

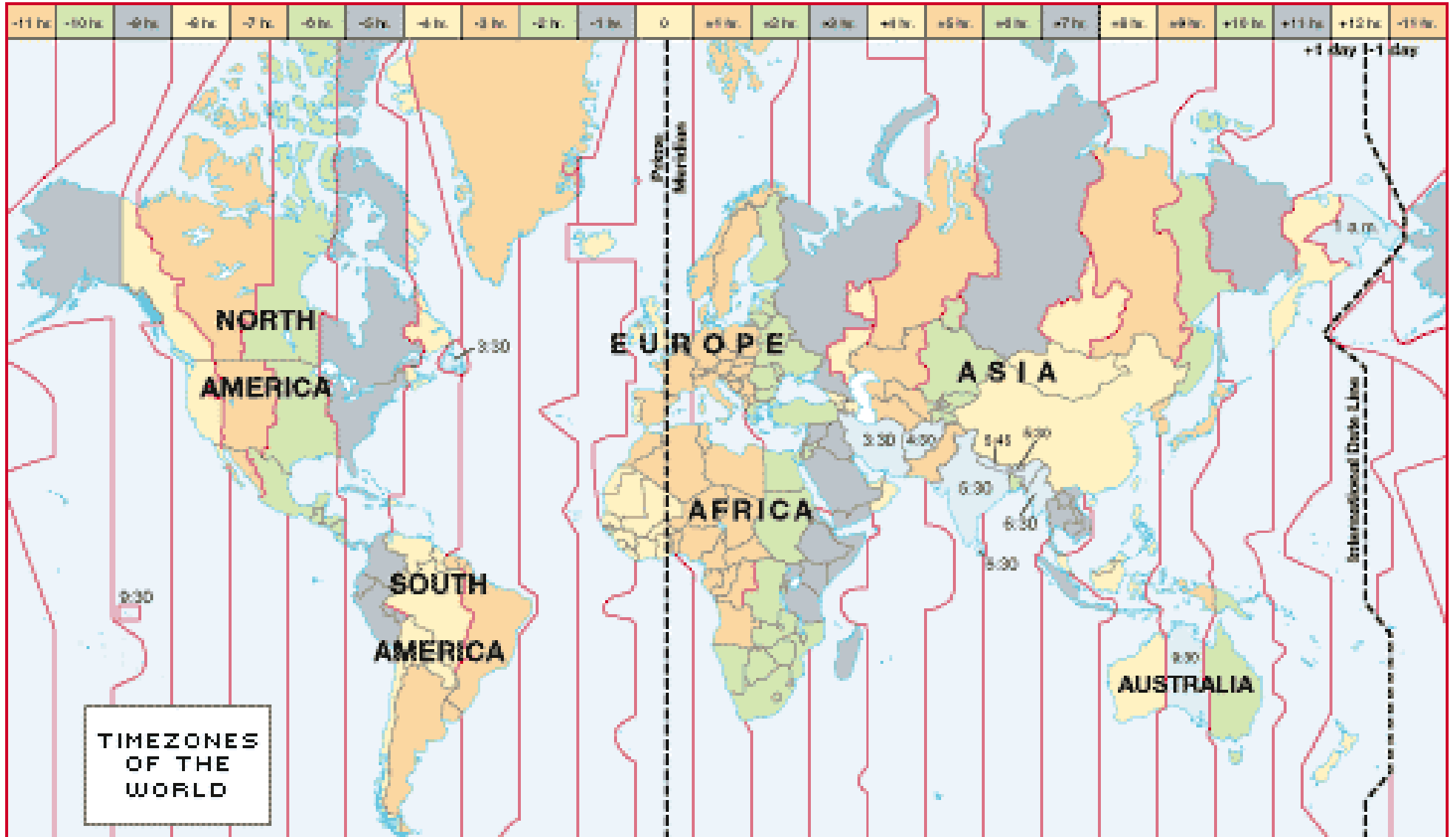


# GMT to +2 or How Can TimeML Be Used in Romanian

Corina Forăscu

Research Institute for Artificial Intelligence of the Romanian Academy  
&  
Faculty of Computer Science, A.I. Cuza University of Iasi, Romania

[corinfor@info.uaic.ro](mailto:corinfor@info.uaic.ro)



# Outline

1. Basic concepts
2. Standard & initial corpus
3. Corpus creation & processing
4. Analysis
5. Conclusions

# Temporal information in NL

1. **Time-denoting** expressions – references to a calendar or clock system (*NPs, PPs, or AdvPs*)

*the 28<sup>th</sup> of May, 2008; Wednesday; tomorrow; the third month*

2. **Event-denoting** expressions - reference to an event (*sentences, NPs, Adjs, PPs*)

*Jerry is watching the talks.*

*The presenter is prepared for a possible attack.*

*A student, dormant for half of the session, suddenly started to ask questions.*

# Benefits from TIP for NLP

1. CL: lexicon induction, linguistic investigation
2. QA: *when?*, *how often?* or *how long?*
3. IE & IR
4. MT:
  - translated and normalized temporal references
  - mappings between different behavior of tenses from language to language
5. DP: temporal structure of discourse and summarization



# Standard: TimeML

A metadata standard developed especially for news articles, for marking

- events: **EVENT, MAKEINSTANCE**
- temporal anchoring of events: **TIMEX3, SIGNAL**
- links between events and/or timexes: **TLINK, ALINK, SLINK**



10/30/09

# TimeML

McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

10/30/09

# TimeML: EVENTS

```
<EVENT eid="e206" class="I_STATE">
```

McDonalds is so **anxious**<sup>e206</sup> to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.



10/30/09

# TimeML: EVENTS

```
<EVENT eid="e32" class="OCCURRENCE">
```

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
soon will begin selling  
hamburgers for 99 cents.

10/30/09

# TimeML: EVENTS

```
<EVENT eid="e33" class="ASPECTUAL">
```

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
soon will **begin**<sup>e33</sup> selling  
hamburgers for 99 cents.

10/30/09

# TimeML: EVENTS

```
<EVENT eid="e34" class="OCCURRENCE">
```

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
soon will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

10/30/09

# TimeML: INSTANCES

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
soon will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

```
<MAKEINSTANCE aspect="NONE" eiid="ei2019" tense="PRESENT" eventID="e206" />  
<MAKEINSTANCE aspect="NONE" eiid="ei2020" tense="NONE" eventID="e32" />  
<MAKEINSTANCE aspect="NONE" eiid="ei2021" tense="FUTURE" eventID="e33" />  
<MAKEINSTANCE aspect="PROGRESSIVE" eiid="ei2022" tense="NONE" eventID="e34" />
```



# TimeML: TIMEX3

10/30/09<sup>t192</sup>

```
<TIMEX3 tid="t192" type="DATE" temporalFunction="false"  
functionInDocument="CREATION_TIME" value="2009-10-30">10/30/09  
</TIMEX3>
```

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
soon will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

10/30/09<sup>t192</sup>

# TimeML: TIMEX3

```
<TIMEX3 tid="t207" type="DATE" temporalFunction="true"
functionInDocument="NONE" value="FUTURE_REF"
anchorTimeID="t192">
```

McDonalds is so **anxious**<sup>e206</sup> to  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.



10/30/09<sup>t192</sup>

# TimeML: SIGNALS

<SIGNAL sid="s31">

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

# TimeML: TLINKs

10/30/09<sup>t192</sup>

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.


```
<TLINK relatedToTime="t192" eventInstanceId="ei2019" relType="INCLUDES" />
```



10/30/09<sup>t192</sup>

# TimeML: TLINKs

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
 hamburgers for 99 cents.



```
<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES" />
<TLINK relatedToEventInstance="ei2021" eventInstanceID="ei2019"
relType="BEFORE" />
```

10/30/09<sup>t192</sup>

# TimeML: TLINKs

McDonalds is so anxious<sup>e206</sup> to<sup>s31</sup>  
 turn<sup>e32</sup> around KFC sales that it  
 soon<sup>t207</sup> will begin<sup>e33</sup> selling<sup>e34</sup>  
 hamburgers for 99 cents.

```
<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES" />
<TLINK relatedToEventInstance="ei2021" eventInstanceID="ei2019"
relType="BEFORE" />
<TLINK relatedToTime="t207" eventInstanceID="ei2021"
relType="IS_INCLUDED" />
```

# TimeML: TLINKs

10/30/09<sup>t192</sup>

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
 hamburgers for 99 cents.

```
<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES" />
<TLINK relatedToEventInstance="ei2021" eventInstanceID="ei2019"
relType="BEFORE" />
<TLINK relatedToTime="t207" eventInstanceID="ei2021"
relType="IS_INCLUDED" />
<TLINK relatedToTime="t192" eventInstanceID="ei2021" relType="AFTER" />
```

# TimeML: SLINKs

10/30/09<sup>t192</sup>

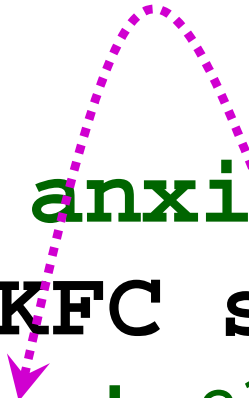
McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

```
<SLINK signalID="s31" subordinatedEventInstance="ei2020"  
eventInstanceID="ei2019" relType="MODAL" />
```

10/30/09<sup>t192</sup>

# TimeML: SLINKs

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.



```
<SLINK signalID="s31" subordinatedEventInstance="ei2020"  
eventInstanceID="ei2019" relType="MODAL" />
```

```
<SLINK signalID="s31" subordinatedEventInstance="ei2020"  
eventInstanceID="ei2021" relType="MODAL" />
```



# TimeML: ALINKs

10/30/09<sup>t192</sup>

McDonalds is so **anxious**<sup>e206</sup> **to**<sup>s31</sup>  
**turn**<sup>e32</sup> around KFC sales that it  
**soon**<sup>t207</sup> will **begin**<sup>e33</sup> **selling**<sup>e34</sup>  
hamburgers for 99 cents.

```
<ALINK relatedToEventInstance="ei2022" eventInstanceID="ei2021"  
relType="INITIATES" />
```

# Corpus: TimeBank

- 183 English news report documents TimeML annotated, freely distributed through LDC
- 4715 sentences with
  - 10586 unique lexical units, from
  - a total of 61042 lexical units

## Non-TimeML Markup in Time Bank 1.1:

- structure information: **header**
- named entity recognition: **<ENAMEX>**, **<NUMEX>**, **<CARDINAL>**
- sentence boundary information: **<s>**



# Corpus: TimeBank – stats

• EVENTS	7935
• INSTANCES	7940
• TIMEX3es	1414
• SIGNALS	688
• TLINKS	6418
• SLINKS	2932
• ALINKS	265
• TOTAL	27592





# Parallel corpus creation & processing

1. translation
2. pre-processing
3. alignment
4. annotation import

# Corpus translation

## 1. Translation

- 2 “trained translators”; one final correction
- translation criteria
- 4715 sentences (translation units)
  - 65375 lexical tokens (61042 in English)
  - 12640 lexical types (10586 in English)

## 2. pre-processing

## 3. alignment

## 4. annotation import

# Pre-processing the parallel corpus

1. Translation
2. Pre-processing (RACAI web services)
  1. Tokenisation – MtSeg, with idiomatic expressions, clitic splitting
  2. POS-tagging – TnT adapted & improved to determine the POS of unknown words
  3. Lemmatisation – probabilistic, based on a lexicon
  4. Chunking – REs over POS tags to determine non-recursive NPs, APs, AdvPs, PPs
3. alignment
4. annotation import

# Aligning the parallel corpus

1. Translation
2. Pre-processing
3. Alignment (RACAI YAWA aligner)
  1. Content words alignment
  2. Inside-Chunks alignment
  3. Alignment in contiguous sequences of unaligned words
  4. Correction phase
    - 91714 alignments, manually checked
4. annotation import



# Aligning the parallel corpus

MTkit  
Aligner Save Align Viewer

Aligner ABC19980120.1830.0957\_msd.align

6

ABC199...			
02--01			
03--02			
03--03			
03--04			
04--05			
05--06			
06--07			
06--08			
06--09			
06--10			
07--12			
08--14			
09--11			
09--13			
10--15			

He<sup>1</sup> 1 A  
has<sup>2</sup> 2 trăit  
outlasted<sup>3</sup> 3 mai  
and<sup>4</sup> 4 mult  
sometimes<sup>5</sup> 5 și  
outsmarted<sup>6</sup> 6 uneori  
eight<sup>7</sup> 7 a  
American<sup>8</sup> 8 fost  
presidents<sup>9</sup> 9 mai  
. <sup>10</sup> 10 inteligent  
. 11 decât  
. 12 opt  
. 13 președinți  
. 14 americani  
. 15 .

**Properties**

**Main**

Ana	Vmps
Chunk	Vp#1,Ap#1
Lemma	outlast
Ortho	outlasted
TAna	1

**WordNet**

**Ana**  
The category/class of the word.

UPEC 2008 mtaarakteh

# Parallel corpus: annotation import

1. Translation
2. Pre-processing
3. Alignment (RACAI YAWA aligner)
4. Annotation import
  1. Inline markup (EVENT, TIMEX3, SIGNAL): sentence level import of XML tags from English to Romanian
  2. Offline markup (MAKEINSTANCE, ALINK, TLINK, SLINK) : the transfer kept only those XML tags whose IDs belong to XML structures that have been transferred to Romanian



# Parallel corpus: annotation import

TimeML tags	#	% transfered
EVENTs	7703	97.07
INSTANCESs	7706	97.05
TIMEX3s	1356	95.89
SIGNALs	668	97.09
TLINKs	6122	95.38
SLINKs	2831	96.55
ALINKs	249	93.96
<b>TOTAL</b>	<b>26635</b>	<b>96.53</b>

# Analysis of the annotation import

A preliminary study using 10% of the parallel corpus in order to identify:

1. Types of temporal annotation import
  1. Perfect transfer
  2. Transfer with some amendments due to TimeML specifications
  3. Transfer with amendments imposed by with language specific phenomena
  4. Impossible transfer
2. Temporal elements not (yet) marked



# Types of temporal annotation import

898 inline markups (`EVENT`, `TIMEX3`, `SIGNAL`)

## 1. Types of temporal annotation import

1. Perfect transfer: 847 (91.41%) situations
2. Transfer with some amendments due to TimeML specifications: 40 (6.4%) situations
3. Transfer with amendments imposed by with language specific phenomena: 3 (0.36%) situations
4. Impossible transfer: 8 (6.3%) situations

2. Temporal elements not (yet) marked in English:  
104 **EVENT**s, 2 **TIMEX3**s, 19 **SIGNAL**s

# Annotation import: EVENTS

Type	#	Reason
Perfect	785	☺
Amendment	37	<p>TimeML rule: in cases of phrases, the EVENT tag should mark only the head of the construction:</p> <ul style="list-style-type: none"> <li>• reflexive verbs: (să) <u>se</u> retragă – (to) <i>withdraw</i></li> <li>• verbal collocations: avea <u>permisiunea</u> – <i>permit</i></li> <li>• compound verb phrases: <u>să se</u> îndoiască – <i>doubt</i></li> </ul>
Language specific	3	<p>intercalation of an adverb/conjunction between the verbs forming a verb phrase: <i>also said</i> – (au) <u>mai</u> spus; (he) <i>also criticised</i> – (a) <u>și</u> criticat</p>
Impossible	4	<ul style="list-style-type: none"> <li>• missing translations: <ul style="list-style-type: none"> <li><i>forces that <u>harbor</u> ill intentions</i> – <i>forțe cu intenții rele</i></li> </ul> </li> <li>• non-lexicalisations: <u>give</u><sup>1</sup> the <u>view</u><sup>2</sup> – <u>arată</u><sup>1</sup></li> <li>• missing alignments (situations corrected)</li> </ul>



# Annotation import: TIMEX3s

Type	#	Reason
Perfect	33	😊
Amendment	3	<ul style="list-style-type: none"><li>•wrong marking of the Romanian prepositions as part of TIMEX3: <i>eight years (war) – (războiul) <u>de</u> opt ani</i></li><li>•missing alignment: <i>some time - un timp <u>mai lung</u></i></li></ul>
Language specific		
Impossible		



# Annotation import: SIGNALS

Type	#	Reason
Perfect	29	😊
Amendment		
Language specific		
Impossible	4	•non-lexicalisations: <i><u>on</u> Thursday – joi</i>

# New temporal elements

- 104 **EVENTS**: 70 OCCURRENCEs (nouns: *missions, training, fight, demarcation*, verbs: *supervising, leading, include*), 5 REPORTING (*say, said*), 21 STATEs (*belongs, look, staying, war, policies*), 1 I\_ACTION (*include*), 7 I\_STATE (*like, think*)

Rationale: each sentence expresses an event, even if not so well temporally-anchored

- 2 **TIMEX3s**: *once, not that long ago*

Rationale: non-specific value but possible to normalize according to ISO 8601 extended

- 19 **SIGNALS**: *several, when, meanwhile, time and again, after, on*

Rationale: identify multiple instantiations for some EVENTS (inevitable manual annotation mistakes)

# Conclusions

1. The automatic import of the temporal annotations from English to Romanian is a worth doing enterprise (96.53% success rate).
2. Human introspection shows few modifications are needed.
3. The automatic transfer of (temporal) annotations represents a solution for having a (temporally) annotated corpus, if a parallel corpus & adequate processing tools exist.
4. Improvements can be done in TimeBank – consistent with TimeML developers (Boguraev, Ando, 2006).

# Future work

## Immediate:

- improvement & evaluation of the annotation transfer
- adequacy of temporal theories to Romanian
- translated and normalized temporal references
- mappings between different behavior of tenses from language to language

## Long-term:

- (semi) automatically mark-up of the temporal information in Romanian texts (news + literature, legislation)

# Acknowledgements

The author is grateful to:

- Dan Tufiş and the RACAI NLP group (especially Radu Ion) for the support and helpful discussions and advices w.r.t. this research
- Dan Cristea, Jerry Hobbs, James Pustejovsky, Marc Verhagen, and Georgiana Puşcaşu for usefull research outcomes coming from discussions
- All **LREC 2008** organizers and reviewers



# Thank you!

## (Temporal) Questions???

