TEMANTEX: A Markup Language for Spanish Temporal Expressions and Indicators

Dina Wonsever, Aiala Rosá, Marisa Malcuori, Mathias Etcheverry

Universidad de la Republica, Instituto de Computacion, Facultad de Ingenieria, Uruguay

{wonsever, aialar, mathiase}@fing.edu.uy, marisamalcuori@gmail.com

Abstract. We describe the TEMANTEX annotation scheme for temporal expressions and other lexical indicators of temporality and we analyze a first annotation experience. TEMANTEX is mainly a revision of the markup language TIMEX3, but with some additions and a different treatment for relative expressions. Our alternative proposal is justified for two reasons. First, our system aims to cover other temporality-related lexical elements by defining annotations for what we call temporal indicators, which do not have an equivalent in the TimeML system. Second, regarding temporal expressions, our scheme has relevant differences that improve the annotation process and the interpretation potential. A first task of corpus annotation on a set of 2.300 words, comprising 33 temporal expressions and 35 temporal indicators, showed encouraging results.

Keywords: Markup language, temporal expressions, annotation.

1 Introduction

This document describes the TEMANTEX annotation scheme for temporal expressions and other lexical indicators of temporality and analyzes some annotation experiences. There exist several annotation schemes for temporal expressions, and specifically TIMEX3, included in the TimeML [2] annotation scheme has been widely used, having adaptations for several languages. There are annotated corpora, mainly for English [3], but other languages have been incorporated more recently: Chinese, French, Italian, Korean and Spanish data were incorporated at the TempEval 2 [4]; and enhanced English and Spanish [6] corpora were provided for TempEval 3 [7].

Our alternative proposal is justified for two reasons. First, our system aims to cover other temporality-related lexical elements by defining annotations for what we call temporal indicators, which do not have an equivalent in the TimeML system. For instance, we annotate as temporal indicators terms like *previo/previous*, *siguiente/subsequent* that are not annotated in the Spanish TimeBank, as stated in the annotation guidelines ([5], section 3.2.2). Second, regarding temporal expressions, our scheme has relevant differences that improve the annotation process and the interpretation potential.

Temporal expressions or eTemps (section 2) are linguistic expressions that refer to timeline allocated intervals (or sets of intervals) or to temporal durations. Temporal expressions may include various types of calendar units (parts of a day, days, months, years, etc.), which can have an accurate or vague, absolute or relative reference. They can also consist of terms denoting lengths of time which are usually presented as vague or generic.

Temporal indicators or mTemps, from *marca temporal* in Spanish, (section 3) are an heterogeneous set of elements that influence the temporal interpretation of the text and which are neither temporal expressions *per se* nor are they included within a temporal expression. Unlike eTemps, which refer to intervals or durations of varying lengths on the timeline, mTemps are relational elements or lexical temporality indicators. We have classified in several types the temporal indicators: relational, stages, ordinal numbering, duration, frequency and relative.

TEMANTEX attempts to capture and categorize all the information that might enhance the task of automatic learning of the expressions and the temporal relations in a text. The temporal anchoring of events has multiple applications in tasks such as multi-document summarization, question answering, information retrieval. One of the distinguishing features of our scheme is that it remains as close as possible to the text, excluding the calculation of absolute values made by the annotator through the attribute VALUE, as it happens with TIMEX3. Our scheme also includes a VALUE attribute for which no exact calculations are needed.

A first task of corpus annotation on a set of 2.300 words, comprising 33 eTemps and 35 mTemps, showed encouraging results (section 4).

2 Temporal Expressions

To annotate temporal expressions, we define an eTemp element with the following possible attributes and values, which are explained below:

ATTRIBUTE	VALUES
Type	location, duration, frequency
Mode of reference	absolute, relative
Degree of accuracy	accurate, vague
Value	expression in LDT language
Focus[1]	deictic, anaphoric, age reference, a different speaker
Granularity	second, minute, hour, day, etc.

To mark up an eTemp in the text we must first consider which text segment represents it. In this case we decided to include, as part of the eTemp, all the elements that might contribute to the expression interpretation, as is the case of prepositions and adverbs: <*en la Edad Media/in the middle Ages*>, <*durante dos horas/during two hours*>.

[1] The focus only appears for relative expressions.

For expressions like *el día del tsunami en Japón* (*the day of the tsunami in Japan*), the segment *el día* (*the day*) will be marked up as eTemp. This is a relative expression where the focus is *el Tsunami en Japón* (*the Tsunami in Japan*). This way of working enables us to exclude from eTemps all the elements that are not part of the temporal language and to which we cannot attribute a value.

1.1 TYPE, ACCURACY and MODE OF REFERENCE Attributes

A type is assigned to eTemps by selecting one of the three values of the attribute TYPE: location (chronological location on a timeline: *el 20 de abril / on 20th April*, *el 20 de abril de 1980 / on 20th April, 1980*, *los últimos diez años / the last ten years*, *hace un año / a year ago*), duration (length of time without anchoring to a given point in time: *Esperó durante una hora / She waited for an hour*), and frequency (location on a timeline for a repeated event: *los jueves de 2 a 4 / Thursdays from 2 to 4*, *a menudo / often*). For all cases, the expression refers to a temporal interval or set of temporal intervals (or point in time as an extreme case).

ACCURACY shows if the temporal interval is fully specified (*hoy/today*, *4 horas / 4 hours*, *todos los martes / every Tuesday*) or if the reference is imprecise (*en estos días / these days*, *durante mucho tiempo / for a long time*, *a veces / sometimes*).
MODE OF REFERENCE shows if the expression is absolute, i.e., it does not require additional elements to be interpreted; or relative, i.e., it depends on an additional element, in the text or context of utterance, for its interpretation. This other element, which we call focus, is one of the main distinguishing features between ours and other annotation schemes.

1.2 The FOCUS Attribute

The FOCUS shows how the relative expression is interpreted, its values are: Deictic, Anaphoric, Age-reference, A-different-speaker.

A relative expression anchored to the utterance of the author of the text is assigned a deictic focus, directing us to a constant that is always the date of the document. This is generally the decision adopted in different annotation models. In this case, no additional elements are needed to interpret the temporal expression correctly.

For the three remaining values, the focus will be a text element indicated through the tag focus, and a link between the eTemp and its focus will be added. The most usual scenario for the anaphoric anchor of an eTemp is the anaphoric-type focus. Such is the case in the following example, where *En 1815 / In 1815* (also an eTemp) is the anaphoric focus of the relative eTemp *ese año / that year*.

> ***En 1815*** *Artigas logró que los porteños devolvieran Montevideo a los orientales, y* ***ese año*** *pudo gobernar todo el país. / In 1815 Artigas convinced the porteños to return Montevideo to the orientales, and that year he was able to rule the whole country.*

The anaphoric anchor of an eTemp is an Age-reference focus in expressions like a los 20 años / at 20 years of age, en su primera infancia / in his early chidhood, which direct us to a focus that may be the person's date of birth, if included in the text, or even the person's name. There are previous studies focused on the syntax and auto-

matic recognition of adverbial expressions that are Age-references according to our scheme [1]. In the following example, the eTemp *a los cinco años / at the age of five* has *el 27 de enero de 1756 / 27th Januray, 1756* as Age-reference type focus.

> *Mozart nació el 27 de enero de 1756. A los cinco años ya componía pequeñas piezas musicales. / Mozart was born on 27th January 1756. At the age of five he was already composing short musical pieces.*

When there is a relative temporal expression with a deictic anchor in reported speech, i.e., with a change of speaker, this expression takes on the value A-different-speaker for the FOCUS attribute. A focus tag is assigned to the expression that introduces the reported speech. In the following example, the eTemp *Hoy / Today* has the reporting verb *dijo / said* as its focus.

> *El pasado jueves el presidente dijo: "Hoy iniciamos una nueva etapa en la política cultural" / Last Thursday the President said: "Today we start a new stage in cultural policy."*

1.3 The GRANULARITY Attribute

The GRANULARITY attribute enables us to consign the temporal magnitude to which the expression refers. A wide range of values has been set for this attribute: second, minute, hour, fraction-day, day, fraction-week, week, fortnight, fraction-month, month, fraction-year, year, decade, century, millennium, historical period.

1.4 The VALUE Attribute: The Temporal Description Language LDT

Temporal expressions are described in abbreviated form through the attribute VALUE, expressly naming elements that are implicit or need to be deduced from the context. A temporal description language (LDT) has been defined [8] to describe the expressions within the VALUE attribute. A literature antecedent for LDT is TCNL, Time Calculus for Natural Language, from Han and Kolhase [9].

The name LDT is an abbreviation for Spanish *Lenguaje de Descripción Temporal*. LDT objects are intervals, interval sequences and points. They are arranged in a time line, on which there is an order. Points and intervals are in fact interchangeable with each other, e.g., the temporal expression *December 21, 1980* is seen as an interval in the example a) while in b) it is preferably seen as a point.

a) On December 21st, 1980 I visited some museums and spent the afternoon in the park.
b) A tax moratorium until 21st December 1980 was granted yesterday.

We distinguish in LDT two types of items: basic objects and complex objects. The basic objects are defined by extension and the complex objects are constructed from basic ones and other elements by means of some operations.

1.4.1 Basic Objects

The basic objects correspond to the usual calendar units as well as to names of special events and historical periods. They are noted by abbreviations. Some special basic

objects, necessary for the definition of complex objects, are also defined. In what follows we define different kinds of basic objects:

Calendar units, culturally recognized units.

- usual elements in our calendar system: day (di), month (me), year (yy), Century (sg), Millennium (mi), etc.
- parts of the day (morning, afternoon, etc.) and seasons.
- units of time system (hour, minute, second, etc.)
- names of festivals, Christmas, Easter, New Year, etc.
- culturally recognized names of historical periods (Middle Age, Industrial Revolution, BC (Before Christ), etc.)

Special items: generic intervals.

- u: universal interval
- x: generic interval

Special items: referents for temporal anchoring.

- fd: deictic focus (unique, the utterance moment)
- fa: anaphoric focus (variable: different events and temporal expressions in text)

1.4.2 Operations in LDT

The basic objects rarely appear isolated in a temporal expression. They usually appear with name (e.g., *September* is the month with basic object with name me9 in our notation), quantified (*the last days of September*) or in more complex constructions (*September 21, 1908*, a combination of several basic named units). Notice that an expression such as *September 21, 1908* can be seen as a successive application of restrictions on basic units:

(a) the basic unit day is restricted to a specific day, 21 → di21,

(b) the basic unit month is restricted to a specific month, 9 → me9,

(c) the basic unit year is restricted to a specific year, 1908 → aa1908,

(d) the term di21 is restricted to a specific month → di21-me9,

(e) di21-me9 expression is restricted to a specific year → di21-me9-aa1908.

The applied operations include on one side naming, defining, quantifying, restriction or selection into larger units, and, on the other side union, definition of regions and shift.

1.4.2.1 Naming

It is used to select a specific unit within a class of calendar units. The possible names are ordinal units:

- day, abbreviated ds if it is referred as a day of the week and dm if it is referred as a day of the month. It has two sets of names: ds - 1,2, ..., 7 (Monday is 1) and dm - (1,2, ..., 31)
- month, abbreviation me- (1,2, ..., 12)
- year, abbreviation aa - 1, 2, ... (optional AC)
- century, abbreviation sg - 1, 2, ... (optional AC)
- millennium abbreviation mi - 1,2, ... (optional AC)

We write the abbreviation of the unit and then the name (e.g., me10), followed by AC if applicable.

1.4.2.2 Selection

The selection is an operation with 3 arguments: selected temporal object, selection type, and unit or temporal object on which the selection is made. This operation is mainly used to express quantification on temporal units.

Notation : sel (o1, selectMode, o2)

We select a subset of objects o1 from o2 objects according to selectMode mode of selection.

Example: *the last days of December* → sel (di, last, me12)

The naming is a particular case of selection where the second object o2 does not intervene.

There is a wide range of selection modes:

- Ordinal
- Cardinal (special case of quantifier)
- Quantifier (one, all, some, many, few, several, most, late, early, middle, averaging, etc.). We also have constructions like *almost all*, etc.

The selected object (o1) may be the special object x (time, "temporal substance") previously defined.

*In **early April** the leaves begin to fall* → sel (x, early, me4)

It may also be necessary to use the universal interval u

*In **some years** we get good crops.* → sel (aa, some, u)

It may also be necessary to use the two special items together.

For a long time it was believed that the sun moved around the earth. → sel(x, mucho, u)

1.4.2.3 Duration Expressions

The basic units mentioned in 2.4.1 serve both to build expressions of location and duration. The duration expressions are basically composed of a calendar unit and a magnitude. We use the syntax Magnitude.TemporalObject

Examples:

*Today I have studied for **4 hours**.* → **4.hh**

*It took **a few days** to paint the house.* → **algunos.di**

There may be more than one unit in the same temporal expression, in this case we construct a compound expression whose denotation is the union of both time amounts, separating by hyphens the subexpressions in the different units.
*The ride to Pando takes **1 hour and 20 minutes**.* → **1.hh-20.mt**

1.4.2.4 Selection by Proximity (closer, mp)

We use the notation mp(Object, P, Dir) to select the temporal unit of type Object closest to the point P, in the direction specified in Dir (previous (neg), posterior (pos) or matching (ig)).

It is primarily used to make explicit the position regarding the textual focus on constructions with an implicit focus.

Examples:

*We will meet **in November.*** → mp (me11, fd, pos)

*I saw him **last Thursday.*** → mp (ds4, fd, neg)

In general terms, the context determines whether there is a reference to a textual (anaphoric) or to the deictic focus (usually the creation date of the document). Notice that if you use the anaphoric focus a temporal link with an explicit focus has to be signaled. The direction (previous, subsequent) is also recovered from the context. The case of coincidence (direction = ig) is used for expressions where there is a direct reference to a temporal expression.

We met many times in that year. → mp (aa, fa, ig)

The **mp** operation includes the access to a temporal coordinate, a deictic or anaphoric focus in the previous examples. In other words, we recovered the year, month or day or other unit of a time point. This is a kind of projection operation, in terms of some temporal unit.

*They met on **April 80** and in **December of that same year** they married.*

→ **me4-aa1980**

→ **me12-mp(aa,fa,ig)**

And there exists also an anaphoric link between *April 80* and *that same year* (not shown in the previous expressions).

1.4.2.5 Regions Definition

Region are time intervals. We can build them specifying a point, a length and a direction (*rd operator*), or by specifying both extremes (*rr operator*).

Regions by point and directions

rd (Point, Duration, Direction)

Definition of a temporal interval (region) from Point, Length and direction (neg, anterior; pos, later; ent, environment-centered in P)

In the last 40 years *we observed some climate changes.* → rd (fd, 40.aa, neg)

Region as a range
rr (Point 1, Point 2)
Build the range from Point 1 to Point 2

(i) *I lived with my brother between April 10 and April 20, 2010.*
→ rr(dm10-me4-aa2010, dm20-me4-aa2010)

(ii) I went to the movies several times between April 10 and April 30, 2010.
→ sel (time, number, rr (me4-aa2010-DM10, DM20-me4-aa2010))

(iii) I went to the movies several times between April 10 and April 30 I this year.
→ *sel (time, number, rr (me4-aa2010-DM10, DM20-me4-aa2010)) rr (DM10-me4-mp (aa, fd, ig), DM20-me4- mp (aa, fd, ig))*

To solve (iii) we must consider a further complication. (iii) it is similar to (ii), except that instead of referring to 2010, we use the term *this year*, referring to the year of the deictic focus.

(iv) I lived with my brother between April 10 and April 30.

This example is again more complex than the example (i). We refer to the closest period between April 10 and 30 that precedes the utterance. The expression is a range expression, both extremes (April 10 and April 30) are defined relative to deictic focus.

→ rr(dm10-mp(me4,fd,neg), dm20-mp(me4,fd,neg))

(v) I am going to the movies since two months ago.

In the example (v) we have a region defined by a start point and a direction. That starting point is defined by a shift operation applied to the deictic focus, we see the example in section 4.2.6

1.4.2.6 Shift

The expression desp(P, distance, direction) defines a point P' at a distance Distance preceding (direction = neg) or succeeding (sirection = pos) point P.

Examples:

*(i) We met t****hree years ago****.* → despl (fd, 3.AA, neg)
(ii) ***Tomorrow*** *we're going to see each other.* → despl (fd, 1.di, pos)
(iii) It is ***three years since*** *we are seeing each other.* → rr (despl(fd, 3.aa, neg), fd)

Notice that the verbal aspect affects the interpretation of the temporal expression. The same temporal expression (*3 years*) was interpreted as a point in time in (i) whereas in (iii) it should be interpreted as a region.

1.4.2.7 Union

To represent expressions like *April 4 and 5* it is desirable to have an operator able to form the aggregate or group containing 4 and 5. We have then defined the **union operator**: **un (O1, O2)**, where O1 and O2 are objects temporary.

The course is held on Tuesdays and Thursdays in the second semester.

→ sel (un (ds2, ds4), todo,sm2)

2 Temporal Indicators, *mTemp*

As mentioned in the Introduction, temporal indicators (*mTemps*) are elements of a relational nature or lexical temporality indicators. Even though an expression such as *hasta el año 1925 / until 1925* will be considered an *eTemp* and the element *hasta / until* is part of it, in an expression such as *hasta la elección / until the election*, the element *hasta / until* will be considered as an *mTemp* that links events, since it is not included in any temporal expression. But the adverb *hasta* / *until* is, without a doubt, a relevant element for the temporal analysis of texts, for example, for the temporal sequencing of events within the text. The vocabulary linked to temporality includes several types of temporal indicators which are classified in one of the following classes: relational, stages, ordinal numbering, duration, frequency and relative.

2.1 Relational and Relative mTemps

mTemps such as *antes de que / before , después / after , mientras / while* take on the value Relational. These determine a relation between two events which could be anteriority, posteriority, simultaneity, inclusion, etc. They could be applied to localized temporal intervals (*antes del jueves / before Thursday*) integrating eTemps, or to events, in which case they would be annotated as relational indicators.

In addition, relative mTemps provide a time reference with regard to the moment of utterance or any other moment expressly or implicitly mentioned in the text. They are analogous to relative eTemps and, as such, their focus can take on the values Deictic, Anaphoric or A-different-speaker. Therefore, in *El mes pasado el presidente dijo: "la **próxima** elección será un éxito" / Last month the President said: "the next election will be a success", próxima /next* is a relative temporal indicator of the type different speaker, with focus on the moment of utterance (*el mes pasado / last month*). Note that *"la semana próxima / next week"* is an eTemp, and in this case *próxima / next* is not analyzed as a temporal indicator.

2.2 Stage and Order mTemps

These mTemps focus on a stage within the development of the event (***Al principio** la guerra fue muy cruenta / At first, the war was a bloodshed*), or express lexically, generally in an indirect manner, the temporal sequencing of the events (*Lo entendí recién en la **segunda** clase / I was able to understand it in the second class*).

2.3 Duration and Frequency mTemps

This class is basically comprised by some verbs and adjectives, such as *durar*/*last* and *asiduas/frequent* in the following examples: *La entrevista **duró** toda la tarde / The interview lasted all afternoon, Se acostumbró a sus **asiduas** visitas / He grew used to his frequent visits.*

3 Annotation: Characteristics and Problems

In order to validate the scheme proposed, two previously trained annotators annotated a corpus of 2,300 words containing 33 eTemps and 35 mTemps.

Regarding eTemps determination, we found consistency between the annotators: annotator 1 (A1) marked 32 eTemps and annotator 2 (A2) marked 38. A2 twice marked as eTemps expressions that correspond to temporal indicators; in one case, he annotated an eTemp as focus; the three remaining expressions, which only A2 marked, do not correspond to eTemps nor mTemps.

In addition, when determining the values of the attributes we did not find significant differences between the annotators.

Annotation of temporal indicators was more problematic. Out of the 35 mTemps present in the corpus, A1 marked 30 correctly and A2 only 23. Moreover, we detected a significant number of false positives: A1 annotated 5 and A2 annotated 10. These values show that it is necessary to adjust the definition of temporal indicator.

In spite of the apparent complexity of the LDT language, the annotation of the value attribute proved simple and errorless. We looked at more than 50 cases, in a corpus of journalistic texts, and we didn't notice any error. Interestingly, almost all cases used the mp (closest) operator with a deictic focus.

4 Discussion, Future Work

We worked on a proposal for modeling temporal expression and other lexical elements that convey temporal meaning. The model is mainly compatible with TimeML, extending it with temporal indicators and new classes for different types of relative expressions and their related focus. A first task of corpus annotation showed encouraging results, suggesting the pertinence of our model.

Our plan is to proceed to the automatic recognition and interpretation of temporal expressions and indicators, as an intermediate task for text understanding. As a first step, we are experimenting with the recognition of the extent of temporal expression, using neural networks over a vector based representation of texts. For the interpretation, our plan is to extract the relevant information form the value attribute, that is, the expression in LDT language.

References

1. Galicia-Haro, S.N., Gelbukh, A.F.: Supervised Recognition of Age-Related Spanish Temporal Phrases. In: Proceedings of the 8th Mexican International Conference on Artificial Intelligence, MICAI, pp. 145–156 (2009)

2. Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Proceedings of the Fifth International Workshop on Computational Semantics. IWCS-5 (2003)
3. Pustejovsky, J., Verhagen, M., Saurí, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, R., Setzer, A.: TimeBank 1.2. LDC catalog ref. LDC2006T08 (2006)
4. Pustejovsky, J., Verhagen, M.: SemEval-2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2). In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics. SEW-2009 (2009)
5. Saurí, R., Saquete, E., Pustejovsky, J.: Annotating Time Expressions in Spanish TimeML. Annotation Guidelines (Version TempEval-2010). Barcelona Media Technical Report 2010-02 (2010)
6. Saurí, R., Badia, T.: Spanish TimeBank 1.0. LDC catalog ref. LDC2012T12 (2012)
7. UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J.F., Pustejovsky, J.: SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In: Proceedings of the 7th International Workshop on Semantic Evaluation, pp. 1–9, ACL SemEval (2013)
8. Wonsever, D., Malcuori, M., Etcheverry, M.: Esquema de anotación de expresiones y marcas temporales. Reporte técnico, serie: 0797–6410, PEDECIBA-Informática, https://www.fing.edu.uy/inco/pedeciba/bibliote/reptec/TR1115.pdf (2011)
9. Han, B., Kohlhase, M.: A Time Calculus for Natural Language. In: Proceedings of the 4th Workshop on Inference in Computational Semantics, Nancy, France (2003)