A System for Extraction of Temporal Expressions from French Texts

Nikolai Vazov*
Groupe sur l'asymétrie des langues naturelles, UQAM
c1144@er.uqam.ca

Résumé

Cet article présente un système pour l'identification automatique des expressions temporelles dans des textes français. La procédure d'identification repose sur une stratégie d'exploration contextuelle qui met en oeuvre deux techniques complémentaires: recherche des patrons (expressions régulières) et *chart parsing* qui est déclenché en fonction des patrons repérés.

Mots-clés: chart parsing, exploration contextuelle

We present a system for extraction of temporal expressions from French texts. The identification of the temporal expressions is based on a context-scanning strategy (CSS) which is carried out by two complementary techniques: search for regular expressios and left-to-right and right-to-left local chart-parsing. A number of semantic and distant-dependency constraints have been integrated to the chart-parsing procedure in order to improve the precision of the system.

Keywords: chart parsing, context-scanning

1 Introduction

The identification and the interpretation of temporal and aspectual information plays an important role in text understanding. This information is encoded in the natural languages by a wide array of linguistic means ranging from grammatical (morpho-syntactic) to lexical (verbs and adverbials) or strictly syntactic phenomena (temporal anaphora (Webber1988) or argument structure of the verb (Verkuyl1993)).

In this paper we present a system for identification of lexical non-verbal means of expressing temporal information in French texts. Such means are the temporal adverbs (e.g. *yesterday*) and adverbials (e.g. *three days before the marriage of his youngest son*) studied in detail in (Bras and Molinès1993; Gagnon and Lapalme1996), as well as the temporal subordinate clauses (e.g. *When Peter arrived, ...*) . The identification of these temporal expressions is based on a context-scanning strategy (CSS) (Charolles1997; Desclés1997) which is carried out by two

*This research has been supported by the Social Sciences and Humanities Research Council of Canada (grant # 412-97-0016), for the "Assymmetry and Natural Language Processing Project", awarded to professor Anne-Marie Di Sciullo, in the Département de Linguistique at Université du Québec à Montréal (UQAM).

complementary techniques: search for regular expressions and left-to-right and right-to-left chart-parsing.

The first technique looks for a particular set of markers (regular expressions) encoding temporal information. These markers can be stand-alone or parsing-triggering. The stand-alone markers are inactive chart elements (not necessarily a single word) which represent autonomous temporal expressions. The adverb *then* in *Then, this was the only solution* is an example of a stand-alone marker.

The second technique of the CSS is launched, if the system identifies a parsing-triggering marker. The parsing-triggering markers are active chart elements which signal the presence of a larger temporal expression (for example *before* in *three weeks before Christmas*) and trigger a chart-parsing procedure for its identification. The chart-parsing scans the left and the right context of the marker in order to determine which lexical units from this context belong to the temporal expression. The identified entire temporal expression represents a temporal marker (inactive chart element) generated incrementally by association of chart elements situated to the left and to the right of the parsing-triggerring marker (the initial active chart element) (Wonsever and Minel2001).

The present paper is organised in 4 sections. Section 2 is a short presentation of the CSS as a method for goal-oriented information extraction. Sections 3 and 4 discuss the central notions of the system, namely the markers and the chart-parsing rules. The architecture of the system is shown in section 5.

2 The underlying principles of the context-scanning strategy

The context-scanning strategy (CSS) (Desclés et al.1997; Berri et al.1996; Minel et al.2000; Minel and Desclés2000) is based on the hypothesis that the semantic representations are not necessarily determined by world knowledge but can be considered as configurations integrating elements of local linguistic information. The interaction of these elements is controlled by a set of mechanisms which build the semantic representation by putting together all the pieces of local information. The elements of this representation can be derived from grammar or lexical semantics and from some syntactic phenomena which together account for the non-atomic (holistic) compositionality of semantic representations (Gosselin1996).

This hypothesis allows to formulate the two distinctive features of the CSS:

- The CSS is designed to carry out a very specific task: to identify and interpret all the surface linguistic elements which account for a particular semantic representation. It does not claim to provide a complete analysis of texts.
- The CSS extracts sets of interactive elements integrated in complex semantic representations. Unlike the approaches based on world knowledge, it *deduces* these representations using exclusively linguistic data found in the text.

The system presented in this paper satisfies both features. The specific task of the system (first feature) is to identify all non-verbal linguistic units which can convey temporal information and to group them in sets. These sets are called temporal expressions and may contain one or more members.

All temporal expressions are considered as semantic functions whose argument is the semantic value of the grammatical tense in the analysed utterance. The result of the application of this function to its argument is the semantic (aspecto-temporal) value of the entire utterance. Thus, the identification of the temporal expressions is an important step in the deduction of a semantic representation which integrates information from different linguistic units (second feature). Utterances in 1 and 2 show the role of the temporal expressions in this deduction. Example 1 denotes an event which took place before the utterance time. In turn, the lack of temporal expression in 2 entails a double interpretation: the utterance can refer either to the fact of Jeanne's departure in a past moment, or to the fact that Jeanne is absent at the moment of utterance. This ambiguity is manifested in the English translation of 2.

(1) *Jeanne est partie trois minutes avant Pierre.*
Jeanne left three minutes before Pierre.

(2) *Jeanne est partie.*
Jeanne left/has left.

The discussion in this paper will focus only on the first feature of the implemented CSS, namely, the identification and regrouping of linguistic units into temporal expressions.

As it was mentioned in the introduction, the identification of the temporal expressions is carried out at two steps: identification of markers (stand-alone and parsing-triggering) and chart-parsing applied to the left/right context of the parsing triggering markers. These steps will be discussed in detail in the following two sections.

3 Selection of the markers (active & inactive chart elements)

All temporal markers (regular expressions) detected by the system at the first step of the analysis fall into two sets Σ and M ($\Sigma \cap M = \emptyset$).

The set Σ contains the stand-alone markers which represent autonomous temporal expressions. These expressions can be represented by the following context-scanning rule:

$$TemporalExpression \rightarrow \sigma \mid \sigma \,\epsilon\, \Sigma \tag{1}$$

The stand-alone markers are grouped in two subsets ($\Sigma_1 \cap \Sigma_2 = \emptyset$), defined with regard to the structure of their elements.

- markers (Σ_1) representing constant strings (for example, *par la suite* 'after that', *le lendemain matin* 'the next day's morning')
- markers (Σ_2) specifying the initial element of a string, considered as a temporal expression, and the set of all subsequent authorised characters (for example, *quand* 'when' followed by $n > 1$ number of characters (including spaces, diacritics and special symbols as &, %, etc.).

The markers included in the set M are both indicators and constituent elements of a larger temporal expression. They are the active elements of a chart-parsing procedure which they launch in order to determine the boundaries of this expression in the analysed sentence.

The temporal expressions signaled by the markers from this set can be represented by the following contextual rule:

$$TemporalExpression \rightarrow LC.m.RC \mid \; m \,\epsilon\, M \tag{2}$$

where

- LC is the left context of the marker *m*. It contains an array of strings (lexical units) whose syntactic categories have been authorised by the right-to-left chart-parser (see section 4)
- LC can be empty
- "." is the operator of concatenation
- RC is the right context of the marker *m*. It contains an array of strings (lexical units) whose syntactic categories have been authorised by the left-to-right chart-parser (see section 4)
- RC can be empty

All the markers in M have been organised in three subsets M_1, M_2 and M_3 ($M_1 \cap M_2 \cap M_3 = \emptyset$), according to the particularities of the chart-parsers they trigger:

- Markers triggering a left-to-right chart parser (M_1). These markers represent the leftmost lexical unit of the larger temporal expression and launch parsing rules applying exclusively to the right context of the marker. This set contains markers like *il y a* 'ago', *au cours* 'during', etc.
- Markers triggering a left-to-right *and* a right-to-left parser (M_2). The markers included in this set never occur at the leftmost or at the rightmost position in the larger temporal expression. Hence, the need to analyse both their left and right context in order to identify the boundaries of the temporal expression they belong to. This is the largest of the three sets and contains elements like the names of the months, temporal units (*minute, seconde,..*), names of the 4 seasons, etc.
- Markers requiring a left-to-right *and* a right-to-left parser (M_3). These are the markers which can occur at any position in the larger temporal expression. However, the left and right parsing rules may eventually rule out the left *and/or* the right context as not belonging to the temporal expression and return the detected marker as the only constructive element of the temporal expression. An example of such a marker is *après* 'after' in the following sentences:

  (3) *Jeanne est arrivée trois minutes après Pierre.*
  Jeanne came three minutes after Pierre.
  (4) *Jeanne est arrivée trois minutes après.*
  Jeanne came three minutes later.
  (5) *Jeanne est arrivée après Pierre.*
  Jeanne came after Pierre.
  (6) *Jeanne est arrivée après.*
  Jeanne came later.

Example 3 shows a larger temporal expression (*trois minutes après Pierre*) which extends to the left and to the right of the detected marker *après*. In example 4, the same triggering marker has only a left context (*trois minutes*), the right one being interpreted by the parser as empty (punctuation mark stands for the end of a temporal expression). Conversely, in example 5, the parsing rules will authorise the association of elements (*Pierre*) from the right context of the marker but will block the association of such elements (*arrivée*) from its left context. Finally, following the same line of reasoning as in examples 4 and 5, the parsing of example 6 will return the marker itself.

The above examples explain why it is important to distinguish between the three sets of triggering markers. The M_1 markers signal temporal expressions which only extend to the right of the marker's position. The M_2 markers signal temporal expressions which extend to the right *and* to the left of the marker's position. Alone, M_2 markers do not constitute a temporal expression. The markers of M_3 can be used alone as temporal expressions *or* constitute such an expression together with other elements from their left *and/or* right context.

The total number of markers detected by the system is 121. They are subdivided as follows: Σ_1 (7), Σ_2 (36), M_1 (13), M_2 (51), M_3 (14).

4 Left and right chart-parser rules

4.1 General principles

The left and right chart-parser rules triggered by the M_1, M_2 and M_3 markers are recursion-based algorithms. These rules scan the tagged[1] right and left context of the marker and control the association of the adjacent lexical units to the body of the temporal expression signaled by this marker. The control is carried out by the following context-scanning parsing rules:

Rule for the left context:

```
while cat$_{w^n}\epsilon$ C$_{left}$
    m = $w^n$.m
    n = n - 1
```

where

- *w* is a lexical unit from the LC
- *n* is the position of *w* in LC
- *cat* is the syntactic category of *w*
- *m* is temporal expression containing only the parsing-triggering marker before the application of the rule and completed at each loop by a new element from the LC
- "." is the operator of concatenation
- C$_{left}$ = {DET, PRE, ADJ, PRO, NUM, NOC} - the set of categories authorised in the positions preceding the marker *m*

Rule for the right context:

```
while cat$_{w^n}\epsilon$ C$_{right}$
    m = m.$w^n$
    n = n + 1
```

where

- *w* is a lexical unit from the RC

[1] The system uses the shareware on-line tagger of LATL.ch S.A.(Société Anonyme) (www.latl.ch). The notation for the syntactic categories in this discussion follows the notation provided by the tagger: DET (determiner), PRE (preposition), ADJ (adjective), PRO (pronoun), NUM (numeral), NOC (common noun), ADV (adverb), INF (infinitive), PPA (past participle), PPR (present participle), CLI (clitic), COJ (conjunction) and NOP (proper noun).

- *n* is the position of *w* in RC
- *cat* is the syntactic category of *w*
- *m* is the temporal expression containing only the parsing-triggering marker before the application of the rule and completed at each loop by a new element from the RC
- "." is the operator of concatenation
- C$_{right}$ = {DET, ADJ, NUM, NOC, ADV, INF, PPA, PPR, CLI, COJ, NOP} - the set of categories authorised in the positions following the marker *m*

Each rule tries to recursively match a category from a built-in set (C$_{left}$ and C$_{right}$) to the category of the next lexical unit in the context. If the two categories match, the lexical unit is associated to the temporal expression and the rule moves on to check the category of the next lexical unit. In turn, if the two categories do not match, the rule blocks the association of the unit to the temporal expression and signals the left, respectively, the right boundary of the temporal expression. The analysis of example 7 below shows the scanning procedure step by step (the marker *heures* has been detected at a previous stage):

(7) *Ils ont quitté Air France vers 10 heures pour une courte promenade sur les Champs-Elysées.*
They left Air France about 10 o'clock for a short walk on Champs-Elysées.

1. **left pass**: the NUM category of the lexical unit '10' matches the built-in category NUM in the scanning rule
2. **left pass**: the PRE category of the lexical unit 'vers' matches the built-in category PRE in the scanning rule and associates it to the already processed left context: *vers 10*
3. **left pass**: the NOP category of the lexical unit 'France' is not in the rule's list of categories - the rule cannot move further left and the system returns: *vers 10*

1. **right pass**: the PRE category of the lexical unit 'pour' is not in the category list of the rule - the rule cannot move furhter right and the system returns an empty string

The string *vers 10 heures* constitutes the larger temporal expression signaled initially by the triggering marker *heures*. Formulated in terms of the rules above it will have the form: $w^6.w^7$.m.$\emptyset$ (7 & 6 are the positions of the lexical units in the LC).

4.2 Constraints in the chart-parsing rules

4.2.1 Constraints on the relative positions of syntactic categories

The matching of categories discussed in the previous section is achieved by context-free rules - one for the left and one for the right context of the parsing-triggering marker, respectively.

However, this context-free algorithm will fail to detect the boundaries of the temporal expression in a number of cases where, following a successful category matching, it will associate to this expression some elements which do not belong to its body. Without additional restrictions to the rules, the system will provide the following wrong right-context analysis for example 8 (the marker *année* was detected at a previous stage):

(8) *L'année dernière les actions de cette entreprise ont augmenté de 30%.*
Last year the shares of this company rose by 30%.

1. **right pass**: category of *dernière* is ADJ : authorised
2. **right pass**: category of *les* is DET (determiner) : authorised
3. **right pass**: category of *actions* is NOC : authorised
4. **right pass**: category of *cette* is DET : authorised
5. **right pass**: category of *entreprise* is NOC : authorised
6. **right pass**: category of *ont* is VAU (auxiliary): blocked

Since the syntactic categories DET and NOC are in the list of the authorised categories, they become accepted by the system regardless of their position. The result of this overgeneration will be the segment *l'année dernière les actions de cette entreprise* 'last year the shares of this company'.

To avoid this problem, the above context-free rules have been modified to check the syntactic categories which are adjacent to the syntactic category under analysis and to verify, if a number of adjacency constraints have been respected. An example of such a constraint is the following rule:

If ADJ is immediately preceded by a NOC or by an empty context,

and if it is immediately followed by DET, the DET is not an element of the temporal expression.

In example 8 this constraint will block the association of new elements to the right of the ADJ *dernière* due to the empty context preceding the ADJ[2] and to the DET category (*les*) following the adjective. Similar, yet a more complex constraint on the number and the position of verbs in the relative clasuses enables the system to extract correctly the temporal expressions (in curly braces) in examples 9 and 10[3]:

(9) *Le ministre est venu {3 minutes après son porte-parole} qui avait déjà annoncé la bonne nouvelle.*
The minister came 3 minutes after his spokesman who had already announced the good news.

(10) *La réunion a commencé et {3 minutes après} son porte-parole qui avait déjà annoncé la bonne nouvelle est parti pour la capitale.*
The meeting began, and 3 minutes later his spokesman who had already announced the good news left for the capital.

4.2.2 Constraints on the semantics of the lexical units

Another constraint in the chart-parsing rules is the condition on the semantics of the lexical units in the context of the marker. This constraint affects exclusively the nouns in the left context of the markers from the set M_3 (see section 3). Actually, the left context of these markers is a modifier of the expression constitued by the marker and its right context. Thus, *three minutes before his departure* can be represented as (`three minutes`(`before his departure`)). The internal organisation of these configurations determines the semantics of the nouns occurring in the left context of the marker: they have to denote a time period, like minutes, seconds, seasons, etc. The set of these nouns is actually a subset of M_2 - the set of markers containing the names of months, temporal intervals, etc...(see section 3). In order to authorise the association of the noun to the temporal expression, the scanning rule checks if this noun is a member of this subset. If this is the case, the rule moves left to the next lexical unit. And conversely, if the noun is not a member of this set, the association is blocked and the rule signals the boundary of the temporal expression.

[2] The adjective is the leftmost element of the right context of the marker *année*.

[3] The punctuation marks, like commas, are an important source of information about the temporal expressions' boundaries. Naturally, the system's scanning rules take into account their position. The constraints discussed in section 4.2.1 are needed to process constructions which do not contain punctuation marks.

5 Architecture of the system and a real example

The overall architecture of the system is shown in Figure 1. The input text is tokenized into sentences and a search for stand-alone markers is launched.

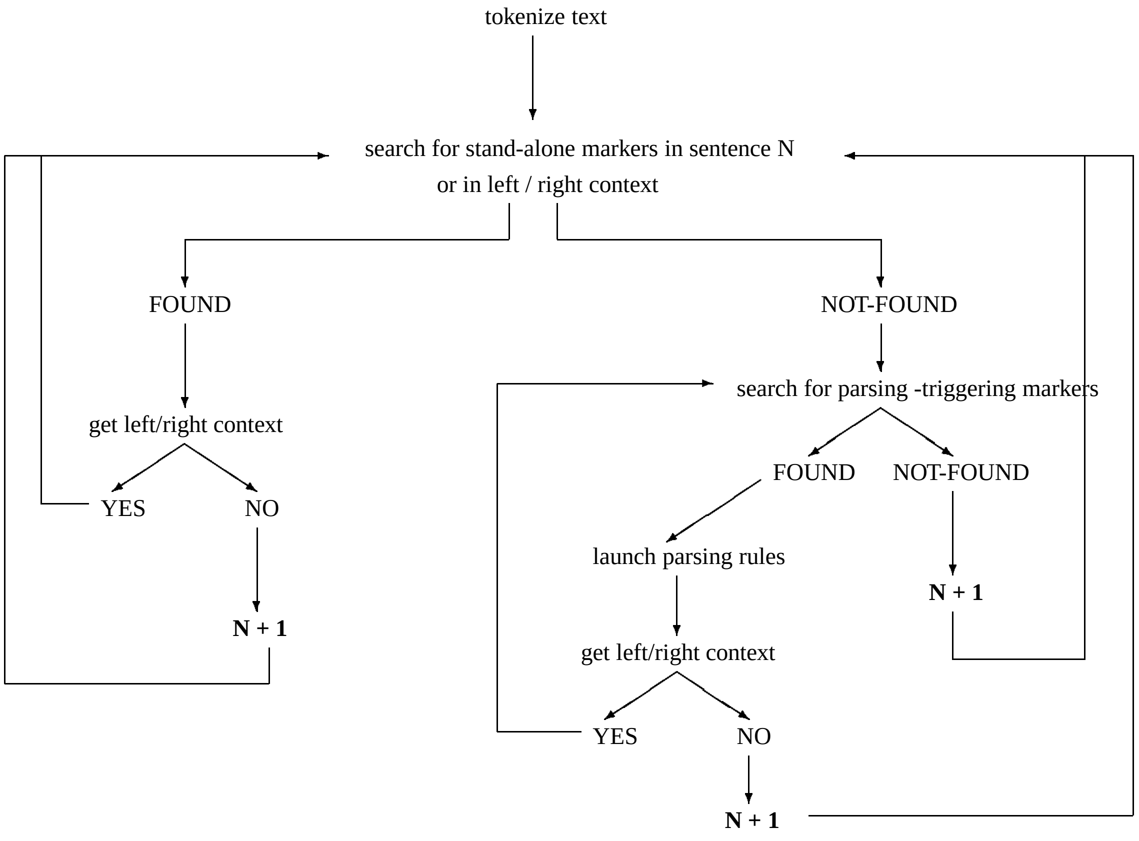

Figure 1: The flowchart of the system

If the system detects such a marker, it extracts it from the sentence and resumes the search within its left and right context. If both contexts are empty, the index of sentences is incremented and the same procedure repeats for sentence $_{n+1}$. In turn, if the system fails to detect a stand-alone marker (during the first or any of the following loops), it scans the sentence for parsing-triggering markers.

If such a marker has been identified, the system launches the chart-parsing rules in order to identifiy the larger temporal expression signaled by the parsing-triggering marker. Once identified, the larger temporal expression is extracted from the sentence and its left and right contexts are scanned again for other markers of the same type. This loop will repeat until there is no remaining left and right context. Then the system increments the index of the sentences and resumes the search for stand-alone markers in the next sentence.

The application of the algorithm can be manifested on the following real example (the comments will follow the output of the system):

```
Trois mille séminaristes, jeunes prêtres et étudiants ont souhaité vendredi un bon anniver-
saire au pape qui fêtera samedi ses 21 ans de pontificat, en lui chantant en polonais "
stolat " ( " cent ans " ) , lors d'une messe dans la basilique Saint-Pierre
```

The first marker detected by the system is 'samedi'. The chart parser launched by this marker finds no other elements which belong to the temporal expression containing 'samedi'. Hence, the system returns the string to the left and to the right of ’samedi’, as the new left and right context, respectively:

```
WANTED is:  samedi
NEW LEFT CONTEXT: Trois mille séminaristes , jeunes prêtres et étudiants ont souhaité ven-
dredi un bon anniversaire au pape qui fêtera
NEW RIGHT CONTEXT: ses 21 ans de pontificat, en lui chantant en polonais stolat ( cent ans
) , lors d’ une messe dans la basilique Saint-Pierre
```

The same procedure repeats for the marker ’vendredi’:

```
WANTED is:  vendredi
NEW LEFT CONTEXT: Trois mille séminaristes , jeunes prêtres et étudiants ont souhaité
NEW RIGHT CONTEXT: un bon anniversaire au pape qui fêtera
```

Since neither of the new contexts contains temporal expressions, the system will resume the scanning from the right context of the first detected marker - ’samedi’:

```
ses 21 ans de pontificat, en lui chantant en polonais stolat ( cent ans ) , lors d’ une
messe dans la basilique Saint-Pierre
```

The system detects the next keyword ’lors d” and the chart parser completes the temporal expression by adding the right context of the marker. The new left and right contexts of the expression are returned to the system:

```
WANTED is:  lors d’
RIGHT CONTEXT: une messe
NEW LEFT CONTEXT: ses 21 ans de pontificat, en lui chantant en polonais stolat (cent ans),
NEW RIGHT CONTEXT: dans la basilique Saint-Pierre
```

The analysis of the remaining context of ’lors d’une messe’ does not detect new markers and the system halts returning the identified temporal expressions[4]

```
samedi / vendredi / lors d’une messe
PHRASE 1 ANALYSEE ** NOMBRE D’EXPRESSIONS REPEREES: 3
```

6 Conclusion

The approach presented in this paper extracts non-verbal temporal expressions from French texts using exclusively linguistic markers and local chart-parsing techniques. The system has been tested over 98 texts containing temporal expressions of different complexity. At present, it is being tested over a number of texts from the corpus Hansard (50 million words of debates in Canadian parliament) in order to evaluate its real precision and recall. The results obtained by the system can be further used in three different fields:

- Information extraction: the temporal expressions retrieved by the system provide substantial information about the temporal localisation of events described in the analysed texts (Ben Hazez and Minel2000).
- Computational semantics: the semantics of the temporal expressions (for example, the set of expressions referring to past events) along with the semantics of grammatical tenses

[4] The goal of the presented system is to detect only the expressions which modify the temporal parameters of the sentence. Hence, unlike some other studies on temporal tagging or annotaiton (Mani et al.2001), we deliberately ruled out expressions which do not answer to the question *when*.

is a constituent element of the aspecto-temporal value of the sentence (Battistelli2000; Vazov1999).

- Parsing: the temporal expressions are syntactic constituents whose identification by parsing techniques is a difficult task. The identification of this constituent will fix its boundaries and hence simplify the task of the parser.

References

D. Battistelli. 2000. *Passer du texte à une séquence d'images, analyse spatio-temporelle de textes, modelisation et réalisation informatique (système SPAT)*. Ph.D. thesis, Université Paris - Sorbonne.

S. Ben Hazez and J.-L. Minel. 2000. Designing tasks of identification of complex patterns used for text-filtering. In *RIAO*, pages 1558–1567.

J. Berri, E. Cartier, J.-P. Desclés, A. Jackiewicz, and J.-L. Minel. 1996. Filtrage automatique des textes. In *Natural Language Processing and Industrial Applications*, pages 22–35, Moncton, N.B., Canada.

M. Bras and F. Molinès. 1993. Adverbials of temporal location: Linguisitic description and automatic processing. In J. Darski and Z. Vetulani, editors, *Sprache-Kommunikation-Informatik, Akten des 26 Linguistischen Kolloquium*, pages 137–146. Max Niemeyer Verlag.

M. Charolles. 1997. Reconnaissance et délimitation des univers d'énonciation introduits par une expresion selon x et contraction automatique de textes. In *Journee d'etude, Filtrage des informations dans les textes*, Paris. Association pour le Traitement Automatique des Langues (ATALA).

J.-P Desclés, E. Cartier, A. Jackiewicz, and J.-L. Minel. 1997. Textual processing and contextual exploration method. In *CONTEXT'97*, pages 189–197, Brasil. Universidade Federal do Rio de Janeiro.

J.-P. Desclés, 1997. *Systèmes d'exploration contextuelle. Co-texte et calcul du sens*, pages 215–232. Presses Universitaires de Caen.

M. Gagnon and G. Lapalme. 1996. From conceptual time to linguistic time. *Computational Linguistics*, 22(1):91–127.

L. Gosselin. 1996. *Sémantique de la temporalité en français, Un modèle calculatoire et cognitif du temps et de l'aspect*. Champs linguistiques. Duculot.

I. Mani, L. Ferro, B. Sundheim, and G. Wilson. 2001. Guidelines for annotating temporal information. In *Notebook Proceedings of Human Language Technology Conference 2001*, pages 299–302, San Diego, California, March 18-21.

J-L. Minel and J-P. Desclés. 2000. *Résumé Automatique et Filtrage des textes*. Ingénierie des langues. Editions Hermès, Paris.

J.-L. Minel, J.-P. Desclés, E. Cartier, G. Crispino, S. Ben Hazez, and A. Jackiewicz. 2000. Résumé automatique par filtrage sémantique d'informaitons dans des textes. *TSI*, X(X):1–23.

N. Vazov. 1999. Context-scanning strategy in temporal reasoning. In *Modeling and Using Context (Second International and Interdisciplinary Conference, CONTEXT'99)*, pages 389–402. Springer-Verlag.

H. Verkuyl. 1993. *A theory of aspectuality. The interaction between temporal and atemporal structure*. Cambridge Studies in Lingustics. Cambridge University Press.

B. Webber. 1988. Tense as discourse anaphor. *Computational Linguistics*, 14(2):61–73.

D. Wonsever and J.-L. Minel. 2001. Contextual rules for text analysis. In *2nd International conference on Intelligent Text Processing and Computational Linguistics CICLing-2001*, Mexico City.