Universidad del País Vasco

Euskal Herriko Unibertsitatea

Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz

Egilea:

Haritz Salaberri Izko

Zuzendariak:

Olatz Arregi Uriarte
Beñat Zapirain Sierra

Hizkuntzaren Azterketa eta Prozesamenduaren Doktorego Programa

Euskal Herriko Unibertsitatea (EHU/UPV)
Konputagailuen Arkitektura eta Teknologia saila. Informatika Fakultatea.

2017

Abstract

The focus of this thesis is on Semantic Role Labeling (SRL) for Basque. As a result of this work, the ability to perform SRL, or shallow semantic parsing, over texts written in Basque has been gained. Additionally, the first steps towards the SRL-related tasks of automatically annotating space and time information contained within Basque texts have been taken. As a matter of fact, spatiotemporal annotation tools that meet current standards have been developed. The performance of the systems designed and implemented during this research have been compared to the performance of analogous tools from other languages.

The other target of this work, apart from extending the processing pipeline for Basque, has been to test the next two hypotheses:

- That in Basque the impact of SRL on temporal annotation is positive, as it is in English and Spanish as well.
- That the linguistic expression of space, like the linguistic expression of time, is a semantic phenomenon and, therefore, the information given by semantic roles has a positive effect on the annotation of spatial information.

In this document, the steps that have led to the meeting of the thesis objectives are described. The results, difficulties and conclusions that have raised from each of these steps are also described. Three systems have been developed in order to accomplish these objectives: an SRL tool by the name of *bRol*, a temporal tagger (*bTime*) and a spatial tagger (*X-Space*).

Laburpena

Tesi honen xede nagusia euskarazko rol semantikoen etiketatze automatikoa da (*Semantic Role Labeling*, SRL). Besteak beste, euskaraz idatzitako testuen analisi-katean SRL edo azaleko analisi semantikoa egitea ahalbidetu dugu. Gainera, SRL atazarekin lotura daukaten euskarazko denbora eta espazio informazioaren etiketatze automatikorako ere aurrerapenak egin ditugu. Izan ere, gaur egungo estandarretara egokitutako denboraren eta espazioaren etiketatzeko tresnak garatu ditugu tesian. Orobat, diseinatu eta inplementatutako sistema guztien emaitzak beste hizkuntza batzuk prozesatzen dituzten tresnen emaitzekin alderatu ditugu.

Gure lanaren beste helburua, euskararen analisi-katea hedatzeaz eta osatzeaz gainera, ondorengo bi hipotesiak baieztatzea izan da:

- Euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek daukaten eragina positiboa dela, ingelesez eta gaztelaniaz bezala.
- Espazioaren adierazpen linguistikoa, denborarena bezala, fenomeno semantikoa dela, eta horregatik semantika eta, zehazkiago, rol semantikoek duten garrantzia nabarmena dela, informazio espazialaren etiketatze eraginkorra egin ahal izateko.

Dokumentu honetan gure ikerketaren hasieran finkatu genituen helburuak betetzeko eman ditugun pausoak zehaztasunez deskribatzen saiatu gara. Ahaleginak egin ditugu, halaber, urrats haietako bakoitzetik sortutako emaitza, zailtasun eta ondorioak deskribatzeko. Aipatu helburuak bete ahal izateko *bRol* izeneko SRL tresna, *bTime* denbora informazioa etiketatzeko sistema eta *X-Space* informazio espaziala markatzeko tresna garatu ditugu besteak beste.

AURKIBIDEA

IRUDIEN ZERRENDA

TAULEN ZERRENDA

1

SARRERA

Rol semantikoen etiketatze automatikoa, azaleko analisi semantikoa ere deitzen dena, hizkuntzaren prozesamenduaren alorrean kokatzen den ataza da. Honetan, testuetako gertaerei (edo predikatuei) antzeman eta horietan parte hartzen duten aktoreak (edo argumentuak) zein diren eta parte nola hartu duten jakin nahi da, baita aipatu gertaerak zer baldintzatan gauzatu diren ere. Lan honen beste xede nagusietako bat euskaraz idatzitako testuetan *nork nori zer* egin dion, *non* eta *noiz* egin dion adieraztea da. Honetarako rolen etiketatze automatikoa aztertzeaz ez ezik testuetan aurkitzen den denbora eta espazio informazioa ustiatzeaz ere arduratu gara, hirurak, hots, rol semantikoen etiketatzea, denbora eta espazio informazioaren erauzketa eta sailkapena estuki lotutako eta elkarren osagarri diren atazak direlako. Izan ere, rol semantikoek *nork nori zer* egin dion eta *non* eta *noiz* finkatzea lortzen badute ere, *non* eta *noiz* galderei zehazkiago erantzuteko, espazioaz eta denboraz espresuki arduratzen diren eskema berezituak erabiltzea beharrezkoa dela deritzogu.

Kapitulu honen lehenengo hiru ataletan kontzeptu orokorrak emango dira (1.1, 1.2 eta 1.3). Jarraian, 1.4 eta 1.5 ataletan aipatutako hiru atazen aurkezpena egingo dugu, 1.6 atalean tesiaren motibazioa azalduko dugu eta, ondoren, 1.7 atalean, gure ekarpenak zerrendatuko ditugu. Azkenik, 1.8 eta 1.9 ataletan, sarrerako kapitulua bukatu aurretik, tesi lanaren egitura aurkeztu eta bertatik sortu diren argitalpenak zein izan diren aipatuko dugu.

1.1 Gertaerak

Gertaerak tesi lan honen oinarri eta ardatz direla esan dezakegu, baina, zer da gertaera? Eta *gertaera* kontzeptuari eman zaizkion definizio guztietatik, zein da guk darabilguna?, alegia, zer da guretzat *gertaera*?

Zientziaren alor bat baino gehiago arduratu izan da gertaerak lantzeaz, besteak beste fisika, filosofia eta hizkuntzalaritza. Fisikan, esate baterako, erlatibitatearen teoria ikertzean, gertaera kokaleku eta une zehatz batean agitzen den egoera fisikoa dela ulertzen da; une jakin batean edalontzi bat lurraren kontra puskatzea, adibidez. Filosofian, berriz, hainbat dira gertaeren inguruko teoriak; hauetatik bost dira ezagunenak eta gainerako zientzietan (hizkuntzaren filosofian eta ondorioz hizkuntzaren prozesaketan) eragin handiena izan dutenak: Kimena (1), Davidsonena (2), Lewisena (3), Badiou eta Felthamena (4) eta Deleuzerena (5).

1. Kimek (1976) aurkeztutako teoriaren arabera, gertaera bat une jakin batean jazotzen den propietate baten gauzatzea da, beste modu batera adierazita, propietate baten une jakin bateko adibide-bihurketa (*property-exemplification*) dela esan daiteke. Sokratesen heriotza, esate baterako, Sokratesek berak une jakin batean (K.a. 399. urtean) parte hartu duen *heriotza* tasunaren gauzatzea da.

2. Davidsonek (1967) proposatutako gertaeren teoriak, ordea, bi irizpide ezartzen ditu haiek definitzeko orduan: kausarena eta espazio-denborarena. Teoria honen arabera, bi gertaera bat eta bera direla ulertzen da baldin eta kausa eta efektu bera badute, edota kokaleku (espazio) eta une (denbora) berean agitu badira.

3. Lewisen (1987) teoriaren arabera, aldiz, gertaera bat leku-denborazko kategoria baten barnean sailka daitekeen edozein jazoera da; kategoria hauetako espazioak bizi garen mundu honetan edo beste edozeinetan kokatuta egon daitezke. Mundu ezberdinen ikuspegi hau bat dator Lewisek berak proposatutako *errealismo modalarekin* (Lewis, 1986). Errealismo modalaren arabera, hainbat mundu daude, eta denak dira bizi garena bezain errealak.

4. Badiou eta Felthamen (2007) ustez, gertaera egoera jakin bat guztiz aldatu eta egoera berria ezartzen duen jazoera edo iraultza da. Gertaeraren definizio hau

gehienbat testuinguru politikoan erabili izan bada ere, edozein testuingurutan erabiltzeko dago formulatuta; esate baterako, Georg Cantor (2006) matematikariak multzoen teoriaren inguruan burututako lana gertaera dela onartzen dute Badiou eta Felthamek, ikerketa matematikoa guztiz aldatzea eragin zuelako. Badiouren gertaerak *gertaera-eremua* izendatutako espazioan eta une jakin batean kokatzen dira.

5. Azkenik, Deleuzek (1988) proposatutako definizioaren arabera gertaerak ez dira, gainerako teoriek defendatzen duten moduan, kausa baten ondorioz egoera jakin bat aldatzen duten espazioan eta denboran kokatutako jazoerak. Deleuzeren arabera, gertaerak ondoriotzat egoerak eguneratzea duten hainbat faktoreren elkarketak dira. Zuhaitzen udazkeneko hosto galera adibide hartuta, gainerako teoriek zuhaitzen hosto galera udazkenak eragindako gertaera dela ulertzen dute; Deleuzeren definizioaren arabera, aldiz, zuhaitzak hostorik gabe gelditzea (egoeraren eguneratzea) eragin duen udazkena (udazkena osatzen duten faktoreen elkarketa) izango litzateke gertaera.

Zerrendatutako teorietatik, hizkuntzaren prozesamenduak, historian zehar, Davidsonek proposatutakoa jarraitu izan du (2). Izan ere, semantika konputazionalean perpausak adierazteko teoria honetan oinarritzen den semantika neo-davidsondarra erabili ohi da. Beraz, hizkuntzaren prozesamenduan, eta ondorioz guretzat, gertaeraren definizioa ondorengoa izango da: denboran eta espazioan kokatua dagoen eta kausa jakin baten ondorioz eragin jakin bat sortzen duen jazoera.

1.1.1 Gertaeren kategorizazioa

Darabilgun gertaeraren definizioa kontuan edukita, gertaerek denborarekin eta espazioarekin duten lotura argia da. Hau dela eta, tesi lan honetan, gertaerak eta hauetan parte hartzen duten entitateak identifikatzeaz gainera, haien propietateak, bereziki denborari eta espazioari dagozkienak, aztertuko ditugu, testuen analisi semantiko egokia burutu ahal izateko. Gertaerek denborarekin duten erlazio semantikoaren kategorizazioa aldatzen duten gramatikaren kategoriak bat baino gehiago dira. Esate baterako *Aktionsart* (Streitberg, 1891) edo aspektu lexikala (iraunkorra, ez-iraunkorra, errepikakorra, ez-errepikakorra, etab.), aspektua (burutua edo burutu gabea), modua (indikatiboa, sub-

juntiboa, optatiboa, etab.) eta denbora gramatikala (orainaldia, lehenaldia, etorkizuna). Gertaerak izaera semantiko-espazialaren arabera kategorizatzeko orduan, berriz, mugimenduzko edo mugimendurik gabeko gertaeren arteko bereizketa egin ohi da. Hau da, jazoera batek espazioko kokaleku edo lekune batetik besterako mugimendua inplikatzen duen edo ez alegia. Muller-ek eta bestek (1998) zerrendatzen dituzten hamar mugimendu kategoriei so eginik kategorizazio espazial zehatzagoa lor badaiteke ere, tesian zehar, Pustejovskyk eta bestek (2010) bezala, gertaera estatikoen eta dinamikoen arteko bereizketa baizik ez dugu egingo. Hain zuzen ere, tesian zehar jarraituko duguna Sauriik eta bestek (2005) proposatutako gertaeren kategorizazio espazio-tenporala da. Hau lortzeko hainbat kategoria gramatikal (aspektua, *Aktionsart*, ebidentzialitatea) eta lexiko izan zituzten kontutan. Kategorizazio honek ondorengo zazpi gertaera motak bereizten ditu:

- `Occurrence`: Iraunkorrak eta burutuak diren gertaera dinamikoak, *ibili* adibidez.
- `State`: Egoerak deskribatzen dituzten gertaera iraunkor, burutu eta estatikoak, *egon* esaterako.
- `Reporting`: Beste gertaera baten berri ematen duten gertaera ez-iraunkor burutu eta estatikoak, *esan* konparaziorako.
- `Aspectual`: Beste gertaera baten hasiera, jarraitutasuna edo amaiera adierazten duten gertaera ez-iraunkor, burutu eta dinamikoak, *hasi* esate baterako.
- `Perception`: Beste gertaera bat zentzuen bitartez hautematea deskribatzen duten gertaera estatikoak, iraunkorrak edo ez-iraunkorrak, burutuak edo burutugabeak, *entzun* argibidez.
- `Intensional action`: Helburuek motibatutako ekintzak deskribatzen dituzten gertaera dinamikoak, iraunkorrak edo ez-iraunkorrak eta burutuak, *agindu* adibidez.
- `Intensional state`: Helburuek motibatutako egoerak deskribatzen dituzten gertaera estatiko, iraunkor, burutu edo burutugabeak, *pentsatu* esaterako.

	BURUTUA	BURUTUGABEA
IRAUNKORRA	**Occurrence** (*ibili*) State (*egon*) Perception (*entzun*) **Intensional action** (*agindu*) Intensional state (*pentsatu*)	Perception (*entzun*) Intensional state (*pentsatu*)
EZ-IRAUNKORRA	Reporting (*esan*) **Aspectual** (*hasi*) Perception (*entzun*) **Intensional action** (*agindu*)	Perception (*entzun*)

Taula 1.1: Gertaeren espazio-denborazko kategorizazioa (dinamikoak beltzez).

1.2 Predikatuak, argumentuak eta adjuntuak

Aurreko azpiatalean gertaera kontzeptua finkatu dugu. Azpiatal honetan, berriz, gertaeren gauzatze gramatikalean (Tenny eta Pustejovsky, 2000) parte hartzen duten *predikatu*, *argumentu* eta *adjuntu* kontzeptuak definituko ditugu. Guretzat predikatuak gertaerak deskribatzen dituzten aditzak, adberbioak, izenak eta adjektiboak dira. Argumentuak, bestalde, predikatuek deskribatzen dituzten gertaeretan parte hartzen duten entitateak (nor, nori, nork) dira. Azkenik, adjuntuak, gertaeren nolakotasunak (non, noiz, nola...) adierazten dituzten propietateak dira.

[Jarraitzaileek$_1$*]* *zelaitik*$_1$ ***ikusi zuten***$_1$ *[partida*$_1$*].*

[[Partida$_2$*]* *zelaitik*$_2$ ***ikusi zuten***$_2$ *[jarraitzaileak*$_2$*]]*$_1$ ***egongo dira***$_1$*[finalean*$_1$*].*

Adibideko lehenbiziko perpausean aditza eta aditz laguntzaileak osatutakoa da predikatua, *ikusi zuten*. *Jarraitzaileek* eta *partida* dira predikatuaren argumentuak eta *zelaitik*, berriz, *nondik (ikusi zuten)?* galderari erantzuna ematen dion predikatuaren adjuntua. Adibideko bigarren perpausean, berriz, bi predikatu daude: *ikusi zuten* eta *egongo dira*. Erlatibozko perpausa den *partida zelaitik ikusi zuten jarraitzaileek* eta *finalean*, *egongo dira* predikatuaren argumentuak dira. Erlatibozko perpausaren barnean, *partida* eta *jarraitzaileak*, bestalde, *ikusi zuten* predikatuaren argumentuak dira, eta *zelaitik* predikatuaren *nondik* galdera erantzuten duen adjuntua.

Uste dugu aurrera jarraitu aurretik garrantzizkoa dela argitzea hizkuntzaren prozesamenduan edo SRL atazan erabiltzen den predikatuaren definizioa eta teoria gramatikal *klasikoetan* ematen dena ez direla berdinak. Izan ere, hizkuntzaren prozesamenduaren

definizioa bat dator egungo teoria gramatikalek defendatzen duten ikuspegiarekin, predikatuen argumentuak eta adjuntuak ez baitira sekula predikatuaren zati izango. Teoria klasikoek, ordea, perpausek bi osagai gramatikal dauzkatela defendatzen dute: subjektua eta predikatua. Subjektua normalean izen sintagmari dagokio, eta predikatua aditz sintagmari. Bereizketa honen ondorioz, subjektua ez diren argumentuak eta adjuntuak predikatuaren barnean kokatzen dira.

[Jarraitzaileek$_1$] ***zelaitik$_1$ ikusi zuten$_1$ [partida$_1$].***

[[Partida$_2$] *zelaitik$_2$* ***ikusi zuten$_2$*** *[jarraitzaileak$_2$]]$_1$* ***egongo dira$_1$ [finalean$_1$].***

Adibideko lehenengo perpausean izen sintagma eta subjektua *jarraitzaileek* da; *zelaitik ikusi zuten partida*, berriz, aditz sintagma eta predikatua. Horren osagaiak *zelaitik* adjuntua, *ikusi* aditza, *zuten* aditz laguntzailea eta *partida* argumentua dira. Bigarren perpausean, *partida zelaitik ikusi zuten jarraitzaileak* erlatibozko perpausa perpaus nagusiaren izen sintagma eta subjektua da; *egongo dira finalean* aditz sintagma eta predikatua. Erlatibozko perpausean *jarraitzaileek* (ergatibo pluralean) da subjektua eta *Partida zelaitik ikusi zuten* predikatua.

1.3 Perpausen semantika: rolen garrantzia

Hizkuntzalaritzaren betidaniko arduretako bat perpausen adierazpen semantikoa izan da, perpaus baten esanahia nola adierazi behar den zehaztea alegia. Horretarako hurbilpen ezberdinak proposatu izan dira: logikoak, kognitiboak eta estatistikoak gehienbat. Perpausen esanahia egituratzen den era garrantzizkoa da, honek baldintzatzen baititu hizkuntza prozesatzeko erabiltzen diren tresnen diseinua eta garapena. Tesi lan honek jarraitzen duen Davidsonen gertaeran oinarritzen den adierazpen neo-davidsondarraz arituko gara atal honetan. Adierazpide hau perpausen semantika azaltzeko rol semantikoak erabiltzen dituen hurbilpen logikoa da. Rolek predikatu baten argumentu eta adjuntuek betetzen duten funtzio semantikoaren berri ematen dute, hau da, predikatuarekiko argumentu eta adjuntuek zer motatako erlazio semantikoa daukaten zehazten dute.

Hurrengo adibidean, *Jarraitzaileek zelaitik ikusi zuten partida* perpausaren semantika adierazi dugu adierazpide neo-davidsondarra erabilita. Goikoa semantikaren forma logikoa da, eta behekoa, berriz, forma logiko horren testuaren gaineko irudikapena.

$$\exists e(\textbf{ikusi}(e) \wedge \texttt{AGENT}(e, Jarraitzaileek) \wedge \texttt{THEME}(e, partida) \wedge \texttt{Location}(e, zelaitik))$$

$$\downarrow$$

$$\overbrace{[\ Jarraitzaileek_e\ ]}^{\textbf{AGENT}_e}\ \overbrace{\underline{zelaitik_e}}^{\texttt{Location}_e}\ \textbf{\textit{ikusi zuten}}_e\ \overbrace{[\ partida_e\ ]}^{\textbf{THEME}_e}.$$

Adierazpide neo-davidsondarrak argi erakusten du *ikusi zuten* predikatuak e izendatutako gertaera deskribatzen duela. Gainera, predikatu honen subjektu eta argumentu den *Jarraitzaileek* AGENT rola betetzen duela adierazten du, hau da, *Jarraitzaileak* direla e gertaeraren abiarazle edo sortzaileak. Honetaz gain, *partida* THEME rola betetzen duen argumentua dela adierazi da, *partida* dela e gertaera jasaten duen entitatea alegia. Azkenik, *zelaitik* predikatuaren adjuntuak gertaera jazo den kokagunearen berri, Location, ematen duela adierazi da.

1.3.1 Bilakaera: Julio Zesarren hilketa

Perpausen semantika adierazteko erabili izan diren erak aldatzen joan dira denborarekin. Jarraian, tesi lan honek oinarri hartzen duen semantika neo-davidsondarrera iritsi arteko teknika logikoen laburpen inkrementala egingo dugu (Goldberg, 1995). Izan ere, teknika horien gabeziek eraman dute egungo adierazpidera. *Brutok Zesar hil zuen* perpausaren semantika adieraztetik abiatuko gara. Gure lanean jarraitzen dugun predikatuaren definizioa kontuan edukita (egungo teoria gramatikaletan oinarritutakoa, eta predikatu logika erabiltzen duena), lehenengo hurbilpena perpausa honela adieraztea izango litzateke:

$$\textbf{\textit{hil}}(Bruto, Zesar)$$

$$\downarrow$$

$$[\ Brutok_1\ ]\ [\ Zesar_1\ ]\ \textbf{\textit{hil zuen}}_1.$$

Adibidean, *hil* predikatuak *Bruto* eta *Zesar* argumentuak hartzen dituela zuzen adierazten da; adierazpide hau hala ere ez da behar bezain zehatza, ez baitu, besteak beste, *nork* hil zuen *nor* zehazten. Gainera, eman dezagun adibideari *non* eta *noiz* galderei erantzuten dieten adjuntuak gehitzen dizkiogula, esate baterako: *Brutok Zesar Koliseoan urrian hil zuen.* Perpaus hau aurreko teknika bera erabiliz adierazita ondokoa izango genuke:

$$\textbf{\textit{hil}}(Bruto, Zesar, Koliseoa, urria)$$

$$\downarrow$$

$$[\ Brutok_1\ ]\ [\ Zesar_1\ ]\ [\ Koliseoan_1\ ]\ [\ urrian_1\ ]\ \textbf{\textit{hil zuen}}_1.$$

Adierazpen honek aditzera ematen du *hil* predikatuaren adjuntuak diren *Koliseoa* eta *urria Bruto* eta *Zesar* argumentuen parekoak direla, adjuntuak izan beharrean argumentuak direla, alegia. Ez da zehazten *nork* hil zuen *nor* eta, honetaz gain, ez da adjuntu moten arteko bereizketarik egiten, hau da, ez da zehazten *Koliseoa* kokapena eta *urria* denbora-adjuntuak direla. Arazo hauek saihesteko aurreko adierazpena honela zati genezake:

$$\textbf{\textit{hil}}(Bruto, Zesar) \wedge non(Koliseoa) \wedge noiz(urria)$$

$$\downarrow$$

$$[\ Brutok_1\ ]\ [\ Zesar_1\ ]\ \underline{Koliseoan_1}\ \underline{urrian_1}\ \textbf{\textit{hil zuen}}_1.$$

Teknika honen bitartez aurreko adierazpidearen hiru arazoetako bi desagertu egiten dira: predikatuaren argumentuen eta adjuntuen arteko bereizketa egiten da eta, gainera, argi geratzen da *Koliseoan* eta *urrian* adjuntuak ez direla mota berekoak. Hala ere, adierazpide hau ez da perpausen semantika egoki adierazteko bezain zehatza. Demagun ondorengo perpausa daukagula: *Brutok Zesar Koliseoan urrian hil eta Grezian atxilotu zuten.*

$$\textbf{\textit{hil}}(Bruto, Zesar) \wedge non(Koliseoa) \wedge noiz(urria) \wedge non(Grezian) \wedge \textbf{\textit{atxilotu}}(Bruto)$$

$$\downarrow$$

$$[Brutok_?][Zesar_?]\ \underline{Koliseoan_?}\ \underline{urrian_?}\ \textbf{\textit{hil}}_1\ \ eta\ \underline{Grezian_?}\ \textbf{\textit{atxilotu zuten}}_2.$$

Adibidean *hil* eta *atxilotu zuten* predikatuak zuzen adierazita daude: argumentu eta adjuntuak bereizten dira, eta gainera adjuntu motak zehaztuta daude. Hala ere, hasierako arazoa ez da konpondu, ez da zehazten *nork* hil zuen *nor*. Gainera, ez da argi gelditzen adjuntu bakoitza zer predikaturi dagokion; ezin esan dezakegu, adibidez, *Grezian* adjuntua *atxilotu zuten* predikatuari dagokiola. Arazo hau izan zen, hain zuzen ere, Davidsonen abiapuntua, eta haren perpausen semantika adierazteko proposamenaren

abiarazlea. Davidsonek (1967), bere lanean, perpaus bakoitzeko predikatuak predikatuek berek deskribatzen dituzten gertaeren bitartez erazagutzea proposatu zuen. Honek argumentu eta adjuntu bakoitza dagokion gertaera edo predikatuarekin lotzea ahalbidetzen du.

$$\exists e_1(\textbf{\textit{hil}}(e_1, Bruto, Zesar) \wedge non(e_1, Koliseoa) \wedge noiz(e_1, urria)) \wedge$$
$$\exists e_2(\textbf{\textit{atxilotu}}(e_2, Bruto) \wedge non(e_2, Grezia))$$

$$\downarrow$$

$$[Brutok_{e_1,e_2}][Zesar_{e_1}]\underline{Koliseoan_{e_1}}\ \underline{urrian_{e_1}}\ \textbf{\textit{hil}}_{e_1}\ eta\ \underline{Grezian_{e_2}}\ \textbf{\textit{atxilotu zuten}}_{e_2}$$

Adierazpide davidsondarra jarraituta argi geratzen da *hil* predikatuaren adjuntuak *Koliseoan* eta *urrian* direla eta *atxilotu zuten* predikatuaren adjuntua *Grezian* dela. Davidsonen proposamenak, beraz, aipatutako arazo eta zehaztasun falta guztiak konpontzea lortzen du, bat izan ezik: *nork* hil zuen *nor* finkatzea. Hau konpontzeko, Parsonsek (1990) Davidsonen proposamena eguneratu zuen, haren adierazpen teknikari rol semantikoak gehituta; teknika honi semantikaren adierazpide neo-davidsondarra deritzo.

$$\exists e_1(\textbf{hil}(e_1) \wedge \texttt{AGENT}(e_1, Bruto) \wedge \texttt{PATIENT}(e_1, Zesar) \wedge$$
$$\texttt{Location}(e_1, Koliseoa) \wedge \texttt{Time}(e_1, urria)) \wedge \exists e_2(\textbf{atxilotu}(e_2) \wedge$$
$$\texttt{PATIENT}(e_2, Bruto) \wedge \texttt{Location}(e_2, Grezia))$$

$$\downarrow$$

$$\overbrace{[\ Brutok_{e_1,e_2}\ ]}^{\substack{\texttt{AGENT}(e_1)\\ \texttt{PATIENT}(e_2)}}\ \overbrace{[\ Zesar_{e_1}\ ]}^{\textbf{PATIENT}_{e_1}}\ \overbrace{\underline{Koliseoan_{e_1}}}^{\texttt{Location}_{e_1}}\ \overbrace{\underline{urrian_{e_1}}}^{\texttt{Time}_{e_1}}\ \textbf{\textit{hil}}_{e_1}\ eta\ \overbrace{\underline{Grezian_{e_2}}}^{\texttt{Location}_{e_2}}$$
$$\textbf{\textit{atxilotu zuten}}_{e_2}.$$

Adierazpide neo-davidsondarraren bitartez argi geratzen da zein den perpausaren semantika, rolek argumentu bakoitzaren eginkizuna predikatuek deskribatzen dituzten gertaeretan adierazten baitute. Adibidean, *Bruto*, *hil* eta *atxilotu zuten* predikatuen argumentua da, lehenbizikoan AGENT rola betetzen du eta bigarrenean PATIENT rola. *Zesar*, *hil* predikatuaren argumentua baizik ez da, eta PATIENT rola betetzen du, AGENTek abiatutako edo eragindako jazoera jasaten du, alegia.

Tesi lan honetan erabili eta garatu ditugun tresnek formalismo edo semantika neo-davidsondarra jarraitzen dute, hau da:

- Predikatu batek Davidsonen gertaera bat eta bakarra deskribatzen du eta predikatua aditza, aditzondoa, izena edo izenondoa izan daiteke.
- Perpaus bakoitzak gutxienez predikatu bat dauka, eta horrek hainbat argumentu eta adjuntu izan ditzake edo bat ere ez.
- Gertaera bakoitzari *predikatu-argumentu-adjuntu* egitura bat dagokio.
- Argumentuek rolak betetzen dituzte eta argumentu bera hainbat predikaturena izan daiteke
- Adjuntuek propietateak adierazten dituzte eta adjuntu bera hainbat predikaturena izan daiteke.

1.4 Rol semantikoen etiketatzea

Gure ikerketa hizkuntzaren prozesamenduaren azpiatala den semantika konputazionalaren barnean kokatzen da. Semantika konputazionala hizkuntza naturalean idatzitako testuen esanahiaren adierazpenak sortu eta prozesatzeaz arduratzen da. Besteak beste, testuetako *predikatu-argumentu-adjuntu* egiturak era automatizatuan detektatzeko ardura bere gain hartzen duen hizkuntzaren prozesamenduaren azpiatala da. Rol semantikoen etiketatzearen (Semantic Role Labeling, SRL) eginbeharra perpaus bateko predikatu, argumentu eta adjuntuen arteko erlazio semantikoak aurkitu eta dagokien rol semantikoa automatikoki esleitzea da. Horretarako lehenengo urratsa perpauseko predikatuak identifikatzea izaten da (1). Ondoren, predikatu bakoitzaren adiera finkatu behar izaten da, adieraren arabera predikatuek hartzen dituzten argumentuak eta horien rolak aldatu egiten direlako (2). Azkenik, predikatuen argumentu eta adjuntuak identifikatzen dira, eta argumentu bakoitzak zer rol jokatzen duen erabakitzen da, aurreko pausoan zehaztutako adiera kontuan izanik (3 eta 4). Analisi mota honek perpausen azaleko adierazpen semantikoa ahalbidetzen du, eta gertaeren propietateak eta parte hartzaileen arteko oinarrizko erlazio semantikoak azaleratzen ditu.

Jarraitzaileek zelaitik ***ikusi zuten***$_e$ *partida.* (1)

↓

$$\textit{Jarraitzaileek zelaitik } \overbrace{\textbf{\textit{ikusi zuten}}_e}^{\textit{ikusi.01}} \textit{ partida.} \tag{2}$$

$$\downarrow$$

$$\textit{[Jarraitzaileek}_e \textit{] } \underline{\textit{zelaitik}_e} \overbrace{\textbf{\textit{ikusi zuten}}_e}^{\textit{ikusi.01}} \textit{ [partida}_e \textit{].} \tag{3}$$

$$\downarrow$$

$$\overbrace{\textbf{AGENT}_e}^{\textit{[Jarraitzaileek}_e \textit{]}} \quad \overbrace{\texttt{Location}_e}^{\underline{\textit{zelaitik}_e}} \quad \overbrace{\textbf{\textit{ikusi zuten}}_e}^{\textit{ikusi.01}} \quad \overbrace{\textbf{THEME}_e}^{\textit{[partida}_e \textit{]}}. \tag{4}$$

Adibidean *Jarraitzaileek zelaitik ikusi zuten partida* perpausaren rol semantikoen etiketatzea pausoz pauso burutu da. Lehenik, *ikusi zuten*, perpausaren predikatu bakarra, identifikatu da. Gero, predikatuari *ikusi.01* izendatutako *ikusi* aditzaren adiera ezarri zaio. *ikusi.01* adiera *norbaitek (Jarraitzaileek) zerbait (partida) ikusi* du(t)ela adierazten duen *ikusi* aditzaren erabilera da. Azkenik, *ikusi.01* adieraren azpikategorizazioan oinarrituta, *Jarraitzaileek* eta *partida* argumentuak eta *zelaitik* adjuntua identifikatu dira, eta AGENT eta THEME rolak eta Location adjuntu-mota esleitu zaizkie. Predikatuen adieren azpikategorizazioaren helburua predikatuek adiera bakoitzean jasotzen ahal dituzten argumentuak eta hauek jokatzen dituzten rolak finkatzea da; honetaz tesi laneko bigarren atalean arituko gara.

1.4.1 Sintaxia eta SRL

Rol semantikoen etiketatze automatikoa estuki lotuta dago analisi sintaktikoarekin, eta hasiera batean analisi sintaktikoa esaldien adierazpen zehatza egiteko nahikoa dela pentsa badaiteke ere, hau ez da horrela. Ondoko adibidean ikus daiteke prozesamendu semantikoaren garrantzia zein den.

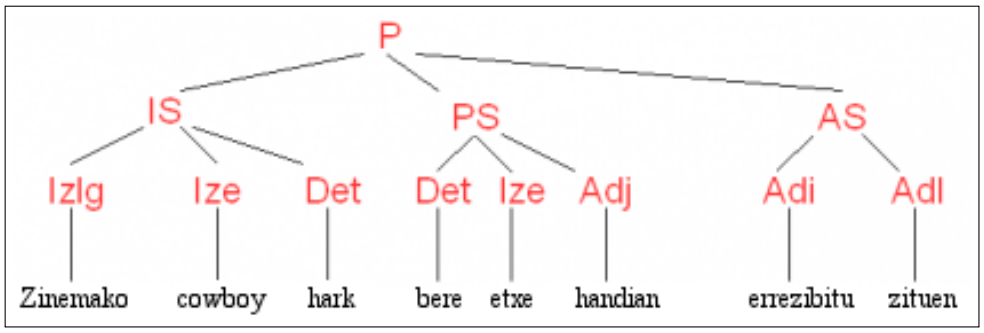
$$\overbrace{\textit{[Mikelek}_e \textit{]}}^{\substack{\textbf{AGENT}\\ \textbf{(Subjektua)}}} \overbrace{\textit{[leihoa}_e \textit{]}}^{\substack{\textbf{PATIENT}\\ \textbf{(Objektua)}}} \textbf{\textit{hautsi zuen}}_e.$$

$$\overbrace{\textit{[Leihoa}_e \textit{]}}^{\substack{\textbf{PATIENT}\\ \textbf{(Subjektua)}}} \textbf{\textit{hautsi egin zen}}_e.$$

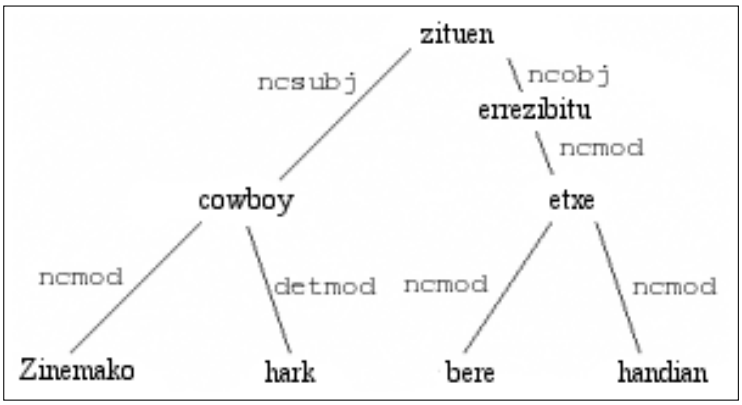
Adibideko lehenengo esaldian *Mikelek* subjektua da eta *leihoa* berriz objektua. Bigarren esaldian, aldiz, *leihoa* subjektua da eta ez objektua. Bi esaldietan, alabaina, *leihoa*

argumentuak rol bera jokatzen du, PATIENT rola. Adibide honek argi erakusten du analisi sintaktikoa ez dela nahikoa esaldien esanahia zehaztasunez adierazteko eta semantika behar beharrezkoa dela.

Aurrera jarraitu aurretik bi puntu argitzea beharrezkoa dela uste dugu: alde batetik, eta arestian aipatutakoak iradokitzen duen moduan, SRL sistemak analisi sintaktikoetan oinarritzen direla rol semantikoak etiketatzeko, eta beste alde batetik, bi analisi sintaktiko mota daudela, osagaietan oinarritutakoa eta dependentzietan oinarritutakoa.



Irudia 1.1: Osagaietan oinarritutako analisi sintaktikoa.



Irudia 1.2: Dependentzietan oinarritutako analisi sintaktikoa[1].

Osagaietan oinarritutako analisi sintaktikoak kategoria sintaktikoen arabera antolatzen ditu esaldiko osagaiak (perpausa (P), izen-sintagma (IS), postposizio-sintagma (PS), eta aditz-sintagma (AS))(1.1). Dependentzietan oinarritutakoak, berriz, elementu lexikoen arabera eratzen ditu, hau da, hitzen arteko erlazioak adierazten dituen

[1]Iturria: Sareko Euskal Gramatika, http://www.ehu.eus/seg/hizk/1/4.

dependentzia-zuhaitza itzultzen du. Dependentzia zuhaitzetan hitz-bikoteen arteko erlazio bitarrak (mendekoa/gobernatzailea) adierazten dira (1.2). Irudietan *Zinemako cowboy hark bere etxe handian errezibitu zituen.* esaldiaren analisi sintaktikoa erakusten da.

1.4.2 SRLren jatorria eta bilakaera

Hizkuntzaren prozesamenduak lengoaia naturala konputazionalki modelatzeko gaitasuna inplikatzen du, haren kategoria gramatikal, topiko eta fenomeno guztiak barne. Oraingoz urrun gaude hizkuntza konputazionalki guztiz modelatu ahal izatetik, nahiz eta, berrogeita hamarreko hamarkadan alor honetako ikerketa hasi zenetik, hainbat gai landu diren eta horietako askotan emaitza onak erdietsi. Jarraian SRLren bilakaera deskribatzen da, horretarako beharrezkoa izan da hizkuntzaren prozesamenduarena ere zeharka azaltzea.

1950-1990

Aipatu dugu lengoaia naturala automatikoki prozesatzeko lehenengo saiakerak berrogeita hamarreko hamarkadan egin zirela, garaiko interes geopolitikoaren ondorioz adimen artifizialak zuen babes ekonomikoak bermatuta. Izan ere, aurrera eramandako lehenbiziko esperimentuak, *The Georgetown Experiment* deitutakoak, errusieraz idatzitako 60 esaldi ingelesera automatikoki itzultzea zeukan helburu. Ordutik eta laurogeita hamarreko hamarkada arte, hizkuntza, eta zehatzago, hizkuntzaren semantika, konputazionalki tratatzean zentratu ziren lanak, batez ere lexikoien, ontologien, gramatiken eta bestelako baliabide semantikoen eskuzko garapenean (adibidez, Hirst, 1987).

1990-2005

Laurogeita hamarreko hamarkadatik aurrera, semantika konputazionalean, hizkuntzaren prozesamenduko gainerako arloetan bezala, ikasketa automatikoa erabiltzen hasi zen. Baliabide konputazionalen ahalmena handitu izanak ekarri zuen ikasketa automatikoa hizkuntza teknologian sartzea. Hasierako ikasketa metodoek teknika gainbegiratuak erabiltzen zituzten, hau da, eskuz etiketatutako corpusetatik ikasten zuten, ereduak sortzeko. Geroago, teknika ez gainbegiratuak, etiketatu gabeko testutik ikasten zutenak, eta teknika hibridoak hasi ziren erabiltzen.

Ikasketa automatikoko metodoek hizkuntza prozesatzeko sistemei egitura linguistiko konplexuak ikasteko eta kudeatzeko gaitasuna ematen die; besteak beste, lan honen ardu-

rakoak diren *predikatu-argumentu-adjuntu* egitura gramatikal eta sintaktiko-semantikoak ikasteko eta kudeatzeko. Hasiera batean, ikasketa automatikoaren eta metodo estatistikoen erabilera morfologiaren eta sintaxiaren prozesamendura mugatu bazen ere, ataza haietan lortutako emaitza onek semantika ere aipatu tekniken bitartez prozesatzen hastea ekarri zuten. Laurogeita hamarreko hamarkadaren bukaeran hasi ziren SRL atazaren oinarri teorikoak finkatu zituzten lanak argitaratzen. Esate baterako, Briscoe eta Carroll-ek (1997) azpikategorizazio egiturak automatikoki erauzteko egin zuten lana, eta Waldek (2000) eta Stevensonek (2001) ikasketa automatikokoak *predikatu-argumentu-adjuntu* egiturak landu nahi badira metodo egokiak zirela erakusteko burututakoak.

2005-2017

Ondorengo urteetan, SRLn erreferentziazkoak diren eta *predikatu-argumentu-adjuntu* egiturak etiketatuta dauzkaten *PropBank* (Palmer et al., 2005) eta *FrameNet* (Baker et al., 1998) ingelesezko corpusak argitaratu ziren. Corpus hauek ingelesezko testuz osatuta baldin badaude ere, duten egitura eta oinarritzat hartzen duten marko teorikoa beste hizkuntza batzuetarako corpusak garatzeko erabili izan da; adibidez, Aldezabalen eta besteren 2013ko lanean euskarazko *EPEC-RolSem* corpusa deskribatzen da.

Gaur egun, rol semantikoen etiketatzea hizkuntzaren prozesamenduko gai ongi definitua da. Azken bost urteotan, alabaina, atazaren ospea apur bat gutxitu da, ingelesez idatzitako testuak rolekin etiketatzeko orduan emaitza onak eskuratu badira ere, tipologikoki ingelesetik urruntzen direnetan emaitzak ez baitira hain onak izan. Erabiltzen diren teknikei dagokionez, gaur egun, baliabide urrikoak ez diren hizkuntzak prozesatzeko (ingelesa esate baterako) gainbegiratu gabeko tekniken eta sare neuronaletan oinarritutako ikasketa *sakonaren* erabilera gailentzen ari da. Baliabide urriak edo baliabide mugatuak dituzten hizkuntzetan, ordea (euskaran adibidez), teknika gainbegiratuak eta ikasketa automatiko *klasikoa* (ez sakona, *azalekoa* baizik) erabiltzen dira gehienbat.

1.5 Denbora eta espazioaren etiketatzea

Testuen semantika egituratzen duten oinarrizko unitateak *predikatu-argumentu-adjuntu* egiturei lotutako gertaerak direla azaldu dugu 1.2 atalean. Argumentu eta adjuntuek gertaeren hainbat propietateren berri ematen dute; besteak beste, gertaerak espazioan eta

denboran kokatzen laguntzen dute. Rol semantikoen etiketatzeak propietate hauek detektatzeko gaitasuna baldin badu ere, uste dugu aberasgarria izan daitekeela gertaerek beste gertaera eta entitateekin dituzten espazio eta denborazko erlazioak automatikoki lortu ahal izatea, baita espazio eta denbora adierazpenak erauzi ahal izatea ere. Horretarako *ISO-Space* (Pustejovsky et al., 2011) eta *ISO-TimeML* (Pustejovsky et al., 2010) estandarretan oinarritutako etiketatzaileak garatu ditugu, izan ere, tesi honen helburuetako bat euskarazko testuen analisi semantikoa automatikoki egiten duten tresnen garapena baita. *ISO-TimeML* testuetako denbora-informazioa etiketatzeko sortutako anotazio eskema eta hizkuntza da; *ISO-Space*, berriz, espazio-informazioa etiketatzeko sortutako anotazio eskema eta hizkuntza, aurreko eskemaren diseinuan oinarritzen da. Anotazio eskemak *hizkuntza naturaleko informazio linguistikoa nola markatu edo bildu behar den ezartzen duten formalismoak dira.* Markaketaren helburua testuetako informazio linguistikoa konputazionalki tratatu edo prozesatzeko gaitasuna eskuratzea da. Gure kasuan, *ISO-TimeML* eta *ISO-Space* eskemak erabili ditugu, denbora eta espazio informazioa erauzten duten tresnak garatu ahal izateko.

1.5.1 Denbora

Informazio tenporala automatikoki prozesatzeko erabiltzen den *ISO-TimeML* eskemak testuetan denborarekin zerikusia duten ondoko osagaiak erauzteko aukera ematen du:

- *Joan*, *pentsatu* eta *dakusat* moduko gertaeren buru lexikalak.
- *Urtarrilaren 31n*, *1923/4/23* eta *ostiral arratsaldean* bezalako adierazpenak.
- *Ondoren*, *baino lehen* eta *aurretik* moduko denbora seinaleak. Seinale hauek gertaeren eta denbora adierazpenen arteko loturak adierazten dituzte, haien artean zer ordenaketa dagoen erakusten dute alegia. *Joan aurretik pentsatu nuen* esaldian, esaterako, *aurretik* seinaleak *pentsatzeko* gertaera *joatekoa* baino lehen jazo dela adierazten du.

Ondoko adibideak erakusten du testua nola etiketatzeko balio dezakeen *ISO-TimeML* eskemak. Adibideak arestian aipatu ditugun gertaeren buru lexikalak, denbora adierazpenak eta seinaleak dauzka markatuta.

Jarraitzaileek $\overbrace{\textbf{\textit{ostiral goizean}}}^{\textit{Adierazpena}}$ $\overbrace{\textbf{\textit{ikusi}}}^{\textit{Gert.}}$ *zuten partida,* $\overbrace{\textbf{\textit{ekaitzaren}}}^{\textit{Gertaera}}$ $\overbrace{\textbf{\textit{ondoren}}}^{\textit{Seinalea}}$.

ISO-TimeML eskemak, gainera, bi gertaeraren artean eta gertaera baten eta denbora adierazpen baten artean izan daitezkeen hiru erlazio mota markatzeko aukera ere ematen du. Lotura hauek denborazkoak, mendekotasunezkoak eta aspektuzkoak izan daitezke. Aurreko adibideko esaldian etiketa daitezkeen denbora loturak hiru dira.

(1) *Jarraitzaileek ostiral goizean* ***ikusi*** *zuten partida,* *ekaitzaren* *[ondoren].*

Lehenengo loturaren kasuan *ondoren* seinaleak adierazten du *ikusi* eta *ekaitzaren* gertaeren arteko denbora erlazioaren gauzatzea. Honek, *ekaitzaren* gertaera *ikusi* baino lehen jazo dela markatzen du.

(2) *Jarraitzaileek* *ostiral goizean* ***ikusi*** *zuten partida, ekaitzaren ondoren.*

Bigarren loturan, ordea, ez dago *ostiral goizean* denbora adierazpenaren eta *ikusi* gertaeraren arteko erlazioaren gauzatzea adierazten duen seinalerik. Erlazio honek *ikusi* gertaera eta *ostiral goizean* adierazpena aldi berean agitu direla markatzen du.

(3) *Jarraitzaileek* ***ostiral goizean*** *ikusi zuten partida,* *ekaitzaren* *ondoren.*

Azkenik, hirugarren denbora erlazioan, *ekaitzaren* gertaera *ostiral goizean* adierazpenak mugatzen duen denbora tartearen barnean kokatzen dela markatzen du.

Ondoko bi adibideetan beste bi erlazio motak, aspektuzkoak eta mendekotasunezkoak, nolakoak diren erakusten da. Lehenengo motakoek aspektuzko gertaera baten (ikus 1.1.1) eta honen argumentuen arteko erlazioak markatzeko balio dute. Bigarren motakoek, aldiz, bi gertaeraren arteko mendekotasun erlazioak markatzeko.

(4) *Jarraitzaileek partida* ***ikusten*** *amaitu zuten.*

Erlazio honetan, *amaitu*, *amaitu zuten* aspektuzko gertaeraren buru lexikaletik *ikusten* gertaerarako erlazio aspektuala markatu da.

(5) *Jarraitzaileek* ***uste*** *dute taldeak partida irabazteko aukera izango duela.*

Beste adibide honek, ordea, mendekotasun erlazioen markaketa erakusten du. Erlazio hauek gertaera nagusia hartzen dute iturritzat, kasu honetan *uste dute*, eta helburutzat berriz mendeko gertaera, kasu honetan *aukera izango dutela*.

Azpiatal honetan *ISO-TimeML* eskema erabilita etiketa daitekeen informazio tenporala deskribatu dugu. Tesi laneko hirugarren atalean zehaztuko ditugu zein diren eskema honek erabiltzen dituen etiketak, etiketa hauek jasotzen dituzten atributuak eta atributu horiek har ditzaketen balioak.

1.5.2 Espazioa

Informazio espaziala automatikoki prozesatzeko erabiltzen den tesi lan honetako *ISO-Space* eskemak testuetan espazioarekin zerikusia duten ondoko osagaiak erauzteko aukera ematen du:

- Lekuak, *eraikina*, *Tokio* eta *Everest mendia* adibidez.
- Bideak, *errepidea*, *A-8 autobidea* eta *ibilbidea* esaterako.
- Gertaera dinamikoak, *joan*, *ibili* eta *erori* konparaziorako.
- Gertaera dinamiko edo mugimenduzkoen argumentuak.
- Gertaera estatiko edo mugimendu gabekoak, *egon*, *pentsatu* eta *ikusi* adibidez.
- Espazio seinaleak, *ezkerrekoan*, *barnean*, *kanpokoan* eta *ezkerretan* argibidez. Seinale hauek osagaien arteko espazio eta norabide erlazioak deriztenak adierazten dituzte.
- Mugimendu seinaleak, *ezkerrekora*, *barnetik*, *kanporantz* eta *ezkerretara* esate baterako. Seinale hauek osagaien arteko mugimendu erlazioak deriztenak adierazten dituzte.

Hurrengo adibideak erakusten du testua nola etiketatzen den *ISO-Space* eskema erabilita. Adibideak arestian aipatu ditugun osagai espazialak dauzka markatuta.

Goiko harmailetan zeuden jarraitzaileek B1 irtenbidea erditik igaro zuten.

Lekua → *harmailetan*
Lekua → *Goiko harmailetan*
Bidea → *irtenbidea*
Bidea → *B1 irtenbidea*
Gertaera dinamikoa → *igaro*
Arg. dinamikoa → *Goiko harmailetan zeuden jarraitzaileek*
Arg. dinamikoa → *B1 irtenbidea*
Gertaera estatikoa → *zeuden*
Seinale espaziala → *Goiko*
Mugimendu seinalea → *erditik*

Adibidean bi gertaera etiketatu dira: *zeuden* estatikoa eta *igaro* dinamikoa. Goian azaldu bezala, *ISO-Space* eskemak mugimenduko gertaeren argumentuak ere markatzen ditu. Kasu honetan *igaro*ren *Goiko harmailetan zeuden jarraitzaileek* eta *B1 irtenbidea* argumentuak markatu dira. Gainera, *Goiko* seinale espaziala eta *erditik* mugimendu seinalea ere etiketatu dira. Baita *harmailetan* eta *Goiko harmailetan* toki izenak eta *irtenbidea* eta *B1 irtenbidea* ibilbide edo bideak ere.

Espazioa etiketatzeko erabiltzen den eskemak, gainera, osagai espazialen artean izan daitezkeen hiru erlazio mota markatzeko aukera ere ematen duela aipatu dugu. Lotura hauek mugimendukoak, espazialak eta norabidekoak izan daitezke. Ondoko esaldietan ematen dira erlazio hauen adibideak.

(1) [*Harmailetako jarraitzaileek*] [*B1 irtenbidea*] *erditik* ***igaro*** *zuten.*

Lehenbiziko adibidean *erditik* seinale dinamikoak adieraz ten duen mugimenduko erlazioa etiketatu da. Honetan *igaro*, *igaro zuten* gertaera dinamikoaren buru lexikalak *Harmailetako jarraitzaileek* argumentuak eta *B1 irtenbidea* bideak hartzen dute parte.

(2) [*Jarraitzaileak*] [*estadioaren*] *barnean* ***zeuden.***

Bigarren adibide honek, berriz, espazio erlazioen erabilera erakusten du. Lotura *barnean* seinale espazialak adierazten du, *zeuden* gertaera estatikoa *Jarraitzaileak* argumentuarekin eta *estadioaren* tokiarekin erlazionatzen dituenak.

(3) [*Jarraitzaileak*] [*estadioaren*] *ezkerretan* ***zeuden***.

Hirugarren adibidean norabideko erlazioen erabilera erakusten da. Bertan, *ezkerretan* seinale espazialak *Jarraitzaileak* argumentuaren eta *estadioaren* tokiaren arteko orientazio erlazioa adierazten du.

Azpiatal honetan *ISO-Space* eskema erabilita etiketa daitekeen informazio espaziala deskribatu dugu. Tesi laneko laugarren atalean zehaztuko ditugu zein diren eskema honek erabiltzen dituen etiketak, etiketa hauek jasotzen dituzten atributuak eta atributu horiek har ditzaketen balioak.

1.6 Motibazioa

Tesi lan honetan zehar jarraian zerrendatzen ditugun bi hipotesiak betetzen diren ala ez aztertzea izan dugu helburu:

1. Euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek daukaten eragina positiboa dela, ingelesez eta gaztelaniaz bezala.

2. Espazioaren adierazpen linguistikoa, denborarena bezala, fenomeno semantikoa dela, eta horregatik semantika eta, zehazkiago, rol semantikoek duten garrantzia nabarmena dela, informazio espazialaren etiketatze eraginkorra egin ahal izateko.

Gainera, bi hipotesi hauek betetzen diren edo ez aztertzeaz gain, euskara semantikoki prozesatu ahal izateko tresnak garatzea ere izan da tesi lan honen motibazioa. Prozesamendu espazialaren kasuan ezin izan da euskarazko tresna sortu baliabide faltaren ondorioz. Dena dela, ingeleserako garatu dugun *X-Space* izeneko tresna, honen egitura eta honetatik atera ahal izan ditugun ondorioak baliozkoak izango zaizkigu etorkizunean garatu gogo den euskarako sistema sortzean.

Gaur egun frogatuta dago rol semantikoek garrantzi handia dutela, beste ezaugarri semantiko batzuekin batera, denboraren adierazpen linguistikoa egoki etiketatzeko, bai

ingelesez eta bai gaztelaniaz (Llorens, 2011). Ondorioz, pentsatzekoa da hau beste hizkuntza batzuetan ere beteko dela. Tesi honetan euskaran ere betetzen dela egiaztatu dugu (lehen hipotesia). Gainera, denboraren adierazpen linguistikoak antzekotasun handia du espazioaren hizkuntza adierazpenarekin. Izan ere, leku informazioa etiketatzeko gailendu den eskema (*ISO-Space*) (Pustejovsky et al., 2011) informazio tenporala etiketatzeko erabiltzen den eskema estandarrean dago oinarrituta (*ISO-TimeML*) (Pustejovsky et al., 2010). Arrazoi horrengatik, besteak beste, da zentzuzkoa espazioaren adierazpen linguistikoa, denborarena bezala, fenomeno semantikoa dela pentsatzea eta rol semantikoek ataza honetan duten eragina neurtzen saiatzea (bigarren hipotesia).

1.6.1 Metodologia

Lehenengo hipotesia frogatzeari dagokionez lehenik eta behin, euskaraz rol semantikoak automatikoki etiketatzen dituen *bRol* tresna (Salaberri et al., 2015a) garatu dugu, eta gero honen emaitzak testuinguru estandarrean balioetsitako beste sei hizkuntzarekin alderatu ditugu. *bRol* SRL tresna garatu aurretik, euskaraz rol semantikoak etiketatzeko tenorean eragin positiboa izan dezaketen ezaugarrien ikerketa eta lehen prototipoaren eraikuntza egin dugu (Salaberri et al., 2014). Rol semantikoak etiketatzeko atazan, ebaluazio estandarra *ConLL-09* (Hajič et al., 2009) lehiaketak ezarritako testuingurua dela ulertzen da. Hurrengo pausua *bTime* (Salaberri et al., 2017) euskaraz idatzitako testuetan denbora adierazpenak etiketatzen dituen sistema sortzea izan da. *bTime*ren emaitzak ingeles, gaztelania eta italierarako erreferentziazkoak diren *TempEval-3* (UzZaman et al., 2012) eta *EVENTI-2014* (Caselli et al., 2014) ebaluazio saioetan lortu zirenekin alderatu ditugu. Honek tesi lan honetako lehenengo hipotesia baieztatzeko balio izan digu.

Gure bigarren hipotesia frogatzeko, ordea, *X-Space* sistema garatu dugu (Salaberri et al., 2015b); honek ingeleseko testuetan ageri den informazio espaziala automatikoki etiketatzen du, besteak beste rol semantikoetatik erauzitako ezaugarriak ikasketa automatikoan erabiliz. *X-Space* aplikazioa *SemEval-2015* lehiaketan (Pustejovsky et al., 2015) parte hartuta ebaluatu dugu eta, ondorioz, testuinguru estandarrean ebaluatu dugula esan daiteke. Bertan lortutako emaitzek gure hipotesia baieztatzen laguntzen dute.

Tesi lanean garatu diren tresnak Euskal Herriko Unibertsitateko *IXA* ikerketa taldearen barnean erabiltzeko sortu dira gehienbat. Hala ere, etorkizun hurbilean tresna horiek

eskuragarri egongo dira edonorentzat *ixaKAT*[2] prozesamendu-katean (*bRol* eskuragarri dago dagoeneko).

1.7 Ekarpenak

Oro har, eta besteak beste hipotesiak frogatzeko asmoz garatu ditugun lau aplikazioak kontuan edukita, tesi lan honen ekarpen nagusiak bi izan direla esan dezakegu: alde batetik, aipatutako hipotesiak berrestea lortu dugu, eta, bestetik, ikerketaren egungo egoeran aurrera egin dugu, ingelesean eta batik bat euskaran, azken hau prozesatzeko dagoen baliabide eta tresna kopurua ingelesa prozesatzeko dagoenera hurbilduta. Gainera, *ISO-TimeML* eta *ISO-Space* eskema jarraituta sortutako etiketatze-fitxategien bisualizaziorako teknika proposatu dugu, eta *VisualTime* eta *VisualSpace* izeneko interfazeak garatu. Tesian egin ditugun hiru urratsak banaka azterturik (SRL eta denboraren eta espazioaren etiketatzea), haietako bakoitzetik sortutako ekarpenak ondorengoak dira:

- **SRL**. Euskaraz rol semantikoak etiketatzen dituen lehenengo tresna garatu dugu, *bRol* izenekoa, eta IXA taldearen euskararen analisi katean txertatu dugu (Otegi et al., 2016). Hortaz, testuen analisirako eta informazioaren erauzketarako ezagutza semantikoaz balia daitezkeen aplikazioen garapena ahalbidetu dugu, esaterako tesian landu dugun denbora-informazio erauzketarako tresnarena (*bTime*).

- **Denboraren etiketatzea**. Atal honetan *ISO-TimeML* estandarrean oinarritzen den *bTime* euskararako denbora informazioaren erauzketa automatikorako tresna sortu dugu. Honen garapenean denbora adierazpenak etiketatzeko dauden bi hurbilpenen erkaketa egin dugu, hots, erregelen bidezkoa (*bTime-rule*) eta ikasketa automatikoaren bidezkoa (*bTime-ML*).

- **Espazioaren etiketatzea**. Atal honetarako ekarpen garrantzizkoena ingelesez idatzitako testuetan aurki daitekeen espazio informazioa etiketatzeko inplementatu dugun *X-Space* tresna da. Etiketatzaile hau *ISO-Space* estandarra jarraitzen duen tresna bakanetakoa da. Izan ere, 2015. urtean *SpaceEval* lehiaketan (Pustejovsky et al., 2015) aurkeztutako sistemetatik (hauek dira mota honetako lehenak), gurea

[2] http://ixa2.si.ehu.es/ixakat/

da bi osatuenetako bat. *X-Space* sisteman proposatzen dugun aplikazioaren diseinua, etorkizunean, euskaraz espazio informazioa etiketatzen duen sistema eraikitzeko orduan baliagarria izango zaigu.

1.8 Dokumentuaren egitura

Tesiaren dokumentua bost ataletan banatu dugu eta horietan jasotzen da egindako ikerkuntza lana. Bost kapituluetan burututakoa jarraian laburbiltzen da:

- **Lehen atala**: *Sarrera.*
- **Bigarren atala**: *Rol semantikoen etiketatze automatikoa.* Atal honen lehenengo eta bigarren azpiataletan SRL atazaren *status quaestionis* delakoa edo ikerketaren egungo egoera azaltzen dugu. Hasteko, ingelesa aztergai harturik burutu diren lanez eta hizkuntza hartan dauden baliabide linguistikoez ariko gara, eta, ondoren, euskarazko rol semantikoen etiketatze automatikoa aurrera eraman ahal izateko eskuragai dauden baliabideak aurkeztuko ditugu. Jarraian, bigarren azpiatalean hain zuzen ere, SRL etiketatzaileen tradiziozko bost mailako arkitektura aztertuko dugu, maila bakoitzaren erronkak zein diren eta normalean nola ebazten diren adieraziz. Bigarren azpiatal honetan SRLn erabili ohi diren hizkuntza ezaugarriak ere zerrendatuko ditugu. Hirugarren azpiatalean, tesian garatu dugun eta rol semantikoak guztiz automatikoki etiketatzen dituen *bRol*era iritsi aurretik sortutako prototipoaz eta honen emaitzez arituko gara. Laugarren azpiatalean berriz *bRol* tresna aurkeztuko dugu, eta honen emaitzak beste sei hizkuntzaren azterketatik erdietsi direnekin alderatuko ditugu.
- **Hirugarren atala**: *Denbora informazioaren erauzketa.* Aurreko atalean egin dugun bezala, atal honetan ere, ikerketaren egungo egoera ikusiz hasiko gara. Lehen azpiatalak tesian jarraitzen dugun *ISO-TimeML* eskema estandarrera iritsi aurretiko ikerketa laburbiltzen du, baita eskemaren gailentze eta estandarizazioa ekarri duten baliabide eta lehiaketena ere. Ondoko azpiatalean, berriz, ingeleserako eskemaren eta guk erabili dugun euskararako egokitzapenaren (Altuna et al., 2016a) arteko aldea azaltzen dugu. Hirugarren azpiatalean denbora informazioa erauzten duten tresnen osagai nagusiez eta denbora adierazpenen erauzketari ematen

zaizkion erregelen eta ikasketa automatikoaren bitartezko hurbilpenez mintzo gara. Gero, laugarren azpiatalean, tesian garatu dugun *bTime* tresnaren diseinua, honen emaitzak, *bTime* erabilita egindako esperimentuak eta tresnaren hobekuntzak azaltzen ditugu. Azkenik, etorkizunean euskaraz denbora etiketatzeko orduan aurreikusten ditugun lanak eta egin beharreko azterketak proposatzen ditugu.

- **Laugarren atala**: *Espazio informazioaren erauzketa.* Atal honetan testuetan aurkitzen den espazio informazioa automatikoki etiketatzeaz dihardugu. Hasierako azpiatalean tesian jarraitu dugun *ISO-Space* eskemaren osaketa prozesua azaltzen dugu. Izan ere, eskema honek jasotzen duen *jaraunspenak* atazaren ikerketaren egungo egoeraren eta azken urteotan izan duen bilakaeraren berri ematen du (rol espazialen etiketatzea, SpRL). Bigarren azpiatala euskararako garatzen ari den eskemaren egokitzapenaren berri emateko darabilgu. Azkeneko azpiatalean garatu dugun *X-Space* ingeleserako etiketatzailearen diseinuaz eta emaitzez mintzo gara.

- **Bosgarren atala**: *Ondorioak eta etorkizuneko lanak.* Tesiaren ondorioak biltzen ditugu atal honetan. Gainera, baliabide eta denbora murriztapenak direla-eta etorkizunerako uzten diren lanak edo ikerlerroak ere biltzen ditugu.

1.9 Argitalpenak

Tesian zehar egindako aurrerapenen berri nazioko eta nazioarteko hainbat kongresutan eman dugu[3]. Jarraian datorrena argitaratutako artikuluen zerrenda da, dagokien tesi atalaren arabera antolatuta. Gainera, tesia egin bitartean hizkuntzaren prozesamenduan kokatzen diren baina tesiaren gaiarekin lotura zuzenik ez duten bi argitalpen ere aipatzen ditugu.

Bigarren atala - *Rol semantikoen etiketatze automatikoa*

- **Rol semantikoen etiketatze automatikoa.**
  Haritz Salaberri, Olatz Arregi eta Beñat Zapirain. 2014.
  Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria-EKAIA, 27. zk., EKAIA. 297-313. ISSN: 0214-9001

[3] http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1355226700

Laburpena: SRL atazaren aurkezpena, dibulgazio zientifiko eta teknikoko *EKAIA* aldizkarian. Ingeleserako eta euskararako orduan eskuragai zeuden baliabideak adibideekin azaldu, eta euskararako eraikitzen ari ginen SRL prototipoaren lehenengo emaitzak aurkeztu genituen bertan.

- **First approach toward Semantic Role Labeling for Basque.**
  Haritz Salaberri, Olatz Arregi and Beñat Zapirain. 2014.
  Language Resources and Evaluation Conference, 9th edition, LREC-2014. Reykjavik, Iceland. 1387-1393. ISBN: 9781632666215

  Laburpena: Euskararako eraikitako SRL prototipoaren azkeneko emaitzak aurkeztu genituen artikulu honetan. Emaitzak ikasketa algoritmo ezberdinak erabilita lortu genituen. Gainera, *Leave-One-out* teknika (Larson, 1931) aplikatu genuen, argumentuei rol semantikoak esleitzeko erabilitako ezaugarri linguistikoen eragina neurtu ahal izateko.

- ***bRol*: The Parser of Syntactic and Semantic Dependencies for Basque.**
  Haritz Salaberri, Olatz Arregi and Beñat Zapirain. 2015.
  Recent Advances in Natural Language Processing, 4th edition, RANLP-2015. Hissar, Bulgaria. 555-562. ISSN: 1313-8502

  Laburpena: Prototipoan oinarrituta sortutako SRL tresna guztiz automatikoa, euskararako lehena eta dagoen bakarra; *bRol* deitu dugu. Tresna honen emaitzak *ConLL-09* lehiaketan beste sei hizkuntzatarako parte hartu zuten sistemen emaitzekin alderatu genituen.

Hirugarren atala - *Denbora informazioaren erauzketa*

- **Euskarazko Gertaeren Etiketatze Automatikoa.**
  Haritz Salaberri, Olatz Arregi eta Beñat Zapirain. 2017.
  IkerGazte-2017 kongresuko artikulu-bilduma (Giza zientziak eta artea), Iruñea, Euskal Herria. 22-29. ISBN: 978-84-8438-628-5

  Laburpena: Argitalpen honetan euskaraz idatzitako testuetan aurki daitezkeen gertaeren etiketatze automatikorako *bEVENT* tresna aurkezten da. Prozesua aurre-

ra eraman ahal izateko gertaerak identifikatzeaz gainera hauei dagozkien atributu linguistikoak ere zehazten dira. *bEVENT* euskararako garatu den mota honetako lehenbiziko sistema da, eta oinarritako *ISO-TimeML* izeneko anotazio eskema estandarra jarraitzen du. Tresnak ikasketa automatikoko metodoak eta *Euskal-TimeBank* izeneko corpusa baliatzen ditu gertaerak etiketatzeko. Ebaluazioa *Train-Test* prozeduraren bitartez egin da eta identifikazioan erdietsitako prezisioa, estaldura eta F_1 neurria 83.92, 72.76 eta 77.94 puntukoak dira.

Laugarren atala - *Espazio informazioaren erauzketa*

- **IXAGroupEHUSpaceEval:(X-Space) A WordNet-based approach towards the Automatic Recognition of Spatial Information following the ISO-Space Annotation Scheme.**
  Haritz Salaberri, Olatz Arregi and Beñat Zapirain. 2015.
  International Workshop on Semantic Evaluation, 9th edition, *SemEval-2015*. Denver, Colorado, USA. 856-861. ISBN: 978-1-941643-40-2

  Laburpena: *SemEval-2015* lehiaketako *SpaceEval* atazarako garatutako *X-Space* sistemari dagokion argitalpena. *X-Space* ingeleseko testuetan espazio informazioa etiketatzen duen sistema da. Etiketatze prozesua aurrera eraman ahal izateko *WordNet* (Miller, 1995) datu-base lexikala eta hainbat ezaugarri linguistiko erabiltzen ditu, hauen artean rol semantikoak.

Beste argitalpenak

- **Simple or Complex? Assessing the readability of Basque Texts.**
  Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Haritz Salaberri. 2014.
  International Conference on Computational Linguistics, 25th edition, *COLING-2014*. Dublin, Ireland. 334-344. ISBN: 978-1-941643-26-6

  Laburpena: Euskarazko testuen kategorizazioaz diharduen artikulua dugu hau. Kategorizazioaren xedea testuak irakurtzen direnean duten zailtasunaren arabera

sailkatzean datza. Kasu honetan bi kategoria multzo egitea izan zen helburua, testu *erraz* edo *sinpleak* eta testu *zail* edo *konplexuak*. Horretarako hainbat ezaugarri linguistiko eta hainbat ikasketa algoritmo erabilita egin genituen esperimentuak, zenbait sailkatzaile sortu eta emaitzarik onenak ematen zituen konfigurazioa lortzeko.

- **IXAGroupEHUDiac: A Multiple Approach System towards the Diachronic Evaluation of Texts.**
  Haritz Salaberri, Iker Salaberri, Olatz Arregi and Beñat Zapirain. 2015.
  International Workshop on Semantic Evaluation, 9th edition, *SemEval-2015*. Denver, Colorado, USA. 840-845. ISBN: 978-1-941643-40-2

  Laburpena: *SemEval-2015* lehiaketako *Diachronic Text Evaluation* atazarako garatutako sistemari dagokion argitalpena. Bertan ingelesez XIX, XX eta XXI. mendeetan idatzitako testuen argitalpen urtea automatikoki eta testuen *hizkeran* oinarrituta esleitzen duen sistemaren diseinu eta emaitzak aurkeztu genituen.

2

ROL SEMANTIKOEN ETIKETATZE AUTOMATIKOA

Atal honetan, rol semantikoen etiketatze automatikoaren inguruan egin dugun lana aurkeztuko dugu. Lehenik, SRLren ikerketaren egungo egoera azalduko dugu; ingeleserako burutu ziren ikerketa eta baliabideetatik hasi eta, tesiari ekin genionean, euskararako eskuragarri zeuden baliabide eta ikerketetarainoko lan-bilduma laburbiltzen saiatuko gara. Gero, euskarazko SRL atazari lotutako aurrerapausoak zerrendatuko ditugu, horiek rol semantikoak etiketatzeko sistema guztiz automatiko baten eraikuntza izan dute helburu. Horregatik, sistema hauek izaten duten ohiko egitura zein den azalduko dugu, eta euskararako eraikitako SRL prototipoa eta honetan oinarritutako azken SRL sistema deskribatuko ditugu. Azkenik, erabilitako ezaugarri linguistikoez, hau da, *ezaugarrien ingeniaritzaz*, eta etorkizunean inplementa genitzakeen hobekuntzez arituko gara atal honen bukaeran.

2.1 Ikerketaren egungo egoera

Lengoaia naturalaren prozesamenduan ataza edo gai jakin baten ikerketaren egungo egoera aztertzeko orduan hiru kategoria hartu ohi dira kontutan: garatutako sistemak, baliabide linguistikoak eta ebaluazio saioak. Jarraian SRLren ikerketaren egungo egoera osatzen duten hiru kategoria hauek aurkeztuko ditugu.

2.1.1 SRL etiketatzaileak

SRL atazaren inguruan argitaratutako lehenengo argitalpen esanguratsua (Gildea eta Jurafsky, 2002) dela esaten da. Bertan, esaldien osagaietan-oinarritutako analisi sintaktikotik abiatuta, analisi sintaktikoko osagai bakoitzari rol semantikoa esleitzen dion ingeleseko SRL sistema aurkezten da. Hau da garatu zen lehendabiziko rol semantikoen etiketatzaile automatikoa. Aipatu behar da SRL ataza sortu zenean, osagaietan oinarritutako sintaxia erabiltzen zela; ondoren, dependentzietan oinarritutakoa gehienbat. Izan ere, dependentzia sintaktikoetan oinarritutako SRL sistemek, entrenamendurako, osagaietan oinarritutako sintaxia erabiltzen dutenek baino denbora eta baliabide konputazional gutxiago behar izaten dituzte, emaitza antzekoak lortzeko.

Rol semantikoen etiketatze automatikoaren hasierako urteetan esanguratsuak izan ziren, aipatutakoaz gainera, (Xue eta Palmer, 2004; Punyakanok et al., 2004; Hacioglu, 2004) argitalpenak. Hirurek deskribatzen dituzte ingelesezko rol semantikoen etiketatzean emaitzak hobetu ahal izateko teknika eta ezaugarriak. Lehenengoak SRL egiteko ezaugarri sintaktiko berriak proposatzen ditu; gainera, SRL egiteko ematen den pauso bakoitzean hainbat ezaugarri erabili beharra adierazten du. Ordura arteko sistema gehienek aplikatzen zuten SVM (*Support Vector Machines*) algoritmoa (Cortes eta Vapnik, 1995) erabili beharrean Entropia Maximoko printzipioan (Jaynes, 1957) oinarritutako algoritmoa eta ezaugarri gutxiago erabilita emaitza berak lor daitezkeela ere erakusten du. Bigarren argitalpenak ILP (*Integer Linear Programming*) probabilitateen optimizazio teknikan (Schrijver, 1998) oinarritzen den inferentzia prozedura eta ikasketa automatikoa elkartzen dituen SRL sistema aurkezten du. Sistema honen emaitzak garaiko sistemarik onenen artean kokatzen dira. Azkenik, hirugarren argitalpenak dependentzia sintaktikoetan oinarritutako lehenengo SRL sistema deskribatzen du. Honetan, hala ere, dependentzien analisi sintaktikoa ez da zuzenean analizatzaile sintaktikotik lortzen, osagaietan oinarritutako analisi sintaktikoko irteera dependentzietara bihurtuta baizik.

2004. urtetik SRL atazak izandako garapena eta erronkak hobekien islatzen dituzten argitalpenak, aldiz, beste hiru hauek direla uste dugu: (Jiang eta Ng, 2006), (Pradhan et al., 2008) eta (Björkelund et al., 2009). Lehenengoak izen-predikatuen argumentuen rolak esleitzen dituen lehendabiziko rol semantikoen etiketatzailea deskribatzen du. Ordura arteko etiketatzaile guztiak aditz-predikatuen argumentuen rolak etiketatzeaz bakarrik arduratzen ziren. Sistema honek Entropia Maximoko printzipioan oinarritutako algoritmoa erabiltzen du. Bigarren argitalpenak, berriz, SRL sistemek domeinuz kanpoko testuak[1] etiketatzean izaten duten emaitzen okertzea aztertzen du. Azkeneko artikuluak hizkuntza batean baino gehiagotan idatzitako testuen SRL egiten duen lehen sistemetakoa deskribatzen du, hain zuzen ere, katalanez, gaztelaniaz, txekieraz, ingelesez, alemanez, japonieraz eta txineraz idatzitako testuen etiketatzea egiten du.

Bukatzeko, aipagarriak direla uste dugu eragin nabarmena izan duten (Màrquez et al., 2008), (Zhou eta Xu, 2015) eta (Foland Jr eta Martin, 2015) lanak ere. Lehenengoak rol semantikoen etiketatze automatikoaren errepasoa egiten du: zer den, SRL etiketatzaileak zer osagai dituen, zer ezaugarri linguistiko erabiltzen diren, etiketatzaileak zer metrika erabilita ebaluatzen diren, e.a. Bigarren eta hirugarren lanek, bestalde, sare neuronaletan (*Neural Networks*) oinarritutako edo ikasketa sakoneko (*Deep Learning*) algoritmoak erabiltzen dituzten ingeleseko rol semantikoen etiketatzaileak aurkezten dituzte.

Rol semantikoak

Garrantzizkoa da, gainera, SRL sistemek esleitzen dituzten etiketak edo rolak zein diren zehaztea. Ildo horretatik, eta guk dakigula behintzat, ez dago rol semantiko eta adjuntuen zerrenda itxirik. Hala ere, hizkuntzalari gehienek onartzen dituzte jarraian aurkezten ditugunak. Izan ere, aipatutako sistemek eta tesi lan honetan garatu ditugun SRL tresnek, euskarazko etiketatzaileen garapenerako erabilitako *EPEC-RolSem* corpusak bezala (Estarrona et al., 2015), helburu orokorreko hurrengo hemeretzi rol eta hamar adjuntu hauek erabiltzen dituzte, predikatuekiko erlazio semantikoak adierazteko (rolak larriz eta adjuntuak xehez):

- `ACTOR` eta `AGENT`: Berariaz gertaera eragiten duena. Abiarazlea.
- `ASSET`: Transakzio ekonomikoetan trukatzen dena. Trukagaia.

[1] Etiketatzailearen entrenamendurako erabili den corpusaren domeinukoak ez diren testuak.

- ATTRIBUTE: Beste argumentu baten izaera adierazten duena.
- BENEFICIARY: Gertaeraren ondorioz etekina lortzen duena.
- CAUSE: Oharkabean gertaera eragiten duena.
- LOCATION, DESTINATION, SOURCE: Gertaera dinamikoetan mugimenduari lotutako argumentuak, ezaugarri espazialak. Kokapena, helmuga eta jatorria.
- EXPERIENCER: Gertaera zentzuen edo emozioen bitartez hautematen, sentitzen edo jasaten duena.
- EXTENT: Gertaerak eragiten duen aldaketa-gradua adierazten duena.
- INSTRUMENT: Gertaeran tresnatako erabiltzen dena.
- MATERIAL eta PRODUCT: Eraldaketak adierazten dituzten gertaeretan (*eraiki*) parte hartzen duten argumentuak. Eraldaketan erabilitako materiala eta haren ondorioz sortutako produktua.
- PATIENT: Gertaera jasaten duen argumentua.
- RECIPIENT: Trukatze batean trukagaia jasotzen duena.
- STIMULUS: Gertaera zentzuen edo emozioen bitartez hautematen, sentitzen edo jasaten duen argumentuaren eragina.
- THEME: Gertaera dinamikoetan kokapena aldatzen duena.
- TOPIC: Hizpidea, komunikazioak adierazten dituzten gertaeretan (*esan*).
- Location, Time, Extent, Direction, Manner, Mode, Cause: Gertaeraren *non*, *noiz*, *zenbat denboraz*, *nora*, *zerekin*, *nola* eta *zergatik* galderei erantzuna ematen dieten adjuntuak.
- Negation, Discourse, Adverbial: Gertaera ezeztatzen dela eta gertaera diskurtsozkoa edo adberbiozkoa dela adierazten duten adjuntuak.

2.1.2 Hizkuntza baliabideak

Rol semantikoen etiketatze automatikoa egiten duten sistemak garatzeko orduan, beharbeharrezkoak dira predikatu-lexikoiak eta rol semantikoekin eskuz etiketatutako corpusak. Predikatu-lexikoiak hizkuntza bateko predikatu zerrendak dira, non predikatu bakoitzarentzat zehazten den zein diren izan ditzakeen adiera ezberdinak, adiera hauetako bakoitzean jasotzen ahal dituen argumentuak, eta argumentuek jokatzen dituzten rolak. Predikatu zerrenda hauek sortzeko egin beharreko hizkuntza ikerketari predikatuen azpikategorizazioa deritzo.

Sarreran adierazi dugun moduan, predikatuak aditzak, izenak, adberbioak eta adjektiboak izan daitezke. Gaur egun, aditzezko eta izenezko predikatu-lexikoiak daude gehienbat; izan ere, aditzak eta izenak dira testuetan agertzen diren predikatu gehienak. Ingeleserako rol semantikoak automatikoki etiketatzeko tresnak garatzea ahalbidetzen duten baliabide zerrenda luzea bada ere, erabilienak lau hauek dira: *PropBank* (Palmer et al., 2005), *NomBank* (Meyers et al., 2004), *FrameNet* (Fillmore et al., 2004) eta *VerbNet* (Kipper et al., 2000). Baliabide hauek ingelesezko SRL etiketatzaileak garatzeko erabiltzen badira ere, beharrezkoa da duten egitura eta erabiltzen dituzten formalismoak ezagutzea beste hizkuntzetarako garatutako baliabideen garapenean erreferentziatzat erabili direlako, euskararako baliabideen garapenean besteak beste.

PropBank

PropBank corpusa *The Wall Street Journal* egunkarian argitaratutako berriez dago osatuta eta milioi bat hitz dauzka. Corpusean predikatu-argumentu-adjuntu egiturak eta argumentuen rol semantikoak daude etiketatuta. Argumentuak anotatzeko ez dira 2.1.1 azpiatalean zerrendatu ditugun rolak erabiltzen, *teorikoki neutrala* den rol multzoa baizik[2]. Predikatu guztien argumentuak etiketatzeko rol berak erabiltzen direnez, predikatuaren eta honen adieraren arabera rol bakoitzaren interpretazio semantikoa ezberdina da; rolen esanahia predikatuaren eta adieraren menpekoa da; alegia. Argumentuak anotatzeko `arg0`, `arg1`, `arg2`, `arg3`, `arg4` eta `arg5` etiketak erabiltzen dira. Hauei *core* rolak esaten zaie. Predikatuen adjuntuak etiketatzeko, ordea, ondoko etiketak erabiltzen dira.

[2] *PropBank*en egileen arabera, rol multzo honek ez du inongo teoria linguistikorik jarraitzen.

ArgM-LOC: lekuzkoa
ArgM-EXT: hedadura
ArgM-DIS: diskurtso markatzailea
ArgM-ADV: adberbiala
ArgM-NEG: ezeztapen marka
ArgM-MOD: modua
ArgM-CAU: kausa
ArgM-TMP: denbora
ArgM-PNC: helburua
ArgM-MNR: moduzkoa
ArgM-DIR: norabidea

PropBank garatu zenean, `arg0` etiketarekin *egile* (`AGENT`, `ACTOR`) rola jokatzen zuten argumentuak etiketatzeko ahaleginak egin ziren, eta `arg1` delakoarekin *gaia* (`THEME`, `PATIENT`, `EXPERIENCER`) rola jokatzen zutenak etiketatzen saiatu ziren. Ondorioz, *PropBank*en `arg0` etiketa daramana, gehienetan, *egilea* izango da, eta `arg1` etiketa daramana, berriz, *gaia*. Saiakera honen bidez, emaitza hobeak lortu ahal izatea zen helburua, ataza ahal zen neurrian behintzat erraztea, alegia. *PropBank* corpusa garatzerakoan, eginbehar garrantzizkoenetakoa etiketatu nahi ziren aditzen azpikategorizazioa egitea izan zen. Ikerketa prozesu horretatik *PropBank*eko aditzen lexikoia edo *verb-index*a sortu zen. Gaur egungo bertsioak 6308 predikatu dauzka[3]. Lexikoi honetan aditz bakoitzari *frameset* izeneko adiera identifikatzaile multzo bat eta bakarra dagokio, eta *frameset* bakoitzari gutxienez *roleset* deritzon rol multzo bat, aditzak duen adiera bakoitzeko bat[4]. *Roleset*ean aditz batek adiera jakin batean onartzen dituen *core* argumentuen multzoa biltzen da. Gainera, *roleset* bakoitzaren ondoren, adibidetako, rol semantikoak etiketatuta dauzkaten eta aditzaren adiera hori erabiltzen duten esaldiak ematen dira. Adibide hauen helburua *roleset* edo rol multzo horrekin bateragarriak diren egitura sintaktikoak zerrendatzea da. Aditz baten *roleset*ek aditzaren *frameset*a osatzen dute. Ondorengo adibidean *abandon* (alde batera utzi) aditz predikatuaren *frameset*a ikus daiteke.

Adibideak erakusten duen moduan, *abandon* aditz predikatuarentzat hiru adiera ezberdin daudela erabaki zen, *PropBank*eko azpikategorizazioa egiterakoan. Lehenengoak zerbait (alde batera) utzi dela adierazten du (`abandon.01`); hau da aditzaren adierarik arruntena. Bigarrenak, berriz, zerbait beste zerbaiten truke, *trukagaitako*, eman dela (`abandon.02`). Azkenik, hirugarren adierak zerbait, errenditzeagatik, galdu dela adierazten du (`abandon.03`).

[3]Hasieran aditz-predikatuak bakarrik landu baziren ere, orain izen eta adjektibo predikatuak ere landuta daude.

[4]`https://verbs.colorado.edu/propbank/framesets-english/`

Roleset id: abandon.01 - *To leave something behind.*

arg0: *abandoner*
arg1: *thing abandoned, left behind*
arg2: *attribute of* arg1

Example: typical_transitive

They think $\overbrace{\textit{[the Board}_e \textit{]}}^{\textbf{arg0}_e}$ $\overbrace{\textbf{\textit{has abandoned}}_e}^{\texttt{abandon.01}}$ $\overbrace{\textit{[their interest}_e \textit{]}}^{\textbf{arg1}_e}$.

Example: with_attribute

$\overbrace{\textit{[John}_e \textit{]}}^{\textbf{arg0}_e}$ $\overbrace{\textbf{\textit{abandoned}}_e}^{\texttt{abandon.01}}$ $\overbrace{\textit{[his pursuit}_e \textit{]}}^{\textbf{arg1}_e}$ $\overbrace{\textit{[as a waste of time}_e \textit{]}}^{\textbf{arg2}_e}$.

Roleset id: abandon.02 - *To exchange for something.*

arg0: *abandoner*
arg1: *thing abandoned, left behind*
arg2: *preferred item*

Example: with_preferred_item

$\overbrace{\textit{[General Noriega}_e \textit{]}}^{\textbf{arg0}_e}$ *wanted* $\overbrace{\textbf{\textit{to abandon}}_e}^{\texttt{abandon.02}}$ $\overbrace{\textit{[his command}_e\textit{]}}^{\textbf{arg1}_e}$ $\overbrace{\textit{[for exile}_e \textit{]}}^{\textbf{arg2}_e}$.

Roleset id: abandon.03 - *To surrender or give over something.*

arg0: *entity abandoning something*
arg1: *thing abandoned*
arg2: *benefactive, abandoned-to*

Example: give_up_to

Once $\overbrace{\textit{[he}_e\textit{]}}^{\textbf{arg0}_e}$ $\overbrace{\textbf{\textit{had abandoned}}_e}^{\texttt{abandon.02}}$ $\overbrace{\textit{[himself}_e\textit{]}}^{\textbf{arg1}_e}$ $\overbrace{\textit{[to the very worst}_e \textit{]}}^{\textbf{arg2}_e}$.

Aditz honen kasuan, hiru adierei dagozkien *roleset*ek `arg0`, `arg1` eta `arg2` *core* rolak jasotzen dituzte. `arg0` eta `arg1` rolek, nahiz eta *abandoner, entity abandoning something* eta *thing abandoned, left behind* bezala definitzen diren, `AGENT` eta `THEME` rol prototipikoekin bat egiten dutela ikusten da. `arg2` rolaren kasuan, ordea, ez da antzekotasun semantiko argirik ikusten; aditzaren adieraren arabera esanahi bat edo beste dauka alegia: `abandon.01`en kasuan `arg1` rolaren atributua adierazten du, `abandon.02`ren kasuan, berriz, alde batera uzten dena zerekin trukatzen den, eta `abandon.03`n nori edo zeri uzten zaion alde batera utzitakoa. *Roleset* bakoitzaren azpian, predikatuaren adiera horiek izan ditzaketen gauzatze sintaktiko ezberdinen adibideak daude.

NomBank

NomBank, *PropBank* bezala, rol semantikoekin etiketatutako corpusa eta predikatu-lexikoia da. Ezberdintasuna da, *NomBank*en kasuan, landutakoak izen predikatuak direla, eta ez *tradiziozko PropBank*en modura aditz-predikatuak. Gaur egun, eta aurretik esan bezala, *PropBank*en izen eta adjektibo predikatuei dagozkien *frameset*ak daude txertatuta; izen predikatuenak *NomBank* proiektuan sortutakoak dira.

NomBank 2004. urtean hasi zen eta 2007. urtera arte iraun zuen. Proiektuaren ondorioz, 4.707 izen-predikaturen azpikategorizaziotik sortutako predikatu-lexikoia osatu zen eta, honetan oinarriturik, *The Wall Street Journal* corpuseko izen predikatuen argumentuen rolen etiketatzea burutu zen. Horretarako hainbat baliabide lexikal behar izan ziren; aipagarriena *NOMLEX-PLUS* deitutako aditz, adjektibo eta adberbioen nominalizazio hiztegia. Hau *NOMLEX* nominalizazio hiztegiaren (Macleod et al., 1998) hedapena da. Hasierakoak 1.000 sarrera zeuzkan eta *NOMLEX-PLUS* hiztegiak, berriz, 6.000 sarrera gehiago zeuzkan.

Hurrengo adibidea *NOMLEX-PLUS* hiztegiko *claim* (kexa, baieztapena, eskaera) aditz nominalizazioaren sarrerari dagokio. *Claim*, *to claim* (kexatu, baieztatu, eskatu) aditzaren nominalizazioa da, sarreraren laugarren lerroan adierazten den moduan. Honen harira, aipatzekoa da *NomBank* sortzerakoan ondorengo prozesu erdi automatikoa jarraitu zela: aditz-nominalizazioa zen izen-predikatu bakoitzak nominalizatutako aditz-predikatuaren *roleset*ak jaraunsten zituen, adieraz adiera, baldin eta hizkuntzalariek izen-predikatuak zeuzkan adierak eta aditz-predikatuak zeuzkanak bat zetozela erabakitzen baldin bazuten.

```
(NOM:ORTH "claim"
    :PLURAL "claims"
    :PLURAL-FREQ "not-rare"
    :VERB "claim"
    :NOUN ((RARE-NOUN))
    :NOUN-SUBC ((NOUN-PP :PVAL ("about")))
    :NOM-TYPE ((VERB-NOM))
    :VERB-SUBJ ((N-N-MOD)
        (DET)
        (DET-POSS)
        (PP :PVAL ("by""of")))
    :SUBJ-ATTRIBUTE ((COMMUNICATOR))
    :VERB-SUBC ((NOM-NP :SUBJECT ((N-N-MOD)
                (DET)
                (DET-POSS)
                (PP :PVAL ("of""by")))
            :OBJECT ((PP :PVAL ("of")))
            :REQUIRED ((SUBJECT :DET-POSS-ONLY T
                    :N-N-MOD-ONLY T)))
        (NOM-S :SUBJECT ((N-N-MOD)
                (DET-POSS)
                (PP :PVAL ("of""by")))
            :NOM-SUBC ((SENT :THAT-S T)))
```

Jarraian, eta ondorengo adibideak erakusten duen moduan, *claim* izen predikatuarentzat bi adiera ezberdin definitu ziren *NomBank* garatu zenean: `claim.01` eta `claim.02`. Gainera, bi adiera hauek *claim* aditz-predikatuaren bi adierekin bat zetozela erabaki zen. Ondorioz, `claim.01` izen-predikatuaren adierak `claim.01` aditz-predikatuaren adieraren *roleseta* jarauntsi zuen, eta `claim.02` izen-predikatuaren adierak, berriz, `claim.02` aditz-predikatuaren adieraren *roleseta*. Adibidean, *NomBank*eko *claim* predikatuari dagokion *frameseta* aurkezteaz gainera, *claim* izen eta aditz predikatuak erabilita gertaera berak deskribatzen dituzten esaldiak erakusten dira.

Lehen adiera *norbaitek zerbait* baieztatzeari dagokio eta bigarrena berriz *norbaitek norbaitentzat zerbait* eskatu, eskuratu edo bereganatzeari. *PropBank*eko *abandon* predikatuaren kasuan bezala, hemen ere, `arg0` eta `arg1` rolak bat datoz, hurrenez hurren, `AGENT` eta `THEME` rol prototipikoekin. `arg2` rola, ordea, lehen adieraren kasuan (`claim.01`) *entzule*ari dagokio eta bigarrenaren kasuan (`claim.02`) *onuradunari*.

Roleset id: `claim.01` - *To assert something.*

`arg0`: *claimer, asserter*
`arg1`: *thing claimed, asserted*
`arg2`: *hearer*

Noun example: *NomBank*

$$\overbrace{[\ Her_e\ ]}^{\textbf{arg0}_e}\ \overbrace{\textbf{\textit{claim}}_e}^{\texttt{claim.01}}\ \overbrace{[\ to\ Pilail_e\ ]}^{\textbf{arg2}_e}\ was\ \overbrace{[\ that\ Fred\ can\ fly_e\ ]}^{\textbf{arg1}_e}.$$

Verb example: *PropBank*

$$\overbrace{[\ She_e\ ]}^{\textbf{arg0}_e}\ \overbrace{\textbf{\textit{claimed}}_e}^{\texttt{claim.01}}\ \overbrace{[\ to\ Pilail_e\ ]}^{\textbf{arg2}_e}\ \overbrace{[\ that\ Fred\ can\ fly_e\ ]}^{\textbf{arg1}_e}.$$

Roleset id: `claim.02` - *To seize something.*

`arg0`: *claimer, seizer*
`arg1`: *object claimed, seized*
`arg2`: *beneficiary*

Noun example: *NomBank*

$$\overbrace{[\ Hernán\ Cortés_e\ ]}^{\textbf{arg0}_e}\ laid\ \overbrace{\textbf{\textit{claim}}_e}^{\texttt{claim.02}}\ \overbrace{[\ to\ Mexico_e\ ]}^{\textbf{arg1}_e}\ \overbrace{[\ for\ Spain_e\ ]}^{\textbf{arg2}_e}.$$

Verb example: *PropBank*

$$\overbrace{[\ Hernán\ Cortés_e\ ]}^{\textbf{arg0}_e}\ \overbrace{\textbf{\textit{claimed}}_e}^{\texttt{claim.02}}\ \overbrace{[\ Mexico_e\ ]}^{\textbf{arg1}_e}\ \overbrace{[\ for\ Spain_e\ ]}^{\textbf{arg2}_e}.$$

FrameNet

FrameNet, *marko semantika* paradigma (Fillmore, 1976) gisa daukan ingeleseko datubase lexikala da. Honen arabera, esaldietan zenbait hitzek gaitasuna daukate jakintza semantikoko egiturak *aktibatzeko*. Hitz hauei unitate lexikal esaten zaie. *Frame*ak gertaerak eta haien parte hartzaileak deskribatzen dituzten egitura kontzeptualak dira.

*[Mike$_e$] **sold**$_e$ [the book$_e$] to [Mary$_e$].*
*[Mary$_e$] **bought**$_e$ [the book$_e$] from [Mike$_e$].*

Adibidean gertaera bera, *Mike* eta *Mary*ren arteko liburuaren salerosketa, adierazten duten bi esaldi ageri dira. Esaldietako bakoitzean erabiltzen den aditz-predikatua ezberdina bada ere, *sold* eta *bought* (*saldu* eta *erosi*), bi predikatuek gertaera bera deskribatzen dutela ulertzen dugu. Esaldi hauek *PropBank/NomBank* eredua jarraituta etiketatuko bagenitu, predikatuetako bakoitzari adiera ezberdina esleituko genioke, hain zuzen ere lehenengoak `sell.01` adiera jasoko luke, eta bigarrenak ordea `buy.01`. Markoen semantika eta *FrameNet* eredua jarraituta, aldiz, *sold* eta *bought* aditz-predikatuek *FrameNet*eko `economic_transaction` *frame*a aktibatuko lukete; *sold* eta *bought* unitate lexikalak dira. Argumentuen rolei dagokienez, *PropBank/NomBank* ereduan *Mike* `arg0` etiketatuko genuke lehenengo esaldian eta `arg2` etiketarekin bigarrenean. *Mary* `arg2` izango litzateke lehenengo esaldian, eta `arg0` bigarrenean. *the book* argumentua `arg1` rolarekin etiketatuko genuke bi esaldietan.

$$\overbrace{\textbf{Seller}_e}^{[\ Mike_e\ ]}\ \overbrace{\overset{\textbf{economic\_}}{\textbf{transaction}}}^{\textbf{sold}_e}\ \overbrace{\textbf{Good}_e}^{[\ the\ book_e\ ]}\ to\ \overbrace{\textbf{Buyer}_e}^{[Mary_e\ ]}.$$

$$\overbrace{\textbf{Buyer}_e}^{[\ Mary_e\ ]}\ \overbrace{\overset{\textbf{economic\_}}{\textbf{transaction}}}^{\textbf{bought}_e}\ \overbrace{\textbf{Good}_e}^{[\ the\ book_e\ ]}\ from\ \overbrace{\textbf{Seller}_e}^{[Mike_e\ ]}.$$

Adibidean esaldiak *FrameNet* eredua jarraituta anotatuta ageri dira. Bertan, eta aurretik esan bezala, *sold* eta *bought* aditz-predikatuek `economic_transaction` *markoa* aktibatzen eta jasotzen dute. Argumentuen rolak direla eta, *FrameNet*en, *PropBank/NomBank*en ez bezala, *frame* bakoitzari rol multzo espezifiko bat dagokio. `economic_transaction`entzat, *salerosketa* gertaera batean parte hartzen duten entitate edo argumentuek jokatzen dituzten rolak, *eroslearena* (`Buyer`), *saltzailearena*

(`Seller`) eta *salgaiarena/erosgaiarena* (`Good`) direla zehazten da *FrameNet*en. Adibidean ikusten da eredu honetaz baliaturik ez dela *PropBank/NomBank* eredua jarraituta bezala *Mary* eta *Mike* argumentuen rolen arteko aldizkatzerik gertatzen. *Mary* bi esaldietan etiketatu da `BUYER` moduan, eta *Mike* `SELLER` moduan. *FrameNet*en rol semantikoei *frame elementuak* (*frame elements*) esaten zaie. Hauek *frame semantikoen* menpekoak dira eta, horregatik, *frame elementuen* kopurua oso handia da. 1.200 *frame* semantiko eta 13.000 unitate lexikal daude osotara. Gainera, datu-base lexikalak eskuz etiketatuta dauden eta *FrameNet* eredua jarraitzen duten esaldiak biltzen ditu. *Frame elementu* kopurua ez bezala, esaldi kopurua ez da oso handia eta, ondorioz, rolen sakabanaketa nabarmena da *FrameNet*en. Honek zailtasunak sor ditzake, estatistika tresnak aplikatuta baliabide honetatik ikasten duten etiketatzaileak sortu nahi direnean.

VerbNet

VerbNet deritzana *Levinen aditz klaseak* (*Levin's Verb Classes*) deitutako aditz-predikatuen sailkapenean (Levin, 1993) oinarritzen den predikatu-lexikoia da. Aditzen klasekako sailkapena *Levin*ek proposatutako ezagutza lexikalaren teoriari jarraikiz egin zen; honen arabera, aditzen semantikaren eta sintaxiaren arteko harremanek aditz multzoak osatzeko aukera ematen dute. Onartu ere, izaera semantiko bereko aditzek egitura sintaktiko berak onartzen dituzte. Ondorioz, egitura sintaktiko haiek identifikatu eta haietan oinarrituta, aditzak klasetan multzoka daitezke, semantika mailan betiere. Aditz multzo hauei deritze *Levinen aditz klaseak*. Ikuspegi honen bitartez klase bereko aditzek gauzatze sintaktiko eta semantiko bera, eta horregatik azpikategorizazio bera, daukatela ulertzen da. *VerbNet* predikatu-lexikoiko sarrera bakoitza *Levin* klase bat da. Klase haietan izaera sintaktiko eta semantiko bera duten aditzak multzokatzeaz gainera, aditz horiek jasotzen dituzten argumentuen rolak, dauzkaten gauzatze sintaktiko eta semantikoen *frame* edo egitura zerrenda, eta argumentuen rolen hautapen murriztapenak ematen dira. *VerbNet*en rol semantikoei *rol tematiko* esaten zaie, eta erabiltzen diren etiketak rol semantiko ezagunenak dira (ikus 2.1.1). *VerbNet*en azkeneko bertsioan (`v3.2b`) 3.769 aditz biltzen dira, 274 aditz-klasetan banatuta.

2.1 irudiko adibidean `Transfer_message-37.1.1` *VerbNet* klasea erakusten da. Bertan, *Kideak* atalean, klasea osatzen duten aditzak zerrendatu dira. Klase hau *norbaitek beste norbaiti mezu bat* helarazteari dagokionez, bertako kideak *communicate* (*ko-*

munikatu), *corroborate* (*berretsi*), *demonstrate* (*frogatu*), *elucidate* (*argitu*) eta antzeko aditzak dira. *Rol tematikoak* atalean aditz hauen argumentuek jokatzen dituzten *rol tematikoak* ematen dira.

Klasea	Transfer_mesg-37.1.1		
Guraso nodoa	–		
Kideak	communicate, corroborate, demonstrate, elucidate, explain...		
Rol tematikoak	Agent Topic Recipient Source		
Hautapen murrizt.	Agent[+ANIMATE |+ORGANIZATION] Recipient[+ANIMATE |+ORGANIZATION]		
Frameak	Izena	Sintaxia	Predikatu semantikoak
	NP V how S	Agent V Topic	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)
	NP V PP	Agent V from Source	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)
	NP V	Agent V	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)

Irudia 2.1: *VerbNet*eko `Transfer_message-37.1.1` klasea (Zapirain, 2010).

Klase honen kasuan lau argumentu hartzen ahal dituzte bertako aditzek edo kideek: mezua ematen duen argumentuak `AGENT` rola jokatzen du, mezua jasotzen duenak `RECIPIENT` rola, mezuak berak `TOPIC` rola eta, azkenik, mezuaren jatorria adierazten duen argumentuak `SOURCE` rola jokatzen du. Hurrengo adibideak `Transfer_message-37.1.1` klaseko informazioa erabilita rol semantikoak etiketatuak dituen esaldia erakusten du. *PropBank/NomBank* ereduko roletan ez bezala, *VerbNet*eko rol tematikoen interpretazio semantikoa ez da aditzaren eta adieraren menpekoa. Hori dela eta, ez da beharrezkoa, *PropBank/NomBank*eko *frameset*etan egiten denaz bestera, rolen semantika adiera bakoitzerako definitzea.

$$\overbrace{[\ \textit{The major}_e\ ]}^{\textbf{AGENT}_e}\ \overbrace{\textbf{\textit{communicated}}_e}^{\texttt{Transfer\_message-37.1.1}}\ \textit{to}\ \overbrace{[\textit{the villagers}_e\ ]}^{\textbf{RECIPIENT}_e}\ \overbrace{[\textit{the king's order}_e\ ]}^{\textbf{SOURCE}_e}\ \overbrace{[\textit{to close the gate}_e\ ]}^{\textbf{TOPIC}_e}.$$

Hautapen murriztapenak atalean aditzen argumentuek rol tematiko jakin bat jokatu ahal izateko bete beharreko baldintzak ezartzen dira. Baldintza hauei esaten zaie *hautapen murriztapenak* eta argumentuaren izaeraren ingurukoak izaten dira: mugimendua daukan edo ez daukan, gizakia, animalia edo makina den, e.a. Adibideko `Transfer_message-37.1.1` klasearen kasuan, `AGENT` eta `RECIPIENT` rol tematikoak jaso beharreko argumentuentzat definitu dira soilik hautapen murriztapenak. Biek

dituzte [+ANIMATE] eta [+ORGANIZATION] murriztapenak. Hauek adierazten dute mezua emateaz eta jasotzeaz arduratzen diren argumentuek bizidunak izan behar dutela, eta erakunde, autoritate edo nolabaiteko multzoa, taldea, izan (*The major*, *the villagers*). *VerbNet*en erabiltzen diren hautapen murriztapenak modu hierarkikoan daude antolatuta[5]. Hierarkia hau *EuroWordNet*en (Vossen, 1998), Europako hainbat hizkuntzatarako sortutako *WordNet* (Fellbaum, 1998) motako datu-base lexikalak biltzen dituen proiektuan erabilitako hierarkian, dago oinarrituta.

Azkenik, *Frame*ak atalean, *Sintaxia* izeneko azpiatalean, aditz klaseak daukan gauzatze sintaktiko bakoitzerako rol tematikoak argumentu sintaktikoekin nola lotzen diren adierazten da. Gainera, *Predikatu semantikoak* azpiatalean, *VerbNet* klasea osatzen duten aditz-predikatuek deskribatzen dituzten gertaeretan argumentu bakoitzaren fasekako deskribapena egiten da (start, during, end edo result). 2.1 irudiko adibidean, *informazio transferentzia* gertaera (E) AGENT rola jokatzen duen argumentuak sortzen duela dio (cause(Ag., E)). Horretaz gainera, transferentzia hori gauzatzen diren rol tematikoen artean jazotzen dela zehazten da, gertaerak irauten duen bitartean (during(E)).

Euskararako baliabideak

Pentsa daitekeen bezala, euskarazko SRL tresnen garapena ahalbidetzen duten baliabide linguistikoen kopurua eta tamaina ez da ingelesekoen parekoa. Izan ere, gaur egun hedadura mugatuko bi baliabide besterik ez daudela esan dezakegu: *Basque Verb Index-BVI* predikatu-lexikoia (Estarrona, 2014) eta *EPEC-RolSem* corpusa (Estarrona et al., 2015).

Basque Verb Index

*Basque Verb Index*a *PropBank* eta *VerbNet* ereduak jarraitzen dituen euskarazko aditzen predikatu-lexikoia da. 244 aditz sarrera eta 364 adiera biltzen ditu. Bi eredu hauek jarraitzen dituela esaten dugunean, adierazi nahi dugu, besteak beste, lexikoiko aditzen adiera bakoitzerako argumentuek jokatzen dituzten rolak *PropBank* ereduko (arg0...arg5) etiketak eta *VerbNet* ereduko *rol tematikoak* erabilita markatu direla.

[5] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html

Gainera, predikatuen adieretarako ingelesezko *PropBank*eko etiketak erabili dira. Ondorengo adibidean euskarazko *hautatu* aditz-predikatuari BVIn dagokion sarrera ikus daiteke.

HAUTATU

select.01

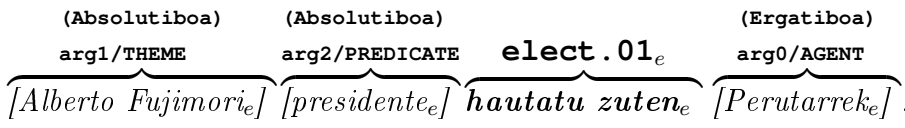
arg0: AGENT - Ergatiboa
arg1: THEME - Absolutiboa
arg2: SOURCE - Ablatiboa
arg3: ATTRIBUTE - Absolutiboa

elect.01

arg0: AGENT - Ergatiboa
arg1: THEME - Absolutiboa
arg2: PREDICATE - Absolutiboa

Hautatu aditzak bi adiera ezberdin ditu identifikatuta. Lehenengoa, arruntena, *norbaitek hainbat aukeraren artean bat hautatzeari dagokio*; bigarrena, berriz, hauteskunde testuinguruan erabiltzen dena da, *postu baterako hautagai bat aukeratzen denekoa.* BVI garatu zenean euskarazko aditzen adierak ingelesezko *PropBank*eko adierekin lotzea erabaki zen. Beste hizkuntzetarako garatu diren predikatu-lexikoietan egin izan den bezala. Ondorioz, euskarazko *hautatu* aditzaren lehen adiera ingelesezko *select* aditzaren lehen adierarekin (`select.01`) dator bat. *Hautatu* aditzaren bigarren adiera, aldiz, *elect* aditzaren lehen adierarekin (`elect.01`). Hizkuntzen arteko adieren *mapaketa* honek, *PropBank*eko adieren *RoleSet*ak zuzenean euskarara igarotzea ahalbidetu zuen. Adibidean ikus daiteke *PropBank*eko *core* rolak erabiltzeaz gainera, *BVI* garatzean *VerbNet* eredua ere jarraitu zela, *rol tematikoak* ere erabili zirela, alegia. Esan beharra dago euskararako garatutako BVI lexikoiak ezberdintasun nabarmena duela ingelesekoaren aldean, euskara morfologia aberatseko hizkuntza baita (*Morphologically rich language-MRL*): edozein rol jokatzen duen argumentuaren deklinabide kasua zehazten da.

(Ablatiboa) arg2/SOURCE: *[Aukeretatik$_e$]*
(Absolutiboa) arg1/THEME: *[txanpon aldaketa$_e$]*
izan zen
(Ergatiboa) arg0/AGENT: *[Fujimorik$_e$]*
select.01$_e$ ***hautatutakoa$_e$*** .



Ikus daitekeenez, BVIn *hautatu* aditzaren adierentzat zehazten diren deklinabide kasuak eta adibideko esaldietakoak bat datoz, lehen esaldian esate baterako: *Aukeretatik* argumentuaren deklinabide kasua ablatiboa da, *txanpon aldaketa* argumentuarena absolutiboa eta *Fujimorik*ena ergatiboa.

EPEC-RolSem

Tesian prestatutako SRL tresna garatzea ahalbidetu duen beste baliabidea *EPEC-RolSem* corpusa izan da. Hau BVI lexikoia erabilita *EPEC* corpusa (*Euskararen Prozesamendurako Erreferentzia Corpusa*) (Aduriz et al., 2006) predikatu-mailan etiketatzetik sortutako corpusa da. Izenak dioen bezala, *EPEC* IXA taldean[6] osatutako (eta oraindik ere osatze prozesuan den) euskararen prozesamendurako erreferentziazko corpusa da, euskara batuan idatzitako 300.000 hitzek osatzen dutena. Corpuseko testuen heren bat *UZEI Terminologia eta Lexikografia Zentroak*[7] egindako *XX. mendeko euskararen corpus estatistikotik* hartuak dira, eta beste bi herenak *Euskaldunon Egunkaria*tik. Lehen herenari dagokionez, *XX. mendeko euskararen corpus estatistikoa* delakoak osotara 4.658.036 hitz ditu eta horietatik 48.000 hartu ziren *EPEC* corpusa osatzeko, 1991tik eta 1999ra bitarteko testuetakoak. Beste bi herenei dagokienez, *Euskaldunon Egunkaria*k 1999ko urtarriletik 2000ko maiatzera bitartean argitaratutako ale guztietako testuak hartu ziren. *EPEC* corpusa hainbat hizkuntza mailatan etiketatuta dago; maila horietatik hiru dira gaur egun eskuz eta corpus guztian etiketatuta daudenak.

- **Segmentazioa/*EPEC-SEG***: Maila honetan corpuseko hitz bakoitzaren lema, kategoria, azpikategoria, postposizio atzizkia eta numeroa etiketatuta dago. Hau da, testuko hitzen analisi morfologikoa egina dago.

- **Sintaxia/*EPEC-DEP***: Dependentzietan oinarritutako etiketatzeari dagokion maila. Dependentzia gramatika (Tesnière, 1959) jarraituz, aditzen menpeko elementuak etiketatuta daude, dependentzia zuhaitzak osatzeko. Maila honetan eti-

[6] http://ixa.si.ehu.es/Ixa
[7] http://www.uzei.eus/en/

ketatutako *EPEC* corpusari *Basque Dependency TreeBank* deitu ohi zaio (Aduriz et al., 2003).

- **Semantika/*EPEC-Eusemcor***: Maila honetan corpuseko izen arruntak *Euskal WordNet*en (Pociello et al., 2011) dagozkien adierekin (*synset*) lotuta daude.

Semantika mailan, gainera, tesian erabili dugun *EPEC-Rolsem* predikatuen etiketatze mailako corpusa dago. Maila semantiko hau, dena den, ez da oraingoz *EPEC* corpus guztian zehar etiketatu *EPEC-SEG*, *EPEC-DEP* eta *EPEC-Eusemcor* mailekin egin denaz bestera. Izan ere *EPEC* osoan 1.211 aditz-predikatu identifikatu baldin badira ere, hauetatik BVIko 244 predikatuei dagozkien rol semantikoak baizik ez dira etiketatu, hain zuzen ere corpusean 30 agerraldi edo gehiago dauzkatenenak. Gainera, bestelako predikatuak (izen, adjektibo eta adberbio predikatuak) eta hauei dagozkien argumentu-adjuntu egiturak ez dira etiketatu. Atal honetan aipagarria da, halaber, *e-ROLda* web aplikazioa[8] (Estarrona, 2014). *e-ROLda*k, berez, SRL tresnen garapenerako baliabide linguistikoa ez bada ere, BVI lexikoia eta *EPEC-Rolsem* corpusa modu bisual eta egituratuan egikaritzeko aukera ematen du.

2.1.3 Ebaluazio saioak

Azpiatal honetan SRL atazarekin zerikusia daukaten ebaluazio saioak zerrendatuko ditugu. Orain arte izan diren saioetan ezagunenak eta atazaren ikerketan inpakturik handiena izan dutenak CoNLL (*Conference on Natural Language Learning*) kongresuaren barnean antolatutakoak izan direla uste dugu[9]. CoNLL ebaluazio saioak 1999. urtean hasi ziren, eta urtero-urtero hizkuntzaren prozesamenduarekin lotutako ataza desberdinak jorratu dituzte. Hauen helburua atazen ikerketa eta interes zientifikoa sustatu eta horietan erabilitako emaitza eta konfigurazioak (metrikak, datu multzoak...) erreferentziazko bihurtzea da. Zerrendatuko ditugun CoNLL saio guztiek *PropBank/NomBank* eredua erabili zuten. SRL atazaren gaineko edizioak ondorengoak izan dira:

- ***CoNLL-2004*** (Carreras eta Márquez, 2004): Ingelesez idatzitako testuetan rol semantikoen etiketatze automatikoaz arduratu zen lehenengo lehiaketa izan zen.

[8] http://ixa2.si.ehu.es/e-rolda/index.php?lang=eu
[9] http://www.signll.org/conll

Bertan bi modalitate eskaini zitzaizkien parte hartzaileei: *open challenge* eta *closed challenge* ebaluazioak. Lehenengoan, parte hartzaileek, eraiki beharreko SRL etiketatzaileak inplementatzeko orduan, ebaluazio saioan emandako entrenamendu corpusaz gainera beste baliabide linguistiko batzuk erabiltzeko aukera zuten. Bigarrenean, ordea, ebaluazio saioko entrenamendu corpusa baizik ezin erabil zezaketen. Osotara hamar taldek hartu zuten parte, denek *closed challenge* modalitatean. Emaitzarik onenak lortu zituen SRL etiketatzaileak (Hacioglu et al., 2004) 72.43 puntuko eta 66.77 puntuko doitasuna eta estaldura eta 69.49 puntuko F_1 balioa iritsi zituen.

Edizio honetan erabili zen adierazpen sintaktikoa *partziala* eta osagaietan oinarritua izan zen. Gainera, aditz-predikatuak baizik ez ziren landu. Entrenamendurako eta ebaluaziorako *The Wall Street Journal Corpus-WSJ* corpusa erabili zen.

- ***CoNLL-2005*** (Carreras eta Márquez, 2005): Hurrengo urteko edizioan ere ingelesezko rol semantikoen etiketatze automatikoaz arduratu zen CoNLL ebaluazio saioa. Aurreko urtekoaren aldean hainbat ezberdintasun izan zituen, esanguratsuenak hauek: informazio sintaktiko aberatsagoa erabili zela (osagaietan oinarritutako zuhaitz sintaktiko osoa ematen zen), eta ebaluaziorako hiru corpus erabili zirela. Ondorengoak izan ziren ebaluaziorako corpusak: (1) Entrenamendurako, WSJ corpusetik sortutakoa, (2) Brown corpusetik[10] sortutakoa (WSJren domeinu berekoa), eta (3) WSJ eta Brown ebaluazio corpusak elkartzetik osatutakoa. Aldaketa hauen helburua bikoitza izan zen: lehenik informazio sintaktikoaren SRLren gaineko onura neurtu nahi zen, eta bigarrenik rol semantikoen etiketatzaileek, entrenamenduko corpusa ez zena etiketatzean, jasaten zuten eraginkortasunaren galera zenbaterainokoa zen jakin nahi zen. *CoNLL-2005* ebaluazio saioan aditz-predikatuak besterik ez ziren landu.

  Edizio honetan ere *open challenge* eta *closed challenge* modalitateak eskaini baziren ere parte hartu zuten hemeretzi sistemek *closed challenge* modalitatean jardun zuten. 2.1 taulan biltzen dira emaitzarik onenak lortu zituen sistemaren emaitzak (Punyakanok et al., 2004).

[10] `http://clu.uni.no/icame/brown/bcm.html`

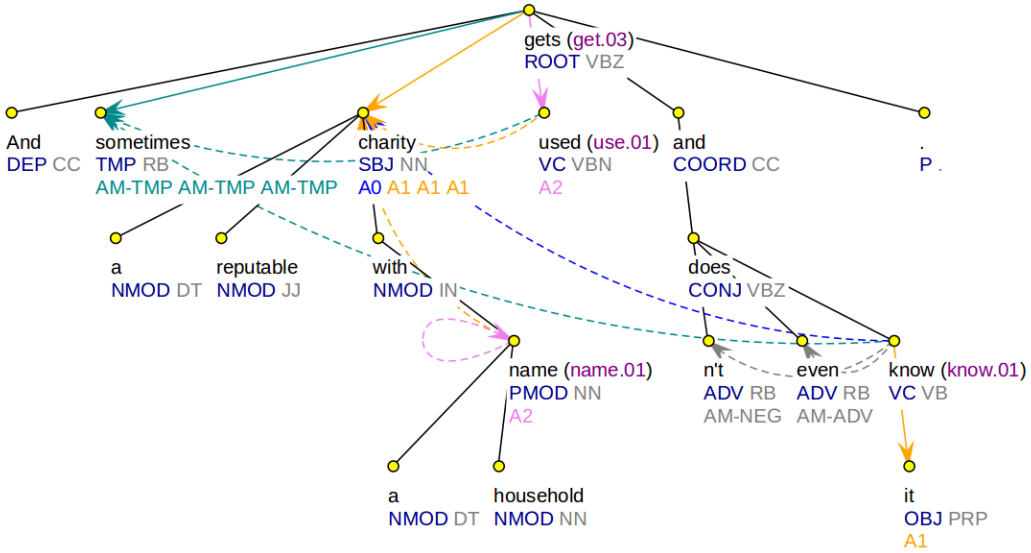
		Doitasuna	Estaldura	F1-Scorea
1	**WSJ**	82.28	76.78	79.44
2	**Brown**	73.38	62.93	67.75
3	**WSJ+Brown**	81.18	74.92	77.92

Taula 2.1: *CoNLL-2005* ebaluazio saioko emaitzarik onenak.

*CoNLL-2005*eko emaitzak *CoNLL-2004*koekin alderagarriak ez badira ere (bi saioetako konfigurazioak ezberdinak izateagatik), argi ikusten da bateko eta besteko emaitzarik altuenak alderatzean (WSJ) sintaxiaren eragina nabarmena dela. Ia 10 puntuko hobekuntza dago F_1-Scorean (69.49 eta 79.44). Hobekuntza honen kausa nagusia informazio sintaktikoaren aberastasunean egon arren, *CoNLL-2005* edizioan erabilitako entrenamendu corpusaren tamainak ere (*CoNLL-2004* ediziokoa baino handiagoa) eragina dauka. Gainera, 2.1 taulako WSJ lerroa, Brown eta WSJ+Brown lerroekin alderatzean ikusten dugu SRL tresnen eraginkortasuna jaitsi egiten dela entrenamendurako erabili ez den corpusak etiketatzerakoan. Hau arazo ezaguna da ikasketa automatikoan eta hizkuntzaren prozesamenduan.

- ***CoNLL-2008*** (Surdeanu et al., 2008): CoNLL ebaluazio saioaren edizio honek SRLren ikerketaren egungo egoeran bigarren etapa ireki zuen, dependentzia sintaktiko eta semantikoena hain zuzen ere. Aurreko bi ebaluazio saioen aldean ezberdintasun nabarmenak dauzka. Garrantzizkoena, dependentzietan oinarritutako formalismoa hartzen duela. Modu honetara, SRL ataza esaldiak osatzen dituzten tokenen arteko erlazio semantikoak ezartzean datza. Formalismo honek, gainera, dependentzia gramatika jarraitzen du sintaxiaren adierazpenerako. Ondorioz, analisi sintaktiko-semantiko konputazionala ataza bakar batean biltzea lortzen du, *dependency parsing* izenekoan, hain zuzen. Tesian garatu dugun euskararako SRL etiketatzaileak formalismo hau jarraitzen du, gaur egungo sistemetan eskemarik erabiliena delako. Hori dela eta gure SRL sistemari dependentzia *parser*a dei dakioke. Argitu beharra daukagu, berez, SRL ataza dependentzia semantikoen identifikazioari baizik ez badagokio ere (`SRL-only`), ez beraz dependentzia sintaktikoen identifikazioari, ataza burutu ahal izateko informazio sintaktikoa ezinbestekoa dela, eta horregatik rol semantikoen etiketatze automatikoa dependentzia sintaktikoak eta semantikoak, biak, identifikatzean datzala.

2.2 irudiko adibidean *And sometimes a reputable charity with a household name gets used and doesn't even know it* esaldiaren dependentzia sintaktiko eta semantikoen zuhaitza ikus daiteke. Bertan izen eta aditz predikatuen *PropBank/Nombank* adierak (`name.01` eta `get.03`, `use.01`, `know.01`) eta hauei dagozkien argumentu eta adjuntuak etiketatu dira. Gainera, dependentzia sintaktikoak ere etiketatuta daude.



Irudia 2.2: Dependentzia sintaktiko-semantikoen zuhaitza (Hajič et al., 2009).

Aipatutako formalismoa erabiltzeaz gainera, *CoNLL-2008*n beste berrikuntza batzuk ere izan ziren. Esate baterako, lehenengo aldia izan zen izen-predikatuak eta hauei zegozkien argumentuak etiketatzeko eskatzen zena. Gainera, sistemen ebaluaziorako *Labeled Attachment Score-LAS*, *Labeled* F_1 eta *Labeled Macro* F_1 *Score* metrikak erabili ziren, hurrenez hurren sintaxiaren, semantikaren eta sistema osoaren (*sintaxia* + *semantika*) eraginkortasuna neurtzeko. Entrenamenduarako eta ebaluaziorako testuak direla-eta, *PropBank*, *NomBank* eta *Penn Treebank 3* (Marcus et al., 1993) corpusak erabili ziren. 2.1.2 azpiatalean aipatu dugun bezala, lehenengo biak *WSJ* corpuseko testuz osatuta daude, hirugarrena, berriz, *WSJ* corpusa eta *Brown* corpusaren sail baten elkarketa da. *CoNLL-2008*n, *WSJ*, *Brown* eta *WSJ+Brown* corpusak erabili ziren sistemen ebaluaziorako. Aurreko edizioetan bezala, honetan ere ingelesa bakarrik landu zen, eta *open challenge* eta

closed challenge modalitateak eskaini ziren. Hemeretzi taldek hartu zuten parte *closed challenge*an eta bostek *open challenge*an. Edizio honetako emaitzak aurreko bietako emaitzekin alderagarriak ez badira ere (corpus desberdinak, beste metrika batzuk... erabili direlako), *closed challenge* modalitatean, emaitzarik onenak lortu zituen sistemaren emaitzak (Johansson eta Nugues, 2008) baliagarriak izan daitezke SRLren ikerketaren egungo egoera garaian zein zen jakiteko (2.2 taulan).

		LAS	**Labeled F_1**	**Labeled Macro F_1**
1	**WSJ**	90.13	81.75	85.95
2	**Brown**	82.81	69.06	75.95
3	**WSJ+Brown**	89.32	80.37	84.86

Taula 2.2: *CoNLL-2008*ko *closed challenge* emaitzarik onenak.

- ***CoNLL-2009*** (Hajič et al., 2009): Hau izan zen SRL atazaz arduratu zen azkeneko CoNLL ebaluazio saioa. *CoNLL-2008* edizioak bezala, dependentzietan oinarritutako formalismoa jarraitzen zuen sintaxiaren eta semantikaren adierazpenerako. Bestalde, eta aurreko edizioan ez bezala, parte hartzaileei ingeleserako ez ezik, beste sei hizkuntzetarako ere dependentziak etiketatzeko gaitasuna zuten SRL sistema edo *dependency parser*ak gauzatzeko eskatu zitzaien. Hauek izan ziren hizkuntzak: katalana, gaztelera, ingelesa, alemana, txekiera, txinera eta japoniera.

  Sistemen eraikuntzarako eta ebaluaziorako erabili ziren corpusak hizkuntza guztietarako tamainak ahalik eta antzekoenak, hots, orekatuenak, izaten ahalegindu ziren. Hainbat *orekatze* edo *zuzenketa* egin ziren helburu hori betetzeko. Japonieraren kasuan, adibidez, entrenamendurako eskuragai zegoen dependentzia semantikoen corpus bakarra, *Kyoto University Text Corpus*a (Kawahara et al., 2002), beste hizkuntzetakoa baino askoz ere mugatuagoa zen. Horregatik, eta sintaxiak SRLn duen eragina aintzat harturik, dependentzia sintaktikoekin etiketatuta zegoen entrenamendurako beste corpus bat erabiltzeko aukera izan zen, beste hizkuntzetakoa baino handiagoa. Hizkuntza guztietan landu ziren izen eta aditz predikatuak eta hauen argumentu-adjuntu egiturak, gazteleran, katalanean eta alemanean izan ezik. *open challenge* eta *closed challenge* modalitateak eskaini ziren eta parte hartu zuten hogei taldeetatik bik bakarrik hartu zuten parte *open challenge*an.

2.3 taulan *CoNLL-2009* ebaluazio saioan hizkuntza bakoitzerako eta metrika bakoitzean lortu ziren emaitzarik altuenak biltzen dira. Emaitza hauek *closed challenge* modalitateari dagozkio. Ebaluaziorako erabilitako corpusak entrenamendurako erabilitako berak dira (*in-domain*). Taularen emaitzek hizkuntzen araberako SRL atazaren zailtasuna erakusten dute. Argi ikus daiteke hizkuntzek daukaten tipologiaren eta jatorriaren arabera balio batzuk edo beste lortzen dituztela. Alde batetik, ingelesaren eta alemanaren emaitzak oso antzekoak dira, biak hizkuntza germaniarrak direlako, eta beste horrenbeste agitzen da katalanaren eta gaztelerarenen emaitzekin, biak ere latin hizkuntzak direlako.

	Katalana	Txinera	Txekiera	Ingelesa	Alemana	Japoniera	Gaztelera
LAS	87.86 (2)	79.17 (5)	80.38 (2)	89.88 (1)	87.48 (1)	92.57 (3)	87.64 (2)
Labeled F1	80.10 (4)	77.15 (3)	86.51 (3)	86.15 (4)	78.61 (3)	78.26 (3)	80.29 (4)
Labeled Macro F1	83.01 (4)	76.38 (3)	83.27 (3)	87.69 (4)	82.44 (3)	85.65 (3)	83.31 (4)

Taula 2.3: *CoNLL-2009* saioko *closed challenge*, *in-domain*, emaitzarik onenak [(1): (Bohnet, 2009), (2): (Gesmundo et al., 2009), (3): (Che et al., 2009), (4): (Zhao et al., 2009), (5): (Ren et al., 2009)].

2.2 SRL arkitektura

Azpiatal honetan rol semantikoak automatikoki etiketatzen dituzten gaur egungo sistemen ohiko egitura azalduko dugu. Tesian garatu dugun euskararako SRL etiketatzailea *dependency parser* motakoa da, dependentzietan oinarritutako formalismoa erabiltzen duelako, formalismo hau erabiltzearen arrazoia beste hizkuntzetan izan duen erabilera izan da. Atal honetan, beraz, dependentzia *parser*en arkitektura aurkeztuko dugu. Bukaeran, gainera, sistema hauek itzultzen dituzten zuhaitz sintaktiko eta semantikoen ebaluazioa egiteko erabiltzen diren *Labeled Attachment Score*, *Labeled* F_1 eta *Labeled Macro* F_1 *Score* metriken azalpena egingo dugu.

Dependentzia *parser*ek testuak edo esaldi soilak jasotzen dituzte sarreratzat, eta etiketatutako zuhaitz sintaktiko-semantikoak itzultzen dituzte irteeratzat (ikus 2.2 irudia). Bost urratsetan banatu ohi da dependentzietan oinarritutako SRL prozesua: (1) dependentzia sintaktikoen etiketatzea, (2) predikatuen identifikazioa, (3) predikatuen desanbiguazioa, (4) argumentuen identifikazioa eta (5) argumentuen sailkapena. Aurreko azpiatalean argitu dugu SRL ataza, berez, dependentzia semantikoen identifikazioari bai-

zik ez badagokio ere, ataza burutu ahal izateko informazio sintaktikoa ezinbestekoa dela eta, horregatik, rol semantikoen etiketatze automatikoa dependentzia sintaktikoak eta semantikoak, biak, identifikatzean datzala. Aipatu ditugun bost urratsetatik azkeneko laurak (2-5) dira berezko SRL prozesuari dagozkionak. Dependentzia sintaktikoen urratsa (1) azpiurratsetan bana bagenezakeen ere, iruditu zaigu ez dela beharrezkoa burutu dugun lana ulertzeko, gauzak urratsez urrats eta xehe-xehe azaltzea.

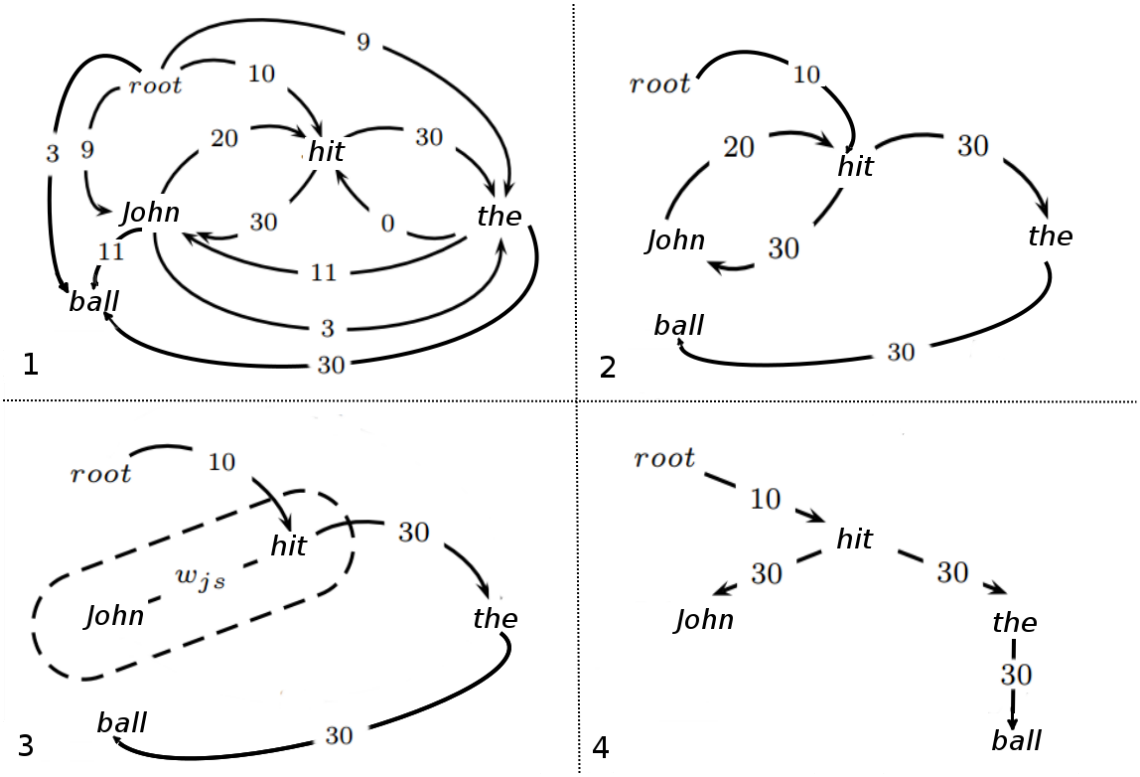
2.2.1 Bost urratsetako prozesua

Rol semantikoen etiketatze automatikoaren helburua esaldietako predikatuen argumentuek haien predikatuen aldera jokatzen duten rola zehaztea da. Hau betetzeko aipatu berri ditugun bost urratsak erabiltzen dira. Horiek azalduko ditugu jarraian:

1. **Dependentzia sintaktikoen etiketatzea**: Bi hurbilpen ezberdin erabil daitezke esaldiak osatzen dituzten tokenen arteko dependentzia sintaktikoak zehazteko: trantsizioetan oinarritutakoa (*transition-based dependency parsing*) (Nivre, 2003) eta grafoetan oinarritutakoa. Bigarren hurbilpenean gehienetan erabiltzen den metodoa hedadura maximoko zuhaitzetan oinarritutakoa da (*Maximum Spanning Tree/MST-based dependency parsing*) (McDonald eta Pereira, 2006).

   2.3 irudiak *John hit the ball* (Johnek pilota jo zuen) esaldiaren hedadura gehieneko zuhaitzetan oinarritutako dependentzia sintaktikoen analisia erakusten du. Esan bezala, hau da grafoetan oinarritutako dependentzia sintaktikoetan metodorik erabiliena, eta lan honetan jarraitzen dena. Metodo honetan x esaldi bakoitzerako $G_x = (V_x, E_x)$ grafo zuzendua definitzen da.

   V_x multzoa esaldiko tokenek eta `root` elementu *artifizialak* osatzen dute, hauek grafoaren erpinak dira. `root` etiketa analisiaren ondorioz sortuko den zuhaitz sintaktikoaren erroa izango da. E_x multzoa, berriz, grafoaren ertzek osatzen dute. Analisiaren hasierako egoeran erpin bakoitzetik irten eta gainerako erpin guztietara iristen diren ertz bideratuek osatuko dute E_x multzo hau. Ertz bakoitzari analizatzaileak probabilitate bat esleituko dio eta ondoren erpin loturagaberik izan gabe probabilitate guztien batura maximoa izatea ahalbidetzen duten ertzak bakarrik mantenduko dira. Trantsizioetan oinarritutako analisian ez bezala, honetan maximo globala bilatzen da. Gerta daiteke probabilitate maximoa lortzeko man-



Irudia 2.3: *Maximum Spanning Tree dependency parsing* adibidea.

tendu diren ertzek zikloak osatzea, hau da, e_1 erpinetik e_2 erpinera eta e_2 erpinetik e_1 erpinera doazen ertzak izatea. Helburua zuhaitz sintaktikoa sortzea denez eta zuhaitz egiturek ziklorik onartzen ez dutenez, zikloak desegin egiten dira. Horretarako, MST oinarriko analisian erabili ohi den *Chu-Liu-Edmonds* algoritmoak (Chu eta Liu, 1965) zikloa osatzen duten erpinak erpin bakartzat hartu eta ertz guztien probabilitateak berriz ere kalkulatzen ditu. Horrela, dependentzietan oinarritutako zuhaitz sintaktikoa lortzen da. 2.3 irudiko adibidean pauso hauek eman dira:

- (1) $G_x = (V_x, E_x)$ osatu da, non $V_x = \{\texttt{root}, John, hit, the, ball\}$ eta $E_X = \{\texttt{root} \rightarrow hit(10), \texttt{root} \rightarrow John(9), ..., the \rightarrow John(11)\}$ multzoak diren.
- (2) Erpin loturagaberik izan gabe osotara zuhaitzaren probabilitatea maximoa izan dadin erabiltzen diren ertzak mantendu eta gainerakoak kendu dira. $E_X = \{\texttt{root} \rightarrow hit(10), John \rightarrow hit(20), hit \rightarrow John(30), hit \rightarrow the(30), the \rightarrow ball(30)\}$

- (3) *John* eta *hit* erpinen artean zikloa dagoenez gero ($John \rightarrow hit(20)$ eta $hit \rightarrow John(30)$), bi erpin hauek bakartzat hartu eta lehen urratsa errepikatzen da, zikloa desagerrarazteko.
- (4) Bigarren urratsa errepikatu eta ziklorik gabeko dependentzia zuhaitza lortzen da. $V_x = \{\texttt{root}, John, hit, the, ball\}$ eta $E_X = \{\texttt{root} \rightarrow hit(10), hit \rightarrow John(30), hit \rightarrow the(30), the \rightarrow ball(30)\}$

2. **Predikatuen identifikazioa**: Rolen etiketatzea egiteko bigarren urratsa da hau, eta dependentzia semantikoak lortzeko lehenengoa. Identifikazioa ikasketa automatikoko metodoekin egiten da normalean. Horretarako, esaldi baten tokenak eta haien gaineko informazio linguistikoa emanda (analisi sintaktikoa, morfologikoa, e.a.) token bakoitza predikatua den edo ez erabakitzen duen sailkatzaile bitarra eraikitzen da. Badago aukera token guztiak zuzenean sailkatzaileari eman baino lehen *garbiketa prozesua* deritzana egiteko. Bertan heuristikoen bitartez eta eskura dagoen informazioa erabilita, predikatuak ezin izan daitezkeen tokenak baztertzen dira. Adibidean 2.2 irudiko esaldiaren predikatuen identifikazioa ikus daiteke.

And sometimes a reputable charity with a household name gets used and doesn't even know it.

↓

(1) ~~*And*~~ *sometimes* ~~*a*~~ *reputable charity* ~~*with a*~~ *household name gets used* ~~*and*~~ *doesn't even know* ~~*it.*~~

↓

(2) ~~*And*~~ $\overbrace{\textit{sometimes}}^{?}$ ~~*a*~~ $\overbrace{\textit{reputable}}^{?}$ $\overbrace{\textit{charity}}^{?}$ ~~*with a*~~ $\overbrace{\textit{household}}^{?}$ $\overbrace{\textit{name}}^{?}$ $\overbrace{\textit{gets}}^{?}$ $\overbrace{\textit{used}}^{?}$ ~~*and*~~ $\overbrace{\textit{doesn't}}^{?}$ $\overbrace{\textit{even}}^{?}$ $\overbrace{\textit{know}}^{?}$ ~~*it.*~~

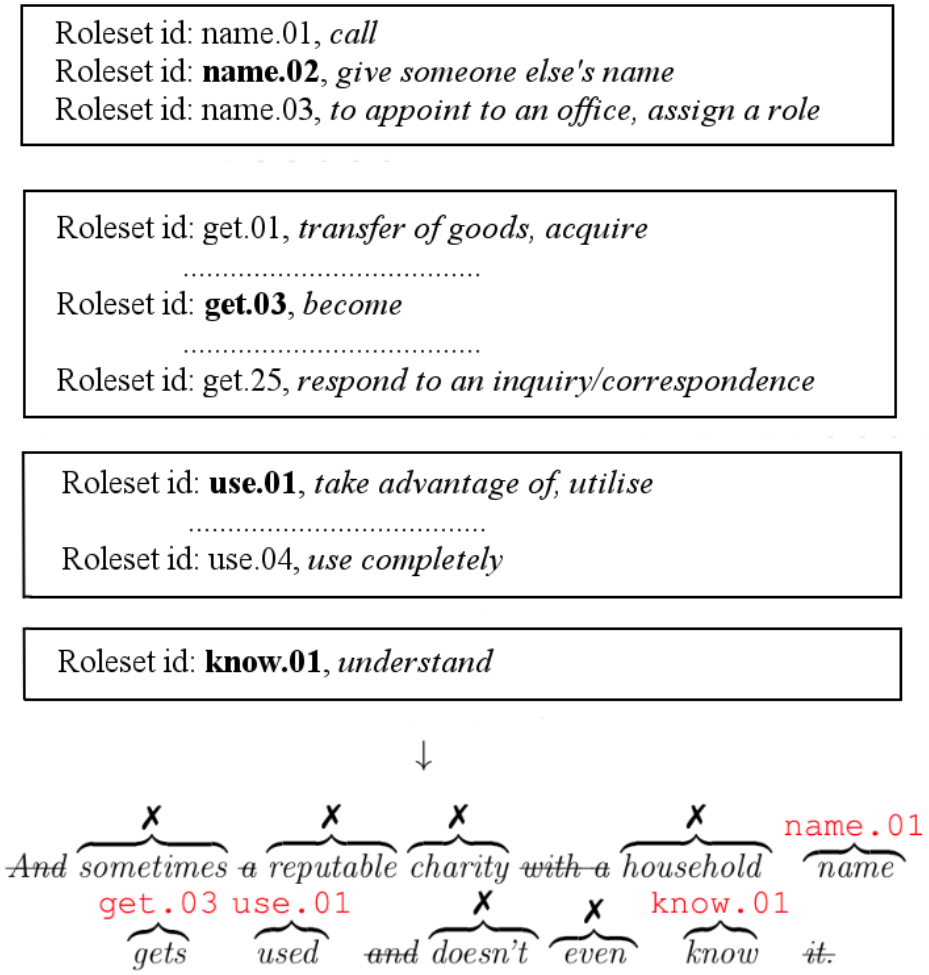
↓

(3) ~~*And*~~ $\overbrace{\textit{sometimes}}^{\times}$ ~~*a*~~ $\overbrace{\textit{reputable}}^{\times}$ $\overbrace{\textit{charity}}^{\times}$ ~~*with a*~~ $\overbrace{\textit{household}}^{\times}$ $\overbrace{\textit{name}}^{\checkmark}$ $\overbrace{\textit{gets}}^{\checkmark}$ $\overbrace{\textit{used}}^{\checkmark}$ ~~*and*~~ $\overbrace{\textit{doesn't}}^{\times}$ $\overbrace{\textit{even}}^{\times}$ $\overbrace{\textit{know}}^{\checkmark}$ ~~*it.*~~

Lehenengo pausoan garbiketa prozesua aplikatu da (1). Honek, PoS (*Part-of-Speech*) kategoriari erreparatuta, predikatuak izan ezin daitezkeen tokenak baztertu ditu, kasu honetan *and* juntagailuak, *with* preposizioa, *a* determinatzaileak eta *it* izenordaina. Ondoren, gainerako tokenak predikatu identifikazioaz arduratzen den sailkatzaile bitarrari eman zaizkio (2). Sailkatzaileak, tokenen gaineko informazioa eta *PropBank* eta *NomBank* bezalako lexikoiak erabilita, predikatuak diren eta ez diren tokenak iragarri ditu. Adibidean *name*, *gets*, *used* eta *know* aditzezko predikatuak identifikatu ditu (3).

3. **Predikatuen desanbiguazioa**: SRL prozesuaren hirugarren urratsak predikatutzat identifikatu diren tokenak jasotzen ditu sarreratako. Urrats honetan predikatu hauei dagozkien adierak zehazten dira horretarako predikatu-lexikoiak erabilita. Ingeleserako garatzen diren egungo sistema ia guztietan erabiltzen diren lexikoiak *PropBank* eta *NomBank* dira. Gainerako hizkuntzetarako SRL tresnak garatzeko orduan ere *PropBank* eredua jarraitzen duten lexikoiak erabiltzen dira. Predikatu batek adiera bakarra edo bat baino gehiago izan ditzake lexikoian; adiera bakarra izatekotan, predikatuari adiera hori esleitzen zaio zuzenean. Adiera bat baino gehiago izatekotan, ordea, sailkatzaile batek arduratu behar du kasu bakoitzean predikatuari dagokion adiera zein den erabakitzeaz.

   2.4 irudiak predikatuen identifikazioan adibidetako erabili dugun esaldiko predikatuen desanbiguazio prozesua erakusten du. Bertan ikus daitekeen bezala, identifikatutako lau aditzetatik batek baizik ez du adiera bakarra *PropBank*en, `know` aditzak hain justu. Ondorioz, honi `know.01` adiera esleitzen zaio automatikoki. Gainerako hiru kasuek adiera bat baino gehiago onartzen dute, eta hauetarako ikasketa automatikoa erabiltzea beharrezkoa izaten da. Bi aukera daude ikasketa algoritmoak aplikatzeko garaian: predikatu bakoitzarentzat aparteko sailkatzaile bat eraikitzea, edo adiera anitzeko predikatu guztientzat sailkatzaile bakarra sortzea. Lehen teknikak emaitza hobeak izaten ditu, oro har, baina duen arazoa da eraginkorra izan dadin hagitz corpus handiak behar izaten direla. Adibidean sailkatzaileak *name* aditzari lehenengo adiera esleitu dio, `name.01` (*norbait edo zerbait izendatu, izena jarri*); *get* predikatuari, ordea, `get.03` adiera ezarri dio (*norbait edo zerbait nolabaitekoa bihurtzea, bilakatzea*), eta *use*ri `use.01` (*zerbait erabili, probestu*).

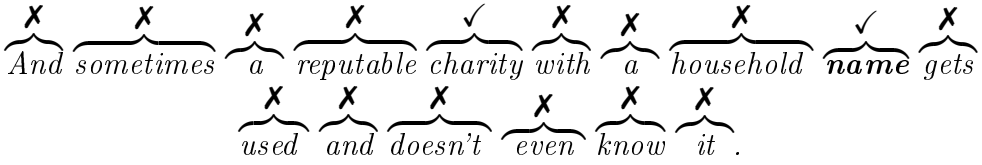


Irudia 2.4: Predikatuen desanbiguazioaren adibidea.

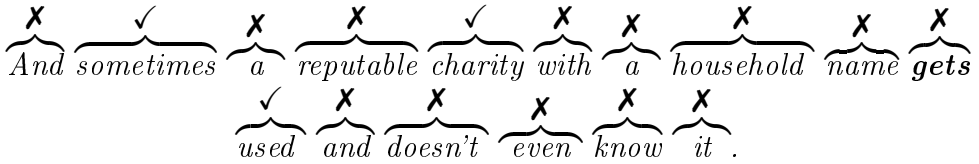
4. **Argumentuen identifikazioa**: Hau azken-aurreko urratsa da. Bertan, predikatu bakoitzari dagozkion argumentuen buru lexikalak identifikatzen dira eta, horretarako, ohikoena heuristikoak edo sailkatzaile bitarrak erabiltzea izaten da. Esaldiko predikatu bakoitzarentzat, esaldiko token bakoitza haren argumentua den edo ez erabaki behar izaten da. Berez argumentu bat hainbat tokenez osatuta egon badaiteke ere, SRL sistemek argumentuen buru lexikalak besterik ez dituzte etiketatzen. Buru lexikalaz gain argumentu bat zer beste tokenek osatzen duten jakiteko, nahikoa izaten da lehen urratseko analisi sintaktikoari erreparatu eta buru lexikalaren menpeko azpi-zuhaitza eratzen duten tokenak zuhaitzak ezartzen dituen dependentzien arabera antolatzea.

Hurrengo adibidean aurretik erabili dugun esaldiko predikatuen argumentuen identifikazio prozesua ikus daiteke.

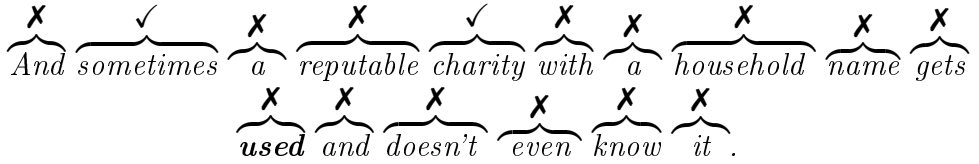
name



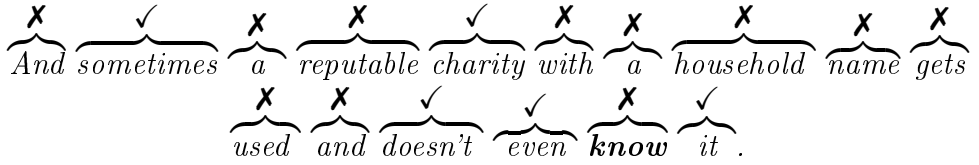
gets



used



know



Adibideetan ikusten da esaldiko lehenbiziko predikatuarentzat, *name*rentzat, bi argumenturen buru lexikalak identifikatu direla: *charity* eta *name* bera. Lehen argumentu osoa *a reputable charity with a household name* da, eta bigarrena berriz *a household name*. Predikatu bat bere buruaren argumentua izatea (*name*) ez da

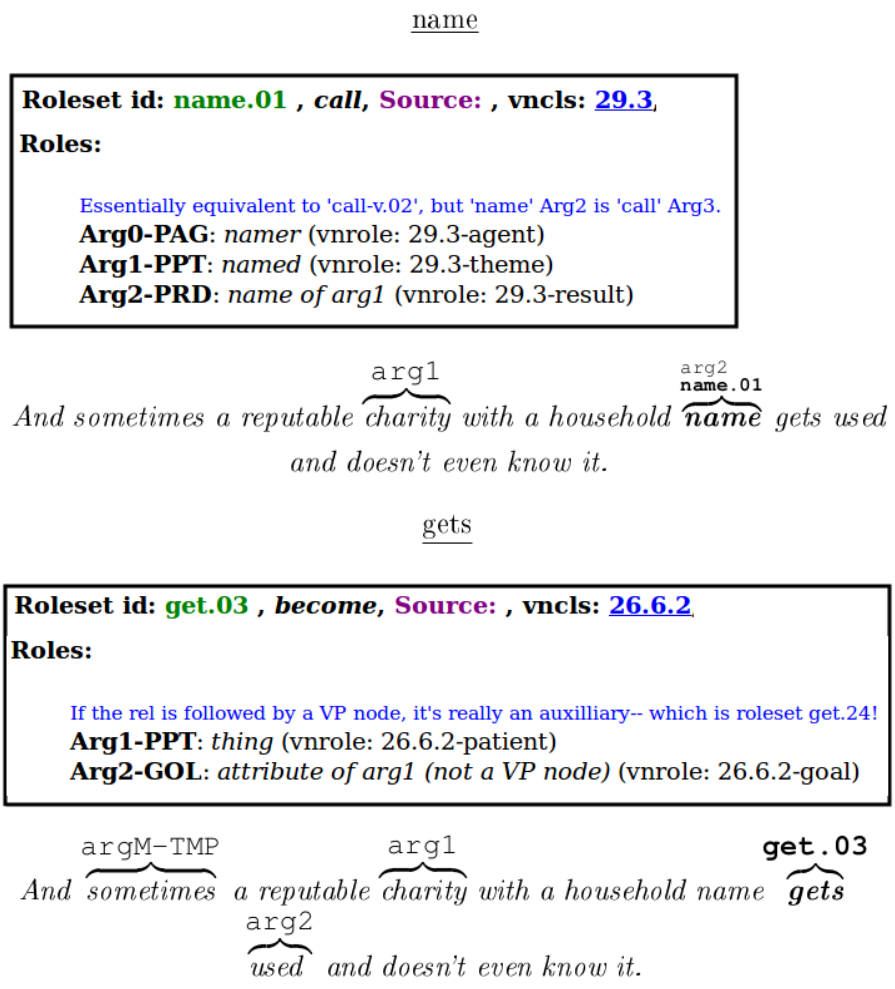
arrunta, baina hizkuntza batzuetan, ingelesean esate baterako, mintzaira honetako *PropBank* anotatzerakoan hartu ziren erabakien ondorioz gerta daiteke. Bigarren predikatuarentzat, *gets* aditzarentzat, hiru argumenturen buru lexikalak identifikatu direla ikusten da: *sometimes*, *charity* eta *used*. Hiru argumentu osoak *sometimes*, *a reputable charity with a household name* eta *used* aditza dira. Hurrengo predikatuarentzat (*used*) bi argumentu identifikatu dira, *sometimes* eta *charity*. Azkenik, *know* aditzaren lau argumentu eskuratu dira: *sometimes*, *charity*, *doesn't* eta *even*.

5. **Argumentuen sailkapena**: Hau da SRL atazaren azkeneko urratsa. Honetan predikatu bakoitzarentzat identifikatu diren argumentuen rolak zein diren erabakitzen da. Argitu beharra daukagu aurreko urratsari *argumentuen identifikazioa* deitzen zaion arren, adjuntuak ere identifikatzen direla. Izan ere, *argumentuen sailkapena* deritzan urrats honetan erabakitzen da aurreko pausoan identifikatu diren buru lexikaletatik zein diren argumentuenak eta zein adjuntuenak. Argumentuen (eta adjuntuen) sailkapena egiteko ikasketa automatikoa erabiltzen da. *Multiclass* motako sailkatzailea (klase anitz esleitzen dituena) entrenatzen da horretarako. Sailkatzaile hauek *PropBank* eta *NomBank* corpusetatik, edo eredu hau jarraitzen duten beste hizkuntza batzuetako corpusetatik, ikasi ohi dute normalean.

   Hizkuntzaren arabera, eta horretan erabilitako corpusaren arabera, gerta daiteke predikatu baten adiera batek rol bera jokatzen duten bi argumentu ezberdin jasotzea, ingelesez gertatzen den bezala. Hala ere, hau jazotzea ezinezkoa den hizkuntzetan, euskaran adibidez, inkoherentzia hauek sailkatzailearen emaitzak jaso eta gero zuzendu behar izaten dira. Zuzenketa hau egiteko erabil daitekeen prozeduretako bat ILP teknika (*Integer Linear Programming*) da. ILP optimizazioan, predikatuen argumentuek rol (edo adjuntu) etiketa bakoitza jokatzeko daukaten probabilitatea itzultzen duen sailkatzaile bat erabilita, predikatu-argumentu-adjuntu egituren probabilitate maximoko rol semantikoen konbinazioa bilatzen da.

   2.5 eta 2.6 irudietan aurreko adibideko predikatuen argumentu eta adjuntuen rol semantiko eta adjuntu-etiketen esleipena ikus daiteke. Sailkatzaileak *name* aditzarentzat identifikatutako buru lexikaletatik biak argumentuei dagozkiela (eta ez adjuntuei) erabaki du; izan ere, *charity*ri `arg1` esleitu dio, eta `name`i `arg2`. 2.5 irudian ikus daiteke *PropBank*en arabera `arg1` rola *izena jaso duen entitateari* dagokiola, eta `arg2` izenari berari. Kasu honetan izena ez da esplizituki aipatzen,

baina *name* buru lexikala daukan *a household name* argumentuak hari egiten dio erreferentzia. Horregatik jasotzen du `arg2`. Euskararako erabili dugun corpusak ez du horrelako kasurik, predikatu batek ezin du bere buruaren argumentua izan.



Irudia 2.5: Argumentuen sailkapenaren adibidea (*name* eta *gets*).

Bigarren predikatuarentzat (*gets*) bi argumentu eta adjuntu bat etiketatu dira: *charity*k `arg1` jokatzen du eta *used* aditzak `arg2`, gainera, *sometimes* adjuntua (`argM-TMP`) etiketatu da. Predikatu bat beste baten argumentua izatea ohikoa da. Kasu honetan *charity*/`arg1` *probestua izan den entitateari* eta *used*/`arg2` aldiz `arg1`en atributuari dagozkiola ikus dezakegu, 2.5 irudiko *PropBank* sarreran. Ingelesez `get.03` adieraren `arg2` argumentua `arg1` jokatzen duen argumentuaren atribututzat ikusten da, adiera honetan *get* aditzak duen erabilerarengatik.

used

Roleset id: use.01 , *Take advantage of, utilise*, Source: , vncls: 66 105 54.3

Roles:

Arg0-PAG: *User* (vnrole: 66-agent, 105-agent, 54.3-agent)
Arg1-PPT: *thing used* (vnrole: 66-asset, 105-theme, 54.3-value)
Arg2-PRP: *purpose* (vnrole: 105-predicate, 66-goal, 54.3-location)

argM−TMP arg1 use.01
And sometimes a reputable charity with a household name gets used and doesn't even know it.

know

Roleset id: know.01 , *understand*, Source: , vncls: 29.5-1 29.9-1-1 29.2-1
Roles:

Arg0-PAG: *knower* (vnrole: 29.5-1-agent, 29.9-1-1-agent, 29.2-1-agent)
Arg1-PPT: *thing known or thought* (vnrole: 29.5-1-theme, 29.9-1-1-theme, 29.2-1-theme)
Arg2-PRD: *attributive of arg1 (known as, or known about)*

argM−TMP arg0
And sometimes a reputable charity with a household name gets used
argM−NEG argM−ADV know.01 arg1
and doesn't even know it .

Irudia 2.6: Argumentuen sailkapenaren adibidea (*used* eta *know*).

Bestalde, *used* predikatuarentzat *sometimes* denbora adjuntua (`argM-TMP`) eta *charity* (`arg1`) argumentua etiketatu ditu sailkatzaileak. Azkenik, *know* aditzak hiru adjuntu eta bi argumentu dauzkala ikus daiteke 2.6 irudian.

Aurrera jarraitu aurretik argitu beharra daukagu predikatu-lexikoietan adiera bakoitzarentzat zehazten diren argumentuek ez dutela zertan beti agertu testuetan dituzten gauzatzeetan. *know* aditzaren kasuan (`know.01`), esate baterako, `arg0`, `arg1` eta `arg2` rolak jasotzen dituela azaltzen da, baina adibidetako erabili dugun esaldian ikus daiteke `arg2` rola jokatzen duen argumentua ez dela agertzen. Adjuntuei dagokienez, ordea, ez dago adiera bakoitzak izan ditzakeen adjuntuen zerrendarik eta, beraz, printzipioz behintzat, edozein adjuntu jaso ditzakete.

2.2.2 Ebaluaziorako metrikak

Azpiatal honetan dependentzia etiketatzaileen eraginkortasuna neurtzeko erabiltzen diren *Labeled Attachment Score*, *Labeled* F_1 eta *Labeled Macro* F_1 *Score* metriken aurkezpena eginen dugu.

- *Labeled Attachment Score-LAS*: Dependentzia sintaktikoak etiketatzerakoan sistemak izan duen eraginkortasuna adierazten du. Zehazkiago esan, esaldi bat eta dagokion zuhaitz sintaktikoa emanda, token bakoitzari zuhaitzean dagokion buruko tokena (`HEAD`), eta tokenaren eta buruaren arteko erlazio sintaktiko mota (`DEPREL`) ongi definituta daukaten tokenen ehunekoa da. Buruko tokena edo erlazio sintaktiko mota gaizki etiketatuta badago erlazio guztia gaizki dagoela ulertzen da. Eman dezagun s_1 eta s_2 esaldien dependentziak etiketatu direla eta lehenengoak 10 token dituela, eta bigarrenak 45. Pentsa dezagun, gainera, etiketatu diren dependentzia hauetatik burua eta etiketa ongi eskuratu direla $s_1$eko 9 tokenentzat eta $s_2$ko 15 tokenentzat. Orduan bi aukera izango lirateke LAS neurria kalkulatzeko: (1) LAS *micro* edo (2) LAS *macro* neurriak.

  1. LAS *micro*: tokenetan oinarrituta kalkulatzen da LAS metrika. Adibidean $(9+15)/(10+45) = 0.436$ egingo litzateke, alegia $100 * 0.436 = \%43.6$ izango litzateke sistemaren eraginkortasuna, dependentzia sintaktikoak eskuratzeko orduan.
  2. LAS *macro*: esaldietan oinarrituta kalkulatzen da LAS metrika. Adibidean $(9/10 + 15/45)/2 = 0.617$ egingo litzateke, alegia $100 * 0.617 = \%61.7$ izango litzateke sistemaren eraginkortasuna, dependentzia sintaktikoak eskuratzeko orduan.

- *Labeled* F_1: Neurri honek dependentzia semantikoak etiketatzean sistemak izan duen eraginkortasuna adierazten du. Hau kalkulatu ahal izateko bi dependentzia semantiko mota hartzen dira kontutan: (1) predikatu baten eta honen argumentu edo adjuntuen artekoak, eta (2) predikatuen eta esaldiaren zuhaitz semantikoaren erro *artifiziala* den `root` tokenaren artekoak. Dependentzia sintaktikoekin egin den moduan, erlazio semantiko bat ongi egongo da burua den tokena (predikatua edo erroa) eta etiketa semantikoa ((1) motako erlazioen kasuan rol semantikoa edo

adjuntu-etiketa eta (2) motakoenean predikatuaren adiera) zuzen etiketatuta badaude. Neurri honek ahalbidetzen du predikatu-argumentu-adjuntu egitura batean predikatuari dagokion adiera gaizki esleituta egonda ere, argumentu edo adjuntuak ongi etiketatuta baldin badaude, hauengatik puntuak jasotzea. Eman dezagun s_1 eta s_2 esaldien dependentzia semantikoak etiketatu direla, eta bertan e_{11}, e_{12} eta e_{21} predikatu-argumentu-adjuntu egiturak etiketatu direla. Honela:

s_1:
e_{11} = `predikatua.01`: `arg0`, `arg1`, `argM-TMP`
e_{12} = `predikatua.04`: `arg0`, `arg1`
s_2:
e_{21} = `predikatua.01`: `arg0`, `arg1`, `arg2`, `argM-LOC`

Eman dezagun, gainera, dependentzia semantiko hauetako batzuk gaizki daudela, buruko tokena edo erlazio semantiko mota ongi eskuratu ez direlako. Adibidean ezabatuta agertzen direnak gaizki dauden dependentzia semantikoei dagozkie.

s_1:
e_{11} = ~~`predikatua.01`~~: `arg0`, `arg1`, ~~`argM-TMP`~~
e_{12} = `predikatua.04`: `arg0`, `arg1`
s_2:
e_{21} = `predikatua.01`: ~~`arg0`~~, ~~`arg1`~~, `arg2`, `argM-LOC`

e_{11} egituraren kasuan ikusten den bezala, adiera zuzena ez izatea ere dependentzia okertzat ulertzen da. Estrategia hau baliaturik, hemen ere bi aukera daude, *micro* edo *macro*:

1. *Micro*: tokenetan oinarrituta kalkulatzen da. Adibidean (2+3+3)/(4+3+5) = 0.666 egingo litzateke, alegia $100 * 0.666 = \%66.6$ izango litzateke sistemaren eraginkortasuna, dependentzia semantikoak eskuratzeko orduan.
2. *Macro*: Predikatu-argumentu-adjuntu egituretan oinarrituta kalkulatzen da. Adibidean $((2/4 + 3/3) + (3/5))/3 = 0.7$ egingo litzateke, alegia $100 * 0.7 = \%70$ izango litzateke sistemaren eraginkortasuna, dependentzia semantikoak eskuratzeko garaian.

Ondoren, *micro* eta *macro* aukerei dagozkien doitasuna, estaldura eta F_1 neurriak kalkulatuko lirateke. Hauei *micro* eta *macro Labeled doitasuna*, *Labeled estaldura* eta *Labeled* F_1 neurriak esaten zaie, hurrenez hurren.

- *Labeled Macro* F_1 *Score*: Metrika honek aurreko biak konbinatzen ditu. Dependentzia etiketatzaileen kasuan, LAS *macro* eta *macro Labeled* F_1 elkartzen dituen *Labeled Macro* F_1 *Score* neurria erabiltzen da sistemaren eraginkortasuna neurtzeko. Hala ere, aukera dago LAS *micro* eta *micro Labeled* F_1 uztartu eta *Labeled Micro* F_1 *Score* neurria kalkulatzeko. Bi kasuetan LAS eta *Labeled* F_1en konbinatzea berdin egiten da. Erabiltzen den formula, *macro* aukerara egokitua, hau da:

  $$\text{LMP} = W_{sem} * LP_{sem} + (1 - W_{sem}) * \text{LAS}$$
  $$\text{LME} = W_{sem} * LE_{sem} + (1 - W_{sem}) * \text{LAS}$$
  $$\textit{Labeled Macro } F_1 \textit{ Score} = \texttt{batezbesteko\_harmonikoa}\,(\text{LMP}, \text{LME})$$

  LMP *Labeled Macro Doitasuna* da, eta LME, berriz, *Labeled Macro Estaldura*. W_{sem} aldagaiak dependentzia semantikoen azpiatazari, eta ondorioz dependentzia sintaktikoenari ere (1 - W_{sem}), ematen zaion pisua adierazten du. Eskuarki (*CoNLL-2008, 2009*) biei garrantzia bera ematen zaie, erdia eta erdia ($W_{sem} = 0.5$). LP_{sem} eta LE_{sem} aldagaiak dependentzia semantikoen ebaluazioaren atalean kalkulatu ditugun *macro Labeled doitasuna* eta *macro Labeled estaldura* neurriak dira. Behin LMP eta LME kalkulatuz gero bien arteko batezbesteko harmonikoa kalkulatzen da, sistema osoaren eraginkortasuna adierazten duen *Labeled Macro* F_1 *Score* neurria lortzeko. (Surdeanu et al., 2008) artikuluan esaten denez, LMP eta LME lortzeko, LAS doitasuna LP doitasunarekin eta LE estaldurarekin uztar daiteke, LAS neurria doitasun eta estaldura kasu berezitzat ikus daitekeelako. Honetan sistemak etiketatutako dependentzia sintaktiko kopurua eta eskuzko etiketatze kopurua (*gold*) bera da.

2.3 Euskararako SRL prototipoa

Atal honetan tesi lanean garatu dugun euskarazko SRL sistemaren lehen eredua deskribatzen da (Salaberri et al., 2014); hemendik aurrera SRL prototipoa deituko diogu tresna honi. 2.2 atalean esan dugun bezala, dependentzietan oinarritzen den rol semantikoen

etiketatze automatikoak bost urrats ditu. Prototipo honetan, hala ere, predikatuen argumentu eta adjuntuei dagozkien rolak eta adjuntu-etiketak esleitzera mugatu gara, hau da, prozesuaren azkeneko urratsera.

Metodologiaren aldetik egokia iruditu zaigu lehenengo aldiz SRL atazari era honetara heltzea. Izan ere, tesi lanaren helburuetako bat euskararako SRL sistema guztiz automatikoa garatzea da, arkitekturako bost urratsak automatikoki egiten dituena. Horregatik, pentsatu dugu lehenik urrats bakar batean zentratuta algoritmorik egokienak eta emaitzarik onenak itzultzen dituzten ezaugarriak zein diren identifikatzea beharrezkoa dela.

Tesi lanean zehar erabili dugun corpusak ahalbidetzen duenez, prototipoan *PropBank* eta *VerbNet* ereduak jarraitzen dituen argumentuen sailkapena egin dugu. Honek bi rol multzoen arteko alderaketa egin eta azken sistemarako egokiena aukeratzea ahalbidetu digu. Prototipoaren garapenetik ateratako ondorioak kontuan izan ditugu *bRol*, SRL etiketatzaile guztiz automatikoa garatzeko orduan.

2.3.1 Informazioaren adierazpidea

EPEC-RolSem corpuseko fitxategiak *CoNLL* formatura bihurtu genituen, besteak beste irakurterrazagoak eta prozesatzeko erosoagoak izan zitezen. Modu honetara dependentzia etiketatzaileen arloan *estandarra* den formatuarekin lan egiteko aukera izan dugu. Gainera, adierazpide hau erabiltzeak ahalbidetu du SRL prototipoan oinarrituta garatu dugun *bRol* tresna *CoNLL-2008, 2009* saioetako ebaluaziorako scriptak erabilita ebaluatu ahal izatea, testuinguru estandarrean alegia. 2.7 irudian *Egiptora Axtarteren hilobia bilatzera zihoala esan zion norbaiti* esaldiaren *CoNLL* formatuan adierazitako fitxategia ikus daiteke.

Adibideak adierazten duenez *CoNLL* izeneko formatu edo adierazpidea den honetan informazioa zutabetan dago antolatuta (Z1-Z16). Azkeneko sei zutabeetan antzeman daitekeen moduan, esaldian identifikatu den predikatu bakoitzarentzat bi rol zutabe sortzen dira, bat *PropBank* ereduko rol eta adjuntuentzat (Z11, Z13, Z15) eta beste bat *VerbNet* ereduko rolentzat (Z12, Z14, Z16). `look.05` adiera daukan predikatuari Z11 eta Z12 zutabeak dagozkio, `go.01` adierakoari Z13 eta Z14 zutabeak eta, azkenik, `say.01/tell.01` adiera duen predikatuari Z15 eta Z16. Hortaz, formatu honetan zutabe kopurua predikatu kopuruaren araberakoa da. Gainerako zutabeei dagokienez (Z1-Z10), hau da gordetzen duten informazioa:

Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
1	Egiptora	Egipto	NNP	LIB	ala	Place	5
2	Axtarteren	Axtarte	NNP	PIB	gen	Person	3
3	hilobia	hilobi	NN	ARR	abs	-	4
4	bilatzera	bilatu	VBD	SIN	helb	-	5
5	zihoala	joan	VBD	SIN	konpl	-	6
6	esan	esan	VBD	SIN	-	-	0
7	zion	izan	MD	-	-	-	6
8	norbaiti	norbait	IN	ARR	dat	-	6
9	.	.	PUNC	-	-	-	0

Z9	Z10	Z11	Z12	Z13	Z14	Z15	Z16
ncmod	-	-	-	A2	Dest.	-	-
ncmod	-	-	-	-	-	-	-
ncobj	-	A1	Theme	-	-	-	-
xmod	look.05	-	-	AM-PRP	-	-	-
ccomp_obj	go.01	-	-	-	-	A1	Topic
ROOT	say.01/tell.01	-	-	-	-	-	-
auxmod	-	-	-	-	-	-	-
nczobj	-	-	-	-	-	A2	Recip.
p	-	-	-	-	-	-	-

Irudia 2.7: CoNLL formatuaren adibidea.

- Z1: Tokenen indizeak, batetik hasten da esaldi berri bakoitzarentzat.
- Z2: Tokenak berak (formak).
- Z3: Tokenen lemak.
- Z4: Tokenen *Part-of-Speech* (PoS) kategoriak.
- Z5: Tokenen *Part-of-Speech* azpikategoriak.
- Z6: Tokenen deklinabide kasuak.
- Z7: Izen entitateen etiketak, NER. Hiru etiketa mota daude, `Place`, `Person` eta `Organization` (lekua, pertsona eta erakundea).
- Z8: Tokenen buru sintaktikoei dagozkien Z1 zutabeko balioak (`HEAD`). Tokena esaldiko erroa edo puntuazio marka izatekotan zero izango da.

- Z9: Tokenek beren buru sintaktikoekin (Z8) daukaten erlazio mota (`DEPREL`). Euskararako erlazioak: `ncsubj`, `ncobj`, `nczobj`, `ncmod`, `ncpred`, `ccomp_obj`, `ccomp_subj`, `cmod`, `xcomp_obj`, `xcomp_subj`, `xcomp_zobj`, `xmod`, `xpred`.
- Z10: Esaldiko predikatuen *PropBank* adierak.

Erabiltzen dugun *CoNLL* formatuak zenbait ezberdintasun ditu ebaluazio saioetan erabili zirenekin. Formatua, euskararako egokitzeko asmoz, deklinabide kasuaren zutabea (Z6) gehitu dugu. Gainera, *Eihera*k, izen entitateen euskararako identifikatzaileak (Alegria et al., 2003), itzultzen duen informazioa Z7 zutabean gehitu dugu.

Corpusaren hasierako azterketa: rol multzoen alderaketa

Argumentuen sailkapena egiteko *EPEC-Rolsem* corpuseko eskuzko dependentzia sintaktikoak, eskuz identifikatutako predikatuak eta hauen adierak eta argumentu eta adjuntuak erabili ditugu. Argumentuen sailkapena *PropBank* eta *VerbNet* ereduak jarraituta egin dugu. Honek aukera eman digu bi rol multzoen arteko alderaketa kuantitatiboa egin eta azken SRL sistema guztiz automatikorako egokiena zein den erabakitzeko. Corpusa rol mailan aztertu eta ingelesezko *PropBank* corpusarekin alderatu dugu. 2.4 taulan ikus daiteke ingeleseko *PropBank* corpusean eta euskarako *EPEC-Rolsem* corpusean dauden argumentuen rolak zer proportziotan datozen bat *VerbNet* corpusekoekin.

	`arg0`	%	`arg1`	%	`arg2`	%	`arg3`	%	`arg4`	%
PropBank	Agent	85	Theme	47	Recipient	22	Asset	33	Location	89
	Theme	2	Topic	23	Extent	15	Theme2	14	Beneficiary	5
	Experiencer	7	Patient	11	Predicate	14	Recipient	13	-	-
EPEC	Agent	77	Theme	52	Attribute	41	Location	41	Destination	41
	Theme	18	Topic	22	Destination	14	Destination	20	Location	30
	Topic	2	Product	10	Location	13	Beneficiary	18	Attribute	16

Taula 2.4: PB-VN rolen korrespondentzia *PropBank* eta *EPEC-Rolsem* corpusetan.

(Loper et al., 2007) lanean esaten denez, *PropBank* etiketatu zenean ahalegin berezia egin zen `arg0` rolarekin *prototypical-agent* izateko eta `arg1` rolarekin aldiz *prototypical-patient* (Dowty, 1989) izateko irizpideak betetzen zituzten argumentuak etiketatzeko. Ondorioz, bi hauek `arg2`, `arg3` eta `arg4` rolek duten baino sakabanaketa txikiagoa izango dute, *VerbNet*ekoen artean: `arg0` rolen % 85 *Agent* dira, eta `arg1`en % 11 *Patient*. *EPEC-Rolsem* corpusa dela eta, *Theme* rolen % 18 `arg0`ri dagokio, eta % 52 `arg1`

rolari. Honek argi erakusten du *EPEC-Rolsem* corpusean dagoen `arg0` eta `arg1` rolen sakabanaketa *PropBank*ekoa baino handiagoa dela. Honen arrazoia da euskarazko corpusa anotatu zenean beste irizpide batzuk jarraitu zirela. Honek eragina izango du garatu dugun SRLaren emaitzak ingelesezko SRL sistemen emaitzekin alderatzeko orduan.

2.3.2 Prototipoaren garapena: *argumentuen sailkapena*

SRL prototipoa garatzeko *PropBank* eta *VerbNet*eko rolak erabiltzen dituzten argumentuen sailkatzaileak eraiki eta ebaluatu ditugu, ikasketa automatikoa erabiliz (ML). Prototipoa ebaluatzeko *10-fold cross-validation* jarraitzea erabaki dugu. Ikasketa ebaluatzean teknika hau (ez *Train-Test* teknika) erabiltzearen arrazoia corpusaren tamaina mugatua izanda. Gure kasuan 35.500 argumentu eta adjuntu instantzia izan ditugu eskura.

Sailkatzaileak eraikitzeko hartu dugun ezaugarri multzoari dagokionez, beste hizkuntzetan SRL egiterakoan erabili ohi izan diren ezaugarriak eta euskararako bereziki prestatutakoak uztartu ditugu. Beste hizkuntzetako ezaugarriei dagozkien adibide batzuk (Palmer et al., 2010) eta (Carreras eta Márquez, 2005) argitalpenetako ingeleseko SRL sistemek erabiltzen dituztenak, txinerako (Xue eta Palmer, 2005) sistemakoak eta suedierako (Johansson et al., 2012) etiketatzailearenak dira. *EPEC-Rolsem* corpusean dauden argumentu eta adjuntuentzat erauzi dira ezaugarri linguistikoak.

Hurrengo zerrendan, prototipoak erabiltzen dituen ezaugarriak aurkezten dira. Ulergarritasun arrazoiak direla eta, *Atzo, Mikel Argentinara joan zen Bilbotik* esaldian *Mikel*ek, *joan* predikatuaren argumentuak, ezaugarri bakoitzean hartzen dituen balioak ageri dira. Izartxo batekin adierazi ditugu euskararako bereziki prestatutako ezaugarriak.

- Argumentuaren lema (*Mikel*).
- Argumentuari dagokion predikatuaren lema (*joan*).
- Argumentuaren *Part-of-Speech* kategoria (`IZE`-Izena).
- Argumentuaren *Part-of-Speech* azpikategoria* (`IZE/B`-Izen berezia).
- Argumentuaren deklinabide kasua* (`ABS`-Absolutiboa).
- Argumentuak bere buru sintaktikoarekin duen funtzioa (`ncsubj`-Subjektua).
- Argumentuak esaldian predikatuarekiko daukan posizioa (aurretik).

- Argumentuaren eta predikatuaren artean dagoen token kopurua (1).
- Argumentuaren eta predikatuaren artean dagoen argumentu kopurua (1).
- *Frame* semantikoa (`arg_arg_arg_PRED_arg`).
- *Frame* sintaktikoa (Xue eta Palmer, 2004).
- Izen entitatea, argumentua halakoa bada* (`Person`).
- Zenbaki entitatea* (`Date`, `Price` e.a.), argumentua halakoa bada (–).

Argumentu eta adjuntuentzat erauzitako ezaugarrietatik sailkatzaileak sortzeko hiru algoritmo ezagunekin egin ditugu saiakerak: *Support Vector Machines-SVM* (Cortes eta Vapnik, 1995), *Random Decision Trees-RDT* (Breiman, 2001) eta *Decision Trees-DT* (Quinlan, 1996). Modu honetara emaitzarik onenak itzultzen dituen ikasketa algoritmoa aukeratu ahal izan dugu, SRL sistema osoa eraikitzerakoan kontuan izateko.

2.3.3 Emaitzak

Garatutako sailkatzaileen ebaluazioaren emaitzak erdiesteko doitasun, estaldura eta F_1 neurri estandarrak erabili ditugu. 2.5 taulan *EPEC-Rolsem* corpusean eskuz identifikatutako argumentu eta adjuntuei *PropBank* eta *VerbNet* ereduetako etiketak automatikoki esleitzerakoan lortu ditugun emaitzak ikus daitezke.

PropBank	Doitasuna	Estaldura	F_1
SVM	84.30	84.60	84.30
DT	84.00	84.20	83.90
RDT	77.40	78.30	77.70
VerbNet	Doitasuna	Estaldura	F_1
SVM	83.10	83.10	82.90
DT	81.70	81.80	81.50
RDT	72.20	72.90	72.10

Taula 2.5: Argumentuen sailkapena, *PropBank* eta *VerbNet* rol multzoekin.

Ikus daitekeenez, *PropBank* rolak esleitzen dituen sailkatzailerik eraginkorrena *SVM* algoritmoarekin lortu dugu, eta 84.3 puntuko F_1 dauka. *VerbNet* sailkatzailea dela eta,

honek ere *SVM* erabilita lortzen ditu emaitzarik onenak; zehazki esan, 82.9 puntuko F_1 neurria du. Taulako balioek argi uzten dute emaitzarik apalenak itzultzen dituena *RDT* dela. Izan ere, eredu bakoitzeko emaitzarik altuenekin alderatzean ikusten dugu *PropBank*entzat ia zazpi puntu jaisten dela F_1 neurria, eta *VerbNet*entzat hamar baino gehiago. 2.6 taulak *SVM* sailkatzaileak etiketa bakoitzerako lortutako emaitzak erakusten ditu eta ingeleserako (Zapirain eta Agirre, 2008) argitalpenean lortutakoekin alderatzen ditu.

	Ingelesa		**Euskara**	
	PropBank	***VerbNet***	***PropBank***	***VerbNet***
arg0	88.49		95.00	
arg1	79.81		93.70	
arg2	65.44		81.60	
arg3	52.63		57.90	
ACTOR		85.44		89.70
AGENT		87.31		96.20
ATTRIBUTE		71.43		92.40
CAUSE		62.20		79.20
EXPERIENCER		87.76		66.90
LOCATION		64.58		80.10
PATIENT		78.64		80.60
PREDICATE		62.88		74.60
PRODUCT		61.97		91.60
RECIPIENT		79.81		83.20
SOURCE		60.42		74.40
STIMULUS		63.93		87.30
THEME		75.46		88.00
TOPIC		85.70		87.70
ADV	53.44		50.80	
CAU	53.06		80.50	
DIS	77.78		41.60	
LOC	61.76		73.90	
MNR	58.29		67.80	
MOD	96.14		54.30	
NEG	98.41		99.20	
TMP	75.00		78.90	
Osotara	**78.93**	**76.99**	**84.30**	**82.90**

Taula 2.6: *SVM*rekin eraikitako argumentu sailkatzaileen F_1, etiketa bakoitzerako.

2.6 taulan ikus daiteke *PropBank*eko *core* rolak deriztenentzat (arg*i*) F_1ek behera egiten duela, *i* handitzearekin. arg0k 95 puntu lortzen ditu, eta arg4ek berriz 15.4. *PropBank* adjuntuei doakienez, motaren arabera emaitzak aldatzen direla antzeman daiteke. Ezeztapeneko (NEG) eta kausazko (CAU) adjuntuek, esate baterako, 99.2 eta 80.5 puntu lortzen dituzte hurrenez hurren. Aditzondo (ADV) eta dislokazio (DIS) adjuntuek, aldiz, 50.8 eta 41.6 puntuko F_1 lortzen dute. *VerbNet* roletan erdietsitako emaitzak 96.2 eta 66.9 puntuen bitartean kokatzen dira, eta lau dira 80tik beherako emaitza daukatenak, CAUSE, EXPERIENCER, PREDICATE eta SOURCEk, hain justu.

2.3.4 Analisia

SRL prototiporako lortutako emaitzak hobeki interpretatu ahal izateko (Zapirain eta Agirre, 2008) argitalpenean *CoNLL-2005* saiorako aurkeztu zirenekin alderatu ditugu, zuzenki alderagarriak ez badira ere. Erkaketa 2.6 taulan ikus daiteke. Taulan horretan antzeman daitekeen lehen gauza da gure emaitzak, oro har, ingelesekoak baino altuagoak direla. Honen arrazoia da, ingelesekoetan, SRL egiteko beharrezkoak diren predikatuen identifikazio eta desanbiguazioa batetik, eta argumentuen identifikazioa bestetik automatikoak direla, gure prototipoaren kasuan ez bezala. Gainera, kontuan eduki beharra dago *EPEC-Rolsem* corpusean sintaxiak dependentzia formalismoa jarraitzen duela, eta ingeleseko (Zapirain eta Agirre, 2008) argitalpeneko SRLk, berriz, osagaietan oinarritutakoa. Argumentuen identifikazioa egitea, esate baterako, konplexuagoa izaten da osagaietan oinarritutako sintaxia erabiltzen denean, eta honek eragin negatiboa du.

Bi hizkuntzetako *PropBank* ereduetako *core* rolen emaitzak alderatzean, ikusten da euskarazko hobekuntzaren neurria ez dela bera rol guztientzat: arg1 eta arg2 rolek 15 puntuko hobekuntza dute gutxi gorabehera, eta arg0 eta arg4 rolek, aldiz, 5 eta 6.5 puntuko hobekuntza. Emaitza hauek interpretatu ahal izateko kontuan izan behar da arg0/AGENT korrespondentzia ingelesez ematen dela euskaraz baino gehiago. Euskaraz, aditz inakusatiboen core rolak arg0 etiketatik hasten dira. Mota honetako aditzek ez dute agenterik, beraz, aditz horiek direla eta, euskaraz arg0 baten atzean ez da zertan agente prototipiko bat egongo. Ingelesez, aldiz, aditz inakusatiboak etiketatzeko garaian arg1 etiketatik hasten dira (aditz horiek ez dute arg0rik). Honek, konparazioan, euskarazko core argumentuak detektatzea zailtzen du, ez dagoelako argi zein den arg0 eta arg1en arteko diferentzia. Ingelesez, bai. Euskarazko rolak ingelesezkoak baino neutrala-

goak dira beraz. Ingelesez `arg0` *proto-agentea* da eta `arg1`, berriz, *proto-tema*. Euskaraz ez da hori bilatu, neutralagoak dira alegia (*theory neutral*).

Ingeleseko eta euskarako *VerbNet* rolen emaitzak alderatzerakoan, eskuzko informazioaren eraginez sortutako 5 eta 15 puntu bitarteko hobekuntza ikusten dugu orokorrean euskaraz. Badira, dena dela, salbuespen batzuk: `PRODUCT` rolak 30 puntuko hobekuntza du, eta `STIMULUS` rolak 24 puntukoa; `EXPERIENCER`, bestalde, hobetu beharrean 20 puntu okertzen da. 2.4 taulan ikus dezakegu ingelesez (*PropBank*) `EXPERIENCER` `arg0` rolari dagokion hirugarren *VerbNet*eko rola dela arruntena. Gainera, ingelesez `EXPERIENCER` rolak ez du sakabanaketa handiegirik *PropBank* rolen artean. Beraz, pentsa dezakegu, gainerako *PropBank* rolen korrespondentziak ere kontuan edukita, `EXPERIENCER` rolaren sakabanaketa ingelesez baino nabarmen handiagoa dela euskaraz, eta horregatik dagoela 20 puntuko okertzea rol honentzat. Kontuan izan behar da, gainera, `EXPERIENCER` aditz *sentsitiboentzako* rola dela (*ikusi* adibidez), paziente moduko bat, sentitzen duena. Hori dela eta aditz gutxi batzuek baizik ez dute izango `EXPERIENCER` rola. Euskarazko emaitza apalaren arrazoia beraz corpusean rol honen adibide gutxi edota esanguratsuak ez direnak edukitzea izan daiteke. Azkenik, adjuntuen karietara, euskaraz ez da hobekuntza orokorrik nabari.

Ezaugarrien aukeraketa

Ezaugarrien aukeraketa funtsezkoa izaten da ikasketa automatikoa erabiltzen denean. Aukeraketaren bitartez *ezaugarri-espazioaren* dimentsioa murriztea lortzen da, eragin positiborik ez duten ezaugarriak ikasketa automatikoko prozesutik aterata. Ondorioz, ikasketa aplikatu ahal izateko beharrezkoak diren baliabideen kopurua gutxitu egiten da. Hau kontuan edukita, SRL prototipoaren eraikuntzarako erabilitako ezaugarri multzoaren gaineko aukeraketa egin dugu, zehazkiago esan, ezaugarri horietako bakoitzak argumentuen sailkapenean duen eragina aztertu dugu. Horretarako *Leave-One-Out* (LOO) izendatutako teknika erabili dugu. Honetan ezaugarriak banan-banan ezaugarri multzotik ateratzen dira, eta gainerakoekin sailkatzailea entrenatu eta ebaluatzen da. 2.7 taulan ikus daitezke ezaugarrien eragina adierazten duten sailkatzaileen ebaluazioaren emaitzak. Kendutakoan emaitzak gehien apaltzen dituzten ezaugarriak azpimarrarekin daude adierazita taulan eta gehien hobetzen dituzten ezaugarriak, aldiz, azpigakoarekin daude markatuta.

	PropBank			***VerbNet***		
Ezaugarri hau gabe	Doitasuna	Estaldura	F_1	Doitasuna	Estaldura	F_1
Predikatuaren lema	78.30	77.50	77.10	67.40	68.20	66.10
Argumentuaren lema	79.90	80.40	79.90	78.70	79.00	78.50
Argumentuaren *PoS*	84.20	84.50	84.20	83.00	83.00	82.80
Argumentuaren azpi-*PoS*	84.00	84.20	83.90	82.60	82.50	82.30
Deklinabide kasua	75.20	76.10	75.30	73.60	73.90	73.40
Funtzio sintaktikoa	82.00	82.20	81.90	80.90	80.90	80.60
Argumentuaren posizioa	84.30	84.60	84.30	83.10	83.10	82.90
Distantzia hitzetan	84.30	84.60	84.30	83.10	83.10	82.90
Distantzia argumentutan	84.30	84.50	84.30	83.10	83.10	82.90
Frame semantikoa	84.40	84.70	84.40	83.30	83.30	83.10
Frame sintaktikoa	84.50	84.60	84.30	83.40	83.30	83.10
Izen entitatea	84.30	84.60	84.30	83.20	83.20	83.00
Zenbaki entitatea	84.40	84.60	84.30	83.10	83.10	82.90
Ezaugarri guztiekin	**84.30**	**84.60**	**84.30**	**83.10**	**83.10**	**82.90**

Taula 2.7: *Leave-One-Out* teknikaren bitartezko ezaugarrien aukeraketa, *PropBank* eta *VerbNet* ereduak jarraitzen dituzten argumentu sailkatzaileentzat.

2.7 taulako balioetan nabari da badirela bi ereduetan emaitzak okertzen dituzten ezaugarriak: *VerbNet* ereduko sailkatzailearentzat eragin negatiboa daukatenak *frame* semantikoa, *frame* sintaktikoa (F_1 82.9 puntutatik 83.1 puntutara) eta izen entitatea (F_1 82.9 puntutatik 83 puntutara) dira. *PropBank* sailkatzailearentzat, aldiz, emaitza beherarazten duen ezaugarri bakarra *frame* semantikoa da (F_1 84.3 puntutatik 84.4 puntutara).

Taulan ikus daiteke, orobat, ezaugarri bakoitzak argumentuen sailkapenean duen eragina eredu batetik bestera aldatzen dela; *PropBank*en eraginik positiboena duten lauak, handienetik txikienera ordenaturik, hauek dira: deklinabide kasua, predikatuaren lema, argumentuaren lema eta funtzio sintaktikoa. *VerbNet*en aldiz: predikatuaren lema, deklinabide kasua, argumentuaren lema eta funtzio sintaktikoa.

Ezaugarrien aukeraketaren emaitzak kontuan hartuta sailkatzaileen ikasketa prozesuan eragin negatiboa duten ezaugarri guztiak baztertu ditugu ezaugarri multzotik eta eragin positiboko eta neutroko ezaugarriak erabilita eraiki eta ebaluatu ditugu, berriro ere, sailkatzaileak. Hauek dira erabilitako ezaugarriak: predikatuaren lema, argumentuaren lema, *Part-of-Speech* kategoria eta azpikategoria, deklinabide kasua eta funtzio sintaktikoa. Ebaluazio berriaren emaitzak 2.8 taulan ikus daitezke.

	Doitasuna	Estaldura	F_1
PropBank	84.20	84.30	84.00
Guztiak *PropBank*	84.30	84.60	84.30
VerbNet	82.90	82.80	82.60
Guztiak *VerbNet*	83.10	83.10	82.90

Taula 2.8: Argumentuen sailkapena, ezaugarrien aukeraketarekin

Taulako balioek erakusten duten moduan, ezaugarri multzotik baztertutako ezaugarriek banaka kenduz gero sailkatzaileen emaitza hobetzen dutela ikusi badugu ere, guztiak aldi berean kentzean sailkatzaileen emaitza ez da hobetzen, okertzen baizik. Bi ereduen kasuan emaitza (F_1 neurria) 0.3 puntu jaisten da. Beraz, esan dezakegu banaka harturik eragin negatiboa duten ezaugarri baztertuen konbinazioak eragin positiboa duela, argumentuen sailkapenean.

2.3.5 Eskuzko ebaluazioa

SRL prototiporako garatu ditugun bi sailkatzaileen ebaluazioek hauen benetako eraginkortasuna zer neurritan adierazten duten egiaztatu nahi izan dugu. Horretarako, *EPEC-Rolsem* corpuseko argumentu eta adjuntu instantzia multzo berri baten eskuzko ebaluazio egin dugu.

	Doitasuna	Estaldura	F_1
`arg0`	99.20	86.60	92.50
`arg1`	93.00	85.70	89.20
`arg2`	50.90	41.80	45.90
`arg3`	100.00	100.00	100.00
`argM`	81.90	94.30	87.70
Osotara	**85.80**	**87.00**	**85.90**

Taula 2.9: *PropBank* argumentu sailkatzailearen eskuzko ebaluazioa.

Multzo berria osatzeko *EPEC-Rolsem* corpusean lan honen hasieran eskuragarri ez zeuden 2.558 instantzia hartu ditugu ausaz. Eskuzko ebaluazioaren emaitzak 2.9 eta 2.10 tauletan daude bilduta. Lehenbiziko taulan, *PropBank* sailkatzaileari dagokionean, 85.8 eta 87 puntuko doitasuna eta estaldura, eta 85.9 puntutako F_1 neurria lortu ditugu. Ebaluazio automatikoan, ordea, 84.3 eta 84.6 puntuko doitasuna eta estaldura, eta 84.3

	Doitasuna	Estaldura	F1
ACTOR	0	0	0
AGENT	94.00	92.70	93.40
ATTRIBUTE	97.30	40.8	57.50
CAUSE	98.30	98.30	98.30
EXPERIENCER	0	0	0
LOCATION	100.00	29.60	45.70
PATIENT	78.60	10.10	17.90
PREDICATE	62.50	23.80	34.50
PRODUCT	0	0	0
RECIPIENT	35.40	100.00	52.30
SOURCE	35.70	12.50	18.50
STIMULUS	0	0	0
THEME	71.50	81.10	76.00
TOPIC	85.60	85.60	85.60
Osotara	**76.20**	**83.60**	**78.20**

Taula 2.10: *VerbNet* argumentu sailkatzailearen eskuzko ebaluazioa.

puntuko F_1 neurria erdietsi ditugu. Gure irudiko, eskuzko ebaluazioak 2.558 instantzia besterik ez dituela erabili aintzat hartzen badugu (*10-fold cross-validation* ebaluazioan erabili ziren 35.500ak baino franko gutxiago, hortaz), eskuzkoan lortu emaitzak ebaluazio automatikoan erdietsitakoekin bat datoz. Kontuan izan beharra dago corpus bereko instantziak direla eskuzko ebaluaziokoak eta ez dela beraz emaitzen okertzerik izango domeinu edo corpus aldaketak eraginda. 2.9 taulan ikusten da eskuz egindako ebaluazioan ez dela *PropBank* adjuntuen arteko bereizketarik egin, adjuntuak diren predikatuen instantziak adjuntu bezala ongi etiketatu diren edo ez baizik ez dela egiaztatu, alegia.

VerbNet ereduko sailkatzailearentzat eskuz egin den ebaluazioak 76.2 puntuko doitasuna, 83.6 puntuko estaldura eta 78.2 puntuko F_1 neurria itzuli du. Ebaluazio automatikoak, aldiz, 83.1 puntuko doitasuna eta estaldura eta 82.9 puntuko F_1 neurria. Eredu honen kasuan bi ebaluazioen arteko aldea *PropBank* ereduko bi ebaluazioen artekoa baino handiagoa da: F_1 neurrian 1.6 puntuko aldea dago *PropBank*en, eta 4.7 puntukoa *VerbNet*en. Uste dugu aurretik aipatu ditugun faktoreak kontuan edukita (ebaluazio multzoaren tamaina aise txikiagoa dela eta domeinu bereko instantziak erabili direla), *VerbNet*en espero bezalako emaitzak lortu ditugula, eskuzko ebaluazioan ere. Ondorioz, egindako *10-fold cross-validation* ebaluazioa esanguratsua eta baliagarria dela uste dugu.

2.10 taulan ACTOR, EXPERIENCER eta STIMULUS rolei dagozkien balioak zero dira. Honen arrazoia da ausaz hartutako instantzietan rol hauei dagozkienak oso balio txikiak edo zero direla. *EPEC-Rolsem* corpuseko rol arruntenei dagozkien instantziak arruntenak eta gehiengoa direlako.

2.4 *bRol*: euskararako SRL etiketatzaile automatikoa

Atal honetan *bRol* izendatu dugun euskararako rol semantikoen etiketatzaile guztiz automatikoa aurkezten dugu. *bRol* euskarazko lehen SRL sistema da, eta honen berri (Salaberri et al., 2015a) argitalpenean eman genuen. Etiketatzailearen diseinuan zehar hartutako erabakietako asko (jarraitu beharreko eredua, sailkatzaileak sortzeko ikasketa algoritmoa, ezaugarri linguistikoak e.a.) SRL prototipoaren garapenean lortutako emaitzak ikusirik hartu ditugu. *bRol* sistemaren deskribapena egiten hasi baino lehen 2.11 taulan biltzen ditugu SRL prototipoarekin dauden ezberdintasun eta antzekotasun esanguratsuenak.

	Prototipoa	***bRol***
Dependentzia sintaktiko automatikoak	✗	✓
Predikatuen identifikazio automatikoa	✗	✓
Predikatuen sailkapen automatikoa	✗	✓
Argumentuen identifikazio automatikoa	✗	✓
Argumentuen sailkapen automatikoa	✓	✓
Dependentzietan oinarritutako formalismoa	✓	✓
CoNLL formatuko *dataset*a	✓	✓
CoNLL ebaluazioa (testuinguru estandarra)	✗	✓
Doitasuna, estaldura eta F_1 neurria ebaluatzeko	✓	✗
Train-Test ebaluazioa	✗	✓
k-fold cross-validation ebaluazioa	✓	✗
PropBank eredua	✓	✓
VerbNet eredua	✓	✗

Taula 2.11: *bRol* eta SRL prototipoaren arteko alderaketa.

Taula honetan argi ikusten da zein den SRL prototipoaren eta *bRol* etiketatzailearen arteko alde nagusia: prototipoak ez bezala, *bRol* sistemak testu soila hartzen du sarreratako, eta dependentzia sintaktiko-semantikoak ezartzen ditu modu guztiz automatikoan. Ebaluazioa testuinguru estandarrean egin da *CoNLL-2009* ebaluazio saioko *konfigurazioa* jarraitu dugulako: *dataset*aren formatua, ebaluaziorako erabilitako metrikak eta teknika,

rol multzoaren eredua, *scorer* ofiziala, e.a. Gainera, *bRol* sistemarentzat lortu ditugun emaitzak ebaluazio saioan erabili ziren beste zazpi hizkuntzarentzat argitaratutako emaitzekin alderatu ditugu.

2.4.1 Corpusen alderaketa kuantitatiboa

Azpiatal honetan *bRol* etiketatzailearen garapenean erabili dugun *EPEC-Rolsem* corpusaren bertsioa *CoNLL-2009* ebaluazio saioko zazpi hizkuntzen corpusekin alderatzen dugu. Ebaluazio saioko hizkuntzak hauek dira: katalana, gaztelera, ingelesa, alemana, txekiera, txinera eta japoniera. Corpusen alderaketa kuantitatiboa beharrezkoa da, emaitzen analisi eta interpretazio egokia egin ahal izateko. Jarraian, ebaluazio saioko hizkuntzetan erabili ziren domeinu-barneko corpusak laburki deskribatzen ditugu.

- **Katalana eta Gaztelera**: Bi hizkuntza hauen *dataset*ak *AnCora* corpusean (Taulé et al., 2008) oinarrituta egin ziren. Corpuseko sintaxiak osagaietan oinarritutako formalismoa erabiltzen du; *CoNLL-2009* saioan, berriz, dependentzietan oinarritutakoa jarraitu zen, horregatik corpuseko informazioari bihurketa prozesu automatikoa ezarri behar izan zitzaion, *dataset*ak osatu ahal izateko.

- **Ingelesa**: Ingeleserako erabili zen *dataset*a *CoNLL-2008* ebaluazio saioan erabilitako corpusa izan zen (Surdeanu et al., 2008). Corpusa beste lau corpusen elkarketatik sortu zen, hain zuzen ere *Penn Treebank 3* (Marcus et al., 1993), *BBN Pronoun Coreference and Entity Type Corpus* (Weischedel eta Brunstein, 2005), *PropBank* (Palmer et al., 2005) eta *NomBank* (Meyers et al., 2004) corpusen elkarketatik.

- **Alemana**: Hizkuntza honetarako *dataset*a SALSA corpuseko (Burchardt et al., 2006) esaldietan anotaturik dauden aditzen instantzietatik abiatuta sortu zen. Osotara 40.000 esaldi inguru daude SALSA corpusean, sintaktikoki etiketatutako egunkari berriez osatuta dagoen TIGER corpusetik (Brants et al., 2002) datozenak.

- **Txekiera**: Txekierako *dataset*a dependentzien formalismoan oinarrituta sintaktikoki anotatuak zeuden *Prague Dependency TreeBank 2.0.* corpuseko (Hajic et al., 2006) esaldiez dago osatuta. *Dataset*aren geruza semantikoa osatzeko esaldi hauen adierazpide tektogramatikoa (Hajičová, 1998) hartu zen oinarritako.

- **Txinera**: *CoNLL-2009* ebaluazio saioan txinerarako erabili zen *dataset*a osatzeko *Treebank* (Xue et al., 2005) eta *PropBank* (Xue eta Palmer, 2009) txinatar baliabideak erabili ziren. Lehenengotik esaldien informazio sintaktikoa hartu zen, osagaietan oinarritutako formalismoa dependentzietara bihurtuta, eta bigarrenetik esaldien dependentzia semantikoak.

- **Japoniera**: Hizkuntza honetarako *Kyoto University Text Corpus* delakoa (Kawahara et al., 2002) erabili zen. Bertan *Mainichi* egunkariko[11] 40.000 esaldi biltzen dira. Hauetatik 5.000k bakarrik dauzkate dependentzia sintaktiko eta semantikoak etiketatuta; gainerakoek dependentzia sintaktikoak besterik ez dituzte etiketatuak.

Deskribatu ditugun saioko zazpi *dataset*ak domeinu-barneko *modalitateari* dagozkio; egunkari-berriak dira hain zuzen. SRL sistemaren ebaluazioan gainera, sistemaren garapenerako eta ebaluaziorako corpus bera erabili dugu, *EPEC-RolSem* (bi zati ezberdin, *Train* eta *Test*). Gure corpusaren domeinua ere egunkari-berriak dira. Domeinu-kanpoko ebaluazioa egitea gure interesekoa baldin bada ere, ezin izan dugu horrelakorik aurrera eraman, baliabide faltarengatik. Ondoren *EPEC-RolSem* eta deskribatutako zazpi corpusak kuantitatiboki alderatzen dituzten 2.12, 2.13 eta 2.14 taulak ikus daitezke.

Ezaugarriak	**Euskara**	Katalana	Txinera	Txekiera	Ingelesa	Alemana	Japoniera	Gaztelera
Train esaldiak	**6941**	13200	22277	38727	39279	36020	4393	14329
Train tokenak	**108003**	390302	609060	652544	958167	648677	112555	427442
Batazbesteko esaldi luzera	**15.56**	29.6	27.3	16.8	24.4	18.0	25.6	29.8
Tokenak argumentuekin (%)	**10.75**	9.6	16.9	63.5	18.7	2.7	22.8	10.3
DEPREL motak	**30**	50	41	49	69	46	5	49
PoS motak	**26**	12	41	12	48	56	40	12
FEAT motak	**298**	237	1	1811	1	267	302	264
FORM hiztegiaren tamaina	**20051**	33890	40878	86332	39782	72084	36043	40964
LEMMA hiztegiaren tamaina	**9042**	24143	40878	37580	28376	51993	30402	26926
Test esaldiak	**3438**	1862	2556	4213	2399	2000	500	1725
Test tokenak	**53809**	53355	73153	70348	57676	31622	13615	50630
Test FORM OOV	**12.41**	5.40	3.92	7.98	1.58	7.93	6.07	5.63
Test LEMMA OOV	**6.38**	4.14	3.92	3.03	1.08	5.83	5.21	3.69

Taula 2.12: *EPEC-RolSem* eta *CoNLL-2009* saioko *in-domain* corpusen alderaketa.

2.12 taulan *CoNLL-2009* saioko SRL etiketatzaileen garapenerako eta ebaluaziorako erabilitako *train* eta *test* corpus zatien artean dauden ezberdintasunak ikus daitezke, hizkuntza bakoitzerako. Esaldi eta token kopuruari erreparatuta euskararako erabili dugun corpusa beste hizkuntzetakoa baino murritzagoa dela antzeman daiteke, izan ere

[11] http://mainichi.jp

gure corpusak 108.003 token baititu *train* zatian, eta ebaluazio saioko corpusik txikienak aldiz, japonierakoak, 112.555 token. Handiena ingelesekoa da, *train* zatian 958.167 token baititu. 2.12 taulak, gainera, token kopuruan esaldien batezbesteko luzera zein den erakusten du, datu garrantzitsua, bi arrazoirengatik: (1) orokorrean esaldi luzeen dependentzia sintaktikoak etiketatzea zailagoa izaten da esaldi motzenak etiketatzea baino; honek eragina dauka dependentzia semantikoak ezartzean, semantikak sintaxiarekin daukan erlazio estuarengatik, eta (2) esaldi luzeak aberatsagoak izaten dira semantikoki, predikatu gehiago eta argumentu gehiago izaten dituzte etiketatuta (Peterson et al., 2014) eta, ondorioz, ikasketa algoritmoak eraginkorragoak izaten dira. *EPEC-RolSem* corpusean esaldien 15.56 tokeneko batezbesteko luzerarik apalena da, beste zazpi hizkuntzetako esaldien luzeraren aldean.

	Euskara	%	Katalana	%	Txinera	%	Txekiera	%	Ingelesa	%	Alemana	%	Japoniera	%	Gaztelera	%
	ncmod	**26**	sn	16	COMP	21	Atr	26	NMOD	27	NK	31	D	93	sn	16
	PUNC	**15**	spec	15	NMOD	14	Aux	10	P	11	PUNC	14	ROOT	4	spec	15
DEPREL	**lot**	**9**	f	11	ADV	10	Adv	10	PMOD	10	MO	12	P	3	f	12
	auxmod	**8**	sp	9	UNK	9	Obj	7	SBJ	7	SB	7	A	0	sp	8
	ncsubj	**7**	suj	7	SBJ	8	Sb	6	OBJ	6	ROOT	6	I	0	suj	8
Osotara	% 65		% 58		% 62		% 59		% 61		% 70		% 100		% 59	

Taula 2.13: *Train* zatietako bost DEPREL etiketarik ohikoenak (funtzio sintaktikoak).

2.13 taulak corpusetako tokenek beren buru sintaktikoekin dauzkaten bost erlazio sintaktiko arruntenak zerrendatzen ditu. Taularen azkeneko lerroak hizkuntza bakoitzeko bost etiketa hauek hizkuntza horretako corpusetako tokenen zer ehunekori dagokion adierazten du. Portzentajea zenbat eta altuagoa, orduan eta errazagoa izango da dependentzia sintaktikoak ezartzea. Japonieraren kasuan, esate baterako, taulan corpuseko tokenek hiru etiketa sintaktiko baizik ez dituztela jasotzen ikus daiteke. Euskararako dagoen etiketa sintaktikoen banaketa (% 65) beste hizkuntzenaren antzekoa da.

	Euskara	%	Katalana	%	Txinera	%	Txekiera	%	Ingelesa	%	Alemana	%	Japoniera	%	Gaztelera	%
	A1	**21**	arg1-pat	22	A1	30	RSTR	30	A1	37	A0	40	GA	33	arg1-pat	20
	A2	**15**	arg0-agt	18	A0	27	PAT	18	A0	25	A1	39	WO	15	arg0-agt	19
ROLA	**A0**	**14**	arg1-tem	15	ADV	20	ACT	17	A2	12	A2	12	NO	15	arg1-tem	15
	AM-TMP	**8**	argM-tmp	8	TMP	7	APP	6	AM-TMP	6	A3	6	NI	9	arg2-atr	8
	AM-MNR	**7**	arg2-atr	8	DIS	4	LOC	4	AM-MNR	3	A4	1	DE	6	argM-tmp	8
Osotara	% 65		% 71		% 91		% 75		% 83		% 97		% 78		% 70	

Taula 2.14: *Train* zatietako bost rol semantikoen etiketarik ohikoenak.

Azkeneko taulak, 2.14k, corpusetako predikatuen argumentuek jasotzen dituzten rol semantikoen bost etiketa arruntenak biltzen ditu. Kasu honetan, sintaxiko etiketekin ez

bezala, euskararako dagoen rol etiketen banaketa beste corpusetakoa baino handiagoa da. Izan ere, euskarazko rol semantikoen bost etiketarik arruntenak instantzien % 65i dagozkie eta horrek esan nahi du instantzien % 35 beste rol etiketei dagozkiela. Honek euskararako SRL ataza zailtzen du. Alemanean, esate baterako, kontrakoa da egoera, bost etiketarik arruntenak instantzien % 97ri dagozkie.

2.4.2 *bRol* etiketatzailearen garapena

Azpiatal honetan *bRol* sistemaren diseinua eta garapen prozesua pausoz pauso azalduko dugu. Horretarako, 2.2 atalean aurkeztu ditugun sei urratsak banan-banan aztertuko ditugu: dependentzia sintaktikoen etiketatzea, predikatuen identifikazioa eta desanbiguazioa, argumentuen identifikazioa eta sailkapena eta post-prozesua.

Dependentzia sintaktikoen etiketatzea

Grafoetan oinarritutako dependentzien sintaxiko hedadura maximoko zuhaitzak erabiltzen dituen metodoa aplikatzea erabaki dugu. Hau (Atutxa et al., 2015) lanean aurkezten diren esperimentuetan euskararako emaitzarik onenak eman zituena izan zen eta horregatik erabili dugu *bRol* etiketatzailearen garapenean. Hain zuzen ere, metodo honetan oinarritutako MATE (Bohnet, 2010) tresnaren bitartez etiketatzen ditugu dependentzia sintaktikoak.

Gure SRL etiketatzailean *MATE* erabili ahal izateko lehenik ezaugarri linguistikoak aukeratu eta hauekin analizatzailea entrenatu behar izan dugu. Ezaugarririk egokienak hautatu ahal izateko, kontuan eduki dugu ingelesa, txinera, gaztelera eta katalana ez bezala, euskara morfologia aberatseko hizkuntza dela (*Morphologically Rich Language-MRL*). Gainera, euskarak inflexio eta eratorpen-morfologia maila gora erakusten duela ere aintzat hartu dugu. Nilssonen (2007) arabera, ikerketaren egungo egoerako analizatzaile sintaktikoek, morfologia aberatseko hizkuntzak prozesatzean, ez dituzte hizkuntza ez-eransleak prozesatzean lortzen dituzten bezalako emaitzak eskuratzen. *MRL* hizkuntzen emaitzak hobetu eta morfologikoki aberatsak ez diren hizkuntzen eraginkortasunera hurbiltzeko, informazio morfologikoa biltzen duten ezaugarri linguistikoak erabili ohi dira. (Goenaga et al., 2013) lanean argitaratutako emaitzetan oinarrituta ezaugarri hauek aukeratu ditugu, besteak beste, *bRol* etiketatzailean txertatu dugun *MATE* analizatzailea entrenatzeko: deklinabide kasua, numeroa eta menpeko esaldi mota.

Predikatuen identifikazioa

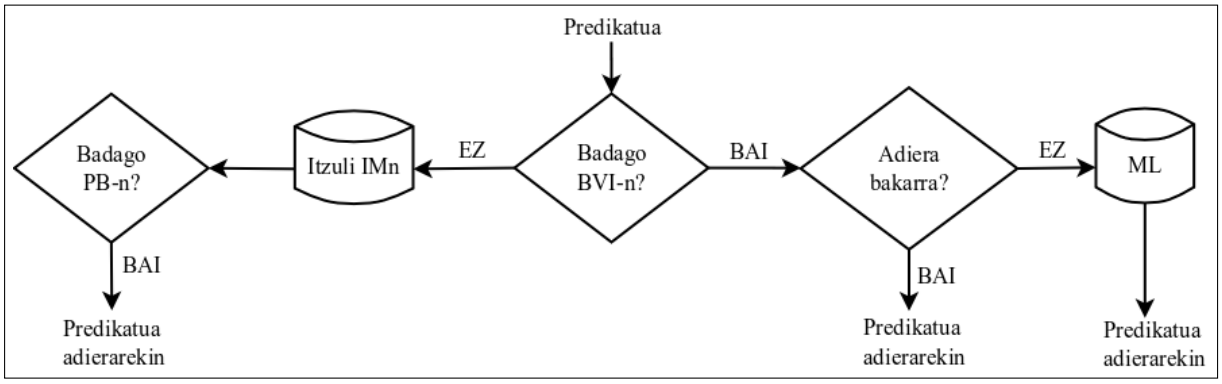
Predikatuen identifikazioa dependentzia semantikoak etiketatzeko lehenengo urratsa da. Honetan, esaldietako token bakoitza predikatua izateko hautagai bihurtzen da, eta sailkatzaile bitar baten bitartez hautagai hauek predikatuak diren edo ez erabakitzen da. *bRol* etiketatzailerako aditz predikatuak baizik ez ditugu izan kontuan garapenerako erabili ditugun *BVI* predikatu-lexikoian eta *EPEC-RolSem* corpusean daudenak hauek direlako. Hortaz, ez ditugu izen, adjektibo eta adberbio predikatuak identifikatu. Erabilitako sailkatzaile bitarra sortzeko *SVM* algoritmoa aplikatu dugu, SRL prototipoan emaitzarik onenak eskuratu dituena izan delako. Gainera, sailkatzaileari predikatu izateko hautagai zerrenda eman baino lehen garbiketa prozesua jarraitu dugu, puntuazio markak diren tokenak hautagai zerrendatik ateratzeko. Sailkatzailea sortzeko ezaugarriak token bakoitzarentzat erauzi dira; ondorengoak dira ezaugarri hauek:

- Forma
- Lema
- *Part-of-Speech* kategoria
- *Part-of-Speech* azpikategoria
- Dependentzia sintaktikoa (`DEPREL`)
- Buru sintaktikoaren lema
- Buru sintaktikoaren *Part-of-Speech* kategoria
- Buru sintaktikoaren *Part-of-Speech* azpikategoria
- Zuhaitz sintaktikoan tokenaren umeak diren tokenen formak
- Zuhaitz sintaktikoan tokenaren umeak diren tokenen lemak
- Zuhaitz sintaktikoan tokenaren umeekin tokenak berak dituen dependentziak

Predikatuen desanbiguazioa

Urrats honetan aurretik identifikatu ditugun aditz predikatuen adierak zehazten ditugu. Hau lortzeko *multiclass* sailkatzaile bat sortu dugu *SVM* algoritmoan oinarrituta, adiera bat baino gehiago daukaten predikatuentzat. Adiera bakarrekoentzat prozesua zuzena da eta ez da beharrezkoa ikasketa automatikoa erabiltzea. Adiera anitz dituzten predikatu guztientzat sailkatzaile bakarra sortzen da *bRol* etiketatzailean. *BVI*n dauden 244 aditzetatik 80 aditzek adiera bat baino gehiago dute, eta gainerako 164 aditzek bat bakarra. *EPEC-RolSem* corpusean 11.740 predikatu instantzia daude, eta horietatik 8.166 adiera anitzeko 80 predikatuei dagozkie (% 69.55). Predikatuen adieren desanbiguazioan erabili ditugun ezaugarriak predikatuen identifikazioko berak eta deklinabide kasua dira.

2.1.2 atalean aipatu dugu *EPEC-RolSem* corpusean 1.211 aditz ezberdin daudela, eta *BVI* lexikoian, berriz, 30 agerraldi edo gehiago dituzten 244 aditz baizik ez daudela. Honek esan nahi du 967 aditzen agerraldiak ez daudela corpusean etiketatuta eta, beraz, *bRol* tresnak ez duela hauen dependentzia semantikoak corpusetik etiketatzen ikasten. Sistemaren estalduran dagoen galera hau gutxitzen saiatzeko *bRol* etiketatzaileari, predikatuen desanbiguazioaren urratsean, *Itzulpen Modulua-IM* deitu dugun osagaia gehitu diogu. Izan ere, aurreko urratsean, predikatuen identifikazioan, *BVI* lexikoian ez dauden aditzak ere identifikatzen dira, baina desanbiguatzean adiera egokia ezartzea lortzen ez bada, haien argumentuak ez dira identifikatuko eta, beraz, rolak ere ez zaizkie esleituko.



Irudia 2.8: Predikatuen desanbiguazio prozesua *bRol* etiketatzailean.

2.8 irudian predikatuen desanbiguazio prozesu osoa ikus daiteke. Desanbiguatzaileak identifikatutako predikatuak jasotzen ditu sarreratzat, *BVI* lexikoian predikatua aurki badaiteke eta adiera bakarra badu, hau zuzenean esleitzen zaio. Adiera bat baino gehiago

izatekotan sailkatzaileak erabakiko du *BVI*n dauden adieretako zein dagokion predikatuari. Aditza lexikoian aurkitzen ez bada, ordea, (IM) itzulpen moduluan euskaratik ingelesera itzultzen da, eta itzulitako predikatua ingeleseko *PropBank* predikatu-lexikoian bilatzen da. Bertan itzulpena aurkituz gero, honi lexikoian dagokion lehenengo adiera ezartzen zaio predikatuari.

Pausu honen zuzentasuna ulertu ahal izateko bi gauza izan behar dira kontutan: (1) 2.1.2 atalean azaldu dugun moduan *EPEC-RolSem* corpusean predikatuen adierak ingeleseko *PropBank* adierekin etiketatuta daudela, (2) *BVI*n ez egoteak aditzak *arruntak* ez direla esan nahi duela (`agerraldi_kopurua` < 30), eta arruntak ez diren aditzek normalean adiera bat bakarra izaten dutela (*usteldu* adibidez). Predikatuak itzultzeko IMk euskara-ingelesa *Elhuyar hiztegia*[12] erabiltzen du.

bRol ebaluatzeko orduan kontuan izan dugu itzulpen moduluaren ondorioz etiketatutako dependentzia semantikoak ezin ebalua daitezkeela, hauek corpusean eskuz anotatuta ez daudelako. Beraz, ezin jakin dezakegu, IM aktibatuta dagoenean, zein den *BRol* sistemaren benetako eraginkortasuna. Aurrerago aurkezten dugun *bRol* etiketatzailearen ebaluazioan itzulpen modulua desaktibatu egin dugu. Predikatuen desanbiguazioa egiten duen sailkatzailea *SVM-multiclass* tresnaren (Joachims, 1999) bitartez sortu dugu. Sailkatzailearen parametroei dagokienez, erabilitako *kernel funtzioa* lineala da, eta entrenamendu errorearen eta marjinaren arteko hartu-emana *batezbetekoa*$(x*x)^{-1}$ formularen bitartez kalkulatzen da.

Argumentuen identifikazioa

Pauso honetan aurreko bi urrats semantikoen ondorioz identifikatu eta desanbiguatu diren aditz predikatuen argumentuak eta adjuntuak identifikatzen dira. Bi eratako hurbilpena egin dugu: predikatuekin egin dugun bezala ikasketa automatikoan oinarritzen den sailkatzaile bitar bat erabilita, eta heuristikoen bitartez. Gure sistemarako bi hurbilpenetatik egokiena aukeratzeko, zenbait esperimentu egin eta gero, heuristikoak erabiltzea erabaki dugu. *bRol* etiketatzaileak erabiltzen duen heuristikoak *Part-of-Speech* kategoria, buru sintaktikoaren inguruko informazioa (`HEAD`) eta dependentzia sintaktiko mota (`DEPREL`) erabiltzen ditu, t_i tokena P_j predikatuaren argumentua edo adjuntua den edo ez erabakitzeko. Heuristikoa ondorengoa da:

[12] `http://hiztegiak.elhuyar.eus/eu_en`

```
//t_i:tokena
//P_j:predikatua
IF (HEAD(t_i) == P_j){
    //auxmod:laguntzailea
    //haos:hitz anitzeko unitate lexikal baten zatia
    //postos:hitz anitzeko postposizio baten zatia
    //entios:hitz anitzeko entitate baten zatia
    //PUNC:puntuazioa
    IF (DEPREL(t_i,P_j)!=[auxmod,haos,postos,entios,PUNC]){
        //ADK:aditz konposatua
        IF (PoS(P_j) != ADK){
            Argumentua_edo_Adjuntua_da(t_i,P_j);
        }
    }
}
```

Heuristikoaren hainbat bertsiorekin saiakerak egin eta gero hau da emaitzarik onenak itzultzen dituena. Honek gainera ikasketa automatikoko hurbilpenak baino doitasun altuagoa lortzen du. Argumentuen identifikazioan sailkatzailerik behar ez izateak *bRol* sistemaren exekuzio denbora gutxitzea dakar berekin, honetan ezaugarriak erauzten dituen moduluaren beharrik ez dagoelako. Eskuarki hauek izaten dira sistema hauetako osagairik mantsoenak.

Argumentuen sailkapena

Identifikatu ditugun predikatuen argumentuei (eta adjuntuei) rol semantikoak esleitzeko, entropia maximoan oinarritutako *multiclass* sailkatzailea inplementatu dugu, *MEGA* tresnaren (Daumé III, 2004) bitartez. Pauso honetan entropia maximoko sailkatzailea (eta ez *SVM* sailkatzailea) erabiltzearen arrazoia jarraian azaltzen dugu.

2.2 atalean aipatu dugu hizkuntzaren eta erabiltzen den corpusaren arabera, gerta

daitekeela predikatu baten adiera batek rol semantiko bera jokatzen duten bi argumentu ezberdin jasotzea, ingelesez *PropBank* corpusean gertatzen den bezala. Aipatu dugu, gainera, hau euskaraz ezinezkoa dela eta horregatik mota honetako inkoherentziak, argumentuen sailkatzailearen emaitzak jaso ondoren, zuzendu egin behar izaten direla.

Zuzenketarako *Integer Linear Programming-ILP* optimizaziorako teknikak erabili ohi dira. Teknika hauek predikatuen argumentuek rol semantiko (edo adjuntu etiketa) bakoitza jokatzeko daukaten probabilitatea ($P(klasea|argumentua)$) itzultzen duen sailkatzailea erabiltzen dute predikatu-argumentu-adjuntu egituren probabilitate maximoko rol semantikoen konbinazioa bilatzeko. Arazoa da *SVM* algoritmoarekin entrenatutako sailkatzaileek ez dituztela probabilitate hauek itzultzen; entropia maximokoek, ordea, bai. Berez *SVM* sailkatzaileek itzultzen dituzten balioetatik behar ditugun probabilitateak lortzea posible bada ere, sailkatzaile bitarretan esate baterako erregresio logistikoa erabilita (Platt et al., 1999), eta *multiclass* sailkatzaileetan (Wu et al., 2004) argitalpenean proposatzen den metodoa erabilita, haiek lortzeko prozesua nahiko konplexua eta konputazionalki garestia da.

Ondorioz, eta *bRol* etiketatzailearen behar konputazionalak ez handitzeko asmoz, argumentuen eta adjuntuen sailkatzailea sortzeko entropia maximoko algoritmoa erabiltzea erabaki dugu. Jarraian argumentu edo adjuntu bakoitzarentzat erauzi eta urrats honetan erabili ditugun hizkuntza ezaugarriak zerrendatzen ditugu:

- Lema
- *Part-of-Speech* kategoria
- *Part-of-Speech* azpikategoria
- Dependentzia sintaktikoa (`DEPREL`)
- Deklinabide kasua
- Argumentuari dagokion predikatuaren adiera
- Argumentuari dagokion predikatuaren lema

Argumentuen sailkapena egiten duen sailkatzailearen parametroei dagokienez, zehaztutako zalantza koefizientearen (*perplexity*) aldaketa minimoa -999999 da, eta priore Gaussiarraren (*Gaussian prior*) doitasuna 1.

Post-prozesua

Hau da *bRol* etiketatzaileak sarreratako jaso dituen esaldien dependentzia sintaktiko eta semantikoak itzuli aurretik burutzen duen azkeneko urratsa. Esan dugu *ILP* optimizazio metodoaren helburua, gure kasuan, rol semantiko errepikaturik ez daukaten predikatu-argumentu-adjuntu egituren probabilitate maximoko rol semantikoen konbinazioak bilatzea dela. Rol semantikoen etiketatze automatikoan teknika hau erabiltzeko ideia (Che et al., 2008) argitalpenetik hartu dugu. Teknikaren inplementazioari *rol-semantiko-errepikaturik-ez* izendatu dugun murriztapena gehitu diogu. Maximizatzen den helburuko funtzioa (*objective function*) hau da:

$$f = \sum log(p_{ir}.v_{ir})$$

Funtzio hau errepikatutako rol semantikoen etiketak dituen predikatu-argumentu-adjuntu egitura bakoitzarentzat maximizatzen da. v_{ir} aldagaia bitarra da, eta i indizeko tokenak (`Token ID`) $r \in R$ rola jasotzen duen edo ez adierazten du. R rol semantikoen etiketen multzoa da $R = \{\texttt{arg0}, \texttt{arg1}, \texttt{arg2}...\}$. p_{ir} aldagaiari dagokionez, honek i tokenak r rola jokatzeko daukan probabilitatea adierazten du, aldagai erreala da ($p_{ir} \in \mathbb{R}$). Optimizazio prozesua bukatzean, v_{ir} aldagaiari esleitu zaizkion balioetatik rol errepikaturik ez daukan konbinaziorik probableena eskuratzen dugu.

2.4.3 Emaitzak

bRol sistemaren eraginkortasuna neurtzeko *EPEC-RolSem* corpusaren *test* zatian sistemak etiketatu dituen dependentziak ebaluatu ditugu, *CoNLL-2009* saioko *scorer* programa erabilita (`eval09.pl`); 2.15 taulan ikus daitezke programa honek itzulitako balioak. Gainera, ebaluazio saioko *closed challenge* modalitatean beste hizkuntzetarako bertan parte hartu zuten sistemek lortu zituzten emaitzarik onenak ikus daitezke. Emaitza hauek dira gure sistemarenekin alderagarriak direnak, jarraitu dugun garapena entrenamendu corpusean oinarrituta bakarrik egin dugulako, eta ez beste baliabide batzuk ere erabilita (*open challenge*).

2.16 taulan *bRol* etiketatzaileak dependentzia semantikoak lortzeko, hau da, rol semantikoak etiketatzeko, egin dituen urratsen banakako emaitzak bildu ditugu. Hauek doitasun, estaldura eta F_1 neurri estandarrak dira.

	LAS	Labeled F1	Labeled Macro F1
Euskara	**80.51**	**75.10**	**77.80**
Katalana	87.86 (2)	80.10 (4)	83.01 (4)
Txinera	79.17 (5)	77.15 (3)	76.38 (3)
Txekiera	80.38 (2)	86.51 (3)	83.27 (3)
Ingelesa	89.88 (1)	86.15 (4)	87.69 (4)
Alemana	87.48 (1)	78.61 (3)	82.44 (3)
Japoniera	92.57 (3)	78.26 (3)	85.65 (3)
Gaztelera	87.64 (2)	80.29 (4)	83.31 (4)

Taula 2.15: *bRol* sistemaren emaitzak CoNLL-2009ko *closed challenge*, *in-domain*, emaitzarik onenekin alderatuta [(1):`Bohnet`, (2): `Merlo`, (3):`Che`, (4):`Chen`, (5):`Ren`].

Urratsa	Doitasuna	Estaldura	F_1
Predikatuen identifikazioa	87.00	88.00	87.50
Predikatuen desanbiguazioa	79.41	81.29	79.82
Argumentuen identifikazioa	72.70	86.10	78.80
Argumentuen sailkapena	77.60	77.80	77.50

Taula 2.16: *bRol* sistemaren urrats semantiko bakoitzeko emaitzak.

2.4.4 Analisia

Azpiatal honetan 2.15 eta 2.16 tauletan aurkeztu ditugun *bRol* sistemaren emaitzak analizatzen ditugu.

Dependentzia sintaktikoak

bRol sistemak 80.51 puntuko LAS neurria dauka. Balio hau dela eta, 2.15 taulan gainerako zazpi hizkuntzei dagozkien LAS neurriekin alderatzean, gurea txinerarako lortutakoa baino 1.34 puntu eta txekierakoa baino 0.13 puntu hobea dela ikus daiteke. Kontrako aldean, gure emaitza japonierarako LAS neurria (92.57) baino ia hamabi puntu, ingelesekoa (89.88) baino bederatzi puntu eta katalana (87.86), gaztelania (87.64) eta alemana (87.48) baino zazpi puntu apalagoa dela ikusten da.

Gure iritziz, *bRol* sistemaren LAS neurriaren eta gainerako hizkuntzen LAS neurrien artean dagoen aldea corpusen arteko ezberdintasun nabarmenen ondorioa da, baita hizkuntza bakoitzaren izaerari lotutako faktoreen ondorioa ere. Lehenik eta behin, kontuan izan beharra dago analisi sintaktikoak, oro har, emaitza okerragoak izaten dituela euskara

bezalako *MRL* diren hizkuntzak prozesatzean, haietan ezaugarri morfologikoak erabiltzen badira ere. *CoNLL-2009* ebaluazio saioan morfologikoki aberatsak diren txekieraren eta alemanaren analisi sintaktikoa egiteko orduan, adibidez, *MRL* ez diren ingelesa, gaztelera eta katalana analizatzean baino emaitza txarragoak lortzen dira, 2.15 taulan ikus daitekeen bezala. Aipagarria da, baita ere, eranslea den japonieraren 92.57 puntuko LAS neurri altua. Honen arrazoia da japonierako corpusean, funtzio sintaktiko guztiak etiketatzeko, bost etiketa sintaktiko besterik ez direla erabili (ikus 2.12 eta 2.13 taulak).

Euskara eta ebaluazio saioko zazpi hizkuntzak aintzat hartzen baditugu, txinera da denetan LAS baxuena daukana, 79.17 puntu. Txinera ez da *MRL* hizkuntza, txinera morfologia pobreko hizkuntzatzat (*Morphologically Poor Language-MPR*) aitortzen baita. Honen arrazoia txineraren izaera tipologikoa da: morfologia isolatzailea du, hots, morfema bakoitza hitz edo unitate lexikal bati dagokio eta, horregatik, ez dago ia ageriko morfologiarik. Seddah eta besteren (2013) arabera, ingelesetik tipologikoki urrunen dauden hizkuntzak (asiarrak eta semitikoak esaterako) dira gaur egun ere sintaktikoki (eta semantikoki) analizatzeko zailenak direnak. *bRol*en LAS neurria interpretatzeko garaian kontuan eduki beharreko faktore bat euskararen hitz ordena librearen eta morfologia aberatsaren arteko konbinazioa da. Donelaicio eta bestek (2013) diotenez, inflexio maila gora duten hizkuntzek askotan hitz ordena erlatiboki librea ere izaten dute, eta hauen analisi sintaktikoaren doitasuna ingelesekoarena baino apalagoa izaten da.

Dependentzia semantikoak

Labeled F_1 neurriak dependentzia semantikoak etiketatzerakoan *parser*ak daukan eraginkortasuna adierazten du. *bRol* etiketatzaileak 75.1 puntuko *Labeled* F_1 neurria du. 2.15 taulak erakusten duen moduan, gure emaitza *CoNLL-2009* saioko hizkuntzen artean emaitzarik okerrena lortu zuen txinerarako emaitza (77.15) baino bi puntu baxuagoa da. Gure ustetan euskararako lortu dugun *Labeled* F_1 neurri baxuaren arrazoia, beste zazpi hizkuntzena baino txarragoa, erabilitako corpusaren tamaina (oso) mugatua da. *EPEC-RolSem* corpusean esaldi kopurua % 71.1 eta token kopurua % 80.1 txikiagoak direla egiazta daiteke. Uste dugu rol eta adjuntu etiketek *EPEC-RolSem* corpusean daukaten banaketak (ikus 2.14 taula) eta *Part-of-Speech* kategoria mota edo `FEAT` ezaugarri kopuruek (ikus 2.12 taula) ez dutela zailtasun berezirik inplikatzen, beste hizkuntzetakoenen aldean, euskaraz dependentzia semantikoak etiketatzerakoan.

Urrats semantikoak banaka

*bRol*ek gauzatzen dituen urrats semantikoen banakako emaitzak aztertuta, emaitzarik onena erdietsi duen urratsa (87.5) predikatuen identifikazioa dela ikusten da, predikatuen desanbiguazioa (79.82) eta argumentuen identifikazioa (78.80) ondoren, eta argumentuen sailkapena buruenik, emaitzarik baxuenarekin (77.5). Gure iritziz emaitza hauek lotura estua dute urrats bakoitzak daukan berezko zailtasunarekin. Predikatu identifikazioa, esate baterako, sailkapen bitarra da, eta predikatuen desanbiguazioa eta argumentuen sailkapena, berriz, klase anitzekoak (*multiclass*) dira. *bRol* sistemaren kasuan, adiera anitzeko predikatuen desanbiguazioa egiten duen sailkatzaileak 200 klase ezberdin dauzka, eta argumentuen sailkapena egiten duenak, berriz, 15.

Klase kopuruari erreparatuta, predikatuen desanbiguazioaren F_1 neurriak argumentuen sailkapenarena baino baxuagoa izan beharko lukeela pentsa daiteke, hein handi batean sailkatzaile baten klase kopuruak lantzen ari den urratsaren zailtasuna adierazten duelako. Kasu honetan kontuan izan beharra dago, hala ere, 2.16 taulan predikatuen desanbiguaziorako ematen den balioa adiera bakarreko eta anitzeko aditzen desanbiguaziorako lortutako emaitzen elkarketatik kalkulatu dela. Lehenengoaren F_1 neurria 100 da (% 100 asmatze tasa adieraren esleipena zuzena delako) eta bigarrenarena 70.01.

Osotara

bRol etiketatzaile osoaren eraginkortasuna zein den jakiteko, dependentzia sintaktiko eta semantikoen etiketatzean erdietsitako emaitzak uztartzen dituen *Labeled Macro* F_1 *Score* neurria kalkulatu dugu. Gure sistemak 77.8 puntuko *Labeled Macro* F_1 *Score* neurria lortzen du. Balio hau *CoNLL-2009* ebaluazio saioko emaitzekin alderatzen denean (ikus 2.15 taula), gure sistema txinerarako etiketatzailea (76.38) baino 1.42 puntu hobea dela ikus dezakegu. Ebaluazio saioko beste hizkuntzen emaitzekin alderatzean ikus daiteke gurea katalaneko (83.01), gaztelerako (83.31), txekierako (83.27) eta alemaneko (82.44) emaitzak baino bost puntu, japonierako emaitza (85.65) baino zortzi puntu eta ingeleseko emaitza (87.69) baino hamar puntu okerragoa dela. 2.4.1 atalean azaldu dugun bezala, ezin izan dugu domeinuz-kanpoko *bRol* sistemaren ebaluaziorik egin, baliabide faltarengatik, horrelako ebaluazioa aurrera eraman ahal izateko dependentzia sintaktiko-semantikoekin eskuz anotatutako eta *EPEC-RolSem* corpusaren domeinu berekoa ez den euskarazko corpusa beharko baikenuke.

2.4.2 atalean predikatuen desanbiguazioaren urratsaren barnean txertatu dugun itzulpen moduluaren azalpena egin dugunean aipatu dugu *bRol* ebaluatu ahal izateko itzulpen modulua desaktibatu egin dugula, honek etiketatutako dependentzia semantikoak ezin ebalua ditzakeelako. Uste izatekoa da honen ebaluazioa egin ahal izan bagenu, sistemaren emaitzak hobetu egingo ziratekeela, itzulpen moduluak etiketatzen dituen predikatu-argumentu-adjuntu egitura gehienetan predikatua adiera bakarrekoa izaten delako.

2.5 Ondorioak eta etorkizuneko lanak

Atal honetan euskararako lehenengo SRL etiketatzailearen garapen prozesua deskribatu dugu. Horretarako SRL atazaren ikerketaren egungo egoeraren erakusgarri diren beste hizkuntzetako etiketatzaileak, SRL egin ahal izateko eskuragarri dauden baliabide linguistikoak eta *CoNLL* ebaluazio saioak aurkeztu ditugu lehenik (2.1 azpiatala). Gero, gaur egungo SRL etiketatzaileek, eta gure *bRol* sistemak, jarraitzen duten arkitektura eta urratsak deskribatu ditugu, 2.2 azpiatalean. Jarraian, eta SRL etiketatzaile guztiz automatikoa den *bRol* sistema deskribatu baino lehen, *SRL prototipoa* deitu dugun argumentuen sailkapenerako etiketatzailea aurkeztu dugu (2.3 azpiatala). Honen garapenak azken SRL sistema garatzean baliagarriak izan zaizkigun hainbat baliabide eta parametro zehazteko balio izan digu: ikasketa algoritmorik egokiena, ezaugarrien aukeraketa, corpusaren formatua eta azterketa, roletarako eredua eta beste ikasteko. Azkenik, 2.4 azpiatalean *bRol* etiketatzailearen deskribapena egin dugu, eta gure sistemarenak *CoNLL-2009* ebaluazio saioko zazpi hizkuntzetarako lortu ziren emaitzekin alderatu ditugu. *bRol* tresnak 80.51 puntuko LAS, 75.1 puntuko *Labeled* F_1 eta 77.8 puntuko *Labeled Macro* F_1 *Score* neurriak erdietsi ditu. Aipatu behar da, bukatzeko, tesi lan honetan sortutako SRL tresna edonoren eskura dagoen *ixaKAT*[13] izeneko prozesamendu-katean gehitu dela.

Etorkizuneko lanen artean hurrengoak aurreikusi ditugu: *bRol* sistemaren domeinuz kanpoko ebaluazioa egitea, predikatuen desanbiguazio urratseko itzulpen moduluaren eraginkortasuna aztertzea, *bRol VerbNet* ereduko rolak esleitzeko egokitzea, izen, adjektibo eta adberbio predikatuak eta hauetatik sortzen diren dependentzia semantikoak etiketatzeko gaitasuna gehitzea eta, azkenik, sailkatzaileen eraikuntzarako teknika berriak aplikatzea.

[13] `http://ixa2.si.ehu.es/ixakat/`

3

DENBORA INFORMAZIOAREN ETIKETATZE AUTOMATIKOA

Hirugarren atal honetan euskaraz idatzitako testuetako denbora informazioaren etiketatze automatikoaz arduratzen gara. Sarreran azaldu dugunez, bi helburu dauzka ardura honek: euskararen analisi-katea aberastea batetik, eta tesian planteatzen ditugun bi hipotesietatik lehenbizikoa betetzen den edo ez aztertzea bestetik. Hipotesi honen arabera, euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek duten eragina positiboa da, ingelesez eta gaztelaniaz den bezala.

Atal honen lehenengo azpiatalean atazari dagokion ikerketaren egungo egoera laburbiltzen saiatuko gara. Horretarako, euskara ez den beste hizkuntzetarako eskuragarri dauden sistemak eta baliabide linguistikoak aztertuko ditugu, euskararako denbora informazioa etiketatzen duen sistema garatu ahal izateko erabili dugun *Euskal-TimeBank* corpusa deskribatuko dugu eta ataza horretaz arduratu diren *TempEval* ebaluazio saioak aurkeztuko ditugu. Gero, bigarren azpiatalean, denboraren etiketatzaileek izan ohi duten arkitektura azalduko dugu. Atala bukatu aurretik, *bTime*, sortu dugun euskarazko denbora etiketatzeko sistema azaltzeaz gain, egin ditugun esperimentuak eta lortutako emaitzak aurkeztuko ditugu. *bTime* etiketatzailearen emaitzak beste hizkuntza batzuenekin alderatu ditugu.

3.1 Ikerketaren egungo egoera

Rol semantikoez ez ezik, denboraren eta espazioaren etiketatze automatikoaz ere arduratzen garela azaldu dugu. Izan ere, predikatu-argumentu-adjuntu egituretako AM−LOC eta AM−TMP adjuntuek eta mugimenduko aditz batzuen kasuan argumentuek (*joan* adibidez) *non* eta *noiz* galderak erantzuteko balio baldin badute ere, hauek eskaintzen duten espazioaren eta denboraren informazioa mugatua dela uste dugu. SRLn eskuratzea lortzen den espazioaren eta denboraren informazioa predikatu baten argumentu edo adjuntuek ematen dutena da, eta hau gertaeraren unearen eta lekuaren adierazpenera mugatzen da. Ez da zehazten, esate baterako, zein diren gertaeren artean dauden lotura tenporal eta espazialak, ez dira adierazten informazioaren interpretazio automatikoan baliagarriak izan daitezkeen denbora adierazpenak zein diren, eta kokalekuak ez dira motaren arabera sailkatzen. Informazio hau guztia ere automatikoki eskuratzeko, tesian espazioaz eta denboraz espresuki arduratzen diren eskema berezituak erabiltzea aukeratu dugu. Atal honi dagokion denbora informazioaren kasuan *ISO-TimeML* eskema (Pustejovsky et al., 2010) erabili dugu.

3.1.1 Denbora markatzeko hizkuntzak

Azkeneko hamarkadan testuetako informazio tenporalaren anotazioaz egin den lan gehiena *TimeML* (Pustejovsky et al., 2003a) eta *ISO-TimeML* markaketa lengoaien inguruan burutu da. *ISO-TimeML* 2010. urtean aurkeztu zen eta *TimeML* lengoaiaren hobetutako bertsiotzat hartzen da. *TimeML* lengoaiaren oinarriak bi izan ziren: *TIDES TIMEX2* denbora adierazpenen anotaziorako gidalerroak (Ferro et al., 2001) eta Setzer-ek (2001) bere tesian proposatzen duen egunkari berrietako informazio tenporalaren anotaziorako lengoaia. *ISO-TimeML* lengoaiak, *TimeML*ren bertsio hobetua denez gero, denbora adierazpenak *TIMEX3* gidalerroei (Pustejovsky et al., 2005) jarraituta anotatzen ditu. Jarraian zehaztasun osoz aurkezten dugu tesi lan honetan garatutako *bTime* tresnak oinarritzat hartzen duen *ISO-TimeML* eskema.

ISO-TimeML eskema

Denborari dagokion eskemak sei etiketa bereizten ditu:

- <EVENT>: Gertaeren buru lexikalak markatzen ditu (*joan*, *pentsatu*, *dakusat*).

- <TIMEX3>: Denbora adierazpenak markatzen ditu (*urtarrilaren 31n*, *1923/4/23*, *ostiral arratsaldean*).
- <SIGNAL>: Denbora seinaleak markatzen ditu (*ondoren*, *baino lehen*, *aurretik*). Seinaleek denbora erlazioak (<TLINK>) adierazten dituzte, baina ez dute nahitaez agertu behar, denbora erlazioak gauzaturik ere.
- <TLINK>: Bi gertaeraren arteko edo gertaera baten eta denbora adierazpen baten arteko denbora erlazioak markatzen ditu (<TLINK> = *Temporal link*).
- <ALINK>: Aspektuzko gertaera baten (ikus 1.1.1) eta argumentuen arteko erlazioak markatzen ditu (<ALINK> = *Aspectual link*).
- <SLINK>: Bi gertaeraren arteko mendekotasun erlazioak markatzen ditu (<SLINK> = *Subordination link*).

ISO-TimeML eskemaren hasierako bertsioa ingelesez idatzitako testuak etiketatzeko garatu bazen ere, ordutik hainbat hizkuntza prozesatzeko egokitu da, besteak beste gure tesian erabili dugun euskararako egokitzapena garatzeko (Altuna et al., 2016a). Ondorengo adibideek euskarazko *ISO-TimeML* eskemaren erabilera erakusten dute:

Jarraitzaileek <TIMEX3> ***ostiral goizean*** <EVENT> ***ikusi*** *zuten partida,* <EVENT> ***ekaitzaren*** <SIGNAL> ***ondoren***.

Adibidean bi gertaera etiketatu dira: *ikusi* eta *ekaitzaren*. Lehenengoaren kasuan gertaera *ikusi zuten* bada ere, *ikusi* bakarrik markatu da, *ISO-TimeML* eskemaren egokitzapenean gertaeren buru lexikala besterik ez markatzeko erabakia hartu zelako. Etiketatu den bigarren gertaera berriz, *ekaitzaren*, izen predikatu baten bitartez deskribatutakoa da. Adibidean *ostiral goizean* denbora adierazpena eta *ondoren* seinale tenporala ere etiketatu dira.

ISO-TimeML eskemak, gainera, bi gertaeraren artean eta gertaera baten eta denbora adierazpen baten artean izan daitezkeen hiru erlazio mota markatzeko aukera ere ematen du (<TLINK>, <ALINK> eta <SLINK>). Aurreko adibideko esaldian etiketa daitezkeen denbora loturak (<TLINK>) hiru dira, horien markaketa honela egiten da:

(1) *Jarraitzaileek ostiral goizean* ***ikusi*** *zuten partida,* *ekaitzaren* *[ondoren].*

<TLINK>

Lehenengo esaldian *ondoren* seinaleak adierazten du *ikusi* eta *ekaitzaren* gertaeren arteko denbora erlazioaren gauzatzea.

(2) *Jarraitzaileek ostiral goizean* ***ikusi*** *zuten partida, ekaitzaren ondoren.*
<TLINK>

Bigarren esaldian, ordea, *ostiral goizean* eta *ikusi* gertaeraren arteko erlazioa da markatzen dena.

(3) *Jarraitzaileek* ***ostiral goizean*** *ikusi zuten partida, ekaitzaren ondoren.*
<TLINK>

Azkenik, hirugarren esaldian, *ekaitzaren* gertaera eta *ostiral goizean* adierazpena lotzen dira.

Ondoko bi adibideetan (4 eta 5) beste bi erlazio motak, aspektuzkoak (<ALINK>) eta mendekotasunezkoak (<SLINK>), nola markatzen diren erakusten da. Lehenengo motakoek aspektuzko gertaera baten eta honen argumentuen arteko erlazioak markatzeko balio dute. Bigarren motakoek, aldiz, bi gertaeraren arteko mendekotasun erlazioak markatzeko.

(4) *Jarraitzaileek partida* ***ikusten*** *amaitu zuten.*
<ALINK>

Lotura honetan, *amaitu*, *amaitu zuten* ASPECTUAL motako gertaeraren buru lexikaletik *ikusten* gertaerarako erlazio aspektuala markatu da.

(5) *Jarraitzaileek* ***uste*** *dute taldeak partida irabazteko aukera izango duela.*
<SLINK>

Beste lotura honek mendekotasun erlazioen markaketa erakusten du. Erlazio hauek gertaera nagusia hartzen dute iturritzat, kasu honetan *uste dute*, eta helburutzat berriz mendeko gertaera, kasu honetan *aukera izango dutela*.

ISO-TimeML eskemaren euskararako egokitzapena

Ingeleserako eta euskararako eskemen artean dauden ezberdintasunak aipatuko ditugu orain. Gainera, euskaraz gertaeretan eta denbora adierazpenetan dagoen kasuistika erakusten duten adibideak emango ditugu.

- **Gertaerak** (**<EVENT>**): Tesian zehar esan dugu gertaerak aditzak, adberbioak, izenak edo adjektiboak diren predikatuen bitartez deskribatutako jazoerak direla. Definizio orokor hau baliagarria izan zaigu SRLn predikatu-argumentu-adjuntu egiturak etiketatzeko motibazioa azaltzeko orduan. Denboraz ari garenean, hala ere, eta zehazki *ISO-TimeML* eskemaz, gertaeraren definizioa zertxobait aldatzen da izan ere anotazio eskema honek berezko definizioa baitu. Ingelesezko bertsioan, Sauriik eta bestek (2005) *TimeML* eskemaren gidalerroetan diotenez, gertaerak hurrengoek deskribatzen dituzten jazoerak dira: aldidun edo aldirik gabeko aditzek (jokatuek edo jokatugabeek), nominalizazioek, adjektiboek, esaldi predikatiboek eta preposizio sintagmek. Euskaraz ordea, Altunak eta bestek (2016) adierazten dutenez, beste hauek deskribatzen dituzten jazoerak dira: aldidun edo aldirik gabeko aditzek (1), nominalizazioek, gertaera jakinei erreferentzia egiten dieten izen arruntek eta bereziek (2), predikatu lana egiten duten adjektiboek (3), aditzei joskera konplexuko gertaerak adierazten laguntzen dieten adberbioek (4) eta esaldi predikatiboek, joskera generikoek, metaforek edo esapide idiomatikoak diren predikatu konplexuek (5). Ingelesez eta euskaraz *ISO-TimeML* eskemako gertaeren definizioa bi eratara egin dela ikusten dugu. Bi kasuetan, hala ere, emandako definizioak, oro har, oso antzekoak direla ikus daiteke.

  1. *NASAk urteak* ***daramatza*** *Marten ur bila.* [Aditz jokatua]
  2. ***UEFA 2016 Europar Txapelketa*** *Frantzian ospatuko da.* [Izen berezia]
  3. *Zer egin behar da filma* ***gustagarri*** *egiteko?* [Adjektiboa]
  4. *Emaitzak* ***harro*** *egoteko modukoak dira.* [Adberbioa]
  5. *Ikusleak* ***barrez lehertu ziren.*** [Metafora]

  Hiru aditz forma edo adizki daude euskaraz: trinkoa, perifrastikoa eta jokatugabea. Lehenengo biak jokatuak dira. Adizki trinkoak hitz bakarreko unitateak

dira, non erro lexikoak (hitz erroak) semantika informazioa baitakar. Forma trinko horietan aspektu, aldi, pertsona eta moduaren gaineko informazioa bideratzen duten morfemak aipatu erroari atxikitzen zaizkio. Adizki perifrastikoak, ordea, hitz beregain bat baino gehiago dituztenak dira. Azkenik, euskaraz, adizki jokatugabeak perpaus burutzat erabil daitezke, esapide ihartuetan, harridura perpausetan eta galderetan, adibidez (Altuna et al., 2016b). Testuetan identifikatutako gertaerek hartzen dituzten atributuei dagokienez, badira ezberdintasunak euskaraz eta ingelesez. Hain zuzen ere, aurrerago aurkeztuko dugun *TempEval-3* ebaluazio saioan, ingeleserako erabilitako corpusean, `class`, `tense`, `aspect`, `polarity` eta `pos` atributuak hartzen dituzte gertaerek; euskaraz, ordea, `class`, `tense1`, `tense2`, `aspect1`, `aspect2`, `polarity`, `pos` eta `modality` atributuak hartzen dituzte. Jarraian atributu hauetako bakoitzaren azalpena egiten dugu.

- `class`: Gertaerak esanahiaren arabera sailkatzen ditu. Sailkapena egiteko tesian jarraitzen dugun Sauriik eta bestek (2005) proposatutako kategorizazioa erabiltzen da euskaraz eta ingelesez.
- `tense1`: Gertaera orainaldikoa den edo ez.
- `tense2`: Gertaera iraganekoa den edo ez.
- `aspect1`: Gertaera burutua edo burutugabea den.
- `aspect2`: Gertaera etorkizunekoa den edo ez.
- `polarity`: Gertaera positiboa edo negatiboa den (sintaktikoki erabakitzen da). Ezeztatuta dauden gertaerak negatiboak dira eta gainerakoak positiboak.
- `pos`: Gertaeraren buru lexikala den tokenaren kategoria gramatikala.
- `modality`: Aditz modalen bitartez deskribatutako gertaeren buru lexikaletan aditzaren informazioa adierazteko erabiltzen da.

Euskaraz, ingelesez ez bezala, `tense1`, `tense2`, `aspect1`, `aspect2` eta `modality` atributuak izatearen arrazoia hau da: euskaraz aldia, aspektua eta modalitatea, sintetikoki adierazten direla, eta ez indoeuropar hizkuntza gehienetan bezala aditz askeen bitartez. Xehetasunak euskarazko *ISO-TimeML* gidalerroetan aurki daitezke[1]. 3.1 irudian atributuek hartzen dituzten balioak ikus daitezke.

[1] `https://addi.ehu.es/handle/10810/17305`

```
class ::= REPORTING | PERCEPTION | ASPECTUAL | I_ACTION |
          I_STATE | STATE | OCCURRENCE
pos ::=  VERB | NOUN | ADJECTIVE | ADVERB | PRONOUN | OTHER
tense1 ::= PRESENT | -PRESENT | NONE
tense2 ::= PAST | -PAST | NONE
aspect1 ::= PERFECT | -PERFECT | NONE
aspect2 ::= FUTURE | -FUTURE | NONE
polarity ::= NEG | POS {default, if absent, is POS}
modality ::= AHAL | NAHI | BEHAR | NONE
```

Irudia 3.1: Gertaeren atributuek hartzen ahal dituzten balioak.

Atributu hauek adierazten duten gertaeren inguruko informazioa hobeki ulertzeko segidan ageri den adibidea aztertuko dugu.

Zerbitzu bereziek urriaren 28tik aurrera jarraituko dute.

↓

<EVENT>*Zerbitzu***</EVENT>** *bereziek* **<TIMEX3>***urriaren 28tik***</TIMEX3> <SIGNAL>***aurrera***</SIGNAL> <EVENT>***jarraituko***</EVENT>** *dute.*

Esaldian bi gertaera egon arren, *jarraituko dute* gertaera hartuko dugu aintzat (buru lexikala bakarrik markatuta). Gertaera hori bere atributu guztiekin etiketatuta hemen ikus daiteke:

<EVENT m_id="19**" tense2="**-PAST**"tense1="**PRESENT**" aspect1="**-PERFECT**"aspect2="**FUTURE**" polarity="**POS**" class="**ASPECTUAL**" modality="**NONE**" pos="**VERB**">***jarraituko***</EVENT>**

`class` atributuak `ASPECTUAL` balioa jasotzen du. Horrek esan nahi du *jarraituko* beste gertaera baten hasiera, jarraitutasuna edo amaiera adierazten ari dela, kasu honetan *zerbitzu*ren jarraitutasuna. `tense1` eta `tense2` atributuek, aldiz, `PRESENT` eta `-PAST` balioak jasotzen dituzte, hurrenez hurren. Aditz denborak bi dimentsio ditu euskarazko *ISO-TimeML* eskeman (±`PRESENT` eta ±`PAST`),

eta horiek uztartuz lortzen da denboraren anotazioa. Kasu honetan *jarraituko dute* gertaerak ez du lehenaldia adierazten (−PAST), baina bai orainaldia (PRESENT).

Altunak eta bestek (2016) diotenez, tempusa markatzeko orduan aditz laguntzaileari begiratuko zaio aditz perifrastikoen kasuan, eta tempusa adierazten duen morfemari aditz trinkoenean. Aditzak ez diren predikatuen bitartez deskribatutako gertaeretan tense1 eta tense2 atributuek NONE balioa jasotzen dute, denborarik adierazten ez delako. Adibidean aspect1 eta aspect2 atributuek −PERFECT eta FUTURE balioak hartzen dituzte. Denborarekin gertatzen den bezala, aspektua ere bi dimentsioren bitartez adierazten da euskarazko *ISO-TimeML* eskeman (±PERFECT eta ±FUTURE). Adibidearen kasuan −PERFECT eta FUTURE balioek *jarraituko dute* gertaera burutugabea eta etorkizuna adierazten duena dela markatzen dute. polarity atributuari dagokionez POS balioa (positiboa) hartzen du adibidean sintaktikoki ezeztatua ez dagoelako. pos atributuak, berriz, VERB balioa jasotzen du *jarraituko* tokenaren *Part-Of-Speech* kategoria delako. Azkenik, modality atributuaren balioa NONE dela ikus dezakegu, honek esan nahi du esaldian ez dagoela adibideko gertaerari dagokion aditz modalik (*nahi*, *behar* edo *ahal*) eta gertaera deskribatzen duen adizkia ez dela ahalerakoa.

- **Denbora adierazpenak** (**<TIMEX3>**): Hirugarren atal honetan jarraitzen dugun euskararako *ISO-TimeML* eskemak testuetako denbora adierazpenak nola etiketatu behar diren ere zehazten du. Honen arabera, euskaraz adierazpen tenporalak ondorengoak izan daitezke: izen, adjektibo eta adberbio sintagmak (1,2 eta 3 adibideak), postposizioa duten izen sintagmak (4), esaldi berezi batzuk (5) eta daten eta orduen berri ematen duten adierazpenak (6).

1. ***Igande goizean*** *paseotara joan ginen.* [Izen sintagma]
2. *The Black Dwarf argitalpen* ***astekaria*** *izan zen.* [Adjektibo sintagma]
3. *Garaipena* ***berandu*** *iritsi zen.* [Adberbio sintagma]
4. *Moskurako trena* ***bostak aldera*** *abiatuko da.* [Izen sintagma + Postposizioa]
5. *Abioirako atea* ***duela hamar minutu*** *itxi dute.* [Esaldi berezia]
6. *Uda* ***2016-06-20an*** *hasten da.* [Data]

Denbora adierazpenek jasotzen dituzten atributuei dagokienez, esan beharra dago euskaraz eta ingelesez berberak erabiltzen direla: `type`, `value`, `mod`, `temporalFunction`, `anchorTimeID`, `functionInDocument`, `valueFromFunction`, `beginPoint`, `endPoint`, `freq`, `quant` eta `comment`. Hamabi atributu hauetatik, halarik ere, bi bakarrik izan dira tesian landu direnak: `type` eta `value`. Izan ere, atributu hauek dira denboraren ebaluazio saioek (*TempEval*) *tradizioz* landu dituztenak, eta ondorioz, garatu dugun *bTime* denbora etiketatzaileak automatikoki tratatu dituenak, besteak beste, adierazpenen gaineko izaera linguistikoaz informazioa atxiki duten bakarrak direlako. Beste hamar atributuen gaineko informazioa ingeleseko edo euskarako gidalerroetan aurki daiteke.

- `type`: Adierazpenak motaren arabera sailkatzen ditu. 3.2 irudian ikus daitezke atributu honek hartzen dituen lau motak. `DATE` (eguna baino handiagoa), `TIME` (eguna baino txikiagoa), `DURATION` (irupena) eta `SET` (errepikapena).
- `value`: Adierazpenaren normalizazioa, `ISO-8601` estandarrean oinarrituta.

```
type ::= 'DATE' | 'TIME' | 'DURATION' | 'SET'
value ::= CDATA
```

Irudia 3.2: Euskarazko *ISO-TimeML* eskeman denbora adierazpenen `type` eta `value` atributuek hartzen ahal dituzten balioak.

Esaldiko *urriaren 28tik* denbora adierazpenaren atributuen balioei dagokienez hurrengo adibidean ikusten dugu `type` atributuak `DATE` balioa hartzen duela. Honek esan nahi du denbora adierazpena data motakoa dela. `value` atributuak, bestalde, `2007-10-28` balioa jasotzen du. Hau adibideko esaldiaren sorrera data erabilirik ($DCT = 2007$) sortutako adierazpenaren forma normalizatu edo *eskematikoa* da.

<TIMEX3 m_id="31**" type="**DATE**"**

value="2007-10-28**">***urriaren 28tik***</TIMEX3>**

- **Denbora erlazioak** (**<TLINK>**): Lehenago azaldu dugun moduan, *ISO-TimeML* eskemak hiru erlazio mota nagusi bereizten ditu: bi gertaeraren edo gertaera ba-

ten eta denbora adierazpen baten artekoak (<TLINK>), aspektuzko gertaera baten (class="ASPECTUAL") eta bere argumentuen artekoak (<ALINK>), eta bi gertaeraren arteko mendekotasunezkoak (<SLINK>). Hiru erlazioentzat ingeleseko eta euskarako eskemetan zehazten den anotatzeko modua eta hauek jasotzen dituzten atributuak berak dira. Tesian landu duguna <TLINK> erlazioa mota da. Izan ere, erlazio mota hau baita *bTime* tresna automatikoa garatzeko erabilitako *Euskal-TimeBank* corpusean kopuru minimo bat anotatuta duen erlazio mota bakarra. Honek jasotzen dituen atributuei doakienez, esan beharra dugu erlazioaren izaera linguistikoaz informazioa atxiki duen bakarra relType izenekoa dela. Honez gain denborazko erlazioek dituzten atributuen inguruko informazioa euskarako edo ingeleseko gidalerroetan aurki daiteke.

– relType: Denborazko erlazioak motaren arabera sailkatzen ditu. 3.3 irudian ikus daitezke zerrendatuta dauden hamahiru motak.

```
relType ::= 'BEFORE' | 'AFTER' | 'IBEFORE' | 'IAFTER' | 'INCLUDES' |
            'IS_INCLUDED' | 'MEASURE' | 'SIMULTANEOUS' | 'BEGINS' |
            'BEGUN_BY' |'ENDS' | 'ENDED_BY' | 'IDENTITY'
```

Irudia 3.3: relType atributuak hartzen ahal dituen balioak.

Jarraian aurreko azpiataleko erlazio tenporalen adibideak erabiliko ditugu relType atributuek hartzen dituzten balioak azaltzeko. Denbora osagaien artean ondoko hiru denborazko erlazioak daude.

(1) *Ostiral goizean* ***ikusi*** *zuten partida,* *ekaitzaren* *[ondoren].*

<TLINK relType="BEFORE**">**

Lehenbizikoan *ondoren* seinaleak adierazten du *ikusi* eta *ekaitzaren* gertaeren arteko denbora erlazioaren gauzatzea. Honek *ekaitzaren* gertaera *ikusi* baino lehen jazo dela markatzen du, eta horregatik da relType atributuaren balioa BEFORE.

(2) *Ostiral goizean* ***ikusi*** *zuten partida, ekaitzaren ondoren.*

<TLINK relType="INCLUDES**">**

Bigarrenean, ordea, ez dago *ostiral goizean* denbora adierazpenaren eta *ikusi* gertaeraren arteko erlazioaren gauzatzea adierazten duen seinalerik. Erlazio honek

ikusi gertaera *ostiral goizean* adierazpenaren barnean agitu dela markatzen du, eta horregatik dauka `relType` atributuak `INCLUDES` balioa.

(3) ***Ostiral goizean*** *ikusi zuten partida*, *ekaitzaren* *ondoren*.
<TLINK relType="INCLUDES**">**

Azkenekoan *ekaitzaren* gertaera *ostiral goizean* adierazpenak mugatzen duen denbora tartearen barnean kokatzen dela markatzen da, eta horregatik du honek ere `INCLUDES` balioa.

3.1.2 Etiketatzaileak

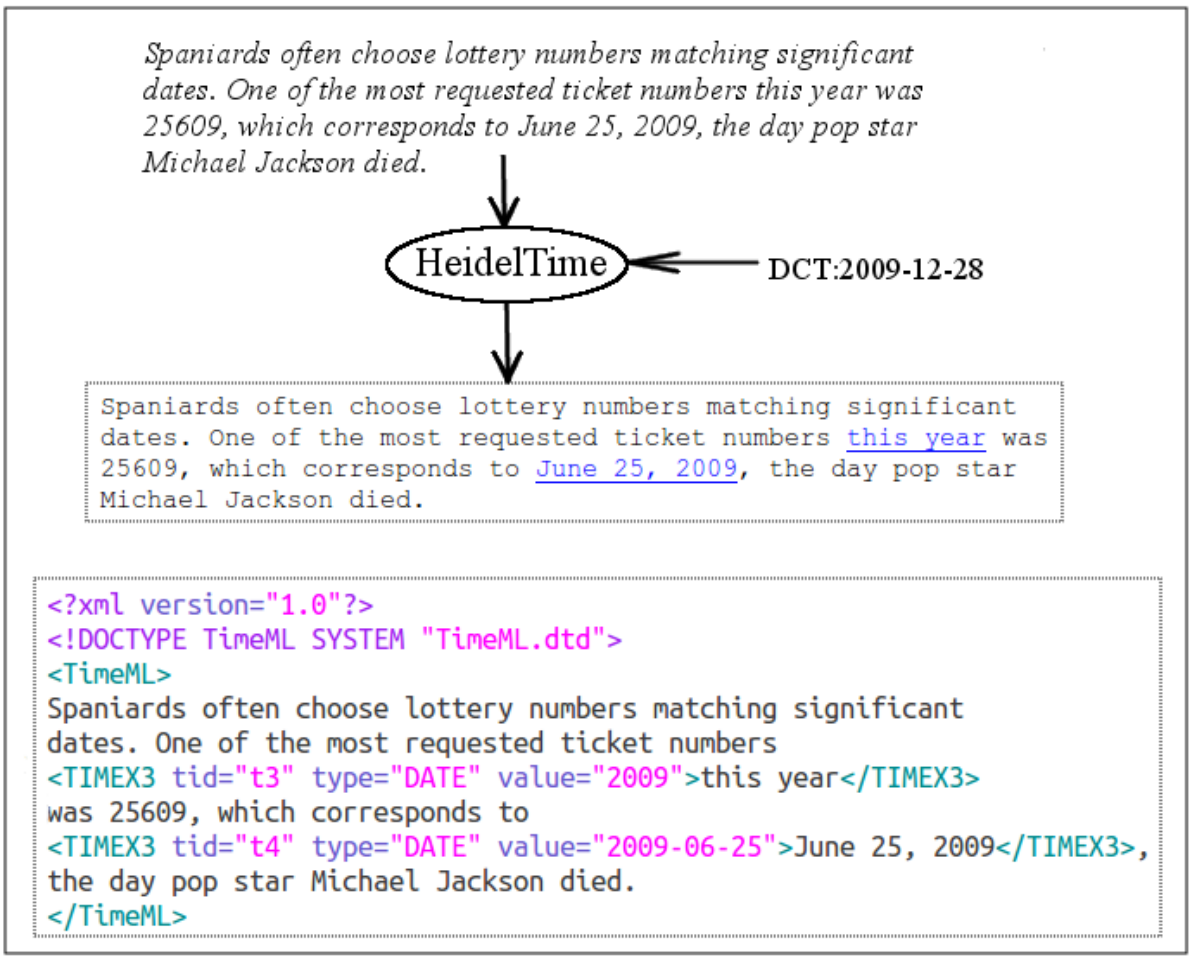
2003. urtean *TimeML* eskema aurkeztu zenetik hainbat sistema garatu dira denbora informazioaren etiketatze automatikoa egiteko, gehienak ingeleserako. Guk dakigula ez dira oso arruntak edo ugariak *TimeML*/*ISO-TimeML* gidalerroetako hiru osagai nagusiak (`<EVENT>`, `<TIMEX3>` eta `<TLINK>`) etiketatzen dituzten edota hizkuntza bat baino gehiago prozesatzeko gaitasuna daukaten sistemak (*end-to-end* motakoak). Badago, esate baterako, bakarrik denbora adierazpenak etiketatzen dituen eta hizkuntza bat baino gehiagorako balio duen etiketatzaile ezagun bat: *HeidelTime* (Strötgen eta Gertz, 2010).

Denbora adierazpenak etiketatzeko tresnak

HeidelTime ez da denbora adierazpenak soilik etiketatzen dituen tresna bakarra, baina bai ezagunena eta hizkuntza gehien lantzeko aukera ematen duena. Denbora adierazpenak etiketatzeko sistemen beste adibide batzuk hauek dira: erregeletan eta ikasketa automatikoan oinarritzen den (Ramrakhiyani eta Majumder, 2015) argitalpenean aurkeztutako hindirako sistema eta Mirzak-ek (2015) deskribatzen duen indonesierarako etiketatzailea.

HeidelTime adierazpenen etiketatzailearen lehen bertsioa argitaratu zenean hizkuntza bakoitzerako eskuz idatzitako erregela multzoak erabiltzen ziren testuetako denbora adierazpenak identifikatu eta normalizatzeko. Normalizazio prozesuaren helburua adierazpenen forma eskematikoa lortzea da (1993ko abenduaren 16a eta `1993/12/16` esate

baterako). Gaur egun *HeidelTime*k eskuz eta automatikoki sortutako erregela multzoak ditu. Eskuzkoak hiztun kopuru handiak dauzkaten hamahiru hizkuntzari dagozkie (ingelesari, alemanari, frantsesari, gaztelerari, txinerari eta abarri) eta hauek erabilita doitasun eta hedadura maila altuak lortzen dira orokorrean. Automatikoki sortutako erregela multzoak, aldiz, 200 hizkuntza[2] baino gehiagotarako daude (albanierarako, armenierarako, faroerarako eta bestetarako), eta haiekin lortzen diren emaitzetan estaldura oso apala izaten da, erregela multzoen tamaina mugatuaren eraginez (Strötgen eta Gertz, 2015). 3.4 irudian *HeidelTime* tresnaren funtzionamendua irudikatzen duen adibidea ikus daiteke.



Irudia 3.4: *HeidelTime* etiketatzaileak daukan funtzionamenduaren adibidea.

[2] https://github.com/HeidelTime/heideltime/tree/master/resources

Irudian sarrera moduan emandako testua *HeidelTime*k nola etiketatu duen ageri da. Testua prozesatu baino lehen, sarrera dokumentuaren sorrera data, *Document Creation Time* (*DCT*) delakoa eman zaio, 2009ko abenduaren 28a hain zuzen ere. Sistemak bi denbora adierazpen identifikatu ditu: *this year* (aurten) eta *June 25, 2009* (2009ko ekainaren 25a). Biak sailkatu dira data bezala (`type="DATE"`). Identifikatutako lehenbiziko adierazpenaren normalizazioa 2009 da (`value="2009"`) sarrera dokumentuaren sorrera data 2009ko abenduaren 28a dela zehaztu dugulako. Izan ere, denbora informazioarekin etiketatu nahi diren testuetan *DCT*ak finkatzea ezinbestekoa izaten da, denboran (*timeline*) sistemak identifikatzen dituen gertaerak eta adierazpenak hauen arabera ordenatzen direlako. Adibideko bigarren adierazpenaren normalizazioa, ordea, 2009-06-25 da (`value="2009-06-25"`).

end-to-end motako tresnak

Adierazpenen etiketatzaileez gainera badira denbora informazioa eskuratzeaz arduratzen diren beste zenbait sistema ere. *end-to-end* tresnak direlakoak hain zuzen ere, gorago aipatu ditugun *TimeML/ISO-TimeML* gidalerroetako hiru osagai nagusiak etiketatzen dituztenak. *end-to-end* sistemen adibideak EVENTI saioan (Caselli et al., 2014) parte hartu zuen italierarako *Fbkhlt-time* sistema (Mirza eta Minard, 2014), eta *TempEval-2* (Verhagen et al., 2010) eta *TempEval-3* (UzZaman et al., 2012) ebaluazio saioetan parte hartu zuen ingeleserako eta gaztelerarako *TipSem* sistema (Llorens et al., 2010) dira.

3.1.3 Corpusak

Testuetan informazio tenporala etiketatzen duten sistemak garatzeko beharrezkoak izaten dira informazio mota hau eskuz anotatuta daukaten corpusak. Denboraren etiketatze automatikorako dauden baliabide linguistikoak zerrendatzerakoan kontuan izan beharra dago ez dagoela SRLn adina baliabide eta eredu edo eskema ezberdin. Arestian aurkeztu dugun *TimeML/ISO-TimeML* eskemaz gain besterik ez dago. Jarraian *ISO-TimeML* eskeman oinarrituta anotatutako ingeleseko *TimeBank* corpusa (Pustejovsky et al., 2003b) deskribatuko dugu. Hau izan zen denbora informazioarekin anotatutako lehenengo corpus ezaguna, eta gainerako hizkuntzetarako ondoren sortu diren mota bereko corpusen erreferentzia. Honetaz landa, euskararako *ISO-TimeML* eskemaren egokitzapena jarraituta garatutako *Euskal-TimeBank* corpusa ere aurkeztuko dugu.

TimeBank

TimeBank corpusa ingelesez idatzitako egunkari berriez osatuta dago. Berri hauetan, *TimeML* eskema erabilita, gertaerak (<EVENT>), denbora adierazpenak (<TIMEX2> edo <TIMEX3> corpusaren bertsioaren arabera), erlazio tenporalak (<TLINK>, <ALINK> eta <SLINK>) eta seinaleak (<SIGNAL>) daude anotatuta (ikus 3.5 irudia). Corpusaren lehen bertsioa 2003. urtean argitaratu zen eta hiru iturritatik hartutako 300 testuk osatuta dago. Testuen % 12 *Document Understanding Conference-DUC*[3] ebaluazio saioetan (gaur egungo *Text Analysis Conference-TAC*[4]) erabilitako corpusetatik datoz, % 33 *Automatic Content Extraction-ACE*[5] ikerketa proiektutik eta gainerako % 55 *PropBank* corpusetik.

```
There was <SIGNAL sid="8"> no </SIGNAL> <EVENT EID="57" CLASS="STATE"
TENSE="PAST" ASPECT="NONE"> hint of trouble </EVENT> <SIGNAL id="11">
in </SIGNAL> the last <EVENT class="OCCURRENCE" aspect="NONE"
eid="10" tense="NONE"> conversation </EVENT> between controllers and
TWA pilot Steven Snyder. But <TIMEX3 TID="58" val="PT1M30S"
type="DURATION" temporalFunction="false"> a minute and a half
</TIMEX3> <SIGNAL SID="59"> later </SIGNAL>, a pilot from a nearby
flight <EVENT aspect="NONE" eid="18" tense="PRESENT"
CLASS="REPORTING"> calls </EVENT> in.

<SLINK eventInstanceID="eiid57" signalID="8" relType="NEGATIVE"/>
<TLINK eventInstanceID="eiid57" relatedToEvent="eiid10"
       signalID="s11" relType="IS_INCLUDED"/>
<TLINK eventInstanceID="eiid18" relatedToEvent="eiid10"
       signalID="s59" relType="AFTER" magnitude="t58"/>
```

Irudia 3.5: *TimeBank* corpusaren adibidea (Pustejovsky et al., 2003b).

Corpusaren bigarren eta azken bertsioa 2006. urtean argitaratu zen. Honetan 183 testu edo agiri aurki daitezke. Boguraevek eta bestek (2007) diotenez, corpus honetan anotatutako denbora informazioa lehenbiziko argitalpenekoa baino nabarmen garbiagoa eta fidagarriagoa da. *TimeBank* orain *Linguistic Data Consorcium-LDC* izeneko kontsortzioak eskaintzen duen hizkuntza baliabideen katalogoaren barnean[6] dago eskuragarri.

[3] http://duc.nist.gov/
[4] https://tac.nist.gov/tac/
[5] https://www.ldc.upenn.edu/collaborations/past-projects/ace
[6] https://catalog.ldc.upenn.edu/LDC2006T08

Euskal-TimeBank

Euskal-TimeBank da gaur egun euskararako denbora informazioarekin anotatuta dagoen corpus bakarra. Honen garapenerako *MEANTIME* corpusaren (Minard et al., 2016) euskarazko bertsiotik hartutako 30 dokumentu *ISO-TimeML* gidalerroen[7] egokitzapenaren arabera anotatu ziren. *MEANTIME* corpusa *NewsReader* proiektuan (Agerri et al., 2014) erabili zen, eta semantikoki anotatutako (SRL, NER, etab.) *Wikinews*[8] berriek osatzen dute.

Euskal-TimeBank corpusaren garapena bi fasetan egin zen, ingeleseko *TimeBank*en garapen prozesuan oinarrituta. Lehenik, gertaera, denbora adierazpen eta erlazio tenporal kopuru mugatu bat eskuz anotatu zen. Altunak eta bestek (2016) diotenez, hiru pertsonak hartu zuten parte lehenengo fasean. Hauen arteko adostasuna (*Inter-Annotator Agreement-IAA*) neurtzeko erabiltzen den *Dice*-en koefizientean (Dice, 1945), anotatzaile bikotearen arabera, 0.864 eta 0.947 bitarteko batezbesteko balioak iritsi ziren. Bigarrenik, gidalerroak eguneratu egin ziren, anotazioen analisi gramatikala eginda, eta analisi honen ondorioz sortutako gidalerroen bigarren bertsioarekin corpuseko 30 dokumentuak osorik anotatu ziren. Bigarren fasean lau pertsonak hartu zuten parte, eta 0.812 eta 0.883 bitarteko batezbesteko adostasuna lortu zen.

3.1 taulan *Euskal-TimeBank* corpuseko zenbait datu bildu dira: dokumentu eta token kopuruak, anotatutako gertaera eta denbora adierazpenak eta gertaeren eta dokumentuen sorrera daten arteko erlazio tenporalak.

	Train	**Test**
Dokumentuak	20	10
Tokenak	5,749	3,884
Gertaerak	1,133	760
Adierazpenak	343	208
TLINK [Gertaera/DCT]	552	388

Taula 3.1: *Euskal-TimeBank* corpuseko informazio kopuruaren datuak.

[7] https://addi.ehu.es/handle/10810/13803
[8] https://www.wikinews.org/

3.1.4 Ebaluazio saioak

Azpiatal honetan testuetako denborazko informazioa automatikoki etiketatzeaz arduratu diren ebaluazio saioak aurkeztuko ditugu. Hauek garrantzia handia izan dute atazaren garapenean. Gaur egun arte antolatutako nagusienak bost izan dira. Horietako hiru, garrantzitsuenak izan duten eraginarengatik, *SemEval* barnean antolatutako *TempEval-1*, *TempEval-2* eta *TempEval-3* saioak izan dira. Beste biak, berriz, *EVALITA* ebaluazio saioaren[9] (*SemEval* italiarra) barneko *TERN-2007* eta *EVENTI-2014* izan dira. Saio hauen helburua atazaren ikerketa eta interes zientifikoa sustatzea eta erabilitako konfigurazioak (metrikak, datu multzoak, etab.) eta lortutako emaitzak atazarentzat erreferentziazko bihurtzea da. Jarraian, aipatutako bost saioak kronologikoki ordenatuta aurkeztuko ditugu.

- ***TERN-2007*** (Lenzi eta Sprugnoli, 2007): Saioa hau italieraz idatzitako testuetan dauden denbora adierazpenen identifikazioaz eta normalizazioaz arduratu zen. Horretarako *ACE TIMEX2* gidalerroak jarraitu ziren. Ebaluazio saio honetan, denbora adierazpenek *ACE TIMEX2* gidalerroetan hartzen dituzten bost atributuei balioak ezartzeari deitu zitzaion *normalizazio*. Atributuak ondorengoak dira: `value`, `anchor_val`, `anchor_dir`, `mod` eta `set`. Parte hartu zuten sistemen entrenamendu eta ebaluaziorako erabili zen corpusari dagokionez, 525 egunkari berriez osatutako *I-CAB* (*Italian-Content Annotation Bank*) corpusa erabili zen (Magnini et al., 2006). Ebaluazio saioan parte hartzaileei bi modalitate eskaini zitzaizkien: (1) denbora adierazpenen identifikazioa bakarrik egitea (`id`) edo (2), adierazpenen identifikazioa ez ezik, haien normalizazioa ere egitea (`id` + bost atributuak).

  Denetara, lau taldek hartu zuten parte. Horietako hiruk aurkeztutako sistemek identifikazioa eta normalizazioa egiten zuten eta laugarrenak identifikazioa baizik ez. Ebaluazioaren emaitzak itzultzeko doitasun, estaldura eta F_1 neurri estandarrak eta *relaxed* eskema erabili ziren. Horrela, sistema batek *domenica mattina* (igande goiza) denbora adierazpenetik *domenica* tokena bakarrik identifikatzea lortzen bazuen, adibidez, puntuak jasoko zituzkeen bi tokenetako bat ongi identifikatzeagatik. Parte hartu zuten sistemek iragarritako adierazpenengatik puntuak jaso ahal izateko, iragarpenek eta corpusean zegozkien eskuzko anotazioek % 30ean

[9] `http://www.evalita.it/`

bat etorri behar zutela finkatu zen. Normalizazioa ere bat etortze hau betetzen zuten adierazpenen iragarpenentzat baizik ez zen ebaluatu. Bi modalitateetan emaitzarik onenak lortu zituen sistemaren balioak (Negri, 2007) 3.2 taulan bildu dira.

Osagaia/Ataza		Doitasuna	Estaldura	F_1
<TIMEX>	**id**	95.7	89.8	92.6
	id + 5 atr.	68.5	63.3	67.4

Taula 3.2: *TERN-2007* saioko emaitzarik onenak.

Taulan ikus daitekeenez, F_1 neurria 25 puntu apalagoa da denbora adierazpenen identifikazioa ez ezik haien normalizazioa ere egiten denean (92.6 eta 67.4 puntu). Honek denbora adierazpenen normalizazio zuzena egiteak daukan zailtasuna adierazten du.

- ***TempEval-1*** (Verhagen et al., 2007): Ebaluazio saio honetan ingelesez idatzitako testuetan eskuz anotatutako denborazko erlazioen sailkapena egiteaz (mota erabakitzeaz alegia; `relType`) arduratu ziren parte hartu zuten sistemak. Hauen entrenamendurako eta ebaluaziorako erabilitako corpusa *TimeBank* corpusaren bigarren bertsioa izan zen.

  TempEval-1 saioak A, B eta C izendatutako hiru azpiataza izan zituen, erlazio mota bakoitzeko bat. Atazak errazteko edo mugatzeko asmoz *Event Target List-ETL* deitutako zerrenda osatu zen. Bertan, *TimeBank* corpusean 20 agerpen edo gehiago zeuzkaten gertaerak bildu ziren (*ETL-gertaerak*). Azpiatazen helburua ondorengoen arteko erlazio tenporalaren sailkapena egitea izan zen: (A) esaldi bereko denbora adierazpenenen eta *ETL*-gertaeren artekoa, (B) *ETL*-gertaeren eta dokumentuaren sorrera dataren (*DCT*) artekoa eta (C) ondoz-ondoko esaldietako ($s_{i\pm1}$, s_i esaldi bakoitzarentzat) *gertaera nagusien* artekoa. *Gertaera nagusia* esaldi bateko erro sintaktikoa den predikatuak deskribatzen duen jazoerari esaten zaio.

  Ebaluazio saioan metodo estatistikoak, erregelak eta hurbilpen hibridoak erabiltzen zituzten sei taldek hartu zuten parte. 3.3 taulan ikus daitezke ataza bakoitzerako lortutako emaitzarik onenak (Puşcaşu, 2007). Ebaluazio saio honetan ere, atazak doitasun, estaldura eta F_1 neurri estandarrak erabilita ebaluatu ziren. Taulan ikusten den bezala, *strict* (S) eta *relaxed* (R) eskemen bitartez egin zen erlazioen kategorizazioaren ebaluazioa. Modu honetara `BEFORE` motakoa zen erlazio

tenporal bat sistema batek, adibidez, IBEFORE gisa sailkatzen bazuen, puntuak jasotzen zituen, *relaxed* ebaluazio eskeman IBEFORE, BEFORE erlazio motaren antzekoa izateagatik (semantikoki ere bai). Izan ere, BEFORE etiketak gertaera edo denbora adierazpen bat beste baten aurretik agitzen dela adierazten du, eta IBEFORE etiketak, berriz, gertaera edo denbora adierazpen bat justu beste baten aurretik agitzen dela. *Strict* ebaluazio eskeman adibideko sistemak ez zukeen punturik jasoko, erlazioa BEFORE gisa sailkatu ez zuelako. Emaitzei dagokienez, ikusten ahal da hiruretan emaitzarik onenak dauzkan azpiataza B dela.

Osagaia/Ataza		Eskema	Doitasuna	Estaldura	F_1
<TLINK>	A	S	62	62	62
		R	64	64	64
	B	S	80	80	80
		R	81	81	81
	C	S	54	54	54
		R	64	64	64

Taula 3.3: *TempEval-1* saioko emaitzarik onenak azpiataza bakoitzerako.

- ***TempEval-2*** (Verhagen et al., 2010): Denbora informazioaren etiketatzeaz arduratu zen hirugarren ebaluazio saioa izan zen. *TempEval-1* saioa hartzen zuen oinarritako, eta honekiko ondorengo bi berritasunak zeuzkan: ingelesaren sistema garatu ahal izateko ez ezik, beste bost hizkuntzarenak ere egiteko corpusak eskaintzen zituela eta hiru azpiataza berriri heltzen ziela. Denetara sei hizkuntza eskaini baziren ere (ingelesa, gaztelera, frantsesa, italiera, koreera eta txinera), parte hartu zuten hemezortzi sistemetatik hamabostek ingelesa besterik ez zuten aztergai, batek gaztelania, eta bik baizik ez hizkuntza bat baino gehiago (ingelesa eta gaztelera). Gainera, sistemetako batzuek ez zuten ebaluazio saioko sei azpiatazetan parte hartu.

  Ingeleserako erabili zen corpusa *TempEval-1*eko bera izan zen, errebisio eta aldaketa gutxi batzuekin. Gaztelerarako, berriz, une horretan garapen prozesuan zegoen *Spanish TimeBank* corpusaren zati bat erabili zen. Kontuan izan beharra dago *TempEval-1* saioan A, B eta C izeneko atazei *TempEval-2*n C, D eta E deitu zitzaiela. Berriak ziren hirurei, berriz, A, B eta F. Azkeneko hiru hauen ardurak ondorengoak izan ziren:

– A: Testuetako denbora adierazpenak identifikatu (`id`) eta hauen `type` eta `value` atributuei balioak esleitzea.

– B: Gertaerak identifikatu (`id`) eta `class`, `tense`, `aspect` eta `polarity` atributuen balioak esleitzea.

– F: Gertaera batek beste bat sintaktikoki menperatzen zuen gertaerez osatutako denbora erlazioen mota (`relType`) zehaztea.

*TempEval-2*ko C, D, E eta F azpiatazetan *strict* eskema baizik ez zen erabili. Adierazpenen eta gertaeren identifikazioa ebaluatzeko (`id`), berriz, *relaxed* eskema erabili zen. Gainera, *TempEval-2*n erlazioen sailkapenaz arduratzen ziren azpiatazak eta gertaeren eta adierazpenen atributuak ebaluatzerakoan (A eta B-ren bigarren zatiak) doitasuna bakarrik kalkulatu zen. 3.4 eta 3.5 tauletan bildu ditugu ingeleserako eta gaztelerarako lortu ziren emaitzarik onenak.

Osagaia/Ataza/Atrib			Eskema	Doitasuna	Estaldura	F_1
<EVENT>	**B**	**id**	R	81 (3)	86 (3)	83 (3)
		class	-	79 (3)	-	-
		tense	-	92 (7)	-	-
		aspect	-	98 (6)(7)	-	-
		polarity	-	99 (4)(6)(7)	-	-
<TIMEX>	**A**	**id**	R	90 (1)	82 (1)	86 (1)
		type	-	98 (2)	-	-
		value	-	85 (1)	-	-
<TLINK>	**C**	**relType**	S	65 (6)	-	-
	D			82 (3)	-	-
	E			58 (4)	-	-
	F			66 (5)	-	-

Taula 3.4: *TempEval-2* saioko ingeleserako emaitzarik onenak [(1):(Strotgen eta Gertz, 2010), (2):(Saquete Boro, 2010), (3):(Llorens et al., 2010), (4):(UzZaman eta Allen, 2010), (5): (Ha et al., 2010), (6): (UzZaman eta Allen, 2010), (7): (Grover et al., 2010)]

Ingeleserako C, D eta E azpiatazetan erdietsitako emaitzak (3.4 taula) alderagarriak dira *TempEval-1* ebaluazio saioan lortutakoekin. Tauletan adierazten den bezala, *TempEval-2*ko hiru azpiatazetan iristen da emaitzak hobetzea. Bi mintzairak alderatzean, ikusten da gertaeretan eta denbora adierazpenetan (An eta Bn)

Osagaia/Ataza/Atrib			Eskema	Doitasuna	Estaldura	F_1
<EVENT>	**B**	**id**	R	90 (3)	86 (3)	88 (3)
		class	-	66 (3)	-	-
		tense	-	96 (3)	-	-
		aspect	-	89 (3)	-	-
		polarity	-	92 (3)	-	-
<TIMEX>	**A**	**id**	R	95 (3)	87 (3)	91 (3)
		type	-	99 (8)	-	-
		value	-	83 (9)	-	-
<TLINK>	**C**	**relType**	S	81 (3)	-	-
	D			59 (3)	-	-
	E			-	-	-
	F			-	-	-

Taula 3.5: *TempEval-2* ebaluazio saioko gaztelerarako emaitzarik onenak [(3):(Llorens et al., 2010), (8):(Llorens et al., 2010), (9):Vicente-Díez et al. (2010)]

gaztelerarako erdietsitako emaitzak ingeleserakoak baino hobeak direla. Verhagenek eta bestek (2010) diotenez, gaztelerarako lortutako emaitzak hobeak izatearen arrazoia zein den ez dago argi, haien iritzian, ebaluazio saioan erabilitako corpusak txikiegiak dira, ondorio zuzenak atera ahal izateko.

- ***TempEval-3*** (UzZaman et al., 2012): Hau izan zen oraindainoko ebaluazio saioetatik laugarrena eta *TempEval* edizioetatik azkenekoa. *TempEval-3*k aurreko bi saioetan landutako azpiatazak hartzen zituen oinarritako. Aurreko saioan izan zen parte hartzearen ondorioz ingeleserako eta gaztelerarako corpusak baizik ez ziren eskaini edizio honetan. Ingeleserako corpus berria gehitu zen. Hau sortzeko eskuz anotatutako 6.000 tokeneko *TempEval-3 Platinum* deitu zuten corpusa eta automatikoki anotatutako 600.000 tokeneko *TempEval-3 Silver* corpusa elkartu ziren. Azken hau garatzeko, egunkari berriz osatutako *Gigaword* corpusa (Parker et al., 2011) anotatu zen *TempEval-2*n parte hartu zuten *TIPSem* (Llorens et al., 2013) eta *TRIOS* (UzZaman eta Allen, 2010) etiketatzaileen bitartez. Gaztelerarako *TempEval-2*ko *Spanish TimeBank* corpusaren lehen bertsio bukatua (Sauri eta Badia, 2012) erabili zen.

  *TempEval-3*n hiru azpiataza nagusi (A, B eta ABC) eta bi azpiataza *extra* (C eta C-*relation-only*) landu ziren. *TempEval-3*ko A eta B azpiatazak *TempEval-2*ko

A eta Bren antzekoak ziren, baina ez guztiz berdinak, ondoko ezberdintasun hauek baitzituzten: lehenik denbora adierazpenentzat ez zen `type` atributua zehazteko eskatu (A), eta gero gertaerentzat ez zen `tense`, `aspect` eta `polarity` atributuak zehazteko galdegin (B). ABC deitutako azpiatazan sistemek A-ko eta B-ko denbora adierazpenen eta gertaeren arteko erlazioak identifikatu behar zituzten, eta sailkapena egin (`relType`). ABC azpiataza da 3.1.2 atalean definitu ditugun *end-to-end* sistemek burutzen dutena, hau da, testu soiletik abiatuta *TimeML/ISO-TimeML* gidalerroetan zehazten diren hiru osagai nagusiak identifikatzen eta etiketatzen dituzte (`<EVENT>`, `<TIMEX>` eta `<TLINK>`). ABC-n honako denborazko erlazio hauek hartu ziren kontuan:

1. Ondoko esaldietako *gertaera nagusiak*. C_{TE1} eta E_{TE2} bezalakoa.
2. Esaldi bereko gertaerak. E_{TE2} bezalakoa.
3. Esaldi bereko gertaerak eta denbora adierazpenak. A_{TE1} eta C_{TE2} bezalakoa.
4. Gertaerak eta dokumentuaren sorrera data. B_{TE1} eta D_{TE2} bezalakoa.

C izendatutako *extra* azpiatazetako lehenbizikoan, bestalde, parte hartzaileei eskuz anotatutako gertaerak eta denbora adierazpenak eman zitzaizkien. Parte hartzaileek hauen arteko denbora erlazioak identifikatu eta motaren arabera sailkatu behar izan zituzten (`relType`). Bigarren *extra* azpiatazan (C-*relation-only*), berriz, eskuz anotatutako gertaerak, denbora adierazpenak eta hauen arteko erlazioak emanda, erlazio hauen mota (`relType`) erabaki behar izan zuten. Azkeneko hau *TempEval-1*eko A, B eta C eta *TempEval-2*ko C, D, E eta F azpiatazak bezalakoa zen. Aurrera baino lehen ondorengoa argitu beharra dagoela uste dugu: ABC atazan identifikatu beharreko erlazio motak (1-4) aitzineko *TempEval* edizioetako zer azpiatazarekin zetozen bat azaldu dugun arren, kontuan izan beharra dago bat etortze honek ez dituela alderagarri egiten edizio-arteko emaitzak, ABC-n testu soiletik abiatzen zelako eta besteetan, berriz, eskuzko anotazioetatik.

Saio honetan parte hartu zuten sistemen ebaluaziorako erabili ziren neurriak doitasuna, estaldura eta F_1 izan ziren, aurreko edizioetan bezala. Hala ere, *TempEval-3*k aurreko bi edizioekiko ezberdintasun nabarmenak izan zituen. A eta B azpiatazetan, esate baterako, *strict* eta *relaxed* ebaluazio eskemak, biak, erabili ziren gertaeren eta denbora adierazpenen identifikazioaren ebaluazioa egiteko

(id). Gertaeren kasuan, *strict* eta *relaxed* ebaluazioen emaitzak berdinak izan ziren, *TempEval-3*n erabilitako corpusetan gertaeren token bakarreko buru lexikalak zeudelako anotatuta. Emaitzak aurkezteko orduan F_1 neurria bakarrik eman zen *strict* eskemarako.

Osagaia/Ataza/Atrib			Eskema	Doitasuna	Estaldura	F_1
<EVENT>	**B**	**id**	S/R	81.44 (1)	80.67 (1)	81.05 (1)
		class	-	-	-	71.88 (1)
<TIMEX>	**A**	**id**	S	-	-	82.71 (3)
		id	R	89.36 (2)(5)	91.3 (2)(5)	90.32 (2)(5)
		value	-	-	-	77.61 (4)
<TLINK>	**ABC**	**relType**	TAS	34.08 (3)	28.4 (3)	30.98 (3)
	C			37.32 (3)	35.25 (3)	36.26 (3)
	C-ro			55.58 (8)	57.35 (8)	56.45 (8)

Taula 3.6: *TempEval-3* ebaluazio saioko ingeleserako emaitzarik onenak [(1):(Jung eta Stent, 2013), (2):(Chambers, 2013), (3):(Bethard, 2013), (4):(Strötgen et al., 2013), (5):(Chang eta Manning, 2013), (8):(Laokulrat et al., 2013)]

Osagaia/Ataza/Atrib			Eskema	Doitasuna	Estaldura	F_1
<EVENT>	**B**	**id**	S/R	91.7 (7)	86 (7)	88.8 (7)
		class	-	-	-	57.6 (7)
<TIMEX>	**A**	**id**	S	-	-	85.3 (6)
		id	R	96 (6)	84.9 (9)	90.1 (6)
		value	-	-	-	87.5 (6)
<TLINK>	**ABC**	**relType**	TAS	37.8 (7)	46.2 (7)	41.6 (7)
	C			-	-	-
	C-ro			-	-	-

Taula 3.7: *TempEval-3* ebaluazio saioko A eta B azpiatazetako gaztelerarako emaitzarik onenak [(6):(Strötgen et al., 2013), (7):(Llorens et al., 2010)]

Ebaluazio saio honetan, gainera, denbora adierazpenen eta gertaeren atributuak etiketatzerakoan sistemek lortzen zuten eraginkortasuna F_1 neurriaren bitartez adierazi zen, eta ez *TempEval-2*n bezala doitasunaren bitartez. ABC eta bi *extra* azpiatazak ebaluatzeko UzZaman eta Allenek (2011) proposatutako *Temporal Awareness Score-TAS* metrika erabili zen. Metrika honek erlazio tenporalen

etiketatzearen ebaluazioa egiten du, bertan parte hartzen duten entitateen identifikazioa (`<EVENT>`, `<TIMEX3>`) eta erlazioen kategorizazioa (`relType`) zuzena den edo ez kontuan edukita. Metrika honek doitasuna, estaldura eta F_1 neurriak itzultzen ditu. 3.6 eta 3.7 tauletan aurkezten ditugu ingeleserako eta gaztelerarako *TempEval-3*ko atazetan lortu ziren emaitzarik onenak.

Aipatutako tauletan biltzen diren balioek adierazten dutenez gaztelerarako lortutako emaitza gehienak altuagoak izan ziren ingeleserako lortutakoak baino, gertaeren `class` atributuaren etiketatzean eta denbora adierazpenen *relaxed* eskemako identifikazioan (`id`) izan ezik.

- ***EVENTI-2014*** (Caselli et al., 2014): Ebaluazio saio hau, *TERN-2007* bezala, italieraz idatzitako egunkari berrien prozesamenduaz arduratu zen. Horretarako, *Ita-TimeBank* corpuseko (Caselli et al., 2011) testuez osatutako *EVENTI* corpusa erabili zen. *EVENTI-2014* saioan *ISO-TimeML* gidalerroak jarraitu ziren. Ondorioz, denbora adierazpenen `value` eta `type` atributuak zehazteko eskatu zen, adierazpenen identifikazioaz gainera. Honi A azpiataza deitu zitzaion. Honetaz gain, B, C eta D izeneko beste hiru azpiataza ere landu ziren: B-n gertaerak identifikatu eta hauen `class` atributuen balioak esleitu behar ziren; C-n, aldiz, A eta B azpiatazetan identifikatutakoen arteko erlazio tenporalak ezarri eta hauek kategorizatu (`relType`) beharra zegoen; azkenik, D ataza C bezalakoa zen baina eskuz anotatutako gertaerak eta denbora adierazpenak emanda. C-n eta D-n ondorengoen arteko erlazio tenporalak hartu ziren aintzat:

  1. Esaldi bereko *gertaera nagusiak*.
  2. Esaldi bereko *gertaera nagusiak* eta *menpeko gertaerak*.
  3. Esaldi bereko gertaerak eta denbora adierazpenak. A_{TE1} eta C_{TE2} bezalakoa.

  *EVENTI-2014*n hiru sistemak hartu zuten parte eta hauetako bakoitzak hainbat exekuzio aurkeztu zituen. 3.8 taulan aurkezten ditugu ebaluazio saio honetako azpiataza bakoitzean lorturako emaitzarik onenak. Emaitzak alderagarriak dira *TempEval-3* saioan erdietsitakoekin. Casellik eta bestek (2014) diotenez, *TERN-2007*ko emaitzak eta *EVENTI-2014*koak ez dira erkagarriak, ondorengo bi arrazoiengatik: (1) *TERN-2007*ko denbora adierazpenen etiketatzea *ACE TIMEX2*

gidalerroak jarraituta egin zelako eta *EVENTI-2014*koena, ordea, *TimeML/ISO-TimeML* gidalerroak jarraituta; (2) batean eta bestean erabilitako ebaluaziorako metodoak, *id + 5 atributuak* azpiatazaren kasukoak, ezberdinak izan zirelako. Balioak *TempEval-3*koekin erkatzean (*relaxed-*F_1) ikus dezakegu italieraz denbora adierazpenak detektatzeko denboran lortzen diren emaitzak gazteleraz eta ingelesez lortzen direnak baino okerragoak direla. Gertaerak harrapatzeko orduan, aldiz, italierakoak ingelesekoak baino hobeak dira, baina gaztelerakoak baino apalagoak.

Osagaia/Ataza/Atrib			Eskema	Doitasuna	Estaldura	F_1
<EVENT>	**B**	**id**	S	-	-	86.7 (4)
		id	R	90.2 (4)	86.8 (4)	88.4 (4)
		class	-	-	-	67.1 (4)
<TIMEX>	**A**	**id**	S	-	-	82.7 (1)
		id	R	93.5 (2)	85.4 (2)	89.3 (2)
		value	-	-	-	70.9 (2)
		type	-	-	-	77.5 (3)
<TLINK>	**C**	**relType**	TAS	29.6 (5)	23.8 (5)	26.4 (5)
	D			74 (6)	73.1 (6)	73.6 (6)

Taula 3.8: *EVENTI-2014* ebaluazio saioko emaitzarik onenak [(1):(Mirza eta Minard,2014)(A1), (2):(Manfredi et al., 2014), (3):(Manfredi et al., 2014)(no ET), (4):(Mirza eta Minard,2014)(B1), (5):(Mirza eta Minard,2014)(C1), (6):(Mirza eta Minard,2014)(D1)]

3.2 Denbora etiketatzeko *end-to-end* arkitektura

Azpiatal honetan testuetako informazio tenporala automatikoki etiketatzen duten sistemen ohiko arkitektura aurkeztuko dugu. Zehazkiago, *end-to-end* etiketatzaileak izenekoen egitura deskribatuko dugu. Aurrera jarraitu aurretik, dena den, ondorengoa argitzea beharrezkoa dela uste dugu: *ISO-TimeML* aurkeztu dugunean, `<EVENT>`, `<TIMEX3>` eta `<TLINK>` etiketez gainera `<SIGNAL>`, `<ALINK>` eta `<SLINK>` etiketak ere zerrendatu ditugu, baina gero, euskarazko *ISO-TimeML*ren egokitzapena eta ebaluazio saioak deskribatu ditugunean azkeneko hirurak ez ditugu aipatu. Honen arrazoia da, informazio tenporala etiketatzeko garaian, beste ataza batzuetan agitzen den bezala, hasierako urteetako lanak atazak oinarri hartzen duen eskemaren zati (edo etiketa) batean bakarrik zentratzen direla, eta gero ikerketa lan hauek eskemak zehazten dituen

gainerako etiketetara progresiboki zabaltzen direla. Informazio tenporalaren eta *ISO-TimeML* eskemaren kasuan, lehenengo sistemak eta argitalpenak denbora adierazpenetan (<TIMEX2>/<TIMEX3> etiketan) zentratu ziren. Ildo beretik, *ISO-TimeML* eskemako etiketa gehien lantzen dituzten sistemak, *end-to-end* deriztenak dira, eta hauek, orain arte behintzat, <EVENT>, <TIMEX3> eta <TLINK> etiketez baizik ez dira arduratu. Pentsatzekoa da etorkizunean *ISO-TimeML*n zehazten diren sei etiketak landuko dituzten *end-to-end* etiketatzaileak garatuko direla. Dena den, kontuan izan beharra dago, etiketa guztiak lantzen dituzten sistemak garatu ahal izateko (normalean) hauek eskuz anotatuta dauzkaten corpusak behar direla, eta hauek sortzeko denbora eta kostu handia behar izaten dela. Corpusak izaten dira, hortaz, atazen modu bateko edo besteko garapena gehien baldintzatzen duten faktoreak.

Gure kasuan, tesian garatu dugun *bTime* denbora etiketatzailea sortzeko *Euskal-TimeBank* corpusa erabili dugu. Honen ondorioz *bTime* deskribatzean azalduko dugun bezala, gertaerak, denbora adierazpenak eta gertaeren eta dokumentuen sorrera daten arteko (*DCT*) erlazio tenporalak lantzen ditugu.

3.2.1 Hiru urratsetako prozesua

Jarraian aurkeztuko dugun *end-to-end* arkitekturak hiru urrats nagusi ditu, arestian azaldutakoaren ondorioz: gertaeren eta denbora adierazpenen etiketatzea (1,2) eta denbora erlazioen identifikazio eta kategorizazioa (3).

1. **Gertaeren etiketatzea**: Urrats honetan *ISO-TimeML* eskeman definitu diren gertaerak etiketatzen dira. Predikatuen, esaldien, sintagmen eta esapideen identifikazioa egiten da lehenbizi, eta gero hauek jasotzen dituzten atributuen balioak zehazten dira. Esan bezala bi aukera daude gertaerak etiketatzeko garaian: (a) gertaera osoak kontuan izatea edo (b) gertaeren buru lexikalak bakarrik.

a

Some historians **<EVENT>***state*</EVENT> *that* **<EVENT>***World War I*</EVENT> **<EVENT>***was caused*</EVENT> *by the* **<EVENT>***Industrial Revolution*</EVENT>.

b

Some historians **<EVENT>***state*</EVENT>** *that World* **<EVENT>***War*</EVENT>** *I was* **<EVENT>***caused*</EVENT>** *by the Industrial* **<EVENT>***Revolution*</EVENT>**.

Adibideko esaldian lau gertaera etiketatu dira: *state* (diote), *Industrial Revolution* (Industria Iraultza), *was caused* (eragin zuen) eta *World War I* (Lehen Mundu Gerra). Aukera hauetako bat edo beste hartzeak eragina dauka, besteak beste, *end-to-end* sistemaren ebaluazioan. Izan ere, *b* aukera jarraitzen denean gertaera bakoitzeko token bakarra etiketatzen da eta, ondorioz, *strict* eta *relaxed* ebaluazio eskemen arteko bereizketa egiteak zentzua galtzen du, biek balio bera itzultzen dutelako.

Testuetan gertaerak identifikatu eta hauen atributuei balioak esleitzeko erabiltzen den hurbilpenik arruntena ikasketa automatikoko teknikak erabiltzean oinarritzen da. Buru lexikala lantzen duten etiketatzaileen kasuan (*b*) identifikaziorako sailkatzaile bitarrak erabili ohi direla. Horiek testuetako token bakoitza (puntuazio markak eta antzekoak kenduta) gertaera baten buru lexikala den edo ez erabakitzen dute. Beste kasuan (*a*), ordea, identifikaziorako `BIO/IOB` motako sailkatzaileak erabiltzen dira. Hauetan testuetako token bakoitza gertaera baten hasierakoa (`B`-*Begin*), barnekoa (`I`-*Inside*) edo gertaera baten zati ez dena, kanpokoa (`O`-*Outside*), den erabakitzen da. Adibideko esaldiko (*a*) *World War I*-en, esate baterako, *World* tokenak `B` etiketa jasoko luke eta *War* eta *I* tokenek, berriz, `I` etiketak jasoko lituzkete.

Gertaeren atributuen esleipenerako ere ikasketa automatikoa erabiltzea izaten da arruntena. Bi balio posible bakarrik dauzkatenetan (`polarity`-n adibidez) sailkatzaile bitarrak erabiltzen dira eta gainerakoetan (`pos`-en esate baterako), berriz, *multiclass* motako sailkatzaileak.

2. **Denbora adierazpenen etiketatzea**: Hurrengo urratsa izaten da denbora adierazpenak identifikatzea eta hauen atributuen etiketatzea. Sistema batzuek gertaeren etiketatzearekin batera, paraleloan egin ohi dute, eta beste batzuek, aldiz, lehenago edo beranduago, bi pausuen arteko menpekotasunik ez dagoelako.

ISO-TimeML eskemako denbora adierazpenen etiketatzea aldatu egin daiteke hizkuntza batetik bestera. Izan ere, denbora informazioa adierazteko era bera, hein batean behintzat, mintzaira batetik bestera aldatu egiten baita.

Denbora adierazpenak identifikatu (`id`) eta kategorizatzeko (`type`) bi teknika daude: ikasketa automatiko bidezkoa eta denboraren fenomeno semantikoa aztertzetik hizkuntzalariek inferitutako erregelen bidezkoa. Normalizazioa (`value`), aldiz, beti egiten da heuristikoak erabilita, daukan kasuistika zabalaren ondorioz. Uste izatekoa den bezala, adierazpen tenporalen identifikazioan lehenbiziko teknika aplikatzearen zuzentasuna (eta honen arrakasta) corpusaren tamainaren araberakoa izaten da neurri handi batean. Corpusak ahalbidetzekotan, beraz, ML teknikak erabil daitezke.

Denbora adierazpenen identifikaziorako eta kategorizaziorako erregelen teknika erabiltzea erabakitzen baldin bada, ordea, bi bide ibili izan dira gaur egun arteko ikerketa lanetan: erregelak idatzi eta testuari haiek aplikatzen dizkion sistemak sortzea edo erregelak *HeidelTime* tresnak aplikatzeko moduan idaztea. Lehenbiziko aukera erabiltzen duten sistemen adibideak *FSS-TimEx* (Zavarella eta Tanev, 2013) eta *NavyTime* (Chambers, 2013) dira. Bigarren aukeraren etsenplua, berriz, txinerarako (Li et al., 2014) argitalpenean deskribatzen den etiketatzailea da.

3. **Denbora erlazioen identifikazioa eta kategorizazioa**: Urrats honetan aurreko bi urratsetan detektatu diren gertaeren eta adierazpenen arteko erlazioak bilatu eta kategorizatzen dira. *End-to-end* sistemek, normalean, denbora erlazioen identifikazioa eta kategorizazioa urrats bakar batean egiten dute. Horretarako, *multiclass* motako sailkatzaile bat erabiltzen dute. Honek, lehenbiziko eta bigarren urratsetan identifikatutako gertaerak eta denborazko adierazpenak eta dokumentuaren *DCT*a hartuta, sor daitezkeen erlazio tenporal posible guztiak sailkatzen ditu. Bi osagairen arteko lotura eza adierazteko, sailkatzaileari, `NO_RELATION` (erlaziorik ez) izeneko kategoria gehitzen zaio. Horrela, bi osagairen artean dagoen erlazioa mota honetakoa dela zehazten badu sailkatzaileak, bi hauen artean erlaziorik ez dagoela ulertzen da. Gainerako kasuetan, teknika hau baliaturik, erlazioa identifikatzea ez ezik, hura kategorizatzea ere erdiesten da.

3.2.2 Ebaluaziorako metrikak

Azpiatal honetan informazio tenporala etiketatzeko *ISO-TimeML* eskema oinarritako hartzen duten *end-to-end* sistemen ebaluaziorako metrikak deskribatuko ditugu, ondoko hauek hain zuzen: *strict* eta *relaxed* ebaluazio eskemen araberako doitasuna, estaldura eta F_1 neurriak, eta *Temporal Awareness Score-TAS* metrika.

- Osagaien identifikazioa: *Strict* eta *Relaxed* eskemak *end-to-end* sistemetan gertaeren eta denbora adierazpenen identifikazioaren ebaluazioa egiteko erabiltzen dira. *Strict* eta *Relaxed* eskemen araberako gertaeren eta adierazpenen identifikazioan lortutako neurriak nola kalkulatzen diren azaldu ahal izateko ondorengo adibide etiketatua ekar daiteke hona:

Historialariek diotenez, Lehen Mundu Gerraren ondorengo negua 1901eko negua baino luzeagoa izan zen.

↓

Historialariek **<EVENT id="e1">***diotenez***</EVENT>**, **<EVENT id="e2">***Lehen Mundu Gerraren***</EVENT>** *ondorengo* **<TIMEX3 id="t1">***negua***</TIMEX3> <TIMEX3 id="t2">***1901eko negua***</TIMEX3>** *baino luzeagoa* **<EVENT id="e3">***izan zen***</EVENT>**.

Adibidean hiru gertaera (`e1`, `e2` eta `e3`) eta bi denbora adierazpen (`t1` eta `t2`) etiketatu dira. Batzuk token bakarrekoak dira (`e1` eta `t1`) eta besteak, aldiz, token bat baino gehiagokoak (`e2`, `e3` eta `t2`). Adibidean ongi etiketatuta ageri dira gertaera eta denbora adierazpen guztiak; eman dezagun, hala ere, sistemak ez dituela denak ongi identifikatu eta hau dela itzulitakoa:

Historialariek **<EVENT id="***e1′***">***diotenez***</EVENT>**, **<EVENT id="***e2′***">***Lehen Mundu***</EVENT>** *Gerraren ondorengo* **<TIMEX3 id="***t1′***">***negua 1901eko***</TIMEX3> <TIMEX3 id="***t2′***">***negua***</TIMEX3>** *baino* **<TIMEX3 id="***t3′***">***luzeagoa***</TIMEX3>** *izan zen.*

Hainbat ezberdintasun ikus daitezke ongi etiketatutakoarekin erkatzean: $e2'$n ez da gertaera osoa identifikatzea lortu, $t2'$ adierazpeneko *1901eko* (w8) tokena

"$t1'$"adierazpenaren zatitzat identifikatu da, *luzeagoa* (w11) denbora adierazpentzat etiketatu eta $t3'$ bezala erazagutu da. Gainera, ez da `e3` izeneko gertaera identifikatzea lortu (ikus 3.9 taula).

Tokenak	**GOLD osagaiak**	**IRAGARRITAKO osagaiak**
w1=Historialariek w2=diotenez w3=Lehen w4=Mundu w5=Gerraren w6=ondorengo w7=negua w8=1901eko w9=negua w10=baino w11=luzeagoa w12=izan w13=zen	e1={w2} e2={w3, w4, w5} e3={w12, w13} t1={w7} t2={w8, w9}	$e1'$={w2} $e2'$={w3, w4} $t1'$={w7, w8} $t2'$={w9} $t3'$={w11}

Taula 3.9: Ebaluaziorako metriken adibideetan erabiltzeko aldagaiak.

Ezberdintasun hauek kontuan edukita, honela kalkulatuko genituzke gertaeren eta denbora adierazpenen identifikazioari dagozkion doitasuna, estaldura eta F_1 neurria.

$$Doitasuna = P = \frac{TP_{tok}}{TP_{tok} + FP_{tok}}$$

$$Estaldura = R = \frac{TP_{tok}}{TP_{tok} + FN_{tok}}$$

$$F_1 = 2 * \frac{P * R}{P + R}$$

Token mailako (*relaxed*) ebaluazioari dagozkion formuletan TP_{tok} aldagaiak adierazten du zein izan den automatikoki ongi etiketatua izan den token kopurua.

FN_{tok} aldagaiak, ordea, tresna automatikoak etiketatu ez duen baina etiketatzea behar duen token kopurua adierazten du. Azkenik, FP_{tok} tresnak automatikoki etiketatu duen baina etiketatzea behar ez duen token kopurua biltzen du.

Gertaerak

$$P = \frac{\#\{w2, w3, w4\}}{\#\{w2, w3, w4\} + \#\{\emptyset\}} = \frac{3}{3+0} = 1$$

$$R = \frac{\#\{w2, w3, w4\}}{\#\{w2, w3, w4\} + \#\{w5, w12, w13\}} = \frac{3}{3+3} = 0.5$$

$$F_1 = 2 * \frac{1 * 0.5}{1 + 0.5} = 0.67$$

Atal honetan darabilgun adibidean oinarrituta ikusten da gertaerei dagozkien sei tokenetatik (*w2*, *w3*, *w4*, *w5*, *w12* eta *w13*) hiru bakarrik identifikatu direla behar bezala (*w2*, *w3* eta *w4*). Automatikoki etiketatu den baina etiketatzea behar ez duen token kopuruari dagokionez, aldiz, ez da izan horrelako egoeran egon den tokenik. Azkenik, tresna automatikoak etiketatu ez duen baina etiketatzea behar duen token kopurua hirukoa izan da (*w5*, *w12* eta *w13*).

Adierazpenak

$$P = \frac{\#\{w7, w9\}}{\#\{w7, w9\} + \#\{w8, w11\}} = \frac{2}{2+2} = 0.5$$

$$R = \frac{\#\{w7, w9\}}{\#\{w7, w9\} + \#\{w8\}} = \frac{2}{2+1} = 0.67$$

$$F_1 = 2 * \frac{0.5 * 0.67}{0.5 + 0.67} = 0.57$$

Entitate mailako (*strict*) ebaluazioaren formuletan (ikus hurrengo orrialdean) *Iragarritakoak* aldagaiak adierazten du zein izan diren automatikoki etiketatutako entitateak. *Anotatutakoak* aldagaiak, berriz, eskuz markatutakoak biltzen ditu. Formuletan horien arteko ebakiduran gelditzen diren entitateak zenbatzen dira, eta gero, iragarritako edo anotatutako entitate kopuruekin zatitu.

$$Doitasuna = P = \frac{\#\{Iragarritakoak \cap Anotatutakoak\}}{\#\{Iragarritakoak\}}$$

$$Estaldura = R = \frac{\#\{Iragarritakoak \cap Anotatutakoak\}}{\#\{Anotatutakoak\}}$$

$$F_1 = 2 * \frac{P * R}{P + R}$$

Gertaerak

$$P = \frac{\#\{\{e1', e2'\} \cap \{\texttt{e1}, \texttt{e2}, \texttt{e3}\}\}}{\#\{e1', e2'\}} = \frac{1}{2} = 0.5$$

$$R = \frac{\#\{\{e1', e2'\} \cap \{\texttt{e1}, \texttt{e2}, \texttt{e3}\}\}}{\#\{\texttt{e1}, \texttt{e2}, \texttt{e3}\}} = \frac{1}{3} = 0.3$$

$$F_1 = 2 * \frac{0.5 * 0.3}{0.5 + 0.3} = 0.37$$

Atal honetan aurkeztu dugun egoeran bi gertaera identifikatu ditu sistemak: $e1'$ eta $e2'$. Horiek eskuz anotatutako hiru gertaerekin alderatzerakoan (e1, e2 eta e3) ikusten da lehenbiziko entitatea bakarrik lortu dela ongi identifikatzea (entitateak osatzen dituzten token guztiak kontuan izanda).

Adierazpenak

$$P = \frac{\#\{\{t1', t2', t3'\} \cap \{\texttt{t1}, \texttt{t2}\}\}}{\#\{t1', t2', t3'\}} = \frac{0}{3} = 0$$

$$R = \frac{\#\{\{t1', t2', t3'\} \cap \{\texttt{t1}, \texttt{t2}\}\}}{\#\{\texttt{t1}, \texttt{t2}\}} = \frac{0}{2} = 0$$

$$F_1 = 2 * \frac{0 * 0}{0 + 0} = 0$$

Adierazpenen kasuan ebaluazioa berdin egin da baina denbora adierazpenei dagozkien entitateak kontuan izanda.

Laburbilduz, beraz, eta adibidean ikus daitekeen moduan, *relaxed* eskema jarraitzen denean, identifikazioaren ebaluazioa egiteko asmatutako eta asmatu gabeko tokenak edukitzen dira kontuan. *Strict* eskema jarraitzen denean, berriz, gertaera edo denbora adierazpen osoa identifikatzea lortu den edo ez hartzen da aintzat. Gertaera edo denbora adierazpena osatzen duten token guztiak identifikatzen ez baditu, sistemak entitatea identifikatzea erdietsi ez duela ulertzen da.

- Atributuen sailkapena: Lehenik eta behin, jakin beharra dago ebaluazio hau egiteko ez dela beharrezkoa izaten entitateetako token guztiak ongi identifikatzea. Atributu esleipenaren doitasuna, estaldura eta F_1 neurria nola kalkulatzen diren azaldu ahal izateko, eta ulerterraztasuna dela medio, aurreko adibideko gertaerak (`e1`, `e2` eta `e3`) ongi identifikatu direla pentsatuko dugu. Gertaerek zortzi atributu har ditzakete euskaraz, adibide honetan, hala ere, atributu bakarra ebaluatuko dugu, `class` delakoa, atributu guztien ebaluazioa era berean egiten baita. Eman dezagun, beraz, ondorengoak direla adibideko gertaeren `class` atributuei dagozkien benetako balioak.

```
<EVENT id="e1"class="REPORTING"/>
<EVENT id="e2"class="OCCURRENCE"/>
<EVENT id="e3"class="OCCURRENCE"/>
```

Eta demagun beste hauek direla ebaluatu nahi dugun *end-to-end* sistemak itzultzen dituenak.

```
<EVENT id="e1"class="ASPECTUAL"/>
<EVENT id="e2"class="OCCURRENCE"/>
<EVENT id="e3"class="PERCEPTION"/>
```

Ikusten den bezala, hiru gertaeretatik bakar bati baizik (e2ri) ez dio esleitu etiketatzaileak `class` atributuaren balio egokia. Hori kontuan edukita honela kalkulatuko lirateke atributu honen doitasuna, estaldura eta F_1 neurria:

$$P_{\texttt{class}} = \frac{\#\{\forall x | x \in (Irag \cap Anot) \wedge Irag_{\texttt{class}}(x) == Anot_{\texttt{class}}(x)\}}{\#\{Irag\}}$$

$$R_{\texttt{class}} = \frac{\#\{\forall x | x \in (Irag \cap Anot) \wedge Irag_{\texttt{class}}(x) == Anot_{\texttt{class}}(x)\}}{\#\{Anot\}}$$

$$F_{1\texttt{class}} = 2 * \frac{P_{\texttt{class}} * R_{\texttt{class}}}{P_{\texttt{class}} + R_{\texttt{class}}}$$

Atributuen sailkapenaren ebaluazioa egiteko formuletan ongi identifikatu diren entitateak (gertaerak edo adierazpenak) hartzen dira kontuan ($\{\forall x | x \in (Irag \cap Anot)\}$). Horietan sistemak iragarritako atributuaren balioa eskuz markatutakoa bezalakoa den edo ez egiaztatzen da ($Irag_{\texttt{class}}(x) == Anot_{\texttt{class}}(x)$).

Adibidea

$$P_{\texttt{class}} = \frac{\#\{Irag_{\texttt{class}}(\texttt{e}_2) == Anot_{\texttt{class}}(\texttt{e}_2)\}}{\#\{\texttt{e}_1, \texttt{e}_2, \texttt{e}_3\}} = \frac{1}{3} = 0.3$$

$$R_{\texttt{class}} = \frac{\#\{Irag_{\texttt{class}}(\texttt{e}_2) == Anot_{\texttt{class}}(\texttt{e}_2)\}}{\#\{\texttt{e}_1, \texttt{e}_2, \texttt{e}_3\}} = \frac{1}{3} = 0.3$$

$$F_{1\texttt{class}} = 2 * \frac{0.3 * 0.3}{0.3 + 0.3} = 0.3$$

Adibidean ikus daitekeen moduan, ebaluazioan 0.3 puntu lortzen dira neurri guztietan hiru gertaeretatik `class` atributuaren balio egokia bakar batek jasotzen duelako (eta hiru gertaerak ongi identifikatu direla pentsatu dugulako).

- Denborazko loturen identifikazioa eta kategorizazioa: *TAS* izeneko metrikak erlazio tenporalen etiketatzearen ebaluazioa egiten du, bertan parte hartzen duten entitateen identifikazioa eta erlazioen kategorizazioa zuzena den edo ez kontuan edukita. Gainera, esan dugu metrika honek doitasuna, estaldura eta F_1 neurria itzultzen dituela. *TAS* metrikaren inplementazioaren inguruko zehaztasunak eta hura kalkulatzeko era (UzZaman eta Allen, 2011) argitalpenean daude bilduta.

3.3 *bTime*: euskararako denbora etiketatzailea

Atal honetan euskaraz idatzitako testuetako denbora informazioa etiketatzen duen lehenengo *end-to-end* sistema aurkezten dugu: *bTime*. Etiketatzaile honek *ISO-TimeML* eskemaren euskarako egokitzapena jarraitzen du. Sistemaren garapenerako *Euskal-TimeBank* corpusa erabili dugu, eta etiketatzailearen emaitzak testuinguru estandarrean ebaluatu dira. Bestalde, ebaluaziorako *TempEval-3* saioko *scorer*a exekutatu dugu. Gainera, ebaluazio honen emaitzak ingeleserako, gaztelerarako eta italierarako aurkeztutako balioekin alderatuko ditugu.

3.3.1 Informazioaren adierazpidea

Euskal-TimeBank corpuseko fitxategien formatuari dagokionez, testuak eta hauen eskuzko anotazioak *ISO-TimeML* eskemak ezartzen duenaren arabera daude adierazita. 3.6 irudian ikus daiteke honen adibidea. Bertan *Superjumboa gaur entregatu zaio Iberiari* esaldiari dagokion *ISO-TimeML* formatuko fitxategia erakusten da.

```
<Document doc_name="A380_delivered.txt">
    <token t_id="1">Superjumboa</token>
    <token t_id="2">gaur</token>
    <token t_id="3">entregatu</token>
    <token t_id="4">zaio</token>
    <token t_id="5">Iberiari</token>
    <token t_id="6">.</token>
    <Markables>
      <EVENT m_id="1"><token_anchor t_id="3"/></EVENT>
      <TIMEX3 m_id="2" funcInDoc="DCT" value="2007-1-31"/>
      <TIMEX3 m_id="3" funcInDoc="" value="2007-1-31">
        <token_anchor t_id="2"/>
      </TIMEX3>
    </Markables>
    <Relations>
      <TLINK r_id="1" relType="BEFORE">
        <source m_id="1" /><target m_id="2" />
      </TLINK>
      <TLINK r_id="2" relType="IS_INCLUDED">
        <source m_id="1" /><target m_id="3" />
      </TLINK>
    </Relations>
</Document>
```

Irudia 3.6: *ISO-TimeML* formatuaren adibidea.

Fitxategian ikusten denez, lehenik, esaldia osatzen duten tokenak (tokenizazioa) biltzen dira. Gero, `<Markables>` etiketaren barnean, gertaerak eta denbora adierazpenak (*DCT*a ere bai). Adibideko esaldian gertaera baten buru lexikala (`m_id="1"`, *entregatu*) eta denbora adierazpen bat (`m_id="3"`, *gaur*) anotatu dira. Gainera, testuaren *DCT*a, `2007-1-31` da eta hau `m_id="2"`bezala erazagututa dago fitxategian. Denbora adierazpenen `value` atributuak haien forma normalizatua biltzen du. Normalizazioa egiteko *DCT*a erabiltzen da. Segidan erlazio tenporalak biltzen dituen `<Relations>` etiketa dago. Adibidean bi erlazio anotatu dira: *entregatu* gertaeraren eta *DCT*aren artekoa (`r_id="1"`) eta gertaeraren eta *gaur* adierazpenaren artekoa (`r_id="2"`). Loturetako `relType` atributuak erlazio bakoitza osatzen duten elementuen arteko ordena tenporala finkatzen du. Informazio hau guztia `<Document>` izeneko etiketen barnean biltzen da. Irakurterraztasuna dela medio, irudian, gertaeren, denbora adierazpenen eta erlazio tenporalen atributuetako batzuk ez dira ageri.

Corpusen alderaketa kuantitatiboa

Orain, *bTime*n erabili dugun *Euskal-TimeBank* corpusa *TempEval-3* eta *EVENTI-2014* saioetako corpusekin (ingelesekoarekin, gaztelerakoarekin eta italierakoarekin) alderatzen dugu. Bi saio hauetan zentratuko gara batez ere hauek direlako *end-to-end* sistemak garatzera iritsi ziren bi saioak. Alderaketa hau, gainera, baliagarria da *bTime*kin euskararako lortutako emaitzak beste hizkuntzetarako lortutakoen aldean non kokatzen diren hobeki ulertzeko. 3.10n bildu ditugu corpusen tamainak adierazten dituzten hainbat datu.

	Train				**Test**			
#	**Euskara**	**Ingelesa**	**Gaztelera**	**Italiera**	**Euskara**	**Ingelesa**	**Gaztelera**	**Italiera**
Dokumentuak	20	2,708	175	373	10	21	35	92
Tokenak	5,749	761,700	57,977	103,593	3,884	6,375	9,833	26,686
Gertaerak	1,133	96,193	10,449	23,964	760	746	-	3,798
Adierazpenak	343	17,269	1,269	2,906	208	158	-	482
TLINK [Gertaera/DCT]	552	30,033	10,545	-	388	183	-	-

Taula 3.10: *Euskal-TimeBank* corpusa (euskara), *TempEval-3* eta *EVENTI-2014* ebaluazio saioetako corpusak (ingelesa, gaztelera eta italiera).

Taulan ikus daitekeen moduan, lau corpusetatik txikiena *Euskal-TimeBank* da. Izan ere, token kopurutan, ingeleseko, italierako eta gaztelerako entrenamendurako zatiak (*train*) euskarakoa ahalako 132, 18 eta 10 baitira, hurrenez hurren. Gertaera kopuru-

tan, euskarakoa ahalako 85, 21 eta 9 dira, eta adierazpen kopurutan, berriz, euskarakoa ahalako 50, 8 eta 4. Gertaeren eta *DCT*en artean dauden erlazio tenporalei dagokienez, ingeleseko eta gaztelerako entrenamendurako zatietan anotatuta daudenak euskarakoan anotatutakoak ahalako 54 eta 19 dira.

Argitu beharra daukagu beste erlazio tenporal motak, gertaeren eta denbora adierazpenen artekoak, *TempEval-3* eta *EVENTI-2014* saioetako corpusetan etiketaturik daudela, baina ez ditugula hemen adierazi. Izan ere, *Euskal-TimeBank*en etiketatuta dauden mota honetako erlazio kopuruak txikiegiak dira. *Euskal-TimeBank* corpusaren tamaina mugatuak eragina izan du *bTime* sistemaren diseinuan eta honen funtzionalitateetan.

3.3.2 *bTime* etiketatzailearen garapena

bTime tresnaren diseinua eta garapen prozesua pausoka azalduko ditugu. Horretarako arkitekturako hiru urratsak banan-banan aztertuko ditugu: gertaeren eta denbora adierazpenen etiketatzea eta erlazioen identifikazioa eta kategorizazioa. Gainera, *bTime*n sailkatzaileak eraikitzeko inplementatutako ezaugarriak ere zerrendatuko ditugu.

Prozesua modu guztiz automatikoan egiten da testu soiletik abiatuta. Lehenengo urratsean gertaerak identifikatu eta hauek euskaraz jasotzen dituzten zortzi atributuak etiketatzen dira; bigarrenean denbora adierazpenak identifikatu eta hauen `type` eta `value` atributuak etiketatzen dira. Azkeneko urratsean, berriz, gertaeren eta sarrerako testuen *DCT*en arteko erlazioak ezarri eta kategorizatzen dira (`relType`).

Sarrerako dokumentuetan aipatutako hiru urratsak egin ahal izateko beharrezkoa da lehenbizi dokumentu horiek aurreprozesatzea. Aurreprozesamenduan euskararako sortuta dauden hainbat tresna erabiltzen dira. Tokenizaziorako, lematizaziorako eta *Part-of-Speech* kategoriak etiketatzeko *Eustagger* tresna (Alegria et al., 2002) erabiltzen da. Dokumentuen analisi sintaktiko eta semantikorako, berriz, tesi lan honetan garatu dugun euskararako *bRol* dependentzia *parser*a erabiltzen dugu (Salaberri et al., 2015a).

Ezaugarri linguistikoak

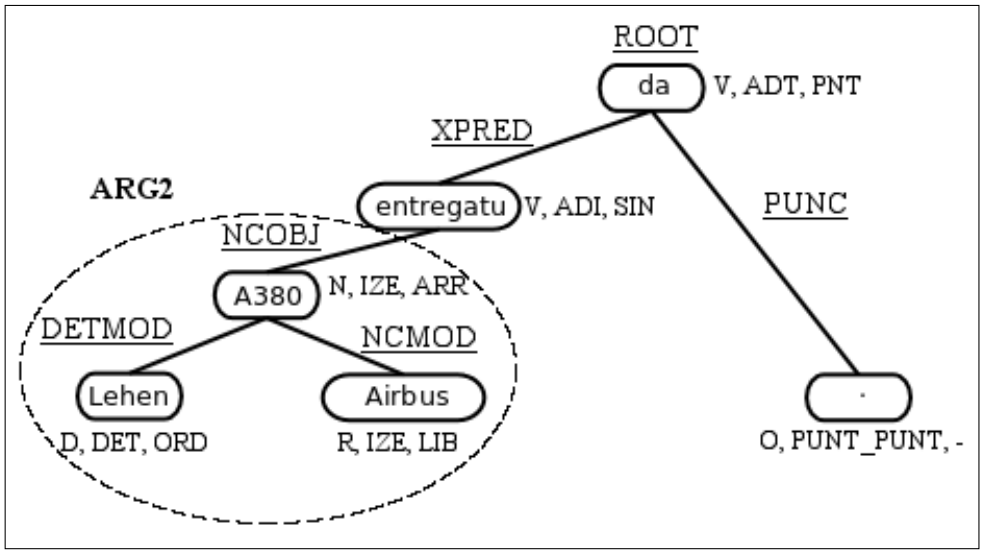
*bTime*k ikasketa automatikoko eta erregeletan oinarritutako metodoak erabiltzen ditu. 3.11 taulan etiketatzailea eraikitzeko erabili ditugun ezaugarriak bildu ditugu. Denetara 24 ezaugarri inplementatu dira. Ezaugarri hauekin, motaren arabera, lau talde hauek egin ditugu: lexikalak, *Part-of-Speech*, sintaktikoak eta semantikoak. Hauetako batzuk

beste hizkuntzetan denbora etiketatzen duten *end-to-end* sistemak garatzeko erabili izan dira (Jung eta Stent, 2013). Beste batzuk, ordea, *Part-of-Speech* azpikategoria adibidez, lehenengo aldiz inplementatu dira ataza honetan.

Ezaugarri mota	Ezaugarriak
Lexikalak	Forma eta lema hauentzat: uneko tokenarentzat (1/2), guraso tokenarentzat (13/14) eta uneko tokenaren aditz gobernatzailearentzat (19/20).
PoS	*PoS* kategoria *estandarra*, euskarazko *PoS* kategoria eta azpikategoria hauentzat: uneko tokenarentzat (3/4/5), guraso tokenarentzat (15/17/18) eta uneko tokenaren aditz gobernatzailearentzat (21/23/24).
Sintaktikoak	Uneko tokenetik esaldiaren erro sintaktikorainoko tokenei dagozkien hurrengo ezaugarriek osatutako multzoak: lemek (7), dependentzia sintaktiko motek (6), *PoS* kategoria *estandarrek* (8), euskarazko *PoS* kategoriek (9) eta azpikategoriek (10). Elkarren arteko dependentzia sintaktiko mota hauentzat: uneko tokenaren gurasoaren eta zuhaitz sintaktikoan honen gurasoa den tokenaren artean (16), eta uneko tokenaren aditz gobernatzailea eta zuhaitz sintaktikoan honen gurasoa den tokenaren artean (22).
Semantikoak	Uneko tokenaren rol semantikoa (12) eta uneko tokenetik esaldiaren erro semantikorainoko tokenei dagozkien rol semantikoek osatutako multzoa (11).

Taula 3.11: *bTime* etiketatzailearen garapenean erabilitako ezaugarrien zerrenda.

Ezaugarri hauek nolakoak diren ulertzeko 3.7 eta 3.8 irudiak aurkezten ditugu. Lehenbizikoan, *Lehen Airbus A380a entregatu da* esaldiarentzat, aurreprozesamenduan egindako etiketatzetik, esaldiaren token bakoitzarentzat erdietsitako informazioa adierazi dugu. Etiketa horiek dira goiko taulan zerrendatu ditugun ezaugarriak adierazteko erabili direnak (ikus 3.8 irudia). 3.7 irudian hiru etiketa ageri dira: lehenbizikoa *Part-of-Speech* kategoria estandarrari dagokio, bigarrena euskarazko kategoriari eta hirugarrena azpikategoriari. Tokenen arteko dependentzia sintaktikoak azpimarratu egin dira.



Irudia 3.7: Ezaugarriak azaltzeko egindako irudikapena.

Jarraian zerrendatzen ditugu *Airbus* tokenaren 3.11 taulako 24 ezaugarriek hartzen dituzten balioak.

1. Airbus	12. -	23. ADI
2. Airbus	13. A380a	24. SIN
3. R	14. a380	
4. IZE	15. N	
5. IB	16. ncobj	
6. ncmod|ncobj|xpred|Root	17. IZE	
7. Airbus|a380|entregatu|izan	18. ARR	
8. R|N|V|V	19. entregatu	
9. IZE|IZE|ADI|ADT	20. entregatu	
10. LIB|ARR|SIN|PNT	21. V	
11. -|-|ARG2|-	22. xpred	

Irudia 3.8: *Airbus* tokenari dagozkion ezaugarriak.

Gertaeren etiketatzea

Gertaerak identifikatzeko garaian *bTime*k buru lexikalak besterik ez ditu hartzen kontuan, ez gertaerak osatzen dituzten token guztiak (*Euskal-TimeBank* corpusean horrela daudelako etiketatuta). Buru lexikalak token bakarrekoak izaten direnez gero, gertaerak identifikatzeko *Support Vector Machines-SVM* algoritmoarekin sortutako sailkatzaile bitarra erabiltzen da. Honek sarreratzat jasotako testuko token bakoitza gertaera baten buru lexikala den edo ez erabakitzen du.

Euskaraz gertaerek hartzen dituzten zortzi atributuen etiketatzailea dela eta, hauetan ere ikasketa automatikoa, eta beraz, sailkatzaileak, erabili dira. Bi balio bakarrik har ditzaketen atributuentzat (ikus 3.1.3), identifikaziorako bezala, sailkatzaile bitarrak erabili dira. Gainerakoentzat, berriz, *multiclass* motakoak. Sailkatzaileen eraikuntzarako baliatutako algoritmoa *SVM* izan da hemen ere, hizkuntzaren prozesamenduko ataza askori algoritmo hau ongi egokitzen zaielako jakina delako, tesi lan honen lehenbiziko atalean landutako rol semantikoen etiketatzeari, esate baterako. Sailkatzaileak garatzeko 3.11 taulan aurkeztutako ezaugarriak erabiltzen dira.

Denbora adierazpenen etiketatzea

bTime sisteman denbora adierazpenen identifikazioa (`id`) eta kategorizazioa (`type`) egiteko erabil daitezkeen bi hurbilpenak inplementatu ditugu: ikasketa automatikoa eta heuristikoak. Azken sistemarako bietatik emaitzarik onenak itzuli dituena erabili dugu, heuristikoak. Adierazpenen normalizazioa egiteko (`value`), aldiz, heuristikoak baizik ez ditugu baliatu, dagoen kasuistika zabalaren eraginez.

Ikasketa automatikoko hurbilpenean identifikazioa `BIO/IOB` motako sailkatzailea erabilita egin dugu. Mota honetako sailkatzaileetan inkoherentziak sor daitezke, esate baterako, `B-O-I` moduko tokenen etiketatze-sekuentziak sor daitezke. Kasu hauek *zuzentzeko* post-prozesuko teknikak erabiltzen dira. Sailkatzaileak eraikitzeko, gertaerekin bezala, *SVM* algoritmoa aplikatu dugu.

Heuristikoen bitartezko hurbilpenari dagokionez, aldiz, *HeidelTime* (Strötgen eta Gertz, 2010) tresnarako *IXA* taldearen barnean garatu diren erregelak aplikatu ditugu. Euskarara egokitzeko, gainera (sarrerako eta irteerako formatuak, tokenizazioa, etab.), *HeidelTime*n jatorrizko inplementaziori zenbait aldaketa egin behar izan dizkiogu. Kopuruen karietara, normalizaziorako 27 fitxategi (`normalization`), adierazpen erregula-

rrak biltzen dituzten 50 fitxategi (`repattern`) eta 334 erregela (`rules`) baliatzen dira. *HeidelTime*k euskarazko testuetan denbora adierazpenak etiketatu ahal izateko behar dituen fitxategiak hiru direktoriotan zehar banatzen ditu:

- `normalization`: Normalizaziorako fitxategiak biltzen ditu. Hauetan denbora adierazpenak osatzen dituzten tokenen balio normalizatuak zein diren zerrendatzen da.
- `repattern`: Denbora adierazpenak identifikatzeko erregeletan erabiltzen diren adierazpen erregularrak dauzkaten fitxategiak biltzen ditu.
- `rules`: Denbora adierazpenak identifikatzeko eta normalizatzeko erregelak dituzten fitxategiak biltzen ditu.

Jarraian erakusten dugun adibidea oinarri hartuta, *HeidelTime*k daukan funtzionamendua, eta nola erabiltzen dituen hiru direktorio hauetako fitxategiak adierazpenen identifikaziorako eta normalizaziorako azalduko dugu. Adibidean etiketatuko den esaldia ondorengoa izango da: *Atzo eman zitzaien hasiera aurtengo Neguko Olinpiar Jokoei.*

Beharrezkoa da *HeidelTime*eri, testuak prozesatu baino lehen, haien sorrera datak zehaztea (*DCT*), izan ere, testu bateko adierazpen tenporalen normalizazio prozesua testuaren *DCT*an oinarritzen baita. Adibideko esaldiaren sorrera data `2016-11-19` da. Uste dugu aurrera jarraitu aurretik garrantzizkoa dela ondorengoa argitzea: testu bat *ISO-TimeML* eskemaren arabera etiketatzen denean, honen *DCT*a testuaren barneko denbora adierazpentzat hartzen dela (`<TIMEX3>`). Beraz, testuaren prozesamenduaren eraginez sortzen den etiketatutako fitxategian denbora adierazpen gisa aurki daiteke *DCT*a. Denbora adierazpen hori besteetatik bereizteko, `functionInDocument` izeneko atributuari `Document_Creation_Time` balioa esleitzen zaio.

*HeidelTime*k adibideko esaldia prozesatzeko era ordenatuan aplikatuko ditu `rules` direktorioan bildutako erregelak (orokorrenak azkenak). Hauetako batzuk aktibatu egingo dira, mota bateko edo besteko adierazpenak detektatzen direnean. 3.9 irudian bildu ditugu adibideko esaldirako aktibatzen diren hirurak.

`R1` gisa erazagutu dugun erregelak esaldiko *Aurtengo* denbora adierazpena detektatzen du. `R2` eta `R3` gisa ezagutarazi ditugun erregelek, aldiz, *Neguko* eta *Atzo* adierazpenak detektatzen dituzte, hurrenez hurren. 3.9 irudian ikusten den bezala, erregela bakoitzak bi osagai nagusi ditu: `EXTRACTION` eta `NORM_VALUE`. Lehenbiziko osagaian erre-

```
(R1) EXTRACTION="%reHauHurrengoAzken urte",
     NORM_VALUE="UNDEF-%normHauHurrengoAzken(group(1))-urte"

(R2) EXTRACTION="%reUrtaro",
     NORM_VALUE="UNDEF-urte-%normUrtaro(group(2))"

(R3) EXTRACTION="%reDataHitza",
     NORM_VALUE="%normDataHitza(group(1))"
```

Irudia 3.9: Adibideko esaldia etiketatzeko aktibatzen diren erregelak (`rule`).

gelak detektatu beharreko denbora adierazpena deskribatzen da, horretarako adierazpen erregularrak erabilita. Adierazpen erregular hauek `repattern` direktorioan bildutakoak dira. Bigarren osagaian, berriz, erregelak identifikatzen dituen denbora adierazpenak nola normalizatu behar diren zehazten da. Normalizaziorako `normalization` direktorioko fitxategiak hartzen dira. Erregela hauetan erabilitako adierazpen erregularrak eta normalizaziorako fitxategien edukiak 3.10 eta 3.11 irudietan ikus daitezke hurrenez hurren.

%reHauHurrengoAzken	%reDataHitza	%reUrtaro
[Aa]urtengo	[Bb]ihar	[Uu]daberriko
[Aa]zkeneko	[Aa]tzo	[Nn]eguko
[Hh]urrengo	[Gg]aur	[Uu]dako
...	...	...
[Ll]ehenengo	[Oo]rain	[Uu]dazkeneko

Irudia 3.10: Esaldia etiketatzeko erabiltzen diren adierazpen erregularrak (`repattern`).

%normHauHurrengoAzken	%normDataHitza	%normUrtaro
"Aurtengo", "this"	"Bihar", "UNDEF-next-day"	"Udaberriko", "SP"
"aurtengo", "this"	"bihar", "UNDEF-next-day"	"udaberriko", "SP"
"Azkeneko", "last"	"Atzo", "UNDEF-last-day"	"Neguko", "WI"
"azkeneko", "last"	"atzo", "UNDEF-last-day"	"neguko", "WI"
"Hurrengo", "next"	"Gaur", "UNDEF-present-day"	"Udako", "SU"
"hurrengo", "next"	"gaur", "UNDEF-present-day"	"udako", "SU"
...	...	...
"Lehenengo", "first"	"Orain", "PRESENT_REF"	"Udazkeneko", "FA"
"lehenengo", "first"	"orain", "PRESENT_REF"	"udazkeneko", "FA"

Irudia 3.11: Esaldia etiketatzeko normalizaziorako fitxategiak (`normalization`).

3.10 irudian ikus daitekeenez, R1 erregelarentzat EXTRACTION osagaian finkatzen den adierazpen erregularren fitxategiak (%reHauHurrengoAzken) [Aa]urtengo dauka barnean. Hori dela-eta aktibatu da *HeidelTime*ko R1 erregela esaldiko *aurtengo* tokena prozesatzean. Normalizazioari dagokionez, %normHauHurrengoAzken izeneko fitxategia erabiltzen du R1 erregelak. 3.11 irudian ikus daiteke fitxategi honen arabera *aurtengo*, this moduan normalizatzen dela. *HeidelTime* arduratzen da ondoren honetatik, testuaren *DCT*a aintzat harturik, *aurtengo* denbora adierazpenaren balio normalizatu osoa lortzeaz (kasu honetan value=2016). *Neguko* eta *Atzo* adierazpenentzat prozesua bera da, baina R2 eta R3 erregeletan zehazten diren fitxategiak erabilita. Honako hau izango litzateke *HeidelTime*ko irteera, adibideko esaldiarentzat:

<TIMEX3> <TIMEX3> <TIMEX3>
Atzo$_{(3)}$ *eman zitzaien hasiera* ***aurtengo***$_{(1)}$ ***Neguko***$_{(2)}$ *Olinpiar Jokoei.*

↕

```
<TIMEX3 m_id="1" type="DATE" value="2016">
<TIMEX3 m_id="2" type="DATE" value="2016-WI">
<TIMEX3 m_id="3" type="DATE" value="2016-11-18">
```

Denbora erlazioen identifikazioa eta kategorizazioa

Euskal-TimeBank corpusean etiketatutako denbora erlazio mota gehienen kopuruak txikiegiak dira sistema automatiko bat eraikitzeko. Ondorioz, *bTime* sarrerako dokumentuen *DCT* denbora adierazpenen eta gertaeren arteko erlazio tenporalen identifikaziora eta kategorizaziora (relType) mugaturik dago.

Erabiltzen dugun teknika 3.2.1 atalean deskribatutakoa da. Hau da, aipatu erlazioen identifikazioa eta kategorizazioa urrats bakar batean egiten da. *multiclass* motako sailkatzaile batek sistemaren lehenbiziko urratsean etiketatutako gertaerak eta sarrerako dokumentuen *DCT*ak hartuta, sor daitezkeen erlazio tenporal posible guztiak sailkatzen ditu. 3.3 irudiko kategoriak (relType) ez ezik, NO_RELATION (loturarik ez) izeneko kategoria ere esleitzen du sailkatzaile honek. *bTime*ek urrats honetan erabiltzen duen sailkatzailea ere *SVM* algoritmoaren bidez sortu dugu.

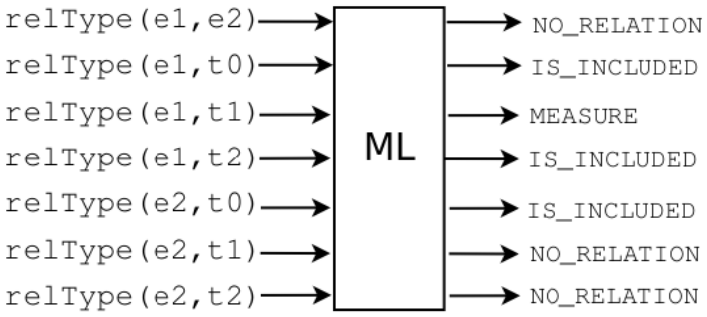
Erabili ditugun ezaugarriak (Bethard, 2013) argitalpenean oinarrituta daude eta honako hauek dira: `class`, `tense1`, `tense2`, `aspect1`, `aspect2`, `polarity`, `modality`, `pos` eta gertaera bera (testua). Hurrengo adibidean ikus daiteke erlazioen identifikazio eta kategorizazio prozesua nolakoa den. Adibidearen *DCT*a `2016-3-5` da eta, lehenago aipatu dugun gisan, denbora adierazpentzat baliatzen da. Kasu honetan `t0` gakoaren bitartez ezagutarazi dugu *DCT*a. Guztira hiru adierazpen (`t0`, `t1`, `t2`) eta bi gertaera (`e1`, `e2`) daude etiketatuta. Sistemak bikote posible guztiak hartzen ditu sarreratako eta horietako bakoitzak zein erlazio mota duen ondorioztatuko du.

Jonek 20 minutu irakurtzen zuen astelehenetan. Futbola gustatzen zaio.

↓

Jonek <TIMEX3> ***20 minutu***$_{(t1)}$ <EVENT> ***irakurtzen***$_{(e1)}$ *zuen* <TIMEX3> ***astelehenetan***$_{(t2)}$. *Futbola* <EVENT> ***gustatzen***$_{(e2)}$ *zaio.*

↓



↓

```
<TLINK event="e1" time="t0" relType="IS_INCLUDED"/>
  <TLINK event="e1" time="t1" relType="MEASURE"/>
<TLINK event="e1" time="t2" relType="IS_INCLUDED"/>
<TLINK event="e2" time="t0" relType="IS_INCLUDED"/>
```

Adibideko testuan lau denbora erlazio detektatu eta kategorizatu direla ikus daiteke. Hauetako bi gertaeren eta *DCT*aren artekoak dira (`e1-t0` eta `e2-t0`) eta beste biak, aldiz, gertaera baten eta denbora adierazpenen artekoak (`e1-t1` eta `e1-t2`). Erlazio motei dagokienez, hiru `IS_INCLUDED` kategoriakoak dira (`e1-t0`, `e1-t2` eta `e2-t0`) eta bat `MEASURE` kategoriakoa (`e1-t1`).

3.3.3 Esperimentazioa

Sistema eraginkorrena eta onena lortzeko asmotan *bTime* tresnaren barnean, gertaeren, denbora adierazpenen eta erlazio tenporalen etiketatzeari dagozkien esperimentuak egin ditugu hainbat neurritako testuinguru leihoekin (bat, hiru, zazpi eta hamabost hitzetakoak). Honekin adierazi nahi dugu *bTime*k prozesatzen duen token bakoitzarentzat ezaugarriak tokenarentzat berarentzat erauzi ditugula, baita haren eskuinetan eta ezkerretan dauden beste zero, bat, hiru eta zazpi tokenentzat ere. Honek badu eragina prozesatutako token bakoitzarentzat erauzitako ezaugarri kopuruan. Hitz bakarreko leihoentzat 24 ezaugarri erauzi dira (tokenarenak berarenak), hirukoentzat 72, zazpikoentzat 168 eta, azkenik, hamabostekoentzat, 360 ezaugarri.

Esperimentuak, testuinguru leihoekin ez ezik, ezaugarri multzoekin ere egin ditugu. Horrela, azken hauek euskarazko denboraren etiketatze automatikoan daukaten eragina neurtu ahal izan dugu. Besteak beste, esperimentu hau egitea tesi lan honen hipotesietako bat betetzen den edo ez ikusteko ere baliagarria izan zaigu: euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek duten eragina positiboa dela, ingelesez eta gaztelaniaz bezala. Ezaugarrien eraginaren inguruko ikerketa hau burutu ahal izateko 3.11 taulan aurkeztutako ezaugarriekin ondoko zazpi multzoak egin ditugu. Bertan azaltzen dugu multzo hauetako bakoitzarekin aztertu nahi genuena:

- `ALL`: 1-24, ezaugarri guztiak erabiltzen dira.
- `FLP`: 1, 2, 3, 4 eta 5. Une bakoitzean *bTime*k prozesatzen duen tokenak eskainitako informazioa erabiltzeak daukan eragina aztertzeko. Inguruko tokenek eta hauekiko erlazio sintaktiko-semantikoek eskaintzen duten informazioa kontuan izan gabe.
- `SIN`: 6, 7, 8, 9 eta 10. Une bakoitzean *bTime*k prozesatzen duen tokenetik esaldiaren erro sintaktikorainoko tokenen informazioa erabiltzeak duen eragina aztertzeko. Honek esaldiaren egituraketa sintaktikoaren garrantzia erakutsiko du.

- SEM: 11 eta 12. Rol semantikoen eragina neurtzeko. Tesiaren lehenbiziko hipotesia betetzen den edo ez egiaztatzeko balio du.
- PRE: 13, 14, 15, 16, 17 eta 18. Une bakoitzean *bTime*k prozesatzen duen tokenaren gurasoak eskaintzen duen informazioa erabiltzeak daukan eragina aztertzeko.
- GOV: 19, 20, 21, 22, 23 eta 24. Une bakoitzean *bTime*k prozesatzen duen tokenaren aditz gobernatzaileak eskaintzen duen informazioa erabiltzeak daukan eragina aztertzeko.
- BPO: 4, 5, 9, 10, 17, 18, 23, 24. Une bakoitzean *bTime*k prozesatzen duen tokenaren *Part-of-Speech* kategoriak eskaintzen duen informazioa erabiltzeak daukan eragina aztertzeko.

Egin dugun ezaugarrien multzokatzeak informazio sintaktikoak eta semantikoak euskarazko denboraren etiketatzean duen eragina positiboa den edo ez ikusteko balio izan digu (SIN eta SEM), besteak beste. Izan ere, Jung eta Stenten (2013) arabera, mota honetako ezaugarriek ingelesezko etiketatze tenporalaren emaitzak hobetzen dituzte. Multzokatze honetan, gainera, uneko tokenaren aditz gobernatzailearekin, aurreko tokenarekin eta euskarazko analizatzaile morfologikoarekin zerikusia duten ezaugarrien (GOV, PRE, BPO eta FLP) eragina ere aztertu dugu.

Multzo hauen eragina neurtu ahal izateko *Leave-One-Out* (LOO) teknika baliatu dugu. Prozedura honen bitartez ezaugarriak banan-banan ateratzen dira, guztiak biltzen dituen multzotik, eta gainerako ezaugarriekin sailkatzailea entrenatu eta ebaluatzen da, ateratakoen eragina neurtu ahal izateko. Kasu honetan, ezaugarriak banaka ateratzen joan beharrean, ezaugarri multzoak, arestian aurkeztutakoak, atera ditugu. Jarraian dagoen 3.12 taulan bildu ditugu *bTime*n emaitzak.

Taula honetan sistemaren azpiataza bakoitzean konfiguraziorik egokienarekin, hau da, baliorik altuenak itzuli dituen testuinguru leihoaren tamaina eta ezaugarri multzoarekin, lortutako emaitzak biltzen dira[10]. *Eskema* izeneko zutabeak adierazten du taulako lerro (azpiataza) bakoitzean jasotako emaitzak zer ebaluazio eskema jarraituta konputatu diren (*strict* edo *relaxed*). Taulako lehenbiziko lerroko *Eskema* zutabean, gertaeren identifikazioari dagokionean, S/R balioa ikus daiteke. Honekin adierazi nahi dugu *bTime*k egiten duen gertaeren identifikazioaren ebaluazioak bi eskemetarako emaitza berak

[10]Emaitza guztiak ikusteko: https://ehubox.ehu.eus/index.php/s/SCHA7MsTQO90lEg

Entitatea	Ataza	Eskema	Hurbilpena	Konfiguraziorik onena		Doitasuna	Estaldura	F_1
<EVENT>	id	S/R		1	ALL - BPO	85.1	70.66	77.21
	class	-		3	ALL - BPO	-	-	59.74
	tense1	-		3	ALL - PRE	-	-	64.15
	tense2	-		3	ALL - PRE	-	-	65.02
	aspect1	-	ML	3	ALL	-	-	64.25
	aspect2	-		3	ALL - BPO	-	-	65.12
	polarity	-		3	ALL - GOV	-	-	74.86
	pos	-		3	ALL - GOV	-	-	70.09
	modality	-		3	ALL - BPO	-	-	73.98
<TIMEX3>	id	S	ML	1	ALL	60.56	36.13	45.26
	id	S	HEUR	-	-	76.15	69.75	72.81
	id	R	ML	1	ALL - PRE	84.81	56.3	67.68
	id	R	HEUR	-	-	87.16	79.83	83.33
	type	-	ML	1	ALL	-	-	48.42
	type	-	HEUR	-	-	-	-	73.68
	value	-	HEUR	-	-	-	-	54.39
<TLINK>	relType	TAS	ML	3	ALL - SIN	67.16	69.59	68.35

Taula 3.12: *bTime* etiketatzailearentzat lortutako emaitzak (S: Strict, R: Relaxed).

itzultzen dituela, esan dugun moduan gertaeren token bakarreko buru lexikalak etiketatzen direlako. *Hurbilpena* zutabeak, aldiz, emaitza hauek ikasketa automatikoaren (ML) edo heuristikoen (HEUR) bitartez lortu diren zehazten du.

3.3.4 Analisia

Azpiatal honetan *bTime* sistemarentzat lortutako emaitzak analizatuko ditugu. Hiru gauza nabarmen ikus daitezke 3.12 taulan:

1. Ikasketa automatikoaren bitartezko gertaeren eta adierazpenen identifikazioan emaitzarik onenak itzultzen dituen testuinguru leihoa token bakarrekoa dela.

2. Gertaeren atributuak etiketatzeko leihorik egokiena hiru hitzekoa dela.

3. Denbora adierazpenen etiketatzea erregelen bitartez eginda ikasketa automatikoa baliaturik baino emaitza hobeak eskuratzen direla.

Lehenbiziko puntuaren inguruan, esan beharra daukagu gure emaitzak ez datozela bat Jung eta Stentek (2013) ingeleserako aurkeztutakoekin. Izan ere, aipatutako argitalpenean, hamabost eta zazpi hitzeko testuinguru leihoekin hitz bateko eta hiru hitzeko

leihoekin baino emaitza hobeak lortu zirela azaltzen da. Kontuan izan beharra dago aipatu argitalpenean *TempEval-3* saioan parte hartu zuen `att` sistema eta hau erabilita egindako testuinguru leihoen inguruko esperimentuen emaitzak aurkezten direla.

Token bakarreko gertaeren buru lexikalak daude anotatuta. Alde horretatik, beraz, ez dago arrazoirik euskaraz bat eta hiru tokeneko testuinguru leihoek zazpikoek eta hamabostekoek baino emaitza hobeak itzultzeko. Denbora adierazpenen token kopuruari dagokionez, aldiz, ez dugu uste ingelesekoek, oro har, euskarazkoek baino token kopuru handiagoa daukatenik eta horrek ezin azal dezake lortutako emaitzak ingelesez zergatik diren hobeak. Izan ere, denbora adierazpenetako asko, gehienak, token bakarrekoak izaten dira aztertu ditugun lau hizkuntzetan.

Gure iritziz, *bTime*n kasuan, bat eta hiru hitzeko leihoekin balio altuagoak eskuratzeko arrazoi nagusia ondorengoa da: *bTime* barneko sailkatzaileak sortzeko erabili diren ezaugarrietako asko (forma, lema, uneko tokenetik esaldiaren erro sintaktikorainoko tokenen formak, etab.) *string* motakoak dira, eta *Euskal-TimeBank*en tamaina mugatuaren ondorioz ezaugarri hauetako askok hartzen dituzten balioak oso gutxitan errepikatzen dira corpusean zehar. Daitekeena da, beraz, ezaugarri horiek jasotzen dituzten balioetako batzuk corpus guztian behin baizik ez agertzea. Jakina da balio ezberdin asko edota gutxitan errepikatzen diren balioak dauzkaten ezaugarriek *zarata* sortzen dutela ikasketa prozesuan. *Zarata* honek eragin negatiboa du sailkatzaileen eraginkortasunean eta, ondorioz, baita sistemaren emaitzetan ere. Testuinguru leiho txikiak erabilita eragin negatibo hau sortzen duten ezaugarrien kopurua gutxitu egiten da.

Bigarren puntuan esan bezala, gertaeren atributuak etiketatzean emaitzarik onenak itzultzen dituzten testuinguru leihoen tamaina hiru hitzekoa da. Atributuek gertaeren izaera gramatikala adierazten dute. Izan ere, 1.1.1 azpiatalean azaldu dugunez, gertaerek denborarekin duten erlazio semantikoaren kategorizazioa aldatzen duten gramatikaren kategoriak bat baino gehiago dira (aspektu lexikala, aspektua, modua, denbora gramatikala, etab.), eta tesi lan honetan jarraitzen dugun *ISO-TimeML*n atributuen bitartez daude adierazita kategoria hauek. Hau kontuan edukita uste dugu atributuen esleipenak hiru hitzeko leihoekin emaitzarik onenak itzultzeko arrazoia gertaeren izaera gramatikala finkatu ahal izateko beharrezkoa den testuinguru linguistikoa osatzen duten token kopuruaren ondorio dela. Hau da, atributuak finkatu ahal izateko garrantzizkoa dela gertaera barnean hartzen duen sintagmari, bertako tokenei eta orokorrean hauek osatzen duten egitura sintagmatikoari, erreparatzea. Gertaera hauek askotan hiru edo lau tokenez osa-

tutako sintagmen zati izaten dira. Euskaraz aditz forma perifrastikoaren erabilera (aditza eta aditz laguntzailea) oso arrunta da, eta, gainera, aditz trinkoak ere hiru eta lau hitzeko sintagmen zati izaten dira, hurrengo adibidean bildu ditugun sintagmetan ikus daitekeen moduan. Kontuan izan beharra dago, gainera, *bTime* sistema garatzeko erabili dugun *Euskal-TimeBank* corpusean anotatutako predikatu gehienak aditzak direla.

1. *Lagunak etorri dira.*
2. *Mikel joan zen.*
3. *Liburua ez zen iritsi.*

Adibideetan ikus daiteke hiru eta lau tokeneko sintagma hauek bertako gertaeren izaera gramatikala finkatu ahal izateko beharrezkoa den informazio linguistikoa eskaintzen dutela. Esate baterako, aditz laguntzaileak adierazten du zein den denbora (`tense1` eta `tense2`). (3) sintagman, bestalde, *ez* tokenak polaritatea (`polarity`) adierazten du. Kontuan hartu beharra dago, maila lexikoan, hizkuntza baten egitura sintagmatikoa hizkuntza horren erregela sintaktikoen araberako tokenen konbinazioa dela.

*bTime*n analisiaren hirugarren puntuan esan dugun moldean, denbora adierazpenen etiketatzea, erregelen bitartez eginik, ikasketa automatikoa erabilita baino emaitza hobeak eskuratzen dira. Argi dago, kasu honetan, *Euskal-TimeBank*en tamaina eta bertan anotatutako denbora adierazpen kopurua direla bi hurbilpenekin lortutako emaitzen arteko aldearen arrazoia. Hurbilpen bien artean 27.55 puntuko ezberdintasuna dago *strict* F_1 neurrian, eta 33.62 eta 15.59 puntukoa estalduran eta doitasunean.

3.10 taulan ikusten den gisa, euskarazko corpusaren *train* zatian 343 adierazpen daude anotatuta. Ingeleserako, gaztelerarako eta italierarako *TempEval-3* eta *EVENTI-2014* saioetan erabilitako *train* corpusetan, aldiz, 17269, 2906 eta 1269 denbora adierazpen daude. Hau da, euskarakoan halako 50, 8 eta 4. *TempEval-3* saioan ingeleseko denbora adierazpenen identifikazioan emaitzarik onenak lortu zituen sistemak (`ClearTK`) ikasketa automatikoko teknikak erabili zituen. Honek 82.71 puntuko F_1 neurria lortu zuen *Strict* ebaluazio eskema jarraituta. Balio hau saio berean eta erregelak erabilita emaitzarik onena erdietsi zuen sistemaren emaitza baino 1.37 puntu altuagoa da (81.34 *Strict* F_1). Gainera, Uzzaman-ek eta bestek (2012) *TempEval-2* ebaluazio saioko denbora adierazpenen identifikazioaren inguruan diotenez, saio eta ataza horretan baliorik gorenak iritsi zituen sistemak ere (*TIPSem*) ikasketa automatikoa erabili zuen. Honek, corpusaren

tamaina egokia denean, ikasketa automatikoa denbora adierazpenen etiketatzerako hurbilpen egokia izan daitekeela adierazten du. Ondorioz, *bTime* sistemarentzat lortutako balioek agerian uzten dute *Euskal-TimeBank* ez dela denbora adierazpenen identifikazioa ikasketa automatikoko teknikak erabilirik burutzeko behar bezain handia.

Ezaugarri multzoei dagokienez, uste dugu garrantzizkoa dela azpimarratzea *bTime* sistemaren ardurakoak diren hamahiru azpiatazatatik lautan baliorik onenak `BPO` ezaugarri multzoa erabili gabe lortzen direla, hirutan `PRE` ibili gabe, bitan `GOV` multzoa baliatu gabe eta batean `SIN` gabe. Lau multzo hauek dira aurretik aipatu ditugun balio ezberdin askotako ezaugarriak gehien biltzen dituzten multzoak, edota gutxitan errepikatzen diren balioak gehien jasotzen dituztenak. Mota honetako ezaugarrien agerpena *bTime* inplementatzeko baliatu *Euskal-TimeBank* corpusaren izari mugatuaren ondorioa dela esan dugu, eta haiek, ezaugarriek alegia ikasketa prozesuari negatiboki eragiten diotela. Horregatik uste dugu hamahiru atazatatik hamarretan haiek kenduta iristen direla balio gorenak. Ikasketa automatikoko hurbilpena inplementatzen duten hamahiru azpiatazatatik hiruk baizik ez dituzte erdiesten baliorik altuenak, 3.11 taulan zehaztu ezaugarri guztiak (`ALL`) baliaturik, hurrengo hauek, hain zuzen: gertaeren `aspect1` atributuaren etiketatzeak, denbora adierazpenen *Strict* ebaluazio eskemaren araberako identifikazioak eta `type` atributuaren etiketatzeak.

Beste hizkuntzekiko alderaketa

3.13 taulan *bTime* sistemarekin euskararako lortutako emaitzak *TempEval-3* eta *EVENTI-2014* ebaluazio saioetan ingeleserako, gaztelerarako eta italierarako erdietsitako emaitzarik hoberenekin erkatzen ditugu.

			Euskara			Gaztelera (TE3)			Italiera (EV14)			Ingelesa (TE3)		
Entitatea	**Ataza**	**Eskema**	**Prez.**	**Est.**	**F1**	**Prez.**	**Est.**	**F1**	**Prez.**	**Est.**	**F1**	**Prez.**	**Est.**	**F1**
<EVENT>	id	**S**	85.1	70.66	77.21	91.7	86	88.8	-	-	86.7	81.44	80.67	81.05
	id	**R**	85.1	70.66	77.21	91.7	86	88.8	90.2	86.8	88.4	81.44	80.67	81.05
	class	**-**	-	-	59.74	-	-	57.6	-	-	67.1	-	-	71.88
<TIMEX>	id	**S**	76.15	69.75	72.81	-	-	85.3	-	-	82.7	-	-	82.71
	id	**R**	87.16	79.83	83.33	96	84.9	90.1	93.5	85.4	89.3	89.36	91.3	90.32
	type	**-**	-	-	73.68	-	-	-	-	-	77.5	-	-	-
	value	**-**	-	-	54.39	-	-	87.5	-	-	70.9	-	-	77.61
<TLINK>	relType	**TAS**	67.16	69.59	68.35	37.8	46.2	41.6	29.6	23.8	26.4	34.08	28.4	30.98

Taula 3.13: Euskararako, ingeleserako, gaztelerarako eta italierarako lortutako emaitzen alderaketa (S: Strict, R: Relaxed).

(Caselli et al., 2014) argitalpenean adierazten denez, anotazio kontuak direla-eta, ezin da alderaketa zuzenik egin *TempEval-3* eta *EVENTI-2014* saioetako hizkuntzen artean. Hortaz, euskararekiko erkaketa ere ez da zuzena izango. Hala ere, interesgarria iruditzen zaigu hizkuntza ezberdinetan *ISO-TimeML* anotazio eskema jarraituta eta ebaluaziorako metodo berak aplikatuta lortzen diren emaitzen arteko aldeak zein diren ikustea.

3.13 taulan ebaluazio saio guztietan landu diren azpiatazetako emaitzak besterik ez ditugu jaso. Bertako balioei erreparatuta ikusten da euskararako lortutakoak direla baliorik apalenak, gertaerak eta denbora adierazpenak etiketatzean (*Relaxed* F_1 77.21 eta 83.33 puntu), baina altuenak erlazio tenporalak kategorizatzeko orduan (*TAS* F_1 68.35 puntu). Hau espero izatekoa zen gertaerentzat eta erlazioentzat. Izan ere, gertaerak ikasketa automatiko bitartez etiketatzen dira eta gorago azaldu dugun gisa, hurbilpen honen eraginkortasuna sailkatzaileak sortzeko baliatu corpusaren izariaren mendekoa da. Azaldu dugu, gainera, euskararako hartutako *Euskal-TimeBank* corpusaren tamaina oso mugatua dela, beste hizkuntzetako corpusen tamainaren aldean.

Ingeleseko, gaztelerako eta italierako emaitzak erkatzen direnean ikus daiteke, gertaeren eta denbora adierazpenen identifikazioan (`id` *Strict* eta *Relaxed* F_1), hiruretan 80 puntutik gorako balioak erdiesten direla, eta emaitzak, oro har, aski antzekoak direla (7.75 puntuko aldea dago gehienez ere). Casellik eta bestek (2014) diotenez, aldiz, denbora adierazpenen normalizazioari dagokion `value` atributuarentzako balioak ez daude hain hurbil: italierakoa da baliorik apalena eta gaztelerakoa, bestalde, gorena (F_1 70.9 eta 87.5 puntu). Haien arabera, italierako emaitza baxua, ebaluatzean, *TempEval-3*n ez bezala, `TIMEX3` etiketa hutsak kontuan hartu izanaren ondorio da. Gertaeren `class` atributuaren esleipenean ere badira ezberdintasun nabarmenak; adibidez, gaztelerako emaitzatik (F_1 57.6, okerrena) ingelesekora (F_1 77.88, hoberena) 14.28 puntuko aldea dago. Azpimarratu behar da azpiataza honetan euskararako lortutako emaitza gaztelerakoa baino hobea dela (F_1 59.74 puntu).

TempEval-2 saioa deskribatu dugunean aipatu dugu Verhagenek eta bestek (2010) diotenez ez dagoela garbi saio horretan gaztelerarako gertaeren eta adierazpenen identifikazioan erdietsi emaitzak zergatik diren ingelesekoak baino hobeak, bertan erabilitako corpusak ondorio zuzenak atera ahal izateko txikiegiak direlako. *TempEval-3* saioko emaitzek ere bi hizkuntza hauen arteko aldea erakusten dute, baina ez da honen arrazoirik aipatzen, saioari dagokion argitalpenean. Gure ustez, eta italierak gaztelerarekin duen antzekotasun tipologikoa oinarri hartuta, badago arrazoi linguistikoren bat gazteleraz

eta italieraz gertaeren identifikazioa ingelesez baino *errazagoa* izateko. Taulari oharturik, ikus daiteke bi hizkuntza hauetan lortutako balioak oso antzekoak direla (88.8 eta 88.4 puntu, *Relaxed* F_1), eta ingeleseko balioa baino 7 puntu hobeak direla. Denbora adierazpenen identifikazioari dagokionez, aldiz, ematen du aldea corpusen, hauek anotatzeko moduaren edo etiketatzeko sistemen ondorio dela. Izan ere, gaztelera eta italiera tipologikoki oso antzekoak baldin badira ere, italierako emaitza ingelesekoaren oso antzekoa da (82.7 eta 82.71 puntu, *Strict* F_1), ez gaztelerakoaren antzekoa (85.3 puntu, *Strict* F_1).

bTime tresnarekin denbora adierazpenak etiketatzean lortutako emaitzak direla eta, euskarakoak beste hizkuntzetakoak baino baxuagoak izan zitezkeela pentsatzen bagenuen ere, ingeleserako, gaztelerarako eta italierarako erdietsi balioetatik hurbilago egotea espero genuela esan beharra dugu. Izan ere, 3.13 taulan denbora adierazpenen identifikazioan euskararako aurkezten dugun balioa erregeletan oinarritzen den hurbilpenetik lortutakoa da (ikus 3.12 taula) eta jakina denez, hurbilpen hau ez da, ikasketa automatikoarekikoan agitzen denaz bestera, neurri handi batean behintzat, corpusaren tamainaren mendekoa.

Denbora erlazioen kategorizazioan euskaraz lortutako emaitzei dagokienez, euskarazko balioak altuenak izateko arrazoia ondorengoa da: *bTime* sistemak sarreratzat jasotzen dituen dokumentuen `DCT`en eta hauetan identifikatutako gertaeren arteko erlazioak besterik ez ditu aztertzen. *TempEval-3* eta *EVENTI-2014* ebaluazio saioetako sistemek, aldiz, haiek ez ezik gertaeren arteko eta gertaeren eta denbora adierazpenen arteko erlazioak ere lantzen dituzte. Ebaluazio saio hauetan eta *bTime* sistemaren ebaluazioan baliaturiko metodoan erlazio tenporal mota guztiak aldi berean ebaluatzen dira, eta honetatik *TAS* balio bakarra itzultzen da. Ondorioz, ezinezkoa da jakitea zein den euskaraz gainerako hizkuntzetan `DCT`en eta gertaeren arteko erlazioak kategorizatzean iristen diren benetako *TAS* balioak.

Erroreen analisia eta sistemaren hobekuntza

bTime sistema hobetzeko asmoz erroreen analisia egin dugu. Gertaeren eta denbora adierazpenen identifikazioan zentratu gara batez ere, ez hauek jasotzen dituzten atributuen etiketatzean. Erroreen analisi honetan ez dugu detektatu *bTime* barnean denbora adierazpenen identifikazioa egiten duten erregelen hobekuntzarako interesgarria izan daitekeen faktorerik. Hala ere, gai izan gara gertaeren identifikazioan lagungarria izan den faktore bat aurkitzeko: *bTime* etiketatzaileak identifikatzea erdiesten ez duen ger-

taeretako asko adjektiboetatiko izenak, hau da, adjektiboen nominalizazioak diren izen predikatuen bitartez deskribatutako gertaerak dira.

Hau jakinda, gure sistemari adjektiboen nominalizazioak diren 2.728 izen predikatuk osatutako zerrenda bat gehitu diogu[11]. Zerrenda hau osatu ahal izateko *IXA* taldearen barnean garatzen ari den euskarazko *NomBank* baliabidetik hornitu gara. Hain zuzen ere, *-(t)asun* eta *-(k)eria* atzizki-amaierak baliatuta EDBL datu base lexikaletik erauzi diren izenak hartu ditugu. Kontuan hartu behar izan dugu, gainera, izen hauetako batzuk adjektiboen nominalizazioak baldin badira ere, ez zaizkiola *ISO-TimeML* eskemak zehazten duen gertaeraren definiziora egokitzen eta beraz horrelakoak kendu behar izan ditugu aipatutako zerrenda sortu ahal izateko.

3.14 taulan bildu ditugu *bTime* sistemari gertaeren identifikazioa hobetu ahal izateko zerrenda gehitu ondoren egindako *train-test* ebaluaziotik iritsitako emaitzak[12]. Sistemaren aldaketak, zerrenda gehitzeak, ez dio denbora adierazpenen etiketatzeari eragiten. 3.14 taulan, *bTime* sistemaren aldaketa egin aurretik lortutako emaitza ageri da ezkerrean eta aldaketa egin ondorengoa eskuinean.

Entitatea	Ataza	Eskema	Konfiguraziorik onena		Doitasuna	Estaldura	F_1
<EVENT>	id	S/R	1/1	ALL - BPO/ALL - BPO	85.1/85.3	70.66/74.08	77.21/79.30
	class	-	3/1	ALL - BPO/ALL - BPO	-	-	59.74/60.00
	tense1	-	3/3	ALL - PRE/ALL - PRE	-	-	64.15/63.47
	tense2	-	3/3	ALL - PRE/ALL - PRE	-	-	65.02/64.33
	aspect1	-	3/3	ALL/ALL	-	-	64.25/63.53
	aspect2	-	3/3	ALL - BPO/ALL - BPO	-	-	65.12/64.38
	polarity	-	3/1	ALL - GOV/ALL - BPO	-	-	74.86/76.90
	pos	-	3/1	ALL - GOV/ALL - GOV	-	-	70.09/71.19
	modality	-	3/1	ALL - BPO/ALL - BPO	-	-	73.98/75.35
<TLINK>	relType	TAS	3/3	ALL - SIN/ALL - SIN	67.16/67.08	69.59/69.85	68.35/68.43

Taula 3.14: *bTime* etiketatzailearentzat lortutako emaitzak, sistemaren hobekuntzaren ondoren (S: Strict, R: Relaxed).

Taulako balioek adierazten dutenez, gertaeren identifikazioari dagokion F_1 neurria 2.09 puntu hobetzen da, 77.21 puntutatik 79.3 puntutara igarotzen da. Hobekuntza hau estaldurari zor zaio neurri handi batean, estaldura ia lau puntu hobetzen baita, 70.66 puntutatik 74.08 puntutara. Doitasuna berriz 0.2 bakarrik hobetzen da zerrendarekin. Gertaeren atributuak direla eta, ikusten da zortzi atributututatik lauk emaitza hobea itzul-

[11] Zerrenda eskuratzeko: https://ehubox.ehu.eus/index.php/s/LrvSOMMpstcO8BD

[12] Emaitza guztiak ikusteko: https://ehubox.ehu.eus/index.php/s/cCp8OZFZJq0xPc1

tzen dutela (`class`, `polarity`, `pos`, `modality`), eta beste lauk, berriz, okerragoa (`tense1`, `tense2`, `aspect1`, `aspect2`).

Erlazio tenporalen emaitzari dagokionez esan beharra dago sistemari zerrenda gehitzearekin hobetu egiten dela: *TAS* F_1 neurria 68.35 puntutatik 68.43 puntutara igotzen da. Kasu honetan ere emaitzaren hobekuntza estaldurari zor zaio. Gogoratu beharra dago, denbora erlazioen identifikazioa eta kategorizazioa egiten duen sailkatzaileak gertaeren atributuek hartzen dituzten balioak eta gertaera bera (testua) hartzen dituela ezaugarritako. Gainera, kontuan hartu behar da *bTime* etiketatzaileak `DCT`en eta identifikatutako gertaeren arteko erlazioak besterik ez dituela lantzen. Hortaz, esan dezakegu zerrenda gehitzeak, oro har, sistema hobetu egiten duela. Izan ere, erlazioen kategorizazioa gertaerekin lotura daukaten azpiataza guztiek eragiten dioten ataza da, eta emaitzek eragin hau positiboa dela adierazten dute.

Aipatu beharra dago *bTime* etiketatzaileari zerrenda txertatzearekin, aldaketaren aurretik baliorik altuenak itzultzen zituzten konfigurazioak, hau da, testuinguru leihoen token kopuruak eta ezaugarri multzoak, aldatu egiten direla azpiataza batzuentzat. Hauek gutxienak dira: `class`, `polarity`, `pos` eta `modality` atributuen esleipena, zerrendarekin hobetzen diren lau atributuak hain zuzen ere.

3.3.5 SRLren eraginaren azterketa

Tesi lan honetan zehar esan dugu baieztatu nahi diren bi hipotesietako bat ondorengoa dela: euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek daukaten eragina positiboa dela, ingelesez eta gaztelaniaz bezala. Hau betetzen den edo ez ikusi ahal izateko, lehenbiziko atalean aurkeztutako *bRol* tresnak *bTime* etiketatzailean duen efektua aztertu dugu. Zehazkiago, honetarako 3.3.3 azpiatalean deskribatu dugun *Leave-One-Out* (LOO) prozedurarako zerrendatu multzoetatik `SEM` ezagutaraziak duen eragina kalkulatu dugu.

*bTime*ren ebaluazioaren kasuan, ezaugarriak banaka ateratzen joan beharrean, ezaugarri-multzoak, `SEM` besteak beste, ateratzen dira. 3.11 taulan adierazten den bezala, azkeneko horrek *bRol* tresnaren bidez lortutako bi ezaugarri biltzen ditu, *bTime*k prozesatu beharreko token bakoitzarentzat erauzten direnak: tokenaren rol semantikoa (hala dagokionean), eta tokenetik esaldiaren erro semantikorainoko tokenen rol semantikoak. `SEM` bildumaren efektua *bTime*n ikasketa automatikoa baliatzen duten azpiataza

guztientzat aztertu da. 3.15 taulan ikus daiteke zein diren SEMeko ezaugarriak erabiltzen ez direnean lortzen diren balioak[13] (ALL−SEM). Gainera, balio hauek *bTime* sistemako sailkatzaileak sortzeko ezaugarri guztiak inplementatzerakoan lortutako emaitzekin alderatzen dira (ALL).

Entitatea	Ataza	Eskema	Konfigurazioa		F_1	Hobekuntza
<EVENT>	id	S/R	1	ALL / ALL−SEM	75.79 / 76.00	-0.21
	class	−	3	ALL / ALL−SEM	58.16 / 58.06	+0.10
	tense1	−	3	ALL / ALL−SEM	62.94 / 62.26	+0.68
	tense2	−	3	ALL / ALL−SEM	63.96 / 63.43	+0.53
	aspect1	−	3	ALL / ALL−SEM	64.25 / 64.01	+0.24
	aspect2	−	3	ALL / ALL−SEM	64.83 / 64.88	-0.05
	polarity	−	3	ALL / ALL−SEM	74.55 / 73.88	+0.67
	pos	−	3	ALL / ALL−SEM	69.62 / 68.51	+1.11
	modality	−	3	ALL / ALL−SEM	73.1 / 72.13	+0.97
<TIMEX3>	id	S	1	ALL / ALL−SEM	45.26 / 41.05	+4.21
	id	R	1	ALL / ALL−SEM	65.26 / 64.21	+1.05
	type	−	1	ALL / ALL−SEM	48.42 / 48.42	0
<TLINK>	relType	TAS	3	ALL / ALL−SEM	59.55 / 59.70	-0.15

Taula 3.15: *bTime* etiketatzailearentzat lortutako emaitzak (S: Strict, R: Relaxed).

3.15 taulan ikus daitekeen moduan, rol semantikoek eskaintzen duten informazioa ibiltzeak eragin positiboa du *bTime*k ikasketa automatiko bitartez burutzen dituen hamabi azpiatazatatik bederatzitan. Batean ez du efekturik eta gainerako hiruretan hura negatiboa da. *ISO-TimeML* eskemako entitateei erreparatuta, berriz, ikus dezakegu rol semantikoen eragina gertaeren identifikazioan ez dela positiboa, baina bai denbora adierazpenenean. Izan ere, gertaeren detekzioan, SEM darabilgunean, F_1 neurria 0.21 puntu okertzen da, eta adierazpenen identifikazioa, ordea, 4.21 eta 1.05 puntu hobetzen da (*Strict* eta *Relaxed* F_1). Hau da *bTime* tresnako azpiataza guztietatik SEM multzoaren ondorioz hobekuntzarik nabarmenena jasaten duena. Gure iritziz, *bRol*ek predikatuen adjuntu tenporalei ezartzen dizkien etiketek, TMP motakoek alegia (ikus 2.1.1 eta 2.1.2 azpiatalak), *bTime*ri franko laguntzen diote une bakoitzean prozesatzen ari den tokenak adierazpenen zati diren edo ez erabakitzen. Denbora erlazioen kategorizaziori dagokionez, taulan emaitza 0.15 puntu okertzen dela ikus daiteke. Emaitza hau bi faktoreren elkarketaren ondorioz jaisten dela uste dugu: *bTime* DCTeen eta gertaeren arteko erla-

[13] Oharra: eraginaren azterketa 3.3.4 azpiatalean deskribatutako hobekuntzaren aurretik egin dugu.

zioez bakarrik arduratzen delako batetik, eta SEM erabilita gertaeren identifikazioa okertu egiten delako bestetik.

Hipotesiaren egiaztapenean aurrera jarraitu ahal izateko, kontuan izan behar da euskarazko informazio tenporala etiketatzean rol semantikoek duten efektua kalkulatzeaz gainera, ingelesez eta gaztelaniaz dutena ere aztertu behar dela. Beraz, rol semantikoen eraginaz euskararako kalkulatu ditugun balioak Llorensek (2011) *TempEval-2* ebaluazio saioko corpusak eta *TIPSem* izeneko tresnaren bidez neurtutako emaitzekin alderatuko ditugu. Argitalpen horretan hiru multzotan sailkatu ezaugarri semantikoek (LS, TS eta SR) ingelesezko eta gaztelerazko informazio tenporalaren etiketatzean duten eragina aztertzen da. Ezaugarri hauek osatzen dituzte hiru multzoak: semantika lexikokoek LS, semantika tenporalekoek TS eta rol semantikoek SR. Gure interesekoa azkenekoaren eragina da. Alderaketa hau, esan dugun moduan, ez da izango behar bezain esanguratsua, eta, beraz, ez du balioko hipotesia betetzen den ala ez zalantzarik gabe ikusteko. Balekoa izanen da, ordea, hipotesia guk deskribatutako egoeran, hots, gure tresna eta corpusekin betetzen denetz jakiteko. Aldi berean, eta neurri batean behintzat, ebaluazio-egoera esanguratsuan hipotesia beteko litzatekeen edo ez zantzua iradokiko digu.

3.16 eta 3.17 tauletan Llorensen argitalpenean aurkezten diren emaitzak bildu ditugu. Aipatu dugun bezala, sistema horretan rolekin lotutako ezaugarriak gure SEM multzoaren iruditsua den SR izeneko multzoan daude kokatuak. Aipatu bi hizkuntzetan SRLren eragina positiboa den edo ez aztertzeko *TIPSem* tresnaren bi bertsio alderatu ziren: *TIPSem-B* eta *TIPSem-SR*. Lehenbizikoak ezaugarri morfosintaktikoak bakarrik erabiltzen ditu. Bigarrenak, aldiz, ezaugarri morfosintaktikoez gainera rolekin lotutakoak ere erabiltzen ditu, SR multzokoak. Hurrengo tauletako *Doitasuna*, *Estaldura* eta F_1 zutabeetan *TIPSem* sistemaren bi bertsio hauek *TempEval-2* ebaluazio saioan erdietsitako emaitzak biltzen dira, *TIPSem-B*renak ezkerrean eta *TIPSem-SR*renak eskuinean. Taulek agerian uzten duten moduan, ebaluaziorako erabili diren metrikak ez dira *bTime*en ebaluaziorako berak, hau da, *TempEval-3* eta *EVENTI-2014* saioetakoak, *TempEval-2* saiokoak baizik.

Denbora etiketatze ia osoan ingelesez eta gazteleraz ere rolen eragina positiboa da: lehen mintzairan lau azpiatazatatik hirutan (gertaeren eta adierazpenen identifikazioan eta erlazio tenporalen kategorizazioan) SR multzoaren eragina positiboa da; beste batean (gertaeren sailkapenean) ez du eragiten. Gazteleraz, berriz, erdietan (gertaeren eta adierazpenen identifikazioan) rolek eragin positiboa dute eta, gainerakoetan (gertaeren eta

Entitatea	Ataza	Eskema	Doitasuna	Estaldura	F_1	Hobekuntza (F_1)
<EVENT>	id	R	82 / 82	80 / 86	81 / 84	+3
	class	–	79 / 79	- / -	- / -	-
<TIMEX3>	id	R	89 / 91	68 / 76	77 / 83	+6
<TLINK>	relType	S	80 / 82	- / -	- / -	-

Taula 3.16: *TIPSem-B* eta *TIPSem-SR* etiketatzaileek ingelesez lortutako emaitzak.

Entitatea	Ataza	Eskema	Doitasuna	Estaldura	F_1	Hobekuntza (F_1)
<EVENT>	id	R	90 / 92	86 / 87	88 / 89	+1
	class	–	66 / 66	- / -	- / -	-
<TIMEX3>	id	R	97 / 96	81 / 88	88 / 91	+3
<TLINK>	relType	S	59 / 59	- / -	- / -	-

Taula 3.17: *TIPSem-B* eta *TIPSem-SR* etiketatzaileek gazteleraz lortutako emaitzak.

erlazioen sailkapenean), ez dute inolako efekturik. Erlazioen karietara, aipatu beharra daukagu 3.16 eta 3.17 tauletan bildutako balioak ez dagozkiela testu soiletik abiatuta kategorizatutako *DCT*en eta gertaeren arteko erlazioei, *TempEval-2* saioko C azpiatazako *DCT*en eta gertaeren arteko erlazioen kategorizazioei baizik (ikus 3.1.4 azpiatala). Ezberdintasuna da *TempEval-2* saioan erlazioen kategorizazioa egiten zela bakarrik eta ez identifikazioa.

Rol semantikoek denboraren etiketatzean duten efektua aztertzen denean, bestalde, aintzat hartu beharra dago hizkuntza bakoitzeko SRL tresnaren eraginkortasunak ere zerikusia baduela. *TIPSem-SR* sisteman *CCG* izenekoa (Punyakanok et al., 2004) erabili zen ingeleserako, eta *AnCora* corpusa (Taulé et al., 2008) gaztelerarako. Lehenbizikoak 70.07, 63.07 eta 66.39 puntuko doitasuna, estaldura eta F_1 neurriak dauzka. *AnCora*n, berriz, rolen etiketatzea erdi automatikoki egin zen, eta ez zen prozedura honen ebaluaziorik burutu. Hortaz, ezin jakin daiteke gaztelerarako baliatutako rolak etiketatzean erdietsi doitasuna, estaldura eta F_1 neurria zein diren. Hauek ezin ditugu, bestalde, zuzenean euskarazko *bRol* tresnaren emaitzekin alderatu, 2.1.3 azpiatalean azaldu dugun gisan, gure sistemak, semantika (eta sintaxia) adierazteko, dependentzietan oinarrituriko formalismoa baliatzen baitu. *CCG* tresnak eta *AnCora* corpusak, berriz, osagaietan oinarritutako adierazpidea jarraitzen dute. Zernahi gisaz, ideia bat egiteko baizik ez bada ere, ingeleseko eta gaztelerako SRLren eraginkortasuna *bRol* sistemak argumentuen sail-

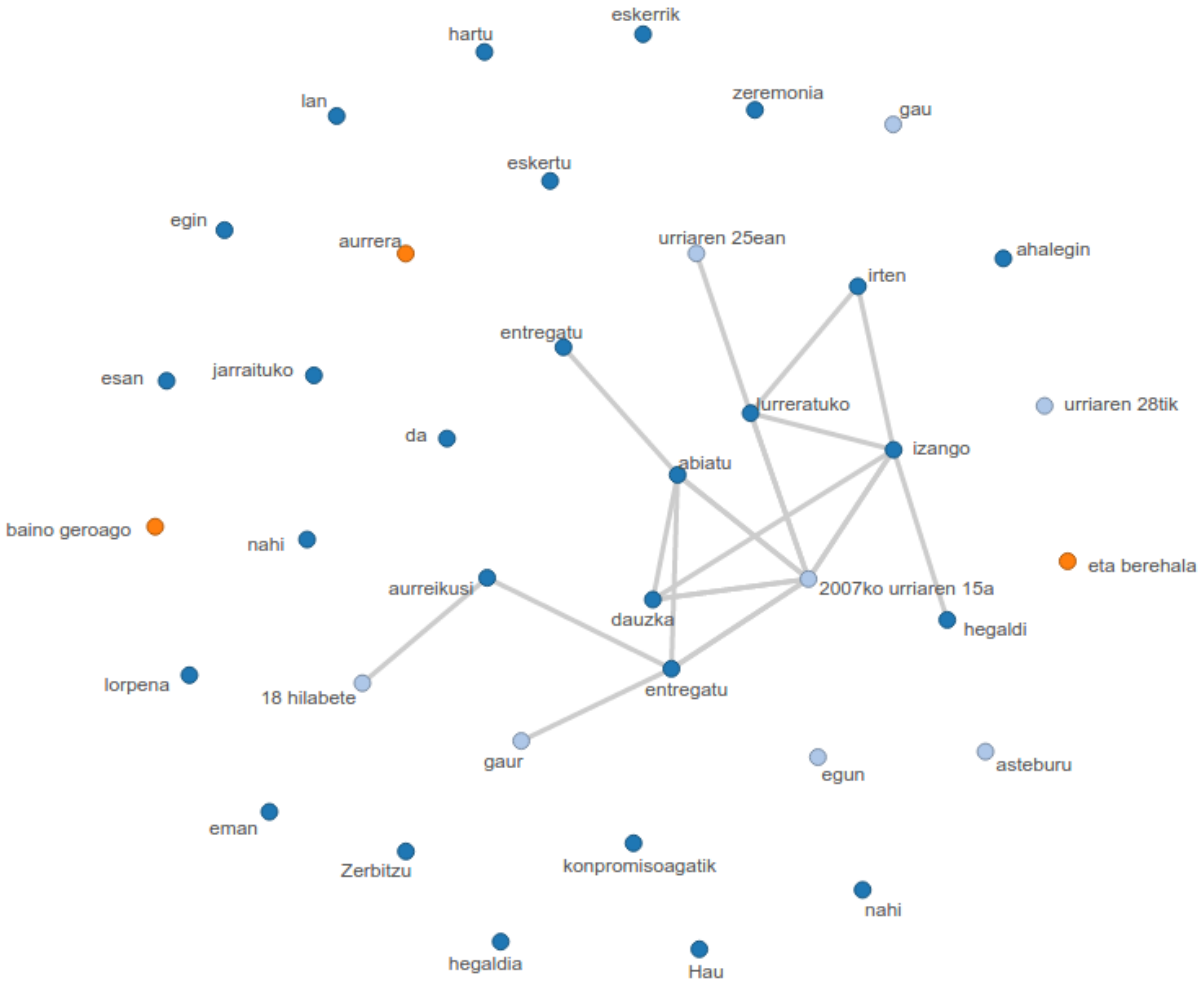
kapenean erdiesten dituen emaitzekin aldera daiteke. Bertan, 77.6, 77.8 eta 77.5 puntuko doitasuna, estaldura eta F_1 neurriak lortu zituen *bRol*ek.

3.4 *VisualTime*: bisualizaziorako interfazea

Tesi lanaren atal honetan aurkeztu dugun *bTime* tresnaren garapenean, eta honekin lotuta egin ditugun esperimentu eta alderaketetan, sarritan aztertu behar izan dugu *ISO-TimeML* formatuko fitxategietan bildutako informazioa. Azterketa hau erraztu eta informazioa lehen begi kolpean hobeki interpretatu ahal izateko, *VisualTime* izendatu dugun interfazea inplementatu dugu. *ISO-TimeML* formatuko fitxategien bisualizaziorako interfaze hau *JavaScript* lengoaia[14] eta *D3* liburutegia[15] erabilita garatu da. Modu honetara eta aparteko aplikaziorik instalatzeko beharrik gabe, *ISO-TimeML* formatuko informazioa zuzenean nabigatzailearen bitartez ikuskatzeko gai izan gara. 3.12 irudian ageri da *VisualTime* interfazearen adibidea. Bertan *Euskal-TimeBank* corpuseko `10021-First_Airbus_A380_delivered` fitxategiari dagokion edukia aurkezten da.

> *Lehen Airbus A380a entregatu da. 2007ko urriaren 15a. Airbus-en lehen A380 superjumboa gaur entregatu zaio Singapore Airlines-i (SIA), aurreikusi baino 18 hilabete geroago. Hegazkina Singapurren entregatu eta berehala, Tolosara abiatu da (Frantzia), 500 gonbidatu inguruko zeremonia batera. Hegazkinak lau Rolls-Royce Trent 900 motor dauzka. Lehen hegaldia ongintzako hegaldi berezi bat izango da, eta Singapurretik irten eta Sydney-n lurreratuko da, urriaren 25ean. Zerbitzu bereziek urriaren 28tik aurrera jarraituko dute. Hau sekulako lorpena da A380 programarentzat, eta eskerrik beroenak eman nahi dizkiet parte hartu duten guztiei. Gure taldeak eskertu nahi nituzke, eta baita egun, gau eta asteburu gogor lan egin duten horiek guztiak ere, beren ahalegin eta konpromisoagatik, esan du Tom Riddle Airbus-eko presidente eta buru exekutiboak.*

[14] `https://www.javascript.com/`
[15] `https://d3js.org/`



Irudia 3.12: `10021-First_Airbus_A380_delivered` fitxategiaren bisualizazioa, *VisualTime* interfazea erabilita.

Irudian ikus daitekeenez, gertaerak, denbora adierazpenak eta seinaleak (`<SIGNAL>`) kolore ezberdinetako puntuen bitartez eman dira aditzera: urdin ilunez, urdin argiz eta laranjaz. Gainera, hauen arteko erlazio tenporalak ere markatuta daude, puntuen arteko lerroak baliaturik[16]. Hau jakinda antzeman daitekeenez, dokumentuaren sorrera data (`DCTa`) *2007ko urriaren 15a* denbora adierazpena da, normalki sarearen erdian kokatzen dena. Erlazioak osatzen dituzten puntuen koloreei erreparatuta jakin daiteke erlazioetako bakoitza zer motatakoa den, hau da, bi gertaeraren, `DCTaren` eta gertaera baten, edo denbora adierazpen baten eta gertaera baten artekoa den.

[16]Oharra: adibideko fitxategian erlazioak bosgarren esaldiraino baizik ez daude anotatuak.

3.5 Ondorioak eta etorkizuneko lanak

Atal honetan *bTime* tresna aurkezten da, euskararako inoiz izan den lehenbiziko denboraren etiketatzaile automatikoa, gaur egun estandartzat hartzen den *ISO-TimeML* denbora informaziorako eskema baliaturik egina, eta, horregatik, *ISO-TimeML* jarraitzen duten beste hizkuntzetako sistemen esparru berean kokatzeko aukera izan dugu. Honek euskarazko emaitzak gaztelerako, ingeleseko eta italierako *TempEval-3* eta *EVENTI-2014* ebaluazio saioetakoekin erkatzea ahalbidetu du.

Emaitzen analisitik eta hauen alderaketatik etorkizuneko *bTime*n bertsio hobetuek kontuan izan beharko lituzketen ondorioak atera ditugu. Besteak beste, ikusi da inplementatutako hizkuntza ezaugarrietako batzuek, zazpi eta hamabost tokeneko testuinguru leihoak hartzen dituztenean, ikasketa prozesuan *zarata* egiten dutela, hein batean corpusaren izariarengatik. Etorkizuneko *bTime*en bertsioetan, beraz, ondorengo jokabidea inplementatuko da emaitzak hobetzen ahal diren aztertzeko: tokenentzat zazpi edo hamabost hitzeko leihoak hartuko dira, baina ezaugarri guztiak haien barneko bateko edo hiruko leihoentzat bakarrik erauziko dira. Gainerakoentzat, berriz, balio diskretuak jasotzen dituzten ezaugarriak baizik ez dira hartuko.

Bigarren atal honetan tesi lanaren hipotesietako bat egiaztatzen ere saiatu gara: euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek daukaten eragina positiboa dela, ingelesez eta gaztelaniaz bezala. Hau hasieran adierazitako egoeran ezin baieztatu izan bada ere, *bTime* tresnan *bRol*ek esleitutako rolek oro har positiboki eragiten dutela ikusi da.

Azkenik, aipatu nahi dugu etorkizunean, euskararako eskuragarri dagoen baliabide kopurua handitzen bada, aukera izango dela *bTime ISO-TimeML* eskemak zehazten dituen gainerako etiketak landu ahal izateko egokitzea. Gainera, *Euskal-TimeBank* corpusa zabalduko balitz gertaeren eta adierazpenen arteko denbora erlazioak ere etiketa litezke, *bTime*ren bidez.

4

ESPAZIO INFORMAZIOAREN ETIKETATZE AUTOMATIKOA

Gertaerak, denboran eta espazioan kokatuta dauden jazoerak, tesi lan honen ardatz direla aipatu dugu. Gainera, hauen *predikatu-argumentu-adjuntu* egiturak eta espazio-denborarekin lotutako propietateak testuetako semantikaren egituraketa automatikoki mapatu ahal izateko baliagarriak direla erakutsi dugu.

Atal hau ingelesezko testuetan bildutako espazio informazioaren etiketatzeaz arduratzen da. Horretarako, *ISO-Space* (Pustejovsky et al., 2011) eskema hartu da. *ISO-Space* sortu zenean ingelesa prozesatzeko gaitasuna besterik ez zuen, *ISO-TimeML*k hasiera batean bezala. Gainera, urteekin *ISO-TimeML* hainbat hizkuntzarako egokitzea lortu bada ere, euskararako besteak beste, *ISO-Space*k ingelesa prozesatzeko gaitasuna bakarrik dauka oraindik ere. Euskararako egokitze prozesuaren lehenengo urratsak egin direla zehaztu dugu, baina, ezin esan daiteke egokitzapena bukatuta dagoenik. Izan ere, oraindik ez dugu, besteak beste, euskararako espazio informazioa etiketatzen duen tresna garatu ahal izateko ezinbestekoa den corpus etiketaturik (*Euskal-SpaceBank*). Baliabideen egoera hau denez, tesi honen helburuen artean euskararako garatu gogo den etiketatzailerako aurrekaria ezartzea dago eta, horretarako, ingelesez espazioa markatzen duen *X-Space* (Salaberri et al., 2015b) garatu da. Etorkizunean burutu beharreko lanen artean kokatu dugu, beraz, *bSpace* izendatuko genukeen euskararako espazioaren etiketatzailea sortzea.

4.1 Ikerketaren egungo egoera

Atal honetan leku informazioaren anotaziorako *ISO-Space* jarraitzen duen ingeleseko *X-Space* tresna aurkezten da. Honen ebaluazioa *SemEval-2015* saioko[1] *SpaceEval* (Pustejovsky et al., 2015) atazaren barnean egin zen. Honek *X-Space* testuinguru estandarrean ebaluatu eta pareko sistemekin alderatzea ahalbidetu zuen. Ikerketaren egungo egoeraren gainbegiratua ematean kontuan izango dugu, hortaz, *SpaceEval* saioa: parte hartu zuten gainerako leku-etiketatzaileak, corpusa, ebaluaziorako metrikak eta emaitzak. Izan ere, zehaztu beharra dago *SpaceEval* izan zela *ISO-Space* oinarritako hartuta leku informazioaren etiketatzea egiteko sistemak garatzeko erronka bota zuen lehenengoa.

Rol semantikoei eta denborari dagozkien ataletan egin dugun bezala, hemen ere ikerketaren egungo egoera hiru azpiataletan banatuko dugu: *espazioa markatzeko hizkuntzak eta etiketatzaileak* (4.1.1 azpiatalean), *hizkuntza baliabideak* (4.1.2n) eta *ebaluazio saioak* (4.1.3n). Lehenengoan *ISO-Space* eskemaren sorreraren aurretik espazioa anotatzeko erabili izan diren *SpatialML* (Mani et al., 2008) eta *Spatial Role Labeling*-SpRL (Kordjamshidi et al., 2010) eskemak eta hauek jarraituta garatutako sistemak deskribatu eta zerrendatuko ditugu. Gainera, *ISO-Space* ere atal honetan aurkezten da. Bigarrenean, berriz, *SpaceEval* saioan erabilitako corpusa eta aurreko eskemak jarraituta osatu ziren corpusak azalduko ditugu. Azkenik, hirugarren azpiatalean, *SemEval-2012*[2] eta *SemEval-2013*[3] ebaluazio saioetan SpRL eskema erabilita egindako atazak aurkeztuko dira. Gainera, *SpaceEval* saioa bera ere zerrendatu eta deskribatuko dugu hemen.

4.1.1 Espazioa markatzeko hizkuntzak eta etiketatzaileak

Hiru dira hizkuntzaren prozesamenduan espazio informazioa markatzeko erabili izan ohi diren hizkuntzak: *SpatialML*, *Spatial Role Labeling* eta *ISO-Space*. Zerrenda hau kronologikoki ordenatuta dago, sorrera dataren arabera (zaharrenetik berrienera).

SpatialML

SpatialML 2008. urtean aurkeztu zen eta leku informazioa anotatzeko lehenbiziko eskema esanguratsua izan zen. Eskema honek ondokoak hartu zituen oinarritako: toponimoen

[1] http://alt.qcri.org/semeval2015/
[2] https://www.cs.york.ac.uk/semeval-2012/
[3] https://www.cs.york.ac.uk/semeval-2013/

etiketatzeaz diharduen Leidner-en (2006) lana, Schilder eta besteren (2004) ikerketa eta (Garbin eta Mani, 2005) argitalpenean aurkeztutako eskema. *SpatialML*ren xedea toki izenak eta hauen arteko erlazioak etiketatzea da. Leku izenendako <PLACE> etiketa erabiltzen da eta erlazioendako <LINK> eta <PATH> etiketak bereizten dira; <SIGNAL> ere erabiltzen da, *seinale espazialak* anotatzeko. Hurrengo adibideak *SpatialML* eskemaren erabilera erakusten du.

A town 50 miles south of Salzburg in the central Austrian Alps.

<PLACE>town**</PLACE>**
<PLACE>Salzburg**</PLACE>**
<PLACE>Austrian**</PLACE>**
<PLACE>Alps**</PLACE>**
<SIGNAL>50 miles**</SIGNAL>**
<SIGNAL>south**</SIGNAL>**

Adibideko esaldian toki izenak eta seinaleak etiketatu dira. Aipatu beharra daukagu Mani eta besteren (2008) arabera *SpatialML*k <PLACE> etiketarekin toki izenei egiten zaizkien erreferentzia absolutuak (*Rome*, *Rochester, NY*, *southern Kerala district of Cudallah*) eta erlatiboak (*an underpass beneath Pushkin Square*, *in the vicinity of Georgetown University*), biak, markatzen dituela. Gainera, leku izen bereziak eta arruntak bereizten dira (NOM eta NAM erazagutuak); aipatu adibidean, esaterako, *town* arrunta da eta *Salzburg*, *Austrian* eta *Alps* bereziak. Jarraian, etsenpluko esaldian *SpatialML*k finkatzen dituen <LINK> eta <PATH> motako erlazioak zehazten ditugu.

A [*town*] **50 miles** *south of* [*Salzburg*] *in the central Austrian Alps.*
<PATH>

A [*town*] 50 *miles south of Salzburg* **in** *the central Austrian* [*Alps*].
<LINK>

A town 50 *miles south of Salzburg in the central* [*Austrian*] [*Alps*].
<LINK>

Hiru lekuzko erlazio identifikatu dira adibideko esaldian: <PATH> motako bat eta <LINK> motako bi. Lehenbizikoa *50 miles* eta *south* seinaleek *markatzen* dutela esan daiteke; adibidean, *town* eta *Salzburg* lekuen artean berrogeita hamar mila luze den bidea dagoela adierazten da. Beste bi erlazioek, aldiz, *town* eta *Alps* eta *Austrian* eta *Alps* tokiak erlazionatzen dituzte. Adibidean ageri ez bada ere <LINK> erlazioak sei motakoak izan daitezkeela jakin beharra dago, eta mota `type` izeneko atributuan zehazten dela. Kasu honetan, bi <LINK> erlazioak IN motakoak dira, erlazionatutako lekuak bata bestean kokatuta daude, alegia, hiria (*town*) Alpeen (*Alps*) barnean eta Alpeak Austrian (*Austrian*). Lehenengo <LINK> erlazioan *in* tokenak adierazten du mota; bigarrenean, berriz, *Austrian* tokeneko *n*-ak.

Gaurdaino *SpatialML* jarraitzen duen sistema automatiko bakarra garatu den ziurtasuna dugu; hau (Mani et al., 2008) argitalpenean deskribatzen da, *SpatialML* eskemarekin batera. Bertan esaten den eran, bi elementuk osatzen dute sistema: toki izenak, seinaleak eta erlazio espazialak etiketatzen dituen moduluak (*tagger*) eta leku izenen desanbiguatzaileak (*disambiguator*). Lehenbizikoak entitateak identifikatu eta hauen atributuak esleitzen ditu; bigarrenak aurreneko osagaiak identifikatu dituen toki izen bereziak desanbiguatzen ditu. Demagun *Plaza Gorria* leku izena identifikatu dela eta bi hiritan daudela izen hori duten plazak, Moskun eta Iruñean. Kasu honetan desanbiguatzaileak zehaztuko du zer hiritako plaza den identifikatu dena.

Sistema honen emaitzak hauek dira: leku izen arruntak (NAM) identifikatzean 85 puntuko F_1 neurria lortu zen, eta toki izen bereziak (NOM) identifikatzerakoan 72 puntukoa. <SIGNAL>, <PATH> eta <LINK> etiketentzat ez zen ebaluaziorik burutu. Toki izen bereziak desanbiguatzean 93 puntuko F_1 neurria erdietsi zen.

Spatial Role Labeling-SpRL

Spatial Role Labeling deritzon eskema (SpRL) 2010. urtean aurkeztu zen. SpRL Kordjamshidi eta bestek (2010) proposatu zuten, eta *ISO-Space*n sorreraz geroztik azken honek barneratu duela esan daiteke. Izan ere, jarraian *ISO-Space* deskribatuko dugunean adieraziko den moduan *ISO-Space*ko hiru erlazio motetan parte hartzen duten osagaiei rol espazialak esleitzen zaizkie. Rol horiek osagai espazialek aipatu hiru erlazioetan jokatzen duten papera adierazten dute. SRL atazaren alderako analogiaz osagai espazialak SRLko argumentuak izango lirateke, eta rol espazialak, aldiz, rol semantikoak.

SpRL eskeman bost etiketa definitzen dira: <TRAJECTOR> eta <LANDMARK> rol espazialak, <SPATIAL-INDICATOR> eta <MOTION-INDICATOR> erlazioen adierazleak eta <SR> erlazioak. Hurrengo adibideak SpRLren erabilera erakusten du.

She is playing in her room.

<TRAJECTOR>She**</TRAJECTOR>**
<LANDMARK>her room**</LANDMARK>**
<SPATIAL-INDICATOR>in**</SPATIAL-INDICATOR>**
<MOTION-INDICATOR>playing**</MOTION-INDICATOR>**

Adibideko esaldian rol espazialak eta erlazioen adierazleak etiketatu dira, bi rol eta bi adierazle denetara. *Trajector* rola *She* argumentu edo entitateak jokatzen duela erabaki da, enuntziatuko subjektua den osagaia delako (*gaia* edota *pazientea* SRLn). *Landmark* rola, aldiz, *her room* lekuak betetzen duela ezarri da, esaldiko gertaera (*playing*) non gauzatzen den adierazten duelako, kokalekua (AM-LOC adjuntua SRLn) alegia. Gainera, *in* eta *playing* espazio eta mugimendu adierazleak etiketatu dira. Lehenengoak gertaeraren eta kokalekuaren arteko lotura espaziala zehazten du, non jazo den; bigarrenak gertaera bera, mugimendukoa den *playing*, adierazten du. Jarraian, adibideko esaldian SpRL eskemak etiketatutako osagaien artean finkatzen duen <SR> erlazioa azaltzen da.

[*She*] *is* *playing* **in** [*her room*].
<SR>

Ikus daitekeenez, erlazio honetan aurretik etiketatu ditugun rolen eta adierazleen arteko menpekotasunak ezartzen dira. Erlazioak adierazten duen bezala, mugimendu bat (kokaleku aldaketa normalean) leku baten barnean (*in* eta *her room*) jazo da (*playing*), eta mugimendu honetan parte hartu duena *She* izenordainaren bitartez erreferentzia egiten zaion argumentua edo entitatea da.

SpRL eskema baliaturik garatu diren sistema automatikoei dagokienez, esan beharra dugu hainbat izan direla gaur arte sortutakoak. Izan ere, kontuan hartu behar da *SemEval-2012* eta *SemEval-2013* ebaluazio saioen barnean eskema hau jarraitzen zuten tresnak garatzean zentratu ziren atazak antolatu zirela. 4.1.3 azpiatalean bi ebaluazio saio hauek deskribatzean aipatuko ditugu haietan garatutako tresnarik esanguratsuenak.

ISO-Space

Tesi lan honetan espazioa etiketatzeko erabiltzen den *ISO-Space* eskemak ondoko hamar etiketak bereizten ditu:

- <PLACE>: Toki izenak markatzen ditu (*eraikina*, *Tokio*, *Everest mendia*).
- <PATH>: Bide izenak markatzen ditu (*errepidea*, *A-8 autobidea*, *ibilbidea*).
- <SPATIAL_ENTITY>: Gertaera dinamiko edo mugimenduzkoen argumentuak markatzen ditu (*Mikel A-8 autobidean barna etorri da* esaldiko *Mikel*).
- <MOTION_EVENT>: Gertaera dinamiko edo mugimenduzkoen buru lexikalak markatzen ditu (*joan*, *ibili*, *erori*).
- <NONMOTION_EVENT>: Gertaera estatiko edo mugimendu gabekoen buru lexikalak markatzen ditu (*egon*, *pentsatu*, *ikusi*).
- <SPATIAL_SIGNAL>: *Seinale espazialak* markatzen ditu (*ezkerrekoan*, *barnean*, *ezkerretan*). Seinale hauek espazio eta orientazio erlazioak (<QSLINK> eta <OLINK>) adierazten dituzte. Hiru espazio-seinale mota daude: aldi berean norabidekoak eta topologikoak direnak (*ezkerrekoan*), eta norabidekoak (*ezkerretan*) edo topologikoak (*barnean*) baizik ez direnak, hots, bata edo bestea, ez biak batean.
- <MOTION_SIGNAL>: *Mugimendu seinaleak* markatzen ditu (*ezkerrekora*, *barnetik*, *ezkerretara*). Hauek mugimendu erlazioak (<MOVELINK>) adierazten dituzte.
- <MOVELINK>: Gertaera dinamikoetan parte hartzen duten osagai espazialen (<PLACE>, <PATH>, <SPATIAL_ENTITY>) arteko erlazioak markatzen ditu (<MOVELINK> = *Movement link*).
- <QSLINK>: Seinale espazial topologikoen menpekoak diren osagai espazialen arteko erlazioak markatzen ditu (<QSLINK> = *Qualitative spatial link*).
- <OLINK>: Norabide seinale espazialen menpekoak diren osagai espazialen arteko erlazioak markatzen ditu (<OLINK> = *Orientation link*).

ISO-Space eskema (*ISO-TimeML* bezala) ingeleserako garatu zen. Gaur egun, euskararako egokitze prozesuaren lehenengo urratsak eman badira ere, ezin esan daiteke egokitzapena bukatuta dagoenik[4]. Hurrengo adibideek euskarazko *ISO-Space* eskemak izango lukeen erabilera erakusten dute.

Goiko harmailetan zeuden jarraitzaileek B1 irtenbidea erditik igaro zuten.

<SPATIAL_SIGNAL>Goiko**</SPATIAL_SIGNAL>**
<PLACE>harmailetan**</PLACE>**
<PLACE>Goiko harmailetan**</PLACE>**
<NONMOTION_EVENT>zeuden**</NONMOTION_EVENT>**
<PATH>irtenbidea**</PATH>**
<PATH>B1 irtenbidea**</PATH>**
<MOTION_SIGNAL>erditik**</MOTION_SIGNAL>**
<MOTION_EVENT>igaro**</MOTION_EVENT>**
<SPATIAL_ENTITY>Goiko harmailetan zeuden jarraitzaileek**</SPATIAL_ENTITY>**
<SPATIAL_ENTITY>B1 irtenbidea**</SPATIAL_ENTITY>**

Adibidean bi gertaera etiketatu dira: *zeuden* estatikoa eta *igaro* dinamikoa. Goian azaldu bezala, *ISO-Space* eskemak mugimenduko gertaeren argumentuak <SPATIAL_ENTITY> moduan markatzen ditu. Kasu honetan *igaro*ren *Goiko harmailetan zeuden jarraitzaileek* eta *B1 irtenbidea* argumentuak markatu dira. Gainera, *Goiko* seinale espaziala eta *erditik* mugimendu seinalea ere etiketatu dira. Baita *harmailetan* eta *Goiko harmailetan* toki izenak eta *irtenbidea* eta *B1 irtenbidea* ibilbideak ere.

[*Goiko harmailetan zeuden jarraitzaileek*] [*B1 irtenbidea*] *erditik* **igaro** *zuten.*
<MOVELINK>

[4] http://ixa.si.ehu.es/master/master_tesiak/1487284481/publikoak/HAP_proiektua_Jasone.pdf

Adibidean *erditik* mugimendu seinaleak adierazten duen <MOVELINK> motako erlazioa etiketatu da. Honetan *igaro*, *igaro zuten* gertaera dinamikoaren buru lexikalak *Goiko harmailetan zeuden jarraitzaileek* entitate espazialak eta *B1 irtenbidea* bideak hartzen dute parte. *ISO-Space* eskemak <MOVELINK>, <QSLINK> eta <OLINK> erlazioetan parte hartzen duten osagai espazialei *rol espazialak* izenekoak esleitzen dizkie. Rol espazialek *osagai espazialek erlazio espazialetan jokatzen duten paper (espaziala) adierazten dute*. Kordjamshidik eta bestek (2010) proposatu zituzten lehenengo aldiz, eta ondoren *ISO-Space* eskemak barneratu zituen. Rol semantikoen etiketatzearen alderako analogiaz osagai espazialak SRLko argumentuak izango lirateke, eta rol espazialak, aldiz, rol semantikoak. Erlazio motaren arabera osagaiek jokatzen ahal dituzten rol espazialak aldatu egiten dira; <MOVELINK> motako erlazioen kasuan lau hauek dira:

- `source`: Mugimendua hasten den lekua, abiapuntua.
- `goal`: Mugimendua bukatzen den lekua, helmuga.
- `mover`: Mugimenduan kokalekua aldatzen duen osagaia, mugitzen dena.
- `landmark`: Mugimendua non barna egiten den adierazten duen osagaia, bidea.

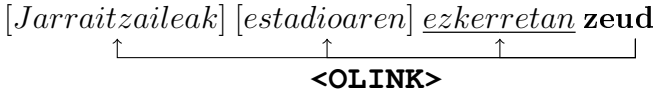
Gure adibidean *goiko harmailetan zeuden jarraitzaileek* osagaiak `mover` rola jokatzen du, eta *B1 irtenbidea* osagaiak, aldiz, `landmark` rola.

[*Jarraitzaileak*] [*estadioaren*] *barnean* **zeuden**.
<QSLINK>

Adibide honek, berriz, <QSLINK> erlazioen erabilera erakusten du. Erlazioa *barnean* seinale espazial topologikoak adierazten du, *zeuden* gertaera estatikoa *Jarraitzaileak* entitate espazialarekin eta *estadioaren* tokiarekin erlazionatzen dituenak. <QSLINK> erlazioek bi rol espazial hauek jasotzen dituzte:

- `trajector`: Erlazioaren subjektua den osagaia, gaia, pazientea.
- `landmark`: Erlazioa non gauzatzen den adierazten duen osagaia, kokalekua.

Adibidean *Jarraitzaileak* osagaiak `trajector` rola jokatzen du eta *estadioaren* osagaiak berriz `landmark` rola.



Azkeneko adibidean <OLINK> erlazioaren erabilera erakusten da. Bertan, *ezkerretan* norabidezko seinale espazialak *Jarraitzaileak* entitate espazialaren eta *estadioaren* tokiaren arteko orientazio erlazioa adierazten du. <OLINK> erlazioek <QSLINK> erlazioen rol espazial berak hartzen dituzte. Adibide honen kasuan, aurrekoarenean bezala, *Jarraitzaileak* osagaiak `trajector` rola betetzen du eta *estadioa* osagaiak, berriz, `landmark`.

Aurrera jarraitu aurretik, azpimarratu nahi dugu denbora eta espazio informazioa etiketatzeko atazak, zehazkiago esan *ISO-TimeML* eta *ISO-Space* anotazio eskemak, antzekoak baldin badira ere, badituztela kontuan eduki beharreko ezberdintasunak. Gure ustez nabarmenenak eta tesian garatutako tresnen diseinuan eragin handiena izan dutenak ondorengoak dira:

1. *ISO-TimeML*n, erlazioak gauzatzeko, ez da beharrezkoa seinaleak agertzea; *ISO-Space*n bai. Gainera, *ISO-TimeML*eko erlazioentzat azpimotak definitzen dira eta *ISO-Space*ko erlazioentzat, aldiz, ez.

2. Badira rol espazialak, baina ez dago rol tenporalik.

3. *ISO-Space*n gertaeren mugimenduaren araberako kategorizazioa etiketa mailan gauzatzen da; *ISO-TimeML*n, berriz, kategorizazioa atributu mailan egiten da.

SpatialML, *SpRL* eta *ISO-Space* eskemen alderaketa

Azpiatal honetan espazio informazioa etiketatzeko erabili ohi diren eskemen alderaketa egiten dugu. Zehazkiago esan, *SpatialML* eta *SpRL* eskemek *ISO-Space*kin dituzten antzekotasun eta ezberdintasunak eta haien etiketen artean aurki daitezkeen baliokidetasunak aipatzen ditugu. Espazio erlazioei erreparatuta, antzeman daiteke *SpatialML*k badituela *ISO-Space* eskematik bereizten duten hainbat faktore nabarmen, *SpatialML* aurkeztu dugunean erabilitako adibidea hartuta esate baterako: *Austrian* eta *Alps* <PLACE> bezala etiketatzeaz gainera, *ISO-Space* eskemak *Austrian Alps* ere etiketatuko luke. Hau *Austrian* eta *Alps* tokenen artean <LINK> moduko erlazio bat ezarrita egiten da *SpatialML*n. Gainera, *ISO-Space* eskeman *in*, *central* edota *in the central* tokenak ere seinaletako etiketatuko lirateke; *SpatialML*n, berriz, *south* bakarrik. Kontuan izan beharra dago bi

eskemen arteko ezberdintasunetan eragina izan dezakeela hurrengoak: Mani eta besteek (2008) diotenez *SpatialML*, ikuspegi sintaktikotik, etiketen hedadura ahalik eta laburren mantentzen saiatzen da, eskema honen araberako corpusen eskuzko anotazioa errazagoa egiteko. Hau lortzeko adjektiboak, determinatzaileak eta bestelako modifikatzaileak ez dira etiketen barnean kokatzen, izen berezien zati ez badira behintzat.

Eskema	Etiketa	Baliokidetasuna	Etiketa	Eskema
SpatialML	<PLACE>	≡	<PLACE>	***ISO-Space***
	<SIGNAL>	≈	<SPATIAL_SIGNAL>, <MOTION_SIGNAL>	
SpRL	<TRAJECTOR>	≈	<PLACE>, <PATH>, <SPATIAL_ENTITY>	***ISO-Space***
	<LANDMARK>	≈	<PLACE>, <PATH>, <SPATIAL_ENTITY>	
	<SPATIAL-INDICATOR>	≡	<SPATIAL_SIGNAL>, <MOTION_SIGNAL>, <MOTION_EVENT>	
	<MOTION-INDICATOR>	≡	<SPATIAL_SIGNAL>, <MOTION_SIGNAL>, <MOTION_EVENT>	
	<SR>	≡	<MOVELINK>, <QSLINK>, <OLINK>	

Taula 4.1: Espazioa etiketatzeko eskemen baliokidetasunak. ≈:Antzekoa, ≡:Berdina.

SpRL eta *ISO-Space* eskemetako adibideak alderatuta, aldiz, ikus dezakegu nabarmena dela ondoko ezberdintasuna: SpRLn <MOTION_INDICATOR> bezala etiketatzen dira *ISO-Space*ko mugimenduko gertaeren buru lexikalak eta mugimendua adierazten duten seinaleak (<MOTION_EVENT> eta <MOTION_SIGNAL>), baina <SPATIAL_INDICATOR> etiketa *ISO-Space*ko <SPATIAL_SIGNAL> etiketari bakarrik dagokio, ez <NONMOTION_EVENT> etiketari. Kontua da SpRL eskeman mota honetako gertaerak, mugimendukoak ez direnak alegia, ez direla etiketatzen. 4.1 taulak hiru eskemetako etiketen artean aurkitu ahal izan ditugun baliokidetasunak biltzen ditu.

Tesi lanean, hiruetatik, *ISO-Space* erabiltzea erabaki dugu. Izan ere, etorkizunean eskema hau, *ISO-TimeML* denborarako bezala, estandar bihurtzea espero dugulako. Kontuan izan beharra dago gainera, *ISO-Space*k *SpRL* barnean hartzen duela eta *SpatialML*, aldiz, proposatu zenetik ez dela baliabideak edo ebaluazio saioak egiteko erabili.

4.1.2 Corpusak

Orain, espazio informazioaren erauzketarako corpusak aurkezten ditugu. Horretarako corpus bakoitzak oinarri hartzen duen eskemaren araberako sailkapena jarraitzen dugu.

SpatialML

Gaur egun corpus bakarra dago eskuragarri *SpatialML* baliaturik anotatua izan dena: *Automatic Content Extraction*-ACE[5] programako ASC corpusa[6] (Mani et al., 2008). ASC zuzeneko elkarrizketen transkripzioek, berriek eta *weblog* motako webguneetatik eskuratutako ingelesezko 428 testuk osatzen dute. Kopuruei dagokionez, 6.338 <PLACE> etiketa daude, eta hauetatik 4.783 koordenadak dauzkaten toki izen bereziak dira.

ASC corpusaz gainera eskuragarri ez dagoen beste bat ere bada, *ProMED* corpusa[7] hain zuzen, *SpatialML* anotazioak dituena. *ProMED* sistema *International Society for Infectious Diseases*-ISID[8] erakundeak kudeatzen du, kutsakorrak diren eritasunen monitorizazioa egin eta posta bitartez abisuak eman ahal izateko. *ProMED* corpusa mota honetako 100 abisu-dokumentuk osatzen dute. Hasiera batean dokumentu guztiak anotatzaile bakar batek etiketatu zituen, eta ondoren, hauetako 41 beste anotatzaile batek etiketatu zituen berriro ere. ASC eta *ProMED* baliabideen kasuan <PLACE> etiketak markatzean lortutako *Inter-Annotator Agreement-IAA* neurtu zen. Lehenbizikoarentzat 77 puntuko F_1 eskuratu zen, eta bigarrenarentzat 92.3 puntukoa. Ikusten den gisan, aldea handia da bi corpusen artean. (Mani et al., 2008) argitalpenean adierazten denez, honen arrazoi nagusiak ondorengo biak dira: lehenbizikoa *SpatialML*ren aurreneko bertsioa erabilita, ikerketaren hasieran, garatu zela ASC, eta *ProMED*, aldiz, bukaeran; bigarrena, berriz, *ProMED*en, bestean ez bezala, anotatzaileak etiketatzaile adituak izan zirela.

SpRL

Denen eskura dauden *SpRL* eskemaren arabera anotatutako corpusak bi dira, *SemEval-2012* eta *SemEval-2013* ebaluazio saioetan erabilitakoak hain zuzen ere (Kordjamshidi et al., 2012) eta (Kolomiyets et al., 2013) hurrenez hurren. Lehenbiziko saioko corpusa

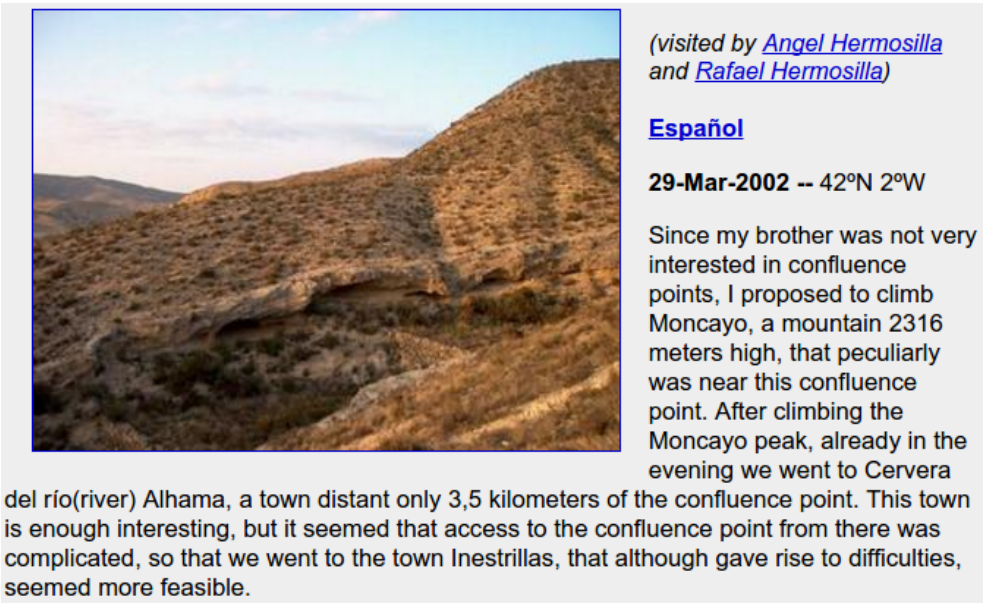
[5] https://www.ldc.upenn.edu/collaborations/past-projects/ace
[6] https://catalog.ldc.upenn.edu/LDC2008T03
[7] http://www.promedmail.org
[8] http://www.isid.org/

(Kordjamshidi et al., 2011) argitalpenean aurkeztutakoaren hedapena da, eta hau, aldi berean, *IAPR TC-12* (Grubinger et al., 2006) izeneko irudi-deskribapenen bildumaren azpimultzoa da. Bilduma horren edukiaren adibidea 4.1 irudian ikus daiteke.

Irudia 4.1: *IAPR TC-12* corpusaren adibidea (Grubinger et al., 2006).

Azpimultzo honetan 1.213 esaldi dituzten 613 testu biltzen dira eta hauek izan ziren *SemEval-2012* saioko corpusa osatu zutenak. Irudiak eta hauen deskribapenak turismo agentzia batek antolatutako bidaietan parte hartu zuten turistek egindakoak dira. Kontuan izan beharra dago, gainera, corpus hau osatu ahal izateko objektuak eta irudian hauen posizio absolutu edota erlatiboa azaltzen zuten irudi-deskribapenak bakarrik hartu zirela; deskribapen hauek beste mota bateko informazioa ere, espaziala ez dena, biltzen dute kasu batzuetan.

SemEval-2013 saioan erabilitako hizkuntza baliabidea, berriz, bi corpusen elkarketatik sortu zen: aipatu berri dugun *SemEval-2012* saiokoa eta *Degree Confluence Project*[9] izeneko proiektutik sortutakoa (*Confluence Project Corpus*-CPC) batzetik. Proiektu honen helburua mundu mailako latitude eta longitude osoen ($\mathbb{Z}$) arteko ebakidura-lekuen (*confluence points* deitzen dutenen) argazkiak eta deskribapenak egitea eta biltzea da. Halakoetara joan, argazkiak atera eta deskribapenak idazteaz arduratzen direnak boluntarioak izaten dira. Corpusak 1.789 esaldiz osatutako 117 testu dauzka, 40.000 token inguru. 4.2 irudian ikus daiteke 42 gradu iparraldera eta 2 gradu mendebaldera latitudea eta longitudea dituen puntuaren irudia eta deskribapena.



Irudia 4.2: CPC corpusaren adibidea (42°N 2°W).

Aurkeztu ditugun SpRL eskemarekin anotatutako corpusetatik, *SemEval-2012* saiokoaren anotazio prozesuaren datuak bakarrik dira ezagunak, hau da, ez da argitaratu CPC corpusaren anotazio prozesua nola egin zen eta zein den bertan lortutako *Inter-Annotator Agreement* neurria. *SemEval-2012* saioko corpusa bi anotatzailek etiketatu zuten hasiera batean (325 esaldi); bat aditua zen eta bestea ez. Hauen arteko IAA 89 puntukoa izan zen. Ondoren, corpuseko gainerako esaldiak anotatzeko (888 esaldi) hirugarren anotatzaile bat erabili zen. Hau ez zen aditua, baina hasi aurretik rol espazialak (SpRL eskema) nola anotatu behar zituen erakutsi zitzaion.

[9] http://confluence.org/

ISO-Space

*ISO-Space*n arabera anotatutako corpus bakarra dago eskuragarri: *SemEval-2015*[10] saioko *SpaceEval* (Pustejovsky et al., 2015) atazan erabilitako *SpaceBank* (Pustejovsky eta Yocum, 2013) corpusa hain zuzen ere. Baliabide hau CPC corpusaren, *American National Corpus*-ANC (Reppen et al., 2005) barneko *Berlitz Travel Guides* direlako bidaia giden eta *Ride For Climate*-RFC izeneko blogeko[11] sarreren elkarketatik sortu zen. CPCtik hartutako testuak *SemEval-2012* eta *SemEval-2013*n SpRL atazarako erabilitako berak izan ziren; aldea da *ISO-Space*n arabera anotatu zirela, *SpaceBank* sortzean.

Tesi lan honetan garatutako *X-Space* espazio informazioaren etiketatze automatikorako tresna *SpaceEval* saioren barnean sortu eta ebaluatu da. Horregatik, eta atal honen bukaeran egingo den analisian beharrezkoa izango delako, *SpaceBank* corpusaren datu estatistikoak biltzen dituen (Pustejovsky et al., 2015) argitalpeneko datuen taula aurkezten da jarraian (4.2 taula). Bertan *SpaceEval* atazarako baliatu corpusaren token, esaldi eta dokumentu kopuruak biltzen dira. Hauek, gainera, *SpaceBank* corpusa osatzen duten ANC, CPC eta RFC corpusetatik hartuak direnez gero, bakoitzetik hartutako kopurua zehazten da. Kopuruak *ISO-Space* eskemaren etiketetarako banatuta ere ageri dira.

	ANC	CPC	RFC	Train	Test	Denera
Tokenak	1577	7673	21048	24150	6148	30298
Esaldiak	61	369	821	1001	250	1251
Dokumentuak	3	22	44	55	14	69
<PLACE>	148	691	1250	1661	428	2089
<PATH>	19	246	278	415	128	543
<SPATIAL_ENTITY>	34	461	1175	1347	323	1670
<MOTION_EVENT>	16	330	588	751	183	934
<NONMOTION_EVENT>	14	66	301	321	60	381
<SPATIAL_SIGNAL>	39	216	550	653	152	805
<MOTION_SIGNAL>	17	260	365	508	134	642
<MOVELINK>	15	345	614	779	195	974
<QSLINK>	69	348	693	886	224	1110
<OLINK>	14	82	191	225	62	287

Taula 4.2: *SpaceBank* corpuseko kopuru edo estatistikak (Pustejovsky et al., 2015).

[10] http://alt.qcri.org/semeval2015/
[11] http://rideforclimate.com/blog/

Taulan ikusten den gisa, *SpaceBank* sortzean informazio gehien RFC corpusetik eskuratu zen eta gutxien, berriz, ANCtik. Etiketa kopuruari dagokionez hiru gauza azpimarra daitezkeela uste dugu: (1) <PLACE> dela etiketarik arruntena (2.089 agerpen), (2) corpusean gehien eta gutxien agertzen diren erlazio espazialak <QSLINK> eta <OLINK> direla (1.110 eta 287 agerpen), eta (3) mugimenduko gertaerak (<MOTION_EVENT>) mugimendukoak ez direnak (<NONMOTION_EVENT>) baino nabarmen arruntagoak direla (934 eta 381 agerpen). Kopuru hauek osagai espazial batzuen etiketatze prozesuaren zailtasunaren erakusle dira.

Anotazio prozesuari dagokionez (Pustejovsky et al., 2015), argitalpenean adierazten den bezala, anotazioa lau fasetan bereizi zela jakitea interesgarria da. Hau horrela egiteko arrazoia erlazio espazialen etiketek (<MOVELINK>, <QSLINK> eta <OLINK>) eskemako gainerako elementuen alderako dependentzia izan zen. Lehenbiziko fasean osagai espazialen, seinaleen eta gertaeren identifikazioa eta hauen atributuen esleipena egin zen, bakoitza zer motatakoa zen zehaztu gabe; haien mota edo etiketa bigarrenean finkatu zen. Hirugarren fasean, aldiz, aurreko faseko osagai, seinale eta gertaeren arteko erlazioak finkatu eta atributuak esleitu ziren, erlazio motak erabaki gabe. Azkenik, laugarrenean, erlazio horien motak, etiketen esleipenak alegia, burutu ziren.

Deskribatu berri dugun *SpaceBank* corpusaren anotazio prozesuan hainbat anotatzailek hartu zuten parte eta gutxienez horietako hiruk etiketatu zuten testuetako bakoitza. Puntu honetan azaldu behar da guztiek ingelesa zutela ama hizkuntza. Lortutako IAA neurria erlazio espazialei ez zegozkien elementuen identifikazioan 85 puntukoa izan zen, 91 puntukoa <MOVELINK> erlazioetan, eta 39 eta 33 puntukoa <OLINK> eta <QSLINK> etiketetan. Azkeneko hiru hauei dagozkien adostasunaren balioen arteko ezberdintasunak agerian uzten du <OLINK> eta <QSLINK> erlazioak ezartzeak duen zailtasuna.

4.1.3 Ebaluazio saioak

Testuetako espazio informazioaren etiketatze automatikoarekin zerikusia daukaten hiru ebaluazio saio antolatu dira gaurdaino: SpRL eskemaren araberako *SemEval-2012* eta *SemEval-2013* saioetakoak, eta *ISO-Space*en araberako *SemEval-2015*eko *SpaceEval* izenekoa. Jarraian hauetako bakoitzaren azalpena egiten dugu.

- ***SemEval-2012*** (Kordjamshidi et al., 2012): saio hau izan zen lehenbizikoa espazio informazioaren etiketatze automatikoaz arduratu zena. Honetan SpRL eskema

hartu zen oinarritako eta bertan parte hartu zuten sistemek egin beharrekoa hiru azpiatazatan banatu zen. Lehenbizikoan <TRAJECTOR> eta <LANDMARK> rol espazialak (hauek jokatzen dituzten argumentuak) eta <SPATIAL-INDICATOR> motako erlazioen adierazleak identifikatu behar ziren (A). Bigarrenean aurreko azpiatalean eskuratutako rolen eta adierazleen hirukoteek osatu <SR> erlazioak identifikatzea izan zen helburua (ikus 4.1.1 azpiataleko adibidea) (B). Azkenik, hirugarren azpiatazak aurreko pausoan finkaturiko erlazioen `type` atributua esleitzea zeukan arduratako (C). Kordjamshidi eta bestek (2012) adierazten dutenez, azkeneko hori lortu ahal izateko beharrezkoa da lehenago erlazioek SpRLn jasotzen dituzten `region`, `direction` eta `distance` atributuak zehaztea.

Ebaluazio saio honetan parte hartu zuen sistema bakarra (Roberts eta Harabagiu, 2012) argitalpenean aurkezten den `UTD-SpRL` izeneko tresna izan zen. Gainera, ebaluazio saioaren antolatzaileek haiek berek (Kordjamshidi et al., 2011) aurretik garatutako etiketatzaile batekin ebaluazio saioko *dataset*a baliaturik erdietsitako emaitzak ere aurkeztu zituzten (`KUL-SKIP-CHAIN-CRF`). Ebaluaziorako doitasun, estaldura eta F_1 neurri estandarrak erabili ziren saioko hiru azpiatazetan. Lehenengoan, A-n, *Relaxed* ebaluazio eskema erabili zen. Rolek eta adierazleek osatutako hirukoteak diren <SR> erlazioak ezartzean, B azpiatazan, ordea, hiru osagaiek ongi identifikatuta egon behar zuten (*Strict*). C azpiatazan, azkenekoz, Bn ongi identifikatzea lortutako erlazioen motaren ebaluazioa baizik ez zen egin, hau da, Cn erlazioen identifikazioa eta sailkapena, biak batean, ebaluatu ziren. Horretarako guztiz bat etorri behar zuten (*Strict*) iragarritako erlazioen motek eta anotatutakoenek. 4.3 taulan ebaluazio saio honetan lortu ziren emaitzarik onenak bildu ditugu.

Osagaia/Ataza		Eskema	Doitasuna	Estaldura	F_1
<TRAJECTOR>	**A**	R	78.2 (1)	64.6 (1)	70.7 (1)
<LANDMARK>			89.4 (1)	68 (1)	77.2 (1)
<SPATIAL-INDICATOR>			91.3 (2)	88.7 (2)	90 (2)
<SR>	**B**	S	61 (1)	54 (1)	57.3 (1)
<SR> + type	**C**	S	60.3 (1)	53.4 (1)	56.6 (1)

Taula 4.3: *SemEval-2012* saioko emaitzarik onenak [(1):(Roberts eta Harabagiu, 2012), (2):(Kordjamshidi et al., 2011)].

Taulan ikus daitekeen moduan <SPATIAL-INDICATOR> adierazleen identifikazioan ez, baina saioko gainerako azpiatazetan UTD-SpRL-SUPERVISED2 izan zen emaitzarik gorenak erdietsi zituen sistema. A-ko osagaiei erreparatuta, bestalde, ikus daiteke erlazio adierazleak detektatzean lortu balioak rol espazialen identifikazioan eskuratutakoak baino aise hobeak izan zirela (F_1 90, 70.7 eta 77.2 puntu). Gure ustez honen arrazoia honako hau da: <SPATIAL-INDICATOR> motako adierazleak token bakarreko preposizioak izaten dira normalean (*in*, *at*, *about*, *around...*). <TRAJECTOR> eta <LANDMARK> rolek jasotzen dituzten argumentuek, berriz, askotan token bat baino gehiago dituzte eta, ondorioz, zailagoa izaten da hauek osorik detektatzea. B-n eta C-n lortutako balioek (F_1 57.3 eta 56.6 puntu) erlazio espazialak finkatu eta kategorizatzeak daukan zailtasuna adierazten dutela uste dugu.

- ***SemEval-2013*** (Kolomiyets et al., 2013): saio honetan ere SpRL hartu zen oinarritzat, espazio informazioa etiketatu ahal izateko; ebaluazio saio hau izan zen eskema hori erabilita gaurdaino antolatutako azkenekoa. Aurrekoarekin zuen ezberdintasunik nabarmenena, *CPC* corpusaren erabileraz gainera, mugimenduaren (*motion*) tratamendu automatikoa izan zen. Horretarako, *SemEval-2012* saioko anotazio eskema aldatu eta SpRL barneko <MOTION-INDICATOR> eta <PATH> etiketa berriak erabili behar izan zirela adierazten dute Kolomiyets eta bestek (2013). Saioa A, B, C, D eta E erazagututako ondoko bost azpiatazatan banatu zen:
  - A, B: A_{SE12} eta B_{SE12} bezalakoak ($_{SE12}$ = *SemEval-2012*).
  - C, D: A_{SE12} eta B_{SE12} atazen parekoak baina <TRAJECTOR>, <LANDMARK> eta <SPATIAL-INDICATOR> etiketak ez ezik, berriki aipatutako <PATH> rol espaziala jokatzen zuten argumentuak eta <MOTION-INDICATOR> mugimendu erlazioen adierazleak ere etiketatu (C) eta erlazionatu (D) zituzten.
  - E: C_{SE12} atazaren parekoa baina <PATH> eta <MOTION-INDICATOR> ere hartzen zituzten <SR> hirukoteentzat.

*SemEval-2013*ko SpRL atazan UNITOR-HMM-TK sistemak (Bastianelli et al., 2013) baizik ez zuen parte hartu. Honek hiru exekuzioren emaitzak itzuli zituen. Horietako bi *IAPR TC-12* corpusaren gainean egin ziren konfigurazio ezberdinak erabilita eta beste exekuzioa, berriz, *CPC* corpusaren gainean. Aipatu beharra

daukagu sistema honek ez zituela emaitzak bost azpiatazetarako itzuli, lehenbiziko eta bigarren exekuzioetan A eta B azpiatazetan baizik ez baitzuen parte hartu. Azkenekoan, ordea, A-n eta C-n.

Ebaluaziorako erabilitako metrikak eta metodoak *SemEval-2012*ko berak izan ziren, ezberdintasun bakar batekin, erlazioen identifikazioan *Relaxed* eskema ere aplikatu zela (B eta D azpiatazak). Horrela, <SR> lotura bateko osagaietakoren bat (rol espaziala edo adierazlea) zuzen etiketatuta ez bazegoen ere, sistemak puntuak jasoko zituen gainerakoengatik, ongi identifikatutakoengatik alegia. 4.4 eta 4.5 tauletan aurkezten ditugu UNITOR-HMM-TK etiketatzaileak *IAPR TC-12* eta *CPC* corpusetan lortutako emaitzarik onenak.

Osagaia/Ataza		Eskema	Doitasuna	Estaldura	F_1
<TRAJECTOR>	**A**	R	68.4	68.1	68.2
<LANDMARK>			74.1	83.5	78.5
<SPATIAL-INDICATOR>			96.7	88.9	92.6
<SR>	**B**	S	43.1	30.6	35.8
		R	55.1	39.1	45.8

Taula 4.4: *SemEval-2013* saioko emaitzarik onenak *IAPR TC-12* corpusean.

Osagaia/Ataza		Eskema	Doitasuna	Estaldura	F_1
<TRAJECTOR>	**A**	R	56.5	31.7	40.6
<LANDMARK>			66.1	47.6	55.4
<SPATIAL-INDICATOR>			61.2	48.1	53.8
<TRAJECTOR>	**C**	R	56.5	31.7	40.6
<LANDMARK>			66.2	47.6	55.4
<PATH>			77.5	29.5	42.7
<SPATIAL-INDICATOR>			60.9	47.9	53.6
<MOTION-INDICATOR>			89.2	29.4	44.3

Taula 4.5: *SemEval-2013* saioko emaitzarik onenak *CPC* corpusean.

Tauletan aurkezten ditugun balioak (Kolomiyets et al., 2013) direla eta, argitalpenean esaten den moduan *SemEval-2013*ko emaitzak ezin dira *SemEval-2012* saiokoekin zuzenean alderatu. Hala ere, badago aukera bi saioetako balioen arteko amankomuneko zenbait joera detektatzeko, esate baterako bietan <LANDMARK> eta <SPATIAL-INDICATOR> etiketentzat emaitzak hobeak dira <TRAJECTOR>

rol espazialarentzat baino. Gainera, bi saioetan <SPATIAL-INDICATOR> etiketak erdietsi zituen emaitzarik onenak (*IAPR TC-12* corpusean).

Bi *dataset*en balioak alderatzeari dagokionez (A azpiataza), lehen begi kolpean ageri da UNITOR-HMM-TK sistemak izan duen eraginkortasuna *CPC* corpusean *IAPR TC-12*n baino nabarmen apalagoa izan dela. Izan ere, <TRAJECTOR> eta <LANDMARK> roletan dagoen F_1 neurriaren aldea 28 eta 23 puntukoa da, eta <SPATIAL-INDICATOR> adierazlearenean, aldiz, ia 40 puntukoa. Kolomiyets eta bestek (2013) adierazten dutenez, ezberdintasun hau corpus bateko eta besteko testuen konplexutasunean dagoen aldearen eragina izan daiteke. Oro har, *CPC*koak beste corpusekoak baino zehatzagoak dira.

- ***SemEval-2015*** (Pustejovsky et al., 2015): *SemEval-2015* barnean kokatutako *SpaceEval* saioa izan zen *ISO-Space* eskema baliaturik antolaturiko lehenbizikoa. Lan honetan aurkezten den *X-Space* espazio informazioaren etiketatze automatikorako tresna haren barnean garatu eta ebaluatu genuen (Salaberri et al., 2015b).

  Saio honetan parte hartu zuten sistemen entrenamendurako eta ebaluaziorako *SpaceBank* corpusa hartu zen. Tresna hauen eginbeharra banatzeko hiru konfigurazio ezberdin eta hainbat azpiataza erabili ziren: sistemek, lehen konfigurazioan (1), inongo anotaziorik gabeko testu soila jaso zuten sarreratako. Bigarrenean (2), aldiz, atributuak esleitu gabe zeuzkaten eskuz identifikatutako osagai espazialak jaso zituzten (<PATH>, <PLACE>, <SPATIAL_ENTITY>, <MOTION_EVENT>, <NONMOTION_EVENT>, <SPATIAL_SIGNAL> eta <MOTION_SIGNAL>). Azkenik, hirugarren konfigurazioan (3), eskuz identifikatutako osagaiak ez ezik, hauek hartzen zituzten atributuen balioak ere atxiki zitzaizkien. Konfigurazio hauek kontuan edukita, hurrengo azpiatazak izan ziren etiketatzaileek egin beharrekoa:

  - **1_Osag_ID**: Osagai espazialen identifikazioa.
  - **1_Osag_SAILK**: Osagai espazialen sailkapena, etiketaren zehaztapena.
  - **1_Osag_ATR**: Motaren araberako osagai espazialen atributuen esleipena.
  - **1_Erla_ID**: Erlazio espazialen identifikazioa.
  - **1_Erla_ATR**: Erlazio espazialen atributuen esleipena, rol espazialena.
  - **2_Osag_ATR**: Motaren araberako osagai espazialen atributuen esleipena.

- **2_Erla_ID**: Erlazio espazialen identifikazioa.
- **2_Erla_ATR**: Erlazio espazialen atributuen esleipena, rol espazialena.
- **3_Erla_ID**: Erlazio espazialen identifikazioa.
- **3_Erla_ATR**: Erlazio espazialen atributuen esleipena, rol espazialena.

SpaceEval saioan BRANDEIS-CRF (Pustejovsky et al., 2015), HRIJP-CRF-VW (Nichols eta Botros, 2015), UTD (D'Souza eta Ng, 2015) eta hemen azalduko dugun *X-Space* (Salaberri et al., 2015b) izeneko lau sistemek hartu zuten parte. Atazaren antolatzaileek, gainera, *baseline* sistema (Pustejovsky et al., 2015) batek ebaluazio saioko corpusa baliatuz itzulitako emaitzak ere aurkeztu zituzten. Aintzat hartu behar da aipatutako bostetatik bik bakarrik itzuli zituztela emaitzak konfigurazio eta azpiataza guztietarako, *baseline* eta *X-Space* sistemek hain zuzen ere.

Ebaluaziorako metrikei dagokienez, doitasun, estaldura, F_1 eta *accuracy* neurri estandarrak aplikatu ziren. Pustejovskyk eta bestek (2015) diotenez, azpiataza bakoitzeko doitasuna eta estaldura kalkulatu ahal izateko, hauetan landutako etiketa bakoitzean erdietsi emaitzen batezbesteko aritmetikoak konputatu ziren. 1_Osag_IDn esate baterako, osagai espazial bakoitza identifikatzean lortutako doitasunen eta estalduren batezbestekoa kalkulatu zen, azpiataza osoarenak zein ziren jakin ahal izateko. F_1 neurriak konputatzeko, berriz, aurretik kalkulatutako batezbesteko doitasun eta estaldura hauen batezbesteko harmonikoa egin zen. *Accuracy* neurria, azkenik, sistemek zuzen identifikatutako etiketen edo ongi esleitutako atributuen kopurua, alegia eskuzko anotazioetako etiketa edo atributu guztien kopuruarekin zatituta kalkulatu zen. 4.6 taulan bildu ditugu *SpaceEval* saioko azpiatazetan iritsi ziren emaitzarik onenak.

Taula horretan ikusten den gisa, ebaluazio saioko hamar azpiatazatatik lautan baliorik altuenak eskuratu zituena *SpaceEval* atazaren antolatzaileek aurkeztutako *baseline* sistema izan zen (2_Erla_ID, 2_Erla_ATR, 3_Erla_ID eta 3_Erla_ATR). Beste lautan, aldiz, tesi lan honetan garatu dugun *X-Space* (1_Osag_ATR, 1_Erla_ID, 1_Erla_ATR eta 2_Osag_ATR) tresna, eta gainerako bietan *Brandeis-CRF* etiketatzailea (1_Osag_ID eta 2_Erla_ID).

Ataza		Doitasuna	Estaldura	F_1	Accuracy
1	**Osag_ID**	85 (2)	80 (2)	83 (2)	89 (2)
	Osag_SAILK	78 (2)	76 (2)	77 (2)	92 (2)
	Osag_ATR	18 (4)	15 (4)	16 (4)	30 (4)
	Erla_ID	54 (4)	51 (4)	53 (4)	55 (4)
	Erla_ATR	6 (4)	5 (4)	5 (4)	25 (4)
2	**Osag_ATR**	26 (4)	33 (4)	29 (4)	63 (4)
	Erla_ID	79 (1)	58 (1)	67 (1)	90 (1)
	Erla_ATR	19 (1)	20 (1)	19 (1)	66 (1)
3	**Erla_ID**	86 (1)	84 (1)	85 (1)	98 (1)
	Erla_ATR	26 (1)	26 (1)	26 (1)	79 (1)

Taula 4.6: *SemEval-2015* saioko emaitzarik onenak [(1):(Pustejovsky et al., 2015), (2):(Pustejovsky et al., 2015), (3):(Nichols eta Botros, 2015), (4):(Salaberri et al., 2015b), (5):(D'Souza eta Ng, 2015)].

Saio honetan lortu ziren emaitzek argi uzten dute osagaiak identifikatzea eta sailkatzea (`1_Osag_ID` eta `1_Osag_SAILK`) eraginkortasun onargarria erdiesten duten azpiatazak direla. Izan ere, hauetan lortutako F_1 neurririk onenak 83 eta 77 puntukoak dira, hurrenez hurren. Osagaien motaren araberako atributuen esleipenari dagokionean, aldiz, emaitzak apalagoak dira: `1_Osag_ATR`n 16 puntukoa da F_1 baliorik esanguratsuena, eta `2_Osag_ATR`n, ordea, 29koa. Erlazioen ezarpenari dagozkion azpiatazetako balioak aztertzean, hauek nahiko onak direla ikusten da; `1_Erla_ID`n 57 puntuko F_1 lortzen da, eta `2_Erla_ID`n eta `3_Erla_ID`n 67 eta 85 puntukoak. Neurri hauek, gainera, argi uzten dute erlazioen identifikazioak aurreko azpiatazekin duen dependentzia; eskuz anotatutako osagai espazialekin hobeak dira emaitzak. Erlazio espazialen atributuak, rol espazialak etiketatzeari dagokionez, ikus dezakegu eginbehar konplexua dela, hauetaz arduratzen diren azpiatazak baitira denetan emaitzarik apalenak lortzen dituztenak: `1_Erla_ATR` azpiatazako F_1 neurria 5 da, `2_Erla_ATR`koa 19 eta `3_Erla_ATR`koa 26.

4.2 Espazioa etiketatzeko arkitektura: *X-Space*

*X-Space*n egitura *ISO-Space* eskemaren araberako etiketatzaileek etorkizunean izan dezaketen arkitekturaren proposamena dela ulertzen dugu. Azpiatal honetan *X-Space* tresna-

ren garapenerako eta ebaluaziorako hartutako *SpaceBank* corpuseko fitxategien egitura aurkezten da lehenik, tresnaren ebaluazioan baliatu metrikak ere, adibideen laguntzaz, azaldu bidenabar. Gero, etiketatzailearen garapen prozesua eta egitura bera deskribatu, *SpaceEval* saioan iritsi emaitzak aurkeztu, eta hauen analisia egiten da. Bukatzeko, tesiko bigarren hipotesia aztertu ahal izateko SRLk ataza honetan duen eragina neurtu dugu.

4.2.1 Informazioaren adierazpidea

Corpusak biltzen dituen testuak eta anotazioak *ISO-Space*k ezartzen duenaren arabera daude adierazita. Adierazpen honen adibidea 4.3 irudian ikus daiteke.

```
<SpaceEvalTask>
   <TEXT>
      Jerry is in Grosvenor Square, he is
      sitting on a bench next to the statue.
   </TEXT>
   <TAGS>
      <!--SPATIAL ELEMENTS-->
      <PLACE id="pl0" text="Grosvenor Square"/>
      <PLACE id="pl1" text="bench"/>
      <PLACE id="pl2" text="statue"/>
      <SPATIAL_ENTITY id="se0" text="Jerry"/>
      <SPATIAL_ENTITY id="se1" text="he"/>
      <SPATIAL_SIGNAL id="s0" text="in"
                              semantic_type="TOPOLOGICAL"/>
      <SPATIAL_SIGNAL id="s1" text="on"
                              semantic_type="TOPOLOGICAL"/>
      <SPATIAL_SIGNAL id="s2" text="next to"
                              semantic_type="DIR_TOP"/>
      <NONMOTION_EVENT id="e0" text="sitting"/>
      <!--SPATIAL RELATIONS-->
      <QSLINK id="qsl0"
              fromText="Jerry" toText="Grosvenor Square"
              trajector="se0" landmark="pl0" trigger="s0"/>
      <QSLINK id="qsl1"
              fromText="sitting" toText="bench"
              trajector="e0" landmark="pl1" trigger="s1"/>
      <OLINK id="ol0" fromText="bench" toText="statue"
             trajector="pl1" landmark="pl2" trigger="s2"/>
   </TAGS>
</SpaceEvalTask>
```

Irudia 4.3: *ISO-Space* formatuaren adibidea.

Adibidean *Jerry is in Grosvenor Square, he is sitting next to the statue* esaldiari dagokion *ISO-Space* fitxategia erakusten da. Honetan ikusten den eran, lehenik dokumentuko

testua <TEXT> barnean biltzen da, eta gero, anotazioak zerrendatzen dira, <TAGS> etiketen artean. Adibideko esaldian hiru leku identifikatu dira: *Grosvenor Square*, *bench* eta *statue*. Gainera, bi entitate espazial (*Jerry* eta *he*), gertaera estatiko baten buru lexikala (*sitting*) eta hiru espazio seinale ere detektatu ahal izan dira (*in*, *on* eta *next to*). Seinaleetako biren izaera semantikoa (semantic_type) topologikoa da (TOPOLOGICAL) eta hirugarrenarena, berriz, topologikoa eta norabidekoa aldi berean (DIR_TOP).

Hiru seinale hauetako bakoitzak (izaeraren arabera) erlazio espazial bat adierazten edo abiarazten du: izaera topologikoko biek <QSLINK> motako qsl0 eta qsl1 erlazioak adierazten ditu, eta aldi berean topologikoa eta norabidekoa den seinaleak <OLINK> motako ol0 erlazioa. qsl0k *Jerry* entitate espaziala *Grosvenor Square* leku izenarekin lotzen du, *in* preposizioaren bitartez (<SPATIAL_SIGNAL>), qsl1 erlazioak *sitting* gertaera estatikoa *bench* tokiarekin estekatzen du, *on* baliaturik, eta ol0k, azkenik, *bench* eta *statue next to* seinalearekin.

4.2.2 Ebaluaziorako metrikak

ISO-Space eskema jarraituta informazio espaziala etiketatzen duten sistemen ebaluaziorako aplikatutako metrikak 4.1.3 azpiatalean *SpaceEval* saioa aurkeztean aipatu ditugu. Izan ere, saio hau eta bertan erabilitako neurriak dira mota honetako tresnen ebaluaziorako oraindainoko erreferentzia bakarrak eta, beraz, *X-Space* tresnarako guk hartu ditugunak. Azaldu dugun moduan, sistemek burutzen dituzten azpiatazen doitasunak eta estaldurak kalkulatu ahal izateko haietan landutako etiketetan erdietsi doitasunen eta estalduren batezbesteko aritmetikoak konputatzen dira, eta F_1 neurriak lortzeko, berriz, aurreko hauen batezbesteko harmonikoak. *Accuracy* neurria tresnek zuzen identifikatutako etiketen kopurua eskuzko anotazioetako etiketa guztien kopuruarekin zatituta kalkulatzen da.

Rol semantikoen eta denboraren etiketatze automatikoaz arduratzen diren sistemen ebaluaziorako aplikatutako metrikentzat egin dugun moduan, hemen ere espaziorako erabili diren metriken kalkuluen adibideak ematen ditugu, hauen funtzionamendua hobeki azaltzeko asmoz. 4.2.1 azpiataleko adibidea hartzen dugu horretarako; bertan, hiru leku, hiru espazio seinale, bi entitate espazial eta gertaera estatiko bat identifikatu dira (ikus 4.3 irudia). Orain, hala ere, eman dezagun sistema automatiko batek etiketatu duela adibidea eta hau dela honek itzulitakoa:

Jerry is in Grosvenor Square, he is sitting in a bench next to the statue.

↓

Jerry is **<SPATIAL_SIGNAL id="s0">***in***</SPATIAL_SIGNAL>** **<PLACE id="pl0">***Grosvenor***</PLACE>** *Square,* **<SPATIAL_ENTITY id="se0">***he***</SPATIAL_ENTITY>** *is* **<NONMOTION_EVENT id="e0">***sitting***</NONMOTION_EVENT>** *on* **<SPATIAL_SIGNAL id="s1">***a***</SPATIAL_SIGNAL>** **<PLACE id="pl1">***bench***</PLACE>** **<SPATIAL_SIGNAL id="s2">***next***</SPATIAL_SIGNAL>** **<SPATIAL_SIGNAL id="s3">***to***</SPATIAL_SIGNAL>** *the statue.*

Hainbat ezberdintasun ikus daitezke eskuz etiketatutako adibidearekin alderatzean: *Jerry* entitatea, *on* seinalea eta *statue* lekua ez dira identifikatu; gainera, *Grosvenor Square* tokiko *Square* tokena ez da detektatu, eta *next to* aparteko bi seinale balira bezala etiketatu da. Azkenik, *"a"* seinaletzat hartu da, eta hau ere ez da zuzena. Deskribatu egoeran jarraian azaltzen den moduan kalkulatuko lirateke osagai espazialen identifikazioari eta sailkapenari dagozkion doitasuna, estaldura, F_1 neurria eta *accuracy* neurria.

Ebaluazioa token mailan

$$P_i = P(< \texttt{ETIK}_\texttt{i} >) = \frac{TP_{tok}(< \texttt{ETIK}_\texttt{i} >)}{TP_{tok}(< \texttt{ETIK}_\texttt{i} >) + FP_{tok}(< \texttt{ETIK}_\texttt{i} >)}$$

$$R_i = R(< \texttt{ETIK}_\texttt{i} >) = \frac{TP_{tok}(< \texttt{ETIK}_\texttt{i} >)}{TP_{tok}(< \texttt{ETIK}_\texttt{i} >) + FN_{tok}(< \texttt{ETIK}_\texttt{i} >)}$$

$$P = \frac{P_1 + ... + P_n}{n}$$

$$R = \frac{R_1 + ... + R_n}{n}$$

$$F_1 = 2 * \frac{P * R}{P + R}$$

$$Acc = \frac{TP_{tok}(< \texttt{ETIK}_\texttt{1} >) + ... + TP_{tok}(< \texttt{ETIK}_\texttt{n} >)}{\#\{< \texttt{ETIK}_\texttt{1} >\} + ... + \#\{< \texttt{ETIK}_\texttt{n} >\}}$$

Osagai espazialen identifikazioaren eta sailkapenaren ebaluazioa

$$P_1 = P(< \texttt{PLACE} >) = \frac{\#\{Grosvenor, bench\}}{\#\{Grosvenor, bench\} + \#\{\emptyset\}} = 1$$

$$R_1 = R(< \texttt{PLACE} >) = \frac{\#\{Grosvenor, bench\}}{\#\{Grosvenor, bench\} + \#\{Square, statue\}} = 0.5$$

$$P_2 = P(< \texttt{SPATIAL\_SIGNAL} >) = \frac{\#\{in, next, to\}}{\#\{in, next, to\} + \#\{a\}} = 0.75$$

$$R_2 = R(< \texttt{SPATIAL\_SIGNAL} >) = \frac{\#\{in, next, to\}}{\#\{in, next, to\} + \#\{on\}} = 0.75$$

$$P_3 = P(< \texttt{SPATIAL\_ENTITY} >) = \frac{\#\{he\}}{\#\{he\} + \#\{\emptyset\}} = 1$$

$$R_3 = R(< \texttt{SPATIAL\_ENTITY} >) = \frac{\#\{he\}}{\#\{he\} + \#\{Jerry\}} = 0.5$$

$$P_4 = P(< \texttt{NONMOTION\_EVENT} >) = \frac{\#\{sitting\}}{\#\{sitting\} + \#\{\emptyset\}} = 1$$

$$R_4 = R(< \texttt{NONMOTION\_EVENT} >) = \frac{\#\{sitting\}}{\#\{sitting\} + \#\{\emptyset\}} = 1$$

$$P = \frac{P_1 + P_2 + P_3 + P_4}{4} = 0.94$$

$$R = \frac{R_1 + R_2 + R_3 + R_4}{4} = 0.69$$

$$F_1 = 2 * \frac{P * R}{P + R} = 0.79$$

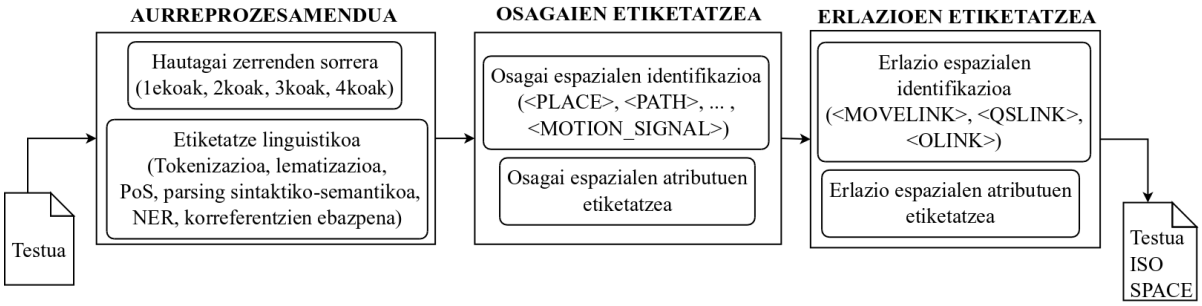
$$Acc = \frac{2 + 3 + 1 + 1}{4 + 4 + 2 + 1} = 0.64$$

Ikusten den gisan, 94 eta 69 puntuko doitasuna eta estaldura kalkulatu dira adibideko sistemarako. Hauek erabilita, gainera, 79 puntuko F_1 neurria erdietsi da. *Accuracy*a, aldiz, % 64koa izan da. Adibide honek argi uzten du *relaxed* dela osagai espazialen identifikazioaren eta sailkapenaren ebaluazioa egiteko jarraitzen den eskema (*token-mailakoa*). Izan ere, emaitzak kalkulatzeko tenorean ongi eta gaizki identifikatutako tokenak hartzen dira kontuan, ez osagaiak berak (*osagai-mailakoa*). Erlazio espazialak ebaluatzeko, ordea, *SemEval-2013* saioko teknika bera erabiltzen da.

4.2.3 *X-Space* etiketatzailearen garapena

Azpiatal honetan *X-Space* deskribatzen dugu. Segidan zehaztasun osoz azalduko dugun gisan, sistema hau berritzailea da, besteak beste, osagai espazialen identifikaziorako eta sailkapenerako erabiltzen duen metodologiarengatik.

X-Space sistemaren arkitektura hiru pauso nagusitan banatzen den *pipeline* motako egitura da (ikus 4.4 irudian). Honekin adierazi nahi da testua era sekuentzialean prozesatzen dela, eta bigarren eta hirugarren urratsek lehenbizikoaren eta bigarrenaren irteerak jasotzen dituztela sarreratako; lehenbizikoak, aldiz, prozesatu beharreko testua jasotzen du (testu soila). Hiru pauso hauetan burutzen diren azpiatazak hurrengoak dira: sarrerako testuaren aurreprozesamendua, osagai espazialen etiketatzea eta erlazioen ezarpena. Bigarren eta hirugarren urratsetan osagaien eta erlazioen identifikazioa egiteaz gainera, hauen atributuen etiketatzea ere egiten da, erlazioen kasuan rol espazialena ere bai.



Irudia 4.4: *X-Space* sistemaren arkitektura.

A. Aurreprozesamendua

Lehenbiziko urrats honetan sarrera moduan jasotako testuak bi eratara aurreprozesatzen dira: osagai espazialak izateko hautagai zerrendak sortzen dira batetik, eta testuak linguistikoki etiketatzen dira bestetik. Lehenengo aurreprozesuko hautagai zerrendak sortu ahal izateko sarrerako testuetako tokenak banakako, binakako, hirunakako eta launakako leihoetan biltzen dira, hurrengo adibideak erakusten duen bezala:

Police have been fining cabbies in Madrid.

↓

Police	Police have	Police have been	Police have been fining
have	have been	have been fining	have been fining cabbies
been	been fining	been fining cabbies	been fining cabbies in
fining	fining cabbies	fining cabbies in	fining cabbies in Madrid
cabbies	cabbies in	cabbies in Madrid	cabbies in Madrid .
in	in Madrid	in Madrid .	-
Madrid	Madrid .	-	-
.	-	-	-

Zerrenda hauetako hautagaien artean identifikatzen ditu *X-Space* sistemak osagai espazialak, adibidean *Madrid* leku izena esate baterako. Gehienez ere lau tokeneko osagai-hautagaiak sortzeko erabakia etiketatzailearen garapenerako erabilitako *SpaceBank* corpusean oinarrituta hartu dugu, bertan anotaturiko osagaietan lau token baino gehiagokorik ez baitago, eta lau tokenekoak berak ez baitira batere arruntak.

Aurreprozesamenduaren beste eginbeharra, testuak linguistikoki etiketatzea da. Hain zuzen ere, egiten dena testuen tokenizazioa, lematizazioa, tokenen kategoria gramatikalen etiketatzea, *parsing* sintaktiko-semantikoa, entitate izendunen identifikazioa eta korreferentzia-ebazpena da. Maila linguistiko desberdinetako etiketatze hau baliagarria da *X-Space* tresnaren bigarren eta hirugarren urratsetako atazak burutzeko gai diren sailkatzaileak eraiki ahal izateko.

Entitate izendunen detekziorako (*Apache*[12] software fundazioaren) *OpenNLP* (Baldridge, 2014) tresna erabili da. Honek lekuak, erakundeak eta pertsona izenak identifika-

[12] https://www.apache.org/

tzeko gaitasuna du, eta eskaintzen duen informazioa baliagarria izan da *X-Space* barnean leku izenak (<PLACE> eta <PATH>) detektatzen laguntzeko. Jakina den bezala, zaila izaten baita, kasu batzuetan, erakunde eta pertsona izenak leku izenetatik bereiztea. Korreferentzien ebazpenerako, ordea, *Standford* unibertsitateko[13] *CoreNLP* sistema (Manning et al., 2014) hartu da. Azkenik, analisi sintaktiko eta semantikorako *ClearNLP* (Choi, 2012) dependentzia *parser*a baliatu da. Tresna horiek erabiltzea erabaki dugu lortzen dituzten emaitza onengatik.

B. Osagai espazialak (identifikazioa + atributuen sailkapena)

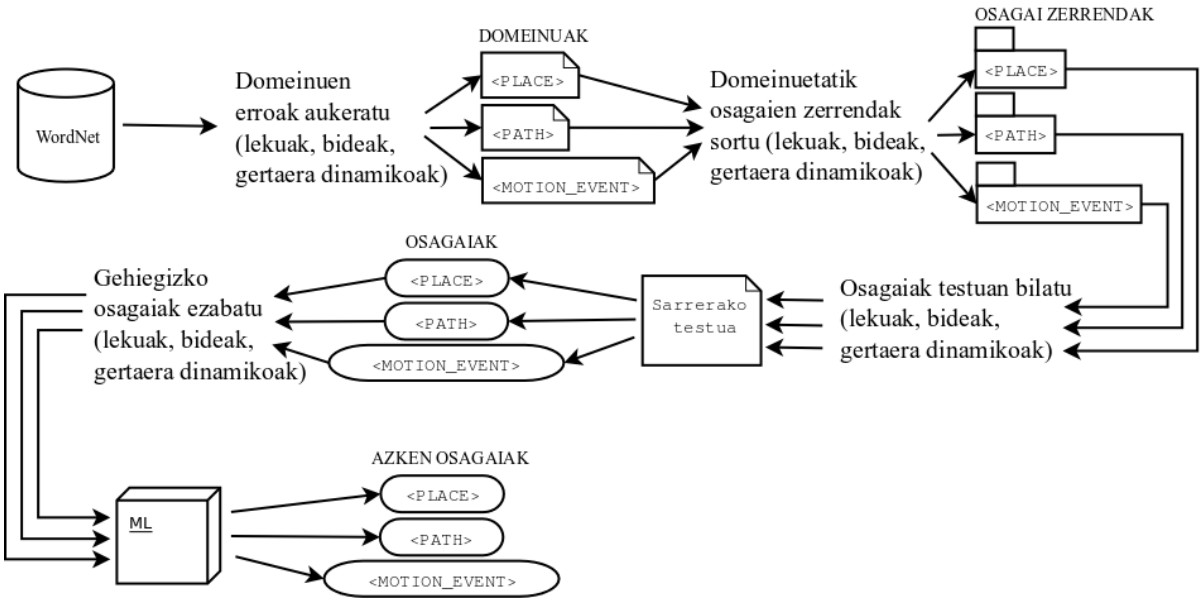
X-Space sistemaren bigarren urratsean osagai espazialak etiketatzen dira; etiketatze honetan, osagaiak identifikatzea eta hauen atributuak esleitzea da helburua.

Zazpi dira osagai motak: lekuak (<PLACE>), bideak (<PATH>), espazio entitateak (<SPATIAL_ENTITY>), gertaera dinamikoak (<MOTION_EVENT>), gertaera estatikoak (<NONMOTION_EVENT>), seinale espazialak (<SPATIAL_SIGNAL>) eta mugimendu seinaleak (<MOTION_SIGNAL>). Osagai hauek identifikatu ahal izateko motaren araberako hurbilpenak aplikatzen dira; aipagarriena lekuetarako, bideetarako eta gertaera dinamikoetarako erabiltzen den *WordNet* baliabidean oinarritutako hurbilpena da. Jarraian deskribatzen dugu *WordNet* erabiltzen duen metodologia zertan datzan eta gainerako etiketen identifikazioa *X-Space*n nola burutzen den.

B.1. Lekuak, bideak eta gertaera dinamikoak

Arestian esan dugun bezala, *X-Space* sisteman sarreratako jasotzen diren testuetan lekuak, bideak eta gertaera dinamikoak etiketatzeko *WordNet* datu-base lexikala erabiltzen da. Azpiatal honetan hau nola egiten den zehaztuko dugu pausoz pauso. Prozesu osoa 4.5 irudian ikus daiteke. Bertan antzeman daitekeen moduan, hiru pausotan banatzen da prozesua: *WordNet*en oinarrituta lekuen, bideen eta mugimenduzko gertaeren *domeinuak* finkatzen dira lehenik, gero, hauetatik *osagai zerrendak* deitu ditugunak eratzen dira testuan aipatu osagaiak bilatu ahal izateko. Azkenik, eta aurreko pausoan osagai gehiegi identifikatzen direla kontuan edukita, ikasketa automatikoko modulu bat erabiltzen da osagai kopuru horiek murrizteko.

[13] https://www.stanford.edu/



Irudia 4.5: *X-Space*n lekuak, bideak eta gertaera dinamikoak etiketatzeko arkitektura.

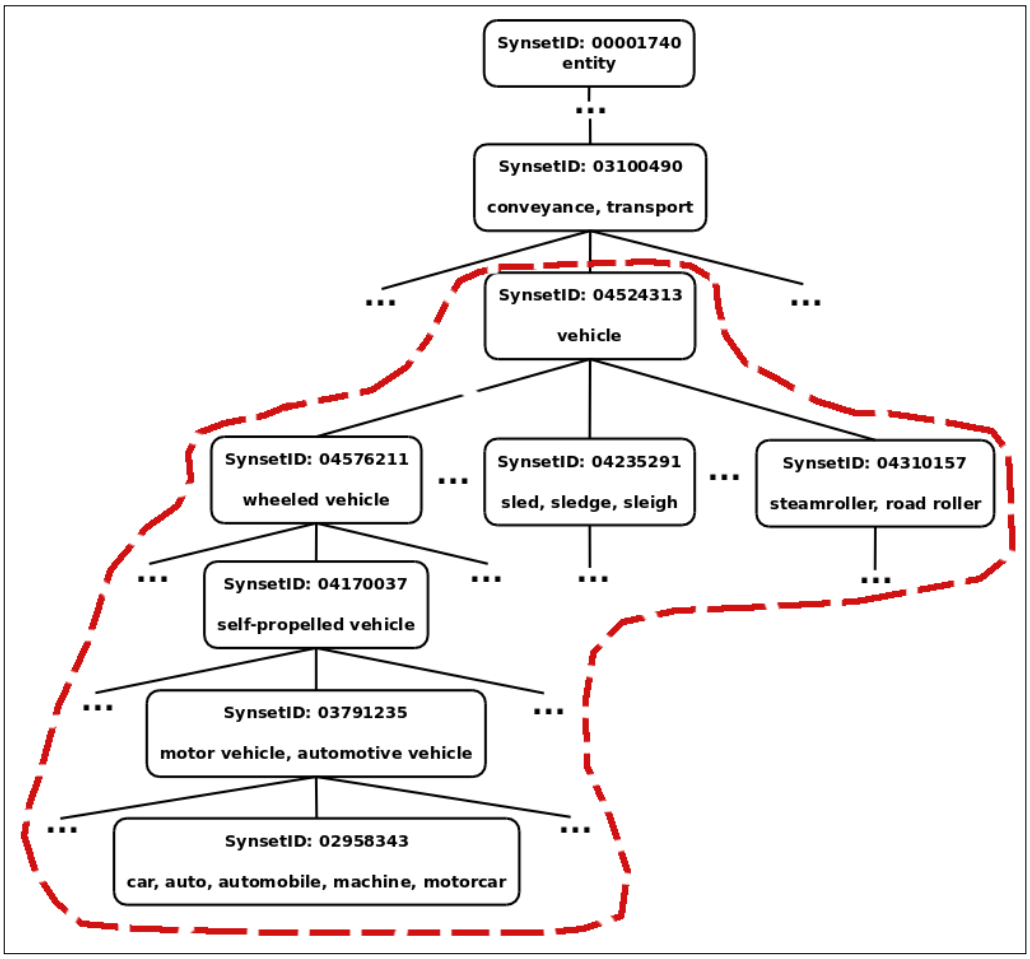
B.1.1. Zer da *WordNet* eta nola dago antolatuta

WordNet (Fellbaum, 1998) ingeleserako eskuz sortutako datu-base lexikala da. Honetan, izenak, aditzak, adjektiboak eta adberbioak *synset* deituriko sinonimo multzoetan daude bilduta; *car* izenari (*autoa*) dagokion *synset*ean, esate baterako, honen sinonimoak diren *auto*, *automobile*, *machine* eta *motorcar* izenak ere biltzen dira[14]. Sinonimo multzo hauen artean, bestalde, hainbat erlazio lexiko edota semantiko zehazten dira[15], besteak beste, hiperonimia, hiponimia, meronimia, holonimia eta antonimia erlazioak. Aipatutako adibidearen hiperonimoa den *synset*ak, esate baterako, *motor vehicle* eta *automotive vehicle* izenak biltzen ditu. Adibide honetan oinarrituta antzeman daiteke, gainera, *WordNet* datu-basearen sinonimo multzoen arteko hiperonimia eta hiponimia erlazioek zuhaitz moduko egitura osatzen dutela. Izan ere, edozein multzotatik abiatuta honen hiperonimoak diren multzoak era iteratiboan ibiltzen baldin badira, *WordNet*en (izenen) erroa den *synset*eraino iritsi arte egingo da gora zuhaitzean. *WordNet*eko izenen erroan[16] *entity* dago, izen guztiak (hiponimiaz) barneratzeko gai den kontzeptu zabala.

[14] *WordNet*en `3.0` bertsioan `02958343` identifikadorea daukan *synset*a.
[15] *Synset*en barnekoa sinonimia erlazioa da.
[16] *WordNet*en `3.0` bertsioan `00001740` identifikadorea daukan *synset*a.

B.1.2. Zer dira *domeinuak* eta nola mugatzen ahal dira *WordNet*ekin

Azaldu berri dugunetik ondoriozta daiteke *WordNet*en zuhaitz egiturak *domeinuak* (osatzen dituzten hitzak) zehaztu ahal izateko balio dezakeela. Aurreko adibideko *car* izenaren ildoa jarraituta aukera dago, esate baterako, *vehicles* (ibilgailuak) deritzan domeinua finkatzen saiatzeko. Horretarako beharrezkoa da ibilgailu guztiak barneratzeko bezain zabala den kontzeptuari dagokion *synset*a (edo *synset*ak) identifikatu eta honen (edo hauen) azpizuhaitza osatzen duten sinonimo multzoetako hitz guztiak hartzea. Kontuan eduki behar da, gainera, identifikatzen den erroko kontzeptua zabalegia baldin bada domeinuz kanpoko izenak ere barneratuko direla; hori da domeinuak finkatzeko erronka. 4.6 irudiak era grafikoan erakusten du prozesua.



Irudia 4.6: *Vehicles* domeinuaren mugatzea *WordNet* datu-basea erabilita.

B.1.3. Zergatik ez da erabili teknika hau gainerako osagaiak etiketatzeko

Domeinuak mugatzeko teknika lekuak, bideak eta gertaera dinamikoen buru lexikalak identifikatzeko baizik ez erabiltzea erabaki dugu, ez gainerako etiketetarako. Hasiera batean zazpi osagai espazial motak metodo hau baliaturik detekta daitezkeela pentsa badaiteke ere, gainerako osagaien kasuan zailtasunak daudela ikusi dugu. Entitate espazialek, esate baterako, gertaera dinamiko edo mugimenduzkoen argumentuak markatzen dituztela adierazi dugu, eta honek esan nahi du entitate hauen izaera gramatikala oso ezberdina izango dela *predikatu-argumentu* egitura batetik bestera; ez du zentzurik, hortaz, *WordNet*en entitate espazialen domeinua mugatzen saiatzeak. Mugimendu seinaleen eta seinale espazialen kasuan, aldiz, kontuan eduki behar da seinale hauek preposizioak izaten direla eta, gorago aipatu dugun gisan, *WordNet* datu-baseak izenak, aditzak, adjektiboak eta adberbioak baizik ez dituela barneratzen. Bukatzeko, eta gertaera estatikoen buru lexikalentzat, dinamikoentzat egin den bezala, *domeinua* finkatzen saiatzeak berez zentzua baldin badauka ere, ezin izan dugu hau zehaztu, ez baitugu datu-basean honi dagokion erroa zuzen identifikatzea lortu.

B.1.4. Zein izan dira domeinuak finkatzeko aukeratu diren erroak

WordNet erabilita domeinuak finkatu ahal izateko domeinu hauen erroak diren *synset*ak aukeratu ditugu. Aukeraketa honetatik lekuen, bideen eta gertaera dinamikoen zerrendak sortu dira. Gure kasuan eskuz identifikatu ditugu domeinuen erroak. Jarraian aurkezten dira osagai bakoitzerako hartutakoak.

- **Lekuak**: "*topographic point, place spot*" (08664443-N), "*place, property*" (08513718-N), "*position, place*" (08621598-N), "*location*" (00027167-N), "*state, nation, country, land, commonwealth, res publica, body politic*" (08168978-N), "*country, state, land*" (08544813-N), "*country, rural area*" (08644722-N) eta "*area, country*" (08497294-N).

- **Bideak**: "*path*" (03899328-N), "*path, route, itinerary*" (08616311-N), "*path, track, course*" (09387222-N) eta "*way*" (04564698-N).

- **Gertaera dinamikoak**: "*move, locomote, travel, go*" (01835496-V) eta "*to be*" (02604760-V).

B.1.5. Nola sortzen dira domeinuak aukeratutako erroetatik

Lekuen eta bideen kasuan erro hauek dauzkaten zuhaitzetatik zuzenean sortzen dira aipatu zerrendak zuhaitzeko *synset*ak (ikus 4.6 irudia) osatzen dituzten hitzak hartuta. Zerrenda hauek, gainera, *X-Space* garatzeko erabili den *SpaceBank* corpusaren entrenamendurako zatian (*train*) anotatu ziren (eta *WordNet*en ez dauden) leku eta bide izenekin osatu ditugu. Osotara, corpusetik eta datu-base lexikaletik hartuta, 5.572 leku eta 664 bide dauzkaten zerrendak osatu ahal izan ditugu.

Gertaera dinamikoendako, berriz, prozesua ez da zuzenekoa izan. Hauentzat beharrezkoa izan da *WordNet*eko domeinuan eskuratutako gertaera dinamikoen adieren desanbiguazioa egitea. Izan ere, tesi lanean zehar azaldu dugun moduan, gertaerak predikatuek deskribatzen dituzte, eta jakina denez predikatuek hainbat adiera izan ditzakete. Argitu behar da, bestalde, erabilitako corpusean predikatu baten adiera bat dinamikoa izan daitekeela eta beste bat aldiz ez; *run* aditza, esate baterako, mugimendukoa da *James ran the Boston Marathon* esaldian, baina estatikoa *James runs the family business* esaldian (*PropBank*eko `run.02` eta `run.01` adierak).

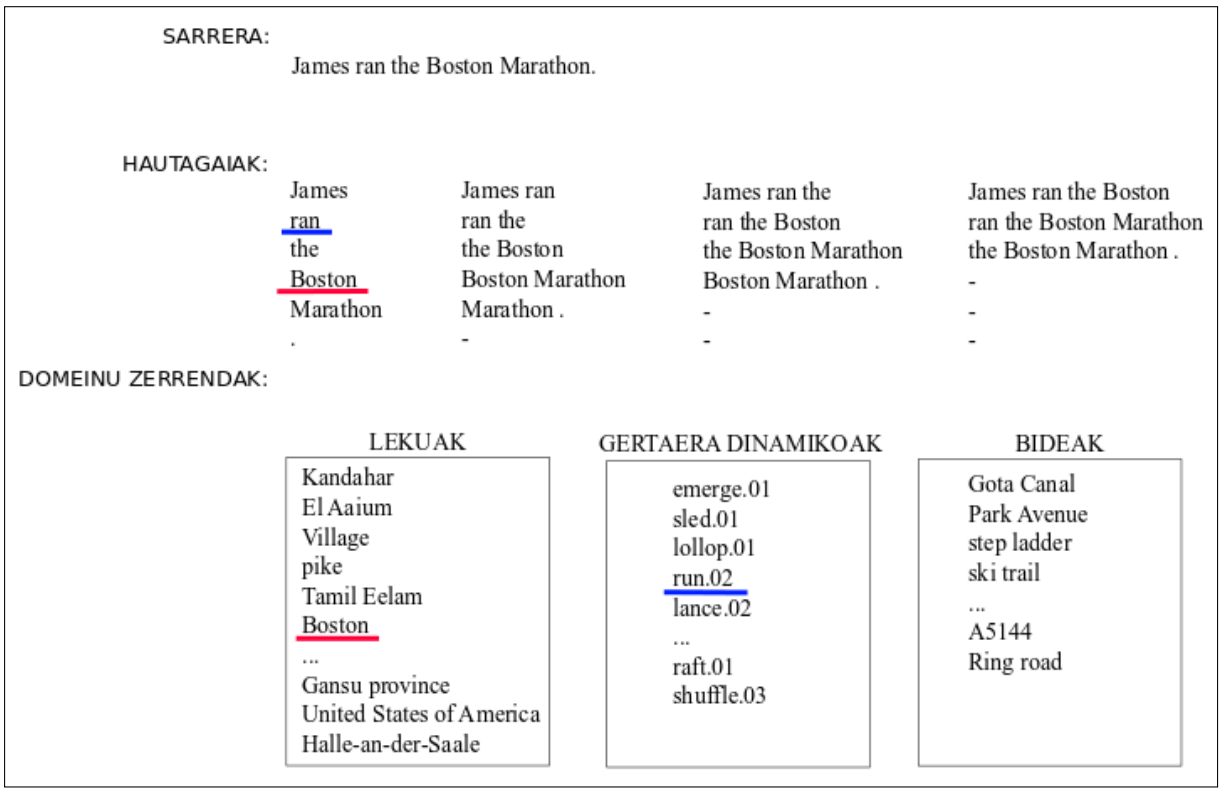
Adieren desanbiguazioa aurrera eramateko *Predicate Matrix*[17] (De Lacalle et al., 2014) izeneko baliabide lexikoa erabili da. Honek *FrameNet* (Fillmore et al., 2004), *PropBank* (Palmer et al., 2005) eta *WordNet* (ikus 2.1.2 atala) baliabideetatik eskuratutako predikatuen inguruko informazioa uztartzen du. *X-Space*n gertaera dinamikoen domeinuko *synset*en gakoak baliaturik multzo hauen barnean kokatutako gertaeren *PropBank* adiera finkatzeko erabili da. Horrela, dagozkien adierekin desanbiguatutako 511 mugimenduko gertaerak osatu zerrenda sortu ahal izan da.

B.1.6. Nola egiten da osagaien identifikazioa domeinuen zerrendekin

Lekuen, bideen eta gertaera dinamikoen identifikazioa egiteko domeinuen zerrendak aurreprozesamenduan eratutako hautagaiekin alderatzen dira. Prozesu hau 4.7 irudian ikus daiteke. Lekuentzat eta bideentzat erkaketa zuzenekoa da, osagai mota hauen domeinuetatik sorturiko zerrendetan aurki daitezkeen hautagaiak `<PLACE>` edo `<PATH>` bezala etiketatzen dira. Gertaera dinamikoentzat, aldiz, aurreprozesamenduan egindako analisi semantikotik token bakarreko hautagaien *PropBank* adierak hartzen dira (predikatuak

[17] *X-Space*ek hartzen duen bertsioa `1.1`a da.

izatekotan) eta hauek *WordNet*etik sortu den desanbiguatutako mugimendu gertaeren zerrendakoekin erkatzen dira. Hautagaien artean aurki daitezkeenei <MOTION_EVENT> etiketa esleitzen zaie. Argitu behar da gertaerak token bakarreko hautagaietan bilatzeko arrazoia honako hau dela: *X-Space* sistemaren garapenerako erabili den *SpaceBank* corpusean gertaeren token bakarreko buru lexikalak daudela anotatuta.



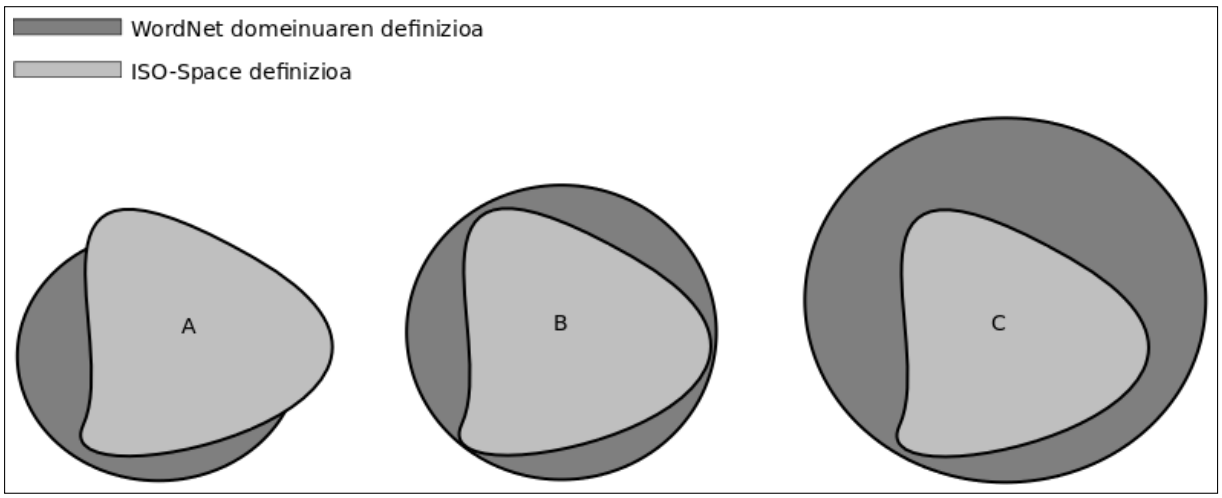
Irudia 4.7: Lekuen, bideen eta gertaera dinamikoen identifikazioa domeinuen eta osagai izateko hautagaien zerrendak erkatuta.

B.1.7. Zergatik identifikatzen dira hautagai gehiegi domeinuak erabilita

Identifikazioaren azkeneko urratsa domeinuak erabilita detektatzen den osagai kopuru handiak motibatzen du. Lehenago aipatu dugun bezala, domeinuak mugatu ahal izateko hauek osatzen dituzten elementu guztiak barneratzeko bezain zabalak diren kontzeptuei dagozkien *synset*ak aukeratu behar dira errotzat, eta kontuan eduki behar da, gainera, hautatu diren erroek adierazitako kontzeptuak zabalegiak baldin badira domeinuz kanpoko izenak ere barneratuko direla. Gorago zerrendatu ditugun erroen kasuan hauek

zuzenak direla egiaztatu ahal izan dugu, domeinuak ahal den hobekienik mugatzen dituztela alegia.

Gure ustez *ISO-Space* eskemak eta *WordNet* datu baseak lekuei, bideei eta mugimenduko gertaerei emandako definizioen arteko ezberdintasuna da identifikatutako mota horietako osagai kopurua behar baino handiagoa izatea. Honekin adierazi nahi dugu *WordNet*eko domeinua ahal den ongiena mugaturik ere, beti izango direla *ISO-Space* eskemako definizioaren arabera domeinutik kanpokoak diren osagaiak, bi baliabideek darabiltzaten definizioak, defektuz, ehuneko ehunean bat ez datozelako. Esate baterako `08644722-N` *synset*aren barneko *rural area* izeneko lekua ez dago anotatuta *SpaceBank* corpusean. 4.8 irudiak erakusten du definizioen ezberdintasuna kontuan edukita domeinuak mugatzeko izan dugun erronka.



Irudia 4.8: *WordNet* datu-basearen eta *ISO-Space* eskemako domeinuen arteko ezberdintasunak.

Hautatu erroekin *X-Space*n egoera erdikoan deskribatzen dena da (B). Bertan ikus daitekeenez, aukeratutako erroak dauzkaten *WordNet*en domeinuekin *ISO-Space* eskemaren definizioak ahal den gehiena hurbildurik ere, badira kanpoan gelditzen diren osagaiak (eremu gris ilunetakoak), aipatutako *rural area* lekua esate baterako. Hauek dira, hain zuzen, detektatutako gehiegizko osagaiak. A-n erakusten den egoera emango litzateke domeinua finkatzerakoan aukeratutako erroek *ISO-Space*n definizioa barnean hartuko ez balute eta C egoera, berriz, domeinua behar baino handiagoa izango balitz.

B.1.8. Nola ekidin domeinuekin hautagai gehiegi detektatzea

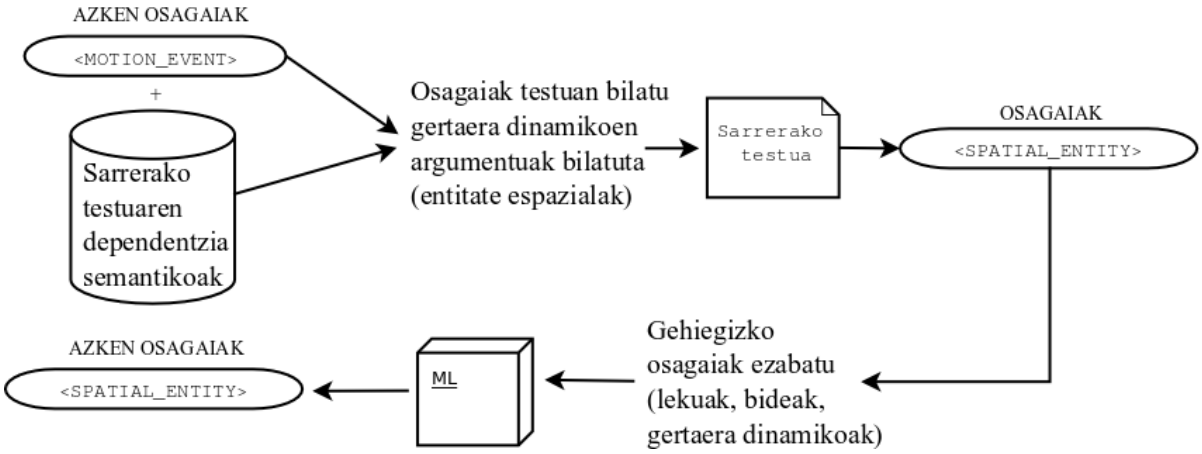
*ISO-Space*n definizioetara ahalik eta gehien hurreratzeko asmoz, beraz, *WordNet* domeinuak erabilita detektatutako hautagaiak eskemaren arabera ongi identifikaturik dauden edo ez erabakitzeaz arduratzen diren sailkatzaileak erabili dira. Horrela, oreka aurkitu ahal izan da *WordNet* domeinuen zabaltasunaren eta *ISO-Space*n definiziorako egokitzapen mailaren artean. Hiru sailkatzaile bitar sortu dira denetara, osagai bakoitzerako bat (lekuetarako, bideetarako eta gertaera dinamikoetarako). Horretarako *Support Vector Machines* algoritmoa eta segidan zerrendatzen diren ezaugarriak inplementatu dira. Ezaugarri hauek hautagai bakoitzerako erauzi dira, *X-Space*n aurreprozesamendu fasean burutako oinarrizko etiketatze linguistikotik. Zerrendan antzeman daitekeenez hiruko leihoa hartu dugu, token bakoitzaren aurreko eta hurrengo tokenek eskaintzen duten informazioa erabili dugu alegia. Horiek erabiltzeko arrazoia proba-errore prozesu baten ondorio da, bertan ikusi ahal izan dugu ezaugarri horiek lagungarriak direla.

- Forma
- Lema
- *Part-of-Speech* kategoria
- Dependentzia sintaktikoa (`DEPREL`)
- Rol semantikoa (rol, adjuntu edo "–" etiketa)
- Hautagaiari dagokion predikatuaren *PropBank* adiera.
- Aurreko tokenaren forma
- Aurreko tokenaren *Part-of-Speech* kategoria
- Aurreko tokenaren dependentzia sintaktikoa (`DEPREL`)
- Hurrengo tokenaren forma
- Hurrengo tokenaren *Part-of-Speech* kategoria
- Hurrengo tokenaren dependentzia sintaktikoa (`DEPREL`)

X-Space etiketatzaileak lekuen, bideen eta gertaera dinamikoen identifikaziorako aplikatzen duen metodologiaren deskribapena bukatzeko, hurbilpen hau ataza honetan aplikatzen den lehen aldia dela azpimarratu nahi dugu. *WordNet* erabilita domeinuak mugatzeko ideia berria ez bada ere, hots, beste ataza batzuetarako inplementatu izan bada ere, ez da inoiz espazioa etiketatzeko baliatu (*SpatialML*, *SpRL* eta *ISO-Space* eskemak kontuan izanda).

B.2. Entitate espazialak

Entitate espazialen kasuan esan dugu gertaera dinamiko edo mugimenduzkoen argumentuak markatzen dituztela, hori dela eta, *WordNet* baliatuz detektatutako mugimenduko gertaeren argumentuak diren hautagaiak bilatzen dira entitateak etiketatu ahal izateko. Bilaketa aurreprozesamendu fasean burutu analisi semantikotik sortutako dependentzia zuhaitzak aztertuta egiten da, eta gertaera dinamikotzat etiketatutako hautagaien argumentuei `<SPATIAL_ENTITY>` etiketa esleitzen zaie. Prozesu osoa 4.9 irudian ikus daiteke.



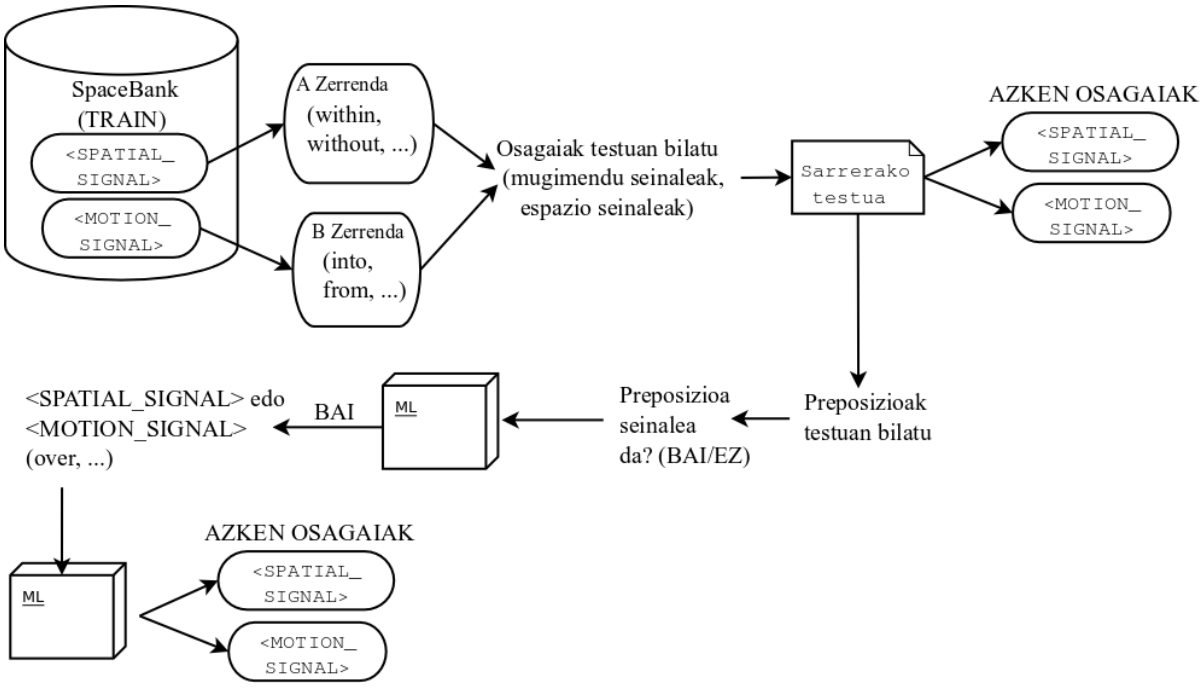
Irudia 4.9: *X-Space*n entitate espazialak etiketatzeko arkitektura.

Argitu nahi dugu, entitate espazialen detekziorako arestian azaldu dugun bilaketa teknika erabilita identifikatutako entitate kopurua hasiera batean handia izan zela, eta honek estaldura altua izatea ekartzen zuela, doitasunaren kaltetan betiere. Hortaz, bi neurrien arteko desoreka arindu, eta sistemaren eraginkortasuna hobetzen saiatzeko asmoz, ikasketa automatiko bidezko sailkatzaile bitarra gehitu genuen. Sailkatzaile

honek entitate espazialak bilaketa teknikaren bitartez ongi edo gaizki etiketatzen diren erabakitzea dauka helburu. Sailkatzaile honek xedea iristea erdiesten du, doitasunaren eta estalduraren arteko oreka aurkitzea alegia. Sailkatzailea sortzeko *WordNet* hurbilpenean aurkeztutako ezaugarriak erabili dira.

B.3. Mugimendu eta espazio seinaleak

Seinaleak detektatzeko, bi zerrenda sortu dira, *SpaceBank* corpusaren entrenamendu zatiko anotazioetan oinarrituta. Zerrendetako batek (A) espazio seinaleak baizik ezin izan daitezkeenak biltzen ditu, *within* eta *without* esate baterako, eta besteak (B), berriz, mugimendu seinaleak baizik ezin izan daitezkeenak, *into* eta *from* adibidez. Kontuan izan behar da, preposizio asko egoera batzuetan mugimenduko seinaleak izan daitezkeela, baina espaziokoak beste zenbaitetan, *over* esaterako. Aipatu nahi dugu, gainera, *SpaceBank* corpusa azterturik gehienez ere bi token luze diren seinaleak aurki daitezkeela, edozein motatakoak izanda ere. Seinaleak identifikatu ahal izateko 4.10 irudian azaltzen den teknika erabiltzen da.

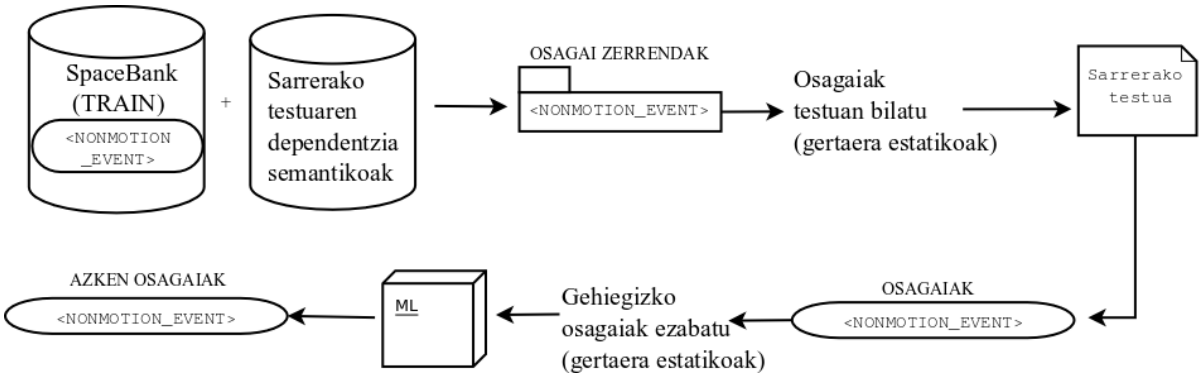


Irudia 4.10: Seinaleak detektatzeko teknika *X-Space* tresnan.

Lehenik, espazio seinaleak eta mugimenduko seinaleak bilatzen dira, token bakarreko eta bi tokeneko hautagaien artean aipatu zerrendak (`A` eta `B`) erabilita; `An` aurki daitezkeenei `<MOTION_SIGNAL>` etiketa esleitzen zaie eta `Bn` aurki daitezkeenei `<SPATIAL_SIGNAL>` etiketa. Bigarrenik, token bateko eta bi tokeneko hautagaietan preposizioak bilatzen dira, aurreprozesamenduan egindako etiketatze linguistikoa baliatuz eta, ondoren, *WordNet* hurbilpenean aurkeztutako ezaugarriak inplementatzen dituen sailkatzaile baten bitartez, preposizio haietako bakoitza etiketatu edo ez (seinalea den edo ez), eta nola (seinalea bada zer motatakoa den) erabakitzen da. Sailkatzaile honi hautagaiak eman baino lehen `A` eta `B` zerrenden bitartez etiketatutakoak hautagaitzatik kentzen dira.

B.4. Gertaera estatikoak

Gertaera mota hauen identifikazioa lortzeko *SpaceBank* corpusaren entrenamendurako (*train*) zatian anotatutakoak eta hauei dagozkien adierak hartu dira. Gertaera horiekin eta horien adierekin zerrenda sortu da. Adierak erdiesteko, aurreprozesamendu faseko dependentzia semantikoen zuhaitzak baliatu dira. Identifikazioa burutzeko token bakarreko hautagaiak eta hauen adierak (izatekotan) hartu, arestian aipatu zerrendan xerkatu eta hautagairik aurkitzen bada, `<NONMOTION_EVENT>` etiketa eransten zaie. Etiketatze honetan mugimenduko gertaeren zerrenda ere hartzen da aintzat, eta bertan edireten direnei gertaera estatiko izateko hautagaitza zuzenean kentzen zaie. Gertaera estatikoak identifikatu ahal izateko 4.11 irudian azaltzen den teknika erabiltzen da.



Irudia 4.11: Gertaera estatikoak detektatzeko teknika *X-Space* tresnan.

Teknika hau erabilita identifikatutako gertaera estatiko kopurua handia izan zen hasiera batean, eta honek estaldura altua izatea ekartzen zuen, doitasunaren kaltetan betiere. Hortaz, bi neurrien arteko desoreka arindu, eta sistemaren eraginkortasuna hobetzen saiatzeko asmoz, ikasketa automatiko bidezko sailkatzaile bitarra gehitu genuen. Honek gertaera estatikoak azaldutako teknikaren bitartez ongi edo gaizki etiketatzen diren erabakitzea dauka helburu. Sailkatzaile honek xedea iristea erdiesten du, doitasunaren eta estalduraren arteko oreka aurkitzea alegia.

C. Erlazio espazialak (identifikazioa + atributuen sailkapena)

X-Space tresnak aurrera eramaten duen azkeneko urratsa osagai espazialen arteko erlazioak etiketatzea da. Hiru motakoak dira erlazio hauek: mugimendukoak (<MOVELINK>), kualitatiboak (<QSLINK>) eta norabidekoak (<OLINK>). Lehenbizikoek gertaera dinamikoetan parte hartzen duten osagai espazialen arteko erlazioak markatzen dituzte. Bigarrenek, berriz, seinale espazial topologikoen (TOPOLOGICAL) edo aldi berean norabidekoak eta topologikoak diren seinale espazialen (DIR_TOP) menpe dauden osagaien artekoak. Hirugarren motakoek, bukatzeko, norabide seinale espazialen (DIRECTIONAL) menpeko osagaien arteko erlazioak markatzen dituzte.

X-Space sistemaren aurreko urratsa bezala, erlazioen etiketatzea ere sekuentzialki burutzen diren bi pausotan egiten da. Erlazioak identifikatzen dira lehenbizi, eta hauek jasotzen dituzten atributuen balioak esleitzen dira ondoren. Kasu honetan, dena dela, identifikazioa zuzenean egiten da. Izan ere, Pustejovsky eta Yocumek (2013) adierazten duten gisan, mugimenduko erlazio bat eragiten edo abiarazten du identifikatutako gertaera dinamiko bakoitzak. Gainera, topologikoa edo aldi berean norabidekoa eta topologikoa den seinale espazial bakoitzak erlazio kualitatibo bat adierazten duela eta norabideko seinale bakoitzak norabideko lotura bat sortzen duela ere esaten dute. Hauek kontuan izanda identifikatzen dira, beraz, lotura edo erlazio espazialak *X-Space*n.

Atributuen balioak esleitzeari dagokionez, ordea, prozesua ez da, orokorrean, identifikaziokoa bezain zuzena. Izan ere, *ISO-Space* aurkeztean azaldu bezala, eskema honek <MOVELINK>, <QSLINK> eta <OLINK> erlazioetan parte hartzen duten osagai espazialei rol espazialak deriztenak esleitzen dizkie besteak beste. *ISO-Space* eskemak zehaztutakoa kontuan edukita, balioak teknika desberdinak baliaturik jasotzen dituzten hiru atributu mota bereizi ditugu: abiarazleak, rol espazialak eta ohiko atributuak.

Erlazio bakoitzak haren abiarazle papera jokatzen duen osagai espazial bat jasotzen du atribututzat. Mugimenduko erlazioen kasuan gertaera dinamikoak izaten dira eta gainerakoetan, aldiz, seinale espazialak. Loturen detekzio prozesuan azaldu dugunetik ondoriozta daitekeen bezala, erlazio bakoitzaren abiarazlea den atributuaren zehaztapena *X-Space*n zuzenekoa da; identifikazio fasean egindakotik hartzen baita.

Rol espazialak diren atributuen kasuan, aldiz, kontuan izan dugu hauen jatorri teorikoa, *SpRL* eskema alegia, eta horregatik hauen esleipena segidan azaltzen den bezala egitea erabaki dugu. `<MOVELINK>` lotura bakoitzak hartzen dituen `source`, `goal`, `mover` eta `landmark` atributuak loturaren abiarazlea den gertaera dinamikoaren (predikatuaren) argumentuak direla uler daiteke, horrela ikus daitezke. `<QSLINK>` eta `<OLINK>` loturetako bakoitzaren atributuak diren `trajector` eta `landmark`, berriz, erlazio hauek abiarazten dituzten seinale espazialen menderatzaileak diren predikatuen argumentutzat hartzen dira.

Aipatu ikuspegietan oinarriturik, `<MOVELINK>` erlazio bakoitzaren rol espazialak esleitzeko, lehenik, hura abiarazten duen gertaera dinamikoaren argumentuak eta hauei esleitutako rol semantikoak hartzen dira aurreprozesamendu faseko dagokion dependentzia semantikoen zuhaitzetik. Bigarrenik, rol etiketa horietako bakoitza (*PropBank* ereduko SRLrena, ikus 2.1.2) egoera edo *predikatu-argumentu* egitura bakoitzean *SpRL* ereduko zer rol espaziali dagokion erabakitzeaz arduratzen den sailkatzailea baliatzen da. `<QSLINK>` eta `<OLINK>` erlazioen rol espazialak esleitzeko ere teknika bera aplikatzen da, baina lotura horien abiarazleei dagozkien (menderatzaileak diren) predikatuen argumentuak hartu dira, eta ez abiarazleenak berenak (seinaleak eta preposizioak baitira besteak beste). Azkenik, erlazio espazialen ohiko atributuak direnak etiketatzeko, sailkatzaile bana aplikatu da.

4.2.4 Emaitzak eta analisia

*X-Space*n eraginkortasuna neurtzeko *SpaceBank*en *test* zatian sistemak egindako etiketatzeak ebaluatu dira. *SpaceEval* deskribatzean azaldu bezala, hiru konfigurazio erabilita egindako hainbat azpiataza landu ziren saio honetan, eta hauek dira, ondorioz, *X-Space*n ebaluazioan ere kontuan hartu direnak. Lehenbizikoan testu soila jasotzen da sarreratako (1). Bigarrenean (2), aldiz, atributuak esleitu gabe dauzkaten eskuz identifikatutako osagai espazialak. Azkenik, hirugarrenean (3), eskuz identifikatutako osagai espazialez gaine-

ra haiek hartzen dituzten atributuen balioak ere jasotzen dira. Hiru konfigurazioetakoak bildurik, hamar dira lantzen diren azpiatazak. 4.7 taulan ikus daiteke hauetan *X-Space*k iritsitako emaitzak zein diren. Analisi egokia egiteko, gainera, 4.8 taulan `1_Osag_SAILK` azpiatazan osagai bakoitzerako erdietsitakoak ere aurkezten ditugu.

Ataza	Doitasuna	Estaldura	F_1	Accuracy
1_Osag_ID	81	72	76	88
1_Osag_SAILK	75	72	74	90
1_Osag_ATR	18	15	16	30
1_Erla_ID	54	51	53	55
1_Erla_ATR	6	5	5	25
2_Osag_ATR	26	33	29	63
2_Erla_ID	55	51	53	89
2_Erla_ATR	6	8	7	46
3_Erla_ID	63	51	56	89
3_Erla_ATR	7	9	8	48

Taula 4.7: *X-Space* sistemaren emaitzak *SpaceEval* (*SemEval-2015*) ebaluazio saioko konfigurazio eta azpiataza bakoitzerako.

1_Osag_SAILK	Doitasuna	Estaldura	F_1	Accuracy
<PLACE>	75.45	75.07	75.26	84.72
<PATH>	72.20	77.61	74.54	93.94
<SPATIAL_ENTITY>	85.94	80.75	82.96	89.88
<MOTION_EVENT>	80.30	82.00	81.11	92.17
<NONMOTION_EVENT>	69.01	64.78	66.59	95.94
<SPATIAL_SIGNAL>	73.09	69.15	71.06	87.20
<MOTION_SIGNAL>	76.17	72.45	74.26	85.13

Taula 4.8: `1_Osag_SAILK` azpiatazan osagai bakoitzarentzat iritsitako emaitzak.

4.8 taularen inguruan argitu nahi dugu, `1_Osag_SAILK` azpiatazaren ebaluaziorako erabilitako programak ez dituela itzultzen `<SPATIAL_SIGNAL>` eta `<MOTION_SIGNAL>` osagaiei dagozkien emaitza banakatuak, eta horregatik haien balioak lortu ahal izateko ebaluazio saioan erabilitako *script*a egokitu behar izan dugula.

4.7 eta 4.8 tauletako balioez gainera, 4.9n erlazio espazialen identifikazioari hiru konfigurazioetan dagozkion emaitzak ere aurkezten ditugu (`1_Erla_ID`, `2_Erla_ID` eta

`3_Erla_ID`). Emaitza hauek aski erabilgarriak dira, osagai espazialen eta hauen atributuen etiketatzeek loturen identifikazioan daukaten eragina islatzen dutelako.

		Doitasuna	Estaldura	F_1	Accuracy
1_Erla_ID	**<MOVELINK>**	61.22	53.08	56.86	61.08
	<QSLINK>	50.89	50.05	50.46	52.83
	<OLINK>	51.47	50.02	50.73	52.93
2_Erla_ID	**<MOVELINK>**	74.31	51.91	61.12	84.77
	<QSLINK>	45.18	50.00	47.46	90.36
	<OLINK>	45.68	50.00	47.74	91.37
3_Erla_ID	**<MOVELINK>**	74.31	51.91	61.12	84.77
	<QSLINK>	45.18	50.00	47.46	90.36
	<OLINK>	70.69	50.05	58.60	91.37

Taula 4.9: Erlazio espazialen identifikazioa konfigurazio desberdinetan.

Analisia

Azpiatal honetan *X-Space* sistemaren emaitzak analizatzen ditugu. Hori egin ahal izateko, arestian aurkeztutako 4.7, 4.8 eta 4.9 taulak ez ezik, 4.2 taula ere hartzen da kontuan. Azkeneko honetan *X-Space* sistemaren garapenean eta ebaluazioan erabili den *SpaceBank* corpuseko etiketa bakoitzaren kopurua biltzen da.

1. Analisiarekin hasteko, gure ustez ondokoa ikus daiteke 4.7 taulan: osagai espazialen identifikazioan (`1_Osag_ID`) eta kategorizazioan (`1_Osag_SAILK`) erdiesten diren emaitzek adierazten dutenez, azpiataza hauek teknika egokien bitartez hurbiltzea lortu dela. Izan ere, nabarmendu behar da bai identifikazioan eta baita kategorizazioan ere osagai espazialentzat lortutako F_1 neurria batez beste 75 puntukoa dela, eta *accuracy* neurria 90 puntutik hurbil dagoela. Gure iritziz hauek emaitza onak dira, eta *X-Space* tresnak erabiltzen dituen identifikazio eta sailkapen teknikak egokiak direla adierazten dute. Etiketatzailearen diseinua azaltzean aipatu bezala, gure sistemak *WordNet* baliabidean oinarritzen den teknika berritzailea erabiltzen du, lekuen, bideen eta gertaera dinamikoen identifikazioa eta sailkapena burutzeko. 4.7 taulako balioak ez ezik, 4.8 taulakoak ere interesgarriak dira, *WordNet* teknikaren eraginkortasuna egiaztatzeko garaian. Izan ere, bertan ikusten den moduan, lekuen F_1 neurria eta *accuracy*a 75.26 eta 84.72 dira, bideena 74.54 eta 93.94, eta mugimenduko gertaerena, azkenik, 81.11 eta 92.17.

2. Gainerako osagai espazialen identifikazio eta kategorizazioari erreparatzen badiogu, ordea, haietan ere balioak nahiko altuak direla ikus daiteke 4.8 taulan. Esate baterako, entitate espazialena da F_1 neurririk gorena, 82.96, lortzen duen osagai mota. *X-Space* tresnaren deskribapenean azaldu dugunez, entitate espazialak etiketatu ahal izateko, lehenbizi gertaera dinamiko edo mugimenduzkoen argumentuak markatzen dira eta, horretarako, *WordNet* hurbilpena baliatuz detektatutako mugimenduko gertaeren argumentuetan hautagaiak direnak bilatzen dira. Gero, bigarrenik, ikasketa automatiko bidezko sailkatzaile bitarra erabiltzen da, lehenbiziko pausuaren bitartez detektatutako entitate kopurua handiegia delako, eta honek estalduraren eta doitasunaren desoreka sortzen duelako. Ikusten den eran, orekatze hau lortu egiten da (85.94 puntuko doitasuna eta 80.75 puntuko estaldura) eta argi gelditzen da bi pausoko entitateen etiketatzerako teknika hau eraginkorra dela.

3. Osagai mota bakoitzetik *SpaceBank*en anotaturik dagoen kopuruari (4.2 taula) kasu eginda 4.8 taulako balioak aztertzen direnean, oro har emaitzarik onenak iristen dituztenak kopururik handiena anotatuta duten osagaiak direla ikusten da, eta gutxien dauzkatenak baliorik apalenak erdiesten dituztenak. Izan ere, 1.347 entitate espazial eta 751 gertaera dinamiko daude anotatuta *Train* zatian, eta hauen etiketatzean 82.96 eta 81.11 puntuko F_1 neurria lortzen da hurrenez hurren. Gertaera estatikoetan, bestalde, 321 baizik ez daude anotatuak, eta 66.59 puntuko F_1 iristen da. Etiketatzeaz arduratzen diren sailkatzaileen jokabide hau ohikoa izaten da, landu beharreko osagai guztien zailtasun maila eta haiek etiketatzeko erabili metodoen eraginkortasuna antzekoak izanik ikasketa teknikak aplikatzen direnean.

4. Emaitzen analisiarekin jarraitzean ikus daiteke osagai espazialei dagozkien atributuen esleipenarekin zerikusia duten azpiatazak (`1_Osag_ATR` eta `2_Osag_ATR`) oraindik ere landu beharra duten azpiatazak direla. Honekin, handi-handika, adierazi nahi dugu bertan lortzen diren emaitzak behar baino apalagoak direla iruditzen zaigula. *X-Space*k, sistemaren deskribapenean azaldu den moduan eta, teorian behintzat, ez ditu atributu hauetako bakoitzari hobekien egokitzen zaizkion ezaugarri multzoak inplementatzen. Hori dela eta, pentsa daiteke hau dela emaitza apalak erdiesteko arrazoia. Aipatu behar da, dena den, azpiataza hauek landu dituen beste sistemari erreparatzean (`BASELINE`) *X-Space*nak azken honenak baino hobeak direla ikus daitekeela (4.10 taula). `1_Osag_ATR` azpiatazan *X-Space*n F_1 neurria

16 da eta BASELINEn, aldiz, 4. Bestalde, 2_Osag_ATR azpiatazan *X-Space*n F_1 neurria 29 da, eta BASELINEn, berriz, 27.

	Sistema																			
	BASELINE				**IXA**				**BRANDEIS-CRF**				**HRIJP_CRF_VW**				**UTD**			
Ataza	P	E	F_1	D	P	E	F_1	D	P	E	F_1	D	P	E	F1	D	P	E	F_1	D
1_Osag_ID	55	52	53	75	81	72	76	88	85	80	83	89	**84**	**83**	**83**	**89**	-	-	-	-
1_Osag_SAILK	55	51	53	86	75	72	74	90	**78**	**76**	**77**	**92**	77	76	76	91	-	-	-	-
1_Osag_ATR	10	2	4	5	**18**	**15**	**16**	**30**	-	-	-	-	-	-	-	-	-	-	-	-
1_Erla_ID	50	50	50	50	54	51	53	55	-	-	-	-	**56**	**51**	**53**	**57**	-	-	-	-
1_Erla_ATR	5	2	2	6	**6**	**5**	**5**	**25**	-	-	-	-	3	4	3	25	-	-	-	-
2_Osag_ATR	27	28	27	76	**26**	**33**	**29**	**63**	-	-	-	-	-	-	-	-	-	-	-	-
2_Erla_ID	**79**	**58**	**67**	**90**	55	51	53	89	-	-	-	-	-	-	-	-	-	-	-	-
2_Erla_ATR	**19**	**20**	**19**	**66**	6	8	7	46	-	-	-	-	-	-	-	-	-	-	-	-
3_Erla_ID	**86**	**84**	**85**	**98**	63	51	56	89	-	-	-	-	78	57	66	86	87	82	85	98
3_Erla_ATR	**26**	**26**	**26**	**79**	7	9	8	48	-	-	-	-	5	6	5	48	5	9	7	51

Taula 4.10: *SpaceEval* ebaluazio saioan parte hartu zuten etiketatzaileen emaitzak. Onenak beltzez (P: Doitasuna, E: Estaldura, D: *Accuracy*).

(Pustejovsky et al., 2015) argitalpenean adierazten den moldean, BASELINE tresnak osagai espazialen atributu guztiak esleitzeko osagaiaren testua baizik ez du erabiltzen ezaugarritako, eta ikasketa metodotzat erregresio logistikoa aplikatzen du. Gure sistemak, aldiz, 4.2.3 azpiatalean zerrendatzen diren ezaugarriak (lexikoak, sintaktikoak, semantikoak, etab.) eta *SVM* algoritmoa erabiltzen ditu. Emaitza hauek alderatzetik atera daitekeen ondorioa, hortaz, hau da gure ustez: osagaien atributuen etiketatzean, emaitzak hobetzeko, beharrezkoa izango dela atributu bakoitzari hobekien egokitzen zaizkion ezaugarrien azterketa egitea.

Emaitzen analisiaren puntu honekin amaitzeko, aipatu nahi dugu *ISO-Space*n atributu kopuru handia, eta orokorrean, *SpaceEval* atazaren hedadura zabala izan direla, gure iritziz, osagaien atributuekin lotutako 1_Osag_ATR eta 2_Osag_ATR azpiatazak bi sistemak bakarrik lantzearen arrazoia. Gainera, aipatu sistemek atributu guztientzat ezaugarri berak erabiltzeko arrazoia ere hau dela deritzogu.

5. *X-Space*n emaitzen analisiarekin jarraitzean ikus daitekeen moduan, erlazioen (<MOVELINK>, <QSLINK> eta <OLINK>) identifikazioan erdietsitako balioak (1_Erla_ID, 2_Erla_ID eta 3_Erla_ID) ez dira *ISO-Space* eskemaren gidalerroetan aipatzen dena kontuan izanda espero zitezkeenak. Sistemaren diseinuan azaldu dugun gisa, gidalerro horiek (Pustejovsky eta Yocum, 2013) ondokoak zehazten dituzte: a) identifikatutako gertaera dinamiko bakoitzerako <MOVELINK>

motako lotura sortzen dela, b) identifikaturiko TOPOLOGICAL edo DIR_TOP seinale espazial bakoitzerako <QSLINK> motako lotura ezartzen dela, eta c) DIRECTIONAL seinale espazial bakoitzerako <OLINK> lotura finkatzen dela.

*X-Space*k, erlazioak identifikatzeko, aurrekoa hartzen du oinarritako, baina 4.7 eta 4.9 tauletako balioak kontuan izanda, argi gelditzen da hori ez dela nahikoa. 1_Erla_ID azpiatazaren kasuan, pentsa liteke lortutako emaitzak seinale espazialen, hauen atributuen eta gertaera dinamikoen etiketatzean sortutako erroreen ondorio direla (ikus 4.7 taula). Kontua da 2_Erla_IDn sisteman osagai hauen identifikazioa eskuz anotaturik jasotzen dela, eta 3_Erla_IDn, identifikazioaz gainera, hauen atributuen esleipena ere eskuz anotatuta jasotzen dela. Eta, hori horrela izanda ere, 2_Erla_IDn eta 3_Erla_IDn loturen identifikazioan erdietsi emaitzak 1_Erla_IDkoen oso antzekoak dira: 1_Erla_IDn F_1 neurria 53 da, 2_Erla_IDn ere 53, eta 3_Erla_IDn 56. Honetatik ondorioztatzen dugu beraz, *SpaceBank* anotatu zenean *ISO-Space*n ez zela guztiz jarraitu gidalerroetan zehazten dena, eta horregatik erdiesten direla, loturak identifikatzean, emaitza horiek.

6. Azkenik, analisiarekin bukatzeko, antzematen da erlazio espazialei dagozkien atributuen esleipenean (1_Erla_ATR, 2_Erla_ATR eta 3_Erla_ATR) emaitza oso apalak iristen direla. Izan ere, *X-Space*k 1_Erla_ATR azpiatazan 5 puntuko F_1 neurria eta 25 puntuko *accuracy*a lortzen ditu, 2_Erla_ATR azpiatazan 7 eta 46 puntukoak, eta 3_Erla_ATRn 8 eta 48koak. Emaitza hauek, beraz, arras apalak izan arren, aipatu azpiatazak landu dituzten *SpaceEval* saioko gainerako sistemen emaitzekin alderatzen ditugunean, ikusten dugu *X-Space*k erdietsitakoa dela saio horretako 1_Erla_ATR azpiatazako F_1 emaitza gorena. 3_Erla_ATR azpiatazan, gainera, gure sistemaren F_1 emaitza ere azpiataza hau landu zuten HRIJP-CRF-VW eta UTD taldeena baino hobea bada ere (8 puntu), urrun gelditzen da BASELINE sistemak itzulitako 26 puntuko F_1 neurritik.

(Pustejovsky et al., 2015) argitalpenean azaltzen den moduan BASELINE sistemak, erlazioak etiketatzeko, atributu motaren araberako hurbilpenak erabiltzen ditu: rol espazialetarako, hau da, rolak jokatzen dituzten argumentuak identifikatzeko, loturaren abiarazlearen esaldian identifikatutako osagaien artean bilatzen da, eta ohiko atributuetarako, aldiz, rol espazialak jokatzen dituzten argumentuen testuak eta abiarazlearen mota zehazten duen atributua hartzen dira. Argi dago,

beraz, emaitzetan oinarrituta, `BASELINE` tresnak inplementatzen dituenak direla, egundaino, erlazio espazialen atributuen esleipenak egiteko metodorik egokienak.

4.2.5 SRLren eraginaren azterketa

Sarrerako atalean tesi lan honek dituen helburuen artean bi hipotesi egiaztatzea dagoela adierazi dugu. Hauetako batek euskaraz denboraren adierazpen linguistikoa etiketatzeko orduan rol semantikoek daukaten eragina, ingelesez eta gaztelaniaz agitzen den eran, positiboa dela dio. Azpiatal honetan, tesi laneko beste hipotesian zentratzen gara. Honen arabera espazioaren adierazpena, denborarena bezala, fenomeno semantikoa da, eta horregatik semantika eta, zehazkiago, rol semantikoek duten garrantzia nabarmena da, informazio espazialaren etiketatze eraginkorra egin ahal izateko. Hau egiaztatzeko *X-Space*n erabiltzen diren sailkatzaileak rol semantikoen ezaugarriak erabili gabe entrenatu dira, eta sistema ondoren berriz ebaluatu da. SRLren eraginaren azterketa hiru modutara egin dugu: azpiatazaka, osagai espazial bakoitzerako eta erlazioen identifikaziorako.

Azpiatazak

Jarraian aurkezten ditugu, 4.11 taulan, azpiataza bakoitzean SRL erabilita eta erabili gabe erdietsitako balioak (zutabeen ezkerreko aldean rolekin eta eskuinekoan rolik gabe).

	Doitasuna		**Estaldura**		F_1		**Accuracy**		
Ataza	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	**Hobekuntza (F_1)**
1_Osag_ID	81	78	72	72	76	75	88	85	+1
1_Osag_SAILK	75	68	72	70	74	69	90	82	+5
1_Osag_ATR	18	18	15	14	16	16	30	30	0
1_Erla_ID	54	50	51	48	53	49	55	52	+4
1_Erla_ATR	6	7	5	5	5	6	25	29	-1
2_Osag_ATR	26	25	33	32	29	28	63	62	+1
2_Erla_ID	55	54	51	51	53	52	89	87	+1
2_Erla_ATR	6	6	8	8	7	7	46	45	0
3_Erla_ID	63	63	51	51	56	56	89	89	0
3_Erla_ATR	7	8	9	11	8	9	48	51	+1

Taula 4.11: Rol semantikoek *X-Space* tresnan duten eraginaren neurketaren emaitzak ($\neg$*SRL*: Rolik gabe, *SRL*: Rolekin).

X-Space sistemak burutzen dituen azpiatazetan rol semantikoek duten eraginari erreparatzen baldin badiogu (4.11 taula), ikus daiteke hobekuntzarik nabarmenena testu soi-

letik abiatuta egindako osagai espazialen kategorizazioan (1_Osag_SAILK) eta loturen identifikazioan iristen (1_Erla_ID) dela. Bi ataza hauek bost eta lau puntuko hobekuntza izan dute F_1 neurrian rol semantikoei esker, lehenbizikoak 69 puntutik 74 puntura eta bigarrenak 49 puntutik 53 puntura. Gure iritziz, 1_Osag_SAILK azpiatazaren hobekuntza da 1_Erla_ID azpiatazarena eragiten duena, neurri batean behintzat. Izan ere, *X-Space*k lotura espazialak detektatzeko erabiltzen duen teknikak 1_Osag_SAILKen kategorizatutako gertaera dinamikoak eta seinale espazialak hartzen ditu oinarritako.

X-Space etiketatzaileak lantzen dituen gainerako zortzi azpiatazei dagokienez, aldiz, 4.11 taulako balioek adierazten dute SRLk lautan puntu bat igo arazten dituela emaitzak (1_Osag_ID, 2_Osag_ATR, 2_Erla_ID eta 3_Erla_ATR), hirutan ez duela (ia) eraginik (1_Osag_ATR, 2_Erla_ATR eta 3_Erla_ID), eta beste batean emaitza (F_1) puntu bat okertzen duela (1_Erla_ATR). Nabarmendu nahi dugu lotura espazialen atributuak esleitzeaz arduratzen diren azpiatazetatik lehenbizikoak, aurreneko konfiguraziokoak (1_Erla_ATRek), rol semantikoekin puntu bat galtzen duela, bigarren konfiguraziokoak (2_Erla_ATRek) ez duela ez hobetzerik ez okertzerik, eta hirugarren konfiguraziokoak (3_Erla_ATRek) puntu bateko hobekuntza lortzen duela. Ezin izan dugu zehaztu jokabide hau zeren ondorio den, baina emaitzak, oro har, aski apalak dira aipatu hiru azpiatazetan. Bestalde, konfigurazio batetik besterako emaitzen aldaketa hau ongi edo gaizki esleitutako atributu gutxi batzuen kontua dela egiaztatu ahal izan dugu.

Osagai espazialak

*X-Space*k burutzen dituen azpiatazetan rol semantikoek daukaten eragina neurtzeaz gainera, osagai espazial mota bakoitzean dutena ere neurtu dugu. Neurketa hauek 4.12 taulan bildu dira.

1_Osag_SAILK	**Doitasuna**		**Estaldura**		F_1		**Accuracy**		
Osagaia	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	*SRL*	$\neg$*SRL*	**Hobekuntza (F_1)**
<PLACE>	75.45	65.11	75.07	68.33	75.26	66.68	84.72	71.23	+8.58
<PATH>	72.20	64.03	77.61	71.89	74.54	67.73	93.94	77.96	+6.81
<SPATIAL_ENTITY>	85.94	81.76	80.75	77.08	82.96	79.35	89.88	84.07	+3.61
<MOTION_EVENT>	80.30	78.41	82.00	80.01	81.11	79.20	92.17	90.45	+1.91
<NONMOTION_EVENT>	69.01	67.53	64.78	62.25	66.59	64.78	95.94	91.00	+1.81
<SPATIAL_SIGNAL>	73.09	71.17	69.15	67.87	71.06	69.48	87.20	86.11	+1.58
<MOTION_SIGNAL>	76.17	75.80	72.45	73.01	74.26	74.37	85.13	84.43	-0.11

Taula 4.12: Rol semantikoek osagai espazialen kategorizazioan (1_Osag_SAILK) daukaten eraginaren neurketaren emaitzak.

4.12 taulan ikus daitekeenez, aldaketarik nabarmenenak jasaten dituzten osagai motak lekuak (<PLACE>), bideak (<PATH>) eta entitate espazialak (SPATIAL_ENTITY) dira. Lehenbiziko eta bigarren motakoentzat 8.58 eta 6.81 puntuko hobekuntza dago F_1 neurrian, eta hirugarren motakoentzat 3.61 puntukoa. Espero izatekoa zen rolen eraginik handiena jasotzen dutenak hauek izatea, lekuen eta bideen kasuan SRL sistemak esleitzen duen AM-LOC adjuntu etiketa ia beti mota honetako osagaitan baliatzen delako eta, entitate espazialen kasuan, gertaera dinamikoen argumentuak direlako.

Gainerako osagai espazialek rol semantikoen ondorioz duten eraginaren kariaz, argi gelditzen da, 4.12 taula ikusirik, rolen efektua orokorrean positiboa dela. Seinale dinamikoak (<MOTION_SIGNAL>) dira beren F_1 neurriaren balioa hobetzea lortzen ez duten bakarrak. Hauetan 0.11 puntu okertzen da aipatu neurria, aldaketa oso apala. Pentsatzekoa da seinale dinamikoak kategorizatu ahal izateko (preposizioak izaten direla kontuan edukita) *Part-of-Speech* kategoria eta tokenaren forma edota lema nahikoa izaten direla gehienetan, eta ondorioz, rol semantikoek eskaintzen duten informazioak nolabaiteko *zarata* sortzen duela mugimenduko seinaleen kategorizazio prozesuan.

Erlazioen identifikazioa

Azkenik, 4.13 taulan erlazio espazialen identifikazioan (1_Erla_ID, 2_Erla_ID eta 3_Erla_ID) SRLren eragina adierazten duten balioak bildu dira.

		Doitasuna		Estaldura		F_1		Accuracy		
Ataza	**Lotura**	*SRL*	*¬SRL*	*SRL*	*¬SRL*	*SRL*	*¬SRL*	*SRL*	*¬SRL*	**Hobekuntza (F_1)**
	<MOVELINK>	61.22	56.12	53.08	48.98	56.86	52.30	61.08	56.32	+4.56
1_Erla_ID	**<QSLINK>**	50.89	47.76	50.05	46.88	50.46	47.31	52.83	49.76	+3.15
	<OLINK>	51.47	46.51	50.02	46.91	50.73	46.70	52.93	47.01	+4.03
	<MOVELINK>	74.31	74.31	51.91	51.91	61.12	61.12	84.77	84.77	0
2_Erla_ID	**<QSLINK>**	45.18	44.11	50.00	48.62	47.46	46.25	90.36	90.12	+1.21
	<OLINK>	45.68	44.91	50.00	48.72	47.74	46.73	91.37	91.01	+1.01
	<MOVELINK>	74.31	74.31	51.91	51.91	61.12	61.12	84.77	84.77	0
3_Erla_ID	**<QSLINK>**	45.18	45.18	50.00	50.00	47.46	47.46	90.36	90.36	0
	<OLINK>	70.69	70.69	50.05	50.05	58.60	58.60	91.37	91.37	0

Taula 4.13: Rol semantikoek lotura espazialen identifikazioan (1_Erla_ID, 2_Erla_ID eta 3_Erla_ID azpiatazetan) duten eraginaren neurketaren emaitzak.

4.13 taulan lehen begiratuan antzeman daitekeenez, hirugarren konfigurazioko (3_Erla_ID) <MOVELINK>, <QSLINK> eta <OLINK> lotura moten detekzioan eta bigarren konfigurazioko (2_Erla_ID) <MOVELINK> loturen identifikazioan rol semantikoek ez dute inongo eraginik. Hori horrela da ondoko hiru arrazoiengatik:

- <MOVELINK> motako erlazioak ezartzeko <MOTION_EVENT> etiketak erabiltzen direlako.

- <QSLINK> eta <OLINK> erlazioak ezartzeko <SPATIAL_SIGNAL> etiketak eta hauen semantic_type atributuak baliatzen direlako.

- Bigarren konfigurazioan osagai espazialak, atributurik gabe, anotatuta ematen direlako eta hirugarrenean, aldiz, osagaiez gainera hauen atributuak ere ematen direlako.

1_Erla_ID azpiatazari eta 2_Erla_ID azpiatazako <QSLINK> eta <OLINK> erlazio espazialen emaitzei egingo diegu kasu beraz. 4.13 taulak adierazten duen moduan, lehenbiziko konfigurazioan <MOVELINK> da hiru erlazio motetatik rolei esker hobekuntzarik handiena jasaten duena, 4.56 puntu F_1 neurrian. Gure iritziz hau horrela da mota horretako loturak, beste biak ez bezala, osagai motan oinarrituta (<MOTION_EVENT>) etiketatzen direlako, osagaiaren inongo atributuren balioari erreparatzeko beharrik izan gabe. Bigarren konfigurazioari dagokionez, aldiz, taulan <QSLINK> eta <OLINK> loturek, rolei esker, antzeko hobekuntza jasaten dutela ikusten ahal da, 1.21 eta 1.01 puntu F_1 neurrian.

4.3 *VisualSpace*: bisualizaziorako interfazea

X-Space tresnaren garapenean sarritan aztertu behar izan dugu *ISO-Space* formatuko fitxategietan bildutako informazioa. Azterketa hau erraztu eta informazioa lehen begi kolpean hobeki interpretatu ahal izateko, *VisualSpace* izendatu dugun interfazea inplementatu dugu, zein *VisualTime* interfazearen espaziorako egokitzapena baita. Modu honetara eta aparteko aplikaziorik instalatzeko beharrik izan gabe, *ISO-Space* formatuko informazioa zuzenean nabigatzailearen bitartez ikuskatzeko aukera eduki dugu. 4.12 irudian ageri da *VisualSpace* interfazearen adibidea. Bertan *TripAdvisor*[18] web orrialdetik jasotako bezero baten iritzia ageri da. Hain zuzen ere *Arizona Grand resort & Spa* izeneko hotelaren inguruan bezero horrek idatzitako iritzia *X-Space* tresnaren bitartez prozesatu ondoren sortutako *ISO-Space* formatuko fitxategiaren edukia aurkezten da.

[18] https://www.tripadvisor.com

We had a sales conference at the Arizona Grand resort & Spa. I landed and plopped my bags in my room and headed to the top of the hill to the Spa. A wonderful, relaxing, citrus scrub (April Special) and massage and I was set for the week. I loved my ground level suite with seperate sleeping area (so quiet!) Beautiful golf course and grounds. The food was less than stellar, Las Palmas & the Steakhouse, not impressive at all. The catering was plentiful and fresh. I love the outdoor seating and fire pits, very relaxing stay.

Irudia 4.12: *VisualSpace* interfazearen adibidea.

Irudian ikusten den bezala, osagai espazialak kolore ezberdinetako puntuen bitartez eman dira aditzera: lekuak berde ilunez, bideak laranja argiz, entitate espazialak

gorriz, gertaera dinamikoak urdin argiz, gertaera estatikoak urdin ilunez, espazio seinaleak laranja ilunez eta mugimendu seinaleak berde argiz. Gainera, hauen arteko erlazio espazialak ere markatuta daude, puntuen arteko lerroak baliaturik.

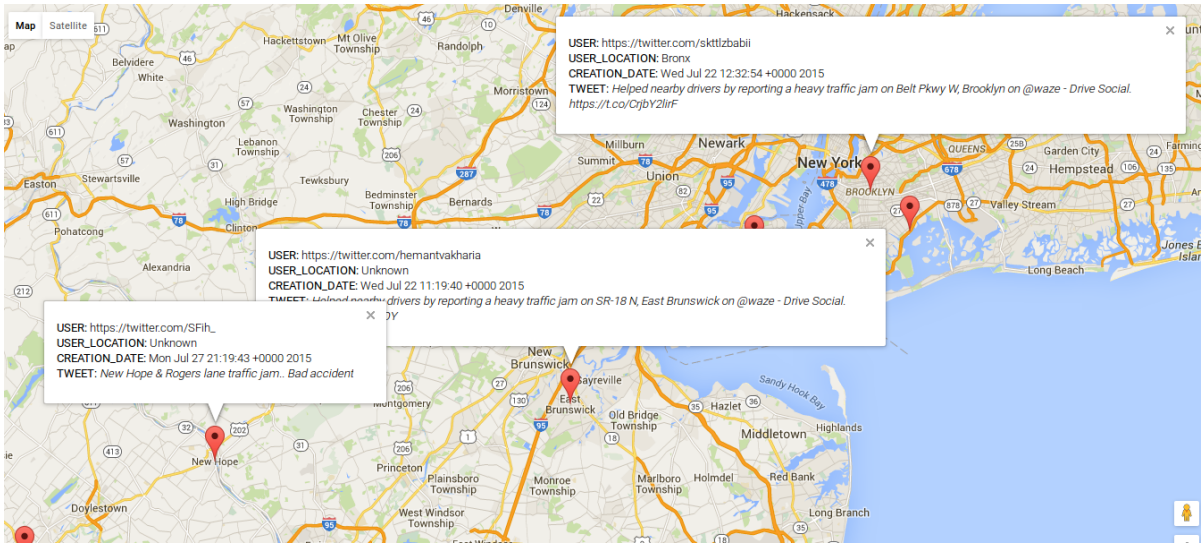
Kontuan izan behar da irudian ageri dena *X-Space* tresnaren erabilera erakusten duen kasu erreala dela, eta ez dela, hortaz, eskuz anotaturik balego bezain akasgabea izango. Etorkizunean, *VisualSpace* `<MOVELINK>`, `<QSLINK>` eta `<OLINK>` erlazio motak ere kolore desberdinen bidezko lerroak erabilita adierazteko egokitu beharko dugu. Egokitzapen honek interfazean hiru loturak elkarrengandik bereizi ahal izatea ahalbidetuko du.

4.4 *ARTSSID*: jazoerak identifikatzeko tresna

*X-Space*k etiketatzen duen informazio espazialak izan ditzakeen aplikazioak erakusteko asmoarekin garatu dugu *ARTSSID* izeneko tresna (*A Real-Time Spatially-Smart Incident Detector*). Honek eremu geografiko jakin batean (edo mundu guztian) jazotzen diren gertakariak denbora errealean detektatzeko gaitasuna dauka (*trafiko buxadurak*, *istiluak*, *auto istripuak*, *suteak* eta abar). Hori lortzeko *Twitter*[19] sare sozialean argitaratzen diren *txioek* eskaintzen duten informazio espaziala *X-Space* baliatuz etiketatu, etiketa hauek geokodetu eta gertakariak mapa batean kokatzen ditu. *ARTSSID*ek sarrerako bi parametro jasotzen ditu: detektatu nahi den gertakari mota eta eremua. Horrela, tresna baliagarria izan daiteke, esate baterako, New Yorken metropoli-barrutiko trafiko buxadurak identifikatzeko edo Europan barnako gripe kasuak detektatzeko.

Sistemak hiru pauso ematen ditu exekuzio bakoitzean. Lehenbizi *Twitter* sare sozialera konektatzen da honek eskaintzen duen *Streaming Public API*[20] izeneko interfazea erabilita. Gero, *X-Space* erabiltzen da aurreko pausuan eskuratutako (sarrera parametroen araberako) txioen espazio informazioa etiketatzeko. Azkenik, *X-Space*k ezarritako etiketak geokodetzen dira *Google*n *Geocoding API*[21] interfazearekin. 4.13 irudian ikus daitezke New Yorken metropoli-barrutian 2015. urteko uztailaren bigarren hamabostaldian detektatutako trafiko buxadura batzuk.

[19] `https://twitter.com/`
[20] `https://dev.twitter.com/streaming/public`
[21] `https://developers.google.com/maps/-/documentation/geocoding/intro`



Irudia 4.13: *ARTSSID* tresnarekin New Yorken identifikatutako trafiko buxadurak.

4.4.1 Ebaluazioa

Sistema ebaluatzeko *traffic jam* gakoarekin *ARTSSID*ek identifikatutako 150 txio aukeratu ditugu. Hauek hogeita lau ordutan zehar mundu guztian barna detektatutako trafiko buxadurei dagozkie. Txioak eskuratu ahal izateko lau eta bost ordu bitartean banatutako bost exekuzio egin ditugu (30 txio exekuzio bakoitzean). Txio bilketa bost exekuziotan banatzeko arrazoia herrialde ezberdinetako txioak detektatzea izan da. Izan ere, denbora zonaldeak direla eta *ARTSSID* exekutatzen den orduaren arabera aldatu egiten da herrialde bakoitzetik jasotako txio kopurua.

Ebaluazioa eskuz egin dugu. Honetan trafiko buxadura bakoitza mapan ongi kokatu den edo ez, hau da, *ARTSSID* tresnak ongi identifikatu duen edo ez egiaztatu da. Egiaztapena egin ahal izateko gure sistemak itzulitakoa *HERE*[22] konpainiak eskaintzen duen trafiko egoeraren inguruko informazioarekin alderatu dugu. Konpainia honen zerbitzuak 33 herrialderen hiririk handienetako trafikoaren denbora errealeko egoera erakusten du. Ebaluazioaren emaitzak 4.14 taulan bildu ditugu. Bertan ikus daitekeenez sistemarekin eskuratutako 150 txioetatik 72 dira espazio informazioa daukatenak eta *X-Space* tresnarekin etiketatu eta mapan kokatu ahal izan ditugunak. Horrek esan nahi du gainerako 78 txioek deskribatzen dituzten buxadurak ezin izan direla mapan kokatu. Hiru arrazoi izan

[22] `https://maps.here.com/traffic`

daitezke horretarako: espazio informaziorik ez edukitzea (46), espazio informazioa izan baina *X-Space*k hau ezin etiketatzea (20) edo espazio informazioa izan eta *X-Space*kin etiketatuta ere hau ezin geokodetzea (12).

Txio multzoen deskribapena	**# (%)**		
Osotara	150 (% 100)		
(Informazio espaziala)∧(*X-Space*)∧(Geokodetzea)∧(¬zuzena)	30 (% 20)	42 (% 28)	72 (% 48)
(Informazio espaziala)∧(*X-Space*)∧(Geokodetzea)∧(zuzena)∧(*HERE*)	24 (% 16)		
(Informazio espaziala)∧(*X-Space*)∧(Geokodetzea)∧(zuzena)∧(¬*HERE*)	18 (% 12)		
(¬Informazio espaziala)	46 (% 31)	32 (% 21)	78 (% 52)
(Informazio espaziala)∧(¬*X-Space*)	20 (% 13)		
(Informazio espaziala)∧(*X-Space*)∧(¬Geokodetzea)	12 (% 8)		

Taula 4.14: ARTSSID tresnaren ebaluazioa buxadurak identifikatzen. (¬: ez, ∧: eta)

Espazio informazioa daukaten, *X-Space* tresnarekin etiketatu eta mapan kokatu diren 72 txioei dagokienez, horietako 42 ongi eta 30 gaizki identifikatu direla egiaztatu ahal izan dugu. Ongi dauden 42 txioetatik 24 *HERE*k eskainitako zerbitzuan aurkitu ahal izan ditugu eta beste 18 ez. Azkeneko horiek *HERE*k lantzen ez dituen zonaldeetan kokatu dira (garatzeko bidean dauden herrialdeak, herriak, hiri txikiak eta abar).

Ebaluazioaren emaitzetatik, beraz, bi gauza ondoriozta daitezke: *ARTSSID* oraindik ere hobetzea behar duen sistema dela, oraindik ere lantzeko beharra daukana eta *HERE* moduko zerbitzuek kontuan hartzen ez dituzten urruneko edota garatu gabeko lekuetan gerta daitezkeen buxadurak identifikatzeko baliagarria dela.

4.5 Ondorioak eta etorkizuneko lanak

Atal honetan *X-Space* aurkeztu da, ingelesezko testuetan aurki daitekeen informazio espaziala automatikoki eta *ISO-Space* anotazio eskema baliaturik etiketatzen duen tresna. Sistema *SpaceEval* izeneko saioan ebaluatu da eta lortutako emaitzak onak izan dira.

Tresnaren emaitzak analizatu ditugunean esan dugun moduan, aipatutako ebaluazio saioan lantzen diren hamar azpiatazatatik lautan eman ditu gure sistemak baliorik gorenak. Hori lortu ahal izateko, besteak beste, *WordNet* hurbilpena deitu dugun teknika aplikatu dugu. Metodo hau berritzailea da baliabide linguistiko hau (*WordNet*) ibiltzen den lehen aldia delako, *ISO-Space* eskema jarraituta informazio espaziala eskuratu ahal izateko. Gainera, atal honen hasieran adierazi dugun eran, tesian *ISO-Space* hartzeko arrazoietako bat denboraren etiketatze automatikoaren arlotik ikasitakoa da. Izan ere,

aurreko atalean aipatu denez, ataza tenporalean, diseinu aldetik *ISO-Space*ren aurrekaria den *ISO-TimeML* arloko estandar bihurtu zen urte gutxitan. Hori kontuan edukita alde batetik, eta eskema espazialen arloan izandako mugimenduak ikusita bestetik ($SpRL \subset$ *ISO-Space*), *X-Space* tresnaren inplementaziorako *ISO-Space* erabiltzea erabaki genuen.

Azaldu dugu tesiaren hasieran planteatutako helburuetako bat euskarazko testuetan egin nahi zen informazio espazialaren etiketatze automatikoan rolek zuten eragina neurtzea izan zela, denborarekin egindako moduan. Hala ere, baliabide kontuak direla medio, hori ezinezkoa gertatu zaigu, eta horregatik garatu dugu *X-Space* ingeleserako. Etorkizunerako asmoa da beharrezkoak diren gidalerroak eta corpusa prest daudenean, *X-Space* euskararako egokitzea, era horretara SRLren eragina euskaraz ere neurtu ahal izateko. *X-Space* tresnak ingeleserako balio izan digu, baina ez euskararako, hots, tesi lan honetako bi hipotesietako bat egiaztatzeko baizik ez digu balio izan. Gure neurketek erakutsi duten gisan, rol semantikoen eragina esanguratsua da etiketatze espazialak dauzkan azpiataza guztietan; eraginik esanguratsuenekoak, bestalde, osagaien kategorizazioa eta erlazioen identifikazioa izan direla ere ikusi ahal izan da.

*X-Space*ren hurrengo bertsioaren garapena ere etorkizuneko lanetan kokatzen da. Honek hemen aurkeztutako bertsioaren aldean izango duen berrikuntzarik azpimarragarriena atributuen esleipenerako ezaugarri multzo egokituak inplementatzea izango da.

5

ONDORIOAK ETA ETORKIZUNEKO LANAK

Azkeneko atal honetan tesian ikertu ditugun hiru aztergaietatik atera ahal izandako ondorioak aurkezten dira. Ataza hauek hizkuntzaren prozesamenduaren barnean zeresana eman duten, eta oraindik ere ematen ari diren, SRL eta denboraren eta espazioaren etiketatze automatikoa izan dira. Haietako bakoitzari atal bana eskaini diegu ikerketa lan honetan. Horregatik, ondorioak ere, aztergai bakoitzari dagozkionak, horrela aurkeztuko dira, azpiataletan banatuta. Etorkizuneko eginbeharrak, berriz, bukaeran zerrendatzen ditugu aparteko azpiatal batean.

Ikerlan hau euskaraz idatzitako testuen prozesamendu semantiko automatikoan aurrera egiteko asmoarekin bideratu dugu. Lanaren ardatza *gertaera* kontzeptua izan da. Gertaerak dira, hain zuzen, SRLren eta leku-denborazko etiketatze konputazionalaren oinarria. Izan ere, azaldu dugu jazoerak eta haiei dagozkien *predikatu-argumentu-adjuntu* egiturak eta hauen propietate espazial eta tenporalak direla testuetako semantika egituratzen dutenak.

5.1 Rol semantikoen etiketatze automatikoa

SRL ataleko ikerketaren fruitu den *bRol* dependentzia sintaktiko-semantikoen etiketatzailearen garapenaren aurretik *SRL prototipoa* deitu duguna inplementatu dugu. Azken hau, besteak beste, *bRol* sistemaren diseinuan munta handikoak izan zaizkigun ondorioak ateratzeko baliatu ahal izan dugu. *bRol* eta honen prototipoa garatzetik ateratako ondorioak garrantzizkoak dira tresna horiek euskarazko SRL atazaren ikerketaren egungo egoera lehen aldiz finkatzeko balio izan dutelako. Euskaraz rol semantikoak automatikoki etiketatzeko ezaugarririk baliotsuenak zein izan diren ondorioztatu ahal izan da besteak beste. Atal honen ondorioak jarraian zerrendatutakoak dira:

- Gure esperimentuetan *PropBank* corpusak proposatzen duen ereduak, tesi lanean zehar *PropBank eredua* deitu dugunak, itzuli ditu emaitzarik altuenak. Rol semantikoak etiketatzeaz arduratzen diren sistemak dituzten hizkuntza gehienek (ingelesak, alemanak, gaztelerak, italierak, txinerak eta abarrek) eredu hau erabili izan dute azkeneko urteotan. Hau ingeleseko *PropBank* corpusak gainerako hizkuntzetan (eta zehazki hauen baliabideen garapenean) izan duen eraginaren eta ikasketa automatikoko metodoek *PropBank*eko rol multzoarekin (arg$i : 0 \leq i \leq 4$), *VerbNet* ereduko multzoaren aldean, duten eraginkortasunaren ondorio da. Hori jakinda, eta euskarako SRL prototipoaren garapenean ere *PropBank* rol multzoarekin *VerbNet* rol multzoarekin baino emaitza hobeak lortzen direla kontuan hartuta (F_1 neurria 84.3 eta 82.9 puntu hurrenez hurren), euskararako ere *PropBank* eredua jarraitzea erabaki dugu. Eredu hori hartzeak, gainera, tesi lanean sortu dugun *bRol* SRL etiketatzaile automatikoa beste hizkuntza batzuenekin alderatzea ahalbidetu du.

- Rol semantikoak etiketatzeko prozesuaren azkeneko urratsa den argumentuen sailkapenean, *Support Vector Machines-SVM* ikasketa automatikoko algoritmoak *Decision Trees-DT* eta *Random Decision Trees-RDT* algoritmoek baino emaitza hobeak itzultzen dituzte egindako ikerlanean, bai *PropBank* ereduan eta baita *VerbNet* ereduan ere. Gure esperimentuetan *PropBank* eta *VerbNet* rolak esleitzen dituzten sailkatzailerik eraginkorrenak *SVM* algoritmoarekin lortu ditugu. Emaitzarik apalenak itzultzen dituen ikasketa algoritmoa, berriz, *RDT* izan da. Kalkulatu ahal izan dugunez, *SVM*ren aldean *PropBank* ereduarentzat ia zazpi puntu jaisten da $F1$ neurria, eta *VerbNet*entzat hamar puntu baino gehiago.

- Gure ikerketan euskaraz argumentuen sailkapena egitean eraginik positiboena daukaten ezaugarriak *predikatuaren lema*, *deklinabide kasua*, *funtzio sintaktikoa* eta *argumentuaren lema* dira, edozein rol multzo erabilita ere. Ezaugarri linguistikoek rol etiketen esleipen prozesuan duten eragina neurtzean, gainera, ikusi ahal izan dugu *frame semantikoak* bi ereduei negatiboki eragiten diela, eta *VerbNet* ereduan *frame sintaktikoa* eta *izen entitatea* ezaugarriek ere emaitzak okertzen dituztela. Hori jakinda, esperimentua egin dugu, eragin negatibokoak ezaugarri multzotik kenduta, eta ezaugarri guztiak baliaturik baino emaitza kaskarragoak erdiesten direla egiaztatu dugu. Horregatik, ezaugarri guztien uztardurak argumentuen sailkapenean eragin positiboa daukala ondorioztatu ahal izan dugu.

- *PropBank* ereduko SRL sistemek emaitza hobeak lortzen dituzte entrenamendu corpuseko `arg0` eta `arg1` *core* rolek agente eta paziente prototipikoak (Dowty, 1989) adierazten dituztenean. Ingeleserako *PropBank* corpusean horrela egin zen, baina euskararako *EPEC-Rolsem* corpusean ez. Ondorioz, `arg0` eta `arg1` rolak etiketatzea zailagoa izan da euskaraz ingelesez baino, eta emaitzak bestela baino apalagoak izan dira.

Lau horiek izan dira *SRL prototipoaren* garapenetik eta honekin lotutako esperimentuetatik atera ahal izan ditugun ondorioak. Jarraian *bRol*en inplementaziotik, hots, euskararako garatu den aurreneko SRL tresna guztiz automatikoaren inplementaziotik ondorioztatu duguna azaltzen da:

- Euskarazko SRLren kasuan bezala, lagungarria da, tamaina mugatuko baliabideak dauzkaten hizkuntzetarako tresnak garatzeko garaian teknika berritzaileak aplikatzea, hauen emaitzak beste hizkuntzenekin parekatzen saiatzeko. Teknika berritzaile hauek corpus txikiak handitzeko, adibide gehiago sortzeko edo, gure sistemaren kasuan egin dugun bezala, estaldura handitzeko erabil daitezke. *bRol* sistema inplementatzean, predikatuen desanbiguazioaren urratsean, *Itzulpen Modulua* izenekoa gehitu dugu eta, guk dakigula, lehenbiziko aldia izan da SRLri teknika hau aplikatu zaiona. Baliabide faltaren ondorioz modulu horren irteera ebaluatu ezin izan badugu ere, eskuz egindako zenbait etiketatzeren eskuzko ebaluazio partzialarekin modulu horrek sistemaren estaldura hobetzen duela ikusi ahal izan da. Horren arrazoia da itzulpen moduluak bestela etiketatuko ez liratekeen predikatuen desanbiguazioa eta haien argumentuen eta adjuntuen etiketatzea eragiten duela.

Kontuan izan behar da ondorengo aldea dagoela *bRol* garatzeko erabili den corpusaren eta *CoNLL-2009* ebaluazio saioan beste hizkuntza batzuetarako erabili zirenen artean: *EPEC-RolSem* corpuseko *train* zatian esaldi kopurua % 71.1 eta token kopurua % 80.1 txikiagoak dira *CoNLL-2009* saioko entrenamendu corpusen batezbestekoarekin alderatuta.

- Orokorrean, euskara ez da morfologikoki aberatsak diren beste hizkuntza batzuk baino zailagoa semantikoki analizatzeko garaian. Euskararako corpusa *CoNLL-2009* ebaluazio saioko MRL hizkuntzen corpusekin alderatzen dugunean, rol eta adjuntu etiketek *EPEC-RolSem* corpusean daukaten banaketak eta *Part-of-Speech* kategoria motek edo `FEAT` ezaugarri kopuruek ez dute zailtasun berezirik, beste hizkuntzetakoenen aldean, euskaraz dependentzia semantikoak etiketatzean. Izan ere, etiketek corpusean duten banaketak hizkuntza bakoitza etiketatzearen zailtasuna adierazten dute. Japonieraren kasuan, adibidez, hiru dependentzia sintaktiko mota (hiru `DEPREL` etiketa) baizik ez dira erabiltzen sintaxian eta ondorioz japoniera da ebaluazio saio osoan sintaktikoki emaitzarik altuenak erdiesten dituen hizkuntza.

5.2 Denbora informazioaren etiketatze automatikoa

Tesi lan honetako hurrengo atala euskaraz idatzitako testuetan aurki daitekeen informazio tenporalaren etiketatze automatikotik ateratako ondorioei eskaini diogu. Ataza honetan *bTime*, *end-to-end* motako tresna, garatu dugu. Honek *bRol*ek eskaintzen duen informazio semantikoa erabili du ezaugarritako, eta ondorioz, rol semantikoek euskarazko denboraren etiketatze automatikoan izan duten eragina neurtu ahal izan dugu. Jarraian aurkezten dira atal honetan egindako esperimentuetatik ateratako ondorenak:

- Gaur egun euskarazko etiketatze tenporal automatikorako dagoen corpus bakarraren tamaina (oso) mugatua kontuan edukirik (*Euskal-TimeBank*) erregeletan oinarritutakoa da denbora adierazpenen etiketatzerako dagoen hurbilpenik egokiena. Tesi lanean garatu dugun *bTime* sisteman ikasketa automatikoa eta (erregeletan oinarritzen den) *HeidelTime* teknikak erabiltzen dituen tresnaren bi bertsio inplementatu dira adierazpen tenporalak etiketatzeko. Hauen emaitzak erkatzean ikusi ahal izan da 27.55 puntuko aldea dagoela (F_1 neurrian) bi tekniken artean, adierazpen tenporalak identifikatu eta *strict* eskemaren arabera ebaluatzeko garaian

(*HeidelTime*ek hobekien). Adierazpenek jasotzen duten `type` atributuaren kasuan ere aldea nabarmena da bi metodoen artean: F_1 neurrian 25 puntuko desberdintasuna neurtu ahal izan dugu gure esperimentuetan. Azkenean, ikasketa corpusak oso txikiak direnean, zaila da ikasketa automatikoko algoritmoek ezer esanguratsurik ikastea, eta gure esperimentuetan hori gertatu dela uste dugu. ML sistema automatikoek emaitza apalak lortu dituzte ikasketa corpus txikiekin elikatu ditugulako, eta hortaz emaitza onenak, diferentzia handiarekin gainera, eskuzko erregeletan oinarritutako sistemak eman ditu.

- Gertaerek euskaraz hartzen dituzten zortzi atributuei balioak esleitzeko tenorean erabiltzeko testuinguru leihoen tamainarik egokienak hiru tokenek osatutakoak dira. *bTime* tresnan probak egin ditugu, ingeleserako (Jung eta Stent, 2013) argitalpenean bezala bat, hiru, zazpi eta hamabost hitzeko testuinguru leihoekin, eta gure emaitzek argi erakutsi dute hirukoak direla egokienak, gure esperimentuek deskribatutako egoeran. Ondorioztatu ahal izan dugu atributuen esleipenak hiru hitzeko leihoekin emaitzarik onenak itzultzeko arrazoia gertaeren izaera gramatikala finkatu ahal izateko beharrezkoa den testuinguru linguistikoa osatzen duten token kopuruaren ondorena dela. Hau da, atributuak finkatu ahal izateko garrantzizkoa dela gertaera barnean hartzen duen sintagmari ohartzea, baita bertako tokenei eta, oro har, horiek osatzen duten egitura sintagmatikoari ere.

- *bTime* tresnari adjektiboen nominalizazioak diren izen-predikatuek osatu zerrenda gehitzeak sistemaren emaitzen hobekuntza dakar berekin. Zerrenda osatzeko *IXA* taldean garatzen ari den euskarazko *NomBank* baliabidea erabili dugu, eta zehazkiago, honen barnean daudenetatik, *-(t)asun* eta *-(k)eria* bukaerak baliatuta EDBL datu base lexikaletik erauzi diren 2.728 izen hartu ditugu. Horrela, gertaeren identifikazioan iritsitako emaitzak bi puntu baino gehiago hobetzea lortu da. Gainera, erlazio tenporalen kategorizazioan eta gertaerek hartzen dituzten zortzi atributuetatik lautan ere emaitzak hobetu egin dira.

- Ingelesez eta gaztelaniaz gertatzen den bezala, rol semantikoek denboraren etiketatze automatikoan duten eragina, euskaraz ere, oro har, eta tesi lanean planteatu dugun egoeran, positiboa da. Gure neurketek erakutsi dute *bRol* sistemak etiketatzen dituen rol semantikoak *bTime*n ezaugarritako erabilita emaitzak hobetzea

lortzen dela. Izan ere, ikasketa automatikoaren bitartez hurbildutako *end-to-end* arkitekturaren hamabi azpiatazatatik bederatzitan balioak altuagoak izan dira, rolei esker.

5.3 Espazio informazioaren etiketatze automatikoa

Tesi laneko azken atalean *X-Space* tresna garatu dugu, ingelesez idatzitako testuetan aurki daitekeen informazio espaziala etiketatzeko. Azaldu dugun moduan, ezin izan da etiketatzaile hau euskararako ere sortu gaur egungo baliabideen egoerak ahalbidetzen ez duelako. Alabaina, ingeleserako proposatzen dugun sistemarentzako arkitektura, eta bertan gauzatzen den hurbilpenak euskararako ere erabiltzen ahalko dira etorkizunean. *X-Space* sistema baliatuta, aurreko atalean denborarekin egin dugun moduan, rol semantikoek etiketatze espazialaren atazan duten eragina neurtu ahal izan dugu. Ondoren aurkezten direnak dira atal honetatik ateratako ondorioak:

- *ISO-Space*k zehazten dituen osagai espazialen kategorizazioa egin ahal izateko *WordNet* datu-basea eta ikasketa automatikoa uztartzen dituen hurbilpena baliagarria dela uste dugu. *X-Space* inplementatu dugunean aipatu datu-basea erabili dugu lekuen, bideen eta gertaera dinamikoen *domeinuak* zehazte aldera. Hautagaien artetik mota horietako osagaiak identifikatu ahal izan ditugu domeinu hauekin.

  Kategorizazioan gure sistemarenak *SpaceEval* ebaluazio saioan parte hartu zuten gainerako sistemen emaitzekin erkatzean ikusten da gure tresnarenak, emaitza gorenak iritsi dituen sistematik hurbil gelditzen direla. Hori dela eta, guk proposatu dugun *WordNet* hurbilpen berritzailea osagai espazialak kategorizatu ahal izateko teknika egokia dela pentsa dezakegu.

- Osagai eta erlazio espazialen atributuak etiketatzeko, beharrezkoa da haietako bakoitzari hobekien egokitzen zaion ezaugarri linguistikoen multzoaren edota tekniken azterketa egitea. Adierazi dugunez, epe kontuak direla eta, osagai espazialen atributuen kasuan *X-Space*k ezaugarri eta ikasketa algoritmo berak erabiltzen ditu denentzat. Bestalde, azpiataza hau burutzeko gai den beste sistemak ere metodo hau bera erabiltzen du, baina ezaugarri linguistiko bat bakarra erabilita. Gure tresnaren emaitzak azken honenak baino hobeak gertatu badira ere, oso apalak

izan dira bienak. Horregatik ondorioztatzen dugu atributu bakoitzari egokitutako ezaugarriak bilatzea behar-beharrezkoa dela, haien etiketatzean lortutako balioak hobetu nahi badira.

- Rol semantikoek ingelesez idatzitako testuetan aurki daitekeen informazio espazialaren etiketatze automatikoan duten eragina positiboa da. Gure neurketen arabera, etiketatze mota honetatik eraginik handiena jasotzen duten azpiatazak osagai espazialen kategorizazioa eta loturen identifikazioa dira. Lehen azpiatazaren barnean, gainera, lekuak, bideak eta entitate espazialak dira rolei esker, emaitzak gehien hobetzen dituzten osagaiak. Denborarentzat, euskararako ikusi duguna eta denboraren etiketatzerako erabiltzen den *ISO-TimeML* eskemak espazioko *ISO-Space* eskemarekin dauzkan antzekotasunak aintzat edukirik, hau ondoriozta dezakegu: seguruenik euskarazko etiketatze espazialean ere, denborarenean bezala, rolek izango duten eragina positiboa izango dela.

5.4 Etorkizuneko lanak

Etorkizuneko eginbeharrak hainbat dira. Lan hauetako gehienek lotura daukate aurkeztu ditugun *bRol*, *bTime* eta *X-Space* sistemei funtzionalitateak gehitzearekin, edo horien barneko sailkatzaile edo teknikak hobetzearekin. Beste lanek, ordea, esperimentu berriak egitearekin edo aipatu sistemetan oinarriturik beste tresna batzuk sortzearekin dute zerikusia. Jarraian zerrendatzen ditugu etorkizunerako aurreikusten ditugun egitekorik esanguratsuenak:

1. *bRol*ek izen predikatuekin lan egiteko gaitasuna inplementatu ahal izatea gustatuko litzaiguke. Horrela, predikatu horien argumentu eta adjuntuak, eta hauei dagozkien rol semantiko eta adjuntu etiketak automatikoki etiketa litezke, eta euskararako SRL tresnaren estaldura handitu, horren ondorioz. Hau lortzeko bi gauza dira beharrezkoak: *Basque Verb Index-BVI* euskarazko aditzen predikatu-lexikoian izen predikatuei dagozkien azpikategorizazioak ere gehitzea batetik, eta azpikategorizazio hauen araberako euskarazko *NomBank* corpusa anotatzea bestetik. Aipatu dugu *Euskal-NomBank* baliabidea garapen prozesuan dagoela gaur egun; garapena amaitzen denean, *EPEC-RolSem* corpuseko izen predikatuak eta hauen *argumentu-adjuntu* egiturak izango ditu anotatuta.

2. Tesi laneko SRL atalarekin lotuta egin nahiko genituzkeen beste gauza batzuk hauek dira: a) *bRol* sistemaren domeinuz kanpoko ebaluazioa egitea, b) predikatuen desanbiguazio urratseko *Itzulpen Moduluaren* eraginkortasuna aztertzea, c) *bRol VerbNet* ereduko rolak esleitu ahal izateko ere egokitzea, d) aditz eta izen predikatuak ez ezik, adjektibo eta adberbio predikatuak eta hauetatik sortzen diren dependentzia semantikoak ere etiketatzeko gaitasuna gehitzea, eta e) sailkatzaileen eraikuntzarako teknika berriak aplikatzea (*deep learning* edo ikasketa sakona esate baterako).

3. *bTime* tresnak dokumentu bateko gertaeren eta *DCT*aren arteko lotura tenporalak (<TLINK>) etiketatzeko gaitasuna izateaz gainera, beste osagai batzuen artekoak eta beste mota bateko erlazioak etiketatzeko aukera izatea ere gustatuko litzaiguke. *ISO-TimeML* eskema aurkeztean esan dugun moduan, erlazio tenporalak gertaeren artekoak izan daitezke, edota gertaera baten eta denbora adierazpen baten artekoak. Bestalde, badira aspektuzko eta mendekotasunezko loturak deriztenak (<ALINK> eta <SLINK>). Etorkizunean hauek ere automatikoki etiketatzea gustatuko litzaiguke, *bTime* erabilita. Horretarako beharrezkoa izango da *Euskal-TimeBank* corpusa zabaldu eta horrela lotura hauei dagozkien anotazioak ugaritzea. Denbora seinaleak markatzen dituzten <SIGNAL> etiketak ezartzeko gaitasuna ere gehitu nahi diogu *bTime*ri.

4. *bRol* tresnarako planteatu dugun gisa, *bTime* etiketatzailerako ere badago aukera etorkizunean euskara ez diren beste hizkuntzak prozesatu ahal izateko egokitzeko. Printzipioz behintzat, eta gaur egungo baliabideei erreparatuta, hizkuntza hauek ingelesa, gaztelania eta italiera izango lirateke.

5. *X-Space* etiketatzailea euskarazko testuetan informazio espaziala etiketatzeko egokitu nahi genuke. Hau ezinezkoa balitz, etiketatzaile honen arkitektura eta hurbilpenak erabiltzen dituen (eta *bSpace* deituko genukeen) euskararako tresna inplementatzen saiatuko ginateke. Hori egin ahal izateko, baliabide linguistikoei dagokienez, beharrezkoa izango da *ISO-Space* eskema euskararako egokitu eta honen araberako *Euskal-SpaceBank* corpusa anotatzea. Honek euskarazko testuetako informazio espazialaren etiketatze automatikoan *bRol*ek esleitutako rol semantikoek daukaten eragina neurtzeko balioko liguke.

6. Hizkuntzaren prozesamenduaren arloko egungo joera kontuan izanda, sarrera *multimodala*, zehazki irudia eta testua uztartzen dituen sarrerako informazioa, prozesatzeko gaitasuna daukan *X-Space* tresnaren bertsioa garatu gogo genuke. Gainera, hizkuntzaren tratamendu konputazionaleko ikerkuntza joeren ildotik jarraituta, ikasketa sakona (*deep learning*) edo sare neuronaletan oinarritutako teknikak gure sisteman inplementatzea ere gustatuko litzaiguke.

BIBLIOGRAFIA

Aduriz, I., Aranzabe, M. J., Arriola, J. M., Atutxa, A., de Ilarraza, D. A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. (2006). Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15.

Aduriz, I. et al. (2003). Construction of a basque dependency treebank.

Agerri, R., Agirre, E., Aldabe, I., Altuna, B., Beloki, Z., Laparra, E., de Lacalle, M. L., Rigau, G., Soroa, A., et al. (2014). Newsreader project. *Procesamiento del Lenguaje Natural*, 53:155–158.

Aldezabal, I., Aranzabe, M., Ilarraza, A. D., and Estarrona, A. (2013). A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicative level following the propbank-verb net model. Technical report, UPV/EHU/LSI/TR.

Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., and Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for basque. In *Proceedings of Workshop on"Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation.*

Alegria, I., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named entity recognition and classification for texts in basque. *Proceedings of JOTRI II.*

Altuna, B., Aranzabe Urruzola, M. J., and Díaz de Ilarraza Sánchez, A. (2016a). Euskarazko denbora-egiturak etiketatzeko gidalerroak v2. 0.

Altuna, B.ñ., Aranzabe, M. J., and Diaz de Ilarraza, A. (2016b). Adapting timeml to basque: Event annotation. In *CICLing.*

Atutxa, A., Ezeiza, N., Goenaga, I., and Gojenola, K. (2015). Experiments on semi-supervised dependency parsing of a morphologically rich language.

Badiou, A. and Feltham, O. (2007). *Being and event*. A&C Black.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Baldridge, J. (2014). The opennlp project. 2005.

Bastianelli, E., Croce, D., Nardi, D., and Basili, R. (2013). Unitor-hmm-tk: Structured kernel-based learning for spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 573–579.

Bethard, S. (2013). Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

Boguraev, B., Pustejovsky, J., Ando, R., and Verhagen, M. (2007). Timebank evolution as a community resource for timeml parsing. *Language Resources and Evaluation*, 41(1):91–115.

Bohnet, B. (2009). Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 67–72. Association for Computational Linguistics.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97. Association for Computational Linguistics.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, pages 356–363. Association for Computational Linguistics.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 969–974.

Cantor, G. (2006). Fundamentos para una teoría general de conjuntos. escritos y correspondencia selecta. *LLULL*, 30:344.

Carreras, X. and Màrquez, L. (2004). Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eigth Conference on Computational Natural Language Learning*, pages 106–109. Association for Computational Linguistics.

Carreras, X. and Màrquez, L. (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics.

Caselli, T., Lenzi, V. B., Sprugnoli, R., Pianta, E., and Prodanof, I. (2011). Annotating events, temporal expressions and relations in italian: the it-timeml experience for the ita-timebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151. Association for Computational Linguistics.

Caselli, T., Sprugnoli, R., Speranza, M., and Monachini, M. (2014). Eventi: Evaluation of events and temporal information at evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 27–34. Pisa University Press.

Chambers, N. (2013). Navytime: Event and time ordering from raw text. Technical report, DTIC Document.

Chang, A. and Manning, C. D. (2013). Sutime: Evaluation in tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 78–82, Atlanta, Georgia, USA. Association for Computational Linguistics.

Che, W., Li, Z., Hu, Y., Li, Y., Qin, B., Liu, T., and Li, S. (2008). A cascaded syntactic and semantic dependency parsing system. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 238–242. Association for Computational Linguistics.

Che, W., Li, Z., Li, Y., Guo, Y., Qin, B., and Liu, T. (2009). Multilingual dependency-based syntactic and semantic parsing. In *Proceedings of the thirteenth conference on computational natural language learning: shared task*, pages 49–54. Association for Computational Linguistics.

Choi, J. D. (2012). Optimization of natural language processing components for robustness and scalability.

Chu, Y.-J. and Liu, T.-H. (1965). On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Daumé III, H. (2004). Notes on cg and lm-bfgs optimization of logistic regression. *Paper available at http://pub. hal3. name# daume04cg-bfgs, implementation available at http://hal3. name/megam*, 198:282.

Davidson, D. (1967). The logical form of action sentences.

De Lacalle, M. L., Laparra, E., and Rigau, G. (2014). Predicate matrix: extending semlink through wordnet mappings. In *LREC*, pages 903–909.

Deleuze, G. (1988). Signes et événements. *Magazine littéraire*, 257.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Donelaicio, K., Nivre, J., and Krupavicius, A. (2013). Lithuanian dependency parsing with rich morphological features. In *Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, page 12.

Dowty, D. R. (1989). On the semantic content of the notion of thematic role. In *Properties, types and meaning*, pages 69–129. Springer.

DSouza, J. and Ng, V. (2015). Utd: Ensemble-based spatial relation extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 862–869.

Estarrona, A. (2014). Epec corpusa predikatu-mailan etiketatzeko oinarriak: Epec-rolsem, bvi eta e-rolda.

Estarrona, A., Aldezabal, I., de Ilarraza, A. D., and Aranzabe, M. J. (2015). A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicate level following the propbank-verbnet model. *Digital Scholarship in the Humanities*, page fqv001.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Ferro, L., Mani, I., Sundheim, B., and Wilson, G. (2001). Tides temporal annotation guidelines-version 1.0.2. *The MITRE Corporation, McLean-VG-USA*.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Fillmore, C. J., Ruppenhofer, J., and Baker, C. F. (2004). Framenet and representing the link between semantic and syntactic relations. *Frontiers in linguistics*, 1:19–59.

Foland Jr, W. R. and Martin, J. H. (2015). Dependency-based semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288.

Garbin, E. and Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 363–370. Association for Computational Linguistics.

Gesmundo, A., Henderson, J., Merlo, P., and Titov, I. (2009). A latent variable model of synchronous syntactic-semantic parsing for multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 37–42. Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Goenaga, I., Gojenola, K., and Ezeiza, N. (2013). Exploiting the contribution of morphological information to parsing: the basque team system in the sprml2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 61–67.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Grover, C., Tobin, R., Alex, B., and Byrne, K. (2010). Edinburgh-ltg: Tempeval-2 system description. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 333–336, Uppsala, Sweden. Association for Computational Linguistics.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 5, page 10.

Ha, E., Baikadi, A., Licata, C., and Lester, J. (2010). Ncsu: Modeling temporal relations with markov logic and lexical ontology. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 341–344, Uppsala, Sweden. Association for Computational Linguistics.

Hacioglu, K., Pradhan, S., Ward, W. H., Martin, J. H., Jurafsky, D., et al. (2004). Semantic role labeling by tagging syntactic chunks. In *CoNLL*, pages 110–113.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., et al. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Hajic, J., Panevová, J., Hajicová, E., Sgall, P., Pajas, P., Štepánek, J., Havelka, J., Mikulová, M., Zabokrtskỳ, Z., and Ševcıková-Razımová, M. (2006). Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Hajičová, E. (1998). Prague dependency treebank: From analytic to tectogrammatical annotations. *Proceedings of 2nd TST, Brno, Springer-Verlag Berlin Heidelberg New York*, pages 45–50.

Hirst, G. (1987). Semantic interpretation and the resolution of ambiguity.

Im Walde, S. S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 747–753. Association for Computational Linguistics.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jiang, Z. P. and Ng, H. T. (2006). Semantic role labeling of nombank: A maximum entropy approach. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 138–145. Association for Computational Linguistics.

Joachims, T. (1999). Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4).

Johansson, R., Heppin, K. F., and Kokkinakis, D. (2012). Semantic role labeling with the swedish framenet. In *LREC*, pages 3697–3700.

Johansson, R. and Nugues, P. (2008). Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187. Association for Computational Linguistics.

Jung, H. and Stent, A. (2013). Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 20–24.

Kawahara, D., Kurohashi, S., and Hasida, K. (2002). Construction of a japanese relevance-tagged corpus. In *LREC*.

Kim, J. (1976). Events as property exemplifications. In *Action theory*, pages 159–177. Springer.

Kipper, K., Dang, H. T., Palmer, M., et al. (2000). Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

Kolomiyets, O., Kordjamshidi, P., Bethard, S., and Moens, M.-F. (2013). Semeval-2013 task 3: Spatial role labeling. In *Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 255–266. ACL.

Kordjamshidi, P., Bethard, S., and Moens, M.-F. (2012). Semeval-2012 task 3: Spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 365–373. Association for Computational Linguistics.

Kordjamshidi, P., Moens, M.-F., and van Otterlo, M. (2010). Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 413–420. European Language Resources Association (ELRA).

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4.

Laokulrat, N., Miwa, M., Tsuruoka, Y., and Chikayama, T. (2013). Uttime: Temporal relation classification using deep syntactic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 88–92, Atlanta, Georgia, USA. Association for Computational Linguistics.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45.

Leidner, J. L. (2006). Toponym resolution: A first large-scale comparative evaluation. *Institute for Communicating and Collaborative Systems*.

Lenzi, V. B. and Sprugnoli, R. (2007). Evalita 2007: Description and results of the tern task. In *Proceedings of the Evalita Workshop*.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Lewis, D. K. (1986). *On the plurality of worlds*, volume 322. Cambridge Univ Press.

Lewis, D. K. (1987). *Philosophical Papers: Volume II*. Oxford university press.

Li, H., Strötgen, J., Zell, J., and Gertz, M. (2014). Chinese temporal tagging with heideltime. In *EACL*, pages 133–137.

Llorens, H. (2011). *A semantic approach to temporal information processing*. Universidad de Alicante.

Llorens, H., Saquete, E., and Navarro, B. (2010). Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics.

Llorens, H., Saquete, E., and Navarro-Colorado, B. (2013). Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, 49(1):179–197.

Loper, E., Yi, S.-T., and Palmer, M. (2007). Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX*, volume 98, pages 187–193. Citeseer.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. (2006). I-cab: the italian content annotation bank. In *Proceedings of LREC*, pages 963–968. Citeseer.

Manfredi, G., Strötgen, J., Zell, J., and Gertz, M. (2014). Heideltime at eventi: Tuning italian resources and addressing timemls empty tags. In *Proceedings of the Forth International Workshop EVALITA*, pages 39–43.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools. In *LREC*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

McDonald, R. T. and Pereira, F. C. (2006). Online learning of approximate dependency parsing algorithms. In *EACL*.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). Meantime, the newsreader multilingual event and time corpus. *Proceedings of LREC2016*.

Mirza, P. (2015). Recognizing and normalizing temporal expressions in indonesian texts. In *International Conference of the Pacific Association for Computational Linguistics*, pages 135–147. Springer.

Mirza, P. and Minard, A.-L. (2014). Fbkhlt-time: a complete italian temporal processing system for eventi-evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Muller, P. et al. (1998). A qualitative theory of motion based on spatio-temporal primitives. *KR*, 98:131–141.

Negri, M. (2007). Trattamento di espressioni temporali in italiano: Ita-chronos dealing with italian temporal expressions: The ita-chronos system. pages 58–59.

Nichols, E. and Botros, F. (2015). Sprl-cww: Spatial relation classification with independent multi-class models. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 895–901.

Nilsson, J., Riedel, S., and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932. sn.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer.

Otegi, A., Ezeiza, N., Goenaga, I., and Labaka, G. (2016). A modular chain of nlp tools for basque. In *International Conference on Text, Speech, and Dialogue*, pages 93–100. Springer.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition, linguistic data consortium. Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

Parsons, T. (1990). Events in the semantics of english: A study in subatomic semantics.

Peterson, D., Palmer, M., and Wu, S. (2014). Focusing annotation for semantic role labeling. In *LREC*, pages 4467–4471.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

Pradhan, S. S., Ward, W., and Martin, J. H. (2008). Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310.

Punyakanok, V., Roth, D., Yih, W.-t., and Zimak, D. (2004a). Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1346. Association for Computational Linguistics.

Punyakanok, V., Roth, D., Yih, W.-t., Zimak, D., and Tu, Y. (2004b). Semantic role labeling via generalized inference over classifiers. In *CoNLL*, pages 130–133.

Puşcaşu, G. (2007). Wvali: Temporal relation identification by syntactico-semantic analysis. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 484–487. Association for Computational Linguistics.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005). The specification language timeml. *The language of time: A reader*, pages 545–557.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 884–894.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). Iso-timeml: An international standard for semantic annotation. In *LREC*.

Pustejovsky, J., Moszkowicz, J. L., and Verhagen, M. (2011). Iso-space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 1–9.

Pustejovsky, J. and Yocum, Z. (2013). Capturing motion in iso-spacebank. In *Workshop on Interoperable Semantic Annotation*, page 25.

Quinlan, J. R. (1996). Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730.

Ramrakhiyani, N. and Majumder, P. (2015). Approaches to temporal expression recognition in hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14(1):2.

Ren, H., Ji, D., Wan, J., and Zhang, M. (2009). Parsing syntactic and semantic dependencies for multiple languages with a pipeline approach. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 97–102. Association for Computational Linguistics.

Reppen, R., Ide, N., and Suderman, K. (2005). American national corpus (anc) second release. *Linguistic Data Consortium*.

Roberts, K. and Harabagiu, S. M. (2012). Utd-sprl: A joint approach to spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 419–424. Association for Computational Linguistics.

Salaberri, H., Arregi, O., and Zapirain, B. (2014). First approach toward semantic role labeling for basque. In *LREC*, pages 1387–1393.

Salaberri, H., Arregi, O., and Zapirain, B. (2015a). brol: The parser of syntactic and semantic dependencies for basque. pages 555–562.

Salaberri, H., Arregi, O., and Zapirain, B. (2015b). Ixagroupehuspaceeval:(x-space) a wordnet-based approach towards the automatic recognition of spatial information following the iso-space annotation scheme. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 856–861.

Salaberri, H., Arregi, O., and Zapirain, B.ñ. (2017). Euskarazko gertaeren etiketatze automatikoa. In *IkerGazte-2017 kongresuko artikulu-bilduma (Giza zientziak eta artea)*, pages 22–29.

Saquete Boro, E. (2010). Id 392:terseo + t2t3 transducer. a systems for recognizing and normalizing timex3. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 317–320, Uppsala, Sweden. Association for Computational Linguistics.

Sauri, R. and Badia, T. (2012). Spanish timebank 1.0. *LDC catalog ref. LDC2012T12*.

Saurii, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2005). Timeml annotation guidelines.

Schilder, F., Versley, Y., and Habel, C. (2004). Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proceedings of the workshop on geographic information retrieval at SIGIR 2004*.

Schrijver, A. (1998). *Theory of linear and integer programming*. John Wiley & Sons.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J., Farkas, R., Foster, J., Goenaga, I., Gojenola, K., Goldberg, Y., et al. (2013). Overview of the spmrl 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. Association for Computational Linguistics.

Setzer, A. (2001). *Temporal information in newswire articles: an annotation scheme and corpus study*. PhD thesis, University of Sheffield Sheffield, UK.

Streitberg, W. (1891). Perfective und imperfective actionsart im germanischen. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)*, 1891(15):70–177.

Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

Strötgen, J. and Gertz, M. (2015). A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics*, pages 541–547.

Strötgen, J., Zell, J., and Gertz, M. (2013). Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, USA. Association for Computational Linguistics.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.

Tenny, C. and Pustejovsky, J. (2000). Events as grammatical objects the converging perspectives of lexical semantics and syntax.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Librairie C. Klincksieck.

UzZaman, N. and Allen, J. F. (2010). Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283. Association for Computational Linguistics.

UzZaman, N. and Allen, J. F. (2011). Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 351–356. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., and Pustejovsky, J. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Vicente-Díez, M. T., Moreno-Schneider, J., and Martínez, P. (2010). Uc3m system: Determining the extent, type and value of time expressions in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 329–332, Uppsala, Sweden. Association for Computational Linguistics.

Vossen, P. (1998). *A multilingual database with lexical semantic networks*. Springer.

Weischedel, R. and Brunstein, A. (2005). Bbn pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

Xue, N. and Palmer, M. (2004). Calibrating features for semantic role labeling. In *EMNLP*, pages 88–94.

Xue, N. and Palmer, M. (2005). Automatic semantic role labeling for chinese verbs. In *IJCAI*, volume 5, pages 1160–1165. Citeseer.

Xue, N. and Palmer, M. (2009). Adding semantic roles to the chinese treebank. *Natural Language Engineering*, 15(01):143–172.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238.

Zapirain, B. (2010). *Rol semantikoen etiketatze automatikoa: rol multzoak eta hautapen murriztapenak*. PhD thesis, Universidad del País Vasco (UPV/EHU).

Zapirain, B.ñ. and Agirre, E. (2008). Robustness and generalization of role sets: Propbank vs. verbnet. *ACL-08: HLT*, page 550.

Zavarella, V. and Tanev, H. (2013). Fss-timex for tempeval-3: Extracting temporal information from text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 58–63. Citeseer.

Zhao, H., Chen, W., Kazama, J., Uchimoto, K., and Torisawa, K. (2009). Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 61–66. Association for Computational Linguistics.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.