French TimeBank: An ISO-TimeML Annotated Reference Corpus

André Bittar
Alpage
Univ. Paris Diderot
andre.bittar@linguist.jussieu.fr

Pascal Amsili
LLF
Univ. Paris Diderot
amsili@linguist.jussieu.fr

Pascal Denis
Alpage
INRIA
pascal.denis@inria.fr

Laurence Danlos
Alpage
Univ. Paris Diderot
danlos@linguist.jussieu.fr

Abstract

This article presents the main points in the creation of the French TimeBank (Bittar, 2010), a reference corpus annotated according to the ISO-TimeML standard for temporal annotation. A number of improvements were made to the markup language to deal with linguistic phenomena not yet covered by ISO-TimeML, including cross-language modifications and others specific to French. An automatic pre-annotation system was used to speed up the annotation process. A preliminary evaluation of the methodology adopted for this project yields positive results in terms of data quality and annotation time.

1 Introduction

The processing of temporal information (events, time expressions and relations between these entities) is essential for overall comprehension of natural language discourse. Determining the temporal structure of a text can bring added value to numerous NLP applications (information extraction, Q&A systems, summarization...). Progress has been made in recent years in the processing of temporal data, notably through the ISO-TimeML standard (ISO, 2008) and the creation of the TimeBank 1.2 corpus (Pustejovsky et al, 2006) for English. Here we present the French TimeBank (FTiB), a corpus for French annotated in ISO-TimeML. We also present the methodology adopted for the creation of this resource, which may be generalized to other annotation tasks. We evaluate the effects of our methodology on the quality of the corpus and the time taken in the task.

2 ISO-TimeML

ISO-TimeML (ISO, 2008) is a surface-based language for the marking of events (`<EVENT>` tag) and temporal expressions (`<TIMEX3>`), as well as the realization of the temporal (`<TLINK>`), aspectual (`<ALINK>`) and modal subordination (`<SLINK>`) relations that exist among these entities. The tags' attributes capture semantic and grammatical features such as event class, tense, aspect and modality, and the type and normalized interpretative value of temporal expressions. The `<SIGNAL>` tag is used to annotate relation markers, such as *before* and *after*. A set of resources for English has been developed over the years, including an annotated corpus, TimeBank 1.2 (TB1.2)[1], which has become a reference for temporal annotation in English.

3 Improving ISO-TimeML

We propose a number of improvements to ISO-TimeML to deal with as yet untreated phenomena. These include both cross-language annotation guidelines, as well as guidelines specific to French. All these guidelines are implemented in the FTiB.

Cross-language Improvements : ISO-TimeML currently provides for the annotation of event **modality** by capturing the lemma of a modal on a subordinated event tag in the `modality` attribute. Inspired by the fact that in French, modality is expressed by fully inflected verbs, we propose that those verbs be tagged as modal, and we

[1] Annotated according to the TimeML 1.2 specification, as opposed to the more recent ISO-TimeML.

provide a set of normalized values for the modality attribute, within a manual annotation context, that reflect the classic classes of linguistic modality (Palmer, 1986): NECESSITY and POSSIBILITY (epistemic), OBLIGATION and PERMISSION (deontic). We also provide a way of capturing the difference between **support verb constructions** with a neutral aspectual value (*mener une attaque* (*carry out an attack*)) and those with an inchoative aspectual value (*lancer une attaque* (*launch an attack*)). ISO-TimeML encodes the relation between the verb and its nominal argument via a <TLINK> of type IDENTITY. We encode aspectual variants in the FTiB by using an <ALINK>. A significant proportion (13/36) of the annotated <ALINK> tags in the FTiB (36%) are used in this case. A third improvement we propose is the introduction of the **event class** EVENT_CONTAINER[2] to distinguish predicates that take an event nominal as subject. In TB1.2, these predicates were sometimes marked, but not distinguished from the OCCURRENCE class. The distinction is appropriate as these predicates have events as arguments, unlike OCCURRENCEs. The relative frequency of this class (19 occurrences) compared to the standard PERCEPTION class (10) also justifies its use. Although not yet dealt with in ISO-TimeML, **aspectual periphrases**, such as *en train de* + V_{inf} (*akin to the English progressive -ing*), adding an aspectual value to an event, are captured in the FTiB in the aspect attribute for events. We also propose a **new value** for aspect, PROSPECTIVE, encoding the value of the construction *aller* + V_{inf} (*going to* + V_{inf}), as in *le soleil va exploser* (*the sun is going to explode*).

Improvements for French : a correspondence had to be made between the ISO-TimeML schema and the grammatical tense system of French, in particular, to account for tenses such as the *passé composé* (PAST tense value, as opposed to the present perfect used in English) and *imparfait* (IMPERFECT, not present in English as a morphological tense). French modal verbs behave differently to English modal auxiliaries as they can be conjugated in all tenses, fall within the scope of aspectual, negative polarity and other modal operators. Unlike in TB1.2, modal verbs (and adjectives), are marked <EVENT> in FTiB and have the class MODAL. 72 events (3.4%) are annotated with this class in the FTiB.

4 Methodology

Text sampling : the source texts for the FTiB were selected from the *Est Républicain* corpus of journalistic texts.[3] The journalistic genre was chosen for its relatively high frequency of events and temporal expressions. Texts were sampled from 7 different sub-genres[4], the distributions of which are shown in Table 1. Certain sub-genres appear in higher proportions than others, for two main reasons. Firstly, to favor comparison with TB1.2 (which is made up of news articles). Secondly, because the news genres are relatively diverse in style compared to the other sub-genres, which follow a certain format (e.g. obituaries). We present some of the correlations between sub-genre and linguistic content in Section 5.

Sub-genre	**Doc #**	**Doc %**	**Token #**	**Token %**
Annmt.	22	20.2%	1 679	10.4%
Bio.	1	0.9%	186	1.1%
Intl. news	32	29.4%	5 171	31.9%
Loc. news	19	17.5%	4 370	27.0%
Natl. news	25	22.9%	3 347	20.7%
Obituary	2	1.8%	313	1.9%
Sport	8	7.3%	1 142	7.0%
Total	109	100%	16 208	100%

Table 1: Proportions of sub-genres in the FTiB.

Automatic pre-annotation : To speed up the annotation process, we carried out an automatic pre-annotation of markables (events, temporal expressions and some relation markers), followed by manual correction. Relations were annotated entirely by hand, as this task remains very difficult to automate. Below we describe the two modules developed for pre-annotation.

The **TempEx Tagger** marks temporal expressions <TIMEX3> and sets the tag's attributes, and annotates certain <SIGNAL> tags. This module consists of a set of Unitex (Paumier, 2008) transducers that are applied to raw text. We adapted and

[2] After the terminology of (Vendler, 1967)

[3] Available at http://www.cnrtl.fr.

[4] These are *announcement, biography, international news, local news, national news, obituary* and *sport*.

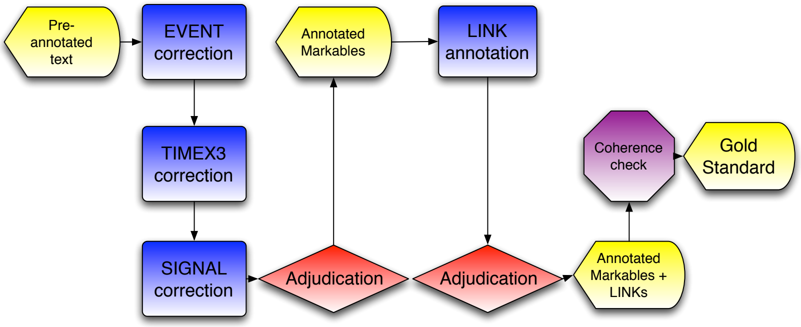

Figure 1: Schema of the annotation strategy.

enriched a pre-existing set of transducers for annotating temporal expressions in French (Gross, 2002) for our purposes. Marked expressions are classified according to their ISO-TimeML type[5] and the values of certain attributes are calculated. The `value` attribute is only set during normalization, carried out after the detection phase. A script calculates normalized values for marked expressions, including indexicals, such as *lundi dernier* (*last Monday*) or *l'année prochaine* (*next year*) (with the article's publication date as reference point). A comparative evaluation with the DEDO system of (Parent et al, 2008) shows very similar performance (for exact match on tag span and for the `value` attribute) over the same evaluation corpus (Table 2).

	System	Prec.	Rec.	F-sc.
Match	TempEx	**84.2**	**81.8**	**83.0**
	DEDO	83.0	79.0	81.0
Value	TempEx	55.0	44.9	49.4
	DEDO	**56.0**	**45.0**	**50.0**

Table 2: Comparative evaluation of the TempEx Tagger for exact match on tag span and `value` calculation.

The **Event Tagger** marks up events (`<EVENT>` tag) and certain relation markers through the application of a sequence of rules acting on the local chunk context. The rules eliminate unlikely candidates or tag appropriate ones, based on detailed lexical resources and various contextual criteria. Input is a text preprocessed with POS tags, morphological analysis and chunking (carried out with the Macaon processing pipeline (Nasr et al, 2010)). A reliable comparison with the DEDO system, to our knowledge the only other system for this task in French, was unfortunately not possible. Evaluations were made on different, yet comparable, corpora, so results are merely indicative. For event tagging, our system scored a precision of 62.5 (62.5 for DEDO), recall of 89.4 (77.7) and an F-score of 75.8 (69.3). There is room for improvement, although the system still yields significant gains in total annotation time and quality. An experiment to evaluate the effects of the pre-annotation showed a near halving of annotation time compared to manual annotation, as well as a significant reduction of human errors (Bittar, 2010).

Unfortunately, it was not possible to reliably compare the performance of the **Event Tagger** with the similar module by (Parent et al, 2008) (DEDO), to our knowledge the only other system developed for this task for French. Evaluations of each system were carried out on different, although similar, corpora. Thus, results remain merely indicative. For the task of event recognition, our system scored a precision of 62.5 (62.5 for DEDO), recall of 89.4 (77.7) and an F-score of 75.8 (69.3).

Manual annotation and validation : after pre-annotation of markables, texts were corrected by 3 human annotators (2 per text), using the Callisto[6] and Tango[7] tools, designed for this task. Figure 1 shows the process undergone by each document.

The final step of the process is a coherence check of the temporal graph in each document, carried out

[5] `DATE` (e.g. *15/01/2001, le 15 janvier 1010, jeudi, demain*), `TIME` (ex. *15h30, midi*), `DURATION` (ex. *trois jours, un an*) ou `SET` (ex. *tous les jours, chaque mardi*)

[6] `http://callisto.mitre.org/`

[7] `http://timeml.org/site/tango/tool.html`

via application of Allen's algorithm (Allen, 1983) and graph saturation (Tannier & Muller, 2008). Using the same method, we found 18 incoherent graphs among the 183 files of the TB1.2 corpus for English. At this stage, the corpus contained 8 incoherencies, which were all eliminated by hand. Manually eliminating incoherencies is an arduous task, and performing an online coherence check during annotation of relations would be extremely useful in a manual annotation tool. All files were validated against a DTD, provided with the corpus.

5 French TimeBank

Our aim for the FTiB is to provide a corpus of comparable size to TB1.2 (approx. 61 000 tokens). Version 1.0 of FTiB, presented here and made available online[8] in January 2011, represents about $\frac{1}{4}$ of the target tokens. Figure 2 shows that proportions of annotated elements for French are mostly very similar to those in TB1.2. This suggests the annotation guidelines were applied in a similar way in both corpora and that, for the journalistic genre, the distributions of the various marked elements are similar in French and English. By far the most common relation type in the French corpus is the <TLINK>. Among these, 1 175 are marked between two event arguments (EVENT-EVENT), 722 between an event and a temporal expression (EVENT-TIMEX3), and 486 between two temporal expressions (TIMEX3-TIMEX3).

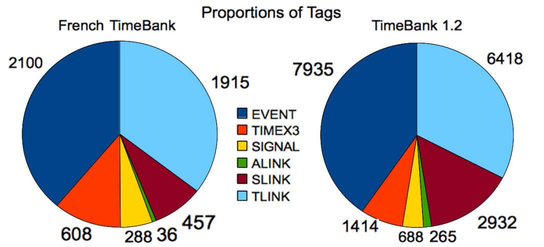

Figure 2: Annotated content of the FTiB and TB1.2.

Inter-annotator agreement was measured over the entire FTiB corpus and compared with reported agreement for TB1.2.[9]. F-scores for agreement are significantly higher for the French corpus on <EVENT> and <TIMEX3> tag spans than for TB1.2, and very slightly lower for <SIGNAL>. Figures for tag attributes are higher for TB1.2, as a much looser metric[10] was used for agreement, so comparison is not yet possible. The same measure will need to be implemented to afford an accurate comparison.

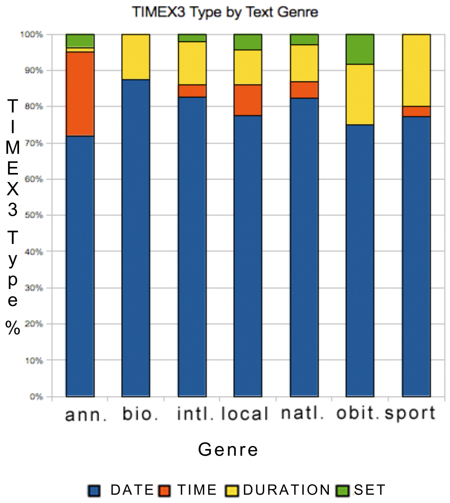

Figure 3: Distribution of <TIMEX3> types by sub-genre.

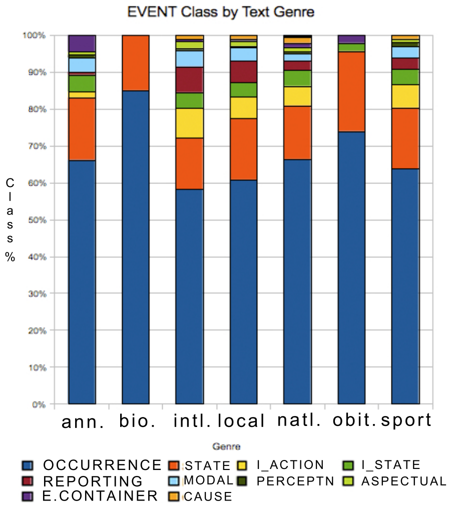

Figure 4: Distribution of <EVENT> classes by sub-genre.

[8] Via the INRIA GForge at https://gforge.inria.fr/projects/fr-timebank/.

[9] Available at http://www.timeml.org/site/timebank/documentation-1.2.html Note that figures were only calculated for a small subset of the entire corpus, unlike for the FTiB, for which all data was used.

[10] Agreement for TB1.2 was only calculated over tags with matching spans and wrong attributes on non-matching spans were not penalized. For the FTiB, all tags were considered and all attributes for non-matching tag spans were penalized.

Corpus	<TIMEX3>		<EVENT>		<SIGNAL>
	Span	Attr	Span	Attr	Span
FTiB	**.89**	.86	**.86**	.85	.75
TB 1.2	.83	(.95)	.78	(.95)	**.77**

Table 3: Inter-annotator agreement (F-scores).

Sub-genre and linguistic content : a preliminary study showed correlations between the various sub-genres chosen for the corpus and the annotations in the texts. For example, Figure 3 shows a high proportion of TIMEs in announcement texts (46% of the corpus total)[11], while DURATIONs are infrequent (2%), but appear in higher proportions in news (21–32%) and sports (13,5%). DATEs are by far the most frequently marked (80%), with SETs being the least. In Figure 4, the preponderance of the OCCURRENCE class is obvious (62.1% of all events). REPORTING is most frequent in local and international news. Announcements stand out yet again, with the highest number and highest proportion of the class EVENT_CONTAINER. These initial observations argue in favor of text sampling to achieve a diversity of temporal information in a corpus and suggest such features may prove useful in text classification.

6 Conclusion

Our experiences show ISO-TimeML is a stable language and, with some modification, is applicable to French. The FTiB is a valuable resource that will surely stimulate development and evaluation of French temporal processing systems, providing essential data for training machine learning systems. An initial survey of the data suggests temporal information may be useful for text classification. Our methodology is time-efficient and ensures data quality and usability (coherence). It could be adopted to create temporally annotated corpora for other languages as well as being adapted and generalized to other annotation tasks.

References

ISO 2008. *ISO DIS 24617-1: 2008 Language Resource Management - Semantic Annotation Framework - Part 1: Time and Events.* International Organization for Standardization, Geneva, Switzerland.

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

André Bittar 2009. *Annotation of Temporal Information in French Texts..* Computational Linguistics in the Netherlands (CLIN 19).

Sébastien Paumier 2008. *Unitex 2.0 User Manual..* Université Paris Est Marne-la-Vallée, Marne-la-Vallée, France.

Gabriel Parent, Michel Gagnon and Philippe Muller 2008. *Annotation d'expressions temporelles et d'événements en français.* Actes de TALN 2008. Avignon, France.

Alexis Nasr, Frédéric Béchet and Jean-François Rey 2010. *MACAON : Une chaîne linguistique pour le traitmeent de graphes de mots.* Actes de TALN 2010. Montreal, Canada.

James F. Allen. 1983. *Maintaining Knowledge About Temporal Intervals.* Communications of the ACM. 26:11 832-843.

Xavier Tannier and Philippe Muller 2008. *Evaluation Metrics for Automatic Temporal Annotation of Texts.* Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) Marrakech, Morocco.

Frank Robert Palmer 1986. *Mood and Modality* Cambridge University Press Cambridge, UK.

James Pustejovsky, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen and Andrea Setzer 2006. *TimeBank 1.2* Linguistic Data Consortium

Nabil Hathout, Fiammetta Namer and Georgette Dal 2002. *An Experimental Constructional Database: The MorTAL Project* Many Morphologies 178–209 Paul Boucher ed. Somerville, Mass., USA

Zeno Vendler 1967 *Linguistics and Philosophy* Cornell University Press Ithaca, NY, USA

Maurice Gross 2002 *Les déterminants numéraux, un exemple : les dates horaires* Langages 145 Larousse Paris, France

[11] This is particularly significant given the low proportion of the total corpus tokens in this sub-genre.