Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTimeren bidez

Basque Temporal Information Processing Using TimeML and HeidelTime

Begoña Altuna, M.ª Jesús Aranzabe, Arantza Díaz de Ilarraza*

Euskal Herriko Unibertsitatea

* begona.altuna@ehu.eus

DOI: 10.1387/ekaia.16362

Jasoa: 2016-05-15 Onartua: 2016-06-29

Laburpena: hizkuntzaren prozesamenduan (HP), denbora-informazioa beharrezkoa da testuak ulertzeko, testuko gertaerak noiz jazotzen diren edo zenbat irauten duten adierazten baitu. Artikulu honetan, euskarazko denbora-informazioaren azterketa eta prozesamendua aurkezten dira. Lehenik, denbora-egituren deskribapena egin da. Bigarren, informazio egituratua emateko markaketa-lengoaia eta horren bidez etiketatutako corpusak azaldu dira. Ondoren, etiketatzeko tresna automatikoa ere deskribatzen da eta lehen etiketatze automatikoaren saiakera bat eta horren emaitzak ere ematen dira.

Hitz gakoak: hikuntzaren prozesamendua, denbora-informazioa, markaketa-lengoaia, corpus etiketatua.

Abstract: Temporal information is compulsory for textual comprehension, since it describes when the events in text happen or their duration. In this article temporal information and processing are presented. First the temporal constructions are described. Secondly, the mark-up language that structures the data and the annotated corpora following it are shown. Finally, we describe automatic tool for annotation and a first automatic annotation effort and its results are also described.

Keywords: Natural Language Processing, temporal information, mark-up language, annotated corpus.

1. SARRERA

Denborak ekintzak eta egoerak gertatzea eta gizakiok gertaera horien ondorio diren aldaketak edo aldaketa ezak nabaritzea ahalbidetzen

du. Denborako une bakoitzari izen zehatz bat ematen zaio eta iraupenen luzera denbora-unitatetan adierazten da. Balio eta izen hauek, ordea, hiztun-komunitate bakoitzaren araberakoak dira: nola ulertu denbora, hala adierazi. Horrez gain, hizkuntzaren prozesamenduan (HP) informazio linguistikoa baliatzen da testuak automatikoki ulertu edota sortzeko. Horregatik, hizkuntza bakoitzean denbora-informazio hori nola gauzatzen den aztertu behar da.

HPn denbora-informazioa hainbat sistematan izan daiteke erabilgarria, hala nola kronologien sorrera automatikoan [1], gertaeren aurreikuspenean [2] eta etorkizunaren iragarpenean [3]. Horretarako, hiztunok denbora adierazteko erabiltzen ditugun egiturak identifikatu, normalizatu eta horien ezaugarriak azaleratu behar dira. Informazio hori markaketa-lengoaiek tresna automatikoentzat atzigarri den formatuan antolatzen dute eta horien bidez etiketatutako corpusetan gordetzen da informazioa.

Euskararako ere denbora-informazioaren azterketa eta prozesamendua egiten ari gara, euskarazko testuetako informazioa automatikoki erauzi eta tresna automatikoetan baliatu ahal izateko. Euskarazko gertaerak, denborako une eta iraupenak eta horien artean sortzen diren denbora-erlazioak adierazten dituzten egiturak aztertu eta haien ezaugarriak identifikatu ditugu 2. atalean. Ondoren, 3. atalean, informazio horren EusTimeML markaketa-lengoaiaren [4] bidezko kodetzea eta horren araberako euskarazko corpusen etiketatzea aurkezten dira. 4. atalean, HeidelTime tresnaren euskararako moldaketa eta lehen etiketatze automatikoaren saiakera azaltzen dira. Amaitzeko, ikerketaren ondorioak adierazten dira 5. atalean.

2. **DENBORA-INFORMAZIOA EUSKARAZ**

Euskaraz, aztertutako beste hizkuntzetan bezala, denbora-informazioa adierazteko hiru elementu nagusi daude: gertaerak, denbora-adierazpenak eta denborazko erlazio-eraikuntzak. Horiez gain, TimeML markaketa-lengoaiak [5] horien arteko hiru erlazio mota ere proposatzen ditu. Jarraian deskribatuko dira elementu eta erlazio horiek.

2.1. **Gertaerak**

Gertaerak gertatzen diren ekintzak (1) eta egoerak (2) dira [6] (kurtsibaz adibideetan), eta euskaraz nagusiki aditzen (*erabaki zuen*) eta izenen (*su-etena*) bidez adierazten dira.

(1) Gobernuak joan den astean *erabaki zuen* helegitea jartzea.
(2) Gauerditik indarrean da *su-etena* Sirian.

Gertaerak esanahiaren eta beren denborarekiko gauzatzearen arabera sailka daitezke. Hainbat azterketa egin dira, baina lan honetan Croften [7] sailkapenari jarraituko zaio (1. taula) egindako sintesi-lanagatik:

1. taula. Gertaeren sailkapena denboran duten gauzatzearen arabera.

Gertaera-mota	Gertaera-azpimota		Adibideak
egoerak (*states*)	iragankorrak		Bidea *bustita* dago.
	iraunkorrak	bereganatutakoak	Mugikorra *apurtuta* dago.
		berezkoak	Marie *frantsesa* da.
	puntuak		*Arratsaldeko bostak* dira.
lorpenak (*achievements*)	zuzendutako itzulgarriak		Leihoa *zabaldu da*.
	zuzendutako itzuliezinak		Leihoa *puskatu da*.
	ziklikoak		Txakurrak *zaunka egin du*.
ekintzak (*activities*)	zuzenduak		Zopa *hoztu da*.
	zuzendu gabeak		Neskek *abestu zuten*.
jarduerak (*performances*)	hazten diren lorpenak		Sagar bat *jan dut*.
	hazten ez diren lorpenak		Mikelek ordenagailua *konpondu du*.

Gertaera guztiak ez dira berdin gauzatzen denboran 1. taulan ikus daitekeenez; batzuk une batean gertatzen dira (egoera-puntuak eta lorpenak) eta beste batzuek denbora-tarte bat irauten dute (egoera iragankor eta iraunkorrak, eta ekintzak). Halaber, horietako batzuen (jarduerak) hasiera- eta amaiera-uneak ezagut ditzakegu, baina ez beste batzuenak (ekintzak). Bereizketa honen analisiak testuko denbora-adierazpenekiko harremana erabakitzen lagunduko du; iraupenik gabeko gertaerak iraupena adierazten duten denbora-adierazpenekin ezin lotzea, adibidez.

2.2. **Denbora-adierazpenak**

Denbora-adierazpenek kronologiako une edo tarte bati egiten diote erreferentzia eta gertaerak denboran kokatzeko baliatzen dira, arestian esan bezala. Euskaraz postposizio-sintagmen bidez (3) edo adberbioen bidez (4) adierazten dira nagusiki. Denbora-erreferentzia batzuk absolutuak (3) dira, baina beste batzuen kokapen kronologikoa zehaztu ahal izateko, testuaren sorrera-unea ere ezagutu behar da (4).

(3) Estatu Batuak *1941eko abenduaren 7an* sartu ziren gerran Ardatzaren kontra.
(4) Askatasunaren plaza izena jartzea *atzo* erabaki zuen udalbatzak.

2.3. **Denborazko erlazio-eraikuntzak**

Hizkuntzek gertaeren eta denbora-adierazpenen arteko erlazioak esplizituki adierazteko denborazko erlazio-eraikuntzak ((5) eta (6)), nagusiki postposizio sinple eta konplexuak erabiltzen dituzte. Horiek gertaera bat noiz gertatu den edo zenbat luzatu den adierazten dute.

(5) Orain *arteko* terminal guztiak prest daude A380a hartzeko.
(6) Japoniako Nikkei indizea izan da egun*ean zehar* hazi den gutxietako bat.

2.4. **Erlazioak**

TimeML markaketa-lengoaiak proposatzen dituen erlazioak euskararako baliatzea erabaki da. Hiru erlazio-mota bereizten dira, eta erlazio bakoitzak lotura jakin bat adierazten du:

— **Mendekotasun-erlazioek** (7) gertaera nagusi baten (beltzez) eta mendeko baten (kurtsibaz) arteko mendekotasuna adierazten dute.

(7) Erreserba Federalak *ez duela* mailegu gehiago *egingo* **adierazi zuen**.

— **Aspektu-erlazioek** (8) aspektuzko gertaera baten (beltzez) eta bere argumentu den beste gertaera baten (kurtsibaz) arteko erlazioa agertzen dute.

(8) Errusiako bankuek *itxita* **jarraituko** dute hirugarren egunez.

— **Denbora-erlazioak** (9) gertaeren (beltzez) arteko edo gertaera baten eta denbora-adierazpen baten (kurtsibaz) arteko erlazioak dira.

(9) Errusiako burtsak *ostiral goizean* **zabalduko dira** berriro.

Hurrengo atalean egitura eta erlazio hauen tratamendua nola gauzatzen den azalduko da.

3. **DENBORA-INFORMAZIOAREN ETIKETATZEA ETA CORPUS ANOTATUEN SORRERA**

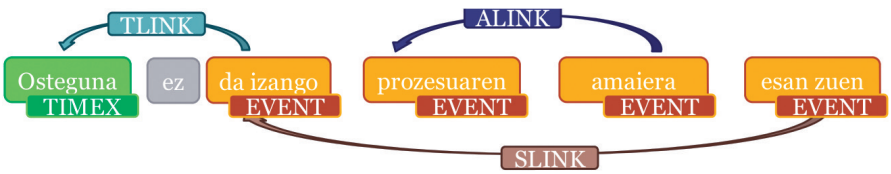
Denbora-egituren azterketa egin ondoren ([8] eta [9]), egitura horien informazio linguistikoa (morfosintaxia eta semantika) normalizatu eta in-

terpretatzeko moduan jarri behar da automatikoki tratatzeko. Euskarazko denbora-informazioa etiketatzeko, EusTimeML markaketa-lengoaia garatu dugu, ingelesezko ISO-TimeML denbora markaketa-lengoaian [5] oinarrituta. Horrez gain, denbora-informazioa etiketatuta duten corpusak (EusMEANTIME, WikiWarsEu eta FaCor) ere osatu ditugu. Corpus horiek hizkuntzaren analisirako eta tresna automatikoen etiketatzea ebaluatzeko *gold standard* bezala baliatuko dira.

3.1. **EusTimeML markaketa-lengoaia**

TimeML denbora-egiturentzat *de facto* estandarra bihurtu da [10] denbora-informazioaren adierazpenean parte-hartzen duten elementu nagusiak etiketatzeko balio izan zuen lehena izan zen-eta. Euskararen denbora-egiturentzat ere hori moldatzea erabaki da; hortik EusTimeML izena. EusTimeMLren bidez, gertaerak, denbora-adierazpenak, beren artean sortzen diren erlazioak eta erlazio horiek esplizitu egiten dituzten erlazio-eraikuntzak etiketatu eta horien ezaugarriak azaleratu dira.

Aurreko atalean deskribatutako denbora-egiturak etiketatzeko, hiru kategoria eskaintzen ditu: EVENT gertaerentzat, TIMEX3 denbora-adierazpenentzat, eta SIGNAL denborazko erlazio-eraikuntzentzat. Horien artean sortzen diren hiru erlazio motak bistaratzeko ere hiru kategoria eskaintzen ditu: 2.4 atalean deskribatutako mendekotasun-erlazioak (SLINK), aspektu-erlazioak (ALINK) eta denbora-erlazioak (TLINK).



1. irudia. EusTimeMLren etiketa bidezko esaldi baten etiketatzea.

Aurretik deskribatutako elementu batzuk ikus daitezke 1. irudiko *«Osteguna ez da izango prozesuaren amaiera» esan zuen* esaldian. TLINK-aren bidez *(ez) da izango* gertaera denboran (*Osteguna*) kokatzen da. Bestalde, *esan zuen* gertaeraren eta esandakoaren (*Osteguna ez da izango prozesuaren amaiera*) arteko mendekotasun-erlazioa *esan zuen* eta *(ez) da izango* gertaeren arteko SLINK-aren bidez azaltzen da. *Amaiera*-ren eta *prozesuaren*-en arteko aspektuzko erlazioa, hau da, prozesua amaitu dela, ALINK-aren bidez adierazten da.

EusTimeML XML lengoaian oinarritzen da eta etiketak testutik kanpo gordetzen ditu. Token bakoitzaren identifikazioaz gain, denbora-egitura-

ren ezaugarriak TIMEX3 etiketaren barruan adierazten dira, aurrez definitutako atributuen bidez (morez) eta atributu horien balioen bidez (arrosaz) (2. irudia).

```
<token t_id="26" sentence="2" number="25">Osteguna</token>

<TIMEX3 m_id="15" functionInDocument="NONE" endPoint="" anchorTimeID="13"
beginPoint="" quant="" freq="" mod="" value="2014-03-20" type="DATE">
<token_anchor t_id="26"/>
</TIMEX3>
```

2. irudia. «Osteguna» denbora-adierazpenaren etiketatze-adibidea.

Osteguna denbora-adierazpenaren etiketatzea agertzen da 2. irudian. Testuko hitza (osteguna) <token> etiketaren bidez adierazten da eta identifikatzaile uniboko bat esleitzen zaio (t_id= "26"). Denbora-egituraren etiketa, <TIMEX3>, denbora- adierazpena denez, token horri lotzen zaio; hala ikus daiteke <token_anchor> etiketa hutsaren bidez. Unearen balio normalizatua (value="2014-03-20") eta denbora-adierazpena zein motatakoa den (type="DATE") value eta type atributuen bidez adierazten da. Denbora-adierazpenaren identifikatzailea (m_id) identifikatzaile unibokoa da eta denbora-adierazpenak dokumentuan duen funtzioak (functionInDocument) testutik kanpoko funtziorik duen adierazten du, esaterako, denbora-adierazpena dokumentuaren sorrera-data den. Denbora-adierazpenaren balioa kalkulatzeko erreferentzia den denbora-ainguraren identifikatzailea (anchorTimeID) ere adierazten da. Beste atributu batzuk ere agertzen dira: iraupenentzat hasiera- (beginPoint) eta amaiera-puntuak (endPoint), kantitatea (quant), maiztasuna (freq) eta denbora-adierazpenaren modifikatzaileak (*hasiera*, *baino gehiago*, etab.) adierazteko (mod). Horiek hutsik daude ez baitagozkio adibideko adierazpenari.

3.2. **Euskarazko corpusak**

FaCor, EusMEANTIME eta WikiWarsEu corpusak denbora-informazioaren azterketan eta tresnen garapenean erabili dira. Horretaz gain, etiketatzaileen trebakuntzarako eta *gold standard*aren sorrerarako ere baliatu dira. *Gold standard* hori tresna automatikoen ebaluaziorako erabili da.

FaCor corpusa Fagorren itxierari buruzko prentsako 25 albistez osatuta dago. Corpus hori, alde batetik, etiketatzaileak trebatzeko, euskarazko denbora-egiturak identifikatzeko eta etiketatze-gidalerroak sortzeko erabili da. Beste aldetik, EusHeidelTime tresna automatikoarentzat erregelak sortzeko eta tresna bera ebaluatzeko ere erabili da.

NewsReader proiektuko MEANTIME corpusa [11] euskarara itzuli da. MEANTIME corpusak 120 dokumentu ditu; zehazki ekonomiari buruzko

prentsa-albisteak dira. Horiek EusTimeML markaketa-lengoaiaren bidez eskuz etiketatzen ari gara. Eskuz etiketatutako lagin baten azterketaren bidez, denbora-informazioaren gaineko etiketatze-erabakien argitasuna eta egokitasuna ebaluatu dira. Etiketatutako lagin hori EusHeidelTime tresna garatzeko eta horren funtzionamendua neurtzeko *gold standard* moduan ere erabili da.

WikiWars corpusa [12] historian gertatutako gerrei buruzko ingeleseko Wikipediako hogei dokumentuz osatuta dago. Jatorrizko dokumentuek erreferentzia egiten dieten hogei gerretatik hemeretziren euskarazko ordaina aurkitu da eta euskarazko narrazio-testuen corpusa sortu da. Corpus horretako 12 artikulu erabili dira narrazio-testuetan ohikoak diren denbora-adierazpenak identifikatzeko, eta beste 7, EusHeidelTimeren ebaluaziorako.

Corpus bakoitzaren ezaugarri nagusiak, hala nola, tamaina osoa, etiketatuta dagoena eta entrenamendu eta ebaluaziorako erabilitako laginak 2. taulan agertzen dira:

2. taula. Euskarazko corpusen deskribapena.

Corpusak	Tamaina	Anotazio-mota	Atazak eta erabilerak
FaCor (albisteak)	25 dokumentu 6 K hitz	TIMEX3 SIGNAL EVENT	14 dok/3,8 K hitz etiketatzaileen trebaketa
			4 dok/ 939 hitz TIMEX eta SIGNAL etiketatzaileen arteko adostasuna gidalerroen egokitasunaren ebaluazioa
			17 dok/4,5 K hitz HeidelTime patroien sorkuntza
			8 dok/1,5 K hitz HeidelTime ebaluazioa
EusMEANTIME (albisteak)	120 dokumentu (30 dokumentuko lagina/8,7 K hitz)	EusTimeML maila guztiak (dokumentuen lehen 5 lerroak)	15 dok/3,9 K hitz EVENT etiketatzaileen arteko adostasuna
			20 dok/5,2 K hitz garapenerako *gold standard* (HeidelTime patroien sorkuntza)
			10 dok/3,5 K hitz testerako *gold standard* (HeidelTime ebaluazioa)
WikiWarsEu (narrazioak)	19 dokumentu 35,9 K hitz	TIMEX3	12 dok/22,3 K hitz HeidelTime patroien sorkuntza
			7 dok/13,6 K HeidelTime ebaluazioa

3.3. **Corpusen eskuzko etiketatzea**

Eskuzko etiketatzean hiru fase bereizten dira: etiketatzaileen trebakuntza, erabaki linguistikoen ebaluaziorako etiketatze-saiakerak eta *gold standard* corpusaren etiketatzea. Ataza horiek CELCT Annotation Tool (CAT) [13] tresnaren bidez egin dira, tresnak atazak sortu eta moldatzeko eskaintzen duen erraztasunagatik.

Etiketatzaileen trebakuntza-fasean, etiketatzaileek EusTimeML etiketatze-gidalerroak eta CAT tresna ezagutu dituzte. Bigarren fasean, denbora-egituren gainean hartutako erabakien egokitasuna eta etiketatze-gidalerroen estaldura eta argitasuna ebaluatu dira anotatzaileen arteko adostasuna neurtuz ([8] eta [9]). Urrats honi analisi linguistikoaren eta etiketatzeko erabakien errebisio-fase batek jarraitu dio eta gidalerroak moldatu dira. Behin etiketatze-irizpideak finkatuta, *gold standard*a izango den corpusa etiketatu da.

4. **SAIAKERA AUTOMATIKOAK**

Hizkuntza prozesatzeko metodoak hiru multzotan bana daitezke: i) ezagutza linguistikoan oinarritzen diren erregela bidezkoak [14], ii) ikasketa automatikoan oinarritzen direnak [15] eta iii) bien arteko hibridoak [16]. Lehenengo multzokoetan, ezagutza linguistikoan oinarrituta, hizkuntzaren egiturak identifikatuko dituzten erregelak sortzen dira. Bigarrenetan, ordea, etiketatutako corpus handietatik entitateen ezaugarriak baliatuz ikasten dute tresnek entitate berriak identifikatzen. Hirugarrenetan erregelak eta metodo estatistikoak konbinatzen dira. Entitate-motaren arabera, metodo bat besteak baino egokiagoa izan daiteke.

Gertaerentzat ikasketa automatikoa da irtenbiderik egokiena, gertaerak oso ugariak, askotarikoak eta multzokatzen zailak baitira eta mota guztietako gertaeren lagin erabilgarria erraz bil baitaiteke. Corpus etiketatuan, gertaera-etiketa (EVENT) jasotzen duten entitateetatik eta haien ezaugarrietatik ikasten du tresna automatikoak, eta testu berrietan gertaerak identifikatzen saiatzen da.

Denbora-adierazpenak automatikoki prozesatzeko, hizkuntzaren azterketa teorikoa egin ohi da eta horretan oinarrituta denbora-adierazpenen egitura posibleak identifikatzen eta normalizatzen dituzten erregelak sortu ohi dira. Denbora-adierazpenek, oro har, egitura-multzo murritza osatzen dute eta ez dute aldakortasun handirik. Ondoko adibideetan ikus daiteke egiturok eredu bera mantentzen dutela, elementuak gehituz zabal badaitezke ere (10) edota mota bereko elementuak trukatuz alda badaitezke ere (11).

(10) *1937ko apirilean* izan zen Gernikako bonbardaketa, hain zuzen ere, *1937ko apirilaren 26ko arratsaldeko hiru eta erdietan*.

(11) *Orain dela ia ordu bete* nago autobusaren zain; *orain dela hiru ordu laurden* igaro behar zuen hemendik.

Denbora-adierazpenak ez dira oso ugariak testuetan, eta hainbat genero batzen dituzten tamaina handiko corpusak behar dira hizkuntzan ager daitezkeen denbora-adierazpenen laginak lortu ahal izateko.

Denborazko erlazio-eraikuntzekin ere gauza bera gertatzen da. *Eta gero*, *ondoren*, *baino lehen*, *arte* eta moduko erlazio-eraikuntzek osatzen duten multzoa are murritzagoa da eta testuetako agerpenak oso urriak dira. Horiek automatikoki antzemateko eskuz sortutako zerrenda bat aski izan daiteke.

Erlazioak automatikoki sortzeko garrantzitsua da gertaeren informazio sintaktikoari eta semantikoari erreparatzea. Informazio hori erauzi eta erlazioaren iturri eta xedeak aukeratu behar dira. Horretarako, ikasketa automatikoa balia daiteke [15], eta ikasketa automatikotik ikasitako elementuen ezaugarri-kopuru handia eraginkortasunez erabili ahal izateko, metodo estatistikoak erabil daitezke.

4.1. **EusHeidelTime tresna**

Denbora-adierazpenen identifikaziorako eta erauzketa automatikorako, HeidelTime [14] tresna euskararako moldatu da. Ezagutza linguistikoan oinarritzen den tresna honek hizkuntzalariek definitutako erregelak baliatzen ditu informazioa kodetzeko. Bertan euskarazko testu-prozesatzaile automatikoa txertatu da [17] eta prozesaturiko testuaren gainean denbora-adierazpenen prozesatzailea abiarazi da.

(12)
```
RULENAME="eta_hamar",EXTRACTION="%reOrduak(rak|ak)
eta %reMinutuak%reMarkak?",NORM_VALUE="UNDEF-REF-
day-PLUS-0T%normOrduak(group(1)):%normMinutuak
(group(4))"
```

Bostak eta hogeian moduko egiturak erauzteko erregela dago (12) adibidean, eta erregelaren izenak (RULENAME), bilatu behar den patroiak (EXTRACTION) eta emango zaion balio normalizatuak (NORM_VALUE) osatzen dute. Erauzketa-patroian ikusten da ordua adieraziko duen katea (*bost*), *rak* edo *ak* kateak (*ak*), *eta* juntagailua, minutuak adierazten dituen katea (*hogei*) eta postposizio-markak hartzen dituen segida (*an*) identifikatu behar dela. Normalizazioan ikus daiteke orduak (*bost*) adierazten dituen katearen (group (1), patroiaren lehen postuan dagoelako) eta minutuak (*hogei*) adierazten dituen katearen (group (4), patroiaren laugarren postuan dagoelako) balio normalizatua kate osoaren balio normalizatuan txertatuko direla, %normOrduak eta %normMinutuak normalizazio-fitxategien bidez.

Erregelak sortu ondoren, tresnaren funtzionamendua ebaluatu da TempEval-3ko [18] formatuari jarraituz. Lehen saiakera honetan FaCor, EusMEANTIME eta WikiWarsEU corpusetako 49 dokumentuko lagina erabili da erregelak sortu eta beren estaldura neurtzeko. (13) adibidean etiketatze automatikoaren eta *gold standard*aren arteko bat etortze osoa (*strict match*) erakusten da. Bai denbora-adierazpenaren forman bai balio normalizatuan bat etortzea dago. (14) adibidean, ordea, formaren gaineko bat etortzea partziala baino ez da eta ez dago bat etortzerik balioan; horregatik, bat etortze partziala (*relaxed match*) dagoela esaten da.

(13) `gold annotation: <TIMEX3 type=»DATE» value="2014-W11" tid="t6">Joan den astean</TIMEX3>`
`system annotation: <TIMEX3 type="DATE" value="2014-W11» tid="t6">Joan den astean</TIMEX3> -- strict match`

(14) `system annotation: <TIMEX3 type="DATE" value="UNDEF-REF-day" tid="t19">une</TIMEX3> -- relaxed match`
`-> gold value: <TIMEX3 type="DATE" value="PRESENT_REF" tid="t19">une honetan</TIMEX3>`
`-> system wrong value: <TIMEX3 type="DATE" value="UNDEF-REF-day" tid="t19">une</TIMEX3>`

Garapen-corpusaren gaineko EusHeidelTimeren emaitzak 3. taulan agertzen dira, bat etortze osoarentzat, bat etortze partzialarentzat, balio normalizatuarentzat (*attribute value*) eta motarentzat (*attribute type*).

3. taula. EusHeidelTimeren errendimendua garapen-corpusaren gainean.

	WikiWarsEu			EusMEANTIME			FaCor		
	P	R	F1	P	R	F1	P	R	F1
Strict match	68,16	79,71	73,49	69,54	65,05	67,22	83,87	74,82	79,09
Relaxed match	76,28	89,2	82,23	79,89	74,73	77,22	88,71	79,14	83,65
Attribute value			65,28			49,44			55,51
Attribute type			79,95			68,33			80,61

Corpus bakoitzean lortutako doitasuna (P), estaldura (R) eta F neurria (F1) ere agertzen dira 3. taulan. Ikus daitekeen moduan, FaCor corpusean lortu dira emaitzarik onenak: % 79,09ko F1 balioa bat etortze osoan, eta % 83,65 bat etortze partzialean. Mota atributuan ere (*attribute type*) corpus honetan lortu dira emaitzarik onenak. Balioan (*attribute value*), ordea, WikiWarsEu corpusean daude emaitzarik onenak (% 65,28 F1ean), corpus honetan data absolutu asko baitago.

5. ONDORIOAK ETA ETORKIZUNEKO LANAK

Artikulu honetan, euskaraz denbora-informazioa daramaten zenbait egitura aztertu eta horien ezaugarriak azaleratu dira hizkuntzaren prozesamenduari begira. Halaber, EusTimeML moldatu da, ISO-TimeML oinarritzat hartuta, euskarazko denbora-egiturak eta horien ezaugarriak egoki kodetzeko.

Tresnei dagokienez, euskarazko denbora-informazioaren prozesamenduan CELCT Annotation Tool (CAT) egokitu da eskuzko etiketatzerako, markaketa-lengoaiak txertatzeko duen erraztasunagatik. Tresna hori denbora-informazioa batzen duten corpusak sortzeko erabili da. Corpus horiek tresna automatikoen funtzionamendua ebaluatzeko *gold standard* moduan baliatu dira. Horrez gain, HeidelTimen euskarazko denbora-adierazpenentzat erregelak sortu dira adierazpen horien identifikazio eta normalizazio automatikorako, denbora-adierazpenek egitura zehatzei jarraitzen baitiete, oro har. Erregelen estaldura eta doitasuna ebaluatu dira asmatze-tasari erreparatuz eta denbora-adierazpenen identifikazio eta normalizazioan. Denbora-adierazpenen % 73,27 zuzen identifikatu da (bat etortze osoan batez beste % 73,27ko adostasuna), balio zuzena % 56,74ri esleitu zaio eta % 76,3ri mota egokia eman zaio batez beste. Hasierako emaitza horiek ontzat hartzen ditugu.

Etorkizuneko lanen artean, EusHeidelTime garatzen jarraituko da eta emaitzak hobetzeko, erroreen analisia egingo da. Izan ere, EusHeidelTime corpus anotatua zabaltzeko lanetan erabiltzea dugu helburu.

Horrez gain, denborazko erlazio-eraikuntzak eta gertaerak identifikatzeko eta beren ezaugarriak erauzteko tresnak ere garatzen ari gara. Ezaugarri horiek denbora-erlazioak sortzeko ere baliatuko dira. Bukatzeko, baliabide guztiak bateratuz euskarazko denbora-informazioaren prozesamendurako sistema eraikiko da.

6. ESKER ONAK

Lan hau Eusko Jaurlaritzako Hezkuntza Sailaren PRE_2015_2_0284 bekaren bidez finantzatu da.

Lan hau I. IkerGazte Nazioarteko Ikerketa Euskaraz kongresuko Giza zientzien arloan «Euskarazko denbora-egituren tratamendu automatikorako azterketa» artikuluari emandako sariari esker argitaratu da.

7. **BIBLIOGRAFIA**

[1] BAUER, S., CLARK, S. eta GRAEPEL, T. 2014. «Learning to identify historical figures for timeline creation from wikipedia articles». In *Proceedings of HistoInformatics2014 - the 2nd International Workshop on Computational History*, Barcelona, Spain.

[2] RADINSKY, K. eta HORVITZ, E. 2013. «Mining the web to predict future events». In *Proceedings of the sixth ACM international conference on Web search and data mining*, 255-264. ACM.

[3] KAWAI, H., JATOWT, A., TANAKA, K., KUNIEDA, K. eta YAMADA, K. 2010. «Chronoseeker: Search engine for future and past events». In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication* ICUIMC '10, ACM, http://doi.acm.org/10.1145/2108616.2108647, 25:1-25:10. New York, NY, USA.

[4] ALTUNA, B., ARANZABE, M.J. eta DÍAZ DE ILARRAZA, A. 2016. «Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0». Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016 https://addi.ehu.es/handle/10810/17305

[5] ISO-TimeML WORKING GROUP. 2008. «Language Resource Management — Semantic Annotation Framework (SemAF) — Part 1: Time and Events. International Standard ISO/CD 24617-1(E). ISO».

[6] SAURÍ, R., BATIUKOVA, O. eta PUSTEJOVSKY, J. 2009. «Annotating Events in Spanish. TimeML Annotation Guidelines. Technical Report Version TempEval-2010». Barcelona Media-Innovation Center.

[7] CROFT, W. 2015. «Force dynamics and directed change in event lexicalization and argument realization». In *Cognitive Science Perspectives on Verb Representation and Processing*, 103-129. Springer International Publishing.

[8] ALTUNA, B., ARANZABE, M.J. eta DÍAZ DE ILARRAZA, A. 2014. «Euskarazko denbora-egiturak. Azterketa eta etiketatze esperimentua». *Linguamática*, **6**, 13-24.

[9] ALTUNA, B., ARANZABE, M.J. eta DÍAZ DE ILARRAZA, A. 2016. «Adapting TimeML to Basque: Event Annotation ». In *CICLing 2016, 17th International Conference on Intelligent Text Processing and Computational Linguistics*. Konya, Turkey. Springer.

[10] SCHILDER, F., KATZ, G. eta PUSTEJOVSKY, J. 2007. «Annotating, Extracting and Reasoning about Time and Events». In *Annotating, Extracting and Reasoning about Time and Events*, 1-6 . Springer.

[11] MINARD, A.M. SPERANZA, R. URIZAR, B. ALTUNA, M. VAN ERP, A. SCHOEN eta C. VAN SON. 2016. «MEANTIME, the NewsReader Multilingual Event and Time Corpus». In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portoroz, Slovenia. European Language Resources Association.

[12] MAZUR, P. & DALE, R. 2010. «WikiWars: A New Corpus for Research on Temporal Expressions». In *Proceedings of the 2010 Conference on Empiri-*

cal Methods in Natural Language Processing, 913-922. EMNLP '10. Massachusetts, USA: Association for Computational Linguistics.

[13] BARTELESI LENZI, V., MORETTI, G. eta SPRUGNOLI, R. 2012. «CAT: the CELCT Annotation Tool». In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* 333-338. Istanbul, Turkey: European Language Resources Association (ELRA).

[14] STRÖTGEN, J. eta GERTZ, M. 2010. «HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions». In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, 321-324. Association for Computational Linguistics, Stroudsburg, PA, USA.

[15] BETHARD, S. 2013. «ClearTK-TimeML: A minimalist approach to TempEval 2013». In S. Manandhar & D. Yuret (Eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM) 2: Seventh International Workshop on semantic Evaluation (SemEval 2013)*, 10-14. Atlanta, Georgia, USA: Association for Computational Linguistics.

[16] JEONG, Y.S. eta CHOI, H.J. 2015. «Language Independent Feature Extractor». In B. Bonet & S. Koenig (Eds.), *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 4170-4171. Austin, Texas, USA: AAAI Press.

[17] AGERRI, R., BERMÚDEZ, J. eta RIGAU, G. 2014. «IXA pipeline: Efficient and Ready to Use Multilingual NLP tools». In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* 3823-3828. Reykjavik, Islandia.

[18] UZZAMAN, N., LLORENS, H., ALLEN, J.F., DERCZYNSKI L., VERHAGEN, M. eta PUSTEJOVSKY, J. 2012. «TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations», in *Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval 2013*, 1-9. Association for Computational Linguistics, Stroudsburg, PA, USA.