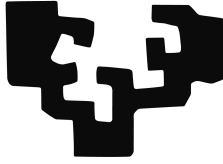


eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Euskal Hizkuntza eta Komunikazioa Saila

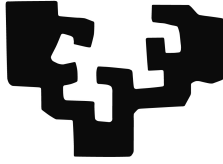
Doktorego-tesia

Euskarazko denbora-egituren azterketa eta corpusaren sorrera

Begoña Altuna Díaz

2018

eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Euskal Hizkuntza eta Komunikazioa Saila

Euskarazko denbora-egituren azterketa eta corpusaren sorrera

Begoña Altuna Díazek Arantza Díaz de Ilarraza eta María Jesús Aranzaberen zuzendaritzapean egindako tesi-txostena, Euskal Herriko Unibertsitatean Hizkuntzalaritzan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2018ko urriaren 2a.

You never learn to swim until you are in the water.

Jack Ma

v

Amari

Eskerrak

Hasteko, eskerrak eman nahi dizkiet Maxuxi eta Arantzari tesi-lan hau egin bitartean emandako laguntza guztiagatik eta ikerkuntza zer den irakasteagatik. Eskerrik asko bost urtean ni aguantatzeagatik.

Eskerrik asko Ixa taldean nirekin egon zareten guztioi! Bereziki Kortako jendeari, zuek izan zaretelako nire zalantza gehienak argitu dituzuenak eta zuek egin duzuelako (Korta mundutik aparte badago ere) inoiz bulegoz aldatu nahi ez izatea.

Ezin ditut ahaztu tuppertxokokoak. Elkarriketa bizarro guztiengatik, informazio extraofizial guztiagatik, bazkalosteko atsedengatik, harrikoa egiteko teknikarik optimizatuena irakasteagatik eta beste hainbatengatik, eskerrik asko!

Dena ez da Kortan eta fakultatean gertatzen, ordea. Eskerrik asko pintxopotera animatu izan zareten guztiei (Pintxopote Buru-Batzarrari eta Pintxopotearen Zaindariei bereziki) askotan osteguna nire asteko egunik gogokoena bihurtzeagatik. Eskerrik asko Matia kaletik kanpo egin ditugun planetara etorri zareten guztiei ere; Orhiseak ez bagintuen akabatu, ez gaitu lanak hobiratuko!

Eskerrik asko nirekin bizi izan zareten guztiei ere, bai Casa Mertxeko nire arkitekto maiteei, bai Tolosa 87koei (alfabetikoki, inor haserretu ez dadin, Ander, Manex, Mara eta Olatz). Oso guay egon dira pisu-poteak, afariak eta karaokeak.

Eskerrik asko familiari azken urtea baino ez duelako pasatu tesia noiz defendatuko dudan galdetzen. Plazera da lanari buruz hitz egiten ez duen

jendearekin egotea, azken urtean pelma batzuk izan bazarete ere. Noski, eskerrik asko Pedromariri tupperrak egiteagatik eta Itsasori nire Deustuko gela oraindik bere billar gela ez bihurtzeagatik.

Eskerrik asko Deustuko (eta Bakioko) jendeari, kuadrillari, mendi taldeari (eta Izabako jendeari) eta filologoei (eta Marruskilariei). Nahiz eta AP-8ak banatu zuek beti egon zaretelako ni hartzeko prest.

Vorrei ringraziare la gente della FBK, soprattutto Manuela e Anne-Lyse, ma anche Mara, Mohammed e Roldano e tutti gli altri. Vorrei salutare anche la gente della FB3 al Sanba (e tanti altri), perché mi hanno accolto dal primo momento.

Tesia amaitzen ari naizen honetan inor ahaztu badut, barkatu. Zuei ere, eskerrik asko!

Esker instituzionalak

Tesi-lan hau Eusko Jaurlaritzaren Hezkuntza, Unibertsitate eta Ikerketa Sail-
leko PRE_2013_1_959, PRE_2014_2_242, PRE_2015_2_0284 eta
PRE_2016_2_294 diru-laguntzei esker finantzatu da, baita DETEAMI (Eus-
ko Jaurlaritza, ref.: 2014111003) eta PROSA-MED (MINECO, ref.: TIN2016-
77820-C3-1-R) proiektuen bidez ere.

Laburpena

Ikerketa-lan honetan euskarazko denbora-informazioaren prozesamenduan lehen urratsak egin ditugu. Horretarako, beste hizkuntzetan egin diren lanetan eta euskarazko denbora-egituren analisi linguistikoan oinarritu gara. Informazio hori baliatuta, euskarazko denbora-egiturak automatikoki tratatzeko ezaugarri linguistiko esanguratsuenak identifikatu ditugu eta horiek kodetzeko EusTimeML markaketa-lengoaia sortu dugu. Era berean, EusTimeMLri jarraituta denbora-informazioa eskuz etiketatuta duen EusTimeBank corpusa sortu dugu.

Corpus hori, euskarazko fenomenoak aztertzeke erabiltzeaz gain, tresna automatikoen garapenerako eta ebaluaziorako erabili dugu. Hain zuzen ere, tesi-lan honetan denbora-adierazpenak identifikatzeko eta normalizatzeko EusHeidelTime tresna garatu dugu eta denbora-lerroak automatikoki eratzeko KroniXa sistema sortu dugu. Tresna horiek euskararen prozesamendukateetan integratzeko eta beste tresna batzuekin uztartzeko urratsak egin ditugu, euskararen ulermen eta tratamendu automatikoan denbora-informazioa ere baliatu ahal izateko.

Gaien aurkibidea

Laburpena	ix
Gaien aurkibidea	xi
SARRERA	3
1 Proiektuaren nondik norakoak	3
1.1 Denbora-informazioa	4
1.2 Motibazioa eta helburuak	7
1.3 Tesi-lanaren antolakuntza	8
1.4 Tesiari lotutako argitalpenak	9
2 Ikergaiaren egungo egoera	13
2.1 Denbora-informazioko elementuak eta horien izenak	13
2.1.1 Gertaerak	14
2.1.2 Denbora-adierazpenak	15
2.1.3 Seinaleak	16
2.1.4 Erlazioak	16
2.2 Aurrekari teorikoak	17
2.2.1 Gertaeren sailkapenak	17
2.2.1.1 Venderren sailkapena	17
2.2.1.2 Croften sailkapena	19

2.2.1.3	Coterónen sailkapena	19
2.2.1.4	Hallidayren sailkapena	20
2.2.1.5	Setzerren sailkapena	21
2.2.1.6	TimeMLren sailkapena	22
2.2.2	Denboraren tratamendu teorikoa	23
2.2.2.1	Denbora-uneak vs interbaloak	24
2.2.2.2	Allenen denbora-tarteen teoria	24
2.2.2.3	Denbora-adierazpenen sailkapena	25
2.2.2.4	Aditz-aspektuaren eta tempusaren teoriak	27
2.3	Markaketa-lengoiak eta eskuz etiketatze tresnak	28
2.3.1	Denbora-informazioa etiketatze markaketa-lengoiak	28
2.3.2	Eskuz etiketatze tresnak	30
2.4	Denbora-informazioa prozesatzeko sistemak	34
2.4.1	Erregeletan oinarritutako sistemak	35
2.4.2	Metodo estatistikoetan oinarritutako sistemak	38
2.4.3	Sistema hibridoak	40
2.4.4	Ikasketa sakoneko teknikak	41
2.5	Corpusak	43
2.5.1	Denbora-informaziodun lehen corpusak	43
2.5.2	Denbora-informazioa etiketatuta duten albiste corpusak	45
2.5.3	Beste domeinuetako denbora-informaziodun corpusak	51
2.6	Denbora-lerroen sorreran egindako beste lan batzuk	56
2.7	Laburpena	59

ETIKETATZEA ETA CORPUSAREN SORRERA 63

3	Euskarazko gertaerak, denbora-adierazpenak eta horien ar- teko erlazioak	63
3.1	Gertaerak	64
3.1.1	Gertaera-adierazpenak euskaraz	64
3.1.1.1	Aditzak	65
3.1.1.2	Izenak	67
3.1.1.3	Adjektiboak	74
3.1.1.4	Adberbioak	75
3.1.1.5	Izenordainak	75
3.1.1.6	Bestelakoak	76

3.1.1.7	Gertaera-adierazpen konplexuetako kategoria- konbinazioak	77
3.1.2	Gertaeren ezaugarriak	79
3.1.2.1	Gertaera-klasea	79
3.1.2.2	Tempusa	81
3.1.2.3	Aspektua	83
3.1.2.4	Modalitatea	84
3.1.2.5	Polaritatea	85
3.1.3	Gertaeren ezaugarri osagarriak	86
3.1.3.1	Ziurtasuna	87
3.1.3.2	Faktualitatea	87
3.1.3.3	Faktualitatearen kasu bereziak	88
3.2	Denbora-adierazpenak	89
3.2.1	Denbora-adierazpenak euskaraz	89
3.2.1.1	Adberbioak	90
3.2.1.2	Izenak	91
3.2.1.3	Adjektiboak	91
3.2.1.4	Perpaus bereziak	92
3.2.1.5	Zenbakizko egiturak	92
3.2.2	Abiarazle lexikoak	93
3.2.3	Denbora-adierazpenen ezaugarri linguistikoak	95
3.2.3.1	Mota	95
3.2.3.2	Balio normalizatua	97
3.2.3.3	Denbora-adierazpenak testuan betetzen duen funtzioa	97
3.3	Erlazioak	98
3.3.1	Denbora-erlazioak	101
3.3.1.1	Denbora-erlazio motak	102
3.3.1.2	Denbora-erlazioen seinaleak	104
3.3.2	Mendekotasun-erlazioak	105
3.3.3	Aspektuzko erlazioak	108
3.4	Laburpena	110
4	EusTimeML bidezko eskuzko etiketatzea eta EusTimeBank Corpusaren osaketa	111
4.1	EusTimeML markaketa-lengoaia	112
4.1.1	TimeML	112
4.1.2	EusTimeMLn jarraitutako metodologia	114

4.2	Euskarazko denbora-egituren etiketatzea	116
4.2.1	Gertaeren etiketatzea, EVENT	116
4.2.1.1	Gramatika-kategoria, pos	117
4.2.1.2	Gertaera-klaseak, class	118
4.2.1.3	Tempusa, tense	119
4.2.1.4	Aspektua, aspect	121
4.2.1.5	Modalitatea, modality	122
4.2.1.6	Polaritatea, polarity	123
4.2.1.7	Ziurtasuna, certainty	124
4.2.1.8	Faktualitatea, factuality	125
4.2.1.9	Kasu bereziak, specialCases	126
4.2.1.10	Gertaeren beste atributuak eta MAKEINSTANCE etiketa	127
4.2.2	Denbora-adierazpenen etiketatzea, TIMEX3	128
4.2.2.1	Denbora-adierazpen motak, type	128
4.2.2.2	Balio normalizatua, value	129
4.2.2.3	Iraupenen hasiera- eta amaiera-puntuak, beginPoint eta endPoint	135
4.2.2.4	Errepikapenen atributu bereziak: kuantifika- tzaileak eta frekuentzia-adierazleak, freq eta quant	135
4.2.2.5	Modifikatzaileak, mod	137
4.2.2.6	Funtzioa dokumentuan, functionInDocument	139
4.2.2.7	Beste atributu batzuk	140
4.2.3	Denbora-erlazioen etiketatzea, TLINK	140
4.2.3.1	TLINKen beste atributuak	143
4.2.3.2	Seinaleen etiketatzea, SIGNAL	143
4.2.4	Mendekotasun-erlazioen etiketatzea, SLINK	144
4.2.4.1	SLINKen beste atributuak	146
4.2.5	Aspektuzko erlazioen etiketatzea, ALINK	146
4.2.5.1	ALINKen beste atributuak	147
4.2.6	TimeMLrekiko egokitzapenak eta aldaketak	147
4.2.6.1	Predikatu generikoen tratamendua	148
4.2.6.2	Euskarazko aditzen aspektu eta tempusa erre- presentatzeko balioak	148
4.2.6.3	Euskarazko modalitatea adierazteko balioak	148
4.2.6.4	Euskarazko postposizioen tratamendua	149
4.2.6.5	Ezeztapenaren tratamendua	150

4.2.6.6	Alde batera utzitako atributuak	150
4.2.6.7	Faktualitate-informazioa gehitzea	151
4.2.6.8	EusTimeMLren uztarketa beste etiketatzeekin	151
4.3	EusTimeBank corpora	153
4.4	EusTimeBank eskuzko etiketatzeko metodologia	156
4.5	Denbora-informazioa etiketatzeko esperimentuak	159
4.5.1	Denbora-adierazpenak eta seinaleak etiketatzeko esperimentua	160
4.5.1.1	Anotatzaileen arteko adostasuna	160
4.5.1.2	Emaitzen ebaluazioa	162
4.5.2	Gertaerak etiketatzeko esperimentuak	163
4.5.2.1	Lehen etiketatzeko esperimentua	163
4.5.2.2	Bigarren etiketatzeko esperimentua	164
4.5.2.3	Faktualitatea etiketatzeko esperimentua	168
4.5.3	Erlazioak etiketatzeko esperimentua	169
4.6	EusTimeBank corpusaren beste erabilerak: ezeztapenaren analisia eta etiketatzeko	171
4.6.1	Ezeztapena etiketatzeko eskema	171
4.6.2	Corpusa	172
4.7	Laburpena	173

TRESNAK ETA APLIKAZIOAK 177

5 EusHeideltime eta KroniXa 177

5.1	Aurreprozesamendua: euskararen prozesamendurako analisi-katea	177
5.1.1	Tokenizazioa	179
5.1.2	Analisi morfosintaktikoa	179
5.1.3	Desanbiguazio morfologikoa	181
5.1.4	Funtzio sintaktikoen osaketa	182
5.1.5	Desanbiguazio sintaktikoa	183
5.1.6	Analisi sintaktiko partziala	183
5.1.7	Analisi sintaktiko osoa	184
5.1.8	Analisi semantikoa	184
5.2	EusHeidelTime: denbora-adierazpenak identifikatzeko eta normalizatzeko tresna	186

5.2.1	EusHeidelTimerako erregelak, patroiak eta balio normalizatuak	187
5.2.1.1	Patroiak	187
5.2.1.2	Balio normalizatuak	190
5.2.1.3	Erregelak	191
5.2.2	EusHeidelTimeren integrazioa beste tresnetan eta funtzionamendua	196
5.3	KroniXa: denbora-lerroak sortzeko tresna	200
5.3.1	Denbora-informazioaren tratamendu automatikoa	201
5.3.2	Denbora-lerroen sorrera	202
5.3.3	Esperimentazioa	207
5.3.3.1	Ebaluaziorako eskuzko denbora-lerroak	207
5.3.3.2	Ebaluaziorako metodologia eta metrikak	209
5.3.3.3	Denbora-lerroak dependentzietatik ateratako informazioaren bidez hobetzea	210
5.4	Laburpena	214

ONDORIOAK ETA ETORKIZUNEKO LANAK 217

6	Ondorioak eta etorkizuneko lanak	217
6.1	Ekarpenak	217
6.1.1	Denbora-informazioa adierazten duten elementuen azterketa	218
6.1.2	Euskarazko denbora-informazioa kodetzeko markaketa-lengoaiaren sorrera	219
6.1.3	EusTimeBank corpusa	221
6.1.4	EusHeidelTime tresna	222
6.1.5	KroniXa: denbora-lerroak sortzeko tresna	223
6.2	Ondorioak	223
6.3	Etorkizuneko lanak	224

Bibliografia 227

ERANSKINAK 261

A Abiarazle lexikoen zerrenda 261

B	EusTimeMLko etiketatze-eskema	265
B.1	Gertaeren BNFa	265
B.2	Denbora-adierazpenen BNFa	266
B.3	Seinaleen BNFa	267
B.4	TLINKen BNFa	267
B.5	SLINKen BNFa	268
B.6	ALINKen BNFa	268

SARRERA

Proiektuaren nondik norakoak

Tesi-lan hau Hizkuntzaren Prozesamenduan (HP) kokatzen da; hain zuzen ere, Ixa taldearen¹ jardunean. Ixa taldea euskararen prozesamenduan lan egiten duen UPV/EHUko erreferentziako taldea da eta azken hiru hamarkadetan HPn euskararen analisisian ikerketa egin ez ezik, baliabide aurreratuak ere sortu ditu. Tesi-lan honen bidez, euskararen tratamendu automatikoan beste urrats bat aurrera egin nahi dugu. Hala, Ixa taldean garatu diren tresnak eta baliabideak abiapuntutzat hartuta, euskarazko testuak interpretatzen laguntzen diguten tresnak eta baliabideak sortu ditugu testu horiek duten denbora-informazioan oinarrituta.

Zehazki, ikerketa-lan hau euskarazko denbora-informazioaren azterketan eta prozesamenduan zentratzen da. Gizakiok hitzezko produkzioan tarte handia eskaintzen diegu igarotako gertaeren narrazioari zein etorkizuneko planen kontaktuari. Narrazio-testuetan, denboran zehar gertatzen diren gertaeren berri ematen da eta, ondorioz, denbora-informazioak testu horietako informazioaren antolakuntzan garrantzi berezia du, testuetako gertaerak denbora-ardatzean kokatzen eta testuko ideiak antolatzen laguntzen baitu.

Gizakiontzat erraza da testuetako denbora-informazioa erauztea, gai baikara gertaerak beren artean ordenatzeko eta zein unetan gertatu diren inferitzeko. Informazio hori, ordea, ez da gizakiontzat bezain ulergarria tresna automatikoentzat, testuko gertaerak eta denborak interpretatzeaz gain, horien mundu errealeko erreferentziak ere ulertu behar baitira.

HPn, beraz, denbora-informazioa ulertu egin behar da testuaren ulermen sakontzeko edota informazio hori baliatzen duten tresna aurreratuak

¹<http://ixa.si.ehu.eus/>

sortzeko. Aplikazio horiek euskararako garatu ahal izateko, i) euskarazko denbora-informazioa nola gauzatzen den aztertu behar da, ii) denbora-informazioko elementu bakoitzak zein informazio adierazten duen ezagutu behar da eta iii) informazio hori esplizitu egiten duten tresnak eta baliabi-deak sortu behar dira.

Tesi-lan honetan, gertaeren, denbora-adierazpenen eta horien artean sortzen diren erlazioen identifikazioa eta beren ezaugarrien kodeketa deskribatuko ditugu. Informazio hori baliatuta, denbora-adierazpenak automatikoki erauzteko sistema nola sortu dugun eta denbora-lerroen sorrera automatikoki nola egin dugun deskribatuko dugu. Horretarako, horiek sortu bitartean, euskarazko denbora-informazioaren azterketa eta prozesamenduan egin ditugun lanen berri emango dugu.

1.1 Denbora-informazioa

Aipatu dugunez, denbora-informazioa HPn baliatu ahal izatea du tesi-lan honek helburu. Horretarako, denbora-informazioa tresna automatikoentzat atzigarri egin behar da.

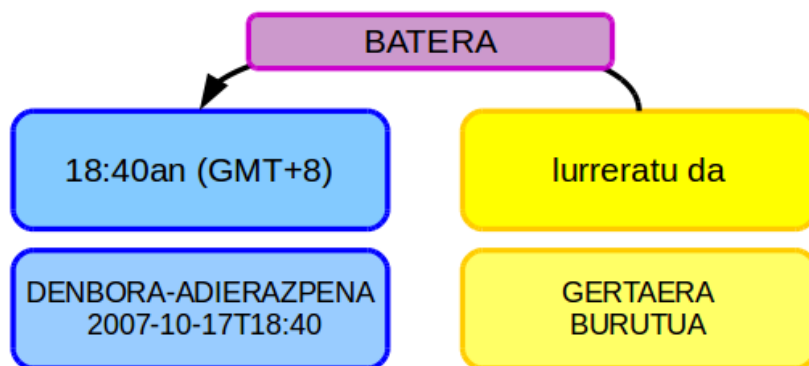
Testuetako denbora-erreferentziak eta mundu errealeko kronologia eta gertaerak denbora-informazioa dira. Bestela esan, denbora-informazioa zer noiz gertatzen den adierazten duen informazioa da. Adibidez, (1) adibideko esaldian, denbora-une batean gertatu den gertaera baten berri ematen da: *lurreratzea 18:40an* gertatu da.

(1) Lehen Airbus A380a *18:40an (GMT+8)* **lurreratu da** Singapurren.

Denbora-informazioa adierazteko, hainbat denbora-egitura baliatzen ditugu gizakiok. Zer gertatzen den adierazteko, *gertaera* deritzen elementuak erabiltzen ditugu. Denborako une eta iraupenak adierazteko, berriz, *denbora-adierazpenak* ditugu. Era berean, gizakiok badakigu gertaeraren eta denbora-adierazpenaren artean lotura edo harreman bat dagoela: *lurreratzea 18:40an* gertatu da. Denbora-informazioa duten erlazioak *denbora-erlazioak* direla esaten dugu.

Denbora-informazioa interpretatu ahal izateko, informazio hori adierazten duten ezaugarriak eta bakoitzaren balioak identifikatu behar dira. Esaterako, 1.1 irudian, (1) adibideko esaldiaren denbora-informazioa adierazi dugu. Adibide horretan, gertaera horiz adierazi dugu, denbora-adierazpena urdinez eta horien arteko denbora-erlazioa morez.

Lehen Airbus A380a 18:40an (GMT+8) lurreratu da.



1.1 irudia – (1) adibideko esaldiaren etiketatzea

1.1 irudian, denbora adierazten duten elementuez eta horien artean sortutako erlazioez gain, elementu bakoitzaren informazio esanguratsuena ere azaleratu dugu. Alde batetik, denbora-adierazpenaren balio normalizatuaren bidez, *2007-10-17T18:40*, esaldiko *18:40 (GMT+8)* adierazpenak kronologiko une zehatz horri erreferentzia egiten diola adierazi dugu. Beste aldetik, *lurreratu da* gertaera burutua dela adierazi dugu. Horrek esan nahi du errealitatean gertatu dela eta dagoeneko amaitu dela. Gainera, lurreratzea eta ordu hori batera gertatu direla ere adierazi dugu.

Hala eta guztiz ere, gizakiok sortzen ditugun mezuak, normalean, luzeagoak izaten dira; esaldi bat baino gehiagoko testuak dira. Ondorioz, denbora-informazioaren azterketa ezin da esaldi mailara mugatu. Sarrera honetan denbora-informazioa aurkezteko baliatu dugun esaldia euskarazko denbora-informazioa aztertzeke sortu dugun corpuseko testu honen parte da:

Lehen A380a hasi da zerbitzu komertziala ematen

2007ko urriaren 17a

Hegazkingintzaren historian, sekulako lorpena izan da: *lehen Airbus A380a 18:40an (GMT +8) lurreratu da Singapur-en*, Changi nazioarteko aireportuan, Airbus-en bidalketa-zentrotik atera eta 12 orduko hegaldia egin ostean. Hegazkinari 400 bat gonbidatuk egin zioten ongietorria laster zabalduko den Changi Nazioarteko aireportuko 3. terminalean.

Tartean zen Lee Hsien Loong, Singapur-eko lehen ministroa. Hitzaldian, lehen ministroak Singapore Airlines (SIA) aire-linea hartzailea txalotu zuen, eta aipatu zuen arrakasta Singapur-eko biztanle guztiak harro egoteko modukoa zela. Gehitu zuen lurreratzea gertaera gogoangarria zela, ez bakarrik SIArentzat, baita Changi nazioarteko aireportuarentzat ere, zeina herrialdeko aireporturik handiena baita.

Orain arteko terminal guztiak prest daude A380a hartzeko, eta 3. terminalak (1,75 milioi dolar singapurtar kostatu da) datorren urteko urtarrilean zabalduko ditu atea.

A380ak ongintzazko hegaldi berezi bat egingo du urriaren 25ean, Singapur-etik Sydney-ra, eta Singapur-Sydney ibilbideari dagokion ohiko zerbitzua urriaren 28an hasiko du.

Ikus daitekeenez, testuan A380 hegazkinaren lurreratzeaz (2) gain, inguruko gertaeren berri ere ematen da, hala nola 400 gonbidatu ingururen ongietorria (3) edota Changi aireportuko terminalen egoera (4).

- (2) Lehen Airbus A380a 18:40an (GMT+8) **lurreratu da** Singapurren.
- (3) Hegazkinari 400 bat gonbidatuk egin zioten **ongi etorria**.
- (4) Orain arteko terminal guztiak **prest daude** A380a hartzeko.

Denbora-uneei dagokienez, lurreratzea *18:40an* gertatu dela badakigu (2), baita hegaldiak *12 ordu* iraun duela (5) eta 3. terminala *2008ko urtarrilean* zabalduko dela (6) ere.

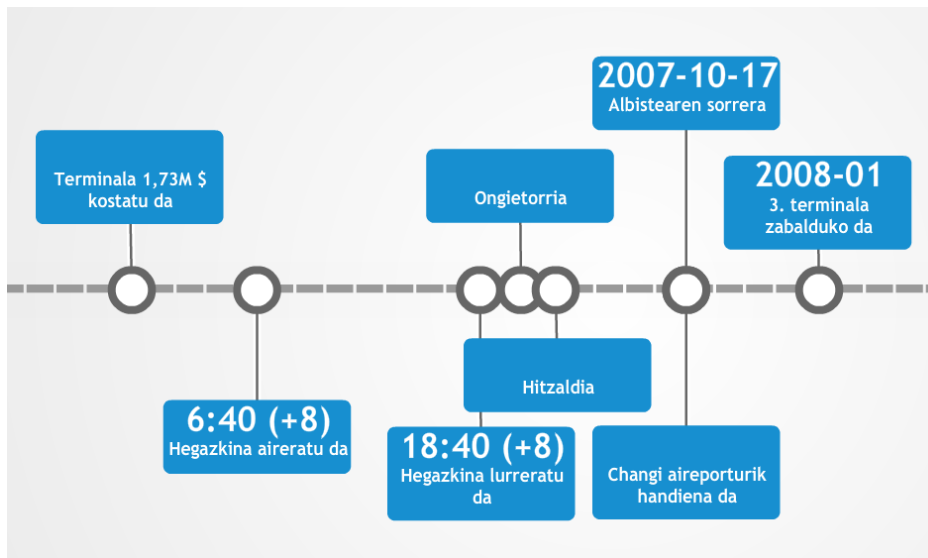
- (5) Lehen Airbus A380a **18:40an (GMT+8)** lurreratu da Singapurren, (...) **12 orduko** hegaldia egin ostean.
- (6) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu atea

Halaber, aurreko albistean, denbora-informazioak testua antolatzen laguntzen duela ikus dezakegu. Hiru une nagusi bereiz daitezke testu horretan: i) hegaldia eta lurreratzea, ii) terminalak prest egotea eta iii) ongintzazko hegaldi berezia. Horiek kronologikoki adierazi dira testuan eta denbora-informazioaren bidez gertaerak ordenatzeko gai gara.

Era berean, gertaerak beren artean kronologikoki antolatzen direla edo kronologiako une zehatz batean ainguratzen direla ere ikus dezakegu. Hau da, gertaeren eta denboren artean erlazioak sortzen dira. Erlazio horiek

aldiberekotasuna adierazten dute, *3. terminala hurrengo urteko urtarrilean zabalduko dela* esatean bezala, edo bi elementuren arteko ordena adierazten dute, *lurreratzea hegaldiaren ondoren gertatu zela* esatean bezala.

Testuko denbora-informazioa errepresentatzeari dagokionez, denbora lehenalditik etorkizunera doan ardatz gisa irudika daiteke. Testuetako gertaera batzuk denborako une zehatzetan aingura ditzakegu une horietan gertatu direla badakigulako. Gertaerak besteekiko ordenatzen direla ere aipatu berri dugu. Ainguratuta dauden gertaera horiek oinarritzat hartuta, beste gertaerak ere denboran koka ditzakegu. Ardatz horretan, ainguratze-datak eta gertaerak irudikatzean, denbora-lerroak sortzen dira. 1.2 irudian, aurreko testuko denbora-lerroa agertzen da eta bertan gertaerak kronologiako une zehatzetan ainguratuta agertzen dira.



1.2 irudia – *Lehen A380a hasi da zerbitzu komertziala ematen* albistetik sortutako denbora-lerroa

1.2 Motibazioa eta helburuak

Lan honen motibazio nagusia denbora-informazioa testuen interpretazioan baliatu ahal izatea da, euskararen tratamendu automatikoan aurrera egiteko. Izan ere, aipatu bezala, denbora-informazioak testuetako informazioa ardatz

kronologikoan antolatzen du. Ondorioz, HPn denbora-informazioa interpreta badaiteke, informazio hori testuak ulertzeko baliatu ahal izango da. Hain zuzen ere, gertaerak kronologiaren arabera kokatzeaz gain (Yan *et al.*, 2011; Minard *et al.*, 2015; Bauer *et al.*, 2015), denbora-informazioa kausa-ondorio erlazioen prozesamenduan (Mirza eta Tonelli, 2014), galdera-erantzun sistemetan (Pustejovsky *et al.*, 2003a; Saurí *et al.*, 2005; Mirza eta Minard, 2015) edota bilaketa-sistemetan (Chieu eta Lee, 2004; Setty *et al.*, 2010) baliatu ahal izango da. Halaber, denbora-informazioa aplikazio aurreratueta ere erabil daiteke, hala nola gertaeren aurreikuspenetan (Radinsky eta Horvitz, 2013) eta etorkizunaren iragarpenean (Kawai *et al.*, 2010).

Ikerketa-lan honen helburu nagusia euskarazko denbora-informazioaren tratamendu automatikorako baliabideak garatzea da, horiek gero denbora-informazioa baliatzen duten tresna aurreratueta erabili ahal izateko. Helburu nagusi hori lortzeko, tarteko helburu hauek definitu ditugu:

- Denbora-informazioa adierazten duten elementuek zein forma hartzen duten aztertuko dugu.
- Euskarazko denbora-informazioa kodetzeko lengoia garatuko dugu.
- Kazetaritza-testu eta historia-testu corpora abiapuntutzat hartuta, HPn baliatu ahal izango den denbora-informaziodun corpora sortuko dugu.
- Denbora-informaziodun corpora erabilita, denbora-informazioko elementuak automatikoki identifikatzen eta beren ezaugarriak erauzten dituzten tresnak sortuko ditugu.
- Denbora-informazioa baliatuta denbora-lerroak sortzeko tresna garatuko dugu.

1.3 Tesi-lanaren antolakuntza

Sarrera honen ostean, denbora-informazioaren prozesamenduan egin diren lan esanguratsuenen berri emango dugu 2. kapituluari, euskarazko denbora-informazioaren tratamendua nondik bideratu erabaki ahal izateko.

Jarraian, euskarazko denbora-informazioa nola adierazten den eta denbora-egitura horiek zer motatako informazioa adierazten duten aztertuko dugu 3. kapituluari. Denbora-informaziodun EusTimeBank corpora nola sortu dugun 4. kapituluari azalduko dugu, baita horretarako denbora-informazioa nola

kodetu dugun eta testuan nola txertatu dugun ere. 5. kapituluan, denbora-informazioaren prozesamendurako tresnen berri emango dugu. Alde batetik, EusHeidelTime denbora-adierazpenak erauzteko eta normalizatzeko tresna aurkeztuko dugu eta, beste aldetik, KroniXa, denbora-informazioa baliatuta denbora-lerroak sortzeko tresna, deskribatuko dugu.

Amaitzeko, tesi-lan honen ekarpenak, ondorioak eta etorkizuneko lanak aurkeztuko ditugu 6. kapituluan.

1.4 Tesiari lotutako argitalpenak

Ikerketa-lan honetan zehar egindako lan batzuk hainbat biltzar eta aldizkarietan aurkeztu ditugu. Jarraian, tesi-lanari lotutako argitalpenak argitalpen-motaren arabera eta kronologikoki berrienetik zaharreneira aurkezten ditugu. Argitalpen horiek tesi-txosten honetako zein kapitulurekin lotzen diren 1.1. taulan ikus daiteke.

Aldizkariak

- Altuna B., Soraluze A., Aranzabe M.J., Arregi O., eta Díaz de Ilarraza A. KroniXa: Timeline Creation from Basque Texts. *Journal of Information Processing and Management*, Bidalita
- Salaberri H., Altuna B.A., Aranzabe M.J., Arregi O., eta Díaz de Ilarraza A. bTime: a Hybrid Architecture for Capturing Temporal Information in Basque. *Knowledge-Based Systems*, Bidalita
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, 59(0):15–22, 2017a. ISSN 1989-7553. <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488>
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTime-ren bidez. *Ekaia*, 30:153–165, 2016b. ISSN 0214-9001 e-ISSN: 2444-3255. URL http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8334/Denboraren%20tratamendua_TimeML_HeidelTime.pdf
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):

13–24, 2014a. ISSN 1647-0818. <http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1>

Liburuetako kapituluak

- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. An Event Factuality Annotation Proposal for Basque. In Frank A.U., Ivanovic C., Mambrini F., Passarotti M., eta Sporleder C., editoreak, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, 1 lib., 15–24, Vienna, Austria, 2018b. Gerastree Proceedings. ISBN 978-3-901716-43-0. <https://www.oeaw.ac.at/ac/crh2/proceedings/>
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In Alegria I., Latatu A., Ormaetxebarria M.J., eta Salaberri P., editoreak, *II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Arteak*, 127–134, Bilbo, 2017b. Udako Euskal Unibertsitatea (UEU). ISBN 978-84-8438-628-5. <http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8863/IKERGAZTE.2017.GIZAZIENTZIAKetaARTEA.pdf>
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Adapting TimeML to Basque: Event Annotation. In Gelbukh A., editorea, *Computational Linguistics and Intelligent Text Processing*, 565–577, Cham, Switzerland, 2018a. Springer International Publishing. ISBN 978-3-319-75487-1. URL https://link.springer.com/chapter/10.1007/978-3-319-75487-1_43
- Altuna B. Análisis de la información temporal en euskera. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., eta Romá Ferri M.T., editoreak, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015a. Sociedad Española para el Procesamiento del Lenguaje Natural. ISBN 978-84-608-1989-9. URL https://gplsi.dlsi.ua.es/sepln15/sites/gplsi.dlsi.ua.es.sepln15/files/attachments/begona_altuna.pdf
- Altuna B. Euskarazko denbora-egituren tratamendu automatikorako azterketa. In Alegria I., Latatu A., eta Ormaetxebarria M.J., editoreak, *I. IkerGazte, Nazioarteko ikerketa euskaraz*, 46–53, Bilbao, Spain, 2015b. Udako Euskal Unibertsitatea. ISBN 978-84-8438-540-0. URL <http://>

ixa.si.ehu.es/sites/default/files/dokumentuak/3395/IKERGAZTE.2015.pdf

Txosten teknikoak

- Altuna B., Aranzabe M., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Barne-txostena, Lengoia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016, 2016a. <https://addi.ehu.es/handle/10810/17305>
- Altuna B., Aranzabe M., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Barne-txostena, Lengoia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2014, 2014b. http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etiketatzeko%20gidalerroak

1.1 taula – Kapituluekin lotutako artikuluak

Kapituluak	Argitalpenak
3	Altuna <i>et al.</i> (2014a, 2018a, 2017b, 2018b)
4	Altuna <i>et al.</i> (2014b, 2016a)
5	Altuna <i>et al.</i> (2016b, 2017a); Salaberri <i>et al.</i> (Bidalita); Altuna <i>et al.</i> (Bidalita)

Tesi-lana egin bitartean, ikerketa-lanaren osagarri izan diren beste ikerketa batzuk ere egin ditugu. Horiek jarraian agertzen dira kronologikoki antolatuta berrienetik zaharreneira.

- Altuna B., Minard A.L., eta Speranza M. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, 34–42. Association for Computational Linguistics, 2017c. <http://aclweb.org/anthology/W17-1806>
- Minard A.L., Speranza M., Urizar R., Altuna B., van Erp M., Schoen A., eta van Son C. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., eta Piperidis. S., editoreak, *Proceedings of LREC 2016, Tenth*

International Conference on Language Resources and Evaluation. European Language Resources Association, 2016. ISBN 978-2-9517408-9-1. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/488_Paper.pdf

- Agerri R., Agirre E., Aldabe I., Altuna B., Beloki Z., Laparra E., de Lacalle M.L., Rigau G., Soroa A., eta Urizar R. NewsReader project. *Procesamiento del Lenguaje Natural*, 53:155–158, 2014. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5063>

Amaitzeko, aurretik deskribatutako lanetako batzuek jaso dituzten sarien berri emango dugu. Sariak berrienetik zaharreneira ordenatu ditugu:

- SEPLN 2017 artikulua onenari saria *XXXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* kongresuan (Murztzia, 2017), *EusHeidelTime: Time Expression Extraction and Normalisation for Basque* artikulua.
- II. Ikergazte Giza Zientzietako ahozko aurkezpen onenari saria *II. Ikergazte, Nazioarteko ikerketa euskaraz* kongresuan (Iruñea, 2017), *Euskarazko ezeztapenaren tratamendu automatikorako azterketa* atikulua.
- I. Ikergazte Giza Zientzietako artikulua onenari saria *I. Ikergazte, Nazioarteko ikerketa euskaraz* kongresuan (Durango, 2015), *Euskarazko denbora-egituren tratamendu automatikorako azterketa* artikulua.

Ikergaiaren egungo egoera

Azken bi hamarkadetan aurrerapen handiak egin dira denbora-informazioaren tratamendu automatikoan. XX. mendearen erdialdeko eredu teorikoak baliatuz, denbora-informazioa identifikatzeko eta normalizatzeko markaketa-lengoaiak eta tresnak garatu dira. Horien berri emango dugu atal honetan. Ikerlan honetan erabiliko dugun terminologia nagusia definituko dugu 2.1 atalean. Jarraian, denbora-informazioaren prozesamenduan egin diren lanik garrantzitsuenak deskribatuko ditugu. Zehazki, 2.2 atalean, denbora-informazioaren arloko aurrekari teorikoen berri emango dugu. 2.3 atalean, denbora-informazioa etiketatzeko eskemen eta eskuzko tresnen berri emango dugu. Ondoren, denbora-informazioaren tratamendu automatikoa egiten duten tresnak deskribatuko ditugu 2.4 atalean eta denbora-informazioa duten corpusak aurkeztuko ditugu 2.5 atalean. Amaitzeko, 2.6 atalean denbora-informazioa baliatuta sortu diren denbora-lerroak sortzeko tresnen berri emango dugu.

2.1 Denbora-informazioko elementuak eta horien izenak

Denbora-informazioaren azterketan gertaerak, kronologiako uneen eta tarteen adierazpenak, horien artean sortzen diren erlazioak eta erlazio horiek esplizitu egiten dituzten elementuak hartuko ditugu kontuan. Denbora-informazioan parte hartzen duten elementu horiek izendatzeko hainbat termino erabili dira. Batzuk, *event* eta *eventuality* bezalakoak, semantikaren eta logi-

karen tradizioetatik datoz, baina beste batzuk elementu berriak gehitu ahala sortu behar izan dira. Atal honetan denbora-informazioaren azterketan identifikatu ditugun elementuen izendapenak eta terminologiak batu ditugu gure ikerketa-lanean zein zentzurekin erabili erabakitzeko.

2.1.1 Gertaerak

Gertaera terminoa gertatzen den zerbait adierazteko erabiltzen badugu ere, HPn, hainbat erataria definitu izan da. Hain zuzen ere, gertaeren definizioa denborarekin aldatu dela eta konplexutasuna irabazi duela esan daiteke. Bachek (1986) gertaerei *eventuality* esaten die eta hala definitu ditu: gertaerak gertatzen, hasten edo irauten duten ekintza eta egoerak dira. Bayerek ere (1986) definizio orokorra eman du eta gertaerak denbora-espazioan kokatutako gorputzak direla dio.

Wonsever *et al.*entzat (2008) gertaera predikatu batek adierazten duen edozein egoera edo gertakari da. Saurí *et al.*-ek (2009), Bachén (1986) il-doari jarraituz, horrela definitzen dituzte gertaerak: beren esanetan gertaera gertatzen diren, gertatzen, izaten edo mantentzen diren situazioak eta egia diren edo egia bilakatzen diren egoerak edo kondizioak adierazten dituen termino orokorra da. Dowdenek (2009), ordea, gertaerak eta denbora lotzen ditu. Bere aburuz, eguneroko diskurtsoan, gertaera iraupen finitua duen gertakaritza hartzen da eta iraupen horretan objektu batek bere ezaugarriak aldatzen ditu.

Denbora-informazioaren prozesamenduaren atazari lotuago dauden definizioak ere garatu dira. Setzerrek (2001) gertaerak eta egoerak definitu ditu. Setzerren arabera, gertaerek eta egoerek denboran irauten dute, errealitatearen barruan daude eta proposizioek adierazten dituzte. Gertaerei dagokienez dio gertaera, intuitiboki, hasiera eta amaiera mugatuak dituen gertatzen den zerbait dela. Egoerak, ordea, bi entitateren arteko erlazioa edo, normalean, hasiera edo amaiera definiturik ez dituen denbora-tarte batean gertatzen den eta eraldatzeko gaitasuna duten entitate baten atributuaren gauzatzeak dira.

Azkenik, gure gertaeraren kontzeptualizaziotik hurbilen dago Wonsever *et al.*en (2012, 207 or.) definizioa:

Events can be actions (carried out voluntarily by an agent), processes (events spontaneously set off or caused by a force external to the process, which can, in both cases, be punctual or have duration), or states (situations maintained along a period or that

are permanent). Generic predications will also be considered as events for they refer to states of things, states about which it is asserted that they take place.

Gure ikerketa-lanerako definizio hori iruditu zaigu egokiena predikatu generikoak (*generic predications*) gertaeratzat hartzen baititu. Adibidez, *Txinako bandera gorria izatea* (7) oso denbora luzean aldaketarik gabe izan den egoera da eta *gorria izatea* berezko ezaugarria dela pentsa daiteke. Hala ere, *gorria izateak* errealitateko egoera bati egiten dio erreferentzia eta proposizioak egia balioa du. Ondorioz, gertatzen da eta gertaeratzat hartuko dugu.

(7) Txinako bandera **gorria da**.

Euskarazko terminoa definitzeari dagokionez, *gertakari* eta *gertaera* terminoak baliatu izan dira. *Euskal Gramatika Lehen Urratsak II* lanean (Altuna *et al.*, 1987) bi terminoak erabili dira errealitatean gertatzen diren ekintza edo egoerei erreferentzia egiteko. Hona hemen gramatika horretako bi pasarte: i) "... adizki aspektudunek *gertaera* gertagarriago bat edo adierazten bide baitute" (429 or.); eta ii) "... testuinguru berezi batzutan iragan burutua, aspektu perfektua adieraztean, gaurko *gertakari* bati buruz erabil dezakegula" (425 or.).

Sareko Euskal Gramatikan¹ ere antzera gertatzen da eta *gertaera* eta *gertakari* testuinguru bertsuetan kontzeptu berari erreferentzia egiteko baliatzen dira. Ikerketa-lan honetan *gertaera* terminoa baliatuko dugu gertatzen diren ekintzak eta egoerak adierazteko.

2.1.2 Denbora-adierazpenak

Denbora-adierazpenak denbora-informazioaren prozesamenduko entitate garrantzitsuenak dira gertaerekin batera, gertaerak gertatzen diren denborak adierazten baitituzte. Hainbat autorek denbora-adierazpenen definizioa eman dute. Esaterako, Schilderrentzat eta Habelentzat (2001) denbora-adierazpenak nolabaiteko denbora-informazio zuzen edo inferitutakoa adierazten duten testu-zatiak dira. Ahn *et al.*ek (2005) diotenez, denbora-adierazpenak denbora-uneei edo tarteei erreferentzia zuzena egiten dieten hizkuntza naturaleko sintagmak dira. Azkenik, Bittarek (2010) definizio sinple bat ematen

¹<http://www.ehu.eus/seg/aurkezpena>

du: denbora-adierazpenak denbora adierazteko erabiltzen diren adierazpenak dira.

Aurreko definizioetan *temporal expression* terminoa erabili izan bada ere, bibliografian arrunta da *temporal expression* eta *time expression* balio berarekin erabiltzea. Hala egiten dute, esaterako, Ahn *et al.*ek (2005) eta Hacıoglu *et al.*-ek (2005). Guk *time expression*en alde egin dugu, uste baitugu hobeto irudikatzen duela kontzeptua eta euskaraz *denbora-adierazpen* terminoa erabiltzen dugu. Lehenengo terminoa irekiagoa da, bibliografian denbora-informazioa adierazten duten egiturak (gertaerak eta seinaleak, esaterako) era orokorrean adierazteko ere erabili baita.

2.1.3 Seinaleak

Denbora-informazioaren prozesamenduan ingelesezko *signal* terminoa baliatu da gertaeren edota denbora-adierazpenen arteko erlazioak esplizitu egiten dituzten testuetako elementuak izendatzeko. Zehazki, seinaleak dira denbora-objektuak beren artean nola erlazionatzen diren adierazten duten testu-zatiak, normalean funtzio-hitzak (Pustejovsky *et al.*, 2003a). Beraz, *seinale* terminoa denbora-erlazioak adierazten dituzten postposizioak eta karaktereak adierazteko baliatu dugu.

Berez, *signal* edozein erlazio esplizitu egiten duen elementua adierazteko baliatu da denbora-informazioaren prozesamenduan. Adibidez, TimeMLren lehen bertsioan (Pustejovsky *et al.*, 2003a) mendekotasun-erlazioetako seinaleak aztertu ziren eta NewsReader proiektuan (Tonelli *et al.*, 2014) kausazko erlazioak adierazten zituzten seinaleei C-SIGNAL etiketa esleitzen zitzairen. Ikerketa lan honetan, ordea, denbora-erlazioetako seinaleak baino ez ditugu kontuan hartuko.

2.1.4 Erlazioak

Tesi-lan honetan gertaeren edota denbora-adierazpenen arteko erlazioak landuko ditugu. Hiru motako erlazio semantikoak definitu ditugu: denbora-erlazioak, mendekotasun-erlazioak eta aspektu-erlazioak. Denbora-erlazioak gertaeren edota denbora-adierazpenen artean sortzen diren erlazioak dira. Horien bidez gertaerak edo iraupenak kronologian ainguratzen dira edo gertaeren iraupena adierazten da. Mendekotasun-erlazioak gertaera baten eta bere argumentu edo mendeko den beste gertaera baten artean sortzen direnak dira. Aspektu-erlazioak gertaera baten fasea (hasiera, amaiera, etab.)

adierazten duen gertaera baten eta beste gertaera baten artean sortzen dira eta bigarren horren fasea adierazteko baliatzen dira.

2.2 Aurrekari teorikoak

Azken hamarkadetan, ikerketa ugari egin dira denbora-informazioaren alorrean, eta horiek tratamendu automatikoan erabiltzen den denbora-informazioari buruzko analisiaren oinarriak ezarri dituzte. Gertaerak dira esaldiaren muina, eta denbora-informazio asko dakarte. Ikertzaile batzuek —esaterako, Vendlerrek (1957) eta Smithek (1991)— gertaeren tipologian jarri dute arreta; beste aditu batzuek, berriz, aditz-aspektuan eta tempusean (Reichenbach, 1947). Beste ikerketa-alor bat denbora-tarteen arteko erlazioen analisia izan da (Allen, 1983). Hori dela eta, gertaerak ez ezik, gizakiok gertaera horiek denbora-lerroan nola kokatzen ditugun ere aztertu da.

Ikusten denez, alderdi teoriko horiek hainbat eratako gaiak aztertzen dituzte; horiek guztiak ondorengo ataletan aurkeztuko ditugu. Lehenik, 2.2.1 atalean, denbora-informazioaren prozesamenduarekin lotura estuena duten gertaeren sailkapenetako batzuk deskribatuko ditugu eta bigarrenik, 2.2.2 atalean, denboraren kontzeptualizazioari buruzko teoriak deskribatuko ditugu.

2.2.1 Gertaeren sailkapenak

Aurreko definizioetan ikusi bezala, gertaerak ez dira soilik gertazen diren gauzak, nola gertatzen diren ere kontuan izan behar da. Definizio askotan jaso da gertaerek denborarekin duten erlazioa, erlazio horren arabera gertaerak mota batekoak edo bestekoak izan baitaitezke. Bestalde, autore batzuek eraldaketa ere aipatu dute, gertaerak aldaketa adierazi edo ez mota desberdinekoak izango baitira.

Hizkuntzalari askok gertaerak edo predikatuak beren ezaugarri linguistikoaren arabera sailkatu dituzte. Guk sailkapen sintaktikoak alde batera utzita, denborarekin lotura duten gertaeren sailkapen semantikoak aztertu ditugu, horiek baitira gure atazari gehien egokitzen zaizkionak. Jarraian emango dugu horien berri eta 2.2 taulan laburbilduko ditugu.

2.2.1.1 Vendlerren sailkapena

Vendlerrek (1957) aditzak aztertzen ditu eta denboran gertatzen diren moduaren arabera sailkatzen ditu. Vendlerren arabera, aditz-denborak² ez ditu soilik lehenaldia, orainaldia eta etorkizuna adierazten, aditz horiek adierazten dituzten gertaerak denboran nola gertatzen diren ere adierazten du. Aditzen sailkapena egin zuen aditzek har ditzaketen aditz-denboretan oinarrituta eta lau aditz-mota identifikatu zituen:

- **Egoerak** (*states*): denboran luzatzen diren gertaerak eta aspektu jarraituko aditzik onartzen ez dutenak. Ez dira denbora-tarte zehatz batean gertatzen.

(8) Oso **maite dut** opera.

- **Jarduerak** (*activities*): denboran luzatzen diren baina amaierarik ez duten eta aspektu jarraituko aditzak onartzen dituzten ekintzak dira. Ez dira denbora-tarte zehatz batean gertatzen.

(9) **Irakurtzen** ari naiz.

- **Burutzapenak**³ (*accomplishments*): hainbat fase eta amaiera dituzten eta aspektu jarraituko aditzik onartzen ez duten prozesuak dira. Denbora-tarte zehatz batean gertatzen dira.

(10) Orriak pasa eta pasa liburua **irakurri dut**.

- **Lorpenak** (*achievements*): une batean gertatzen diren eta aspektu jarraituko aditzik onartzen ez duten egoera-aldaketak dira. Denbora-tarte zehatz batean gertatzen dira.

(11) Mendian gora nindoan eta tontorrera **heldu naiz**.

Aditz-mota horietatik jarduera, erdiespen eta lorpen motakoak gertaeratzat hartzen ditu Vendlerrek. Egoerak, ordea, ez ditu gertaeratzat hartzen. Hala ere, Vendlerren arabera, egoera batzuek testuinguru batzuetan gertaeren ezaugarriak hartzen dituzte, eta gertaerak direla esan daiteke (*zutik egon, eserita egon, etzanda egon*, etab.).

²Kontuan izan behar da ingelesezko *tensek* tempusa eta aspektua hartzen dituela.

³Hala erabiltzen du Azkaratek (2008a).

Hala ere, Carlsonen (1977) gaineratzen duenez, Vendlerren zerrendako aditz guztiek izan dezakete ez-gertaerazko irakurketa bat. Adibidez, *Jonek erretzen du* perpausa *Jon erretzailea da* egoeraren parekoa izan daiteke lehenaren ohiturazko zentzua kontuan hartuz gero.

2.2.1.2 Croften sailkapena

Croftrek (2015) gertaeren sailkapen berri bat proposatu zuen, gertaerek denboran duten gauzatzearen arabera. Lau kategoria proposatu zituen: egoerak, lorpenak (berehalako egoera-aldaketak), ekintzak (irauten duten prozesu mugagabeak) eta jarduerak (prozesu mugatuak). Gainera, azpisailkapen bat ere eskaini zuen. 2.1 taulan ageri dira klase horiek eta horien azpiklaseak.

2.1 taula – Croften gertaeren sailkapena

Gertaera mota	Gertaera azpimota		Adibideak
Egoerak	iragankorrak		Bidea bustita dago.
	iraunkorrak	eskuratutakoak	Leihoa apurtuta dago.
		berezkoak	Marie frantsesa da.
	puntua		Goizeko 5ak dira.
Lorpenak	itzulgarri zuzendua		Leihoa zabalik dago.
	itzuliez in zuzendua		Leihoa birrinduta dago.
	ziklikoa		Txakurrak zaunka egin zuen.
Ariketak	zuzendua		Zopa hoztu da.
	ez-zuzendua		Neskek abestu zuten.
Ekintzak	erdiespen progresiboak		Sagarra jan nuen.
	erdiespen ez-progresiboak		Harryk ordenagailua konpondu zuen.

2.2.1.3 Coterónen sailkapena

Coterónen (2015) aditz jokatuaren eta gertaeren arteko erlazioa lantzen du. Semantika aldetik aditzek eta horien inguruan egituraturiko predikatuek oro har gertaera bat errepresentatzen dutela adierazten du. Gertaeraren ideia (*eventuality*) Bach (1986)-en lanetik hartzen du eta gertaerentzat ondoko sailkapena proposatzen du:

- Mementoak:

(12) Garbigailuak **eztanda** egin du.

- Ez-mementoak:

(13) Azken hilabeteak doktoretza-tesia **idazten** igaro ditut.

- Aldaketa bat barneratzen dutenak:

(14) Ikusleak **eseri dira** saioa hastear zegoela.

- Ez dutenak aldaketarik barneratzen:

(15) Egunean zehar **esna** egoten naiz.

- Dinamismoa txertatzen dutenak:

(16) Larunbat goizetan menditik **ibiltzea** gustatzen zait.

- Ez dutenak dinamismorik txertatzen:

(17) Gure bulegoan guztiok **gara** euskaldunak.

Sailkapenean nabari da gertaera-unearen eta gertaeraren arteko harremana, mementoekoak eta ez-mementoekoak bereizten dituenen.

2.2.1.4 Hallidayren sailkapena

Hallidayren (2004) arabera, prozesuak denboran kokatu eta denboran zehar garatzen dira eta partaideak dituzte. Esan daiteke Hallidayren arabera prozesu izena hartzen dutenak eta beste autoreen arabera gertaera edo ebento izena hartzen dutenak hurbilekoak direla. Hala, guk ere *prozesu* eta *gertaera* hurbileko esanahia duten terminotzat hartuko ditugu.

Hallidayk 6 prozesu mota desberdintzen ditu:

- **Materialak** (*material*): norberarengandik kanpo gertatzen direnak dira.

(18) 25 milioi **ordainduko** ditu talde aljeriarrak.

- **Jokabidezkoak** (*behavioural*): norberaren barruan gertatzen direnak eta kanpoan isla dutenak. Kontzientzia-prozesuak eta egoera psikologikoak.

(19) Umeak pailazoaren keinuekin **barre egiten** ari dira.

- **Mentalak** (*mental*): norberaren barruan gertatzen direnak dira.

(20) Langileek ez dituzte baldintzak **onartu**.

- **Hitzezkoak** (*verbal*): hizkuntzaren bidez adierazten diren norberaren kontzientzian eraikitako erlazio sinbolikoak dira.

(21) “Lehentasuna lanera itzultzea da”, **ziurtatu du** Prevostek.

- **Erlaziozkoak** (*relational*): orokortzeak adierazten dituztenak dira eta zerbait identifikatzeko eta horren ezaugarriak adierazteko balio dute.

(22) Txinako bandera gorria **da**.

- **Izatezkoak** (*existential*): edozein motatako fenomenoak badirela aitortzeko erabiltzen direnak.

(23) Gaur euria **ari du**.

2.2.1.5 Setzerren sailkapena

Setzerrek (2001) Vendlerren (1957) eta Hallidayren (2004) sailkapenak aztertzen ditu eta aspektuetan oinarritutako gertaeren sailkapena denbora-informazioaren prozesamenduan ez dela egokia ondorioztatzen du. Bere hitzetan, Vendlerren aspektuan oinarritutako analisiak esaldiz esaldiko analisi eskatzen du eta hori ez da posible paragrafo laburrak dituzten testuetan. Halaber, etiketatzailearentzat gertaera bakoitza puntua, kulminazioa, kulminatutako prozesua edo prozesua den erabakitzea lan handiegia litzatekeela gaineratzen du.

Hallidayren hurbilpena ere ez zaio egoki iruditzen Hallidayrentzat *relational processes* diren horiek korapilatsuak izan baitaitezke. Izan ere, multzo horretakoak dira *azoka ostegunean izango da* eta *Mikelek piano bat du* modukoak. Lehenak une konkretu batean (*ostegunean*) gertatuko den gertaera baten (*azoka*) berri ematen du, baina bigarrenak egoera estatiko baten (*pianoa izatea*) berri ematen du.

Ondorioz, sailkapen horien ordeztu, gertaerak denboran kokatzen lagunduko zuen sailkapen bat proposatu zuen:

- **Jazoerak** (*occurrence events*): denbora-grafoan agertuko diren gertaerak.

- **Diskurtsiboak** (*reporting events*): informazio-iturria eta gertaera bat lotzen dituzten gertaerak. Normalean, adierazitako gertaeraren ondotik gertatzen dira eta adierazitako gertaera hori denboran kokatzen laguntzen dute.
- **Hautematezkoak** (*perception events*): gertaera bat “sentitu” dela adierazten dute. Normalean, gertatutako gertaera eta hori sentitzea batera gertatzen dira.
- **Aspektuzkoak** (*aspectual events*): gertaera bat hartzen dute argumentutzat eta gertaera horren hasiera edo amaiera adierazten dute.
- Beste gertaera mota batzuk:
 - **Jarrerazkoak** (*attitude events*): gertaera bat hartzen dute argumentutzat, baina ez dute gertaera horren errealitatea bermatzen, ezta gertaera horrekiko denbora-erlazio argirik ere adierazten.
 - **Hipotetikoak** (*hypothetical events*): ondorioen arabera denbora-lerro hipotetikoak sortzen dituzte.

Ikusten denez, gertaerak sailkatzeko modu horrek gertaera mota bakoitzaren eta denboraren arteko harremana azaleratzen zuen. Nabarmenezkoa da ere ez zituela egoerak aintzat hartu, horiek denboran kokaezinak zirela uste baitzuen. Gerora ere TimeMLko gertaeren sailkapena eraikitzeke oinarritzat hartu zen Setzerren sailkapena.

2.2.1.6 TimeMLren sailkapena

TimeML markaketa-lengoaia (Pustejovsky *et al.*, 2003a) garatzean, gertaerak sailkatzeko kategoriak definitu ziren. Oinarrian argia da Setzerren eragina, baina desberdintasun nabarmen bat ere badago: TimeMLn egoerak gertaera motatzat hartu ziren, egileen arabera horiek ere denboran koka baitaitezke. Hala, ondoko gertaeren sailkapena proposatu zen TimeMLrako:

- **Jazoerak**: gertatzen diren ekintzak dira orokorrean.
- **Egoerak**: zerbaitek egia balioa lortu edo mantentzen duen zirkunstantziak adierazten dituzte.

- **Diskurtsiboak:** pertsona batek edo erakunde batek zerbait aldarrikatzen duela, gertaera bat kontatzen duela edota hartzaileei informazioa ematen diela, eta abar adierazten duten gertaerak.
- **Hautematezkoak:** gertaera bat fisikoki antzematea adierazten duten gertaerak.
- **Aspektuzkoak:** predikatu horiek gertaera bat hartzen dute argumentutzat eta bere denbora-egituraren (gertaera edo denbora-adierazpena) aspekturen bati (hasiera, jarraipena edo amaiera) egiten dio erreferentzia.
- **Ekintza intentsionalak:** testuan esplizituki agertzen den gertaera bat argumentutzat aukeratzen duten gertaera dinamikoak dira.
- **Egoera intentsionalak:** ekintza intentsionalek bezala, gertaera bat hartzen dute argumentu gisa eta zirkunstantziak adierazten dituzte.

TimeML markaketa-lengoaia denbora-informazioaren etiketatzerako estandarra bihurtu denez, hainbat autorek aldaketak proposatu dizkiote. Bittarek (2010), esaterako, beste hiru gertaera-klase proposatzen du: modalak (*modal*), kausa adierazten dutenak (*cause*) eta informazio gramatikala adierazten dutenak (*event container*). Hala eta guztiz ere, jatorrizko sailkapena da zabalduena.

2.2.2 Denboraren tratamendu teorikoa

Kronologia era formalean nola landu ere aztertu da. Zehazki, denbora nola kontzeptualizatu erabaki da eta denbora hizkuntzan nola adierazten den aztertu da.

2.2.2.1 Denbora-uneak vs interbaloak

Denbora adierazteko hainbat modu formalko dira. Hala ere, guk bi ikuspuntu konparatuko ditugu: denbora-uneetan oinarritutako teoriak eta interbaloetan oinarritutakoak.

Denbora-uneetan oinarritutako denbora-modeloetan, oinarrizko denbora-entitateak denbora-uneak dira eta horien arteko oinarrizko erlazioa (berdintasunezkoaz gain) lehentasunezkoa da (Goranko eta Galton, 2015). Hala,

2.2 taula – Gertaeren sailkapenerako irizpideak

Autoreak	Sailkapenerako irizpideak
Vendler (1957)	Aditzak denboran gertatzen diren moduaren arabera sailkatzen ditu. Aspektuaren arabera sailkatzen ditu.
Croft (2015)	Gertaerak denboran duten gauzatzearen arabera sailkatzen ditu.
Coterón Agorria (2015)	Aditz jokatuaren eta gertaeren arteko erlazioa erlazioa lantzen du.
Halliday (2004)	Prozesuak denboran nola gertatzen diren arabera sailkatzen ditu. Aspektuaren arabera sailkatzen ditu.
Setzer (2001)	Gertaerak denboran kokatzen laguntzen duen sailkapena proposatzen du.
Pustejovsky <i>et al.</i> (2003a)	Setzerri jarraituta denboran kokatzen laguntzen duen sailkapena proposatzen du.

denbora-uneak batera edo bata bestearen atzetik gertatzen direla baino ezin da onartu.

Denbora-uneetan oinarritutako denboraren tratamendua, ordea, ez da gai mundu errealeko gertaera guztiak adierazteko, ez baitu bi uneren gainjartzea onartzen. Denbora-tarteetan oinarritutako denbora-modeloak uneetan oinarritutakoak baino ontologikoki aberatsagoak dira erlazio posible gehiago baitago tarteen artean uneen artean baino (Goranko eta Galton, 2015). Interbaloek hasiera- eta amaiera-uneak dituzte eta horien bitartean interbaloak ez dauka etenik. Gainera, errepresentazio mota horrek denbora-uneak adieraztea ere ahalbidetzen du: uneak hasiera- eta amaiera-une berberak dituzten interbaloak dira.








Setzerrek (2001) ere, interbaloetan oinarritutako errepresentazioa defendatzen du bi arrazoiengatik:

- gizakia ez da gai enpirikoki oso denbora laburrean eta une bakarrean gertatzen diren gertaeren artean bereizketa egiteko. Ondorioz, ez da informaziorik galtzen interbaloetan oinarritutako tratamendua eginez gero.
- Denbora-uneak kronometriaren artefaktua baino ez dira eta horiek zenbakiz adieraztea guztiz erabaki arbitrarioa da (Galton, 1990).

Gainera, Setzerren arabera, denbora-unea beti irudika daiteke bi interbaloren arteko lotura-puntu gisa. Amaitzeko, denbora interbalotan irudikatzean, Allenen denbora-tarteen teoria (Allen, 1983) ahalbidetzen da.

2.2.2.2 Allenen denbora-tarteen teoria

Orobat, Allenen denbora-tarteen teoria (1983) oinarri teoriko nagusietako bat izan da denbora-informazioa aztertzeko. Gizakiok egiten ditugun denbora-inferentziak deskribatzea zen Allenen helburua. Denbora-tarteez osatuta zegoen eredua, munduan aldaketaren bat baldin badago soilik detektatzen baitugu gizakiok denbora. Gainera, Allenek zioen unitate txikiagoetan zatitu zitezkeela denbora eta gertaerak azpigertaeretan bereiz zitezkeela. 13 denbora-erlazio proposatu zituen: sei eta horien kontrakoak, eta identitate-erlazio bat (ikusi 2.1 irudia). Denbora-erlazio horiek denbora-tarte pareen (eta zegozkien gertaeren) arteko erlazio posible guztiak adierazten zituzten.

Relation	Symbol	Inverse	Meaning
X BEFORE Y	b	bi	
X MEETS Y	m	mi	
X OVERLAPS Y	o	oi	
X DURING Y	d	di	
X STARTS Y	s	si	
X FINISHES Y	f	fi	
X EQUAL Y	eq	eq	

2.1 irudia – Allenen denbora-tarteak⁴

Azken hamarkadetan, gertaeren tipologiari, gertaeren denbora-ezaugarriari eta kronologiari buruzko informazio hori hartu da kontuan denbora-informazioaren prozesamenduan.

⁴<http://stackoverflow.com/questions/12069082/allens-interval-algebra-operations-in-sql>

2.2.2.3 Denbora-adierazpenen sailkapena

Ikusi dugunez, denbora puntuka edo tarteka uler daiteke eta denbora-informazioaren prozesamenduan bi joerak uztartzen dituen bidea hautatu izan da denbora kontzeptualizatzeko. Puntu eta tarte horiek, ordea, ez dira beti berdinak, handiagoak edo txikiagoak izan daitezke, hau da, granularitate ezberdina izan dezakete. Setzerrek (2001), adibidez, hiru denbora-adierazpenak hiru multzotan sailkatu zituen:

- **Date:** eguna baino granularitate handiagoa adierazten duten denbora-adierazpenak

(24) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu ateak.

- **Time:** eguna baino granularitate txikiagoa adierazten duten denbora-adierazpenak

(25) Lehen Airbus A380a **18:40an** lurreratu da Singapur-en.

- **Complex:** Berez gertaera bati erreferentzia egiten dioten denbora-adierazpenak

(26) Lehen Airbus A380a Singapur-en lurreratu da **aireratu eta 12 ordura**.

Gaur egun, ordea, denbora-adierazpenen prozesamendurako nahiko orokortuta dago lau kategoriako denbora-entitateen sailkapena:

- **Datak:** eguna baino granularitate handiagoa adierazten duten denbora-adierazpenak

(27) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu ateak.

- **Orduak:** eguna baino granularitate txikiagoa adierazten duten denbora-adierazpenak

(28) Lehen Airbus A380a **18:40an** lurreratu da Singapur-en.

- **Iraupenak:** denbora-tarteak adieraztekoak

- (29) Lehen Airbus A380a Singapur-en lurreratu da **12 orduko** hegaldia egin ostean.

- **Errepikapenak**: denbora-tarte batean errepikatzen diren uneak

- (30) Igerilekura joaten naiz **asteen hirutan**.

Sailkapen horretan uneak eta interbaloak batera erabiltzea bi arrazoiengatik gertatzen da: i) una eta hasiera eta amaiera bera dituen interbaloa gauza bera direlako eta ii) denborazko balio-normalizatuekin era errazagoan bat etortzeko. Izan ere, 4.2.2.2 atalean ikus daitekeenez, TimeML eske-man ISO-8601 denbora-adierazpen normalizatuak erabiltzen dira denbora-adierazpenak errepresentatzeko. Hain zuzen ere, ISO-8601 arauak uneen zein iraupenen balioak normalizatzeko aukera ematen du.

2.2.2.4 Aditz-aspektuaren eta tempusaren teoriak

Gizakiok denbora era naturalean uneko mementoan, aurrekoan eta ondokoan banatzen dugu. Uneko mementoa orainalditzat hartuko dugu, aurrekoa lehenalditzat eta ondokoa etorkizuntzat. Halaber, denbora-tarte horiek tarte txikiagotan ere bana daitezke eta tarte bakoitzak aurreko eta ondoko tarteak ditu. Gertaerak horietako edozein mementotan gerta daitezke eta zein mementotan gertatu, halako aditz-denboraren bidez adierazten da.

Aditz-aspektuaren eta tempusaren azterketari dagokionez, Reichenbachek (1947) aditz-tempusari eta aditz-formen eta hizketa-unearen arteko erlazioari buruzko azterlan sakona aurkeztu zuen. Hark adierazi zuenez, hiru alderdik baldintzatzen dute aditz-denboraren aukeraketa: hizketa-unea (S), gertaera noiz jazo den (E), eta erreferentzia-unea (R) —hiztunak erreferentzia gisa hartu duen una.

Ingeleseko aditz-tempusen eskema bat proposatu zuen, gertaerak esate-unearekin eta hiztunak aukeratutako erreferentzia-unearekin duen erlazioan oinarrituta (2.2 irudia). Irudi horretan ikusten denez, hizketa-unearen (S), gertaeraren (E) eta erreferentzia-unearen (R) arteko erlazioan oinarrituta eraikitzen dira aditz-tempusak.

Reichenbachen arabera, ingelesez esate-unearen esparrua baino lehen gertatutako gertaerak lehenaldiaren bidez adierazten dira. Hizgai den gertaera esate-unearen esparruan gertatzen bada, orainaldiko forma baliatzen du hiztunak. Azkenik, gertaera esate-unearen ondoren gertatzen bada, etorkizuneko formak erabiltzen dira gertaera adierazteko.

Structure	Reichenbach's name	Traditional English
E-R-S	Anterior past	Past perfect (or pluperfect)
E,R-S	Simple past	Simple past
R-E-S	Posterior past	-
R-S,E		
R-S-E		
E-S,R	Anterior present	Present perfect
S,R,E	Simple present	Present
S,R-E	Posterior present	Simple future
S-E-R	Anterior future	Future perfect
S,E-R		
E-S-R		
S-R,E	Simple future	Simple future
S-R-E	Posterior future	-

2.2 irudia – Reichenbachen denbora-tarteak

Atal honetan denbora-informazioa nola kontzeptualizatu den aztertu dugu. Hala ere, beste urrats bat egin behar da informazio hori automatikoki prozesatzeko: informazioa ordenagailu bidez irakurtzeko moduan jartzea. Hurrengo ataletan, ataza horretarako sortutako baliabide nagusiak aurkeztuko ditugu.

2.3 Markaketa-lengoiak eta eskuz etiketatze-ko tresnak

Denbora-informazioa ordenagailu bidez irakurtzeko moduan jarri ahal izateko, denbora-informazioa dakarten tokenak —hutsuneek edo puntuazioak bereizten dituzten karaktere segidak— etiketatu egin behar dira, eta horien ezaugarriak erauzi. Token horiek zenbait denbora-egitura adieraz ditzakete —hain zuzen ere, denbora-adierazpenak, gertaerak, seinaleak eta erlazioak (ikusi 3. kapitulua)—, eta, horiek etiketatzeko, markaketa-lengoiak garatu dira. Atal honetan, denbora-informazioa egituratzeko eta normalizatzeko sortu diren markaketa-lengoiak aurkeztuko ditugu (2.3.1 atala), baita denbora-informazioa eskuz etiketatzeko erabiltzen diren tresnak ere (2.3.2 atala).

2.3.1 Denbora-informazioa etiketatzeako markaketa-lengoaiak

Denbora-informazioa etiketatzeako markaketa-lengoaiak etiketa bat esleitzen diete denbora-egiturei, eta atributu eta balio batzuk ematen dizkiete beren ezaugarriak deskribatzeko. Hasieran, denbora-adierazpenak eta horien ezaugarri nagusietako batzuk soilik etiketatzen ziren, baina markaketa-lengoaiak konplexuagoak bihurtu ziren berehala, eta, gaur egun, denbora-informazioa-rekin erlazionatutako alderdi gehienak biltzen dituzte.

Denbora-informazioa etiketatzeako lehenengo markaketa-lengoaietako bat MUC (Message Understanding Conferences) kongresuetarako garatu zen (Grishman eta Sundheim, 1996), SGMLen⁵ oinarrituta. MUC kongresuen lehenengo ediziotik bosgarrenera, entitateak eta gertaerak identifikatu ziren. MUCen seigarren edizioan (1996), aldiz, denbora-informazioa era osotuagoan etiketatu zen, gertaera noiz gertatu den identifikatu ahal izatea oso garrantzitsua baitzen testuak automatikoki ulertzeko.

Orduraz geroztik, ugaritu egin dira denbora-informazioari buruzko ikerketak. Lehenik, Manik eta Wilsonek (2000) TIMEX etiketa proposatu zuten denbora-adierazpenetarako, baita ISO-8601 arauan (ISO Committee, 1997) oinarritutako denbora-informazioa normalizatzeko eskema ere. Etiketak atributu batzuk zituen —adibidez, **VAL** (*value*) eta **POS** (*part-of-speech*)—, eta horiek agerian uzten zituzten denbora-adierazpenen ezaugarriak.

TIDES (Translingual Information Detection, Extraction and Summarisation) programan (Cieri eta Liberman, 2002), TIMEX2 etiketa (Ferro *et al.*, 2003) proposatu zen denbora-adierazpenak etiketatzeako. Etiketa horrek hainbat ezaugarri zituen; besteak beste, **VAL** (balioa), **MOD** (modifikatzaileak), **SET**, **PERIODICITY** eta **GRANULARITY**. Setzerrek ere (2001) etiketa hori erabili zuen, baina urrunago joan zen, eta denbora-informazioa etiketatzeako lehenengo eskema osoa proposatu zuen. Etiketa bat proposatu zuen gertaeretarako eta denbora-erlazioetarako, eta horien atributuak eta balio posibleak zehaztu zituen gertaeren eta erlazioen ezaugarriak azaltzeko. Beste hizkuntza batzuetako denbora-informazioa etiketatzeako ere erabili zen TIMEX2; besteak beste, koreerakoa (Jang *et al.*, 2004) eta alemanekoa (Strötgen eta Gertz, 2011).

TimeML (Pustejovsky *et al.*, 2003a) desberdina zen ordura arte proposatutako markaketa-lengoaien aldean: i) gertaerak denbora-adierazpenekin

⁵Standard Generalized Markup Language

ainguratzen ditu; ii) gertaera-adierazpenak elkarren artean ordenatzen ditu; eta iii) denbora-adierazpen zehaztugabeen bi urratseko interpretazioa ahalbidetzen du. Aitzindarien ezaugarriak ere hedatu zituen: etiketen atributuak eta gertaeren sailkapena zabaldu zituen, denbora-adierazpen erlatiboak prozesatzeko aukera zabaldu zuen eta gertaeren eta denbora-adierazpenen arteko erlazioak sortu zituen. TimeBank corpusarekin batera proposatu zen lehen aldiz TimeML (Pustejovsky *et al.*, 2003b), denborazko galderak aztertuko zituen galderei erantzuteko sistema bat sortzeko.

Ordutik, TimeML hobetu egin da. Nazioarteko ISO arau bihurtu da (ISO-TimeML working group, 2008), eta azken bertsioa (TimeML Working Group, 2010) markaketa-lengoaia fidagarri bihurtu da denbora-informazioa etiketatzeko. 4.1.1 atalean deskribatuko dugu xehetasunez TimeML.

Geroago, beste markaketa-lengoaia batzuk sortu ziren TimeML lengoaiaren oinarrituta. SIBILA proiektuan (*Sistema automático de respuestas basado en un modelo de discurso*) (Wonsever *et al.*, 2012), gertaerak eta horien faktualitatea aztertzen ziren. TimeML lengoaiak gertaerak noiz gertatzen diren aztertzen zuen; eta SIBILAk, berriz, gertaera bat gauzatu den ala ez berretsi ote daitekeen.

PLIMEX ere (Kocoń *et al.*, 2015) TimeMLn oinarritu da. Zehatzago, polonierazko denbora-adierazpenetarako TIDESeko TIMEX2 eta TimeMLko TIMEX3 etiketen egokitzapenak konbinatzen ditu. Denbora-adierazpenak TimeMLko mota berberetan sailkatzen dira: data, ordua, iraupena eta errepikapena. Gertaerak identifikatzeko eta sailkatzeko eskema bat ere proposatu zuten (Marciniczuk *et al.*, 2015); eskema horrek TimeMLko gertaeren sailkapenari jarraitzen zion.

NewsReader proiektuan⁶ (Vossen *et al.*, 2014), denbora-informazioa etiketatzeko egokitu zen TimeML (Tonelli *et al.*, 2014). Nahiz eta etiketa nagusiak gorde, horien ezaugarrietako batzuk aldatu egin ziren ataza horretarako, eta kausa- eta gramatika-erlazioak gehitu. Gertaeren eta entitateen arteko korreferentziarekin eta rol semantikoei buruzko informazioarekin konbinatu zen denbora-informazioaren etiketatzea.

TEMANTEX (Wonsever *et al.*, 2015) denbora-informazioa etiketatzeko eskema bat da, eta, denbora-adierazpenak ez ezik, denborazkotasunaren beste adierazle lexiko batzuk ere etiketatzen ditu, “denbora-adierazle” deituri-koak; hau da, testuetako denboraren interpretazioan eragina duten elementuak. Nahiz eta nagusiki bateragarria izan TimeMLrekin, ez dio jarraitzen

⁶<http://www.newsreader-project.eu/>

ISO-8601 arauari (ISO Commitee, 1997), eta denbora-adierazpenen eta atributuen beste sailkapen bat eskaintzen du.

2.3 taula – Denbora-informazioa etiketatzeko markaketa-lengoaiak

Markaketa-lengoaia	Etiketatzeko entitateak	Lehen proiektua
TIMEX (Mani eta Wilson, 2000)	Denbora-adierazpenak	MUC kongresuak
TIMEX2 (Ferro <i>et al.</i> , 2003) eta (Setzer, 2001)	Denbora-adierazpenak	TIDES (Cieri eta Liberman, 2002)
TimeML (Pustejovsky <i>et al.</i> , 2003a)	Gertaerak, denbora-adierazpenak, seinaleak eta denbora-, aspektu- eta mendekotasun-erlazioak	TimeBank
SIBILA (Wonsever <i>et al.</i> , 2012)	Gertaerak eta horien faktualitatea	SIBILA
PLIMEX (Kocoń <i>et al.</i> , 2015)	Denbora-adierazpenak (eta gertaerak)	
NewsReader (Tonelli <i>et al.</i> , 2014)	Gertaerak, denbora-adierazpenak, seinaleak eta denbora-, gramatika-, kausa- eta mendekotasun-erlazioak	NewsReader (Vossen <i>et al.</i> , 2014)
TEMANTEX (Wonsever <i>et al.</i> , 2015)	Denbora-adierazpenak eta denbora-adierazleak	

2.3 taulan, goian aurkeztutako markaketa-lengoaien laburpen bat dago. 2.5 ataleko corpusak markaketa-eskema horiei jarraikiz etiketatu dira.

2.3.2 Eskuz etiketatzeko tresnak

Urre-patroi corpusak funtsezkoak dira HPn, kalitate handiko etiketatzeak eskaintzen baitituzte tresnen garapenerako, prestakuntzarako eta ebaluaziorako. Denbora-informazioaren prozesamenduan, corpusek garrantzi berezia dute, gertaerak eta denbora-adierazpenak beren testuinguruen arabera soilik interpreta baitaitezke zehatz-mehatz. Hain zuzen ere, gertaeren eta denbora-adierazpenen arteko erlazioak oso garrantzitsuak dira alor honetan,

denbora-informazioaren ordena erlazio horietatik ondorioztatzen baita. Ondorioz, denbora-egiturak ez ezik, horien testuinguruak ere hartu behar dira kontuan. Ondorioz, denbora-egiturez gain, horien artean sortutako erlazioak ere etiketatu behar dira.

Aski jakina da kalitate oneneko corpusak eskuz etiketatuta lortutakoak direla. Hala eta guztiz ere, corpusak formatu digitaleko testu-bilduma gisa ulertzen direnez, etiketatze-tresnen bidez gehitzen dira etiketatzeak. Tresna horiek eskaintzen duten interfazeaz, etiketatzailleek eskuz gehitzen dute informazio linguistikoa dokumentuetan eta normalizatu eta tresna automatikoen zat atzigarri ere egin ditzakete etiketatzeak. Oro har, etiketa eta atributu batzuk definitzen dira tresnan; etiketatzailleek aukeratzen dute zer testu zati etiketatu nahi duten, eta etiketa jakin bat esleitzen diote. Eskuz etiketatzeko oinarritzko interfazeaz gain (bertan, etiketatzeko eskema bat kargatzen da, eta etiketak eta atributuen balioak eskuz gehitzen dira), tresnek ezaugarri espezifiko batzuk ere eskaintzen dituzte.

Denbora-informazioa etiketatuta duten corpusak sortzeko erabili zen lehenengo etiketatze-tresna Alembic Workbench izan zen (Day *et al.*, 1997). Ataz horretarako garatu zen berariaz, MUC kongresuetan, eta etiketatze automatikoa eta eskuzkoa konbinatzeko aukera ematen zuen. Hain zuzen ere, *bootstrap* prozedura bat zerabilen, eta, haren bidez, eskuzko lehenengo etiketatzea arauak sortzeko eta ondorengo dokumentuak automatikoki etiketatzeko berrerabiltzen zen. Ondorioz, pixkanaka, etiketatzeko ahalegina eskuzko etiketatzea izatetik zuzenketa izatera pasatzen zen. Hurbilketa horrek izugarri bizkortzen zuen corpusak eraikitzeke lana. Horretaz gain, eskuen mugimendua murrizteko prestatu zen tresna, eskuzko etiketatzea bizkortzeko.

Eskuzko eta arauetan oinarritutako etiketatzeak konbinatzeaz gain, tresna hibrido honek puntuagailu bat ere bazekarren SGML markaketa ausaz hautatu eta puntuatzeko. Hala ere, ez zen erabili MUC kongresuetan soilik, baita Schilderren eta Habelen (2001) esperimentuan ere. Libreki banatu zen, baina dagoeneko ez dago erabilgarri⁷.

Callisto⁸ (Day *et al.*, 2004) eskuz etiketatzeko tresna bat da, eta modularitatea eta konfiguragarritasuna ziren ezaugarri nagusiak. Eredua-Ikuspegia-Kontrolagailua diseinua eskaintzen zuen, erabiltzaile-interfazeke osagaien independentzia bermatzeko eta, aldi berean, beti informazio koherentea erakus-

⁷http://annotation.exmaralda.org/index.php?title=Alembic_Workbench

⁸<https://www.mitre.org/research/technology-transfer/open-source-software/callisto-0>

ten zuen. Bestela esanda: etiketatzean egindako aldaketa oro osagai grafikoe-tara zabaltzen zen, etiketatzailari aurkezten zaion ikuspegia eguneratzeko. Atazetarako eredu espezifikoak zuzenean ezartzen ziren iturburu-kodean; horri esker, atazetarako logika espezifikoak eta erabiltzaile-interfaze grafikoak txerta zitezkeen.

Callisto ATLAS⁹ (Bird *et al.*, 2000) delakoaren testuinguruan sortu zen, seinale linguistikoak eta horien etiketatzeen abstrakzioa lortzeko. Hizkuntza anitzetan etiketatzeak tresna gisa eraiki zen, Javak ezin hobeki kudeatzen baitzituen Unicode karaktere-kodeketak eta tresnak automatikoki aukeratzeko letraz-tipoa kodeketa bakoitzerako. Gainera, Callisto Java tresna gisa sortu zenez, atazak bereiz gorde eta exekuta zitezkeen. Horretaz gain, atazak ez ziren zertan zerotik sortu, antzeko zereginak aurrekoetan aldaketa txikiak eginda sor zitezkeelako. Azkenik, Callistok XMLn¹⁰ oinarritutako Atlas Interchange Format (AIF) kudea zezakeen, baita SGMLn edo XMLn oinarritutako beste markaketa batzuk ere. Hortaz, tresna sarri erabili zen denbora-informazioa etiketatzeak, eta antzeko zereginak definitu ziren; besteak beste, TIMEX2 normalizazioa, TimeML etiketatzea eta gertaeren etiketatzea. Tresna kode irekikoa da, eta, egun, libreki deskarga daiteke¹¹.

Tango (Verhagen *et al.*, 2006) denbora-erlazioak etiketatzeak eta frantse-seko TimeBank sortzeko erabili zuten. Orduko etiketatzeak tresnen aldean, Tangok grafikoki erakusten zuen denbora-informazioa. Bertan, gertaerak eta denbora-adierazpenak nodoak ziren; eta haien arteko denbora-erlazioak, berriz, arkuak. Etiketatzeak beste tresnek, ordea, taulen bidez irudikatzen zuten informazio hori. Nodoak pantailako edozein lekutan jar zitezkeen, eta denbora-lerro bisualak sortzeko antolatatu. Laguntzaile bat ere bazuen —hark erlazio posibleak iradokitzen zituen, eskuz etiketatutako loturen kopurua handitzeko—, baita T-BOX irudikapena ere —haren bidez, denbora-entitateak kutxa gisa irudikatzen dira—. Erlazioak kutxa gisa irudikatzeari dagokionez, inklusioak denbora-inklusiok esan nahi du; metatutako kutxek, berriz, aldiberekotasuna.

Tango Callisto ingurunean exekutatzen zen, ataza zehatz gisa. Callistoren jatorrizko atazak gertaerak, denbora-adierazpenak eta seinaleak etiketatzeak balio zuen; aitzitik, Tangoen atazak denbora-, mendekotasun- eta aspektu-erlazioak etiketatzeak ingurunea eskaintzen zuen. Tango tresna libreki esku-

⁹Architecture and Tools for Linguistic Analysis Systems

¹⁰Extensible mark-up language

¹¹<https://github.com/mitre/callisto>

ratu daiteke, <http://timeml.org/tango/tool.html> helbidean.

Brandeis Annotation Tool (BAT) (Verhagen, 2010) etiketatze-geruzetan eta atazen deskonposizioan oinarrituta zegoen. Horrela, errazagoa zen etiketatze-atazak bateratzea eta ataza konplexuak azpiataza sinpleetan zatitzea. Adibidez, bereiz mantentzen zituen hedaduraren eta atributuen geruzak, eta, horrela, errazagoa zen beste iturri batzuetatik hartutako etiketak irudikatzea.

Aurretik aurkeztutako etiketatze-tresnetan ez bezala, administratzailearen zeregina funtsezkoa zen BATen. Administratzaileak corpusak, geruzak eta atzak sortu eta edita zitzakeen, etiketatzaileak eslei ziezazkiokeen ataza bakoitzari, eta etiketatzaileen aurrerapena kontrola zezakeen.

Lehenengo aldiz, Tempeval ebaluazio-atazarako erabili zen, denbora-tarte labur batean etiketatu behar izan baitziren hainbat testu. Informazio teknikoari dagokionez, PHP¹²n eta Javascripten oinarrituta dago eta MySQL datu-baseetan gordetzen dira datuak. Sarean zegoen eskuragarri, eta, horri esker, askoz ere errazagoa zen eguneratzeak eta bertsio-berritzeak kudeatzea. Tamalez, dagoeneko ez dago erabilgarri.

Content Annotation Tool¹³ (CAT) (Lenzi *et al.*, 2012) sarean erabilgarri dagoen etiketatze-tresna bat da, eta etiketatze linguistikoa egiteko balio du. Geruza anitzetan etiketatzeko aukera ematen du; hau da, dokumentu bakarrean hainbat etiketatze konbina ditzake. Haren indargune nagusiak intuitibotasuna eta pertsonalizazio-ahalmena dira. Gainera, bilaketak egiteko eta adostasuna neurtzeko aukera eskaintzen du, baita tokenizatzailea ere; hori dela eta, testu gordina nahiz tokenizatutako testu-fitxategiak onartzen ditu sarrera gisa. CATek XML fitxategiak sortzen ditu irteera gisa, eta DTD¹⁴ fitxategiak automatikoki sortzen dira etiketatzea koherentea izan dadin. Interfazean, etiketatzeak testu gisa bistaratzen dira; eta erlazioak, berriz, taula-formatuan.

Maiz, etiketatze-ahaleginak atazei oso lotuta daude, eta erraza izan behar luke etiketatze-eskemak tresnetan integratzea. CAT erabiltzeko ez da behar programazioan trebea izatea, eta erraz sor daitezke etiketatze-eskemak interfazean. Ataza askotan erabili da CAT, besteak beste, italierako TimeBank sortzerakoan, EVENTI (EVALITA 2014) atazan eta testuak etiketatzeko TERENCE¹⁵, Excitement¹⁶ eta NewsReader proiektuetan. Libreki eskuratu dai-

¹²PHP: Hypertext Preprocessor

¹³Lehen CELCT Annotation Tool

¹⁴Document Type Declaration

¹⁵http://vittorini.univaq.it/?page_id=869

¹⁶<https://sites.google.com/site/excitementproject/home>

teke¹⁷, erabiltzaile-profila aurretiaz eskatu behar den arren.

Inforex etiketatze-tresna (Marcinčuk *et al.*, 2017) JavaScripten oinarrituta dago, eta XML dokumentuak eta XMLn oinarritutako markaketa-lengoaiak kudea ditzake. Hizkuntzekiko independentea da, eta erraz sor daitezke etiketatze-atazak. Polonierazko testuak eskuz etiketatzeko hainbat atazatan erabili da. Denbora-informazioa etiketatzeari dagokionez, polonierazko denbora-adierazpenak (Kocoń *et al.*, 2015) eta gertaerak (Marcinčuk *et al.*, 2015) eskuz etiketatzeko erabili da Inforex. Polonierazko denbora-informazioa geruza gehigarri gisa gehitu da KPWr corpusean (Broda *et al.*, 2012), TimeMLn inspiratutako eskema bati jarraiki.

Goian aipatutako etiketatze-tresnak 2.4 taulan laburtuta ageri dira .

2.4 taula – Denbora-informazioa eskuz etiketatzeko tresnak

Etiketatzetresna	Erabilpenak	Eskuz auto	vs	Markaketa-lengoaia	Erlazioen erreprezentazioa	Motorea	Eskuragarritasuna
Alembic Workbench	MUC kongresuak Schilder eta Habelen lana	Eskuz Erregeletan oinarritua bootstrap		SGML	Taula	Sun	Doan Ez (2009)
Callisto	TimeML TIMEX2 Gertaerak	Eskuz		SGML XML AIF	Taula	Java	Kode irekia Doan
Tango	TimeML Frantseseko TimeBank	Eskuz + laguntza		XML (TimeML)	Grafoak	Java (Callisto atazaz)	Doan
Brandeis Annotation Tool	TempEval	Eskuz		TimeML XML	Taula	PHP Javascript	Webean Ez-eskuragarri
Content Annotation Tool	Ita-TimeBank EVENTI TERENCE Excitement NewsReader	Eskuz		XML	Taula		Webean Doan (eskaera eginez)
Inforex	KPWr corpusa Polonierako TimeML	Eskuz		XML		JavaScript	Webean sarrera librea eta harpidedunentzat

2.4 Denbora-informazioa prozesatzeko sistemak

Denbora-informazioa automatikoki tratatzeko sistemak corpusetako informazio linguistikoa eta algoritmoak erabiltzen dituzte testuetako denbora-

¹⁷http://celct.fbk.eu:8080/CAT_WEB_APP/checkUsers

informazioa erazteko. Denbora-informazioa automatikoki prozesatzeko sistema horiek lau taldetan sailka daitezke, haien izaeraren arabera: i) erregeletan oinarritutakoak, ii) estatistikoak, iii) hibridoak, eta iv) ikasketa sakoneko metodoetan oinarritutako esperimentuetarako sortutakoak. Era berean, tresna automatiko horiek denbora-informazioa erazteko prozesuan duten zereginaren arabera ordenatu daitezke. Izan ere, tresna automatikoak hainbat lanetarako erabil daitezke; besteak beste, denbora-adierazpenak eta gertaerak erazteko, denbora-informazioa etiketatzeke eta denbora-erlazioak erazteko.

2.4.1 Erregeletan oinarritutako sistemak

Erregeletan oinarrituta dauden denbora-informazioa erazteko sistema ugari daude. Era horretako sistemak batez ere denbora-informazioa prozesatzen hasi zirenean garatu ziren, etiketatutako datu gutxi zegoelako. Lehenengo esperimentuak denbora-adierazpenak identifikatzeari buruzkoak izan ziren, eta, gero, denbora-adierazpen erlatiboak ebazteari buruzkoak (esaterako, *gaur* edo *bost minutuan*). Manik eta Wilsonek (2000) TempEX garatu zuten; alegia, data, ordu eta denbora-adierazpentzat dituztenak erazteko eta etiketatzeke gai den etiketatzailerak. Lehenik, adierazpen independenteak landu zituen tresnak, eta, gero, testuinguruaren mendekoak. Softwareak arrasto sintaktikoak erabiltzen zituen adierazpen bakoitzari balio bat esleitzeko.

Schilderrek eta Habelek (2001) egoera finituko transduktoreak (FST) erabili zituzten denbora-informazioa erazteko. Lehenik, dokumentuak aurreprozesatu egiten ziren; hau da, sintaktikoki etiketatzen ziren, eta haietako lemak erazten ziren. Ondoren, denbora-adierazpenetan espezializatutako FST transduktoreak sortutako zatiak denbora-adierazpen gisa, gertaera (aditzak) gisa eta gertaerak adierazten dituzten izen gisa etiketatzen zituzten. Balio normalizatu bat ere ematen zitzairen denbora-adierazpenei, indexazio-denboran oinarrituta (dokumentuaren sorrera-data).

TERQAS proiektuan ere erabili ziren FST transduktoreak, gertaerak eta denbora-adierazpenak erazteko; denbora-erlazioak erazteko, aldiz, arau heuristiko gutxi batzuk erabili ziren. Orobat, Alembic Workbench (Day *et al.*, 2004) erabili zen ataza horretan, informazioa automatikoki eta eskuz etiketatzeke aukera ematen duelako.

KTX (Korean Temporal eXpression) etiketatzailerak (Jang *et al.*, 2004) entrenamendu-datuetatik abiatuta egin zen hiztegi batean oinarritutako ikasketa-metodo bat erabili zuen, eta, ondoren, gizakiak sortutako patroiak gehi-

tu zitzaizkion; guztira, 460. Ondoren, denbora-adierazpenak bilatzen ziren, analisi morfologikoa eta hitz-zerrenda bat erabilita. Halaber, KTX etiketa-tzaileak denbora-adierazpen absolutuak eta erlatiboak iraupenetatik desanbiguatzeko erregela-modulu bat zuen.

CTEMP denbora-analizatzaileak (Wu *et al.*, 2005) bi modulu zituen: bata, denbora-informazioa erauzteko; eta bestea, denbora-informazioa normalizatzeko. Taulak erabiltzen zituen analisirako, baita testuingururik gabeko gramatika eta murriztapen-erregelak ere. Analisiak atributu batzuk ematen zizkion identifikatutako denbora-adierazpen bakoitzari, eta, ondoren, atributu horiek denbora-adierazpenak normalizatzeko erabiltzen ziren.

DANTEn ere (Mazur eta Dale, 2007) denbora-adierazpenak erauzteko eta normalizatzeko bi moduluko arkitektura erabili zen. Erregeletan oinarritutako sistema horrek Javan oinarritutako JAPE gramatika bat erabiliz detektatzen zituen denbora-adierazpenak. Gramatika hori 31 hiztegi geografikoz, 80 makroz eta 250 erregelaz osatuta zegoen. Erregelak patroia-ekintza moduko erregela tradizionalak ziren. Erregelaren lehenengo zatiak aurkitu beharreko testu-patroia deskribatzen zuen; bigarrenak, berriz, esleitu beharreko normalizazio-patroia. TIMEX2 eskemari jarraituta identifikatzen eta normalizatzen zituen DANTEk denbora-adierazpenak.

Hona hemen erregeletan oinarritutako beste hiru denbora-analizatzaile: Bittarek (2009), Robaldo *et al.*ek (2011) eta Grover *et al.*ek (2010) sortutakoak. Gertaerak eta horiei buruzko informazioa erauzteko erabili ziren hiru tresna horiek. Lehenengoa frantseserako garatu zen eta gertaerak detektatzeko eta sailkatzeko eskuz egindako hiztegi espezifikoekin konbinatzen zituen erregelak. Bigarrenak, berriz, mendekotasunak eta gertaeren zerrenda bat erabiltzen zituen italierazko gertaerak identifikatzeko eta sailkatzeko. Hirugarren tresnak —Edinburgh-LTG— aurretik etiketatutako corpusetatik erauzitako zerrendak WordNet-eko informazioarekin elkartzen zituen.

HeidelTime (Strötgen eta Gertz, 2013) erregeletan oinarritutako sistema bat da, eta denbora-adierazpenak erauzteko eta normalizatzeko balio du. Erregelak, patroiak eta normalizazio-informazioa hizkuntzaren mende daude, baina iturburu-kodea bera da hizkuntza guztietarako. HeidelTime UIMA gehigarri gisa eraiki da; beraz, UIMA prozesuetan integra daiteke. Denbora-adierazpenetako elementuak identifikatzeko, HeidelTimek tokenak eta definitutako adierazpen erregularretako elementuak (astegunak, hilabeteak, zenbakiak, eguneko zatiak, etab.) elkartzen ditu. Ondoren, erregelen arabera identifikatzen dira denbora-adierazpenak, eta klasea (data, ordua, iraupena eta errepikapena) eta balio normalizatua ematen zaie. TimeML-

ko TIMEX3 etiketa erabiliz etiketatzen dira denbora-adierazpenak. Ingeleserako ez ezik, beste hizkuntza batzuetarako ere erabili da HeidelTime denbora-adierazpenak erauzteko eta normalizatzeko; besteak beste, alemanerako (Strötgen eta Gertz, 2011), nederlanderako (van de Camp eta Christiansen, 2013), frantseserako (Moriceau eta Tannier, 2014) eta kroazierarako (Skukan *et al.*, 2014).

Azkenik, ParsTime (Mansouri *et al.*, 2018) erregeletan oinarrituta dagoen persierazko denbora-adierazpenak erauzteko eta normalizatzeko sistema bat da, eta denbora-informazioa etiketatuta daukaten testuetako data-patroiak detektatzen ditu. Bestela esanda: testuetako tokenei aurretik definitutako 12 etiketetako bat egokitzen zaie (eguna, hilabetea, garaia, zenbakia, etab.), eta patroiei horiek identifikatzeko erregelak sortzen dira. 346 erregela definitu dira, eta horiek TimeML TIMEX3 etiketa bat esleitzen diote denbora-adierazpen bakoitzari, baita klase balio bat (data, ordua, iraupena eta errepikapena) eta balio normalizatu bat ere. ParsTimek egutegi gregorianokoak ez ezik, Hejira eta Jalali egutegietako denbora-adierazpenak ere identifika eta normaliza ditzake.

Erregeletan oinarrituta dauden denbora-informazioa etiketatzeko tresnak 2.5 taulan ageri dira.

2.5 taula – Denbora-informazioa prozesatzeko erregeletan oinarritutako tresnak

Tresna	Garatzaileak	Ataza	Mota
TempEX	Mani eta Wilson (2000)	datak, orduak eta denbora-adierazpenak erauztea eta etiketatzea	Erregeletan oinarrituta (aztarna sintaktikoak)
	Schilder eta Habel (2001)	Denbora-informazioaren erauzketa	Egoera finituetako transduktoreak (FST)
Alembic Workbench	Day <i>et al.</i> (2004)	Gertaerak eta denbora-adierazpenak erauztea	FST
KTX	Jang <i>et al.</i> (2004)	Denbora-informazioaren erauzketa eta normalizazioa	errepikapenezko ikasketa + patroiak + erregelak
CTEMP	Wu <i>et al.</i> (2005)	Erauzketa eta normalizazioa	Erregeletan oinarrituta

DANTE	Mazur eta Dale (2007)	Denbora-adierazpenen erauzketa eta normalizazioa	Erregeletan oinarrituta
	Bittar (2009)	Gertaeren erauzketa eta sailkapena	Erregeletan oinarrituta + eskuz egindako hiztegiak
	Robaldo <i>et al.</i> (2011)	Gertaeren erauzketa eta sailkapena	Erregeletan oinarrituta. Dependentsiak eta gertaera zerrendak
Edinburgh-LTG	Grover <i>et al.</i> (2010)	Gertaeren erauzketa eta sailkapena	Erregeletan oinarrituta. Zerrendak eta WordNet
HeidelTime	Strötgen eta Gertz (2013)	Denbora-adierazpenen erauzketa eta normalizazioa	Erregeletan oinarrituta
ParsTime	Mansouri <i>et al.</i> (2018)	Denbora-adierazpenen erauzketa eta normalizazioa	Erregeletan oinarrituta

2.4.2 Metodo estatistikoetan oinarritutako sistemak

Beste tresna automatiko batzuek metodo estatistikoak aplikatzen dituzte. Bethardek eta Martinek (2006) denbora-informazioa klase anitzetan sailkatzeko metodo bat proposatzen dute. Metodo horrek *inside* (barruan), *outside* (kanpoan) edo *begin* (hasiera) etiketa esleitzen dio token bakoitzari, gertaerak diren ala ez kontuan hartuta. Bestalde, Marchek eta Baldwin (2008) gertaerak erauzteko eta sailkatzeko algoritmo bat proposatzen dute euskarri bektoredun makinetan (*Support Vector Machines*) (Boser *et al.*, 1992) oinarrituta.

TIPSem (Temporal Information Processing based on Semantic information) (Llorens *et al.*, 2010) sistema bat da. Denbora-egiturak identifikatzeko, sailkatzeko, normalizatzeko eta esteka bidez kategorizatzeko erabiltzen da eta informazio semantikoa erabiltzen du lau ataza horiek egiteko. Sistemaren oinarria ausazko baldintza-eremuak (Conditional Random Fields) (Lafferty *et al.*, 2001) delakoa da, ikasketa automatikoko teknika bat.

ClearTK-TimeMLk (Bethard, 2013) Bethardek eta Martinek (2006) egindako lanari jarraitzen dio. ClearTK egituraren oinarrituta dago eta ikas-

keta automatikoaren bidez sailkatzen du informazioa gertaerak, denbora-adierazpenak eta denbora-erlazioak identifikatzeko. Ausazko baldintza-eremuak (CRF), euskarri bektoredun makinak (SVM) eta erregresio logistikoa (LR) aplikatzen ditu.

Kocońek eta Marcińczuk (2017) polonierazko denbora-adierazpenak erazuteko eta normalizatzeko tresna bat garatu dute. Tresna hori Liner2 tresnan (Marcińczuk *et al.*, 2013) oinarrituta dago eta CRF++ tresna multzoa darabil. Hurbilketa zuzena erabiltzen du, ezaugarri linguistiko egokienak hautatzeko.

Murat *et al.*ek (2017) CRFen oinarrituta dagoen uigurrerako denbora-adierazpenak identifikatzeko sistema bat aurkeztu dute. Lehenik, lematizazioa eta hitzen segmentazioa egiten dute, eta, ondoren, bost ezaugarri multzo aplikatzen dituzte sekuentzialki: tokena, gramatika-kategoria, zenbakia (zenbakizko denbora-adierazpenak identifikatzeko), karakterea (karakter bereziak dituzten denbora-adierazpenak identifikatzeko) eta denborazko abiarazle lexikoak.

TOMN (Zhong eta Cambria, 2018) denbora-adierazpenak modelatzen dituen ikasketa automatikoko metodo bat da. CRFen bidez identifikatzen ditu denbora-adierazpenak, eta TOMN etiketak esleitzen dizkie testuetako tokenei. Zehazki, TOMNek denbora-tokenak (T), modifikatzaileak (M) eta zenbakizko egiturak (N) identifikatzen ditu eta *out* (O) etiketa bat esleitzen die gainerako tokenei. Denbora-tokenak, modifikatzaileak eta zenbakizko egiturak Tmn-Regex adierazpen erregularretan biltzen dira, eta aurretik prozesatutako ezaugarriak erazuteko fasean elkartzen dira.

2.6 taulan laburbildu ditugu ikasketa automatikoan oinarrituta dauden denbora-informazioa etiketatze tresnak.

2.6 taula – Denbora-informazioa prozesatzeko ikasketa automatikoko tresnak

Tresna	Garatzaileak	Ataza	Mota
	Bethard eta Martin (2006)	Gertaeren erazketa eta sailkapena	Multi-class classification
	March eta Baldwin (2008)	Gertaeren erazketa eta sailkapena	Support Vector Machines
TIPSem	Llorens <i>et al.</i> (2010)	Erazagupena, sailkapena, normalizazioa eta erlazio-sailkapena	Datuetan oinarritutako sistema. CRF (Conditional Random Fields) modeloak

ClearTK-TimeML Bethard (2013)	(Bethard, 2013)	Denbora-informazioaren erauzketa	Conditional Random Fields, Support Vector Machines eta Logistic Regression
	Kocoń eta Marcińczuk (2017)	Denbora-adierazpenen erauzketa	Conditional Random Fields
	Murat <i>et al.</i> (2017)	Denbora-adierazpenen erauzketa	Conditional Random Fields
TOMN	Zhong eta Cambria (2018)	Denbora-adierazpenen erauzketa	Conditional Random Fields

2.4.3 Sistema hibridoak

Ikus daitekeenez, egokia da sistemak erregeletan oinarritutakoen multzoan eta tresna estatistikoen multzoan banatzea. Hala ere, metodo hibridoak ere garatu izan dira. Verhagen *et al.*ek (2005) GUTime, GUTenLINK eta Evita etiketatzaile partzialak, besteak beste, aurkeztu zituzten TARSQI¹⁸ Toolkit paketearen (Verhagen eta Pustejovsky, 2008) barruan. Etiketatzaile horiek berez etiketatze-ataza ezberdinak egiten dituzten moduluak dira. GUTimek denbora-adierazpenak erauzi eta normalizatzen ditu, Evitak (Events in Text Analyzer) gertaerak adierazten dituzten aditzak, izenak, adjektiboak eta horien ezaugarri gramatikalak erauzten ditu. Aditzak identifikatzeko hiztegietan bilatzen dira, izenak ikasketa automatikoaren bidez erauzten dira eta adjektiboak TimeBanken gertaera bezala agertzen badira baino ez dira gertaeratzat hartzen.

TARSQI sisteman, GUTenLINK, Slinket eta Sputlink garatu ziren erlazioak etiketatzeko. GUTenLINK denbora-erlazioen etiketatzailea da; egoera finituko transduktoreetan (*Finite State Transducers*) oinarritzen da erlazioak sortzeko eta gertaerak erregeletan oinarrituta ordenatzen ditu. Slinket-ek esaldiko aditz nagusiaren eta bere mendekoen arteko erlazioak sortzen ditu ezagutza linguistikoan eta erregeletan oinarrituta. Sputlinkek denbora-erlazio inplizituak tratatzen ditu eta horretarako, Allenen denbora-tarteen algebraren (Allen, 1983) murriztapenak kontuan hartzen ditu.

TRIOSeK (UzZaman eta Allen, 2010) lehen fase batean, gertaerak identifikatzeko erregeletan oinarritutako sistema baliatzen du. Emaitzak hobetzeko,

¹⁸Temporal Awareness and Reasoning Systems for Question Interpretation

Markoven sare logikoa (Markov Logic Network) erabiltzen du sailkapenerako emaitzak hobetzeko.

Language Independent Feature Extractor (LIFE) (Jeong *et al.*, 2015) ere denbora-egiturak eta beren ezaugarriak etiketatzeko erabili da. LIFEk analisia letra-mailan egiten du, token mailan egin ordez, eta Conditional Random Fields (CRF) modeloak erabiltzen ditu. Ikasketa automatikoa erabiltzen bada ere, analisia eskuzko erregela bidez hobe daiteke, denbora-informazioaren analisi osoagoa lortzeko. Sistema orain dela gutxi hobetu zen ExoTime sisteman (Jeong *et al.*, 2017). Horrek kanpoko ezagutza basea erabiltzen du denbora-adierazpenak normalizatzeko.

2.4.4 Ikasketa sakoneko teknikak

Azken boladan, ikasketa sakoneko teknikak erabili dira denbora-informazioaren prozesamenduan. Etcheverry eta Wonseverrek (2017) adierazpide banatuak eta sare neuronal artifizialak erabiltzen dituzte denbora-adierazpenak identifikatzeko. GloVe-ren (Pennington *et al.*, 2014) bidez gaztelaniazko Wikipediatik erauzitako hitzen errepresentazioak inferitu zituzten. Modelizaziorako, kopuru eta tamaina ezberdindun ezkutuko geruzak eta erregularizazio teknikak zituzten aurre-elikatze modeloak eta modelo errepikariak baliatu zituzten eta etiketatu beharreko hitzetik bi aldeetarako lehoaren bektore-errepresentazioarekin lotzen zen uneko testuinguruko informazioa gehitu zuten. Gupta *et al.*ek (2015) antzeko hurbilpena egin zuten ingelesezko testu klinikoetan denbora-adierazpenak identifikatzeko.

Sistema hibridoak eta sare neuronaletan oinarritutako esperimentuak 2.7 taulan laburbildu ditugu.

2.5 Corpusak

HPn gaur egun corpora ordenagailu bidez prozesa daitekeen formatu elektronikoan dagoen eta hizkuntza-ikerketarako eta hizkuntzaren prozesamenduan balia daitekeen testu-gorputz gisa ulertzen da (Setzer, 2001, 64 or.). Hau da, corpora etiketatze linguistikoa duen eta automatikoki irakur daitekeen testu-bilduma da. Corpusak hainbat hizkuntzatarako sortu dira eta testu mota ezberdinak dituzte. Denbora-informaziodun corpusen kasuan, denbora-informazioa adierazten duten elementuak etiketatzen dira eta horien ezaugarriak azaleratzen dira.

2.7 taula – Denbora-informazioaren prozesamendurako tresna hibridoak

Tresna	Garatzaileak	Ataza	Mota
GUTime	Verhagen <i>et al.</i> (2005)	Time expression extraction and normalisation	
GUTenLINK, Slinket and Sputlink	Verhagen <i>et al.</i> (2005)	Temporal and subordination link creation	FST
Evita	Verhagen <i>et al.</i> (2005)	Event recognition and feature extraction	lexical look-up + machine learning
TRIOS	UzZaman eta Allen (2010)	Event extraction and classification	Combined: rules + Markov Logic Network
LIFE and ExoTime	Jeong <i>et al.</i> (2015, 2017)	Temporal information extraction and classification	Conditional Random Fields + hand-written rules
	Etcheverry eta Wonsever (2017)	Time expression extraction and classification	Distributed representations and neural networks

Denbora-informaziodun corpusak nagusiki urre-patroi bezala erabiltzen dira denbora-informazioa prozesatzeko tresnak entrenatzeko eta ebaluatzeko. Beste helburu orokor bat aurretik informazio linguistikoa duten corpusei denbora-informazioa gehitzea da. Horrelakoak dira Katalanezko TimeBank (Saurí eta Badia, 2012), Estonierazko TimeBank (Orasmaa, 2016) eta Chinese Temporal Annotation Corpus (Cheng *et al.*, 2007). Horietan, denbora-informazioa informazio morfosintaktikoari eta lexiko-semantikoari gehitzen zaie eta hainbat mailatan etiketatutako corpora lortzen da.

Atal honetan denbora-informaziodun zenbait corpus aurkeztuko dira, baita horiek sortzeko erabili diren hizkuntzak eta ezaugarriak ere; adibidez, zein motatako dokumentuak dituen corpusak eta zein tresna erabili den etiketatzeko. Denbora-informazioaz etiketatutako lehen corpusak aurkeztuko ditugu 2.5.1 atalean, denbora-informazioadun albiste-corpusak 2.5.2 atalean eta beste domeinuetako corpusak 2.5.3 atalean.

2.5.1 Denbora-informaziodun lehen corpusak

Denbora-informazioa prozesatzeko lehen saiakerak MUC kongresuetan egin ziren. Kongresu horietan informazio-erazketarako hainbat ataza proposatu ziren eta horietan parte hartu zuten tresnak ebaluatzeko corpusak sortu ziren. MUC-6n Wall Street Journaleko 318 albiste zituen etiketatutako corpora zabaldu zen tresnen entrenamendu eta ebaluaziorako (Chinchor eta Sundheim, 2003). Corpus hori eta denbora-informazioaren prozesamenduan erabilitako beste corpus goiztiar batzuk 2.8 taulan laburbildu ditugu.

Hurrengo urteetan hainbat esperimentu egin ziren denbora-informazioaren erauzketan. Mani eta Wilsonek (2000) erregeletan oinarritutako denbora-adierazpenen identifikaziorako algoritmoa garatu zuten eta tresna albiste corpus baten aurka ebaluatu zuten. Corpus horrek TDT-2 bildumako (Fiscus *et al.*, 1999) idatzizko 22 albiste eta telebistako 199 transkribapen zituen.

DARPAk (Defense Advanced Research Projects Agency) sustatutako MUC kongresuen ondotik TIDES proiektua garatu zen. Proiektu horretan “TIMEX2 estandarra” (Ferro *et al.*, 2003) sortu zen denbora-adierazpenak etiketatzeko. Estandar horrek denbora-adierazpenak eta horien atributuak SGML¹⁹ markaketa-lengoiaren bidez etiketatzea ahalbidetu zuen. Horrekin batera, TIDES temporal corpus corpora (Ferro *et al.*, 2002) sortu zen. Corpus horrek denbora-adierazpenak eta denbora-erlazioak zituen eta eskuz

¹⁹Standard Generalized Markup Language

2.8 taula – Denbora-informazioa duten lehen corpusak

Corpusa	Hizkuntza	Konposaketa	Markaketa-lengoaia	Etiketatzetresna	Eskuz vs. auto
MUC-6 (Chinchor eta Sundheim, 2003)	Ingelesa	Wall Street journal-eko 318 artikulua	SGML		Auto
(Mani eta Wilson, 2000)	Ingelesa	TDT-2ko idatzizko 22 artikulua eta 199 transkribapen	TIMEX		Eskuz
Tides Temporal Corpus (Ferro <i>et al.</i> , 2002)	Ingelesa Gaztelania	95 elkarrizketa (gaztelania) + itzulpenak (ingelese) and TDT-2ko 193 dokumentu	TIMEX2	Alembic Workbench	Eskuz
(Setzer, 2001)	Ingelesa	MUC-7ko 6 artikulua	SGML	Perl-en oinarritutako tresna	Eskuz

etiketatu zen Alembic Workbench tresnaren (Day *et al.*, 1997) bidez. Guztira, TIDES corpusak 95 gaztelania-ingelese elkarrizketa-dokumentu eta TDT-2ko 193 dokumentu zituen eta TIDES programan garatutako tresnak ebaluatzeko erabili zen.

Garai berean, Setzerrek (2001) SGML baliatuta denbora-adierazpenak, gertaerak eta denbora-erlazioak etiketatzeko proposamena egin zuen. Setzerrek testuetako denbora-informazioa esplizitu egiteko etiketak eta atributuak aurkeztu zituen eta Perl programazio-lengoaian oinarritutako etiketatze-tresna garatu zuen. Etiketatzegidalerroen egokitasuna ebaluatzeko, etiketatzaileen arteko adostasuna neurtu zuen MUC-7ko 6 dokumenturen gainean. Corpusaren tamaina txikia gorabehera, Setzerren ikerketan hartutako hainbat erabaki berebizikoak izan ziren TimeMLren (Pustejovsky *et al.*, 2003a) garapenerako.

2.5.2 Denbora-informazioa etiketatuta duten albiste corpusak

Azken hamarkadetan, denbora-informazioaren prozesamendua biziki zentratu da albisteetan eta, ondorioz, albiste osatutako hainbat corpusi denbora-

informazioa gehitu zaio. TimeML markaketa-lengoaia erabilienera bihurtu da, baita estandartzat hartu ere. TimeBank 1.1 (Pustejovsky *et al.*, 2003b) eta TimeBank 1.2 (Pustejovsky *et al.*, 2006) corpusak horren bidez etiketatu dira.

TimeBank corpusak TERQAS²⁰ programan sortu ziren galdera-erantzun sistemak entrenatu eta ebaluatzeko. Hala ere, azken urteetan beste gauza batzuetarako ere erabili dira: i) TempEvaleko (Verhagen *et al.*, 2007) hasierako atazetan erabili dira eta ii) TimeBank 1.2 ingelesezko denbora-egituren analisirako eta beste hizkuntzetarako corpusen garapenerako oinarri bilakatu da. Gaztelaniazko bertsioa guztiz parekagarria da, baina beste hizkuntza batzuetan antzerako corpusak sortzeko oinarria ere izan da: frantsesa (Bittar, 2010), italiera (Caselli *et al.*, 2011), errumaniera (Forăscu eta Tufiş, 2012), portugesa (Costa eta Branco, 2012), katalana (Saurí eta Badia, 2012), persiera (Yaghoobzadeh *et al.*, 2012), japoniera (Asahara *et al.*, 2013) eta koreera (Jeong *et al.*, 2015), orain arte.

Etiketatzeko dentsitatea duen TimeBanken bertsioa ere eskuragarri dago: TimeBank-Dense (Cassidy *et al.*, 2014). Corpus horrek TimeBanketik ausaz hautatutako 36 dokumentu ditu. Dokumentu horiek TimeMLren bidez etiketatuta zeuden eta TimeML Dense etiketatze-lengoaian oinarrituta aberastu ziren. Lortutako dokumentuek erlazio-grafo deskribatzaileagoak zituzten.

Frantseseko TimeBank (Bittar, 2010) TimeBank 1.2rekin konparagarria izatea nahi izan zen eta horregatik, TimeMLren bidez etiketatutako 109 albiste dokumentu aukeratu ziren. Datu-kopuru handiagatik etiketatzea erdiautomatikoki egin zen. Etiketatzeko automatikoan TempEx etiketatzailerak (Mani eta Wilson, 2000) erabili zen denbora-adierazpenak etiketatzeko. Ondoren, eskuz gainbegiratu zen etiketatzea Callisto (Day *et al.*, 2004) eta Tango (Verhagen *et al.*, 2006) etiketatze-tresnen bidez.

Italierako TimeBanka (Ita-TimeBank) (Caselli *et al.*, 2011) CELCT corpora²¹ eta ILC corpora (Bindi *et al.*, 1989) batzetik sortu zen. Corpus horiek italierarako TimeMLri (Caselli, 2010) jarraituta etiketatu ziren urre-patroi bihurtzeko helburuz. CELCT corpora Brandeis Annotation Tool (BAT) tresna (Verhagen, 2010) eta CELCT Annotation Tool (CAT) tresna (Lenzi *et al.*, 2012) baliatuz etiketatu zen. ILC corpora, ordea, soilik BAT erabiliz. Bi corpusak eskuz etiketatu ziren.

Errumanierako (Forăscu eta Tufiş, 2012) eta portugesezko TimeBankPT

²⁰Temporal and Event Recognition for QA Systems

²¹<http://www.livememories.org/Home.aspx>

(Costa eta Branco, 2012) TimeBankak ingeleseko bertsioaren itzulpenak dira. Hala eta guztiz ere, bi hizkuntza horietan etiketatzea desberdin egin zen: errumanierako TimeBankean TimeML etiketak tokenen lerrokatzearekin transferitu ziren, eta TimeBankPT eskuz etiketatu zen.

Katalaneko TimeBankak (Saurí eta Badia, 2012) Ancora-Ca corpuseko (Taulé *et al.*, 2008) 210 dokumentu ditu (75.800 tokenetik gora). Automatikoki etiketatu zen TIPSeimen (Llorens *et al.*, 2010) bidez katalaneko TimeMLren bertsioari jarraituz.

Persierako PersTimeBank (Yaghoobzadeh *et al.*, 2012) 43 dokumentuko corpusa da. Dokumentuak Peykareh corpusetik (Bijankhan *et al.*, 2011) ateratu ziren eta erdiautomatikoki etiketatu ziren. Gertaerak Persian Event Tagger (PET) tresnaren (Yaghoobzadeh *et al.*, 2012) bidez etiketatu ziren, eta etiketatze hori Multipurpose Annotation Environment (MAE) tresnaren²² laguntzaz gainbegiratu zen.

BCCWJ-TimeBank corpusa (Asahara *et al.*, 2013) TimeMLri jarraituta etiketatutako japonierako 1,3 milioi hitzeko corpusa da. Corpusa XML Editor oXygen²³ tresnaren bidez eskuz etiketatu zen. Etiketatze-tresna horren bidez XML balidazioa ere egin zen.

TimeBank estiloko corpusak 2.9 taulan laburbildu ditugu.

2.9 taula – TimeBank estiloko denbora-informaziodun albiste-corpusak

Corpusa	Hizkuntza	Edukia	Markaketa-lengoaia	Etiketatzetresna	Eskuz vs auto
TimeBank 1.1, TimeBank 1.2 ETA TimeBank QA CorpusA (Pustejovsky <i>et al.</i> , 2003b, 2006)	Ingelesa	Albiste testuak (183 dokumentu)	TimeML	Alembic Workbench (Day <i>et al.</i> , 1997)	Eskuz
TimeBank-Dense (Cassidy <i>et al.</i> , 2014)	Ingelesa	TimeBank 183 dokumentu	TimeML-Dense		Eskuz

²²<https://code.google.com/p/mae-annotation/>

²³<http://www.oxygenxml.com/>

Frantseseko TimeBank (Bittar, 2010)	Frantsesa	109 albiste	TimeML	TempEx tagger (Mani eta Wilson, 2000), Callisto (Day <i>et al.</i> , 2004) eta Tango (Verhagen <i>et al.</i> , 2005)	Erdiautomatikoa
Ita-TimeBank (Caselli <i>et al.</i> , 2011)	Italiera	CELCT Corpus + ILC Corpus	TimeML		Eskuz
CELCT Corpora ¹	Italiera	I-CAB corpuseko albisteak	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010) eta CELCT Annotation Tool (CAT) (Lenzi <i>et al.</i> , 2012)	Eskuz
ILC Corpus (Bindi <i>et al.</i> , 1989)	Italiera	albisteetako 171 istorio	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010)	Eskuz
Errumanierako TimeBanka (Forăscu eta Tufiş, 2012)	Errumaniera (ingelesetik itzulita)	183 albiste	ISO-TimeML		Auto
TimeBankPT (Costa eta Branco, 2012)	Portugesak (TempEval corpusaren itzulpena)	182 dokumentu	TimeML		Eskuz
Katalaneko TimeBanka (Saurí eta Badia, 2012)	Katalana	Ancora-Ca corpuseko (Taulé <i>et al.</i> , 2008) 210 dokumentu (albisteak eta fikzioa)	TimeML	TIPSem (Llorens <i>et al.</i> , 2010)	Auto

PersTimeBank (Yaghoobzadeh <i>et al.</i> , 2012)	Persiera	Peykareh corpuseko 43 testu (hainbat gaitako albisteak)	ISO-TimeML	Persian Event Tagger (PET) eta Multi-purpose Annotation Environment (MAE) ²	Erdiau-tomatikoa
BCCWJ-TimeBank (Asahara <i>et al.</i> , 2013)	Japoniera	1,3 milioi hitz, hainbat gai	ISO-TimeML	XML Editor oXygen ³	Erdiau-tomatikoa

Corpus horiez gain, txinerarako Chinese Temporal Annotation Corpusa (Cheng *et al.*, 2007) ere sortu da. Corpusak denbora-informazioa du, baina badu desberdintasun bat ere: etiketatze-saiakera dependentzietatik erauzitako informazioan oinarritu zen etiketatzailen lana errazteko. Corpusaren edukia Penn Chinese TreeBank corpusaren (Xue *et al.*, 2005) % 10 da. Corpus hori dependentzia-egituretara moldatu zen eta TimeMLri jarraituta etiketatu zen.

Li *et al.*ek (2014) NISTeko²⁴ MT 2008ko ebaluazio corpuseko 105 dokumentu TimeMLri jarraituta etiketatu zituzten. Corpusak 1.357 txinera-inglese esaldi-bikote ditu. Chinese TimeBank izena eman zioten, beste hizkuntzetako TimeBanekin alderatuta ezberdintasunak baditu ere. Gertaeren, denbora-adierazpenen eta horien arteko erlazioak ditu eta SUTime (Chang eta Manning, 2012) erabiliz etiketatu da eta etiketatzailer batek gainbegiratu ditu etiketatzeak.

Chronolines proiektuan (Bittar *et al.*, 2012) urrats bat haratago joan ziren eta denbora-informazioaren bisualizazioan erabiltzeko frantsese eta ingeleseko albiste corpusa aurkeztu zuten. Etiketatzeko TimeMLn oinarritzen den eta horrekin bateragarria den markaketa-lengoaia bat erabili zuten. Etiketatzeari eskuz egin zen Glozz (Widlöcher eta Mathet, 2009) tresna baliatuta. Emaizta gisa denbora-informazioaren etiketatzerako urre-patroia lortu zuten.

¹<http://www.livememories.org/Home.aspx>

²<https://code.google.com/p/mae-annotation/>

³<http://www.oxygenxml.com/>

²⁴National Institute of Standards and Technology

TIDES proiektua urtetan luzatu zen, eta horretatik ACE TERN corpora (Ferro *et al.*, 2010) sortu zen. Ingeleseko ahozko 127 albistez eta idatzizko 65 albistez osatuta zegoen, eta TIMEX2ren bidez etiketatutako denbora-adierazpenak etiketatuta zituen.

Orasmaak (2014) estonierazko egunkarietako 80 albisteko bilduma (22.000 token) erabili zuen gertaeren etiketatze-modeloak ebaluatzeko. Dokumentuak erdi-automatikoki etiketatu zituzten 3 etiketatzailerik eta epaile batek Brandeis Annotation Tool (BAT) tresna (Verhagen, 2010) erabilia eta urrepatroia sortu zuten. Corpusak informazio morfologikoa eta dependentzia-koak zuen gertaeren arteko erlazioak erauzten laguntzeko. Denbora-informazioaren etiketatzea estonierako TimeMLri jarraituta egin zen.

NewsReader proiektuan (Agerri *et al.*, 2014) WikiNewsetik ateratako 120 albisteko corpora sortu zen: MEANTIME (Minard *et al.*, 2016). Proiektuaren helburu nagusia albisteetatik denbora-lerroak sortzeko sistema garatzea zen. Horregatik, dokumentuek denbora-informazioa etiketatuta zuten denbora-lerroak sortzeko tresnak ebaluatzeko. Corpus horrek bi ezaugarri interesgarri ditu: i) lau hizkuntzatan (ingelese, gaztelania, italiera, nederlandera) etiketa-mailan lerrokatuta dago eta ii) testuen arteko denbora-informazioaren etiketatzea eskaintzen du. Eskuz etiketatu zen CAT (Lenzi *et al.*, 2012) eta CROMER (CROss-document Main Events and entities Recognition) (Girardi *et al.*, 2014) tresnen bidez.

Event StoryLine Corpora (ESC) (Caselli eta Vossen, 2017) ECB+ corpusetik (Cybulska eta Vossen, 2014) eskuratutako hondamendiei buruzko 258 dokumenturen bilduma da. Denbora-informazioa eta kausalitate-erlazioak etiketatuta ditu. Hain zuzen ere, etiketatzeko TimeMLri jarraitu zioten eta kausalitatea-informazioa gehitu zioten. Horiez gain, gertaeren arteko korreferentzia ere jasotzen du. Corpora CAT (Lenzi *et al.*, 2012) erabiliz etiketatu zen.

Denbora-informaziodun corpus horiek 2.10 taulan laburbildu ditugu.

2.10 taula – Denbora-informaziodun albiste-corpusak

Corpusa	Hizkuntza	Edukia	Markaketa-lengoaia	Etiketatzetresna	Eskuz vs. auto
Chinese Temporal Annotation Corpus (Cheng <i>et al.</i> , 2007)	Txinera	Penn Chinese TreeBank corpuseko (Xue <i>et al.</i> , 2005) % 10	TimeML + dependents-ziak		Auto (dependents-ziak) + eskuz
Chinese TimeBank (Li <i>et al.</i> , 2014)	Txinera-Ingelesa	NIST open MT 2008 corpuseko 105 dokumentu. Txinera-ingelesa esaldien 1.357 pare.	TimeML	SUTime (Chang eta Manning, 2012)	Erdiautomatikoa
Chronolines (Bittar <i>et al.</i> , 2012)	Frantsesa	AFPko ⁴ testu idatziak	TimeMLn oinarrituta	Glozz annotation tool (Widlöcher eta Mathet, 2009)	Eskuz
ACE TERN (Ferro <i>et al.</i> , 2010)	Ingelesa	Ahozko 127 albiste eta idatzizko 65 albiste	TIMEX2 ⁵		Eskuz
(Orasmaa, 2014)	Estoniera	80 albiste, 22.000 token	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010)	Erdiautomatikoa
NewsReader (Minard <i>et al.</i> , 2016)	Ingelesa, gaztelania, italiera, nederlandera	120 albiste	TimeMLn oinarrituta	CAT (Lenzi <i>et al.</i> , 2012) eta CROMER (Girardi <i>et al.</i> , 2014)	Eskuz
Event Story-Line (Caselli eta Vossen, 2017)	Ingelesa	258 albiste	TimeML eta kausalitatea	CAT (Lenzi <i>et al.</i> , 2012)	Eskuz

2.5.3 Beste domeinuetako denbora-informaziodun corpusak

Denbora-informazioaren azterketan eta prozesamenduan esperimentu gehienak albisteen domeinuan egin badira ere, azken hamarkadan ikerketa beste domeinu batzuetara zabaldu da. Ondorioz, domeinu berri horietarako corpusak sortu dira. Segidan, hainbat generotako testuez osatutako corpusak deskribatuko ditugu eta 2.11 taulan laburbilduko ditugu.

WikiWars corpora (Mazur eta Dale, 2010) guduak bezalako gertaera historikoei buruzko Wikipediako 22 artikuluz osatuta dago. Autoreek historia-testuetako denbora-erlazioak konplexuagoak zirela demostratu nahi izan zuten, testu luzeagoetan narrazio-estiloan aldaketak izaten baitira. Alemanezko bertsioa ere sortu da: WikiWarsDE (Strötgen eta Gertz, 2011). Corpus horretan ingelesezko WikiWarseko gertaera historikoen alemanezko bertsioa jasotzen da. Bi corpusak TIMEX2 eskemari jarraituta etiketatu ziren eta denbora-adierazpenak baino ez dituzte etiketatuta. Corpusak erdiautomatikoki etiketatu ziren eta eskuz zuzendu ziren. WikiWars etiketatzeko, DAN-TE (Mazur eta Dale, 2007) eta Callisto (Day *et al.*, 2004) erabili ziren eta WikiWarsDE etiketatzeko, HeidelTime (Strötgen eta Gertz, 2013).

Beste hizkuntza batzuetarako ere WikiWarsen bertsioak sortu dira. WikiWarsHr (Skukan *et al.*, 2014) kroazierazko bertsioa da eta Wikipediatik ateratako gudei buruzko 22 artikulua ditu, gehienak WikiWarsekoekin bat datozenak. Ia 60.000 token ditu (puntuazio-markak alde batera utzita) eta 1.440 denbora-adierazpen jasotzen dira bertan. Ezberdintasun handiena corpora TimeMLko TIMEX3 etiketekin etiketatu dela da.

ModeS TimeBank (Guerrero Nieto *et al.*, 2011) beste corpus historiko bat da. Corpus horretan XVIII. mendeko nabigazioari eta informazioaren zabal-kundeari buruzko 102 dokumentu historiko bildu dira. Testuak gaztelania modernoan idatzita daude eta horregatik, denbora-informazioaren etiketatzea egin baino lehen normalizaziorako aurreprozesamendua egin behar izan da. Denbora-informazioa eskuz etiketatu da TimeMLri jarraituta Brandeis Annotation Tool (BAT) tresnaren bidez.

⁴Agence France Presse

⁵Denbora-adierazpenak soilik.

THYME proiektuaren (Styler *et al.*, 2014) barruan eta ISO-TimeML markaketa-lengoaian oinarrituta, historia kliniko elektronikoak batzen dituen urre-patroia sortu da: THYME corpusa. Corpusak Mayo klinikako onkologiako bi esparrutako 1.254 nota ditu. Nota horiek medikuen eta gaixoen arteko elkarrekintzak batzen dituzte eta zenbait ezaugarri berezi dituzte: i) nahiko estandarizatuta daude, ii) nota bakoitzak elkarrekintza bakarra jasotzen du, iii) denbora-tarte batean gertatutako gertaerak jasotzen dituzte eta iv) testuinguruari lotutako adierazpenak, laburtzapenak eta terminologia oso espezializatua dute. Corpusa erdiautomatikoki etiketatu da. 600 dokumentuko lagina Clinical TempEval atazan (Bethard *et al.*, 2015) erabili zen 2015ean.

i2b2 Temporal Relation Corpusa (Sun *et al.*, 2013) ere osasunaren domeinuko denbora-informaziodun corpusa da. Anonimizatutako 310 alta-txosten ditu eta 178.000 token inguru. Eskuz etiketatu zen TimeMLri jarraituta. Etiketatuko THYMEko gidalerroak kontsultatu zituzten, baina bide sinpleagoa hartu zuten eta ez zituzten seinaleak, aspektu-erlazioak eta mendekotasun-erlazioak etiketatu. Zortzi etiketatzailer hartu zuten parte lanean eta Multi-purpose Annotation Environment (MAE) toolkit eta Multi-document Adjudication Interface (Stubbs, 2011) tresnak erabili zituzten.

SIBILA proiektuan (Wonsever *et al.*, 2012) erabilitako corpusak gaztelaniazko albisteak eta dokumentu historikoak batzen ditu. 11.986 token eta 408 esaldi ditu eta horietatik albisteei dagozkienak TempEval2 corpusetik (Verhagen *et al.*, 2007) atera dira. Eskuz etiketatu da eta ikasketa-automatikoko metodoak (euskarri bektoredun makinak eta ausazko baldintza-eremuak) entrenatzeko eta ebaluatzeko erabili da.

Koreerako TimeBanka (Jeong *et al.*, 2015) pertsonen, musikari, unibertsitateari eta historiari buruzko Wikipediako artikuluz eta eskuz sortutako 3.700 esalditik gorako galdera-erantzun bikotez osatutako corpusa da. Eskuz etiketatu da TimeMLri jarraituta, baina markaketa-lengoiari aldaketa batzuk egin zaizkio koreerazko denbora-egituren ezaugarriak hartu ahal izateko. Denbora-egiturak erauzteko tresna automatikoak entrenatzeko erabili da LIFEn (Jeong eta Choi, 2015) oinarritutako O-LIFE ezaugarri erauzlea erabilia.

SoNaR-corpusak (Reynaert *et al.*, 2010) nederlanderazko 500 milioi hitz ditu eta horietatik milioi bat aukeratu dira corpus espezializatuagoa egiteko. Corpus horrek izendun entitateen arteko korreferentzia-erlazioak, rol semantikoak eta espazio- eta denbora-egiturak etiketatuta ditu informazio morfo-sintaktikoz gain. Denbora-informazioa STEx markaketa-lengoaia (Schoor-

man *et al.*, 2010) baliatuta etiketatu zen.

KPWr corpora (Broda *et al.*, 2012) polonierazko ahozko eta idatzizko 1.637 dokumentuko corpora da. Corpusaren helburua polonieraren corpus erakusgarria osatzea da eta, ondorioz, testu formalak eta ez-formalak, zaharrak eta garaikideak eta teknikoak eta orokorrak batzen ditu. Maila morfosintaktikoko eta lexiko-semantikoko etiketatzea du. Denbora-informazioari dagokionez, denbora-adierazpenak, gertaerak eta horien sailkapena (Kocoń eta Marcińczuk, 2017) jasotzen ditu. Denbora-informazioa eskuz etiketatu zen PLIMEX markaketa-lengoiaren (Kocoń *et al.*, 2015) eta Inforex etiketatze-tresnaren (Marcińczuk *et al.*, 2017) bidez.

PHEME proiektuan²⁵ (Derczynski eta Bontcheva, 2014), txutxumutxuak identifikatzeko tresnak entrenatzeko 400 txioko corpora sortu zen. Txioak ingelesez idatzita daude eta tiroketei eta istripuei buruzkoak dira. Corpusak ISO-TimeML etiketatzea du gertaerentzat eta denbora-adierazpenentzat, eta ISO Space (Pustejovsky *et al.*, 2011) lekuentzat eta espazio-entitateentzat. Corpusak 605 gertaera, 122 denbora-adierazpen, 139 espazio-entitate eta 223 leku ditu. Corpora gerora TempEval2ko corpusarekin uztartu zen denbora-entitateak etiketatzeko.

ROCStories corpusak (Mostafazadeh *et al.*, 2016) sen oneko istorioak batzen ditu. Zehazki, Amazon Mechanical Turkeko langileek idatzitako bost esaldiko 40.000 istorio ditu. Gertaerak eta horien arteko denbora-erlazioak eta kausa-erlazioak etiketatu zituzten eskuz adituek TimeMLren bertsio muriztu bat erabilita.

RussianFlu-DE corpora (Canh *et al.*, 2017) ANNO gordailuan²⁶ jasotako Austrian eta Alemanian, 1889-1893 urteen artean izandako gripe-epidemiari buruzko albistez osatuta dago. Corpuseko 639 dokumentuak alemanez daude eta jatorrizko ortografia gordetzen dute, baina normalizatu ere egin dira. Dokumentuak SGML formatura bihurtu dira eta dokumentuen sorrera-data gisa albistearen publikazio-data hartu da. Denbora-adierazpenak erdiautomatikoki etiketatu dira HeidelTimeren bidez eta TIMEX2 etiketak esleitu zaizkie. Guztira, corpusak 453.000 token eta 7.492 denbora-adierazpen ditu.

²⁵<https://www.pheme.eu/>

²⁶<http://anno.onb.ac.at/>

2.11 taula – Denbora-informaziodun beste domeinuetako corpusak

Corpusa	Hizkuntza	Edukia	Markaketa-lengoaia	Etiketatzetresna	Eskuz vs. auto
WikiWars (Mazur eta Dale, 2010)	Ingelesa	Wikipediako 22 testu historiko	TIMEX2	DANTE (Mazur eta Dale, 2007) + Callisto	Erdiau-tomatikoa
WikiWarsDE (Strötgen eta Gertz, 2011)	Alemana	Wikipediako 22 testu historiko	TIMEX2	HeidelTime (Strötgen eta Gertz, 2013)	Erdiau-tomatikoa
WikiWarsHR (Skukan <i>et al.</i> , 2014)	Kroaziera	Wikipediako 22 testu historiko (60.000 token inguru, 1,440 denbora-adierazpen)	TimeML TIMEX3	HeidelTime (Strötgen eta Gertz, 2013)	Erdiau-tomatikoa
ModeS TimeBank (Guerrero Nieto <i>et al.</i> , 2011)	Gaztelania modernoa	XVIII. mendeko 102 dokumentu	TimeML	Brandeis Annotation Tool (BAT)	Eskuz
THYME corpus (Styler <i>et al.</i> , 2014)	Ingelesa	Onkologiako 1.254 nota (Clinical TempEval 2015-eko 600 dokumentu)	ISO-TimeML	Erregeletan oinarritutako sistema	Erdiau-tomatikoa
i2b2 corpora (Sun <i>et al.</i> , 2013)	Ingelesa	310 altatxosten (178.000 token inguru)	TimeML	Multi-purpose Annotation Environment and Multi-document Adjudication Interface (Stubbs, 2011)	Eskuz

SIBILA (Wonsever <i>et al.</i> , 2012)	Gaztelania	Albisteak: 11.986 token eta 408 esaldi (Tem- pEval2 corpu- setik) eta testu historikoak	SIBILA		Eskuz
Korean Time- Bank (Jeong <i>et al.</i> , 2015)	Koreera	Wikipedia orriak eta galdera- erantzun bikoteak (3.700 esaldi)	TimeML		Eskuz
SoNar corpus (Reynaert <i>et al.</i> , 2010)	Nederlandera	milioi 1 hitz	STEx denbora- adierazpe- nentzat (+ espazio- informazioa eta in- formazio linguistikoa)		
KPW corpus (Broda <i>et al.</i> , 2012)	Poloniera	1.637 do- kumentu	PLIMEX (Kocoń <i>et al.</i> , 2015)	Inforex (Mar- cińczuk <i>et al.</i> , 2017)	Eskuz
PHEME tweet corpus (Derczynski eta Bontche- va, 2014)	Ingelesa	400 txio (605 ger- taera, 122 denbora- adierazpen, 139 espazio- entitate eta 223 leku)	ISO- TimeML eta ISO- Space		
ROCStories (Mostafaza- deh <i>et al.</i> , 2016)	Ingelesa	5 esaldiko 40.000 is- torio	Murritzutako TimeML		Eskuz

RussianFlu-DE (Canh <i>et al.</i> , 2017)	Alemana	639 do- kumentu 453.000 token eta 7.492 denbora- adierazpen	TIMEX2	HeidelTime (Strötgen eta Gertz, 2013)	Erdiau- tomatikoa
---	---------	---	--------	---	----------------------

Tresnen garapenak aurrera jarraitzen duenez, corpusak arinago etiketa-
tuko dira eta denbora-informazioaren prozesamenduan orain pentsaezinak
diren helburuak beteko dira. Adibidez, corpusak erabilgarriak dira denbora-
lerroen sorrera automatikoan.

2.6 Denbora-lerroen sorreran egindako beste lan batzuk

Corpusetan batutako denbora-informazio kopuru handiak eta denbora-infor-
mazioa erauzteko tresnak informazio berria sortzen duten tresnetan erabil
daitezke. Denbora-informazioa erabilita sortutako baliabide esanguratsu ba-
tzuk atal honetan deskribatuko ditugu.

Denbora-informazioa egituratua denboraren arabera informazioaren mul-
tzokatzerako erabil daiteke. Zehazki, Alonso *et al.*ek (2009) testuetako den-
bora-adierazpenak testuen denbora-profila sortzeko erabiltzea proposatzen
dute. Profil horiek, denbora-adierazpenak zerrendatzen dituzten denbora-
lerro modukoak, dokumentuak ardatz kronologikoan antolatzeke erabiltzen
dira. Autoreen arabera, informazio-bilaketa atazari denbora-informazioa gehi-
tzean erabiltzaileek beren eskaeratik hurbilago dagoen emaitzak jasoko dituz-
te.

InZeit (Setty *et al.*, 2010) sortzean ere denbora-informazioa kontuan har-
tu zen. InZeit-ek esanguratsuak diren denbora-uneak aukeratuta gai bati
buruzko gertaera-mugarriak identifikatzen laguntzen du. Izan ere, norma-
lean bilaketa-emaitzak adierazgarritasunaren edo publikazio-dataren arabe-
ra ordenatzen dira, baina gaiaren eboluzioa ezin da erraz ikusi. InZeit-ek
bilaketa-emaitzak berrantolatzen ditu testu bakoitzaren denbora-mugarrien
adierazgarritasunaren arabera.

InZeit New York Times corpusean (Sandhaus, 2008) probatu zuten. Corpus horrek New York Times egunkarian 1987-2007 tartean publikatutako 1,8 milioi artikulua batzen ditu. Lucene-n²⁷ oinarritutako bilaketa-interfaze baten bidez, InZeit-ek bilaketaren emaitzak ardatz kronologiko batean erakusten ditu gaiaren eboluzioa ikusi ahal izateko. Gainera, InZeit-ek gertaerarik esanguratsuenak ere nabarmentzen ditu.

Denbora-informazioa baliatu duten beste sistema batzuek informazio hori denbora-lerroak sortzeko baliatu dute. Horrelakoak dira albisteak laburbiltzeko sistemak edota testu historikoetatik denbora-lerroak sortzeko tresnak.

Dokumentuen arteko laburpenak egiteko lehen saiakeretako bat Allan *et al.*ena (2001) izan zen. Beren helburua denboran zehar gertaerek pairatzen zituzten aldaketak ikusten lagunduko zuten metodoak sortzea zen. Beren hipotesiaren arabera, albisteetako gai bat hainbat gertaeraz osatuta zegoen eta gairaren inguruko dokumentuek gertaera horietako batzuk aipatzen zituzten. Gainera, gaiaren barruko gertaeren ordenaren eta dokumentu barruko esaldien ordenaren konbinaketa eginez esaldien ordena absolutua lortzen zen.

Sistemaren eginkizuna esaldi bakoitzari laburpenean zuen garrantziaren araberrako puntuazioa esleitzea zen, kontuan izanda atalasea gainditzeko zuten esaldi guztiak laburpenean agertzen zirela. Gainera, esaldiaren erabilgarritasuna (gaiarekiko lotura) eta gertaeraren informazioaren berritasuna kontuan hartu zituzten. Gaiaren eta gertaeren hizkuntza-modeloen errepresentazioan oinarritutako esperimendazioa egin zuten. Esperimendu horretarako TDT-2 corpuseko tamaina ertaineko 22 dokumentu aukeratu zituzten. Ebaluaziorako corpusean (11 gai) egindako neurketek erakutsi zuten dokumentuko lehen esaldiei puntuazio altuena ematen zieten hizkuntza-modeloei laburpenik onenak lortzen zituztela.

Chieu eta Lee (2004) galderetan oinarritutako bilaketatik lortutako dokumentu-bilduma handi bat laburbiltzen saiatu ziren. Horretarako, gaiari lotutako gertaera esanguratsuak denbora-lerroan irudikatu zituzten. Haien-tzat gertaera esanguratsuak gertaera gertatu zeneko datatik hurbil albistee-tan hainbat aldiz errepikatutako gertaerak ziren. Denbora-lerroak sortzeko, gaiarekin harremana zuten esaldiak identifikatzen zituzten eta esaldi horietako gertaeren datak identifikatzen zituzten. Hori egiteko, esaldian agertzen zen lehen denbora-adierazpena gertaeraren denbora-aingura zela deliberatu zuten eta denbora-adierazpen horiek normalizatzeko erregeletan oinarritutako sistema sinple bat sortu zuten. Ondoren, esaldiak ordenatzen ziren eta

²⁷<https://lucene.apache.org/>

errepikatutakoak ezabatzen ziren, gertaerak beren denbora-ainguren arabera multzokatuz.

Sistema garatzeko, Reuters Volume 1 corpora (Rose *et al.*, 2002) erabili zen eta horretan *earthquake* eta *quake* hitzak bilatu ziren. Sistema ebaluatzeke, automatikoki sortutako denbora-lerroak eskuz sortutakoekin alderatu zituzten. Ebaluazio-corpora English Gigaword corpusetik²⁸ sortu zuten eta pertsona-izenak erabili zituzten bilaketetarako. Ebaluazioan ikusi zen ebaluatzaileek automatikoki sortutako denbora-lerroak komunikabideetan agertzen zen informazioaren adierazgarri zirela eta batzuetan, eskuz egindakoak baino hobeak zirela aldarrikatu zuten.

Evolutionary Timeline Summarization (ETS) (Yan *et al.*, 2011) sistemak bilaketa batean lortutako dokumentu-bilduma bateko data zehatzetako gertaerekin egindako denbora-lerroak sortzen zituen. Zehazki, gertaera horien garapena islatzen zen denbora-lerroetan. Denbora-lerro horiek sortzeko, bilaketa-emaitzetako esaldi-bilduma lortzen zuten eta era berean esaldi horiek beren publikazio-datari lotuta zeuden. Sistemaren emaitza elkarri lotuta dauden laburpenen denbora-lerro ebolutiboa zen. Horiek lortzeko, laburpenetako gertaeren adierazgarritasuna, estaldura, koherentzia eta berritasuna hartzen zen kontuan eta horietatik *erabilgarritasunaren* arabera balioa lortzen zen laburpen bakoitzarentzat.

Sistema ebaluatzeke eskuz sortutako denbora-lerroak sortu zituzten. Hainbat sistema probatu zituzten ETSren egokitasuna ebaluatu ahal izateko. Horretarako, laburpen-sistemak ebaluatzeke ROUGE neurriak (Lin, 2004) baliatu zituzten eta deskubritu zuten ETS zela sistemarik onena bai dokumentu kopuru mugatuan, bai handitzen zen dokumentu-bilduman. Probatu zuten sistemetako bat Chieu eta Leeren (2004) izan zen.

Tran *et al.* (2015) denbora-lerroak sortzeko daten aukeraketan zentratu ziren. Beren saiakera gertaera nagusi baten denbora-tartean gertatzen ziren bigarren mailako gertaeren datak identifikatzean zetzan. Horretarako, testuetako data guztiak hartzen zituen grafo bat sortzen zuten eta daten arteko erlazioak horien maiztasunaren, denborarekiko adierazgarritasunaren eta gaiarekiko adierazgarritasunaren arabera antolatzen zituzten. Ondoren, daten arteko ausazko ibilaldiak egiten zituzten datak ordenatzeko.

Datasetak²⁹ lau gertaerari buruzko 21 denbora-lerro ditu, baita gertaera bakoitzari buruzko albiste corpora ere. Ausazko ibilaldien emaitzak ebalua-

²⁸<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05>

²⁹<http://l3s.de/~gtran/timeline/>

tzean ikusi zutenez, eragina duten faktoreen konbinaketa linealak sistemaren funtzionamendua nabarmen hobetzen zuen. Hain zuzen ere, dataren agerpenen maiztasuna, data beste informazio esanguratsuen ondoan agertzea eta denbora-tarte luzean data aipatzea atazan garrantzitsuak zirela ondorioztatu zuten.

2.7 Laburpena

Atal honetan denbora-informazioaren analisisian eta prozesamenduan egin diren lanik garrantzitsuenak aztertu ditugu. Lehenik, gure lanean baliatuko ditugun kontzeptu garrantzitsuenak identifikatu ditugu eta horiei zein izen eman erabaki dugu. Bigarrenik, denbora-informazioaren prozesamendurako oinarri teorikoak aztertu ditugu. Ondoren, beste hizkuntzetan denbora-informazioaren prozesamendurako sortutako baliabideak, hala nola etiketatze-eskemak, corpusak eta etiketatze tresnak eta denbora-informazioa erauzteko tresna automatikoak, aztertu ditugu.

Aurreko lanen azterketak gure ikerketa lanean zein ildori jarraitu erabakitzen lagundu digu. Denbora-informazioko zein elementuri emango diogun garrantzia erabaki dugu eta elementu horiek TimeML etiketatze-eskemaren euskararako moldaketaren bidez kodetuko ditugula deliberatu dugu. Izan ere, TimeML hizkuntza askotan baliatu den estandarra da eta hori baliatuta euskarazko etiketatzea beste hizkuntzekiko konparagarri egin ahal izango dugu. Halaber, beste hizkuntzetan egin diren corpusekin konparagarria izango den corpusa etiketatuko dugula erabaki dugu eta denbora-informazioa zein baliabide aurreratutan erabili ahal izango dugun ikusi dugu.

EUSKARAZKO ETIKETATZEA ETA CORPUSAREN SORRERA

Euskarazko gertaerak, denbora-adierazpenak eta horien arteko erlazioak

Lan honen sarreran aipatu bezala, denbora-informazioa zer noiz gertatzen den adierazten duen informazioa da. Hain zuzen ere, testuetako gertaerak gertatzen diren ekintzak edo egoerak dira, eta denbora-adierazpenak, gertaera horiek gertatzen diren kronologiako uneak edo iraupenak adierazten dituzten testu-zatiak.

Gertaerak eta denbora-adierazpenak ezin dira batzuk besteak gabe ulertu. Hau da, horien artean sortzen diren erlazioak ezagutu behar dira testuko denbora-informazioa ulertzeko. Izan ere, gertaeren arteko ordena, baita mendekotasuna ere, eta gertaeren kronologiako kokapena ezinbestekoak dira testua ulertzeko.

Atal honetan, denbora-informazioan parte hartzen duten elementuak aztertuko ditugu. Kontuan izan behar da ikerketa-lan honen helburua ez dela analisi linguistiko sakona egitea; beraz, denbora-informazioari dagozkion ezaugarriei ari garenean, ikuspegi konputazionaletik baliagarria izango zaigun informazioaz ari gara. Hain zuzen ere, denbora-informazioaren tratamendurako erabilgarriak diren ezaugarriak eta horien formak baino ez ditugu aztertu.

Gertaerak eta denbora-adierazpenak deskribatuko ditugu 3.1 atalean eta 3.2 atalean, hurrenez hurren; euskaraz zer forma hartzen duten aztertuko dugu eta denbora-informazioa adierazteko esanguratsuak diren beren ezaugarri linguistikoak deskribatuko ditugu. Halaber, horien artean sortzen diren erlazioak aztertu eta deskribatuko ditugu 3.3 atalean.

3.1 Gertaerak

Gertaerak, zentzu zabalean, gertatzen diren ekintzak eta egoerak dira. Tesi-lan honetan, Wonsever *et al.*-ek (2012) proposatu bezala, esango dugu gertaerak direla agente batek borondatez egiten dituen ekintzak, kanpoko indar batek abiatutako prozesuak eta egoerak.

Jarraian, gertaerak euskaraz zein gramatika-kategoriaren bidez adierazten diren eta beren ezaugarrietatik denbora-informazioaren adierazpenean zeresana dutenak deskribatuko ditugu.

3.1.1 Gertaera-adierazpenak euskaraz

Gertaerak ekintzak eta egoerak direnez, pentsatzekoa da nagusiki aditzen bidez adieraziko direla. Hain zuzen ere, Merriam Webster hiztegiak¹ hala definitzen du aditza:

a word that characteristically is the grammatical center of a predicate and expresses an act, occurrence, or mode of being, that in various languages is inflected for agreement with the subject, for tense, for voice, for mood, or for aspect, and that typically has rather full descriptive meaning and characterizing quality but is sometimes nearly devoid of these especially when used as an auxiliary or linking verb.

Hau da, ekintza, gertaera edo izaera adierazten duen predikatu baten buru gramatikala da aditza.

Denbora-informazioa etiketatuta duen EusTimeBank urre-patroi corpusean (4. kapitulua) egindako kontaketek argi uzten dute aditzak direla gertaerak adierazteko gehien erabiltzen direnak. 3.1. taulan ikus daitekeenez, EusTimeBank corpusean, gertaeren ia bi heren aditzen bidez adierazten dira. Hala ere, beste gramatika-kategoria batzuek ere adieraz ditzakete gertaerak: izenak, adjektiboak, adberbioak, izenordainak eta bestelakoak. Esaterako, izenen bidez gertaeren ia laurdena adierazten da eta beste kategorien bidez gertaeren % 11 inguru adierazten da.

Era berean, kontuan hartzekoa da gertaera-adierazpenak egitura konplexuagoetan ere ager daitezkeela eta kategoria desberdinetako bi gertaera-adierazpenek gertaera bakarra adieraz dezaketela. Euskarazko gertaerak kategoriaren arabera nola gauzatzen diren deskribatuko dugu 3.1.1.1–3.1.1.6

¹<https://www.merriam-webster.com/dictionary/verb>

3.1 taula – Gertaerak gramatika-kategoriaren arabera EusTimeBank corpusean

Gramatika-kategoria	Kopurua	%
Aditzak	2.095	64,3
Izenak	791	24,3
Adjektiboak	168	5,2
Adberbioak	159	4,9
Izenordainak	36	1,1
Beste batzuk	6	0,2
GUZTIRA	3.255	100

ataletan eta 3.1.1.7 atalean gertaera-adierazpen konplexuen berri emangu dugu.

3.1.1.1 Aditzak

Ikusi dugunez, EusTimeBank corpusean aditza da gertaerak adierazteko mai-zen erabiltzen den gramatika-kategoria. Zehazki, euskaraz, beste hainbat hizkuntzatan bezala, aditz jokatua (31) izaten da arruntena perpausetan eta, gehienetan, horien bidez adierazten dira gertaerak. Kasu batzuetan, ordea, bai egitura koordinatu batean daudelako (32), bai perpaus ez-jokatuko burua direlako (33–34), aditz ez-jokatuek adierazten dituzte gertaerak.

Atal honetako adibideetan, nabarmendu nahi dugun egitura beltzez adieraziko dugu.

- (31) Montxo Armendarizi **emango diote** Eusko Ikaskuntza saria.
- (32) Pentsamenduak **irakurri** eta mezutara itzultzen dituen sistema bat sortzeko asmoa du Facebook enpresak.
- (33) Ez **jarri** oinak mahaiaren gainean!
- (34) Theresa May lehen ministroari eskatu dio meskiten inguruan segurtasuna **handitzeko**.

Segidan sakonago aztertuko ditugu aditz jokatua zein aditz ez-jokatua.

Aditz jokatua Euskarako aditz jokatuak pertsonarekiko eta numeroarekiko komunztadura eduki eta islatzen dutenak dira. Halaber, denborazko, aspektuzko eta moduzko flexio-markak hartzen dituzte. Esaterako, (35) adibi-

deko *emango diote* aditz jokatuak subjektua hirugarren pertsona pluralekoa, objektu zuzena hirugarren pertsona singularrekoa, eta zeharkako objektua hirugarren pertsona singularrekoa direla adierazten du. Aspektu-denborei dagokienez, etorkizuneko gertaera bat adierazten dute etorkizuneko aspektuak eta orainaldiko denborak. Azkenik, *emango diote* indikatiboko aditza da eta, ondorioz, errealtzat dugun gertaera bati egiten dio erreferentzia.

(35) Montxo Armendarizi **emango diote** Eusko Ikaskuntza saria.

Euskaraz bi motatako aditz jokatuak daude, (35) adibideko *emango diote* aditzaren moduko perifrastikoak eta (36) adibideko *dauka* moduko trinkoak. Aditz perifrastikoak karga semantikoa duen aditz nagusi batez (*emango*) eta informazio gramatikala gehitzen duen aditz laguntzaile batez (*diote*) osatzen dira. Aditz trinkoak, ordea, hitz bakarrekoak dira eta forma bakar horrek hartzen du informazio guztia. Hala, (36) adibideko *dauka* aditzak hainbat motatako informazioa adierazten du: i) *eduki* aditzaren informazio lexikoa, ii) subjektua eta objektu zuzena hirugarren pertsona singularra direla, eta iii) orainaldian gertatzen ari den ekintza dela.

(36) Kooperatibak zail **dauka** negozioarekin jarraitzea.

Ikus daitekeenez, aditz jokatuak informazio asko hartzen dute beren gain. Ez soilik gertaeraren semantika, parte-hartzaileei buruzko informazioa eta gertaera noiz gertatu edo gertatuko den (gertatu edo gertatuko bada) ere.

Aditz ez-jokatua aldi edo pertsona markarik ez daukana da. Alegia, ez dute parte-hartzaileei edo denborari, aspektuari eta moduari buruzko informaziorik ematen. Oso testuinguru bakanetan agertzen dira ikusiko dugunez. Adibidez, arrunta da aditz ez-jokatua aginduzko perpausaren buru izatea (37).

(37) **Ekarri** arropa zikinak hona!

Hala ere, maizago ikusten dira aditz ez-jokatuak jokatutako aditz baten mende, perifrastiko buruaren mende daudelako zein mendeko perpausaren buru direlako. Forma ez-jokatua den partizipioa, esaterako, mendeko perpausetako aditz nagusi gisa (38) edota aditz modal baten osagarri gisa (39) ager daiteke.

(38) Ezer **erabaki** aurretik, gobernuak zer gertatuko den jakin nahi du.

- (39) Kopuru zehatzak datorren astean **jakinarazi** behar ditu multinazionalak.

Aditz ez-jokatuak laguntzailerik hartzen ez badu ere, postposizioak har ditzake. Esaterako, partizipioari instrumentala (-z) gehituz gero, gaztelaniazko gerundioaren parekoa lortzen da testuinguru batzuetan. (40) adibideko *alderatuz* gaztelaniazko “comparando” gerundioaren ordaintzat har daiteke eta (41) adibideko *baliatuz* aditzaren gaztelaniazko itzulpena “aprovechando” izan daiteke testuinguru horretan.

- (40) Orain kontsumitzen dutenarekin **alderatuz**, % 30 murrizketa izango dutela espero da.
- (41) Soldatapekoek barne-agiri bat kaleratu dute, ekarpenen auzia **baliatuz**.

Gertaeraren muga kokatzen dira *partizipioa+-z* formako aditzak, zaila baita horiek denboran ainguratzea. Hizkuntza erromanikoetarako moldatutako TimeML eskemak aztertu ostean, gertaeratzat hartzea erabaki dugu berez ekintza bat (*alderaketa* (40) edo *baliatzea* (41)) adierazten baitute. Frantseserako (Bittar, 2010) eta italierarako (Caselli, 2010) erabaki bera hartu da, eta gaztelaniarako (Saurí *et al.*, 2009), etiketatze-gidalerroetan adibiderik agertzen ez bada ere, ez da gerundioak gertaeratzat ez hartzeko arrazoirik adierazten.

3.1.1.2 Izenak

Izenak, arruntak (42–43) eta nagusiki nominalizazioak (44–45), gertaerak adierazteko erabiltzen dituzte hitzunez.

- (42) Ekarpunen **auzia** baliatuz taldeari **kalte** egin nahi izatea leporatu diote.
- (43) Lasserre-David, Lassalle eta Bru dira Ipar Euskal Herriko hiru **diputatu** berriak.
- (44) Txinako Gobernua kexu da AEBek enpresa baten **erosketa** trabatu dutelako.
- (45) Pobreziaren atzean den **edertasuna** dute lanek haritzat.

Sakonago aztertuko ditugu horiek hurrengo lerroetan.

Izen arruntak Badira gertaerak adierazten dituzten izen arruntak. Horiek ekintza bat adierazten dute normalean (*eraso* (46) eta *otoitz* (47)).

- (46) Erresuma Batuko Musulmanen Kontseiluak islamofobiari egotzi dio **erasoa**.
- (47) Oinezkoak meskitatik atera berriak ziren gauerdian, ramadam garai-ko **otoitzaren** ondoren.

Aditz kopulatiboekin atributu posizioan agertzen diren izen batzuk ere gertaeratzat hartu ditugu. Posizio horretan, normalean entitateak adierazteko erabiltzen diren izenak agertzen dira, zehazki, lanpostu edo ardura bati erreferentzia egin ohi diotenak. Lanpostu edo ardura hori egoeratzat ulertuz gero, gertaeratzat hartu behar dira izen horiek. (48) adibidean, *hiru diputatu berriak* predikatua *dira* aditz kopulatiboarekin agertzen da. *Diputatu izatea* subjektuaren funtzio edo egoera denez, gertaeratzat hartu dugu *Ipar Euskal Herriko hiru diputatu berriak* predikatua.

- (48) Lasserre-David, Lassalle eta Bru dira Ipar Euskal Herriko hiru **diputatu** berriak.

Predikatu inkoatiboekin ere antzekoa gertatzen da. *Bihurtu* eta *bilakatu* moduko predikatuarekin agertzen diren izenak ere gertaeratzat hartu ditugu ondoko (49) adibideko *zuzendari bihurtu* bezalako kasuetan, esaterako. Horrelakoak dira *moja sartu*, *artzain joan* eta *lehendakari izendaturen* gisakoak. Horietan guztietan, predikatu inkoatiboek subjektuaren rola betetzen duen entitatearen egoera berria abiarazten dute.

- (49) Agurtzane eskolako **zuzendari** bihurtu da, beste hautagaiak atzera egin ostean.

Magnitude-izenak (*tenperatura*, *luzera*, *biztanleria*, etab.) gertaeratzat hartzea erabaki dugu egoera adierazten duten kasuetan. Hala, (50) adibideko *biztanleria* gertaeratzat hartuko dugu, denboran koka daitekeen Euskal Herriaren egoera bat adierazten baitu. Magnitude izenak gertaeratzat hartzea TimeML definitzean hartutako erabakia da eta guk euskarara ekarri dugu.

- (50) Euskal Herriak hiru milioiko **biztanleria** du.

Izen eratorriak eta nominalizazioak Euskaraz izen eratorriak edo nominalizazioak sortzeko modu ugari dago; bai izen eratorriak hainbat gramatika-kategorietatik sor daitezkeelako, bai ugariak direlako eratorpen-atzizkiak. Ondoko adibideetan ikus daitekeenez, (51) adibideko *erosi* aditzetik, (52) adibideko *erosketa* eratorri da aditz-oina eta *-keta* eratorpen-atzizkia uztartuz. Era berean, (53) adibideko *eder* adjektibotik (54) adibideko *edertasun* sortu da, adjektiboari *-tasun* eratorpen-atzizkia gehitzean.

- (51) Txinako Gobernua kexu da ezin izan duelako enpresa bat **erosi** AEBren trabagatik.
- (52) Txinako Gobernua kexu da AEBek enpresa baten **erosketa** traba dutelako.
- (53) Pobreziaren atzeko hori **ederra** da eta lanek haritzat hartzen dute.
- (54) Pobreziaren atzean den **edertasuna** dute lanek haritzat.

Hala, nominalizazioari esker, beste kasu batzuetan aditzen eta adjektiboen bidez adierazten diren gertaerak izenez adieraz daitezke. Izan ere, “nominalizazioa aldaketa da, kategoria-aldaketa; eta haren ondorioz, hitz edo elementu lexiko bat (dela aditza, dela izenondoa) izen bihurtzen da” (Garcia, 2001, 101 or.). Ondoko adibideetan, *erosi* aditzaren (55) eta *eder* adjektiboaren (56) nominalizazioak irudikatu ditugu. Nominalizazio horiek *-te/-tze* atzizkiaren bidez egin dira.

- (55) Txinako Gobernuak ez du enpresa bat **erosterik** izan AEBren trabagatik.
- (56) Pobreziaren atzean den **edertze** prozesua dute lanek haritzat.

Ikusi dugunez, *-keta*, *-tasun* eta *-te/-tze* atzikiaren bidez sor daitezke izenak, besteak beste. SGen², “egoera, gertaera, ondorioa adierazten duten izenak sortzen dituzten atzizkiak” (3.2 taula) agertzen dira.

Taulan ikus daitekeenez, alde batetik, nominalizazioak sortzeko *-te/-tze* atzizkia dago. Beste alde batetik, euskarazko jatorria duten *-men*, *-keta* eta *-tasun* (57) gisako atzizkiak daude, baita latindar jatorriko *-zio* (58) eta *-mendu* bezalakoak ere. Halaber, nominalizazioa ez da beti atzizki bidez egiten; atzizkirik gabeko nominalizazioa (zero atzizkia) ere badago (59). Kasu horretan, jatorrizko gramatika-kategoriako formak eta izenak forma bera dute.

²<http://www.ehu.eus/seg/morf/4/5>

3.2 taula – SEGeko egoera-, gertaera-, ondorio-izenak sortzeko atzizkiak

Egoera, gertaera, ondorioa adierazten duten izenak sor- tzen dituzten atzizkiak	-men/-pen, -mendu, -tze/-te, -dura, -era, -zio, -keta, -ginan, -ginean, - kuntza, -ada, -antza, -antzia, -ada, zero atzizkia
--	---

- (57) Kudeaketarako **zailtasunak** negatiboki eragin du azken urteetan.
 (58) Jakinarazi du zorra berregituratzeko **negoziazioei** ekin diela
 (59) 1990. urteko isurien aldean, % 4,1 urritu dira berotegi gasen **isuriak**.

Azkaratek (2008b) ere nominalizazioa eta eratorpena aztertu ditu eta aditzetik gertaera-izenak sortzeko ondoko atzizkiak identifikatzen ditu (3.3 taula).

3.3 taula – Azkarate Villar (2008b) laneko nominalizazio-atzizkiak

Egoera, gertaera, ondorioa adierazten duten izenak sor- tzen dituzten atzizkiak	-tze/-te, -era, -keta, -men/-pen, - kuntza, -kunde, -dura, -tza
--	--

3.2 eta 3.3 taulak alderatuz gero, ondoriozta daiteke Azkaratek euskal jatorridun atzizkiei baino ez diela erreparatu. Hala ere, bi iturrietan atzizki horien bidez sortutako izenak gertaerak adierazten dituztela argi uzten da.

Azkaratek beste lan batean (Azkarate Villar, 2008a) nabarmentzen duenez, gertaera-izen batzuk adierazteko *-aldi* atzizkia erabil daiteke, *hegaldi* (60) moduko gertaera-izenak sortzeko. Atzizki horrek, izatez denbora-tartea adierazten badu ere, hiztunek denbora-tarte bat irauten duten gertaerak adierazteko ere baliatzen dute.

- (60) Lehen **hegaldia** ongintzako **hegaldi** berezi bat izango da.

Atzizki horien bidez sortutako izenak ez dute beti gertaera-izaerarik, ordea. Azkaratek nabarmentzen duenez, euskaraz gertaerak adierazten dituzten izen batzuek gertaera horien ondorioak ere adieraz ditzakete. Zehazki, batzuek gertaera irakurketa soilik onartuko dute eta beste batzuek gertaera eta emaitza irakurketak. Esaterako, (61) adibidean *gonbidapena* ekintza da, (62) adibidean, ordea, *gonbidatzearen* ekintzaren ondorio edo produktua. Gertaera irakurketa dutenak gertaeratzat hartuko ditugu; ez, ordea, gertaeraren ondorioa adierazten dutenak.

- (61) Jaurilaritzak dio ETAk egin dion mintzatzeko **gonbidapena** ez dela serioa.
- (62) **Gonbidapena** jasota, Bilbo Basketek baieztatu du Eurokopen ariko dela.

Gertaeren prozesamenduan, gertaeren eta gertaeren ondorioen arteko bereizketa berebizikoa da, gertaerak zuzen identifikatzea baita urratsik garrantzitsuenetako bat. Horregatik gertaeren eta gertaera-ondorioen arteko desberdintasunak aztertu ditugu.

Pentsatzekoa da gertaera adierazten duen aditz batetik eratorritako izenak ere gertaera bera adieraziko duela. Hori gertatzeko, ordea, izenak aditzaren argumentu-egitura heredatu behar du. Gainera, gertaera- eta emaitza-irakurketa onartzen duten izen horiek testuinguruaren arabera baino ezin dira bereizi. Adibidez, (63–64) adibideetako *erosketak* gertaera-irakurketa du lehenengoan eta emaitza edo produktuarena bigarren. Argumentu-egiturari dagokionez, (63) adibideko *erosketak* gaia (*enpresa bat*) agerian du. (64) adibidean, ordea, zailagoa da *erosiren* argumentu-egituraren isla ikustea.

- (63) Txinako Gobernuak kexu da AEBek enpresa baten **erosketa** tratatu dutelako.
- (64) Mesedez, lagundu **erosketa** etxera igotzen.

Azkaratek morfologian oinarrituta, atzizkien araberako analisia egiten du. Ondoko eratorpen atzizkiez osatutako izenak zein argumentu har dezaketen deskribatzen du:

- Aditz-oina+TZE motakoak berez dira “gertaera-izenak” aditzetik zuzenean hartuak. San Martinek (2009) ere, *-tze* motako nominalizazioak aztertzen ditu eta aditzetik hurbilen daudenak direla adierazten du. Izan ere, ergatibodun, absolutibodun eta datibodun argumentu-egitura arrunta onartzen dute, baita beste adjuntu batzuk ere.
- Aditz-oina+ERA motakoen artean, ordea, bereizketa dago NOR motako aditzetatik eratorritako izenen eta beste aditzetatik eratorritako izenen artean. Lehenek gertaerak adierazten dituzte (*hasiera*, *igoera*, *joanera*, *sarrera*) eta bigarrenek “gertaera-modua” (*janzkera*, *siniskera*, *idazkera*, *erabilera*, *joskera*, *egoera*, *ibilera*, *izaera*). Dena den, kontuan izan behar da bereizketa horrek baduela hainbat ñabardura, adibidez,

ur-sarrera ez da ura sartzearen ekintza, baizik eta ura sartzen den to-
kia. *Egoera* izenarekin ere, gure hasierako hipotesia zalantzan jartzen
da guk gertaeratzat hartu baititugu “gertatzen diren ekintzak edo egoe-
rak”.

- Aditz-oina+KETA eta Aditz-oina+PEN motakoez argumentu-egiturari
buruzko informazioa baino ez du ematen.
- Aditz-oina+KUNTZA/KUNDE/DURA motakoen kasuan, ordea, ger-
taera-izaeraren zalantzazkotasuna nabarmentzen du, horien argumentu-
egitura argia bada ere.
- Aditz-oina+TZA motakoak eta Aditz-oina+TZE motakoak kontrajar-
tzen ditu eta lehenek ekintza soilaz gaindiko esanahia dutela dio eta
eskaintze-eskaintza bikotea jartzen du adibide gisa.

Azkaratek (2008b) planteatzen duen gertaeren eta gertaeren ondorioen
arteko banaketa hori askotan aztertu da literaturan. Azkuek (1923) gertaera
adierazten dutenei “derivativos operativos” deritze eta multzo horretan sar-
tzen ditu *-era* eta *-keta* atzizkiaz sortutako izenak. Emaizaren berri ematen
dutenei, ordea, “derivativos nominales efectos” deritze eta *-ada* eta *-dura* har-
tzen duten izenak multzo horretan sartzen ditu. Azkenik, *-kuntza* eta *-men*
izen abstraktuak sortzeko baliatzen direla dio.

Lafittek (1944) nominalizazioak sortzeko atzizkiak bi multzotan banatzen
ditu: ekintza edo gertaera adierazten dutenak eta ondorioa adierazten dute-
nak. Lehen multzokoak dira *-keta*, *-kunde*, *-kuntza*, *-men/-mendu/-pen*, *-te*,
-tza eta *-tze*. Bigarrenekoak *-men/-mendu/-pen* eta *-dura* dira.

Azkenik, Hualde eta Ortíz de Urbinak (2003) batera ematen dituzte ekin-
tza eta ondorioa adierazten dituzten izenak, beren esanetan banaketa ez bai-
ta zorrotza. Hala ere, *-pen*, *-keta* eta *-mendu* atzizkiek gehienbat ekintza
adierazten dutela, eta *-dura* ondorioa adierazteko erabili ohi dela diote.

Autore horiek mota horietako izen eratorriak gertaeratzat (G) edo ger-
taeraren ondoriotzat (O) hartzen dituzten 3.4 taulan laburbildu dugu.

Forma berak gertaera zein gertaeraren ondorioa (edo produktua) adie-
raz ditzakeenez, gertaerak identifikatzeko prozedurak bilatu behar izan ditu-
gu. Grimshawk (1990) eta Artiagotiak (2000), gertaera-egitura duten izenak
eta “ondorio” izenak bereizten dituzte. Artiagoitiaren lanean gertaera-izenen
ezaugarri hauek deskribatzen dira:

- gertaera-izenek ez dute plurala onartzen.

3.4 taula – Gertaera-adierazpen (G)/ondorio-adierazpen (O) sailkapena

	-te/-tze	-era	-keta	-men/-pen/ (-mendu)	-kuntza	-kunde	-dura	-tza	-ada
Azkue (1923)		G	G				O		O
Lafitte (1944)	G		G	G/O	G	G	O	G	
Hualde (2003)			G(/O)	G(/O)			(G/)O		
Azkarate (2008b)	G	G ez beti	G	G				G/O	

- gertaera-izenek nekez onartzen dute erakuslea.
- gertaera-izenak ez dira predikatu gisa agertzen.
- gertaera-izenek, iraupena adierazten duten modifikatzaileak erraz onartzen dituzte, ondorio izenak ez.
- gertaera-izenek nekez onartzen dituzte adjektiboak.
- gertaera-izenek ez dute *bat* zenbatzailea onartzen.

Azkarateren (2008b) arabera, ordea, zenbatzailea har dezakete eta horren adibidetzat hartzen du *sorgin-bilerak* izen elkartua. Ondorioz, ezbaian jartzen du Artiagoitiak proposatutako gertaera-izenen lehen ezaugarria.

Denbora-informazioaren prozesamenduan gertaerak identifikatzeak duen garrantziagatik, atazari lotuta dauden azterketak egin dira. Esaterako, TimeMLn (TimeML Working Group, 2010) izen baten bidez adierazitakoa gertaera den erabakitzeke *testa* proposatzen da:

- IZENAK minutuak/egunak/urteak... iraun zuen.

(65) [Galletak] bi ordu iraun zuen.

(66) [Uholdeak] bi ordu iraun zuen.

- IZENA oso azkar/bat-batean gertatu zen.

(67) * [Galleta] oso azkar gertatu zen.

(68) [Uholdea] oso azkar gertatu zen.

- IZENA *denbora-adierazpenean* gertatuko da.
 - (69) * [Galleta] etzi gertatuko da.
 - (70) [Uholdea] etzi gertatuko da.
- IZENA/IZENAK *denbora-adierazpenean* hasi/jarraitu/bukatu zen/zuen.
 - (71) * [Galleta] atzo hasi zen.
 - (72) [Uholdea] atzo hasi zen.

Goiko (67–72) adibideetan ikus daiteke *galleta* hitza testuinguru baten baino ez dela egokia, eta, ondorioz, ez da gertaeratzat hartuko automatikoki. *Uholde* hitza, ordea, egokia da lau testuinguruetan eta, bi baino gehiagotan (lauretan) egokia denez, gertaera izango da.

Guk *test* hori aplikatu dugu etiketatze-lanetan.

3.1.1.3 Adjektiboak

Adjektiboa izenaren inguruan ageri den perpausaren osagai bat da. Izenaren kualitate bat adierazten du edota izenaren ezaugarri bat adierazteko erabiltzen da. Adjektiboek kualitate edo ezaugarri bat adierazten dutenez, egoera motako gertaerak adierazten dituzte. Adibidez, (73) adibideko *beldurgarriak* adierazten du gertatutakoak beldurra eragiten duela.

- (73) Gaineratu du **beldurgarria** dela gertatutakoa.

Adjektibo batzuk izenen ondoan agertzen dira (74). Beste batzuk, aldiz, predikatuak dira eta aditz kopulatiboarekin agertzen dira atributu posizioan (75) (Zabala, 1999).

- (74) Erabaki **traumatikoak** hartu beharko ditu.
- (75) Ziurgabetasuna eta beldurra dira **nagusi**.

Adjektiboek entitate baten egoera adierazten badute ere, gure lanean ez ditugu guztiak gertaeratzat hartuko. Izenaren adjuntu diren adjektiboak, hala nola (74) adibideko *traumatikoak*, alde batera utzi ditugu zaila baita batzuetan horien gertaera-izaera identifikatzea. (75) adibideko *nagusi* modukoei, ordea, gertaera-trataera eman diegu. Aditz kopulatiboaren atributu gisa agertzen diren adjektiboak (75) predikazioaren parte direnez, errazagoa

da beren gertaera esanahia antzematea. Subjektua den entitatearen egoera bat adierazten dute.

Adjektiboen inguruan prozesamendurako hartu dugun erabaki bati erreparatu behar zaio. Euskararen prozesamendurako sortu diren baliabideetan adjektibo kategoria esleitu zaie beste lan batzuetan izenondotzat eta izenlaguntzat hartu direnei. Esaterako Altuna *et al.* (1985) lanean hala adierazten da. Hala ere, sailkapen hori eztabaidatua izan da eta hainbat lanetan (Goenaga, 1991) izenlagun eta adjektibo/izenondo bereizketa hobesten da.

Denbora-egiturak prozesatzeko beste hizkuntza-maila batzuen prozesamendutik ateratako analisiak hartzen ditugunez, lan honetan ere, izenlagun zein izenondoak adjektibo kategoriaren barruan batuko ditugu. Izan ere, denbora-informazioaren tratamendu automatikoa egitean, analisi morfosintaxikotik lortutako informazioa erabiltzen dugu.

3.1.1.4 Adberbioak

Adberbioek ere gertaerak adieraz ditzakete. Kasu batzuetan, egoera bat adierazten dute, (76) adibideko *haserre* eta (77) adibideko *beteta* bezala. Beste kasu batzuetan, ekintza bat adierazten dute, esaterako, *korrika* (*egin*) (78).

(76) Langileak **haserre** daude.

(77) Eskari-poltsak **beteta** daude.

(78) Astean behin **korrika** egiten dutenek hobeto egiten dute lo.

Predikatu-osagarriaren funtzioa hartzen duten adberbioek subjektu edo objektu posizioan dauden entitateen egoera bat adierazten dute. Adibidez, (77) adibideko *beteta* adberbioak eskari-poltsen egoera adierazten du. Predikatu-osagarriaren funtzioa betetzen duten adberbioen eta aditz kopulatioarekin batera agertzen diren eta atributu funtzioa betetzen duten adjektiboen funtzionamendua antzekoa dela esan dezakegu, biek adierazten baitute perpauseko entitate baten egoera.

3.1.1.5 Izenordainak

Sareko Euskal Gramatikan³ eta EGLUn (Altuna *et al.*, 1985) izenordainek definizio bertsua jasotzen dute: izenak ordezkatzeko ditzakeen hitza. Gainera,

³<http://www.ehu.eus/seg/morf/5/4>

biak bat datoz izenordainak izen berezitzat jo behar direla esaten dutenean, ez baitute mugatzailerik onartzen eta berek bakarrik osa baitezakete izen sin-tagma. Guk, ordea, izenetatik aparte aztertuko ditugu gertaerak adierazteko duten berezitasunagatik. Izan ere, izenordainek, berez, ez dute gertaerarik adierazten, aurretik aipatutako gertaera bati erreferentzia egiten diote, bigar-ren aipamen horretan gertaera elidituta badago ere. Hau da, korreferentzia-erlazio batean izenordain baten aurrekariak gertaera bat adierazten badu, izenordainak ere gertaera horri egingo dio erreferentzia. Adibidez (79) adibi-deko *hori* izenordainak “legebiltzarrak Segurtasun sailari bi hizkuntzak era-biltzeko eskubidea bermatzea” adierazten du, hau da, eskatzearen gertaerari egiten dio erreferentzia.

- (79) ...legebiltzarrak Segurtasun Sailari eskatzea berma dezala bi hizkun-tzak erabiltzeko eskubidea. Ez zuen **hori** erabaki mahaiak.

Gertaerak adierazten dituzten izenordainek orokorrean erakusleen forma hartzen dute, erakusleak determinatzaileak badira ere (Aduriz *et al.*, 2005) eta horregatik sailkatu ditugu multzo honetan. EusTimeBank corpusean (ikus 4. kapitulua) *horren* (80) eta *horrek* (81) formak aurkitu ditugu bes-teak beste. Adibideetan ikus daitekeenez, erakusle horiek (beltzez) kurtsibaz markatutako gertaera bati egiten diote erreferentzia.

- (80) Turkiako gobernuak Turkiako tropak Irakera bidaltzeko neurri bat *onartu zuen* urriaren 18an. **Horren** ondorioz, petrolioaren prezioak inoiz baino gehiago igo ziren.
- (81) Hegazkin europar berriak erregistratu baino lehen Indonesiara bidali *beharrak* kezkatzen du Djama, **horrek** kostu handiagoak eragingo bailituzke.

3.1.1.6 Bestelakoak

Multzo honetan sartu ditugu aurreko kategorietan kokatzeko arazoak dituz-ten gertaera-adierazpenak. 3.1 taulan ikus daiteke eskuzko etiketatzean, 6 kasutan baino ez dela gertaera bat bestelakoen multzoan sailkatu. Ondo-ko (82–85) adibideetan, multzo honetan sailkatu diren gertaera-adierazpen “mota” bakoitzeko adibideak agertzen dira.

- (82) Itxiko diren kontzesionarioak enpresaren kontzesionarioen % **18** dira, baina diru-sarreraren % **7** baino ez dira.

(83) Senatuko botoen kontaketa **52-35** izan zen.

(84) Ekoizlearen serie-zenbakia **001** da.

(85) **Ezetz** esan zuten.

(82–85) adibideetan ikus daitekeenez, zenbakizko egiturak dira multzo honetan sailkatu diren seitek lau. (82) adibideko % 18 eta % 7 izenordaintzat har bazitezkeen ere, (83–84) adibideetako kasuak sailkatzen zailagoak dira. Bestalde, (85) adibideko *ezetz* adberbioa da, baina perpaus osoa ordezkatzeko du.

3.1.1.7 Gertaera-adierazpen konplexuetako kategoria-konbinazioak

Aurreko 3.1.1.1–3.1.1.6 ataletan ikusi dugunez, gertaera-adierazpenak ez dira beti isolatuta agertzen; beste gertaera-adierazpen batzuekin batera ager daitezke. Esaterako, aditzak jorratzean, aditz ez-jokatuz adierazitako gertaera aditz modalez adierazitako gertaera batekin ager daitekeela ikusi dugu.

Hori, ordea, ez da gertaera-konbinazio bakarra. Gertaerak sarri agertzen dira egitura handiagoetan konbinatuta. Atal honetan, gramatika kategoriazberdinetako gertaera-adierazpenen konbinazioak deskribatzeko ditugu horien konposizionaltasunari erreparatuz.

Hitz Anitzeko Unitate Lexikoak (HAUL) maiz agertzen dira testuetan. Euskaraz, HAUL horietako batzuk, ziurrak edota maiztasun handikoak, Euskararen Datu-Base Lexikalean (EDBL) (Aldezabal *et al.*, 2001) batzen dira. EPEC corpusean (Euskararen Prozesamendurako Erreferentzia Corpora) (Aduriz *et al.*, 2006) ere HAULei buruzko informazioa adierazten da eta, corpus horri erreparatuta ikusi dugu *so egin* (86), *uste izan* (87) edo *behar izan* (88) modukoak bi gertaera-adierazpenez osatutako HAULak direla (Urizar, 2012).

(86) Lantzean behin gortinen atzetik kalera **so egiten genuen**.

(87) Mundu zibilizatuko jendeak **uste du** selbako erregea lehoia dela.

(88) Askok kostatzen da neurria hartzea, eta kalkulu ugari egin **behar izaten ditugu**.

HAUL horiek hiru multzotan sailka daitezke Corpasen (1996) eta Urizarren (2012) arabera: kolokazioak, lokuzioak eta enuntziatu fraseologikoak. Normalean esaldi osoak diren eta fosilizatuta dauden enuntziatu fraseologikoak alde batera utzita, kolokazioak eta lokuzioak bereizteko bi ezaugarri

nagusi erabiltzen dira: konposizionaltasun semantikoa eta murriztapen morfosintaktikoak. Kolokazioak erlazio sintaktikoren bat duten bi unitate lexikok osatzen dituzte (kolokazioaren oinarriak eta kokakideak) eta kolokazioaren oinarriak kokakidearen adiera berezi bat aukeratzen du. (89) adibideko *zarata atera* kolokazioan *zarata* oinarriak *atera* kokakidea aukeratzen du eta *zarata sartu* modukoak ekiditen ditu.

- (89) Metroaren lanek **zarata** handia *ateratzen dute*.

Era berean, kolokazioak morfologikoki erregularrak dira eta nominalizazioa eta elementuen txertaketa bezalako bariazioak onartzen dituzte. Hala, posible da *zarata atera* kolokaziotik, (89) adibidean bezala, *zarata handia atera* modukoak sortzea.

Lokuzioen kasuan, idiomatikotasuna da ezaugarri nagusia; hau da, semantikoki ez-konposizionalak dira eta nekez uler daiteke esanahi osoa osagai bakoitzaren esanahitik abiatuta. Horrelakoak dira *adarra jo* modukoak. Beste batzuk murriztapen sintaktikoek edota irregulartasun morfologikoek egiten dituzte lokuzio. Bigarren mota horretakoak dira *aurre egin* (90), *ahal izan* (91) eta *irrist egin* (92).

- (90) Peronismoak ezin izan dio ofizialismoari **aurre egin**.
(91) Medikuek 68 urtera atzeratu **ahal izango dute** erretiro adina
(92) Bi zezenek **irrist egin**, eta laguntza behar izan dute altxatzeko.

Formari erreparatuz, (90–92) adibideetan informazio lexikala duen elementu bati informazio gramatikala ematen dion beste elementu bat, aditza, gehitzen zaio. Egitura horiei aditz arineko egitura izena eman ohi zaie. Grossek (1981) adierazten duenez, aditz arina duten predikatu konplexuetan gertaera bakarra adierazten da. Horrelakoak dira *irrist egin* (92), baina baita *min hartu* (93) edo *musu eman* (94) modukoak ere (Urizar, 2012).

- (93) Bertak **min hartu du** buruan adar batekin.
(94) Ona da lagunek elkarri **musu ematea**.

Azken horiei emango diegu trataerarik sakonena bi gertaera-adierazpen dituztelako.

Amaitzeko, aipamen berezia egin behar diegu *adarra jo* moduko lokuzioei. Horiek ez dira bi gertaera-adierazpenez osatzen, *adarra* izena ez baita

gertaera-izena. Hala ere, horrelako lokuzioek tratamendu berezia behar dute, parte hartzen duten elementuek ez baitute beti beren oinarrizko esanahia mantentzen (Iñurrieta *et al.*, 2016). Esaterako, *adarra joko*, *txistua joko* eta *laguna joko jo* gertaerek esanahi desberdina dute testuinguruaren arabera. Gertaeren sailkapena egitean esanahia kontuan izan behar da, forma berak ez baitu beti kategoria bereko gertaera adierazten.

3.1.2 Gertaeren ezaugarriak

Denbora-informazioaren prozesamenduan baliagarriak izango diren gertaeren ezaugarriak identifikatu ditugu. Ezaugarri horietako gehienak semantikaren ingurukoak dira; zehazki gertaeren eta denboraren arteko erlazioa adierazten dute askok. Gertaera guztiak ez dira berdin gauzatzen denboran eta, gertaera-klasearen, aditzen tempusaren eta aspektuaren, modalitatearen eta polaritatearen analisiaren bidez, gertaera mota bakoitza kronologian nola txertatzen den azaleratu nahi dugu. Izan ere, gertaera batzuk puntu batean gertatzen dira eta beste batzuk denboran luzatzen dira, batzuk burututa daude eta beste batzuk gertatzeke daude, adibidez. Horien berri emango dugu atal honetan.

3.1.2.1 Gertaera-klasea

Ez da erraza gertaerak nola sailkatu erabakitzea; atazaren arabera sailkapen-irizpideak aldatzen dira. 2.2.1 azpiatalean ikusi bezala, hainbat sailkapen proposatu dira gertaerak antolatzeke gertaeren zenbait ezaugarri kontuan hartuta. Guk TimeMLn proposatutako sailkapena aukeratu dugu. Aukera hori egin dugu, alde batetik, TimeML denbora-informazioaren prozesamenduan estandarra bihurtu delako eta, sailkapen bera mantenduz, gure etiketatzea beste hizkuntzetakoekin konparatu ahal izango dugulako, eta, bestetik, TimeMLko gertaeren sailkapena denbora-informazioa islatzeko egokia iruditu zaigulako. Jarraian, gertaerak sailkatzeko baliatu ditugun zazpi kategoriak deskribatuko ditugu. Eskaintzen ditugun definizioak TimeML Working Group (2010) lanekoak dira.

- **Diskurtsiboak.** Pertsona batek edo erakunde batek zerbait aldarrikatzen duela, gertaera bat kontatzen duela edota hartzaileei informazioa ematen diela adierazten duten gertaerak, adibidez: *esan*, *adierazi*, *aipatu*, *iragarri*, *gogoratu*, *adierazpen*, *konfirmazio*...

- (95) Endesak **iragarri du** Garoña berriro zabaltzeko baimena eskatuko duela.
- (96) Rousseffek **gogoratu du** gaztea zenean ia hiru urte eman zituela preso.
- (97) Tenientea generalaren **konfirmazioaren** zain dago eraso hasteko.
- **Hautematezkoak.** Gertaera bat fisikoki antzematea adierazten duten gertaerak, adibidez: *ikusi, entzun...*

(98) Eztandak **entzun dira** geltoki barruan.
 - **Aspektuzkoak.** Aspektuzko predikazioa gertaera baten denbora-egituran zentratzen da, hau da, gertaeraren hasiera, jarraipena edo amaieran. Predikatu horiek gertaera bat hartzen dute argumentutzat eta bere denbora-egituraren aspekturen bati egiten dio erreferentzia. Horien artean ditugu:
 - Hasierazkoak: *hasi, ekin, hasiera...*
 - Berrasierazkoak: *berrabiarazi...*
 - Amaierazkoak: *gelditu, bertan behera utzi, geratu, geldialdi...*
 - Burututasunezkoak: *burutu, amaitu, amaiera...*
 - Jarraipenezkoak: *jarraitu, eutsi, jarraipena...*
- (99) Omenaldian izan zen Nafarroako Parlamentuko presidentea, eta irainka eta mehatxuka **hasi zitzaion** talde bat.
- (100) Asteburuko biharamuna gainditu ostean, jaien **amaiera** bertan da.
- **Ekintza intentsionalak** (*intensional actions*). Argumentutzat testuan esplizituki agertzen den gertaera bat hartzen duten ekintzak dira. Argumentua den gertaeraren eta ekintza intentsionalaren erlaziotik hainbat gauza inferi daitezke, esaterako, i) argumentua den gertaera hori ez da zertan gertatu ekintza intentsionala gertatzen den unean (Saurí *et al.*, 2006) eta ii) ekintza intentsionalak argumentu duen gertaeraren gauzatzeari buruzko informazioa ematen du. (101) adibidean ekintza intentsionala beltzez eta bere argumentua dena kurtsibaz adierazi dira.

(101) Joane Getariako txakolina *edatea* **proposatu du**.

- **Egoera intentsionalak** (*intensional states*). Ekintza intentsionalek bezala, gertaera bat hartzen dute argumentu gisa eta egoera horiek babilzko errealitateak edota errealitate alternatiboak ahalbidetzen dituzte. (102) adibidean ekintza intentsionala beltzez eta bere argumentua dena kurtsibaz adierazi dira.

(102) Orain arteko terminal guztiak **prest daude** A380a *hartzeko*.

- **Egoerak** (zirkunstantziak). Zerbaitek egia balioa lortu edo mantentzen duelako zirkunstantziak adierazten dituzte egoerek. Egoera intentsionalak ez diren beste egoera guztiak sartzen dira multzo honetan.

(103) Eguerdian, % 19,43koa **izan da** parte hartze datua.

- **Jazoerak**. Gertatzen diren ekintzak orokorrean. Aurreko multzoetan sailkatu ezin direnak mota honetakoak dira.

(104) Bart **sutea** piztu zen lantegi baten inguruan.

(105) Bart Athleticek **irabazi zuen**.

(106) Itxasok aitortu du berak eta Elorzak **kendu zutela** Igeldoko AHTaren kontrako pankarta.

3.1.2.2 Tempusa

Aditzen tempusa, “aditz aldia” (Coterón Agorria, 2015) edo aditz-denbora euskaraz aditz jokatuak adierazten duten ezaugarria da, eta aditzak adierazten duen ekintza denboran kokatzen du, noiz gertatu den zehaztuz (Zubiri eta Zubiri, 1995). Reichenbachek (1947) esaten duen bezala, tempusek denbora zehazten dute hizketa-ekintzaren denbora-unearekiko, bestela esan, tempusak gertaerak esate-unearen arabera antolatzen ditu.

Aditz jokatuaren tempusa aditz trinko edo sintetikoetan (107) forman bertan adierazten den gramatika-ezaugarria da; aditz perifrastiko edo analitikoetan (108), ordea, aditz laguntzailean adierazten da. Denbora objektiboa adierazteko balio du.

(107) Hobekuntza **dator** astebururako.

(108) Euskararen Legea aldatzea onartu **du** Nafarroako Parlamentuak.

Aditzak hartzen duen tempusa erabakitzeke, hizketa-unea hartu behar da kontuan; igorleak esaldia sortzen duen una, hain zuzen ere. Une horretatik atzera gertatutakoa lehenaldikoa izango da eta tempusaren lehenaldiko formaren bidez adieraziko da aditz jokatueta; une horretan gertatzen dena edo etorkizunean gertatuko dena orainaldiarekin adierazten da, ordea.

Euskarazko tempusak lehenaldikoa (109), lehenaldikoa ez dena (110) eta balizko denboran dagoena (111) bereizten ditu. Hala batzen du Sareko Euskal Gramatikak: “‘Aldi nagusiak’ hiru dira: orainaldia, iraganaldia eta alegialdia”⁴.

(109) Bart sutea **piztu zen** lantegi baten inguruan.

(110) Hobekuntza **dator** astebururako.

(111) Euskal artea munduratzeko sarea osatuko **litzateke**, akordioa gauzatu-
tuko **balitz**.

Alegiazko tempusa ez da ez orainaldia ez iragana. Alegiazko tempusa duten aditzek errealitatetik kanpo dauden gertaerak adierazten dituzte. EGLU II-n (Altuna *et al.*, 1987) adierazten denez, alegiazko formak ezin koka daitezke denboran eta, ondorioz, ezin sailka daitezke orainaldian edo lehenaldian. (111) adibideko *osatuko litzateke* eta *gauzatuko balitz* aditzek denbora hipotetikoari egiten diote erreferentzia.

Aditzek eskaintzen duten denbora-informazio osoa lortzeko, tempusa aspektuarekin konbinatu behar da. Ondoko hiru adibideetan (112–114) orainaldiko forma duten hiru aditz ikus ditzakegu, baina ezin esan daiteke hirurak denboran berdin kokatzen direnik. (112) adibidean ohiko ekintza bati buruz hitz egiten dugu, denboran zehar luzatu eta orainaldian ere gertatzen dena. (113) adibidean, ordea, hizketa-unea baino lehenago burutu den gertaera bati buruz ari gara. Azkenik, (114) adibidean, hizketa-unearen ondoren kokatzen da gertaera.

(112) Nerabeekin eta haien familiekin lan egiten **du** Mendigutxiak.

(113) Euskararen Legea aldatzea onartu **du** Nafarroako Parlamentuak.

(114) Amara Berriko krosaren laugarren edizioa egingo **dute** 19:00etan.

Ondorioz, atal honetan deskribatzen ari garen tempusa eta 3.1.2.3 ataleko aspektua kontuan hartu behar izango ditugu aditz jokatz adierazitako gertaerak aztertzean.

⁴<http://www.ehu.eus/seg/morf/5/15/2/2/1>

3.1.2.3 Aspektua

Aipatu bezala, aspektuak gertaera bat denboran nola gauzatzen den adierazten du. Hain zuzen ere, aditzak markatzen duen ekintza amaituta dagoen ala ez adierazten du, eta hiru aspektu motaz hitz egiten da: burutua, burutugabea eta geroa. Euskaraz aspektuaren informazioa aditz nagusian agertzen da (Altuna *et al.*, 1987).

Aspektu burutuak ekintza bukatua, perfektua adierazten du. Partizipioaren bidez adierazten da; aditz perifrastikoetan aditz nagusiak partizipioaren forma hartzen badu, aditz horrek aspektu burutua duela esaten dugu. Esaterako, (115) adibideko *entregatu zaio* aditza *entregatu* partizipioa eta orainaldiko *zaio* aditz laguntzailearen bidez osatuta dago eta entregatzea dagoeneko egin dela adierazten du.

- (115) Airbus-en lehen A380 superjumboa gaur **entregatu zaio** Singapur Airlinesi.

Aspektu burutugabea ekintza burutugabea, ez-perfektua adierazteko baliatzen da. Euskaraz aditz trinkoen eta aditzoina+*-ten/tzen (ari izan)* forma duten perifrastikoen bidez adierazten da gertaeraren ez-burututasuna. Aditz trinkoek puntualtasuna, unean gertatzen ari dena, adierazten dute eta perifrastikoek, ohikotasuna. Aditz guztiek, ordea, ez dute aditz trinko paradigmatik eta kasu horietan puntualtasuna aditz perifrastikoen bidez adierazten da. (116) adibideko *dirau* forma *iraun* aditzaren hirugarren pertsona singularrereko orainaldiko aditz trinkoa da eta esate-unean indarrean dagoen gertaera bat adierazten du. (117) adibideko *jarraitzen dute* aditza *jarrai* aditzoina eta *-tzen* aspektu ez-burutuaren marka duen aditz nagusiz osatutako aditz perifrastikoa da. Horrek ere esate-unean indarrean dagoen gertaera baten berri ematen du. Adibideetan antzeman daitekeenez, azpimarratutako gertaerak gertatzen ari dira, baina ezin esan noiz hasi edo bukatuko diren.

- (116) Txapelak buru berean **dirau**.

- (117) Indarrean **jarraitzen dute** aire-linea indonesiarrarentzako zigorrek.

Geroa, semantikoki aldia bada ere, euskaraz aspektuaren bidez adierazten da. Geroak ekintza gertakizun dela adierazten du, baina ezin daiteke aldiak kokatu, geroaren *-ko* eta *-ren* aspektu markak aise konbina baitaitezke orainaldiko eta lehenaldiko aditz laguntzaileekin. Hala adierazten dira erreferentzia-unearen ondoren gertatzen diren gertaerak euskaraz. Esaterako, (118) adibideko *zabalduko ditu* aditz perifrastikoan *zabald* aditzari *-ko*

etorkizuneko marka gehitu zaio eta *ditu* orainaldiko aditz laguntzailearekin agertzen da.

- (118) Hirugarren terminalak datorren urteko urtarrilean **zabalduko ditu** ateak.

Aspektu-markarik gabeko aditzak ere badaude. Halakotzat hartzen dira, esaterako, aditzoinak. Aspektu-markarik gabeko aditzak euskaraz denboran kokatzen zailak diren gertaerak adierazteko baliatzen dira, hala nola aginterazko formak (119) eta subjunktiboko formak (120). Hain zuzen ere, aditz aspektudunek gertaera gertagarriago bat adierazten dute (Altuna *et al.*, 1987, 429 or.), agintera eta subjunktiboko formek ez bezala. (120) adibideko *izan dezan* aditz perifrastikoa subjunktibokoa da eta errealitatetik kanpoko gertaera bat adierazten du. *Izan dezan* aditza *izan* aditzoinaren bidez eta subjunktiboko orainaldiko *dezan* laguntzaileaz osatuta dago.

- (119) **Etor** bedi zure erreinua.

- (120) Espetxeratzea bertan behera uzteko galdegin die Auzitegi Gorenari eta Auzitegi Konstituzionalari, Generalitateko presidente gisa jarduteko aukera **izan** dezan.

3.1.2.4 Modalitatea

Modalitatea beharrarekin eta posibilitatearekin lotzen den informazio linguistikoa da. (Zubeldia, 2010, 22 or.) arabera, modalitateak “beharrezkotasunarekin eta posibilitatearekin du zerikusia tradizioko logika modalean, proposizioaren egiarekin (edo faltsutasunarekin) dago lotuta; eta modalitate aletikoa esaten zaio”. Bittarek (2010), ordea, modalitatea hiztunaren ikuspuntu edo iritziaren adierazpentzat hartzen du modalitate epistemikoaren ildoari jarraituta. Hala eta guztiz ere, Bittarek ere beharraren eta posibilitatearen nozioak modalitatean sartzen ditu.

Guk modalitatearen ikuspegi tradizionalari helduko diogu, gramatiketan horrela tratatu izan baita, eta hori euskaraz nola gauzatzen den aztertuko dugu. Euskaltzaindiaren arabera (Eguskitza, 1985), hiru dira euskaraz modalitatea adierazteko bideak:

- modalitate-partikulak: *gogo*, *omen*, *ote*, etab.
- *ahal/ezin izan* aditz modalak

- *nahi izan* eta *behar izan* aditz semimodalak

Guk ahalerazko *-kedun* aditz laguntzailea ere modalitate adierazletzat hartzea erabaki dugu, Altuna *et al.*-en (1987) adierazten denez, *ahal izan* forma perifrastikoaren baliokidea baita.

Ikerketa-lan honetan aditz eta izen modalek adierazten duten modalitatea baino ez dugu kontuan hartuko; zehazki, *-kedun* formak eta *ahal/ezin (izan)* (121), *behar (izan)* (122) eta *nahi/gogo/gura/gei (izan)* (123) izenek edo aditz-multzoek adierazten dutena.

- (121) Sindikatuek laguntza eskatu dute ahal bezain laster lanean hasi **ahal izateko**.
- (122) Ordurako epaileak salmentari baietza edo ezetza eman **behar izango dio**.
- (123) Aljeriarrek egozketa adarra eta etxe-tresna txikiena erosi **nahi dute**.

3.1.2.5 Polaritatea

Polaritateak gertaera baiezkia edo ezezka agertzen den adierazten du. Semantikaren aldetik, ezeztapena bere eraginpean dagoen proposizioaren egia-balioa aldatzen duen hizkuntzako fenomeno da (Martí *et al.*, 2016). Ondorioz, gertaera bat ezezka agertzen bada, horren egia-balioa aldatuko da.

Ezeztapenaren eraginaz ari garenean, bi elementu mota definitu izan dira ezeztapenaren azterketan: ezeztapenaren irismena eta fokoa. Irismena ezeztapen-markek eragiten dioten testu-zatia da. Ezeztapenaren fokoa ezeztapenaren irismenean argien edo indar handienez ezeztatutako elementua da (Huddleston eta Pullum, 2002).

Euskaraz, gertaerarekin batera ezeztapen-markarik ez badago (*ez, ezin, gabe, etab.*) (Altuna *et al.*, 2017b), gertaerak polaritate positiboa duela esango dugu. Aitzitik, gertaera ezeztapen-markaren eraginpean badago, ezezka agertzen dela eta polaritate negatiboa duela esango dugu.

Hurrengo (124–125) adibideetan polaritatearen eta ezeztapenaren arteko eragina zehatzago adierazten dugu. Hasteko, (124) adibidean ezeztapen-markarik ez duen perpausa dago. Bertan *erosi* eta *nahi dute* gertaerak polaritate positiboz agertzen direla esango dugu. (125) adibidean, aldiz, *ez* ezeztapen-marka ageri da. Horrek bere eraginpean dauden gertaeren polaritatea aldatuko du (Altuna *et al.*, 2017c); hala, (125) adibidean *erosteko* da ezezka agertzen den gertaera eta horri bakarrik eragiten dio ezeztapenak.

Eskatu die, ordea, baiezkata dago ezeztapenaren eragina mendeko perpausari mugatzen baitzaio.

- (124) Aljeriarrek egozketa adarra eta etxe-tresna txikiena **erosi nahi dute**.
(125) Garraio-ministroak hegazkin europarrak *ez* **erosteko** eskatu die airelineei.

Polaritateak eta ezeztapenak garrantzi handia dute gertaeren prozesamenduan, gertaeren faktualitatea (3.1.3.2 azpiatala) deskribatzean ikusiko dugunez. Gure hipotesiaren arabera, gertaera bat ezeztapenaren irismenaren barruan badago, bere faktualitatea alda daiteke. Esaterako, (126) adibidean baiezkata ematen den (polaritate positiboa duen) eta faktualitate positiboa (gertatu dela edo gertatuko delako balioa) duen *erosteko* gertaera (127) adibidean ezeztapenaren irismenaren barruan agertzen da eta ondorioz, bere polaritatea eta faktualitate balioa aldatu egin dira.

- (126) Garraio-ministroak hegazkin europarrak **erosteko** eskatu die airelineei.
(127) Garraio-ministroak hegazkin europarrak *ez* **erosteko** eskatu die airelineei.

3.1.3 Gertaeren ezaugarri osagarriak

Denbora-informazioaren prozesamenduan erabilgarria izan daitekeen informazio osagarria tratatzeko beharra ikusi dugu ikerketak aurrera egitean. Hasiera batean, denbora-informazioa aztertzean, gertaeren eta kronologiaren arteko erlazioa baino ez dugu aztertu. Azterketan aurrera egin ahala, ordea, testuko gertaerek errealtatean gertatu diren gertaerei erreferentzia egiten dieten aztertu nahi izan dugu.

Azterketan sakontzean, denbora-informazioaren prozesamendurako informazio linguistiko gehigarria azaleratu behar zela ikusi genuen. Informazio horren berri emango dugu atal honetan. Lehenik eta behin, igorleak aipatzen duen gertaerarekiko konpromiso edo ziurtasun maila aztertuko dugu (3.1.3.1 atala). Horrekin lotuta, gertaeren faktualitatea aztertuko dugu; gertaerak errealtatean gertatu diren edo ez (3.1.3.2 atala). Informazio hori beharrezkoa da errealtatearen narratiboki deskribatzeko tresnetan, hala nola denbora-lerroak sortzekoetan edo gertaera bati buruzko albisteen laburpean. Azkenik 3.1.3.3 atalean ziurtasunari eta faktualitateari eragin diezaizketen kasu bereziak aztertuko ditugu.

3.1.3.1 Ziurtasuna

Ziurtasuna testuetako informazio subjektibotzat eta esplizituki adierazten den modalitate epistemikoko formatzat (Rubin *et al.*, 2006) hartzen dugu. Ziurtasunak, beraz, igorleak adierazitako mezuaren egiazkotasunarekiko duen konpromisoa adierazten du. Zubeldiak (2010), *ebidentzialitatearekin* lotzen du ziurtasuna eta esaten du ebidentzialitateak ezagutzari buruzko jarrreraren ezein hizkuntz adierazpen biltzen duela. Hau da, ebidentzialitatea ziurtasunaren adierazpentzat du.

Gertaerak errealitatean gertatu direlako informazioa nola adierazten den ikertu dugunez, igorlearen gertaerarekiko konpromisoa edo ziurtasuna euskaraz nola adierazten den aztertu dugu. Arau orokor bezala, lehen urrats honetan, gertaerak ziurtzat hartuko ditugu (128), baldin eta ez badago ziurgabetasun-marka espliziturik edo ez bada ezinezkoa ziurtasun balio bat ematea. *Ote* ziurgabetasun partikulak (129) *galdu* gertaeraren ziurtasunarekiko igorlearen konpromiso falta adierazten du.

(128) Boeing-ek 11.000 milioi dolar **lortu ditu** akordioetan.

(129) Hegazkin-merkatuaren kontrola *galdu* **ote** zuen eztabaida piztu zen.

Ote (129) edo *omen* eta *ei* bezalako partikula modalak ziurtasun eza adierazteko modu bat dira, esanahi hori testuinguruari eta pragmatikari lotuta badago ere (Zubeldia, 2010). Era berean, badira beste baliabide lexiko batzuk ere. Baliabide lexiko horien artean, gertaera batzuek (*ez jakin*, *zalantza izan*) esplizituki adierazten dute hizlariaren ziurgabetasuna gertaeran horien mendeko den gertaerari dagokionez. Adibidez, (130) adibidean balearen iraupenarekiko ziurgabetasuna eta zalantza ageri dira.

(130) 'Euskal balea' ere esaten zaion espezie horrek apenas duen hazkunderik, eta *zalantzan jarria* da **haren iraupena**.

3.1.3.2 Faktualitatea

Sauríren (2008) arabera, faktualitatea testuan aipatutako gertaeren izaera faktuala adierazten duen informazio-maila da; hau da, testuan agertzen diren gertaerak errealitateko egoera errealekin edo ziurrak ez diren egoerekin bat datozen edo ez aztarna emango dute. Ondorioz, testuan agertzen diren gertaera guztiak ez dira gertatu edo gertatuko, eta horren arabera

faktualitate-balio bat hartuko dute, esaterako, gertaera faktuala edo kontra-faktuala. Saurik lan berriago batean (Sauri *et al.*, 2009) iturriaren konpromisoa ere gehitzen dio eta diskurtsoan aipatutako gertaeren izaera faktualarekiko iturri esanguratsuek duten konpromisoa adierazten duen informazio-maila dela esaten du. Bigarren definizio horrek agerian uzten du faktualitatea ebazteko, faktualitateaz gaindiko ikuspegia behar dela.

Faktualitatea ebazteko, denbora-informazioa baliatu behar da behinik behin. Lehenik, aditz jokatuaren aspektuak eta tempusak gertaera dagoeneko gertatu den edo ez adieraziko dute. Informazio hori, ordea, ez da kontuan izan behar den bakarra. Bigarrenik, esaldiaren polaritatea, gertaera baiezkak (131) edo ezezka (*ez*) (132) agertzen den, baldintzazko egitura batean agertzen diren (133) eta modalitate markarik hartzen duten (134) ere kontuan izateko ezaugarri linguistikoak dira. Azkenik, igorlearen gertaeren faktualitatearekiko ziurtasuna ere kontuan hartzekoa da faktualitatea erabakitzeke orduan; igorleak gertatu delako zalantzarik badu (135), kasu honetan *ez daki*, ezin izango baitzaio faktualitate balio ziur bat eskaini.

- (131) Orain arteko terminal guztiak **prest daude** A380a hartzeko.
- (132) *Ez* **zuen** 12 orduko hegaldirik **egin**.
- (133) 600 milioi euro **ordainduko ditu** blokeoak 2019ra arte jarraitzen badu.
- (134) Beren habitatean **ikus daitezke** animaliak safarian.
- (135) Poliziak *ez daki* oraindik lapurrek banku zentraleko atea **bortxatu zuten**.

Hala, faktualitateari dagokionez, gertaerak gertatu direla (131), iraganean ez direla gertatu (132), oraindik gertatzeke daudela (133) edo ezinezkoa dela gertatu edo gertatuko diren jakitea (134–135) esango dugu.

3.1.3.3 Faktualitatearen kasu bereziak

Testuinguru batzuetan denbora-informazioa eta ezeztapena ez dira nahikoak gertaeren faktualitatea ebazteko. Guk Minard *et al.*-en (2014) ereduari jarraituz, baldintzazko egituretan eta perpaus generikoetan gertaeren faktualitate-balioa aztertu nahi izan dugu kasu berezitzat. Izan ere, testuinguru berezi horietan—bereziki aditz jokatzat adierazitako gertaerekin—gertaerak ez du espero zitekeen faktualitate-balioa hartzen.

Adibidez, *etorri izan ez banintz* baldintzazko perpausean (136), *etorri* gertaera ezezka eta lehenaldi burutuan agertzen da. Normalean, ezezka eta lehenaldi burutuan adierazten den gertaera ez da gertatu; baldintzazko testuinguruaren kasuan, ordea, faktualitate balioa aldatzen da, hiztunak baitaki *etorri izana* gertatu dela.

(136) Etorri **izan** ez **banintz**, ez **nuke** lehendakaria **ezagutu izango**.

Perpaus generikoek kategoriei buruzko orokortzeak adierazten dituzte. Horregatik, perpaus horietako gertaerak (137) ezin koka daitezke denborako une zehatz batean, eta ondorioz ezin dute faktualitate-balio zehatzik hartu errealitatean gertatzen ari badira ere. Ez da, ordea, gauza bera gertatzen egoera guztietan: orokorrean egoerak deskribatzen dituzten gertaerak denboran koka daitezke eta gauzatzen direla esan daiteke.

(137) Airbus A320a korridore bakarreko hegazkina **da**.

Gertaerak gertatzen diren ekintzak eta egoera badira, kronologiako punturen batean gertatzen dira. Kronologiako puntu horiek adierazteko denbora-adierazpenak baliatzen ditugu. Jarraian ikusiko ditugu denbora-adierazpenen definizioa eta ezaugarri nagusiak.

3.2 Denbora-adierazpenak

Denbora-informazioaren elementuen definizioak eman ditugunean, ikusi dugu hainbat autorek diotela denbora-adierazpenak kronologiako uneen eta iraupenen adierazpenak direla. Hala, denbora-adierazpenek denbora esanahia dute eta perpauseko gertaerak denboran kokatzen dituzte, hau da, eurek seinalatzen dute perpauseko gertaerari dagokion denbora-unea edo luzapena. Adierazpide horiek guztiek, formaz ugariak badira ere, *noiz?*, *noiztik?*, *noiz arte?*, *noizko?* eta *zenbat denbora?* galderei erantzuten diete eta denbora une edo tarte mugatu bat (iraupen bat) irudikatzen dute.

3.2.1 Denbora-adierazpenak euskaraz

Euskaraz denbora-adierazpenak nagusiki adberbio-sintagmen bidez eta postposizio-sintagmen bidez adierazten dira, baina adjektibo-sintagmek, zenbait perpaus berezik eta zenbakizko egiturek ere adieraz dezakete denbora.

3.2.1.1 Adberbioak

Denbora adierazten duten adberbioek, euskal gramatika tradizionalen arabera, *aditzondo* funtzioa bete izan dute. Aditzondoak berez eta postposizio-atzizkirik gabe aditzari laguntzen dioten aditz-sintagmako elementuak dira (Altuna *et al.*, 1987). Adberbioek, normalean, zirkunstantzia bat adierazten dute. Denborazko adberbioek zehazki, aditzak adierazten duen gertaera denboran kokatzen dute.

Adberbioak formaz anitzak dira; Euskaltzaindiak (Altuna *et al.*, 1985) hainbat motatako aditzondoak batzen ditu⁵: adberbio bakunak (*atzo* (138) eta *maiz* (139)), adberbio eratorriak (*hilabeteka* (140)), zehaztugabeko konposatuak ((141) adibideko *noizbait* eta (142) adibideko *inoiz*) eta bestelakoak ((143) adibideko *dagoeneko*). Gertaeretan egin bezala, atal honetako adibideetan, nabarmendu nahi dugun egitura beltzez adieraziko dugu.

- (138) **Atzo** programa berria estreinatu zuten telebistan.
- (139) **Maiz** joaten gara Frantziara ostrak jatera.
- (140) Alokairua **hilabeteka** ordaindu behar da.
- (141) **Noizbait** entzuna nuen Mikelek gaur kontatu duen istorioa.
- (142) **Inoiz** horrelako soineko politik dendan ikusten baduzu, eros ieza dazu.
- (143) **Dagoeneko** euritakoa eta neguko arropa atera behar izan dugu.

(138–143) adibideetan adberbioak beren oinarritzko forman agertu badira ere, adberbioak egitura handiagoetan, postposizio-sintagmetan, ager daitezke eta *gaurtik* (144) edo *noizean behin* moduko esamoldeak (145) sortu.

- (144) **Gaurtik** aurrera ez duela gehiago erreko zin egin du.
- (145) **Noizean behin** jatetxe garestietan bazkaltzea gustatzen zaio Maitei.

Horiez gain, adizlagun itxura duten *askotan* (146), *sasoiz* (147) eta gisakoak aipatu behar ditugu, adizlagun itxura izanagatik ere, lexikalizaturiko aditzondoak baitira.

- (146) **Askotan** ahazten ditut eguzkitarako betaurrekoak hondartzara noanean.
- (147) Lanera **sasoiz** heltzeko ordubete lehenago esnatu ohi naiz.

⁵EGLUko aditzondoan sailkapena adberbioentzat baliatu dugu aditzondoak eta adberbioak parekoak baitira.

3.2.1.2 Izenak

Izenek ere denbora adieraz dezakete. Adberbioek bezala, perpauseko aditza denboran kokatzen dute gehienek. Orokorrean leku-denborazko atzizkiak (inesiboa nagusiki (148), adlatiboa (149), ablatiboa (150), leku genitiboa (151)) hartzen dituzte izenak, baina batzuetan sozietiboa (152) eta instrumentalak (153) ere hartzen dituzte. Postposizio askeak ere har ditzakete (154).

- (148) **Goizean** irten zen etxetik.
- (149) Iluntzetik **egunsentira** eten behar izan dituzte erreskate operazioak.
- (150) **Gabonetatik** egon zen Mikelen eskutitzaren zain.
- (151) Eskatutakoa **sanferminetarako** prest izango duzu.
- (152) Andra Mari eguna **ostegunarekin** jausten da aurten⁶.
- (153) **Bi orduz** egon naiz zure zain.
- (154) **Eguerdira arte** iraungo du emanaldiak.

Denbora adierazten duten izenak normalean postposizioak hartzen dituzte. Postposizio-sintagma horiek gehienetan adizlagun funtzioa betetzen badute ere, posible da perpausean funtzio ezberdina izatea, hala nola subjektua (155), atributua (156) edota izenlaguna (157).

- (155) **Gaurko goiza** oso luzea egin zait.
- (156) Izena emateko azken eguna **gaur** da.
- (157) Lehen Airbus A380a Singapur-en lurreratu da **12 orduko** hegaldia egin ostean.

3.2.1.3 Adjektiboak

Adjektibo batzuek⁷ ere denbora-informazioa adieraz dezakete. Adjektiboek perpauseko izen baten ezaugarri bat adierazten dute, kasu honetan denborazkoa. Adjektibo horiek izenlagun funtzioa (158) edota atributu funtzioa (159) bete dezakete.

- (158) Gure **eguneroko** borroka metroan leku librea topatzea da.
- (159) Margolan horiek gotikoak bai, baina **berantiarak** dira.

⁶ *Andra Mari eguna* ere denbora-adierazpena da, baina adibide honekin sozietiboaren denbora balioa nabarmendu nahi dugu.

⁷ 3.1.1.3 atalean azaldu bezala, euskararen tratamendu automatikoan izenlagun eta ize-nondoak adjektibo kategoriaren barruan hartu izan dira.

3.2.1.4 Perpaus bereziak

Azpiatal honen sarreran aipatu bezala, gertaera gehienak aditzen bidez adierazten dira, baina horrek ez du esan nahi aditz guztiek gertaerarik adierazten dutenik. Denbora adierazten duten perpaus berezi batzuetan, aditza perpausaren nukleoa bada ere, perpausaren denbora-esanahia gertaera-esanahiari gailentzen zaio. Perpaus horietan aditzak ez du gertaera bat adierazten, denbora-adierazpenaren parte da. Mota horretakoak dira *badira hiru aste, duela 20 urte* (160), *orain dela 25 urte*, *joan den astean* eta *datozen egunetan* modukoak.

- (160) **Duela 20 urte** AEBtako DowJones merkatuaren hondamendia izan zen.

3.2.1.5 Zenbakizko egiturak

Denbora adierazteko oso arruntak dira zenbakizko egiturak (161). Zenbakiz adierazitako orduak nahiz datek egitura finakoak dituzte euskaraz, eta erraz identifika daitezke. Euskaltzaindiaren 35.⁸ eta 37. arauetan⁹ datak eta orduak nola adierazi deskribatzen da, hurrenez hurren, baita zenbakiz nola adierazten diren ere.

- (161) Lehen Airbus A380a **18:40an** lurreratu da Singapur-en.

Datak zenbakiz adierazteko, Euskaltzaindiak jarraian datozen formak proposatzen ditu (162–165).

- (162) 1983/03/07; 1983/III/07
(163) 1983-03-07; 1983-III-07
(164) 2018-01; 01-22; 21ean
(165) 1993-03-07ko; 1993ko; 20an

Orduak zenbakiz adierazteko, Euskaltzaindiaren 35. arauak jasotzen duenez, orduak zein minutuak bi zifra baliatuz adierazi behar dira eta orduen eta minutuen arteko tarteak bi puntuez (:) (166–167). Era berean, zenbakizko orduak hitzezko ñabardurekin ere adieraz daitezke (168).

- (166) 18:43

⁸http://www.euskaltzaindia.eus/dok/arauak/Araua_0035.pdf

⁹http://www.euskaltzaindia.eus/dok/arauak/Araua_0037.pdf

(167) 18:00etan; 18:43an

(168) Arratsaldeko 6etan

Zenbakizko egiturak aparte deskribatu baditugu ere, postposizioak hartzean izenek bezala funtzionatzen dute eta morfema-multzo berak hartzen dituzte analisi morfosintaktikoan. Zenbakiz idatzitako *18:43* moduko forma horiek (166) leku-denborazko postposizio-atzizkiak (inesiboa, ablatiboa, adlatiboa...) eta *arte* bezalako postposizio askeak ere har ditzakete *18:43an* (167) moduko formak sortuz.

3.2.2 Abiarazle lexikoak

Euskaraz denbora adierazten duten egiturak ikusi berri ditugu. Batzuek berez egiten diote erreferentzia denborari, beste batzuen denbora-adierazpen izaera, ordea, testuinguruaren arabera da. Denborari berez erreferentzia egiten dioten elementu horiei abiarazle lexiko deritze. Abiarazle lexikoak, Ferro *et al.*-en (2003) arabera, denborazko unitate edo kontzeptu bat adierazten duten hitzezko edo zenbakizko adierazpenak dira. Denbora-informazioaren prozesamenduan garrantzitsua da abiarazle lexiko horiek identifikatzea, horien presentziak denbora-adierazpenaren presentzia adieraziko baitu kasu gehienetan.

Euskaraz ere, berez denbora adierazten duten formak aurki ditzakegu: *ordu*, *gaur*, *ekain*, eta abar. Aurreko ataletako adibideetan ikusi ahal izan denez, denbora-adierazpenek berez denbora adierazten duen unitatea izango dute buru askotan (*goizean* eta *gauera* (148), *sanferminetarako* (151), *ostegunarekin* (152) eta *orduz*) (153).

Forma horiek orokorrean denbora adierazten duten esapideetan agertzen dira. Ondorioz, abiarazle lexiko horiek aurretik deskribatutako egituraren batean agertuz gero, denbora-adierazpentzat hartuko dira, izan denbora-une edo iraupen bat. Hartara, euskarazko hainbat abiarazle lexiko zerrendatu ditugu (ikus A eranskina). Denbora adierazten duten hitz horien zerrenda zabaldu eta osatzeko, euskarazko WordNet (Pociello, 2008) kontsultatu dugu, ingelesezko *time* eta euskarazko *denbora* hitzen euskarazko hiponimoak bilatuz. Besteak Altuna *et al.* (1987) gramatikatik eta Ferro *et al.* (2003)-ko adibideetatik itzulita edo moldatuta lortu ditugu. 3.5. taulan abiarazleotako lagin bat gramatika-kategoriaren arabera sailkatu dugu.

Bestalde, 3.5. taulan adierazitako abiarazle lexikoak denbora adierazle moduan hartu izan diren arren, gerta daiteke testuinguru batzuetan hitz

3.5 taula – Kategoriaren arabera abiarazle lexikoen sailkapena

Kategoria	Abiarazle lexikoak
Izen arruntak	aro, arratsalde, asteburu, astelehen, eguerdi, etorki-zun, mende, minutu, ordu, otsail, sanfermin, sasoi, seiurteko, solstizio, udazken, etab.
Izen bereziak	Gabon, San Joan, etab.
Adjektiboak	eguneroko, hilabetekari, bienal, etab.
Denbora-patroiak	13:03, 2014/02/12, 1992ko, etab.
Adberbioak	gaur, berandu, orain, lehen, berehala, egundo, etab.
Zenbakiak	5etan, 6an, etab.

horiek denborarik ez adieraztea. (169) adibideko *arratsalde* denborazkotzat ez hartzea erraza izan daiteke unitate fraseologikoen azterketa eginez gero, esaldiko bi tokenak egitura fosilizatu batean agertzen baitira.

(169) **Arratsalde** on!

Beste kasu batzuetan anbiguitasun lexikoari edo polisemiari arreta jarri behar diogu. Esaterako, urteko hamargarren hilabetea adierazten duen *urri* formak (170) *bakan* edo *eskas* adierazten du (171) adibidean.

(170) Leihaketa beriro abiatuko da **urrian**.

(171) Eskolak psikomotrizitatea lantzeko dituzten baliabide **urriak** salatu ditu.

Amaitzeko, kontuan izan behar da denbora-adierazpen guztiek ez dutela abiarazle lexiko bat hartzen eta, ondorioz, egiturek denborazkoak direla erabakitzeko testuinguruaren beharra izango dugu. Denbora-adierazpenak prozesatzean ezingo gara forma horien semantikaz baliatu eta hori ere kontuan hartzekoa da egiturek automatikoki nola erauzi aztertzean. Adibidez, (172) adibidean *bostetan* denborazko adizlaguna da eta ordua adierazten du; (173) adibideko *bostetan* zenbatzaile izenordainak, ordea, lekua adierazten du:

(172) **Bostetan** geratu ginen kafea hartzeko.

(173) Bost etxe zituen eta **bostetan** jarri zuen jacuzzia.

3.2.3 Denbora-adierazpenen ezaugarri linguistikoak

Denbora-adierazpenek kronologiako denbora uneak eta iraupenak adierazten dituztenez, zein motatakoak diren eta kronologiako zein une edota zenbateko tartea adierazten duten identifikatzea da garrantzitsuena. Hau da, denbora-adierazpenaren mota eta kronologiako balio zehatza dira ezaugarri nagusiak. Hala ere, badira informazio osagarria eskaintzen duten beste ezaugarri batzuk ere. Horiek ondoko ataletan deskribatu ditugu.

3.2.3.1 Mota

Denbora-uneak edozein granularitate duten puntutzat hartuko ditugu, eta iraupenak esplizituki denbora-tarte bati dagozkien tartetzat. Adibidez, (174) adibidean agertzen den *datorren urteko urtarrilean* denbora-adierazpenak hila-bete oso bati egiten dio erreferentzia, baina hilabete hori unetzat hartzen da. (175) adibideko *12 orduko*, ordea, denbora-tarte baten luzerari buruzko erreferentzia zuzena da.

(174) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu atea.

(175) Lehen Airbus A380a Singapur-en lurreratu da **12 orduko** hegaldia egin ostean.

Lehen bereizketa hori gertaeren sailkapenari estuki lotuta dago. Izan ere, gertaera guztiak ez dira berdin gauzatzen denboran. Adibidez, (176) adibideko esaldia onartzea zaila egiten da, *tontorrera heltzea* une bakarrean gertatzen den gertaera baita eta ezin baita tarte batez luzatu.

(176) **Hiru orduz** heldu zen tontorrera.

2.2.2. atalean ikusi dugunez, denbora-une multzo bezala edo iraupen multzo bezala uler daiteke. Guk tarteko bidea hartuko dugu eta denbora-uneen zein tarteen bidez adieraziko dugu, euskaraz ere denbora horrela adierazteko bideak baitaude. Horien adierazpenak dira datak, orduak, iraupenak eta errepikapenak. Sailkapen hau aztertutako beste hizkuntzetan era bertsuan jokatu delako aukeratu dugu. Era berean, sailkapen hori bat dator normalizaziorako aukeratutako ISO-8601 arauarekin eta horrek erraztu egiten digu denbora-adierazpenen prozesamendua.

Datak Daten multzoan eguna baino granularitate handiagoa duten denbora-adierazpenak (*eguna* (177), *hilabetea* (178), *urtea* (179), etab.) batzen ditugu.

- (177) Autogile estatubatuarrak **ostiralean** jakinarazi zuen 1000 kontzesionario ixteko asmoa duela.
- (178) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu ateak.
- (179) Zifra hori 3.605era jaisteko asmoa du **2010erako**.

Denbora-adierazpen horiek gizakiok antzeman dezakegun iraupena badute ere, denbora-une bezala ere trata daitezke eta denbora-lerroan puntutzat irudika daitezke.

Orduak Orduen multzoko denbora-adierazpenek eguna baino granularitate txikiagoa duten denbora-uneak adierazten dituzte, hala nola egunaren parteak (*goiza* (180)), orduak (*18:40an* (181)), minutuak, segundoak, baita unitate txikiagoak ere.

- (180) **Gaur goizean** bi orduko bilera egin dute sindikatueta ordezkariek.
- (181) Lehen Airbus A380a **18:40an** lurreratu da Singapur-en.

Iraupenak Iraupen-adierazpenek denbora-tarteak adierazten dituzte. Lan honetan esplizituki denbora-tartearen iraupena adierazten dutenak hartu ditugu iraupentzat. Esaterako, (182) adibidean *12 orduko* denbora-adierazpena agertzen da. Denbora-adierazpen horrek denbora-tarte baten iraupena esplizituki adierazten du eta, ondorioz, iraupenen multzoan hartu dugu. (183) adibideko, *1792–1795* tarteak, ordea, ez dugu iraupentzat hartu ez baita iraupenaren luzera esplizituki adierazten; hau da, hasiera- eta amaiera-uneak baino ez dira adierazten. Horrelako kasuetan, bi denbora-adierazpen desberdintzat hartuko ditugu une horiek.

- (182) Lehen Airbus A380a Singapur-en lurreratu da **12 orduko** hegaldia egin ostean.
- (183) Konbentzio Nazionalea (**1792–1795**).

Errepikapenak Errepikatzen diren denborak edo denbora-multzoak adierazten dituzten denbora-adierazpenak dira. Multzo honetan sartzen ditugu maiztasuna (184) eta kuantifikazioa (185) adierazten dituztenak. Maiztasunek gertaera baten maiztasuna adierazten dute eta kuantifikazioek denbora tarte batean gertaera zenbatetan gertatzen den.

(184) 1970eko hamarkadan 300.000 ale saltzen zituen **astero**.

(185) **Hamabostean behin** egiten duten elkarretaratzean hainbat hautesiren babesak jaso dute gaur.

3.2.3.2 Balio normalizatua

Denbora-adierazpenek adierazten dituzten une eta iraupen horiei balio zehatza eman behar zaie denbora-informazioaren prozesamenduan. Hau da, bakoitzak zehazki kronologiako zein punturi egiten dion erreferentzia edo zenbateko luzapena hartzen duen adierazi behar da, eta hori ordenagailu bidez interpretagarria izango den eran egin behar da (ikus 4.2.2.2 atala).

Kasu batzuetan erraza da denbora-adierazpenak kronologiako zein uneri (186) edo zenbateko tarteari (187) egiten dioten erreferentzia ebaztea, esplizituki adierazten baita testuan. Beste denbora-adierazpen batzuk, ordea, ez dira hain errazak normalizatzeko, denbora-adierazpen erlatiboak (188) direlako edo lausoak direlako (189). Denbora-adierazpen erlatiboen kasuan esate-unearen edo testuaren sorrera-dataren arabera esleia dakieke balio normalizatu bat. Denbora-adierazpen lausoei, ordea, ezin zaie balio normalizatu zehatzik esleitu, denbora-adierazpenak eta testuinguruak ez dutelako horretarako informazio nahikoa adierazten.

(186) iTunes Music Store onlineko musika-denda merkaturatu zuen **2003ko apirilaren 28an**.

(187) Lehen Airbus A380a Singapur-en lurreratu da **12 orduko** hegaldia egin ostean.

(188) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu ateak.

(189) Hegazkinari 400 bat gonbidatuk egin zioten ongietorria **laster** zabalduko den 3. terminalean.

3.2.3.3 Denbora-adierazpenak testuan betetzen duen funtzioa

Denbora-adierazpen guztiak ez dira berdin gauzatzen testuetan: batzuk testuko perpausetan agertzen diren denbora-adierazpen *arruntak* dira eta tes-

tuaren baitan ulertzen dira. Aitzitik, beste batzuek testua ulertzeko informazio gehigarria dute.

Aurreko ataletan esaldietan agertzen diren denbora-adierazpenak *arruntak* direla esan dezakegu perpauseko gertaerak denboran kokatzen laguntzen baitute eta testuan lortzen baitute beren esanahia. Adibidez, (190) adibideko *2003ko apirilaren 28an* perpauseko informazioarekin ulertzen da.

(190) iTunes Music Store onlineko musika-denda merkaturatu zuen **2003ko apirilaren 28an**.

Hala ere, aipatu berri dugunez, denbora-adierazpen batzuei ezin zaie berez balio bat esleitu denbora-adierazpen anbiguoak edo lausoak direlako. Esaterako, (191) adibideko *datorren urteko urtarrilean* denbora-adierazpenari balioa esleitu ahal izateko, perpausa noiz idatzi zen ezagutzea baino ez da behar.

(191) 3. terminalak **datorren urteko urtarrilean** zabalduko ditu ateak.

(191) adibideko denbora-adierazpena ebazteko testuz kanpoko erreferentzia une bat behar da. Testu batzuetan sorrera-data edo argitalpen-data modukoak adierazten dira denbora-adierazpenen bidez. Denbora-adierazpen horiek ez dute testuan parte hartzen, baina badituzte funtzio berezituak: testuaren denbora-markoa definitzea testuko denbora-adierazpen erlatiboak desanbiguatzeke informazioa ematea. Adibidez, 3.1 irudian *2007ko urriaren 17a* denbora-adierazpenak *datorren urteko urtarrilean* unea *2008ko urtarrila* izango dela erabakitzen laguntzen du.

3.3 Erlazioak

Aipatutako gertaera eta denbora-adierazpen horiek bere horretan denbora-informazioaren zati bat baino ez dute adierazten. Esaterako, (192) adibidean *merkaturatzea* gertatu dela eta *2003ko apirilaren 28ko* data agertzen dira. Hala ere, intuitiboki gertaeraren eta denbora-adierazpenaren arteko lotura egiten dugu eta *merkaturatzea 2003ko apirilaren 28an* gertatu dela jakin dezakegu. Izan ere, testuko elementuak elkarri lotzen zaizkie. Lotura horiek *erlaziotzat* hartu ditugu.

Atal honetan denbora-informazioan esanguratsuak direnak aztertuko ditugu eta erlazioen adibideak ematean erlazioaren iturria edo burua beltzez

10026-First_A380_enters_commercial_service.txt

s0 Lehen A380a hasi da zerbitzu komertziala ematen .

s1 2007ko urriaren 17a .

Hegazkingintzaren historian , sekulako lorpena izan da : lehen Airbus A380a

18:40an (GMT + 8) lurreratu da Singapur-en , Changi nazioarteko

s2 aireportuan , Airbus-en bidalketa-zentrotik atera eta 12 orduko hegaldia egin ostean .

Hegazkinari 400 bat gonbidatuk egin zioten ongietorria laster zabalduko den

s3 Changi Nazioarteko aireportuko 3. terminalean . Tartean zen Lee Hsien Loong , Singapur-eko lehen ministroa .

Hitzaldian , lehen ministroak Singapore Airlines (SIA) aire-linea hartzailea

s4 txalotu zuen , eta aipatu zuen arrakasta Singapur-eko biztanle guztiak harro egoteko modukoa zela .

Gehitu zuen lurreratzea gertaera gogoangarria zela , ez bakarrik SIArentzat ,

s5 baita Changi nazioarteko aireportuarentzat ere , zeina herrialdeko aireporturik handiena baita .

Orain arteko terminal guztiak prest daude A380a hartzeko , eta 3.

s6 terminalak (1,75 milioi dolar singapurtar kostatu da) datorren urteko urtarrian zabalduko ditu ateak .

3.1 irudia – Testu bateko denbora-adierazpenak

emango dugu; erlazioaren xedea edo mendeko elementua, kurtsibaz; eta erlazioa testuan esplizituki adierazten duen seinalea azpimarraz.

(192) iTunes Music Store onlineko musika-denda **merkaturatu zuen 2003-ko apirilaren 28an**.

Perpaus bereko gertaerak edota denbora-adierazpenak lotzen dituzten erlazioak gertaera horiek kronologian kokatzeko informazio asko adierazten

badute ere, ez dira denbora-informazioaren prozesamenduan balia daitezkeen bakarrak. Hala, atazaren berezitasunen arabera, atazarako ahalik eta informazio gehien eman dezaketen erlazioak azaleratuko dira. Denbora-informazioaren inguruko atazetan baliatu diren erlazioak aztertu ditugu, euskarazko denbora-informazioaren prozesamenduan zein erlazio mota landu erabakitzeke.

NewsReaderren (Agerri *et al.*, 2014) proposatutako erlazioak ekonomia-albisteetako informazioaren erauzketarekin estuki lotuta daude. Proiektuaren helburua ekonomia-albisteak enpresentzat baliagarri egitea zenez, gertaeretakoko parte-hartzaileak eta zenbakizko egiturak identifikatu zituzten, horiek garrantzia handia baitute ekonomiari buruzko albisteetan. Era berean, ondoko sei erlazio motak definitu ziren:

- dokumentu barruko gertaeren korreferentzia-erlazioak
- gertaeren eta gertaeretan parte hartzen duten entitate eta zenbakien arteko erlazioak
- gertaeren arteko kausazko erlazioak
- mendekotasun-erlazioak
- gramatika-erlazioak (aspektuzko erlazioak, HAULetako elementuen arteko erlazioak eta gertaerak ahalbidetzen dituzten gertaeren eta ahalbidetutako gertaeren artekoak)
- denbora-erlazioak

Ikusten denez, NewsReaderren denbora-erlazioez gain, mendekotasun- eta gramatika-erlazioak ere aztertu dituzte eta kausazko erlazioei garrantzi berezia eman diete. Bestalde, gertaera-instantzien arteko korreferentzia ere aztertu dute, hori aztertzeak testu kantitate handietan gertaerak identifikatzen laguntzen baitu, gertaera-instantzia horiek errealitateko gertaera bakarrari egiten baitiote erreferentzia. Bistan da NewsReaderren proposatutako erlazio-eskemaren bidez informazio gehiago adierazten dela, baina etiketatzaileen lan handia eskatzen du eta denbora-informazioa baino esparru zabalgia hartzen du.

SIBILA proiektuan (Wonsever *et al.*, 2008), ordea, mendekotasunezko eta aspektuzko erlazioak baino ez dira kontuan hartzen. SIBILaren helburua gertaeren faktualitatea aztertzea da eta horregatik alde batera uzten dira

denbora-erlazioak. Proposatzen dituzten mendekotasunezko eta aspektuzko erlazioak oso hurbil daude TimeMLn proposatutako erlazioetatik.

Hain zuzen ere, TimeMLn hiru erlazio hauek definitu ziren:

- denbora-erlazioak
- mendekotasun-erlazioak
- aspektuzko erlazioak

Ikus daitekeenez, erlazio horiek testuko elementuen arteko semantikaren esparruko erlazioak adierazten dituzte. Zehatzago esanda, denboraren irudikapenarekin harremana duten erlazioak dira: denbora-erlazioek gertaerak ordenatzen laguntzen dute eta aspektuzkoek gertaeren faseak (hasiera, amaiera, etab.) adierazten dituzte. Mendekotasun-erlazioek, berriz, gertaerak errealtatean gauzatzeari buruzko informazioa ematen digute, besteak beste.

Ikerketa-lan honetan, TimeMLn definitu diren erlazio horietan oinarritu gara. Jarraian deskribatuko ditugu erlazio horiek, baita euskaraz zein elementuk parte hartzen duten ere.

3.3.1 Denbora-erlazioak

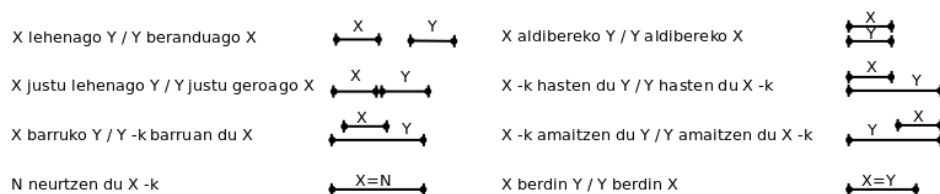
Denbora-erlazioak gertaeren, denbora-adierazpenen edota gertaeren eta denbora-adierazpenen artean sortzen diren erlazioak dira. Erlazio horiek lotzen dituzten elementuen arteko hurrenkera kronologikoa adierazten dute. Denbora-erlazioa bi gertaeraren edo bi denbora-adierazpenen artean gertatzen bada, biak beren artean denboraren arabera ordenatzen dira, eta gertaera baten eta denbora-adierazpen baten artekoa bada, gertaera denborari ainguratzen zaio.

Gertaerak kronologian kokatzeko, gertaera bakoitza zein denborari lotuta dagoen edota zein gertaeraren aurretik edo ondoren gertatzen den erabaki behar da. Adibidez, (193) adibidean adierazten da une batean *Superjumboa zerbitzuan hasiko dela* eta, geroago, *2011* urtea hasiko dela. Esan daiteke *hastearen* eta *2011*ren artean denbora-erlazioa sortu dela.

(193) Superjumboa *2011* baino lehen **hasiko da** zerbitzuan.

Gertaerak zein kronologiako une eta tarteak ez dira beti sekuentzialki gertatzen, ordea. Hain zuzen ere, gertaeren edota denbora-adierazpenen artean

13 motatako erlazioak sor daitezke 3.2 Allenen denbora-tarteen teoriaren arabera. irudian ikus daitekeenez. Errepresentazio hori denbora-adierazpenak eta gertaerak interbalo bezala ikusiz gero da posible, batzuek besteak guztiz edo partzialki beren baitan har baititzaizkete.



3.2 irudia – Allenen denbora-tarteen teoriako erlazioak

Testuetako gertaeren eta denbora-adierazpenen arteko erlazioei dagokienez, denbora-erlazio batzuk inplizituak dira (194), baina beste asko esplizituki adierazten dira denborazko seinaleen (195) (ikus 3.3.1.2 atala) bitartez.

- (194) Lehen ministroak aire-linea **txalotu zuen** eta *aipatu zuen* arrakasta Singapur-eko biztanle guztiak harro egoteko modukoa zela.
- (195) Lehen airbus A380a Singapur-en **lurreratu da** 12 orduko hegaldia *egin ostean*.

Esaterako, (194) adibidean *txalotzea aipatzearen* aurretik gertatu zela on-doriozta daiteke, horretarako ebidentziarik ez badago ere. (195) adibidean, ordea, *ostean* hitzak lurreratzea *12 orduko hegaldiaren* ondoren gertatu zela esplizituki adierazten digu. Hala, *ostean* denborazko seinale bat dela esan dezakegu.

3.3.1.1 Denbora-erlazio motak

Allenen teoriari jarraituz, denbora-tarteen (edo gertaeraren gauzatzeen) artean 6 konbinaketa pare posible eta berdintasun erlazio bat sortzen dira eta horietan oinarrituta definitu dira TimeMLn proposatutako erlazioak. Erlazio horiek hainbat eratara gauzatuko dira testuan, baina beti izango dira berak. Honakoak dira denbora-erlazio horiek eta beren adibideak:

- Gertaera edo denbora-tarte bat beste gertaera edo denbora-tarte bat baino lehenago/beranduago gertatzea

- (196) Lehen airbus A380a Singapur-en **lurreratu da** 12 orduko hegaldia *egin ostean*.
- Gertaera edo denbora tarte bat justu beste gertaera edo denbora-tarte bat baino lehenago/beranduago gertatzea
- (197) Kamioi batek bi txirrindulari **harrapatu** eta *hil ditu*.
- Gertaera edo denbora-tarte bat beste gertaera edo denbora-tarte baten barruan gertatzea
- (198) Petro Poroxenko Ukrainako presidenteak astebeteko su-etena **iragarri zuen** *joan den astean*.
- Gertaera edota denbora-tarte bi aldiberean gertatzea
- (199) Sorginak *gauerdian* **irteten dira** akelarreetarantz.
- Gertaera edo denbora-tarte batek beste gertaera edo denbora-tarte bat hastea
- (200) Grebalariak *astelehenetik* **daude** lantegian itxita.
- Gertaera edo denbora-tarte batek beste gertaera edo denbora-tarte bat amaitzea
- (201) Grebalariak *astelehenera* arte **egongo dira** lantegian itxita.
- Gertaera edo denbora-tarte bat beste gertaera edo denbora-tarte baten berdina izatea
- (202) Fundazio horrek Madrilen baso-eskola **irekitzeko** asmoa du. 2015–2016 ikasturtean *abiatuko* litzateke Bosquescuola ekimena.
- Gertaera batek iraupena izatea
- (203) Lehen airbus A380a Singapur-en lurreratu da *12 orduko* **hegal-dia** egin ostean.

3.3.1.2 Denbora-erlazioen seinaleak

Arestian aipatu bezala, denbora-erlazio batzuk esplizituki adierazten dira testuetan seinaleak deritzen elementuen bidez. Elementu horiek ez dute soilik denbora-erlazioa markatzen, zein erlazio mota dagoen ere adierazten dute.

Seinaleak definitzean, etiketatzeari eta prozesamendu automatikoari erreparatu diogu, seinaleen definizioa eta etiketatuko duguna bat etortzea nahi dugulako. Etiketatzea token mailan egitea erabaki dugunez, ondoko egiturak baino ezingo ditugu seinaleztat hartu. Egitura horiek gramatika kategoriaren arabera sailkatu ditugu.

- Denborazko postposizio askeak
 - (204) Lehen Airbus A380a Singapurren lurreratu da 12 orduko hegaldiaren ostean.
 - (205) Cata 16 milioiko inbertsioa egiteko prest zegoen epaitegian bertan erregistratu ondoren.
 - (206) Enpresak 60,4 milioiko galera izan du ekainera arte.
 - (207) Hor da negoziazio epea. Hortik aurrera konkurtso legeak aurrera jarraituko du.
- Denborazko postposizio-atzizkiak
 - (208) Bahituak aurkitu zituztenean, oso egoera txarrean zeudela konturatu ziren.
 - (209) Esan zitzaien ez erabiltzeko “betierekotasun” hitza menpeko ekarpenak adieraztean.
 - (210) Lehen Airbus A380a Singapurren lurreratu da 12 orduko hegaldiaren ostean.
- Karaktere bereziak, hala nola -, : eta /
 - (211) Irekiera ordutegia: 15:00_18:00.
 - (212) 2017_2020 urteetarako aurrekontuak onartu zituen batzordeak.

3.3 irudiko testuan *ostean* eta *arteko* postposizioek markatzen dituzte denbora-erlazioak esplizituki.

10026-First_A380_enters_commercial_service.txt

s0 Lehen A380a hasi da zerbitzu komertziala ematen .

s1 2007ko urriaren 17a .

Hegazkingintzaren historian , sekulako lorpena izan da : lehen Airbus A380a

18:40an (GMT + 8) lurreratu da Singapur-en , Changi nazioarteko

s2 aireportuan , Airbus-en bidalketa-zentrotik atera eta 12 orduko hegaldia egin
ostean.

Hegazkinari 400 bat gonbidatuk egin zioten ongietorria laster zabalduko den

s3 Changi Nazioarteko aireportuko 3. terminalean . Tartean zen Lee Hsien Loong ,
Singapur-eko lehen ministroa .

Hitzaldian , lehen ministroak Singapore Airlines (SIA) aire-linea hartzailea

s4 txalotu zuen , eta aipatu zuen arrakasta Singapur-eko biztanle guztiak harro
egoteko modukoa zela .

Gehitu zuen lurreratzea gertaera gogoangarria zela , ez bakarrik SIArentzat ,
s5 baita Changi nazioarteko aireportuarentzat ere , zeina herrialdeko aireporturik
handiena baita .

Orain arteko terminal guztiak prest daude A380a hartzeko , eta 3.

s6 terminalak (1,75 milioi dolar singapurtar kostatu da) datorren urteko
urtarrilean zabalduko ditu ateak .

3.3 irudia – Testu bateko seinaleak**3.3.2 Mendekotasun-erlazioak**

Mendekotasun erlazioak bi gertaeraren artean eratzen diren erlazioak dira. Erlazio horietan gertaera batek beste bat (edo proposizio oso bat) argumentutzat hartzen du. Esaterako (213) adibidean, *diote* gertaerak *historiak errepikatzeko joera du* proposizioa hartzen du argumentutzat.

(213) Historiak errepikatzeko joera *duela diote* zenbaitek.

Hizkuntzalaritzan, normalean mendekotasun-erlazioez ari garenean, syntaxi-mailako mendekotasunaz ari gara. Mendekotasun sintaktikoak, ordea, ez du denbora-informazioaren analisisian eragin handirik. Esparru horretan mendekotasun semantikoari erreparatu behar zaio, jakina bada ere mendekotasun semantikoa askotan mendekotasun sintaktikoaren bidez adierazten dela. Hala, denbora-informazioa aztertzean, gertaeren errealitateko gauzatzeari buruzko informazioa ematen duten gertaeren arteko erlazioak aztertu behar dira.

Bi mendekotasun-erlazio eskema aztertu ditugu: SIBILA proiektuan proposatutakoa (Wonsever *et al.*, 2008) eta TimeMLkoa (Pustejovsky *et al.*, 2003a). Bi horietaz gain, NewsReader proiektuan proposatutakoa ere aztertu dugu, baina alde batera utzi dugu, eskaintzen dituen mendekotasun-erlazio motak oso murrizak direlako. SIBILAk eta TimeMLk, ordea, mendekotasun-erlazio eskema zabala dute. Hain zuzen ere, SIBILAk zortzi mendekotasun-erlazio mota hauek eskaintzen ditu:

- **Modalak** (*modal*): esperientzia subjektiboak adierazten dituzten gertaerek sortutako erlazioak dira. Gertaera horiek balizko mundu edo egoerak hartzen dituzte argumentu gisa.

(214) Endesak **uste du** Garoña berriro zabaltzeko *gai izango dela*.

- **Ebidentzialak** (*evidencial*): gertaera diskurtsibo edo hautematezkoek sortzen dituzten erlazioak dira.

(215) Kapitainak bera ere inoiz marinel *izan zela* **esan zuen**.

- **Ezezkako ebidentzialak** (*negativo evidencial*): lexikoki ezeztapena edo ukapena adierazten duten gertaera diskurtsiboek sortzen dituzten erlazioak dira.

(216) Kapitainak bera ere inoiz marinel *izana* **ukatu zuen**.

- **Faktikoak** (*fáctico*): mendeko gertaeraren egiazkotasuna ontzat hartzen duten predikatuek sortzen dituzten erlazioak dira.

(217) Kapitainak bera ere inoiz marinel *izan zela* **ahaztu zuen**.

- **Kontrafaktikoak** (*contrafáctico*): mendeko gertaeraren egiazkotasun eza ontzat hartzen duten predikatuek sortzen dituzten erlazioak dira.

(218) Kapitainak bera ere inoiz marinel *hegaldia* **kantzelatu zuen**.

- **Baldintzazkoak** (*condicional*): bi gertaeren artean baldintzazko tarreetan (*periodos condicionales*) sortzen diren erlazioak dira.

(219) Zuei horrelakorik **gertatzen bazaizue**, ez zenukete atarramentu onik *aterako*, duzuen eskarmentu urriagatik.

- **Existentzialak** (*existencial*): aditz kopulatiboen, existentzialen eta aurkezpenezkoen edo antzerako nominalizazioen eta gertaerak adierazten dituzten beste batzuen arteko erlazioak dira.

(220) Gaur Bilboko Euskaraldiaren *aurkezpena* **egin dute**.

- **Kausazkoak** (*causal*): kausa adierazten duten predikatuek sortutako erlazioak dira.

(221) Euriteek *uholdeak* **eragin zituzten**.

Ikus daitekeenez, SIBILako sailkapenean mendekotasun semantikoa eta sintaktikoa nahasten dira, erlazio existentzialen kasua sintaxi mailakoa baita. TimeMLn, ordea, sei erlazio mota hauek baino ez dira definitu:

- **Modalak**: gertaera batek unibertso posible bati buruzko erreferentzia egiten du. Bestela esan, buru den gertaerak balizko errealitate baten aukera zabaltzen du.

(222) Endesak **uste du** Garoña berriro zabaltzeko *gai izango dela*.

- **Faktiboak** eta **ez faktiboak**: buru diren gertaerek beren argumentu diren gertaeren egiazkotasunaren edo egiazkotasun ezaren ondorio logikoa aurkezten dute. Zehatzago, erlazio faktiboetan, gertaera nagusiek argumentu diren gertaerak gertatu direla adierazten dute eta, ez-faktiboetan, ez direla gertatu adierazten dute.

(223) Kapitainak bera ere inoiz marinel *izan zela* **ahaztu zuen**.

- **Ebidentzialak** eta **ezezko ebidentzialak**: gertaera nagusiak bere mendeko gertaeraren ebidentzia edo ebidentziaren ukapena adierazten du.

(224) Kapitainak bera ere inoiz marinel *izan zela* **esan zuen**.

(225) Kapitainak bera ere inoiz marinel *izana* **ukatu zuen**.

- **Baldintzazko erlazioak:** baldintzazko erlazioetan protasia eta apodosiaren arteko erlazioak sortzen dira.

(226) Zuei horrelakorik **gertatzen bazaizue**, ez zenukete atarramentu onik *aterako*, duzuen eskarmentu urriagatik.

TimeMLn definitzen diren erlazioetan, mendekotasun semantikoa baino ez dago. Aitortzen dugu erlazio-semantiko gehiago ere defini daitezkeela, kausa esaterako, baina gure lanean TimeMLn proposatutako mendekotasun-erlazioen sailkapenari jarraituko diogu. Alde batetik, gertaeren sailkapen semantikoa ere TimeMLtik hartu dugu eta erlazioak horietan oinarrituko baditugu, pentsatzekoa da gertaera mota beretatik erlazio mota berak sortuko direla. Beste aldetik, gure etiketatzea beste hizkuntzetako TimeML etiketatzeekin konparatu ahal izateko, erlazio mota berdinak mantendu behar ditugu euskararako ere egokiak direla iruditzen zaigulako. Hala ere, etorkizunean mendekotasun-erlazioen azterketan sakontzeari eta horien kopurua zabaltzeari interesgarri deritzegu.

Horretaz gain, kontuan izan behar da mendekotasun-erlazioek faktualitatearekin (ikus 3.1.3.2 atala) harreman estua dutela, erlazioaren buruek mendeko gertaeraren egia-balioari eragiten baitiote. Ondorioz, mendekotasun-erlazioak tratatzeak faktualitatearen ebazpenean lagunduko du.

3.3.3 Aspektuzko erlazioak

Aspektuzko erlazioak aspektuzko gertaera baten eta beste gertaera baten artean eratzen direnak dira. Aspektuzko gertaerek denboran luzatzen den gertaera baten (edo iraupen baten) faseak adierazten dituzte eta aspektuzko erlazioetan gertaera horren fasea adierazten dute. Adibidez, (227) adibidean *saltzearen* hasieraren berri ematen da eta *hasi zen* eta *saltzen* gertaeren artean aspektuzko erlazioa dagoela esango dugu.

(227) Apple multimedia produktu batzuk *saltzen* **hasi zen**.

Aztertu ditugun etiketatze-eskemen artean, SIBILAk eta TimeMLk eskaintzen dituzte aspektuzko erlazioen sailkapenak. SIBILAn ondoko bost aspektuzko erlazio mota hauek definitu dira:

- **Hasiera** (*inicio*)

(228) 1956. urtearen bueltan, *irakasle* **hasi zen**.

- **Berrartzea/errepikapena** (*reiteración*)

(229) Isunak *jartzeari* **berrekingo dio** udalak datorren hilabetetik aurrera.

- **Amaiera** (*terminación*)

(230) Mikelek liburua *irakurtzen* **amaitu du**.

- **Jarraipena** (*continuación*)

(231) Autobide berriaren *lanek* aurrera **diraute**.

- **Beste aspektuzko erlazio batzuk** (*otros*). Horiek ez dira SIBILAn zehazten.

Lan honetan definitutako beste erlazioekin egin bezala, aspektuzko erlazioak sailkatzeko, TimeMLn proposatutako sailkapena ekarri dugu euskarazko denbora-informazioaren prozesamendura. TimeMLn aspektuzko erlazioak erlazioaren buru diren aspektuzko gertaerek ahalbidetzen dituzte, eta ondoko erlazioak adieraziko dituzte:

- **Hasiera**: aspektuzko gertaerak argumentutzat duen gertaeraren hasiera adierazten du.

(232) 1956. urtearen bueltan, *irakasle* **hasi zen**.

- **Amaiera**: aspektuzko gertaerak argumentutzat duen gertaeraren amaiera naturala adierazten du; hau da, ez da gertaera hori behar baino lehenago eteten.

(233) Mikelek liburua *irakurtzen* **amaitu du**.

- **Etetea**: gertaera bat bere berezko amaiera baino lehenago eteten dena. Aspektuzko gertaerak argumentutzat duen gertaeraren etetea adierazten du. Bestela esan, argumentua den gertaera amaiera naturalera heldu baino lehen amaitu da.

(234) Gobernuak azalpenak *emateari* **utzi dio**.

- **Jarraipena:** aspektuzko gertaerak argumentutzat duen gertaeraren jarraipena adierazten du.

(235) Autobide berriaren *lanek* aurrera **diraute**.

- **Berrabiaketa:** aspektuzko gertaerak argumentutzat duen gertaera berriro abiatu dela adierazten du. Hau da, etenda zegoen gertaera bat berriz martxan jarri dela adierazten da.

(236) Isunak *jartzeari* **berrekingo dio** udalak datorren hilabetetik aurrera.

3.4 Laburpena

Kapitulu honetan, denbora-informazioa euskaraz zein gertaerak eta denbora-adierazpenek adierazten duten aztertuko dugu. Zehazki, elementu horiek zein forma hartzen duten eta beren zein ezaugarri linguistikok laguntzen duten denbora-informazioan deskribatu dugu. Era berean, gertaeren edota denbora-adierazpenen artean sortzen diren erlazioak definitu ditugu. Erlazioen bidez gertaerak denboran kokatzen dira, bai kronologian ainguratuta, bai beren artean ordenatuta.

Azterketa egin ondoren, informazio hori guztia nola kodetu aztertuko dugu hurrengo atalean. HPn informazioa ordenagailu bidez prozesagarria izan behar denez, denbora-informazioa kodetzeko eskema definitu behar da. Euskarazko denbora-informazioa kodetzeko, EusTimeML etiketatze-lengoaia definitu dugu.

EusTimeML bidezko eskuzko etiketatzea eta EusTimeBank Corpusaren osaketa

Hizkuntzaren prozesamenduan (HP) ezinbestekoak dira corpusak hizkuntzaren azterketarako zein tresnen garapenerako. Izan ere, corpusak hizkuntzen edo hizkuntzen ezaugarri edo egitura konkretu batzuen lagin adierazgarriak dira eta espero izatekoa da ondo eraikitako corpus batean aurkitzen den informazioak hizkuntza horren ezaugarri edo egitura nagusiak jasoko dituela. Hala definitzen dituzte corpusak McEnery *et al.*-ek (2006, 5 or.): “a corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety”.

Hizkuntzaren prozesamendurako tresnen garapenari dagokionez, zenbat eta lagin adierazgarriagoa izan, orduan eta hizkuntza-modelo osotuagoak sortuko dira eta tresnek egitura gehiago tratatu ahal izango dituzte. HPn zehazki corpusek bi erabilera nagusi dituzte i) baliabide linguistikoak garatzen laguntzen dute, hizkuntzako fenomenoak islatzen baitituzte, eta ii) tresna automatikoen garapenean laguntzen dute, tresna horiek entrenatzeko informazioa batzen baitute eta tresna horien ebaluazioa egiteko urre-patroia baitira.

HPn erabiltzen diren corpus horiek, testu gordinaz gain, horri atxikitzen zaion informazio linguistiko egituratua dute. Horren adibide dugu, esaterako, Euskararen Prozesamendurako Erreferentzia Corpora (EPEC) (Aduriz *et al.*, 2006; Aldezabal *et al.*, 2009). Informazio hori testutik bereiz ematen da eta markaketa-lengoiaren bidez adierazten da. Markaketa-lengoiaren bi-

dez, testuko hitzei ordenagailu bidez prozesa daitekeen informazioa gehitzen zaie. Bestela esan, hitzei adierazten duten informazioa esplizitu egiten duten etiketak esleitzen zaizkie.

Corpus etiketatuak eskuz, automatikoki edo erdi-automatikoki osa daitezke. Automatikoki sortzean, tresnak berak txertatzen ditu etiketak eta horien balioak. Eskuzko etiketatzea egiteko, ordea, etiketatzailerak lagunduko dituen interfaze bat behar da. Interfaze horren bidez, eskuz gehitutako informazio linguistikoa era egituratuan antolatuko da. Etiketatze erdi-automatikoa egitean, bi iturri horietatik lortutako etiketatzeak uztartzen dira.

Ikerketa-lan honetan, denbora-informazioa ordenagailu bidez eskuragarri izateko, EusTimeML markaketa-lengoaia definitu dugu (4.1 atala). Horretarako denbora-giturak etiketatzeko eskema definitu dugu (4.2 atala), eta hori baliatuta, corpora etiketatu dugu. Corpus hori EusTimeBank dugu (4.3 atala) eta corpusaren etiketatze-metodologia ere deskribatzen dugu. Zehazki, 4.4 atalean eskuzko etiketatzea egiteko sistema aurkezten dugu eta 4.5 atalean definitutako etiketatze-irizpideak ebaluatze egin ditugun esperimentuak deskribatzen ditugu. Amaitzeko, corpusari eman dizkiogun beste erabilerei buruz 4.6 atalean ematen dugu.

4.1 EusTimeML markaketa-lengoaia

Aurreko kapituluan aurkeztu ditugun gertaerak, denbora-adierazpenak eta horien artean sortzen diren erlazioak esplizitu egin behar dira testuetan, testu horietako informazioa automatikoki erabili ahal izateko. Euskarazko denbora-informazioa etiketatzeko, EusTimeML baliatu dugu. EusTimeML TimeMLren euskararako moldaketa da eta ezaugarri batzuk zuzenean hartu baditugu ere, euskarazko denbora-informazioa errepresentatzeko aldaketak ere egin ditugu.

Jarraian, TimeML deskribatuko dugu (4.1.1 atala). Gero, 4.1.2 atalean EusTimeML sortzeko jarraitu dugun metodologiaren berri emango dugu eta euskarazko denbora-informazioa kodetzeko definitutako etiketa, atributu eta horien balioak deskribatuko ditugu.

4.1.1 TimeML

TimeML markaketa-lengoaia gaur egun denbora-informazioa etiketatzeko estandarra da eta hainbat esperimentu eta atazatan baliatu da 2.3.1 atalean

aurreratu dugun bezala. TimeML testuan esplizituki ageri diren gertaerak, denbora-adierazpenak eta beren arteko erlazioak adierazten dituzten hitzak etiketatzeke erabiltzen da. Token mailako etiketatzea egiten duenez, azaleko analisisian oinarritutako markaketa-lengoaia da (Im *et al.*, 2009; Bittar, 2009), baina denbora-egituren identifikazioa azalekoa bada ere, horien kategorizazioa beren ezaugarri semantikoetan oinarritzen da.

TimeML lehen aldiz NRRC (National Regional Research Center) udako tailerrean (Pustejovsky *et al.*, 2002) aurkeztu zen, gertaerei eta entitatei buruzko testuetan oinarritutako denborazko galderei erantzuna emateko baliabide gisa. TimeMLk ordura arteko markaketa-lengoaiekin alderatuta, desberdintasun batzuk ditu. TimeMLk aurreko markaketa-lengoaiei ezaugarriak hedatu zituen: i) TIMEX2 etiketaren atributu multzoa zabaldu zuen, TIMEX3 etiketa sortuz, ii) gertaeren sailkapena proposatu zuen, iii) denborazko funtzio berriak aurkeztu zituen (*azken astea* edo *bi eguneko* denbora-adierazpen erlatiboak tratatzea ahalbidetu zuen) eta iv) gertaeren eta denbora-adierazpenen arteko erlazioak sortu zituen.

Ondorioz, TimeMLk ondoko ezaugarri bereizgarriak lortu ditu: i) gertaerak denbora-adierazpenekin ainguratzen ditu, ii) gertaera-adierazpenak elkarren artean ordenatzen ditu, iii) testutik kanpoko etiketen sorrera ahalbidetzen du eta iv) denbora-adierazpen zehaztugabeen bi urratseko interpretazioa ahalbidetzen du (Pustejovsky *et al.*, 2003a).

TimeML hobetuz joan da informazioa gehituz, etiketatzea sinplifikatuz eta modelo unibertsalagoa sortuz. Bertsio berriak egin ahala, TimeML hobetu da eta azken bertsioa (TimeML Working Group, 2010) markaketa-lengoaia fidagarria bihurtu da denbora-informazioa etiketatzeke. Gainera, TimeML nazioarteko ISO arau bihurtu da (ISO-TimeML working group, 2008), eta horrek kalitate handia eta erabilera zabala egiaztatzen du.

Hain zuzen ere, TimeML hizkuntza askotarako egokitu da: frantsesa (Bittar, 2010), italiera (Caselli *et al.*, 2011), portugesa (Costa eta Branco, 2012), errumaniera (Forăscu eta Tufiş, 2012) eta koreera (Jeong *et al.*, 2015), eza gutzen ditugunen artean. Bertsio horiek TimeML eskemaren egitura mantentzen dute, baina aldaketa txikiak dituzte hizkuntzaren mendeko ezaugarri batzuk tratatzeko.

Erabilera orokorrerako TimeMLz gain, beste aldaketa eta bertsio batzuk ere proposatu dira; esaterako, TimeML-strict (Derczynski *et al.*, 2013). TimeML-strict TimeMLren denbora-informazioaren prozesamendurako azpimultzo praktikoa, ez-anbigua eta erraza da autoreen arabera. Eskema mugatuagoa denez, prozesatzeko errazagoa da eta denbora-informazioaren

prozesamenduan denbora gutxiago kontsumitzen da. Aldaketa nagusiak dira: i) dokumentuaren metadatuaren eta testu-gorputzaren arteko aldea argi adierazten da, ii) soilik testuko gertaeren eta denbora-adierazpenen arteko erlazioak sortzeko gaitasuna dago eta iii) denbora-erlazioen multzo murriztua da.

TimeML testuinguru zehatzetara ere egokitu da. Azken hamarkadan, txosten klinikoaren interpretazio automatikoa interes handiko gai bilakatu da. Oso arlo zehatza da eta erabiltzen den hizkuntza oso espezifiko: i) esaldiak ez dira beti hizkuntza naturalekoak, ii) denbora-erlazio inplizituak daude, iii) botiken izenek tratamendua adierazten dute, eta abar. Ezaugarri horiek testu klinikoak prozesamendurako gai interesgarri bihurtzen dituzte ez baitira aurretik landutako testu-motak bezalakoak. Desberdintasun horiek islatuta, ulermen hobe lortzen da. Hala, THYME-TimeML (Styler *et al.*, 2014) medikuntzaren arlora egokitu da, narrazio klinikoan denbora-informazioa etiketatzeko eta denbora-lerro klinikoak eskaintzeko:

- ISO-TimeMLko gertaerak etiketatzeko, *EVENT* etiketaren atributu batzuk domeinuaren beharrianetara egokitu dira.
- Denbora-adierazpenak etiketatzeko gidalerroak eta atributuak *ebakuntza aurreko* eta *ebakuntza ondoko* denbora-erreferentzien modukoak harrapatzeko moldatu dira.
- Halaber, denbora-erlazioak murriztu dira eta gertaerak gertatu diren unearen arabera ordenatzen dira.

Euskararako ere TimeMLren ildoari jarraitu diogu.

4.1.2 EusTimeMLn jarraitutako metodologia

Euskarazko denbora-informazioa prozesatu ahal izateko, hizkuntzaren azterketa egin ondoren, EusTimeMLren bidezko etiketatzea egiteko irizpideak definitu ditugu. Hain zuzen ere, EusTimeBank corpusa etiketatzeko, markaketa-gidalerroak definitu ditugu (Altuna *et al.*, 2014b, 2016a), EusTimeMLri jarraiki. Gidalerroen lehen bertsioan (Altuna *et al.*, 2014b) gertaerak, denbora-adierazpenak eta seinaleak etiketatzeko gidak eta oharra argitaratu ditugu. Bigarrenean (Altuna *et al.*, 2016a), lehen bertsioa gidalerro hobetuak eta erlazioak etiketatzeko gidalerroak argitaratu ditugu. Etiketatzea eta gidalerroen sorrera batera garatu dira; corpusak hizkuntzako fenomeno berri

ematen du eta horiek nola etiketatu erabaki behar da, eta erabakiek hizkuntzaren informazioa ondo islatzen duten ebaluatu behar da ondoren. Hala, corpusa etiketatzeko eta gidalerro horiek sortzeko urrats hauek egin ditugu:

- Behin-behineko gidalerroak idatzi ditugu corpusean eta gramatiketan oinarrituta. Zehazki, 3. kapituluak deskribatu bezala, euskarazko denbora-informazioa adierazten duten elementuak eta beren arteko loturak aztertu ditugu, eta informazio horretan oinarrituta EusTimeML markaketa-lengoaia definitu dugu.
- Etiketak eta atributuak nola eta zein testuingurutan esleitu deskribatu dugu, markaketa-lengoaia definitzeaz gain.
- Hainbat esperimentutan (ikus 4.5 atala) etiketatzailerik corpuseko lagin batzuk etiketatu dituzte eta sortutako zalantzak eta desadostasunak eztabaidatu dira definitutako markaketa-lengoiari jarraituta.
- Gidalerroen bertsio berria (behin betikoa) idatzi dugu gidalerroak ebaluatu ondoren. Horri jarraituta esperimentuetako corpusa zuzendu dugu eskuz. Corpus hori urre-patroitzat hartu da. Urre-patroi hori tresna automatikoak entrenatu eta ebaluatzeko erabili dugu gerora.

Amaitzeko, eskuzko etiketatzeari dagokionez, gogoeta orokor batzuk erauzi ditugu. Izan ere, eskuzko etiketatzeak etiketatzailerik lan eskergea eskatzen du:

- Gidalerro argiak sortu behar dira.
- Etiketatzailerik gidalerroak ondo barneratu behar dituzte.
- Etiketatzailerik tresna ondo ezagutu behar dute.
- Etiketatzailerik abiadura mugatua da.

Esperimentazio horren bidez, markaketa-lengoaia definitu dugu. Aipatu bezala, EusTimeML markaketa-lengoaia TimeMLren egokitzaia da. Euskararako markaketa-lengoaia definitzean, nahiko hertsiki jarraitu diogu TimeMLn proposatutako eskemari. Izan ere, denbora-informazioa etiketatzeke, hasierako etiketatze osotua eskaintzen baitu, gero, erraz zabal daitekeena. Etiketei dagokienez, TimeMLn proposatutako etiketak baliatu ditugu EusTimeMLn:

- EVENT: gertaerentzat
- TIMEX3: denbora-adierazpenentzat
- SIGNAL: seinaleentzat
- TLINK: denbora-erlazioentzat
- SLINK: mendekotasun-erlazioentzat
- ALINK: aspektu-erlazioentzat

Etiketa horietako bakoitzeko atributuak erabakitzeko, TimeMLren hainbat bertsiotan proposatutako atributuak aztertu ditugu eta euskarazko gertaeren, denbora-adierazpenen eta horien arteko erlazioen ezaugarriak ondoen adierazten dituztenak aukeratu ditugu. Atributu bakoitzari balio normalizatuak emateko, euskarazko ezaugarriak aztertu ditugu, bai gramatiketan, bai gure analisi-katearen (Ezeiza *et al.*, 1998; Alegria *et al.*, 2002) irteeran eta TimeMLn aurretik definituta zeudenekin adieraz zitezkeen aztertu dugu. Gehien bat, TimeMLko eskemetan orokortutako balioak baliatu ditugu, baina kasu batzuetan balio berriak sortu behar izan ditugu euskararen berezitasunak adierazteko. Jarraian ikusiko ditugu EusTimeMLn definitutako atributuak eta beren balioak.

4.2 Euskarazko denbora-egituren etiketatzea

Euskarazko denbora-egiturak EusTimeMLri jarraituta etiketatu ditugu. Atal honetan EusTimeMLko etiketen, atributuen eta horien balio garrantzitsuenak deskribatu ditugu eta eskema osoaren laburpena egin dugu B eranskinen. Atal honetan hartutako erabakiak irudikatzeko adibideak emango ditugu eta aurreko adibideetan egin bezala, atal honetako adibideetan ere nabarmendu nahi dugun egitura beltzez adieraziko dugu. Zehazki, etiketak hartuko dituzten tokenak baino ez ditugu nabarmenduko.

4.2.1 Gertaeren etiketatzea, EVENT

Arestian aipatu bezala, EusTimeMLn gertaerek EVENT etiketa hartzen dute. Etiketa horrek token bakarreko hedapena du eta gertaera-adierazpenaren

10026-First_A380_enters_commercial_service.txt

s0 Lehen A380a hasi da zerbitzu komertziala ematen .

s1 2007ko urriaren 17a .

Hegazkingintzaren historian , sekulako lorpena izan da : lehen Airbus A380a 18:40an (GMT + 8) lurreratu da Singapur-en , Changi nazioarteko aireportuan , Airbus-en bidalketa-zentrotik atera eta 12 orduko hegaldia egin ostean .

Hegazkinari 400 bat gonbidatuk egin zioten ongietorria laster zabalduko den s3 Changi Nazioarteko aireportuko 3. terminalean . Tartean zen Lee Hsien Loong , Singapur-eko lehen ministroa .

Hitzaldian , lehen ministroak Singapore Airlines (SIA) aire-linea hartzailea s4 txalotu zuen , eta aipatu zuen arrakasta Singapur-eko biztanle guztiak harro egoteko modukoa zela .

Gehitu zuen lurreratzea gertaera gogoangarria zela , ez bakarrik SIarentzat , s5 baita Changi nazioarteko aireportuarentzat ere , zeina herrialdeko aireporturik handiena baita .

Orain arteko terminal guztiak prest daude A380a hartzeko , eta 3. s6 terminalak (1,75 milioi dolar singapurtar kostatu da) datorren urteko urtarrian zabalduko ditu ateak .

4.1 irudia – Gertaeren etiketatze adibidea

buru lexikoari esleitzen zaio. Adibidez, 4.1. irudian ikus daitekeenez, aditz perifrastikoen kasuan aditz nagusiak hartzen du gertaera-etiketa (horiz).

Gertaeren informazio linguistikoa azaleratzeko atributu batzuk ere definitu ditugu. Hurrengo ataletan euskarazko gertaeren ezaugarriak (ikus 3.1.2 atala) adierazteko EusTimeMLn aukeratu ditugun atributuak, baita horiei esleitu dizkiegun balioak ere, deskribatuko ditugu.

4.2.1.1 Gramatika-kategoria, pos

Euskarazko gertaera-adierazpenak aztertzean ikusi dugunez, gertaerak euskaraz hainbat gramatika-kategoriaren bidez adieraz daitezke. Gramatika-kategoria horiek `pos` (*part-of-speech*) atributuaren bidez adierazi ditugu EusTimeMLn eta gramatika-kategoria bakoitzari esleitu dizkiogun balioak hauek dira:

- `VERB`: aditzez adierazitako gertaerentzat
- `NOUN`: izenez adierazitako gertaerentzat
- `ADJECTIVE`: adjektiboz adierazitako gertaerentzat
- `ADVERB`: adberbioz adierazitako gertaerentzat
- `PRONOUN`: izenordainez adierazitako gertaerentzat
- `OTHER`: beste gramatika-kategoria batzuen bidez adierazitako gertaerentzat

Gertaerek token bateko hedapena izanik, gertaeren buru lexikoak baino ez du etiketa hori hartuko eta gertaeraren gramatika-kategoria token horri dagokiona izango da.

4.2.1.2 Gertaera-klaseak, class

Euskarazko gertaeren ezaugarri linguistikoak azaltzean (3.1.2.1 atala) esan bezala, gertaerak sailkatzeko TimeMLn proposatutako sailkapenari jarraitu diogu. Sailkapen horretan, klasea adierazteko EusTimeMLn TimeMLtik hartu dugun `class` atributua baliatu dugu. Atributu horrek ondoko balioak hartzen ditu:

- `REPORTING`: gertaera diskurtsiboentzat
- `PERCEPTION`: hautematezko gertaerentzat
- `ASPECTUAL`: aspektuzko gertaerentzat
- `I_ACTION`: ekintza intentsionalentzat
- `I_STATE`: egoera intentsionalentzat

- STATE: egoerentzat edo zirkunstantzientzat
- OCCURRENCE: jazoerentzat

4.2.1.3 Tempusa, tense

Tempusa aditz jokatuak adierazten duen denborarekin lotutako informazioa da eta aditzaren eta hizketaren erreferentzia-unearen arteko erlazioa adierazten du. Tempusaren tratamendu automatikoari dagokionez, TimeMLren beste hizkuntzetarako bertsioetan, **tense** atributuaren bidez normalizatu da aditzen tempusa. Guk ere etiketa bera baliatu dugu eta ondoko balioak esleitu dizkiogu:

- PRESENT: orainaldiko tempusa duten aditz jokatzeko gertaerentzat
 - (237) Euskararen Legea aldatzea **onartu** du Nafarroako Parlamentuak.
- PAST: lehenaldiko tempusa duten aditz jokatzeko gertaerentzat
 - (238) Bart sutea **piztu** zen lantegi baten inguruan.
- ALEGIAZKOA: alegiazko tempusa duten aditz jokatzeko gertaerentzat
 - (239) Euskal artea munduratzeko sarea **osatuko** litzateke, akordioa **gauzatuko** balitz.
 - (240) Ondo **legoke** unibertsitatean bizikletak aparkatzeko leku estali gehiago jartzea.
- NONE tempus markarik ez duten gertaera-adierazpenentzat, zehazki, aditz ez-jokatuentzat (241) eta beste gramatika-kategorien bidez adierazitako gertaerentzat, esaterako, (242) adibideko izenarentzat.
 - (241) Bigarren seihilabetekoan irekiko du galeriak, baina txillidarrren jabetza **izaten** segituko du.
 - (242) Erakundeek oso berri ontzat jo dute Txillida Lekuren **irekiera**.

Balioak erabakitzeke, euskararen ezaugarriak izan ditugu kontuan, nagusiki alegiazko denbora —ez-oraina eta ez-lehenaldia— esplizituki adierazten

dela euskaraz, baita gertaera guztiak ez direla aditz jokatu bidez adierazten ere. Kasu horietan, tempusik eza **NONE** balioaren bidez adierazi dugu.

Tempusa etiketatze balio horiek erabakitzeko, euskarazko gramatikei (Altuna *et al.*, 1985, 1987) erreparatu diegu eta beste hizkuntzetan hartutako etiketatze-erabakiak aztertu ditugu. Hizkuntza horietan tempusa adierazteko hartu diren balioak aurkeztu ditugu 4.1. taulan eta jarraian, horien azalpena emango dugu.

4.1 taula – Erdarazko TimeML eskemetan aditz-tempusa (**tense**) etiketatze proposatutako balioak

Ingelesa	Gaztelania	Frantsesa	Italiera
PRESENT			
PAST			
FUTURE			
NONE			
INFINITIVE			
PRESPART (teaching)			
PASTPART (taught)			
	PRESENT	PRESENT	PRESENT
	PAST	PAST	PAST
	FUTURE	FUTURE	FUTURE
	NONE	NONE	NONE
		IMPERFECT	

Aditz-denborak, orokorrean, orainaldia (**PRESENT**), lehenaldia (**PAST**) eta etorkizuna (**FUTURE**) adierazten ditu aztertutako hizkuntzetan, 4.1. taulan ikus daitekeenez. Tempus balio eza (**NONE**) ere aukera daiteke TimeMLren hainbat bertsiotan aditzezkoak ez diren gertaerak etiketatze edo, aditzezkoak izanda ere, tempusik adierazten ez duten aditz-formak etiketatze.

Ingelesaren zutabearen, aditz ez-jokatuen formak (**INFINITIVE**, **PRESPART** eta **PASTPART**) ere adierazten dira **tense** atributuaren bidez. Aztertutako beste hizkuntzetarako, ordea, aditz-forma ez-jokatu horiek adierazteko, **vform** atributua sortu da. 4.2. taulan ikus daitezke gaztelaniako, frantseseko eta italierako aditz-oinak, gerundioak eta partizipioak adierazteko balioak.

EusTimeMLn etiketatzea errazteko, ingelesean egin bezala, aditz jokatu eta ez-jokatu balioak batera eman ditugu **tense** atributuaren bidez.

4.2.1.4 Aspektua, aspect

Aspektua aditz jokatuak duten informazioa da eta aditzek adierazten duten gertaeraren burututasunari buruzko informazioa ematen du. Aditz jokatuaren aspektua adierazteko, **aspect** etiketa definitu dugu EusTimeMLn. Hala ere,

4.2 taula – Erdarazko TimeML eskemetan aditz-forma (**vform**) etiketatzeko proposatutako balioak

Ingelesa	Gaztelania	Frantsesa	Italiera
	INFINITIVE GERUNDIVE PARTICIPLE NONE (aditz jokatuak edo beste POS)	INFINITIVE PASTPART PRESPART GERUNDIVE NONE	INFINITIVE GERUND PARTICIPLE NONE

kontuan izan dugu gertaera guztiak ez direla aditz jokatu bidez adierazten eta beste gertaera-adierazpenei **aspect** atributuan balio bat esleitzeko bidea ere sortu dugu. Euskarazko aspektua errepresentatzeko, lau balio definitu ditugu.

- **PERFECT**: aspektu burutua adierazteko

(243) Airbus-en lehen A380 superjumboa gaur **entregatu** zaio Singapuri Airlinesi.

- **-PERFECT**: aspektu ez-burutua adierazteko

(244) Indarrean **jarraitzen** dute aire-linea indonesiarrarentzako zigorrak.

- **FUTURE**: geroaren aspektu-marka duten aditzentzat

(245) Hirugarren terminalak datorren urteko urtarrilean **zabalduko** ditu ateak.

- **NONE**: aspektu-markarik gabeko formentzat; bai aditzoinaz adierazitako gertaerentzat (246), bai beste gramatika-kategorien bidez (izenak, esaterako) adierazitako gertaerentzat (247).

(246) Espetxeratzea bertan behera uzteko galdegin die Auzitegi Gorenari eta Auzitegi Konstituzionalari, Generalitateko presidente gisa jarduteko aukera **izan** dezan.

(247) Aurten, emakume guztiok **planto!**

4.3 taula – Erdarazko TimeML eskemetan aspektua etiketatzeko proposatutako balioak

Ingelesa	Gaztelania	Frantsesa	Italiera
PROGRESSIVE PERFECTIVE PERFECTIVE_PROGRESSIVE NONE	PERFECTIVE IMPERFECTIVE IMPERFECTIVE_PROGRESSIVE PERFECTIVE_PROGRESSIVE	PROGRESSIVE PERFECTIVE PERFECTIVE_PROGRESSIVE PROSPECTIVE NONE	PROGRESSIVE PERFECTIVE IMPERFECTIVE NONE

Etiketa horren balioak definitzeko, alde batetik, euskarazko aspektuak adierazten duen informazioa aztertu dugu eta, bestetik, beste hizkuntzetako TimeML eskemetan definitutako balioak aztertu ditugu.

Aztertutako hizkuntzetan (ingelesa, gaztelania, frantsesa eta italiera), aspektu burutua (*perfective*) eta burutugabea (*imperfective*) adierazteko balioak kodetu dira, 4.3 taulan ikus daitekeenez. Aspektu progresiboa ere agertzen da aztertutako lau eskemetan. Aspektu horren bidez gaztelaniazko *estar + gerundio* modukoak markatzen dira.

Beste hizkuntzetako aspektuarentzako eskemekin konparatuta, euskarazko aspektuaren etiketatzeak bi berezitasun ditu. Alde batetik, euskaraz aspektu progresiborik ez dugu esplizituki markatzen; izan ere, aspektua aztertzean ikusi dugunez, euskarazko gramatiketan tradizionaliki aspektu ezburutuaren barruan hartu da progresibotasuna. Beste aldetik, etorkizuna adierazteko balioa definitu dugu; aztertutako hizkuntzetan ez bezala, euskaraz etorkizuna aspektuaren bidez adierazten baita.

4.2.1.5 Modalitatea, modality

Modalitatea kodetzeko, **modality** atributua baliatu dugu EusTimeMLn eta predikatu modalentzat hiru modalitate-balio definitu ditugu: **AHAL**, **BEHAR** eta **NAHI**. Modalitate markarik ez duten egiturentzat **NONE** balioa ere definitu dugu. Ondoko (248–250) adibideetan forma modala beltzez eta bere argumentua kurtzibaz nabarmendu ditugu.

- (248) Sindikatuak laguntza eskatu dute ahal bezain laster lanean *hasi* **ahal** izateko.
- (249) Ordurako epaileak salmentari baietza edo ezetza *eman* **behar** izango dio.
- (250) Aljeriarrek egozketa adarra eta etxe-tresna txikiak *erosi* **nahi** dute.

4.4 taula – Beste hizkuntzetako TimeMLn modalitatea etiketatzeko hartutako erabakiak

	Ingelesa	Gaztelania/ Katalana	Frantsesa	Italiera	Estoniera
Forma modalak gertaeratzat hartu	EZ	BAI	BAI MODAL klasea	BAI	BAI MODAL klasea
Modalitate balioak 1. zeren bidez? 2. non adierazi?	1. Forma modala 2. Aditz nagusian	1. Ez markatu 2. mendekotasuna SLINK bidez	NECESSITY POSSIBILITY OBLIGATION PERMISION	1. Forma modala 2. Modalean eta aditz nagusian	1. Forma modala 2. Aditz nagusian

Ahalerazko laguntzailea duten aditzen kasuan, aditz nagusia etiketatu dugu, hori baita modalitatearen adierazlea (251).

(251) Kontratuaren balioa 100.000 milioi dolarrera **igo** dezakete.

Euskarazko predikatu modalak gertaeratzat hartu ditugu eta etiketatu ditugu. Gertaera diren aldetik, klasea ere esleitu diegu eta beste gertaera bat argumentutzat hartzen duten gertaera estatikoak direnez, egoera intensionalen (I_STATE) multzoan sailkatu ditugu. Modalitate-informazioa nola kodetu erabakitzeko, hainbat hizkuntzako TimeML eskemak aztertu ditugu. Ingeleserako TimeML eskeman (TimeML Working Group, 2010), modalitatea karga semantikoa (*stand*) duen aditzean adierazten da eta forma modala (*may*) ez da gertaeratzat hartzen (252).

(252) I *may stand* before you in six months and say 'hey I was wrong'.

Aztertutako beste hizkuntzetan, aldiz, forma modalak gertaeratzat hartzen dira eta hala etiketatzen dira. Nabarmentzekoa da estonierarako (Orasmaa, 2016), bai forma modalari, bai bere mendekoari esleitzen zaiela balio modala gertaera modalaren eragina esplizitu egiteko. Guk ere hala egin dugu euskararako (248–250).

Balioei dagokienez, frantseserako (Bittar, 2010) modalitatea adierazteko balio zehatzak definitu dira. Ingeleserako (TimeML Working Group, 2010), italierarako (Caselli *et al.*, 2011) eta estonierarako (Orasmaa, 2016), ordea, modalitatea adierazten duen forma da **modality** atributuaren balioa. Azkenik, gaztelania eta katalanerako (Saurí *et al.*, 2009), ez da **modality** atributurik definitu eta mendekotasun-erlazio bidez adierazten da modalitatea. 4.4 taulan laburbildu ditugu aztertutako hizkuntzetan modalitateari egiten zaion tratamendua.

4.2.1.6 Polaritatea, polarity

Polaritatea kodetzeko, **polarity** atributua definitu dugu eta bi balio eman dizkiogu: POS polaritate positiboa duten gertaerentzat (253) eta NEG polaritate negatiboa dutenentzat (254).

(253) Aljeriarrek egozketa adarra eta etxe-tresna txikiena **erosi nahi** dute.

(254) Garraio-ministroak hegazkin europarrak ez **erosteko** eskatu die aire-lineei.

Orasmaak (2016) zehazten duenez, polaritatea gramatika mailan tratatzen da TimeML eskemetan; bestela esan, gertaerak gramatika mailan ezezka agertzen diren edo ez hartzen da kontuan. Hau da, ez da ezeztapen lexikoa (*ukatu*, *baztertu*) kontuan hartzen. EusTimeMLn ere, gramatika mailako tratamendua eman diogu polaritateari.

4.2.1.7 Ziurtasuna, certainty

Aurreko kapituluan aipatu dugunez, TimeMLn aztertutako gertaeren ezauzgarriez gain, faktualitateari dagozkionak ere aztertu ditugu. Horiek normalizatzeko, EusTimeMLn **certainty**, **factuality** eta **specialCases** atributuak definitu ditugu. Jarraian, horiek deskribatuko ditugu eta beren balio normalizatuak aurkeztuko ditugu.

Hiztunak gertaera bat gertatu delako duen ziurtasuna edo konpromisoa **certainty** atributuaren bidez kodetu dugu. EusTimeMLko **certainty** atributuari hiru balio posible esleitu dizkiogu:

- **CERTAIN**: igorlea gertaera batekiko ziur dagoenerako.

(255) Boeing-ek 11.000 milioi dolar **lortu** ditu akordioetan.

- **UNCERTAIN**: igorleak gertaera batekiko ziurtasun eza adierazten duenerako.

(256) Hegazkin-merkatuaren kontrola **galdu** ote zuen eztabaida piztu zen.

- **UNDERSPECIFIED**: hizlariaren ziurtasuna erabaki ezin denerako.

(257) Espetxeratzea bertan behera **uzteko** galdegina die Auzitegi Gorenari eta Auzitegi Konstituzionalari.

4.5 taula – Beste etiketatze-saiakeretan ziurtasuna etiketatzeko proposatutako balioak

FactBank (Saurí, 2008)	van Son <i>et al.</i> (2014)	Minard <i>et al.</i> (2014)
Certain	Certain	Certain
Possible Probable	Uncertain	Non certain
Underspecified	Other	Underspecified

Beste etiketatze-saiakera batzuetan, ziurtasuna etiketatzeko erabakiak aztertutitu euskarazko ziurtasuna nola kodetu erabakitzeke. Zehazki, igorleak gertaerarekin duen konpromisoa eta ziurtasun edo ziurtasun eza kodeatzeko eskemak proposatu dira. 4.5. taulan ikusten denez, ziurtasuna, ziurgabetasuna eta ezin definitutako ziurtasuna dira balio nagusiak aztertutako hiru eskemetan. Guk ere joera berari jarraitu diogu euskararako.

4.2.1.8 Faktualitatea, factuality

EusTimeMLren barruan **factuality** atributua definitu dugu faktualitatea islatzeko, eta bost faktualitate-balio definitu ditugu:

- **FACT**: gertatu diren gertaerak

(258) Orain arteko terminal guztiak **prest daude** A380a hartzeko.
- **COUNTERFACT**: lehenaldian gertatu ez diren gertaerak

(259) Ez zuen 12 orduko hegaldirik **egin**.
- **NON_FACTUAL**: etorkizunean gertatuko diren gertaerak

(260) 3. terminalak datorren urteko urtarrilean **zabalduko** ditu atearak.
- **NO_FACTUALITY_VALUE**: gertaera berezirik adierazten ez duten gertaerak

(261) **Irakurtzea** gustatzen zait.
- **UNDERSPECIFIED**: faktualitate baliorik ezin esleitu dakiekeen gertaerak

(262) Orain arteko terminal guztiak prest daude A380a **hartzeko**.

Beste hizkuntzetarako TimeML eskemetan ziurtasuna eta faktualitatea tratatu ez badira ere, TimeML bidezko etiketatzeak (edo antzekoak) oinarritzat hartzen dituzten ziurtasuna eta faktualitatea etiketatzeko proposamenak egin dira 4.6. taulan¹ ikus daitekeenez.

4.6 taula – Faktualitatea ebazteko ezaugarriak hainbat faktualitatearen markaketa-eskematari

Ezaugarriak	FactBank (Saurí, 2008)	SIBILA (Wonsever <i>et al.</i> , 2012)	van Son <i>et al.</i> (2014)	NewsReader (Minard <i>et al.</i> , 2014)
Polaritatea	✓	(✓)	✓	✓
Ziurtasuna	✓	(✓)	✓	✓
Denbora-info.	(✓)	(✓)	✓	✓
Kasu bereziak				✓
Faktualitatea	Factual Probable Possible Counterfactual Not probable Not possible Certain but unknown output Unknown or uncommitted	Yes No Programmed_future Negated_future Possible Indefinite	Fact Counterfact Possibility (uncertain) Possibility (future)	Factual Counterfactual Non factual

Faktualitatea ebazteko kontuan hartutako ezaugarri linguistikoak (polaritatea, ziurtasuna, denbora-informazioa eta kasu berezien tratamendua) eta eskema bakoitzean definitutako faktualitate-balio posibleak agertzen dira 4.6. taulan. Ikus daitekeenez, eskema guztietan faktualitate positiboa (FACT, FACTUAL edo YES) eta faktualitate negatiboa (COUNTERFACT, COUNTERFACTUAL edo NO) agertzen dira, baita tarteko faktualitate-balioak (probabilitatea edo ziurtasun eza) ere.

Euskarazko faktualitatearen markaketa-eskema sortzeko, gehienbat Minard *et al.*-ek (2014) proposatutako eskeman oinarritu gara, informazio gutxi galduta sinpletasun handiena eskaintzen duena baita. Aldaketa bat egin diogu, ordea: NO_FACTUALITY_VALUE balioa gehitu diogu faktualitate baliorik aplikatu ezin denean, predikatu generikoen kasuan esaterako, erabiltzeko.

4.2.1.9 Kasu bereziak, specialCases

Faktualitatea aztertzean ikusi dugunez, testuinguru batzuetan gertaeren formek ez dute normalean adierazten duten faktualitate-informazioa adierazten. Horrelakoak dira baldintzazko testuinguruak eta adierazpen generikoak. Kasu berezi horiek etiketatzeko, **specialcases** atributua definitu dugu eta hiru balio esleitu dizkiogu:

¹Parentesi artean dauden ezaugarriak zeharka baliatu dira faktualitatea definitzeko.

- **CONDITONAL_CONDITION**: baldintzaren protasiko gertaerentzat

(263) Bilaketak fruiturik **ematen** ez badu, BEAk jarraitzeko dirua bilatuko du.

- **CONDITONAL_MAIN**: baldintzaren apodosiko gertaerentzat

(264) Bilaketak fruiturik ematen ez badu, BEAk jarraitzeko dirua **bi-latuko** du.

- **GENERIC_STATEMENT**: predikatu generikoentzat

(265) Airbus A320a korridore bakarreko hegazkina **da**.

Balio horiek Minard *et al.*-en (2014) lanetik hartu ditugu, faktualitatean egin bezala, baliozkoak baitira faktualitatearen ebazpenerako.

4.2.1.10 Gertaeren beste atributuak eta MAKEINSTANCE etiketa

Informazio linguistikoa adierazten duten atributuez gain, EusTimeMLk prozesamendurako atributuak ere eskaintzen ditu. Gertaeren kasuan, **eid** atributuak gertaeraren identifikatzailea adierazten du eta **eiid** atributuak gertaera-instantziaren identifikatzailea. Horiek gertaerak erlazioetan ainguratzeko baliatzen dira nagusiki.

Gainera, gertaerek prozesamendurako MAKEINSTANCE etiketa hartzen dute. Etiketa horrek tokenen eta gertaeren gauzatzea bereiz etiketatzea du helburu (Saurí *et al.*, 2006). Hau da, token batek gertaera baten gauzatze bat baino gehiago adieraz ditzake, eta horiek kronologiako bi gertaera indibidual bezala hartu behar dira. Adibidez, (266) adibidean, *gimnasia joatea* behin aipatzen da, baina bi memento ezberdinetan gertatzen da, *astearteetan* eta *ostegunetan*. Ondorioz, gertaera horrek bi instantzia adierazten ditu.

(266) Gimnasia **joaten** naiz astearte eta ostegunetan.

Hala eta guztiz ere, guk (266) *joaten* tokenari MAKEINSTANCE etiketa bakarra esleitzen diogu, gertaeraren *astearte* eta *osteguneko* iterazioei erreferentzia egiten badie ere, ez baititugu horrelako kasuak sakontasunean tratatu.

Atributuen eta balioen zerrenda osoa B.1 atalean aurkeztu dugu.

4.2.2 Denbora-adierazpenen etiketatzea, TIMEX3

Denbora-adierazpenek **TIMEX3** etiketa hartzen dute EusTimeMLn. Etiketaren hedapenari dagokionez, gertaerak ez bezala, denbora-adierazpenetako token guztiak sartzen ditugu etiketaren barruan, beste hizkuntzetarako TimeML eskemetan egin den bezala. Esaterako, (267) adibideko *atzo* adberbioa bere horretan agertzen da. Beste kasu batzuetan, ordea, denbora adierazten duen formaz gain, denbora-erlazioa adierazten duen postposizio-atzizkia agertzen da. Adibidez, (268) adibideko *goizean* tokenean, *goiz* denborazko abiarazle-lexikoa eta *-ean* inesibozko postposizioa agertzen dira. Hala eta guztiz ere, etiketa bera hartzen dute biek eta ez da tokena moztu.

(267) **Atzo** likidazio plana aurkeztu zuten.

(268) Bazkide kooperatibistek batzarra egin dute **goizean**.

Ondorioz, etiketatze-lana token mailan egin dugunez, kasu batzuetan denbora-unea edo iraupena adierazten duen hitz-zatiaz gain, beste informazio motaren bat ere etiketaren barruan jaso dugu.

Postposizio-sintagma bidez adierazitako denbora-adierazpenak nola etiketatatu erabakitzeke, antzerako inflexioa duen beste hizkuntza bateko TimeML eskemari erreparatu diogu, hain zuzen ere, koreerakoari (Im *et al.*, 2009). Koreerak euskararekin bi ezaugarri nagusi partekatzen ditu: i) flexio handiko hizkuntza da eta ii) alfabetoko hizki bakoitzak hizkuntzako soinu bat adierazten du. Ondorioz, pentsatzekoa da denbora-adierazpenak etiketatzeke, euskaraz bezala, postposizioak denbora-adierazpenetan tratatzeko moduz zalantza berak izatea. Aitzitik, euskaraz ez bezala, koreerazko flexioa denbora-adierazpenaren etiketatik kanpo utzi dute (Jang *et al.*, 2004). Euskararako kontrako erabakia hartu da, hein batean CAT etiketatze-tresnak (4.4 atala) token mailako etiketatzea soilik ahalbidetzen duelako.

Denbora-adierazpenak etiketatzeke bi ezaugarri nagusi definitu ditugu EusTimeMLn: mota (4.2.2.1 atala) eta balioa (4.2.2.2). Horiez gain, informazio gehigarria adierazten duten beste ezaugarri batzuk ere identifikatu ditugu. Hurrengo ataletan deskribatuko ditugu atributu horiek eta beren balioak.

4.2.2.1 Denbora-adierazpen motak, type

Denbora-adierazpenak sailkatzean esan dugu gaur egun TimeMLn proposatutako lau kategoriako sailkapena dela orokortuena. Euskararako sailkapen

horri jarraitu diogunez, TimeMLko balioak baliatu ditugu kategoria horiek kodetzeko. EusTimeMLn mota `type` atributuaren bidez adierazten da eta ondoko lau balioak definitu ditugu lau denbora-adierazpen motentzat:

- DATE: datentzat
- TIME: orduentzat
- DURATION: iraupenentzat
- SET: errepikapenentzat

4.2.2.2 Balio normalizatua, `value`

Denbora-adierazpenek kronologiako une edo tarte bati erreferentzia egiten diote. Batzuek kronologiako puntu zehatz bati egiten diote erreferentzia, baina beste batzuek erreferentzia lausoa baino ez dute adierazten. Hala ere, denei eman dakieke balio normalizatu bat. EusTimeMLn balio normalizatua adierazteko, `value` atributua definitu da. Balio normalizatuen bidez anbigutasunik gabe adierazten da zein uneri edo denbora-tarteri egiten zaion erreferentzia. Formatuari dagokionez, EusTimeMLn balio normalizatu horiek ISO-8601 arauari (ISO Committee, 1997) jarraitzen diote eta denbora-informazioaren prozesamenduan nagusiki baliatu direnak dira.

Datak adierazteko era normalizatua (edo normalizatu antzekoa) dela modurik argiena irudikatzen da, umorez, 4.2. irudian.

Atributu horrek jarraian adierazten diren balio normalizatuak hartuko ditu.

Datak Datek YYYY-MM-DD balio normalizatua hartuko dute normalean; hau da, urtea lau zifraz adierazita, hilabetea bi zifraz eta eguna bi zifraz, eta unitate bakoitza gidoiz bereizita (269). Datako elementu guztiek ez dute zertan agertu: urteari soilik egin dakioke erreferentzia (270) edo hilabeteari (271). Kasu horietan, YYYY eta YYYY-MM formatuak erabiliko dira hurrenez hurren. Azkenik, gerta daiteke denbora-adierazpen erlatibo baten balio zehatza ezin eman izatea. Kasu horietan, X hizkiaz ordezkatu dira ezagutzen ez diren zifrak betiere YYYY-MM-DD formatua mantenduz (272). Unitateetako bat ezagutuz gero, ez dira unitate txikiagoak X hizkiaz ordezkatu. Hala, (271) adibideko *datorren urteko urriaren* denbora-adierazpenak ez du inoiz 2010-10-XX balio normalizatua hartuko, esaterako.



4.2 irudia – - Zer da zuretzat data/zita perfektu bat? + EE/HH/UUUU, besteak benetan nahasgarriak izan daitezke.

Atal honetako adibideetan, beltzez adierazitako denbora-adierazpenaren balio normalizatua parentesi artean adierazi dugu.

- (269) Autogile estatubatuarrak **ostiralean** (2009-05-15) jakinarazi zuen 1000 kontzesionario ixteko asmoa duela.
- (270) Zifra hori 3.605era jaisteko asmoa du **2010erako** (2010).
- (271) **Datorren urteko urriaren** (2010-10) ondoren frankizia berrituko ez duten kontzesionarioek gutuna jasoko dute.
- (272) **Hilabete haren** (XXXX-10) ondoren frankizia berrituko ez duten kontzesionarioek gutuna jasoko dute.

YYYY-MM-DD formatuari jarraituta normalizatu ezin diren denbora-adierazpenak ere badaude. Alde batetik, asteen arabera denbora-adierazpenak daude (273). Hilabete eta egunen arabera formatuan ezin da astea unitateztat tratatu eta arazoa saihesteko YYYY-WXX formatu normalizatua erabiltzen da. Formatu horretan, urtea lau zifratu adierazten da eta ondoren astea bi zifratu W astearen identifikatzailea aurretik duela. Hala, (273) adibideko *aste honetan* denbora-adierazpenak 2013ko 42. asteari egiten dio erreferentzia.

Asteburuak adierazteko ere formatu horretan oinarrituko gara eta YYYY-WXX-WE balio normalizatua esleituko diogu asteburuari (274). Izan ere, WE ingelesezko *weekenden* (“asteburua”) laburdura da.

(273) Enpresa bigarren erreskatearen zain egon da **aste honetan** (2013-W42).

(274) **Asteburuan** (2008-W38-WE) Bank of America negoziatutik aterazena.

Beste aldetik, urtearen atalak, urtaroak zein hiruhileko eta seiuhilekoak ezin adieraz daitezke YYYY-MM-DD formatuaren bidez, beren unitate izaera islatu nahi izanez gero. Hala, hiruhilekoak adierazteko Q1 (275), Q2, Q3, eta Q4 balioak baliatuko ditugu eta seiuhilekoak adierazteko, H1 eta H2 (276). Urtaroei dagokienez, urtaro bakoitza normalizatzeko balioak ere definitu ditugu: udaberriarentzat SP (277), udarentzat SU, udazkenarentzat AU eta neguarentzat WI.

(275) Konpainiak 6.000 milioi dolar galdu ditu **2009ko lehen hiruhilekoan** (2009-Q1).

(276) Lehia handia izango du Holden Commodorearen aurka, zeina **aurtengo urtearen bigarren erdian** (2006-H2) aterako baita.

(277) Goizegi da, baina **udaberria** (2009-SP) itxaropentsu hasi dugu.

Era berean, aurretik azaldutako formatuen bidez normalizatu ezin diren denbora-adierazpen lausoak daude. Horrelakoak dira (278) adibideko *orain*, *lehen* edo *etorkizunean*. Horiek PRESENT_REF, PAST_REF eta FUTURE_REF balioak hartuko dituzte, hurrenez hurren.

(278) Prozesu horretan ez dagoela traba berezirik berretsi du epaileak **orain** (PRESENT_REF).

Azkenik, urtea baino handiagoak diren unitateak aipatu behar ditugu: milurtekoak (279), mendeak (280) eta hamarkadak (281). Horiek normalizatzeko Y, YY eta YYYY formatuak erabiliko ditugu hurrenez hurren. Bestela

esan, milurtekoetan milakoen zifra emango dugu, mendeetan milakoena eta ehunekoena eta hamarkadetan milakoena, ehunekoena eta hamarrekoea.

- (279) **Milurtekoaren** (2) Garapenerako Helburuak ez dira bete.
- (280) Amerikako Iraultza **XVIII. mendean** (17) izandako asaldura politikoa izan zen.
- (281) **1980ko hamarkadaren amaieran** (198), Apple Computer-ek audioa grabatzeko aukera gehitu zien MacIntosh-ei.

Orduak Orduak berez eguna baino granularitate txikiagoa duten unitateak dira eta, ondorioz, balio normalizatu zehatzagoa izango dute. ISO-8601 erregelaren arabera, YYYY-MM-DDTHH:MM:SS formatua hartuko dute, hori da datentzat formatu normalizatua, orduaren formatua markatzen duen T marka eta orduak, minutuak eta segundoak bi zifraz eta bi puntuz bereizita (282). Datekin gertatzen den bezala, ordu erlatibo bati balio zehatza esleitu ezin zaionean, desanbiguatu ezinezko zifrak X bidez ordezkatzeko dira (283) eta, unitate bat ezagututa, txikiagoak ez badira ezagutzen ez dira X bidez ordezkatzeko (284).

- (282) **2014-03-07 - 23:33:32** (2014-03-07T23:33:32).
- (283) Lehen Airbus A380a **18:40an** (XXXX-XX-XXT18:40) hurreratu da Singapuren.
- (284) Akzio nagusien indizea 4.211 puntutan zegoen **hiruretan** (XXXX-XX-XXT15:00).

Egunaren atalak ezin dira YYYY-MM-DDTHH:MM:SS formatuaren bidez normalizatu ez baitira orduen arabera, granularitate handiagoa dute. Horiek adierazteko YYYY-MM-DDTAA formatua erabili dugu. (285) adibideko *gaur goizean* denbora-adierazpenari 2013-11-04TMO esleitu diogu: data osoa, T orduaren marka eta MO goizaren balio normalizatua.

- (285) **Gaur goizean** bi orduko bilera egin dute. (2013-11-04TMO)

Egunaren ataletarako ondoko balioak definitu ditugu:

- MO goizarentzat

- (286) **Gaur goizean** (2013-11-04TMO) bi orduko bilera egin dute.

- AF eguerdiarentzat (12:00–16:00)

(287) **Eguerdian** (2013-11-04TAF) jende asko hurbildu da merkatu-
ra.

- EV arratsaldearentzat

(288) **Arratsaldeko** (2013-11-04TEV) txandetan 5.000 kg elikagai
batu dira.

- NI gauarentzat

(289) **Astelehenetik astearterako gauean** (2013-11-04TNI) me-
tro koadroko 100 litro bota zituen Euskal Herriko hainbat toki-
tan.

Iraupenak Iraupenen balio normalizatuak beti izango du P (*period*) adie-
razlea iraupen baten balio normalizatua dela adierazteko. Eguna baino gra-
nularitate handiagoa duten iraupenetan, iraupenaren balioa denbora-unita-
tearen kuantifikatzailea eta denbora-unitatearen hizkia adierazten dira. Adi-
bidez, (290) adibideko *lau hilabeteko epea* denbora-adierazpenak P4M balio
normalizatua hartuko du: P iraupena delako, 4 lau hilabete direlako eta M
iraupenaren unitatea hilabeteak (*month*) direlako.

(290) Enpresak **lau hilabeteko epea** (P4M) izango du sindikatuekin akor-
diora erdiesteko.

Iraupenaren granularitatea eguna baino txikiagoa bada, ordea, balio nor-
malizatuaren formatua ondokoa izango da: P iraupena dela adierazteko, T
eguna baino txikiagoa dela adierazteko, unitatearen kuantifikatzailea zenba-
kiz eta unitatearen hizkia. Hala, (291) adibideko *12 orduko* iraupenak PT12H
balio normalizatua hartuko du. P-k iraupena dela adierazten du eta T-ren
bidez, eguna baino unitate txikiagoa duela adierazten da, 12ak hamabi ordu
adierazten ditu eta, azkenik, H orduaren unitateak unitatea adierazten du.

(291) Lehen Airbus A380a Singapur-en hurreratu da **12 orduko** (PT12H)
hegaldia egin ostean.

Denbora-unitateen eta beren identifikazio-hizkien zerrenda aurkeztu du-
gu 4.7 taulan. Ikus daitekeenez, hilabeteek eta minutuek M hizkia (*month*,

minute) hartzen dute, baina minutuko granularitatea duen iraupen batek beti T marka hartzen duenez, ez da anbiguotasunik izango. Ondorioz, (292) adibideko *lau hilabeteko epea* denbora-adierazpenak P4M balioa hartuko du eta (293) adibideko *lau minutuko* denbora-adierazpenak, PT4M.

4.7 taula – Denbora-unitateak eta beren adierazpideak

Unitatea	Hizkia
Mendea	C
Hamarkada	DE
Urtea	Y
Hilabetea	M
Astea	W
Eguna	D
Ordua	H
Minutua	M
Segundoa	S

(292) Enpresak **lau hilabeteko epea** (P4M) izango du sindikatuarekin akordioa erdiesteko.

(293) Trena **lau minutuko** (PT4M) atzerapenarekin ibiliko da egun osoan.

Errepikapenak Errepikapenak denbora-tarte batean gertatzen diren iterazioak dira, 3.2.3.1 atalean esan bezala. Bi motatakoak izan daitezke: i) denbora-tarte zehatz batean iterazio kopuru zehatza adierazten dutenak eta ii) denbora-tarte zehaztugabearen iterazioaren kokapen kronologikoa adierazten dutenak.

Lehen multzokoek, denbora-tarte mugatuari erreferentzia egiten diotenez, iraupenen antzerako balio normalizatuak izango dituzte. Zehatzago esanda, mota horretako errepikapenek **value** atributuan iraupenen balio berdinak izango dituzte, baina hori ez da nahikoa, kuantifikatzaileak eta maiztasun adierazleak ere behar dira. Horiek **freq** eta **quant** atributuen bidez (ikus 4.2.2.4. atala) adieraziko dira.

Adibidez, (294) adibideko *bi astean behin* denbora adierazpenak P2W bali-no normalizatua hartuko du, bi asteko (2W) denbora-tartea (P) hartzen duelako.

(294) Akusatua **bi astean behin** (P2W) azaldu beharko da epaitegian.

Bigarren multzokoek iterazioaren kokapen kronologikoa dutenez informazioarik garrantzitsuena, balio normalizatuan data (295) edo orduaren (296) formatua izango dute.

- (295) Akusatua **ostiralero** (XXXX-XX-WXX-5) azaldu beharko da epaitegian.
- (296) **Segundo bakoitzeko** (XXXX-XX-XXTXX:XX:XX) sei ume jaiotzen dira munduan.

4.2.2.3 Iraupenen hasiera- eta amaiera-puntuak, `beginPoint` eta `endPoint`

Iraupenak denbora-tarte mugatuak direnez, hasiera- eta amaiera-puntuak dituzte. Horiek adierazteko, EustimeMLn `beginPoint` eta `endPoint` atributuak definitu dira hurrenez hurren. Atributu horiek iraupenaren hasiera- eta amaiera-uneen balio normalizatua adierazten dute, betiere ISO-8601 formatuan. Adibidez, badakigu (297) adibideko *12 orduko* iraupena noiz amaitzen den; hain zuzen ere, *18:40an*, testuan esplizituki agertzen baita. Horrez gain, iraupenaren hasiera-unea ere identifika dezakegu iraupena eta amaiera-unea ezagututa, kasu honetan, *6:40*, hegaldia 12 ordukoa baita.

Hala, (297) adibideko **12 orduko** iraupena horrela adieraziko dugu:

- (297) Lehen Airbus A380a Singapur-en 18:40an lurreratu da **12 orduko** hegaldia egin ostean.
`value="PT12H"`
`beginPoint="2007-10-17T06:40"`
`endPoint="2007-10-17T18:40"`

Beti ez da posible iraupen baten hasiera- eta amaiera-uneak identifikatzea, ezin baita iraupena kronologian kokatu. Posible izanez gero, ordea, bi atributu horien balioak beteko dira.

4.2.2.4 Errepikapenen atributu bereziak: kuantifikatzaileak eta frekuentzia-adierazleak, `freq` eta `quant`

Ikusi dugunez, errepikapenen kasuan, balio normalizatua ez da nahikoa denbora-adierazpenak adierazten duen informazio guztia azalarazteko. Errepikapenetan, iterazioen tarteaz gain, iterazioen maiztasuna edo kopurua ere adierazi behar da.

Errepikapenak adierazten duen maiztasuna—denbora-tarte zehaztu bateko iterazio kopurua—esplizitu egiten da **freq** atributuaren bidez . Horren bidez, zenbakizko balio bat eta denbora-granularitate bat baliatuta, denbora-egitura baten gauzatze-frekuentzia adierazten da. (294) adibideko *bi asteen behin* denbora-adierazpena errepikapenari buruzko informazio osoaz agertzen da: **value** atributuaren bidez errepikapenaren denbora-tartea normalizatzen da. Denbora-tarte bat denez, iraupenen formatua du.

- (298) Akusatua **bi asteen behin** azaldu beharko da epaitegian.
type="SET"
value="P2W"
freq="1"

quant atributuak, berriz, gertaera bat tarte zehaztugabe batean zer une kronologikotan errepikatzen den adierazten du. Bestela esan, gertaera zenbatero errepikatzen den esplizitu egiten du. Hiru balio nagusi har ditzake: **EVERY** (guztiak) (299), **EACH** (bakoitz) (300) edo **SOME** (batzuk) (301), nahiz eta beste batzuk ere har ditzakeen, esaterako (302) adibideko **FEW** (gutxi batzuk).

- (299) Akusatua **ostiralero** azaldu beharko da epaitegian.
type="SET"
value="XXXX-XX-WXX-5"
quant="EVERY"
- (300) **Segundo bakoitzeko** sei ume jaiotzen dira munduan.
type="SET"
value="XXXX-XX-XXTXX:XX:XX"
quant="EACH"
- (301) Arazoak egoten dira **ostiral eguerdi batzuetan** Bilbora San Mamesetik sartzeko.
type="SET"
value="XXXX-XX-WXX-5TAF"
quant="SOME"
- (302) Arazoak egoten dira **ostiral gutxi batzuetan** Bilbora San Mamesetik sartzeko.
type="SET"
value="XXXX-XX-WXX-5"
quant="FEW"

4.2.2.5 Modifikatzaileak, mod

Denbora-adierazpenen balio normalizatua definitzeko ISO-8601 eskemari jarraitu badiogu ere, eskema horrek ezin ditu denbora-adierazpen batzuen ñabardurak islatu. Esaterako, (303) eta (304) adibideetako denbora-adierazpenek ISO-8601 eskemaren arabera balio normalizatu bera hartzen dute.

(303) **Gaur goizean** bi bilera izan ditut.

value='2018-03-20TMO'

(304) **Goizaren amaieran** neka-neka eginda nengoen hainbeste bileraren ostean.

value='2018-03-20TMO'

mod='END'

(304) adibideko *goizaren amaieran* denbora-adierazpenak, ordea, informazio gehigarria du: goizaren une zehatz bati, kasu horretan amaierari, egiten dio erreferentzia. Informazio hori irudikatu ahal izateko denbora-adierazpenen “modifikatzaileak” kodetzeko balioak definitu ditugu:

- LESS_THAN: iraupen adierazpenak adierazten duen balioa murrizteko

(305) **Hamarkada bat baino gutxiago** behar izan zuten eraikin hura altxatzeko.

type="DURATION"

value="P1DE"

mod="LESS_THAN"

- MORE_THAN: iraupen-adierazpenak adierazten duen balioa handitzeko

(306) **Hamarkada bat baino gehiago** behar izan zuten eraikin hura altxatzeko.

type="DURATION"

value="P1DE"

mod="MORE_THAN"

- EQUAL_OR_LESS: iraupen-adierazpenak adierazten duen balioaren berdina edo txikiagoa adierazteko

(307) **Hamarkada bat edo gutxiago** behar izan zuten eraikin hura altxatzeko.

type="DURATION"
value="P1DE"
mod="EQUAL_OR_LESS"

- EQUAL_OR_MORE: iraupen-adierazpenak adierazten duen balioaren berdina edo handiagoa adierazteko

(308) **Hamarkada bat edo gehiago** behar izan zuten eraikin hura altxatzeko.
type="DURATION"
value="P1DE"
mod="EQUAL_OR_MORE"

- START: denbora-adierazpenak adierazten duen unearen edo denbora-tartearen hasiera adierazteko

(309) **Hilabete *hasieran*** langabezia % 25ekoa zen.
type="DATE"
value="2014-03"
mod="START"

- MID: denbora-adierazpenak adierazten duen unearen edo denbora-tartearen erdialdea adierazteko

(310) **Hilabete *erdialdean*** langabezia % 25ekoa zen.
type="DATE"
value="2014-03"
mod="MID"

- END: denbora-adierazpenak adierazten duen unearen edo denbora-tartearen amaiera adierazteko

(311) **Hilabete *amaieran*** langabezia % 25ekoa zen.
type="DATE"
value="2014-03"
mod="END"

- APPROX: gutxi gorabeherako balioak adierazteko

- (312) Filma **hamarrak inguruan** amaitu zen.
type="TIME"
value="XXXX-XX-XXT22H"
mod="APPROX"

4.2.2.6 Funtzioa dokumentuan, `functionInDocument`

Testuetan bi motatako denbora-adierazpenak daude, i) soilik testu barruan esanahia dutenak eta ii) testuaren ulermeneari funtzio berezia betetzen dutenak. Lehenak testu bakoitzaren testuinguruan baino ezin dira ulertu. Bigarrenak testu barruko denbora-adierazpenak ulertzeko eta normalizatzeko informazioa ematen dute. Bestela esan, bigarren multzoko denbora-adierazpenek testuko denbora-adierazpen erlatiboak ebatzeko informazioa ematen dute.

Denbora-adierazpen bakoitzak testuan zer funtzio betetzen duen `functionInDocument` atributuaren bidez adierazten da. Aipatutako lehenengo multzokoek NONE balioa hartuko dute, ez baitute testuko denbora-erreferentzia egiteaz gain beste funtziorik. Bigarren multzokoek, ordea, balio hauek har ditzakete, adierazten duten informazioaren arabera:

- `CREATION_TIME`: dokumentuaren sorrera-data adierazteko
- `MODIFICATION_TIME`: dokumentuaren azken aldaketaren data adierazteko
- `PUBLICATION_TIME`: dokumentuaren argitalpen-data adierazteko
- `RELEASE_TIME`: dokumentuaren zabalpen-data adierazteko
- `RECEPTION_TIME`: dokumentuaren jasotze-data adierazteko
- `EXPIRATION_TIME`: dokumentuaren iraungitze-data adierazteko

Datu hori metadatu gisa adierazi ohi da Interneteko testuetan, baita data edo ordu esplizitu bezala ere albisteen kasuan. Esplizitu agertuz gero, dokumentuaren sorrera-datak denbora-adierazpenaren etiketa hartzen du, denbora-adierazpena baita. Etiketatzea errazteko, gure lanean ez dugu sorrera-, modifikazio-, publikazio- eta zabalpen-daten arteko desberdintasunik egin eta guztiei `CREATION_TIME` esleitu diegu `functionInDocument` atributuan.

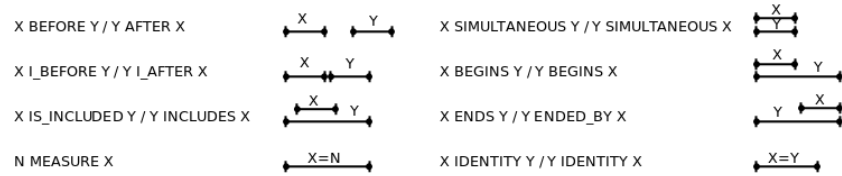
4.2.2.7 Beste atributu batzuk

Konputaziorako automatikoki sortutako atributuak esleitzen zaizkie denbora-adierazpenei, hala nola **tid** identifikatzailea eta **temporalFunction** denbora-adierazpena beste denbora-adierazpen baten balioa ebazteko erabili dela adierazteko. Atributuen zerrenda osoa B.2 eranskinean definitu dugu.

4.2.3 Denbora-erlazioen etiketatzea, TLINK

EusTimeMLn denbora-erlazioak TLINK etiketaren bidez adierazten dira, TimeMLn bezala. Etiketa horiek testutik kanpo kokatzen dira eta parte hartzen duten gertaera, denbora-adierazpen eta seinaleak beren identifikatzaileen bidez adierazten dira.

Denbora-erlazioen ezaugarri nagusiak bi dira: zer erlazio mota adierazten duten eta erlazio hori zer elementuk adierazten duen. Erlazio motaz ari garenean, bi elementuen arteko hurrenkera kronologikoaz ari gara, eta horiek errepresentatzeko Allenen Denbora-tarteen Teorian (ikus 2.2.2.2 atala) oinarritutako balioak erabiltzen ditugu euskarazko denbora-erlazioen prozesamenduan. Erlazio bakoitzak hartzen duen balio normalizatua 4.3. irudian ikus daiteke.



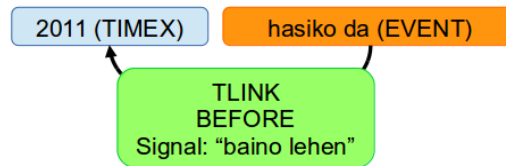
4.3 irudia – EusTimeMLko denbora-erlazioen balioak

Hala, (313) adibideko denbora-erlazioak 4.4. irudiko etiketatzea hartuko du; *hasiko da* gertaeraren eta *2011* denbora-adierazpenaren artean **BEFORE** motako denbora-erlazioa sortu da.

Aurreko atalean egin bezala, adibideetan erlazioaren iturria edo burua beltzez adieraziko dugu eta mendeko elementua edo erlazioaren xedea kurtzibaz. Erlazioan seinalerik badago, azpimarraz nabarmenduko dugu.

(313) Superjumboa *2011* baino lehen **hasiko da** zerbitzuan.

Denbora-erlazio motak TLINK etiketaren **relType** atributuaren bidez adierazten dira. Jarraian azalduko ditugu EusTimeMLn denbora-erlazioek har ditzaketen balioak:



4.4 irudia – (193) adibideko denbora-erlazioaren irudikapena

- BEFORE: gertaera bat edo denbora bat beste baten aurretik gertatzen dela adierazteko

(314) Bilerara *sartu aurretik*, Pavlo Klimkin Ukrainako Atzerri ministroak babesa **eskatu du**.

- AFTER: gertaera bat edo denbora bat beste baten ondoren gertatzen dela adierazteko

(315) Bilerara *sartu ondoren*, Pavlo Klimkin Ukrainako Atzerri ministroak babesa **eskatu du**.

- IBEFORE (*immediately before*): gertaera bat edo denbora bat justu beste baten aurretik gertatzen dela adierazteko. Bestela esan, lehenaren amaiera unea eta bigarrenaren hasiera unea bat datoz.

(316) Kamioi batek bi txirrin dulari **harrapatu ditu** eta *hil*.

- IAFTER (*immediately after*): gertaera bat edo denbora bat justu beste baten ondoren gertatzen dela adierazteko. Hau da, lehenaren hasiera unea eta bigarrenaren amaiera unea bat datoz.

(317) NA-1240 errepideko lehen kilometroan bi txirrindulari **hil dira**, kamioi batek azpian *harrapatuta*.

- INCLUDES: gertaera batek edo denbora batek beste bat bere barne hartzen duela adierazteko

(318) **Iazko epe bereko** galerekin alderatuta, txarrera egin du.

- IS_INCLUDED: gertaera bat edo denbora bat beste baten barne dagoela adierazteko

- (319) Petro Poroxenko Ukrainako presidenteak astebeteko su-etena **iragarri zuen** *joan den astean*.
- (320) Kooperatibistek batzarra **egin dute** *goizean*.
- **BEGINS**: denbora baten edo gertaera baten hasiera adierazten duten denbora edo gertaerak adierazteko².
 - **BEGUN_BY**: denbora-adierazpen baten unean edo beste gertaera bat gertatzen den unean hasten diren gertaerak adierazteko
- (321) Grebalariak elkarriketak *bertan behera utzi zirenetik* **daude** lantegian itxita.
- **ENDS**: denbora baten edo gertaera baten amaiera adierazten duten denbora edo gertaerak adierazteko.
- (322) *Bonbardaketak* amaitu zuen **bake garaia**.
- **ENDED_BY**: denbora-adierazpen baten unean edo beste gertaera bat gertatzen den unean amaitzen diren gertaerak adierazteko
- (323) Eusko Legebiltzarrak *urrira arte* **atzeratu** berri du Durangoko Azokari diru-laguntza handiagoa emateko erabakia.
- **IDENTITY**: gertaera berbera adierazten duten bi testu zati lotzeko, gertaera-adierazpen konplexuak direlako zein bi perpausetan gertaera berari erreferentzia egiten diotelako
- (324) Fundazio horrek Madrilen baso-eskola **irekitzeko** asmoa du. 2015–2016 ikasturtean *abiatuko* litzateke Bosquescuola ekimena.
- **MEASURE**: iraupena adierazten duen denbora-adierazpen baten gertaera batekiko lotura, gertaera tarte horretan gertatzen baita.
- (325) Petro Poroxenko Ukrainako presidenteak **astebeteko** *su-etena* iragarri zuen joan den astean.

²EusTimeML gidalerroetan erlazioaren iturria beti gertaera izango denez, ez dugu **BEGINS** motako loturarik aurkituko ez bada bi esaldiren arteko gertaera nagusien artean, kontzeptualki alde batera utzi baitugu denbora batek gertaera bat hasten duelako ideia.

- **SIMULTANEOUS**: bi denbora edo gertaera une berean (edo ia une berean) gertatzen direla adierazteko.

(326) Leihoak *zabaltzean* haize bortitza **sartzen da** haietatik.

4.2.3.1 TLINKen beste atributuak

Arestian esan bezala, TLINKak denbora-erlazioak adierazteko baliatu dira. Erlazio horiek gertaeren edota denbora-adierazpenen artean gauzatzen dira. Erlazio mota adierazten duen atributuaz gain, erlazioko burua eta xedea, baita erlazioa esplizitu egiten duen seinalea adierazteko atributu hauek ere sortu ditugu:

- **eventInstanceID** edo **timeID**: erlazioan parte hartzen duten elementuen identifikatzaileak. Nahitaezkoa da, tresnak beti esleituko du bat.
- **signalID**: denbora-erlazioa adierazten duen seinale bat testuan agertzen bada, horren identifikatzailea markatzeko. Aukerakoa da.
- **relatedToEventInstance** edo **relatedToTime**: erlazioan parte hartzen duen gertaerari edo denbora-adierazpenari lotutako entitatea adierazteko. Nahitaezkoa da, erlazioan parte-hartzen duten iturri eta xedearen identifikadoreak automatikoki jarriko ditu.

4.2.3.2 Seinaleen etiketatzea, SIGNAL

Denbora-erlazioak esplizitu egiten dituzten seinaleentzat **SIGNAL** etiketa baliatzen dugu EusTimeMLn TimeMLn bezala. Seinalea osorik etiketatuko da, hau da, denbora-erlazioa esplizitu egiten duten testuko token guztiak etiketaren barruan hartzen dira. Esaterako, (327) adibideko *eta berehala* bi tokeneko seinale bakarra da eta hala etiketatu da.

(327) Hegazkina singapurren entregatu **eta berehala**, Tolosara abiatu da.

Baieztapen horrek, ordea, murriztapenak ditu. 3.3.1.2 atalean aipatu legez etiketatzea token-mailakoa denez, ez dugu morfema-mailako elementurik etiketatuko. Zehazki, denbora-adierazpenei atxikitako postposizioak ez ditugu seinale bezala etiketatuko. (328) adibidean ikusten denez, denbora-erlazioa inesiboaren bidez adierazten da eta denbora-adierazpenari atxikita

dago. Kasu horretan, denbora-adierazpena adierazten duen tokenari (*goizean*) denbora-adierazpenaren etiketa baino ez diogu esleituko. (329) adibidean, ordea, *baino lehen* seinaleztat hartu dugu. Amaitzeko, parte askeek **SIGNAL** etiketa hartzen dute. Horrela gertatzen da (330) adibidean *-tik aurrera* seinalearekin: *aurrera* postposizio askeak baino ez du seinale etiketa hartzen.

- (328) Bazkide kooperatibistek batzarra egin dute **goizean**.
TIMEX3
- (329) Hegazkina bostak **baino lehen** lurreratu zen Pariseko Charles de Gaulle aireportuan.
SIGNAL
- (330) Zerbitzu bereziek urtarilaren 28*tik* **aurrera** jarraituko dute.
SIGNAL

Aditzei atxikitako seinaleei dagokienez, antzera jokatu da. Tken batek gertaera etiketa har badezake (331), ez da seinalerik etiketatuko. Gertaera eta seinalea adierazten duten tokenak ez badira berak (332), bi elementuak etiketatuko dira. (333) adibidean ere antzera gertatzen da. Kasu honetan denborazko menderagailua aditz jokatu adizlagunari atxikita agertzen da (*dir-enean*), baina gertaeren buru lexikoa baino ez dugunez etiketatzen, aditz laguntzailea adierazten duen tokena ez da gertaeratzat etiketatzen eta, ondorioz, seinale etiketa har dezake.

- (331) Bazkide kooperatibistek batzarra egin dute zuzendaria heltze**an**.
- (332) Hegazkina lurreratu **baino lehen** pistako argiak piztu ziren.
- (333) Zerbitzu bereziek zuzendaritza eta sindikatuak batzen **direnean** jarraituko dute.

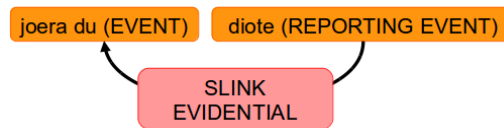
Seinale-etiketak ez du informazio semantikoa adierazten duen atributurik hartzen. Izan ere, duen karga semantikoa denbora-erlazioan adierazten da dagoeneko.

4.2.4 Mendekotasun-erlazioen etiketatzea, SLINK

Bi gertaeraren arteko mendekotasun-erlazioak markatzeko erabiltzen da **SLINK** etiketa. Erlazioak iturri bat eta xede bat izango du; gertaera nagusia eta mendekoa, betiere esaldi beraren barruan. Hala, (334) adibideko esaldiko

gertaeren arteko mendekotasun-erlazioa honela irudikatu da 4.5. irudian: *diote* adierazpenezko gertaera erlazioaren burua da eta *duela* xedea eta beraien artean EVIDENTIAL motako erlazioa sortu da *diote*k bere mendeko proposizioaren ebidentzia ematen baitu.

(334) Historiak errepikatzeko joera *duela* **diote** zenbaitek.



4.5 irudia – Representation of the subordination link in (213) adibideko mendekotasun-erlazioaren irudikapena

mendekotasun-erlazio motak (3.3.2 atala) SLINK etiketaren `relType` atributuaren bidez adierazten dira. Hurrengo adibideetan, gertaera nagusia (erlazioaren iturria) beltzez agertuko da eta mendekoa (erlazioaren xedea) kurtsubaz. Honako erlazioak adierazten dituzte SLINKek:

- MODAL modalentzat

(335) Endesak Garoña berriro zabaltzeko *baimena* **eskatuko du**.

- FACTIVE faktiboentzat

(336) Jaitik inor konturatu gabe *alde egiteko* **gai izan zen**.

- COUNTER_FACTIVE ez-faktiboentzat

(337) **Bertan behera geratu da** gaur hastekoa zen *epaiketa* auzipetuaren abokatua atxilotu dutelako.

- EVIDENTIAL Ebidentzialentzat

(338) Guztiek **ikusi zuten** jaitik *alde egiten*.

- NEG_EVIDENTIAL Ezezko ebidentzialentzat

(339) Kapitainak bera ere inoiz marinel *izana* **ukatu zuen**.

- CONDITIONAL Baldintzazkoentzat

(340) Zuei horrelakorik **gertatuko balitzaizue**, ez zenukete atarramentu onik *aterako*, duzuen eskarmentu urriagatik.

4.2.4.1 SLINKen beste atributuak

Erlazio motaz gain, erlazioaren iturria eta xedea ere adierazten dira SLINKetan:

- **eventInstanceID**: erlazioko buru den gertaera-instantziaren identifikatzailea. Nahitaezkoa da eta automatikoki esleitzen da.
- **subordinatedEventInstance**: mendeko gertaera-instantziaren identifikatzailea. Nahitaezkoa da eta automatikoki esleitzen da.

4.2.5 Aspektuzko erlazioen etiketatzea, ALINK

EusTimeMLn aspektuzko erlazioak adierazteko, ALINK etiketa definitu dugu, TimeMLn proposatutako eskemari jarraituz. Erlazioaren burua beti izango da aspektuzko gertaera bat (EusTimeMLn ASPECTUAL mota hartzen duena) eta erlazioaren xedea edozein motatako gertaera izango da. EusTimeMLn aspektuzko erlazio motak ALINK etiketaren **relType** atributuaren bidez adierazten dira. Atributu horrek ondoko balioak hartuko ditu:

- **INITIATES**: hasiera

(341) Gauzak beste era batera egiteko gogoz **ekin zion** *irakasletzari* Mitxelenak.
- **CULMINATES**: amaiera; gertaera bat berez amaitzen denean, amaiera “naturala”

(342) Tuneleko *lanen* **bukaera** txanpainaz ospatu zuten bi aldeetako langileek.
- **TERMINATES**: etetea; gertaera bat bere berezko amaiera baino lehenago eten egiten denean

(343) Gobernuak azalpenak *emateari* **utzi dio**.
- **CONTINUES**: jarraipena

(344) Autobide berriaren *lanek* aurrera **diraute**.
- **REINITIATES**: berrabiatzea

(345) Isunak *jartzeari* **berrekingo dio** udalak datorren hilabetetik aurrera.

(346) adibideko *hasi zen* eta *saltzen* gertaeren arteko aspektuzko erlazioa irudikatu da 4.6. irudian. Ikus daitekeenez, aspektuzko erlazioak ALINK etiketa hartu du eta INITIATES motakoa dela adierazi dugu, *saltzearen* hasiera adierazten baita.

(346) Apple multimedia produktu batzuk *saltzen* **hasi zen**.



4.6 irudia – (227) adibideko aspektuzko erlazioaren irudikapena

4.2.5.1 ALINKen beste atributuak

Erlazio mota adierazten duen ezaugarria garrantzitsuen bada ere, beste atributu batzuk ere hartzen dituzte aspektuzko erlazioek, hala nola parte hartzen duten gertaeren identifikatzaileak. Ondorioz, hauek dira ALINK etiketako atributuak:

- **eventInstanceID**: erlazioko buru den gertaera-instantziaren (aspektuzkoa) identifikatzailea. Nahitaezkoa da eta automatikoki esleitzen da.
- **relatedToEventInstance**: aspektuzko gertaeraren argumentu den gertaeraren identifikatzailea. Nahitaezkoa eta automatikoa da.

4.2.6 TimeMLrekiko egokitzapenak eta aldaketak

Ikusi bezala, euskarazko denbora-informazioaren ezaugarri nagusiak errepresentatzeko gai da TimeML eta, horregatik, TimeMLren ezaugarri asko erraz transferitu ditugu EusTimeMLra. Hala ere, euskararen ezaugarri linguistiko batzuk hobeto islatzeko, aldaketa batzuk egin behar izan dizkiogu. Era berean, beste hizkuntzetarako hartutako etiketatze-erabakiak aztertzean, ikusi

dugu hizkuntza guztietan daudela hizkuntzaren ezaugarriak hobeto islatzeko aldaketak. Atal honetan, EusTimeML sortzean egin ditugun egokitzapen nagusiak deskribatuko ditugu.

4.2.6.1 Predikatu generikoen tratamendua

Predikatu generikoak kronologian kokatzea zaila da, denboran asko luzatzen diren egoerak adierazten baitituzte. Egia kategorikoen kutsua dutenez, bertiereko egoeratzat har daitezke. Horregatik alde batera uzten dira TimeMLn. EusTimeMLn, ordea, SIBILAn (Wonsever *et al.*, 2012) egin bezala, gertaeratzat hartzen dira izatez errealitateko egoera bati egiten diotelako erreferentzia. (347) adibideko gertaera, oso denbora luzean aldaketarik gabe izan ez duen egoera da.

(347) Txinako bandera **gorria da**.

Hala ere, faktualitatea lantzean horien berezitasuna nabarmendu dugu.

4.2.6.2 Euskarazko aditzen aspektu eta tempusa errepresentatzeko balioak

3.1.2.2 eta 3.1.2.3 ataletan aipatu bezala, euskaraz aditz jokatuaren aspektuak eta tempusak 4.8. taulako balioak hartzen dituzte. Aztertu ditugun beste hizkuntzekin alderatuta, etorkizuna aspektu gisa adieraztea eta ez-oraina eta ez-lehenaldia den alegiazko tempusa dira euskararen ezaugarri nagusiak.

4.8 taula – Euskarazko aditz jokatuaren aspektua eta tempusa

Aspektua	Tempusa
Burutua	Orainaldia
Ez-burutua	Lehenaldia
Etorkizuna	Alegiazkoa
Aspektu markarik gabekoa	

Ondorioz, EusTimeMLn euskararen berezitasun hori kodetu behar da. Hala, euskararako **aspect** eta **tense** atributuentzat aspektuaren eta tempusaren informazioarekin bat datozen balioak definitu ditugu (4.9. taula).

4.9 taula – EusTimeMLn **aspect** eta **tense** atributuek hartzen dituzten balioak

aspect	tense
PERFECT	PRESENT
-PERFECT	PAST
FUTURE	ALEGIAZKOA
NONE	

4.2.6.3 Euskarazko modalitatea adierazteko balioak

3.1.2.4 atalean ikusi dugunez, aztertutako beste hizkuntzetako TimeMLetan hiru modutara adierazi da modalitatea: modalitatearentzako atributu baten bidez i) balio diskretuak definituta edo ii) modalitatea adierazten duen forma baliotzat erabilia, eta iii) mendekotasun-erlazioen bidez.

Euskararako, lehen bidea aukeratu dugu; **modality** atributua eta balio diskretuak definitu ditugu. Balioak euskarazko aditz/izen modal eta semi-modalenak dira: AHAL, NAHI eta BEHAR.

4.2.6.4 Euskarazko postposizioen tratamendua

Aزتututako hizkuntza gehienak ez bezala, euskara hizkuntza eranskaria da eta hitzek hainbat postposizio har ditzakete, postposizio-atzizkiak zein postposizio askeak. Denbora-informazioaren prozesamenduan, postposizio horiek nola tratatu erabaki behar izan dugu, tratatu ditugun egitura askotan informazio erabilgarria adierazten baitute.

3.3.1.2 eta 4.2.3.2 ataletan azaldu dugu postposizioekin izandako kasuistika interesgarriena, seinaleetan baitute postposizioek garrantzirik handiena. Gogoan izan behar dugu EusTimeMLn etiketatzea token mailan egiten dela eta, ondorioz, gertaeren eta denbora-adierazpenen informazioari lehentasuna eman diogula postposizio-atzizkiek adierazten duten seinaleen informazioaren aurretik.

Postposizioen tratamendu horrek beste hizkuntzetan baino nabarmen seinale gutxiago etiketatu izana ekarri du. Adibidez, MEANTIME corpuseko (597 esaldi) seinale kopurua (Minard *et al.*, 2015) eta EusTimeBank corpuseko entrenamendu eta garapenerako testuetako (710 esaldi) seinale kopurua alderatu ditugu 4.10. taulan. Ikus daitekeenez, preposizioak erabiltzen dituzten hizkuntzetan euskaraz baino ia hiru aldiz seinale gehiago etiketatu

dira, hizkuntza horietan esaldi kopurua txikiagoa bada ere.

4.10 taula – MEANTIME eta EusTimeBank corpusetako seinaleak

	Ingelesa	Neerlandera	Italiera	Gaztelania	Euskara
MEANTIME 597 esaldi	291	291	253	280	
EusTimeBank 710 esaldi					98

4.2.6.5 Ezeztapenaren tratamendua

EusTimeMLn esplizituki txertatu ez badugu ere, euskarazko ezeztapenaren azterketa egin dugu (Altuna *et al.*, 2017b), ezeztapenak bereziki gertaere-tan duen eraginagatik. Ezeztapena aztertzean, ezeztapen-markak eta horien eragin esparrua aztertu ditugu, ezeztapenaren eraginpean dauden gertaeren faktualitate-balioa aldatzen baita.

Ezeztapenaren fokoa ezeztapen-markak zuzenean ezeztatzen duen ele-mentua da, eta hori gertaera bada, gertaera horren polaritatea eta faktua-litatea negatiboak izango dira. Hala ere, ezeztapenak proposizio osoari ere eragiten dio, hori da ezeztapenaren irismena deritzona. Ezeztapenaren iris-menaren esparruan dagoen elementu bakoitzaren egia-balioa ez da zertan aldatu, ordea. Hala, ezeztapenaren azterketa egitean, gertaerei faktualitate-balioa esleitzeko irizpide nagusietako bat finkatu ahal izan dugu.

4.2.6.6 Alde batera utzitako atributuak

EusTimeML eraikitzean, eskuzko etiketatzearen optimizazioan zentratu ga-ra; hala, ahalik eta informazio gehien adierazi nahi izan dugu ahalik eta erarik errazenean. Ondorioz, TimeMLn definitutako atributu batzuk alde batera utzi ditugu, ez baitzaizkigu euskarazko denbora-informazioaren erre-presentazioan nahitaezkoak irudit. Atributu horiek hauek dira:

vform atributuak gertaerak adierazten dituzten aditz ez-jokatuen formak adierazten ditu. Euskaraz partizipioak adjektibotzat eta aditz-izenak izen-tzat hartzea erabaki dugu, eta, gaztelaniako gerundioa adberbioen bidez adie-razten denez gehienetan, ez dugu atributu horren beharrik ikusi. Aditz ez-

jokatuei aditz kategoria esleitu diegu, baina aspektu eta tempusik ez dutela adierazi dugu.

mood Euskaraz aditzaren modua adierazten bada ere, alde batera utzi dugu etiketatzea errazteko. Esaterako, indikatiboaren eta subjunktiboaren arteko bereizketa egitea baliagarria da, baina faktualitatea etiketatzeak eta mendekotasun-erlazioek nahikoa informazio ematen dutenez gertaeraren egia balioaz, atributu hau ez erabiltzea erabaki dugu.

4.2.6.7 Faktualitate-informazioa gehitzea

Euskarazko denbora-informazioa aztertzearen helburuetako bat testuetan gertatzen diren gertaerak identifikatzea eta kronologian kokatzea denez, gertaeren faktualitatea aztertu dugu. Izan ere, gertatzen diren gertaerak (faktualitate positiboa dutenak) baino ez dira denbora-lerroetan agertuko. Ondorioz, faktualitate-informazioa informazio gehigarri interesgarria da denbora-informazioaren prozesamenduan.

Beste hizkuntza batzuetan faktualitate-informazioa informazio beregainztat hartu bada ere, guk EusTimeMLn txertatzea erabaki dugu faktualitatea ebazteko eta denbora-informazioa ebazteko, kasu askotan informazio bera erabili behar delako. TimeML eskemak jatorriz aditzen aspektu eta tempusari buruzko informazioa, polaritatea, modalitatea eta mendekotasun-erlazioen informazioa duenez, faktualitatearen, iturriaren ziurtasunaren informazioa eta kasu berezien trataera baino ez da gehitu behar. NewsReader proiektuan ere antzera jokatu dute eta, beren markaketa-lengoaia TimeMLren oso antzekoa bada ere, esan daiteke EusTimeML dela TimeML eskemetan faktualitatea gehitzen lehena.

4.2.6.8 EusTimeMLren uztarketa beste etiketatzeekin

TimeML XMLn oinarritutako formatua da eta etiketak testuan bertan txertatzen dira. 4.7. irudian ikusten denez, **TIMEX3** eta **EVENT** etiketak eta beren atributuak testuko tokenei lotzen zaizkie testuan zuzenean. Formatu hori, gizakiontzat irakurterraza bada ere, ez da prozesamendurako egokia.

NewsReader proiektuan egin bezala, denbora-informazioaren prozesamendua NAF formatudun (Fokkens *et al.*, 2014) fitxategietan egin dugu. Formatu horrek nahibeste maila linguistikoko informazioa prozesatzea errazten du. Gure kasuan, informazio semantikoa, gramatikala eta denbora-informazioa

batera prozesatu ahal izan ditugu. Gainera, NAF formatuak XML formaturako bihurtzea erraza ahalbidetzen du etiketatze-geruza guztiak tokenei lotuta baitaude.

NAF formatuko dokumentu baten zati bat ikus daiteke 4.7 irudian. Koadro gorrian, tokenen formak, horien informazio linguistikoa eta zein tokeni lotzen zaizkion ikus daiteke. Koadro berdean, hiru denbora-adierazpenen etiketatzea nabarmendu dugu. Lehena (**tx1**) testu kanpoko da eta ez dago testuko token bat ainguratua. **tx2** eta **tx3**, ordea, testuko **w3–w5** eta **w20** tokenei lotuta daude, hurrenez hurren. Denbora-adierazpenaren etiketatzean, token-informazioaz gain, denbora-adierazpenen mota (**type**), balioa (**value**) edota testuan betetzen duten funtzioa (**functionInDocument**) ikus daitezke.

```
<!--fakturazioan-->
<term id="t344" lemma="fakturazio" pos="N" morphofeat="NC0IS000" case="IZE ARR INE NUMS MUGM
@ADLG">
  <span>
    <target id="w344"/>
  </span>
</term>
<!--...-->
<term id="t345" lemma="." pos="0" morphofeat="Fp" case="PUNT_PUNT">
  <span>
    <target id="w345"/>
  </span>
</term>
</terms>
<timeExpressions>
  <timex3 id="tx1" type="DATE" functionInDocument="CREATION_TIME" value="2013-10-16"/>
  <timex3 id="tx2" type="TIME" value="2013-10-16T17:05:59">
    <!--2013-10-16 17:05:59-->
    <span>
      <target id="w3"/>
      <target id="w4"/>
      <target id="w5"/>
    </span>
  </timex3>
  <timex3 id="tx3" type="TIME" value="2013-10-16TMO">
    <!--goizean-->
    <span>
      <target id="w20"/>
    </span>
  </timex3>
```

4.7 irudia – Euskarazko denbora-adierazpenen erauzketaren diagrama

Esaterako, 4.7. irudiko *goizean* denbora-adierazpena, (348) adibideko esaldikoa da.

- (348) Fagor Etxetresnen bazkide kooperatibistek batzarra egin dute **goi-zean**.

Behin denbora-informazioa nola etiketatu definituta, informazio hori markatuta duen corpusa azalduko dugu jarraian.

4.3 EusTimeBank corpora

EusTimeBank corpora EusTimeML markaketa-lengoaia baliatuz etiketatu den euskarazko corpora da. Bi domeinuko testuz osatuta dago: ekonomia-albisteak eta historia-testuak. Helburu behinena euskarazko denbora-informazioan baliatzea denez, beste hizkuntzetan ataza bererako erabili diren corpusetan oinarritu gara dokumentuak batzeko. Nagusiki TimeBank (Pustejovsky *et al.*, 2003b) eta MEANTIME (Minard *et al.*, 2016) ingelesezko albiste-corpusak eta WikiWars (Mazur eta Dale, 2010) gerrei buruzko narrazio-corpora izan ditugu inspirazio, horiek izan baitira arloan erreferente izan direnak. EusTimeBank hiru ataletan banatzen da: FaCor, EusMEANTIME eta WikiWarsEU.

- FaCor euskarazko Berria egunkaritik³ eta Argia agerkaritik⁴ jasotako 25 albisteko azpicorpora da. Albisteak Fagor enpresaren krisiari buruzkoak dira eta 2013–2014 urteen artean krisialdi horretan gertatu ziren gertakariak kontatzen dituzte. Bi arrazoi nagusi egon zen FaCor sortzeko i) gai konkretu bati buruzko ekonomia-albisteak zirela eta ii) hizkuntzaren kalitate altua zutela.

Bigarren arrazoi hori ez da inondik inora hutsala; izan ere, hizkuntzaren egiturak identifikatzeko (kasu honetan denbora-egiturak), beharrezko genituen egitura horiek zuzen islatzen zituzten testuak. Arrazoi horregatik, FaCoreko dokumentuak izan dira gidalerroak osatzean gehien baliatu direnak.

Azpicorpus hori euskarazko denbora-egiturak identifikatzeko eta gidalerroak osatzeko zein ebaluatzeko erabiltzeaz gain, EusTimeBank urrepatrio corpora osatzeko baliatu da; zehazki, azpicorpus horretako 9 dokumentu dira urre-patroia osatzen dutenak. Corpus hori tresna automatikoen entrenamendurako eta ebaluaziorako erabili dugu.

- EusMEANTIMEren oinarrian NewsReader proiektuan⁵ egindako lana dago. Ikergaiaren egungo egoera aztertzean ikusi dugunez, proiektu horrek helburu zuen ekonomiako aurreikuspenetan lagunduko zuten baliabideak sortzea eta, horretarako, ekonomia-albisteak prozesatu zituzten. 120 ekonomia-albiste batu zituzten MEANTIME corpusean (Mi-

³<http://www.berria.eus>

⁴<http://www.argia.eus>

⁵<http://www.newsreader-project.eu/>

nard *et al.*, 2016), 30 albiste gai hauetako bakoitzeko: burtsa, Apple, Boeing eta Airbusen arteko konpetentzia, eta General Motors, Chrysler eta Forden krisialdia. 2007–2010 urteen arteko albisteak dira eta WikiNewsetik⁶ erauzi ziren, WikiNewseko albisteak libreak baitira.

Jatorrizko corpusa ingelesez osatu bada ere, 120 dokumentuak gaztelania, italiara eta neerlandera itzuli dira. Etiketatzegidalerroak ere ingelesez osatu dira eta, gero, beste hizkuntzetara moldatu dira. Eskuzko etiketatzeak ere prozesu bera izan du; lehenik, ingelesezko dokumentuak etiketatu dira eta, ondoren, etiketatzea beste hizkuntzetara transferitu da (gaztelania eta italieraren kasuan, zuzenean transmititu da).

Gure ikerlanak ere denbora-informazioa helburu izanik, MEANTIME euskerara itzuli dugu. Ez soilik euskarazko corpusa zabaltzeko, baita euskarazko etiketatzea aurreko beste lau hizkuntzetakoekin konparatzeko ere. Hala sortu dugu EusMEANTIME azpicorpusa. Azpicorpus horrek jatorrizko MEANTIME corpusaren dokumentu berak ditu eta itzultzaile profesionalek itzuli dituzte. Markaketa-lengoaia, ordea, ez da bera bietan; MEANTIME NewsReader proiekturako espreski garatutako gidalerroei jarraituz etiketatu da eta EusMEANTIME EusTimeMLri jarraituz. Desberdintasun batzuk badaude ere (erlazio ezberdinak eta gertaeren sailkapen desberdina batik bat), hein handi batean bi markaketa-lengoaiak konparagarriak dira.

Azpicorpus hori nagusiki urre-patroi bezala erabili dugu tresnak eta baliabideak ebaluatzeko. Zehazki, 51 dokumentu baliatu ditugu EusTimeBank urre-patroi corpusean.

- WikiWarsEU Wikipediatik ateratako 19 gerratako narrazioz osatutako corpusa da, FaCor eta EusMEANTIME ez bezala. Inspirazioa WikiWars (Mazur eta Dale, 2010) corpusean du. Jatorrizko WikiWars ingelesezko Wikipediatik ateratako 20 gerratako dokumentuz osatuta dago. Euskarazkoa ere gerra berdinekin osatu nahi izan dugu, baina WikiWarsen batutakoetatik 19 gerratako dokumentuak baino ez ditugu aurkitu euskarazko Wikipedian.

Narrazio historikoak batzen dituen corpusa sortzeko arrazoia testu horiek albisteeikiko dituzten desberdintasunetan datza. Normalean, histo-

⁶https://en.wikinews.org/wiki/Main_Page

4.11 taula – EusTimeBank osatzen duten azpicorpusak

	Datu teknikoak	Jatorria	Erabilerak
FaCor	25 dokumentu 6.016 token Jatorriz euskaraz	Berria eta Argia	Gidalerroen definizioa 9 dok. EusTimeBanken
EusMEANTIME	120 dokumentu 31.223 token Itzulpenak	MEANTIME	Gidalerroen definizioa 51 dok. EusTimeBanken
WikiWarsEU	19 dokumentu 35.866 token Jatorriz euskaraz	Euskarazko Wikipedia	EusHeidelTimeren garapenerako

ria-testuetan eta albisteetan agertzen diren denbora-adierazpenak ezberdinak izaten dira. Historia-testuetan denbora-une esplizitu gehiago agertzen dira, bestela esan, data zehatz gehiago dago. Albisteetan, ordea, oso ohikoak dira denbora-adierazpen erlatiboak.

Beste hizkuntzetako WikiWars corpusak bezala, WikiWarsEU euskarazko denbora-adierazpenak identifikatu eta normalizatzen dituen EusHeidelTime tresna (Altuna *et al.*, 2017a) garatzeko eta ebaluatzeko erabili dugu. Hain zuzen ere, EusHeidelTime albisteetan zein narrazioetan egoki dabilela probatu nahi izan dugu. Horretarako, 19 dokumentuetako denbora-adierazpenak eskuz etiketatu ditugu eta EusHeidelTimeren etiketatzeekin alderatu dugu.

Hiru corpus horien ezaugarriak 4.11. taulan laburbildu ditugu.

Arestian aipatu bezala, corpus horietatik hartutako dokumentu batzuk EusTimeBank urre-patroi corpusean⁷ bildu ditugu eta tresnen entrenamendu eta ebaluaziorako baliatu ditugu. Tesi-lan hau amaitzeko orduan, EusTimeMLri jarraituta etiketatuta dagoen corpusaren lagina 60 dokumentukoa da. 30 trebakuntzarako (*train*), 15 garapenerako (*development*) eta 15 ebaluaziorako (*test*). 4.12. taulan azpicorpus horietako tokenen eta etiketen kopurua adierazten da.

⁷EusTimeBank urre-patroi corpora helbide honetan eskuratu daiteke: <http://ixa2.si.ehu.es/eusheideltime/>

4.12 taula – EusTimeBank urre-patroi corpusaren token eta etiketa kopurua

	Train (+ Development)	Test
Token kopurua	13.321	3.329
TIMEX3	448	137
SIGNAL	98	26
EVENT	2583	672
ALINK	50	10
SLINK	398	142
TLINK	3310	749

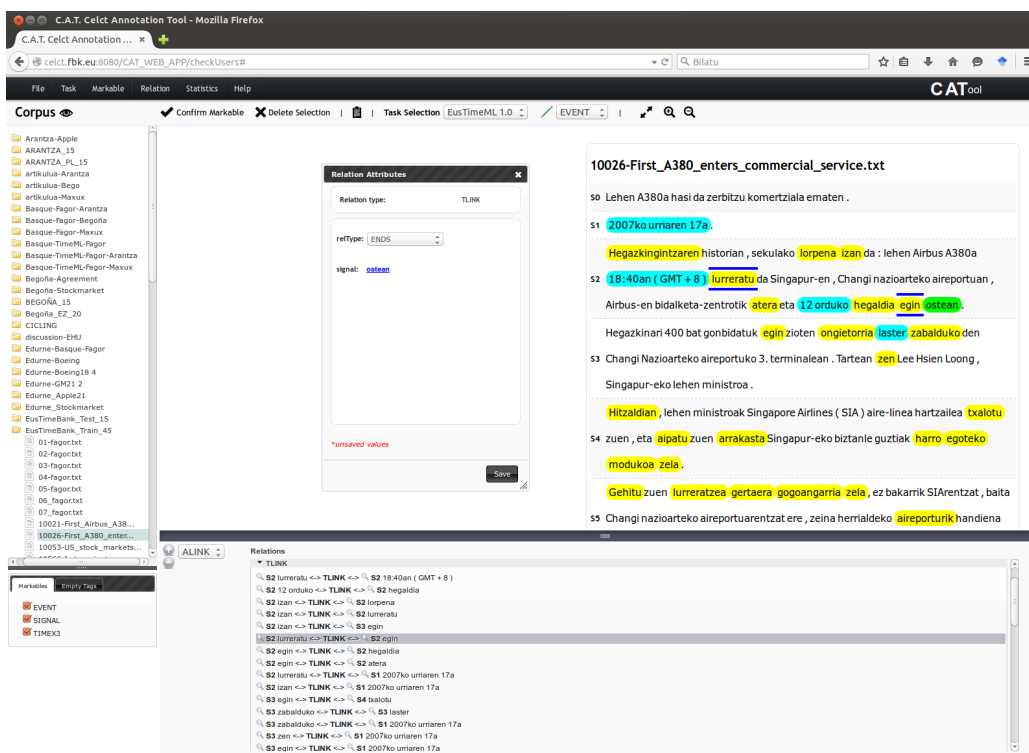
4.4 EusTimeBank eskuzko etiketatze metodologia

Arestian aipatu bezala, EusTimeBank corpora eskuz etiketatu dugu. Etiketatzeko, EustimeML etiketatze-gidalerroak eta CAT (Content Annotation Tool) tresna (Lenzi *et al.*, 2012) baliatu ditugu. Eskuzko etiketatzea eta gidalerroen fintzea batera egin ditugu, izan ere, etiketatze-esperimentuen bidez gidalerroen egokitasuna neurtu dugu eta emaitzen azterketaren arabera gidalerroak findu ditugu.

Bai corpusen azterketa, bai eskuzko etiketatzea CAT tresnaren bidez egin dugu. Eskuzko etiketatzerako tresna egokia dela iritzi dugu ondoko arrazoiengatik:

- CATek testu gordina, tokenizatua zein XML formatuan etiketatua har dezake sarreratzat. Testu gordinarentzat integratutako tokenizatzailea du eta tokenizazioaren eskuzko edizioa ere ahalbidetzen du.
- CAT edozein etiketatze linguistiko egiteko molda daiteke. Etiketatzearazak inporta daitezke edo integratutako ataza-sortzailearen bidez sor daitezke. Hor bi motatako etiketatzeak sortzen dira: etiketak eta erlazioak. Etiketak token batekoak edo hainbat tokenetakoak izan daitezke, eta erlazioetan aurretik etiketatutako bi etiketa lotzen dira. Tokenen eta erlazioen etiketatzea ikusten da 4.8 irudian.

Etiketa eta erlazioez gain, etiketa bakoitzari gehi dakizkiokkeen atributuak eta bere balioak ere defini daitezke. Eskuzko etiketatzea egitean,



4.8 irudia – Tokenen eta erlazioen etiketatzea

etiketa edo erlazio baten gainean klik egitean, atributuak betetzeko leihoa ateratzen da. 4.9 irudian, gertaera bati dagozkion atributuak eta horientzat hautatutako balioak etiketatzeko leihoa irudikatzen da.

- Etiketatzailen arteko adostasuna neurtzeko tresnak ditu.
- CATen bidez etiketatutako testuak XML formatuan esporta daitezke hizkuntzaren prozesamenduan baliatzeko.

4.10 irudian CATen interfazea agertzen da.

CATek eskaintzen dituen aukerak kontuan hartuta, corpora sortzeko ondoko metodologiari jarraitu diogu:

1. Etiketatzeko-ataza definitu dugu. Gure kasuan EusTimeMLri jarraituta etiketatu dugunez, markaketa-lengoiaren etiketak eta eskuz betetzeko atributuak eta beren balioak definitu ditugu.

Markable Attributes [X]

Current Extent: lurrerratu

Markable type: EVENT

class: OCCURRENCE

pos: VERB

tense1: PRESENT

tense2: -PAST

aspect1: PERFECT

aspect2: -FUTURE

polarity: POS

modality: NONE

certainty: CERTAIN

specialCases: NONE

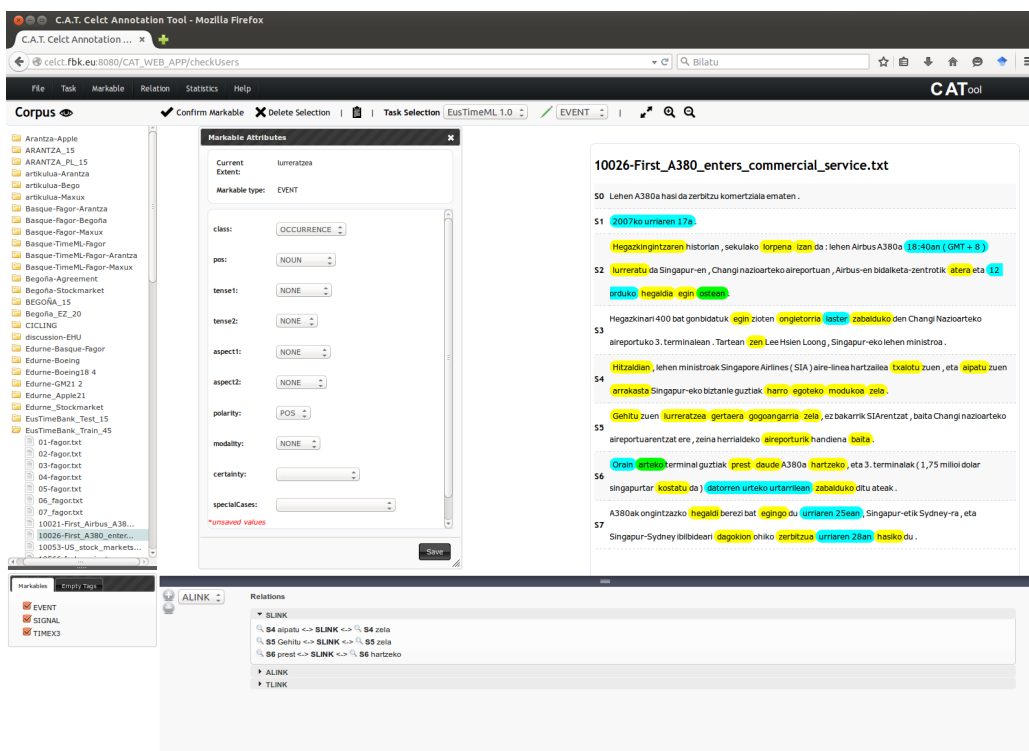
factuality: FACTUAL

*unsaved values

Save

4.9 irudia – Gertaera baten atributuak etiketatzeko leihoa

2. Corputa inportatu dugu. Etiketatzea egiteko aurretik Ixakateko tokenizatzailearen (ikus 5.1.1 atala) bidez tokenizatutako testuak baliatu ditugu. Horrela urre-patroiaren eta automatikoki sortutako dokumentuen tokenizazioa bera dela bermatu dugu.
3. Dokumentuak eskuz etiketatu ditugu.



4.10 irudia – CAT tresnaren ikuspegi orokorra

4. Etiketatzaileen arteko adostasuna neurtu dugu (ikus 4.5 atala) bai tresnan dauden etiketatzaileen arteko adostasuna neurtzeko tresnen bidez, bai beste neurri batzuk baliatuta, .
5. Gidalerroak findu ditugu etiketatzeen emaitzak aztertu ondoren.
6. Urre-patroia zuzendu dugu gidalerro berriei jarraituta.

Aipatu bezala, jarraian etiketatzean egin ditugun esperimentuen berri emango dugu.

4.5 Denbora-informazioa etiketatzeko esperimentuak

Atal honetan, urre-patroia sortzeko EusTimeML gidalerroen ebaluaziorako egin ditugun esperimentuak aurkeztuko ditugu: 4.5.1 atalean denbora-

adierazpenak eta seinaleak etiketatzeko esperimentua, 4.5.2 gertaerak etiketatzeko esperimentuak eta, azkenik, 4.5.3 atalean erlazioak etiketatzeko esperimentua.

4.5.1 Denbora-adierazpenak eta seinaleak etiketatzeko esperimentua

Lehenik, denbora-adierazpenak eta seinaleak etiketatzeko saiakera egin dugu (Altuna *et al.*, 2014a). NewsReader proiektuaren metodologia jarraituz, hiru anotatzailek (A, B eta C) euskarazko 4 kazetaritza-testu etiketatu dituzte denbora-adierazpenak eta seinaleak etiketatuz. Testuak FaCor corpuseko albisteak dira eta etiketatzailleek guztira 56 esaldi aztertu dituzte.

Etiketatzeari dagokionez, etiketatzailleek EusTimeML gidalerroen lehen bertsioari (Altuna *et al.*, 2014b) jarraitu diote eta denbora-adierazpenen zein seinaleen etiketatze osoa egin dute; bestela esan, atributu guztiak bete dituzte. Gidalerroen egokitasuna eta kalitatea neurtzeko, etiketatzailleen arteko adostasuna neurtu dugu eta etiketatzaille bakoitzaren etiketatzeak aztertu ditugu.

4.5.1.1 Anotatzaileen arteko adostasuna

Gidalerroen ebaluazioaren lehen urratsean, anotatzaileen arteko adostasuna neurtu dugu testuan ezarritako etiketa kopurua eta etiketa horien luzera (token berak hartzen dituzten) kontuan hartuta. Horretarako, Diceren koefizientea (Dice, 1945) erabili dugu denbora-adierazpen eta seinaleen gaineko adostasuna neurtzeko (ikusi 4.13 eta 4.14 taulak). Diceren koefizienteak etiketentzat token berak aukeratu diren neurtzen du adostasun osoa islatuz.

4.13 eta 4.14 tauletan, etiketaren gaineko hainbat adostasun neurri islatu ditugu. *Markable*⁸ etiketaren luzera zehatzaren gaineko adostasuna adierazten du eta *tokenek* token indibidualen gaineko adostasuna. Denbora-adierazpenetan oso hurbileko emaitzak lortu izanak adierazten du egiturak identifikatzean etiketatzailleek adostasun handia izan dutela. Seinaleen kasuan, token bakarrekoak izanik orokorrean, *markable* eta *tokenentzat* emaitza berak lortu dira. Tamalez, emaitzak ez dira hain onak.

Testuan ezarritako TIMEX3 etiketekin batera, bost TIMEX3 etiketa huts ere etiketatu dituzte. Etiketa hutsak informazio inplizitua azaleratzeko era-

⁸Markatu nahi den edozein entitate.

Anotatzaile bikoteak	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
A – B	0,96	0,976	0,969	0,977
A – C	0,943	0,965	0,923	0,965
B – C	0,902	0,94	0,892	0,942
Guztira	0,935	0,96	0,928	0,961

4.13 taula – Denbora-adierazpenen (TIME3) etiketatzaileen arteko adostasuna

Anotatzaile bikoteak	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
A – B	0,75	0,75	0,79	0,79
A – C	0,556	0,556	0,479	0,479
B – C	0,444	0,444	0,393	0,393
Guztira	0,583	0,583	0,554	0,554

4.14 taula – Seinaleen (SIGNAL) etiketatzaileen arteko adostasuna

biltzen diren testuaz kanpoko etiketak dira. Horien gaineko adostasuna oso eskasa izan da A anotatzaileak 5, Bk 1 eta Ck 0 etiketatu baititu. Hiru anotatzaileek jarri dituzten etiketak, testukoak zein hutsak, guztira 33 izan dira eta adostasuna % 66,7koa izan da. Seinaleen aldetik, 16 izan dira markatu direnak eta % 31,2ko adostasuna lortu da hiru anotatzaileen artean.

Denbora-adierazpenen balio eta motaren gaineko adostasuna ere neurtu dugu. 4.15 taulan **type** eta **value** atributuen gaineko adostasuna ikus dezakegu, bai bikotekakoa, bai orokorra.

Emaitza orokorrak emateaz aparte, adostasunaren eta desadostasunaren arrazoiak ere aztertu ditugu.

Anotatzaile bikoteak	A–B	A–C	B–C	Orokorra
type	0,76	0,64	0,75	0,55
value	0,76	0,68	0,46	0,45

4.15 taula – type eta value atributuen gaineko adostasuna

4.5.1.2 Emaizten ebaluazioa

Etiketatuak denbora-adierazpen eta seinaleak alderatuz gero, adostasuna non lortu den edo ez ikusi ahal izan dugu. Alde batetik, *markable* bati etiketa bera jarri zaion neurtu dugu eta beste aldetik, etiketa horiek hedapen bera (token berak) hartu duten. Denbora-adierazpenen kasuan, emaitza onak lortu dira, adostasun osoa beti izan da 0,89tik gorakoa testuan jarritako etiketetan, baina etiketa hutsak kontuan hartuz gero, emaitzek behera egin dute. Seinaleen kasuan, emaitzak nabarmen baxuagoak dira. Anotatzaileak bat ez etortzeko arrazoi nagusiak ondokoak izan dira:

- Denbora-adierazpen edo seinale bat etiketatu ez izana edo etiketa huts osagarriak sortu ez izana.
- Etiketatu behar ez zen tokena etiketatu izana. Hau da, denbora-egituratzat hartu da denbora-egitura ez dena.
- Tokenei etiketa ezberdina eman izana; denbora-egitura zuzen identifikatu, baina etiketa okerra eman izana.
- Denbora-adierazpenen atributuei balio ezberdinak eman izana.

Emaizten ebaluaziorako, anotatzaileak batu eta adostasunaren edo ez-adostasunaren arrazoiak aztertu ditugu. Gidalerroen interpretazio okerra izan da arazorik handiena eta, ondorioz, anbiguotasuna edo argitasun falta agertzen zuten atalak berrikusi eta zuzendu ditugu. Halaber, gidalerroak idaztean kontuan hartu ez ziren denbora-adierazpenak eta seinaleak, lagintzat hartutako corpusean agertu ez zirelako, gehitu ditugu, egituren zerrenda osatzeko. Horrekin batera, denbora-egitura batzuk, *jadanik* eta parekoak, esaterako, alde batera uztea erabaki dugu, ez baitute denbora-informazio berririk gehitzen eta beste hizkuntzetan ere albo uzten direlako.

Etiketa hutsen sorrera eta erabilera ere arazo iturri izan da. Horien bidez, testuan esplizitu agertzen ez den, baina deduzi daitekeen denbora-informazioa adierazten da. Gidalerroetan horien erabilera argiago azaldu dugu eta adibide praktikoak ere egin ditugu anotatzailen lana hobetzeko.

4.5.2 Gertaerak etiketatzeko esperimentuak

EusTimeMLren eta gertaerak etiketatzeko gidalerroen zuzentasuna eta uni-bertsaltasuna ebaluatzeko, gertaeren identifikaziorako eta ezaugarrien erauzketarako bi itzuliko etiketatze-esperimentua egin dugu (Altuna *et al.*, 2018a). Lehen etiketatze-esperimentuan, gidalerroak (Altuna *et al.*, 2014b) ebaluatze eta eztabaidatzeko saiakera egin dugu; ondoren, gidalerroak findu ditugu etiketatzeko irizpideak gehituz eta hobetuz.

Bigarren etiketatze-esperimentuan, euskarazko gertaera-adierazpenen urrepatroia sortu dugu gidalerroen bertsio berria (Altuna *et al.*, 2016a) oinarri hartuta. Halaber, hirugarren esperimentu batean, faktualitatea etiketatzeko esperimentua ere egin dugu, faktualitatea EusTimeMLn landu dugunez.

4.5.2.1 Lehen etiketatze-esperimentua

Lehen esperimentu honetan, FaCor corpuseko 172 gertaera inguru⁹ etiketatu ditugu. Gertaerak EusTimeMLri (Altuna *et al.*, 2014b) jarraituta etiketatu ditugu eta hiru etiketatzailek (A, B eta C) hartu dute parte.

Etiketatzeko-esperimentu honetan gertaeren identifikazioaren eta hedaduraren gaineko adostasuna neurtu dugu etiketatzeak binaka ebaluatuta. Adostasun maila Diceren koefizientea (Dice, 1945) aplikatuta, 0,864 eta 0,947 punturen artekoa izan da etiketatzaile bikotearen arabera. Horrez gain, gramatika-kategoriaren eta modalitatearen gaineko adostasuna eta aspektuzko gertaeren gaineko adostasuna ere neurtu dugu.

Ikusi dugunagatik, hiru etiketatzaileek token bakarraren araua errespetatu dute eta token bakarreko gertaerak orokorrean zuzen etiketatu dira. Halaber, ez genuen espero izenezko eta adjektibozko gertaerek adostasun maila altua lortuko zutenik, baina hala gertatu da. Hala ere, orokorrean adostasun maila altua izan bada ere, aspektu korapilatsu batzuk identifikatu ahal izan ditugu:

⁹Etiketatzailen arabera kopurua aldatzen da.

- Token batzuk gertaeratzat hartu dira ez zegokienean; nagusiki denbora-adierazpenetako parte diren aditzak (ikus 3.2.1.4 atala) eta diskurtso markatzaileak.
- Gertaera bat baino gehiagoko gertaera-egitura konplexuetan ez dira beti gertaera guztiak etiketatu.
- Aditzetatik eratorritako gertaera-adierazpenei ez zaie beti gramatika-kategoria bera esleitu.

Hurrengo etiketatze-saiakeran desadostasun horiek ekiditeko, berebizikoa izan da etiketatzailen arteko gidalerroei buruzko eztabaida. Nagusiki gidalerroetako pasarte ilunak eta gramatika-kategorien deskribapenetako anbiguotasunak landu dira. Horretarako, euskarazko gramatikak berraztertutako ditugu eta gidalerroetan informazioa era zehatzagoan adierazi dugu.

4.5.2.2 Bigarren etiketatze-esperimentua

Berranalisi linguistikoaren ondoren, bigarren etiketatze-esperimentu bat egin dugu. Oraingoan, lau etiketatzailerik hartu dute parte: hiruk EusTimeML eta CAT etiketatze-tresna ondo ezagutzen dute eta laugarrenak oso ondo ezagutzen ditu etiketatze-gidalerroak eta tresna. Etiketatzeari EusMEANTIME entrenamendurako azpicorpuseko 15 dokumentutan egin da. Hiru etiketatzailerik 115 esaldi etiketatu dituzte eta beren etiketatzeak laugarren etiketatzailerik etiketatzeekin erkatu dira.

Etiketatzailerik bakoitzaren (A, B eta C) eta laugarren etiketatzailerik (4.a) etiketa kopurua eta bat etorritako gertaeren kopurua adierazi ditugu 4.16 taulan. Kopuruek adostasun maila adierazten dute.

4.16 taula – Etiketatzailerik bakoitzak etiketatutako gertaera kopurua eta bat etorritako etiketatzeak

Dokumentu multzoak	Etiketatzailerik	4. etiketatzailerik	Bat etorritako gertaerak
Lehena (A)	96	74	69
Bigarrena (B)	394	418	358
Hirugarrena (C)	95	99	84

4.17 taula – Adostasuna gertaeren hedapenean

Etiketatzailerik bikoteak	Micro- average (Marka- ble)	Micro- average (Token)	Macro- average (Marka- ble)	Macro- average (Token)
A – 4.a	0,812	0,812	0,819	0,819
B – 4.a	0,877	0,877	0,875	0,875
C – 4.a	0,866	0,866	0,883	0,883

Desadostasunerako arrazoirik handiena token batzuei gertaera etiketa ematea izan da. Esaterako, (349) adibidean, EusMEANTIME corpuseko dokumentu batean *ekoizpenak* gertaera adierazten du; (350) adibidean, ordea, forma berak ez du gertaerarik adierazten. Arazo hori bereziki nabarmena izan da forma berak gertaera eta gertaera baten produktua edo ondoria adierazten dituenean.

(349) **Ekoizpena** AEBra ekartzeko asmoa du.

(350) Nekazariak euren **ekoizpena** salgai jarriko dute.

Egoerak adierazten dituzten gertaerak ere desadostasun iturri izan dira. Batzuetan zaila izan da gertaera diren erabakitzea; gertatzen ari den gertaera estatiko bat (351) edo entitate baten ezaugarri bat (*desengainatuta egotea*) adierazten duten erabakitzea.

(351) Oso **desengainatuta** gaude.

Horretaz gain, kontuan izan behar da euskaraz *egin* aditzak balio fokalizatzailea duela. *Eginek* aditzaren fokalizazioa adierazten duenean, ez du gertaera-informaziorik adierazten. Hala ere, etiketatzailerik batzuek horrelako kasuetan *egin* etiketatu dute.

4.17 taulako emaitzek erakusten dutenez, adostasun handia (Dice, 1945) lortu da etiketatzailerik eta laugarren etiketatzailerik (4.a) artean gertaeren hedapenari (*markable extent*) dagokionez. Hau da, bi etiketatzailerik artean etiketen gainjartze osoa egon da kasu gehienetan. Tokenen hedapena (*token extent*) ebaluatzean, ordea, soilik gainjartzen diren tokenen araberrako gertaeren hedapena ebaluatzen da. Gure kasuan, *markable* zein *token extentek*, biek dute balio bera, gertaerak token bakarrek hedapena dutelako, gidalerroetan

4.18 taula – Adostasuna etiketatutako gramatika-kategorietan

Gertaeren etiketatzea	A-SE	B-SE	C-SE
Aditzak	45	242	51
Izenak	14	58	33
Adjektiboak	1	3	1
Adberbioak	0	9	2
Izenordainak	0	1	0
Bestelakoak	0	0	0
GUZTIRA	60 (% 87)	311 (% 87)	77 (% 92)

definitu bezala. Ondoriozta daiteke kasu honetan ere etiketatzailerik guztiek EusTimeML gidalerroetako token bakarraren araua errespetatu dutela.

Hedapenaz gain, gertaeren ezaugarri linguistikoen etiketatzea ere aztertu dugu. Lehen ezaugarria gramatika kategoria izan da.

4.18 taulan adierazi bezala, gertaeren gramatika-kategoriari buruzko adostasun handia lortu da. Desadostasun gehienak etiketatzailerik batek gertaera bati baliorik eman ez diolako edo balio lehenetsia aldatu ez duelako gertatu dira. Hala ere, beste desadostasun batzuk ondoko arrazoi gramatikalek eragin dituzte:

- *-tea/-tzea* motako nominalizazio batzuek aditz- zein izen-kategoria hartu dute.
- *-tako/-riko* erlatibo markadun partizipioek aditz- zein adjektibo-kategoria hartu dute.
- Aditzondo batzuk aditz formaren zatitzat hartu dira eta aditz-kategoria edo “bestelako” kategoria hartu dute.

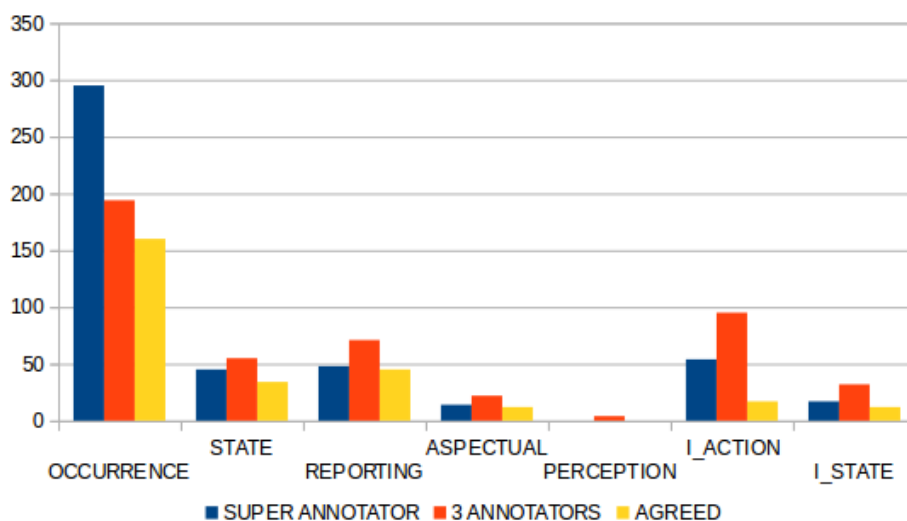
Gertaera modalen etiketatzea ere aztertu dugu. Hiru etiketatzailerik eta laugarren etiketatzailerik etiketatutako aditz modalen bat etortzea 4.19 taulan ikus daiteke. Gertaera modalak identifikatzea erraza izan da, modalitatea adierazten duten formetan aldakortasun gutxi baitago. Gainera, ia ez dago aukerarik erratuta gertaera modal bati balio okerra emateko, gertaera modalen esanahiak oso argi bereizten baitira. Gertaera modalen kopurua txikia bada ere, adostasuna aztertzeke baliagarriak izan zaizkigu. Ikusi ahal izan

4.19 taula – Adostasuna modalitatean

Gertaera modalen etiketatzea	A-SE	B-SE	C-SE
Behar	0	5	2
Nahi	0	3	1
Ahal	3	4	0
GUZTIRA	3	12	3

dugunez, desadostasunak etiketatzailerik etiketatzeko egindako akatsek sortu dituztelako izan da; ez, gertaera horien pertzepzio edo ulermen okerrenetatik.

Amaitzeko, gertaera-kategoriaren gaineko adostasuna neurtu dugu. Emaitzak ez dira espero bezain onak (A-4.a: % 58, B-4.a: % 56 eta C-4.a: % 49). Hala eta guztiz ere, aipatzekoa da adostasuna asko aldatzen dela kategoriaren arabera. Gertaera diskurtsiboak eta aspektuzkoak erraz identifikatu dira, batzuei, ziurrenik akatsak direla medio, kategoria okerra eman bazaie ere. Jazzerak eta ekintza intentsionalak, ordea, hainbatetan izan dira desadostasun iturri 4.11 irudian ikus daitekeenez.



4.11 irudia – Gertaera klaseen adostasunaren arteko konparaketa

Etiketatzeko-emaitzak aztertu ondoren, ondorio batzuk atera ditugu eta

erabaki batzuk hartu ditugu:

- Gertaeraren identifikazioa eta gramatika-kategoriaren araberako sailkapena ez dira menderatzeko zailak.
- Gertaera modalak orokorrean ondo etiketatu badira ere, etorkizunean horiei buruzko eztabaida eta entrenamendu handiagoa espero dugu emaitzak hobetzeko.
- Gertaeren sailkapenaren gaineko adostasuna espero baino txikiagoa izan da. Katgoria batzuk esleitzen errazagoak izan badira ere, katgoria korapilatsuenek beste hausnarketa bat merezi dute.

4.5.2.3 Faktualitatea etiketatzeko esperientua

Aipatu bezala, faktualitatea denbora-informazioaren prozesamenduan oso informazio baliagarria da. Faktualitatea etiketatzeko eskema definitu ostean (ikus 4.2.1.8 atala), gidalerroen egokitasuna neurtzeko esperientua egin dugu (Altuna *et al.*, 2018b). Esperientu horretan bi etiketatzailerak hartu dute parte. Gertaerak osorik etiketatzeko eskatu diegu, EusTimeMLko informazioa faktualitatea ebazteko erabili ahal izateko. Era berean, faktualitatea ebazteko mundu-ezagutza baliatzeko eskatu zaie. Guztira, etiketatzailerak bat etorri dira 734 gertaeren identifikazioan (787tik eta 818tik, hurrenez hurren).

Horretaz gain, gertaera horien faktualitateari dagozkion atributuen etiketatzea aztertu dugu. 4.20 taulan, faktualitate-informazioa adierazten duten atributuen gaineko doitasuna eta κ balioa (Cohen, 1960) adierazten dira. κ neurriak item kualitatiboen (kategorikoak) gaineko etiketatzailerak arteko adostasuna neurtzen du. Ikus daitekeenez, atributu gehienek nahiko doitasun altua lortzen dute eta κ balio altuek adostasun handia adierazten dute. κ neurriak asko penalizatzen ditu balio baterako joera dituzten kategoriak eta horregatik, katgoria batzuetan κ balio baxuak lortu dira. Esaterako, ziurtasunarentzat *certain* eta faktualitatearentzat *factual* bezalako balio batzuk maiz agertu dira etiketatzean, albisteetan gertatu diren eta ziurrak diren gertaerak agertzen baitira gehien.

Desadostasunei dagokienez, desadostasun gehienak gidalerroetako azalpen lausoei eragin dituzte. Adibidez, (352) adibideko *emateko* gertaerari etiketatzailerak UNCERTAIN eta UNDERSPECIFIED esleitu diete. Gidalerroetan adierazi denez, asmoa adierazten duten gertaerak mendeko gertaerak

4.20 taula – Etiketatzaileen arteko adostasuna faktualitatearen etiketatzean

	Polaritatea	Ziurtasuna	Kasu bereziak	Faktualitatea
Doitasuna	0,98	0,89	0,95	0,77
κ	0,68	0,24	0,29	0,53

raren ziurtasun balioa baldintzatzen dute eta hori **UNCERTAIN** da. Hala ere, ez dago argi *prest dago* gertaerak nahia adierazten duen eta etiketatzaileek desberdin etiketatu dute.

(352) 20 milioi dolar arteko laguntza **emateko** prest dago.

Azalpen lauso horiek gidalerroak berririo eztabaidatzean konpondu dira. Hala, gidalerro zehatzagoak definitu ditugu i) ziurtasunarentzat **UNCERTAIN** eta **UNDERSPECIFIED** balioak birdefinituta, ii) perpaus generikoen mugak eta ezaugarriak zehaztuta eta iii) ezeztapenaren fokoa hobeto aztertuta.

4.5.3 Erlazioak etiketatzeko esperimentua

Erlazioen etiketatze-gidalerroak ebaluatzeko, bi etiketatze-saiakera egin ditugu. Lehen, gidalerroen ebaluaziorako eta bigarrena, lehenengotik egindako aldaketak ebaluatzeko.

Lehen etiketatze-esperimentuan, hiru etiketatzailek hartu dute parte eta dokumentu bateko erlazioak identifikatu eta sailkatu dituzte. Lehen etiketatzaileak 102 erlazio sortu ditu, bigarrenak, 96 eta hirugarrenak, 87. Horietatik 43tan bat etorri dira hirurak eta beste 39tan bi etiketatzailearen arteko adostasuna egon da.

4.21 taula – Etiketatzaileen arteko adostasuna erlazioen etiketatzean (1. esperimentua)

Etiketatzailerikoteak	Erlazioen adostasuna	Erlazioen sailkapenean adostasuna
A – B	7	2
A – C	18	10
B – C	14	13
ABC	43	17

Erlazioen sailkapenari dagokionez, 4.21 taulan ikus daitekeenez, hiru etiketatzailerik etiketatu dituzten 43 erlazioetatik 17ri eman diete klase bera. Halaber, bi etiketatzailerik etiketatu dituztenetan, 39tik 25i klase bera esleitu diete. Nabarmentzekoa da B eta C etiketatzailerik arteko adostasuna A etiketatzailerik baino altuagoa dela. Emaitzak ez dira altuak, baina A etiketatzailerik lanean 12 erlaziok ez dute motarik esleituta, ziurrenik informazioa gordetzeko arazoak izan dituelako.

Mendekotasun-erlazioen motaren esleipenean, erlazio ebidentzialen eta faktiboen arteko nahasketa ikusi dugu. Horren atzeko arrazoiak da ebidentzialak gertaera bati buruzko ebidentzia adierazten dutela eta faktiboek gertaera gertatu dela adierazten dutela; hau da, semantikoki hurbil dauden erlazioak dira.

Semantikaren ikuspuntutik hurbil egotea edo testuinguruan ezin bereizi izatea izan dira denbora-erlazioei mota ezberdina esleitzeko arrazoi nagusiak. Adibidez, hainbat aldiz identifikatu ditugu desadostasunerako iturri izana al-diberekotasuna (SIMULTANEOUS) eta gertaera batek beste bat barruan hartzea (IS_INCLUDED eta INCLUDES), baita lehenago (BEFORE) eta justu lehenago (IBEFOR) eta geroago (AFTER), eta justu geroago (IAFTER) balioak esleitzea ere. Etiketatze-saiakera horren ostean, gidalerroak berriro aztertu ditugu eta azalpen eta deskribapenen alde ilunak identifikatu ditugu.

Bigarren etiketaze-saiakera bat ere egin dugu, gidalerroak berriz ebaluatzeke. Bi etiketatzailerik hartu dute parte eta gertaerak, denbora-adierazpenak eta seinaleak aurretiaz etiketatuta dituen dokumentu bateko erlazioak etiketatu dituzte. Esperimentu horretan etiketatzailerik batek 71 erlazio etiketatu ditu eta besteak 61. 61 horiek bi etiketatzailerik identifikatu dituzte eta horietan erlazio-motaren gaineko adostasuna neurtu da. 4.22 taulan ikus daiteke zenbat erlazio etiketatu dituzten bi etiketatzailerik.

4.22 taula – Etiketatzailerik arteko adostasuna erlazioen etiketatzean (2. esperimendua)

	TLINK		SLINK	
	Etiketatuak	Adostuak	Etiketatuak	Adostuak
A etiketatzailerik	68	59	3	2
B etiketatzailerik	59	59	2	2

Erlazioaren klasearen gaineko adostasuna ere aztertu dugu. Bi etiketatzailerik etiketatu dituzten mendekotasun-erlazioei klase bera esleitu diete

(2tik 2ri) eta denbora-erlazioetan 59tik 54ri; hau da, osotara % 92ko adostasuna lortu da denbora-erlazioen sailkapenean. Gertaerak dokumentuaren sorrera datarekin lotzen dituzten erlazioetan are altuagoa izan da adostasuna: % 96; ataza ondo definituta dagoen seinale da hori.

Denbora-erlazioen sailkapeneko desadostasunak aztertzean, ikusi dugu `identity` balioak sortu duela arazo gehien. Erlazio-mota horrek gertaera beraren bi adierazpen lotu direla adierazten du. Desadostasunen testuingurua aztertuta, uste dugu etiketatzailaetako batek gertaera berari erreferentzia egitea eta aldiberekotasuna nahastu dituela.

4.6 EusTimeBank corpusaren beste erabilerak: ezeztapenaren analisia eta etiketatzea

Denbora-informazioaren azterketaz eta tratamenduaz gain, EusTimeBank corpuseko lagin bat euskarazko ezeztapenaren azterketan baliatu dugu. Ezeztapena estuki lotuta dago denbora-informazioaren prozesamenduari gertaeren faktualitatean eragina baitu. Jarraian deskribatuko dugu ezeztapena aztertzeko egindako esperimentua.

4.6.1 Ezeztapena etiketatzeko eskema

Ezeztapenaren inguruko informazioa denbora-informazioaren osagarritzat hartu dugu. Izan ere, denbora-informazioaren baitan gertaeren faktualitatea dago —gertaerak gertatu diren edo gertatuko diren identifikatzen duen informazioa— eta faktualitatea ondo ebazteko, ezinbestekoa da ezeztapena identifikatzea. Orokorrean, ezeztatuta agertzen diren gertaerak ez direla gertatu edo gertatuko uste da, hau da, faktualitate negatiboa izango dutela.

EusTimeMLz aparteko gidalerroak¹⁰ sortu ditugu ezeztapenarentzat. Ezeztapen-informazioa adierazteko hiru entitate definitu ditugu:

- Ezeztapen-markak
- Ezeztapenaren irismena
- Ezeztapenaren fokoa

¹⁰https://docs.google.com/document/d/18tj2TNIjcNs-jeipcPzYo0j_R8ubuMweJHW-JwYz0qQ/edit?usp=sharing

Euskarazko ezeztapenaren gidalerroak definitzeko, italierarako sortutako gidalerroak¹¹ eta eta euskal gramatiketan (Altuna *et al.*, 1987) euskarazko ezeztapenaren gauzatzeaz esaten dena hartu ditugu abiapuntutzat.

4.6.2 Corpora

Esperimentua egiteko EusTimeBank urre-patroi corpuseko 20 dokumentu baliatu ditugu. Lagina albiste ekonomikoz osatuta dago, 5.247 hitz ditu eta ezeztapena etiketatu duten etiketatzaileek 33 ezeztapen-marka etiketatu dituzte. Corpus hori EusTimeMLren bidez ere etiketatuta dago eta espero dugu etorkizunean bi markaketa-lengoaiak uztartu ahal izatea.

4.23 taula – Esperimentuan erabilitako corpusaren datuak

Dokumentu kopurua	20
Token kopurua	5247
Ezeztapen kopurua (ggb)	33

Etiketatzailen arteko adostasuna neurtu da Diceren koefizientea (Dice, 1945) erabiliz. 4.24. taulan agertzen den bezala, etiketa bakoitzeko lau balio lortu ditugu adostasunaren ebaluaziorako: etiketen batez besteko hedapenarekin bat etortzea dokumentu bakoitzeko: *micro-average (markable)*; batez besteko tokenen gainjarketa dokumentu bakoitzeko: *micro-average (token)*; batez besteko etiketen hedapenaren bat etortzea corpusean: *macro-average (markable)*; eta batez besteko tokenen gainjarketa corpusean: *macro-average (token)*.

Emaitzetatik ondoriozta daiteke ezeztapena adierazten duen elementua identifikatzea ez dela lan zaila izan, ia bat etortze osoa (0.969 eta 0.982ko balioak) lortu baitugu. Desberdintasun bakarrak da “proiektua onartu **ez** zelako” eta “jaso **gabe**” egiturak etiketatzaile bakarrak etiketatu dituela eta, ezeztapen argiak izaki, uste dugu etiketatzaile batek ahaztu egin dituela.

Irismenari eta fokoari dagokienez, ordea, emaitzak ez dira hain onak izan, nabarmentzekoa bada ere irismenen gainjartzea 0,8tik gorakoa izan dela. Emaiza horietatik ondoriozta dezakegu irismena identifikatzea ere ez dela

¹¹https://docs.google.com/document/d/116FCW01A_DohTUk-j_k3rTyjoFF-c3Ih172A61y6pLA/edit?usp=sharing

4.24 taula – Ezeztapen-marken etiketatzean etiketatzailen arteko adostasuna

Etiketa	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
NEG_CUE	0,969	0,969	0,982	0,982
SCOPE	0,562	0,806	0,55	0,858
FOCUS	0,381	0,55	0,373	0,549

ataza zaila, desberdintasunak aurki badaitezke ere. Fokoaren kasuan, ordea, emaitzak baxuak izan dira. Fokoaren identifikazioan ez da nahikoa sintaxia, subjektibotasunak eta intonazioak rol garrantzitsua jokatzen baitute.

4.7 Laburpena

Kapitulu honetan, denbora-informazioa etiketatuta duen corpora sortzeko egin ditugun urratsak azaldu ditugu. Lehenik, 3. kapituluaren identifikatu ditugun denbora-informazioaren elementuak kodetzeko, EusTimeML markaketa-lengoaia definitu dugu. EusTimeML TimeMLren ereduari jarraiki eraiki dugu eta denbora-informazioa adierazten duten elementu guztiak errepresentatzeko gai da. Era berean, TimeMLn ez bezala, EusTimeMLn faktualitate-informazioa kodetzeko eskema gehitu dugu, faktualitateak denbora-informazioaren prozesamenduan duen garrantzia dela-eta.

Etiketatzeko erabakiak esperimentuen bidez ebaluatu ditugu. Ez dira markaketa-lengoiako aspektu guztiak aztertu, zailtasun handiena sor zezaketenak baizik. Adibidez, ataza zaila izan da gertaera batzuk sailkatzea. Etiketatzailen arteko desadostasunak aztertuta etiketatze-gidalerroetan korapilatsua den informazioa identifikatu dugu. Hala, gidalerroak findu ahal izan ditugu.

Behin gidalerroak definituta, corpora eskuz etiketatzeari ekin diogu. EusTimeBank corpusak 3 azpicorpus ditu eta albiste eta testu historikoez osatuta dago. EusTimeBankeko 164 dokumentuetatik 60 prozesamendurako urre-patroi bezala baliatu dira. Hurrengo kapituluaren deskribatuko ditugu euskarazko denbora-informazioa prozesatzeko garatu ditugun tresnak, baita informazio hori baliatzen duen denbora-lerroak sortzeko tresna aurreratua ere.

TRESNAK ETA APLIKAZIOAK

EusHeidelTime eta KroniXa

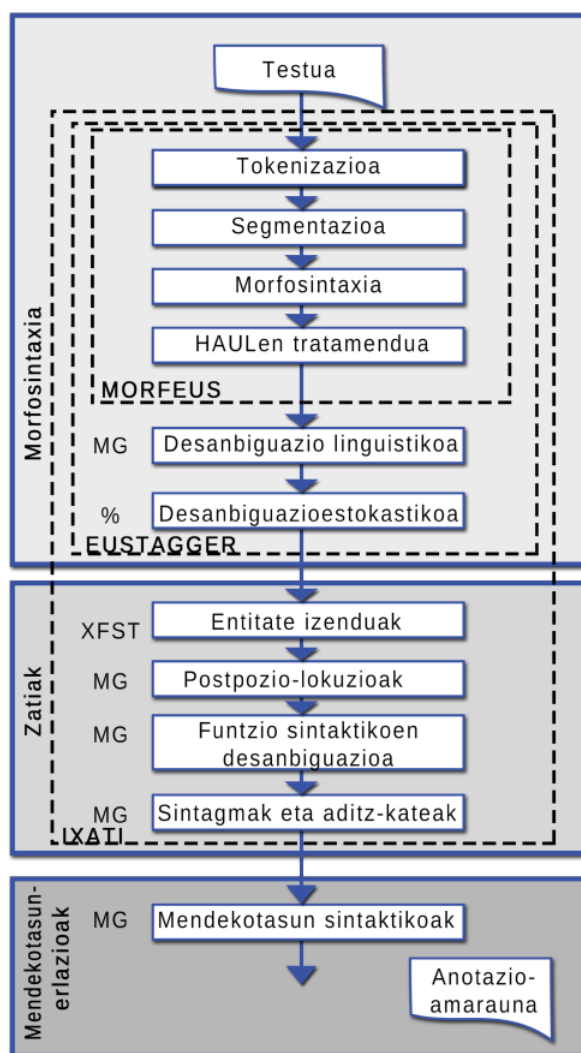
Kapitulu honetan denbora-informazioa automatikoki aztertze sortu ditugun tresnak deskribatuko ditugu. Alde batetik, denbora-adierazpenak identifikatzeko eta normalizatzeko, EusHeidelTime tresna garatu dugu. Beste aldetik, denbora-lerroak sortzeko, KroniXa tresna sortu dugu. Horiek denbora-informazioaren prozesamendurako definitu dugun etiketatze-eskema eta osatu dugun corpusa erabilia garatu ditugu. Era berean, tresna horiek baliatzen duten informazio linguistikoa lortzeko aurreprozesamendurako tresnen berri ere emango dugu.

Kapitulua horrela antolatu dugu: 5.1 atalean, Ixa taldean garatu diren hizkuntzaren tratamendu automatikorako analisi-katearen berri emango dugu. 5.2 atalean, EusHeidelTime tresna azalduko dugu eta 5.3 atalean, denbora-lerroak sortzeko KroniXa tresna aurkeztuko dugu.

5.1 Aurreprozesamendua: euskararen prozesamendurako analisi-katea

Euskararen denbora-informazioaren tratamendu automatikoa gauzatzeko, ezinbestekoa da Ixa ikerketa-taldean sortutako hizkuntzaren prozesamendurako katea, IxaKat, erabiltzea. Hau da, denbora-informazioa automatikoki tratatzeko abiapuntua izango da euskararen prozesamendurako definitutako analisi-katea (5.1 irudia).

Analisi-kate hori modulutan banatuta dago. Modulu bakoitzak aurreko moduluak eskaintzen dion informazioa erabiltzen du sarrera moduan, eta



5.1 irudia – Oronoz (2008) lanean aurkeztutako euskararako prozesamenduaren katearen eskema

jasotako analisia informazio linguistiko berriarekin aberasten du. Jarraian informazio linguistikoaren (aurre)prozesamenduan egiten ditugun lanak eta lan horiek egiteko baliabideak deskribatuko ditugu. Prozesu hori (353) adibideko esaldiaren analisiaren bidez irudikatuko dugu.

(353) Lehen Airbus A380 entregatu da.

Lan honetan bereziki hiru modulu baliatu ditugu denbora-informazioa prozesatu ahal izateko: analisi morfologikoa, dependentzia sintaktikoen analisi eta rol semantikoen analisi. Zehazki EusHeidelTimek (ikus 5.2 atala) analisi morfologikoko informazioa baliatzen du eta KroniXak (ikus 5.3 atala) dependentzia sintaktikoen eta rol semantikoen.

5.1.1 Tokenizazioa

Abiapuntua testu gordina izanik, tokenizazioa Morfeus tresnaren (Aduriz *et al.*, 1998) bidez egin dugu. Tokenizatzaileak testua tokenetan (zuriunez banatutako karaktere-segidetan) eta esaldietan banatzen du; hau da, analisi morfologikoa sarrera gisa erabiliko diren unitateak bereizten dira *token* edo *item* ezagutzailaren bidez. Gainera, testuan hitzak, zenbakiak (arruntak zein erromatarrak; deklinatu gabeak edo deklinatuak), laburdurak, siglak, zuriuneak eta puntuazio-markak identifikatzen ditu. Tokenei ere informazio geografikoa (identifikadorea) gehitzen die. Tokenizazioa egitean, (353) adibideko esaldia, 5.2 irudian bezala geratzen da¹.

Lehen | Airbus | A380 | entregatu | da | .

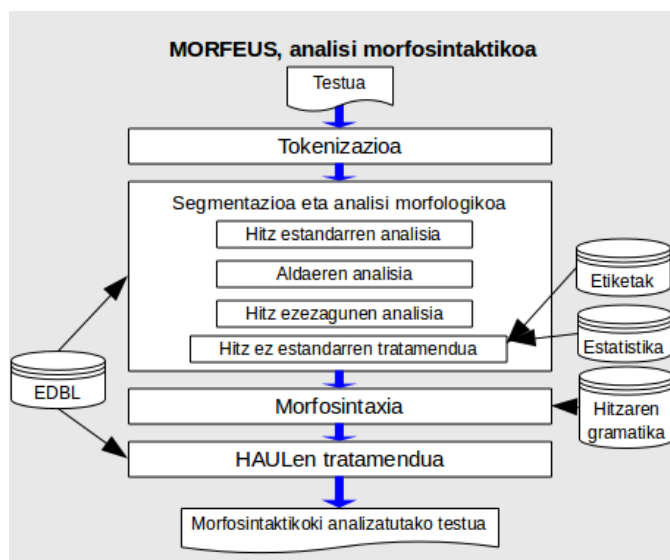
5.2 irudia – (353) adibideko esaldiaren tokenizazioa

5.1.2 Analisi morfosintaktikoa

Tokenizazioaren ondoren, analisi morfosintaktikoa egiten da Morfeus analizatzaile morfologikoaren bidez. Jarraian deskribatuko ditugu Morfeusek egiten dituen prozesamendu-urratsak eta 5.3 irudian tresnaren funtzionamenduaren eskema orokorra aurkeztuko dugu.

1. Lehenik analisi morfologikoa egiten da. Horretarako, segmentatzaileak hitz bakoitza lemetan eta morfemetan banatzen du eta osagai horien informazio morfologikoa ematen du EDBLko (Euskararen Datu Base Lexikala) (Aldezabal *et al.*, 2001) informazioa erabilta.

¹Lerro bertikalak bereizketa argiago adierazteko gehitu ditugu.



5.3 irudia – Morfeus analizatzaile morfologikoa

- Analizatzaile morfosintaktikoak aurreko fasean sortutako morfemetatik abiatuz hitz-formaren interpretazio posible bakoitzarentzako ezaugarri-egitura bat eskaintzen du. Hitzaren egitura deskribatzeko, testuingururik gabeko gramatika baliatzen da (Gojenola, 2000; Aduriz *et al.*, 2000).
- Hitz Anitzeko Unitate Lexikalak (HAUL) (Urizar, 2012) ere tratatzen dira morfosintaxiaren analisian. Hitz elkartuen, lokuzioen eta kolokazio murriztuen tratamendua egiten da eta, horretarako ere, EDBLko informazioa erabiltzen da. HAULak identifikatzeko HABIL (Urizar *et al.*, 2009) baliatzen da eta entitateen izenak mugatzeko eta sailkatzeko, EIHERA (Alegria *et al.*, 2003).

5.4 irudian (353) adibideko esaldiko hitzen analisi morfosintaktikoak agertzen dira. Irudian ikus daitekeenez, token bakoitzaren hitz-forma eta interpretazio bakoitzeko lema, informazio morfologikoa eta, dagokionean, funtzio sintaktikoa² @ ikurrarekin adierazita agertzen dira. Era berean, token ba-

²Funtzio sintaktikoen deskribapena eta horiek etiketatzeko etiketen zerrenda https://www.ix.a.eus/sites/default/files/dokumentuak/4083/Etiketen_eskuliburua_IXA.pdf helbidean aurki daiteke.

koitzari analisi posible guztiak esleitu zaizkio. Hurrengo urratsean desanbiguazio morfosintaktikoa egingo da hitz bakoitzari analisi bakarra esleitzeko.

```

/<Lehen>/<HAS_MAI>/
  ("lehen" ADB ARR w1,M-A-ADB-ARR-2)
  ("lehen" DET ORD w1,M-A-DET-ORD-3)
  ("lehen" DET ORD ABS MG w1,M-A-DET-ORD-4 @OBJ @PRED @SUBJ)
  ("lehen" IZE ARR w1,M-A-IZE-ARR-5)
  ("lehen" IZE ARR ABS MG w1,M-A-IZE-ARR-6 @OBJ @PRED @SUBJ)
/<Airbus>/<HAS_MAI>/
  ("Airbus" IZE LIB PLU- w2,M-A-IZE-LIB-3)
  ("Airbus" IZE LIB PLU- ABS NUMS MUGM w2,M-A-IZE-LIB-4 @OBJ @PRED @SUBJ)
/<A380>/<IDENT>/
  (" " /A380/ IZE ARR w3,M-G-IZE-ARR-7)
  (" " /A380/ IZE ARR ABS MG w3,M-G-IZE-ARR-8 @OBJ @PRED @SUBJ)
/<entregatu>/
  ("entregatu" ADI SIN DU PART w4,M-A-ADI-SIN-4)
  ("entregatu" ADI SIN BURU DU w4,M-A-ADI-SIN-5)
  ("entregatu" ADI SIN ABS MG PART w4,M-A-ADI-SIN-6 @-JADNAG_MP_OBJ @-JADNAG_MP_PRED @-JADNAG_MP_SUBJ)
/<da>/
  ("izan" ADL A1 NR_HURA w5,M-A-ADL-2)
  ("izan" ADT A1 NR_HURA PNT w5,M-A-ADT-2)
/<.>/<PUNT_PUNT>/

```

5.4 irudia – Morfeusen bidezko (353) adibideko esaldiaren analisi

5.1.3 Desanbiguazio morfologikoa

Morfeusek hitz bakoitzari egoki dakizkikeen analisi posible guztiak eskaintzen ditu, baina bakarra aukeratu behar da. Horretarako, EusTagger etiketatzailerak (Ezeiza *et al.*, 1998; Alegria *et al.*, 2002) hitz bakoitzari, testuinguruaren arabera, analisi zuzena esleitzen dio. Anbiguitasuna ebazteko, murritzapen-gramatika eta etiketen bidezko desanbiguazio estokastikoa erabiltzen ditu. Defektuz, gramatika-kategoria eta azpikategoria hartzen dira kontuan desanbiguaziorako. 5.5 irudian 5.4 irudiko hitz bakoitzaren analisi posible guztien artean hitz bakoitzarentzat bat aukeratu da.

EusTaggerren arkitektura 5.6 irudian irudikatu dugu. Ikus daitekeenez, EusTaggerrek desanbiguazio morfologikoa bi mailatan egiten du:

- Murritzapen Gramatiken bidez hizkuntza-ezagutzan oinarritutako desanbiguazioa egiten da (Aduriz *et al.*, 2000).
- Markoven eredu ezkutua (HMM) baliatuta, estatistikan oinarritutako dezanbiguazioa egiten da (Ezeiza, 2002).

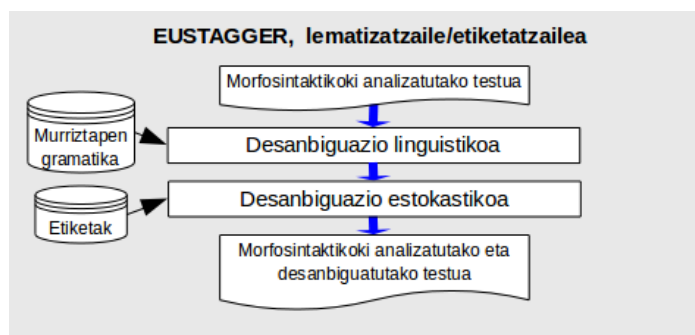
Bi desanbiguatzailerak emaitzak konbinatzen dira aukera zuzena lortzeko eta, bi desanbiguazio-moduluak exekutatu ostean anbiguitateak jarraitzen badu, sarrien lortzen den analisi esleitzen zaio hitzari.

```

/<Lehen>/<HAS_MAI>/
  ("lehen" DET ORD w1,L-A-DET-ORD-3,lsfi1 @ID>)
/<Airbus>/<HAS_MAI>/
  ("Airbus" IZE LIB PLU- w2,L-A-IZE-LIB-3,lsfi2 @KM>)
/<A380>/<IDENT>/
  (" " /A380/ IZE ARR ABS MG w3,L-G-IZE-ARR-8,lsfi3 @SUBJ)
/<entregatu>/
  ("entregatu" ADI SIN BURU DU w4,L-A-ADI-SIN-5,lsfi4 @-JADNAG)
/<da>/
  ("izan" ADL A1 NR_HURA w5,L-A-ADL-2,lsfi5 @+JADLAG)
/<.>/<PUNT_PUNT>/

```

5.5 irudia – EusTaggerren bidezko (353) adibideko esaldiaren analisi morfologiko desanbiguatua



5.6 irudia – EusTagger lematizatzaile/etiketatzailea

5.1.4 Funtzio sintaktikoen osaketa

EDBLko informazioan oinarrituta, aurreko urratsean funtzio sintaktikorik esleitu ez zaion analisi morfologiko bakoitzari funtzio sintaktiko bat esleitzen zaio. Horretarako Murriztapen Gramatikako islapen-erregelak (Arantzabe, 2008) baliatzen dira. Izan ere, Murriztapen Gramatikak esaldiko hitz guztiek funtzio sintaktikoa izatea eskatzen du, horiek hurrengo prozesamendu urratsean erabili ahal izateko.

Funtzio sintaktikoak 5.5 irudian analisi-lerro bakoitzean agertzen diren @-ren ondoko etiketak dira. @SUBJ, @OBJ eta @PRED etiketek subjektua, objektu zuzena eta predikatua adierazten dituzte hurrenez hurren. Horiek funtzio sintaktiko nagusiak dira. Esaldiko hitz guztiek funtzio sintaktiko bat behar

dutenez, ordea, @ID>³, @KM>⁴, @-JADNAG⁵ eta @+JADLAG⁶ etiketak ere sortu dira.

5.1.5 Desanbiguazio sintaktikoa

EusTaggerrek funtzio sintaktikoen desanbiguazioa erregela sintaktikoen birtartez (Aduriz *et al.*, 2004) egiten du. Hitz bati funtzio sintaktiko bat baino gehiago esleitzen zaionean, egokiak ez direnak baztertzen dira eta, ahal izanez gero, bakarra hautatzen da. Esaterako, 5.4 irudian *Airbus* hitzaren bigarren analisisian, @OBJ, @PRED eta @SUBJ esleitu zaizkio analisi horri.

5.1.6 Analisi sintaktiko partziala

Etiketatzeko morfologikoa egin ondoren, aurreko etiketatze sintaktikoan oinarrituta, perpausa zatietan, sintagma mailako kateetan, banatzen da. Urrats honetan hitzak beren sintagmetako edota kateetako beste elementuekin lotzen dira. Azaleko sintaxia Ixatiren bidez (Aduriz *et al.*, 2004) egiten da. Ixatik azaleko analisi sintaktikoa egiten du eta sintagmak, aditz-kateak, postposizio-lokuzioak eta entitate izendunak bereizten ditu. Era berean, sintagmen sailkapena ere egiten du.

```
"<Lehen>"<HAS_MAI>"
    "lehen" DET ORD ZERO HAS_MAI w1,L-A-DET-ORD-3,lsfi1 @ID> %SIH
"<Airbus>"<HAS_MAI>"
    "Airbus" IZE LIB PLU- ZERO HAS_MAI w2,L-A-IZE-LIB-3,lsfi2 @KM>
"<A380>"<IDENT>"
    EZEZAG "A380" IZE ARR ABS MG IDENT w3,L-G-IZE-ARR-8,lsfi3 @SUBJ %SIB
"<entregatu>"
    "entregatu" ADI SIN BURU DU NOTDEK w4,L-A-ADI-SIN-5,lsfi4 @-JADNAG %ADIKATHAS
"<da>"
    "izan" ADL A1 NOR NR_HURA w5,L-A-ADL-2,lsfi5 @+JADLAG %ADIKATBU
"<$.>"<PUNT_PUNT>"
    PUNT_PUNT
```

5.7 irudia – EusTaggerren bidezko (353) adibideko esaldiaren analisi sintaktiko partziala

³Ezkerreko determinatzailea

⁴Kasua daraman formaren modifikatzailea

⁵Aditz nagusi jokatu gabea

⁶Aditz laguntzaile jokatu

5.7 irudian, (353) adibideko esaldiaren analisi desanbiguatua agertzen da. Esaldiaren zatiak edo sintagmak ere agertzen dira %SIH eta %SIB etiketek sintagmaren hasiera eta amaiera adierazten dituzte hurrenez hurren eta %ADIKATHAS eta %ADIKATBU etiketek aditz-katearen hasiera eta amaiera adierazten dituzte.

5.1.7 Analisi sintaktiko osoa

Aurreko urratsetan identifikatu ditugun hitzak eta sintagmak bere gobernatzaileekin lotu behar dira; bestela esan, horien arteko dependentziak sortu behar dira. Dependentzia sintaktikoak perpauseko elementuen arteko mendekotasun-erlazioak dira (Aranzabe, 2008).

Zehazki, dependentziak lantzean, perpauseko zatien (sintagmak eta kateak) arteko mendekotasunak aztertzen dira, baita zati horien barruko elementuen artekoak ere. Sortzen diren erlazioak perpauseko elementuek betetzen duten funtzioaren arabera izendatzen dira. 5.8 irudian ikus daitezke *Lehen Airbus A380 entregatu da* esaldiko dependentziak lerroez adierazita. Lerro bakoitzaren ondoan azpimarratuta agertzen den hitzak dependentzia-mota adierazten du.

Ixa taldean sintaxiaren prozesamendurako metodo hibridoa erabiltzen da (Goenaga, 2017): trantsizioetan oinarritutako etiketatzaile sintaktikoa eta erregela bidezkoa uztartzen dira. Erregela bidezkoa (Bengoetxea eta Gojenola, 2007) Euskararako Dependentzia Gramatika Konputazionala (EDGK) gramatika (Aranzabe, 2008) erabiltzen duen analizatzaile sintaktikoa da.

Analizatzaile sintaktiko horrek hitzak binaka lotzen ditu esaldiaren zuhaitz sintaktikoa (dependentzia-zuhaitza) lortzeko. Esaldi bakoitzerako, esaldiz gaindiko *root* erroa definitzen da eta horri lotzen zaizkio esaldiko tokenak beren arteko dependentziak errespetatuz. Halaber, dependentzien analizatzaileak aldeztu aurretik definituta dauden dependentzia-erlazioei dagozkien etiketak esleitzen dizkie hitzei. Etiketa bakoitza dependentzia-erlazioaren izenarekin eta bere buruaren (gurasoa) norantzarekin (ezker edo eskuin) osatuta dago.

5.1.8 Analisi semantikoa

Dependentzia sintaktikoak bRol bidezko rol semantikoaren etiketatzean (Salaberri *et al.*, 2015) baliatzen dira. Rol semantikoak esaldiko predikatuetan parte-hartzen duten elementuen *eginkizun* edo *paper* semantikoak dira.

Dependentziak irudikatzeko baliatu dugun 5.8 irudian *Lehen Airbus A380 entregatu da* esaldiko rol semantikoak ere irudikatu ditugu. Esaldiak argumentu bakarra du, gaia. Biribilaren barruan *entregatu da* predikatuaren argumentu hori (ARG1) irudikatu dugu.

5.8 irudia – Lehen Airbus A380 entregatu da esaldiaren dependentzia-zuhaitza

5.2 EusHeidelTime: denbora-adierazpenak identifikatzeko eta normalizatzeko tresna

EusHeidelTime euskarazko denbora-adierazpenak identifikatzeko eta normalizatzeko tresna da. EusHeidelTime HeidelTime (Strötgen eta Gertz, 2013) oinarritzat hartuta sortu dugu. Zehazki, euskararako HeidelTimeren ereduari jarraitzea ondoko arrazoiengatik erabaki genuen:

- HeidelTime erregeletan oinarrituta dago. Euskarazko denbora-adierazpenak dituen corpusa erlatiboki txikia da eta corpus txikietan emaitza hobeak lortzen dituzte erregeletan oinarritutako tresnek.
- Iturburu-kodea hizkuntzarekiko independentea da. Euskararako erregelak eta patroiak definitu ditugu soilik.
- Modulu beregaina da eta erraz txerta daiteke prozesamendu-katean.
- TimeMLko **TIMEX3** etiketak sortzen ditu. EusTimeML TimeMLn oinarrituta dago eta denbora-adierazpenentzat **TIMEX3** etiketak ditu.
- Publikoki dago eskuragarri.

EusHeidelTime, beraz, erregeletan oinarritutako sistema da eta denbora-adierazpenei EusTimeMLko **TIMEX3** etiketak (4.2.2 atala) esleitzen dizkie. EusHeidelTimen denbora-adierazpenak hiru dimentsiodun entitatetzat hartzen dira:

- denbora-adierazpenaren forma
- denbora-adierazpenaren mota (data, ordua, iraupena edo errepikapena)
- denbora-adierazpenaren semantika edo balio normalizatua.

Aipatutako hiru elementu horiek nola kodetzen diren 5.9 irudian ikus daiteke. Bertan (354) adibideko *2005eko apirilaren 27a* denbora-adierazpenaren etiketatzea irudikatu da.

(354) 2005eko apirilaren 27a

<TIMEX3 type="DATE" value="2005-04-27" tid="t1"> 2005eko apirilaren 27a </TIMEX3>

5.9 irudia – 2005eko apirilaren 27a denbora-adierazpenaren etiketatzea

Honela etiketatu da (354) adibidea: i) denbora-adierazpenaren hedape-
na bi TIMEX3 etiketaren artean agertzen da, ii) DATE motakoa da eta iii)
2005-04-27 balio normalizatua du. Gainera, EusHeidelTimek tid identifika-
tzailea ere automatikoki esleitzen dio denbora-adierazpenari, identifikatzaile
hori prozesamenduan baliatzeko.

Denbora-adierazpenen informazioa sakonago adierazteko, tresnak badu
TI- MEX3 etiketako mod (4.2.2.5 atala), freq eta quant (4.2.2.4 atala) atribu-
tuak kudeatzeko gaitasuna ere. Hala (355) adibideko *urtero* denbora-adie-
razpenaren kuantifikatzailea (quant) ere islatu ahal izan dugu. 5.10 irudian
urtero denbora-adierazpenaren etiketatzea irudikatu dugu.

(355) Urtero

<TIMEX3 type="SET" value="P1Y" quant="EVERY" tid="t1"> Urtero </TIMEX3>

5.10 irudia – *urtero* denbora-adierazpenaren etiketatzea

Aipatu dugunez, HeidelTimen bertsio guztiek iturburu-kode bera dute.
Ondorioz, HeidelTime hizkuntza bateko denbora-adierazpenak identifikatze-
ko eta normalizatzeko erabili nahi denean, hizkuntza horretako erregelak eta
patroiak sortu behar dira. Jarraian, euskarazko denbora-adierazpenak trata-
tzeko garatu ditugun baliabide horiek deskribatuko ditugu.

5.2.1 EusHeidelTimerako erregelak, patroiak eta balio normalizatuak

EusHeidelTime erregeletan oinarritutako sistema da. Erregelak definitu au-
rretik, kontuan hartu behar dira erregeletan denbora-adierazpenak errepre-
sentatzeko erabiltzen diren patroiak eta horien balio normalizatuak.

5.2.1.1 Patroiak

Patroiak erregela guztietan balia daitezkeen adierazpen erregularrak dira.
Ondorioz, patroieta, antzeko jokaera duten hitzak batu ditugu (asteko egu-

nak, hilabeteak, urtaroak, zenbakiak, adberbioak, etab.), horiek perpausean egitura bertsuetan agertzen baitira. Hala, erregetan multzoaren izena erabil daiteke, multzoko elementu guztiak erregelan bertan aipatu orde. Adibidez, 5.11 irudian ikus daitezke `resources_repattern_reWeekday` patroiko elementuak, kasu horretan asteko egunen izenak maiuskulaz zein minuskulaz hasita.

```
"[Aa]stelehen"  
"[Aa]stearte"  
"[Aa]steazken"  
"[Oo]stegun"  
"[Oo]stiral"  
"[Ll]arunbat"  
"[Ii]gande"
```

5.11 irudia – %reWeekday asteko egunen patroia

Patroietan zerrendako elementu bakoitza lerro batean adierazten da eta Heidelbergeren sintaxi-operatzaileak onartzen ditu:

- // lerroa komentatzeko
- [] eta | hautaketarako

Esaterako, 5.11 irudian, hitz hasierako maiuskula/minuskula alternantzia tratatzeko bi aukerak kortxete artean adierazi dira.

Denbora-adierazpenen formak aztertu behar izan ditugu patroia-bilduma osatzeko. 5.11 irudian ikus daitekeenez, astegunen lema zerrendatu ditugu. Horiek, ordea, ez dira bere horretan gauzatzen euskaraz normalean. Euskara hizkuntza eranskaria denez, lema horiei kasu gehienetan postposizio-atzizkiak gehitzen zaizkie, baina postposizio atzizki horiek ez zaizkie edonola gehitzen lemei. Lema eta postposizioak lotzeko arau morfofonologikoak aztertu ditugu patroiak eraikitzean.

Aipatu dugunez, EusHeidelTimek azaleko analisia egiten du eta erregelek testuan agertzen diren karaktere-segidak identifikatzen dituzte. Horrek ez du esan nahi EusHeidelTimeko erregelak sortzean ezagutza linguistikoa baliatu ez denik. Hain zuzen ere, EusHeidelTimeko patroietan arau morfofonologikoak aintzat hartzean, euskararen morfologiaren tratamenduan (Alegria, 1995; Alegria eta Urkia, 2002) definitu diren bi hizkuntza mailak, morfemen maila eta hitzaren maila, batu ditugu. Lema eta postposizioak morfemen mailako elementuak dira eta testuan aurkitzen ditugun formak hitzen mailakoak. Maila lexikotik azaleko mailara pasatzeko arau morfofonologikoak

baliatzen dira. Horren bidez, lema eta postposizioen forma lexikoek aldaketa morfofonologikoak izaten dituzte.

Maila lexikoko informazioa azaleko mailan nola irudikatzen den aztertu dugu patroiak sortzean. Patroiak sortzean, aldaketa morfofonologiko horiek kontuan hartu eta patroiz-zerrendetan alomorfoak gehitu behar izan ditugu. Patroiak sortzean ondoko arau morfofonologikoak izan ditugu kontuan:

- Postposizio-atzizkiak antolatzeke 4 patroiz-zerrenda sortu ditugu: orokorra **ReMarkak**, artikulua singularrarekin eratzen direnak **ReSingularra**, pluralarekin eratzen direnak **RePlurala** eta mugagabearekin eratzen direnak **ReMugagabea**.
- Postposizio-zerrendetan ondoko ezaugarri morfofonologikoei erreparatu diegu postposizioen alomorfoak sortzean:
 - K-ren ozenketa sudurkarien eta albokarien atzetik: *lehen + ko > lehengo*
 - E epentetikoa kontsonantez amaitutako lemen atzetik: *bost + ko > bosteko*
 - Determinatzailearen forma A organikoa duten hitzetan: *hamaika + a + n > hamaikan; hamaika + ak + ko > hamaiketako*
- Erroen zerrendetan ezaugarri morfofonologiko hauek hartu ditugu kontuan:
 - Hitz amaierako R gogorra duten erroen bi formak sortu ditugu: *hamar* eta *hamarr*
 - *Hiru* eta *lau* zenbakien *hirur* eta *laur* formak gehitu ditugu, zenbaki horiek zenbatzaile bezala funtzionatzean forma horiek hartzen baitituzte⁷.
 - A berezkoa duten lemen bi forma sortu ditugu determinatzailea egoki gehitu ahal izateko: *hamaika* eta *hamaik*

5.12 irudian orduen patroiak adierazi ditugu. Ikus daitekeenez, orduen patroietan aurreko arau morfofonologikoak islatu ditugu.

Patroiek denbora adierazten dutenez, balio normalizatuak hartzen dituzte. Jarraian deskribatuko ditugu patroietako elementuen balio normalizatuak batzen dituzten baliabideak.

⁷http://www.euskaltzaindia.eus/dok/arauak/Araua_0008.pdf

```
"puntu", "00"  
"bat", "01"  
"bi", "02"  
"hiru", "03"  
"hirur", "03"  
"lau", "04"  
"laur", "04"  
"bost", "05"  
"sei", "06"  
"zazpi", "07"  
"zortzi", "08"  
"bederatzi", "09"  
"hamar", "10"  
"hamarr", "10"  
"hamaik", "11"  
"hamaika", "11"  
"hamabi", "12"
```

5.12 irudia – ISO balio normalizatuak

5.2.1.2 Balio normalizatuak

Balio normalizatuen fitxategiak patroieta elementuen balio normalizatuak batzen dituzten baliabideak dira. Denbora-adierazpenak deskribatzean, adierazten duten unea edo tartearen normalizatuan ISO-8601 arauari jarraituta adieraz daitekeela ikusi dugu. Hala, balio normalizatuen zerrendei esker patroieta elementuei ere balio normalizatuak esleitu dizkiegu.

5.13 irudian ikus daitekeenez, asteko egunei posizioaren arabera balio normalizatuak esleitu diegu. Balio normalizatuen zerrenda bakoitzean, patroiak eta beren balio normalizatuak agertzen dira batera; lehenik, patroia komatxo artean eta, bigarrenik, komaren ondoren, bigarren balio normalizatuak komatxo artean. Normalean, balio normalizatuak ISO balioak izaten dira. 5.13 irudiko asteko egunen balio normalizatuak astearen arabera egituretan (XXXX-WXX-1) erabil ditzakegu.

"astelehen", "1"	"Astelehen", "1"
"astearte", "2"	"Astearte", "2"
"asteazken", "3"	"Asteazken", "3"
"ostegun", "4"	"Ostegun", "4"
"ostiral", "5"	"Ostiral", "5"
"larunbat", "6"	"Larunbat", "6"
"igande", "7"	"Igande", "7"

5.13 irudia – Asteko egunen ISO balio normalizatuak

Beste kasu batzuetan, ordea, denbora-adierazpenei ezin zaie zuzenean ba-

lio normalizaturik eman. Horiek prozesamenduko bigarren fasean normalizatu diren adierazpen erlatiboak dira. Esaterako, 5.14. irudian `UNDEF-this-day` modukoak agertzen dira adierazteko behin denbora-aingura definituta horren arabera balioa hartu behar duela denbora-adierazpenak. Hala, *gaur* hitzaren normalizazioa lehen fase batean, `UNDEF-this-day` izango da eta bigarren fase batean, `XXXX-XX-XX` formatua emango zaio. Egunaren mailako granularitatedun denbora-adierazpena denez, *gauri day* esleitzen zaio tarteko normalizazio-fasean.

```
"gaur", "UNDEF-this-day"
"Gaur", "UNDEF-this-day"
"bihar", "UNDEF-next-day"
"Bihar", "UNDEF-next-day"
"atzo", "UNDEF-last-day"
"Atzo", "UNDEF-last-day"

"herenegun", "UNDEF-this-day-MINUS-2"
"Herenegun", "UNDEF-this-day-MINUS-2"
"etzi", "UNDEF-this-day-PLUS-2"
"Etzi", "UNDEF-this-day-PLUS-2"
"etzidamu", "UNDEF-this-day-PLUS-3"
"Etzidamu", "UNDEF-this-day-PLUS-3"
```

5.14 irudia – Data erlatiboentzako balio normalizatuak

Patroi-fitxategietan gertatzen den bezala, patroi-balio normalizatu bikote bakoitza lerro batean adierazten da eta HeidelbergTimeren sintaxi operatzaileak onartzen ditu.

5.2.1.3 Erregelak

Patroi horiek eta balio normalizatu horiek erregeletan erabili ditugu. EusHeidelTimen lau erregela-multzo dago, bat denbora-adierazpen mota bakoitzeko: datak, orduak, iraupenak eta errepikapenak.

Sailkapen horrek ematen dio denbora-adierazpenari lehen informazioa. Izan ere, EusHeidelTimek denbora-adierazpenei klasea erregela hautatu duen multzoaren arabera esleitzen die. Hau da, denbora-adierazpena daten multzoko erregela baten bidez normalizatzen bada, `DATE` esleitzen dio.

Erregelek sintaxi finkoa dute: lehenik, izena adierazten da; bigarrenik, testuan identifikatu behar den segida; hirugarrenik, denbora-adierazpenari eman behar zaion balio normalizatua eta, laugarrenik, murriztapenak, behar izanez gero.

```
RULENAME="Data_partz_beste",  
EXTRACTION="%reMonth%reSingularra %reDayNumber%reSingularra",  
NORM_VALUE="UNDEF-this-year-%normMonth(group(1))-%normDay(group(3))"
```

5.15 irudia – Hiru elementu nagusiak dituen erregela “arrunt” bat

```
RULENAME="egun_parte",  
EXTRACTION="[Ee]gun%reSingularra %rePartWords(%reMarkak)?",  
NORM_VALUE="UNDEF-this-day", NORM_MOD="%normPartWords(group(2))"
```

5.16 irudia – Modifikatzailea normalizatzeko informazioa duen erregela

Hiru elementu nagusiak dituen erregela “arrunt” bat ikus daiteke 5.15 irudian; zehazki *apirilaren 7ko* moduko denbora-adierazpenak identifikatzeko. Aurretik aipatutako erregelaren izena (RULENAME), identifikatzeko segida (EXTRACTION) eta balio normalizatuak (NORM_VALUE) agertzen dira.

Erregelaren sintaxiari dagokionez, identifikatzeko segidan %reMonth eta %reDayNumber patroiak agertzen dira, hilabeteak eta zenbakiz adierazitako egunak identifikatzeko. Balio normalizatuari erreparatuz gero, %normMonth (group(1)) eta %normDay(group(3)) elementuak agertzen direla ikus daiteke. Horiek esan nahi dute balio normalizatuan posizio horietan denbora-adierazpeneko lehen eta hirugarren elementuen balio normalizatuak jarriko direla, hurrenez hurren. group horiek hutsuneen arteko karaktere segidak errepresentatzen dituzte. Kasu horretan, group(1)-ek %reMonth patroia balio normalizatua hartuko du %normMonth balio normalizatuen zerrendaren arabera, eta gauza bera gertatuko da group(3)-rekin: %reDayNumber patroiak identifikatzen duen elementua %normDay zerrendako balio normalizatuaz normalizatuko da.

Erregela arrunten bidez denbora-adierazpenei mota eta balio normalizatua esleitzen zaie, baina informazio gehiago identifikatzeko erregela konplexuagoak ere badaude. Adibidez, modifikatzailea (mod) (ikus 4.2.2.5 azpiatala) ere normalizatzen duen erregela bat agertzen da 5.16 irudian. Erregela horrek *egunaren hasieran* moduko denbora-adierazpenak identifikatzen ditu, *hasieran* moduko modifikatzaileak dituenak. Adibide horretan, NORM_MOD murriztapena agertzen da zeinaren bidez modifikatzailea gehitzen baitzaio denbora-adierazpenari.

Modifikatzaileaz gain, errepikapenen frekuentzia (**freq**) eta errepikapenen kopurua (**quant**) ere adieraz daitezke. Hala, informazio gehigarria gehitzea ahalbidetzen duten hiru erregela mota daude:

- **NORM_MOD**: **mod** atributua definitzeko, **NORM_MOD** gehigarria gehitzen da (5.17 irudia).

```
RULENAME="hilabetearen_hasieran",  
EXTRACTION="%reMonth%reSingularra %rePartWords%reMarkak",  
NORM_VALUE="UNDEF-year-%normMonth(group(1))",  
NORM_MOD="%normPartWords(group(3))"
```

5.17 irudia – Modifikatzailea normalizatzeko **NORM_MOD** gehigarria duen erregela

- **NORM_FREQ**: **freq** atributua definitzeko, **NORM_FREQ** gehitzen da (5.18 irudia).

```
RULENAME="set_r2a",  
EXTRACTION="%reUnit(an|ean) %reFrek",  
NORM_VALUE="P1%normUnit(group(1))",  
NORM_FREQ="%normFrek(group(3))"
```

5.18 irudia – Maiztasuna normalizatzeko **NORM_FREQ** gehigarria duen erregela

- **NORM_QUANT**: **quant** atributua definitzeko, **NORM_QUANT** gehitzen da (5.19 irudia).

```
RULENAME="set_r1a",  
EXTRACTION="%reUnit(ero|ro)",  
NORM_VALUE="P1%normUnit(group(1))",  
NORM_QUANT="EVERY"
```

5.19 irudia – Errepikapenen kopurua normalizatzeko **NORM_QUANT** gehigarria duen erregela

```
RULENAME="date_r2a_negation",  
EXTRACTION="(urri)%reMarkak",  
NORM_VALUE="REMOVE",POS_CONSTRAINT="group(1):G:"
```

5.20 irudia – Gramatika-kategoriaren murriztapena duen erregela

Informazio gehigarria normalizatzen duten elementu horiek elkarrekin konbina daitezke eta erregela batean bat baino gehiago ager daitezke.

Era berean, EusHeidelTimek denbora-adierazpenen identifikaziorako murriztapenak ere onartzen ditu:

- gramatika-kategoriaren murriztapena: `POS_CONSTRAINT`
(`group(x):y:`): x taldearen gramatika-kategoria y izan behar da. Esaterako, 5.20 irudian, erregelak soilik funtzionatuko du, baldin eta denbora-adierazpenaren lehen elementua, *urri*, adjektiboa (**G**) bada.
- denbora-adierazpenaren hasieraren murriztapena: `OFFSET(group(x)-group(y))`. Erauzi behar den segida x taldearekin hasi eta y taldearekin amaitzen da. EusHeidelTimen, ordea, ez dugu `OFFSET` murriztapenik baliatu erregeletan, ez baitugu horren beharrik izan euskarazko denbora-adierazpenak identifikatzeko erregelak sortzeko.

Azkenik, erregela negatiboak sortu ditugu. Horien kasuan, segida bat aurkitzean ezer ez egitea agintzen du erregelak. 5.20 irudiko erregelaren arabera, *urri* adjektiboa aurkitzean ez da ezer egin behar. Ikusten denez, balio normalizatua `REMOVE` da, hau da, ez du balio normalizaturik. Azaldu bezala, erregela horrek *urri* adjektiboa denean bakarrik eragin dezan, kategoria gramatikalkaren araberrako murriztapena ezarri diogu (`POS_CONSTRAINT`). Hala, segidako lehen elementuak “urri” forma badu eta adjektiboa (**G**) bada, ez du etiketarik hartuko. Murriztapen hori jarrita, *urri* erroak “urteko hamargarren hilabetea” adierazten duenean bakarrik hartuko du `TIMEX3` etiketa, ez, ordea, “eskas” adierazten duenean.

Egun, EusHeidelTimek 313 erregela, 58 patroi-zerrenda eta 29 normalizazio-fitxategi baliatzen ditu. Horien banaketa 5.1 taulan ageri da.

5.1 taula – EusHeidelTimerako baliabide linguistikoak

Baliabide mota	Kopurua
Erregelak	
DATE	142
TIME	64
DURATION	101
SET	6
Patroi-zerrendak	58
Normalizazio-balioen zerrendak	29

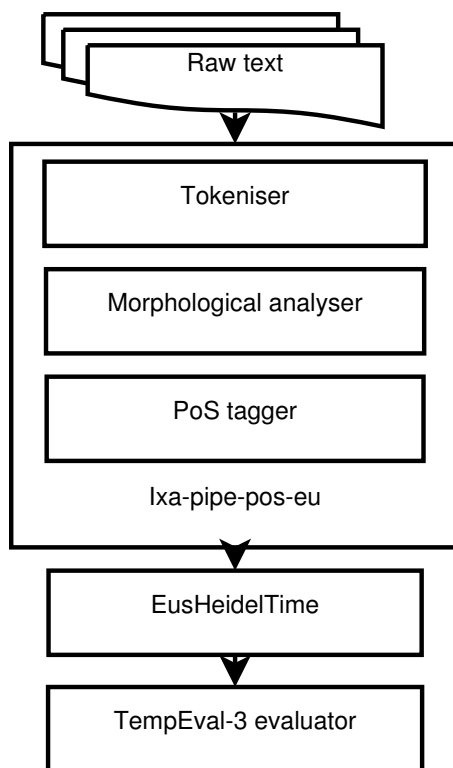
5.2.2 EusHeidelTimeren integrazioa beste tresnetan eta funtzionamendua

HeidelTime jatorriz UIMA (Unstructured Information Management Architecture) (Ferrucci eta Lally, 2004) osagaitzat garatu zuten, UIMA estiloko prozesamendu-kateetan txertatzeko. Strötgenen eta Gertzen lanean (2010) azaltzen denez, ingeleserako UIMA kateak esaldi-banatzailea, tokenizatzailea eta OpenNLP analizatzaile morfologikoa ditu denbora-informazioaren etiketatzaileak balia ditzan. Euskararako, aitzitik, aurreprozesamendurako modulu horiek ixaKat katean integratuta daudenez, EusHeidelTime ixaKaten txertatzeko egokitu dugu.

Zehazki, EusHeidelTime euskararen prozesamendurako bi katetan txertatu dugu: ixaKaten eta ixa-pipe-pos-eu prozesamendu-katean. Bigarren kate horrek ez du hain kalitate oneko analisia lortzen, baina askoz ere prozesamendu azkarragoa eskaintzen du. EusHeidelTime era isolatuan erabili dugunean, ixa-pipe-pos-eu-ren⁸ analisia baliatu dugu, prozesamendu azkarragoa ahalbidetzen baitu. 5.21 irudian EusHeidelTime ixa-pipe-pos-eu katean nola txertatu dugun irudikatu dugu. Halaber, EusHeidelTime bTime denbora-informazioa erauzteko sisteman (Salaberri Izko, 2017) integratu dugu. EusHeidelTime bTimeren barruan erabili dugunean (ikus 5.3.1), ordea, ixaKateko analisia baliatu dugu. Hala ere, bi kateek antzeko funtzionamendua dute eta bi kateetako modulu guztiek NAF formatua (Fokkens *et al.*, 2014), HPrako kate konplexuetarako garatutako etiketatze formatua, irakurri eta idazten dute.

EusHeidelTimeren funtzionamenduari dagokionez, EusHeidelTimek gra-

⁸<http://ixa2.si.ehu.es/ixakat/>



5.21 irudia – EusHeidelTime Ixa-pipe-pos-eu katean integratuta

matika-kategoriak markatuta dituzten dokumentuak hartzen ditu sarrera gisa. EusHeidelTimek testuak tokenka irakurtzen ditu eta testuko tokenekin bat datozen erregela guztiak aplikatzen ditu, azkenean erregela “egokiena” soilik aplikatzen badu ere. Erregelen hautapen-parametroak iturburu-kodean defini daitezke. Defektuz, erregelak modu sekuentzialean aplikatzen ditu eta erregela multzoei ondoko lehentasuna ematen die DATE > TIME > DURATION > SET. EusHeidelTime garatzean defektuzko parametroak mantendu ditugu.

Ddenbora-adierazpenen informazioaren erauzketari dagokionez, lehenik denbora-adierazpenaren mota, **type** atributuaren balioa, fitxategiak ematen dio. Hau da, iraupenen erregela-fitxategiko erregela batek denbora-adierazpena identifikatzen badu, denbora-adierazpen horri **DURATION** mota esleituko zaio. Ondoren, normalizazio-erregelen bidez, denbora-adierazpeneko elementu bakoitzari bere balio normalizatua ematen zaio.

Azkeneko emaitza denbora-adierazpenak etiketatuta dituen testua da.

EusHeidelTime parametrizatu dugu denbora-informazioa NAF zein TimeML formatuetan lortzeko. Bi formatuak aurkeztu ditugu 5.22 eta 5.23 irudietan. 5.22 irudian NAF formatuko dokumentu baten zati bat agertzen da. Goian tokenen informazio linguistikoa ikusten da eta behean denbora-adierazpenen etiketatzea agertzen da. 5.23 irudian, ordea, TimeML formatuan etiketatutako testua agertzen da. Goiko aldean testuaren metainformazioa azaltzen da eta jarraian testua eta testuan txertatutako etiketak.

EusHeidelTimek bi formatuak bi zereginetarako eskaintzen ditu. Bate-tik, arestian aipatu dugunez, euskararen prozesamendurako moduluek, bai-ta EusHeidelTimek ere, NAF formatuko fitxategiak hartzen dituzte sarrera edo irteera gisa. Ondorioz, EusHeidelTimek ere NAF irakurtzen eta sortzen du, denbora-adierazpenak tresna aurreratuetan interpretatu ahal izateko. Bestetik, TimeML EusHeidelTimeren etiketatzea ebaluatzeko erabili dugu. Denbora-adierazpenen erauzketa eta normalizazioa ebaluatzeko, TempEval-3 (UzZaman *et al.*, 2012) atazako ebaluatzailea estandar bihurtu da eta horrek TimeML formatuko fitxategiak hartzen ditu.

Ebaluazioaz ari garenez, EusHeidelTimeren erregelen egokitasuna neur-tzeko, tresna ebaluatu dugu. Horretarako, EusTimeBank corpuseko ebalua-ziorako azpicorpuseko 17 dokumentuko lagina hartu dugu. Hain zuzen ere, azpicorpus horretako albisteetan (FaCor eta EusMEANTIME) eta historia-testuetan (WikiWarsEU) zein emaitza lortzen duen aztertu dugu. Denbora-adierazpenen identifikazio osoan (*strict match*) eta partzialean (*relaxed match*), eta balio normalizatua eta mota esleitzean lortzen duen asmatze-tasa neur-tu ditugu. Horretarako, doitasuna (*precision*, P), estaldura (*recall*, R) eta F-neurria neurtu ditugu. EusHeidelTimeren emaitzak 5.2 taulan adierazi ditugu.

5.2 taula – EusHeidelTimeren emaitzak

	FaCor			EusMEANTIME			WikiWarsEU		
	P	R	F_1	P	R	F_1	P	R	F_1
Strict match	79,39	83,64	81,42	81,4	74,47	77,78	77,98	87,8	82,6
Relaxed match	87,93	92,73	90,27	93,02	85,11	88,89	82,67	93,09	87,57
value			58,41			64,44			74,57
type			83,19			82,22			86,81

5.2 taulan ikusten denez, EusHeildeTime denbora-adierazpenen % 80 in-guru osorik identifikatzeko gai da. Denbora-adierazpenak partzialki identifi-katuz gero, ordea, denbora-adierazpenen % 90 inguru aurkitzen ditu. Balioa-

```

<!--atzero-->
<term id="t343" lemma="atze" pos="N" morphofeat="NC0AS000" case="IZE ARR ALA NUMS MUGM @ADLG">
  <span>
    <target id="w343"/>
  </span>
</term>
<!--fakturazioan-->
<term id="t344" lemma="fakturazio" pos="N" morphofeat="NC0IS000" case="IZE ARR INE NUMS MUGM @ADLG">
  <span>
    <target id="w344"/>
  </span>
</term>
<!--.-->
<term id="t345" lemma="." pos="O" morphofeat="Fp" case="PUNT_PUNT">
  <span>
    <target id="w345"/>
  </span>
</term>
</terms>
<timeExpressions>
  <timex3 id="tx1" type="DATE" functionInDocument="CREATION_TIME" value="2013-10-16"/>
  <timex3 id="tx2" type="TIME" value="2013-10-16T17:05:59">
    <!--2013-10-16- 17:05:59-->
    <span>
      <target id="w3"/>
      <target id="w4"/>
      <target id="w5"/>
    </span>
  </timex3>
  <timex3 id="tx3" type="TIME" value="2013-10-16TMO">
    <!--goizean-->
    <span>
      <target id="w20"/>
    </span>
  </timex3>
  <timex3 id="tx4" type="DATE" value="FUTURE_REF">
    <!--etorkizuna-->
    <span>
      <target id="w72"/>
    </span>
  </timex3>
  <timex3 id="tx5" type="DATE" value="2013-W42">
    <!--aste honetan-->
    <span>
      <target id="w116"/>
      <target id="w117"/>
    </span>
  </timex3>
  <timex3 id="tx6" type="DATE" value="2013-05">

```

5.22 irudia – EusHeidelTimeren bidez etiketatutako testua NAF formatuan

ri dagokionez, emaitzarik onenak testu historikoetan lortzen ditu (% 74,57), testu horietan agertzen baitira denbora-adierazpen absolutu gehien. Kaze-taritzat testuetan (% 58,41 eta % 64,44), aldiz, denbora-adierazpen erlatibo gehiago eta zailagoa da horiei balio normalizatua esleitzea. Mota esleitzea ondo egiten du kasuen % 80tik gora.

```

<?xml version="1.0" ?>
<TimeML xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="http://
timeML.org/timeMLdocs/TimeML_1.2.1.xsd">
<DOCID>01-fagor.txt.xml</DOCID>

<DCT><TIMEX3 tid="t0" type="TIME" value="2013-10-16" temporalFunction="false"
functionInDocument="CREATION_TIME">2013-10-16</TIMEX3></DCT>

<TEXT>

Xabier Martin <TIMEX3 type="TIME" value="2013-10-16T17:05:59" tid="t1">2013-10-16 - 17:05:59</
TIMEX3> Fagor Etxetresnak kooperatibak konkurtso legera jo du Fagor Etxetresnen bazkide
kooperatibistek batzarra egin dute <TIMEX3 type="TIME" value="2013-10-16T00" tid="t2">goizean</
TIMEX3> . Mondragon Inbertsioak erakundearen erabakiak aztertu ditu Enpresak beharko lukeen diru
laguntzaren inguruan . Ez da bigarren erreskatea izango , eta Fagorrek ordaintzeari utzi dio .
Mondragon taldeko ikurra oso larri dago , 800 milioi euroko zorrari buelta eman ezinik . Itxieraren
atarian egon daiteke , beraz . Fagor Etxetresnak kooperatibaren <TIMEX3 type="DATE"
value="FUTURE_REF" tid="t3">etorkizuna</TIMEX3> kolokan dago . " Fagor Etxetresnek , Lehiaketa
legearen 5.3 artikulua babesean , Donostiako Merkataritza arloko epaitegian jakinarazi du zorra
berregituratzeko alde aurretiko negoziazioei ekin diela " , azaltzen du kooperatibak hedabideei
igorritako oharra . Fagor Etxetresnak Mondragonen bigarren erreskatearen zain egon da <TIMEX3
type="DATE" value="2013-W42" tid="t4">aste honetan</TIMEX3> , haren bidez hartzekodunekin dituen
zorrak ordaintzen hasteko esperantzan , eta jadanik zenbait plantatan produkzioa etenda dagoenez ,
haiek berriro martxan jartzeko . Baina , hura iritsi ez , eta konkurtso legera jo du . <TIMEX3
type="DATE" value="2013-05" tid="t5">Maiatzean</TIMEX3> 70 milioi euroko funtsa eratu zuen
Mondragon taldeak kooperatiba guztien oniritziarekin , Fagorren alde jartzeko . Bigarren erreskate
hori ez da gauzatu <TIMEX3 type="DATE" value="PRESENT_REF" tid="t6">orain</TIMEX3> , eta Fagorri
urgentziatzeko likidez hori iritsiko ez zaionez , erabaki traumatikoak hartu beharko ditu
seguruenera , planta ixteari eta lanpostuak galtzeari dagokionez . Konkurtso legera jo duenez ,
<TIMEX3 type="DURATION" value="P4M" tid="t7">lau hilabeteko epea</TIMEX3> izango du hartzekodunekin
akordioa erdiesteko . Hori da negoziazio <TIMEX3 type="DURATION" value="PXX" tid="t8">epea</
TIMEX3> . Hortik aurrera konkurtso legeak aurrera egingo du . Fagor Etxetresnak 5.600 langile ditu
mundu osoko hamahiru herrialdetan , 1.600dik gora Euskal Herriko bost lantegietan ; horietako
gehiak bazkide dira , gainera . Lanpostu horiek guztiak kolokan daude <TIMEX3 type="DATE"
value="PRESENT_REF" tid="t9">gaurtik</TIMEX3> , Arrasateko kooperatibak oso zail daukalako
jarduerarekin jarraitzea . Lanpostua galduko balute , langile bazkideek ez lukete langabezia saria
jasotzerik izango , baina Mondrago taldeko aseguratzaile Lagun Aroren funts baten bidez euren
soldaten 80ko ordainketak jasoko lituzkete <TIMEX3 type="DURATION" value="PXX" tid="t10">epe batez</
TIMEX3> Enpresak 60,4 milioi euroko galera izan du <TIMEX3 type="DATE" value="2013-06"
tid="t11">ekainera</TIMEX3> arte etxetresnegileak , <TIMEX3 type="DATE" value="2012"
tid="t12">iazko</TIMEX3> <TIMEX3 type="DURATION" value="PXX" tid="t13">epe</TIMEX3> bereko galera
( 20,3 milioi ) halako hiru , eta negozio kopuruak 19 egin du behera , 609 milioi eurotik 491
milioira . <TIMEX3 type="DATE" value="2012" tid="t14">2012an</TIMEX3> 89 milioi galera izan zuen
eta 9 egin zuen atzera fakturazioan . </TEXT>

</TimeML>

```

5.23 irudia – EusHeidelTimeren bidez etiketatutako testua TimeML formatuan

5.3 KroniXa: denbora-lerroak sortzeko tresna

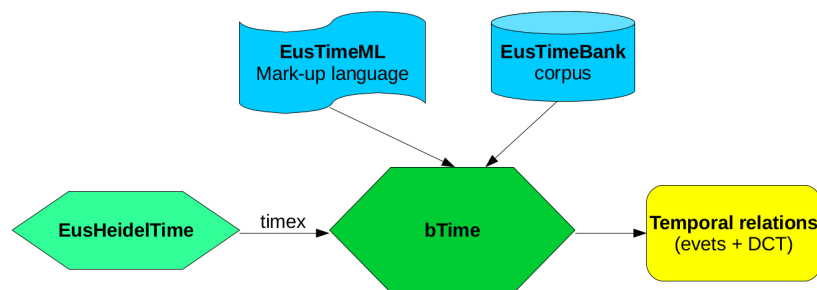
Lan honen sarreran aipatu bezala, denbora-informazioa erauzi ondoren, informazio hori baliatuko duten tresna aurreratuak garatzea dugu helburu. Tresna aurreratu horietako bat KroniXa da, denbora-lerroak sortzeko baliatu duguna. Minard *et al.*-ek (2015) definitzen dutenez, *storyline*ak zer, nori, noiz eta non gertatu den adierazten duten laburpenak dira. Laburpen horietan, denbora-lerroetan gertaerak kronologikoki antolatzen dira gertaeren eta gertatu diren denbora-uneen arteko erlazioak baliatuta.

Jarraian, 5.3.1 atalean, denbora-lerroak automatikoki sortzeko esperimentuan baliatutako denbora-informazioa nola lortu dugun deskribatuko dugu. 5.3.2 atalean, informazio hori baliatuta lortutako denbora-lerroak deskribatuko ditugu. 5.3.3 atalean, hasieran sortutako denbora-lerro horiek ebaluatzeko eta hobetzeko egin dugun esperimentazioa aurkeztuko dugu.

5.3.1 Denbora-informazioaren tratamendu automatikoa

Denbora-informazioaren tratamendu automatikoa EusHeidelTimeren bidez eta bTimeren (Salaberri Izko, 2017) bidez egin da. EusHeidelTimek denbora-adierazpenak erauzi eta normalizatzen ditu eta bTimek gertaerak eta horien eta denbora-adierazpenen artean sortzen diren denbora-erlazioak tratatzen ditu. Zehazki, aditz jokatz adierazitako gertaeren eta DCTaren arteko denbora-erlazioak baino ez dira kontuan hartzen bTimen. bTime sistemen arkitektura irudikatu dugu 5.24 irudian.

bTimen, gertaerak eta denbora-erlazioak erauzteko ikasketa automatikoko moduluak sortu dira. Bi modulu horiek euskarri-bektoredun makinetan (Cortes eta Vapnik, 1995) oinarritzen dira eta gertaeren formak eta gertaeren ezaugarriak (ikus 3.1.2 atala) erabiltzen dituzte sailkapena egiteko. bTime EusTimeBank corpusarekin (4.3 atala) entrenatu eta ebaluatu dugu.



5.24 irudia – bTime sistemen arkitektura

Denbora-lerroak sortzeko esperimentuan, EusTimeBank corpusaren entrenamendurako azpicorpusa (ikus 4.3 atala) erabili dugu bTime entrenatzeko. Entrenamendurako corpusean bTimeren hainbat konfigurazio ebaluatu ditugu eta, emaitzarik onenak lortzen dituen baliatuta, ebaluazio-corpusa etiketatu dugu. Halaber EusHeidelTimeren bidezko ebaluazio-corpusaren etiketatzea egin dugu denbora-adierazpenak lortzeko. Gero, EusHeidelTime eta

bTime EusTimeBankeko ebaluazio corpusean neurtu ditugu KroniXak jasoko duen informazioaren kalitatea neurtzeko. Jarraian aurkeztuko ditugu emaitzak.

EusHeidelTimek denbora-adierazpenetan, 82,4 F_1 puntu lortzen ditu identifikazio hertsian eta 90,99 denbora-adierazpenen identifikazio partzialean. EusHeidelTimeren emaitzak 5.3 taulan ikus daitezke.

5.3 taula – EusHeidelTimeren emaitzak EusTimeBanken ebaluaziorako azpicorpusean

Elementua	Ezaugarria	Doitasuna	Estaldura	F_1
TIMEX	hedapena (identifikazio hertsia)	87.27	78.05	82.4
	hedapena (identifikazio partziala)	96.36	86.18	90.99
	type			84.98
	value			64.38

bTimeren konfigurazioak ebaluatzeko F_1 neurria baliatu dugu. F_1 doitasunaren eta estalduraren batez besteko harmonikoa da (5.25 irudia). Gertaeren prozesamendurako, bTimeren konfiguraziorik onenak 76,59 F_1 puntu lortzen ditu gertaeren identifikazioan.

$$F_1 = 2 \cdot \frac{\text{doitasuna} \cdot \text{estaldura}}{\text{doitasuna} + \text{estaldura}}.$$

5.25 irudia – F_1 neurriaren formula

Denbora-erlazioetan, 66,57 F_1 puntu lortzen ditu *Temporal Awareness Score* (UzZaman *et al.*, 2012) ebaluazioan. 5.4 taulan esperimendu honetan bTimek gertaeretakako eta denbora-erlazioetakako lortutako emaitzak ikus daitezke.

Informazio hori baliatuta denbora-lerroak sortu ditugu.

5.4 taula – bTimeren konfigurazio onenaren emaitzak EusTimeBank cor-pusean

Elementua	Ezaugarria	Doitasuna	Estaldura	F_1
EVENT	hedapena	86.22	68.9	76.59
	class			56.71
	tense1			65.51
	tense2			67.0
	aspect1			64.78
	aspect2			63.89
	polarity			71.48
	pos			69.15
	modality			73.78
TLINK	Event-DCT	67.76	65.42	66.57

5.3.2 Denbora-lerroen sorrera

Gertaerak, denbora-adierazpenak eta horien artean sortutako erlazioak lortu ostean, gertaerak denbora-lerroetan antolatu ditugu. Denbora-lerroak nola sortu ditugun azaltzeko, EusTimeBankeko ebaluaziorako azpicorpusetik ondoko testua aukeratu dugu. Gertaerak beltzez nabarmendu ditugu eta denbora-adierazpenak, kurtzibaz. Gertaeren ondoan agertzen diren azpiindizeek token horren posizioa identifikatzen dute.

Apple Inc.-ek bikoiztu egin ditu irabaziak

2007ko urtarrilaren 18

Apple-ren irabaziak ehuneko 78 **hazi**₁₇ ziren *azkeneko hiruhilekoan, zeina 2006ko abenduaren 30ean* **amaitu**₂₆ baitzen. **Igoera**₂₉ horren arrazoia *Gabonetako eta Urte Berriko* **salmentetako**₃₆ Apple-ren iPod musika irakurgailu digitalaren **eskari**₄₂ handia **izan**₄₄ zen. *Gabonetako* **salmenta**₄₈ arrakastatsuen ondorioz, *urte ekonomikoaren hasieran* **bikoiztu**₅₅ egin ziren Apple-ren **salmentak**₅₉, *iaz tarte berean* **irabazitako**₆₄ 565 milioi dolarrekin **konparatuz**₆₈ gero. Konpainiak, 1.000 milioi dolar **irabazita**₇₆, errekorra **hautsi**₇₉ zuen. Konpainiaren *urteko* diru-sarrera orokorrak 7.100 milioi dolarrean **igo**₈₉ ziren *2006an, 2005eko* 5.800 milioi dolarretatik. Konpainiaren diru-sarrerak **hazi**₁₀₀ izana Apple-k **ekoiztutako**₁₀₃ ordenagailu eramangarriak salduenen artean **egon**₁₀₈ izanari **zor**₁₁₀ zaio nolabait.

Wall Street-eko adituek ere ez zuten hain emaitza harrigarririk **iragarri**₁₂₃ Apple ordenagailu-konpainiarentzat. Apple-ren emaitzak merkatuko iragarpenak baino askoz **hobeak**₁₃₃ **izan**₁₃₄ ziren. Irabazi garbiak akzioko 1,14 dolarrekoak **izan**₁₄₂ ziren; Wall Street-ek, berriz, soilik 77 zentabokoak **izango**₁₅₃ zirela **iragarria**₁₅₅ **zuen**₁₅₆.

Urte ekonomikoko lehen hiruhilekoan, Apple Inc.-ek 21 milioi iPod irakurgailu baino gehiago eta 1,6 milioi ordenagailu inguru **saldu**₁₇₉ zituen; **horrek**₁₈₂ **esan**₁₈₃ **nahi**₁₈₄ du **salmentak**₁₈₆ %28 eta %50 **hazi**₁₉₀ direla beste behin *aurreko urtekoekin* **konparatuz**₁₉₆, hurrenez hurren. New Yorkeko Burtsan emaitzak **jakin**₂₀₅ bezain laster, Estatu Batuetako ordenagailu-konpainiaren akzioak ehuneko 5 **hazi**₂₁₅ ziren.

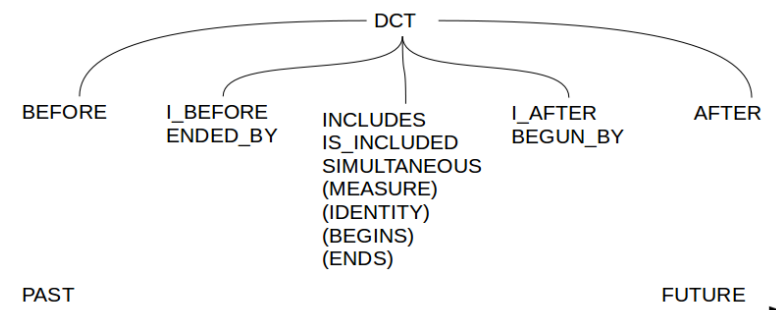
Aipatzekoa₂₁₈ **da**₂₁₉ Apple Inc.-en akzio-kapitala **hirukoiztu**₂₂₆ egin zela *2004an*, **bikoiztu**₂₃₁ *2005ean* eta ehuneko 16 **hazi**₂₃₆ *2006an*. **Hazkunde**₂₃₉ jarraitua iPod eta ordenagailuen **salmenta**₂₄₄ handien ondorio ere **bada**₂₄₈. *Joan den astean*, konpainiak iPhone **aurkeztu**₂₅₆ zuen.

Aipatu dugunez, KroniXak EusHeidelTimek eta bTimek erauzitako denbora-informazioa hartzen du sarrera gisa. Horren bidez, i) denbora-lerroak sortzeko aingurak, ii) gertaerak, eta iii) gertaera batzuen dokumentuaren sorrera-datarekiko (DCT) erlazioak lortzen dira. Denbora-erlazioak lortu ostean, gertaerak gertatzen diren unearen arabera ordenatu ditugu. Horretarako, jarraian aztertuko ditugun hainbat urratsetako prozesua egin dugu.

Lehenik, gertaeraren eta DCTaren arteko erlazio-mota kontuan hartuta, gertaerak multzoka antolatu ditugu. EusTimeMLko denbora-erlazioen baliotetan oinarrituta ondoko bost multzoak egin ditugu:

- Gertaera dokumentua sortu baino lehen gertatzen da (**BEFORE**).
- Gertaera dokumentua sortu baino justu lehenago gertatzen da (**IBEFOR**₁**E**, **ENDE**₂**D_BY**).
- Gertaera dokumentua sortzean gertatzen da (**IS_INCLUDED**, **INCLUDES**, **SIMULTANEOUS**).
- Gertaera justu dokumentua sortu ondoren gertatzen da (**IAFTER**, **BEGUN_BY**).
- Gertaera dokumentua sortu ondoren gertatzen da (**AFTER**).

Erlazio horiek 5.26 irudian irudikatu ditugu. Irudian, EusTimeMLren araberako izendapena ematen zaie erlazioei.



5.26 irudia – DCTaren araberako denbora-erlazioen multzokatzea

Ikusten denez, denbora-erlazio motek gertaera dokumentuaren sorrera-datatik hurbilago edo urrunago jarriko dute; hau da, **BEFORE** etiketa hartzen dutenak **IBEFORE** etiketa hartzen dutenak baino lehenago gertatu dira eta, ondorioz, tarte handiagoa dute dokumentuaren sorrera-datarekiko.

Denbora-lerroak gertaeren ainguratzean oinarrituta eraiki nahi ditugunez, gertaerak kronologiako uneetara ainguratu ditugu. Gertaera bat kronologiako une batean ainguratu dugula esaten dugu bien artean adiberekotasun-erlazioa dagoenean. Hau da, ainguratzeak gertaera une horretan gertatu (edo gertatuko) dela adierazten du. Hala, denbora-lerroak eraikitzeke aldi-berekotasun-erlazioak behar ditugu, baina bTimek lortutako denbora-erlazio gehienek ez dute eraiki nahi ditugun denbora-lerroek eskatzen duten aldi-berekotasunik adierazten. Adibidez, (356) adibideko *hautsi zuen* gertaera DCTarekiko erlazio batean dago, baina ez dago une batean ainguratuta; DCTarekiko posizio erlatibo batean dago, kasu honetan DCTa baino lehenago.

(356) Konpainiak, 1.000 milioi dolar irabazita, errekorra *hautsi zuen*.

Ondorioz, lortutako denbora-erlazioak egokitzeko metodologia garatu dugu. Prozesuak urrats hauek ditu:

- DCTarekiko sortu ditugun erlazioak **EVENT - IS_INCLUDED - TIMEX3** formatuan jarri ditugu. Horretarako,
 - DCTarekin batera gertatzen diren (**SIMULTANEOUS**) edo DCTaren barruan (**IS_INCLUDED**) dauden gertaeren eta DCTaren arteko er-

lazioak ez ditugu eraldatu, gertaera horiek dagoeneko denbora-une batean ainguratuta baitaude.

- DCTarekiko lehenaldian edo geroaldian dauden gertaerentzat transformazio bat proposatu dugu. EusTimeMLk `XXXX-XX-XX` balioa eskaintzen du zehaztugabeko denbora-erreferentzientzat. Hala, erreferentzia lauso horiek denbora-lerroan kokatu ditugu eta DCTarekiko lehenaldian edo geroaldian dauden gertaerak horietan ainguratu ditugu. Hala, moldaketa egin ostean, (356) adibideko erlazioak `hautsi - IS_INCLUDED - XXXX-XX-XX` formatua hartu du.

Denbora-erlazio guztiei formatu egokia eman ondoren, erlazio bakoitzari posizio-identifikatzaile bat esleitu diogu; hain zuzen ere, aingura bera duten denbora-erlazio guztiei, identifikatzaile bera. Lortutako denbora-lerroa 5.27 irudian ikus daiteke.

5.3.3 Esperimentazioa

KroniXaren funtzionamendua ebaluatu dugu. Horretarako, eskuz egindako denbora-lerroak sortu ditugu (5.3.3.1 atala) eta ebaluaziorako metodologia eta metrikak definitu ditugu (5.3.3.2 atala). Hala ere, aipatu dugunez, EusHeidelTimetik eta bTimetik lortutako denbora-informazioa ez da nahi izango genukeen bezain deskribagarria eta, ondorioz, denbora-informazioa dependentzietatik inferitutako informazioaz osatu dugu denbora-lerro adierazgarriagoak sortzeko (5.3.3.3 atala).

5.3.3.1 Ebaluaziorako eskuzko denbora-lerroak

KroniXa ebaluatzeko denbora-lerroen dataseta sortu dugu. Horretarako, EusTimeBankeko ebaluaziorako azpicorpuseko 15 dokumentuen denbora-lerroak sortu ditugu eskuz. Denbora-lerroen sorreran bi etiketatzailek parte hartu zuten eta dokumentuetan etiketatuta zeuden gertaera guztiak denbora-lerroan kokatzeko eskatu genien. Denbora-lerroak eraikitzean ondokoak kontuan hartzeko eskatu genien:

- Soilik testuko informazioa erabili behar zen; ezin zen ezer inferitu ezta mundu-ezagutzarik erabili ere.
- Testuko gertaera guztiak agertu behar ziren denbora-lerroan.

Position ID	Time expression ISO value	Doc ID-token ID-event
1	XXXX-XX-XX	7539-190- <i>hazi</i> (to grow/have grown)
1	XXXX-XX-XX	7539-215- <i>hazi</i> (to grow/grew)
1	XXXX-XX-XX	7539-89- <i>igo</i> (to increase/increased)
1	XXXX-XX-XX	7539-179- <i>saldu</i> (to sell/sold)
1	XXXX-XX-XX	7539-155- <i>iragarria</i> (to announce/announced)
1	XXXX-XX-XX	7539-134- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-79- <i>hautsi</i> (to break/broke)
1	XXXX-XX-XX	7539-256- <i>aurkeztu</i> (to present/presented)
1	XXXX-XX-XX	7539-226- <i>hirukoiztu</i> (to triple/had tripled)
1	XXXX-XX-XX	7539-55- <i>bikoiztu</i> (to double/had doubled)
1	XXXX-XX-XX	7539-44- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-142- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-6- <i>bikoiztu</i> (to double/have doubled)
1	XXXX-XX-XX	7539-17- <i>hazi</i> (to increase/have increased)
2	2007-01-18	7539-108- <i>egon</i> (to be/to be)
2	2007-01-18	7539-111- <i>zaio</i> (to be-AUX)
2	2007-01-18	7539-123- <i>iragarri</i> (to announce/to announce)
2	2007-01-18	7539-184- <i>nahi</i> (to want/wants)
2	2007-01-18	7539-205- <i>jakin</i> (to know/to know)
2	2007-01-18	7539-219- <i>da</i> (to be/is)
2	2007-01-18	7539-248- <i>bada</i> (to be/is)
2	2007-01-18	7539-9- <i>irabaziak</i> (n. profits)
3	XXXX-XX-XX	7539-153- <i>izango</i> (to be/will be)
3	XXXX-XX-XX	7539-26- <i>amaitu</i> (to end/finished)

5.27 irudia – bTimetik lortutako erlazioetatik automatikoki sortutako denbora-lerroa

- Datak eta orduak adierazten zituzten denbora-adierazpenak baino ezin ziren ainguratzat hartu, uneak baino ezin direlako aingura izan.
- Batera gertatzen ziren gertaera guztiak denbora-aingura berari lotuko zitzaizkion eta posizio bera emango zitzaien.
- Gertaera erreferentziakideak banaka tratatuko ziren, baina batera gertatzen zirela adieraziko zen.

5.28 irudian atal honen hasierako testutik sortutako denbora-lerroa irudikatu dugu. Ikus daitekeenez, denbora-lerroek ondoko informazioa dute: i) aingura-gertaera erlazioaren posizioaren identifikatzailea, ii) denbora-ainguraren balio normalizatua eta iii) gertaera (dokumentuaren identifikatzailea, tokena eta gertaeraren forma).

Position ID	Time expression ISO value	Doc ID-token ID-event
1	2004	7539-w226-hirukoiztu
2	2005	7539-w231-bikoiztu
2	2005	7539-w103-ekoiztutako
2	2005	7539-w108-egon
2	2005	7539-w123-iragarri
2	2005	7539-w155-iragarria
2	2005	7539-w156-zuen
3	2006	7539-w64-irabazitako (adj. earned)
3	2006	7539-w89-igo (to increase/increased)
3	2006	7539-w236-hazi (to grow/grew)
4	FY2006-Q1	7539-w179-saldu (to sell/sold)
4	FY2006-Q1	7539-w182-horrek (pron. that)
4	FY2006-Q1	7539-w186-salmentak (n. sales)
4	FY2006-Q1	7539-w190-hazi (to grow/have grown)
5	2006-Q4	7539-w17-hazi (to increase/have increased)
5	2006-Q4	7539-w100-hazi (to grow/have grown)
5	2006-Q4	7539-w115-zor (to owe/owe)
6	2006-12	7539-w48-salmenta (n. sale)
7	2006-12-30	7539-w26-amaitu (to end/finished)
8	2007-01	7539-w36-salmentetako (adj. sold)
8	2007-01	7539-w42-eskari (n. demand)
8	2007-01	7539-w44-izan (to be/were)
9	FY2007	7539-w55-bikoiztu (to double/had doubled)
9	FY2007	7539-w59-salmentak (n. sales)
9	FY2007	7539-w68-konparatuz (to compare/comparing)
9	FY2007	7539-w76-irabazita (adj. earned)
9	FY2007	7539-w79-hautsi (to break/broke)
9	FY2007	7539-w133-hobeak (adj. better)
9	FY2007	7539-w134-izan (to be/were)
9	FY2007	7539-w142-izan (to be/were)
9	FY2007	7539-w153-izango (to be/will be)
10	XXXX-XX-XX	7539-w205-jakin (to know/to know)
10	XXXX-XX-XX	7539-w215-hazi (to grow/grew)
11	2007-W2	7539-w256-aurkeztu (to present/presented)
12	2007-01-18	7539-w183-esan (to say/to say)
12	2007-01-18	7539-w184-nahi
12	2007-01-18	7539-w196-konparatuz
12	2007-01-18	7539-w218-aipatzekoa
12	2007-01-18	7539-w219-da
12	2007-01-18	7539-w239-hazkunde
12	2007-01-18	7539-w244-salmenta
12	2007-01-18	7539-w248-bada

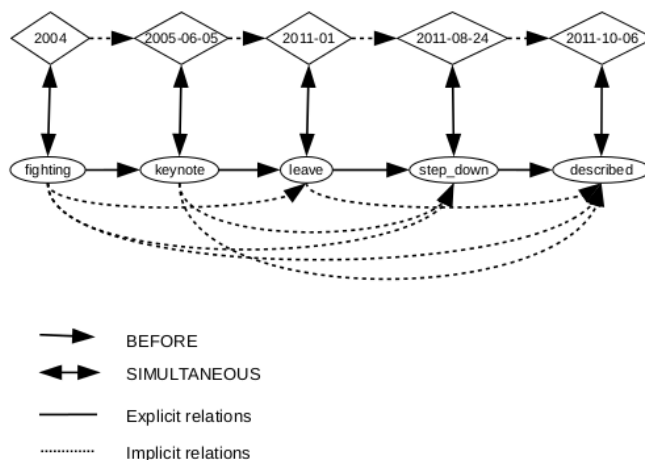
5.28 irudia – Eskuz sortutako denbora-lerroa

Denbora-lerroak sortzeko gidalerroak eskuragarri daude⁹.

5.3.3.2 Ebaluaziorako metodologia eta metrikak

kroniXa ebaluatzeko, SemEval 2015 Task 4 atazan (Minard *et al.*, 2015) inspiratutako metodologia jarraitu dugu. Metodologia horretan TempEval 3ko (UzZaman *et al.*, 2012) erabiltzen dira eta denbora-ezagutza maila (*temporal awareness score*) kontuan hartzen da. Hain zuzen ere, metrika horrek etiketatzearen zuzentasuna neurtzen du denbora-erlazioen identifikazioari eta sailkapenari dagokionez, zeinak denbora-erlazioetan parte hartzen duten denbora-entitateen erazagupen eta sailkapen zuzena hartzen baititu.

Ebaluaziorako, denbora-lerroak denbora-erlazio grafo bihurtzen dira eta lehenetsun eta aldibereketasun erlazioak esplizitu egiten dira (5.29 irudia). Grafo horiek urre-patroitik eskuratutako grafoekin konparatzen dira. Ebaluazio-puntuazioa denbora-lerro bakoitzeko doitasun, estaldura eta F1 neurrien mikro-batezbestekoan oinarritzen dira. Horiek kalkulatzeko urre-patroiko denbora-grafoko erlazioak eta automatikoki lortutako denbora-lerroko grafoak alderatzen dira.



5.29 irudia – Minard *et al.*-etik (2015) hartutako denbora-lerro baten grafoa

KroniXa bi corpusen gainean probatu dugu: alde batetik, EusHeidelTi-

⁹<https://docs.google.com/document/d/1nvY9uU1JrKr-97RD7Vgt3IvHxVSlrUTL0abUIITQCWQ/edit?usp=sharing>

meren eta bTimeren bidez automatikoki lortutako denbora-informazioa baliatuz eraiki ditugu denbora-lerroak (Auto) eta, bestetik, eskuz egindako etiketatzea erabili dugu KroniXaren oinarri gisa (Gold). Horiek ebaluatzeke eskuz egindako denbora-lerroekin alderatu ditugu. 5.5 taulan ikus daitezke KroniXak bi saiakera horietan lortutako emaitzak. Sistema guztiz automatikoari dagokionez, KroniXak 24,65 lortzen du F_1 neurrian eta 34,38, eskuzko etiketatzea baliatzean.

5.5 taula – Kronixaren emaitzak denbora-ezagutza mailaren arabera

	Doitasuna	Estaldura	F_1
Auto	37,79	18,29	24,65
Gold	52,50	25,55	34,38

Emaitzak nahiko onargarriak dira kontuan izanda urrats honetan KroniXak aditz jotatuz adierazitako gertaerak baino ez dituela tratatzen eta hiru denbora-aingura baino ez dituztela sortutako denbora-lerroek: i) DCTa baino lehen, ii) DCTaren unea eta iii) DCTaren ondoren. Denbora-aingurei dagokionez, KroniXak guztiz automatikoki sortzen duen denbora-lerroetan 42 denbora-aingura dago eta eskuz etiketatutako dokumentuetatik sortutakoe-tan 44. Kasu honetan balizko denbora-ainguren kopuru maximoa 45 denez (3 dokumentuko), pentsa daiteke bi sistemek aingura guztiak edo ia guztiak identifikatu dituztela.

Taulara itzulita, pentsa daiteke estaldura hobetuta emaitzak hobetuko direla. Ondorioz, uste dugu, grafoa aberastuz gero, emaitzak hobetuko direla. Horretarako, dependentzietatik inferitutako denbora-erlazioak denbora-lerroetan gehitzeko esperimentua egin dugu.

5.3.3.3 Denbora-lerroak dependentzietatik ateratako informazioaren bidez hobetzea

Aipatu bezala, bTimetik erauzitako informazioarekin informazio gutxiko denbora-lerroak eraiki daitezke. Denbora-lerroak elikatzeko, dependentzia sintaktikoen informazioa baliatu dugu denbora-erlazio kopurua handitzeko: denbora-erlazioen multzoa perpaus barruko denbora-erlazioez handitu dugu. Horretarako, aurreprozesamendu atazan erauzitako dependentzia sintaktikoen informazioa berreskuratu dugu.

Gure hipotesiaren arabera, gertaera-adierazpen baten eta denbora-adierazpen baten artean maila sintaktikoan dependentzia-erlazio bat badago,

denbora-erlazio bat ere egongo da. Hau da, EustimeBank corpora aztertu ostean, uste dugu perpaus barruko denbora-erlazio bat egoteko probabilitate handia dagoela gertaera adierazten duen sintagma baten eta denbora-adierazpena ere baden sintagma baten artean. Adibidez, (357) adibideko esaldian, *aurkeztu zuten* sintagmak gertaera bat adierazten du eta *atzo* sintagma denbora-adierazpena ere bada. Bien artean dependentzia sintaktikoa dago, baita denbora-erlazio bat ere: aurkeztuena testua idatzi baino egun bat lehenago gertatu zela adierazten da.

(357) *Atzo* likidazio plana **aurkeztu zuten**.

Denbora-lerroetan uneak baino ez ditugunez ainguratzat hartzen, TIME edo DATE balioa hartzen duten denbora-adierazpenak baino ez ditugu kontuan hartu dependentzia sintaktikoetatik denbora-erlazioak sortzean. Hala, kronologian ainguratu ez daitezkeen denbora-adierazpenak aukeratzea saihestu dugu.

Era berean, dependentzia sintaktikoetatik erauzitako denbora-erlazioak sortzean, denbora-lerroan agertzen diren gertaeren kopurua handitu dugu. Izan ere, dependentzia-erlazioetan ez dira soilik jokaturako aditzak agertzen; dependentzia-erlazioetan gertaerak adierazten dituzten aditz ez-jokatuek, izenek edota adjektiboek ere hartzen dute parte.

Esaldi bereko gertaeren eta denbora-adierazpenen erlazio-motari dagokionez, uste dugu bien artean gainjartze kronologiko osoa edo partziala dagoela. Hau da, uste dugu gertaera hori eta denbora-adierazpenak adierazten duen une edo tarte kronologiako tarte berean gertatzen direla, tarte hori osorik elkarbanatzen ez badute ere. Ondorioz, dependentzietatik lortutako denbora-erlazioei IS_INCLUDED mota esleitu diegu lehen urrats honetan.

Denbora-erlazio berri horiek bTimek lortutakoei gehitu dizkiegu eta, horrela, erlazio gehiago duen denbora-grafo bat lortu ahal izan dugu. Lortutako erlazio berriak denbora-lerroetan txertatu ditugu. Horretarako, DCTarekiko denbora-erlazioak eta dependentzia sintaktikoetan oinarrituta sortutakoak zerrenda berean jarri ditugu eta identifikatutako denbora-erlazioak denbora-adierazpenaren balio normalizatuaren arabera antolatu ditugu:

- Dependentzietatik erauzitako PAST_REF, PRESENT_REF eta FUTURE_REF aingurak dituzten denbora-erlazioak normalizatu ditugu. PRESENT_REF balioa DCTaren balioarekin ordeztu dugu dokumentua sortu zeneko unea adierazten baitu. PAST_REF eta FUTURE_REF balioak XXXX-XX-XX bihurtu ditugu eta dagokien tokian kokatu ditugu.

- XXXX-XX-XX balioa erreferentzia goiztiarrentzat eta berantiarrentzat hartu dugu eta beste balioak horien artean kokatu ditugu kronologikoki.

Lortutako denbora-lerroen adibidea 5.30 irudian aurkeztu dugu.

Position ID	Time expression ISO value	Doc ID-token ID-event
1	XXXX-XX-XX	7539-190- <i>hazi</i> (to grow up/have grown)
1	XXXX-XX-XX	7539-215- <i>hazi</i> (to grow up/grew)
1	XXXX-XX-XX	7539-89- <i>igo</i> (to increase/increased)
1	XXXX-XX-XX	7539-179- <i>saldu</i> (to sell/sold)
1	XXXX-XX-XX	7539-155- <i>iragarria</i> (to announce/announced)
1	XXXX-XX-XX	7539-134- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-79- <i>hautsi</i> (to break/broke)
1	XXXX-XX-XX	7539-256- <i>aurkeztu</i> (to present/presented)
1	XXXX-XX-XX	7539-226- <i>hirukoiztu</i> (to triple/had tripled)
1	XXXX-XX-XX	7539-55- <i>bikoiztu</i> (to double/had doubled)
1	XXXX-XX-XX	7539-44- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-142- <i>izan</i> (to be/were)
1	XXXX-XX-XX	7539-6- <i>bikoiztu</i> (to double/have doubled)
1	XXXX-XX-XX	7539-17- <i>hazi</i> (to increase/have increased)
2	2005	7539-231- <i>bikoiztu</i> (to double/had doubled)
2	2005	7539-68- <i>konparatuz</i> (to compare/comparing)
3	2006-12-30	7539-26- <i>amaitu</i> (to finish/finished)
4	2006	7539-89- <i>igo</i> (to increase/increased)
5	2006-W51	7539-256- <i>aurkeztu</i> (to present/presented)
6	2007-01-18	7539-108- <i>egon</i> (to be/to be)
6	2007-01-18	7539-111- <i>zaio</i> (to be-AUX)
6	2007-01-18	7539-123- <i>iragarri</i> (to announce/to announce)
6	2007-01-18	7539-184- <i>nahi</i> (to want/wants)
6	2007-01-18	7539-205- <i>jakin</i> (to know/to know)
6	2007-01-18	7539-219- <i>da</i> (to be/is)
6	2007-01-18	7539-248- <i>bada</i> (to be/is)
6	2007-01-18	7539-9- <i>irabaziak</i> (n. profits)
7	XXXX-XX-XX	7539-196- <i>konparatuz</i> (to compare/comparing)
8	XXXX-XX-XX	7539-153- <i>izango</i> (to be/will be)
8	XXXX-XX-XX	7539-26- <i>amaitu</i> (to end/finished)

5.30 irudia – Automatikoki lortutako denbora-lerroa

DCTarekiko erlazioen eta esaldi barruko denbora-erlazioekin lan egiten dugunez, gertatu izan da gertaera bera bi denbora-erlaziotan agertzea: DC-Tari lotuta eta esaldiko denbora-adierazpenari lotuta. Esaterako, *7539-w256-aurkeztu*, *7539-w26-amaitu* eta *7539-w89-igo* gertareak birritan agertzen dira (kursibaz). Horren arrazoiak dira i) aditz jokatuak direla eta ii) denbora-adierazpen batekiko dependentzia erlazio batean daudenez, bigarren esperimendu honetan denbora-lerrora gehitu dela.

Gertaera bera bi aingurari lotuta agertzen denean, gertaera beraren bi iteraziok ekintza indibidualei erreferentzia egiten diotela pentsa dezakegu. Hala ere, hori ez da gure corpusean gertatzen eta gertaera-adierazpen batek gertaera bakarrari egiten dio erreferentzia. Ondorioz, denbora-lerroetan errepikatutako gertaerak akatsak dira eta agerpenetatik adierazgarriena aukeratu behar izango litzateke; gure kasuan dependentzietatik lortutakoa. Adibidez, *7539-w89-igo* gertaerak *XXXX-XX-XX* eta 2006 aingurak hartu ditu eta bigarrenak informazio gehiago ematen du.

Era berean, ikus daitekeenez, dependentzietatik lortutako erlazioak ez daude kronologikoki ordenatuta. Ondo osatutako denbora-lerro batean granularitatearen arabera antolatuko lirateke aingura horiek. Adibidez, 2006 aingura 2006-12-30 aingura baino lehenago agertuko litzateke.

Dependentzietatik lortutako erlazioak gehituta, KroniXa berriz ebaluatu dugu. Horretarako, aurreko esperimentuan bezala, era guztiz automatikoan lortutako denbora-lerroak (Auto) eta eskuz etiketatutako denbora-informaziotik lortutako denbora-lerroak (Gold) ebaluatu ditugu.

5.6 taula – KroniXaren emaitzak dependentzietatiko informazioa gehitzean

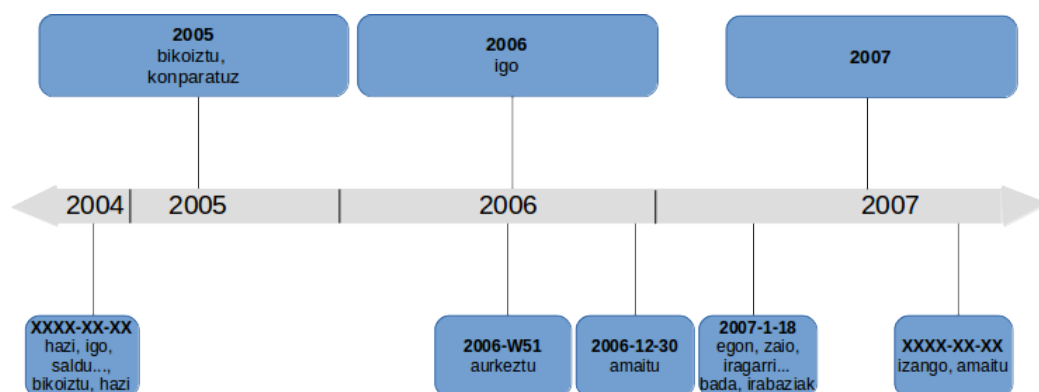
	Doitasuna	Estaldura	F_1
Auto	33,09	19,27	24,36
Gold	45,61	24,02	31,47

Sistema hobetuak lortzen dituen emaitzak 5.6 taulan aurkeztu ditugu. Ikus daitekeenez, dependentzietatik lortutako erlazioak gehitzean emaitzek behera egin dute orokorrean. Doitasunak nabarmen egin du behera, 37,79tik 33,09ra eta 52,5etik 45,61era sistema guztiz automatikoan eta erdiautomatikokoan hurrenez hurren. Estaldurak, ordea, gora egin du sistema guztiz automatikoan, 18,15etik 19,27ra.

Dependentzietatiko denbora-erlazioak denbora-lerroetara gehitzean batezbeste denbora-lerroetako aingura kopurua bikoiztu da guztiz automatikoki lortutako denbora-lerroetan. Erdiautomatikoki lortutako denbora-lerroetan, aingura kopurua 2,2 aldiz biderkatu da. Argi ikusten da 5.30 irudiko denbora-lerroan 5.27 irudian baino askoz ere denbora-aingura gehiago dagoela. Hala ere, sortutako denbora-lerroen azterketa sakonak erakutsi duenez, aingurak zuzen ezin ordenatu ahal izateak (granularitate handitik txikira) drastikoki eragiten die emaitzei.

Amaitzeko, kontuan izan behar da guztiz automatikoki sortutako denbora-lerroetan automatikoki lortutako gertaerak daudela eta horiek ez direla beti zuzenak. Esaterako, 6. posizioan dagoen *7539-9-irabaziak* ez da gertaera-adierazpen bat eta ez luke denbora-lerroan agertu behar izango. Ondorioz, pentsatzekoa da denbora-informazioaren erauzketa hobetu ahala, KroniXaren funtzionamendua ere hobetuko dela.

5.31 irudian 5.30 irudiko denbora-lerroko gertaerak eta denbora-aingurak ardatz kronologikoan kokatu ditugu. Bistan denez, errepresentazioan denbora-adierazpenak une gisa irudikatu ditugu.



5.31 irudia – 5.30 irudiko denbora-lerroaren errepresentazio grafikoa

Denbora-lerroen sorrera automatikoan lanean dihardugu. KroniXa garatzean, horretarako lehen pausoa eman dugu eta etorkizunean espero dugu entitate baten inguruko denbora-lerroak ere sortu ahal izatea.

5.4 Laburpena

Atal honetan, Ixa taldean garatutako euskararen prozesamendu-katea deskribatu dugu, kate horretako analizatzaile morfosintaktikoa erabiltzen baita denbora-informazioaren prozesamenduan baliatzen den informazioa linguistikoa erauzteko. Era berean, EusTimeBank corpora baliatuta sortu ditugun tresnak deskribatu ditugu: alde batetik, euskarazko denbora-adierazpenak identifikatzeko, sailkatzeko eta normalizatzeko erregeletan oinarritutako Eus-HeidelTime tresna, eta beste aldetik, KroniXa denbora-lerroak sortzeko tresna. KroniXak denbora-informazioa dependentzia sintaktikoen erlazioak baliatzen ditu testuko gertaerak denbora-adierazpenekin lotzeko eta denboran

ainguratzeko.

ONDORIOAK ETA ETORKIZUNEKO LANAK

Ondorioak eta etorkizuneko lanak

Tesi-lan honen sarreran esan bezala, denbora-informazioa ulertzeak testua era osotuagoan ulertzea dakar. Hala, HPn testuen ulermenean aurrera egiteko, denbora-informazioa euskararen tratamendu automatikoan eskuragarri egiteko baliabideak garatu ditugu. Informazio hori era egituratuan prozesatu ahal izateak, automatikoki erauzitako denbora-informazioa baliatuta, denbora-lerroak automatikoki sortzeko tresna garatzea ahalbidetu digu.

Jarraian, tesi-lan honetan euskarazko denbora-informazioaren prozesamenduan egindako ekarpenak eta ateratako ondorioak aurkeztuko ditugu eta ildo horretatik sortu zaizkigun erronka berriak azaleratuko ditugu.

6.1 Ekarpenak

Tesi-lan hau testuetako denbora-informazioa euskaraz automatikoki tratatzeko egin den lehen lana da. Denbora-informazioaren tratamendu automatikoan denbora-informazioko elementuak erauzten dira, horien arteko erlazioak garatzen dira eta informazio hori automatikoki erauzten duten edota baliabide aurreratuagoetan erabiltzen duten tresnak sortzen dira. Ekintza horien ondorio dira sortu ditugun eta jarraian aipatuko ditugun baliabideak.

6.1.1 Denbora-informazioa adierazten duten elementuen azterketa

Denbora-informazioaren analisia egiteko, lehenik, informazio horretan zein elementuk parte hartzen duten aztertu dugu.

- Gertaerak eta denbora-adierazpenak aztertu ditugu; hau da, horiek euskaraz nola adierazten diren eta zein informazio adierazten duten aztertu dugu. Orokorrean, denbora-informazioaren azterketan egindako lanetan hartutako ildo berei jarraitu diegu, baina ildo bat baino gehiago egon direnean, nondik jarraitu erabaki behar izan dugu. Adibidez, gertaeratzat zer hartu erabakitzean, zabal jokatzea erabaki dugu. Literaturan, gertaera generikoen tratamenduan bi joera nabarmendu dira: batetik, predikatu generikoak gertaera bezala ez tratatzea kronologian kokatzeko zailak direlako, eta, bestetik, predikatu generikoak gertaeratzat hartzea errealitateko egoera bat adierazten dutelako. Guk bigarrenaren aldeko hautua egin dugu, gure helburua edozein testuingurutan erabil daitekeen denbora-informazioaren analisia egitea delako.
- Gertaeren eta denbora-adierazpenen artean sortzen diren erlazioak ere aztertu ditugu. Hau da, erlazioak zein elementuren artean era daitezkeen aztertu dugu, baita zein motatakoak izan daitezkeen ere. Zehazki, beste hizkuntza batzuetan egin den bezala, denbora-erlazioak, mende-kotasun-erlazioak eta aspektuzko erlazioak aztertu ditugu. Era berean, denbora-erlazioak testuan esplizitu egiten dituzten euskarazko seinaleak identifikatu ditugu.

Erlazioei dagokienez, hiru erlazio mota horiek aztertzea atazaren eta denbora-informazioa etiketatze markaketa-lengoiaren hautapenek baldintzatu dute. Aipatu bezala, edozein motatako testuetako denbora-informazioa aztertu nahi genuen eta, aurreko lanei erreparatuta, esan daiteke erlazio horiek gertaeren denboraren arabera antolakuntzan informazio gehien adierazten dutenak direla. Hala eta guztiz ere, nabarmena da, esaterako, kausazko erlazioek adierazten duten informazioa oso baliagarria izatea denbora-informazioaren azterketan. Ondorioz, erlazio berriak gehitzeko bideak zabalik utzi ditugu.

- Gure ustez, ezinbestekoa da faktualitate-informazioa denbora-informazioarekin batera aztertzea. Batetik, faktualitatea ebazteko eta denbo-

ra-informazioa interpretatzeko antzerako ezaugarri linguistikoei erreparatu behar zaie. Bestetik, gertaerak gertatu diren edo ez jakiteak denbora-informazioaren prozesamenduan laguntzen du, esaterako, denboralerroen sorreran, gertatutako gertaerak baino ez baitira horietan agertzen.

6.1.2 Euskarazko denbora-informazioa kodetzeko markaketa-lengoiaren sorrera

Euskarazko denbora-informazioa kodetzeko EusTimeML markaketa-lengoaia garatu dugu TimeMLn oinarrituta. Jarraian, etiketatze-eskema sortzean egindako ekarpenak zerrendatu ditugu:

- Denbora-informazioa automatikoki tratatu ahal izateko, EusTimeML etiketatze-eskema definitu dugu TimeMLn proposatutako ereduari jarraituz. Hala, errazago konparatu ahal izango dugu TimeML baliatuta egin diren beste hizkuntzetako etiketatze-lanekin. Gainera, uste dugu horrela errazagoa izango dela etiketatutako corpus paraleloak sortzea. HPn testuetako informazioa era egituratuan adierazi behar da ordenagailu bidez informazio hori atzitu ahal izateko. EusTimeML definitzean, euskarazko denbora-informazioa HPrako erabilgarri egin dugu.
- Token mailako etiketatzea definitu dugu prozesamendua errazteko. Ondorioz, token barruko informazioa etiketatze erabakiak hartu behar izan ditugu eta ahalik eta informazio gehien eskaintzen duen etiketatze-eskema definitu behar izan dugu. TimeML eskemak dituzten hizkuntza gehienak ez bezala, euskara inflexio handiko hizkuntza da. TimeML eskemetan tokena da etiketatzearen oinarri eta, EusTimeML sortzean, euskararen kasuan tokenaren barruko informazioa nola azaleratu pentsatu behar izan dugu. Euskararen ezaugarri hori denbora-erlazioak aztertzean izan dugu bereziki kontuan denbora-erlazio asko gertaerei edo denbora-adierazpenei atxikitako postposizioen bidez egiten baitira esplizitu.
- EusTimeMLk TimeMLn proposatutako bideari jarraitzen badio ere, euskarazko denbora-informazioa ahalik eta ondoen islatzeko, etiketak, atributuak eta horien balioak aukeratu edota definitu ditugu. Orokorrean, TimeMLn proposatutako denborazko elementuen (gertaera,

denbora-adierazpen, seinale zein erlazio) atributuak errespetatu ditugu. Izan ere, beste hizkuntzetako etiketatze-saiakeretan egokiak direla ikusi da. Euskararako, ordea, atributu batzuk alde batera uztea erabaki dugu, adibidez gertaeren aditz-forma edota modua adierazten dituztenak, informazio gehigarri adierazgarririk adierazten ez dutelako.

Atributuen balio batzuk ere aldatu ditugu, nagusiki aditzen tempusari eta aspektuari dagozkienak, aztertutako beste hizkuntzetan ez bezala, euskaraz horiek era ezberdinean adierazten baitira bai gramatika klasikoetan, bai hizkuntza automatikoki tratatzeko baliatutako eskemetan. Hain zuzen ere, denbora-informazioaren prozesamendua euskarazko prozesamendu katearen amaieran kokatzen denez, aurreko moduetan informazio linguistikoa nola kodetzen den izan behar da kontuan eta era koherentean adierazi behar da.

- Faktualitatea EusTimeMLren baitan kodetzea erabaki dugunez, faktualitateari dagozkion atributuak gertaeren atributuen artean gehitu ditugu.
- EusTimeML etiketatze-gidalerroak definitu ditugu. Gidalerro horiek sortzeko, i) denbora-informazioa adierazten duten elementuen analisia egin dugu, ii) elementu horiek etiketatzeko atributuak definitu ditugu eta iii) gidalerroen egokitasuna etiketatze-esperimentuen bidez neurtu dugu.

Gidalerroak ondo definituta dauden neurtzeko etiketatzailerik era berean interpretatu dituzten aztertu dugu. Gertaerak identifikatzeko gidalerroei dagokienez, % 80tik gorako bat etortzea lortu dugu eta gertaeren atributu batzuetan ere adostasun maila altua erdietsi dugu, % 87tik gora gramatika-kategoriaren esleipenean, % 98 polaritatean eta % 77 faktualitatean besteak beste. Denbora-adierazpenen identifikazioan % 92tik gorako adostasuna lortu dugu eta mota eta balio normalizatuaren esleipenean kasuen % 70 inguru berdinean etiketatu dira. Erlazioen gaineko etiketatzailerik adostasuna ere neurtu dugu. Erlazioei dagokienez, % 92tik gorako adostasuna lortu dugu erlazioen sailkapenean. Adostasun-maila egokia lortzean (4.5 atala) erabaki dugu gidalerroak ondo definitu ditugula.

6.1.3 EusTimeBank corpora

EusTimeBank corpusean, denbora-informazioak garrantzi berezia duen bi testu mota batu ditugu: i) albisteak, testuetako gertaerak oso lotuta baitaude testua idazten den mementoari, eta ii) gertaera historikoei buruzko narrazioak, denboran kokatutako gertaeren berri ematen baitute. Guztira 164 dokumentuko corpora bildu eta etiketatu dugu eta horietatik 60 dokumentu baliatu ditugu esperimentaziorako urre-patroi azpicorpora definitzeko.

EusTimeBankeko testuak EusTimeMLri jarraiki eskuz etiketatu ditugu eta etiketatzea XML formatuan egin dugu erraz esportatu ahal izateko, besteak beste. Hain zuzen ere, formatu hori erraz bihur daiteke euskararen prozesamendu-katean erabiltzen den NAF etiketatze-formatura. NAF eta XML formatuek hainbat mailatako informazio linguistikoa batera prozesatzeko ahalbidetzen dute. Ondorioz, EusTimeBanken EusTimeMLko informazioa eta beste maila batzuetako informazio linguistikoak erraz konbina daitezke.

EusTimeBank urre-patroi corpusak ditu i) 30 dokumentuko azpicorpora entrenamendurako, ii) 15 dokumentuko azpicorpora garapenerako eta iii) 15 dokumentuko azpicorpora ebaluaziorako. Corpus hori erabilgarri dago denbora-informazioaren prozesamenduan esperimentuak egiteko.

EusTimeBank denbora-informazioa tratatzeko tresnekin erabili dugu. EusTimeBank urre-patroi azpicorpora bTimeren (Salaberri *et al.*, Bidalita) garapenerako eta ebaluaziorako baliatu dugu lehenik. bTimen gertaeren eta erlazioen moduluak SVM motako ikasketa automatikoan oinarritzen dira eta etiketatutako corpusak behar dituzte sailkatzaileak entrenatzeko. Bigarrenik, urre-patroia bTimeren (eta HeidelTimeren) ebaluazioan ere erabili dugu.

Hala, EusTimeBank corpora ez dugu soilik denbora-informazioaren etiketatzeri mugatu. Denbora-informazioaren azterketa eta prozesamenduan interesgarriak diren beste fenomeno batzuk aztertze ere erabili dugu, hala nola ezeztapena. Ezeztapenak garrantzi handia du faktualitatearen ebazpenean eta, horregatik, ezeztapena etiketatze eta etiketatze-eskema definitu dugu eta EusTimeBankeko testu sorta bat horri jarraituta etiketatu dugu. EusTimeBankeko testuen gaineko etiketatzea egin dugunez, erraza izango da etorkizunean ezeztapenaren eta denbora-informazioaren etiketatzeak bateratzeko.

6.1.4 EusHeidelTime tresna

Euskarazko denbora-adierazpenak automatikoki identifikatu, sailkatu eta normalizatzeko, EusHeidelTime tresna garatu dugu. Zehazki, euskararako erregeletan oinarritutako gramatika definitu dugu. Orotara, 313 erregela, 58 patroiz-zerrenda eta 29 normalizazio-fitxategi sortu ditugu. Erregela kopuru handia badirudi ere, erregela asko moldaketa txikiak eginaz definitu ditugu. Hala ere, egia da euskaraz erregela nahiko “errepikatu” behar izan ditugula; izan ere, euskaraz *bat* eta *bi* zenbatzaileak izen multzoaren atzean doazenez, iraupenetan erregela bikoitzak egin behar izan ditugu, bat *bat* zenbakidun iraupen-adierazpenentzat eta beste bat beste zenbakiak dituztenentzat.

Euskararen kasuan, erregela bidezko sistemari jarraitzea zen denbora-adierazpenak identifikatzeko biderik egokiena etiketatutako corpusaren taimainagatik. Halaber, denbora-adierazpenen azterketan ikusi dugunez, denbora-adierazpenek ez dute euskaraz aldakortasun handirik eta antzeko erregelekin denbora-adierazpen asko identifikatzea egingarria izan da. Egia da euskara inflexio handiko hizkuntza dela, baina postposizioak tratatzeko alomorfo zerrendak egitea erabilgarria izan dela ikusi dugu.

Kocońek eta Marcińczuk (2017) esaten dute ez zutela HeidelTime erabili inflexio handiko hizkuntzetan erregela asko sortu behar direlako. Skukan *et al.*-ek (2014) ere kroazierarako 199 erregela erabiltzen dituztela argudiatzen dute. Guk euskararako 313 erregela sortu ditugu, baina erregelen errepikakortasunagatik ez dugu uste lan zaila izan denik.

Amaitzeko, EusHeidelTimek denbora-adierazpenen % 80tik gora ondo identifikatzen dituela azpimarratu nahi dugu. Eraitza hori beste hizkuntzetan lortzen direnen mailan dago eta, ondorioz, EusTimeMLn garatutako hizkuntza-baliabideen kalitatea nabarmena da. Denbora-adierazpenen identifikazio partzialari dagokionez, % 90etik gorako eraitzak lortu ditugu eta horrek EusHeidelTime denbora-adierazpenen etiketatze erdi-automatikoan erabiltzeko prest dagoela erakusten du.

EusHeidelTimeren baliagarritasuna beste aplikazio batzuetan ere frogatu da. EusHeidelTimek NAF formatuko fitxategiak hartzen dituenaz, erraza izan da euskararen prozesamendu-katean era independentean txertatzea. Arrazoi beragatik, EusHeidelTime denbora-informazioa prozesatzeko bTime tresnarekin batera erabili ahal izan dugu.

6.1.5 KroniXa: denbora-lerroak sortzeko tresna

Aipatu bezala, denbora-informazioaren erauzketak hizkuntza-baliabide aurreratuaren sorreran laguntzen du. Tesi-lan honetan, automatikoki erauzitako denbora-informazioa denbora-lerroen sorrera automatikoan baliatu dugu. Hain zuzen ere, euskarazko denbora-informazioa etiketatuta duten testue-tatik denbora-lerroak sortzen dituen KroniXa tresna sortu dugu. Kronixa-ren bidez, bTimetik eta EusHeidelTimetik erauzitako informazioa oinarritzat hartuta, gertaerak denbora-uneetan ainguratu ditugu. Hala eta guztiz ere, bTimek oraindik ez duenez denbora-informazio guztia prozesatzeko gaitasunik, hortik lortutako denbora-informazioa dependentzia sintaktikoetatik lortutako informazioarekin osatu dugu.

Dependentzietatik denbora-informazioa lortzea ez da berria, Chinese Temporal Annotation Corpora (Cheng *et al.*, 2007) etiketatzeke ere erabili baita. Euskaraz ere, dependentziak baliatuta, gertaera gehiago denbora-une gehiagotara ainguratzea lortu dugu.

KroniXak lortzen dituen emaitzak oraindik altuak ez badira ere, abiapuntu ona dira. Alde batetik, dependentzietatik erlazioak gehitzean denbora-lerroetako denbora-ainguren kopurua bikoizten da, are informazio adierazgarriagoa eskainiz. KroniXak F_1 neurrirako 24,36 puntu lortu du sistema guztiz automatikoan eta 31,47 eskuz etiketatutako denbora-informazioa jasotzean.

6.2 Ondorioak

Lan honen sarreran euskarazko denbora-informazioaren prozesamendurako helburuak finkatu ditugu eta tesia egin bitartean, horiek bete ditugu. Ikerketak aurrera egin ahala, gure eginbidetik ondokoak ondorioztatu ditugu.

- Gertaerak definitzean zabal jokatzean, edozein testuingurutan balio duen gertaeraren definizioa lortu dugu.
- Denbora-informazioaren etiketatzeari dagokionez, EusTimeML markaketa-lengoaiak euskarazko denbora-informazioa era osotuan etiketatzea ahalbidetu du. Hain zuzen ere, EusTimeMLn definitutako etiketek eta horien atributuek euskarazko denbora-informazioa zehaztasunez adierazteko balio izan dute.
- EusTimeML TimeMLren beste hizkuntzetako bertsioekin aldera daiteke, egitura bera jarraitzen baitute. Hala, hizkuntzen arteko konpa-

raketak egin daitezke. Batetik, hizkuntza desberdinetako denbora-informazioko fenomenoak aztertzea eta konparatzea posible egiten du. Bestetik, corpus paraleloetan hizkuntza ezberdinetako etiketak alderatzea eta transferitzea ahalbidetzen du.

- Denbora-informazioa erauzteko eta normalizatzeko tresnak garatu ditugu. Zehazki, tesi-lan honetan EusHeidelTime tresna sortu dugu denbora-adierazpenak etiketatzeke. EusHeidelTimek lortutako emaitzek (ikus 5.2.2 atala) erakusten dute erregela bidezko sistema egokia dela euskarazko denbora-adierazpenak erauzteko.
- Denbora-informazioaren analisi eta prozesamendu osotuak tresna aurreratuak sortzeko oinarri sendoak dira. KroniXa denbora-lerroak sortzeko tresna garatzean ikusi dugunez, zenbat eta denbora-informazio gehiago eskura izan, orduan eta denbora-lerro deskribatzaileagoak sortzen dira.

Horiek kontuan hartuta, euskarazko denbora-informazioaren prozesamenduan lehen urratsak egin ditugula ondoriozta daiteke. Hala ere, badira oraindik ere sakonago aztertu beharreko aspektuak, baita denbora-informazioa osatzen duten beste elementu eta ezaugarriak aztertzeke beharra ere.

6.3 Etorkizuneko lanak

Denbora-informazioaren azterketaren eta tratamendu automatikoaren esparruetan ondoko lan-ildoak aurreikusten ditugu etorkizunerako:

- EusTimeML garatzen jarraituko dugu. Lan honetan aurkeztu dugun bertsioa edozein testutan erabiltzeko definitu dugu, baina domeinuaren arabeko egokitzeak ere gehi dakizkioke. Corpora zabaltu ahala, domeinuen arabeko etiketatzea garatu ahal izango dugu.
- EusTimeBank corpora zabaltzea: tesi-lan honetan nabarmendu dugunez, EusTimeBank corpora zabaltzeko bagenu, ikerketarako bide berriak zabaltzeko lirateke. Esaterako, denbora-informazioaren tratamendu automatikoan ikasketa automatikoko teknikak era eraginkorragoan erabiltzeko gai izango ginateke. Hala, gaur egun etiketatzen lagun diezaguketen tresnak ditugula baliatuta, corpora zabaltzeari ekin behar diogu.

- Gertaeren definizioan sakontzea: gertaera-predikatu erlazioa aztertuko dugu eta gure gertaeraren definizioa PropBank (Palmer *et al.*, 2005) eta NomBank (Meyers *et al.*, 2004) corpusetan erabiltzen denarekin alderatuko dugu. Ildo horretatik, gure ikerketa euskarazko predikatuen azterketarekin uztartu nahi dugu.
- Korreferentziaren azterketa: tresnen garapenean aurrera egiteko, gertaeren arteko korreferentzia ere aztergai izango dugu etorkizunean. Testuetako gertaera-adierazpenek errealitateko zein gertaerari egiten dioten erreferentzia identifikatuz gero, gertaera-adierazpen guztiei denbora-aingura bera esleitzeko gai izango gara, adibidez. Hori bereziki baliagarria izango da dokumentuen arteko denbora-lerroen sorreran, hainbat testutatik erauzitako gertaerak antolatzean, bikoizketak saihestuko baititugu.
- Kausalitatearen azterketa: denbora-informazioaren osagarri den informazioa aztertzen jarraitu behar dugu. Adibidez, orain arte kausalitatearen azterketa alde batera utzi badugu ere, ezin dugu ukatu kausalitateak eta denbora-informazioak duten harremana, nagusiki kausa-ondorio sekuentziak gertaeren arteko ordenan duen eraginagatik. Kausalitatea aztertuz gero, albisteetako informazioa enpresentzat erabilgarria izango den informazioa sortzeko balia dezakegu NewsReader proiektuan egin zen bezala. Era berean, euskarazko mediku-txostenak sortu ahala, horiek prozesatzeko aukera izango dugu eta denbora-informazioa eta kausalitatea horietan ere landu ahal izango dugu.
- Denbora-informazioa tresna aurreratuetan baliatuko dugu. Lan hone-tan garatu ditugun baliabideak bereziki interesgarriak izan daitezke ari-keten sorrera automatikoan, denbora-informazioak testua antolatze-ko duen garrantziagatik. Denbora-informazioa galdera-erantzun ari-keten sorrera automatikoan edota testu-sorkuntza automatikoan erabili ahal izango dugu.

Bibliografia

- Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., eta Urkia M. A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada*, 1998. URL <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C8B7C1403C94E6B479E7EE3DA605069D?doi=10.1.1.37.8087&rep=rep1&type=pdf>.
- Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Gojenola K., Sarasola K., eta Urkia M. A Word-grammar Based Morphological Analyzer for Agglutinative Languages. *Proceedings of the 18th Conference on Computational Linguistics - Volume 1, COLING '00*, 1–7, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. URL <https://doi.org/10.3115/990820.990821>.
- Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. In Wilson A., Rayson P., eta Archer D., editoreak, *Corpus Linguistics Around the World*, 1–15. Rodopi, Amsterdam (Netherland), 2006. URL http://npv.edu.ua/!e-book/book/djvu/A/iif_kgpm_Corpus%20Linguistics.pdf.

BIBLIOGRAFIA

- Aduriz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., eta Oronoz M. A Cascaded Syntactic Analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, 124–135. LNCS Series. Springer Verlag, 2004. URL <https://artxiker.ccsd.cnrs.fr/artxibo-00080501>.
- Aduriz I., Ceberio K., eta Díaz de Ilarraza A. Euskarazko anafora pronominala: ikuspuntu konputazionala eta corpus baten garapena. In Logika, Kognizio, Hizkuntza eta Informaziorako Institutua, editorea, *GOGOA*, 5 lib., 91–116, Elhuyar plaza 2, 20018, Donostia, 2005. Euskal Herriko Unibertsitatea/Universidad del País Vasco. URL <http://ixa.si.ehu.eus/sites/default/files/dokumentuak/3564/Euskarazko%20anafora%20pronominala.pdf>.
- Agerri R., Agirre E., Aldabe I., Altuna B., Beloki Z., Laparra E., de Lacalle M.L., Rigau G., Soroa A., eta Urizar R. NewsReader project. *Procesamiento del Lenguaje Natural*, 53:155–158, 2014. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5063>.
- Ahn D., Adafre S.F., eta de Rijke M. Extracting Temporal Information from Open Domain Text: A Comparative Exploration. *Journal of Digital Information Management*, 3(1):14–20, 2005. ISSN 0972-7272. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.7959&rep=rep1&type=pdf>.
- Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. In Bird S., Liberman M., eta Buneman P., editoreak, *Proceedings of the RCS Workshop on linguistic databases*, 1 lib., Philadelphia, 2001. University of Pennsylvania. <https://artxiker.ccsd.cnrs.fr/artxibo-00080504/document>.
- Aldezabal I., Aranzabe M.J., Díaz de Ilarraza A., Estarrona A., Ezeiza N., eta Uria L. Corpusen etiketatze linguistikoa. *A Festschrift for Bernard Oyharçabal. Volumen especial del Anuario del Seminario de Filología Vasca Julio de Urquijo (ASJU)*, 1–2(XLIII):37–50, 2009. URL <http://www.ehu.eus/ojs/index.php/ASJU/article/view/1672>.
- Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsita-

- tea (UPV/EHU), Donostia, 1995. URL <http://ixa.si.ehu.eus/sites/default/files/dokumentuak/4111/tesia-pdf-karpeta-zip.zip>.
- Alegria I., Aranzabe M.J., Ezeiza A., Ezeiza N., eta Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. *Proceedings of Workshop on Customizing knowledge in NLP applications. Third International Conference on Language Resources and Evaluation*, 2002. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.9112&rep=rep1&type=pdf>.
- Alegria I., Balza I., Ezeiza N., Fernández I., eta Urizar R. Named Entity Recognition and Classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI*, Madrid, 2003. ISBN 84-89315-33-7. URL <http://ixa.si.ehu.es/sites/default/files/dokumentuak/3797/JotriII.pdf>.
- Alegria I. eta Urkia M. *Morfologia Konputazionala. Euskararen morfologiaren deskribapena*. Udako Euskal Unibertsitatea, Bilbo, 2002.
- Allan J., Gupta R., eta Khandelwal V. Temporal Summaries of New Topics. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01*, 10–18, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. URL <http://doi.acm.org/10.1145/383952.383954>.
- Allen J.F. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983. URL <http://cse.unl.edu/~choueiry/Documents/Allen-CACM1983.pdf>.
- Alonso O., Gertz M., eta Baeza-Yates R. Clustering and exploring search results using timeline constructions. *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, 97–106, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. URL <http://doi.acm.org/10.1145/1645953.1645968>.
- Altuna B. Análisis de la información temporal en euskera. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., eta Romá Ferri M.T., editoreak, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015a. Sociedad Española para el Procesamiento del Lenguaje Natural. ISBN 978-84-608-1989-9.

BIBLIOGRAFIA

URL https://gplsi.dlsi.ua.es/sepln15/sites/gplsi.dlsi.ua.es/sepln15/files/attachments/begona_altuna.pdf.

Altuna B. Euskarazko denbora-egituren tratamendu automatikorako azterketa. In Alegria I., Latatu A., eta Omaetxebarria M.J., editoreak, *I. Ikergazte, Nazioarteko ikerketa euskaraz*, 46–53, Bilbao, Spain, 2015b. Udako Euskal Unibertsitatea. ISBN 978-84-8438-540-0. URL <http://ixa.si.ehu.es/sites/default/files/dokumentuak/3395/IKERGAZTE.2015.pdf>.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):13–24, 2014a. ISSN 1647-0818. <http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1>.

Altuna B., Aranzabe M., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Barne-txostena, Lengoia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2014, 2014b. http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etikeatzeko%20gidalerroak.

Altuna B., Aranzabe M., eta Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Barne-txostena, Lengoia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016, 2016a. <https://addi.ehu.es/handle/10810/17305>.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTimeren bidez. *Ekaia*, 30:153–165, 2016b. ISSN 0214-9001 e-ISSN: 2444-3255. URL http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8334/Denboraren%20tratamendua_TimeML_HeidelTime.pdf.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, 59(0):15–22, 2017a. ISSN 1989-7553. <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488>.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In Alegria I., Latatu A., Ormaetxebarria M.J., eta Salaberri P., editoreak, *II. IkerGazte*,

- Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Arteak*, 127–134, Bilbo, 2017b. Udako Euskal Unibertsitatea (UEU). ISBN 978-84-8438-628-5. <http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8863/IKERGAZTE.2017.GIZAZIENTZIAKetaARTEA.pdf>.
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Adapting TimeML to Basque: Event Annotation. In Gelbukh A., editorea, *Computational Linguistics and Intelligent Text Processing*, 565–577, Cham, Switzerland, 2018a. Springer International Publishing. ISBN 978-3-319-75487-1. URL https://link.springer.com/chapter/10.1007/978-3-319-75487-1_43.
- Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. An Event Factuality Annotation Proposal for Basque. In Frank A.U., Ivanovic C., Mambriani F., Passarotti M., eta Sporleder C., editoreak, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, 1 lib., 15–24, Vienna, Austria, 2018b. Gerastree Proceedings. ISBN 978-3-901716-43-0. <https://www.oeaw.ac.at/ac/crh2/proceedings/>.
- Altuna B., Minard A.L., eta Speranza M. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, 34–42. Association for Computational Linguistics, 2017c. <http://aclweb.org/anthology/W17-1806>.
- Altuna B., Soraluze A., Aranzabe M.J., Arregi O., eta Díaz de Ilarraza A. KroniXa: Timeline Creation from Basque Texts. *Journal of Information Processing and Management*, Bidalita.
- Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B., eta Rotaetxe K. *Euskal Gramatika Lehen Urratsak (EGLU) II*. Euskaltzaindiko Gramatika Batzordea, Euskaltzaindia, Bilbao, 1987.
- Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B., eta Rotaetxe K. *Euskal Gramatika Lehen urratsak (EGLU) I*. Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo, 1985.

BIBLIOGRAFIA

- Aranzabe M.J. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, 2008. URL http://ixa.si.ehu.eus/sites/default/files/dokumentuak/4118/tesi-txostena_eranskinak.pdf.
- Artiagoitia X. *Hatsarreak eta Parametroak Lantzen*. Arabako Foru Aldundia and Euskal Herriko Unibertsitatea, Gasteiz, 2000.
- Asahara M., Yasuda S., Konishi H., Imada M., eta Maekawa K. BCCWJ-Timebank: Temporal and Event Information Annotation on Japanese Text. *27th Pacific Asia Conference on Language, Information, and Computation*, 206–214, Wenshan, Taipei, Taiwan, 2013. URL <http://www.aclweb.org/anthology/Y13-1019>.
- Azkarate Villar M. Gertaera- eta emaitza-izenak. In Artiagoitia X. eta Laka-rra J.A., editoreak, *Gramatika jaietan. Patxi Goenagaren omenez*, 93–106. Euskal Herriko Unibertsitatea/Universidad del País Vasco (UPV/EHU), Bilbo, 2008a. ISBN 978-84-9860-085-8.
- Azkarate Villar M. Hitz-elkarte sintetikoen argumentuak eta gertaerak. *Jean Haritschelhar-i omenaldia*, 117–134. Euskaltzaindia, Bilbo, 2008b. ISBN 978-84-95438-33-1. URL www.euskaltzaindia.eus/dok/ikerbilduma/69585.pdf.
- Bach E. The algebra of events. *Linguistics and philosophy*, 9(1):5–16, 1986.
- Bauer S., Clark S., eta Graepel T. *Learning to Identify Historical Figures for Timeline Creation from Wikipedia Articles*, 234–243. Springer International Publishing, Cham, Switzerland, 2015. URL https://doi.org/10.1007/978-3-319-15168-7_30.
- Bayer J. The Role of Event Expression in Grammar. *Studies in Language*, 1(10):1–52, 1986. URL https://kops.uni-konstanz.de/bitstream/handle/123456789/3699/The_Role_of_event_expression_in_grammar.pdf?sequence=1.
- Bengoetxea K. eta Gojenola K. Desarrollo de un Analizador Sintáctico Estadístico basado en Dependencias para el Euskera. *Procesamiento del Lenguaje Natural*, 39:5–12, 2007. ISSN 1989-7553. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2651>.

- Bethard S. ClearTK-TimeML: A minimalist approach to TempEval 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 10–14. Association for Computational Linguistics, 2013. URL <http://aclanthology.coli.uni-saarland.de/pdf/S/S13/S13-2002.pdf>.
- Bethard S., Derczynski L., Savova G., Pustejovsky J., et al. Verhagen M. SemEval-2015 Task 6: Clinical TempEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 806–814, Denver, Colorado, June 2015. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S15-2136>.
- Bethard S. et al. Martin J.H. Identification of Event Mentions and Their Semantic Class. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06*, 146–154, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. <http://dl.acm.org/citation.cfm?id=1610075.1610098>.
- Bijankhan M., Sheykhzadegan J., Bahrani M., et al. Ghayoomi M. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2):143–164, 2011. <http://dx.doi.org/10.1007/s10579-010-9132-x>.
- Bindi R., Monachini M., et al. Orsolini P. Il Corpus di Riferimento della lingua italiana contemporanea, Documentazione. Barne-txostena, ILC-TLN-1, 1989.
- Bird S., Day D., Garofolo J., Henderson J., Laprun C., et al. Liberman M. ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May 2000. European Language Resources Association (ELRA). URL <http://www.lrec-conf.org/proceedings/lrec2000/pdf/184.pdf>. ACL Anthology Identifier: L00-1138.
- Bittar A. Annotation of Events and Temporal Expressions in French Texts. *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*

BIBLIOGRAFIA

- '09, 48–51, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-52-7. <http://dl.acm.org/citation.cfm?id=1698381.1698388>.
- Bittar A. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Doktoretza-tesia, Université Paris Diderot, Paris, 2010. <http://www.linguist.jussieu.fr/~abittar/docs/Bittar-PhD.pdf>.
- Bittar A., Hagège C., Moriceau V., Tannier X., eta Teissèdre C. Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3741–3745, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/308_Paper.pdf. ACL Anthology Identifier: L12-1142.
- Boser B.E., Guyon I.M., eta Vapnik V.N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. <http://doi.acm.org/10.1145/130385.130401>.
- Broda B., Marcińczuk M., Maziarz M., Radziszewski A., eta Wardyński A. KPWr: Towards a Free Corpus of Polish. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3218–3222, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/965_Paper.pdf. ACL Anthology Identifier: L12-1574.
- Canh T.V., Markert K., eta Nejdl W. RussianFlu-DE: A German Corpus for a Historical Epidemic with Temporal Annotation. *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessalo-*

- niki, Greece, September 18-21, 2017, Proceedings*, 61–73, 2017. URL https://doi.org/10.1007/978-3-319-67008-9_6.
- Carlson G.N. *Reference to Kinds in English*. Doktoretza-tesia, University of Massachusetts Amherst, Amherst, Massachusetts, 1977. <https://scholarworks.umass.edu/dissertations/AAI7726414>.
- Caselli T. It-TimeML: TimeML Annotation Scheme for Italian. Barne-txostena, Istituto di Linguistica Computazionale, 2010. Version 1.3. 1. Technical Report.
- Caselli T., Lenzi V.B., Sprugnoli R., Pianta E., eta Prodanof I. Annotating Events, Temporal Expressions and Relations in Italian: The It-TimeML Experience for the Ita-TimeBank. *Proceedings of the 5th Linguistic An-notation Workshop, LAW V '11*, 143–151, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-93-0. URL <http://dl.acm.org/citation.cfm?id=2018966.2018984>.
- Caselli T. eta Vossen P. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. *Proceedings of the Events and Stories in the News Workshop*, 77–86, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W17-2711>.
- Cassidy T., McDowell B., Chambers N., eta Bethard S. An Annotation Fra-mework for Dense Event Ordering. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Pa-pers)*, 501–506, Baltimore, Maryland, June 2014. Association for Computa-tional Linguistics. URL <http://www.aclweb.org/anthology/P14-2082>.
- Chang A.X. eta Manning C. SUTime: A library for recognizing and norma-lizing time expressions. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Pro-ceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3735–3740, Istanbul, Turkey, May 2012. Euro-pean Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/284_Paper.pdf. ACL Anthology Identifier: L12-1122.

BIBLIOGRAFIA

- Cheng Y., Asahara M., eta Matsumoto Y. Constructing a Temporal Relation Tagged Corpus of Chinese Based on Dependency Structure Analysis. *Proceedings of the 14th International Symposium on Temporal Representation and Reasoning*, TIME '07, 59–69, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2836-8. <http://dx.doi.org/10.1109/TIME.2007.47>.
- Chieu H.L. eta Lee Y.K. Query Based Event Extraction Along a Timeline. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, 425–432, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. URL <http://doi.acm.org/10.1145/1008992.1009065>.
- Chinchor N. eta Sundheim B. Message Understanding Conference (MUC) 6 LDC2003T13. Barne-txostena, Linguistic Data Consortium, Philadelphia, USA, 2003. <https://catalog.ldc.upenn.edu/LDC2003T13>.
- Cieri C. eta Liberman M. TIDES Language Resources: A Resource Map for Translingual Information Access. *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 1334–1339. European Language Resources Association, 2002. URL <http://www.lrec-conf.org/proceedings/lrec2002/pdf/245.pdf>.
- Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. URL <https://doi.org/10.1177/001316446002000104>.
- Corpas Pastor G. *Manual de Fraseología Española*. Editorial Gredos, 1996.
- Cortes C. eta Vapnik V. Support-vector networks. *Machine Learning*, 20(3): 273–297, Sep 1995. URL <https://doi.org/10.1007/BF00994018>.
- Costa F. eta Branco A. TimeBankPT: A TimeML Annotated Corpus of Portuguese. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3727–3734, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL <http://>

- www.lrec-conf.org/proceedings/lrec2012/pdf/246_Paper.pdf. ACL Anthology Identifier: L12-1096.
- Coterón Agorria I. *Euskal aditz jokatua: gertaeren gaineko ikuspegi sistema (edo TAM sistema)*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Gasteiz, 2015. <https://addi.ehu.es/handle/10810/16026>.
- Croft W. Force Dynamics and Directed Change in Event Lexicalization and Argument Realization. In de Almeida G.R. eta Manouilidou C., editoreak, *Cognitive Science Perspectives on Verb Representation and Processing*, 103–129. Springer International Publishing, Cham, Switzerland, 2015. ISBN 978-3-319-10112-5. http://dx.doi.org/10.1007/978-3-319-10112-5_5.
- Cybulska A. eta Vossen P. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4545–4552, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/840_Paper.pdf. ACL Anthology Identifier: L14-1646.
- Day D., Aberdeen J., Hirschman L., Kozierok R., Robinson P., eta Vilain M. Mixed-initiative Development of Language Processing Systems. *Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLC '97*, 348–355, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. URL <https://doi.org/10.3115/974557.974608>.
- Day D.S., McHenry C., Kozierok R., eta Riek L. Callisto: A Configurable Annotation Workbench. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2073–2076, Lisboa, Portugal, 2004. URL <http://www.lrec-conf.org/proceedings/lrec2004/pdf/612.pdf>.
- de Azkue R.M. *Morfología Vasca*. Editorial Vasca, Bilbao, 1923. URL <http://www.liburuklik.euskadi.eus/handle/10771/25059>.

BIBLIOGRAFIA

- Derczynski L. eta Bontcheva K. PHEME: Veracity in Digital Social Networks. *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA)*, Lisboa, Portugal, 2014. URL http://ceur-ws.org/Vol-1181/pros2014_paper_05.pdf.
- Derczynski L., Llorens H., eta UzZaman N. TimeML-Strict: Clarifying Temporal Annotation. *Computing Research Repository (CoRR)*, abs/1304.7289, 2013. <http://arxiv.org/abs/1304.7289>.
- Dice L.R. Measures of the Amount of Ecologic Association between Species. *Ecology*, 26:297–302, 1945.
- Dowden B. Time Supplement. In Fieser J. eta Dowden B., editoreak, *Internet Encyclopedia of Philosophy*. 2009. URL <http://www.iep.utm.edu/time-sup/>.
- Eguskitza A. Aditz modalak euskaraz. *Euskera*, 2(30):215–221, 1985. URL www.euskaltzaindia.net/dok/euskera/11319.pdf.
- Etcheverry M. eta Wonsever D. Time Expressions Recognition with Word Vectors and Neural Networks. In Schewe S., Schneider T., eta Wijzen J., editoreak, *24th International Symposium on Temporal Representation and Reasoning (TIME 2017)*, 90 lib. of *Leibniz International Proceedings in Informatics (LIPIcs)*, 12:1–12:20, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-052-1. URL <http://drops.dagstuhl.de/opus/volltexte/2017/7925>.
- Ezeiza N. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzailer sintaktiko sendo eta malgua*. Doktoresetia, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, 2002. URL <http://ixa.si.ehu.es/sites/default/files/dokumentuak/4123/OSOA.pdf>.
- Ezeiza N., Alegria I., Arriola J.M., Urizar R., eta Aduriz I. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 379–379, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P98-1062>.

- Ferro L., Gerber L., Mani I., Sundheim B., eta Wilson G. TIDES 2003 Standard for the Annotation of Temporal Expressions. Barne-txostena, MITRE, McLean, USA, September 2003. http://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf.
- Ferro L., Gerber L., Mani I., Sundheim B., eta Wilson G. ACE Time Normalization (TERN) 2004 English Evaluation Data v1.0. <http://www ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T18>, 2010. Accessed: 2013-07-03.
- Ferro L., Kozierok R., Gerber L., Sundheim B., Mani I., eta Wilson G. Annotating temporal information: From theory to practice. *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, 226–230, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. URL <http://dl.acm.org/citation.cfm?id=1289189.1289202>.
- Ferrucci D. eta Lally A. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4):327–348, 2004. URL <https://dl.acm.org/citation.cfm?id=1030325>.
- Fiscus J.G., Doddington G.R., Garofolo J.S., eta Martin A.F. NIST's 1998 topic detection and tracking evaluation (TDT2). *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999*, Budapest, Hungary, 1999. ISCA. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1611&rep=rep1&type=pdf>.
- Fokkens A., Soroa A., Beloki Z., Ockeloen N., Rigau G., van Hage W.R., eta Vossen P. NAF and GAF: Linking linguistic annotations. *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland, 2014. URL <https://research.vu.nl/en/publications/naf-and-gaf-linking-linguistic-annotations>.
- Forăscu C. eta Tufiş D. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Eighth International Conference on Language Resources*

BIBLIOGRAFIA

- and Evaluation (LREC-2012)*, 3762–3766, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/770_Paper.pdf. ACL Anthology Identifier: L12-1451.
- Galton A. A Critical Examination of Allen’s Theory of Action and Time. *Artificial Intelligence*, 42(2-3):159–188, 1990. ISSN 0004-3702. URL [http://dx.doi.org/10.1016/0004-3702\(90\)90053-3](http://dx.doi.org/10.1016/0004-3702(90)90053-3).
- Garcia J. Nominalizazioa itzulpen teknikoan. *Ekaia*, 14:99–115, 2001. URL http://www.ehu.eus/eustek/Index_archivos/Liburutegia/Nominalizazioa,%20itzulpen%20tek.pdf.
- Girardi C., Speranza M., Sprugnoli R., eta Tonelli S. CROMER: a Tool for Cross-Document Event and Entity Coreference. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, 3204–3208, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/726_Paper.pdf. ACL Anthology Identifier: L14-1568.
- Goenaga I. *ASKHi: Analisi sintaktiko konputazional hibridoa paradigma desberdinen konbinazioan oinarrituta*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, 2017. URL <http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8909/Tesia.pdf>.
- Goenaga P. Izen sintagmaren egituraz. *ASJU, Anuario del Seminario de Filología Vasca “Julio de Urquijo”*, Gehigarria 14:847–865, 1991. ISSN 0582-6152. URL <http://www.ehu.eus/ojs/index.php/ASJU/article/viewFile/9235/8449>.
- Gojenola K. *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta errorearen tratamenduan*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, 2000. http://www.inguma.eus/tesiak/Gojenola_Galletetebeitia_2000.pdf.

- Goranko V. et al Galton A. Temporal logic. In Zalta E.N., editore, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition, 2015. URL <https://plato.stanford.edu/archives/win2015/entries/logic-temporal/>.
- Grimshaw J. *Argument Structure*. MIT Press, Cambridge, Massachusetts, 1990. ISBN 9780262071253.
- Grishman R. et al Sundheim B. Message Understanding Conference-6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 466–471, Center for Sprogteknologi, Copenhagen, Denmark, 1996. URL <https://dl.acm.org/citation.cfm?id=992709>.
- Gross M. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63:7–52, 1981. URL https://www.persee.fr/doc/lgge_0458-726x_1981_num_15_63_1875.
- Grover C., Tobin R., Alex B., et al Byrne K. Edinburgh-LTG: TempEval-2 System Description. *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, 333–336, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1859664.1859738>.
- Guerrero Nieto M., Saurí R., et al Poveda M.Á.B. ModeS TimeBank: A Modern Spanish TimeBank Corpus. *Procesamiento del lenguaje natural*, 47: 259–267, 2011. URL journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/983/736.
- Gupta N., Joshi A., et al Bhattacharyya P. A temporal expression recognition system for medical documents by taking help of news domain corpora. In Sharma D.M., Sangal R., et al Sherly E., editore, *Proceedings of the ICON 2015: 12th International Conference on Natural Language Processing*, Hyderabad, India, 2015. URL <https://www.cse.iitb.ac.in/~pb/papers/icon15-temporal-expression.pdf>.
- Hacioglu K., Chen Y., et al Douglas B. Automatic Time Expression Labeling for English and Chinese Text. In Gelbukh A., editore, *Computational Linguistics and Intelligent Text Processing*, 548–559, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN

BIBLIOGRAFIA

- 978-3-540-30586-6. URL <https://pdfs.semanticscholar.org/f95d/73cbffc2599bd45d8ec7e948a868db7104e1.pdf>.
- Halliday M. *An Introduction to Functional Grammar*. Hodder Arnold, London, 2004. ISBN 9780340761670. Third edition.
- Hualde J.I. eta de Urbina J.O. *A grammar of Basque*, 26 lib. Walter de Gruyter, Boston, USA, 2003.
- Huddleston R.D. eta Pullum G.K. *The Cambridge Grammar of the English Language*. Cambridge University Press, April 2002.
- Iñurrieta U., Aduriz I., Díaz de Ilarraza A., Labaka G., eta Sarasola K. Izen+aditz konbinazioen itzulpenaz eta tratamendu konputazionalaz. *Senez*, 47:237–249, 2016. ISSN 1132-2152. URL http://ixa.si.ehu.eus/sites/default/files/dokumentuak/3648/senez47_inurrieta-et-al.pdf.
- Im S., You H., Jang H., Nam S., eta Shin H. KTimeML: Specification of Temporal and Event Expressions in Korean Text. *Proceedings of the 7th workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009*, 115–122, Suntec City, Singapore, 2009. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W09-3417>.
- ISO Commitee. Information interchange: representation of dates and times. International Standard ISO 8601:2004, International Organization for Standardization, Switzerland, 1997. <https://www.iso.org/standard/40874.html>.
- ISO-TimeML working group. Language resource management – Semantic Annotation Framework (SemAF) – Part 1: Time and events. International Standard ISO/CD 24617-1(E), International Organization for Standardization, Switzerland, 2008. http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf.
- Jang S.B., Baldwin J., eta Mani I. Automatic TIMEX2 Tagging of Korean News. *ACM Transactions on Asian Language Information Processing (TALIP) - Special Issue on Temporal Information Processing*, 3(1):51–65, mar 2004. ISSN 1530-0226. <http://doi.acm.org/10.1145/1017068.1017072>.

- Jeong Y.S. et al Choi H.J. Language independent feature extractor. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 4170–4171, Texas, USA, 2015. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9527/9717>.
- Jeong Y.S., Kim Z.M., Do H.W., Lim C.G., et al Choi H.J. Temporal Information Extraction from Korean Texts. 279–288, July 2015. URL <http://www.aclweb.org/anthology/K15-1028>.
- Jeong Y.S., Lim C.G., et al Choi H.J. ExoTime: Temporal Information Extraction from Korean Texts Using Knowledge Base. *Journal of the Korea Society of Computer and Information*, 22:35–48, 2017. URL <http://www.dbpia.co.kr/Article/NODE07286957>.
- Kawai H., Jatowt A., Tanaka K., Kunieda K., et al Yamada K. Chronoseeker: Search engine for future and past events. *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, 25:1–25:10, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-893-3. <http://doi.acm.org/10.1145/2108616.2108647>.
- Kocoń J. et al Marcińczuk M. Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes. *Natural Language Engineering*, 23(3):385–418, 2017. URL https://www.cambridge.org/core/services/aop-cambridge-core/content/view/D6911072A23A622BFFFF9D0BBEEA405AD/S1351324916000255a.pdf/supervised_approach_to_recognise_polish_temporal_expressions_and_rulebased_interpretation_of_timexes.pdf.
- Kocoń J., Marcińczuk M., Oleksy M., Bernaś T., et al Wolski M. Temporal Expressions in Polish Corpus KPWr. *Cognitive Studies/Études Cognitives*, 15:293–317, 2015. URL <https://doi.org/10.11649/cs.2015.020>.
- Lafferty J., McCallum A., et al Pereira F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. <http://dl.acm.org/citation.cfm?id=645530.655813>.
- Lafitte P. *Grammaire Basque, navarro-labourdin littéraire*. Le Livre, Baiona, 1944. URL <http://hdl.handle.net/10690/468>.

BIBLIOGRAFIA

- Lenzi V.B., Moretti G., eta Sprugnoli R. CAT: the CELCT Annotation Tool. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 333–338, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/216_Paper.pdf. ACL Anthology Identifier: L12-1072.
- Li L., Zhang J., He Y., Zhang Y., eta Huilin W. Chinese Temporal Relation Resolution Based on Chinese-English Parallel Corpus. *New Technology of Library and Information Service*, (31):42–48, 2014. URL <https://ieeexplore.ieee.org/document/7063995/>.
- Lin C.Y. Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens S.S., editorea, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- Llorens H., Saquete E., eta Navarro B. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, 284–291, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1859664.1859727>.
- Mani I. eta Wilson G. Robust Temporal Processing of News. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 69–76, Stroudsburg, PA, USA, 2000. URL <http://aclweb.org/anthology/P/P00/P00-1010.pdf>.
- Mansouri B., Zahedi M., Campos R., Farhoodi M., eta Rahgozar M. Pars-Time: Rule-Based Extraction and Normalization of Persian Temporal Expressions. *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018*, 715–721, Grenoble, France, 01 2018. Springer International Publishing. ISBN 978-3-319-76940-0. URL https://link.springer.com/chapter/10.1007/978-3-319-76941-7_67.
- March O. eta Baldwin T. Automatic Event Reference Identification. *Australasian Language Technology Association Workshop 2008*, 6 lib., 79–87, 2008. URL <http://www.aclweb.org/anthology/U08-1011>.

- Marcińczuk M., Kocoń J., eta Janicki M. *Liner2 – A Customizable Framework for Proper Names Recognition for Polish*, 231–253. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-35647-6. URL https://doi.org/10.1007/978-3-642-35647-6_17.
- Marcińczuk M., Oleksy M., Bernaś T., Kocoń J., eta Wolski M. Towards an Event Annotated corpus of Polish. *Cognitive Studies/Études Cognitives*, 15:253–267, 2015. URL <https://doi.org/10.11649/cs.2015.018>.
- Marcińczuk M., Oleksy M., eta Kocoń J. Inforex — a Collaborative System for Text Corpora Annotation and Analysis. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 473–482, Varna, Bulgaria, September 2017. INCOMA Ltd. URL https://doi.org/10.26615/978-954-452-049-6_063.
- Martí M.A., Martín-Valdivia M.T., Taulé M., Jiménez-Zafra S.M., Nofre M., eta Marsó L. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57:41–48, 2016. ISSN 1989-7553. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5335/3148>.
- Mazur P. eta Dale R. The DANTE Temporal Expression Tagger. In Vetulani Z., editorea, *Proceedings of the 3rd Language And Technology Conference (LTC)*, Poznań, Poland, 2007. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.193.4466&rep=rep1&type=pdf>.
- Mazur P. eta Dale R. WikiWars: A New Corpus for Research on Temporal Expressions. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, 913–922, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1870658.1870747>.
- McEnery A., Xiao R., eta Tono Y. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge Applied Linguistics, Abingdon-on-Thames, UK, 2006. ISBN 978-0415286237.
- Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B., eta Grishman R. The NomBank Project: An Interim Report. In Meyers A., editorea, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for

BIBLIOGRAFIA

- Computational Linguistics. URL <http://www.aclweb.org/anthology/W04-2705>.
- Minard A.L., Marchetti A., eta Speranza M. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. *Proceedings of CLiC-it 2014, First Italian Conference on Computational Linguistic*, 2014. URL <http://www.fileli.unipi.it/projects/clic/proceedings/vol1/CLICIT2014150.pdf>.
- Minard A.L., Speranza M., Agirre E., Aldabe I., van Erp M., Magnini B., Rigau G., eta Urizar R. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 778–786, Denver, Colorado, June 2015. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S15-2132>.
- Minard A.L., Speranza M., Urizar R., Altuna B., van Erp M., Schoen A., eta van Son C. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., eta Piperidis. S., editoreak, *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association, 2016. ISBN 978-2-9517408-9-1. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/488_Paper.pdf.
- Mirza P. eta Minard A.L. HLT-FBK: a Complete Temporal Processing System for QA TempEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 801–805, Denver, Colorado, June 2015. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S15-2135>.
- Mirza P. eta Tonelli S. An Analysis of Causality between Events and its Relation to Temporal Information. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2097–2106. Dublin City University and Association for Computational Linguistics, 2014. URL <http://www.aclweb.org/anthology/C14-1198>.
- Moriceau V. eta Tannier X. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno

- A., Odijk J., eta Piperidis S., editoreak, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/45_Paper.pdf.
- Mostafazadeh N., Chambers N., He X., Parikh D., Batra D., Vanderwende L., Kohli P., eta Allen J. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849, San Diego, California, June 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/N16-1098>.
- Murat A., Li X., Jiang T., Yang Y., Zhou X., eta Wang L. Research on Temporal Expression Recognition: A Case Study of Uyghur. *ICIC Express Letters*, 8(4):703–710, 2017. ISSN 2185-2766.
- Orasmaa S. Towards an Integration of Syntactic and Temporal Annotations in Estonian. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1259–1266, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/530_Paper.pdf.
- Orasmaa S. *Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience*. Doktoretza-tesia, University of Tartu, Estonia, 2016. <http://hdl.handle.net/10062/54752>.
- Oronoz M. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Informatika Fakultatea, Euskal Herriko Unibertsitatea, Donostia, 2008. URL http://www.euskara.euskadi.net/appcont/tesisDoctoral/PDFak/Maite_Oronoz_TESI.pdf.
- Palmer M., Gildea D., eta Kingsbury P. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–

BIBLIOGRAFIA

- 106, March 2005. ISSN 0891-2017. URL <http://dx.doi.org/10.1162/0891201053630264>.
- Pennington J., Socher R., eta Manning C. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/D14-1162>.
- Pociello E. *Euskararen ezagutza-base lexikala: Euskal WordNet*. Doktoretzatesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Leioa, 2008. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1204622545/publikoak/2008Tesi-txostena-eranskinak-aurkezpena.rar>.
- Pustejovsky J., Belanger L., Castaño J., Gaizauskas R., Hanks P., Ingria B., Katz G., Radev D., Rumshisky A., Sanfilippo A., Saurí R., Setzer A., Sundheim B., eta Verhagen M. NRRC Summer Workshop on Temporal and Event Recognition for Question Answering Systems. Barne-txostena, National Regional Research Center (NRRC), 2002. URL <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=487F91DE9A089800706B32FC61168034?doi=10.1.1.128.7428&rep=rep1&type=pdf>.
- Pustejovsky J., Castaño J.M., Ingria R., Saurí R., Gaizauskas R.J., Setzer A., Katz G., eta Radev D.R. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34, 2003a. URL <http://www.timeml.org/publications/timeMLpubs/IWCS-v4.pdf>.
- Pustejovsky J., Hanks P., Saurí R., See A., Gaizauskas R., Setzer A., Radev D., Sundheim B., Day D., Ferro L., eta Lazo. M. The TimeBank Corpus. In Archer D., Rayson P., Wilson A., eta McEnery T., editoreak, *Proceedings of Corpus Linguistics 2003*, number 16, 647–656, Lancaster, UK, 2003b. UCREL, Lancaster University. ISBN 1 86220 131 5. URL https://www.researchgate.net/publication/228559081_The_TimeBank_corpus.
- Pustejovsky J., Moszkowics J., eta Verhagen M. ISO-Space: The Annotation of Spatial Information in Language. In Blunt H., editorea, *Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable*

- Semantic Annotation (ISA-6)*, 1–9, Oxford, UK, 2011. Computing Laboratory, University of Oxford. URL https://sigsem.uvt.nl/isa6/isa-6_proceedings.pdf.
- Pustejovsky J., Verhagen M., Saurí R., Littman J., Gaizauskas R., Katz G., Mani I., Knippen R., eta Setzer A. TimeBank 1.2. Barne-txostena, Linguistic Data Consortium, 2006. <https://catalog.ldc.upenn.edu/LDC2006T08>.
- Radinsky K. eta Horvitz E. Mining the web to predict future events. *Proceedings of the sixth ACM international conference on Web search and data mining*, 255–264. ACM, 2013. URL http://erichorvitz.com/future_news_wsdm.pdf.
- Reichenbach H. The tenses of verbs. *Elements of Symbolic Logic*, 287–298. The Macmillan Company, New York, 1947.
- Reynaert M., Oostdijk N., Clercq O.D., van den Heuvel H., eta de Jong F. Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., eta Tapias D., editoreak, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2693–2398, Valletta, Malta, 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/549_Paper.pdf.
- Robaldo L., Caselli T., Russo I., eta Grella M. From Italian Text to TimeML Document via Dependency Parsing. *Computational Linguistics and Intelligent Text Processing*, 177–187. Springer, 2011. URL https://link.springer.com/chapter/10.1007/978-3-642-19437-5_14.
- Rose T., Stevenson M., eta Whitehead M. The Reuters Corpus Volume 1—from Yesterday’s News to Tomorrow’s Language Resources. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA), 2002. URL <http://www.aclweb.org/anthology/L02-1080>.
- Rubin V.L., Liddy E.D., eta Kando N. *Certainty Identification in Texts: Categorization Model and Manual Tagging Results*, 61–76. Springer Nether-

BIBLIOGRAFIA

- lands, Dordrecht, 2006. ISBN 978-1-4020-4102-0. URL https://doi.org/10.1007/1-4020-4102-0_7.
- Salaberri H., Altuna B.A., Aranzabe M.J., Arregi O., eta Díaz de Ilaraza A. bTime: a Hybrid Architecture for Capturing Temporal Information in Basque. *Knowledge-Based Systems*, Bidalita.
- Salaberri H., Arregi O., eta Zapiain B. bRol: The Parser of Syntactic and Semantic Dependencies for Basque. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 555–562, Hissar, Bulgaria, September 2015. INCOMA Ltd. Shoumen, BULGARIA. URL <http://www.aclweb.org/anthology/R15-1072>.
- Salaberri Izko H. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Donostia, Euskal Herria, 2017. URL http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8953/Tesia_azken_bertsioa_2017_05_21.pdf.
- San Martín I. Derived nominals from the nominal perspective. *Anuario del Seminario de Filología Vasca Julio de Urquijo*, 43(1–2):831–846, 2009. ISSN 0582-6152. <http://www.ehu.eus/ojs/index.php/ASJU/article/view/1764>.
- Sandhaus E. The New York Times Annotated Corpus Overview. Barne-txostena, The New York Times Company, Research and Development, New York, 2008. https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf.
- Saurí R. *A Factuality Profiler for Eventualities in Text*. Doktoretza-tesia, Faculty of the Graduate School of Arts and Sciences, Waltham, MA, USA, 2008. URL http://www.cs.brandeis.edu/~roser/pubs/sauriDiss_1.5.pdf. AAI3304029.
- Saurí R. eta Badia T. Catalan TimeBank 1.0. Barne-txostena, Linguistic Data Consortium, Philadelphia, Pennsylvania, 2012. <https://catalog.ldc.upenn.edu/LDC2012T10>.
- Saurí R., Batiukova O., eta Pustejovsky J. Annotating Events in Spanish. TimeML Annotation Guidelines. Barne-txostena, Technical Report

- Version TempEval-2010., Barcelona Media-Innovation Center, 2009. URL https://www.uam.es/gruposinv/upstairs/upstairs2/curricula/trabajos/batiukova_2009_sauri_et_al_eventGuidelines.pdf.
- Saurí R., Knippen R., Verhagen M., eta Pustejovsky J. Evita: A Robust Event Recognizer for QA Systems. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, 700–707, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL <https://doi.org/10.3115/1220575.1220663>.
- Saurí R., Littman J., Knippen B., Gaizauskas R., Setzer A., eta Pustejovsky J. TimeML Annotation Guidelines, Version 1.2.1, 2006. URL https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf.
- Schilder F. eta Habel C. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. *Proceedings of the Workshop on Temporal and Spatial Information Processing - Volume 13*, TASIP '01, 9:1–9:8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. <http://dx.doi.org/10.3115/1118238.1118247>.
- Schuurman I., Hoste V., eta Monachesi P. Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., eta Tapias D., editoreak, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2471–2477, Valletta, Malta, 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/162_Paper.pdf.
- Setty V., Bedathur S., Berberich K., eta Weikum G. InZeit: Efficiently Identifying Insightful Time Points. *Proceedings of the VLDB Endowment*, 3(1-2):1605–1608, September 2010. ISSN 2150-8097. URL <http://dx.doi.org/10.14778/1920841.1921050>.
- Setzer A. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Doktoretza-tesia, University of Sheffield, Sheffield, UK, 2001. ftp://ftp.dcs.shef.ac.uk/home/robertg/theses/setzer_thesis.pdf.

BIBLIOGRAFIA

- Skukan L., Glavaš G., eta Šnajder J. HeidelTime.HR: Extracting and Normalizing Temporal Expressions in Croatian. *Proceedings of the 9th Language Technologies Conference*, 99–103, 2014. URL http://nl.ijs.si/isjt14/proceedings/isjt2014_17.pdf.
- Smith C.S. *The parameter of aspect*. Kluwer, 1991.
- Strötgen J. eta Gertz M. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, 321–324, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1859664.1859735>.
- Strötgen J. eta Gertz M. WikiWarsDE: a German Corpus of Narratives Annotated with Temporal Expressions. In Hedeland H., Schmidt T., eta Wörner K., editoreak, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, 129–134, Hamburg University, 2011. URL <https://people.mpi-inf.mpg.de/~jstroetge/papers/2011-GSCL-StroetgenGertz-WikiWarsDE.pdf>.
- Strötgen J. eta Gertz M. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013. URL <https://link.springer.com/article/10.1007/s10579-012-9179-y>.
- Stubbs A. MAE and MAI: Lightweight Annotation and Adjudication Tools. *Proceedings of the 5th Linguistic Annotation Workshop, LAW V '11*, 129–133, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-93-0. URL <http://dl.acm.org/citation.cfm?id=2018966.2018982>.
- Styler W.F., Bethard S., Finan S., Palmer M., Pradhan S., de Groen P.C., Erickson B., Miller T., Lin C., Savova G., *et al.*. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014. URL <http://aclweb.org/anthology/Q14-1012>.
- Sun W., Rumshisky A., eta Uzuner O. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Supplement): S5–S12, 2013. ISSN 1532-0464. URL <http://www.sciencedirect.com/>

- science/article/pii/S1532046413001032. 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.
- Taulé M., Martí M.A., eta Recasens M. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., eta Tapias D., editoreak, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf. ACL Anthology Identifier: L08-1222.
- TimeML Working Group. TimeML Annotation Guidelines version 1.3. Manuscript. Barne-txostena, Brandeis University, 2010.
- Tonelli S., Sprugnoli R., Speranza M., eta Minard A.L. NewsReader Guidelines for Annotation at Document Level. version final. Barne-txostena, Fondazione Bruno Kessler, 2014. <http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf>.
- Tran G.B., Herder E., eta Markert K. Joint Graphical Models for Date Selection in Timeline Summarization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, 1: Long Papers lib., 1598–1607, Beijing, China, 2015. URL <http://aclweb.org/anthology/P/P15/P15-1154.pdf>.
- Urizar R. *Euskal lokuzioen tratamendu konputazionala*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Donostia, 2012. URL <http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1342621075/publiakoak/TESIA>.
- Urizar R., Alegria I., Odriozola J.C., eta Ezeiza N. Euskarazko hitz anitzeko unitate lexikalen tratamendu konputazionala. *Beñat Oihartzabali Gorazarre. Anuario del Seminario de Filología Vasca Julio Urquijo*, 1–2(XLIII):891–908, 2009. URL <http://ixa.si.ehu.es/sites/default/files/dokumentuak/3333/Oihartzabal-09.pdf>.

BIBLIOGRAFIA

- UzZaman N. eta Allen J.F. Event and Temporal Expression Extraction from Raw Text: First Step towards a Temporally Aware System. *International Journal of Semantic Computing*, 4(04): 487–508, 2010. URL <https://pdfs.semanticscholar.org/e050/2efdc7fb1b992586885291c573207b14adef.pdf>.
- UzZaman N., Llorens H., Allen J.F., Derczynski L., Verhagen M., eta Pustejovsky J. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *Computing Research Repository (CoRR)*, abs/1206.5333, 2012. <http://arxiv.org/abs/1206.5333>.
- van de Camp M. eta Christiansen H. Resolving relative time expressions in Dutch text with Constraint Handling Rules. *Revised Selected Papers of the 7th International Workshop on Constraint Solving and Language Processing - Volume 8114*, CSLP 2012, 166–177, New York, NY, USA, 2013. Springer-Verlag New York, Inc. ISBN 978-3-642-41577-7. URL http://dx.doi.org/10.1007/978-3-642-41578-4_10.
- van Son C., van Erp M., Fokkens A., eta Vossen P. Hope and Fear: How Opinions Influence Factuality. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editoreak, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3857–3864, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/188_Paper.pdf. ACL Anthology Identifier: L14-1194.
- Vendler Z. Verbs and times. *The philosophical review*, 143–160, 1957.
- Verhagen M. The Brandeis Annotation Tool. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., eta Tapias D., editoreak, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/740_Paper.pdf.
- Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G., eta Pustejovsky J. SemEval-2007 Task 15: TempEval Temporal Relation Identi-

- tification. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 75–80, Prague, Czech Republic, 2007. URL <http://aclweb.org/anthology/S/S07/S07-1014.pdf>.
- Verhagen M., Knippen R., Mani I., et al. Pustejovsky J. Annotation of temporal relations with Tango. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2249–2252, Genoa, Italy, 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/638_pdf.pdf.
- Verhagen M., Mani I., Sauri R., Knippen R., Jang S.B., Littman J., Rumshisky A., Phillips J., et al. Pustejovsky J. Automating Temporal Annotation with TARSQI. *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions, ACLdemo '05*, 81–84, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. <http://dx.doi.org/10.3115/1225753.1225774>.
- Verhagen M. et al. Pustejovsky J. Temporal Processing with the TARSQI Toolkit. *22nd International Conference on Computational Linguistics: Demonstration Papers, COLING '08*, 189–192, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1599288.1599300>.
- Vossen P., Rigau G., Serafini L., Stouten P., Irving F., et al. Hage W.V. News-Reader: recording history from daily news streams. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., et al. Piperidis S., editoreak, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2000–2007, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/436_Paper.pdf. ACL Anthology Identifier: L14-1374.
- Widlöcher A. et al. Mathet Y. La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France, 2009. URL <https://hal.archives-ouvertes.fr/hal-01011969/document>.

BIBLIOGRAFIA

- Wonsever D., Malcuori M., eta Rosá A. SIBILA: Esquema de anotación de eventos. Barne-txostena, Instituto de Computación, Universidad de la República, 2008. <https://www.colibri.udelar.edu.uy/handle/123456789/3419>.
- Wonsever D., Rosá A., Malcuori M., eta Etcheverry M. TEMANTEX: A Markup Language for Spanish Temporal Expressions and Indicators. *Research in Computing Science*, 97:9–19, 2015. ISSN 1870-4069. URL http://www.rcs.cic.ipn.mx/2015_97/TEMANTEX_%20A%20Markup%20Language%20for%20Spanish%20Temporal%20Expressions%20and%20Indicators.pdf.
- Wonsever D., Rosá A., Malcuori M., Moncecchi G., eta Descoins A. Event annotation schemes and event recognition in spanish texts. In Gelbukh A., editorea, *Computational Linguistics and Intelligent Text Processing*, 7182 lib. of *Lecture Notes in Computer Science*, 206–218. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28600-1. URL http://dx.doi.org/10.1007/978-3-642-28601-8_18.
- Wu M., Li W., Lu Q., eta Li B. CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. *Natural Language Processing-IJCNLP 2005*, 694–706. Springer, 2005. URL <http://aclweb.org/anthology/I/I05/I05-1061.pdf>.
- Xue N., Xia F., Chiou F.D., eta Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238, 2005. URL <https://pdfs.semanticscholar.org/2c72/257ae7a4a32dc60569f4e1fe4504b2678112.pdf>.
- Yaghoobzadeh Y., Ghassem-Sani G., Mirroshandel S.A., eta Eshaghzadeh M. ISO-TimeML Event Extraction in Persian Text. *Proceedings of COLING 2012*, 2931–2944, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. <http://www.aclweb.org/anthology/C12-1179>.
- Yan R., Wan X., Otterbacher J., Kong L., Li X., eta Zhang Y. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11*, 745–754. ACM, 2011. URL <http://www.cis.pku.edu.cn/faculty/system/zhangyan/papers/SIGIR2011-yanrui.pdf>.

Zabala I. Gramatika-kategoriak eta eglu. *Euskera*, 14(2):865–896, 1999. ISSN 0210-1564.

Zhong X. eta Cambria E. Time Expression Recognition Using a Time-related Tagging Scheme. *Proceedings of the International World Wide Web Conference 2018*, Lyon, France, April 2018. Association for Computing Machinery. <http://sentic.net/time-expression-recognition.pdf>.

Zubeldia L. “OMEN” partikularen azterketa semantikoa eta pragmatikoa. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Gasteiz, 2010. http://www.euskara.euskadi.eus/appcont/tesisDoctoral/PDFak/Larraitx_Zubeldia_TESIA.pdf.

Zubiri I. eta Zubiri E. *Euskal Gramatika Osoa*. Didaktiker, Bilbao, 1995.

ERANSKINAK



Abiarazle lexikoen zerrenda

A.1 taula – Euskarazko abiarazle lexiko nagusien zerrenda

Mota	Formak ¹
Asteke egunak	astelehen astearte asteazken ostegun ostiral larunbat igande
Hilabeteak	urtarril otsail martxo apiril maiatz ekain uztail abuztu irail urri azaro abendu

¹Euskara batuko formak baino ez ditugu jaso taula honetan.

Eguneko parteak	goiz eguerdi arratsalde iluntze gau goizalde
Urtaroak	udaberri uda udazken negu
Egun seinaleak	Gabon San Joan solstizio
Denborazko adberbioak	gaur atzo bihar etzi etzidamu goiz berandu lehen orain gero
Denbora unitateak	milurteko mende hamarkada urte hilabete aste egun ordu minutu segundo

Iraupenak	aro sasoi garai seiurteko hiruhileko
-----------	--

EusTimeMLko etiketatze-eskema

4. kapituluan EusTimeML etiketatze-eskema deskribatu dugu. Eranskin honetan etiketatze-eskemaren etiketa, atributu eta balio guztiak aurkezten ditugu Backus-Naur formatuan (BNF). Etiketa bakoitzerako, lehenik atributuak zerrendatu ditugu eta gero bakoitzaren balio posibleak adierazi ditugu.

Informazio linguistikoz gain BNFa informazio gehigarria eskaintzen du. Hasteko, **attributes** zerrendan etiketa bakoitzak zein atributu har dezakeen adierazi da. Kortexete artean agertzen direnak hautazkoak dira eta parentesi artean agertzen direnak horietako bat aukeratzekoak. Bigarrenik, informazioa linguistikoa zein prozesamendurako informazioa adierazten duten atributuak agertzen direla kontuan izan behar dugu. Azkenik, atributuen balioak zein motatakoak diren ere definitzen da. Atributu batzuek aurretik definitutako balioak hartzen dituzte, beste batzuek testua (CDATA) hartzen dute eta identifikatzaileak adierazten dituztenek identifikatzaile formatua edo beste identifikatzaile baten erreferentzia (IDREF) hartzen dute baliotzat.

B.1 Gertaeren BNFa

```
attributes ::= eid eiid class tense aspect pos polarity
modality
eid ::= e<integer>
eiid ::= ei<integer>
class ::= 'REPORTING' | 'PERCEPTION' | 'ASPECTUAL' | 'I_ACTION'
        | 'I_STATE' | 'STATE' | 'OCCURRENCE'
```

```
pos ::= 'VERB' | 'NOUN' | 'ADJECTIVE' | 'ADVERB' | 'PRONOUN' |  
      'OTHER'  
tense ::= 'PRESENT' | 'PAST' | 'ALEGIAZKOA' | 'NONE'  
aspect ::= 'PERFECT' | '-PERFECT' | 'FUTURE' | 'NONE'  
polarity ::= 'NEG' | 'POS' {default, if absent, is 'POS'}  
modality ::= 'AHAL' | 'NAHI' | 'BEHAR' | 'NONE'  
certainty ::= 'CERTAIN' | 'UNCERTAIN' | 'UNDERSPECIFIED'  
factuality ::= 'FACTUAL' | 'COUNTERFACTUAL' | 'NON_FACTUAL' |  
              'NONE' | 'UNDERSPECIFIED'  
specialCases ::= 'CONDITIONAL_CONDITION' | 'CONDITIONAL_MAIN' |  
                'GENERIC_STATEMENT' | 'NONE'
```

B.2 Denbora-adierazpenen BNFa

```
attributes ::= tid type [functionInDocument][beginPoint]  
            [endPoint][quant][freq][temporalFunction]  
            (value | valueFromFunction)[mod][anchorTimeID]  
tid ::= ID  
tid ::= TimeID  
TimeID ::= t<integer>  
type ::= 'DATE' | 'TIME' | 'DURATION' | 'SET'  
beginPoint ::= IDREF  
beginPoint ::= TimeID  
endPoint ::= IDREF  
endPoint ::= TimeID  
quant ::= CDATA  
freq ::= CDATA  
functionInDocument ::= 'CREATION_TIME' | 'EXPIRATION_TIME' |  
                      'MODIFICATION_TIME' | 'PUBLICATION_TIME' |  
                      'RELEASE_TIME' | 'RECEPTION_TIME' |  
                      'NONE' (default, if absent, is 'NONE')  
temporalFunction ::= 'true' | 'false'  
                  (default, if absent, is 'false')  
{temporalFunction ::= boolean}  
value ::= CDATA  
valueFromFunction ::= IDREF  
{valueFromFunction ::= TemporalFunctionID}
```

```
TemporalFunctionID ::= tf<integer>}
mod ::= 'BEFORE' | 'AFTER' | 'ON_OR_BEFORE' | 'ON_OR_AFTER' |
'LESS_THAN' | 'MORE_THAN' | 'EQUAL_OR_LESS' | 'EQUAL_OR_MORE'
| 'START' | 'MID' | 'END' | 'APPROX' |
'NONE' (default, if absent, is 'NONE')
anchorTimeID ::= IDREF
anchorTimeID ::= TimeID
```

B.3 Seinaleen BNFa

```
attributes ::= sid
sid ::= s<integer>
```

B.4 TLINKen BNFa

```
attributes ::= [lid] [origin] (eventInstanceID | timeID)
[signalID] [syntax]
                (relatedToEventInstance | relatedToTime)
                relType
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
origin ::= CDATE
eventInstanceID ::= IDREF
{eventInstanceID ::= EventInstanceID}
timeID ::= IDREF
{timeID ::= TimeID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEventInstance ::= IDREF
{relatedToEventInstance ::= EventInstanceID}
relatedToTime ::= IDREF
{relatedToTime ::= TimeID}
relType ::= 'BEFORE' | 'AFTER' | 'IBEFORE' | 'IAFTER' |
'INCLUDES' | 'IS_INCLUDED' | 'MEASURE' |
'SIMULTANEOUS' | 'BEGINS' | 'BEGUN_BY' |
```

```
        'ENDS' | 'ENDED_BY' | 'IDENTITY'
syntax ::= CDATA
```

B.5 SLINKen BNFa

```
attributes ::= [lid] eventInstanceID
[signalID] subordinatedEventInstance relType [syntax]
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
eventInstanceID ::= IDREF
{eventInstanceID ::= EventInstanceID}
subordinatedEventInstance ::= IDREF
{subordinatedEventInstance ::= EventInstanceID}
signalID ::= IDREF
{signalID ::= SignalID}
relType ::= 'MODAL' | 'EVIDENTIAL' | 'NEG_EVIDENTIAL'
| 'FACTIVE' | 'COUNTER_FACTIVE' | 'CONDITIONAL'
syntax ::= CDATA
```

B.6 ALINKen BNFa

```
attributes ::= [lid] eventInstanceID [signalID]
relatedToEventInstance relType [syntax]
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
eventInstanceID ::= ID
{eventInstanceID ::= EventInstanceID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEventInstance ::= IDREF
{relatedToEventInstance ::= EventInstanceID}
relType ::= 'INITIATES' | 'CULMINATES' | 'TERMINATES'
| 'CONTINUES' | 'REINITIATES'
syntax ::= CDATA
```