Romanian TimeBank Corpus

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of:

- 183 files with Romanian news texts (translated from English), with ISO-TimeML and other (name-entities, header, sentence) mark-ups.
- 181 alignment files and 181 parallel English-Romanian files, XCES format, including POS, lemma and chunk attributes.

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

The Ro-TimeBank corpus is represented in XML format.
Each line of an .align file has the format:

TU_ID index_ro index_en S|P

where:

<TU_ID> is the ID of the translation unit (TU) the TEQs belong to;
<index_ro> is the token's ID in the Romanian segment of the TU;
<index_en> is the ID of the English token that is TEQ with the Romanian token;
S | P indicates that the alignment is Sure or Probable.

1.3 Character encoding

UTF8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufiş
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
e-mail: tufis@racai.ro

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The first version of the corpus will be available on the MetaShare platform as an archive. Improved versions will be available upon request.

2.3 Copyright statement and information on IPR

The resource is free, license-based for research purposes and fee license-based for commercial purposes.

The TimeBank corpus (description, IPR and copyright) is detailed in [4].
The principles and methodology to obtain the alignments and XCES files are detailed in [2] and [3]. Therefore, the copyright & IPR are governed by these authors.

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive to be uploaded on the MetaShare platform contains:
Ro-TimeBank/data/

Ro-TimeBank/data/align

It contains the 181 files with the English-Romanian alignments. The alignments were automatically obtained and then validated and corrected.

Ro-TimeBank/data/en-ro-msd

The 181 files with the parallel English-Romanian texts, XCES format.

Ro-TimeBank/data/ro

The 183 Romanian files with temporal (TimeML) and other (NE, sentence, header) markups.

3.2 Data structure of an entry

Please see 1.2. and 3.1. above.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The Ro-TimeBank corpus contains 4715 sentences, 65375 lexical units; the whole corpus (with annotations) needs about 3.00 MB for disk storage.
There are 125625 words in the parallel /en-ro subfolder, with 18720 sentences.
The .align and the XCES files need 1.36 MB and 7.55 MB.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

The corpus is parallel (XCES format, with alignments) [2, 3].

4.2 The natural language(s) of the corpus

The languages of the corpus are standard Romanian, orthography being compliant with the current Romanian Academy norms and English.

4. 3 Domain(s)/register(s) of the corpus

The Romanian texts are translations of English news texts.

4.4 Annotations in the corpus (if an annotated corpus)

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

The Ro-TimeBank corpus [1] is annotated according to TimeML standard [5], including also mark-ups for header, sentence and named-entities information.
The XCES corpus includes mark-ups for POS, lemma, chunks [2].

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

The corpus is (automatically) POS-tagged and (semi-automatically) TIME tagged [1].

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

The alignments were obtained automatically using the YAWA toolkit and then manually checked [2, 3].

4.4.4 Attributes and their values (if annotated)

```
<!ELEMENT ISO-TimeML-Ro ( #PCDATA | s | EVENT | TIMEX3 | SIGNAL | TLINK | SLINK | ALINK )* >
<!ATTLIST ISO-TimeML-Ro xsi:noNamespaceSchemaLocation CDATA #IMPLIED >
<!ATTLIST ISO-TimeML-Ro xmlns:xsi CDATA #IMPLIED >
<!ATTLIST TimeML comment CDATA #IMPLIED >

<!ELEMENT s ( #PCDATA | EVENT | TIMEX3 )* >

<!ELEMENT EVENT ( #PCDATA ) >
<!ATTLIST EVENT eid ID #REQUIRED >
<!ATTLIST EVENT mainevent (YES | NO) #IMPLIED >
<!ATTLIST EVENT pred CDATA #IMPLIED >
<!ATTLIST EVENT class ( OCCURRENCE | PERCEPTION | REPORTING | ASPECTUAL | I_STATE | I_ACTION | STATE ) #REQUIRED >
<!ATTLIST EVENT pos ( ADJECTIVE | NOUN | VERB | PREPOSITION | OTHER ) #REQUIRED >
<!ATTLIST EVENT tense ( FUTURE | PAST | SIM_PAST | PLUS_PAST | PRESENT | NONE ) #REQUIRED >
<!ATTLIST EVENT aspect ( NONE | PERFECTIVE | IMPERFECTIVE ) #REQUIRED >
<!ATTLIST EVENT polarity ( POS | NEG ) #REQUIRED >
<!ATTLIST EVENT mood ( SUBJUNCTIVE | CONDITIONAL | IMPERATIVE | NONE ) #REQUIRED >
```

```
<!ATTLIST EVENT vform ( INFINITIVE | GERUNDIVE | PARTICIPLE | NONE) #REQUIRED
>
<!ATTLIST EVENT modality ( NECESSITY | POSSIBILITY | OBLIGATION | PERMISSION)
#IMPLIED >
<!ATTLIST EVENT comment CDATA #IMPLIED >

<!ELEMENT TIMEX3 ( #PCDATA ) >
<!ATTLIST TIMEX3 tid ID #REQUIRED >
<!ATTLIST TIMEX3 type ( DATE | DURATION | SET | TIME ) #REQUIRED >
<!ATTLIST TIMEX3 value NMTOKEN #REQUIRED >
<!ATTLIST TIMEX3 anchorTimeID IDREF #IMPLIED >
<!ATTLIST TIMEX3 beginPoint IDREF #IMPLIED >
<!ATTLIST TIMEX3 endPoint IDREF #IMPLIED >
<!ATTLIST TIMEX3 freq NMTOKEN #IMPLIED >
<!ATTLIST TIMEX3 functionInDocument ( CREATION_TIME | EXPIRATION_TIME |
MODIFICATION_TIME | PUBLICATION_TIME | RELEASE_TIME | RECEPTION_TIME | NONE )
#IMPLIED>
<!ATTLIST TIMEX3 mod ( BEFORE | AFTER | ON_OR_BEFORE | ON_OR_AFTER | LESS_THAN
| MORE_THAN | EQUAL_OR_LESS | EQUAL_OR_MORE | START | MID | END | APPROX )
#IMPLIED >
<!ATTLIST TIMEX3 quant CDATA #IMPLIED >
<!ATTLIST TIMEX3 temporalFunction ( false | true ) #IMPLIED >
<!ATTLIST TIMEX3 valueFromFunction IDREF #IMPLIED >
<!ATTLIST TIMEX3 comment CDATA #IMPLIED >

<!ELEMENT SIGNAL ( #PCDATA ) >
<!ATTLIST SIGNAL sid ID #REQUIRED >
<!ATTLIST SIGNAL comment CDATA #IMPLIED >

<!ELEMENT TLINK EMPTY >
<!ATTLIST TLINK lid ID #REQUIRED >
<!ATTLIST TLINK relType ( BEFORE | AFTER | INCLUDES | IS_INCLUDED | DURING |
DURING_INV | SIMULTANEOUS | IAFTER | IBEFORE | IDENTITY | BEGINS | ENDS |
BEGUN_BY | ENDED_BY ) #REQUIRED >
<!ATTLIST TLINK eventID IDREF #IMPLIED >
<!ATTLIST TLINK timeID IDREF #IMPLIED >
<!ATTLIST TLINK relatedToEvent IDREF #IMPLIED >
<!ATTLIST TLINK relatedToTime IDREF #IMPLIED >
<!ATTLIST TLINK signalID IDREF #IMPLIED >
<!ATTLIST TLINK origin CDATA #IMPLIED >
<!ATTLIST TLINK syntax CDATA #IMPLIED >
<!ATTLIST TLINK comment CDATA #IMPLIED >

<!ELEMENT SLINK EMPTY >
<!ATTLIST SLINK lid ID #REQUIRED >
<!ATTLIST SLINK relType ( CONDITIONAL | COUNTER_FACTIVE | EVIDENTIAL | FACTIVE
| INTENSIONAL | NEG_EVIDENTIAL ) #REQUIRED >
```

```
<!ATTLIST SLINK eventID NMTOKEN #REQUIRED >
<!ATTLIST SLINK subordinatedEvent NMTOKEN #REQUIRED >
<!ATTLIST SLINK signalID NMTOKEN #IMPLIED >
<!ATTLIST SLINK syntax CDATA #IMPLIED >
<!ATTLIST SLINK comment CDATA #IMPLIED >

<!ELEMENT ALINK EMPTY >
<!ATTLIST ALINK lid ID #REQUIRED >
<!ATTLIST ALINK relType ( CONTINUES | CULMINATES | INITIATES | REINITIATES |
TERMINATES ) #REQUIRED >
<!ATTLIST ALINK eventID IDREF #REQUIRED >
<!ATTLIST ALINK relatedToEvent IDREF #REQUIRED >
<!ATTLIST ALINK signalID IDREF #IMPLIED >
<!ATTLIST ALINK syntax CDATA #IMPLIED >
                  <!ATTLIST ALINK comment CDATA #IMPLIED >
```

-

4.5 Intended application of the corpus

The corpus can be used for NLP applications using temporal information, temporal parsing, machine learning, machine translation, summarization.

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

The corpus was checked for alignments and temporal markups.

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] Corina Forăscu. Contributions to Romanian language processing through discourse analysis methods. (in Romanian). PhD thesis. Romanian Academy, Bucharest. 2011.

[2] Radu Ion.Word Sense Disambiguation Methods Applied to English and Romanian. (in Romanian). PhD thesis. Romanian Academy, Bucharest. 2007.

[3] Dan Tufiș. A Cheap and Fast Way to Build Useful Translation Lexicons. In Proceedings of the 19th Intl. Conf. On Computational Linguistics, Taipei, 2002, pp. 1030-1036.

[4] Pustejovsky, James, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. (2006). TimeBank 1.2. Linguistic Data Consortium, Philadelphia, ISBN: 1-58563-386-0.

[5] ISO: Language Resource Management – Semantic Annotation Framework (SemAF) – Part 1: Time and Events. Secretariat KATS, 2009. ISO Report ISO/TC37/SC4 N269 (ISO/WD 24617-1).