Annotating Time, Event, and Space Information in Modern Spanish Text

ModeS TimeBank Annotation Guidelines (Version 1.0)

Barcelona Media

Technical Report 2012-01

Marta Guerrero Nieto
Mercator Group
Technical University of Madrid
m.guerrero@topografia.upm.es
guerreronieto.marta@gmail.com

Roser Saurí
Voice and Language Group
Barcelona Media
roser.sauri@barcelonamedia.org

The goal of this document is two-fold. On the one hand, summarizing the use of TimeML and SpatialML annotation schemes as applied to the Modern Spanish TimeBank corpus (ModeS TimeBank). On the other, discussing the annotation criteria for time and event constructions specific to the historical variety of Modern Spanish (17^{th}-18^{th} Centuries) that are present in this corpus.

1. Time expressions in the corpus

TimeML basics. In TimeML, time expressions are annotated with the TIMEX3 tag, which can be classified into: DATE (*septiembre de 2011* `September 2011´), TIME (*a las 4 de la tarde* `at 4 pm´), DURATION (*los próximos días* `the next days´) and SET (*todos los días* `every day´).

According to the general guidelines for annotating temporal information in Spanish (Saurí et al., a,b), time expressions in that language can be conveyed by temporal nouns (1a) and noun phrases with temporal meaning (1b-c):

(1) a. *Viernes Santo* (`Good Friday´)
b. *la mañana* (`the morning´)
c. *algunas horas* (`some hours´)

Complex time expressions. Complex time expressions may be annotated into separated TIMEX3 tags based on a set of criteria that take into account both the grammatical structure and the semantic behaviour of the construction. For example, the following time expressions are marked up separately because of a difference in granularity (greater than or less than 24 hours) (2a-b) or entity nature (instants vs. durations) (2c):

(2) a. *ayer a las 4 de la tarde* (`yesterday at 4pm´)
b. *el 31 de julio a las 11 de la mañana* (`July 31st at 11 am´)
c. *hoy todo el día* (`all day today´)

Closed intervals. In the ModeS TimeBank corpus, it is very common to find two time expressions denoting a closed interval. Each of these expressions will be marked up separately:

(3) *dia 17 de dicho hasta el 18* (´day 17th of the aforementioned until the 18th´)

Note that the structure of the NPs present in this kind of constructions tends to differ from NPs in current Spanish (in terms of order, constituents, elisions, etc.). Similarly, here and in the

coming examples, the incorrect spelling (e.g., *dia* instead of *día*) is inherited from the original text.

Anchored time expressions. A time expression which interpretation is anchored to (or, equivalently, depends on) the value of another expression can be of two kinds: it can be anchored to the time of speech/creation of the document (4a), or to another time expression in the text (4b-d).

(4) a. *estos dos días* (`these two days´)
b. *el 27 del mismo* [mes] (`27th of the same [month]´)
c. *a esta* [hora] *de 12* (`at this [hour] of 12´)
d. *a esta hora de las 6* (`at this hour of 6´)

Being a corpus of non-contemporary Spanish, expressions like (4b-c) are quite common in ModeS TimeBank. They are idiosyncratic in terms of structure as well as lexical choice. For instance, nouns expressing time units are typically omitted when preceded by and adjective or demonstrative, as indicated with square brackets in (4b-c) above:

Prep +	**Adj/Dem**	**+**	**Time Unit + (...)**		**> Prep +**	**Adj/Dem +**	**(...)**
a	*esta*		*hora*	*de 12*	> *a*	*esta*	*de 12*
del	*mismo*		*mes*		> *del*	*mismo*	

Spanish time construction: *hacer* + Duration NP. Constructions consisting of the verb *hacer* (`to do/make´) followed by a duration-denoting NP are considered time expressions in TimeML:

(5) *hace tres días*
Literally: it-does/makes three days
Gloss: `three days ago´/`for three days´

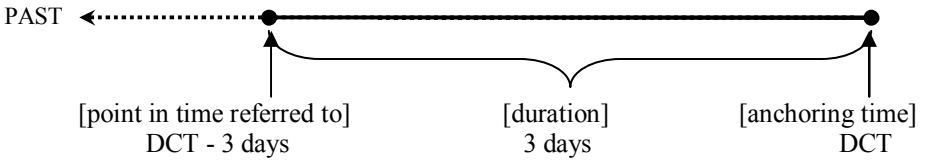
In Modern Spanish, this construction can be found with the verb *haber* ('to have') as well. The structure can be summarized as:

(6) ***havia/ habia/ hace / hacia* + Duration NP**

where the tokens *havia* and *habia* are alternative spellings of an inflected form of the verb *haber* ('to have'), and the tokens *hace* and *hacia* are inflected forms of the verb *hacer* ('to do/make).

This construction can refer to both a point in time or a duration. In some contexts it refers to a point in time by anchoring the duration denoted by the NP complement of *hacer/haber* (e.g., *tres dias*) to a time reference (e.g., the speech time). The following figure illustrates this by using the example in (5), where we assume that the anchor time is the Document Creation Time (DCT):

(7)



In other contexts, it simply denotes the length (or duration) of time conveyed by the NP complement of *hacer/haber*.

In terms of their annotation, these expressions need to be anchored to a second TIMEX3 tag.

Infinitival clauses. Specific to Modes TimeBank is tagging infinitival clauses of time as TIMEX3s:

(8) a. *al romper el alva* (`at the break of sunrise´)

b. *al clariar el día* (`at the lightening of the day´)

The structure of these constructions is*:* ***al* + Infinitive + Time NP**

Verbs expressing day time change. Time expressions in Modes TimeBank also include those verbs referring to moving into a new part of the day:

(9) a. *amanece el dia* (`the day dawns´)
b. *anocheció* (`it nightfalled´)

These expressions function often as: time adverbials, the second member in comparative constructions, or the head of impersonal or intransitive constructions.

Expressions of speed. According to the general TimeML annotation guidelines for Spanish text, expressions like *en cualquier momento* (`at any time´)*, al instante* (`instantly´)*,* etc.*,* are not marked up since they refer to the speed in which the event takes place, and not to the time when it occurs.

(10) *por instantes se esperava llegase de la Habana* (`it was expected to arrive from La Habana momentarily´)

2. Events in the corpus

Most events are expressed by verbs, which can be finite or non-finite:

(11) a. *saltó el viento al Norte* (´the North wind jumped´)
b. *a las nueve avisté una vela al OSO* (´ I sighted a W-SW sailboat at 9h´)
c. *por no haver viento alguno* (´for not having any wind´)
d. *yendo para Montevideo* (´going to Montevideo´)

Moreover, events can be expressed by means of nouns, adjectives, or prepositions. We will see each of these cases in the coming subsections.

2.1 Verbs

Auxiliary verbs are not marked up as part of the EVENT tag, as is the case of the form *haver* ('to have') below.

(12) *el motivo de haver variado el rumbo* (`the reason for having turned´)

Periphrastic constructions. In periphrastic constructions all verbal elements (finite and non-finite) are annotated as independent events:

(13) *bolvimos a ver* (`We came back to see´)

In durative periphrases, only the gerund form is marked up as event:

(14) *vine navegando toda la noche en la mano* (`I came sailing through the night´)

Multiword expressions. In the ModeS TimeBank corpus, it is very common to find the following constructions:

(15) *ir en vuelta* (´return, go back´. Alternative form: *ir enbuelta*)
dar fondo (lit.: ´give deepness´, gloss: 'to anchor')

The first construction contains two candidates to event expression: *ir* (`go') and *en vuelta* (`in return´). We consider *ir* as auxiliary and annotate only the main verbal element: *en vuelta or envuelta*.

(16) *avistamos un navio grande que iva enbuelta del Oeste* (`We sighted a big boat that was in-return from the West´)

The second construction means *fondear* `to anchor´, and it will be annotated as two separate events (13):

(17) *el 27 del mismo dio fondo en este mismo puerto el paquevot correo el Patagon* (`The 27th of that it gave deep in this very port the mailing ship El Patagon´)

Support verb and aspectual constructions. Finally, the two eventive elements in support verb constructions (18) and aspectual constructions (19) are annotated as independent events, following the general annotation guidelines for Spanish:

(18) *hizo demarcación* (`It made demarcation´)
(19) *entro calma* (`The calm came´)
salto una turbonada (´A storm began´)

2.2 Nouns and adjectives

Nominal events. Most of the event-denoting nouns in ModeS TimeBank are derived from verbs:

(20) *fueron causa de su arrivo a este puerto* (`This were the cause of his arrival at this port´)
(21) *salida de la Coruña a las 8 de la mañana del dia 21 de diciembre de 1768* (`Leaving from Corunna at 8 am on day 21 December 1768´)

Predicative complements. In TimeML, adjectives are considered events when functioning as the head of a predicative complement, as in:

(22) a. *se quedo calma* (`It remained calmed´)
b. *bonancible amanecio el dia* (`The day started peaceful´)

It is common to find verbs denoting natural phenomena (*amanecer,* `dawn', *anochecer,* `to darkle') followed by a temporal NP (*el dia* 'the day') and with a predicative adjective preceding or following the verb (*bonancible,* 'peaceful') (22b).

Copulative verbs and their (noun or adjective) predicative complement are marked up independently. In ModeS TimeBank, these constructions mainly express predications over the weather. They are copulative verbs followed by a noun or adjective referring to natural phenomena (e.g., *nubes* `clouds´ in (23)).

(23) *A medio dia no observe el sol por estar cubierto de nubes* (`At noon the sun does not look to be covered with clouds´)

2.3 Prepositions

Event-denoting PPs most typically express states of some sort. PPs are annotated following the general guidelines for Spanish text (Saurí, et al. a-b). That is, tagging their preposition head events if the PP is a predicative complement.

(24) *no vinieron los pliegos [PP a bordo]* (`The sheets did not come [PP on board]´)
(25) *viramos el ancla [PP a pique]* (`We tacked the anchor [PP to the bottom]´)

2.4. Polysemous expressions: event and time

In ModeS TimeBank, there are a number of expressions that simultaneously refer to parts of the day and natural phenomena, such as *amanecer* `dawn´, *anochecer* `dusk', *poner el sol* `sunset´, *romper el alba* `daybreak´, *obscurecer el día,* `darken the day´, etc.

Verbs like *amanener, anochecer,* and *oscurecer* can be followed by a temporal NP such as *the day, the night,* etc. (26).

(26) a. *al romper el alva* (`At the break of sunrise´)
b. *amanecio el dia con orizontes con algunas nubes* (`The day dawned with horizons with some clouds´)

They receive a double interpretation. On the one hand, they refer to a point in time (the time in which the event occurred) while on the other, they express something that happens or takes place –thus, a situation or event (e.g., *vi que amaneció* `I saw that it dawned'). In this respect, they can be considered dot-objects (Pustejovsky, 95).

(27) *anocheció este día con los orizontes aguazerados por todas partes* (`Nightfall this day with the horizons rains everywhere')

(28) *bonancible amaneció el dia con algunas nubes rojas y el tiempo de bastante mal semblante* (`The day started peacefully with some red clouds and the time of bad face')

Because of their dual nature, these expressions will be tagged both as time expressions and events.

4. Spatial expression in the corpus

As can be expected, in ModeS TimeBank it is common to find place names like *Montevideo* o *Sierra de Maldonado*, and NPs such as: *este puerto* (`this port') o *la referida torre* (` the mentioned tower'). However, it is also quite common to find explicit coordinates as: *en la latitud 4 grados y 19 minutos Norte y longitud de 359 grados y 6 minutos* (`at latitude 4 degrees and 19 minutes North and longitude of 359 degrees and 6 minutes'). Given the genre and domain of the present corpus, coordinate-based expressions are the most common here, as they seek to locate events in time and space with high precision.

Acknowledgements

The authors want to thank the Mercator Group at Technical University of Madrid for supporting and funding this piece of research, and very especially to Miguel Ángel Bernabé for his personal involvement. This work was partially supported by an EU Marie Curie grant to R. Saurí, PIRG04-GA-2008-239414.

References

Guerrero Nieto, M., & Saurí, R. (2011). Annotating Time, Event, and Space Information in Modern Spanish Text. ModeS TimeBank Annotation Guidelines. Version 1.0. Barcelona Media Technical Report, BM 2012-01. Available at: http://comunicacio.barcelonamedia.org/technical_reports/BM2012_01.pdf

Pustejovsky, J. (1995). The Generative Lexicon. MIT press.

Saurí, R., Batiukova, O., & Pustejovsky, J. (2009). Annotating Events in Spanish TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2009-01. Available at http://comunicacio.barcelonamedia.org/technical_reports/BM2009_01.pdf

Saurí, R., Saquete, E., Pustejovsky, J. (2010). Annotating Time Expressions in Spanish. TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2010-02. Available at http://comunicacio.barcelonamedia.org/technical_reports/BM2010_02.pdf

Saurí, R. and T. Badia (to appear, a) Spanish TimeBank. Linguistic Data Consortium.

Saurí, R. and T. Badia (to appear, b) Catalan TimeBank. Linguistic Data Consortium.