



IKER
GAZTE
NAZIOARTEKO
IKERKETA EUSKARAZ

I. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK

**Euskarazko denbora-egituren
tratamendu automatikorako
azterketa**

Begoña Altuna

46-53 or.
<https://dx.doi.org/10.26876/ikergazte.i.05>

ANTOLATZAILEA:



BABESLEAK:



LAGUNTZAILEAK:



Euskarazko denbora-egituren tratamendu automatikorako azterketa

Altuna, B.

Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU). begona.altuna@ehu.eus

Laburpena

Hizkuntzaren prozesamenduan (HP) balio handikoa da denbora-informazioa, horrek kokatzen baititu testuko gertaerak kronologian edo horien iraupena adierazten baitu. Erauzitako informazio hori gertaerak noiz jazo diren jakiteko baliatu ahal izango da, baita gertaerak denboran zehar ordenatu edo etorkizunean zer gertatuko den aurrikusteko. Horiek lortzeko, ordea, denbora-egiturak eta euren ezaugarriak identifikatu eta tresna informatikoentzat atzigarri egin behar dira eta zeregin horietarako hizkuntza-baliabideak eta konputazio-baliabideak sortu behar dira.

Hitz gakoak: Denbora-informazioa, hizkuntzaren prozesamendua, corpora, markaketa-lengoaia

Abstract

Temporal information is very valuable on natural language processing (NLP), since it positions the events in the text on a timeline or shows their duration. That information will be employed to know when events have happened, as well as to order events in a timeline or forecast future events. So as to achieve those goals, temporal structures and their features have to be identified and available for automatic tools, and for these tasks linguistic and computational resources have to be developed.

Keywords: Temporal information, natural language processing, corpus, mark-up language

1 Sarrera

Azken urteetan denbora-informazioaren analisia ikerketa ildo nabarmena bilakatu da hizkuntzaren prozesamenduaren (HP) esparruan; esaterako, TempEval lehiaketan ((Verhagen *et al.* 2007, 2010) eta (UzZaman *et al.*, 2012)) helburua informazio hori prozesatzeko sistemak sortzea izan da, denbora-egiturak edo erlazioak erauzteko sistemak, adibidez. Denbora-informazioa erauzteko, testuko gertaerak identifikatu behar dira, hala nola denborako uneak eta tarteak adierazten dituzten egiturak eta horiek erlazionatuko dituzten erlazio-egiturak. Testuko denbora-informazioa prozesatzeko, besteak beste, gertaerak denboraren arabera antolatzea ahalbidetuko du, baita denborazko galderei erantzutea ere informazio-erauzketa sistemetan. Denbora-informazio prozesatu hori baliatuko duten tresna informatikoei aukera eskainiko diote erabiltzaileari, adibidez, bilaketa aurreratuagoak galdera zuzenen bidez egiteko, enpresa bati buruzko iritziak noizkoak diren automatikoki identifikatzeko edo historia klinikoetan kausa-ondorio erlazioak bilatzeko.

Hiztun-komunitate bakoitzak era batera kontzeptualizatzen du denbora eta hala islatzen da erabiltzen dituen egituretan. Adibidez, euskaldunok *goizaldea* eguneko ataltzat hartzen dugu, baina frantsesek *gaue-tik* (nuit) *goizera* (matin) egiten dute jauzia zuzenean. Horregatik, hizkuntza bakoitzean denbora-egituren analisia egin behar da. Ingelesezt (Pustejovsky *et al.*, 2003a), frantsesezt (Bittar, 2010), italieraz (Caselli *et al.*, 2011), errumanieraz (Forăscu eta Tufiş, 2012), etab. lan eskerga egin da denbora-informazioaren prozesamenduan eta, euskarazko denbora-egituren analisisian dihardugu egun, erdaretan bezala, denbora-informazioa prozesatu eta baliatu ahal izateko.

Euskarazko denbora-egituren azterketa eta corpusaren osaketa tesi-lana ildo horretan kokatzen da. IXA

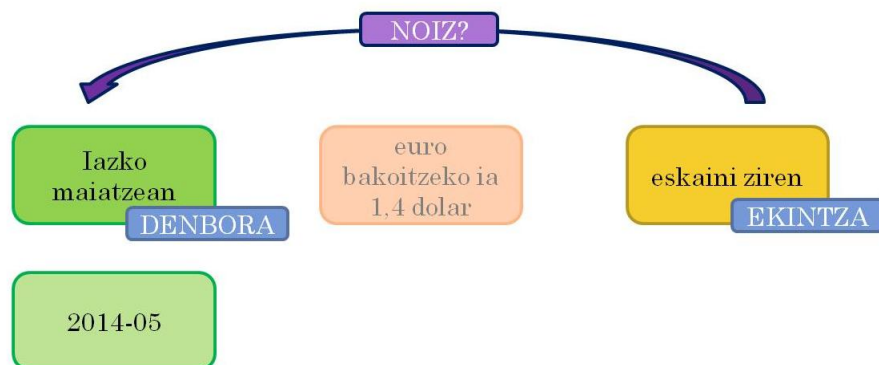
ikerketa-taldearen¹ baitan egiten ari garen ikerketa da. Lanak helburu izango ditu euskarazko denbora-egituren analisia, denbora-informazioa batuko duen corpus baten sorrera eta denbora-informazio hori prozesatuko duten tresnen garapena. Horretarako, nazioarteko NewsReader² proiektuan egin dugun lana euskararako oinarritzat hartuko da eta bai IXA ikerketa-taldean bai Italiako Fondazione Bruno Kesslerren³ existitzen diren tresnak euskarazko denbora-informazioa prozesatzeko egokituko dira edo berriak sortuko dira.

Esan bezala, hizkuntza bakoitzak bere denbora-egitura propioak ditu eta horiek hizkuntzaren arabera aztertu behar dira. Jarraian, (1) adibideko testu-zatian ageri diren euskarazko denbora-egitura batzuk nabarmendu eta horien deskribapena egiteko baliatu ditugu:

- (1) Euroarekin eta Japoniako yenarekin **alderatuz**, hurrenez hurren, % 13 eta % 15 **igo da** billete berdearen **kotizazioa** *urtearen joanean*. *Iazko maiatzean* euro bakoitzeko ia 1,4 dolar **eskaini ziren**, eta *orain* 1,2 dolarren azpitik **mugitzen hasi da**. *Azken bederatzi urtetako* baliorik txikiena. (*Berria*, 2015-01-07)

Goiko (1) adibidean beltzez nabarmendu dira gertaerak, “gertatzen diren ekintzak eta egoerak”. Horiek nagusiki aditzen (*alderatuz*, *igo da*, *eskaini ziren*, *mugitzen* eta *hasi da*) bidez adierazten dira, baina izenez (*kotizazioa*), adjektiboz edo adberbioz ere adieraz daitezke. Beren aldetik, kurtsibaz agertzen diren denbora-adierazpenak “denborako une edo tarte bat adierazten duten esamoldeak” dira eta izen sintagma (*urtearen joanean*, *iazko maiatzean* eta *azken bederatzi urtetako*) edo adberbio (*orain*) bidez gauzatzen dira batik bat. Aurreko denbora-egitura horiek detektatzeaz gain, egitura horien artean gauzatzen diren erlazioak ere esplizitu egin behar dira. 1. irudian ikus daitekeen moduan, “noiz?” galdetzailearen bidez adierazi da erlazio hori.

1 Irudia: Denbora-egiturak eta erlazioak
Iazko maiatzean euro bakoitzeko ia 1,4 dolar eskaini ziren.



Kronologiako unea, gertaera eta horien arteko erlazioa identifikatzeko prozesu hori automatikoki egin ahal izateko, hizkuntzaren prozesamendua egin behar da. Denbora-informazioaren analisi eta prozesamendua egiteko hainbat baliabide nagusi garatu behar dira: denbora-informazioa etiketatzeko markaketa-lengoaia (Pustejovsky *et al.*, 2003a) eta horren arabera etiketatzeko gidalerroak (TimeML Working Group, 2010) eta ikasketa automatikoko algoritmo eta corpus etiketatuak (Pustejovsky *et al.*, 2006). Jarraian ikusiko dira esparru bakoitzean egin diren aurrerapenak.

1.1 Markaketa-lengoiak eta gidalerroak

Denbora-informazioaren prozesamenduan egindako lehenengo saiakerak MUC (Message Understanding Conferences) konferentzietan (Grishman eta Sundheim, 1996) egin ziren eta horien helburua denbora-adierazpenak identifikatzea zen. Gero TIMEX2 markaketa-lengoaia (Ferro *et al.*, 2003) heldu zen, SGML-

¹<http://ixa.si.ehu.es/Ixa>

²<http://www.newsreader-project.eu/>

³<http://http://www.fbk.eu/>

ren⁴ bidez denbora-adierazpenak eta denborazko erlazio-esamoldeak markatzeko. Gertaerak etiketatzen lehena Setzer (2001) izan zen eta bere lana gaur egun estandar bihurtu den XMLn⁵ oinarritutako TimeML (Pustejovsky *et al.*, 2003a) markaketa-lengoiaren aitzindaria da. Egun TimeMLk du erabilerarik orokortuena eta hainbat hizkuntzatar moldatu da: frantsesa Bittar (2010), koreera Im *et al.* (2009), italiera Caselli *et al.* (2011), besteak beste.

TimeMLk etiketa sorta bat definituta du denbora-egiturak eta horien arteko erlazioak markatzeko:

- Denbora-egiturak:
 - EVENT. Gertaerak (ekintzak eta egoerak) etiketatzen.
 - TIMEX3. Denbora-adierazpenak, denborako une edo iraupenak adierazten dituzten esamoldeak, etiketatzen.
 - SIGNAL. Denborazko erlazio-hitz edo esamoldeak etiketatzen.
- Erlazioak:
 - TLINK. Gertaera eta denbora-adierazpenen arteko erlazioak markatzeko.
 - SLINK. Gertaera nagusiaren eta bere mendeakoaren arteko erlazioa markatzeko.
 - ALINK. Aspektuzko erlazioak markatzeko.

Etiketa horiek, halaber, denbora-egitura edo erlazioaren ezaugarriak nabarmenduko dituzten atributuak hartzen dituzte; esaterako, egituraren mota eta balio kronologikoa adierazten dutenak.

Eskuzko etiketatzerako, arestian aipatutako etiketa eta atributuen deskribapen eta azalpen linguistikoa ematen duten etiketatze-gidalerroak sortu dira. TimeMLren ingelesezko gidalerroen bertsiorik berriena 2010ekoa da eta ingelesezko denbora-egitura eta erlazioak markatzeko jarraibideak agertzen dira bertan, baita etiketatze-adibideak eta salbuespenak ere.

1.2 Corpusak

Denborazko markaketa-lengoiak baliatuz corpusak etiketatu dira, TimeBank 1.1 (Pustejovsky *et al.*, 2003b) eta 1.2 (Pustejovsky *et al.*, 2006) corpusak edo WikiWars (Mazur eta Dale, 2010), nagusienak aipatzearen. Corpus horiek ikasketa automatikorako eta *gold standard* izateko sortu dira. Lehenak albisteak batzen ditu eta TimeML jarraituz etiketatu da. Bigarrenak, ordea, testu historiko luzeak batzen ditu, horietan denbora-adierazpen eta erlazioak albiste laburretakoak baino konplexuagoak baitira. WikiWarsen denbora-adierazpenak baino ez dira etiketatu TIMEX2 markaketa-lengoiak baliatuz.

1.3 Tresna automatikoak

Denbora-informazioaren prozesamendurako tresna automatikoen baitan bi multzo egin behar dira; alde batetik, denbora-egiturak eta euren ezaugarriak identifikatzen eta etiketatzen dituzten tresnak eta, bestetik, denbora-informazio hori testuaren ulermenerako baliatzen duten tresnak.

Lehenengo multzoan, esaterako, etiketatzeko tresna automatikoak (Verhagen *et al.*, 2005), denbora-informazioa erauzteko tresnak (Strötgen eta Gertz, 2010), denbora normalizatzeko tresnak (Llorens *et al.*, 2012) edota gertaerak detektatzeko tresnak (Yaghoobzadeh *et al.*, 2012) aurki daitezke. Bestalde, bigarren multzoan, gertaerak iragartzeko tresnak (Radinsky eta Horvitz, 2013), etorkizuna antzemateko tresnak (Kawai *et al.*, 2010) eta denbora-lerroak sortzeko tresnak (Bauer *et al.*, 2014) daude besteak beste.

2 Ikerketaren helburuak

Gure tesi-lanaren helburu nagusia denbora-informazioaren tratamendu automatikoa egitea da, euskararen prozesamenduan urrats bat aurrera egiteko. Gure helburu nagusia lortzeko ondoko helburu zehatz hauek definitu ditugu:

⁴Standard Generalized Markup Language

⁵Extensible Mark-up Language

- Euskarazko denbora-egituren azterketa eta sailkapena
- Euskarazko denbora-egiturak markatzeko etiketatze-eskemaren deskribaketa
- Denbora-egiturak etiketatuta izango dituen corpusaren osaketa
- Euskarazko denbora-egiturak markatzeko tresnen egokitzapena
- Garatutako tresna automatikoen ebaluazioa

Helburu horiek lortzeko, ondoko atalean azalduko diren esperimentuak egin eta baliabideak garatu ditugu edo horiek sortzeko bidean ari gara.

Ingeles eta Europako beste hizkuntza batzuekin alderatuta, euskara baliabide gutxiko hizkuntza da, baina euskarazko testuen prozesamenduan aurrera eginez, IXA ikerketa-taldean baliabide sorta hori etengabe aberasten ari gara. Euskara hizkuntza eranskaria izanik, inguruko hizkuntzetan hartu diren prozesamendu-erabakiak horrentzat egokiak diren aztertu ahal izango dugu eta, koreerarekin batera (Im *et al.*, 2009), adibidez, baliabide gutxiago duten beste hizkuntza eranskari batzuentzat ere eredu izaten saiatuko gara.

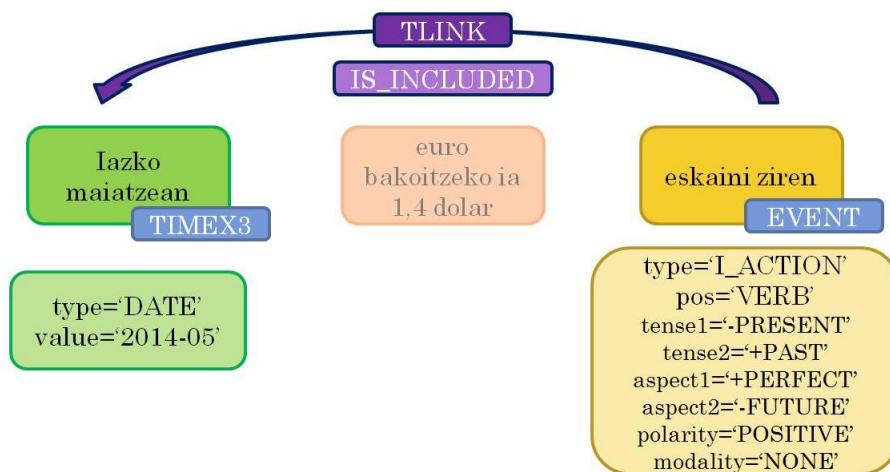
3 Ikerketaren muinak

Euskarazko denbora-informazioaren prozesamendua egiteko baliabideak sortzen ari gara. Lehenik eta behin, euskarazko denbora-egituren informazioa kodetzen duen markaketa-lengoaia osatzen ari gara. Markaketa-lengoaia hori baliatuta albiste ekonomikoak osatutako corpus bat eskuz etiketatzen ari gara, *gold standarda* sortzeko, etorkizunean tresna automatikoak ebaluatu ahal izateko. Jarraian datozen azpiataletan azalduko da egindako lana.

3.1 EusTimeML eta gidalerroak

Sarreran aipatu bezala, denbora-informazioa esplizitu egiteko markaketa-lengoaia bat baliatu behar da. Markaketa-lengoaia horretan denbora-egitura bakoitzari identifikatzen duen etiketa bat ematen zaio eta denbora-egituraren ezaugarriak atributu bidez adierazten dira. Euskarako TimeML markaketa-lengoaia jarraitu da, beste hainbat hizkuntzatan (ingelese, italiara, frantsesa, koreera, errumaniera, etab.) egin den bezala. Halaber, NewsReader proiektuko etiketatze-gidalerroak (Tonelli *et al.*, 2014) eredutzat hartu dira. Euskararako atributu eta balio batzuk egokitu behar izan ditugu, denbora-egituren ezaugarriak ez baitira berdinak hizkuntza guztietan. Hala, EusTimeML eratzen ari gara eta horretan oinarrituta etiketatze-gidalerroak (Altuna *et al.*, 2014b) idatzi ditugu (Altuna *et al.*, 2014b).

2 Irudia: Denbora-egitura eta erlazioen analisisia
Iazko maiatzean euro bakoitzeko 1,4 euro eskaini ziren.



2. irudian ikus daiteke sarrerako (1) adibideko gertaera bat eta denbora adierazpen bat nola etiketatu diren. Denbora-adierazpenek <TIME3> etiketa hartzen dute eta gertaerek <EVENT>. Etiketekin batera, horien atributu esanguratsuenak adierazi dira; denbora-adierazpenaren kasuan, mota (**type**) eta balioa (**value**) eta gertaeraren kasuan mota (**type**), kategoria (**pos**), *tempusa* (**tense1** eta **tense2**) eta aspektua (**aspect1** eta **aspect2**), polaritatea (**polarity**) eta modalitatea (**modality**). Gertaera eta denbora-adierazpenaren arteko denbora-erlazioak <TLINK> etiketa hartu du eta gertaera denborako une horretan jazo dela ikus dezakegu.

3.2 Esperimentazioa

Esperimentazioa hainbat fasetan gauzatu da. Lehen fasean denbora-adierazpenak eta denborazko erlazio-esamoldeak etiketatu dira. Hiru anotatzailek ekonomiaren inguruko 4 albiste, 56 esaldi orotara, baliatu dituzte denbora-adierazpenak eta erlazio-esamoldeak etiketatzeko. Lan hori baliatuta, gidalerroen argitasuna neurtu eta kasuistikaren deskribapen zabala eginda zegoen ebaluatu da (Altuna *et al.*, 2014a). Bigarren fasean gertaerak etiketatu dira eta hirugarren fasean, egituren arteko erlazioak etiketatzen ari gara. Fase bakoitzak eskuzko anotaziorako erabakiak hartzeko, gidalerroen egokitasuna ebaluatzeko eta *gold standard* corpusa sortzeko balio du. Esperimentazioan CELCT Annotation Tool⁶ (CAT) (Bartalesi Lenzi *et al.*, 2012) etiketatzeko tresna erabili da. Tresna hori NewsReader proiektuan erabili dugu eta oso egokia da denbora-informazioaren anotaziorako.

Ondoko 1. eta 2. tauletan ikus daiteke etiketatzaileen arteko adostasuna esperimentazioaren lehen fasean, denbora-adierazpen zein erlazio-esamoldeetarako. Denbora-adierazpenetan % 96tik gorako adostasuna lortu da (1. taula), ataza ondo definitu delako seinale; erlazio-esamoldeetan (2. taula), ordea, emaitza apalagoak lortu dira, gidalerroen zehaztasun eza dela kausa, eta ondorioz, gidalerroak birdefinitu behar izan dira.

1 Taula: Denbora-adierazpenen (TIME3) etiketatzaileen arteko adostasuna

Etiketatzailerikoteak	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
A – B	0.96	0.976	0.969	0.977
A – C	0.943	0.965	0.923	0.965
B – C	0.902	0.94	0.892	0.942
Guztira	0.935	0.96	0.928	0.961

2 Taula: Denbora erlazio-hitzen (SIGNAL) etiketatzaileen arteko adostasuna

Etiketatzailerikoteak	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
A – B	0.75	0.75	0.79	0.79
A – C	0.556	0.556	0.479	0.479
B – C	0.444	0.444	0.393	0.393
Guztira	0.583	0.583	0.554	0.554

3.3 EusTime Corpusa

EusTimeML baliatuta denbora-informazioa batzeko EusTime izeneko corpusa sortzen ari gara. Corpusak egun ekonomiaren esparruko 17 albiste ditu eta esperimentaziorako erabili da. Albiste guztiak gai berari buruzkoak dira, testu barruko zein testu kanpoko informazioaren kudeaketa eta, nagusiki, gertaeren ordena zehaztea baita etorkizuneko helburuetako bat. Egund, ez dugu corpus homogeneorik eskaintzerik, anotazio fase bakoitzean dokumentu sorta bat baino ez baita erabili.

⁶http://celct.fbk.eu:8080/CAT_WEB_APP/checkUsers

4 Ondorioak

Euskarazko denbora-egituren azterketan euskarak berezkoak dituen egiturak eta ezaugarriak identifikatu ditugu ekonomia corpusean. Etiketatzeko esperimentuan ikusi izan denez, ingelesezko TimeML eredua egokia da euskarazko denbora-informazioa adierazteko. Hala ere, egitura horiek adierazi ahal izateko euskararako etiketatze-irizpide berriak sortu behar izan ditugu, izan atributu edo balio berriak.

Egituren etiketen luzeran ere erabakiak hartu behar izan ditugu. Euskara hizkuntza eranskaria da eta TimeMLk token mailako etiketatzea proposatzen du. Ondorioz, ahalik eta informazio gehien nola adierazi adostu behar izan dugu. NewsReader proiektuan gaztelaniarako hartutako irizpide bat euskararako oso erabilgarri gertatu da, proiektu horretan token bati hainbat motatako etiketak esleitzea erabaki baita eta hori egin baita euskarazko etiketatzean ere. Hala, euskarazko denborazko menderagailudun aditz batek bi etiketa hatuko ditu (2) adibidean ikus daitekeenez:

(2) Menditik <SIGNAL><EVENT> zatozenez <\EVENT><\SIGNAL> eros ezazu, mesedez, ogia.

5 Etorkizunerako planteatzen den norabidea

Euskarazko denbora-egituren azterketa eta deskribaketa egin ondoren, hurrengo pausoa denbora-informazioaren prozesamenduan murgiltzea da, denbora-egiturak erazagutzen, lehenik eta behin; gero, euren arteko erlazioak automatikoki sortzen eta, azkenik, gertaerak denboran kokatzen eta ordenatzen. Horretarako, eskuragarri dauden baliabide eta tresnak gure ikerketan nola erabili aztertuko dugu, zuzenean aplikatu edo beharrezko moldaketak egiteko, eta bai IXA taldean euskararako egindako tresnak, bai beste hizkuntzetarako sortutakoak izango ditugu kontuan.

Euskaraz badira hainbat mailatan etiketatutako corpusak; Euskararen Prozesamendurako Erreferentzia Corpora (EPEC) (Aduriz *et al.*, 2006), berbarako, morfologia, sintaxi eta semantika mailetan etiketatua dago, eta horretan denbora-informazioa gehitzea izango da hurrengo urratsa. Denboraren arabera etiketatutako corpus handia sortzea da azken helburua, bere baitan izango duen informazioa tresna automatikoak trebatzeko erabili ahal izango baita.

Euskara hizkuntza eranskaria izanik, euskararako hartutako anotazio-erabakiak, prozesamendu erregelak eta abar beste hizkuntza eranskari batzuentzat baliabide erabilgarriak eskaintzea ere bilatzen jarraituko dugu gure lana nazioartean aurkeztuz.

Erreferentziak

- ADURIZ, ITZIAR, MARÍA JESÚS ARANZABE, JOSE MARI ARRIOLA, AITZIBER ATUTXA, ARANTZA DÍAZ DE ILARRAZA, NEREA EZEIZA, KOLDO GOJENOLA, MAITE OROÑOZ, AITOR SOROA, eta RUBEN 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 *Corpus Linguistics Around the World*, ed. by Andrew Wilson, Paul Rayson, eta Dawn Archer, 1–15. Amsterdam (Netherland): Rodopi.
- ALTUNA, BEGOÑA, MARÍA JESÚS ARANZABE, eta ARANTZA DÍAZ DE ILARRAZA. 2014a. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática* 6.13–24. <http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1>.
- , MARÍA JESÚS ARANZABE, eta ARANTZA DÍAZ DE ILARRAZA. 2014b. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Technical report, Lengoia eta Sistema Informatikoak Saila, UPV/EHU. UPV / EHU LSI / TR 01-2014. http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etikeatzeko%20gidalerroak.
- BARTALESI LENZI, VALENTINA, GIOVANNI MORETTI, eta RACHELE SPRUGNOLI. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis, 333–338, Istanbul, Turkey. European Language Resources Association (ELRA).
- BAUER, SANDRO, STEPHEN CLARK, eta THORE GRAEPEL. 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.

- BITTAR, ANDRÉ, 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Paris: Universit Paris Diderot tesia. <http://www.linguist.jussieu.fr/~abittar/docs/Bittar-PhD.pdf>.
- CASELLI, TOMASSO, VALENTINA BARTALESI LENZI, RACHELE SPRUGNOLI, EMANUELE PIANTA, eta IRINA PRODANOF. 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*, 143–151, Association for Computational Linguistics, Portland, Oregon, USA.
- FERRO, LISA, LAURIE GERBER, INDERJEET MANI, BETH SUNDHEIM, eta GEORGE WILSON. 2003. TIDES 2003 Standard for the Annotation of Temporal Expressions. Technical report, MITRE, McLean, USA. http://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf.
- FORĂSCU, CORINA, eta DAN TUFIŞ. 2012. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis (Eds.), 3762–3766, Istanbul, Turkey.
- GRISHMAN, RALPH, eta BETH SUNDHEIM. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 466–471, Center for Sprogteknologi, Copenhagen, Denmark.
- IM, SEOHYUN, HYUNJO YOU, HAYUN JANG, SEUNGHO NAM, eta 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*, 115–122, Suntec City, Singapore. Association for Computational Linguistics.
- KAWAI, HIDEKI, ADAM JATOWT, KATSUMI TANAKA, KAZUO KUNIEDA, eta KEIJI YAMADA. 2010. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication, ICUIMC '10*, 25:1–25:10, New York, NY, USA. ACM. <http://doi.acm.org/10.1145/2108616.2108647>.
- LLORENS, HECTOR, LEON DERCZYNSKI, ROBERT J GAIZAUSKAS, eta ESTELA SAQUETE. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis, 3044–3051, Istanbul, Turkey. European Language Resources Association (ELRA).
- MAZUR, PAWEŁ, eta ROBERT DALE. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, 913–922, Stroudsburg, PA, USA. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1870658.1870747>.
- PUSTEJOVSKY, JAMES, JOSÉ M CASTAÑO, ROBERT INGRIA, ROSER SAURÍ, ROBERT J GAIZAUSKAS, ANDREA SETZER, GRAHAM KATZ, eta DRAGOMIR R RADEV. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering* 3.28–34.
- , PATRICK HANKS, ROSER SAURÍ, ANDREW SEE, ROBERT GAIZAUSKAS, ANDREA SETZER, DRAGOMIR RADEV, BETH SUNDHEIM, DAVID DAY, LISA FERRO, eta MARCIA LAZO. 2003b. The TimeBank Corpus. In *Proceedings of Corpus Linguistics 2003*, ed. by Dawn Archer, Paul Rayson, Andrew Wilson, eta Tony McEnery, 647–656, Lancaster, UK. UCREL, Lancaster University.
- , MARC VERHAGEN, ROSER SAURÍ, JESSICA LITTMAN, ROBERT GAIZAUSKAS, GRAHAM KATZ, INDERJEET MANI, ROBERT KNIPPEN, eta ANDREA SETZER. 2006. TimeBank 1.2. Technical report, Linguistic Data Consortium. <https://catalog.ldc.upenn.edu/LDC2006T08>.
- RADINSKY, KIRA, eta ERIC HORVITZ. 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.
- SETZER, ANDREA, 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Sheffield, UK: University of Sheffield tesia. ftp://ftp.dcs.shef.ac.uk/home/robertg/theses/setzer_thesis.pdf.

- STRÖTGEN, JANNIK, eta MICHAEL GERTZ. 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, Stroudsburg, PA, USA. Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=1859664.1859735>.
- TIMEML WORKING GROUP. 2010. TimeML Annotation Guidelines version 1.3. Manuscript. Technical report, Brandeis University.
- TONELLI, SARA, RACHELE SPRUGNOLI, eta MANUELA SPERANZA. 2014. NewsReader Guidelines for Annotation at Document Level. version 4.1. Technical report, Fondazione Bruno Kessler. <http://www.newsreader-project.eu/files/2013/01/NWR-2014-2.pdf>.
- UZZAMAN, NAUSHAD, HECTOR LLORENS, JAMES F. ALLEN, LEON DERCZYNSKI, MARC VERHAGEN, eta JAMES PUSTEJOVSKY. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR* abs/1206.5333. <http://arxiv.org/abs/1206.5333>.
- VERHAGEN, MARC, ROBERT GAIZAUSKAS, FRANK SCHILDER, MARK HEPPLE, GRAHAM KATZ, eta JAMES PUSTEJOVSKY. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 75–80, Prague.
- ,INDERJEET MANI, ROSER SAURI, ROBERT KNIPPEN, SEOK BAE JANG, JESSICA LITTMAN, ANNA RUMSHISKY, JOHN PHILLIPS, eta JAMES PUSTEJOVSKY. 2005. Automating Temporal Annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, 81–84. <http://dx.doi.org/10.3115/1225753.1225774>.
- , 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.
- YAGHOOBZADEH, YADOLLAH, GHOLAMREZA GHASSEM-SANI, SEYED ABOLGHASSEM MIRROSHANDEL, eta MAHBANEH ESHAGHZADEH. 2012. ISO-TimeML Event Extraction in Persian Text. In *Proceedings of COLING 2012*, 2931–2944, Mumbai, India. The COLING 2012 Organizing Committee. <http://www.aclweb.org/anthology/C12-1179>.

Eskerrak eta oharrak

Lan hau eusko Jaurkitzako PRE_2014.2.242 bekari esker egin da Arantza Díaz de Ilarraza eta María Jesús Aranzaberen zuzendaritzapean.