TimeBankPT

I. Basic Information

1.1. Corpus information

TimeBankPT, a TimeML annotated corpus of Portuguese, is the first corpus of Portuguese with rich temporal annotations (i.e. it includes annotations not only of temporal expressions but also about events and temporal relations).

The annotation scheme used is similar to TimeML. TimeBankPT is the result of adapting the English corpus used in the first TempEval challenge to the Portuguese language.
TimeML is a rich annotation scheme in so far as it allows for the annotation of several phenomena related to time: the times, dates and periods denoted by temporal expressions, events, temporal relations, etc.

Some of the features of TimeBankPT:

- It uses the new Portuguese spelling (official document describing it, Wikipedia article).
- It was automatically checked for errors using reasoning code.
- It contains around 70,000 words of text, divided in a train set and a test set.
- It contains annotations for events, temporal expressions and temporal relations.

	Train Set	**Test Set**
Sentences	2,281	351
Word Tokens		
According to white space	60,782	8,920
Splitting contractions and detaching punctuation	68,351	9,829
Events	6,790	1,097
Temporal Expressions	1,244	165
Temporal Relations	5,781	758

For more details, see `http://nlx.di.fc.ul.pt/~fcosta/TimeBankPT/`.

1.2. Representation of the corpora (flat files, database, markup)

The corpus is split into several .tml files (in an XML format), and they are all grouped in a compressed .zip file.

1.*3. Character encoding*

The characters are in UTF8 encoding.

II. Administrative Information

2.1. Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de

Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)
This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR
This resource is licensed for research purposes only, with no redistribution, nor derivatives allowed.

III. Technical Information

3.*1. Directories and files*
The archive that can be uploaded on the Meta-Share is a .zip file with two directories (one with the training data and another one with the test data). Each of these directories contains several .tml files (which are in an XML format). Each of the .tml files corresponds to a document (news article, etc.) annotated in TimeML.

3.2. Data structure of an entry
Each .tml file is divided into sentences, and the sentences are segmented into tokens with time annotations.

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*
The corpus is composed by 68,351 words with 476 *KB* compressed (2.4 *MB* uncompressed) for disk storage.

IV. Content Information

4.*1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*
This is a monolingual and annotated corpus.

4.*2. The natural language(s) of the corpus*
TimeBankPT is in Portuguese and adopts the recent spelling reform.

4.*3. Domain(s)/register(s) of the corpus*
Newspaper texts, narratives and other such texts describe events which occur in time and specify the temporal location and order of these events.

4.4. Annotation in the corpus (if an annotated corpus)

4.4.1. Types of annotation (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

```
<s>A polícia <EVENT eid="e76" class="I_ACTION" stem="impedir" aspect="NONE"
tense="PPI" polarity="POS" pos="VERB">impediu</EVENT> a multidão de <EVENT
eid="e49" class="OCCURRENCE" stem="chegar" aspect="NONE" tense="INF" polarity="POS"
```

pos="VERB">**chegar**</EVENT> **ao consulado jugoslavo na baixa de Istambul, mas** <EVENT eid="e77" class="I_ACTION" stem="permitir" aspect="NONE" tense="PPI" polarity="POS" pos="VERB">**permitiu**</EVENT> **que se** <EVENT eid="e13" class="STATE" stem="manifestar" aspect="NONE" tense="PIC" polarity="POS" pos="VERB">**manifestassem**</EVENT> **em ruas próximas**.</s>

4.4.2. Tags (if POS/WSD/TIME/discourse/etc – tagged or parsed)

It does not apply.

4.4.3. Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)
It does not apply.

4.4.4. Attributes and their values (if annotated)
In the annotation scheme that is employed, words that denote events are enclosed in <EVENT> elements. The attributes that are appropriate for these elements are *tense*, *aspect*, *class*, *polarity*, *pos*, *stem*. The *stem* is the term's lemma, and *pos* is its part-of-speech.

The attribute *polarity* takes the value NEG if the event term is in a negative syntactic context, and POS otherwise. The attribute *class* contains several levels of information. It makes a distinction between terms that denote actions of speaking, which take the value REPORTING and those that do not. For these, it distinguishes between states (value STATE) and non-states (value OCCURRENCE), and it also encodes whether they create an intensional context (value I STATE for states and value I ACTION for non-states).

Temporal expressions (timexes) are inside <TIMEX3> elements. The most important features for these elements are *value*, *type* and *mod*. The timex's value encodes a normalized representation of this temporal entity, its *type* can be e.g. DATE, TIME or DURATION. The mod attribute is optional. It is used for expressions like *early this year*, which are annotated with mod="START". There are other attributes for timexes that encode whether it is the document's creation time (*functionInDocument*) and whether its value can be determined from the expression alone or requires other sources of information (*temporalFunction* and *anchorTimeID*).

The <TLINK> elements encode temporal relations. The attribute *relType* of these elements represents the type of relation. Its possible values are BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE, but the last three values occur rarely. The attribute *eventID* is a reference to the first argument of that relation. The second argument is given by the attribute *relatedToTime* (if it is a time, a date or a duration) or *relatedToEvent* (if it is another event).

For more details, see Costa & Branco, 2012.

4.5. Intended application of the corpus
By increasing the set of languages for which this kind of annotated data are available, we hope to stimulate research on temporal information processing, where a lot of progress can still be made.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

TimeBankPT is based on an existing corpus of English, namely the data used in the first TempEval competition (http://www.timeml.org/site/timebank/timebank.html). This English corpus was translated to Portuguese (with Google Translator Toolkit then revised by a human translator manually) and both corpora are aligned by paragraphs: the line breaks in the original collection are simply maintained in the translated corpus. A small script was developed to place all relevant TimeML markup at the end of each paragraph in the Portuguese text. Each TimeML markup element was then manually placed in the correct place in that paragraph. At this point some necessary changes to the annotations were also done manually. These are motivated by language differences. This approach involving manual steps is feasible because the original TempEval corpus is not very large (see Costa & Branco, 2010 e 2012).

V. Relevant References and Other Information

Costa, Francisco and Branco, António. 2012. "TimeBankPT: A TimeML Annotated Corpus of Portuguese". In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Costa, Francisco and Branco, António. 2010. "Temporal Information Processing of a New Language: Fast Porting with Minimal Resources". In *ACL2010-Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Costa, Francisco. to appear. *Processing Temporal Information in Unstructured Documents*. Ph.D.thesis, Universidade de Lisboa, Lisbon.