Euskarazko gertaeren etiketatze automatikoa

Haritz Salaberri, Olatz Arregi eta **Beñat Zapirain**

IXA taldea (UPV/EHU). Manuel Lardizabal Pasealekua 1. 20018 Donostia

Laburpena

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

Hitz gakoak: Hizkuntzaren prozesamendua, adimen artifiziala, semantika, euskara

Abstract

In this paper we present the results obtained by bEVENT, the first system developed for automatically annotating events in Basque that follows the ISO-TimeML standard. A two-step procedure is performed in order to conduct annotation: firstly, event-extents are identified, and secondly, values are given to the linguistic attributes specified for these extents. The Euskal-TimeBank corpus has been used to train and evaluate bEVENT and the system results have been calculated using train-test evaluation. bEVENT implements a machine learning based approach and the precision, recall and F_1 measures that correspond to the extraction of events are the following: 83.92, 72.76 and 77.94.

Keywords: Human language technology, artificial intelligence, semantics, Basque

1 Sarrera eta motibazioa

Gertaerak argitalpen honen oinarria dira, baina, zer da *gertaera*? Eta *gertaera* kontzeptuari eman zaizkion definizio guztietatik, zein da guk darabilguna?, alegia, zer da guretzat *gertaera*?

Zientziaren alor bat baino gehiago arduratu izan dira gertaerak lantzeaz, besteak beste fisika, filosofia eta hizkuntzalaritza. Fisikan, esate baterako, erlatibitatearen teoria ikertzean, gertaera kokaleku eta une zehatz batean agitzen den egoera fisikoa dela ulertzen da; une jakin batean edalontzi bat lurraren kontra puskatzea, adibidez. Filosofian, berriz, hainbat dira gertaeren inguruko teoriak; hauetatik ezagunenak eta hizkuntzaren filosofian eragin handiena izan dutenak hauek: Kimena (Kim, 1976), Lewisena (Lewis, 1987), Badiou eta Felthamena (Badiou eta Feltham, 2007), Deleuzerena (Deleuze, 1988) eta Davidsonena (Davidson, 1967).

Zerrendatutako teorietatik, hizkuntzaren prozesamenduak, historian zehar, Davidsonek proposatutakoa jarraitu izan du (2). Izan ere, semantika konputazionalean perpausak adierazteko teoria honetan oinarritzen den semantika neo-davidsondarra (Parsons, 1990) erabili ohi da. Hortaz, hizkuntzaren prozesamenduan, eta ondorioz guretzat, gertaeraren definizioa teoria honek proposatzen duena izango da: *denboran eta espazioan kokatua dagoen eta kausa jakin baten ondorioz eragin jakin bat sortzen duen jazoera.*

1.1 Gertaeren kategorizazioa

Darabilgun definizioa kontuan edukita, gertaerek denborarekin eta espazioarekin duten lotura argia da. Hau dela eta, lan honetan, gertaerak identifikatzeaz gainera, haien propietateak, bereziki denborari da-

gozkienak, aztertuko ditugu, testuen etiketatze tenporal egokia burutu ahal izateko. Jakina denez, gertaerek denborarekin duten erlazio semantikoaren kategorizazioa aldatzen duten gramatikaren kategoriak bat baino gehiago dira. Esate baterako *Aktionsart* (Streitberg, 1891) edo aspektu lexikala (iraunkorra, ez-iraunkorra, errepikakorra, ez-errepikakorra, etab.), aspektua (burutua edo burutu gabea), modua (indikatiboa, subjuntiboa, optatiboa, etab.) eta denbora gramatikala (orainaldia, lehenaldia, etorkizuna). *bEVENT* tresnaren garapenean jarraituko duguna (Saurii *et al.*, 2005) argitalpenean proposatutako gertaeren kategorizazio espazio-tenporala da. Hau lortzeko hainbat kategoria gramatikal (aspektua, *Aktionsart*, ebidentzialitatea) eta lexiko (`Perception` aditzak, `Intensional state` aditzak eta abar) izan ziren kontutan. Kategorizazio honek ondorengo zazpi gertaera motak bereizten ditu:

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

1.2 Gertaeren gauzatze gramatikala

Aurreko azpiatalean gertaerak zer diren eta *bEVENT* sistemak hauen zer kategorizazio jarraitzen duen finkatu da. Azpiatal honetan, berriz, gertaeren gauzatze gramatikalean (Tenny eta Pustejovsky, 2000) ezinbestekoa den *predikatu* kontzeptua definitzen da, eta hau egin ahal izateko (gauzatze gramatikala azaltzea), gainera, *argumentu* eta *adjuntu* kontzeptuak ere aurkezten dira. Predikatuak garrantzizkoak dira *bEVENT* sistemaren diseinua eta funtzionamendua ulertzeko.

Guretzat predikatuak *gertaerak deskribatzen dituzten aditzak, adberbioak, izenak eta adjektiboak* dira. Argumentuak, bestalde, *predikatuek deskribatzen dituzten gertaeretan parte hartzen duten entitateak (nor, nori, nork)* dira. Azkenik, adjuntuak, *gertaeren nolakotasunak (non, noiz, nola...) adierazten dituzten propietateak* dira.

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

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

Uste dugu aurrera jarraitu aurretik garrantzizkoa dela argitzea predikatuaren gure definizioa eta teoria gramatikal *klasikoetan* ematen dena ez direla berdinak. Izan ere, gure definizioa bat dator egungo teoria gramatikalek defendatzen duten ikuspegiarekin, predikatuen argumentuak eta adjuntuak ez baitira sekula predikatuaren zati izango. Teoria klasikoek, ordea, perpausek bi osagai gramatikal dauzkatela defendatzen dute: subjektua eta predikatua. Subjektua normalean izen sintagmari dagokio, eta predikatua aditz sintagmari. Bereizketa honen ondorioz, subjektua ez diren argumentuak eta adjuntuak predikatuaren barnean kokatzen dira.

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

Adibideko lehenengo perpausean izen sintagma eta subjektua *jarraitzaileek* da; *zelaitik ikusi zuten partida*, berriz, osagaitzat *zelaitik* adjuntua, *ikusi* aditza eta *zuten* aditz laguntzailea dituzten aditz sintagma eta predikatua. Bigarren perpausean, *partida zelaitik ikusi zuten jarraitzaileak* erlatibozko perpausa perpaus nagusiaren izen sintagma eta subjektua da; *egongo dira finalean* aditz sintagma eta predikatua. Erlatibozko perpausean *jarraitzaileek* da subjektua eta *Partida zelaitik ikusi zuten* predikatua.

Predikatu kontzeptuaren bi definizio ezberdin hauen arteko bereizketa egiteko, azken aldian, egungo teoria gramatikalak predikatuari *predikatzailea* deitzen hasi zaizkio. Dena den, eta gure predikatuaren ulermena egungo teorietatik badator ere, ez diogu predikatuari *predikatzaile* deituko, *predikatu* baizik, hizkuntzaren prozesamenduan horrela deitu izan zaiolako.

2 Arloko egoera eta ikerketaren helburuak

Testuen semantika egituratzen duten oinarrizko unitateak *predikatu-argumentu-adjuntu* egiturei lotutako gertaerak direla azaldu dugu 1.2 azpiatalean. Argumentuek eta adjuntuek gertaeren hainbat propietateren berri ematen dute; besteak beste, gertaerak denboran kokatzen laguntzen dute. Jakina denez, hizkuntzaren prozesamenduan rol semantikoen etiketatze automatikoaz arduratzen den atazak, SRL deitutakoak, argumentuak eta adjuntuak, eta ondorioz propietate hauek, detektatzeko gaitasuna dauka. Esan beharra dago, hala ere, SRLk ematen duen gertaeren inguruko informazio tenporala mugatua dela, eta interesgarria dela, gure ustez behintzat, informazio erauzketa sistemetarako adibidez, gertaeren inguruko informazio tenporal ahal den aberatsena automatikoki eskuratu ahal izatea. Hori erdiesteko *ISO-TimeML* (Pustejovsky *et al.*, 2010) estandarrean oinarritutako *bEVENT* etiketatzailea garatu dugu. *ISO-TimeML* testuetako denbora-informazioa etiketatzeko sortutako anotazio eskema eta hizkuntza da. Anotazio eskemak *hizkuntza naturaleko informazio linguistikoa nola markatu edo bildu behar den ezartzen duten formalismoak dira.* Markaketaren helburua testuetako informazio linguistikoa konputazionalki tratatu edo prozesatzeko gaitasuna eskuratzea da.

2.1 *ISO-TimeML* eskema

Hiru osagai tenporal bereizten dira *ISO-TimeML*n:

- `<EVENT>`: Gertaeren buru lexikalak markatzen ditu (*joan*, *pentsatu*, *dakusat*).
- `<TIMEX3>`: Denbora adierazpenak markatzen ditu (*urtarrilaren 31n*, *1923/4/23*, *ostiral arratsaldean*).
- `<SIGNAL>`: Denbora seinaleak markatzen ditu (*ondoren*, *baino lehen*, *aurretik*).

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

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

Adibidean bi gertaera etiketatu dira: *ikusi* eta *ekaitzaren*. Lehenengoaren kasuan gertaera osoa *ikusi zuten* bada ere, *ikusi* bakarrik markatu da, *ISO-TimeML* eskemaren egokitzapenean gertaeren buru lexikala markatzeko erabakia hartu zelako; *ikusi* aditz predikatu batez deskribatutako `PERCEPTION` kategoriako gertaera da. Etiketatu den bigarrena berriz, *ekaitzaren*, izen predikatu baten bitartez deskribatutako `OCCURRENCE` kategoriakoa da. Adibidean *ostiral goizean* denbora adierazpena eta *ondoren* seinale tenporala ere etiketatu dira. Kontuan hartu behar da, dena den, argitalpen honetan aurkezten dugun *bEVENT* tresna gertaerak eta hauek jasotzen dituzten atributuak etiketatzeaz (`<EVENT>`) baizik ez dela arduratzen. Etorkizunean egin beharreko lanen artean dago denbora adierazpenak eta seinaleak ere automatikoki etiketatu ahal izatea. Gertaeren aipatutako atributuak dira hauen propietateak adierazten dituztenak. Ingeleserako *ISO-TimeML* eskeman `class`, `tense`, `aspect`, `polarity` eta `pos` atributuak

hartzen dituzte gertaerek; euskaraz, ordea (egokitzapenaren ondoren), `class`, `tense1`, `tense2`, `aspect1`, `aspect2`, `polarity`, `pos` eta `modality` atributuak hartzen dituzte. Jarraian hauetako bakoitzaren azalpena egiten dugu.

- `class`: Esanahiaren araberako sailkapena 1.1 azpiataleko kategorizazioa erabilita.
- `tense1`: Orainaldikoa edo ez.
- `tense2`: Iraganekoa edo ez.
- `aspect1`: Burutua edo burutugabea.
- `aspect2`: Etorkizunekoa edo ez.
- `polarity`: Positiboa edo negatiboa. Sintaktikoki erabakitzen da, ez semantikoki. Ezeztatuta dauden gertaerak negatiboak dira eta gainerakoak positiboak.
- `pos`: Buru lexikalaren kategoria gramatikala.
- `modality`: Aditz modalen bitartez deskribatutako gertaeren buru lexikaletan aditzaren informazioa adierazteko.

Euskaraz, ingelesez ez bezala, `tense1`, `tense2`, `aspect1`, `aspect2` eta `modality` atributuak izatearen arrazoia ondorengoa da: euskaraz aldia, aspektua eta modalitatea, sintetikoki adierazten direla, eta ez indoeuropar hizkuntza gehienetan bezala aditz askeen bitartez. Aipatu nahi dugu, gainera, aditz denborak bi dimentsio dituela euskarazko *ISO-TimeML* eskeman (±`PRESENT` eta ±`PAST`), eta horiek uztartuz lortzen dela gertaeren denboraren anotazioa. (Altuna *et al.*, 2016) argitalpenean adierazten denez, tempusa markatzeko orduan aditz laguntzaileari begiratuko zaio aditz perifrastikoen kasuan, eta tempusa adierazten duen morfemari aditz trinkoenean. Aditzak ez diren predikatuen bitartez deskribatutako gertaeretan `tense1` eta `tense2` atributuek `NONE` balioa jasotzen dute, denborarik adierazten ez delako. Denborarekin gertatzen den bezala, aspektua ere bi dimentsioren bitartez adierazten da euskarazko *ISO-TimeML* eskeman (±`PERFECT` eta ±`FUTURE`). Xehetasun gehiago euskarazko *ISO-TimeML* anotaziorako gidalerroetan aurki daitezke[1].

2.2 Antzeko tresnak eta *ISO-TimeML* eskemaren jatorria

Azkeneko hamarkadan testuetako informazio tenporalaren anotazioaz egin den lan gehiena *TimeML* (Pustejovsky *et al.*, 2003) eta *ISO-TimeML* markaketa lengoaien inguruan burutu da. *TimeML* 2003. urtean sortu zen *TERQAS* (Pustejovsky, 2002) izeneko proiektuaren ondorioz. Hau galdera-erantzun sistemen eraginkortasuna hobetzeko helburuarekin hainbat ikerketa lan bultzatu zituen *ARDA Aquaint* programak finantzatu zuen. *ISO-TimeML*, aldiz, 2010. urtean aurkeztu zen eta *TimeML* lengoaiaren hobetutako bertsio interoperagarritzat hartzen da. *TimeML* lengoaiaren oinarriak bi izan ziren: *TIDES TIMEX2* denbora adierazpenen anotaziorako gidalerroak (Ferro *et al.*, 2001) eta Setzer (2001) tesian proposatzen den egunkari berrietako informazio tenporalaren anotaziorako lengoaia. *ISO-TimeML* lengoaiak, *TimeML*ren bertsio hobetua denez gero, denbora adierazpenak *TIMEX3* gidalerroei (Pustejovsky *et al.*, 2005) jarraituta anotatzen ditu. Hauek *ACE TIMEX2* gidalerroen (Ferro *et al.*, 2003) bertsio hedatua dira. *ACE TIMEX2* gidalerroak, bestalde, aipatu ditugun *TIDES TIMEX2* gidalerroen hobekuntza dira.

2003. urtean *TimeML* eskema aurkeztu zenetik hainbat sistema garatu dira denbora informazioaren etiketatze automatikoa egiteko, gehienak ingeleserako. Guk dakigula ez da sekula izan *TimeML*/*ISO-TimeML* gidalerroetako osagai tenporal mota guztiak etiketatzen dituen eta hizkuntza bat baino gehiago prozesatzeko gaitasuna daukan sistemarik. Gehienak, denbora adierazpenak besterik etiketatzen ez dituzten tresnak dira. Ingelesa ez diren hizkuntzak prozesatzeaz arduratzen diren etiketatzaileen adibideak EVENTI ebaluazio saioan (Caselli *et al.*, 2014) parte hartu zuen italierarako *Fbkhlt-time* sistema (Mirza eta Minard, 2014), eta *TempEval-2* (Verhagen *et al.*, 2010) eta *TempEval-3* (UzZaman *et al.*, 2012) ebaluazio saioetan parte hartu zuen gaztelerarako *TipSem* sistema (Llorens *et al.*, 2010) dira.

3 *bEVENT* etiketatzailea

Lan honen hasieran esan dugu *bEVENT* euskaraz idatzitako testuetan gertaerak eta hauen propietateak *ISO-TimeML* eskemaren arabera etiketatzeko garatu den lehenengo sistema dela. Etiketatze prozesu

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

hau modu guztiz automatikoan egiten da *Euskal-TimeBank* corpusa oinarritako hartzen duten ikasketa automatikoko teknikak erabilita. *bEVENT* tresna, gainera, testu soiletik abiatzen da eta irteera bezala *ISO-TimeML* formatuko fitxategiak itzultzen ditu. Azpiatal honetan gertaera etiketatzailearen diseinua azaltzen da eta horretarako honen barneko sailkatzaileak, hauek eraikitzeko inplementatutako ezaugarri linguistikoak eta *Euskal-TimeBank* corpusa aurkezten dira.

3.1 Diseinu orokorra eta ezaugarri linguistikoak

Gertaerak identifikatzeko garaian *bEVENT*ek buru lexikalak besterik ez ditu hartzen kontuan, ez gertaerak osatzen dituzten token (hitz) guztiak. Hau etiketatzailearen eraikuntzarako erabili den *Euskal-TimeBank* corpusean gertaerak era honetara anotaturik daudelako egiten da horrela. Buru lexikalak token bakarrekoak izaten direnez gero, gertaerak identifikatzeko sailkatzaile bitarra erabiltzen da. Honek sarreratzat jasotako testuko token bakoitza gertaera baten buru lexikala den edo ez erabakitzen du. Euskaraz gertaerek hartzen dituzten zortzi atributuen etiketatzea dela eta (ikus 2.1), hauetan ere ikasketa automatikoa, eta beraz, sailkatzaileak, erabili dira. Bi balio bakarrik har ditzaketen atributuentzat, identifikaziorako bezala, sailkatzaile bitarrak erabili dira. Gainerakoentzat, berriz, *multiclass* motakoak. Sailkatzaileen eraikuntzarako baliatutako algoritmoa *Support Vector Machines-SVM* (Cortes eta Vapnik, 1995) izan da, hizkuntzaren prozesamenduko ataza askori algoritmo hau ongi egokitzen zaiela jakina delako. Sailkatzaileak garatzeko jarraian aurkezten diren ezaugarri linguistikoak erabiltzen dira.

- **Lexikalak**: Forma eta lema uneko tokenarentzat, aurreko tokenarentzat eta uneko tokenaren aditz gobernatzailearentzat.
- **Part-of-Speech**: *PoS* kategoria *estandarra* eta euskarazko *PoS* kategoria eta azpikategoria uneko tokenarentzat, aurreko tokenarentzat eta uneko tokenaren aditz gobernatzailearentzat.
- **Sintaktikoak**: Uneko tokenetik esaldiaren erro sintaktikorainoko tokenen lemek, dependentzia sintaktiko motek, *PoS* kategoria *estandarrek*, euskarazko *PoS* kategoriek eta azpikategoriek osatutako multzoak. Gainera, dependentzia sintaktiko mota uneko tokenaren eta zuhaitz sintaktikoan honen gurasoa den tokenaren artean, eta uneko tokenaren aditz gobernatzailearen eta zuhaitz sintaktikoan honen gurasoa den tokenaren artean.
- **Semantikoak**: Uneko tokenaren rol semantikoa eta uneko tokenetik esaldiaren erro semantikorainoko tokenei dagozkien rol semantikoek osatutako multzoa.

Ezaugarri hauetako batzuk beste hizkuntzetan denbora etiketatzen duten sistemak eraikitzeko erabili izan dira (Jung eta Stent, 2013). Beste batzuk, ordea, *Part-of-Speech* azpikategoria adibidez, lehenengo aldiz inplementatu dira ataza honetan. Dokumentuetako hitzen tokenizaziorako, lematizaziorako eta *Part-of-Speech* kategoriak etiketatzeko *Eustagger* tresna (Alegria *et al.*, 2002) erabiltzen da. *Eustagger* bi-mailatako formalismoan oinarrituta sortu zen estaldura zabaleko analizatzaile morfologikoa da. Dokumentuen analisi sintaktiko eta semantikorako, berriz, euskararako garatu zen *bRol* dependentzia *parser*a erabiltzen da (Salaberri *et al.*, 2015). *bRol* gaurdaino euskararako eraiki den mota honetako sistema bakarra da, eta bi osagai nagusi dauzka: dependentzia sintaktikoak ezartzen dituen eta bigarren ordenakoa den grafoetan oinarritutako *MATE* analizatzaile sintaktikoa (Bohnet, 2010) batetik, eta dependentzia semantikoak ezartzen dituen osagaia bestetik.

3.2 Euskal-TimeBank corpusa

Euskal-TimeBank da gaur egun euskararako denbora informazioarekin anotatuta dagoen corpus bakarra eta horregatik erabili dugu *bEVENT* garatzeko. *Euskal-TimeBank* eraikitzeko *MEANTIME* corpusaren (Minard *et al.*, 2016) euskarazko bertsiotik hartutako 30 dokumentu *ISO-TimeML* gidalerroen egokitzapenaren arabera anotatu ziren. *MEANTIME* corpusa *NewsReader* proiektuan (Agerri *et al.*, 2014) erabili zen, eta semantikoki anotatutako (SRL, NER, etab.) *Wikinews*[2] berriek osatzen dute. *Euskal-TimeBank* corpusaren garapena bi fasetan egin zen, ingeleseko *TimeBank*en garapen prozesuan oinarrituta. Lehenik, gertaera eta denbora adierazpen kopuru mugatu bat eskuz anotatu zen euskarazko *ISO-TimeML* eskemaren lehenengo bertsioa jarraituta. Bigarrenik, gidalerroak eguneratu egin ziren, anotazioen analisi gramatikala eginda, eta analisi honen ondorioz sortutako gidalerroen bigarren bertsioarekin corpuseko 30

[2] `https://www.wikinews.org/`

dokumentuak osorik anotatu ziren. (Altuna *et al.*, 2016) argitalpenean adierazten denez, hiru pertsonak hartu zuten parte lehenengo fasean. Hauen arteko adostasuna (*Inter-Annotator Agreement-IAA* delakoa) neurtzeko erabiltzen den *Dice*-en koefizientean (Dice, 1945), anotatzaile bikotearen arabera, 0.864 eta 0.947 bitarteko batezbesteko balioak iritsi ziren. Bigarren fasean, aldiz, lau pertsonak hartu zuten parte, eta 0.812 eta 0.883 bitarteko batezbesteko adostasuna lortu zen.

3.3 Emaitzak

bEVENT sistemaren barnean, gertaeren etiketatzeari dagozkien emaitzak lortzerakoan, esperimentuak egin ditugu hainbat neurritako testuinguru leihoekin. Hain zuzen ere, erabilitako leihoen neurriak, bat, hiru, zazpi eta hamabost hitzetakoak izan dira. Honekin adierazi nahi dugu *bEVENT*ek prozesatzen duen token bakoitzarentzat ezaugarriak (3.1 azpiatalekoak) tokenarentzat berarentzat erauzi ditugula, baita haren eskuinetan eta ezkerretan dauden beste zero, bat, hiru eta zazpi tokenentzat ere. Honek badu eragina prozesatutako token bakoitzarentzat erauzitako ezaugarri kopuruan. Hitz bakarreko leihoentzat 24 ezaugarri erauzi dira (tokenarenak berarenak), hirukoentzat 72, zazpikoentzat 168 eta, azkenik, hamabostekoentzat, 360 ezaugarri. 1 taulan aurkezten ditugu *Euskal-TimeBank* corpusaren ebaluaziorako zatia erabilita lortutako emaitzak (*Train-Test* motako ebaluazioa). Taula honetan biltzen direnak *bEVENT* sistemaren azpiataza bakoitzean konfiguraziorik egokienarekin, hau da, baliorik altuenak itzuli dituen testuinguru leihoaren tamainarekin lortutako emaitzak dira.

	ATAZA	LEIHORIK EGOKIENA	PREZISIOA	ESTALDURA	F_1
<EVENT>	`id`	1	83.92	72.76	77.94
	`class`	1	-	-	58.49
	`tense1`	3	-	-	62.24
	`tense2`	3	-	-	63.24
	`aspect1`	3	-	-	63.53
	`aspect2`	3	-	-	64.10
	`polarity`	1	-	-	75.55
	`pos`	1	-	-	70.05
	`modality`	1	-	-	74.14

1 Taula: *bEVENT* etiketatzailearentzat lortutako emaitzak.

4 Ondorioak

Azpiatal honetan *bEVENT* sistemarentzat lortutako emaitzak analizatzen ditugu. Hasteko, eta gure ustez, bi gauza nabarmen ikus daitezke lehen begiratuan 1 taulan: gertaeren identifikazioan emaitzarik onenak itzultzen dituen testuinguru leihoaren tamaina bat dela eta haien atributuak etiketatzeko leihorik egokiena bat edo hiru hitzekoa dela. Lehenbiziko puntuaren inguruan, esan beharra daukagu gure emaitzak ez datozela bat (Jung eta Stent, 2013) argitalpenean ingeleserako aurkeztutakoekin. Izan ere, bertan, hamabost eta zazpi hitzeko testuinguru leihoekin hitz bateko eta hiru hitzeko leihoekin baino emaitza hobeak lortu zirela azaltzen da. Kontuan izan beharra dago aipatu argitalpenean *TempEval-3* (UzZaman *et al.*, 2012) ebaluazio saioan parte hartu zuen `att` sistema eta hau erabilita egindako testuinguru leihoen inguruko esperimentuen emaitzak aurkezten direla. Argitu behar da *TempEval-3* ebaluazio saioko corpusetan ere, *Euskal-TimeBank* corpusean bezala, token bakarreko gertaeren buru lexikalak daude anotatuta. Alde horretatik, beraz, ez dago arrazoirik euskaraz bat eta hiru tokeneko testuinguru leihoek zazpikoek eta hamabostekoek baino emaitza hobeak itzultzeko. Gure iritziz, *bEVENT* sistemaren kasuan, bat eta hiru hitzeko leihoekin balio altuagoak eskuratzeko arrazoi nagusia ondorengoa da: *bEVENT* barneko sailkatzaileak sortzeko erabili diren ezaugarrietako asko (forma, lema, uneko tokenetik esaldiaren erro sintaktikorainoko tokenen formak, etab.) *string* motakoak dira, eta *Euskal-TimeBank* corpusaren tamaina mugatuaren ondorioz ezaugarri hauetako askok hartzen dituzten balioak oso gutxitan errepikatzen dira entrenamendu corpusean zehar. Daitekeena da, beraz, ezaugarri horiek jasotzen dituzten balioetako batzuk corpus guztian behin baizik ez agertzea. Jakina da balio ezberdin asko edota gutxitan errepikatzen diren balioak dauzkaten ezaugarriek *zarata* ateratzen dutela ikasketa automatikoaren prozesuan. *Zarata* honek eragin negatiboa du sailkatzaileen eraginkortasunean eta, ondorioz, baita sistemaren emaitzetan

ere. Testuinguru leiho txikiak erabilita eragin negatibo hau sortzen duten ezaugarrien kopurua gutxitu egiten da.

Bigarren puntuan esan dugun bezala, gertaeren atributuak etiketatzean emaitzarik onenak itzultzen dituzten testuinguru leihoen tamainak bat eta hiru hitzekoak dira. Atributuek gertaeren izaera gramatikala adierazten dute. Izan ere, 1.1 azpiatalean azaldu dugunez, gertaerek denborarekin duten erlazio semantikoaren kategorizazioa aldatzen duten gramatikaren kategoriak bat baino gehiago dira (aspektu lexikala, aspektua, modua, denbora gramatikala, etab.), eta lan honetan jarraitzen dugun *ISO-TimeML* eskeman atributuen bitartez daude adierazita kategoria hauek. Hau kontuan edukita uste dugu atributuen esleipenak bat eta hiru hitzeko leihoekin emaitzarik onenak itzultzeko arrazoia gertaeren izaera gramatikala finkatu ahal izateko beharrezkoa den testuinguru linguistikoa osatzen duten token kopuruaren ondorio dela. Hau da, atributuak finkatu ahal izateko garrantzizkoa dela gertaera barnean hartzen duen sintagmari, bertako tokenei eta orokorrean hauek osatzen duten egitura sintagmatikoari, erreparatzea. Gertaera hauek askotan hiru edo lau tokenez osatutako sintagmen zati izaten dira. Euskaraz aditz forma perifrastikoaren erabilera (aditza eta aditz laguntzailea) oso arrunta da, eta, gainera, hitz bakarreko aditz trinkoak ere erabiltzen dira gertaerak deskribatzeko. Sintagma hauek eta aditz trinkoek gertaeren izaera gramatikala (atributuak) finkatu ahal izateko beharrezkoa den informazio linguistikoa eskaintzen dute eta horregatik lortzen dira, gure ustez, emaitzarik onenak bat eta hiru hitzetako testuinguru leihoekin.

5 Etorkizuneko lanak

Etorkizunerako planteatzen dugun norabidean 2.1 azpiatalean aipatutako eginbeharra dago, hau da, gertaerez gainera denbora adierazpenak (`<TIMEX3>`) eta seinaleak (`<SIGNAL>`) ere automatikoki etiketatu ahal izatea. Gustatuko litzaiguke, bestalde, testuinguru leihoekin egin dugun bezala, 3.1 puntuan aurkeztu ditugun ezaugarriekin ere hauen eragina azpiataza bakoitzean zein den neurtu ahal izatea. Uste dugu neurketa honetatik atributuei hobekien egokitzen zaizkien ezaugarri multzoak zein diren ondoriozta genezakeela. Honek, sistemaren emaitzak hobetu ahal izatea eragingo luke. Bukatzeko, beste hizkuntza batzuenekin (ingelesekoekin, gaztelerakoekin edota italierakoekin) gureak, euskararenak, alderatzea ere gustatuko litzaiguke.

Erreferentziak

Agerri, Rodrigo, Eneko Agirre, Itziar Aldabe, Begona Altuna, Zuhaitz Beloki, Egoitz Laparra, Maddalen López de Lacalle, German Rigau, Aitor Soroa, eta others. 2014. Newsreader project. *Procesamiento del Lenguaje Natural* 53.155–158.

Alegria, I, M Aranzabe, A Ezeiza, N Ezeiza, eta R Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for basque. In *Proceedings of Workshop on"Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation*.

Altuna, Begoña, María Jesús Aranzabe Urruzola, eta Arantza Díaz de Ilarraza Sánchez. 2016. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2. 0.

Badiou, Alain, eta Oliver Feltham. 2007. *Being and event*. A&C Black.

Bohnet, Bernd. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, 89–97. Association for Computational Linguistics.

Caselli, Tommaso, Rachele Sprugnoli, Manuela Speranza, eta Monica Monachini. 2014. Eventi: Evaluation of events and temporal information at evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, 27–34. Pisa University Press.

Cortes, Corinna, eta Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20.273–297.

Davidson, Donald. 1967. The logical form of action sentences.

Deleuze, Gilles. 1988. Signes et événements. *Magazine littéraire* 257.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26.297–302.

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, eta George Wilson. 2003. Tides standard for the annotation of temporal expressions. *200309)[2005-06-10]. http://timex2. mitre, org* .

——, Inderjeet Mani, Beth Sundheim, eta George Wilson. 2001. Tides temporal annotation guidelines-version 1.0.2. *The MITRE Corporation, McLean-VG-USA* .

Jung, Hyuckchul, eta Amanda Stent. 2013. Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, 20–24.

Kim, Jaegwon. 1976. Events as property exemplifications. In *Action theory*, 159–177. Springer.

Lewis, David Kellogg. 1987. *Philosophical Papers: Volume Il*. Oxford university press.

Llorens, Hector, Estela Saquete, eta Borja Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 284–291. Association for Computational Linguistics.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, eta Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. *Proceedings of LREC2016* .

Mirza, Paramita, eta Anne-Lyse Minard. 2014. Fbkhlt-time: a complete italian temporal processing system for eventi-evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Parsons, Terence. 1990. Events in the semantics of english: A study in subatomic semantics.

Pustejovsky, James. 2002. Terqas: time and event recognition for question answering systems. In *ARDA Workshop*.

——, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, eta Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering* 3.28–34.

——, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, eta Inderjeet Mani. 2005. The specification language timeml. *The language of time: A reader* 545–557.

——, Kiyong Lee, Harry Bunt, eta Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*.

Salaberri, Haritz, Olatz Arregi, eta Benat Zapirain. 2015. brol: The parser of syntactic and semantic dependencies for basque. 555–562.

Saurii, Roser, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, eta James Pustejovsky, 2005. Timeml annotation guidelines.

Setzer, Andrea, 2001. *Temporal information in newswire articles: an annotation scheme and corpus study*. University of Sheffield Sheffield, UK tesia.

Streitberg, Wilhelm. 1891. Perfective und imperfective actionsart im germanischen. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)* 1891.70–177.

Tenny, Carol, eta James Pustejovsky. 2000. Events as grammatical objects the converging perspectives of lexical semantics and syntax.

UzZaman, Naushad, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, eta James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333* .

Verhagen, Marc, Roser Sauri, Tommaso Caselli, eta James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, 57–62. Association for Computational Linguistics.