ModeS TimeBank 1.0

Corpus Documentation

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

1. Corpus name: ModeS TimeBank 1.0

2. Authors

Marta Guerrero Nieto **(contact person)**
Mercator Group
Technical University of Madrid
Email: m.guerrero@topografia.upm.es, guerreronieto.marta@gmail.com

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

3. Data type: Text

4. Languages: Spanish (sp)

5. Corpus description

The Modern Spanish TimeBank corpus (**ModeS TimeBank**) is a corpus of Modern Spanish (17th and 18th Centuries) annotated with temporal and event information according to TimeML, and spatial information following the SpatialML scheme.

SpatialML (Mani et al., 2008) is a specification language for annotating and normalizing spatial expressions by means of geographic coordinates. Now, it is being revised and expanded with the addition of semantics for spatial movement (Pustejovsky et al., 2010). To the best of our knowledge, there are no other corpora presently annotated with this specification scheme on (present or historical) Spanish text.

On the other hand, **TimeML** (Pustejovsky et al., 2005) is a specification language for annotating eventualities and time expressions in natural language, as well as the temporal relations these hold among them, thus facilitating tasks of extraction, representation and exchange of temporal information. Recognizing temporal information is essential for any system of Natural Language Processing (NLP) concerned on tasks of information extraction. In recent years, we have seen an increasing development of resources and tools for temporal information extraction and reasoning. However, access to this information still remains a topic of research. The most important properties of TimeML are: the interpretation of time expressions and event mentions in text, and the sequencing of events by means of their temporal anchoring. Currently, there exist TimeML-annotated corpora in seven different languages: English (Pustejovsky et al. 2006), Spanish (Saurí and Badia, to appear-a), Catalan (Saurí and Badia, to appear-b), Italian, French, Korean and Chinese[1], some of which were used in the TempEval-2 evaluation (Verhagen et al., 2010). However, there are no references of equivalent corpora containing historical varieties of these languages, as is the case of the present one.

6. Data source

ModeS TimeBank contains 102 documents reporting a sea crossing cruise by a ship called *La Princesa*, which took place from December 1768 to April 1769. There exist copious logbooks from that period, which not only give information of shipping routes, but also provide valuable data about information flows, commercial agents and social networks.

Focusing on that very specific topic gives a homogeneous character to the data and helps illustrate the linguistic model used in this type of documents in terms of both language register as well as linguistic structures employed. The linguistic representativeness of the data is determined by the corpus strong degree of specialization. In this way, it shows a particular linguistic modality and provides complete and balanced data. The corpus original manuscript is preserved in the *Archivo General de Indias* ("General Archive of the Indies") and is also available online at http://pares.mcu.es/. The documents have been obtained as scanned jpg files and have therefore required a process of manual transcription.

The current corpus was created within the framework of the **DynCoopNet** project (Dynamic Complexity of Self-Organizing Commercial Networks-Based Cooperation in the First Global Age <http://dyncoopnet-pt.org>), focused on the study of trade network cooperation over the 14th-19th Centuries, and which took a wide variety of information sources into consideration, very significantly maps, charts and databases, but also natural language documents. These documents provide valuable information concerning the when, how, and where of the events reported there, but also regarding how long they lasted, how often they took place, etc.

7. Annotated Data

Data in ModeS TimeBank have been tokenized and POS-tagged, and have been annotated with space, time and event information according to SpatialML and TimeML specification schemes. More specifically, the entities annotated in the corpus are the following:

- **Events** (tag EVENT, from TimeML). These include finite and non-finite verbal constructions, nominalizations, nouns, adjectives and prepositional phrases.
- **Temporal expressions** (tag TIMEX3, from TimeML), including expressions of dates, times, durations and frequencies, both precise and vague.
- **Spatial expressions** (tag PLACE, from SpatialML), for proper and common nouns, adjectives, adverbs or spatial coordinates.

The annotation process has been guided by the annotation guidelines for marking up temporal and event information in Spanish text (Saurí et al. 2009, 2010a, 2010b), as well as the specific guidelines developed for the present corpus in order to deal with expressions that are idiosyncratic of Modern Spanish (Guerrero Nieto and Saurí, 2011).

[1] These corpora can be found on http://timeml.org/site/timebank/timebank.html

8. Corpus files

The corpus data (metadata and annotation layers) is structured into tables, each of them stored in an independent file

46 KB	event_extents.txt
75 KB	place_extents.txt
153 KB	timex3_extents.txt
170 KB	place_attributes.txt
349 KB	timex3_attributes.txt
663 KB	base-segmetation.txt

9. Corpus Structure

There are two types of files: extent and attribute files. Extent files contain the following information, displayed in columns:

lex_id	Numeric. Global token ID (i.e., in the whole corpus).
file_name	String. Name of the corpus document.
sentence_id	Numeric. Sentence ID in the corpus document.
token_id	Numeric. Token ID in the sentence.
label	String. Tag name (event, timex3, place)
label_id	Numeric. Tag ID.

On the other hand, attribute files contain the following data, displayed in columns as well:

lex_id	Numeric. Global token ID (i.e., in the whole corpus).
file_name	String. Name of the corpus document.
sentence_id	Numeric. Sentence ID in the corpus document.
token_id	Numeric. Token ID in the sentence.
label	String. Tag name (event, timex3, place).
attribute_name	String. Tag attribute name.
attribute_value	String. Attribute value.

Finally, the corpus textual contents are stored in the *base_segmentation* file, which contains the following information, displayed in columns:

lex_id	Numeric. Global token ID (i.e., in the whole corpus).
file_name	String. Name of the corpus document.
sentence_id	Numeric. Sentence ID in the corpus document.
token_id	Numeric. Token ID in the sentence.
word	String. Corpus token.

10. Directory structure

data/

annotated/

Files containing the corpus annotations. Each file corresponds to one of the tables presented above.

original/

Files containing the original documents constituting the corpus.

doc/ Documentation relative to the present release. Containing: this *readme.pdf* file as well as the annotation guidelines used for marking up the corpus.

Copyright

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 in order to make this linguistic resource possible. 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

Guerrero Nieto, M., Saurí, R., & Bernabé, M.A. (2011). ModeS TimeBank: A Modern Spanish TimeBank Corpus. Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 259-267.

Mani, I., Hitzeman, J., Richer J., Harris, D., Quimby R., & Wellner B. (2008). SpatialML: Annotation scheme, corpora, and tools. Proceedings of the Sixth International Language Resources & Evaluation (LREC'08).Pustejovsky, J. & Moszkowicz, J. (2011). The Qualitative Spatial Dynamics of Motion in Language", Spatial Cognition and Computation, 11:1-32.

Pustejovsky, J., Moszkowicz J., & Verhagen, M. (2010). The recognition and interpretation of motion in language. CICLing2010, 236-256.

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. (2010). Annotating Temporal Relations in Catalan and Spanish. TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2010-04. Available at http://comunicacio.barcelonamedia.org/technical_reports/BM2010_04.pdf

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

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