# GMT to +2 or How Can TimeML Be Used in Romanian

**Corina Forăscu**
Faculty of Computer Science, University "Al. I. Cuza" of Iaşi, Romania
Research Institute in Artificial Intelligence, Romanian Academy, Bucharest
16, Berthelot St., 700483 – Iaşi, Romania
E-mail: corinfor@info.uaic.ro

**Abstract**

The paper describes the construction and usage of the Romanian version of the TimeBank corpus. The success rate of 96.53% for the automatic import of the temporal annotation from English to Romanian shows that the automatic transfer is a worth doing enterprise if temporality is to be studied in another language than the one for which TimeML, the annotation standard used, was developed. A preliminary study identifies the main situations that occurred during the automatic transfer, as well as temporal elements not (yet) marked in the English corpus.

## 1. Introduction

Recently the focus on temporal information in NLP applications has increased (Mani et al., 2005), because this kind of information is useful in question-answering, information extraction or information retrieval, topic detection and tracking, machine translation, linguistic investigation, summarization, and discourse processing.

The temporal elements explicitly present in NL are:

– temporal expressions – references to a calendar or clock system, expressed by NPs, PPs, or AdvPs
– events – syntactically realized through sentences (their syntactic head - the main verb), noun phrases, adjectives, predicative clauses or prepositional phrases.

Implicitly these temporal elements are linked so that the events can be positioned in time, either relatively with respect to other events or on an absolute time axis.

The main objective of our study was to decide how well general temporal theories, developed mainly for English, can be applied to other languages – with emphasis on Romanian. Therefore we used the TimeBank 1.2. corpus (Pustejovsky et al., 2006), an English news corpus manually annotated and widely used in the temporal community together with TimeML the annotation standard (Sauri et al., 2006). These LRs are briefly described in section 2.

Because the existing Romanian LRs do not support temporal annotation (Cristea & Forascu, 2006) and the manual temporal annotation is very time consuming, expensive (Pustejovsky et al., 2002) and error-prone, including for Romanian (Forascu & Solomon, 2004), we decided to translate the English TimeBank and then to use the help or back-up from the same annotation applied to a parallel text. The translation[1], preprocessing and alignment of the corpus are presented in section 3.

In order to have linguistic evidence of how temporal information is really used in Romanian, as source of evidence to inform and substantiate the theory, we automatically imported the temporal annotation from English to Romanian. Together with an evaluation of this import, we identified cases of English temporal elements not marked in the corpus, situation in line with the current status of the TimeBank corpus (Boguraev & Ando, 2006). These activities are described in section 4.

The high success rate of the automatic import and the evaluation study shows that the import is a solution if the temporal information has to be studied based on linguistic evidence. The conclusions, as well as future work, are discussed in the last section.

## 2. Linguistic Resources used

### 2.1 The TimeML annotation standard

The TimeML standard has been developed for the automatic extraction of information about the event-structure of narrative texts, and it has been applied mainly to English news data. The mark-up language consists of a collection of tags intended to explicitly outline the information about the events reported in a given text, as well as about their temporal relations.

The TimeML metadata standard marks:

- Events through the tags:
  – EVENT: for situations that happen or occur, states or circumstances in which something obtains or holds true.
  – MAKEINSTANCE: for tracking the instances or realizations of a given event; the tag also carries the tense and aspect of the verb-denoted event.
- Temporal anchoring of events through the tags:
  – TIMEX3: for times of a day, dates – calendar dates or ranges, and durations.
  – SIGNAL: it marks function words that indicate how temporal objects are to be related to each other.
- Links between events and/or timexes through the tags:
  – TLINK – Temporal Link – indicates 13 types of temporal relations between two temporal elements (event-event, event-timex), similar to Allen's relations.
  – ALINK – Aspectual Link to mark the relationship between an aspectual event and its argument event.
  – SLINK – Subordination Link to mark contexts introducing relations between two events.

[1] We thank to our reviewers for suggesting us to include more details and explanations on this subject.

### 2.2 The TimeBank corpus

The creation of the TimeBank corpus started in 2002 during the TERQAS [2] workshop, and it should be considered preliminary (Boguraev & Ando, 2006): the corpus still needs improvements and reviews. The dimension of the corpus (4715 sentences with 10586 unique lexical units, from a total of 61042 lexical units) might be too small for robust statistical learning and the annotation inconsistencies require corrections. The corpus consists of 183 news report documents, with XML markups for document format and structure information, sentence boundary information, and named entity recognition (ENAMEX, NUMEX, CARDINAL from MUC-7).

## 3. Building the TimeBank parallel corpus

### 3.1 The Romanian version

The TimeBank, corpus was distributed for translation to two CL Master students with strong background in English and Romanian philology and translation. As the next step is the alignment of the English and Romanian versions of the corpus, a minimal set of translation recommendations was elaborated, in order not only to ensure a literal translation, but also to permit a best-possible word-alignment process and, as much as possible, a natural and good-quality target text. The two translators worked separately, sharing the original English corpus; hence an inter-translator agreement could not be estimated. Some basic translation principles are the followings:

- The sentences are translated in a 1:1 correspondence, whenever the language permits it, so that the sentence-alignment is directly obtained through translation.
- The translation equivalents have as much as possible the same part-of-speech; when the English word has a Romanian cognate, this is used in translation, and not its Romanian paraphrase.
- All words are translated and stylistic variations are avoided, so as not to introduce words or expressions without an English equivalent.
- The tense of verbs is mapped onto its corresponding Romanian one, the modifications being accepted only on linguistic grounds, but not stylistic.
- The format of the dates, moments of day and numbers conforms to the norms of written Romanian.

The sentence alignment of the corpus was obtained as a direct output of the translation.

A first automatic import showed that the translation needed more improvements. Therefore a manual check performed on the parallel corpus allowed us to detect and correct also some other lacks and inconsistencies in the way the translators worked.

In the 4715 sentences (translation units) of the current version of the Romanian corpus there are 65375 lexical tokens, including punctuation marks, representing 12640 lexical types.

[2] http://www.timeml.org/site/terqas/index.html

### 3.2 The parallel corpus

The English and Romanian raw texts were preprocessed in order to obtain the corpus in the format required by the lexical aligner. Using the TTL[3] module (Ion, 2007), the texts were tokenized – based on the MtSeg model, POS-tagged – using an adaptation of the TnT tagger, lemmatized – based on a lexicon, and chunked – using regular expressions. This module assembles the bitext in an XML format similar to the XCES one (Ide et al., 2000).

### 3.3 Word-level alignment

The four stage lexical aligner, YAWA, uses bilingual translation lexicons (Tufis & Barbu, 2002) and phrase boundaries detection to align words of a given bitext from Romanian to English. In each of the first three stages (content words alignment, inside-chunks alignment, and alignment in contiguous sequences of unaligned words), YAWA adds new links to those already created in the previous steps. Only in the last phase, correction, it deletes the wrong links (Tufis et al., 2005, 2006).

The automatic alignment performed on 181 files (out of 183) in the TimeBank parallel corpus produced 91714 alignments out of which 25346 are NULL-alignments. Two files were not aligned because of a low translation quality. In order to obtain an optimal transfer of the temporal annotations from the English version onto the Romanian one, all the alignments in the 181 files were manually checked.

## 4. Automatic import and its evaluation

Because of the way the translation was performed, the English corpus was parsed and for every sentence XML tag we could extract its content and replace it with the Romanian translation. We used the Romanian to English lexical alignment to transfer the XML markup from English to Romanian because, otherwise, we could obtain the Romanian translation in a shuffled form if the word order was not preserved. The transfer algorithm (Forascu et al., 2007) goes as follows:

For every pair of sentences ($\mathbf{S}_{ro}$; $\mathbf{S}_{en}$) from the TimeBank parallel corpus with the $\mathbf{T}_{en}$ English equivalent sentence ($\mathbf{T}_{en}$ is the same sentence – same raw text – as $\mathbf{S}_{en}$, with the exception that $\mathbf{T}_{en}$ has the XML structure that we want to transfer) do:

- construct a list **E** of pairs of English text fragments with sequences of English indexes from $\mathbf{S}_{en}$ and $\mathbf{T}_{en}$. Due to the fact that the tokenization of $\mathbf{S}_{en}$ is different from that of $\mathbf{T}_{en}$, the list **E** is needed in order to map English text fragments from $\mathbf{T}_{en}$ with sequences of indexes from $\mathbf{S}_{en}$ so as to be able to use the Romanian lexical alignments which exist relative to these indexes..
- add to every element of **E** the XML context in which that text fragment appeared. For every tag, its attributes – if present – are stored.
- construct the list **RW** of Romanian words along with the transferred XML contexts using **E** and the lexical alignment between $\mathbf{S}_{ro}$ and $\mathbf{S}_{en}$. If a word in $\mathbf{S}_{ro}$ is not

[3] **T**okenizing, **T**agging and **L**emmatizing free running texts

aligned, the top context for it, namely s, is considered.
- construct the final list **R** of Romanian text fragments from **RW** by conflating adjacent elements of **RW** that appear in the same XML context. Output the list in XML format

The transfer procedure is designated for the inline markups in the header and the text parts of a TimeBank document. For the offline temporal markups (MAKEINSTANCE and LINK tags) the transfer kept only those XML tags from the English version whose IDs belong to XML structures that have been transferred to Romanian.
The success rate for the import of the temporal markups altogether is 96.53%. A more detailed statistic, on all tags, is illustrated in Table 1. The 3.47 % of non-transferred tags are due to missing translations (though the Romanian translation was a good and natural one), non-lexicalisations in Romanian, or missing alignments.

| **TimeML tags** | **number** | **% transferred** |
|---|---|---|
| EVENTs | 7703 | 97.07 |
| TIMEXes | 1356 | 95.89 |
| SIGNALs | 668 | 97.09 |
| INSTANCEs | 7706 | 97.05 |
| TLINKs | 6122 | 95.38 |
| ALINKs | 249 | 93.96 |
| SLINKs | 2831 | 96.55 |
| **TOTAL** | **26635** | **96.53** |

Table 1: Tags transferred from the English into the Romanian TimeBank 1.2.

Using about 10% of the Romanian corpus, we performed a preliminary study to analyze the situations of perfect transfer and compare them with those situations in which:
- the temporal annotation transfer has to be done with some amendments when the temporal constructions in the two languages are not similar but they can be transferred using special developed rules; (Amendment);
- it has to deal with language specific phenomena, such as the treatment of clitics or the PRO-drop phenomenon, specific to Romanian but not to English; (Language Specific);
- the transfer can not be performed; (Impossible).

For the statistics in table 2, summarizing the four situations encountered during the temporal annotation import, we will not refer to the offline markups, because they are automatically imported only if the elements they are linked to are present in the text.
The situations encountered in our study for the EVENT tag are the followings:
- The amendments needed to be done in the automatic import are due mainly to the TimeML rule stating that in cases of phrases, the EVENT tag should mark only the head of the construction. This is the case for Romanian reflexive verbs (the reflexive pronoun was marked inside the tag), Romanian verbal collocations, and compound verb phrases.
- The intercalation of an adverb/conjunction between the verbs forming a verb phrase was the only Language Specific phenomenon that occurred in the files we used in our study; these 0.36% cases, when the EVENT tags were automatically imported also on the auxiliary Romanian verb, were corrected for the EVENTs included in our study.
- The 0.48% Impossible transfers, due to missing translations, non-lexicalisations in Romanian, or missing alignments, were also corrected.

| **Tags** / **Transfer** | EVENT | TIMEX3 | SIGNAL |
|---|---|---|---|
| Perfect | 785 | 33 | 29 |
| Amendment | 37 | 3 | - |
| Language Specific | 3 | - | - |
| Impossible | 4 | - | 4 |

Table 2: Types of temporal annotation transfer

The 8.33% situations when amendments were needed for the transfer of the TIMEX3 tag are due to missing alignments or the wrong markings of the Romanian prepositions as part of TIMEX3.
The 12.12% situations of impossible transfer of a SIGNAL tag are due to non-lexicalisation in Romanian.
Even if the main objective of our study was to detect the types and situations encountered during the automatic annotation import, the study permitted, as a side-effect, to identify and mark temporal elements not (yet) marked in the English TimeBank 1.2. We modified and marked only new tags of type EVENT, TIMEX3 and SIGNAL using the Callisto [4] annotation tool for both the Romanian and English parts of the parallel corpus, with the Tango TimeML Importer. In the snapshot illustrated in Figure 1, one can see that there were some problems with missing spaces between words and the Romanian diacritics.
For the EVENTs we classified as such another 104 elements: 70 OCCURENCEs (nouns: *missions, training, fight, (mediation) effort, demarcation, move*, as well as verbs: *supervising, leading, include*), 5 of the REPORTING class (*say, said*), 21 belonging to the STATE class (*belongs, look, ceiling, staying, war, policies*), 1 event of type I_ACTION (*include*), and 7 from the I_STATE class (*like, think, (have the) power*). The rationale behind these modifications is that each sentence expresses an event, even if not so well temporally-anchored.
We have marked two new temporal expressions (TIMEX3 tag) for which the value is PAST_REF – meaning that the expressions do not have a specific value, but they can be normalised according to the extended ISO 8601 standard used in TimeML: *once, not that long ago*.
The 19 new SIGNALs are most probably due to inevitable manual annotation mistakes: *several, when, meanwhile,*

[4] http://callisto.mitre.org/

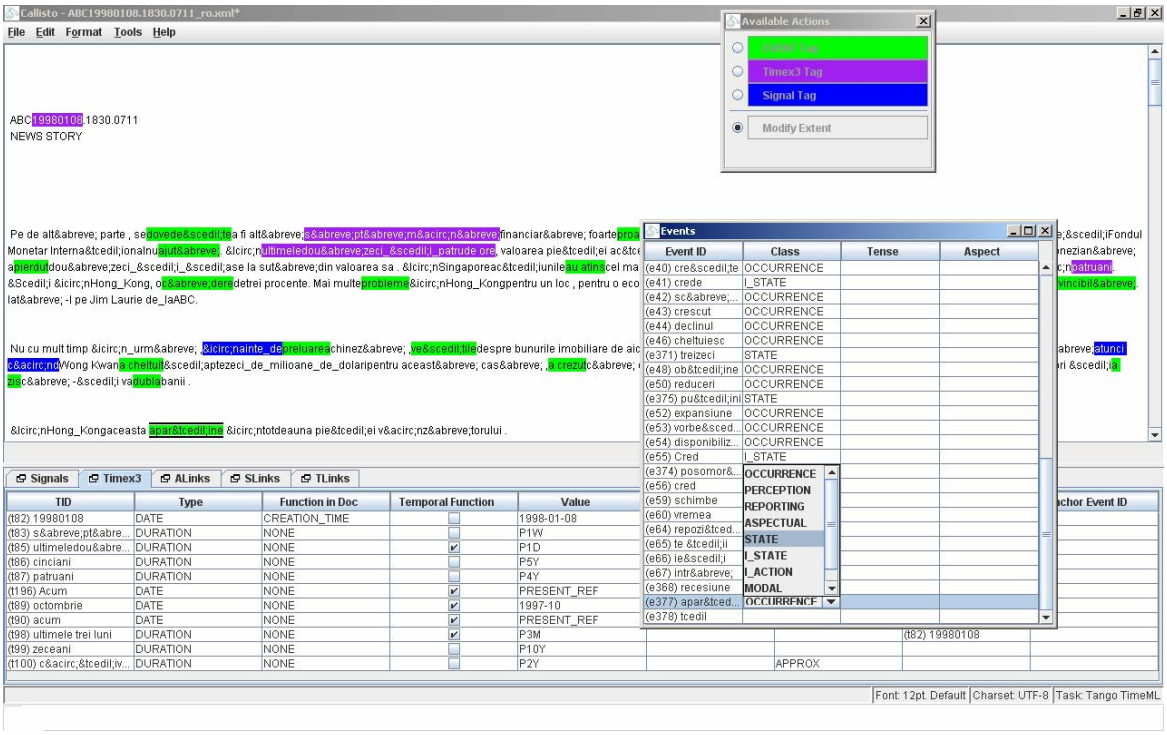


Figure 1. Callisto workspace: a snapshot from a Romanian file

*time and again, after, on*. As we mentioned before, some of these signals are not lexicalized in Romanian, hence we marked them only in the English corpus. Moreover, the absence of some SIGNALs didn't permit to identify multiple instantiations for some EVENTs. For example in *US has tried to hasten it on several occasions* the absence of the SIGNAL on *several* permits to mark only one instance for the *tried* EVENT.

All the above observations are consistent with the conclusions of the TimeBank developers: the corpus still needs improvements and reviews (Boguraev & Ando, 2006) especially with respect to: event classes, incomplete temporal markup and linking, incomplete subordinated linking.

## 5. Conclusions

The research proves that the automatic import of the temporal annotations from English to other language is a worth doing enterprise with a very high success rate. The most important conclusion is that, as the manual annotation of the temporal expressions, events and their links is very time-consuming and expensive, the automatic transfer of annotations represents a solution, provided a parallel corpus involving the target language exists, the source language displays temporal annotation, the target language has similar temporal conceptualizations[5], and adequate processing tools are available. This study opens the possibility to decide, based on corpus-evidence, how well the temporal theories can be applied to other languages, here with emphasis on Romanian.

Future immediate activities include finishing the evaluation and the correction/improvement of the annotations in the parallel Romanian-English TimeBank, especially the temporal annotations. The corpus will be then used to create or adapt a temporal tagger – such as TARSQI (Verhagen et al., 2005) - for Romanian, or even a language independent one, in a combination of rules with statistical information derived from the corpus.

The temporal annotated data together with time ontologies (Hobbs, 2002; Hobbs & Pustejovsky , 2003) will be used to represent the temporal structure of the discourse and its possible relations with other discourse structures, such as, for example, Rhetorical Structure (Mann & Thompson, 1987) or Veins Theory (Cristea et al., 1998), as we already showed there are strong interconnections between the Veins theory and temporal information in texts (Forascu et al., 2006).

## 6. References

Boguraev, B., Ando, R. (2006). Analysis of TimeBank as a Resource for TimeML Parsing. *Proceedings of LREC International Conference 2006*, Genoa, Italy, pp. 71-76.

Cristea, D., Ide, N., Romary, L. (1998). Veins Theory. An Approach to Global Cohesion and Coherence. *Proceedings of COLING/ACL- 98*, Montreal, Canada pp. 281-285.

Cristea, D., Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. In *Journal of Computer Science of Moldova*, Academy of Science of Moldova, vol. 14, nr. 1(40), ISSN 1561-4042, pp. 34-73.

Forăscu, C., Solomon, D. (2004). Towards a Time Tagger for Romanian. *Proceedings of the ESSLLI Student Session*, Nancy, France, pp. 202-213.

Forăscu, C., Pistol, I., Cristea, D. (2006). Temporality in

[5] We acknowledge this suggestion from one of our reviewers.

Relation with Discourse Structure. *Proceedings of LREC International Conference 2006*, ISBN 2-9517408-2-4; Genoa, Italy, pp. 65-70.

Forăscu, C., Ion, R., Tufiş, D. (2007). Semi-automatic Annotation of the Romanian TimeBank 1.2. In *Proceedings of the RANLP 2007 Workshop on Computer-aided language processing - CALP*; Constantin Orăsan, Sandra Kuebler (Eds.). Borovets, Bulgaria, 30 September 2007. ISBN 978-954-452-005-2, pp. 1-7.

Hobbs, J. (2002). Toward an Ontology for Time for the Semantic Web. *Proceedings of the LREC 2002 Workshop Annotation Standards for Temporal Information in Natural Language*, Las Palmas, Spain, pp. 28-35.

Hobbs, J., Pustejovsky, J. (2003). Annotating and Reasoning about Time and Events. *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford, California.

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*, pp. 825-830.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*. (in Romanian). PhD thesis. Romanian Academy, Bucharest.

Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.). (2005). *The Language of Time: A Reader*. Oxford University Press, ISBN-13: 978-0-19-926853-5.

Mann, W. C., Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of texts structures. In: Kempen, G. (ed.): *Natural Language Generation*. Martinus Nijhoff Publisher, Dordrecht, pp. 85–96.

Pustejovsky, J., Belanger, L., Castaño, J., Gaizauskas, R., Hanks, P., Ingria, B., Katz, G., Radev, D., Rumshisky, A., Sanfilippo, A., Sauri, R., Setzer, A., Sundheim, B., Verhagen, M. (2002). *NRRC Summer Workshop on Temporal and Event Recognition for QA Systems*.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, B., Setzer, A. (2006). *TimeBank 1.2.* Linguistic Data Consortium.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J. (2006). *TimeML Annotation Guidelines*, Version 1.2.1, January.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2005). Combined Aligners. *Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan, pp. 107-110.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, Trento, Italy, pp. 153-160.

Tufiş, D., Barbu, A.M. (2002). Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. *International Journal of Speech Technology*. Kluwer Academic Publishers, no.5, pp. 199-209.

Verhagen, M., Mani, I., Sauri, R., Littman, J., Knippen, R., Bae Jang, S., Rumshisky, A., Phillips, J., Pustejovsky, J. (2005). Automating Temporal Annotation with TARSQI. *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, Michigan, pp. 81-84.