzeko era desberdinen adibidea.

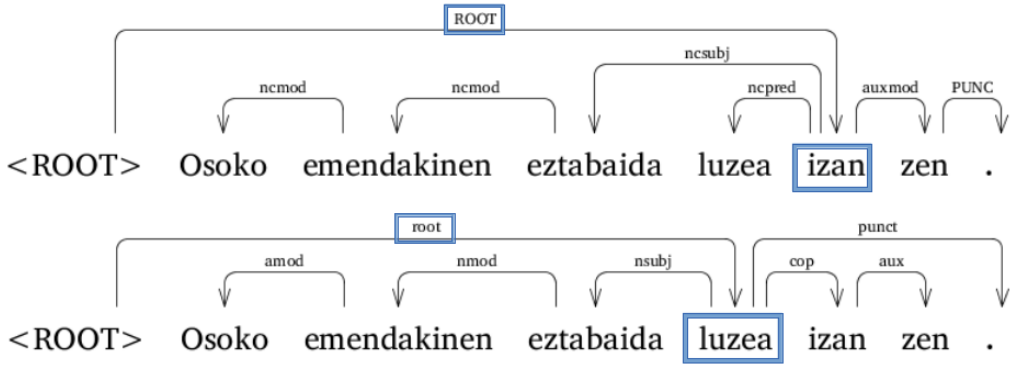
6.3.4.4 Esaldi kopulatiboak

UD gidalerroen arabera esaldi kopulatiboak sortzeko era bakarra kopula bezala *izan* aditza erabiltzea den arren, euskaraz esaldi kopulatiboetan aditz gehiagok parte har dezakete. Are gehiago, esaldi kopulatiboen analisia guztiz desberdina da, aditza ez baita burua. Horrenbestez, esaldi mota horiek azterketa sakonagoa behar dute. Orokorrean komunztadura egoten da kopulazio modifikatzailearen eta subjektuaren artean (6.19 irudiko goiko adibidea), baina aditz predikatiboetan modifikatzailea aditzondoa izaten da eta kasu horretan ez da egoten komunztadurarik (6.19 irudiko beheko adibidea).



6.19 irudia – Goiko adibidean, esaldi kopulatiboa, komunztadura dago subjektuaren eta adjektiboaren artean (urdinez). Beheko adibidean, esaldi predikatiboa, aldiz, ez dago komunztadurarik subjektuaren eta aditzondoaren artean (urdinez).

Esaldi kopulatiboak nola bihurtu diren hobeto azaltzeko azter dezagun 6.20 irudia. Goiko esaldian, jatorrizko esaldian, esaldiaren buruaren papera *izan* aditzak hartzen du. Beheko esaldian, UD formatura bihurtutako esaldian, aurretik aipatu dugun bezala, esaldiaren buruaren papera ez du aditzak hartzen, *luzea* hitzak baizik, eta *izan* aditzak kopularen papera betetzen du.



6.20 irudia – Jatorrizko zuhaitz-bankuko esaldi kopulatibo baten bihurketa adibidea.

6.4 Ondorioak

Kapitulu honetan, euskarazko zuhaitz-bankua Dependentzia Unibertsalak proiektuaren parte izateko bihurketarako jarraitu ditugun urratsak azaldu ditugu. Bihurketa prozesu guztia bukatu ondoren, jatorrizko zuhaitz-bankuak dituen 150.000 hitzetik 121.000 hitz bihurtzea lortu dugu. Modu horretara, euskara, nazio arteko proiektu garrantzitsu horren partaide izatea lortu dugu.

Jarraitu dugun prozesua konplexua eta luzea izan da, eta oraindik ez dugu erdietsi fenomeno linguistiko eta hitz guztiak behar bezala UD formatura pasatzea. Adibidez, hitz anitzeko esapide askoren bihurketa ez da aurrera eraman, ezin izan ditugulako ziurtasunez banatu bere analisi propioa duen hitzetan. Horrenbestez, hitz anitzeko esapide horiek zeuden esaldiak baztertu egin behar izan dira. Tratatu gabe gelditu den beste fenomeno bat elipsia izan da. Oso ohikoa da euskarazko esaldi batean hitz-formaren barruan elipsia aurkitzea eta horrek asko konplikatzen du hitz-forma horren bihurketa zuzena. Izan ere, UD formatuan egokiena esaldi bateko elementu guztiak ikusgai egotea baita. Kasu horretan, elipsia duten hitzak hitz arrunt bezala tratatu ditugu eta gainontzeko hitzak bezala bihurtu dira.

Esandakoak esanda, euskarazko UD zuhaitz-bankuaren kalitatea ona dela uste dugun arren, badakigu hainbeste mila hitz automatikoki bihurtzen diren prozesuan erroreak egon daitezkeela. Ondorioz, gure etorkizunerako lanen artean, lortutako emaitza ondo aztertzeak garrantzi handia izango du. Bestalde, elipsiaren arazoa nola konpondu behar den ere aztertu behar dugu.

Tratatu gabe geratu diren hitz-anitzekoen azterketa sakona ere egin behar da, zuhaitz-bankuaren tamaina handitzeko erarik errazena tratatu gabe geratu diren hitz-anitzeko terminoak modu egokian bihurtzea baita, era horretara ez direlako hainbeste esaldi baztertuko.

Bukatzeko, sortu den zuhaitz-bankuarekin esperimentu desberdinak egitea interesgarria dela iruditzen zaigu. UD proiektuko beste hizkuntzetako zuhaitz-bankuak eskuragarri izanda hainbat proba egin daitezke hizkuntza desberdinen artean. Esaterako, sistema hizkuntza batekin entrenatu daiteke (hitz-formak eta lemak kenduta) eta antzekoa den beste baten gainean aplikatu ikasitakoa. Modu honetara, hizkuntzen arteko antzekotasun sintaktikoa neurtu daiteke. Bestalde, hitz-formen eta lemen informazioa ez galtzeko itzultzaile automatikoak erabili daitezke hitzak, eta bide batez zuhaitz-bankua, hizkuntza batetik bestera itzultzeko. Gauzak horrela, gidalerro berdinak jarraituta bihurtutako zuhaitz-bankuak izango genituzke hizkuntza berean,

honek zabaltzen dituen ikerketa-lerro berri guztiekin: hizkuntzen arteko multzokatzea (*clustering*), hitz-bektoreak (*word vectors, embeddings*)...

Ikerketa lerro berriak zabaltzeko bidearen hasiera bezala jatorrizko euskarazko zuhaitz-bankuaren eta bihurtutako zuhaitz-bankuaren oinarrizko emaitzak konparatzea pentsatu dugu. Modu horretara, bihurketan sintaxirako garrantzitsuak diren ezaugarriak galdu diren ala ez aztertuko dugu. Jatorrizko zuhaitz-bankuarekin testaren gainean lortutako oinarrizko emaitza % 83koa da (LAS) eta bihurtutakoarekin lortutakoa % 78,50ekoa. Emaitza horiek guztiz konparagarriak ez diren arren (ikasketarako eta ebaluaziorako zuhaitz-bankuak ez dira berdinak), bi zuhaitz-bankuek antzeko oinarrizko emaitzak dituztela ondorioztatu daiteke. Izan ere, bihurtutako zuhaitz-bankuaren ikasketarako zuhaitz-bankua txikiagoa baita, eta tamaina berdineko ikasketa corpusak erabilita erdietsitako emaitzetan tarte txikiagoa izatea espero da. Ondorioz, bihurketan zerbait galdu den arren, jatorrizko zuhaitz-bankuaren ezaugarri garrantzitsuenak mantendu direla esan daiteke.

7 Ondorioak eta etorkizuneko lanak

Lehenengo kapituluaren sarreran aipatu dugun bezala, tesi-lan hau sintaxian kokatzen da, zehazki, sintaxi konputazionalean. Sintaxi konputazionalean, esaldien egitura sintaktikoak zehazten dira ordenagailuen bidez. Jakina da sintaxia modu horretan lantzeak bere abantailak eta desabantailak dituela. Abantaila nagusia azkartasuna da, milaka esaldi aztertu baitaitezke minutuetan. Desabantaila nagusia, ostera, lortutako analisi sintaktikoaren kalitatea eskuz aztertutakoarena baino eskasagoa dela da. Hori dela eta, tesi-lan honen helburu nagusia analisi sintaktiko automatikoa hobetzen lagunduko duten teknikak aztertzea da, batez ere euskararen analisi sintaktikoa hobetzeko, baina ikuspegi eleaniztuna ahaztu gabe. Ahal izan den neurrian, esperimentu guztiak hizkuntza desberdinetan probatzen saiatu gara, baina hori ezinezkoa gertatu zaigun kasuetan euskaran zentratu gara, hizkuntza hori baita gure lehentasuna. Ondorioz, euskararako baliagarriak izan daitezkeen baliabideak sortzea ere gure helburuen artean sartu dugu.

Hasieran planteatu diren helburuak betetzeko hainbat eginkizun eraman ditugu aurrera, tesi-lanaren sarreran zehaztutako eginkizun horietako bakoitza atal honetan ekarpen bezala definitu delarik.

7.1 Ekarpenak

Atal honetan tesi-lan honen ekarpen nagusiak zerrendatzen ditugu, ekarpen bakoitzari lotutako ikerkuntza galdera eta erantzunekin batera.

Izaera desberdinetako analizatzaileen hibridazioarekin lortu daitezkeen emaitzak neurtu dira (*4. kapitulua*)

Ondorengo hibridazio motak aztertu ditugu:

- Estatistiketan oinarritutako analizatzaileen eta erregeletan oinarritutako analizatzaileen arteko hibridazioa.

- Estatistiketan oinarritutako analizatzaileen eta erregeletan oinarritutako azaleko analizatzaileen arteko hibridazioa.

- Estatistiketan oinarritutako analizatzaileen, erregeletan oinarritutako analizatzaileen eta erregeletan oinarritutako azaleko analizatzaileen arteko hibridazioa.

- Estatistiketan oinarritutako analizatzaileen eta estatistiketan oinarritutako azaleko analizatzaileen arteko hibridazioa.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

- Erregeletan oinarritutako analizatzaileak analizatzaile estatistikoekin konbinatzen direnean sintaxi osoa lantzen duten erregeletan oinarritutako analizatzaileen ekarpena azaleko sintaxia lantzen dutenena baino handiagoa izan al da? Bai, bien arteko aldea handia ez den arren, esan daiteke sintaxi osoa lantzen duten erregeletan oinarritutako analizatzaileak analizatzaile estatistikoekin konbinatzen direnean hobekuntza handiagoak lortzen direla.

- Azaleko sintaxia lantzen duten analizatzaileen artean erregeletan oinarritutakoen ekarpena estatistiketan oinarritutakoena baino handiagoa izan al da? Konbinaketarako erabili den parserraren arabera; Mate erabiltzen denean estatistiketan oinarritutakoak gehiago laguntzen du, baina MaltOptimizer erabiltzen denean kontrakoa gertatzen da.

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarri morfologiko mota bakoitzak analisi sintaktikoan duen eragina neurtu da (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

- Euskara, frantsesa, alemana, hungariera eta suedierarako ezaugarri morfologikoek dependentzien analisi sintaktikoan duten eragina aztertu da. Horretarako, analizatzaile sintaktikoari ezaugarriak banaka pasatu zaizkio eta erdietsitako emaitzak beste ezaugarriekin erdietsitako emaitzekin konparatu dira. Emaitzarik hoberenak lortu dituzten ezaugarriak analisian pisu handiago dutela ondorioztatu dugu.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

- Morfologia aberatsagoa duten hizkuntzetan, neurtutako ezaugarrien ekarpena handiagoa izan al da? Euskararen eta hungarieraren kasuan esan daiteke ezaugarri batzuek eragin handia dutela analisi sintaktikoan. Frantsesaren, alemanaren eta suedieraren kasuan, aldiz, erabilitako ezaugarriek ez dute eragin handirik. Alemana eta suediera morfologikoki hain aberatsak izanda esan daiteke neurtutako ezaugarrien ekarpena ez dagoela hizkuntzaren aberastasunari hertsiki lotuta. Izan ere, beste faktore batzuek ere eragina izan baitezakete. Esaterako, kategoria etiketatzaile automatikoaren kalitateak.

- Ondorioztatu al daiteke ezaugarri mota batzuek analisi sintaktikoan duten pisua beste batzuena baino handiagoa dela orokorrean? Bai, ikusitako emaitzek kasua oso garrantzitsua dela iradokitzen dute. Gainontzeko ezaugarrietan ez dago adostasun osorik hizkuntza guztien artean.

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarrien ingeniaritza aplikatzeak analisi sintaktikoan duen eragina neurtu da (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

- Euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako hiru ezaugarri morfologiko hoberenak erabili dira analisi sintaktikoan.

- Hitzaren kategoriaren eta azpikategoriaren ordez ezaugarri onena erabili da analisi sintaktikoan.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

– Analisi sintaktikoan pisu gehien duten hiru ezaugarriak erabiltzean, erdietsi al da emaitzarik ezaugarri guztiak erabiltzean lortzen direnarekin konparagarria denik? Bai, hainbat kasutan eta hizkuntza guztietan ikusi da hiru ezaugarririk esanguratsuenak erabiltzean ezaugarri guztiak erabiltzean bezain emaitza onak lortzen direla, kasu batzuetan emaitza horiek gaindituz.

– Parserrari ezaugarri morfologikoak pasatzeko moduak eraginik izan al du analisi sintaktikoan? Euskararen eta suedieraren kasuan bai. Hitzaren kategoriaren edo azpikategoriaren ordez ezaugarri hoberena erabilita oinarrizko emaitzak hobetzea lortu da.

Morfologikoki aberatsak diren zenbait hizkuntzatan ezaugarrien ingeniaritza aplikatzean erdietsitako analisi desberdinen konbinazioak duen eragina neurtu da (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

– Ezaugarrien ingeniaritzako teknikekin euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako lortu diren analisi guztiak bozketaren bidez konbinatu dira.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

– Ondorioztatu al daiteke hizkuntza batek zenbat eta morfologia aberatsagoa izan orduan eta emaitza hobeak lortzen dituela ezaugarrien ingeniaritza aplikatzean erdietsitako analisi desberdinen konbinazioaren bidez? Alemana morfologikoki frantsesa baino aberatsagoa izanda oso antzeko hobekuntza lortu dute biek konbinazioaren bidez. Alemanaren oinarrizko emaitzak oso onak direnez hobekuntzarako tarte txikia geratzeagatik dela uste dugu. Hori dela eta, ez dugu esango konbinazioaren emaitza hizkuntzaren aberastasunari bakarrik lotuta dagoela. Hala ere, euskaran, hungarieran eta suedieran ikusten den bezala, hizkuntzaren aberastasunak badu bere eragina emaitzan.

Morfologikoki aberatsak diren zenbait hizkuntzatan multzokatze mota desberdinek analisi sintaktikoan duten eragina neurtu da (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

- Euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako Brown multzokatzearen eragina neurtu da analisi sintaktikoan.
- Hitz-bektoreetan oinarritutako K-means multzokatzearen eragina neurtu da analisi sintaktikoan.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

- Ondorioztatu al daiteke multzokatze mota bat bestea baino lagungarriagoa dela analisi sintaktikorako? Bai, egindako esperimentuen emaitzen arabera, Brown multzokatzea erabiltzea eraginkorragoa da.

Morfologikoki aberatsak diren zenbait hizkuntzatan multzokatze mota desberdinak aplikatzean erdietsitako analisien konbinazioak duen eragina neurtu (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

- Multzokatze mota desberdinekin euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako lortu diren analisi guztiak bozketaren bidez konbinatu dira.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

- Ondorioztatu al daiteke zenbat eta morfologia aberatsagoa izan hizkuntzak orduan eta emaitza hobeak lortzen dituela multzokatze mota desberdinak aplikatzean erdietsitako analisien konbinazioaren bidez? Beste hizkuntzetarako emaitzek hobekuntza antzekoak erakutsiko dituztela baieztatzea ezinezkoa den arren, guk probatutako hizkuntzetan hizkuntzaren aberastasun morfologikoak erdietsitako emaitzarekin lotura zuzena duela ikusten da. Hizkuntza horietan urritasunaren arazoa (*sparseness*) larriagoa da eta erdi-gainbegiratutako teknikek gehiago lagundu dezakete.

Morfologikoki aberatsak diren zenbait hizkuntzatan meta-ezaugarriek duten eragina neurtu da (*5. kapitulua*)

Euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako ondorengo ekintzak burutu dira:

- Hitz-formen arteko lotura sintaktikoen (bilaketa patroi) maiztasunetatik eratorritako meta-ezaugarrien eragina neurtu da.

– Brown multzoen arteko lotura sintaktikoen (bilaketa patroi) maiztasunetatik eratorritako meta-ezaugarrien eragina neurtu da.

– Hitz-bektoreetan oinarritutako K-means multzoen arteko lotura sintaktikoen (bilaketa patroi) maiztasunetatik eratorritako meta-ezaugarrien eragina neurtu da.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

– Ondorioztatu al daiteke meta-ezaugarriak sortzeko iturri batzuk besteak baino hobeak direla? Bai, Brown multzoen eta hitz-formen artean zalantzak egon daitezkeen arren, K-means multzoetatik eratorritako meta-ezaugarriekin beste biekin baino emaitza okerragoak erdietsi dira.

– Ondorioztatu al daiteke zenbat eta morfologia aberatsagoa izan hizkuntza batek orduan eta emaitza hobeak lortzen direla meta-ezaugarriak erabiltzen direnean? Aurreko kasuetan mota honetako galderak erantzun ditugunean bezala, hizkuntzaren aberastasunaren eta emaitzen artean lotura zuzena ez dagoen arren, hungarieran eta suedieran orokorrean emaitza positiboak erdietsi diren bitartean, frantsesean emaitza negatiboak jaso dira. Euskaran eta alemanean tarteko emaitzak lortu dira. Ondorioz, esan daiteke badagoela aberastasunaren eta emaitzen arteko loturaren bat.

Ezaugarrien ingeniaritzarekin, multzokatze mota desberdinekin eta erdi-gainbegiratutako zuhaitz-bankuetatik erauzitako ezaugarriekin erdietsitako analisien konbinazioak morfologikoki aberatsak diren zenbait hizkuntzatan duen eragina neurtu da (*5. kapitulua*)

Ondorengo ekintzak burutu dira:

– Euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako ezaugarrien ingeniaritzarekin, multzokatze mota desberdinekin eta meta-ezaugarriekin erdietsitako analisiak bozketaren bidez konbinatu dira.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

– Konbinazioan erabilitako iturri desberdinetatik eratorritako analisiak osagarriak izan al dira? Kasu gehienetan ezaugarrien ingeniaritzarekin lortutako analisien konbinazioan erdietsitako emaitza ezin izan da gainditu. Hala ere, konbinazioan hiru bilaketa patroi hoberenetatik eratorritako meta-ezaugarrien batura erabili dugunean frantsesean ezik beste hizkuntza guztietan hobetu dira emaitzak. Kontuan hartuta meta-ezaugarriek frantsesean erakutsi duten eraginkortasun eza, ez da harritzekoa horiek konbinazioan erabiltzean emaitzak ez hobetzea. Ondorioz, esan daiteke hiru bilaketa patroi hoberenetatik eratorritako meta-ezaugarrien batura eta ezaugarrien ingeniaritzan sortu ditugun ezaugarriak osagarriak direla kasu gehienetan.

Dependentzia Unibertsalak proiektuan zehazten diren gidalerroak jarraituz euskarazko zuhaitz-bankua bihurtu da (*6. kapitulua*)

Ondorengo ekintzak burutu dira:

– Hitz anitzeko esapide mota desberdinak banatu dira kasuan kasuko irizpideak aplikatuz.

– Puntuazioa berregituratu da.

– Koordinazioa berregituratu da.

– Esaldi kopulatiboak bihurtzeko soluzio bat planteatu dugu.

– Kategoriak bihurtu dira.

– Ezaugarri morfologikoak bihurtu dira.

– Dependentziak bihurtu dira.

Ekarpen honekin lotutako ikerkuntza galderak eta erantzunak:

– Zehaztutako gidalerroen arabera euskarazko zuhaitz-banku osoa bihurtu ahal izan da? Ez, hitz anitzeko esapide guztiak ezin izan dira banatu, ezta fenomeno guztiak ere. Hala ere, zuhaitz-bankuaren % 81a bihurtzea lortu dugu.

– Bihurtutako zuhaitz-bankuarekin oinarrizkoarekin erdietsitakoak baino emaitza hobeak lortu al dira? Bihurtutako zuhaitz-bankuan emaitza baxuagoak lortu dira. Dena den, bihurtutako zuhaitz-bankuaren ikasketarako zuhaitz-bankua txikiagoa da. Ondorioz, emaitza konparagarriak erdiesten direla esan daiteke.

7.2 Ondorioak

Orain aurkezten ditugun ondorioak aurreko atalean aurkeztutako ekarpenetatik eratorriak dira.

Hibridazioa (*4. kapitulua*)

Kapitulu honetan izaera desberdineko analizatzaileak konbinatu dira euskararen analisi sintaktikoan duten eragina zehazteko. Sintaxi osoa lantzen duten analizatzaile estatistikoei erregeletan oinarritutako azaleko analizatzaileen emaitza eta estatistiketan oinarritutako azaleko analizatzaileen emaitza gehitu diegu. Ikuspegi hibridoa euskararen analisi sintaktikoa hobetzeko jorratu dugun arren, egindako esperimentuetatik jasotako ezagutza euskararekin konparagarriak diren hizkuntza morfologikoki aberatsetan aplikagarria izatea izan da gure helburuetako bat, hizkuntza horietan ere guk erdietsitakoetatik oso urrun ez dauden emaitzak jasotzea espero delarik.

Esan bezala, hibridaziorako bi hurbilpen izan ditugu ardatz: erregeletan eta estatistiketan oinarritutako analizatzaileen hibridazioa, eta estatistiketan oinarritutako analizatzaileen hibridazioa. Erregeletan eta estatistiketan oinarritutako analizatzaileak konbinatu ditugunean ez dira lortu espero ziren emaitzak, batez ere ikasketarako corpus handiagoa erabili den kasuetan. Kasu horietan, grafoetan oinarritutako analizatzailearekin edo trantsizioetan oinarritutako analizatzailearekin oso zaila gertatu zaigu oinarrizko emaitzak gainditzea. Bestalde, ikasketarako corpus txikiagoa erabili dugunean oinarrizko emaitzak gainditu dira hibridazio mota guztietan. Gertaera hori kontuan izanda, hibridazioak ikasketarako corpus txikia dagoen kasuetan lagundu dezakeela uste dugu.

Hibridazioak euskararen analisi sintaktiko orokorrean duen eragina neurtzeaz gain, hibridazioak galderazko esaldietan duen eragina ere neurtu dugu. Helburu hori lortzeko MaltOptimizer eta Mate parserrak EDGK analizatzailearekin konbinatu dira bakoitza bere aldetik. Modu horretara erdietsitako emaitzak itxaropentsuak izan dira, bi hibridazio esperimentuetan gainditu baitira analizatzaile estatistikoen oinarrizko emaitzak. Gainera, dependentzia konkretuetan lortutako emaitzek hobekuntza nabarmenak erakusten dituzte.

Hibridaziorako erabili dugun bigarren hurbilpena estatistiketan oinarritutako analizatzaileen konbinaketa izan da. Kasu honetan, MaltOptimizer eta Mate parserrak MLIxati chunkerrarekin konbinatu dira bakoitza bere

aldetik. Erdietsitako emaitzak positiboak edo negatiboak dira erabilitako parserraren arabera. MaltOptimizer erabiltzen bada, emaitzek (orokorrak edo dependentzia zehatzetakoak) behera egiten dute. Mate erabiltzen bada, aldiz, hobekuntza txikia lortzen da emaitza orokorrean. Hala ere, Mate erabiltzean erdietsitako emaitzen artean garrantzitsuena ez da emaitza orokorrean lortutako hobekuntza, *ncmod* dependentzia erlazioan lortutakoa baizik. Dependentzia horretan 0,75 puntuko hobekuntza erdietsi da eta euskarazko esaldietan gehien agertzen den dependentzia izanik, ekarpen hori oso baliagarria izan daiteke.

Analisi sintaktiko eleaniztuna: ezaugarrien ingeniaritza (*5. kapitulua*)

Euskararako, frantseserako, alemanerako, hungarierarako eta suedierarako ezaugarri morfologikoek analisi sintaktikoan duten eragina zehaztu dugu. Horretarako, ezaugarri guztiak probatu dira banan-banan gainontzeko ezaugarrien eragina saihesteko eta neurketa ahalik eta fidagarriena izateko. Erdietsitako emaitzen arabera, orokorrean ezaugarririk esanguratsuena kasua dela esan genezake, oso garrantzi handia duelarik euskaran eta hungarieran. Ezaugarri bakoitzaren pisua jakinda hiru esperimentu mota eraman dira aurrera: hizkuntza bakoitzerako hiru ezaugarri morfologiko esanguratsuenak erabiltzea, ezaugarririk esanguratsuena kategoriaren ordez jartzea eta ezaugarririk esanguratsuena azpikategoriaren ordez jartzea.

Aipatutako esperimentuekin zenbait hobekuntza lortu ditugu, batez ere euskaran, frantsesean eta suedieran. Hungarieran eta alemanean ez dugu lortu inolako hobekuntzarik ezaugarri esanguratsuenak bakarrik erabilita, ezta ezaugarri garrantzitsuena analizatzaileak haztapen altuagoa ematen dion lekuan jarrita. Alemanaren kasuan ezaugarri morfologikoek analisi sintaktikoan bakarka eraginik ez dutela ikusi da, beharbada hasieratik emaitzak onak zirelako, eta hori izan daiteke emaitza horien arrazoia, aplikatutako esperimentuak ezaugarri morfologikoetan oinarrituta baitaude. Hungarieraren kasuan, emaitza horiek jaso izanaren arrazoia hungarierako ezaugarri morfologiko gehienak beraien artean osagarriak direla izan daiteke. Horrenbestez, ezaugarri guztiak erabili beharrean hiru hoberenak erabiltzean informazioa galdu da. Bestalde, ezaugarririk hoberena kategoriaren eta azpikategoriaren ordez jarri dugunean analizatzaileak ez dio informazio horri haztapen altuagoa eman, ezaugarri morfologikoen zutabean jaso dituen ezaugarriek informazioa oso zabala eta osagarria eskaintzen dutelako.

Hiru ezaugarri morfologiko esanguratsuenak soilik erabiltzea bereziki interesgarria izan da. Kasu gehienetan oinarrizkoen pareko emaitzak lortu dira eta ikasketa prozesua azkartu da erabilitako ezaugarri kopurua txikiagoa baita. Bestalde, esanguratsuak ez diren ezaugarriak kentzeak analizatzaile sintaktikoari lagundu dio kasu batzuetan emaitzak hobetzen, ezaugarri batzuek eragin negatiboa baitute analisian. Hori kontuan hartuta, zuhaitz-bankuan hiru ezaugarri esanguratsuenak soilik gordetzeak zuhaitz-bankuaren tamaina murrizteko balio dezake, aldi berean emaitzak mantenduta.

Bukatzeko, esperimentuetan aplikatutako tekniken eragina asko nabaritzen da konbinaketa esperimentuak lantzen ditugunean, hizkuntza guztietarako hobekuntza sendoak erdiesten direlarik. Ondorioz, uste dugu aplikatutako teknikekin lortutako analisi bakoitzak baduela berezko ezaugarriren bat beste analisiekin osagarria dena eta analisi sintaktiko sendoago bat erdiesteko lagungarria dena.

Analisi sintaktiko eleaniztuna: multzokatzea (*5. kapitulua*)

Hizkuntzaren Prozesamenduko hainbat atazatan erabili dira etiketatu gabeko corpus handiak erabilgarriak diren ezaugarri berriak sortzeko. Hala ere, kasu gehienetan ezaugarri berri horiek corpuseko hitz-forma osoetatik eratortzen dira, hots, hitzak ez dira lema eta morfemetan banatzen. Banaketarik ez egitea nahiko eraginkorra dela ikusi den arren, baliteke hizkuntza batzuetarako erabilera hori egokiena ez izatea, adibidez hauen morfema kopuru altuaren ondorioz ez direlako nahi bezalako hitz multzoak lortzen. Hori dela eta, multzokatze algoritmo desberdinak eta multzoak egiteko modu desberdinak landu dira. Bi multzokatze teknika desberdin erabilita sortu diren ezaugarri berriak gehitu zaizkie lehendik zeudenei, analisi sintaktikoa hobetzeko asmoarekin. Ezaugarri berri horiek era ezberdinetan lortu dira; euskararako, alemanerako eta hungarierarako hitzak lema eta atzizkietan banatuta eta banatu gabe lortu dira multzoak. Frantseserako eta suedierarako, aldiz, ez da inolako banaketarik egin multzokatze algoritmoa aplikatu aurretik, hizkuntza hauetan banaketa egitea oso zaila edo ezinezkoa gertatu zaigulako eskuragarri genituen tresnekin.

Egindako esperimentuen emaitzen arabera Brown multzoak kalkulatzeko garaian hobe da hitza lema eta atzizkietan ez banatzea, modu honetara emaitza hobeak lortzen baitira. K-means multzokatzea erabiltzen denean, aldiz, hitzak banatzearen eta ez banatzearen artean ez dago alde handirik emaitzei dagokienez.

Erabilitako bi multzokatze algoritmoetan zentratzen bagara, esan daiteke dependentzien analisi sintaktikoan aplikatzeko hobe dela Brown multzoak erabiltzea K-means multzoak erabiltzea baino. Nahiz eta corpus txikiagoarekin lortutako multzoak izan emaitza hobeak lortu baitira Brown multzoekin.

Multzokatze mota bakoitzetik eratorritako ezaugarri berriekin lortutako analisiak konbinatu ditugunean, hobekuntza nabariak ikusi dira aztertutako hizkuntza guztietan. Gertaera horretatik, erabili ditugun multzokatze mota biak osagarriak izan daitezkeela ondoriozta daiteke. Gainera, multzokatzearekin lotuta ez dauden analisiak (oinarrizko analisiak) eta multzokatzearekin lotutako analisiak bakoitza bere kabuz konbinatu ditugunean, multzokatzearekin erlazionatutako analisiek konbinaketa osoan (oinarrizkoak + multzokatzekoak) erdietsitako emaitzan zerikusi handia dutela frogatu dugu.

Analisi sintaktiko eleaniztuna: meta-ezaugarriak (*5. kapitulua*)

Iturri desberdinetatik eratorritako meta-ezaugarrien eragina neurtu dugu euskaran, frantsesean, alemanean, hungarieran eta suedieran. Erabilitako hiru iturriak hitz-formak, Brown multzoak eta hitz-bektoreetan oinarritutako K-means multzoak izan dira. Iturri horietako bakoitzerako erdigainbegiratutako zuhaitz-bankuetan kontaketa desberdinak egin dira.

Bilaketa patroi bakoitzeko esperimentu bat gauzatu dugu hizkuntza bakoitzeko, bilaketa patroi horri buruzko kontaketak egin ondoren sortutako meta-ezaugarriak erabiliz. Bilaketa patroi desberdinekin erdietsitako emaitza askotan hobekuntzak erdietsi dira, meta-ezaugarriek hizkuntza bakoitzean eragin desberdina izan dutelarik. Euskaran eta alemanean eragin positiboagoa izan dute hitz-formetatik eratorritako meta-ezaugarriek. Hungarieran eta suedieran, berriz, Brown multzoetatik eratorritako meta-ezaugarriek eragin handiagoa izan dute. Suedieran erdietsi dira hobekuntzarik handienak, hainbat bilaketa patroirekin hobekuntza estatistikoki esanguratsuak lortuz, besteetan ez bezala. Frantsesean, aldiz, ez da inolako hobekuntzarik lortu meta-ezaugarrien bitartez.

Erabilitako hiru iturrien artean lagungarrienak hitz-formak eta Brown multzoak izan dira, hitz-bektoreetan oinarritutako K-means multzoetatik eratorritako meta-ezaugarriek analisi sintaktikoan ia eragin positiborik izan ez dutelarik.

Bestalde, emaitzarik hoberenak lortu dituzten hiru bilaketa patroietatik eratorritako meta-ezaugarrien batura erabili denean, euskararen kasuan soilik lortu da bakarkako emaitzarik hoberena gainditzea. Ondorioz, batu ditugun

bilaketa patroietatik eratorritako meta-ezaugarriak beraien artean osagarriak ez direla uste dugu.

Bukatzeko, emaitzarik hoberenak lortu dituzten hiru bilaketa patroietatik eratorritako meta-ezaugarrien baturak hizkuntzen dependentzia konkretuetan duen eragina aztertu dugu. Dependentzia konkretuetan ikusitako emaitzen arabera, erabilitako meta-ezaugarrien baturak badu nolabaiteko eragina dependentzia batzuetan. Euskararen kasuan, esaldiaren erroa definitzen duen dependentzia puntu batean hobetu da; hungarieraren eta alemanaren kasuan, subjektua definitzen duen dependentzia hobetzea lortu da; eta suedieraren kasuan, objektu zuzena definitzen duen dependentzia izan da hobetu dena. Frantseserako lortutako hobekuntza oso txikia izan den arren, emaitza hori ez da harritzekoa izan, banakako bilaketa patroietatik eratorritako meta-ezaugarriekin ere ez delako inolako hobekuntzarik lortu.

Analisi sintaktiko eleaniztuna: erabilitako tekniken konbinaketa (*5. kapitulua*)

5. kapituluan zehar landu diren tekniken konbinaketa aplikatu da analisi sintaktikoan. Horretarako, ezaugarrien ingeniaritzarekin erdietsitako analisiak, multzokatzearekin erdietsitako analisiak eta meta-ezaugarriekin erdietsitako analisiak konbinatu dira. Meta-ezaugarriak bi eratara konbinatu dira: hiru bilaketa patroi hoberenen analisiak konbinaketan bakarka gehituz eta hiru bilaketa patroi hoberenen baturatik eratorritako meta-ezaugarriekin erdietsitako analisiak gehituz.

Emaitzetan ikusi dugunaren arabera, ez da erraza ezaugarrien ingeniaritzan lortutako analisien konbinaketaren emaitza gainditzea. Ezaugarrien ingeniaritzako analisiei multzokatzearen bidez lortutako analisiak gehitzeak ez du inolako eraginik konbinaketaren emaitzan, ezta hiru bilaketa patroi hoberenetatik eratorritako meta-ezaugarrien analisiak bakarka gehitzeak ere. Hobekuntza bakarra konbinaketan hiru bilaketa patroi hoberenen baturatik eratorritako meta-ezaugarriekin erdietsitako analisiak gehitu direnean lortu da.

Dependentzia Unibertsalak (*6. kapitulua*)

6. kapituluan euskarazko zuhaitz-bankua Dependentzia Unibertsalak (UD) proiektuan erabiltzen den formatura (CoNLL-U) bihurtu dugu. Helburu hori lortzeko hainbat lan burutu ditugu: hitz anitzeko esapideen elementuen ana-

lisiak lortu, puntuazioa moldatu, koordinazioa moldatu, etiketa bakoitzerako kasuak aztertu eta bere UD parea identifikatu, ordena zuzena erabaki eta abar. Bihurketa prozesu guztia bukatu ondoren, jatorrizko zuhaitz-bankuak dituen 150.000 hitzetik 121.000 hitz bihurtu ditugu, euskara, nazioarteko proiektu garrantzitsu horren partaide izatea lortu dugularik.

Zenbakiek erakusten duten bezala, ezin izan dugu euskarazko zuhaitz-banku osoa bihurtu. Izan ere, bihurtu beharreko hitz anitzeko esapide batzuen osagaien analisia falta baitzen eta fenomeno batzuk ez baitira bihurtu. Hala ere, zuhaitz-bankuaren % 81a bihurtu dugu eta bihurtutako zuhaitz-bankuaren kalitatea oso ona dela deritzogu. Kalitate horren erakusle da bihurtutako bankuak analisi sintaktikoan bihurtu gabeko bankuaren pareko emaitzak lortzea.

7.3 Etorkizuneko lerroak

Tesi-lan honetan dependentzien analisi sintaktikoarekin erlazionatutako hainbat ikerketa lerro landu ditugun arren, jorratu nahi ditugun beste batzuk bidean gelditu dira. Hori dela eta, atal honetan etorkizunean landu nahiko genituzkeen esperimentu eta ikerkuntza bideak zerrendatuko ditugu.

Hibridazioarekin erlazionatutako esperimentuak zabaldu

Hibridazio mota desberdinetan egindako esperimentuetan lortutako emaitza orokorrak hobetu nahi ditugu. Ondorioz, gutxienez estatistiketan eta erregeletan oinarritutako hibridazioan proba gehiago egin nahi ditugu. Horretarako hizkuntzalarien laguntzarekin EDGK analizatzaileari gaizki ematen zaizkion dependentzia erlazioak aztertu nahi ditugu, dependentzia horietan lortutako estaldura eta doitasuna hobetzeko asmoarekin.

Bestalde, hibridazioak beste fenomeno linguistiko batzuetan duen eragina neurtu nahi da. Hibridazioari dagokion atalean esan dugun bezala, galderazko esaldietan egindako esperimentuak koordinazioa duten esaldietan eta esaldi konpletiboetan landu nahi dira. Are gehiago, fenomeno horiek ardatz dituzten erregela bereziak landuz gero hobekuntza altuagoak erdiestea espero da.

Bukatzeko, badakigu euskararako erabili ditugun baliabideen parekoak beste hizkuntza batzuetarako aurkitzea ez dela erraza, baina aurkitzen baditugu euskararekin egin ditugun esperimentuak hizkuntza horietara zabaltzea

gustatuko litzaiguke.

Ezaugarrien ingeniaritzarekin erlazionatutako esperimentuak zabaldu

Ezaugarrien ingeniaritzarekin egin diren esperimentuetan emaitzarik hoberenak, bakarkako ezaugarri morfologikoek eragin gehien duten hizkuntzetan erdietsi dira (euskara, hungariera eta suediera). Hori jakinda, antzeko emaitzak espero dira hizkuntza horiekin ezaugarri linguistiko asko partekatzen dituzten hizkuntzetan, esaterako, turkieran eta txekieran.

Bestalde, erabili diren zenbait bakarkako konfiguraziorekin ez da lortu oinarrizko emaitzak gainditzea, baina analisi desberdinak konbinatu ondoren emaitzak asko hobetu dira. Honekin, analisi desberdinen konbinaketaren bidez emaitzak hobetu nahi badira ugaritasuna garrantzitsua dela ondorioztatzen da. Beraz, ezaugarrien ingeniaritza aprobetxatuz sortutako aldaera gehiago erabiltzea interesgarria izango litzateke etorkizunean, baita aldaera horiek eta lan honetan aurkeztutakoak analizatzaile sintaktiko berriekin aztertzea ere.

Multzokatzearekin erlazionatutako esperimentuak zabaldu

Hitza lema eta atzizkietan banatu ez den esperimentuetan, Brown multzoekin erlazionatutako bi ezaugarri berri gehitu zaizkio analizatzaileari: bit katearen lehenengo lau bitak eta bit kate osoa. Etorkizunean orokortze maila desberdinak eskaintzen dituzten Brown multzoak sortu nahi dira konbinaketa aberasteko asmoarekin. Interesgarria litzateke gutxienez lau orokortze maila erabiltzea adibidez, 2, 4 eta 6 biteko kateak bit kate osoarekin.

Multzokatze teknika berriak erabiliko dira ezaugarri berriak sortzeko. Alde batetik multzokatze teknika berriek morfologikoki aberatsak diren hizkuntzetan duten eragina neurtu nahi da; bestetik, jakinda analisiak konbinatu direnean emaitzak asko hobetu direla, logikoa da pentsatzea ideia ona dela, multzokatze teknika berrietatik eratorritako ezaugarriak konbinaketan erabiltzea, aurretik zeudenekin batera.

Meta-ezaugarriekin erlazionatutako esperimentuak zabaldu

Meta-ezaugarriak lortzeko erdi-gainbegiratutako zuhaitz-bankuan hainbat gertaera moten maiztasunak kontatu dira, hau da, patroi desberdinak

bilatu dira eta horien maiztasunak kontatu dira. Erabili ditugun bilaketa patroi batzuetatik eratorritako meta-ezaugarriekin hobekuntzak erdietsi diren arren, modu desberdinean eratutako bilaketa patroiak erabiltzea interesgarria dela uste dugu. Erdi-gainbegiratutako zuhaitz-bankua sintaktikoki analizatuta dagoela aprobetxatuz, bilatuko diren patroietan dependentzia sintaktikoen etiketei buruzko informazioa ere erabili nahi da. Adibidez, gurasoaren hitz-forma, semearen hitz-forma eta bien arteko dependentzia etiketa kontatzen duen bilaketa patroia erabil genezake meta-ezaugarri berriak sortzeko.

Erabilitako zenbait teknika biltzen dituen sistema bat garatu

Tesian zehar hainbat teknika landu dira. Badakigu erabili ditugun teknika guztiak azkarra izan behar duen sistema batean inplementatzea oso zaila dela, baina teknika batzuk erabiltzea posible ikusten dugu. Esaterako, hiru ezaugarri morfologiko hoberenak, abstrakzio maila desberdina eskaintzen duten Brown multzoen bit kate desberdinak eta bilaketa patroi hoberenetatik eratorritako meta-ezaugarriak sistema berean probatu daitezke. Horiekin batera kategoria edo azpikategoria ezaugarri morfologiko esanguratsuenaz ordezkatzen badira, orduan teknika guztiak MaltBlender tresnaren bidez konbinatzean lortutako emaitzekin konparaketa egin ahal izango dugu. Konparaketa horren bidez erabaki ahal izango dugu zein bide den eraginkorrena emaitza aldetik eta emaitza horiek lortzeko erabilitako denborak merezi duen ala ez.

Euskarazko zuhaitz-banku osoa bihurtu UD formatura

Euskarazko zuhaitz-bankuaren zatirik handiena bihurtzea lortu dugun arren, ezin izan dira hitz guztiak bihurtu. Horrenbestez, etorkizunean, bihurtu gabe gelditu diren hitzak eta ezaugarriak bihurtzea da gure asmoa.

Bestalde, euskarazko UD zuhaitz-bankuarekin eta UD proiektuko gainontzeko zuhaitz-bankuekin hizkuntza arteko (*cross-lingual*) hainbat esperimentu egin daitezke. Hasteko, hizkuntza desberdinetako zuhaitz-bankuen antzekotasun sintaktikoa neurtu nahiko genuke. Horretarako, hizkuntza bateko ikasketarako corpusarekin ikasiko luke sistemak (hitz-formarik gabe), eta beste hizkuntza bateko testerako corpusaren gainean (hitz-formarik gabe) aplikatuko luke ikasitakoa. Gauzak horrela, hizkuntza batek bere ikasketarako eta testerako corpusak erabilita (hitz-formarik gabe) erdietsitako emai-

tzarekin konparatuko genuke emaitza. Konparatutako emaitzak antzekoak badira, esan genezake erabilitako bi hizkuntzen zuhaitz-bankuak sintaktikoki antzekoak direla.

Bibliografia

Abeillé A., Clément L., eta Toussenel F. Building a treebank for french. In Abeillé A., editor, *Treebanks*. Kluwer, Dordrecht, 2003.

Abney S.P. Parsing by chunks. *Principle-based parsing*, 257–278. Springer, 1991.

Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Garmendia A., eta Oronoz M. Construction of a Basque dependency treebank. 201–204, 2003.

Aduriz I. *EUSMG: MORFOLOGIATIK SINTAXIRA MURRIZTAPEN GRAMATIKA ERABILIZ. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak.* Doktoretza-tesia, 2000.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., eta Urkia M. A framework for the automatic processing of basque. *Proceedings of Workshop on Lexical Resources for Minority Languages*, 1998.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Sarasola K., *et al.*. A word-grammar based morphological analyzer for agglutinative languages. *Proceedings of the 18th*

conference on Computational linguistics-Volume 1, 1–7. Association for Computational Linguistics, 2000.

Aduriz I., Aranzabe M.J., Arriola J.M., de Ilarraza A.D., Gojenola K., Oronoz M., eta Uria L. A cascaded syntactic analyser for basque. *International Conference on Intelligent Text Processing and Computational Linguistics*, 124–134. Springer, 2004.

Agirre E. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala.* Doktoretza-tesia, 1999.

Agirre E., Baldwin T., eta Martinez D. Improving parsing and pp attachment performance with sense information. *ACL*, 317–325. Citeseer, 2008.

Aldabe I. *Automatic Exercise Generation Based on Corpora and Natural Language Processing Techniques.* Doktoretza-tesia, 2011.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. Edbl: a general lexical basis for the automatic processing of basque. *Proceedings of the IRCS Workshop on linguistic databases.* IRCS Workshop on linguistic databases., 2001.

Aldezabal I. *ADITZ-AZPIKATEGORIZAZIOAREN AZTERKETA SINTAXI PARTZIALETIK SINTAXI OSORAKO BIDEAN. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* Doktoretza-tesia, 2004.

Aldezabal I., Aranzabe M.J., Arriola J.M., eta Díaz de Ilarraza A. Syntactic annotation in the reference corpus for the processing of basque (epec): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory*, 5(2):241–269, 2009.

Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak.* Doktoretza-tesia, 1995.

Ambati B.R., Husain S., Jain S., Sharma D.M., eta Sangal R. Two methods to incorporate local morphosyntactic features in hindi dependency parsing. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 22–30, 2010.

Andreas J. eta Klein D. How much do word embeddings encode about syntax? *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 822–827, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-2133`.

Aranzabe M.J. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. *Doktoretza-tesia. Euskal Herriko Unibertsitatea, UPV/EHU*, 2008.

Aranzabe M.J., Atutxa A., Bengoetxea K., Díaz de Ilarraza A., Goenaga I., Gojenola K., eta Uria L. Automatic conversion of the basque dependency treebank to universal dependencies. *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)*, 233–241. Institute of Computer Science of the Polish Academy of Sciences, Warszawa, Poland. ISBN: 978-83-63159-18-4, 2015.

Arregi X. *ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza.* Doktoretza-tesia, 1995.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* Doktoretza-tesia, 2010a.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* Doktoretza-tesia, 2010b.

Arriola J.M. *EUSKAL HIZTEGIA-REN AZTERKETA ETA EGITURATZEA EZAGUTZA LEXIKALAREN ESKURATZE AUTOMATIKOARI BEGIRA. Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean.* Doktoretza-tesia, 2000.

Artola X. *HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza. Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena. / Conception et construction d'un systÃĺme intelligent d'aide dictionnariale (SIAD). Acquisition et reprÃľsentation des connaissances dictionnariales, Ãľtablissement de mÃľcanismes de dÃľduction et spÃľcification des fonctionnalitÃľs de base.* Doktoretza-tesia, 1993.

Artola X., DÃŋaz de Ilarraza A., Ezeiza N., Gojenola K., Labaka G., Sologaistoa A., eta Soroa A. A framework for representing and managing linguistic annotations based on typed feature structures. *RANLP 2005. ISBN: 954-91743-3-6*, 2005.

Attardi G., Dell'Orletta F., Simi M., Chanev A., eta Ciaramita M. Multilingual dependency parsing and domain adaptation using desr. *EMNLP-CoNLL*, 1112–1118, 2007.

Ballesteros M. eta Nivre J. Maltoptimizer: an optimization tool for maltparser. *Proceedings of the Demonstrations at the 13th Conference of the European Chaptr of the Association for Computational Linguistics*, 58–62, 2012b.

Bansal M., Gimpel K., eta Livescu K. Tailoring continuous word representations for dependency parsing. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

Bengoetxea K. *Estaldura zabaleko euskararako analizatzaile sintaktiko estatistikoa.* Doktoretza-tesia, 2014.

Bengoetxea K., Agirre E., Nivre J., Zhang Y., eta Gojenola K. On wordnet semantic classes and dependency parsing. *ACL (2)*, 649–655, 2014.

Bengoetxea K. eta Gojenola K. Application of different techniques to dependency parsing of basque. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 31–39, 2010.

Bohnet B. Very high accuracy and fast dependency parsing is not a contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics*, 89–97, 2010.

Brants S., Dipper S., Hansen S., Lezius W., eta Smith G. The tiger treebank. *Proceedings of the workshop on treebanks and linguistic theories*, 168 lib., 2002.

Brown P.F., deSouza P.V., Mercer R.L., Pietra V.J.D., eta Lai J.C. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4): 467–479, December 1992. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=176313.176316`.

Buchholz S. eta Marsi E. Conll-x shared task on multilingual dependency parsing. *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164. Association for Computational Linguistics, 2006.

Candito M. eta Seddah D. Parsing word clusters. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 76–84. Association for Computational Linguistics, 2010.

Carreras X. *Learning and inference in phrase recognition: A filtering-ranking architecture using perceptron*. Doktoretza-tesia, 2005.

Carreras X. Experiments with a higher-order projective dependency parser. *EMNLP-CoNLL*, 957–961, 2007.

Çetinoğlu O. eta Kuhn J. Towards joint morphological analysis and dependency parsing of turkish. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, 23–32, Prague, Czech Republic, August 2013. Charles University in Prague, Matfyzpress, Prague, Czech Republic. URL `http://www.aclweb.org/anthology/W13-3704`.

Chang C.C. eta Lin C.J. Training v-support vector classifiers: theory and algorithms. *Neural computation*, 13(9):2119–2147, 2001.

Chen W., Kazama J., Uchimoto K., eta Torisawa K. Improving dependency parsing with subtrees from auto-parsed data. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 570–579. Association for Computational Linguistics, 2009.

Chen W., Zhang M., eta Zhang Y. Semi-supervised feature transformation for dependency parsing. *EMNLP*, 1303–1313, 2013.

Choi K.S., Han Y.S., Han Y.G., eta Kwon O.W. Kaist tree bank project for korean: Present and future development. *Proceedings of the International Workshop on Sharable Natural Language Resources*, 7–14. Citeseer, 1994.

Chomsky N. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.

Chomsky N. Aspects ofa theory of syntax. *Cambridge, Mass.: MI'I'. I'rw Language Journal*, 53:334–341, 1965.

Chomsky N. Lectures on government and binding. *Dordrecht: Foris*, 1981.

Chu Y.J. eta Liu T.H. On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396, 1965.

Cortes C. eta Vapnik V. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Crammer K. eta Singer Y. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3(Jan):951–991, 2003.

Csendes D., Csirik J., eta Gyimóthy T. The szeged corpus: A pos tagged and syntactically annotated hungarian natural language corpus. *International Conference on Text, Speech and Dialogue*, 41–47. Springer, 2004.

De Marneffe M.C., Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., eta Manning C.D. Universal stanford dependencies: A cross-linguistic typology. *LREC*, 14 lib., 4585–4592, 2014.

De Marneffe M.C., MacCartney B., Manning C.D., *et al.*. Generating typed dependency parses from phrase structure parses. *Proceedings of LREC*, 6 lib., 449–454, 2006.

De Marneffe M.C. eta Manning C.D. The stanford typed dependencies representation. *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. Association for Computational Linguistics, 2008.

Domingos P. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

Duan X., Zhao J., eta Xu B. Probabilistic parsing action models for multilingual dependency parsing. *EMNLP-CoNLL*, 940–946, 2007.

Edmonds J. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240, 1967.

Eisner J.M. Three new probabilistic models for dependency parsing: An exploration. *Proceedings of the 16th conference on Computational linguistics*, 340–345. Association for Computational Linguistics, 1996.

Ezeiza N. *CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Doktoretza-tesia, 2002.

Ezeiza N., Alegria I., Arriola J.M., Urizar R., eta Aduriz I. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, 380–384. Association for Computational Linguistics, 1998.

Fan R.E., Chang K.W., Hsieh C.J., Wang X.R., eta Lin C.J. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

Goenaga I., Gojenola K., eta Ezeiza N. Combining clustering approaches for semi-supervised parsing: the basque team system in the sprml2014 shared task. *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages: Shared Task on Statistical Parsing of Morphologically Rich Languages*, 2014.

Gojenola K. *Euskararen sintaxi konputazionalerantz: oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan.* Doktoretza-tesia, 2000a.

Gojenola K. *Euskararen sintaxi konputazionalerantz: oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan.* Doktoretza-tesia, 2000b.

Green S. eta Manning C.D. Better arabic parsing: Baselines, evaluations, and analysis. *Proceedings of the 23rd International Conference on Computational Linguistics*, 394–402. Association for Computational Linguistics, 2010.

Guthmann N., Krymolowski Y., Milea A., eta Winter Y. Automatic annotation of morpho-syntactic dependencies in a modern hebrew treebank. *LOT Occasional Series*, 12:77–90, 2008.

Habash N. eta Roth R.M. Catib: The columbia arabic treebank. *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 221–224. Association for Computational Linguistics, 2009.

Haffari G., Razavi M., eta Sarkar A. An ensemble model that combines syntactic and semantic clustering for discriminative dependency parsing. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 710–714. Association for Computational Linguistics, 2011.

Hall J., Nilsson J., eta Nivre J. Single malt or blended? a study in multilingual parser optimization. *Trends in Parsing Technology*, 19–33. Springer, 2010.

Hall J., Nivre J., eta Nilsson J. A hybrid constituency-dependency parser for swedish. *Proceedings of NODALIDA*, 284–287, 2007.

Hartigan J.A. eta Wong M.A. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, 100–108, 1979.

Iruskieta M. *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean*. Doktoretza-tesia, 2014.

Karlsson F., Voutilainen A., Heikkilae J., eta Anttila A. *Constraint Grammar: a language-independent system for parsing unrestricted text*, 4 lib. Walter de Gruyter, 1995.

Kim Y.B., Chae H., Snyder B., eta Kim Y.S. Training a korean srl system with rich morphological features. 2014.

Koo T., Carreras X., eta Collins M. Simple semi-supervised dependency parsing. Association for Computational Linguistics, 2008.

Liang P. *Semi-supervised learning for natural language*. Doktoretza-tesia, Massachusetts Institute of Technology, 2005.

Lopez de Lacalle O. *Domain-Specific Word Sense Disambiguation*. Doktoretza-tesia, 2009.

Maamouri M., Bies A., Buckwalter T., eta Mekki W. The penn arabic treebank: Building a large-scale annotated arabic corpus. *NEMLAR conference on Arabic language resources and tools*, 27 lib., 466–467, 2004.

Maritxalar M. *Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea*. Doktoretza-tesia, 1999.

Martins A.F., Das D., Smith N.A., eta Xing E.P. Stacking dependency parsers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 157–166. Association for Computational Linguistics, 2008.

Marton Y., Habash N., eta Rambow O. Improving arabic dependency parsing with lexical and inflectional morphological features. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 13–21. Association for Computational Linguistics, 2010.

McClosky D., Charniak E., eta Johnson M. Reranking and self-training for parser adaptation. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 337–344. Association for Computational Linguistics, 2006.

McDonald R., Crammer K., eta Pereira F. Online large-margin training of dependency parsers. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 91–98. Association for Computational Linguistics, 2005.

McDonald R., Lerman K., eta Pereira F. Multilingual dependency analysis with a two-stage discriminative parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 216–220. Association for Computational Linguistics, 2006.

McDonald R. eta Nivre J. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230, 2011.

McDonald R.T., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K.B., Petrov S., Zhang H., Täckström O., *et al.*. Universal dependency annotation for multilingual parsing. *ACL (2)*, 92–97. Citeseer, 2013.

Mikolov T., Chen K., Corrado G., eta Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov T., Sutskever I., Chen K., Corrado G.S., eta Dean J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119, 2013b.

Mikolov T., Yih W.t., eta Zweig G. Linguistic regularities in continuous space word representations. *HLT-NAACL*, 746–751. Citeseer, 2013c.

Miller S., Guinness J., eta Zamanian A. Name tagging with word clusters and discriminative training. *Proceedings of HLT*, 337–342, 2004.

Momtazi S. eta Klakow D. A word clustering approach for language model-based sentence retrieval in question answering systems. *Proceedings of the 18th ACM conference on Information and knowledge management*, 1911–1914. ACM, 2009.

Nguyen M., Shimazu A., Nguyen T.P., eta Phan X.H. A multilingual dependency analysis system using online passive-aggressive learning. *EMNLP-CoNLL*, 1149–1155, 2007.

Nivre J. Constraints on non-projective dependency parsing. *EACL*, 2006.

Nivre J. Non-projective dependency parsing in expected linear time. *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: Volume 1-Volume 1*, 351–359. Association for Computational Linguistics, 2009.

Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., eta Yuret D. The CoNLL 2007 shared task on dependency parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague, Czech Republic, June 2007.

Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., eta Marsi E. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007b.

Nivre J., Kuhlmann M., eta Hall J. An improved oracle for dependency parsing with online reordering. *Proceedings of the 11th international conference on parsing technologies*, 73–76. Association for Computational Linguistics, 2009.

Nivre J. eta McDonald R.T. Integrating graph-based and transition-based dependency parsers. *ACL*, 950–958, 2008.

Nivre J. eta Megyesi B. Bootstrapping a swedish treebank using cross-corpus harmonization and annotation projection. *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*, 97–102, 2007.

Nivre J., Nilsson J., eta Hall J. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. *Proceedings of LREC*, 1392–1395, Genoa, Italy, 2006.

Otegi A. *Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak*. Doktoretza-tesia, 2012.

Petrov S., Das D., eta McDonald R. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2012.

Petrov S. eta Klein D. Improved inference for unlexicalized parsing. *HLT-NAACL*, 7 lib., 404–411, 2007.

Rosa R., Masek J., Marecek D., Popel M., Zeman D., eta Zabokrtskỳ Z. Hamledt 2.0: Thirty dependency treebanks stanfordized. *LREC*, 2334–2341, 2014.

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Sagae K. eta Lavie A. Parser combination by reparsing. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 129–132. Association for Computational Linguistics, 2006.

Sagae K. eta Tsujii J. Dependency parsing and domain adaptation with lr models and parser ensembles. *EMNLP-CoNLL*, 2007 lib., 1044–1050, 2007.

Seddah D., Kübler S., eta Tsarfaty R. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, 103–109, 2014.

Seddah D., Tsarfaty R., Kübler S., Candito M., Choi J., Farkas R., Foster J., Goenaga I., Gojenola K., Goldberg Y., Green S., Habash N., Kuhlmann M., Maier W., Nivre J., Przepiorkowski A., Roth R., Seeker W., Versley Y., Vincze V., Woliński M., Wróblewska A., eta Villemonte de la Clérgerie E. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*, Seattle, WA, 2013.

Seeker W. eta Kuhn J. Making Ellipses Explicit in Dependency Conversion for a German Treebank. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 3132–3139, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Seeker W. eta Kuhn J. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55, 2013.

Shieber S. An introduction to unification-based theories of grammar. *CSLI Lecture Notes*, (4), 1986.

Sima'an K., Itai A., Winter Y., Altman A., eta Nativ N. Building a treebank of modern hebrew text. *Traitement Automatique des Langues*, 42(2): 247–380, 2001.

Spranger M.S.K. Global learning of labelled dependency trees. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, 1156–1160, 2007.

Surdeanu M. eta Manning C.D. Ensemble models for dependency parsing: cheap and good? *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 649–652. Association for Computational Linguistics, 2010.

Świdziński M. eta Woliński M. Towards a bank of constituent parse trees for polish. *International Conference on Text, Speech and Dialogue*, 197–204. Springer, 2010.

Täckström O., McDonald R., eta Uszkoreit J. Cross-lingual word clusters for direct transfer of linguistic structure. *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 477–487. Association for Computational Linguistics, 2012.

Tapanainen P. *The constraint grammar parser CG-2*. University. Department of General Linguistics, 1996.

Tapanainen P. eta Voutilainen A. Tagging accurately: don't guess if you know. *Proceedings of the fourth conference on Applied natural language processing*, 47–52. Association for Computational Linguistics, 1994.

Titov I. eta Henderson J. Fast and robust multilingual dependency parsing with a generative latent variable model. *EMNLP-CoNLL*, 947–951, 2007.

Tsarfaty R. A unified morpho-syntactic scheme of stanford dependencies. *ACL (2)*, 578–584, 2013.

Tsarfaty R., Seddah D., Goldberg Y., Kübler S., Candito M., Foster J., Versley Y., Rehbein I., eta Tounsi L. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. *In Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 2010.

Turian J., Ratinov L., eta Bengio Y. Word representations: a simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394. Association for Computational Linguistics, 2010.

Urizar R. *Euskal lokuzioen tratamendu konputazionala*. Doktoretza-tesia, 2012.

Vincze V., Szauter D., Almási A., Móra G., Alexin Z., eta Csirik J. Hungarian dependency treebank. *LREC*, 2010.

Wolinski M., Glowinska K., eta Swidzinskil M. A preliminary version of skladnicaâĂŤa treebank of polish. *Proceedings of the 5th Language & Technology Conference, Poznan*, 299–303, 2011.

Wróblewska A. Polish dependency bank. *Linguistic Issues in Language Technology*, 7(1):1–15, 2012.

Yamada H. eta Matsumoto Y. Statistical dependency analysis with support vector machines. *Proceedings of IWPT*, 3 lib., 195–206, 2003.

Zeman D. Reusable tagset conversion using tagset drivers. *LREC*, 2008.

Zeman D., Dušek O., Mareček D., Popel M., Ramasamy L., Štěpánek J., Žabokrtskỳ Z., eta Hajič J. Hamledt: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637, 2014.

Zeman D. eta Resnik P. Cross-language parser adaptation between related languages. *IJCNLP*, 35–42, 2008.

Glosategia

adabegi

Grafo bateko puntu berezia da, beste puntu batzuekin lotuta agertzen dena.

analizatzaile sintaktiko estatistikoa (*parser*)

Teknika estatistikoak erabiliz esaldiak sintaktikoki aztertzen dituen analizatzailea. Normalean esaldiaren egitura sintaktikoa eraikitzen da eta egitura horretan agertzen diren lotura motak (etiketak) ere definitzen dira.

bozketa (*voting*)

Sistemaren azken emaitza bezala joan behar dena zein den erabakitzeko beste sistema batzuen emaitzen artean bozketa egiteari deritzo. Bozketa mota desberdinak daude, baina horien artean sinpleena boto gehien jasotzen dituzten emaitzak aukeratzea da.

corpus etiketatua

Informazio linguistikoarekin aberastutako corpusa. Zehaztutako etiketen arabera hainbat zereginetarako erabiltzen dira.

corpus gordina

Informazio linguistikoarekin aberastu gabeko corpusa.

dependentzia egitura

Dependentzietan oinarritutako esaldi egitura.

dependentzia zuhaitza

Zuhaitz bat elementuak hierarkikoki ordenatzen dituen datu egitura ez lineala da. Adabegi multzo bat, berriz, zuzendutako arku talde bat da. Arku bakoitzak bi adabegi konektatzen ditu guraso-seme erlazioaren bidez. Zuhaitz batean gurasorik ez daukan adabegi bakarrari erroa deritzo. Erroa izan ezik, gainontzeko adabegietako bakoitza beste adabegi batera konektatuta dago arku baten bidez eta lotura bakoitzari dependentzia etiketa bat dagokio. Umerik ez duten adabegiei hosto deritze. Bi adabegi guraso berdinaren seme direnean senide deritze. Ume baten gurasoaren gurasoari aitona deituko diogu.

erdi-gainbegiratutako zuhaitz-bankua

Ikasketa gainbegiratua erabiliz ikasi duen sistema batek etiketatu duen corpus ez-gainbegiratua.

ezaugarrien ingeniaritza (*feature engineering*)

Eskuragarri dauden ezaugarrien erabilera jorratu nahi den atazara ahalik eta gehien doitzeko prozesuari deritzo. Ezaugarriak era arruntean erabili beharrean beste modu batean erabilita emaitzak hobetzea da prozesuaren helburua.

garapenerako corpusa

Sistemaren eraginkortasuna doitzeko erabiltzen den corpusaren zatia da. Normalean eskuragarri dagoen corpusaren % 10 inguru erabiltzen da garapenerako.

grafo

Egitura matematikoa, adabegi eta arku deituriko elementuez osatua.

ikasketarako corpusa

Sistemek ikasteko erabiltzen duten corpus zatia da. Normalean eskuragarri dagoen corpusaren % 75 - % 80 inguru erabiltzen da ikasketarako.

kategoria-etiketatzailea (*POS tagger*)

Hitz bakoitzari dagokion kategoria gramatikala edota bestelako marka lexikalen bat esleitzen dion tresna da.

lema (*lemma*)

Hitza atzizki flexiborik gabe, hiztegietan sarrera gisa dagoen hori.

Markov-en eredu ezkutuak (*hidden Markov models*)

Markov-en ereduek Markov-en propietatea betetzen dute: gertaera bat betetzeko probabilitatea bere berehalako aurreko gertaer(ar)en menpe dago soilik. Automata finitu baten moduan ikus daiteke Markov-en eredu ezkutu bat. Egoerek ereduaren aldagaiak errepresentatzen dituzte, eta arkuek egoera batetik bestera joateko dagoen probabilitatea gordetzen dute. Alfabeto bateko sinboloak sortzeko aukera dute egoerek, probabilitate-funtzio baten arabera. Ezkutua dela esaten da ezin delako jakin ereduak aukeratzen duen egoeren segida, ez baita probabilitate maximoa duen trantsizioa beti aukeratuko, maximo globala bilatzen duen egoeren segida baizik.

McNemar testa

McNemar testa bi sailkatzaileen arteko aldea esanguratsua den ala ez erabakitzeko erabiltzen da. Horretarako, corpusa bi zatitan banatu behar da: ikasketa corpusa eta test corpusa. Bi sailkatzaileak (A, B) ikasketa corpus bera erabiliz ikasi ondoren, test corpus beraren gainean ebaluatu behar dira. Hipotesi nuluaren arabera, A sailkatzaileak ondo eta B sailkatzaileak gaizki sailkatutako adibideen kopuruak A sailkatzaileak gaizki eta B sailkatzaileak ondo sailkatutako adibideen kopuruaren berdina izan behar du. Datu hauen arabera, χ^2 testan oinarritzen da McNemar testa, hipotesi nulu uka daitekeen edo ez erabakitzeko. Hipotesia errefusatu baldin badaiteke, bi sailkatzaileen arteko aldea esanguratsua dela esaten da.

meta-ezaugarriak (*meta-features*)

Ezaugarri bakunen arteko erlazioak edo informazioa biltzen duten ezaugarri bereziak dira.

morfologikoki aberatsa den hizkuntza (*morphologically rich language*)

Morfologiaren ikuspuntutik aberatsa den hizkuntza. Normalean hizkuntza horietan lema bakoitza hainbat hitz-forma desberdinekin erlazionatuta egoten da informazio morfologikoa eransten duten hainbat morfemen bidez.

multzokatzea (*clustering*)

Ikasketa ez-gainbegiratua egiteko teknika bat da. Ikasketarako corpuseko instantziak sailkatu gabe ditugunean, instantzia horiek euren arteko zenbait antzekotasunen arabera bil daitezke; horrela lortutako multzo bakoitza klase bat edo *cluster* bat dela esaten da.

murriztapen gramatika (*constraint grammar*)

Patroiak identifikatzeko eta etiketak jarri, kendu edo aldatzeko aukera ematen duen egoera finituko formalismoa.

pilaketa (*stacking*)

Sistema baten irteerako emaitza beste sistema bati sarrerako informazio bezala pasatzeari deritzo.

testerako edo ebaluaziorako corpusa

Sistemaren eraginkortasuna neurtzeko erabiltzen den corpusaren zatia da. Normalean eskuragarri dagoen corpusaren % 10 inguru erabiltzen da testerako.

urre-patroi (*gold standard*)

Automatikoki eskuratutako emaitzak ebaluatzeko erabiltzen diren eskuz sortutako emaitza prototipikoak dira.

zuhaitz-banku (*treebank*)

Sintaktikoki etiketatutako corpusa.

A

Eranskina

Dependentzia etiketa	Deskripzioa
aditz_nagusi	Aditza
aponcmod	Aposizioa (ez-perpausa)
apocmod	Aposizioan dagoen mendeko perpaus jokatua
apoxmod	Aposizioan dagoen mendeko perpaus ezjokatua
arg_mod	Etiketa semantikoa
auxmod	Aditz laguntzailea
ccomp_obj	Mendeko perpaus osagarri jokatua, objektua
ccomp_subj	Mendeko perpaus osagarri jokatua, subjektua
cmod	Mendeko perpaus jokatua; adizlaguna edo izenlaguna
detmod	Determinatzailea
entios	Entitate-osagaia
galdemod	Aditzaren indartzailea
gradmod	Graduatzailea
haos	Hitz anitzekoaren osagaia
itj_out	Interjekzioa
lot	Loturazko elementuak
lot_at	Lokailuak
menos	Menderagailu-osagaia
ncmod	Adizlaguna
ncmod	Modifikatzailea
ncpred	Predikatiboa (ez-perpausa)
ncobj	Objektua (ez-perpausa)
ncsubj	Subjektua (ez-perpausa)
nczobj	Zehar-objektua (ez-perpausa)
postos	Postposizio-osagaia
prtmod	Partikulak; aditzarekin agertu ohi direnak
xcomp_obj	Mendeko perpaus osagarri ezjokatua, objektua
xcomp_subj	Mendeko perpaus osagarri ezjokatua, subjektua
xcomp_zobj	Mendeko perpaus osagarri ezjokatua, zehar-objektua
xmod	Mendeko perpaus ezjokatua; adizlaguna edo izenlaguna
xpred	Mendeko perpaus ezjokatua; predikatiboa

A.1 taula – Euskarako dependentzia etiketak alfabetikoki zerrendatuta.