Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement

André Bittar[1], Caroline Hagège[1], Véronique Moriceau[2], Xavier Tannier[2], Charles Teissèdre[3]

(1) Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan
France
first.last@xrce.xerox.com

(2) LIMSI-CNRS
University Paris-Sud 11
B.P. 133 - 91403 Orsay Cedex
France
first.last@limsi.fr

(3) MoDyCo-CNRS - UMR 7114
Université Paris Ouest Nanterre La Défense
Bâtiment A - Bureau 401B
200, avenue de la République
92001 Nanterre Cedex
France
charles.teissedre@gmail.com

Abstract

This article presents work carried out within the framework of the ongoing ANR (French National Research Agency) project Chronolines, which focuses on the temporal processing of large news-wire corpora in English and French. The aim of the project is to create new and innovative interfaces for visualizing textual content according to temporal criteria. Extracting and normalizing the temporal information in texts through linguistic annotation is an essential step towards attaining this objective. With this goal in mind, we developed a set of guidelines for the annotation of temporal and event expressions that is intended to be compatible with the TimeML markup language, while addressing some of its pitfalls. We provide results of an initial application of these guidelines to real news-wire texts in French over several iterations of the annotation process. These results include inter-annotator agreement figures and an error analysis. Our final inter-annotator agreement figures compare favorably with those reported for the TimeBank 1.2 annotation project.

Keywords: temporal annotation, annotation guidelines, inter-annotator agreement

1. Introduction

The processing of temporal information in natural language texts is essential for full text understanding and this area of research has received increasing attention over recent years. Previous work has resulted in the development of annotation schemata (Ferro et al., 2005), and annotated corpora (Bittar, 2010) (Pustejovsky et al., 2003a; Russo et al., 2011), most notably within the framework of TimeML (Pustejovsky et al., 2003b). The work presented here was carried out as part of the ongoing ANR Chronolines project[1] which focuses on the temporal processing of large news-wire corpora in English and French. The main objective of the project is to create innovative interfaces for visualizing textual content according to temporal criteria. Within this framework, the aim of the work presented here is to render explicit the temporal information that is linguistically realized in texts. Our aim is to annotate all linguistic expressions that convey any kind of temporal information. Expressions that designate a temporal anchoring (unique or multiple) on a timeline, or expressions conveying the idea of temporal duration needed to be considered. One specificity of our approach is that it is not limited simply to date expressions[2]. We also deal with prepositional phrases and embedded clauses headed by nominal or verbal events, when these expressions convey temporal information (e.g. *after he came*, *during the pope's visit*, etc). This work draws on preceding research efforts, and is intended to be compatible with TimeML, while addressing some of its pitfalls. First, we present the main points of our annotation guidelines (the entire annotation guide is available as a deliverable of French ANR project Chronolines) and then report results of an annotation experiment according to these guidelines.

2. Motivation

Our motivation is to focus on what we consider to be important for dealing with temporal objects. In particular:

1. We consider that temporal annotation can only be carried out properly by taking into account the full context of expressions, as opposed to TimeML, which aims for surface-based annotation. This aspect has been stressed in previous work (Ehrmann and Hagège, 2009).

2. Our choices are linguistically founded. We provide linguistic tests in order to help annotators in the annotation task.

3. We provide an annotation schema that is flexible enough to accommodate or integrate other approaches, for example, a functional approach (see Section 5.).

4. As work is being done on both English and French, we ensure that our proposal is applicable to both these languages (and hopefully to other languages as well).

With these elements in mind we defined a typology of temporal expressions, similar to that set out in (Ehrmann and Hagège, 2009) for French. Details are given in the full guidelines. Here, we focus only on the syntactic information concerning temporal expressions, without looking at

[1] ANR-10-CORD-010, `http://www.chronolines.fr`. Thanks to the Agence France Presse (AFP) for providing the corpus.

[2] As "date" we consider calendar dates, referential expressions like *tomorrow*, *three weeks after* and expressions headed by a lexical trigger, such as a day or month name, etc.

interpretation (to be carried out at a later stage). For example, we do not normalize relative dates to their absolute values, we merely indicate if a date is relative (deictic or anaphoric) or absolute.

3. Annotation Guidelines

This section gives a brief summary of the main points of the annotation guidelines.

3.1. Temporal Relation Markers

These are practically identical to the markables annotated with the <SIGNAL> tag in TimeML. These expressions mark a temporal relation between two other elements. They can be stand-alone[3] expressions or be included within a larger temporal expression. For example, in the sentence *Meanwhile, John did the dishes*, the stand-alone expression *meanwhile* indicates a temporal relation (simultaneity) between a previously mentioned event and the event of John's doing the dishes. As in TimeML, the <SIGNAL> tag marks these expressions. In the sentence *Mary arrived after John left*, the conjunction *after* also expresses a temporal relation (sequence) between John's departure and Mary's arrival. In this case, however, as we will see below, *after* in this context is not stand-alone and needs to be related to the departure event for a correct interpretation.

3.1.1. Temporal Expressions

We distinguish different types of temporal expressions: durations (answer the question *for how long?*, equivalent to TimeML DURATION type expressions), temporal aggregates (which answer the question *how often/how frequently?*, which correspond roughly to TimeML SET type expressions), and finally what we call *Temporal Localization Expressions* (TLEs). This last class includes dates (answer the question *when?*) and what we call *Event Temporal Expressions* (ETEs, described below).

Durations and dates have the following common characteristic: their syntactic head is always a temporal lexical trigger (e.g. *January, Thursday, hour, year, week*, etc.) or has an explicit temporal interpretation (e.g. *10/2011, 1989.01.16*). The guidelines contain a list of exactly what these temporal lexical triggers are. Both these types of expressions are **atomic** temporal expressions (contiguous span) and are marked up with the <TEMPEX> tag.

As mentioned above, with dates, we consider ETEs, which also provide an answer to the question *when?*. However, the syntactic head of these expressions is not a temporal lexical trigger, but an event (nominal, verbal or adjectival). For example, in *John left* **after Mary's arrival**, the expression in bold answers the question *when?*, clearly indicating that it has a temporal value. For these kinds of expressions we decided to annotate both the temporal relation marker (<SIGNAL>) that introduces the expression (*after*) as well as the head of the event (*arrival*). This choice was made in order to avoid the difficulties of identifying the boundaries of the expressions, in particular subordinate clauses and event arguments. Thus, these annotations are **non-atomic** (non-contiguous span). The two marked constituents are linked by a relation, <CONNECT>, to show they belong to the same ETE. This differs from current approaches, such as TimeML, that require temporal expressions to be headed by a temporal lexical trigger. We believe that ETEs and dates should be treated in the same way as they both denote a temporal interval (and may, in principle, be normalized) and are both introduced by the same temporal relation markers.

Segmentation of dates, durations and aggregates is carried out according to the criteria set out in (Hagège and Tannier, 2008). For example, for *John arrives on Thursday at 10am*, according to the criteria the temporal expression must be segmented into *on Thursday* and *at 10am*. However, in *John arrived before Thursday at 10am* the temporal expression *before Thursday at 10am* cannot be segmented.

3.1.2. Events

We adopt the same definition for events as that used in TimeML. This definition corresponds to what is usually termed "eventualities" (Bach, 1986) and includes both events and states. It must be kept in mind, however, that we only annotate for the time being those events that are part of an ETE. Events are marked up with the <EVENT> tag.

3.2. Relations

At this stage we are only focused on the annotation of what is linguistically realized in the text, the markables. The only kind of relation we consider for the moment is the <CONNECT> relation mentioned above (relations linking a temporal relation marker and a nominal, verbal or adjectival event head). Temporal relations will be dealt with in future work.

4. Annotation Experiment

A manual annotation exercise was carried out with several aims in mind:

- Test the annotation guidelines on "real" texts in order to get an idea of the schema's coverage and to identify any points which would require modification.
- Constitute a gold standard corpus for the evaluation of an automatic annotation system.
- Measure inter-annotator agreement to determine the human benchmark for the task.

Five annotators, with varying levels of linguistic training and annotation experience, took part in the exercise. Four of the annotators had prior experience with the temporal annotation task, three having good knowledge of TimeML. The remaining annotator was inexperienced with linguistic annotation. Four annotators were French native speakers and one was highly proficient. The Glozz annotation tool (Widlöcher and Mathet, 2009) was used, as it allows for the annotation of both the markables (<TEMPEX>, <EVENT> and <SIGNAL>) and relations (<CONNECT> between <SIGNAL> and <EVENT>). Annotators were provided with the full guidelines as well as a quick reference guide containing the main points and illustrative examples to facilitate and speed up consultation. The

[3] This means that the expression is not a syntactic dependent of a governor in the same clause.

texts were not preprocessed in any way before annotation. Inter-annotator agreement scores were calculated across all annotator pairings. Both F-score and Kappa (Cohen, 1960)[4] were measured for the identification of markables (attributes were not taken into account at this stage). Only exact matches on tag spans were taken as an agreement. To measure agreement on the <CONNECT> relation, only exact matches on start and end offsets of both the relation's arguments were considered as agreement[5]. We carried out two rounds of annotation, for which we will now give a detailed description.

4.1. Annotation Round 1

The first round of annotation represented the first application of the guidelines to real texts. For this round, a total of 50 news-wire articles written in French from the AFP corpus were chosen. Two annotators marked up 16 texts each and the other three had 17 texts each. Each text was marked up at least twice by different annotators. This provided a total of 6 annotator pairings across all documents. Annotators did not consult each other during the process. Once all documents had been annotated, inter-annotator agreement scores were calculated, with differences between two annotated documents being recorded. Table 1 gives the average agreement figures over all annotator pairings.

Average Round 1		
	F_1	κ
Temporal Expressions		
<TEMPEX> (dates, durations, aggregates)	0.80	0.54
<CONNECT> (ETEs)	0.39	0.04
Global	0.60	0.29
	F_1	κ
Other markables		
<SIGNAL>	0.52	-0.07
<EVENT>	0.23	-0.03

Table 1: Average inter-annotator agreement scores (F-score and Kappa) for Annotation Round 1.

Agreement figures for the first round of annotation were much lower than what had been hoped for, especially for events. Upon examination of the documents, it was immediately evident that the very low agreement scores for the <EVENT> tag was due to confusion over how to carry out the task itself, as each annotator had used a different strategy. This is reflected in the negative kappa scores, that indicate that agreement was worse than that expected by chance alone. The poor agreement lead us to review and clarify annotation guidelines and ensure that all annotators agreed on the aim of the different subtasks before carrying out further annotation rounds.

[4]General formulae used are as follows: $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$ and $\kappa = \frac{A_o - A_e}{1 - A_e}$ where A_o = observed agreement and A_e = expected (chance) agreement.

[5]This was done due to the fact that the <CONNECT> relation represents a markable in the text.

4.2. Annotation Round 2

Following the modifications and improvements to the annotation guidelines after Round 1, a second annotation round was carried out. Annotation Round 2 was carried out on a slightly smaller corpus of 30 articles. As before, articles were taken from the French section of the AFP corpus. Four of the five annotators who took part in Round 1 also took part in the second round, giving four pairings. Each annotator marked up a total of 15 texts and each text was annotated by two different annotators. Agreement figures for Round 2, including the observed improvement, are given in Table 2 (page 4).

For Round 2, inter-anntotator agreement across all annotations was significantly higher than for Round 1. These results reflect the fact that annotators were now more familiar with the guidelines, which had been clarified and updated. The dramatic improvement in annotation of events suggests that the task had been approached in the same way this time around. Agreement errors for all markables were due to either (i) complete disagreement through missing or spurious tags (a tag appears in one annotator's document, but not the other's), or (ii) partial disagreement through differing tag span (the extent of annotated tags do not match exactly in both documents).

Complete disagreement: Complete disagreement was due to either **silence** (an annotator missed an annotation) or **noise** (an annotator wrongly added an annotation). Table 3 shows figures for the different types of disagreements for the different markables. For the <TEMPEX> and <SIGNAL> tags a certain proportion of silence was deemed to be attributable to obvious annotator inattention. A disagreement was attributed to inattention if a annotator had missed an apparently obvious expression: an alphanumerical date (e.g. *lundi 3 juin 2003* (Monday 3rd June 2003)), an obvious relation marker (e.g. *avant 16h* (before 4pm)). This heuristic provided ballpark proportions of 41% disagreements due to inattention for <TEMPEX> and 45% for <SIGNAL>. This was not so easily measured for events, which require much more contextual information and are more subjective to annotate.

<TEMPEX>	Count	%
Silence	144	72.3%
Noise	54	27.7%
Total	199	
<SIGNAL>	Count	%
Silence	111	77.6%
Noise	46	22.4%
Total	143	
<EVENT>	Count	%
Silence	57	68.6%
Noise	26	31.4%
Total	83	

Table 3: Types of disagreement for markables in Round 2.

Partial disagreement: Partial disagreements occurred when two annotators marked up the same item with slightly

	Temporal Expressions						<SIGNAL>		<EVENT>	
	GLOBAL		<TEMPEX> (dates, durations, aggregates)		<CONNECT> (ETEs)					
	F_1	κ	F_1	κ	F_1	κ	F_1	κ	F_1	κ
Pair1	0.76	0.44	0.85	0.67	0.67	0.2	0.62	0.15	0.73	0.36
Pair3	0.75	0.41	0.84	0.63	0.66	0.19	0.80	0.55	0.74	0.39
Pair4	0.84	0.62	0.88	0.72	0.79	0.51	0.86	0.68	0.80	0.55
Pair6	0.76	0.45	0.80	0.55	0.72	0.34	0.64	0.14	0.72	0.35
Average Round 2	0.78	0.48	0.84	0.64	0.71	0.31	0.73	0.38	0.75	0.41
Improvement	0.18	0.19	0.04	0.10	0.32	0.27	0.21	0.45	0.52	0.44

Table 2: Inter-annotator agreement scores (F-score and Kappa) for Annotation Round 2.

different boundaries, reflecting a difference in the way each applied guidelines. Several examples of partial disagreements encountered are given in Table 4.

Annotator 1	Annotator 2
dans la nuit de jeudi à vendredi	*dans la nuit*
Mercredi à 01H45 GMT	*à 01H45 GMT*
vendredi	*vendredi et samedi*

Table 4: Examples of annotated expressions yielding partial disagreements between two annotators.

141 partial disagreements occurred for the <TEMPEX> tag. Due to the fact that this tag can span multiple word tokens, it was the most frequent source of errors. 30 partial disagreements were found on the <SIGNAL> tag. No partial disagreements occurred for event expressions as these nearly always covered a single word token.

4.3. Annotation Round 3

A third and final round of annotations was carried out with the same annotators as Round 2 and the same corpus. This time a marked improvement was to be expected, as annotators had already seen the texts, although they were not made aware of the errors they may have made in the previous round. Agreement figures for this final round of annotation are represented in Table 5 (page 5).

A marked increase in agreement was noticed, as was expected. Agreement had reached an acceptable level (¿0.85 F-score) for each of the markables. At this stage, we could start thinking about constituting a reference corpus to use for evaluation purposes.

Complete disagreement: Complete disagreement diminished greatly in Round 3. This indicated a greater consensus on what to annotate and a more coherent application of the guidelines on the part of the annotators. Table 6 shows figures for disagreement types across markables.

Partial disagreement: For temporal expressions there were 59 partial disagreements, due either to segmentation differences (1 annotation or 2), or left or right annotation boundary. Only 8 partial disagreements occurred for relation markers. Finally, there were no partial disagreements on events as events only cover, in the vast majority of cases, a single word token.

<TEMPEX>	Count	%
Silence	48	72.3%
Noise	31	27.7%
Total	79	
<SIGNAL>	Count	%
Silence	59	77.6%
Noise	18	22.4%
Total	26	
<EVENT>	Count	%
Silence	48	90.6%
Noise	5	9.4%
Total	53	

Table 6: Types of disagreement for markables in Round 3.

4.4. Comparison with TimeBank 1.2

Although the elements to be annotated were not exactly the same as the markables of TimeML, there are certain similarities that warrant a comparison with the TimeBank 1.2 corpus. The <TEMPEX> tag (atomic temporal expressions) corresponds roughly to the <TIMEX3> tag in TimeML, while the <SIGNAL> tag is used identically to that in TimeML, the <EVENT> tag has not been used in the same way and therefore is not comparable at this stage. Figures compare favourably to those reported for the TimeBank 1.2 corpus, presented below in Table 7 (the corresponding figures for our experiment are given in brackets for direct comparison).

TimeML tag	Agreement (F_1)
<TIMEX3>	0.83 (0.89)
<SIGNAL>	0.77 (0.92)

Table 7: Reported inter-annotator agreement for TimeBank 1.2.

5. Perspectives and Future Work

The work described is a first step towards a more comprehensive effort for temporal annotation, remaining compatible with the TimeML approach. After three separate annotation rounds, it is clear that the guidelines have become stable and there is a basis of documents from which to constitute an evaluation corpus of good quality. At a later stage, we aim to take into account temporal ordering relations. We also intend to integrate a fine-grained functional approach

	Temporal Expressions						<SIGNAL>		<EVENT>	
	GLOBAL		<TEMPEX> (dates, durations, aggregates)		<CONNECT> (ETEs)					
	F_1	κ	F_1	κ	F_1	κ	F_1	κ	F_1	κ
Pair1	0.83	0.61	0.89	0.75	0.77	0.46	0.82	0.60	0.73	0.36
Pair3	0.90	0.77	0.92	0.83	0.87	0.71	0.91	0.81	0.74	0.39
Pair4	0.91	0.80	0.88	0.88	0.87	0.71	0.95	0.89	0.80	0.55
Pair6	0.88	0.72	0.90	0.77	0.85	0.67	0.89	0.75	0.72	0.35
Average Round 3	0.89	0.76	0.92	0.83	0.86	0.70	0.92	0.82	0.87	0.71
Improvement	0.11	0.28	0.18	0.19	0.15	0.39	0.19	0.44	0.12	0.30

Table 5: Inter-annotator agreement scores (F-score and Kappa) for Annotation Round 3.

(Battistelli et al., 2008). It will also be interesting to see to what point the approach we have presented is transferable to other languages.

6. References

Emmon Bach. 1986. The Algebra of Events. *Linguistics and Philosophy*, 9(1):5–16.

Delphine Battistelli, Javier Couto, Jean-Luc Minel, and Sylviane Schwer. 2008. Representing and Visualizing Calendar Expressions in Texts. In *Proceedings of STEP'08 (Symposium on Semantics in Systems for Text Processing)*, Venice, September.

André Bittar. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Ph.D. thesis, Université Paris Diderot, Paris, November.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 43(6):551–558.

Maud Ehrmann and Caroline Hagège. 2009. Proposition de caractérisation et de typage des expressions temporelles en contexte. In *TALN 2009 (Traitement Automatique des Langues Naturelles)*, Senlis, France, June.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson, 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions*, September.

Caroline Hagège and Xavier Tannier. 2008. XTM: A Robust Temporal Text Processor. In *Computational Linguistics and Intelligent Text Processing, proceedings of 9th International Conference CICLing 2008*, pages 231–240, Haifa, Israel. Springer Berlin / Heidelberg.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

Irene Russo, Tommaso Caselli, and Francesco Rubino. 2011. Recognizing Deverbal Events in Context. In *Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), poster session*, Tokyo, Japan, February. Springer.

Antoine Widlöcher and Yann Mathet. 2009. La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009*, Senlis, June.