eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Euskarazko denbora-egituren azterketa eta lehen etiketatze-eskemaren proposamena

Begoña Altuna
Tutorea: Mª Jesús Aranzabe

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua

lortzeko bukaerako proiektua

2013ko iraila

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

LABURPENA

Testuetan agertzen den denborari buruzko informazioa oso baliagarria da testuaren ulermenerako, gertakariak denboran kokatzen laguntzen baitu. Denboraren araberako antolakuntza hori baliatu ahal izatea oso aurrerakuntza interesgarria da Hizkuntzaren Prozesamenduko hainbat sistemarentzat, hala nola, informazioa erauzteko sistementzat, itzulpen automatikoarentzat, galdera-erantzunen sistementzat eta laburpen automatikoarentzat. Sistema horiek denbora-informazioa atzigarri izan dezaten, testuan agertzen diren denbora-egiturak analizatu, sailkatu eta ordenagailuak prozesatzeko moduan, informazioa etiketen bidez atzigarri bihurtuz, jarri behar dira.

Lan honetan hizkuntza naturaleko testu batzuetako denbora-egitura batzuk aztertu dira eta horien ezaugarririk nabarmenenak etiketatzeko proposamena gauzatu da beste hizkuntzetan egin diren etiketatze-lanetan oinarrituz eta euskararako bide aproposenak aukeratuz.

Hitz gakoak: denbora-egiturak, analisia, etiketatzea, etiketatze-eskema, hizkuntzalaritza, konputazioa.

ABSTRACT

The temporal information in the texts is very useful for the textual understanding as it helps to position events in time. Being able to benefit from this time based organization is a very interesting improvement for computational resources like information extraction, machine translation, question answering and automatic summarization. In order to make temporal information accessible to these resources, the temporal structures in the text have to be analysed, classified and prepared to be automatically processed, making the information accessible by means of tags.

In this essay some temporal expressions of some texts have been analysed and a tagging scheme for their main features has been proposed based on the tagging approaches in other languages and trying to choose the most appropriated options for Basque

Key words: temporal structures, analysis, tagging, tagging scheme, linguistics, computation

Aurkibidea

1 Sarrera

Sarean etengabe ari da informazioa hazten; milioika datu sortzen dira uneoro eta horrek erronka handiak dakarzkio hizkuntza-teknologiaren arloari, informazioari etekin handiagoa lortzeko hizkuntzarentzako konputazio-tresnak beharrezkoak baitira. Hizkuntzaren analisi eta prozesamendua egitean, testuaren ulermen automatikoa bilatzen da behinik behin; zenbat eta testuaren analisi zabalagoa izan, orduan eta hobeto baliatu ahal izango da testuko informazioa ulermena handitu egingo baita. Testua bere osotasunean ulertzeko, denbora-egituren analisia beste urrats bat da. Denbora-egiturek testua egituratzen laguntzen dute: gertakariak kronologian kokatzen dituzte eta testuari kohesioa ematen diote, eta ondorioz, testua nola egituratu den ulertzen laguntzen dute gero beste testu batzuk sortu edo analizatzeko.

(1), (2) eta (3) adibideetan, lan honetarako baliatu den EPEC corpusetik (Euskararen Prozesamendurako Erreferentzia Corpusa) (Aduriz *et al*., 2006) laginetik hartu diren denbora-egitura hauek ikus daitezke azpimarraturik: *arestian*, *esperientzia egiten hasi baino lehen*, *ondoren*, *behaketa hauek egin ondoren*, *orduan*, *udan*, *neguan* eta *batzuetan*:

> (1) Oinarrian bera dela esan dezakegu, arestian esan dugunaren ildotik behintzat.
>
> (2) Esperientzia egiten hasi baino lehen presta itzazu behar dituzun materialak. Ondoren, bi behaketa hauek egin ditzakezu: [...] Behaketa hauek egin ondoren, atera dituzun ondorioak idatz itzazu.
>
> (3) Baina gehien gustatzen zaidan garaia udaberria izaten da, orduan, muinoak berdetu eta lorez estaltzen baitira. Udan, bestalde, ez du euririk egiten eta loreak usteldu egiten dira. Neguan hotz gehiegi egiten du eta batzuetan elurra ere bai.

Adibideotan argi ikus daiteke denbora-adierazpenek testua egituratzeko duten ahalmena gertakariak, gertatzen diren ekintzak edo egoerak, denboraren arabera ordenatzen laguntzen baitute. (1) adibidean *arestian* esamoldeak lehen esandako

zerbaitekin lotzen du esaldia, (2) adibidean gertakariak ordenatzen laguntzen dute denbora-esapideek eta azkenik (3) adibidean denbora zehatz batzuetan zer gertatzen den adierazteko balio dute.

Azken urteotan hainbat esperimentu eta ikerketa garatu izan dira denbora adierazpenen ezagutza eta etiketatze-lanaren gainean, adierazpenon ezagutzak testuen prozesatze automatikoan duen garrantziarengatik.

Denbora-informazioa atzigarri egiteko prozesuak hiru urrats ditu: denbora-egituren analisia, egituron etiketatzea eta denboraren arabera etiketatutako corpusa sortzea. Ehrmann & Hagègeren (2009) arabera, bost atal ditu testuen denboraren araberako analisiak:

a. Denbora-adierazpenen ezagutza eta deskribapena

b. Denbora-adierazpenen normalizazioa eta zeri egiten dioten erreferentzia identifikatzea

c. Gertakarien ezagutza (baita euren denbora gramatikalarena eta aspektuarena ere)

d. Gertakari-gertakari edo gertakari-denbora-adierazpen erlazioak egitea

e. Denboraren inferentzia

Denbora-egiturak etiketatzeko prozesuan etiketatze-eskemak sortu izan dira – Ehrmann & Hagègek (2009) proposatzen duten analisitik ateratako informazioa ordenagailuarentzat irakurgarri egin ahal izateko –, baita denboraren arabera etiketatuta dauden corpusak ere. Horretarako, denbora-egiturak linguistikoki aztertu eta sailkatuko dira, ondoren corpusetan identifikatzeko eta etiketatzeko. Corpus horietan oinarrituta denbora-egituren identifikatzaile automatikoak sortzen dira eta identifikatzaileok baliagarriak izango dira besteak beste Hizkuntzaren Prozesamendurako (HP) hainbat sistematan, hala nola, itzulpen automatikoan, galdera-erantzun sistemetan, testuen laburpen automatikoan eta informazioa erauzteko sistemetan.

Ikerketa-lan honetan denbora-egitura batzuk: denbora-aditzondo eta adizlagunak, denborazko perpausak, denborazko lokailuak, orduak eta datak ezagutu eta deskribatuko dira, baita horientzako formatu normalizatua proposatu ere. Lanaren

egitura osoa ondokoa izango da: lehenik eta behin, sarrera honen ondoren, 2. atalean, denbora nozioa hizkuntzan nola gauzatzen den aztertuko da. Ondoren, 3. atalean hainbat hizkuntzatan egin diren etiketatze-eskemak eta denbora-egiturak etiketatuta dituzten corpusak aurkeztuko dira. Jarraian, 4. atalean, euskarazko denbora-egiturak definitu eta sailkatuko dira eta ondoren, 5. atalean, euskararentzat oinarritzat hartuko den TimeML etiketatze-eskema azalduko da. 6. atalean, aurreko bi ataletako informazioa baliaturik, euskarazko denbora-egitura batzuentzako etiketatze proposamena aurkeztuko da. Eta amaitzeko, ondorioak eta etorkizunerako lanak eta bibliografia azalduko dira 7. eta 8. ataletan hurrenez hurren.

2 Denbora nozioa hizkuntzan

Denbora gertakarien iraupena edo euren arteko bereizketa neurtzeko balio duen magnitude fisikoa da. Denboraz hitz egiten denean, bi denbora bereizi behar dira: denbora kronologikoa, erloju eta egutegien bidez neurtzen edo adierazten dena eta erreferentzia bat emateko balio duena, eta gizakion pertzepzioak neurtzen duen denbora psikologikoa, norberaren araberakoa eta ez datorrena beti bat denbora kronologikoarekin.

Hizkuntzan, bigarren ikuspegia nagusitzen da, gizakion pertzepzioak denboraren araberako leihokatze baten modura funtzionatzen baitu (Evans, 2007), hau da, esperientziak denbora tarte baten barruan ulertzen ditu gizakiak eta horretatik hurbil daude normalean erabiltzen diren denbora-egiturak (4) (5):

(4) *Me pasaré <u>más tarde</u> por tu casa a recoger los libros*. (<u>Geroago</u> pasatuko naiz zure etxetik liburuak hartzeko)

(5) *I had to queue <u>for a long time</u> to buy the tickets*. (<u>Denbora luzez</u> egon nintzen ilaran sarrerak erosteko)

Aurreko (4) eta (5) adibideetan ikus daiteke nola leihokatzen den gure denbora pertzepzioa. (4) adibidean uneko leihotik kanpo dagoen une bati buruz (*más tarde*, geroago) hitz egiten da eta (5) adibidean oso denbora tarte edo leiho luzeari (*for a long time*, denbora luzez) egiten zaio erreferentzia.

Esan gabe doa hizkuntza natural guztiek denbora adierazteko egiturak dituztela, horiek funtsezkoak baitira diskurtsoa antolatzeko. Mota honetako baliabideak dira *lehenik eta behin*, *ondoren* edo *azkenik* modukoak, baita aditzen *tempus*a (denbora gramatikala) ere. Hala eta guztiz ere, denbora erreferentzia ulertzeko esate-unea ezagutu behar dugu edo testuan aurretik agertutako erreferentziak gogora ekarri behar ditugu. Denbora-egitura batzuek balio deiktikoa dute eta esate-unearekin guztiz loturik daude. *Orain* hitzak, adibidez, iterazio-uneari (edo hurbileko bati) egiten dio erreferentzia.

Jarraian datorren (6) adibideko esaldia ezin koka dezakegu esate-unea edo testuaren sorrera unea ezagutu gabe. Beste esapide batzuek (7) (8), aitzitik, ez dute balio deiktikorik ez baitira semantikoki esate-unearen arabera kodetzen; hau da, ez dute esate-unea ezagutzearen beharrik ulertu ahal izateko jarraian datozen (7) eta (8) adibideetan gertatzen den bezala.

(6) *On ira à la montagne le dimanche*. (Mendira joango gara igandean)

(7) *Me limpio los dientes antes de ir a dormir* (Ohera joan aurretik garbitzen ditut hortzak).

(8) *Me limpiaré los dientes antes de ir a dormir* (Ohera joan aurretik garbituko ditut hortzak).

Bestalde, denbora-egiturek bi motatako errealitateak adieraz ditzakete: kronologian koka daitezkeen uneak (9) eta iraupenak (10). Lehenengoekin kronologian koka daitekeen une bati egiten zaio erreferentzia – hizketaren unearekiko erlatiboa izango dena – eta bigarrenekin gertakari baten iraupena adieraziko da kronologian une zehatz batean kokatzeko gai ez bagara ere.

(9) Gaur goizean etxea batu eta liburutegira etorri naiz.

(10) Bi egun eman zituen mendian galdurik.

Boroditskyk (2011) kultura ezberdinetako gizakiek denbora adierazteko antzeko egiturak erabiltzen dituztela adierazten du, fisikoagoak diren gauzetan oinarritutako kontzeptualizazioak egiten dituztela, alegia. Gure inguruko hizkuntzetan eta gurean denbora-espazio erlazioa sortzen da, adibidez:

(11) Epaiketak luze jo zuen.

(12) *Pienso comprarme una casa en un futuro cercano*. (Epe laburrean etxe bat erostea pentsatzen dut)

(13) *He got out of the jungle after some long days of hunger and thirst*. (Oihanetik atera zen egun luze batzuk goseak eta egarri igaro ondoren)

(14) E un altro giorno *è andato*, la sua musica ha finito/ quanto tempo *è* ormai *passato* e *passerà*? (F. Guccini) (Eta beste egun bat joan da, bere musika amaitu da/ zenbat denbora igaro da jada eta igaroko da?)

(11-14) adibideetan ikusten denez, denbora "luzea" da, "hurbil" dago edo "joan" edo "igaro" egiten da. Denak espazioari eta distantziari loturiko kontzeptuak.

Hizkuntza guztietan ez da denbora-espazio erlaziorik sortzen, ordea, hiztunek munduaren kontzeptualizazio ezberdina egiten baitute batzuetan. Gainera, denboraren kontzeptualizazioren aspekturen batean bat datozen hizkuntzek ez dute zertan beste batzuetan etorri. Horren adibide egunaren banaketan aurki dezakegu. Euskarak eta frantsesak denbora-espazio lotura egiten badute ere, euskaraz *goiza*, *eguerdia*, *arratsaldea*, *iluntzea*, *gaua* eta *goizaldea* bereizten dituzte hiztunek. Hiztun frantsesek, ordea, *matin*, *après midi*, *soire* eta *nuit* baino ez dituzte ezberdintzen eta denboraren kontzeptualizazio ezberdin honek denbora nozioaren interpretazioan ere eragina du.

Denbora-egiturek adieraz dezaketen ahalik eta informazio gehien ordenagailuarentzat ulergarri egitea da lan honen helburua, denbora-informazio esplizituaz gain, hiztunek euren mundu-ezagutzaren bidez baino deskodetu ezin dezaketen informazioa baitago. Hurrengo ataletan azalduko da ingelesez, frantsesez, errumanieraz etab. erabili diren informazioa esplizitatzeko teknikak, baita euskararen denbora-egitura batzuentzat proposatu direnak ere.

3 Aurretikoak

Sarrerako atalean esan bezala, batetik, oso garrantzitsua da denbora-egiturak aztertzea eta ulertzea egiturok diskurtsoa antolatzeko duten ahalmenagatik. Bestetik, Setzer & Gaizauskasek (2000) esaten dutenez, gertakari gehienak ez dira betiko egiazko, eta gerta daiteke inolako erabilgarritasunik ez izatea gertakari bat jazo izana jakiteak baldin eta noiz jazo den ez badakigu. Horregatik garrantzitsua da denbora-egiturek barnean daramaten ahalik eta informazio gehien agerian uztea.

Denbora-egituren analisia egin ostean, egiturok ordenagailu bidez prozesatzeko moduan jarri behar dira. Horretarako denbora-informazioa adierazi ahal izango duten etiketa-sistemak eta etiketatze-eskemak sortu izan dira eta etiketaturiko denbora-egiturak corpusetan batu izan dira.

Denbora-egituren analisia XX. mendearen erdialdean hasi zen Reichenbachen (1947) eskutik. Orduko hurbilpena eta ondoren etorri zirenak (Moens & Steedman, 1988; Lascarides & Asher, 1993) Filosofiaren ikuspuntutik egin ziren eta hizkuntzaren kodetze logikoa bilatzen zuten. Ondoren, denbora-egituren arabera etiketatutako corpusak sortzen hasi ziren (Setzer & Gaizauskas, 2000), baita corpus horiek etiketatzeko etiketatze-eskemak ere. Etiketatze-eskema horien adibide eta egungo eskementzat eredu dira TIDES proiektuan (Ferro *et al*., 2005b) definitu den etiketatze-eskema (Ferro *et al*., 2001) eta TimeML bera (Pustejovsky *et al*., 2003a). Azken honetan oinarrituta markatu zen esaterako TIMEBANK corpusa (Pustejovsky *et al*., 2003b). Egun, denbora-egiturak etiketatzeko eskemak zabaltzen eta hobetzen ari dira hainbat ikertzaile (Ehrmann & Hagège, 2009; Bittar et *al*., 2012) testutik ahalik eta informazio gehien eskuratu eta automatikoki tratatu ahal izateko.

XXI. mendeko lehen hamarkadan denbora-egiturak antzematea helburu bazen, egun egitura horiei edozein hizkuntzatarako balioko duen etiketa estandarra ematea da erronka. Italieraz (Caselli *et al*., 2011), koreeraz (Im *et al*., 2009) eta frantsesez (Bittar *et al*., 2011) edota alemanez (Spreyer & Frank, 2008) egin izan dira ingelesarentzako sortutako etiketak moldatzeko saiakerak.

HPn denbora-egituren analisi eta ulermena behar-beharrezkoa gertatu da hainbat aplikazio eta baliabidetan: galdera-erantzun sistemetan (*noiz?*, *noiz arte?* eta abarri erantzuten dioten galderetan edo denborari buruzko informazioa behar duten galderetan), informazio erauzteko eta bilatzeko baliabideetan, itzulpen automatikoan, testuen laburpen automatikoan eta iritzi-meatzaritzan (*opinion mining*) besteak beste. Sistema horiek guztiek corpusetan aurkitzen dute beharrezko informazioa eta erabiltzaileari behar duen informazioa itzultzen saiatzen dira.

Denbora-egituren informazioa programa informatikoentzako era egokian interpretatu ahal izateko, testuan agertzen diren egiturak identifikatuko dituzten etiketak sortzeaz gain, testuaren formatua ere egokitu behar da. Jarraian datozen azpiataletan denbora-egituren etiketatzean arrakastarik handiena izan duten etiketatze-eskemak aurkeztuko dira.

Ondoko (15) adibidean eta 1. irudian denbora-informazioa duen esaldi bat SGML lengoaian[1] etiketaturik ikus daiteke. Denbora-adierazpenak etiketa artean agertzen dira eta etiketa horietan ikus daiteke adierazpenotatik hartu den informazioa, kasu honetan denbora-adierazpenaren balio estandarizatua.

(15) *There were doughnuts at the 8:00 meeting this morning* (Donutsak egon dira goizean 8etako batzarrean).

There were doughnuts at the
<TIMEX2 VAL="1999-07-15-T08:00">8:00</TIMEX2> meeting
<TIMEX2 VAL="1999-07-15-TMO">this morning</TIMEX2>.

1. irudia: *There were doughnuts at the 8:00 meeting this morning* perpauseko denbora-egituren markaketa SGML lengoaia baliatuta

3.1 Lehen etiketatze-lanak

Denbora-egiturak etiketatzeko hainbat proposamen egin izan dira. Jarraian ondoren etorritako TimeML etiketatze-eskemaren oinarri izan ziren TIMEX2 etiketatze-eskema (Ferro *et al*., 2001) eta Setzerren etiketatze-eskema (2001) aurkeztuko dira.

[1] Standard Generalized Markup Language (ISO 8879:1986 SGML) dokumentuetako markatze lengoaia orokorrak definitzeko ISO araberako teknologia estandarra da.

3.1.1 TIMEX2 etiketatze-eskema

Ferrok *et al*.-ek (2001) TIMEX2 denbora-etiketak definitu zituzten. Denbora elementu lexikoren bat (abiarazle lexikoa: *month* (hilabete), *today* (gaur), *currently* (gaur egun)) zuten egiturak baino ez zituzten kontuan hartu eta egutegi gregoriarrean oinarritutako denbora-lerroa erabili zuten denbora-adierazpenak kokatzeko. Adierazpenen balioa adierazteko "YYYYMM[WW]DDhhmmss" (urtea, hilabetea, [astea], eguna, ordua, minutuak, segundoak) formatua baliatu zuten, ISO 8601 formatuan oinarrituz (Wolf & Wicksteed, 1997). Izan ere, formatu honen bidez oso zehaztasun maila handia lortzen da eta adierazpen puntukariak zein iraupenezkoak irudika daitezke. Adierazpen ez-berezituak, kronologiako une zehatz bati erreferentzia egiten ez diotenak (16), etiketatzeko gai zen, baita denbora-egitura bat zuten izen bereziak (*Black September*) (17) ez etiketatzeko ere, azkenok denbora-adierazpen baino entitate izenak baitira.

(16) Ekaineko egun eguzkitsu baten joan zen herritik.

(17) Inazio Mujikaren *Gerezi Denbora* liburua asko gustatu zitzaidan.

TIMEX2 etiketak SGML markaketa-lengoaia erabiliz sortu zituzten. Etiketa bakoitzak hainbat atributu hartzen zituen, garrantzitsuena VAL (balioa) izanik, horrek ematen baitzuen denbora-egitura kronologian kokatzeko informazioa. Etiketatze-eskema horretan, ordea, ez zen denbora-egitura guztientzat etiketa bat eskaintzen; izen (*Monday*, astelehena), izen sintagma (*last week*, aurreko astea), adjektibo (*previous*, aurreko), adberbio (*currently*, gaur egun), adjektibo-sintagmak (*ten weeks long*, hamar asteko) eta adberbio-sintagmak (*two days ago*, duela bi egun) baino ez ziren kontuan hartzen, eta ez preposiziodun egiturak (18) espero zitekeen moduan:

(18) *He wanted all the work done before last Friday* (lan guztia aurreko ostirala baino lehenago amaitua nahi zuen).

Hain zuzen ere, (18) adibideko esaldian "*before last Friday*" denbora-egitura bakartzat hartu behar izango litzateke, denbora-egiturak aurreko ostiralaren aurreko une bati (*before last Frida*") egiten baitio erreferentzia eta ez aurreko ostiralari (*last Friday*). Baina Ferro *et al*.-en (2001) proposamenak ez zituen preposiziodun izen sintagmak aintzakotzat hartzen.

Aitzitik, etiketatze-eskema horretan denbora-egiturak euren artean erlazionatzeko aukera ere deskribatzen zen ondoko adibidean (19) ikus daitekeenez:

(19) *I'm leaving on vacation two weeks from next Tuesday* (Oporretan joango naiz datorren asteartetik hasi eta bi astera).

I'm leaving on vacation **<TIMEX2 VAL="1999-08-03">two weeks from <TIMEX2>VAL="1999-07-20">next Tuesday</TIMEX2></TIMEX2>**.

2. irudia: *I'm leaving on vacation two weeks from next Tuesday* perpauseko denbora-egituraren etiketatzea

(19) adibidean ikusi ahal den bezala, denbora-egitura handi baten (*two weeks fom next Tuesday*) barruan dagoen denbora-egitura laburragoa (*next Tuesday*) aurki daiteke ingelesez (Ferro *et al*., 2001). Horregatik denbora-egitura luzeari etiketa eta balio bat ematen zaio (VAL=”1999-08-03”), egitura osoarena izango dena, eta egitura laburrari bere etiketa propioa (<TIMEX2>VAL=”1999-07-20”>), baina denbora egitura orokorraren eremuaren barrukoa.

3.1.2 Setzerren etiketatze-eskema

Setzerrek (2001) beste etiketatze-eskema bat proposatu zuen, hau ere SGML markaketa-lengoaian idatzitakoa. Etiketatze-eskema horrekin bai denbora-egiturei, bai gertakariei etiketa bat eta hainbat atributu (identifikatzailea, klasea, aspektua, beste gertakariekiko erlazioa eta abar) eman nahi izan zien. Ez ziren kasuistika guztiak kontuan hartu, ordea, eta gertakarien artetik pertzepziozko, aspektuzko eta estilo ez-zuzena adierazten duten gertakariak kanpoan geratu ziren. Gertakariek ondoko (20) adibidea etiketatzeko erabili diren etiketen moduko etiketak hartzen zituzten:

(20) *All 75 people on board the Aeroflot Airbus died when it ploughed into a Siberian mountain* (Aerofloteko Airbusean zihoazen 75 pertsonak hil ziren hegazkinak Siberiako mendi baten kontra talka egin zuenean).

All 75 people on board the Aeroflot Airbus **<event eid=4 relatedToEvent=5 eventRelType=simultaneous> died </event>** when it **<event eid=5> ploughed </event>** into a Siberian mountain.

3. irudia: *All 75 people on board the Aeroflot Airbus died when it ploughed into a Siberian mountain* esaldiko gertakarien etiketatzea

3. irudian ikus daitezke gertakariak adierazteko erabili zuen <event> etiketa eta bere atributuak. Setzerren hurbilpenean gertakari bakoitzak identifikazio zenbaki bat, "eid" hartzen zuen eta gero informazio hori gertakarien arteko harremanak adierazteko baliatzen zuen "relatedToEvent" atributuan.

Denbora-egituren artean ere ez zituen guztiak etiketatu, errealitateko elementu bati (gertakari bat edo denbora bat) erreferentzia egiten diotenak baino ez. Denbora-egitura horien artean bereizketa bat egin zuen Setzerrek: egitura bakunak eta konplexuak euren aldetik tratatzea. Bere aburuz, egitura bakunak, denborako une edo iraupen bati erreferentzia egiten diotenak, orokorrean, preposizio-sintagmetako edo adberbio-sintagmetako izenak dira: *yesterday* (atzo), *in some days* (egun batzuk barru). Denborazko egitura bakunen artean beste bereizketa bat ere egin zuen: egun bat baino luzeagoak zirenei (*5th of June*, ekainaren 5a, adibidez) DATE etiketa eman zien eta eguna baino laburragoei (8:00, *afternoon*, eguerdia) TIME, Message Understanding Conference (MUC) konferentzietako ildoari jarraituz. Denek hartzen zuten identifikazio bakarra (ID) besteetatik bereizi ahal izateko, baita beste bi atributu ere: "type" (mota) eta "calDate" (egutegiko balioa). 4. irudian *Tuesday* (asteartea) denbora-egituraren etiketatzearen adibidea ikus daiteke:

```
<timex tid=5 type=DATE> Tuesday </timex>
```

4. irudia: *Tuesday* hitzaren etiketatzea

Egitura konplexuek, preposizio edo adberbio batez hasitako denborazko perpausek, denborako une bati egiten diote erreferentzia denbora tarte bat gertakari batekin erlazionatuz (21):

(21) *After 17 seconds hearing the sound*.

Egiturok etiketatzeko (*after* (21) adibidean) atributu berezi sorta bat proposatu zuen elementuen arteko erlazioak adierazi ahal izateko eta denbora-egituren eta gertakarien arteko erlazio hitzei (*before*, *after*, *while*...) "*signal*" (seinale) deitu zien. Etiketa horiek 5. irudian jaso dira (22) adibideko esaldia etiketatzeko helburuaz:

(22) *17 seconds <u>after</u> hearing*

<timex tid=5 type=complex eid=3 signalID=7 relType=after> 17 seconds </timex> **<signal sid=7> after </signal>** <event eid=3> hearing

5. irudia: *17 seconds <u>after</u> hearing* egituraren etiketatzea

Etiketatze-lan hori anotazio tresna grafiko baten bidez egin zen eta hiru fase izan zituen:

- Denbora-adierazpen guztiak (denbora-adierazpenak, seinaleak eta gertakariak).
- Esplizituki adierazitako erlazioak etiketatzea
- Inplizituki ulertzen ziren denbora-erlazioak.

Setzerren arabera gertakari guztiek behar dute denbora-adierazpen batekiko edo beste gertakari batekiko erlazioa eta horrela etiketatu zuen osotu zuen corpusa.

3.2 TimeML etiketatze-eskema

Aurreko bi etiketatze-eskemen ildoari jarraituz sortu zen TimeML eskema Pustejovsky *et al*.-en (2002, 2003a) eskutik. Horiek Ferro *et al*.-en (2001) eta Setzerren (2001) proposamenak hobetuz eta osatuz denbora-egitura guztiei etiketa bat esleitzeko etiketatze-eskema sendoa proposatu zuten. XML (eXtensible Markup Language) markaketa-lengoaiaren gainean garatu zen eta hizkuntza guztietan aplikagarri egin zen. 3.3.1 atalean azaltzen diren hainbat hizkuntzatako TIMEBANK corpusek (ingelesekoa, errumanierakoa, italierakoa...) TimeML etiketatze-eskema hainbat hizkuntzatako denbora-egiturak etiketatzeko baliagarri dela erakusten dute. Ondoko adibidean (23) eta 6. irudian TimeML baliatuta etiketaturiko esaldi bat (Pustejovsky *et al*., 2003a) ikus daiteke:

(23) *She was sick after the play*

She was
<EVENT eid="e1" class="STATE" tense="NONE" aspect="NONE">

```
sick
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"/>
<SIGNAL sid="s1">
after
</SIGNAL>
the
<EVENT eid="e2" class="OCCURRENCE" tense="NONE"
aspect="NONE">
play
</EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2"/>
<TLINK eventInstanceID="ei1" signalID="s1" relatedToEvent="ei2"
relType="AFTER"/>
```

6. irudia: *She was sick after the play* perpausaren etiketatze osoa

TimeML edozein denbora-egitura etiketatzeko erabil daiteke, egitura guztientzako etiketak eta atributuak definitu izan baitira. Gertakariak (<EVENT>), denbora-adierazpenak (<TIMEX3>) eta seinaleak (<SIGNAL>) etiketatzeaz aparte, hauen arteko erlazioak ere etiketa ditzake <LINK> etiketen bidez. Etiketa bakoitzak, gainera, hainbat atributu har ditzake hauen bidez denbora-egituren informazioa adierazi ahal izateko. 6. irudian bi gertakari ikus daitezke, *sick* eta *play*, eta euren arteko harremana (<TLINK> bidez adierazia) eta seinale bat, *after*, ikus daitezke.

Hizkuntza guztientzat erabil daiteke TimeML etiketatze-eskema eta horregatik eta duen estaldura zabalagatik erabilia izan da erdal hizkuntzetako corpusak etiketatzeko. Euskarazko denbora-egiturak etiketatzeko ere TimeML eskema erabiliko da eta horregatik eskainiko zaio arreta berezia 5. atalean.

3.3 Denbora-egiturak etiketatuta dituzten corpusak

Aurreko atalean (3.2) azaldutako etiketatze-eskemek ez dute zentzurik eurak baliatuz corpus bat etiketatuko ez bada. Corpus horietan informazio morfosintaktikoaz gain semantikoa ere batzen da, kasu honetan, denborari buruzko informazio semantikoa. Ondoren denbora-egituretan oinarrituta osatu diren hainbat hizkuntzatako corpusak azalduko dira.

3.3.1 TIMEBANK

TIMEBANK corpusa (Pustejovsky *et al*., 2003b) 300 kazetaritza testuz osoturik dago eta TimeML etiketatze-eskema jarraituz etiketatu ziren. Etiketatze-lana gauzatzeko etiketatzeko baliatu zuten tresna Alembic NLP system-en Alembic Workbenchen (Day *et al*., 1997) bertsio modifikatua da. Lehen urrats batean denbora-adierazpenak, gertakariak, seinaleak eta atributu batzuk etiketatu ziren eta bigarren urratsean, gertakarientzako aingurak eta denbora-egitura guztien arteko erlazioak ezarri ziren. TimeMLren XML etiketak zuzenean txertatu ziren testuan; hau da, testua eta XML etiketak batera agertzen ziren 6. irudian agertu modura.

TIMEBANK corpuseko testuak hainbat hizkuntzatara itzuli dira eta hizkuntza horientzako etiketatze-lanak egin izan dira: Forascu eta Tufisek (2012) errumanierarentzat, italierarentzat (Caselli *et al*., 2011), koreerarentzat (Im *et al*., 2009) eta frantsesarentzat (Bittar *et al*., 2011) eta alemanarentzat (Spreyer & Frank, 2008), TimeML markaketa-lengoaiak hizkuntza anitzetako testuak etiketatzeko balio duela frogatzeko.

Errumanieraren corpusari dagokionez, adibidez, lehen urrats batean ingelesezko TIMEBANK corpusa errumanierara itzuli zuten eta denbora-etiketak automatikoki txertatu zituzten lerrokatzeetan, bi hizkuntzetako esaldi eta egituren parekatzeetan, oinarrituta. Bigarren urrats batean, ISO etiketak txertatu zituzten horiek errumanieraren gramatika arauen arabera moldatuz; alegia, atributuak moldatuz eta atributu horientzat aukera egokiak txertatuz. Adibidez, errumanieraz aditzak *tempus* gehiago hartzen ditu eta guztiak irudikatzeko informazioa txertatu zuten, baita modua eta burututasuna irudikatzekoak ere. Azkenik, akatsak zuzendu eta estandarrari egokitu zitzaizkion. Beste hizkuntzetako corpusetan ere antzeko moldaketak egin izan ziren hizkuntza bakoitzaren berezitasunak islatu ahal izateko.

3.3.2 WikiWars corpusa

Aurretik aipatutako TIMEBANK corpusak kazetaritzako testuz osaturik daude eta, ondorioz, nahiko laburrak dira, denbora-egitura gutxi dute eta egiturok nahiko erraz etiketa daitezke dokumentuaren sorrera dataren (DCT, Document Creation Time) informazioa baliatuz. Etiketatze sinple hori, ordea, ez da gertatzen estilo narratiboagoa duten testu luzeagoetan, horietan erabiltzen diren denbora-egiturak eta erlazioak

konplexuagoak baitira. Horregatik, Mazur & Dalek (2010) gerrak narratzen dituzten ingelesezko Wikipediako 22 artikuluz osatutako corpus etiketatua sortu zuten SGML markaketa-lengoaia baliatuz.

Corpus hori Ferro *et al*.-ek (2001) proposatutako TIMEX2 etiketatze-eskemari jarraituz etiketatu zuten DANTE analizatzailea (Mazur & Dale, 2007) baliatuta. Behin analisi erdi-automatikoa gauzatuta, eskuz gainbegiratu ziren emaitzak Callisto etiketatzeko programa (Day *et al.*, 2004) erabilita. Azkenik ACE[2] lehiaketako APF XML (Agency Private Fares eXtensible Markup Laguage) formatura bihurtu zituzten corpuseko testuak.

Lan horrek agerian jarri zuen narraziozko historia testuetan eta kazetaritza testuetan denbora-egitura ezberdinak erabiltzen direla. Historia testuetan abiarazle lexiko ezberdinak erabiltzen dira (*century*, mende, adibidez) eta denbora-egiturak konplexuagoak dira. Hiztegi berria erraz txerta daiteke analizatzaileetan, baina egitura konplexuak prozesatzeak oraindik arazoak ekartzen ditu.

3.3.3 Ebaluaziorako galdera corpusak

Galdera-erantzun sistemak probatzeko, denbora-egiturak etiketatuta dituzten corpusak ez ezik, galdera corpusak ere behar dira (Pustejovsky *et al*., 2003b). Hauetan denbora informazioa duten galderak batzen dira eta etiketatze berezi bat sortu behar da espero den erantzun motaren eta erantzunaren aldakortasunaren araberakoa. Galderak eskatzen duen erantzuna denborazko egitura bat bada (24) eta (25) “esplizitu” izendatzen dituzte Pustejovsky *et al*.-ek (2003b); erantzuna beste edozer bada (26) eta (27), ordea, “inplizitu”.

(24) *When was* The Lord of the Rings: the Two Towers *released?* (Noiz plazaratu zen *Eraztunen jauna: Bi dorreak*?)

(25) *How long did Cromwell rule?* (Zenbat urtez agindu zuen Cromwellek?)

(26) *Who was the president of the USA in 1990?* (Nor zen EE BBetako presidente 1990an?)

[2] Automatic Content Extraction.

(27) *How many Tutsis were killed by Hutus in Rwanda in 1994?* (zenbat Tutsi hil zituzten Hutuek Ruandan 1994an)

3.3.4 Beste corpus batzuk: ACE TERN, ARN Chronolines eta TDT-4

Aurreko atalean hiru corpus mota azaldu dira: TIMEBANK, kazetaritza testu laburrez osatutakoa; WikiWars, testu historiko luzez osatutakoa eta Timebank probatzeko galdera corpusa. Horietaz gain, beste corpus batzuk ere prestatu dira. Ondoko taulan hainbat corpusen deskribaketa laburra jaso da. Alde batetik aurreko puntuetan azaldutako corpusen laburpena egin da eta bestetik beste corpus batzuk ere labur deskribatu dira.

1. taula: Denbora-egiturak etiketatuta dituzte corpusak

IZENA	AUTOREAK	ZEREZ OSATUA	MARKATZE-LENGOAIA	ETIKETATZE-ERREMINTA	HIZKUNTZA
TIMEBANK	Pustejovsky *et al*. (2003b)	300 kazetaritza testu	TimeML	Alembic Workbench (bertsio moldatua)	ingelesa, errumaniera, italiera, koreera, frantsesa
WikiWars	Mazur & Dale (2010)	Wikipediako gerrei buruzko 22 artikulu	TIMEX2 (Denbora-adierazpenak bakarrik)	DANTE eta Callisto	ingelesa, alemana
Timebank galdera corpusa	Pustejovsky *et al*. (2003a)	Galderak			ingelesa
ACE TERN	Ferro *et al*. (2005) Ferro *et al*. (2010)	TDT-4 corpuseko kazetaritza testuak (news-wire eta telebistakoak)	TIMEX2 (Denbora-adierazpenak bakarrik)		ingelesez
French ANR project Chronolines	Bittar *et al*. (2012)	50 kazetaritza (news-wire) artikulu	TimeMLn oinarritua	Glozz annotation tool	ingelesa eta frantsesa
TDT-4	Linguistic Data Consortium (2003)	20 iturritatiko kazetaritza testuak (telebistakoak)			ingelesa, arabiera moderno estandarra eta mandarin txinera

Taulan ageri diren TIMEBANK edo WikiWars moduko corpusak etiketatze-eskema bat probatzeko sortu izan dira. ACE TERN edo TDT4 modukoak, ordea, informazioa eraузteko lehiaketetan aurkeztutako tresnak probatzeko.

Handia izan da erdal hizkuntzetan egindako lana denbora-egituren etiketatzearen gainean. Euskaraz, ordea, ez da denboraren araberako etiketatzerik egin oraindik. Erdal hizkuntzetan egin diren lan horiek guztiak kontuan hartuta, euskarazko denbora-egiturak

aztertu, sailkatu eta horiek markatzeko etiketatze-eskema proposatuko da hurrengo atalean.

4 Denboraren adierazpena euskaraz

Hizkuntzek hainbat era dituzte denboraren erreferentzia egiteko. Euskaraz, alde batetik, aditzen denbora-aspektuak ditugu. Aspektuak aditzak markatzen duen ekintza amaituta dagoen ala ez adierazten du eta hiru aspektu motaz hitz egiten da: burutua , ekintza bukatua, perfektua adierazten duena ((28) adibideko *etorri* aditza esaterako); burutugabea, ekintza burutugabea, ez-perfektua ((29) adibideko *etortzen* aditza esaterako) eta geroa, ekintza gertakizun dela adierazten duena ((29) adibideko *etorriko* aditza). Atal honetan, ordea, denbora adierazten duten adberbioak, denborazko perpausak, postposizio-lokuzioak eta orduak eta datak deskribatuko dira, nahitaezkoa baita hauek ezagutzea gero zer etiketatu eta analizatu erabakitzeko.

(28) Atzo Mikel berandu etorri zen, baina gaur garaiz etorri da.

(29) Iñaki beti etortzen da berandu. Uste dugu ez dela inoiz garaiz etorriko.

Aditz formak alde batera utzita, denbora nozioa lexikalizatua duten adberbio, postposizio sintagma eta mendeko perpausak baliatzen ditu euskal hiztunak denborari erreferentzia egiteko. Egiturok denbora esanahia dute eta perpausaren denbora-kokapena eskaintzen dute; eurek seinalatzen dute perpauseko gertakariari dagokion denbora-unea edo tartea. Adierazpide horiek guztiek, formaz ugariak badira ere, *noiz?*, *noiztik?*, *noiz arte?* eta *noizko?* galderei erantzuten diete eta denbora une edo tarte mugatu bat (iraupen bat) irudikatzen dute, ondoko azpiataletan ikusi ahal izango den moduan.

4.1 Adberbioak

Euskaraz adberbioaren funtzioa bi egitura ezberdinek hartzen dute: aditzondoak eta adizlagunak. Biek aditzari laguntzen diote eta aditzaren testuingurua zehazten dute.

4.1.1 Aditzondoak

Aditzondoak berez eta postposizio-atzizkirik gabe aditzari laguntzen dioten aditz sintagmako elementuak dira. Denborazko aditzondoek, adizlagunekin batera,

zirkunstantzia bat adierazten dute, bestela esanda, zirkunstantzialak dira. *Gaur*, *lehen* edo *sarri* modukoak burura badatozkigu ere, aditzondoak formaz anitzak dira (30), (31) eta (32).

(30) Atzo programa berria estreinatu zuten telebistan.

(31) Maiz joaten gara Frantziara ostrak jatera.

(32) Ebakuntzak luze joko du erizainaren arabera.

Aditzondo bakunez gain, zehaztugabe konposatuak daude ((33) adibideko *noizbait* eta (34) adibideko *inoiz*) eta beste alde batetik bestelakoak (*agudo* edo (35) adibideko *dagoeneko*) beti ere Euskaltzaindiak (Altuna *et al*., 1987) proposaturiko sailkapenaren arabera.

(33) Noizbait entzuna nuen Mikelek gaur kontatu duen istorioa.

(34) Inoiz horrelako soineko politik dendan ikusten baduzu, eros iezadazu.

(35) Dagoeneko euritakoa eta neguko arropa atera behar izan dugu.

Gainera, aditzondook egitura handiagoetan ager daitezke eta *bihartik aurrera* edo *noizean behin* moduko esamoldeak (36) sortu. Azkenik, adizlagun itxura duten *gutxitan*, *askotan*, *orduan*, *garaiz* eta gisakoak (37) (38) aipatu behar dira, adizlagun itxura izanik ere, aditzondoak baitira.

(36) Gaurtik aurrera ez duela gehiago erreko zin egin du

(37) Askotan ahazten ditut eguzkitarako betaurrekoak hondartzara noanean.

(38) Lanera sasoiz heltzeko ordubete lehenago esnatu ohi naiz.

4.1.2 Adizlagunak

Aditzondoen funtzio bera betetzen duten postposizio sintagmak dira adizlagunak. Aurrekoek ez bezala, ordea, hauek postposizio-atzizkiak hartzen dituzten izen sintagmen gainean eraikitzen dira. Orokorrean leku-denborazko atzizkiak (inesiboa nagusiki (39), adlatiboa (40), ablatiboa (41) eta leku genitiboa(42)) hartzen dituzte sintagmok, baina batzuetan absolutiboa (43), instrumentala (44) edo soziatiboa (45) ere hartzen dute. Postposizio askeak ere har ditzakete izen sintagma horiek (39) eta (46),

baina hauek hurrengo 4.2 atalean aztertuko dira. Esan gabe doa denbora-adierazpenek hartzen duten atzizkia hartzen dutela beti hartuko dutela bizigabeei dagokien modura (euskaraz bizidun eta bizigabeen arteko bereizketa egiten dela kontuan hartuta).

(39) Goizean irten zen etxetik eta ez zen gauera arte itzuli.

(40) Iluntze aldera heldu ginen etxera.

(41) Bostetatik egon zen Mikelen zain.

(42) Eskatutakoa astelehenerako prest izango duzu.

(43) Denbora luzea ibili zen oihanetik kobazuloa aurkitu zuen arte.

(44) Bi orduz egon naiz zure zain.

(45) Andra Mari eguna ostegunarekin jausten da aurten.

(46) Mikel berandura arte aritu zen ikasten.

Noiz?, *noizko?* eta *noiztik?* baita *zenbat denboran?* edo *zenbatetan?* bezalako galderei ere erantzungo diete leku-denborazko forma hartzen duten sintagmok, eta askotan denbora adierazten duten abiarazle lexiko bat izango dute buru ((39) adibideko *gauera*, (42) adibideko *astelehenerako*, (43) adibideko *denbora*, (44) adibideko *orduz* eta (45) adibideko *ostegunarekin*), baina hau ez da beti gertatzen (47) (48) eta analisi arazoak gerta daitezke. (47) adibidean *bostetan* denborazko adizlaguna da eta ordua adierazten du; (48) adibideko *bostetan* sintagmak, ordea, lekua.

(47) Bostetan geratu ginen kafea hartzeko.

(48) Bost etxe zituen eta bostetan jarri zuen jacuzzia.

4.2 Postposizio-lokuzioak

Postposizio-lokuzioak “adposizio-sintagma baten burua izateko gauza diren unitate fraseologikoak” (Lorente, 2001) direla esan izan da tradizioan. Sintagma bat eta sintagma horri atxikita dagoen postposizio-atzizki batez edo aditzondo batez eta postposizio aske batez osatzen dira. Bigarren elementua aditzondo bat edo postposiziodun izen bat izan daiteke.

Hainbat dira denbora adizlagunak osatzen dituzten postposizio-lokuzioak. Postposizio-atzizkiak ez dira soilik *arte*, *barru*, etab.; beste batzuk ere denborazko egiturak sortzeko balia daitezke. Postposizio-lokuziook sintagma bat eta sintagma horri atxikita dagoen postposizio atzizki batez (50-52) edo aditzondo batez (49) eta postposizio aske batez osatuta daude.

(49) Gaur eta bihar bitartean amaituko dut.

(50) Bi minutu barru ez bada agertzen, joan egingo gara.

(51) Bostak irian gertatu zen istripua.

(52) Eskabidea epez kanpo aurkeztu zuen eta ez zioten diru-laguntza eman.

Ez dira, ordea, egitura guztiz zurrunak. Postposizio-lokuzioetan kontuan izan behar da forma bat baino gehiago izan ditzaketela nahiz eta esanahiari eusten zaion. Horren adierazle ditugu esaterako (53) eta (54) adibideetako *osteguna arte* eta *ostegunera arte* postposizio-lokuzioak:

(53) Osteguna arte egongo gara Bartzelonan.

(54) Ostegunera arte egongo gara Bartzelonan.

Denboraren adierazpenak berezkoak ditu zenbait postposizio-lokuzio. Ondoko taulan aurki daitezke usuenak eta euren formak (Aduriz *et al*., 2008):

2. taula: Denborazko postposizio-lokuzio ohikoenak

Postposizio beregaina	Forma		Adibidea
	Osagarria	Elementu beregaina	
Alde (IZE)	-ABS	-ra	Hirurak aldera bazkalduko dugu
Arte (IZE)	-ABS - Ø	-Ø/-ko - Ø	Ikastaroa zortziak artekoa da Bihar arte ez dago autobusik
Aurre (IZE)	-tik	-ra	Gaurtik aurrera ez du gehiago erreko
Barru (IZE)	-Ø	- Ø	Bi egun barru entregatuko dut lana
Bitarte (IZE)	-ra -ABS	-Ø/-an/-ko -an	Etxera bitartean kontatuko dizut hori Bostak bitartean hemen egongo naiz
Buru (IZE)	-en	-an	Bi egunen buruan jakin zuen emaitza
Gero/ geroztik (ADB)	-z	-Ø/-ko	Istripuaz gero ez du ezer gogoratzen
Inguru (IZE)	-ABS	-an	Zortziak inguruan esnatu gara

4.3 Orduak eta datak

Denbora une zehatz bat adierazten dituzten egiturak ere existitzen dira euskaraz eta garrantzitsua da horiei arreta berezia eskaintzea euren berezitasunagatik. Orduek nahiz datek egitura finkoak dituzte euskaraz eta erraz identifika daitezke. Normalean aurretik aipatutako adizlagun funtzioa betetzen dute ondoko adibideetan ikus daitekeenez:

(55) Bostetan etorri zen menditik.

(56) Gernikako bonbardaketa 1937ko apirilaren 26an gertatu zen.

(57) Gaur zortzi izango da liburuaren aurkezpena

Baina perpauseko beste funtzio batzuk ere har ditzakete, subjektua eta objektuarena behinik behin jarraian ageri denez:

(58) Ordu biak ziren heldu zirenerako.

(59) Gaur hamalau ditu hilak.

Aurreko adibideetan (55) eta (58) agertu dira ordua adierazteko egituretako batzuk. Euskaltzaindiaren 35. arauaren arabera (1995), normalean orduaren zenbakia eta pluraltasun marka erabiltzen da (ordu bata salbu) orduak adierazteko, baina batzuetan errazago identifikatzen laguntzen duten *ordu*/*oren* hitza ere agertzen da mota horretako denbora adierazpenetan. Orduekin batera minutuak esan behar izanez gero, orduak pluralean (ordu bata izan ezik) eta minutuak mugagabean gehi *minutu* hitza (aukerakoa) esango da (60) adibidean ikus daitekeenez. Orokorrean, ordea, *bostak laurden gutxiago* moduko sintagmak topatuko ditugu ordu frakzioak adieraziz. *Eta laurden, eta erdiak* eta *laurden gutxiago* erabiltzen ditu hiztunak ordua adierazteko, *ordu* hitzaren elipsia eginez (*bostak eta (ordu) laurden gutxiago*), baita *eta bost*, *eta hamar*, *bost gutxiago* eta *hamar gutxiago* ere, *minutu* hitzaren elipsia eginez (*bostak hamar (minutu) gutxiago*). Ezin da ahaztu orduak inesibo marka hartuz gero minutuak singularrean emango direla (61).

(60) Hamabiak eta berrogeita bat dira orain.

(61) Autobusa ordu biak eta hamarrean pasatzen da nire etxe azpitik.

Orduak zenbakiz ere adieraz daitezke eta, Euskaltzaindiaren 35. arauak (1995) jasotzen duenez, orduak zein minutuak bi zifra baliatuz adierazi behar dira eta orduen eta minutuen arteko tartea bi puntuez (:). Zenbakiz idatzitako *15:00* edo *12:34* moduko forma horiek, gainera, leku-denborazko postposizio-atzizkiak (inesiboa, ablatiboa, adlatiboa...) eta “arte” bezalako postposizio askeak ere har ditzakete *15:00etatik* edo *21:37an* moduko formak sortuz. Euskaltzaindiak arau berean zifraz idatzitako orduei letraz idatzitakoei ematen zaien tratamendu bera ematea proposatzen du; hau da, orduak pluralean ematea eta minutuak singularrean ((62) eta (63)), idazteko orduan, hiru aukera onartzen baditu ere: *12:34n*, *12:34an* eta *12:34etan*. Taula eta ordutegietan, ordea, zifrak hutsak ematea gomendatzen du.

(62) 15:00etan hasiko da emanaldia.

(63) Autobusa 15:35ean irtengo da eskolatik.

Ordu zehatza adierazteko, adibidez, errepikapenaz balia gaitezke (64):

(64) Zazpi-zazpietan hasi zen ekitaldia.

Beste batzuetan postposizio-lokuzioak osatzen dituzten egiturak aurkituko ditugu orduak adierazteko egituren artean aurreko 4.3 azpiatalean azaldu bezala ((65) eta (66)):

(65) Zazpiak inguru dira.

(66) Zortziak aldera batuko gara Manuren tabernan.

Adjektibo batek lagunduko du ordua *hirurak jota* edo *laurak pasatxo* moduko esamoldeetan (67). Edo *eta piku* zenbatzaile zehaztugabea hartuko du orduak, ordua eta gehiago dela adierazteko (68).

(67) Laurak jota ziren bazkaldu genuenerako.

(68) Zortziak eta pikuan amaitu genuen lana.

Datei dagokienez, datak adierazteko lau elementu erabiltzen dira (Euskaltzaindia, 1995b) eta ondoko egituran ematea hobesten da: lekua (inesiboan), urtea (leku genitiboan), hilabetea (edutezko genitiboan) eta eguna (absolutiboan edo inesiboan) (69); baten bat falta badaiteke ere. Beste formatu batzuk ere har ditzake, ordea ((70) eta (71)):

(69) Bilbon, 2013ko ekainaren 19an.

(70) 2013ko maiatzak 5

(71) 2013/05/05 edo 2013-05-05

(69) adibideko forma hobesten dena bada ere, hiztunek askotan *ekainak 24* edo *2.000 urtean* moduko esamoldeak erabiltzen dituzte eta horiek ere kontuan izan behar izango dira analisia egiterakoan.

4.4 Denborazko mendeko perpausak

Denborazko mendeko perpausetan, mendeko perpauseko gertakaria eta perpaus nagusikoa denbora-erlazio batean jartzea da euren helburua; perpaus nagusiko gertakaria noiz gertatu den edo zein iraupen izan duen esatea orokorrean. Baina denbora erreferentzia perpaus nagusian duten perpaus elkartuak (74) ere existitzen dira askoz ere urriagoak badira ere. Denborazko mendeko perpaus ia guztiek adizlagunek eta

aditzondoek betetzen duten funtzio bera hartzen dute perpaus zirkunstantzialok eta aurrekoetako batengatik ordezka daitezke. Aditz jokatu zein jokatugabea eraman dezakete, baina denek izango dute denbora adierazten duen menderagailuren bat: -(e)n erlatibozkoaren eta leku-denborazko postposizioen gainean eraikitakoak adibidez:

(72) Mikel etorri zenean bazkaltzen hasi ginen.

(73) Ez naiz hemendik mugituko nire ahotsa entzuna den arte.

(74) Bidea trebeskatzera zoan, noiz-eta ere gainerat ethorri baitzaio Brunet hendaiarraren oloa (*Herria* 1958, 1-16, 2).

Forma aldetik ia guztiek hartzen dute erlatibozko -eN menderagailuaren gainean eraikitako menderagailu bat (-eNEAN, -eNERAKO, -eNETIK, ...), baita "-eLA" edo -eLARIK formakoak ere. Urriagoak dira NOIZ ETA ... BAIT- markatzailea hartzen dutenak. Aurreko (73) adibidean agertu denez, -eN ARTE, -Z GERO, -t(z)eaREKIN BATERA moduko menderagailu konplexuak ere erabil daitezke denbora-erlazioak adierazteko menderagailuari denbora adierazten duen postposizio askea gehitzen baitzaio. Bigarrenok menderagailuari postposizio bat gehituz edo konparazioaren gainean eraikitako forma bat baliatuz (76) eraikitzen dira.

(75) Soldata kobratu zuela ibili zen mundutik.

(76) Zuk bazkaria prestatzen duzun bitartean guk mahaia jarriko dugu.

(77) Etorri zen bezain laster mahaian jesarri eta bazkaltzen hasi zen.

Erreferentearekiko erlazioaren arabera hiru multzotan bereiz daitezke denborazko perpausak Euskaltzaindiak (2011) egiten duen sailkapenaren arabera:

3. taula: Denborazko perpausak erreferentearekiko harremanaren arabera

Aldiberekotasuna	Aldi desberdintasuna	Iraupena
	Aurrekotasuna Ondokotasuna	Noizdanikoa Noiz artekoa Tarte osoa

Erlazio-atzizki ezberdinak baliatuko ditu hiztunak mota bakoitzeko perpausak eraikitzeko, baina ez dago sail bakoitzarentzako menderagailu zehatzik – batzuk bi

multzotarako baliagarri izan baitaitezke – eta horrek, hizkuntzaren ezagutzaz ez beste, munduaren eta testuinguruaren ezagutza ere eskatzen ditu (75).

Orokorrean, ordea, erlazio-atzizki bakoitzak denbora-esanahi bakarra izango du. Horren adibide dira aldiberekotasuna adierazten duen (78) adibidekoa. Kasu horretan -eNEAN atzizkiak adierazten du mendeko perpauseko eta perpaus nagusiko ekintzen arteko aldiberekotasuna. Aldi desberdintasuna adierazten dutenen artean, aurrekotasuna *baino lehen* (79) lokuzioak adieraz dezake eta ondokotasuna *eta gero* (80) lokuzioak. Iraupena (noizdanikoa (81) eta noiz artekoa (82) edo tarte osoa (83)) adierazten dutenen artean, iraupen hori noiz hasi den adierazteko -eNETIK atzizkia (81) dago, iraupenaren amaiera adierazteko -eN ARTE lokuzioa (82) eta iraupenaren luzapena adierazteko -Ø BITARTEAN lokuzioa (83) erabil daitezke beste batzuen artetik.

Sailkapen horretaz aparte, denbora-erreferentzia perpaus nagusian duten perpausak (76) eta perpaus nagusiko gertakaria kronologian kokatzen duten perpausen (84) arteko bereizketa ere egin dezakegu.

(78) Omendua etorri zenean hasi zen jaia.

(79) Lehendakariak hitzaldia hasi baino lehen aretoko ateak itxi zituzten.

(80) Ez da inolako eskakizunik onartuko epea amaitu eta gero.

(81) Ezkondu zenetik asko argaldu da Imanol.

(82) Ikaslerik ez zen eskolatik atera eraikin osoa garbi eta ordenatuta egon zen arte.

(83) Denak egon ziren isilean alkateak bandoa irakurri bitartean.

(84) Duela hogeita bost urte heldu zen argi-indarra auzo honetara.

Kontuan hartzekoa da ere, aurretik azaldutako sailkapena (3. taula) ez dela agian bat etorriko errealitatearekin, adibidez errealitatean kontsekutiboak diren bi gertakari aldi berekotzat har baitaitezke hizkuntzaren ikuspegitik (85):

(85) Ohera sartu orduko geratu zen lo.

Gainera, erlazio-atzizkiak soilik ez du denboraren erreferentzia guztiz markatzen; mendeko perpauseko aditzak ere garrantzia du denboraren erreferentzia

interpretatzerakoan. Aditz nagusiaren formak (aditz-erroa, lehenaldiko partizipioa, geroaldiko partizipioa edo gerundioa) ere denboraren erreferentzia iraganean (86), orainaldian (87) edo etorkizunean (88) kokatzen laguntzen du baita ohikotasuna (87) adierazi ere:

(86) Trena gelditu denean jende pila jaitsi da.

(87) Trena gelditzen denean jende pila jaisten da.

(88) Trena geldituko denean jende pila jaitsiko da.

Horietaz gain kontuan izan behar da perpaus nagusiko aditzak ere denbora kokagunea ematen duela kasu batzuetan (89):

(89) Mahaia jartzean hasiko gara bazkaltzen.

Euskarazko denbora adierazpenak zein diren edo nagusiki zein erabiltzen diren aztertzea izan da atal honen helburua. Zein motatako egitura erabiltzen diren adierazteaz gain, egitura horietako bakoitza nola osatzen den eta zein forma har ditzakeen ere azaldu dugu, nahiz eta etorkizuneko lanetan xehetasun handiagoz aztertzea dugun helburu. Hau guztia, esan bezala, baliagarria izango zaigu azterketa automatikoari ekiten diogunean, garrantzitsua baita denbora adierazten duten elementuak identifikatzea, hauek izango baitira analisi automatikoa abiaraziko dutenak eta garrantzitsua baita ere denbora-egituron barne egitura ezagutzea prozesamendu hobetu eta errazteko.

5 TimeML etiketatze-eskema

Lan honen sarreran eta 3. atalean azaldu modura, hainbat lan egin da denbora-egituren analisian eta prozesamenduan XX. mendean zehar, baina, nabari denez, gailendu den joera Ferrok *et al*.-ek (2001) urratutako eta ondoren Setzer (2001) eta Pustejovsky *et al*.-ek (2002, 2003a) jarraitutako bidea izan da. Hain zuzen ere, SGML markaketa-lengoaian (eta aurrerago XML) oinarritutako etiketatze-eskema izan da eredurik jarraituena.

Jarraian, azken urteotan eredu bihurtu den TimeML etiketatze-eskema (Pustejovsky *et al*., 2003a; Saurì *et al*., 2006) eta beste ikertzaile batzuek (Bittar *et al*., 2012) eredu horrentzat proposatutako hobekuntzak azalduko dira. Euskararen kasuan ere eredu horri jarraitzea erabaki dugu euskarazko denbora-egituren analisi automatikoari ekiteko. Erabaki hori hartu dugu erdal hizkuntza gehienetan jarraitu den eredua izateaz gain, eredurik osatuena eta euskararen ezaugarriak kontuan hartuta egingarria delako.

Pustejovsky *et al*.-ek (2003a) TimeML etiketatze-eskema denbora eta gertakarien identifikazioan sortzen diren arazoei erantzuna emateko sortu zuten AQUAINT programaren barruan. Aditzen denbora eta aspektua kudeatzeaz aparte, honako helburu nagusiak zituen programak:

a. Gertakariei denboraren araberako balio bat ematea

b. Gertakariak beste gertakariekiko ordenatzea

c. Testuinguruak ondo definitu ez dituen denbora-adierazpideen gaineko arrazonamendua

d. Gertakarien iraupenaren gaineko arrazonamendua

Pustejovsky *et al*.en (2003a) proposamenean Ferro *et al*.-ek (2001) proposatutako TIDES TIMEX2 etiketatze-eskema eta Setzerrek (2001) aurkeztutakoa oinarritzat hartu eta lehenen eskema zabaldu eta hobetu eta bigarrenaren proposamenean

azaleratutako arazoei aurre egiten zaie TimeML lengoaiaren bidez. Aurreko lan horiekin alderatuta, TimeMLk honako analisi baliabideak eskaintzen ditu:

a. TIMEX2rentzat proposatutako anotazio atributuak zabaltzen ditu.

b. “Denbora funtzioak” gehitzen dizkio berariaz zehaztutako adierazpideak ahalbidetzeko.

c. Denbora-adierazpenen interpretazioa zehazten duten seinaleak identifikatzen ditu.

d. Edozein motatako gertakariak identifikatzen ditu.

e. Denbora eta gertakarien arteko erlazioak sortzen ditu.

Gainera, gertakarien deskribapenaren eta gertakariok batzen dituzten erlazioen arteko bereizketa ere egiten da. Erlazio horiek beste gertaerekiko edota denbora-adierazpenekiko duten harremanaren arabera sailkatzen dira eta etiketatze lanean <LINK> etiketa hartzen dute (7. irudia).

```
<TLINK eventInstanceID="ei1" signalID="s1" relatedToEvent="ei2"
relType="AFTER" magnitude="t1"/>
```

7. irudia: denbora-egituren arteko erlazioaren gauzatzea TimeMLn

Irudi horretan (7. irudia), denborazko lotura baten etiketa, <TLINK>, ikus daiteke. Etiketaren izenaz aparte (letra larriz idatzita: “TLINK”), atributuak ere aurki ditzakegu (letra larri eta xehez idatzitakoak: “eventInstanceID”). Atributuek lotzen diren gertakari edo denborazko egituren informazioa (komatxoen artekoa: “ei1”) ematen dute, hala nola gertakarien identifikatzailea eta euren arteko harreman mota, kasu honetan “AFTER”; hau da, gertakari kontsekutiboak direla.

Jarraian datozen ataletan azalduko dira TimeML osatzen duten etiketak eta etiketa horiek deskribatu ere egingo dira adibideak erabiliz. Azalpenarekin hasi aurretik adierazi nahi dugu etiketa horiek 7. Irudian agertzen den egituraren antzekoa izango dutela, hau da, egitura horretan etiketaren izena, atributuak eta balioak deskribatuko dira.

5.1 Etiketak

Gertakarien arteko erlazioez aparte, diskurtsoko beste elementu batzuk ere etiketatzeko balio du TimeMLk. TimeML etiketaze-eskema osatzen duten etiketen azalpena hurrenkera honi jarraituz egingo da: lehendabizi, gertakari edo ebentoak (5.1.1 eta 5.1.2 atalak) eta denbora-adierazpenak (5.1.3 atala); ondoren, denbora entitateen arteko harremana adierazten duten erlazio-hitzak (5.1.4 atala); eta jarraian, aurretik aipatutako LINK etiketadunak, gertakariak kronologian antolatzen dituztenak (5.1.5 atala). Azkenik beste etiketa batzuk ditugu (5.1.6 atalean).

Etiketa multzo bakoitzak, bere aldetik, atributu multzo bat ere har dezake adierazten duen erlazioa zehatzago definitu ahal izateko. Atributu horiek Setzerren laneko (2001) irizpideei jarraiki sortu dira, baina Setzerrek bezala SGML markaketa-lengoaia erabili ordez, XML erabili dute TimeMLrako.

5.1.1 <EVENT> etiketa

Gertakariak, <EVENT> etiketa hartzen dutenak, Pustejovsky *et al*.-en (2002) esanetan, gertatzen diren egoerak dira; alegia, "*Situations that happen or occur*" edo zerbaiten egoera edo zirkunstantziak adierazten dituzten egiturak. (90) adibidean gertatzen den ekintza bat, *smash* (talka egin) agertzen da. (91) adibidean, ordea, egoera aldakor edo esplizituki denbora erlazioan parte hartzen duen bat *kidnapped* (bahitu) baino ez da etiketatzen.

(90) *The ship <u>smashed</u> into the concrete and glass tower* (Ontziak hormigoi eta kristalezko dorrearen kontra egin zuen talka).

(91) *The girl was <u>kidnapped</u> while she was waiting for the bus* (Neska autobusaren zain zegoen bitartean bahitu zuten).

Gertakariak puntukariak izan daitezke, denborako une zehatz batean gertatzen badira, edo iraupenezkoak, denboran luzatzen den egoera bat adierazten badute. Jarraian datozen adibideetan ikus daiteke, ingelesez aditz jokatu ((92) adibideko *were reported* aditza) edo jokatugabeen ((93) adibideko *working* aditza), nominalizazioen ((94) adibideko *growth* izena), adjektiboen, perpaus predikatiboen edota preposizio-perpausen bidez adierazten dira.

(92) *Three men <u>were reported</u> missing* (hiru gizon galdutzat eman ziren).

(93) *He went home after <u>working</u> for eight hours* (zortzi orduz lan egun ondoren etxera joan zen)

(94) *The young industry's rapid <u>growth</u> also <u>is attracting</u> regulators eager to <u>police</u> its many facets* (industria gaztearen hazkunde azkarrak bere hainbat aspektu kontrolatzeko prest dauden erregulatzaileak erakartzen ari da).

```
The young industry's rapid
<EVENT eid="e1" class="OCCURRENCE">
growth
</EVENT>
also is
<EVENT eid="e2" class="OCCURRENCE">
attracting
</EVENT>
regulators
<EVENT eid="e4" class="I_STATE">
eager
</EVENT>
to
<EVENT eid="e5" class="OCCURRENCE">
police
</EVENT>
its many facets.
```

8. irudia: *The young industry's rapid <u>growth</u> also <u>is attracting</u> regulators eager to <u>police</u> its many facets* perpauseko gertakarien etiketatzea

Aurreko 8. irudian ikus daiteke gertakarien etiketatze adibide bat. Gertakariek <EVENT> etiketa hartzen dute gertakariak direla adierazteko. Etiketa horrek atributu moduan “eid” identifikazio atributua eta “class” mota atributua hartzen ditu, bigarren hau zein motatako gertakaria den adierazteko.

5.1.2 <MAKEINSTANCE> etiketa

<EVENT> etiketekin batera <MAKEINSTANCE> etiketak proposatzen dituzte Pustejovsky *et al*.-ek (2002). Lehen etiketek gertakaria bera markatzen dute eta bigarrenek gertakariaren errealizazioa. Nahitaezko etiketa da eta gertakari guztiek gutxienez <MAKEINSTANCE> etiketa bat hartuko dute.

(95) *John taught on Monday* (Johnek astelehenean irakatsi zuen)

```
<EVENT eid="e1"> taught </EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"/>
```

9. irudia: *John taught on Monday* perpauseko gertakariaren etiketatzea

(96) *John taught twice on Monday* (Johnek astelehenean birritan irakatsi zuen)

```
<EVENT eid="e1"> taught </EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" signalID="s1"/>
<MAKEINSTANCE eiid="ei2" eventID="e1" signalID="s1"/>
```

10. irudia: *John taught twice on Monday* perpauseko gertakariaren etiketatzea

(96) Adibidean ikus daitekeenez *taught* (irakatsi) gertakaria birritan (*twice*) gertatu da eta horregatik gertakari bakarra markatzen bada ere bi <MAKEINSTANCE> agertzen dira gertakaria gertatu den aldi bakoitza adierazteko. Bi <MAKEINSTANCE> sortu ordez, adibide horretan 2 kardinalitate maila duen <MAKEINSTANCE> bakarra sor daiteke. Bigarren aukera askotan gertatzen den gertakariak erraz etiketatzeko erarik praktikoena da kardinalitatea ere atributu izanik.

Aditz modalek (*should*, *might*) eta ezezko partikulek (*not*) gertakaria gertatzea baldintzatzen dutenez, hauek ere <MAKEINSTANCE> etiketan adierazi behar dira (Saurì *et al*. 2006). Aditz modalak <MAKEINSTANCE> etiketa “modality” atributua hartzera behartuko du eta ezezko partikulak “polarity” atributua.

Azaldutako atributuez aparte, kategoria gramatikalari buruzko informazioa eta, aditzen kasuan, aditzaren aspektuari eta denborari buruzko informazioa ematen duten beste atributu batzuk hartzen ditu <MAKEINSTANCE> etiketak.

5.1.3 <TIMEX3> etiketa

Testuan esplizituki agertzen diren denbora-adierazpenak – denborak, datak, iraupenak eta abar – <TIMEX3> etiketaren bidez markatzen dira. Etiketa hori Setzer (2001) eta

Ferro *et al*.-en (2001) proposamenetan oinarrituta sortu da eta nagusiki hiru motatako adierazpenak bereizten ditu Pustejovsky *et al*.-ek (2003a):

a. Guztiz zehaztutako denbora-adierazpenak: 2013ko apirilaren 25a.

b. Gaizki zehaztutako denbora-adierazpenak: astelehenean, bi hilabete barru.

c. Iraupenak: hiru ordu, lau egun.

Egitura horiek dokumentuaren sorrera datarekiko (DCT, Document Creation Time) ISO balio normalizatu bat hartzen dute eta horrela denbora-lerroan finka daitezke. Ondoko (97) adibidean ikus daiteke nola aurreko astea, *last week*, dokumentuaren sorrera dataren astearen aurreko modura ulertzen den:

(97) *last week* = (predecessor (week DCT)) (Aurreko astea = (aurreko (aste DCT))).

ISO balio normalizatuaz gain, beste atributu batzuk ere har ditzake <TIMEX3> etiketak: identifikatzailea eta mota (DATE, TIME, DURATION edo SET), besteak beste. (98) adibidean, denbora-adierazpen bat eta 11. irudian adierazpen horren etiketatzea ikus daiteke zeinetan denbora-adierazpenaren informazioa atributu bidez ordenagailuarentzat irakurgarri bihurtu den.

(98) *This week* (aste honetan).

```
<TIMEX3 id="1" temporalFunction="true" valueFromFunction="tf1"
temporalAnchorID="
unknown1">
This week
</TIMEX3>
<CoerceTo tfid="tf1" argumentID="unknown1" scale="WEEK"/>
```

11. irudia: *This week* sintagmaren etiketatzea

TimeML markaketa-lengoaiaren 1.2.1 bertsioan (Saurì *et al*. 2006), <TIMEX3> "hutsak" sortzeko aukera azaltzen da. Hauen bidez testuan agertzen ez diren, baina inplizituki ulertzen diren denbora-egiturak etiketatu nahi da. Saurì eta besteren arabera, honek testuak iradokitako, baina esplizituki ez adierazitako denborak markatzeko balio du.

5.1.4 <SIGNAL> etiketa

Setzerrek (2001) proposatutako <SIGNAL> etiketa hartzen duten erlazio-hitzak edo esamoldeak etiketatzen dira. Seinale horiek denbora-egituren (gertakariak eta denbora-adierazpenak) arteko harremanak zehazten dituzten hitz ez-funtsezkoak markatzeko erabiltzen dira. Ingelesez hainbat kategoria morfologikoko elementuek bete dezakete denbora-erlazio hitzen funtzioa: denborazko preposizioak izan daitezke (*on*, *during*, *at*, *from*), denborazko lokailuak (*when*, *while*, *before*), baita menderagailuak ere (*if*) eta ezezko esamoldeak (*no*, *not*, *none*), aditz modalak (*might*, *may*) eta karaktere bereziak (/, -).

TimeMLn, ordea, <SIGNAL> etiketa har dezaketen elementuen multzoa murriztu egiten da eta Setzerrek proposatuetatik denborazko preposizioak, denborazko lokailuak eta karaktere bereziak baino ez dira kontuan hartzen. TimeMLn, aitzitik, beste seinale mota bat gehitzen da: modalitatea adierazten duen *to* preposizioa (Saurì *et al*., 2006). <SIGNAL> etiketek eurak identifikatuko dituen atributu bakarra, "sid", hartzen dute, ondoko (99-101) adibideei dagozkien markaketetan (12, 13 eta 14. irudiak) ikus daitekeen modura:

(99) *John might teach on Monday* (Johnek again astelehenean irakatsiko du).

John **<SIGNAL sid="s1"> might </SIGNAL>** teach Monday.

12. irudia: *John might teach on Monday* perpauseko seinalearen etiketatzea

(100) *John taught on Monday* (Johnek astelehenean irakatsi zuen).

John taught **<SIGNAL sid="s1"> on </SIGNAL>** Monday.

13. irudia: *John taught on Monday* perpauseko seinalearen etiketatzea

(101) *All passengers died when the plane crashed into the mountain* (Bidaiari guztiak hil ziren hegazkinak mendiaren kontra talka egin zuenean).

All passengers died **<SIGNAL sid="s1"> when </SIGNAL>** the plane crashed into the mountain.

14. irudia: *All passengers died when the plane crashed into the mountain* perpauseko seinalearen etiketatzea

5.1.5 <LINK> etiketak

TimeMLren eratzearekin batera egin den berrikuntza handienetakoa 5. atal honen sarreran aipatutako <LINK> etiketak dira. Etiketa horien helburuak testu bateko denborazko egituren arteko erlazioak kodetzea eta gertakarien arteko ordena kronologikoa finkatzea dira. Hiru motatako <LINK> etiketa definitzen dituzte Pustejovsky *et al*.-ek (2003a):

a. <TLINK>: bi denbora-elementuren arteko erlazioa irudikatzen duen lotura.

b. <SLINK>: mendekotasun erlazioan dauden bi gertakariren arteko edo gertakari eta instantzia baten arteko erlazioak adierazten dituzten testuinguruak irudikatzeko mendekotasun lotura.

c. <ALINK>: bi gertakariren aspektuen arteko lotura adierazten duen aspektu lotura. <TLINK> eta <SLINK> baten arteko gurutzaketa moduan ere uler daiteke. Aspektuen arteko mendekotasun egoera ere irudikatzen du.

Esan behar da <LINK> etiketen bidez, lehenik, gertakari moten eta gertakari instantzien arteko bereizketa markatu nahi da, eta bigarrenik egitura modalak eta diskurtso ez-zuzena kudeatu nahi dira. Atributu moduan lotzen dituzten entitateak eta euren arteko loturak agertzen dira, loturaren natura ezagutu ahal izateko. (102) adibidean eta 15. irudian *taught* (irakatsi) eta *explosion* (leherketa) gertakarien arteko lotura ikus daiteke. 15. irudiko <TLINK> (denborazko lotura) etiketaren arabera, lehenik leherketa gertatu zen eta gero irakatsi zuen Johnek.

(102) *John taught five minutes after the explosion* (Johnek leherketa baino bost minutu beranduago irakatsi zuen)

```
<TLINK eventInstanceID="ei1" signalID="s1" relatedToEvent="ei2"
relType="AFTER" magnitude="t1"/>
```

15. irudia: *John taught five minutes after the explosion* perpauseko gertakarien arteko denbora-erlazioa

5.1.6 <CONFIDENCE> etiketa

NRRCn (Northeast Regional Research Center) ospatutako mintegian (Pustejovsky *et al*. 2003a), azkenik beste etiketa osagarri bat aurkeztu zen: <CONFINDENCE>. Etiketa horrek denbora-adierazpenen etiketatzearen fidagarritasuna neurtzeko balio du. Etiketa eta atributuen egokitasuna neurtzeko sortu zen testuinguruaren arabera.

5.2 Hobekuntza eta emendakinak TimeML-ri

TimeML denboran zehar hobetu duten etiketatze-eskema da aurreko azpiatalean ikusi ahal izan denez (Pustejovsky *et al*.-en proposamenetik (2003a) Saurì *et al*.-en (2006) 1.2.1 bertsiora). Egileek egindako zuzenketez gain, beste ikertzaile batzuek ere euren iritzia eman dute etiketatze-eskemaren gainean jarraian azaltzen den moduan.

5.2.1 Ehrmann & Hagège (2009)

Ehrmann & Hagègek (2009) TIMEX2n eta TimeMLn definitu diren etiketak ez dituztela denbora-egitura guztiak estaltzen defendatzen dute. Euren esanetan, Ferro *et al*.-en (2001) eta Pustejovsky *et al*.-en (2003a) etiketatze-eskemetan denbora-adierazpen bakoitza mota batekoa (DATE, TIME, DURATION edo SET) izatea baino ez da onartzen. Eta, ondorioz, baldintza horrek kasurik zailenak (103) ezin tratatzea eragiten du. (103) adibideko perpausean oso denbora-adierazpen konplexua dago; "*au moins*" (gutxienez) egiturak asko zailtzen du analisia eta TIMEX2k eta TimeMLk eskaintzen dituzten analisi aukerak ez dira nahikoak.

> (103) *L'entrée éventuelle de la Turquie dans l'UE, c'est au moins dans quinze ans* (Turkiaren EBean sarrera posiblea gutxienez hamabost urte barru izango da).

Horretaz gain, denbora-adierazpenen zatitzeari negatibo irizten diote. TimeML azaltzean azaldu bezala, ez dira denbora-adierazpenak (<TIMEX3> etiketa hartzen dutenak) eta erlazio-hitzak (<SIGNAL> etiketa hartzen dutenak) unitate bakartzat hartzen. Ehrmann & Hagègek (2009), ordea, sintagmaren burutik ezkerretara dagoen guztia hartzen dute unitatetzat informaziorik ez galtzeko eta analisi sintaktikoa errazteko. (104) adibidean sintagmaren buru "*an*" da eta "*pendant plus d'un*" brutik ezkerretara dagoena.

> (104) *Pendant plus d'un an il a travaillé à Carrefour*. (Urte bat baino gehiago lan egin du Carrefourren)

Azkenik etiketatzearen erdibideko analisien erabilgarritasuna zalantzan jartzen dute. Batzuetan TimeMLk beharrezkoa du erdibideko kalkulu hutsezko analisiak egitea denbora-egitura konplexuen analisia egiteko. Euren aburuz ez da kalkuluaren araberako erdibideko etiketatzerik egin behar, baizik eta karakterizazioaren bidezkoak. Adibidez, TimeMLn *two days before yesterday* (atzo baino bi egun lehenago) moduko esapide luzeak zatitu egiten dira erdiko maila batean *two days* (bi egun) iraupentzat hartuz eta *yesterday* (atzo) datatzat. Azken mailan, aitzitik, esapide osoa TIMEX denbora-adierazpentzat hartzen da.

5.2.2 Bittar *et al*. (2012)

Bittar *et al*.-ek (2012) TimeMLren alderdi ahulak azaleratzen dituzte euren etiketatze-eskema aurkezterakoan. Horrela, denbora-egituren azaleko etiketatzea defendatzen duen TimeMLren aurka, Ehrmann & Hagègek (2009) bat eginik, denbora-etiketatze osoak denbora-egitura euren testuinguru osoan interpretatu behar direla uste dute. Eta, bestalde, hizkuntzan oinarritutako analisia hobesten dute.

Hauek lau denbora-adierazpen mota identifikatzea bilatzen dute: iraupena markatzen dutenak (TimeMLn DURATION etiketa hartzen dutenak), denborazko agregatuak (maiztasuna adierazten dutenak eta TimeMLn SET etiketa hartzen dutenak), denbora lokalizatzen duten adierazpenak (*noiz?* galderari erantzuten diotenak) eta gertakariak.

Euren ikuspegitik TimeMLren etiketatzean aldaketak egin behar dira denborazko egituren errepresentazio fidelaren alde. Azaltzen dutenez, iraupen eta datetan esplizituki agertzen da denborazko balio lexikoa duen elementua eta esplizituki ere interpreta daitezke denborazko moduan: *gaur*, *2013ko apirilaren 29an*, *bi ordu*. Denbora adierazten duten beste egitura batzuk, ordea, ez dira hitz bakarrekoak edo hitz kontsekutiboz osatutakoak (*after Michael came*). Horietan, denborazko lokailuak zein gertakariak etiketa bana (<SIGNAL> eta <EVENT> hurrenez hurren) hartu behar izanagatik, etiketa horiek lotzeko <CONNECT> etiketa ere txertatu behar dela defendatzen dute denbora-egituraren interpretazio zuzena bermatzeko. Hala, Pustejovsky *et al*.-en (2003a) proposamenak denbora-adierazpenak lexikoki denbora adierazten duen buru batek gobernatuak izan behar dutela eskatzen du eta <CONNECT> loturaren bidez hau saihesten da.

TimeML euskararako moldatu da lan honetan aurkezten den euskarazko denbora-egituren etiketatzerako. Pustejovsky *et al*. (2002, 2003a) eta Saurì *et al*.-en (2006) ildoa jarraitu da hurrengo atalean ikusi ahal izango denez.

6 Euskarazko denbora-egiturak etiketatzeko proposamena

Etiketatze proposamena egiteko, TimeML etiketatze-eskema hartu da oinarritzat eta bere etiketa eta atributuak euskararako moldatu eta euskararen ezaugarri bereziak adierazteko atributu berriak sortu behar izan dira 6.1 atalean ikusiko den bezala. Etiketatzea gauzatzeko lagin bateko denbora-egiturak identifikatu eta sailkatu egin dira eta ondoren, horiek etiketatzeko etiketa eta atributu sistema sortu izan da.

Etiketatze-lana gauzatzeko EPEC corpuseko (Aduriz *et al.*, 2006) 8133 tokeneko lagina erabili da. EPEC corpusa euskara batuan idatzitako 300.000 hitzez osaturiko testu-bilduma da. Testu horiek XX. mendeko euskararen corpus estatistikotik[3] eta Euskaldunon Egunkariatik[4] hartu dira eta maila desberdineko informazio linguistikoarekin (morfologia eta sintaxia, egina; semantika eta pragmatika, prozesuan) etiketatu dira eskuz eta automatikoki.

EPEC corpusetik hartu den lagineko esaldietan denbora-egituren bilaketa erraztearren, Constraint Grammar (CG) formalismoari (Karlsson *et al.*, 1995) jarraituz lortzen den esaldien analisia erabili da. Adibide modura, 16. irudian ageri da *Hogeita hamabost urtetan Euskaltzaindiraren Erribera kaleko egoitza garbitu zuen* esaldiaren analisia CGn oinarrituta. Formalismoak eskatzen duen formatuari jarraituz, esaldi horretako informazioa lerrotan antolatuta dago. Lerro bakoitzean hitz bakoitzari dagokion analisi morfologiko (kategoria eta azpikategoria gramatikalak) eta analisi sintaktiko partzial (funtzio sintaktikoa) eta osoa (hitzen arteko dependentzia bidezko lotura) adierazten da euren identifikatzaile-kodeak eta guzti.

Lan honen helburua ez da etiketatutako corpus bat sortzea izan, baizik eta lagin horretan ageri diren denbora-egiturak identifikatzea, sailkatzea eta horiek etiketatzeko etiketa- eta atributu-eskema osatzea eta deskribatzea. Horretarako arrazoi nagusia XML etiketak testu jarraituan txertatzeko pentsatuak daudela dira. Erabili izan den laginean

[3] http://www.euskaracorpusa.net/

[4] http://www.egunero.info/(2002an atzitua)

token bakoitza lerro batean agertzen da bere analisia jarraian duela (16. irudia) eta landu den etiketa-sistema ez da egokia. Etorkizuneko lana izango da aztertzea denbora-egiturak markatzeko definitu diren etiketa horiek EPEC corpuseko hitzen analisiari gehituko zaien edota denbora-egiturak etiketatuta baino izango ez dituen beste corpus bat osatuko den. Hartzen den erabakia hartzen dela, euskarazko testuak automatikoki tratatzeko garatu diren tresnak berrerabiltzea bilatuko da.

```
"<Hogeita_hamabost>"  S:1994/0
      <Correct!> "hogeita_hamabost" DET DZH NMGP ZERO mw1,L-
A-DET-DZH-11,lsfi1 @ID> %SIH S:1994 &DETMOD>
"<urtetan>"  S:1432/0
      <Correct!> "urte" IZE ARR BIZ- INE MG w4,L-A-IZE-ARR-
24,lsfi2 @ADLG %SIB S:1432 &NCMOD>
"<Euskaltzaindiaren>"<HAS_MAI>"  S:2022/0
      <Correct!> "Euskaltzaindia" IZE LIB PLU- GEN NUMS MUGM
ENTI_ORG  AORG  HAS_MAI  w5,L-A-IZE-LIB-8,lsfi4  @IZLG>  %SIH
S:2022 &NCMOD>
"<Erribera>"<HAS_MAI>"   S:891/0
      <Correct!>  "Erribera"  IZE  LIB  PLU-  ENTI_LOC  AORG
HAS_MAI w6,L-A-IZE-LIB-10,lsfi5 @KM> S:891 &NCOBJ>
"<kaleko>"   S:2165/0
      <Correct!> "kale" IZE ARR BIZ- GEL NUMS MUGM w7,L-A-
IZE-ARR-31,lsfi7 @IZLG> S:2165 &NCMOD>
"<egoitza>"  S:801/0
      <Correct!> "egoitza" IZE ARR BIZ- ABS NUMS MUGM AORG
w8,L-A-IZE-ARR-35,lsfi8 @OBJ %SIB S:801 &NCOBJ>
"<garbitu>"  S:60/0
      <Correct!> "garbitu" ADI SIN PART BURU NOTDEK w9,L-A-
ADI-SIN-12,lsfi11 @-JADNAG %ADIKATHAS S:60 &ADITZ_NAGUSI
"<zuen>"  S:136/0, 107/0
      "*edun" ADL B1 NOR_NORK NR_HURA NK_HARK w10,L-A-ADL-
10,lsfi12 @+JADLAG %ADIKATBU S:107 &<AUXMOD
      <Correct!>  "*edun"  ADL  ERLT  B1  NOR_NORK  NR_HURA
NK_HARK  w10,L-A-ADL-18,lsfi104  @+JADLAG_MP_IZLG>  %ADIKATBU
S:136 &<AUXMOD
```

16. irudia: *Euskaltzaindiraren Erribera kaleko egoitza garbitu zuen* esaldiaren analisia CGn.

Aurreko atalean azaldu bezala, XML etiketak baliatu dira denbora-egiturak etiketatzeko TimeML etiketatze-eskemari jarraituz. 17. irudian (105) adibideko denbora-adierazpenarentzako <TIMEX3> etiketa eta horren mota (“type”) eta balio (“value”) atributuak agertzen dira. 18. irudian, bere aldetik, (105) adibideko seinalea nola etiketatu den aurkezten da.

(105) Egun osoa

<TIMEX3 type=”DURATION” value=”P24H”> egun osoa </TIMEX3>.

17. irudia: *Egun osoa* sintagmaren etiketatze proposamena.

(106) Menderagailuak baino lehenago

menderagailuak <SIGNAL> baino lehenago </SIGNAL>.

18. irudia: *Menderagailuak baino lehenago* egituraren etiketatze proposamena.

Jarraian euskarazko denbora-egitura batzuentzako proposatzen diren etiketak eta atributuak deskribatuko dira. Egitura bakoitzak zein etiketa eta atributu hartuko dituen azalduko da eta etiketatze adibideak eskainiko dira.

6.1 Proposatutako etiketak

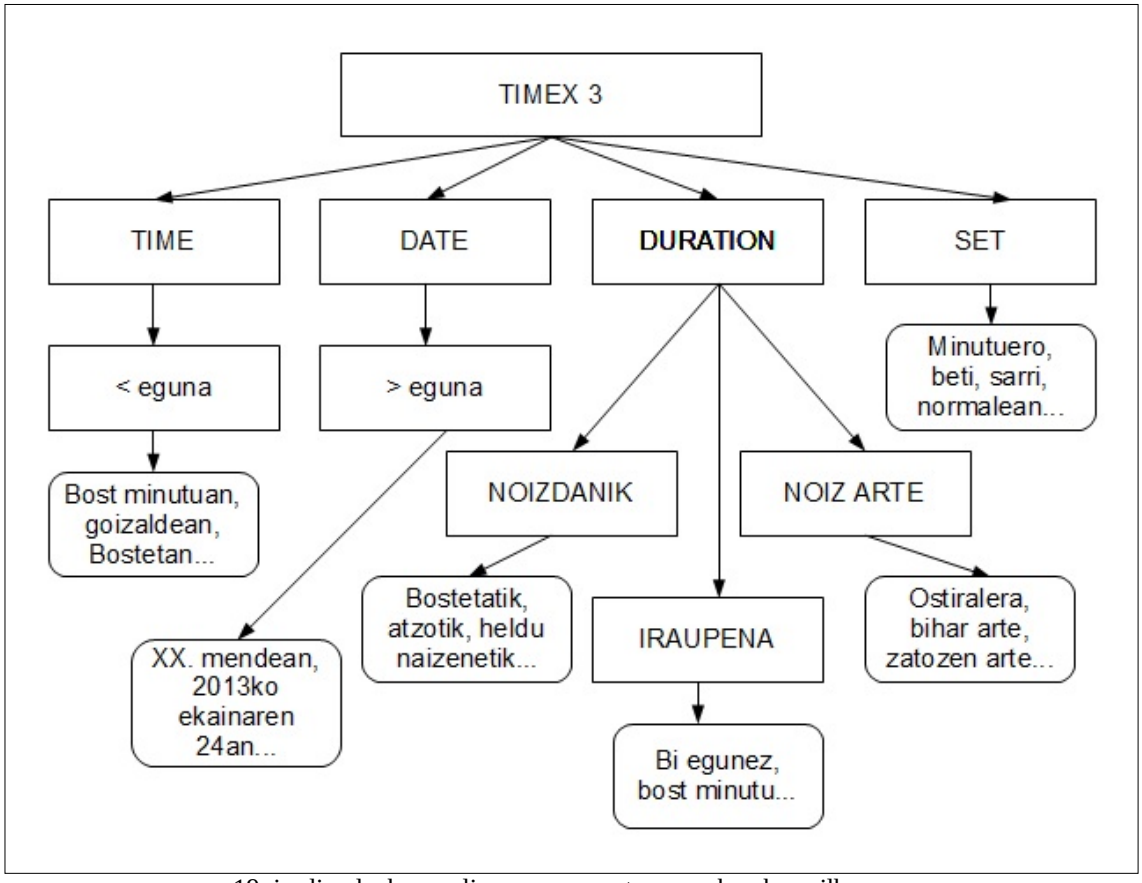
Gurearen tamainako lan batean ezin dira denbora-egiturak osotasunean deskribatuko dituzten etiketak proposatu eta deskribatu. Ondorioz, denbora-egitura batzuetan zentratuko da arreta eta horiek etiketa multzo baten arabera etiketatuko dira. Lan honetan; zehazki, <TIMEX3> etiketa hartzen duten denbora-adierazpen esplizituak eta <SIGNAL> hartzen duten denbora erlaziozko hitzak zein diren eta nola markatuko diren deskribatuko da.

6.1.1 <TIMEX3> hartuko duten egiturak

TimeMLn, 5. atalean azaldu bezala, testuan esplizituki agertzen diren denbora-adierazpenek – denborak, datak, iraupenak eta abar – <TIMEX3> etiketa hartzen dute. Denbora-egitura horiek kronologian koka daitezkeen uneak, denbora-lerroan koka ezin daitezkeenak eta iraupenak dira.

Hori kontuan hartuta, gure proposamena hau da; <TIMEX3> etiketa hartzen duten egiturek bi atributu hartuko dituzte TimeMLko irizpideei jarraiki: balio kronologikoa (“value”) eta mota (“type”). Balio kronologikoa edo iraupenaren luzera (“value”) izango da informazio garrantzitsuena; egiturak kronologiako zein uneri egiten

dion erreferentzia edo zer nolako iraupena adierazten duen baita denbora-adierazpenen daturik nabarmenena. Ondoko lanetan beste ezaugarri batzuk adierazteko atributuak gehituko dira ahalik eta informazio zehatzean eskaini ahal izateko. Mota atributuaren bidez lau motatako denbora-adierazpen identifikatu dira lan honetarako: TIME, eguna baino txikiagoak diren tarteak markatzeko, DATE, eguna baino handiagoak diren tarteak markatzeko, DURATION, iraupena adierazteko, eta SET, maiztasuna adierazteko. Mota bakoitza “type” atributuaren bidez adieraziko da. 19. irudian ikus daiteke motaren araberako sailkapena:



19. irudia: denbora-adierazpenen motaren araberako sailkapena

6.1.1.1 TIME baliodun denbora-adierazpenak

Mota atributuan, TIME balioa eguna baino laburragoak diren denbora-adierazpenak adierazten dituzten egiturei esleitu zaie. Denbora-adierazpen guztiei bezala, balio (“value”) atributu bat ere esleitu zaie honelako adierazpenei euren balioa islatu ahal izateko. Esaterako, 20. irudian ikus daiteke (107) adibideko denbora-adierazpena nola etiketatu izan den.

(107) Goizeko zazpietan

```
<TIMEX3 type=”TIME” value=”T07:00”> goizeko zazpietan </TIMEX3>
```

20. irudia: *Goizeko zazpietan* sintagma etiketatze proposamena

(107) adibideak, denbora-adierazpena izaki, <TIMEX3> etiketa hartu du. Mota atributuan, TIME esleitu zaio eguna baino laburragoa den denbora adierazten baitu eta “value” atributuan denbora-adierazpenaren balio kronologikoa ISO 8601 estandarra (Wolf & Wicksteed, 1997) errespetatuz.

Aurreko (107) adibidean ikus daitekeenez, balioak “T” hartzen du; hau da, eguna baino laburragoa den denbora adierazteko marka. Balioak ordu formatua hartzen badu ere, estandarrak marka hori eskatzen du eta proposamen honetan erabaki hori errespetatu da. Lan honetan orduak adierazteko formatu osoa ondokoa izatea erabaki da: “THH:mm:ss.s” zeinetan orduak (H), minutuak (m), segundoak (ss) eta segundo frakzioak (s) adierazten diren. Bestalde, balioa dokumentuaren sorrera dataren (DCT) araberakoa ere izan daiteke eta kasu horietan erlazio hori islatu behar da. (108) adibideari dagokion 21. irudian “value” atributuari “DCT” balioa eman zaio ezin baitzaio balio kronologiko zehatzik esleitu ez baitago dokumentuaren sorrera datarik. 6.1.3 azpiatalean azaltzen da dokumentuaren sorrera-datari buruz hartu diren erabakiak lan honetarako.

(108) Une honetan

```
<TIMEX3 type=”TIME” value=”DCT”> Une honetan </TIMEX3>
```

21. irudia: *Une honetan* sintagmaren etiketatze proposamena

Denbora-adierazpen guztiek, ordea, ez dute une edo tarte zehatz bat adierazten eta lan honetan balio atributua hutsik uztea erabaki da (109) adibidean gertatzen den modura:

(109) Jateko orduan

```
<TIMEX3 type=”TIME” value=””> jateko orduan </TIMEX3>
```

22. irudia: *Jateko orduan* sintagmaren etiketatze proposamena

Egunaren atalak adierazi behar izana ere gertatu izan da (110) adibidean gertatzen den modura. Kasu hauetan goiza (GO), eguerdia (EG), arratsaldea (AR) eta gaua (GA) bereiztea erabaki da.

(110) Arratsalde edo egun batean

```
<TIMEX3 type=”DURATION” value=”PAR”> arratsalde </TIMEX3> edo
<TIMEX3 type=”DURATION” value=”P1D”> egun batean </TIMEX3>.
```

23. irudia: *Arratsalde edo egun batean* egituraren etiketatze proposamena

23. irudian “value” atributuek “P” hartzen dute iraupenak adierazten dituztelako. Goiko kasuan, eguna baino laburragoak diren denbora tarteak baditugu ere, hauek iraupentzat hartu dira eta horregatik mota atributuan DURATION balioa hartu dute eta balioan “P” iraupenaren marka. 6.1.1.3 atalean azalduko da iraupenak etiketatzeko proposamena.

6.1.1.2 DATE baliodun denbora-adierazpenak

DATE mota-atributua egun bat baino tarte luzeagoa adierazten duten adierazpenei esleitzen zaie. Aurrekoek bezala, honelako egiturek “type” atributuaz aparte, “value” atributua ere hartzen dute balioa adierazteko. Atributu honek aurreko atalean azaldu bezala funtzionatzen du, baina, eguna baino handiagoak diren denborak adierazten direnez, ez da “T” markatzailea erabili behar.

(111) 1995eko azaroaren 18an

```
<TIMEX3 type=”DATE” value=”1995-11-18”> 1995eko azaroaren 18an
</TIMEX3>
```

24. irudia: *1995ko azaroaren 18an* dataren etiketatze proposamena

(112) Asteazkenean

```
<TIMEX3 type=”DATE” value=”XXXX-WXX-3”> Asteazkenean </TIMEX3>
```

25. irudia: *Asteazkenean* sintagmaren etiketatze proposamena

Multzo honetan sartzen diren adierazpenak askoz ere ugariagoak izanik eta islatu beharreko informazioa ezberdina izanik, ez da beti informazio bera aurkituko balio atributuan. (111) adibidearen etiketatzea aurkezten duen 24. irudian, urtea, hilabetea eta egunaz osatutako data bat ikus daiteke eta balioak ere hori bera islatzen du. 25. irudian ((112) adibidearen etiketatzea), ordea, asteko egun baten adierazpena dago, eta urte zehaztugabea, urteko aste (WXX) zehaztugabea eta asteko eguna agertzen dira balio modura (25. irudia). Mendeak ere, beste era batera adieraztea erabaki da, adibidez, mendearen zenbakiaren ostean “C” (century) marka ezarriz (113) adibidean eta 26. irudian ikus daitekeenez:

(113) XVII. eta XVIII. mendeetan

```
<TIMEX3 type=”DATE” value=”17C”> XVII. </TIMEX3> eta <TIMEX3 type=”DATE” value=”18C”> XVIII. mendeetan </TIMEX3>
```

26. irudia: *XVII. eta XVIII. mendeetan* egituraren etiketatze proposamena

Ondorioz, eguna baino handiagoak diren datentzat ondoko formatuak hobetsi dira lan honetan:

- YYYY-MM-DD: urtea (Y), hilabetea (M) eta eguna (D).
- YYYY-UU: urtea (Y) eta urtaroa (U: SP (udaberria), SU (uda), FA (udazkena), WI (negua)).
- YYYY-WXX-X: urtea (Y), urteko astea (WXX) eta asteko eguna (1-7).

- XXC: mendeak.

Milurtekoak eta beste unitate batzuk, euren maiztasun txikia kausa, ez dira lan honetan tratatu eta etorkizuneko lanetan horiei balio estandarra esleitzeko irizpideak zehaztea espero da, baita ordu-eremuak nola adierazi ere. Azkenik, datak eta orduak batera adieraziz gero, dataren eta orduaren artean "T" markatzailea ezarriko da bien arteko bereizketa adierazteko (27. irudia):

```
YYYY-MM-DDThh:mm:ss.s
```

27. irudia: data osoak eta orduak batera emateko formatua.

6.1.1.3 DURATION baliodun denbora-adierazpenak

Iraupenek DURATION hartzen dute mota atribututzat. 23. irudian ageri den bezala, iraupenen multzoan hiru azpi multzo egin daitezke: iraupenaren luzapena, noizdaniko iraupena eta noiz artekoa eta ñabardura horik etiketatze-prozesuan ere azaleratu behar izango dira.

(114) adibideko egituran <TIMEX3> etiketa hartzen duen erabilitako euskarazko denbora-adierazpen markatu baten adibide bat ikus daiteke (28. irudia):

(114) Hogeita hamabost urtetan

```
<TIMEX3 type="DURATION" value="P35Y"> hogeita hamabost urtetan </TIMEX3>
```

28. irudia: *Hogeita hamabost urtetan* egituraren etiketatze proposamena

Adierazgarria da "value" atributuaren balioak 28. irudian hartzen duen formatua; iraupen baten luzera adierazten denez, "P" markatzailea hartzen du lehenik, gero zenbakizko balioa eta azkenik "Y" (year) unitatea (aurreko 6.1.1.2 atalean datekin azaldu modura). Aurreko ataletan bezala, unitate bakoitzak bere hizkia izango du eta horien bidez adieraziko dira iraupenak. Horrela, minutuen kasuan m-ren bitartez

adieraziko da (115) eta (116) adibideko egiturei dagozkien 29. eta 30. irudietan ikusten den bezala.

(115) Bost minutuz

<TIMEX3 type=”DURATION” value=”P5m”> bost minutuz </TIMEX3>

29. irudia: *Bost minutuz* sintagmaren etiketatze proposamena

(116) Yong-koo Jikoo baino hamar minutu zaharragoa da

Yong-koo Jikoo baino **<TIMEX3 type=”DURATION” value=”P10m”> hamar minutu </TIMEX3>** zaharragoa da

30. irudia: *Yong-koo Jikoo baino hamar minutu zaharragoa da* perpauseko denbora-adierazpenaren etiketatze proposamen

Balioei dagokienez, EG balioa sortzea erabaki da *egunez* bezalako egiturak etiketatzeko. Etiketa hori, gainera, 6.1.1.1. atalean azaldutako egunaren beste zatien etiketen multzoan sartzen da. Horrela *egunez*, *goizez* edo *arratsaldez* modukoak etiketatu ahal izango dira

(117) Egunez

<TIMEX3 type=”DURATION” value=”PEG”> egunez </TIMEX3>

31. irudia: *Egunez* egituraren etiketatze proposamena

Aurrera jarraitu baino lehen, gogorarazi nahi dugu etiketatze-eskema hau definitzeko garaian ISO aruak hartu ditugula kontuan. Hala, TIMEX2n (Ferro *et al*., 2001) orduak, minutuak eta segundoak hizki larriz adierazten dira eta hilabete (M) eta minutuen (M) arteko anbiguotasuna sortzen da. Anbiguotasun hori saihesteko, eguna baino txikiagoak diren egiturei “T” markatzailea jartzen zaie (“bost minutu: PT5M”). Saiakera honetan, ordea, ISO 8601 arauari atxikitzea erabaki da esan bezala, eta eguna

baino txikiagoak diren unitateak letra xehez emango dira “T” markatzailerik gabe, anbiguotasunak ez baitu horrela lekurik (“bost minutu: P5m”). TIME eta DATE motetako egiturentzat, ordea, bai gordeko dela “T” markatzailea (6.1.1.1 atala).

Iraupenaren azalpenarekin jarraituz, *1994tik 1999ra* moduko egiturak nola etiketatu ere erabaki behar izan da. Ingelesez *from 1994 to 1999* (118) esapideko hitz bakoitza bere horretan etiketatzen da (Pustejovsky et *al*., 2002) eta ondoko prozesatze urrats batean egitura osoa iraupentzat hartzen da 32. irudian ikus daitekeen modura:

(118) *From 1994 to 1999* (1994tik 1999ra).

```
<SIGNAL sid="s5">from</SIGNAL>
<TIMEX3 tid="t2" type="DATE" value="1994">1994</TIMEX3>
<SIGNAL sid="s6">through</SIGNAL>
<TIMEX3 tid="t3" type="DATE" value="1999">1999</TIMEX3>
```

32. irudia: *From 1994 to 1999* egituraren etiketatzea

Euskaraz, ordea, denbora-adierazpena eta seinalea token berean agertzen dira (1994tik) eta ezin da ingelesezko egiturekin (118) hartu den erabaki bera hartu.

Erromatarren garaian (119) adierazpena ebazteko 33. irudian agertzen den bidea hartu da:

(119) Erromatarren garaian

```
<TIMEX3 type=”DURATION” value=”P10C” beginPoint=”-0027” endPoint=”0476”> erromatarren garaian </TIMEX3>
```

33. irudia: *Erromatarren garaian* egituraren etiketatze proposamena

Aurreko (119) adibideari dagokion 33. irudian ikus daitekeenez, iraupenaren balioaz aparte, hasiera puntua (“beginPoint”) eta amaiera puntua (“endPoint”) adierazten dituzten atributuak ere baliatu dira ahalik eta informazio zehatzena eman ahal izateko. Kasu honetan iraupenaren hasiera eta amaiera ezagutzen dira eta, ondorioz, baita iraupenaren balioa ere. *1994tik 1998ra* egiturarentzat (120) ere antzeko sistema erabil liteke (34. irudian agertzen denaren modukoa), honen gainean lan handia egin

behar bada ere, izatez bi denbora-adierazpenen aurrean baikaude. Hasiera edo amaiera unea ezagun dituzten egiturentzat, aitzitik, azterketa sakonagoa egin behar izango litzateke hasiera eta amaiera une hipotetikoak edo laguntzazkoak finkatzeko.

(120) 1994tik 1998ra

```
<TIMEX3 type=”DURATION” value=”P5Y” beginPoint=”1994” endPoint=”1998”> 1994tik 1998ra </TIMEX3>
```

34. irudia: *1994tik 1998ra* etiketatze proposamena

Iraupenekin ere beste zalantza bat sortzen da iraupena adierazi nahi duten edo kronologiako tarte zabal bati egiten dioten erreferentzia egiten dioten egiturekin. Esaterako, *Erromatarren garaian* moduko esapidea *Erdi Aroan* moduko batekin agertuz gero, argi egongo litzateke iraupena baino, momentuari egiten zaiola erreferentzia. Baina “Erromatarren garaian hazkunde demografiko handia izan zen” moduko perpausean baliteke iraupenari erreferentzia egin nahi izatea. Gauza bera gertatuko litzateke “arratsalde batean egiteko lana da hori” esapidearekin, zeinetan zaila den unea edo luzapena azpimarratu nahi den jakitea.

6.1.1.4 SET baliodun denbora-adierazpenak

SET etiketaren bitartez maiztasuna adierazten duten denbora-egiturak etiketatuko dira. Errepikapenak maiztasuntzat hartu dira lan honetan, eta errepikapena adierazten duten egiturak (121) adibideko esaldiari dagokion 35. irudian bezala etiketatu dira

(121) Behin eta berriz

```
<TIMEX3 type=”SET” value=””> behin eta berriz </TIMEX3>
```

35. irudia: *Behin eta berriz* egituraren etiketatze proposamena

Arazoak izan dira, ordea, euskaraz hain ugariak diren horrelako esapideei (*beti*, *maiz*, *behin eta berriz*) zein balio esleitu erabakitzeko, egitura horiei balio zehatz bat emateko zailtasuna dela kausa, eta balio atributua hutsik uztea erabaki da esperimentu honetan. 35. irudian “value” atributua hutsik utzi da ezin izan baitzaio baliorik eman.

Maiztasun kontuetan, askotan denbora-tarte batean zenbatetan egin edo gertatu den zerbait adierazten da ondoko (122) eta (123) adibideetan eta adibide horiei dagozkien 36 eta 37. irudietan ikus daitekeen modura:

(122) Egunero

```
<TIMEX3 type=”SET” value=”1D” quant=”EVERY”> egunero </TIMEX3>
```

36. irudia: *Egunero* egituraren etiketatze proposamena

(123) Egunean bost aldiz

```
<TIMEX3 type=”SET” value=”P1D” freq=”5X”> egunean bost aldiz </TIMEX3>
```

37. irudia: *Egunean bost aldiz* egituraren etiketatze proposamena

Horrelako esapideetatik, (122) eta (123), ahalik eta informazio gehien adierazteko, TimeML (Pustejovsky *et al*., 2003a) jarraituz hirugarren atributu bat gehitzea erabaki da. Horrela, 36. irudian kantitate atributua, “quant”, gehitu da errepikapenaren kantitatea adierazteko. 37. irudian, aitzitik, denbora-tarte batean gauza bat zenbat aldiz gertatu den adierazi nahi izan da eta horretarako balioari iraupen balioa

eman zaio eta maiztasuna, “freq”, adierazi da ondoren. Horrela, egun bateko tartean bost aldiz gertatu izan dela gertakaria adierazi ahal izan da.

6.1.2 <SIGNAL> hartuko duten egiturak

<SIGNAL> etiketa denbora-erlazio bat adierazten duten egiturek hartzen dute eta gertakariak eta denbora-adierazpenak euren artean lotzeko balio du. Ingelesez hainbat kategoria morfologikoko elementuk bete dezake denbora erlazio-hitzen funtzioa: denborazko preposizioak (*on*, *during*, *at*, *from*), denborazko lokailuak (*when*, *while*, *before*), baita menderagailuak ere (*if*) eta ezezko esamoldeak (*no*, *not*, *none*), aditz modalak (*might*, *may*) eta karaktere bereziak (/, -).

Lan honetan, ingelesezko denborazko preposizioen eta denborazko lokailuen baliokideak baino ez dira hartu seinaletzat. Denbora-egituren analisi sakona egingo denean ikertuko dira Pustejovsky *et al*.-ek (2003a) eta Saurì *et al*.-ek (2006) seinaletzat hartzen dituzten beste egiturak.

Euskarazko denborazko lokailuak nahiko erraz identifikatu eta etiketatu dira lagin honetan multzo murriztua baita eta hitz askez osatuak baitaude orokorrean. Gainera, hurbilpen honetan ez dute inolako atributurik hartu, TimeMLn proposatzen den “sid” atributu identifikatzailea baino ez baita landu, (124) eta (125) adibideei dagozkien 38. eta 39. irudietan agertzen denez.

(124) Ondoren

```
<SIGNAL> ondoren </SIGNAL>
```

38. irudia: *Ondoren* egituraren etiketatze proposamena

(125) Berehala

```
<SIGNAL> berehala </SIGNAL>
```

39. irudia: *Berehala* egituraren etiketatze proposamena

Seinale guztiak, ordea, ez dira hain erraz etiketatu. Euskaraz, hizkuntza eranskaria eta buru-azkena izanik, ez dago denbora adierazten duen preposiziorik; izen multzoari itsatsirik dauden postposizio-atzizkiak erabiltzen dira. Horrek hainbat kasutan token batek bi etiketa hartu behar izana ekarri du "1994tik" kasuan (120) azaldu bezala. Egituren tratamendua errazteko asmotan gramatikari hertsiki jarraituz etiketatzeari uko egin behar izana ere ekarri du seinalea aditz laguntzaileari atxikita agertzen den kasuetan, (126) adibidearen etiketatzean ikus daitekeenez (40. irudia).

(126) Etortzen zarenerako bazkaria prest egongo da.

<EVENT> Etortzen </EVENT> <SIGNAL> zarenerako </SIGNAL> bazkaria prest egongo da.

40. irudia: *Etortzen zarenerako bazkaria prest egongo da* perpausaren etiketatze proposamena

Hori ikusita, token batek bi etiketa hartzen dituenean (127), seinalea baino (kasu honetan inesiboaren bidez emanda datorrena) denbora-adierazpena etiketatzea lehenetsi da, honek ematen baitu informazio gehien. Seinalea ez markatzeak uzten duen hutsunea etorkizuneko lanetan <LINK> etiketen bidez osatzea aurreikusten da, denbora-adierazpena eta gertakaria erlazionatuko direnean. Oraingoz, 41. irudian hartu den irtenbidea proposatzen da:

(127) Gehienak ostiraletan joaten dira meskitara.

<SIGNAL> On </SIGNAL> <TIMEX3> Fridays </TIMEX3> <TIMEX3 type="SET" value="XXXX-WXX-5" quant="EVERY"> ostiraletan </TIMEX3>

41. irudia: *Gehienak ostiraletan joaten dira meskitara* perpauseko denbora-adierazpenaren etiketatze proposamena

Lan honetan, bestalde, postposizio-lokuzioak ahalik eta fidelen adierazteko, beste atributu bat sortzea proposatu da: "postAtzizki". Atributu berri hori 3. taulan aurkeztu diren postposizio-lokuzioak etiketatzeko sortu da. Horrelako egituretan sintagma buruak <TIMEX3> edo <EVENT> etiketa hartuko du eta postposizio askeak,

<SIGNAL>. Baina lokuzioaren parte den postposizio-atzizkia adierazteko, <SIGNAL> etiketari atzizki horren informazioa gehituko zaio "postAtzizki" atributuaren bidez ondoko adibidean ikus daitekeenez:

(128) Berotzen ari den bitartean

<EVENT> berotzen ari den </EVENT> <SIGNAL[5] postAtzizki="en"> bitartean </SIGNAL>

42. irudia: *Berotzen ari den bitartean* egituraren etiketatze proposamena

(129) Ordura arte

<TIMEX3 type="TIME" value=""> ordura </TIMEX3> <SIGNAL postAtzizki="ra"> arte </SIGNAL>

43. irudia: *Ordura arte* egituraren etiketatze proposamena

42 eta 43. irudietan postposizio-lokuzioko elementu askeari (*bitartean* (128) eta *arte* (129)) ezarri zaio seinale-etiketa. Baina arrunta da euskaraz gertakaria eta seinalea token berean adieraztea: *diogunean* (130), *egiterakoan* (131). Kasu horiek, nahiz eta identifikatu diren, ez etiketatzea erabaki da ikerketaren lehen urrats honetan, gertakariak tratatu gabe utzi baitira eta azterketa zehatzagoa behar baita. Aditz perifrastikoz adierazten diren gertakarietan (132), ordea, aditz nagusiak informazio lexikoa hartzen du eta laguntzaileak informazio gramatikala eta denbora-markatzailea. Kasu horietan aditz nagusiari <EVENT> etiketa eta aditz laguntzaileari <SIGNAL> etiketa esleitzea deliberatu da (44. irudia).

(130) hemen sartzen dugun adimendua terminoa diogunean.

(131) Ebbl-a egiterakoan.

[5] Lan honetan ez da "sid" atributu identifikatzailea kontuan hartu.

(132) Artikulu hau argitaratzen denerako

```
Artikulu hau <EVENT> argitaratzen </EVENT> <SIGNAL> denerako
</SIGNAL>
```

44. irudia: *Artikulu hau argitaratzen denerako* perpausaren etiketatze proposamena

6.1.3 Dokumentuaren sorrera data

Erabilitako laginean ez da dokumentuaren sorrera data (DCT) agertzen eta horrek hein handi batean etiketatzea baldintzatu du denbora-adierazpen askoren denbora-erreferentzia horren araberakoak baita. Sorrera data balego, "tid=t0" atributua hartuko luke eta testuko denbora-egiturentzat erreferente litzateke ondoko (133) adibidearen etiketatzea erakusten duen 45. irudian ikus daitekeenez:

(133) *Two weeks from next Tuesday* (Datorren asteartetik bi aste).

```
<TIMEX3 tid="t1" type="DATE" value="2002-07-23"
anchorTimeID="t0" temporalFunction="true" valueFromFunction="tf1">
two weeks from next Tuesday </TIMEX3>
```

45. irudia: *Two weeks from next Tuesday* egituraren etiketatzea

45. irudian agertzen den modura, testuaren sorrera data ezagutzen denean, testuko denbora-egiturak bere balio estandarizatua hartzen du eta sorrera data "anchorTimeID" atributuaren bidez adierazten da. Lan honetarako, ordea, ez dira aingurak baliatu eta 46. irudian agertzen den modura ebatzi dira balioa eta dokumentuaren sorrera data berbera duten denbora-adierazpenen etiketatzea. Lan honetan, dokumentuaren sorrera data ezezaguna izaki, "DCT" balio generikoa esleitu zaie baliotzat dokumentuaren sorrera data hartu behar izango luketen egiturei (134):

(134) Gaur

```
<TIMEX3 type=”DATE” value=”DCT”> gaur </TIMEX3>
```

46. irudia: *Gaur* adberbioaren etiketatze proposamena

6.2 Emaitzak

Etiketatze-eskema deskribatzeaz batera, EPEC corpuseko laginean ageri diren denbora-egitura ezberdinak identifikatu eta sailkatu dira. Esku artean erabili dugun laginean agertu diren denbora-egituren eta moten berri jaso da 4. taulan:

4. taula: Aztertutako denbora-egiturak

Denbora-egitura mota	Kopurua	Ehunekoa
TIMEX		
Time	17	%9,04
Date	43	%22,87
Duration	18	%9,57
Set	31	%16,49
SIGNAL	30	%15,96
EVENT + SIGNAL	45	%23,94
TIMEX + SIGNAL	4	%2,13
Guztira	188	%100

Taula horretan, laginean zein motatako denbora-egiturak azaldu diren esateaz gain, bakoitza zenbateko maiztasunarekin ageri den ere adierazi da. Adibidez, eguna baino luzera handiagoko denbora-tarteak adierazten dituzten egiturak eguna baino laburragoak adierazten dituztenak baino gehiago dira. Seinaleak, <SIGNAL>, ere nahiko maiz agertzen dira. Hori multzo horretan testuetan oso maiz agertzen diren denbora lokailuak (*lehen*, *gero*, *ondoren*...) batzen direlako da. Gertakari eta seinalez osatutako denbora-egituren multzoa ere ugaria da; talde horretan batzen baitira *bazkaldu ostean* edo *bazkaltzen ari den bitartean* modukoak.

Ez da beti erraza izan egitura bat multzo batean sailkatzea batzuetan euren arteko mugak ez baitira garbiak. Esaterako, *Erromatarren garaian* (119) iraupenen multzoan sartu da lan honetan, baina *gure haurtzaroan* edo *handitan* modukoak denbora-adierazpenen multzoan. Argi dago batzuek zein besteek iraupen bat adierazten dutela,

baina agian ez da iraupen hori testuinguru zehatz horretan ezaugarririk behinena. 6.1.1.3 azpiatalean dago *Erromatarren garaian* egiturari iraupen modura eman zaion analisia (33. irudia), baina daten multzoan ere sar litekeela kontuan hartu behar da. Antzera gertatzen da *arratsalde edo egun batean* egitura konposatuarekin, *arratsalde* edo *egun* hitzek beti adierazten baitute iraupena. Zaila da esaten, ordea, *arratsalde edo egun batean* esaten dugunean iraupenari ematen diogun garrantzia.

Bestalde, *zein momentutan* egitura denbora-adierazpenen artean TIME motako atributua hartzen dutenen multzoan sartu da, jakinik zehar-galdera baten testuinguruan agertzen dela. Horrelako egiturentzat aparteko sail bat sortzea aurreikusi da, ez baitute egiatan adierazten ez kronologian koka daitekeen unerik ez iraupenik.

Aurretik azaldutako zalantzazko egitura horietaz gain, beste hainbat egitura, argi sailkatuta egon arren, ezin izan dira osorik etiketatu. Askotan gertatu izan da atributuren bat ezin bete izana informazio falta edo irizpide finkatuak ez izatea kausa. Adibidez, ez da zehaztu zein balio kronologiko eman *jateko orduan* (109) moduko egiturei. TIME eta DATE motetako egiturek ez badute “value” balio atributurik jaso, horietako askotan denbora-adierazpenaren balioa zehaztugabea izan delako da (135) eta (136) dagozkien 47 eta 48. irudietan agertzen den modura.

(135) Hilabete berean

```
<TIMEX3 type=”DATE” value=””> hilabete berean </TIMEX3>
```

47. irudia: *Hilabete berean* egituraren etiketatze proposamena

(136) Gau hartan

```
<TIMEX3 type=”TIME” value=””> gau hartan </TIMEX3>
```

48. irudia: *Gau hartan* egituraren etiketatze proposamena

Amaitzeko esan behar da lehen urrats honetako azterketa denbora-adierazpenak, seinaleak eta denbora-adierazpen eta seinale edo gertakari eta seinale egiturak aztertzera

mugatu dela eta etorkizuneko lanean bestelako egiturak aztertzea eta definitzea dugula helburu, egitura horiek identifikatzeak testuen tratamendu automatikoan beste urrats bat direlako. Horretarako, esan bezala, denbora-egitura posible guztiak identifikatu eta sailkatuko badira, dokumentuaren sorrera datak eskuragarri egon behar izango du behinik behin, baina baita atributu multzo zabalagoak ere.

7 Ondorioak eta etorkizuneko lanak

Testu bat bere osotasunean ulertzeko, nahitaezkoa da hizkuntza naturaleko testuetako denborari buruzko informazioa aztertzea eta prozesatzea. Azken urteotan, garrantzia handia hartu du ikerketa-lerro horrek Hizkuntzaren Prozesamenduan (HP), testuetako denbora-egiturak identifikatu eta esplizitu egiteak testuen ulermen automatikoan duen garrantziarengatik.

Ikerketa-lan honetan erdal hizkuntzetan egin diren denbora-egiturei buruzko lanak aztertu ondoren, euskarazko denbora-egiturak identifikatzeko eta horiek etiketatzeko osatuko den etiketatze-eskemaren lehen urratsak egin dira.

Erdal hizkuntzetako ikerketa-lan horien helburuetako bat testuetan linguistikoki gauzatu diren denbora-egiturak esplizitu egitea izan da, horretarako denborari buruzko informazioa adierazten duten egitura linguistiko guztiak etiketatuz. Etiketatze-lan hori egiteko, eredu bihurtu den TimeML etiketatze-eskema (Pustejovsky *et al.*, 2002) jarraitu dute hizkuntza gehienetan, guztietan ez esatearren. Hori horrela izanik, euskararen kasuan ere eredu bera jarraitzea erabaki da, batetik, denbora-egiturak etiketatzeko eskema estandarra bihurtu delako TimeML eta, bestetik, euskarazko denbora-egiturak etiketatzeko ere baliagarriak direlako bertan deskribatutako etiketak, nahiz eta zenbaitetan etiketako horietako batzuk moldatu edo berriak deskribatu diren euskararen ezaugarriak direla eta. Ondorioz, TimeML ereduaren bidezko lehen etiketatze-eskemaren proposamena egin da euskararako.

Euskal testuetako denbora-egiturak deskribatzeko, EPEC corpuseko lagin bat erabili da. Kontuan izan behar da lagin horretan ez direla euskarazko denbora-egitura guztiak ageri; beraz, etiketatze-eskeman definitu diren etiketak lagin horretan aurkitu diren egiturei bideratuta daude. Guztira, 188 egitura desberdin aztertu dira. Halaber, lanean zehar adierazi den bezala, lehen hurbilpen honetan ez dira denbora-egitura horietako batzuk osorik etiketatu. Horren arrazoia da egitura horiek adierazten duten denbora esplizitu egiteko erabiltzen diren etiketez gain, beste etiketa batzuk ere behar direla atributu osagarriak markatzeko, eta lehen hurbilpen honetan lana mugatzeko helburuarekin adierazgarrienak direnak soilik etiketatzea erabaki da, alegia, balio eta

mota atributuak, kasu batzuetan beste atributu batzuk proposatzeko aukera izan bada ere.

Etorkizunari begira, lehenik eta behin bibliografia aztertzen jarraitu behar da, ikerketa-gai hau garrantzi handikoa delako une hauetan eta lan berriak argitaratzen ari direlako.

Euskarari dagokionez, denbora-egituren azterketa osoagoa egin behar da; alegia, denbora adierazteko euskaraz erabiltzen diren egiturak zein diren eta tamaina handiagoko lagin batean edo corpusean islatzen diren aztertuko da. Azterketa horrek egitura horiek guztiak edo horietako batzuk tratatuko diren erabakitzeko aukera emango digu eta bide batez nola sailkatu ere erakutsiko digu.

Aztertutako denbora-egitura horien markaketarako etiketatze-eskema osatuko da ikerketa-lan honetan egin den proposamena abiapuntu hartuta. Behin etiketatze-eskema definituta, denbora-egiturak eskuz etiketatuta dituen corpusa osatuko da. Ezin ahantz daiteke mota horretako corpusek duten garrantzia HPn, denborari buruzko informazio horrek gertakariak denboraren arabera ordenatzen laguntzen baitu eta hori ezagutzea beharrezkoa delako informazioa erauzteko sistemetan, itzulpen automatikoan edota testuen laburpen automatikoan besteak beste.

Amaitzeko, corpusaren osaketarekin batera etiketatze-prozesuaren ebaluazioa ere egingo da etiketatzaile desberdinen lana irizpide matematikoen bitartez konparatuz.

8 Bibliografia

Aduriz I., Aldezabal I., Aranzabe M. J., Arriola J. M., Ceberio K., Estarrona A., Iruskieta M., Lersundi M., Pociello E., Uria L., Urizar R., Aldasoro, E. 2008. *Euskarazko postposizio-lokuzioen tratamendu konputazionala*. Barne-txostena (UPV / EHU LSI / TR 07-2008). Lengoaia eta Sistema Informatikoak Saila, UPV/EHU

Aduriz I., M.J. Aranzabe, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa and R. Urizar. 2006. 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. and Archer D. editors, *Corpus Linguistics Around the World*, 1-15. Rodopi (Netherland).

Altuna, P., P. Salaburu, P. Goenaga, M. P. Lasarte, L. Akesolo, M. Azkarate, P. Charriton, A. Eguskitza, J. Haritschelhar, A. King, J. M. Larrarte, J. A. Mujika, B. Oyharçabal and K. Rotaetxe. 1987 (1997 berrinprimaldia). *Euskal Gramatika Lehen Urratsak II*. Euskaltzaindia, Bilbo.

Bittar A. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD. Thesis, Université Paris Diderot, Paris.

Bittar, A., Amsili, P., Denis, P., and Danlos, L. (2011). French TimeBank: An ISO-TimeML Annotated Reference Corpus. In Proceedings of the 49th Annual Meeting of ACL, Portland, Oregon, pp. 130--134.

Bittar, A., C. Hagège, V. Moriceau, X. Tannier and C. Teissèdre. 2012. Temporal Annotation: A Proposal for Guidelines and a Experiment with Inter-Annotator Agreement. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis (Eds.), *Proceedings of the Eight*

International Conference on Language Resources and Evaluation (LREC'12), pp. 3741-3745, Istanbul, Turkey.

Boroditsky, L. 2011. How Languages Construct Time. In Dehaene and Brannon (Eds.,) *Space, time and number in the brain: Searching for the foundations of mathematical thought*. Elsevier.

Caselli, T., Bartalesi Lenzi, V., Sprugnoli, R., Pianta, E. and Prodanof, I., 2011. Annotating events, temporal expressions and relations in italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pp 143–151, Association for Computational Linguistics, Portland, Oregon, USA

Day, D., J. Aberdeen, L. Hirschman, R. Kozyerok, P. Robinson, M. Vilain, 1997. "Mixed-Initiative Development of Language Processing Systems". *Proc. of the 5th conference on Applied NLP*, pp 348-355.

Day, D., C. McHenry, R. Kozierok and L. Riek. 2004. Callisto: A configurable annotation workbench. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pp. 2073-2076, Lisboa, Portugal.

Ehrmann, M. and C. Hagège. 2009. Proposition de caractérisation et de typage des expressions temporelles en contexte. In *Actes de TALN 2009*, Senlis, France.

Euskaltzaindia. 1995a. *35. araua: Orduak nola esan*. Bilbo, 1995-06-30.

Euskaltzaindia. 1995b. *37. araua: Data nola adierazi*. Donostia, 1995-07-28.

Evans, V. 2007. How we conceptualise time: language, meaning and temporal coginition. In: Evans, V., B. K. Bergen and J. Zinken (Eds.), *The Cognitive Linguistics Reader*. Ch. 22. Equinox Publishers, Sheffield.

Ferro, L., I. Mani, B. Sundheim and G. Wilson. 2001. TIDES Temporal Annotation Guidelines, Version 1.0.2. MITRE Technological Report, MTR 01W0000041, McLean.

Ferro, L., L. Gerber, J. Hitzeman, E. Lima and B. Sundheim, 2005. "ACE Time Normalization (TERN) 2004 English Training Data v1.0". Accessed: 2013ko

uztailaren 3an http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T07

Ferro, L., L. Gerber, I. Mani, B. Sundheim and G. Wilson. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions*, September

Ferro, L., L. Gerber, I. Mani, B. Sundheim and G. Wilson, 2010. ACE Time Normalization (TERN) 2004 English Evaluation Data v1.0. Accessed: 2013ko uztailaren 3an http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T18

Forascu, C. and D. Tufis. 2012. Romanian TimeBank: An Annotated Parallel Corpufor Temporal Information, In Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Ugur Dogan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 3762-3766, Istanbul, Turkey.

Im Seohyun, Hyunjo You, Hayun Jang, Seungho Nam, and Hyopil Shin. 2009. KTimeML: Specification of Temporal and Event Expressions in Korean Text. In: Proceedings of the 7th workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009, Suntec City, Singapore.

Karlsson F., A. Voutilainen, J. Heikkilä and A. Anttila. 1995. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Lascarides, A. and N. Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment, *Linguistics and Philosophy*, 16 (5), pp. 437-493, Kluwer Academic Publishers, Dordrecht, Holland

Lingusitic Data Consortium, 2005. TDT4 Multilingual Broadcast News Speech, Text and Annotations. Accessed: 2013ko uztailaren 3an http://ssli.ee.washington.edu/people/leixin/TDT4.html

Lorente, M. 2001. *Gramàtica del Català Contemporani*, chapter Altres elements lèxics, 831-888. Empúries, Barcelona.

Mazur, P., and R. Dale. 2007. The DANTE Temporal Expression Tagger. In Zygmunt Vetulani (Ed.), In *Proceedings of the 3rd Language And Technology Conference* (LTC), Poznan, Poland.

Mazur, P. and R. Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 913-922. Association for Computational Linguistics, Stroudsbourg, USA.

Moens, M. and M. Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics – Special issue on tense and aspect*, vol. 14, issue 2, pp. 15-28, MIT Press, Cambridge, USA.

Pustejovsky, J., L. Belanger , J. Castaño , R. Gaizauskas , B. Ingria , G. Katz , D. Radev, A. Rumshisky , R. Saurí , A. Setzer , B. Sundheim and M. Verhagen, 2002. NRRC Summer Workshop on Temporal and Event Recognition for Question Answering Systems, Version: 0.1.0 Release Date: 29-09-02

Pustejovsky, J., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer and G. Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *In Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*, pp. 1-11, Tilburg.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo. 2003b. The TimeBank Corpus. In *Corpus Linguistics*, pp. 647-656, Lancaster, UK.

Reichenbach H. 1947. The tenses of verbs. In *Elements of Symbolic Logic*, pp. 287-298. The Macmillan Company, New York.

Salaburu, P., J. M. Makatzaga, I. Amundarain, M. Azkarate, P. Charriton, B. Fernandez, J. Garzia, P. Goenaga, A. King, I. Laka, M. P. Lasarte, C. Mounole, J. A. Muxika, B. Oyharçabal, P. Rekalde and K. Rotaetxe, 2011. *Euskal Gramatika Lehen Urratsak VII*: *(Perpaus jokatugabeak: denborazkoak, kausazkoak eta helburuzkoak, baldintzazkoak, kontzesiozkoak, moduzkoak, erlatiboak eta osagarriak)*, Euskaltzaindia, Gramatika Batzordea, Bilbo.

Saurí R., J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky, 2006. TimeML Annotation Guidelines Version 1.2.1.

Setzer, A. 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. PhD. Thesis, University of Sheffield.

Setzer, A. and R. Gaizauskas. 2000. Annotating Events and Temporal Information in Newswire Texts. In *Proceedings of the Second International Conference on Language Resources And Evaluation*, pp. 1287–1294, Athens.

Spreyer, K. and A. Frank. 2008. Projection-based Acquisition of a Temporal Labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pp 489-496, Hyderabad, India.

Wolf, M. and C. Wicksteed. 1997. Date and Time Formants. Accesed: 2013ko uztailaren 3an http://www.w3.org/TR/NOTE-datetime