Appendix A
Annotated Corpora and Annotation Tools

A.1 Introduction

TimeML is a standard for annotating time in natural language. It includes annotations for the lexicalised entities TIMEX3, EVENT and SIGNAL, and for the abstract entities TLINK, SLINK, ALINK and MAKEINSTANCE. The syntax is XML-like, with inline annotation. For the temporal link labelling task, one is interested in TIMEX3, EVENT, SIGNAL and TLINK. The MAKEINSTANCE tag gives events extra information and instantiates them for use in TLINKs, and so also contains useful information. TimeML has recently become an ISO standard, ISO-TimeML, which incorporates a few changes to event description and permits stand-off annotation. As almost all prior work and all existing resources use TimeML or an extension thereof, this book considers only TimeML in general.

A.2 Corpora

A.2.1 TimeBank

TimeBank is a human annotated TimeML corpus of 183 newswire texts. TimeBank v1.2 contains 6 418 TLINKs, 1 414 TIMEX3s and 7 935 EVENTs, and is 3004 kB in size. This is tiny compared to some other types of corpus, but is large enough to be useful and has been battered enough by the community through a few versions to be considered robust. TimeBank's creation [1] involved a large human annotator effort and a few different versions [2]; it is currently the largest temporally annotated corpus.

TimeBank 1.2 contains 183 documents, comprising about 64 000 tokens. Over these tokens are:

- 7935 EVENTs
- 6418 TLINKs

L.R.A. Derczynski, *Automatically Ordering Events and Times in Text*,
Studies in Computational Intelligence 677, DOI 10.1007/978-3-319-47241-6

Table A.1 Inter-annotator agreement in TimeBank v1.2; data from [2]

TimeML tag	Exact match IAA
TIMEX3	0.83
EVENT	0.78
TLINK	0.55

Table A.2 Distribution of TIMEX3 type

TIMEX3 type	Frequency	Proportion (%)
DATE	1164	82.3
DURATION	175	12.4
TIME	63	4.46
SET	12	0.849
Total	1414	

Table A.3 Distribution of TIMEX3 mod

TIMEX3 mod	Frequency	Proportion (%)
START	28	30.4
APPROX	16	17.4
END	16	17.4
EQUAL_OR_LESS	8	8.7
MID	7	7.61
EQUAL_OR_MORE	6	6.52
LESS_THAN	4	4.35
MORE_THAN	3	3.26
ON_OR_AFTER	3	3.26
BEFORE	1	1.09
None	0	0.0
Total	92	

- 7940 INSTANCEs
- 688 SIGNALs
- 1414 TIMEX3s
- 2932 SLINKs
- 265 ALINKs

The remainder of this subsection presents summary information over the events, timexes, signals and and temporal relations in TimeBank 1.2 (Tables A.1, A.2 and A.3).

A.2.2 AQUAINT

The second-largest English TimeML corpus is the AQUAINT TimeML corpus. The AQUAINT TimeML corpus consists of around 80 TimeML-annotated newswire documents. These are grouped by the story that they cover, with each group related to the same story, reporting progress on events through time (Table A.4).

Due to repeated text and heavy event co-reference, the AQUAINT corpus requires some care to use correctly. One must firstly maintain document level testing and training set separation, to ensure that evaluation examples are not those found verbatim in training data. Further, due to the corpus' repeated attention to the same story over multiple documents, some event summaries and orderings are repeated using the same text across documents. For this reason, it is best to split datasets by story, so that the background summaries repeated in articles on the same story do not contaminate test and training data. Finally, separately from text re-use, there is re-description of events using later knowledge. Because the news stories contain information on the same topic describing the same events, it is important not to include later articles in the training set for a classifier evaluated on articles published prior. That is to say, evaluation should not be performed using articles that the training data provides hindsight over. This is a common constraint with time-series data [3] and applies to this TimeML corpus because of its repeated coverage of the same super-events (Table A.5).

Table A.4 Distribution of EVENT class

EVENT class	Frequency	Proportion (%)
OCCURRENCE	4215	53.1
STATE	1117	14.1
REPORTING	1028	13.0
I_ACTION	681	8.58
I_STATE	584	7.36
ASPECTUAL	262	3.3
PERCEPTION	48	0.605
Total	7935	

Table A.5 Distribution of EVENT pos

EVENT pos	Frequency	Proportion (%)
VERB	5122	64.5
NOUN	2225	28.0
OTHER	299	3.77
ADJECTIVE	266	3.35
PREPOSITION	28	0.353
Total	7940	

A.2.3 Other TimeML Corpora

There have been other TimeML corpora released, in a range of languages, including French [4], Italian [5] and Romanian [6] (Tables A.6 and A.7).

Table A.6 Distribution of EVENT modality

EVENT modality	Frequency	Proportion (%)
would	127	39.7
could	49	15.3
may	31	9.69
can	26	8.13
none	21	6.56
might	16	5.0
must	14	4.38
should	13	4.06
have to	5	1.56
’d	2	0.625
possible	2	0.625
should have to	2	0.625
close	1	0.313
delete	1	0.313
depending on	1	0.313
have_to	1	0.313
having to	1	0.313
likelihood	1	0.313
potential	1	0.313
to	1	0.313
unlikely	1	0.313
until	1	0.313
would have to	1	0.313
would_be	1	0.313
None	0	0.0
Total	320	

Table A.7 Distribution of EVENT polarity

EVENT polarity	Frequency	Proportion (%)
POS	7651	96.4
NEG	289	3.64
Total	7940	

A.2.4 Other Non-TimeML Corpora

The TempEval corpora [7, 8] feature event, timex and tlink annotations over non-parallel news text in multiple different languages. The set of TLINKs is slightly different from those available in TimeML, being simpler and including a VAGUE relation. TempEval-2 included English, Spanish, French, Italian, Chinese and Korean (Table A.8).

The ACE (Automatic Content Extraction) exercises were based on purpose-built corpora that included a large number of TIMEX2 annotations, comprising almost 26 000 TIMEX2s. For comparison, TimeBank has only 1 414 TIMEX3 annotations (Table A.9).

The WikiWars corpora [9, 10] are derived from WikiPedia articles about wars. These articles tend to contain temporal expressions of a variety of granularities and forms and a generally quite long pieces of connected prose. WikiWars and WikiWars-DE are both annotated according to TIMEX2 and are resources of significant size.

Table A.8 Distribution of TLINK reltype

TLINK reltype	Frequency	Proportion (%)
BEFORE	1408	21.9
IS_INCLUDED	1357	21.1
AFTER	897	14.0
IDENTITY	743	11.6
SIMULTANEOUS	671	10.5
INCLUDES	582	9.07
DURING	302	4.71
ENDED_BY	177	2.76
ENDS	76	1.18
BEGUN_BY	70	1.09
BEGINS	61	0.95
IAFTER	39	0.608
IBEFORE	34	0.53
DURING_INV	1	0.0156
Total	6418	

Table A.9 Transitivity table for the TimeML relation set; an X indicates that no clear inference can be made. Abbreviations: BE = Before, AF = After, IN = Includes, II = Is_included, DU = During, SI = Simultaneous, IA = Iafter, IB = Ibefore, ID = Identity, BG = Begins, EN = Ends, BB = Begun_by, EB = Ended_by, DI = During_inv

A r1 B	B r2 C													
	BE	AF	IN	II	DU	SI	IA	IB	ID	BG	EN	BB	EB	DI
BEFORE	BE	X	BE	X	BE	BE	X	BE	BE	BE	X	BE	BE	BE
AFTER	X	AF	AF	X	AF	AF	AF	X	AF	X	AF	AF	AF	AF
INCLUDES	X	X	IN	X	IN	IN	X	X	IN	X	X	IN	IN	IN
IS_INCLUDED	BE	BE	X	II	II	II	AF	BE	II	II	II	X	X	II
DURING	BE	AF	IN	II	SI	SI	IA	IB	SI	BG	EN	BB	EB	SI
SIMULTANEOUS	BE	AF	IN	II	SI	SI	IA	IB	SI	BG	EN	BB	EB	SI
IAFTER	X	AF	AF	X	IA	IA	AF	X	IA	X	IA	AF	IA	IA
IBEFORE	BE	X	BE	X	IB	IB	X	BE	IB	IB	X	IB	BE	IB
IDENTITY	BE	AF	IN	II	SI	SI	IA	IB	ID	BG	EN	BB	EB	SI
BEGINS	BE	AF	X	II	BG	BG	IA	BE	BG	BG	II	X	X	BG
ENDS	BE	AF	X	II	EN	EN	AF	IB	EN	II	EN	X	X	EN
BEGUN_BY	BE	AF	IN	X	BB	BB	IA	X	BB	X	X	BB	IN	BB
ENDED_BY	BE	X	IN	X	EB	EB	X	IB	EB	X	X	X	EB	EB
DURING_INV	BE	AF	IN	II	SI	SI	IA	IB	SI	BG	EN	BB	EB	SI

A.3 Temporal Annotation Tools

Temporal annotation is a complex task for humans; to this end, we have annotation guidelines to simplify things. Typing XML is also a rather painful experience for us, let alone a specific variant of it that captures abstract information, such as TimeML; and to this end, we have temporal annotation tools that can simplify the task.

In this section, we will first describe TARSQI, a state-of-the-art toolkit containing many components for temporal annotation of text. We will then discuss the problem of visually presenting temporal information.

A.3.1 TARSQI/TTK

A set of tools for automatic TimeML annotation are bundled together in the form of the TARSQI toolkit, TTK [11], which is described as "a modular system for automatic temporal and event annotation of natural language texts" (Fig. A.1). TTK adopts a multi-stage work-flow, beginning with the entry of raw unannotated text, followed by automated annotation and then user correction of machine-produced results. The toolkit ties together a large number of components, including EVITA [12], Slinket [13, 14], SputLink [15] and TBOX [16], using a plugin-based Python framework. It is easy for users to see which plugins have been involved in annotation decisions, making TTK useful for analyzing individual components.

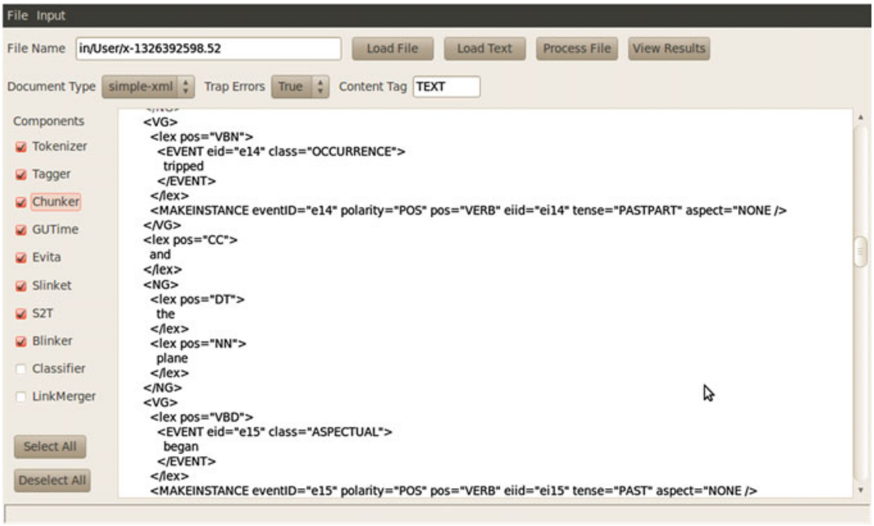

Fig. A.1 Automatically annotating text with TTK

As well as identifying and annotating events and times, TTK also includes sophisticated logic for labelling TLINKs. As far as rule-based relation identification goes, S2T [11, 17] is capable of generating TLINKs from SLINKs and Blinker – based on GutenLink [18] – contains a large set of relation postulations given configurations of EVENTs and TIMEXs and focuses on TLINKs.

Instead of prior versions of the toolkit which permitted co-operation of link annotation components via a voting mechanism [18]. TTK has a separate Link Merger component. The merger uses confidence scores from individual components as well as a pre-set bias (for example, to give low priority to the large number of classifier-generated links) to order candidate links. These are then sequentially tested against a temporal graph of the discourse, with consistency checking between each addition; inconsistent links are not added. This makes it impossible to revoke possibly incorrect information once it has been added, but generates a consistent annotation where high confidence is at least partially rewarded.

A.3.2 Callisto/Tango

TANGO is an assistive annotation tool that helps users build correct annotations from suggestions made by the included automatic temporal annotation systems, as well as a visual representation component. It is integrated within Callisto (Fig. A.2),

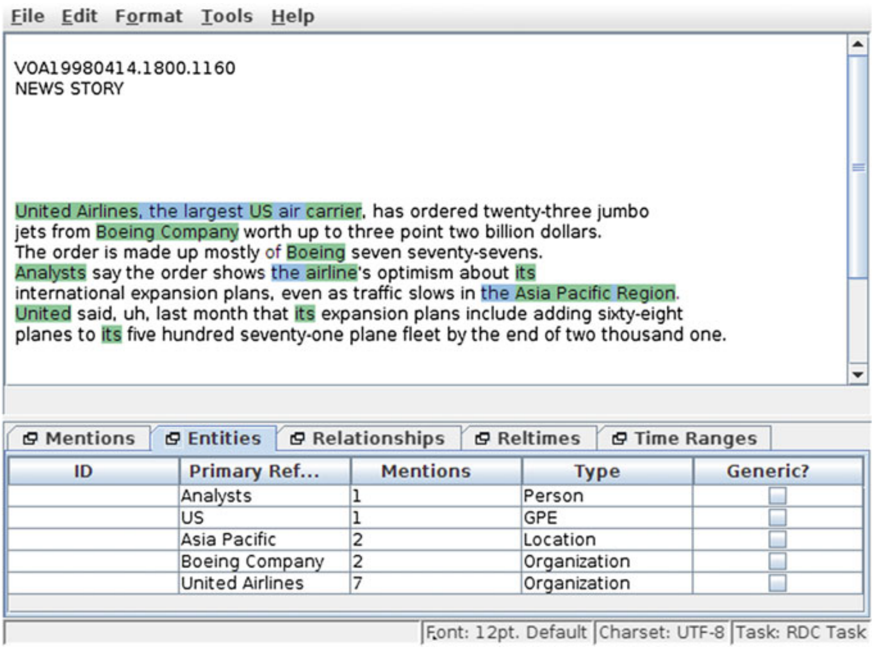

Fig. A.2 Manually annotating text with Callisto a general-purpose manual linguistic annotation tool. Callisto's TANGO component for TimeML annotation is particularly strong for ease of temporal link annotation.

A.3.3 BAT

The Brandeis Annotation Tool, or BAT [19], enables collaborative semantic annotation and breaks down annotation into subtasks. It is a web-based tool, with administrator overview (see Fig. A.3). Multiple asynchronous and concurrent annotations can be made, making BAT a flexible tool for co-ordinating gold standard TimeML annotations. It has been used to create the TempEval-2 and Ita-TimeBank datasets.

A.3.4 Other Tools

Other purpose-built tools exist, such as Dante [20] which concentrates on temporal expression tagging and normalisation across many genres of text but is not are pub-

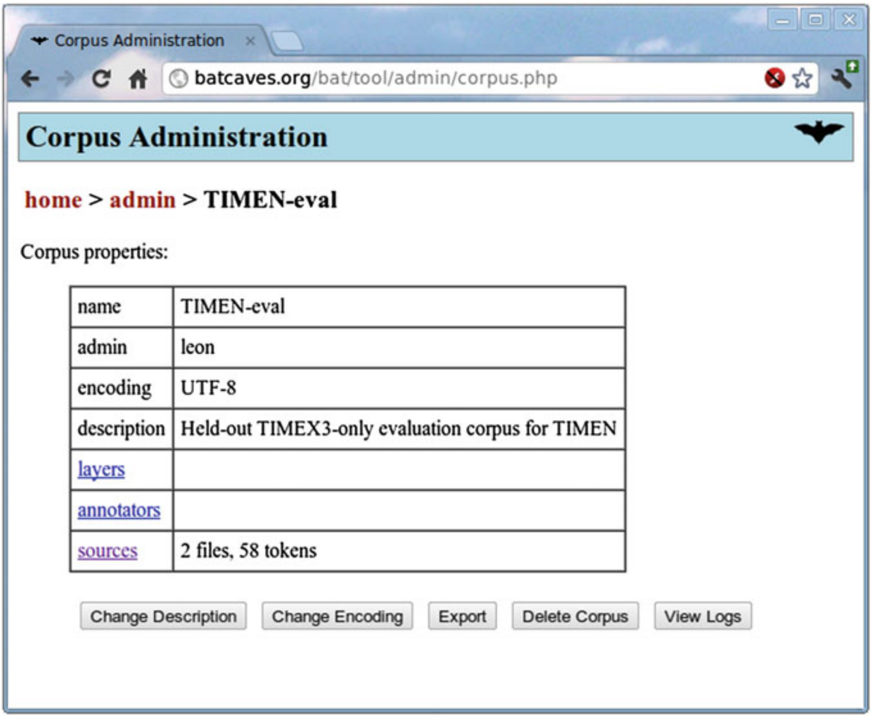

Fig. A.3 Overseeing a BAT annotation project licly available. Existing general purpose language toolkits may also be adapted to cater for TimeML processing, such as NLTK [21], GATE [22] and Xara [23].

References

1. Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al.: The TimeBank Corpus. In: Proceedings of the Corpus Linguistics conference, pp. 647–656 (2003)
2. Boguraev, B., Pustejovsky, J., Ando, R., Verhagen, M.: TimeBank evolution as a community resource for TimeML parsing. Lang. Resour. Eval. **41**(1), 91–115 (2007)
3. Bergmeir, C., Benítez, J.: On the use of cross-validation for time series predictor evaluation. Inf. Sci. **191**, 192–2132 (2012)
4. Bittar, A., Amsili, P., Denis, P., Danlos, L.: French TimeBank: an ISO-TimeML annotated reference corpus. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, vol. 2, pp. 130–134. Association for Computational Linguistics (2011)
5. Caselli, T., Lenzi, V., Sprugnoli, R., Pianta, E., Prodanof, I.: Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In: Proceedings of ACL-HLT 2011, p. 143. Association for Computational Linguistics (2011)
6. Forăscu, C., Ion, R., Tufiş, D.: Semi-automatic Annotation of the Romanian TimeBank 1.2. In: Proceedings of the RANLP 2007 Workshop on Computer-aided language processing (CALP), vol. 30, pp. 1–7 (2007)
7. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J.: The TempEval challenge: identifying temporal relations in text. Lang. Resour. Eval. **43**(2), 161–179 (2009)
8. Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J.: SemEval-2010 task 13: TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 57–62. Association for Computational Linguistics (2010)
9. Mazur, P., Dale, R.: WikiWars: A new corpus for research on temporal expressions. In: Proceedings of the EMNLP, pp. 913–922 (2010)
10. Strötgen, J., Gertz, M.: WikiWarsDE: A German corpus of narratives annotated with temporal expressions. In: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011), pp. 129–134. Hamburg, Germany (2011)
11. Verhagen, M., Pustejovsky, J.: Temporal processing with the TARSQI toolkit. In: Proceedings of CoLing: Posters and Demonstrations, pp. 189–192 (2008)
12. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: a robust event recognizer for QA systems. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p. 707. Association for Computational Linguistics (2005)
13. Saurí, R., Verhagen, M., Pustejovsky, J.: Slinket: A partial modal parser for events. In: Proceedings of the conference on Language Resources and Evaluation Conference (LREC) (2006)
14. Saurı, R., Verhagen, M., Pustejovsky, J.: Annotating and recognizing event modality in text. In: The 19th International FLAIRS Conference, FLAIRS 2006 (2006)
15. Verhagen, M.: Temporal closure in an annotation environment. Lang. Resour. Eval. **39**(2), 211–241 (2005)
16. Verhagen, M.: Drawing TimeML relations with TBox. Lect. Notes Comput. Sci. **4795**, 7 (2007)
17. Verhagen, M.: Times Between the Lines. Ph.D. thesis, Brandeis University (2004)

18. Verhagen, M., Mani, I., Sauri, R., Knippen, R., Jang, S., Littman, J., Rumshisky, A., Phillips, J., Pustejovsky, J.: Automating temporal annotation with TARSQI. In: Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions, p. 84. Association for Computational Linguistics (2005)
19. Verhagen, M.: The Brandeis annotation tool. Lang. Resour. Eval. Conf. LREC **2010**, 3638–3643 (2010)
20. Mazur, P., Dale, R.: The DANTE temporal expression tagger. In: Human Language Technology. Challenges of the Information Society, p. 257. Springer, New York (2009)
21. Loper, E., Bird, S.: NLTK: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, vol. 1, pp. 63–70. Association for Computational Linguistics (2002). http://www.nltk.org/
22. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W., Derczynski, L., et al.: Developing Language Processing Components with GATE Version 8 (a User Guide). University of Sheffield (2001). https://gate.ac.uk/releases/gate-8.2-build5482-ALL/doc/tao/tao.pdf
23. Burnard, L., Todd, T.: Xara: an XML aware tool for corpus searching. Proc. Corpus Linguist. **2003**, 142–144 (2003)

Appendix B
RTMML Reference

This appendix details extensions made to TimeML, so that it may capture extra information helpful for temporal reasoning, based upon Reichenbach's framework of tense and aspect [1].

B.1 Examples

B.1.1 Fiction

From *David Copperfield* by Charles Dickens:

Example 34 When he had put up his things for the night he took out his flute, and blew at it, until I almost thought he would gradually blow his whole being into the large hole at the top, and ooze away at the keys.

We give RTMML for the first five verbal events from Example 34 RTMML in Fig. B.1. The fifth, `v5`, exists in a context that is instantiated by `v4`; its reference time is defined as such. We can use one `link` element to show that `v2`, `v3` and `v4` all use the same reference time as `v1`. The temporal relation between event times of `v1` and `v2` can be inferred from their shared reference time and their tenses; that is, given that `v1` is anterior past and `v2` simple past, we know $E_{v1} < R_{v1}$ and $E_{v2} = R_{v2}$. As our `<rtmlink>` states $R_{v1} = R_{v2}$, then $E_{v1} < E_{v2}$. Finally, `v5` and `v6` happen in the same context, described with a second SAME_TIMEFRAME link.

B.1.2 Editorial News

From an editorial piece in TimeBank [2] (AP900815-0044.tml):

L.R.A. Derczynski, *Automatically Ordering Events and Times in Text*,
Studies in Computational Intelligence 677, DOI 10.1007/978-3-319-47241-6

```
<doc time="1850" mod="BEFORE" />
<!-- had put -->
<verb xml:id="v1"
  target="#range(#token2,#token3)"
  view="anterior" tense="past" />
<!-- took -->
<verb xml:id="v2" target="#token11"
  view="simple" tense="past" />
<!-- blew -->
<verb xml:id="v3" target="#token17"
  view="simple" tense="past" />
<!-- thought -->
<verb xml:id="v4" target="#token24"
  view="simple" tense="past" />
<!-- would gradually blow -->
<verb xml:id="v5"
  target="#range(#token26,#token28)"
  view="posterior" tense="past"
  se="=" er=">" sr=">"
  r="#v4.e" />
<!-- ooze -->
<verb xml:id="v6"
  target="#range(#token26,#token28)"
  view="posterior" tense="past"
  se="=" er=">" sr=">" />
<rtmlink xml:id="l1"
  type="SAME_TIMEFRAME">
  <link target="#v1" />
  <link target="#v2" />
  <link target="#v3" />
  <link target="#v4" />
</rtmlink>
<rtmlink xml:id="l2"
  type="SAME_TIMEFRAME">
  <link target="#v5" />
  <link target="#v6" />
</rtmlink>
```

Fig. B.1 RTMML for a passage from David Copperfield

Example 35 Saddam appeared to accept a border demarcation treaty he had rejected in peace talks following the August 1988 cease-fire of the eight-year war with Iran.

```
<doc time="1990-08-15T00:44" />
<!-- appeared -->
<verb xml:id="v1" target="#token1"
  view="simple" tense="past" />
<!-- had rejected -->
<verb xml:id="v2"
  target="#range(#token9,#token10)"
  view="anterior" tense="past" />
<rtmlink xml:id="l1"
  type="SAME_TIMEFRAME">
  <link target="#v1" />
  <link target="#v2" />
</rtmlink>
```

Here, we relate the simple past verb *appeared* with the anterior past (past perfect) verb *had rejected*, permitting the inference that the first verb occurs temporally after the second. The corresponding TimeML (edited for conciseness) is:

Example 36

```
Saddam <EVENT eid="e74" class="I_STATE">
appeared</EVENT> to accept a border demarcation treaty he had <EVENT eid="e77"
class="OCCURRENCE">rejected</EVENT>

<MAKEINSTANCE eventID="e74" eiid="ei1568"
 tense="PAST" aspect="NONE" polarity="POS"
 pos="VERB"/>
<MAKEINSTANCE eventID="e77" eiid="ei1571"
 tense="PAST" aspect="PERFECTIVE"
 polarity="POS" pos="VERB"/>
```

In this example, we can see that the TimeML annotation includes the same information, but a significant amount of other annotation detail is present, cluttering the information we are trying to see. Further, these two <EVENT> elements are not directly linked, requiring transitive closure of the network described in a later set of <TLINK> elements, which are omitted here for brevity.

B.1.3 Linking Events to Calendar References

RTMML makes it possible to precisely describe the nature of links between verbal events and times, via positional use of the reference point. We will link an event to a temporal expression, and suggest a calendrical reference for that expression, allowing the events to be placed on a calendar. Consider the below text, from wsj_0533.tml in TimeBank.

Example 37 At the close of business Thursday, 5,745,188 shares of Connaught and C$44.3 million face amount of debentures, convertible into 1,826,596 common shares, had been tendered to its offer.

```
<doc time="1989-10-30" />
<!-- close of business Thursday -->
<timerefx xml:id="t1"
  target="#range(#token2,#token5)" />
<!-- had been tendered -->
<verb xml:id="v1"
  target="#range(#token25,#token27)"
  view="anterior" tense="past" />
<rtmlink xml:id="l1" target="#t1 #v1">
  <link target="#t1" />
  <link target="#v1" />
</rtmlink>
```

This shows that the reference time of `v1` is `t1`. As `v1` is anterior, we know that the event mentioned occurred before *close of business Thursday*. Normalisation is not a task that RTMML addresses, but there are existing methods for deciding which Thursday is being referenced given the document creation date [3]; a time of day for *close of business* may be found in a gazetteer.

B.2 Annotation Notes

As can be seen in Table 6.2, there is not a one-to-one mapping from English tenses to the nine specified by Reichenbach. In some annotation cases, it is possible to see how to resolve such ambiguities. Even if view and tense are not clearly determinable, it is possible to define relations between S, E and R; for example, for arrangements corresponding to the simple future, $S < E$. In cases where ambiguities cannot be resolved, one may annotate a disjunction of relation types; in this example, we might say "$S < R$ or $S = R$" with `sr="<="`.

Contexts seem to have a shared speech time, and the $S - R$ relationship seems to be the same throughout a context. Sentences which contravene this (e.g. *"By the time I ran, John will have arrived"*) are rather awkward. Contexts are typically broken by timexes (e.g. positional use of the reference point), shifting of frame of reference by use of *"then"*, use of temporal signals or any boundary of reported speech (e.g. starting and ending quotes).

RTMML annotation is not bound to a particular language. As long as a segmentation scheme (e.g. WordSeg-1) is agreed and there is a compatible system of tense and aspect, the model can be applied and an annotation created.

References

1. Reichenbach, H.: The tenses of verbs. In: Elements of Symbolic Logic. Dover Publications, New York (1947)
2. Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al.: The TimeBank Corpus. In: Proceedings of the Corpus Linguistics conference, pp. 647–656 (2003)
3. Mazur, P., Dale, R.: Whats the date? High accuracy interpretation of weekday. In: 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp. 553–560 (2008)
4. Derczynski, L., Gaizauskas, R.: An Annotation Scheme for Reichenbach's Verbal Tense Structure. In: Proceedings of the 6th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA 6), Oxford, UK (2011)
5. Derczynski, L., Gaizauskas, R.: Empirical Validation of Reichenbach's Tense Framework. In Proceedings of the 10th Conference on Computational Semantics (IWCS), Potsdam, Germany (2013)

Appendix C
CAVaT Reference

This section contains a reference guide for the CAVaT package [1]. Up to date information can always be found at https://github.com/leondz/cavat.

C.1 Installation and Configuration

The first time CAVaT is run, it will attempt to create a directory `$HOME/.cavat/`, where it will store its SQLite files.

C.2 Getting Started

Enter the following to load a TimeML corpus into the "test" database - it's important to include the trailing slash `/` in the path:

```
cavat> corpus import /home/user/corpus/data/ to test
```

Depending on your disk and CPU speeds, this might take about 10–20 seconds per megabyte of TimeML. If it seems to take longer, you can get more information about what CAVaT is doing during import by enabling debug mode before import:

```
cavat> debug on
```

Leave debug mode with a simple:

```
cavat> debug off
```

Once the corpus has loaded, you can use `corpus info` to see metadata about the import, or `corpus list` to see an available list of corpora. To switch between corpora, and to select a newly loaded one, enter `corpus use <name>`.

L.R.A. Derczynski, *Automatically Ordering Events and Times in Text*,
Studies in Computational Intelligence 677, DOI 10.1007/978-3-319-47241-6

C.3 Queries

The show command is used for generating reports on the currently loaded corpus. Reports focus on one tag type, and give information about their attributes. One can view all values for a tag with `list` reports, or the distribution of values with `distribution` reports, or simply see how many instances of that tag list a value for a field with `state` reports.

Reports can be provided in multiple formats; there is:

- `screen` - for screen or fixed-width font output
- `csv` - comma separated values
- `tex` - LaTeX table format

The general format for report generation is:

`show <report type> of <tag> <field> [as <format>]`

To try a simple query, enter:

`cavat> show distribution of tlink reltype`

You should see a table listing the values listed for relType in the current corpus' TLINK tags, as well as their frequencies. To see how many TLINKs use a signal, and use the result in a LaTeX document, you can try:

`cavat> show state of tlink signalid as tex`

Reference

1. Derczynski, L., Gaizauskas, R.: Analysing temporally annotated corpora with CAVaT. In: Proceedings of the Language Resources and Evaluation Conference, pp. 398–404 (2010)

Index

L.R.A. Derczynski, *Automatically Ordering Events and Times in Text*,
Studies in Computational Intelligence 677, DOI 10.1007/978-3-319-47241-6