Representing and Visualizing Calendar Expressions in Texts

Delphine Battistelli
Univ. Paris-Sorbonne (France)
email: Delphine.Battistelli@paris-sorbonne.fr

Javier Couto
INCO, FING, UdelaR (Uruguay)
email: jcouto@fing.edu.uy

Jean-Luc Minel
MoDyCo, CNRS-Univ. ParisX (France)
email: Jean-Luc.Minel@u-paris10.fr

Sylviane R. Schwer
LIPN, CNRS-Univ. ParisXIII (France)
email: Sylviane.Schwer@lipn.univ-paris13.fr

Abstract

Temporal expressions that refer to a part of a calendar area in terms of common calendar divisions are studied. Our claim is that such a “calendar expression" (CE) can be described by a succession of operators operating on a calendar base (CB). These operators are categorized: a pointing operator that transform a CB into a CE; a focalizing/shifting operator that reduces or shifts the CE into another CE, and finally a zoning operator that provides the wanted CE from this last CE. Relying on these operators, a set of annotations is presented which are used to automatically annotate biographic texts. A software application, plugged in the platform Navitext, is described that builds a calendar view of a biographic text.

1 Introduction

Taking into account temporality expressed in texts appears as fundamental, not only in a perspective of global processing of documents, but also in the analysis of the structure of a document.[1] The analysis of temporality within texts has been studied principally by considering verbal times (e.g. Song and Cohen (1991); Hitzeman et al. (1995) and temporal adverbials (see below).

Our approach is focused on temporal adverbials — in French — that refer directly to text units concerning common calendar divisions, that we name "calendar expressions" (CEs for short). Several analyses of this kind of expressions has generated a lot of interest, ranging from their automatic recognition and annotation in texts to their analysis in terms of discursive frames (Charolles, 1997; Tannen, 1997), following work of Halliday (1994) which put the emphasis on the importance of the temporal adverbial expressions as modes of discursive organization.

Nowadays, in the field of temporality processing, automatic identification and annotation tasks of CEs are the most developed, mainly because identifying and annotating expressions which contain calendar units are considered — *a priori* — as trivial tasks. Those tasks have been particularly explored in three contexts:

1. Systems which aim to set events on a time scale depending on their duration and according to a hierarchy of unities called granularities (Schilder and Habel, 2001);
2. Systems for summarizing multi-documents (Barzilay et al., 2001); and
3. QA systems (Pustejovsky et al., 1993; Harabagiu and Bejan, 2005).

Please note that the proposition of the well-known standard temporal meta-language named TimeML (Pustejovsky et al., 2003) initially took place in the context of a QA systems worshop (Pustejovsky, 2002), and mainly integrates two schemes of annotations — namely TIDES TIMEX2 (Ferro et al., 2004) and Sheffield STAG (Setzer and Gaizauskas, 2000) — which were essentially put forward from the analysis of CEs.

In this paper, we propose a formal description of CEs in written French texts, by explicitly distinguishing several classes of linguistic markers which must be interpreted as successive operators. This work is driven in order to propose a set of fine and well-defined annotations which will be used to navigate temporally in an annotated document. Our approach differs from the preceding ones in two crucial ways:

- Our goal is not to link a CE to an event, neither to fix it on a "temporal line", using a set of values relying on ISO 8601 standard format (Mani and Wilson, 2000; Setzer and Gaizauskas, 2000; Filatova and Hovy, 2001); instead our goal is to link CEs between themselves, that is to say to establish their qualitative relative positions (the set of those relations is named "proper text calendar");
- We design CE semantics as algebraic expressions.

[1] This research is funded with an ANR grant (Projet Blanc Conique).

The remainder of this paper is organized as follows. In the next section, we introduce an algebra of CEs. In Section 3 we describe a software application, which exploits functional representation, built with previous way exhibited operators and plugged in the NaviTexte platform, aiming to support text reading. Finally, conclusions and future research directions are presented in Section 4.

2 An Algebra for Calendar Expressions

We postulate that a CE, say E, used to refer to a calendar area can be described by a succession of operators applied on an argument, named *calendar base* (CB), say B, that bears a granulariry and a value for anchoring allowing fixing it in the calendar system used and that gives access at the calendar area described by the CE.

Each operator gives a piece of the processing following a specific order: on B is applied a *pointing* operation, usually expressed by a determinant, whose result is an CE, E_1 part of E. On E_1 is applied a second kind of operator expressing the useful part of this base (all, the beginning, the middle, the end, a fuzzy area around) given as result a new CE E_2 which is part of E and is associated with a piece of the calendar that cuts the time line in three areas (illustrated by Figure 1):

- the former half-line (A),
- the Useful portion (U),
- posterior half-line (P)[2].

The useful part can also be obtained either by shifting, like in “trois semaines plus tard" (three weeks later), or by zooming, as in “l’automne de cette année là" (the autumn of this present year).[3] A third kind of operator gives access at the area described by the complete CE E: selecting one of the three portioned areas, like in “jusqu’en avril 2006" (until April 2006).

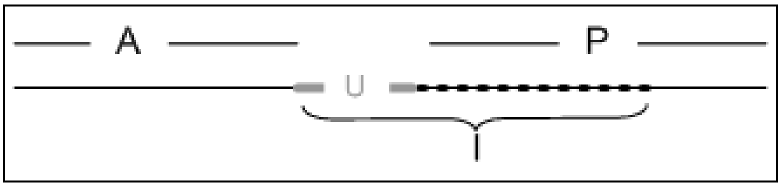


Figure 1: Partition of the time line for a unary CE

The order of operators is the following: a pointing operator *OpPointing*, followed by one or more focalising or shifting operators $OpFocalising/Shifting^{+}$ and finally at least one zoning operator $OpZoning^{\oplus}$.[4] Some operators can be omitted, usually when

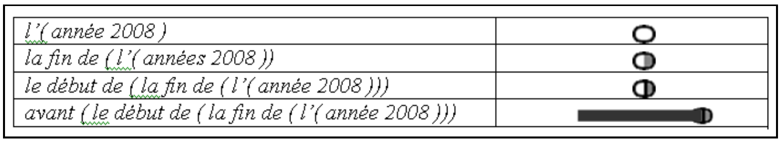
[2] This Time line is pragmatically limited bounded. For instance, (P) can be naturally limited by the present moment, as we do in Figure 1.

[3] For such deictic CEs, the CB has the granularity *year*, and the value *current*.

[4] Usually one, but we also can find two zoning operators, for instance in “jusqu’à avant Noël (until before Christmas"). In this case, the order of the operators is more constraint than the order of Focalising/Shifting operators. Therefore we use the $\oplus$ symbol instead of $+$

they do not provide any new information. In sum, the representation of CEs has the following generic form: $OpZoning^{\oplus}(OpFocalising/Shifting^{+}(OpPoin\text{-}ting(CB)))$.

For instance, let us analyse the CE E="Avant le début de la fin de l'année 2008" (before the beginning of the end of the year 2008). B="année 2008". Firstly, the operator of pointing, triggered by the identification of "l'" (the contraction of "le") is applied, given E_1=L'année 2008".[5] Secondly, two operators of focalising/shifting are applied successively: the first one triggered by "la fin de" , provides E_2' and the second one, triggered by "le début de", provides E_2. Finally an operator of zoning is associated with "avant", provided E. Consequently, the CE "avant le début de la fin de l'année 2008" is produced as *avant (le début de (la fin de(l'(année 2008))))*. The sequence of this CE is depicted and visualized in Figure 2.



l'(année 2008)	○
la fin de (l'(années 2008))	◑
le début de (la fin de (l'(année 2008)))	◐
avant (le début de (la fin de (l'(année 2008)))	▬◐

Figure 2: Computation of "avant le début de la fin de l'année 2008"

Each operator is characterized by its arity (the number of its arguments) and type. With regard to arity, in this paper we focus on unary operators.

2.1 Unary operators

Three types of operators have been defined: pointing, focalising/shifting and zoning. The pointing operator is trivial (it transforms B into a CE of type $E1$) but the two others need some refinements.

Focalising/Shifting operators

Focalising/Shifting operators transform a CE of type $E1$ into a CE of type $E2$. Several kinds of focalising/shifting time may be expressed. For instance, in the expression "au début de mai 2005" (at the beginning of may 2005) the focalising/shifting is localised inside the BC (mai 2005), whereas in the expression "trois semaines avant mai 2005" (three weeks before may 2005) it is outside the BC. Consequently, six sub-operators have been identified and are shown Table 1. It should be noted that ShiftingBeginning and ShiftingAfter operators refers to a family of operators, because for these ones it is necessary to precise two parameters, the granularity and the value of the shifting.

For some reasons of implementation, except for the operator IdShifting, which refers at the identity, all others operators are treated as idempotent. In other words, we consider as equivalent these two expressions "au début du début des années 1980" (at the beginning of the early eighties) and "au début des années 1980" (in the early of eighties). The next version will improve at this point.

Zoning operators

A Zoning operator transforms a CE of type $E2$, associated to the useful portion U of Figure 1, into the CE E analysed. A Zoning operator refers to one of the six possible

[5] This pointing operator, as mentioned previously, is not an operator of the CE algebra, but all the other operators are part of the CE algebra.

Table 1: Focalising/Shifting operators

Operators	Examples
IdShifting	– *en* 1945 – *a*u mois d'août
ZoomBeginning	– *à l'aube d*es années 1980 – *au début d*e mai 1945
ZoomMiddle	– *au milieu d*es années 1980
ZoomEnding	– *à la fin d*es années 1980
ShiftingBefore (granularity, -n)	– *10 jours avant* le 14 juillet 2005
ShiftingAfter (granularity, +n)	– *10 jours après* le 14 juillet 2005

zones[6] built from A, P and U: that is A, A+U, U, U+P, P, A+P. These six kinds of zoning are associated with a set of prepositions, whose prototypes are shown Table 2. Fuzzy expressions like "peu avant" (short before) can double this number. Table 2 also illustrates the the ZoningAbout operator <U>. Further, note that ZoningId is not expressed, but has to be taken into account.

Table 2: Zoning operators

Operators	Expression
ZoningBefore [A]	*avant* fin avril 2008
ZoningUntil [A+U]	*jusqu'*à fin avril 2008
ZoningId [U]	[∅] fin avril 2008
ZoningAbout <U>	*vers* la fin avril 2008
ZoningSince [U+P]	*depuis* la fin avril 2008
ZoningAfter [P]	*après* fin avril 2008
ZoningApart [A+P]	*excepté* fin avril 2008

2.2 N-ary or sequence operators

As mentioned before, it is necessary to use several N-ary operators to represent some CE. For instance, a binary operator is used for representing an expression like "entre fin mai 2005 et avril 2006" (between the end of may 2005 and april 2006). This operator, *Between*, applies to two CEs, so for the preceding expression the representation is Between ((ZoomEnding(Pointing(may 2005), Pointing(april 2006)). Moreover, a sequence operator is needed to represent a CE like "le mardi 21, le mercredi 22 et le vendredi 24 mai 1980" (on Tuesday 21, Wednesday 22 and Friday 24 of May). The study of these operators, associated with even more complex CEs with quantifications, is currently under investigation.

[6] The empty zone, expressed by "jamais" (never) and the full zone, that is A+U+P, expressed by "toujours" (always) are CE, but not associated with unary operators associated to a BC, as defined here, hence excluded of our precedent study.

3 Application

Many applications which exploit temporal expressions in texts, in particular in the area of information extraction, have been implemented (Pazienza, 1999). Our application is plugged into the textual navigation workstation NaviTexte (Couto, 2006; Couto and Minel, 2007), in order to combine a traditional linear reading with a chronological one. With this intention, we have undertaken the construction of a computerized aided reading of biographies. Consequently, we have addressed two issues. First, identifying temporal expressions and ordering chronologically text segments in which they are included. Second, building calendar views of the text and navigating through these views.

3.1 Identifying and ordering calendar expressions

From the linguistic study presented above, we have defined a set of annotations which are used to automatically annotate biographic texts. This process is carried out by transducers which put XML[7] annotations through the processed text. These annotations describe on the one hand, the granularity of CEs, and on the other hand, the kind of identified operator. For instance, the following XML code illustrates how the temporal expression “avant le début de la fin de l'année 2008" (Before the beginning of the end of the year 2008) will be annotated:

```
<UT Type="Expression Calendaire" Nro="7">
  <Annotation Nom="Grain">Annee</Annotation>
  <Annotation Nom="Annee">2008</Annotation>
  <Annotation Nom="RelationCalendrier">Absolue</Annotation>
  <Annotation Nom="OpTempRÂŐgion1">Avant</Annotation>
  <Annotation Nom="OpTempDÂŐplacement1">FocalFin</Annotation>
  <Annotation Nom="OpTempDÂŐplacement2">FocalDebut</Annotation>
  <Chaine>
  avant le debut de la fin de l'annee 2008
  </Chaine>
</UT>
```

From these annotations, an automatic ordering relying on values of CEs can be carried out. A first implementation took only into account disjoined CEs, because they are linearly ordered. Intersecting CEs, like “En juin 2007 (...) en été 2007" (in June 2007 (...) in summer 2007) requires a more powerful formalism. A formalism relying both on S-Languages (Schwer, 2002b) and granules (Schwer, 2002a) is required to provide a full automatic ordering.

3.2 Building a text calendar view

A new kind of view, a calendar one, has been built in the NaviTexte platform. This view is built from texts which contain CEs annotated as described above. An example is shown in Figure 3. Conceptually, a calendar view is a graph coordinated with a two-dimensional grid. In the left part of the view, lexical chains of various occurrences of CEs in the text are displayed. By default, those are ordered according to their order of appearance in the text, but it is possible to display a chronological order,

[7] A DTD is defined in Couto (2006)

by using options offered in the panel located in bottom of the view. Nodes in the graph represent these lexical chains. The visual representation of a CE depends of the functional representation computed as described before Figure 2.

A simple CE, with only a pointing operator like in “l’année 2008" (the year 2008) is always visualised like a white ellipse. An operator of focalising/Shifting like “la fin de" (the end of) selects an area of the ellipse and blackens it. Finally, a zoning operator like “avant" (before) is visualised by a bold line displaying the area that is referred to.

The plug-in is implemented with the JGaph package and we largely use some of its functionalities, like zooming or the partial layout cache. We also use html tooltip text in Swing to contextualise a CE in the original text. For example, in Figure 3, the whole paragraph which contains the CE “en 1953” (in 1953) is displayed and the occurrence of a CE is highlighted.

3.3 Evaluation

Two kinds of evaluation could be performed on this work: (i) evaluation of automatic recognition and semantic annotation of CEs in text, (ii) evaluation of the calendar view. The former calls for a classical protocol in NLP, whereas the latter is more complex to carry out.

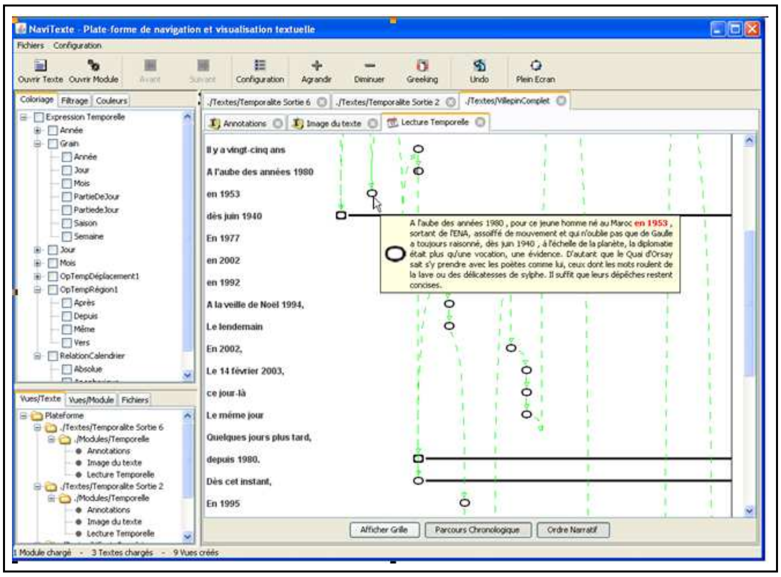
So far, only recognition has been carried out by Teissedre (2007) who computed recall and precision on three kinds of corpora. Due to the fact that an annotation is made up of several fields the recall has been computed like this: a score zero when a CE is not identified, a score 1 when the identification is total, and 0.5 when the identification is partial. Applying these rules, recall is 0.8 and precision is 0.9.

We would like to make two remarks on this result. First, quantified CEs like “tous les mardis” (every Tuesday) or “un mardi sur deux” (one Tuesday out of two) and n-aries ($n \geq 3$) CEs like "entre 2008 et 2009 et en juin 2010" (between 2008 and 2009 and in june 2010) are identified but are not yet taken into account in the semantic annotation process. Second, syntactic ambiguities like in “il a dormi deux jours avant Noël” (he slept two days before Christmas) are not taken into account either. However in this example, there are two possible syntactic structures. In the first case, "avant Noël" is the CE and the operator is the Regionalisation one; in the second case, "deux jours avant Noël" is the CE and the operator is the Shifting one. Presently, our analysis provides only the second one like in Aunargue et al. (2001) but we intend to upgrade it in order to provide both analyses.

Evaluation of the calendar view should be studied from a cognitive point of view and is highly dependent on the application. We plan to work with cognitive scientists to build a relevant protocol to study this aspect of evaluation which calls for the specification of a set of navigation operations based on the algebra of operators.

4 Conclusion

We proposed an algebra of CEs with three kinds of operators to analyse calendar expressions and build a functional representation of these expressions. We described an implementation of this approach in the platform NaviTexte and we have shown how the functional representation is used to visualise a calendar view of a text. In future work, we will rely on a methodology presented in Battistelli and Chagnoux (2007) in



Il y a vingt-cinq ans
A l'aube des années 1980
en 1953
dès juin 1940
En 1977
en 2002
en 1992
A la veille de Noël 1994,
Le lendemain
En 2002,
Le 14 février 2003,
ce jour-là
Le même jour
Quelques jours plus tard,
depuis 1980.
Dès cet instant,
En 1995

A l'aube des années 1980 , pour ce jeune homme né au Maroc **en 1953** , sortant de l'ENA, assoiffé de mouvement et qui n'oublie pas que de Gaulle a toujours raisonné, dès juin 1940 , à l'échelle de la planète, la diplomatie était plus qu'une vocation, une évidence. D'autant que le Quai d'Orsay sait s'y prendre avec les poètes comme lui, ceux dont les mots roulent de la lave ou des délicatesses de sylphe. Il suffit que leurs dépêches restent concises.

Figure 3: Example of calendar view in NaviTexte

order to take into account several temporal axis, and thus several calendar structures, which are expressed in texts by different levels of enunciations, like citations.

References

Aunargue, M., M. Bras, L. Vieu, and N. Asher (2001). The syntax and semantics of locating adverbials. *Cahiers de Grammaire 26*, 11–35.

Barzilay, R., N. Elhadad, and K. McKeown (2001). Sentence ordering in multidocument summarization. In *First International Conference on Human Language Technology Research (HLT-01)*, pp. 149–156.

Battistelli, D. and M. Chagnoux (2007). Représenter la dynamique énonciative et modale de textes. In *actes TALNÕ07 (Traitement automatique du langage naturel*, pp. 13–23.

Charolles, M. (1997). LÕencadrement du discours – univers, champs, domaines et espaces. In *Cahiers de recherche linguistique*, Volume 6 of *LANDISCO*, pp. 1–73. Université Nancy 2.

Couto, J. (2006). *Modélisation des connaissances pour une navigation textuelle assistée. La plate-forme logicielle NaviTexte.* Ph. D. thesis, Université Paris-Sorbonne.

Couto, J. and J.-L. Minel (2007). Navitexte, a text navigation tool. In *, Lecture Notes in Artificial Intelligence 4733*, pp. 251–259. Springer-Verlag.

Ferro, L., L. Gerber, I. Mani, B. Sundheim, and G. Wilson (2004). Standard for the annotation of temporal expressions. Technical report, timex2.mitre.org, MITRE Corporation.

Filatova, E. and E. Hovy (2001). Assigning time-stamps to event-clauses. In *Workshop on Temporal and Spatial Information Processing, ACLÕ2001*, pp. 88–95.

Halliday, M. A. K. (1994). *An introduction to functional grammar*. London: Edward Arnold.

Harabagiu, S. and C. A. Bejan (2005). Question answering based on temporal inference. In *AAAI-2005 Workshop on Inference for Textual Question Answering*.

Hitzeman, J., M. Moens, and C. Grover (1995). Algorithms for analyzing the temporal structure of discourse. In *EACLÕ95*, pp. 253–260.

Mani, I. and G. Wilson (2000). Robust temporal processing of news. In *Proceedings 38th ACL*, pp. 69–76.

Pazienza, M. T. (1999). *Information Extraction, toward scalable, adaptable systems*. New York: Springer-Verlag.

Pustejovsky, J. (Ed.) (2002). *TERQAS 2002: An ARDA Workshop on Advanced Question Answering Technology*.

Pustejovsky, J., J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, and G. Katz (2003). Timeml: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*.

Pustejovsky, J., R. Knippen, J. Lintman, and R. Sauri (1993). Temporal and event information in natural language text. *Lexique 11*, 123–164.

Schilder, F. and C. Habel (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of ACL'01 workshop on temporal and spatial information processing*, pp. 65–72.

Schwer, S. R. (2002a). Reasoning with intervals on granules. *Journal of Universal Computer Science 8 (8)*, 793–808.

Schwer, S. R. (2002b). S-arrangements avec répétitions. *Comptes Rendus de l'Académie des Sciences de Paris Série I 334*, 261–266.

Setzer, A. and R. Gaizauskas (2000). Annotating events and temporal information in newswire texts. In *Proceeedings 2rd LRC*, pp. 64–66.

Song, F. and R. Cohen (1991). Tense interpretation in the context of narrative. In *9th AAAI*, pp. 131–136.

Tannen, D. (1997). *Framing in Discourse*. Oxford: Oxford University Press.

Teissedre, C. (2007). La temporalité dans les textes : de lÕannotation sémantique à la navigation textuelle. Master's thesis, Université Paris-Sorbonne.