1. **Corpus name**: Spanish TimeBank 1.0.

2. **Authors**
   - *Roser Saurí* (contact person)
     - email: roser.sauri@barcelonamedia.org
     - phone: +34 93 238 1400
   - *Toni Badia*
     - email: toni.badia@barcelonamedia.org
     - phone: +34 93 238 1400

3. **Data type**: Text.

4. **Languages**: Spanish (spa).

5. **Description of the corpus**

   The Spanish TimeBank Corpus contains 210 documents (mostly news reports), which have been annotated with time and eventuality information according to the TimeML scheme (Pustejovsky et al., 2005), now accepted as an international cross-language ISO standard (ISO WD 24617-1:200). Specifically, time information in the Spanish TimeBank is annotated with the following levels:

   - **Events** (tag **EVENT**): Marking up different types of actions (activities, transitions, etc.) as well as states. The annotated expressions can belonging to different parts of speech, such as verbs, nouns, or adjectives. Event entities are further specified with attributes concerning grammatical aspects of the tagged expression (e.g., part of speech, verb form, tense, aspect, mood), as well as semantic information (e.g., event class).
   - **Time expressions** (tag **TIMEX3**): This tag includes expressions of calendar dates, times of day (TOD), durations, and sets. Their interpretation is normalized according to an extension of the ISO 8601 format for dates and time-related data. Other pieces of relevant information (for example, the presence of modifiers or the reference to other temporal expressions in the text) are encoded in attributes.
   - **Temporal relations** among events and time expressions (tag **TLINK**), which essentially signal: precedence (before, after), inclusion (is included, includes), and simultaneity. Temporal relations in TimeML can hold between any event and timex entity. However, in the Spanish TimeBank they have been constrained to the 4 types of relations targeted in the TempEval competition (Verhagen et al., 2007, 2010), which are:
• Temporal relations between an event and a time expression appearing in a strong syntactic relation. That is, either when the event syntactically dominates the time expression, or when both event and time expression occur within the same noun phrase.
• Temporal relations between each event and the Document Creation Time (DCT).
• Temporal relations between the two main events in consecutive sentences.
• Temporal relations between two events in a relation of syntactic dominance.

For marking up the Spanish TimeBank, we tailored the TimeML annotation scheme and guidelines, originally for English data, into the specifics of the Spanish language. For example, event expressions in Spanish present distinctions of verbal mood (e.g., indicative, subjunctive, conditional, etc.) and grammatical aspect (e.g., imperfective) which are absent in English. Therefore, the following annotation guidelines have been developed:

– For annotating events: Saurí et al. (2009).
– For annotating time expressions: Saurí et al. (2010)
– For annotating temporal relations: Saurí (2010).

In terms of both the amount and the nature of the annotated data, the present corpus is the Spanish correlate of TimeBank 1.2, developed for English text (Pustejovsky et al., 2006). Furthermore, it has a Catalan twin, the Catalan TimeBank (Saurí & Badia, 2012), a brother for Modern Spanish text, also annotated with spatial information (Guerrero Nieto & Saurí, 2012), and belongs to the family of TimeBanks developed within the TimeML framework for other languages, such as: French (Bittar, 2010);1 Italian (Caselli et al., 2011); Korean (Im et al., 2009); and Chinese (under development). The existence of these corpus resources with a common layer of annotated information can be of great benefit to the community, specially for work involving multilingual temporal extraction and processing, such as multilingual text entailment, opinion mining, or question answering.

An earlier version of this corpus has been used in TempEval-2, one of the tasks in the SemEval International Workshop on Semantic Evaluations (Verhagen et al., 2010).

6. Data Sources

The texts constituting the Spanish TimeBank are mainly news reports but include some fiction as well. They have been obtained form Ancora-Sp, the Spanish part of the AnCorA corpus (Taulé et al., 2008). AnCorA is a remarkable resource in that it provides annotation for a number of linguistic levels, including constituent structure, syntactic functions, dependencies, verb semantic class, argument structure, and thematic roles. This information is not included in the current release, but can be easily mapped to the present annotations.

7. Annotated Data

The Spanish TimeBank contains 210 documents with over 75,800 tokens (including punctuation marks) and 68,000 tokens (excluding punctuation). Table 1 shows the total number of tokens and annotated entities (events, timexes and tlinks of each type).

1See also: http://www.linguist.univ-paris-diderot.fr/abittar/french-timebank/.
Tags **EVENT** and **TIMEX3** have different frequencies for *annotated tokens* and *actual entities*. *Annotated tokens* indicate how many tokens in the text have been marked up with the tag in question, whereas *actual entities* provide the actual number of tags. For events, this is higher than the number of annotated tokens, given that the same event mention can express multiple events, encoded each in an independent **EVENT** tag. By contrast, the number of actual **TIMEX3** entities is lower than the number of tokens marked up as such, given that a time expression may include several tokens.

The corpus has been entirely double-annotated, except for the tlinks in a subordination relation. The double annotation has been carried out by graduate linguistics students using the Brandeis Annotation Tool (BAT)\(^2\), and cases of disagreement have been adjudicated by a third person.

### 8. Corpus structure and data attributes

The whole corpus markup is standoff, represented through a set of 11 tables which can be easily loaded into a DB. Each table is contained in an independent tab-separated file. The following lists each table file and its size in bytes, by alphabetical order.

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Size (in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>base-segmentation.csv</td>
<td>188,487</td>
</tr>
<tr>
<td>dct.csv</td>
<td>5,932</td>
</tr>
<tr>
<td>event-attributes.csv</td>
<td>105,124</td>
</tr>
<tr>
<td>event-extents.csv</td>
<td>45,866</td>
</tr>
<tr>
<td>sentences.csv</td>
<td>42,128</td>
</tr>
<tr>
<td>timex-attributes.csv</td>
<td>7,289</td>
</tr>
<tr>
<td>timex-extents.csv</td>
<td>9,577</td>
</tr>
<tr>
<td>tlinks-dct-events.csv</td>
<td>52,265</td>
</tr>
<tr>
<td>tlinks-event-timex.csv</td>
<td>47,353</td>
</tr>
<tr>
<td>tlinks-main-events.csv</td>
<td>90,251</td>
</tr>
<tr>
<td>tlinks-subordinated-events.csv</td>
<td>281,456</td>
</tr>
</tbody>
</table>

\(^2\)http://www.timeml.org/site/bat/
The structure of each table is described next. Attributes sharing the name across tables encode the same information.

Table: **sentences.csv**  
Containing data relative to each sentence in the corpus documents. Data attributes:

- docId       Document ID.
- sentId      Sentence ID (relative to each document).
- sentTxt     Sentence text.

Note that the sentence text does satisfy the standard punctuation conventions but presents each token, including punctuation marks, separated by blank spaces from its neighbors. This is due to the fact that the original data obtained from AnCora was only in verticalized format and did not preserve the original formatting of the text. Hence, we opted for reconstructing the text into one-line sentences for the sake of readability, while avoiding to restore the text into its presumed original format.

Table: **base-segmentation.csv**  
Containing the corpus tokens in a verticalized format, as inherited from the original files in AnCora. Corpus tokens include words and punctuation marks, but not blank spaces and other formatting characters. Each token specifies the document and sentence where it belongs. Data attributes:

- docId       Document ID.
- sentId      Sentence ID (relative to each document).
- tokId       Token ID (relative to each sentence in the document).
- tokTxt      Token text.

Table: **dct.csv**  
Presenting the Document Creation Time (DCT) of each document in the corpus. Data attributes:

- docId       Document ID.
- dct         Document Creation Time in YYYY-MM-DD format.

Table: **event-extents**  
Information relative to EVENT entity extents. In TimeML, event extents span over only one token, but the annotation can indicate if more tokens are linguistically involved (e.g., when the event expression is a phrasal verb or a multiword construction) by means of the attribute isMultiWord. Similarly, a further attribute (cardinality) indicates whether the expression refers to actually more than one event in the world. Data attributes:

- docId       Document ID.
- sentId      Sentence ID.
- tokId       ID of the token affected by the tag.
- tag         Tag name (here, EVENT).
- tagId       Tag ID.
- **cardinality**: Integer expressing the number of events in the world that are referred to by the current tag.
- **isMultiWord**: Boolean value (y/n) indicating whether the event expression includes further tokens in the text, in addition to the current one. See the guidelines for further details.

**Table: event-attributes**  
Presenting event attributes information. Data attributes:

- **docId**: Document ID.
- **sentId**: Sentence ID.
- **tokId**: Token ID.
- **tag**: Tag name (here, EVENT).
- **tagId**: Tag ID.
- **tagInstanceId**: Event instance ID. Recall that one event expression may refer to several events in the world.
- **pos**: Part of speech. Possible values: ADJECTIVE, NOUN, PREP, VERB, OTHER.
- **vform**: Verbal form. Distinguishing among non-finite verbal forms. Possible values: GERUNDIVE, INFINITIVE, PARTICIPLE, NONE.
- **tense**: Grammatical tense. Possible values: PAST, PRESENT, FUTURE, NONE.
- **aspect**: Grammatical aspect. Possible values: PERFECTIVE, IMPERFECTIVE, PERFECTIVE_PROGRESSIVE, IMPERFECTIVE_PROGRESSIVE, NONE.
- **mood**: Verbal mood. Possible values: INDICATIVE, SUBJUNCTIVE, CONDITIONAL, NONE.
- **polarity**: Polarity of the event expression. Possible values: POS, NEG.
- **class**: Event class. Possible values: ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE, PERCEPTION, REPORTING, STATE.

**Table: timex-extents.csv**  
Information relative to TIMEX3 entity extents, which can consume several tokens, contrary to the annotation of EVENTS. Data attributes:

- **docId**: Document ID.
- **sentId**: Sentence ID.
- **tokId**: ID of the token affected by the tag.
- **tag**: Tag name (here, TIMEX3).
- **tagId**: Tag ID.

**Table: timex-attributes.csv**  
Presenting the information concerning TIMEX3 attributes.

- **docId**: Document ID.
- **sentId**: Sentence ID.
- tokId Token ID.
- tag Tag name (here, TIMEX3).
- tagId Tag ID.
- type Type of time expression. Possible values: DATE, TIME, DURATION, SET.
- val Time expression value, normalized according to an extended version of the ISO 8601.
- mod Modifier. Possible values: APPROX, BEFORE, AFTER START, MID, END, EQUAL OR MORE, EQUAL OR LESS, MORE THAN, LESS THAN.
- anchorTimeId ID of the time expression which the current TIMEX3 is anchored to.
- beginPoint ID of the time expression denoting the beginning point of the duration expressed by the current TIMEX3.
- endPoint ID of the time expression denoting the ending point of the duration expressed by the current TIMEX3.
- quant Quantifier over the set denoted by the current TIMEX3.
- freq Frequency of the temporal set denoted by the current TIMEX3.

Table: tlinks-dct-events.csv Presenting the tlinks that hold between each event and the Document Creation Time. Data attributes:

- docId Document ID.
- eId Event ID.
- eiId Event instance ID.
- timexId ID of the DCT.
- linkId Tlink ID.
- relType Type of temporal relation.
  Possible values: after, before, before-or-overlap, overlap, overlap-or-after, vague.

Table: tlinks-event-timex.csv Presenting the tlinks that hold between events and timex3 in specific syntactic contexts. Data attributes:

- docId Document ID.
- eId Event ID.
- eiId Event instance ID.
- timexId ID of the related timex.
- linkId Tlink ID.
- relType Type of temporal relation.
  Possible values: after, before, before-or-overlap, overlap, overlap-or-after, vague.

Table: tlinks-main-events.csv Presenting the tlinks that hold between two main events in consecutive sentences. Data attributes:
- **docId**: Document ID.
- **eId_1**: Event ID of the first event in the TLINK.
- **eiId_1**: Event instance ID of the first event.
- **eId_2**: Event ID of the second event in the TLINK.
- **eiId_2**: Event instance ID of the second event.
- **linkId**: Tlink ID.
- **relType**: Type of temporal relation.
  Possible values: after, before, before-or-overlap, overlap, overlap-or-after, vague.

**Table: tlinks-subordinated-events.csv** Presenting the tlinks that hold between two main events in a relation of syntactic subordination. Data attributes:

- **docId**: Document ID.
- **eId_1**: Event ID of the subordinating event.
- **eiId_1**: Event instance ID of the subordinating event.
- **eId_2**: Event ID of the subordinated event.
- **eiId_2**: Event instance ID of the subordinated event.
- **linkId**: Tlink ID.
- **relType**: Type of temporal relation.
  Possible values: after, before, before-or-overlap, overlap, overlap-or-after, vague.

9. **Directory structure**

  **doc/** Documentation related to the present release. Containing: this readme file as well as the annotation guidelines used for annotating the corpus.

  **data/** Files containing the corpus annotation. Each file corresponds to one of the tables presented above.

10. **Copyright**

    The annotations in this data collection are copyrighted by the Authors. User acknowledges and agrees that: (i) as between User and Authors, Authors own all the right, title and interest in the Annotated Content, unless expressly stated otherwise; (ii) nothing in this Agreement shall confer in User any right of ownership in the Annotated Content; and (iii) User is granted a non-exclusive, royalty free, worldwide license (with no right to sublicense) to use the Annotated Content solely for academic and research purposes.

    Note: The textual news documents annotated in this corpus have been collected from a wide range of sources and are not copyrighted by Authors. User acknowledges that the use of these documents is restricted to research and/or academic purposes only.
11. Acknowledgements

The current data have been annotated by the following contributors (alphabetically sorted): Joan Banach, Judith Domingo, Pau Giménez, Jimena del Solar, and Teresa Suñol. The adjudication has been in charge of Jimena del Solar and Roser Saurí.

This work has been supported by a EU Marie Curie International Reintegration Grant (PIRG04-GA-2008-239414) to Roser Saurí.

References


