RiCL Research in Corpus Linguistics

EusTimeML: A mark-up language for temporal information in Basque

Begoña Altuna - María Jesús Aranzabe - Arantza Díaz de Ilarraza
University of the Basque Country / Spain

Abstract – We present EusTimeML, a mark-up language for temporal information in texts written in Basque. It is compliant with the TimeML specifications, while offering some adapted attributes and attribute values in order to represent the language-specific features of Basque. In particular, alterations have been carried out for verb tense, aspect and modality coding, as well as for time expression and signal annotation. EusTimeML also provides a major extension to the existing TimeML schemes, since the attributes and values for factuality annotation have been added to the existing temporal information annotation scheme. EusTimeML has been used to annotate the *EusTimeBank Corpus*, the news and history narratives corpus that has been used as the gold standard in temporal information processing in Basque.

Keywords – temporal information processing; Basque; mark-up language; annotation; TimeML

1. Introduction

Natural Language Processing (NLP) aims at getting the deepest textual understanding, for which, after mastering morphosyntactic analysis, the focus has been put on semantic and discourse information. Temporal information is an integral part of those areas as it conveys the information of what is narrated in text while providing information to arrange narratives along a temporal axis. This information is of utmost relevance to the development of automatic systems that benefit from knowing the chronological ordering of events in texts, such as chronology creation (Bauer *et al*. 2015), event prediction (Radinsky and Horvitz 2013) and event forecasting systems (Kawai *et al*. 2010), among others.

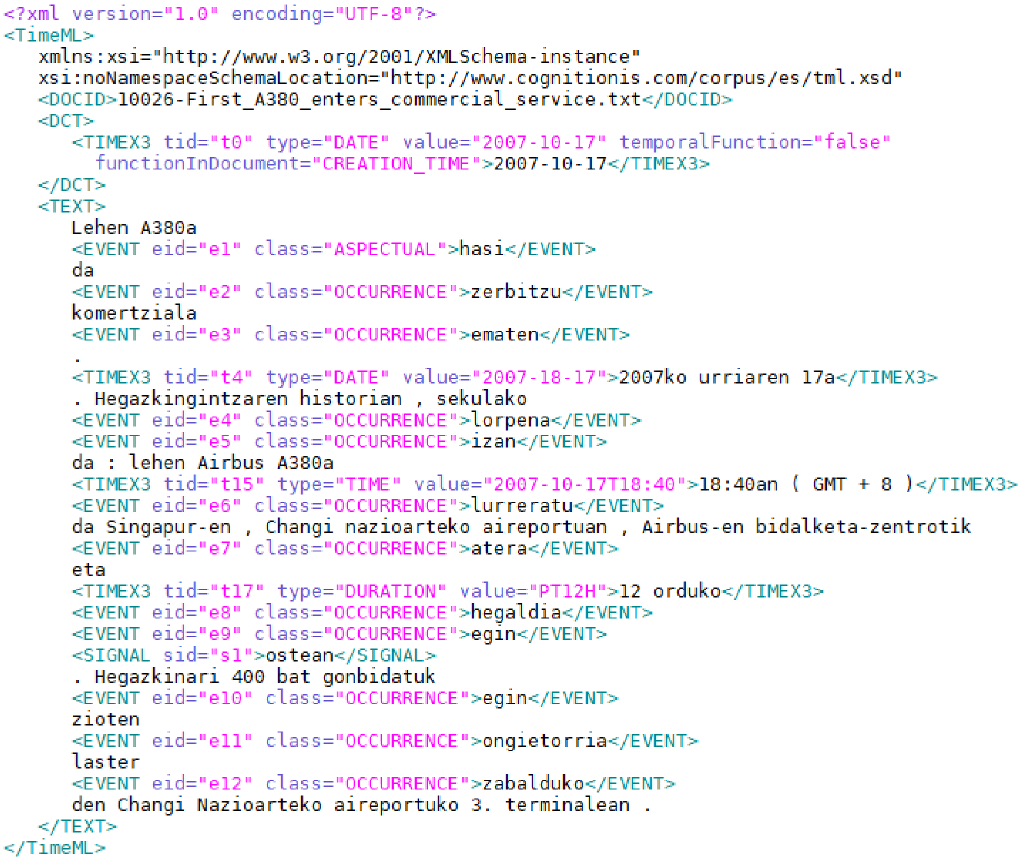
Specifically, temporal information conveys the information of what happens (events narrated) and the times in which they happen (time expressions), as well as the temporal relations (simultaneity, precedence, etc.) between them. For example, in the sentence in (1), one can learn that there was a toilet paper theft (event) last month (time expression) after (temporal relation) there were shortages (event).

Research in Corpus Linguistics 8: 86–104 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.01.06

(1) Last month, armed robbers stole pallets of toilet paper in Hong Kong following panic-buying induced shortages.

That temporal information is collected in corpora that are annotated following structured formats, e.g., the eXtended Mark-up Language (XML), which make the information in the texts machine-readable. Mark-up languages provide a set of tags to classify the different elements in the text, as well as a set of attributes to describe the relevant linguistic features of those elements.

For the annotation of temporal information in Basque, we have created EusTimeML, a TimeML-compliant mark-up scheme (Pustejovsky *et al*. 2003a). It provides tags for events, time expressions and the relations that hold between them in XML format. As Figure 1 shows, some text strings have been assigned a tag (in green) since those are the elements in text that express temporal information. Additionally, a set of attributes (in purple) represents the main information (in pink) those strings convey.



```
<?xml version="1.0" encoding="UTF-8"?>
<TimeML>
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:noNamespaceSchemaLocation="http://www.cognitionis.com/corpus/es/tml.xsd"
   <DOCID>10026-First_A380_enters_commercial_service.txt</DOCID>
   <DCT>
      <TIMEX3 tid="t0" type="DATE" value="2007-10-17" temporalFunction="false"
        functionInDocument="CREATION_TIME">2007-10-17</TIMEX3>
   </DCT>
   <TEXT>
      Lehen A380a
      <EVENT eid="e1" class="ASPECTUAL">hasi</EVENT>
      da
      <EVENT eid="e2" class="OCCURRENCE">zerbitzu</EVENT>
      komertziala
      <EVENT eid="e3" class="OCCURRENCE">ematen</EVENT>
      .
      <TIMEX3 tid="t4" type="DATE" value="2007-18-17">2007ko urriaren 17a</TIMEX3>
      . Hegazkingintzaren historian , sekulako
      <EVENT eid="e4" class="OCCURRENCE">lorpena</EVENT>
      <EVENT eid="e5" class="OCCURRENCE">izan</EVENT>
      da : lehen Airbus A380a
      <TIMEX3 tid="t15" type="TIME" value="2007-10-17T18:40">18:40an ( GMT + 8 )</TIMEX3>
      <EVENT eid="e6" class="OCCURRENCE">lurreratu</EVENT>
      da Singapur-en , Changi nazioarteko aireportuan , Airbus-en bidalketa-zentrotik
      <EVENT eid="e7" class="OCCURRENCE">atera</EVENT>
      eta
      <TIMEX3 tid="t17" type="DURATION" value="PT12H">12 orduko</TIMEX3>
      <EVENT eid="e8" class="OCCURRENCE">hegaldia</EVENT>
      <EVENT eid="e9" class="OCCURRENCE">egin</EVENT>
      <SIGNAL sid="s1">ostean</SIGNAL>
      . Hegazkinari 400 bat gonbidatuk
      <EVENT eid="e10" class="OCCURRENCE">egin</EVENT>
      zioten
      <EVENT eid="e11" class="OCCURRENCE">ongietorria</EVENT>
      laster
      <EVENT eid="e12" class="OCCURRENCE">zabalduko</EVENT>
      den Changi Nazioarteko aireportuko 3. terminalean .
   </TEXT>
</TimeML>
```

Figure 1: A text annotated following the EusTimeML mark-up language (simplified annotation)

The text in Figure 1 is part of the *EusTimeBank Corpus* (Altuna *et al*. under revision a) which, in turn, has been used to train and evaluate temporal information processing tools. The Basque language has a long tradition of linguistic analysis and automatic processing (Alegria and Sarasola 2017) and integrating temporal information processing in the Basque processing pipeline (Otegi *et al*. 2016) has been the major motivation of this work.

This paper is structured as follows. We revisit the most relevant work on temporal information mark-up languages in Section 2. In Section 3, we present the basic features of TimeML, and in Section 4 we describe the most relevant linguistic features of Basque and the adaptations of TimeML that we have instituted to accommodate those features. We discuss the strengths and weaknesses of EusTimeML in Section 5, and we conclude our work in Section 6.

2. Background

Temporal information processing has attracted the interest of NLP scholars over the last two decades and has experienced a substantial boost since the creation of TimeML (Pustejovsky *et al*. 2003a). In fact, ever since the creation of TimeML, resource generation efforts and system evaluation competitions have multiplied. TimeML has been adapted to multiple languages, tasks and domains, and corpora annotated with TimeML schemes have increased in number.

The first temporal information mark-up languages (Mani and Wilson 2000; Ferro *et al*. 2003) only dealt with time expressions, for which the TIMEX and TIMEX2 tags respectively were created. These two tags also offered a set of basic attributes to code the main information expressed by time expressions, such as the normalised value and the granularity of the time expression. TimeML (Pustejovsky *et al*. 2003a), instead, made a qualitative leap in temporal information annotation, as this mark-up language offered tags for all the elements taking part in the expression of temporal information (see Section 3).

TimeML is now an ISO[1] standard (Pustejovsky *et al*. 2010) used in the annotation of many temporally annotated corpora such as *TimeBank* (Pustejovsky *et al*. 2003b, 2006), the *THYME Corpus* (Styler *et al*. 2014), the *PHEME Tweet Corpus* (Derczynski and Bontcheva 2014) and the *Event StoryLine Corpus* (Caselli and Vossen 2017), among others. Moreover, TimeML has been adapted to address some special annotation

[1] International Standards Organisation.

needs. TimeML-strict (Derczynski *et al*. 2013) aims at reducing annotation ambiguity and TimeML-Dense (Cassidy *et al*. 2014) offers the opportunity to create denser temporal relation graphs, while Mostafazadeh *et al*. (2016) use a reduced version of TimeML to annotate the *ROCStories Corpus*.

TimeML has also been developed for many languages, as it is considered a *de facto* standard. Among other languages, TimeML schemes are available for French (Bittar 2010), Italian (Caselli *et al*. 2011), Portuguese (Costa and Branco 2012) Romanian (Forăscu and Tufiş 2012), Spanish (Saurí *et al*. 2009, 2010; Saurí 2010) and Catalan (Saurí and Pustejovsky 2009, 2010; Saurí 2010) and Korean (Jeong *et al*. 2015).

Nonetheless, TimeML is not the only mark-up language that has been developed to address temporal information. TEMANTEX (Wonsever *et al*. 2015) merges event annotation and factuality annotation. In the mark-up language developed for the *NewsReader project* (Minard *et al*. 2016), in turn, temporal information is tagged as in TimeML, but causality relations and entity co-reference are also considered. PLIMEX (Kocoń and Marcińczuk 2015) addresses time expressions in Polish and follows TimeML guidelines quite narrowly. Finally, Ning *et al*. (2018) created a mark-up language that focuses on the extraction of relevant information for timeline construction. This mark-up language complies with most of the TimeML decisions, while it offers a much richer annotation for intrasentential temporal relations.

3. TIMEML

The TimeML mark-up language was specifically created to annotate events, time expressions and the temporal relations between them in text (Pustejovsky *et al*. 2010). For that, the following set of tags was defined, one for each element concerning temporal information or type of relation:

- <EVENT> for events: actions and situations that happen or occur, as in (2).[2]

  (2) Numerous conspiracies have <EVENT>appeared</EVENT> since the <EVENT>outbreak</EVENT>.

- <TIMEX3> for temporal expressions that convey date, time, duration or set information, as in (3).

[2] TimeML foresees single-token annotations for events.

(3) Cases of the new coronavirus emerged in Wuhan <TIMEX3>late last year</TIMEX3>.

- <SIGNAL> for sections of text, most commonly function words, that indicate the type of relation among temporal objects, as in (4).

(4) Numerous conspiracies have appeared <SIGNAL>since</SIGNAL> the outbreak.

- <TLINK> for temporal relations between two events, two time expressions or an event (in bold) and a time expression (in italics), as in (5).

(5) Numerous conspiracies (ei1) have **appeared** (ei2) since the *outbreak* (ei3). <TLINK eventInstanceID="ei2" relatedToEvent="ei3" relType="BEGUN_BY"/>

- <ALINK> for aspectual relations between an aspectual event (in bold) and its subordinated event (in italics), as in (6).

(6) Several patent documents **started** (ei1) to *circulate* (ei2) on Twitter. <ALINK eventInstanceID="ei1" relatedToEvent="ei2" relType="START"/>

- <SLINK> for subordination relations between a main event (in bold) and its subordinated event (in italics), as in (7).

(7) Ms Mengyun apologised (ei1), **saying** (ei2) she was "just *trying* (ei3) to introduce (ei4) the life of local people". <SLINK eventInstanceID="ei2" relatedToEvent="ei3" relType="EVIDENTIAL"/>

These tags contained a set of attributes that coded or normalised the temporal information conveying features of the temporal objects and relations. Table 1 presents the attributes in TimeML for event features. In this case, four types of attributes can be identified according to the type of information they represent: i) event ID (*eid*) and event instance ID (*eiid*) offer identification information; ii) class (*class*) offers event classification information; iii) tense (*tense*) and aspect (*aspect*) offer temporal information; and, finally iv) part-of-speech (*pos*), polarity (*polarity*) and modality (*modality*) offer other relevant linguistic information. Those attributes may get different types of values. Some attribute values can be strings (CDATA) or integers, such as in *eid*, while others get their values from a list of pre-established options, such as for *class*, *tense* and *aspect*.

Attributes and attribute values for the remaining tags have also been defined in TimeML. In the case of TIMEX3 the most relevant tags express the type and normalised value of the time expressions, and SIGNAL tags do not get any attributes. For the relations, the source, target and relation type are specified. The complete description of the TimeML tags, attributes and attribute values can be found in TimeML Working Group (2010).

Event attributes	Values
Event ID (*eid*)	e<integer>
Event instance ID (*eiid*)	ei<integer>
Class (*class*)	REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE
Tense (*tense*)	PAST, PRESENT, FUTURE, NONE, INFINITIVE, PRESPART, PASTPART
Aspect (*aspect*)	PROGRESSIVE, PERFECTIVE, PERFECTIVE_PROGRESSIVE, NONE
Part of speech (*pos*)	ADJECTIVE, NOUN, VERB, PREP, OTHER
Polarity (*polarity*)	NEG, POS
Modality (*modality*)	CDATA

Table 1: Attributes of the TimeML <EVENT> tag and their possible values

4. Language-specific issues and extensions to TimeML

EusTimeML follows most of the standards in TimeML, namely, it preserves the token-level annotation system and all the tags proposed in TimeML, as well as the attributes and values that code temporal information. Additionally, it also shares almost all the attributes for linguistic features and their values. Nonetheless, Basque has some language-specific issues (see Section 4.1) that have conditioned the adaptation of TimeML to Basque (see Section 4.2), for which a series of attribute values has been altered. Furthermore, EusTimeML offers a set of attributes to address some supplementary information so as to increase the amount and variety of information it encodes (see Section 4.3). These all have contributed to the final version of EusTimeML (see Section 4.4).

4.1. Language-specific issues

Basque is a non-Indo-European language isolate, and thus it does not share many of the linguistic features of its neighbouring languages. In particular, many of its morphosyntactic features differ from the features in neighbouring languages and, hence, specific research for processing Basque is usually needed, as choices made for other languages cannot be applied straightforwardly.

For example, Basque is a highly agglutinative language in which information commonly expressed by prepositions in neighbouring languages is expressed by a rich set of postpositions attached to lemmas, as can be seen in the sentence in (8). This feature is extremely relevant in temporal information processing as lemmas accompanied by spatio-temporal declension cases are very frequent in temporal information expressions.

(8) Sorosleek iluntzetik egunsentira etengo dituzte erreskate-operazioak.
Rescuers sunset.ABL sunrise.ALL stop.FUT aux.PRES rescue.operations.
‘Rescuers will stop rescue operations from sunset to sunrise.’

In (8) there are two time expressions: *iluntzetik* ‘from sunset’ and *egunsentira* ‘to sunrise’. *Iluntze* and *egunsenti* mean ‘sunset’ and ‘sunrise’, respectively, while the suffix *-etik* expresses the ablative case and *-ra* represents the allative case.

Verbal conjugation also represents a major difference between Basque and other languages. In Basque, there is a short list of single-word verb forms, typically to express punctual aspect, whereas most of the tensed verb forms are periphrastic. The lexical meaning of the verb and aspect are expressed in the main verb, while the auxiliary verb expresses tense and agreement with the persons taking part in the event, as well as mood and modality.

Looking at the sentence in (8) again, one may notice that the verb *etengo dituzte* (‘will stop’) also shows the rich morphology of Basque. The suffix *-go* expresses the future aspect of the verb and the auxiliary *dituzte* represents the present time tense (*d-*) as well as the concordance with the object (*erreskate-operazioak.*3PL), *-it-* and *-z-*, and the subject (*sorosleek.*3PL), *-te*.

As just mentioned, the future meaning of a verb is considered an aspectual value in Basque, whereas in many European languages future events are expressed by the future tense. This makes it possible to understand the Basque verbal tense as a bi-dimensional present-past feature (Table 2), and verbal aspect as a perfect-future feature (Table 3).

		Present	
		Present	Non-present
Past	Non-past	Eten dituzte ('They have stopped')	Eten (izan) balituzte ('If they had stopped')
	Past		Eten zitutzen ('They stopped')

Table 2: Representation of verbal tenses in Basque

		Perfectiveness	
		Perfect	Non perfect
Futurity	Non-future	Eten dituzte ('They have stopped')	Eteten dituzte ('They stop')
	Future		Etengo dituzte ('They will stop')

Table 3: Representation of verbal aspect in Basque

4.2. Adaptations to TimeML

Although the TimeML mark-up language is considered to be a standard for temporal information annotation, each version contains subtle variations to address language or task-specific issues. In EusTimeML, some of the attribute values have been modified to accommodate the analysis of Basque grammar.

4.2.1. Time expression and signal annotation

As introduced in Section 4.1, time expressions often get spatio-temporal postpositions and both elements commonly appear as a single token. Those elements are given separated tags in TimeML-styled schemes: one for the time expression (<TIMEX3>) and one for the function word (normally a preposition) expressing a temporal relation (<SIGNAL>). In the case of Basque, instead, as EusTimeML respects the token-level annotation, we decided to annotate the whole word as a time expression, since we believe that the postpositions' relational information can always be recovered from the morphosyntactic parsing.

Nevertheless, free postpositions are also possible in Basque and, in those cases, we decided to assign them a signal tag, as the tags for free postpositions do not interfere with any other tags present in a text. As a consequence, the time expression and signal information annotation according to EusTimeML is represented as in examples (9–10). A similar decision was made for the annotation of events and signals. More precisely, as only main verbs are given the event tag, the auxiliaries of the periphrastic forms may get signal tags when they contain a temporal postposition, since there is no overlapping tag as in (11).

(9) Sorosleek <TIMEX3>iluntzetik</TIMEX3>
<TIMEX3>egunsentira</TIMEX3> etengo dituzte erreskate-operazioak.

‘Rescuers will stop rescue operations from <TIMEX3>sunset</TIMEX3> to <TIMEX3>sunrise</TIMEX3>’

(10) Krimean gotortu eta <TIMEX3>1920ko udazkenera</TIMEX3>
<SIGNAL>arte</SIGNAL> eutsi zuten.

‘[They] hid in Crimea and the endured <SIGNAL>until</SIGNAL>
<TIMEX3>Autumn 1920</TIMEX3>’

(11) Gerra Zibila Armada Zuria <EVENT>menderatu</EVENT>
<SIGNAL>zutenean</SIGNAL> amaitu zen.

‘Civil War ended <SIGNAL>when</SIGNAL> [they] <EVENT>overruled</EVENT> the White Army’

As can be seen in (10), the free postposition *arte* has been assigned a SIGNAL tag. The ablative -*tik* and the allative -*ra* in (9), and the allative -*era* in (10), instead, are part of the TIMEX3 tag to which they are attached. Nevertheless, it should be noted that, in (11), the auxiliary *zutenean* contains the locative suffix -*ean*, but as there is no conflict with any other tags, the token has been assigned a SIGNAL tag according to EusTimeML.

4.2.2. Aspect and tense annotation

The fact that the future is represented by aspect in Basque has led us to define an *ad hoc* set of values for aspect and tense. As in other TimeML-styled schemes, verbal aspect is expressed by the *aspect* attribute and verb tense is represented through the *tense* attribute. The values each attribute can be assigned to and the context have been summarised in Table 4.

TENSE		ASPECT	
Values	**Usages**	**Values**	**Usages**
PRESENT	Events expressed by verbs in the present tense	PERFECT	Events expressed by verbs with perfective aspect
PAST	Events expressed by verbs in the past tense	-PERFECT (NON-PERFECT)	Events expressed by verbs with imperfective aspect
HYPOTHETICAL	Events expressed by verbs in the hypothetical (non-present, non-past) tense	FUTURE	Events expressed by verbs with future aspect
NONE	Events expressed by untensed verbs and non-verbal forms	NONE	Events expressed by verbs with no aspect mark and non-verbal forms

Table 4: Values and usages of the *aspect* and *tense* attributes in EusTimeML

As a consequence, for the sentence in (8) the event *etengo dituzte* ('will stop') would be assigned the *aspect* and *tense* values as illustrated in (12), since this is a future verb form. *Erreskate-operazioak* ('rescue operations'), instead, will be assigned NONE as the value for *aspect* and *tense,* as it is expressed by a noun phrase and the form has no aspect or tense marks.

(12) Sorosleek iluntzetik egunsentira <EVENT aspect="FUTURE" tense="PRESENT">etengo</EVENT> dituzte <EVENT aspect="NONE" tense="NONE">erreskate-operazioak</EVENT>.

'Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT> from sunset to sunrise'

4.2.3. Modality annotation

The annotation of modality information has been tackled in various ways in the different TimeML-styled mark-up languages. While the Spanish and Catalan TimeML schemes (Saurí and Pustejovsky 2009) do not contain any means for modality annotation, the rest of the analysed mark-up languages do offer the option to express modality through the *mod* attribute. However, while a set of definite values has been defined for the French TimeML (Bittar 2010), in the rest of the analysed mark-up languages the forms found in text are used as values.

In the case of EusTimeML, we have followed the Basque grammar tradition, in which the modal verb *ahal izan/ezin izan* (possibility) and the semi-modal verbs *behar izan* (need or obligation) and *nahi izan* (desire) are considered. Taking this into account, the AHAL, BEHAR and NAHI values have been created for the modality attribute, and we have also used the NONE value for the cases in which no modality is expressed. The

NONE value is the one assigned to the events in (13), as they do not convey any modality information.

(13) Sorosleek iluntzetik egunsentira <EVENT modality="NONE">etengo</EVENT> dituzte <EVENT modality="NONE">erreskate-operazioak</EVENT>.

'Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT> from sunset to sunrise'

4.3. Extensions to TimeML: Factuality

The main difference between EusTimeML and other TimeML-styled mark-up schemes relies on the factuality annotation added to EusTimeML (Altuna *et al.* 2018a). Factuality annotation has been closely related to TimeML in works such as Saurí (2008), but EusTimeML is the first TimeML-styled scheme that integrates it. For example, verb aspect and tense, the time expressions related to the events condition, the factuality values of the events, and some subordination relations (evidential, factive or counterfactive, among others) may evidence the factuality value of the subordinated event.

As our final goal is building timelines, factuality information will help us discern between events that effectively do occur and that should, as a consequence, appear on a timeline, events that have not happened, and events that may happen in the future. For this reason, we have opted for a factuality scheme in which we classify events as facts, counterfacts, or non-factual events when possible.

In EusTimeML, factuality information is coded through a set of event attributes. These attributes are *polarity* (defined also in TimeML), *certainty*, *factuality* itself, and *specialCases.* These attributes and their values are illustrated in Table 5.

Polarity	**Certainty**	**Factuality**	**specialCases**
Grammatical polarity (affirmation or negation) expression	*Commitment of the source with the information expressed*	*Information on whether events correspond to a fact in the world, a possibility or a situation that does not hold*	*Marking of conditionals and generic statements*
POS NEG	CERTAIN UNCERTAIN UNDERSPECIFIED	FACTUAL COUNTERFACTUAL NON-FACTUAL NO FACTUALITY VALUE UNDERSPECIFIED	CONDITIONAL_CONDITION CONDITIONAL_MAIN (main clause) GENERIC NONE

Table 5: Factuality specific attributes and values in EusTimeML

We have represented the factuality information of the events in (8) as shown in example (14).

(14) Sorosleek iluntzetik egunsentira <EVENT polarity="POS" certainty="CERTAIN" factuality="NON_FACTUAL" specialCases="NONE">etengo</EVENT> dituzte <EVENT polarity=POS' certainty="CERTAIN" factuality="FACTUAL" specialCases="NONE">erreskate-operazioak</EVENT>.

'Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT> from sunset to sunrise'

4.4. Final EusTimeML definition and usage

Taking into account the decisions we made, the attributes and values for event annotation in EusTimeML are presented in Table 6. The remaining tags preserve the original TimeML attributes and the only differences in annotation are the ones presented in Section 4.2.1. As a consequence, annotations following EusTimeML remain easily transferable and comparable to other annotations carried out following any of the TimeML-styled schemes.

Event attributes	**Values**
Event ID (*eid*)	e<integer>
Event instance ID (*eiid*)	ei<integer>
Class (*class*)	REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE, STATE
Tense (*tense*)	PAST, PRESENT, HYPOTHETICAL, NONE
Aspect (*aspect*)	PERFECT, -PERFECT, FUTURE, NONE
Part of speech (*pos*)	ADJECTIVE, NOUN, VERB, ADVERB, OTHER
Polarity (*polarity*)	NEG, POS
Modality (*modality*)	AHAL, BEHAR, NAHI, NONE
Certainty (*certainty*)	CERTAIN, UNCERTAIN, UNDERSPECIFIED
Factuality (*factuality*)	FACTUAL, COUNTERFACTUAL, NON_FACTUAL, NO_FACTUALITY_VALUE, UNDERSPECIFIED
Special cases (*specialCases*)	CONDITIONAL_CONDITION, CONDITIONAL_MAIN CLAUSE, GENERIC, NONE

Table 6: event attributes and values in EusTimeML

The mark-up language described in the preceding sections has been used for the annotation of *EusTimeBank*, the gold standard corpus for temporal information in Basque. *EusTimeBank* is a 92-document corpus (23,000 tokens) made up of 86 news documents and 6 historical narratives. The corpus has been used for the training and evaluation of bTime[3] (Salaberri 2017) and EusHeidelTime[4] (Altuna *et al*. 2017). Additionally, the annotated documents obtained by those tools have been used as input for KroniXa (Altuna *et al*. under revision b), a tool to build timelines from Basque texts.

News and history texts are especially rich in temporal information, as they commonly narrate past events and offer the necessary information to arrange the events along the temporal axis. Hence, their narrative nature makes these texts an interesting basis for timeline generation. For this reason, a timeline dataset for the evaluation of KroniXa has been created from *EusTimeBank* (Altuna *et al*. 2019).

5. Discussion

The creation of EusTimeML has been the first step towards automatic temporal information extraction from Basque texts. In order to be able to compare the Basque annotated corpora and the results obtained by NLP tools for Basque with the NLP resources for other languages, comparable annotation schemes and evaluation measures should be adopted. Hence, as TimeML schemes are widely used in English, Spanish and French, building the TimeML-compliant EusTimeML has been a convenient option.

The decisions on EusTimeML have been validated by means of a set of manual annotation efforts (Altuna *et al*. 2014, 2018a, 2018b; Altuna 2018), in which inter-annotator agreement has been measured. Manual annotation analysis has shown that EusTimeML annotation guidelines are unambiguous for most of the elements, but we must note that event classification has been a major source of disagreement as annotators have considered some event classes to be virtually indiscernible in some contexts. The discussions after the agreement assessment have led to a wide consensus on EusTimeML and a consistent set of annotation guidelines has been produced (Altuna *et al*. 2016).

As our final goal is generating timelines based on the temporal information contained in texts, we have paid special attention to similar work based on TimeML annotations. In fact, the suitability of TimeML to encode temporal information for timeline building has been called into question. Ning *et al*. (2018) argue that the scarcity of intrasentential temporal relations heavily affects the event-event ordering. This

[3] Event and temporal relation extraction and classification tool.

[4] Time expression extraction and normalisation tool.

opinion is shared by Derczynski *et al*. (2013), as they proposed TimeML-Dense, although timeline building was not their final goal. Laparra *et al*. (2017) are also aware of the data sparsity problem for timeline building provided by TimeML annotations. They thus propose assigning the same time tag to all events a certain entity takes part in if they share the same tense, as a way to increase the number of anchored events. This partially solves the lack of temporal relations between events in the text. In Altuna *et al*. (under revision b) we have also found that some time expressions can have more than one correct normalised value in TimeML, which causes unnecessary time expression ordering problems as simultaneous events can be incorrectly placed in two different time points. For example, the quarters of the year may get different normalised values depending on whether they are referred to as quarters of a natural year or of a fiscal year.

Nonetheless, we consider that EusTimeML still offers sufficient information for timeline building. It should be taken into account that, even if bTime can only deal with a restricted set of temporal relations, experiments with KroniXa have shown very promising results, as a third of the events are correctly placed in the timelines.

Other authors have also highlighted some points in which TimeML struggles to properly encode temporal information. Ehrmann and Hagège (2009) noted that TimeML neither offered precise guidelines for time expression classification nor a clear distinction between characterisation and reference calculation annotations. According to them, a time expression such as *2 days before yesterday* should be considered a date, and *2 days* should be used to calculate its reference; TimeML proposes to annotate a duration (*2 days*) and a date (*yesterday*), instead. This same concern is shared by Bethard (2013) who proposes a scheme (SLATE) that allows machine-learning calculations.

Along the same lines, Laparra *et al*. (2018) identified the incapacity of TimeML to annotate compositional time expressions such as *Saturdays since March 6*, in which a set of dates is bounded by a determined time point. Event annotation through TimeML has also been a matter of discussion among scholars. For example, as Leeuwenberg and Moens (2019) point out, event durations cannot be explicitly tagged through TimeML, as no scheme for marking the durative (or punctual) nature of the events is provided. In spite of these flaws, TimeML is still the most widely used mark-up language for temporal information annotation.

6. CONCLUSIONS

EusTimeML addresses the need for a temporal information mark-up language for Basque that can deal with its language-specific features. Nevertheless, even if it contains some modifications, it is largely comparable to other TimeML-styled schemes. Adding factuality information has contributed to enlarging the amount of relevant information for timeline generation, which is our final goal.

In fact, EusTimeML has been the first step towards temporal information processing in Basque as it has been the mark-up language used for the *EusTimeBank* annotation, the corpus used for the development of the EusHeidelTime and bTime tools for temporal information extraction and normalisation. Furthermore, documents annotated following EusTimeML have also been used to generate timelines for the evaluation of KroniXa.

EusTimeML is now ready to use, although its customisability still allows for improvements and expansions. Addressing duration anchoring and increasing the amount of intrasentential temporal relations should be a goal for the TimeML community.

REFERENCES

Alegria, Iñaki and Kepa Sarasola. 2017. Language technology for language communities: An overview based on our experience. In Nicholas Ostler ed. *FEL XXI Alcanena 2017 Communities in Control*. Hungerford, UK: Foundation for Endangered Languages, DIDLeS, SOAS World Languages Institute and Mercator Research Centre, 91–97.

Altuna, Begoña. 2018. *Euskarazko denbora-egituren azterketa eta corpusaren sorrera / Analysis of Basque temporal constructions and the creation of a corpus*. Donostia: University of the Basque Country dissertation.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2014. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática* 6/2: 13–24.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2016. *Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0* (UPV/EHU/LSI/TR;01-2016). Donostia: University of the Basque Country.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2017. EusHeidelTime: Time expression extraction and normalisation for Basque. *Procesamiento del Lenguaje Natural* 59: 15–22.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2018a. An event factuality annotation proposal for Basque. In Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti and Caroline Sporleder eds.

Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2), Vol. 1. Vienna: Gerastree Proceedings, 15–24.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2018b. Adapting TimeML to Basque: Event annotation. In Alexander Gelbukh ed. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science (LNCS)* 9624. Cham: Springer International Publishing, 565–577.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2019. EusTimeBank-TL corpusa: Denbora-informaziodun testuetatik denbora-lerroetara. In Olatz Arbelaitz, Urtzi Etxeberria, Ainhoa Latatu, Miren Josu Ormaetxebarria eds. *III. Ikergazte. Nazioarteko Ikerketa Euskaraz, Giza Zientziak eta Artea*, Vol. 1. Bilbao: Udako Euskal Unibertsitatea, 83–90.

Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. Under revision a. EusTimeBank: A corpus for temporal information processing in Basque. *Language Resources and Evaluation*. Cham: Springer International Publishing.

Altuna, Begoña, Ander Soraluze, María Jesús Aranzabe, Olatz Arregi and Arantza Díaz de Ilarraza. Under revision b. KroniXa: Timeline creation from Basque texts. *Digital Scholarship in the Humanities*. Oxford: Oxford University Press.

Bauer, Sandro, Stephen Clark and Thore Graepel. 2015. Learning to identify historical figures for timeline creation from Wikipedia articles. In Lucia Aiello and Daniel E. McFarland eds. *SocInfo 2014 International Workshops, Revised Selected Papers*. Barcelona, Spain: Springer, 234–243.

Bethard, Steven. 2013. A synchronous context free grammar for time normalization. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu and Steven Bethard eds. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, USA: Association for Computational Linguistics, 821–826.

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

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: The It-TimeML experience for the Ita-TimeBank. In Nancy Ide, Adam Meyers, Sameer Pradhan and Katrin Tomanek eds. *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, Oregon: Association for Computational Linguistics, 143–151.

Caselli, Tommaso and Piek Vossen. 2017. The Event StoryLine Corpus: A new benchmark for causal and temporal relation extraction. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, Teruko Mitamura and David Caswell eds. *Proceedings of the Events and Stories in the News Workshop.* Vancouver, Canada: Association for Computational Linguistics 77–86.

Cassidy, Taylor, Bill McDowell, Nathanael Chambers and Steven Bethard. 2014. An annotation framework for dense event ordering. In Kristina Toutanova and Hua Wu eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, USA: Association for Computational Linguistics, 501–506.

Costa, Francisco and António Branco. 2012. TimeBankPT: a TimeML annotated corpus of Portuguese. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis eds. *Proceedings of the Eighth International Conference on Language Resources* *and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 3727–3734.

Derczynski, Leon and Kalina Bontcheva. 2014. PHEME: veracity in digital social networks. In Harry Bunt ed. *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA)*. Reykiavic: Association for Computational Linguistics, 65–68.

Derczynski Leon, Héctor Llorens, and Naushad UzZaman. 2013. TimeML-Strict: clarifying temporal annotation. Computing Research Repository (CoRR) abs/1304.7289. http://arxiv.org/abs/1304.7289 (29 December, 2019.)

Ehrmann, Maud and Caroline Hagège. 2009. Proposition de caracterisation et de typage des expressions temporelles en contexte. In Adeline Nazarenko and Thierry Poibeau eds. *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles*. Senlis, France: Association pour le Traitement Automatique des Langues.

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim and George Wilson. 2003. *TIDES 2003 Standard for the Annotation of Temporal Expressions*. McLean, USA: The MITRE Corporation.

Forăscu, Corina and Dan Tufiş. 2012. Romanian TimeBank: An annotated parallel corpus for temporal information. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis eds. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 3762–3766.

Jeong, Young-Seob, Zae Myung Kim, Hyun-Woo Do, Chae-Gyun Lim and Ho-Jin Choi. 2015. Temporal information extraction from Korean texts. In Afra Alishahi and Alessandro Moschitti eds. *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015*. Beijing, China: Association for Computational Linguistics, 279–288.

Kawai, Hideki, Adam Jatowt, Katsumi Tanaka, Kazuo Kunieda, and Keiji Yamada. 2010. Chronoseeker: Search engine for future and past events. In Dongsoo S. Kim, Sang-Wook Kim, Suk-Han Lee, Lajos Hanzo and Roslan Ismail eds. *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication, ICUIMC ’10*. New York, USA: Association for Computing Machinery, 25:1–25:10.

Kocoń, Jan and Michał Marcińczuk. 2015. Recognition of Polish temporal expressions. In Galia Angelova, Kalina Bontcheva and Ruslan Mitkov eds. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2015)*. Hissar, Bulgaria: RANLP, 282–290.

Laparra, Egoitz, Rodrigo Agerri, Itziar Aldabe, German Rigau. 2017. Multi-lingual and cross-lingual timeline extraction. *Knowledge-Based Systems* 133, 77–89.

Laparra, Egoitz, Dongfang Xu and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics* 6, 343–356.

Leeuwenberg, Artuur and Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *The Journal of Artificial Intelligence Research (JAIR)* 66: 341–380.

Mani, Inderjeet and George Wilson. 2000. Robust temporal processing of news. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, 69–76.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), 4417–4422.

Mostafazadeh, Nasrin, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli and James Allen. 2016. A corpus and cloze evaluation framework for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of NAACL-HLT 2016*. San Diego, CA: Association for Computational Linguistics, 839–849.

Ning, Qiang, Hao Wu and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 1318–1328.

Otegi, Arantxa, Nerea Ezeiza, Iakes Goenaga and Gorka Labaka. 2016. A modular chain of NLP tools for Basque. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala eds. *Proceedings of the 19th International Conference on Text, Speech and Dialogue, TSD*. Cham: Springer, 93–100.

Pustejovsky, James, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz and Dragomir Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering* 3, 28–34.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro and Marcia Lazo. 2003b. The TimeBank Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery eds. *Proceedings of Corpus Linguistics 2003*. Lancaster, UK: UCREL, Lancaster University, 647–656.

Pustejovsky, James, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, Andrea Setzer. 2006. *TimeBank 1.2 LDC2006T08*. Web Download. Philadelphia: Linguistic Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/LDC2006T08

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias eds. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. La Valletta: Association for Computational Linguistics, 394–397.

Radinsky, Kira and Eric Horvitz. 2013. Mining the web to predict future events. In Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina and Aristides Gionis eds. *Proceedings of the sixth ACM international conference on Web search and data mining*. New York: Association for Computing Machinery, 255–264.

Salaberri, Haritz. 2017. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Donostia: University of the Basque Country dissertation.

Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Waltham, MA: Brandeis University dissertation.

Saurí, Roser. 2010. *Annotating Temporal Relations in Catalan and Spanish TimeML Annotation Guidelines*. Barcelona: Barcelona Media.

Saurí, Roser and James Pustejovsky. 2009. *Annotating Events in Catalan – TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona: Barcelona Media.

Saurí, Roser and James Pustejovsky. 2010. *Annotating Time Expressions in Catalan – TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona: Barcelona Media.

Saurí, Roser, Olga Batiukova and James Pustejovsky. 2009. *Annotating Events in Spanish. TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona Media.

Saurí, Roser, Estela Saquete and James Pustejovsky. 2010. *Annotating Time Expressions in Spanish. TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona Media.

Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky. 2014. Temporal annotation in the clinical domain. In Ellen Riloff ed. *Transactions of the Association for Computational Linguistics* 2: 143–154.

TimeML Working Group. 2010. *TimeML Annotation Guidelines Version 1.3*. Technical report.

Wonsever, Dina, Aiala Rosá, Marisa Malcuori and Matias Etcheverry. 2015. TEMANTEX: A markup language for Spanish temporal expressions and indicators. *Research in Computing Science* 97: 9–19.

Corresponding author
Begoña Altuna
Faculty of Informatics
University of the Basque Country
Manuel Lardizabal 1
20018 Donostia (Spain)
e-mail: begona.altuna@ehu,eus

received: January 2020
accepted: April 2020