Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies

Delphine Battistelli* — Jean-Luc Minel— Sylviane R. Schwer*****

** LaLIC, FRE 2919 Université Paris IV*
Maison de la Recherche
28, rue Serpente, 75 006 Paris
Delphine.Battistelli@paris4.sorbonne.fr

*** MoDyCo, UMR7114, CNRS*
Université Paris X
200, avenue de la République 92 001, Nanterre France
Jean-Luc.Minel@u-paris10.fr

**** LIPN, UMR 7030, CNRS*
Université Paris XIII
99, avenue Jean-Baptiste-Clément, 93430 Villetaneuse, France.
Sylviane.Schwer@lipn.univ-paris13.fr

RÉSUMÉ. *Cet article aborde le problème de la représentation formelle des expressions calendaires et en présente une analyse fonctionnelle, fondée sur la représentation catégorique des ordinaux. Il expose ensuite un projet d'aide à la lecture d'un texte biographique long selon une navigation temporelle.*

ABSTRACT. *In this paper, first we examine the usual treatment of temporal information and we recall the main features relative to the temporality analysis in texts. Then we deal with the formal representation of calendar expressions, and we provide a functional approach, based on a categorical representation of ordinals. Finally, an ongoing application providing a help for the reading of long biography is presented.*

MOTS-CLÉS : *annotation, temps, discours, calendriers, navigation textuelle.*

KEYWORDS : *annotation, discourse, temporality, calendars, textual navigation.*

1. Introduction

La prise en compte de la temporalité exprimée dans les textes apparaît comme fondamentale non seulement dans une perspective de traitement global de documents[1] mais également dans l'analyse de la structure même d'un document, tout particulièrement quand ce dernier est long. De nombreux travaux depuis l'introduction des cadres de discours de (Charolles 1997) ont ainsi mis en avant l'importance des expressions temporelles en tant que modes d'organisation discursive. L'analyse de la temporalité au sein des textes s'est particulièrement focalisée sur les temps verbaux et sur les adverbiaux temporels.

Notre étude vise les expressions temporelles relatives aux repérages dans un calendrier. Nous les qualifions d' « expressions calendaires » (désormais notées EC). Dans les systèmes actuels d'annotation des EC – qui visent le plus souvent à en calculer les valeurs –, il n'y a pas de véritable consensus sur ce qu'est une information calendaire et sur le traitement des unités utilisées ; en revanche, les calculs visant à expliciter les liens entre les EC et un calendrier donné y sont correctement pris en charge. Aussi, ce n'est pas dans cette problématique que nous nous situons et à laquelle nous cherchons à proposer des éléments d'amélioration. Notre but est de proposer un système de représentation qui, outre le fait qu'il permette de mettre en œuvre ces calculs, soit plus proche de notre « compréhension » d'un texte utilisant différents modes de référenciation à un système calendaire. Nous pensons en effet que, contrairement à l'approche adoptée dans les systèmes actuels, il n'est pas nécessaire de calculer systématiquement et de manière globale toutes les références calendaires, mais seulement de rendre possible ces calculs à l'intérieur du système de représentation. En ce sens, nous nous appliquons à décrire le processus de calcul encodé et non à réaliser le calcul lui-même en privilégiant le caractère local d'une expression.

Dans cet article, nous nous proposons donc de revenir sur ces notions et de décrire formellement les expressions linguistiques relatives aux repérages dans les calendriers afin, d'une part, de proposer un système d'annotation permettant la structuration et la navigation temporelle dans un document et, d'autre part, de lui associer un « système calendaire propre ». Le but de ce travail est d'utiliser ces systèmes calendaires propres pour naviguer temporellement non seulement à travers un document, mais également à l'intérieur d'un corpus. L'article est organisé de la manière suivante : après un rappel des principaux éléments qui participent à l'analyse de la temporalité dans les textes (parties 2 et 3), nous abordons le problème de la représentation formelle des EC et nous présentons notre approche fondée sur la représentation catégorique des ordinaux (partie 4) ; enfin, après une courte synthèse à propos de notre méthodologie d'analyse des EC (partie 5), nous présentons un

[1] Concernant les corpus de documents, les informations temporelles sont en effet essentielles pour organiser les documents selon un (voire plusieurs) ordre(s) chronologique(s).

projet d'application qui consiste à aider à la lecture d'une texte biographique (partie 6). Enfin, nous concluons sur les perspectives offertes par le développement d'une première maquette de navigation temporelle.

2. L'annotation des expressions calendaires dans les textes : quelques remarques préliminaires

La tâche d'annotation dans les textes des expressions qualifiées de temporelles renvoie en réalité principalement à une annotation des seules EC. Elle vise en effet à identifier dans les textes les expressions propres à être ancrées dans un système calendaire, c'est-à-dire dans un système qui permet de situer des événements sur une échelle de temps, en fonction de la durée de ces événements et selon une hiérarchie d'unités – encore appelées « grains ». On distingue alors classiquement deux sous-tâches : une première qui repère automatiquement lesdites EC ; la seconde qui réalise l'ancrage sur une « ligne temporelle » sous forme de valeurs, le plus souvent notées en suivant le format standard ISO 8601 de manière à assurer une certaine « portabilité » des annotations (Mani *et al.,* 2000 ; Setzer *et al.*, 2000 ; Filotava *et al.,* 2001). Ce type de systèmes d'annotation « sémantique » est actuellement principalement développé sur – et à partir de – corpus journalistiques qui font beaucoup usage de ces EC. Il peut être intégré à des applications du domaine de l'ingénierie linguistique ou de l'ingénierie documentaire comme (i) le résumé multi-documents, à l'instar de (Barzilay *et al*., 2001) qui, s'appuyant sur les EC et sur la date d'émission du texte source, cherche à gérer l'ordonnancement temporel des phrases extraites pour une bonne compréhension du résumé final par un lecteur humain ou (ii) la navigation intradocumentaire comme dans (Bilhaut *et al*., 2003) qui s'intéresse plus particulièrement à la délimitation des cadres temporels instaurés par lesdites EC.

Notre approche résulte du travail empirique que nous avons mené sur le repérage et la représentation des expressions temporelles dans un corpus de dépêches fourni par une entreprise et dans un corpus de biographies. Au vu des problèmes rencontrés (concernant essentiellement la circonscription et la classification des expressions considérées comme calendaires) et la façon dont ils sont résolus ou non dans la littérature correspondante, il nous est apparu nécessaire de chercher à rendre compte au mieux de la façon dont l'humain utilise et exprime à travers la langue les informations temporelles à même d'être ancrées dans un système calendaire, plutôt que de focaliser la démarche sur une recherche systématique de référenciation à des valeurs selon une norme qu'est – en la pratique – la norme ISO. Deux remarques peuvent en effet être formulées à l'égard de la manière dont sont appréhendés – ou résolus – les problèmes concernant d'une part, le repérage et, d'autre part, la représentation des informations calendaires dans les systèmes actuels d'annotation des expressions temporelles qui visent à calculer leurs valeurs calendaires, comme dans, par exemple, (Ferro *et al.*, 2003 ; Setzer *et al.* 2000). Pour ce qui concerne tout d'abord la tâche de repérage des EC, nous avons constaté qu'il n'y a pas de véritable

consensus sur ce qu'est une information calendaire et sur le traitement des unités utilisées. Ce fait repose sur la perception hybride de ces unités, vues selon le contexte comme atomiques ou divisibles, la langue utilisant le même terme ('année', par exemple) pour dire la « date » (comme dans « *l'année dernière* ») ou la « durée » (comme dans « *depuis l'année dernière* »). Ce problème n'est pas propre aux EC. Techniquement, il est étudié en intelligence artificielle sous le nom de « problème de granularité » (Bettini *et al.*, 2000 ; Bechet *et al.* 2000). Si l'humain gère parfaitement ce type d' « ambiguïté » (qui n'en est pas vraiment une en réalité, la distinction entre « date » et durée » n'étant pas pertinente selon nous, nous reviendrons sur ce point) et les changements d'échelle, il reste encore à transmettre cette capacité aux systèmes d'annotation et de représentation des EC. Quant à la tâche de représentation des EC, nous avons remarqué que l'ancrage sur une ligne du temps est le plus souvent envisagé d'une manière que nous qualifions de « procédurale » en opposition à une autre manière que nous qualifions de « fonctionnelle ». Cette dernière n'a pas fait l'objet à notre connaissance de travaux approfondis, exception faite de la proposition formulée dans TIMEX3, extension du langage TIMEX2 dans le cadre du projet TimeML (Pustejovsky *et al.*, 2003), mais qui a été abandonnée dans la pratique actuellement, comme le signale (Mani, 2004). C'est pourtant cette dernière qui retient toute notre attention et sur laquelle nous pensons qu'il convient de revenir dans le cadre des traitements automatiques visant à annoter les EC dans les textes.

3. Traitement automatique de la temporalité dans les textes : repérer, annoter et représenter les expressions calendaires

3.1. *Traitement de la temporalité dans les textes*

Actuellement, la temporalité dans les textes est appréhendée dans les traitements automatiques à deux principaux niveaux d'analyse et de représentation[2] : l'un renvoie à la tâche d'ancrage des expressions temporelles dans un système calendaire (mise en relation avec des dates ou des durées) ; l'autre renvoie à la tâche de calcul de l'ordonnancement temporel des événements dans un texte. Historiquement, c'est en fait la deuxième – pourtant plus complexe – qui a fait l'objet des premiers travaux (Song *et al.*, 1991 ; Hitzeman *et al.*, 1995) pour être ensuite peu à peu « abandonnée » au profit de la première considérée comme plus réalisable et surtout comme indispensable dans le cadre du développement des systèmes de Q/R (TERQAS, 2002) ou de résumés multi-documents (Barzilay *et al.*, 2001). Ces deux tâches sont en réalité étroitement liées[3] et renvoient toutes deux à un même problème fondamental, celui qui concerne l'interaction des différents modes

[2] Voir en particulier (Muller *et* Tannier, 2004) pour une synthèse des problèmes rencontrés à ces deux niveaux d'analyse dans le champ de la linguistique computationnelle.

[3] La deuxième tâche s'appuie dans certains systèmes explicitement sur la première, les systèmes concernés cherchant d'abord à calculer les références calendaires des événements pour ensuite chercher à les situer temporellement les uns par rapport aux autres.

d'expression de la temporalité dans la langue (temps grammaticaux, locutions adverbiales, connecteurs, indices typographiques tels que par exemple le guillemet et le point qui ouvrent ou ferment des espaces de validation). Ces différents modes d'expression interagissent entre eux pour renvoyer à une certaine interprétation (aspecto-)temporelle de l'unité textuelle considérée, qu'il s'agisse de l'EC stricte ou de la proposition ou encore d'un segment textuel constitué de plusieurs propositions.

Les tâches de repérage et d'annotation automatiques des EC prévalent actuellement dans le domaine du traitement de la temporalité dans les textes, du fait essentiellement que seules les expressions faisant directement intervenir les unités calendaires sont considérées comme trivialement – *a priori* – repérables et annotables dans leur interprétation temporelle. Ces types de tâches ont donné lieu à nombre de travaux, en particulier ceux instaurés dans le cadre de projets de standardisation comme par exemple TimeML (Pustejovsky *et al.*, 2003). TimeML constitue une proposition de métalangage standard pour l'annotation des événements dans des textes et de leurs relations temporelles. Il a été initialement mis en place dans le cadre du workshop TERQAS (2002), qui lui-même prenait place dans le contexte des systèmes de questions/réponses. Il intègre principalement deux schémas d'annotation, TIDES TIMEX2 (Ferro *et al*., 2003) et Sheffield STAG (Setzer *et al.*, 2000), proposés à partir principalement de l'analyse d'adverbiaux temporels comme, par exemple, « *mardi dernier* ». En vue d'identifier les EC[4], la plupart des systèmes actuels se fondent sur le repérage de certains lexèmes « clés » (ou « indices déclencheurs ») que sont les « grains d'observation classiques » comme 'année', 'mois', ou 'jour'.[5] Des marqueurs linguistiques tels que des articles déterminants (« *le 12 juin* »), des prépositions (« *en 2005* », « *vers 2002* ») ou encore des connecteurs (« *pendant 2002* ») sont ensuite pris en compte dans les systèmes de représentation sous forme d'attributs qui permettent de distinguer les types d'ancrage réalisés[6].

3.2. *Quelques difficultés soulevées à propos des expressions calendaires*

La tâche de repérage des EC n'est cependant pas toujours triviale. En effet, on peut rencontrer : (i) des expressions incorporant un grain sans être calendaires comme dans « *L'année 2002 fut une excellent année* » ; (ii) des expressions

[4] Les EC sont généralement désignées sous le terme de « locutions adverbiales de temps » car elles occupent effectivement une telle fonction du point de vue morpho-syntaxique (voir par exemple pour le français (Maurel, 1989) qui qualifie ces expressions d' « adverbes de date »).
[5] Cela conduit à exclure le plus souvent des expressions telle que « quand le pape est arrivé » *vs.* « le jour où le pape est arrivé ». Ce sera aussi notre choix que d'exclure de l'analyse de telles expressions qui ne rentrent pas *stricto sensu* dans notre définition d'une EC (pour une analyse de ce type d'expressions, voir par exemple (Le Draoulec *et al.*, 2005)).
[6] Dans tous nos exemples, nous soulignerons ce qui fonctionne comme indice déclencheur, c'est-à-dire les unités calendaires, mettant en italique ce qui est interprété comme un attribut, et entre crochets ce qui correspond à l'EC complète.

incorporant un grain non « classique », qui doit donc être identifié au préalable de la tâche de repérage des EC, comme par exemple l'expression « *Au dernier CEBIT* », trouvée dans notre corpus de dépêches[7]. Par ailleurs, se pose le problème de la délimitation de certaines EC, comme le montrent les exemples (1) et (2) tirés de (Vazov, 2001).

(1) Le ministre est venu [3 minutes *après son porte-parole*] qui avait déjà annoncé la bonne nouvelle.

(2) La réunion a commencé et [3 minutes *après*] son porte-parole qui avait déjà annoncé la bonne nouvelle est parti pour la capitale.

Les systèmes rencontrent une autre difficulté, qui est vue en général comme relevant du phénomène d'« ambiguïté » dans la langue : les EC désignent (i) d'une part des dates ou des durées selon le contexte dans lequel elles apparaissent ; (ii) et d'autre part des durées qui ne sont pas de « simples » intervalles. Examinons les exemples (3)-(4), (5)-(6) et (7) infra. Le couple (3)-(4) renvoie à une « ambiguïté » de traitement de la zone temporelle 'deux semaines' : dans (3), l'EC renvoie à une date tandis que dans (4) elle renvoie à une durée. Par ailleurs, l'EC de (3) soulève un autre type de problème : celui de la valeur de 'semaine', qui peut être soit l'unité de comptage soit le nombre de ses jours. Le couple (5)-(6) renvoie à une « ambiguïté » de positionnement de l'occurrence 'samedi' suivant le temps grammatical auquel est conjugué le verbe dans la phrase. Quant à l'exemple (7), l'EC « les jours » est interprétée dans TIDES (Ferro *et al.*, 2003) comme un ensemble d'heures contiguës ou non et le contexte ne permet pas de désambiguïser.

(3) [*Il y a* deux semaines], J. Chirac faisait une intervention télévisée remarquée.

(4) Il a été déprimé [*pendant* deux semaines].

(5) [Samedi], je *viendrai* te voir.

(6) [Samedi], je *suis venue* te voir.

(7) Je suis occupée [les jours] où tu ne l'es pas.

Si (3), (4), (5) et (6) peuvent être considérés comme des cas d' « ambiguïté », correctement traités par les systèmes, nous contestons ce statut à (7) : pour nous, cet énoncé signifie que quelle que soit la période d'occupation de la journée (courte, répartie dans la journée,...), toute rencontre est à oblitérer ce jour-là entre l'énonciateur et le co-énonciateur.

Un autre type de phénomène auquel doivent faire face les systèmes concerne les changements de granularité opérés au sein d'une unité textuelle : notre utilisation des unités calendaires nous permet ainsi des effets de zoom ou d'éloignement, comme dans (8). Nous mettons entre parenthèses le type d'unité calendaire (Jour ou Minute) en jeu dans la proposition.

[7] Dans cet exemple, il est fait référence à une occurrence d'un événement périodique, le « CEBIT », la durée écoulée entre deux « CEBIT » pouvant être considérée comme un grain propre à ce corpus de textes.

(8) [*Le* 8 juin (J)], la machine à café a explosé [*à* 8h (M)]. L'appartement a pris feu [10 minutes (M) *plus tard*]. [*Le* lendemain (J)], j'habitais chez mon voisin.

4. De la datation et des calendriers

Notre but étant de modéliser les EC dans les textes, il nous faut d'abord examiner ce que l'on entend, dans l'usage courant, par les termes de *datation* et de *calendrier*, avant de regarder les systèmes formels susceptibles de les décrire. Dans les dictionnaires anciens (comme les Dictionnaires d'Autrefois (DA) ou le Littré 1994) et dans le *Trésor de la Langue Française* (TLFI), il ressort deux éléments importants : d'une part, la notion de *chronologie*, vue comme une séquence d'événements privilégiés permettant de positionner d'autres événements ; d'autre part, le caractère essentiellement relationnel des calendriers, à savoir une mise en relation de deux unités. Ce sont ces deux points de vue qui ont fondé le modèle formel des systèmes calendaires (Schwer, 2002). Dans ce cadre, nous définissons deux types de chronologies : (i) les C-Chronologies, qui sont liées aux unités usuelles – adoptées par (ISO 8601) comme 'An', 'Mois', 'Semaine', 'Jour', ... – ; (ii) les E-chronologies, correspondant à des événements récurrents comme les couronnements des pharaons égyptiens, des conférences périodiques comme TALN ou CEBIT. De même, nous définissons deux types de dates : les *dates calendaires* et les *dates événementielles*.

Les systèmes d'annotation actuels des EC essaient, dès que le moyen se présente, d'attribuer une date calendaire à une date événementielle. Pour notre part, nous définissons une EC comme étant une expression renvoyant à une date calendaire ou à une date événementielle.

4.1. *Définitions formelles*

4.1.1. *La nature hybride d'un système calendaire*

Formellement, un calendrier, ou système de datation, est un système conventionnel qui permet de se repérer et d'effectuer des mesures dans le temps. En tant que tel, il fonctionne conceptuellement comme tout système de mesure (longueur, masse...), à savoir le choix d'une unité étalon (le mètre pour les longueurs), associé à un système de multiples (décamètre, kilomètre…) et de sous-multiples (décimètre, centimètre…).

Deux problèmes concernent le traitement des unités calendaires. Le premier problème concerne tous les systèmes de mesure : il y a un usage « ambigu » de la langue concernant les unités qui correspondent tantôt à l'unité nombrante, tantôt à la grandeur, c'est-à-dire qui correspondent tantôt à une unité de datation, tantôt à une unité de durée. Le second problème vient de l'échec de rationalisation : les occurrences d'une unité ne correspondent pas toujours au même nombre d'occurrences d'unités plus fines (un mois n'a par exemple pas un nombre fixe de

jours). Mais ce problème est soluble du fait que les occurrences d'une unité sont organisées en une séquence ordonnée, ce qui attribue de fait une identité à chacune d'entre d'elles (son numéro ordinal dans la séquence) et que la distribution cardinale d'une unité à une autre est cyclique (chaque irrégularité observée à une échelle correspondant à une régularité à une échelle supérieure). C'est pourquoi les systèmes de datation actuels possèdent la double dimension de tout système de mesure : ordinale en lecture longitudinale (à une unité donnée) et cardinale en lecture verticale (c'est-à-dire en passant d'une unité à une autre). La dimension ordinale permet de balayer tout le temps social, et de situer tout événement à l'intérieur d'une séquence ordinale d'occurrences d'une même classe d'événements ou, de façon duale, à l'intérieur d'une période correspondant à un segment de temps entre deux occurrences consécutives. La dimension cardinale permet de préciser la situation de l'événement dans des bornes plus précises. Cette double dimension donne également à l'unité utilisée une double nature, soit élémentaire, soit comme séquence ordonnée d'un nombre fini d'occurrences d'une unité inférieure, ce qui lui donne le double aspect du segment pris comme élément simple ou comme intervalle d'éléments simples.

Dans les textes, on observe ces deux types de parcours, ordinal et cardinal, qui correspondent respectivement à des séquencements et à des emboîtements de « cadres temporels » au sens de (Charolles, 1997), comme nous le montrons dans la section 6.1.3. Typiquement, des expressions comme « *la veille* », « *trois ans auparavant* », « *mardi prochain* » induisent un parcours ordinal ; une cascade d'expressions comme « *En 2002, …, au mois de septembre, …, le 21* » induit quant à elle un parcours cardinal.

4.1.2. *Le modèle formel choisi*

De nombreux modèles théoriques des calendriers ont été proposés tant en bases de données qu'en intelligence artificielle (sous le terme de modèles granulaires du temps, parmi lesquels (Niezette *et al.*, 1991 ; Chandra *et al.*, 1994 ; Bettini *et al.*, 2000 ; Béchet *et al.*, 2000)) pour permettre le traitement des informations calendaires. Nous avons choisi pour notre part d'adopter la représentation calendaire fondée sur la représentation catégorique des ordinaux de (Schwer *et al.*, 1998 ; Schwer, 2002), car ce modèle traite de l'aspect ordinal par les ordinaux (ce qui permet le déplacement à granularité constante par simple calcul arithmétique) et de l'aspect cardinal par simple substitution d'expressions « régulières ». Dans la suite de cet article, il ne s'agit pas de présenter *stricto sensu* le modèle de représentation choisi mais d'en expliquer l'essence.

4.2. *Notre proposition de modélisation des expressions calendaires*

La description d'une EC est appréhendée comme un « cheminement » à l'intérieur du système calendaire. À partir d'un point noté *r* (pour référence)

(correspondant à une unité donnée, en général la date de la dépêche, ou le dernier point atteint[8]), ce cheminement s'effectue suivant deux dimensions :

- une dimension horizontale (ou ordinale), quand il s'agit de traduire des déplacements de localisation temporelle au sein d'une même granularité (*cf.* figure 3). Le déplacement du curseur se fait un utilisant les opérateurs d'addition et de soustraction correspondant aux nombres de pas élémentaires (passage d'une occurrence de l'unité à sa voisine) à faire. Nous adoptons les notations suivantes : <*Gr, +n*> pour une addition de *n* pas dans la granularité *Gr* ; <*Gr, -n*> pour une soustraction de *n* pas dans la granularité *Gr* ; <*Gr, +n*> pour une addition de *n* pas dans la granularité *Gr* ; <*Gr*,++ |*Nm*> pour un déplacement positif jusqu'à la première occurrence de *Gr* de nom *Nm*. Ainsi, lorsque le nombre de pas dépend d'une étiquette référentielle ('mardi', 'Noël', 'CEBIT', etc.), nous le traitons comme un *while (*noté + +|) c'est-à-dire comme une opération itérative conditionnelle. Ainsi les expressions « *dans* deux jours » ou « deux jours *plus tard* » correspondent au déplacement <Jour, +2> tandis que l'expression « *le prochain* CEBIT » correspond au déplacement <Jour,++ |CEBIT> ;

- une dimension verticale (ou cardinale), qui change de séquence en suivant les flèches (en montant si l'on passe à une unité plus grossière ou en descendant si l'on passe à une unité plus fine) en utilisant les morphismes f_{xy} (*cf.* figure 4). Le morphisme f_{xy} traduit le passage de la granularité *x* vers la granularité plus fine *y* ; le morphisme $f^{-1}{}_{xy}$ traduit le passage inverse (Schwer, 2002).

Nous allons illustrer sur deux exemples l'intérêt de notre modélisation tout en montrant comment elle se distingue des approches les plus connues dans le domaine du calcul des expressions temporelles.

(9) La commission se réunira [mardi *prochain*].

(10) [*La* semaine *prochaine*], la commission se réunira [mardi].

Pour des exemples tels que (9) et (10), TIDES (Ferro *et al.*, 2003) représente les deux occurrences de « mardi » par une même référence explicite de date, du type : *<TIMEX> 1999-07-22</TIMEX>*, calculée par rapport à la date de la dépêche qui est le 16 juillet 1999. TimeML (Pustejovsky *et al.*, 2003) représente également les deux occurrences de « mardi » de la même façon, mais selon une approche fonctionnelle, du type : (*Tuesday (successor (week (DCT*))), DCT étant le numéro de la semaine de la dépêche. Nous contestons pour notre part le fait que ces deux expressions sont identiques : d'une part, elles peuvent renvoyer à des valeurs différentes (par exemple, dans le cas où le point de référence *r* est un 'lundi') ; d'autre part, traiter les deux expressions de la même façon, c'est ne pas prendre en compte le fait que les jours sont désignés dans la langue usuelle sans référence à la notion de semaine : les noms se substituent aux numéros ordinaux (autrement dit, les jours nominalisés par 'lundi', …, 'dimanche' forment une E-chronologie événementielle dans la C-chronologie des jours). Il convient de noter par ailleurs

[8] Ce point est appelé « focus temporel » dans nombre de travaux afférents (voir par exemple (Muller *et al.*, 2004)).

que nous traiterions de la même façon « *le prochain* CEBIT » ou « Noël *prochain* » ; en cela, nous adoptons aussi une approche différente de (Schilder *et al*., 2001).

Aussi, selon nous, la différence entre l'énoncé (9) et l'énoncé (10) se situe dans le type de cheminement opéré qui doit être représenté pour être à même ensuite de calculer les valeurs correspondantes. Dans le cas de (9), partant du jour de référence *r*, on avance jusqu'à trouver un jour étiqueté par 'mardi' ; autrement dit, l'expression indique d'aller à la prochaine occurrence de M(ardi) dans le sens progressif, c'est-à-dire de suivre la séquence des points dans le sens positif jusqu'au point noté O. Le chemin suivi ignore le passage d'une semaine à une autre dans sa progression. Ainsi, nous représentons (9) par <Jour,++ |Mardi> comme décrit dans la figure 3. Dans le cas de l'exemple (10), pour passer d'un jour à un autre en passant par la granularité semaine, il suffit de suivre le chemin décrit par la figure 4, qui correspond au calcul de l'expression codée par <Jour, f_{sj} [f^{-1}_{sj} (r) + 1] ++|Mardi>.

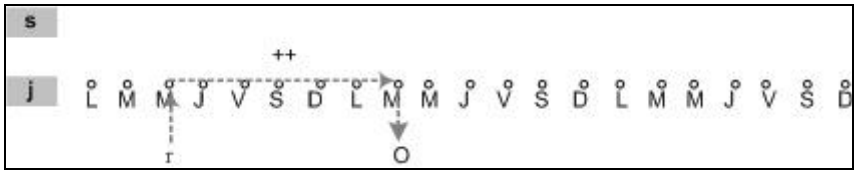

Figure 3. *Cheminement à l'intérieur du système calendaire pour l'énoncé (9)*

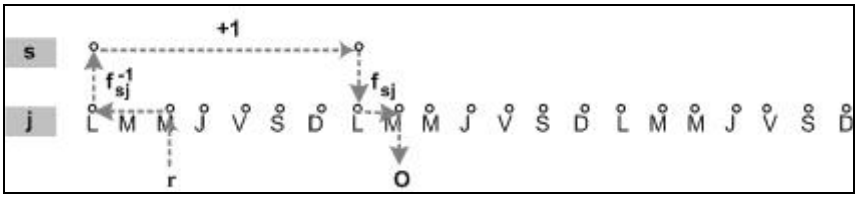

Figure 4. *Cheminement à l'intérieur du système calendaire pour l'énoncé (10)*

5. Synthèse à propos de notre analyse des expressions calendaires et de leur annotation dans les textes

Nous l'avons dit, les équipes impliquées dans la constitution de systèmes d'annotation temporelle des textes ont consacré l'essentiel de leurs efforts aux principes de calculs des liens entre les expressions temporelles dites calendaires et les calendriers (le calendrier grégorien par exemple). Ces travaux ont conduit à des systèmes performants. En revanche, peu d'efforts ont été consacrés à décrire les processus de calcul encodés, de façon à rendre compte de la façon dont l'humain gère les informations temporelles à même d'être ancrées dans un système calendaire.

Notre but est de proposer un tel système de représentation qui, outre le fait qu'il permet de mettre en œuvre ces calculs, soit plus proche de notre « compréhension » d'un texte qui implique différents modes de parcours (vertical, horizontal ou mixte) et de références linguistiques (notions de C-chronologie et de E-Chronologie) à un système calendaire. Nous avons décrit quelques problèmes auxquels sont confrontés les systèmes visant à repérer et à calculer les valeurs des EC ; ce ne sont pas les plus complexes à traiter, tant du point de vue de l'analyse linguistique que du point de vue du traitement informatique. Il reste en particulier à inclure le traitement des discours directs *vs.* indirects (*Il a dit qu'il viendrait demain* vs. *Hier, il a dit « je viendrai demain. »*) ; ce travail est actuellement mené en parallèle par (Battistelli *et al*., 2006).

Par ailleurs, le travail empirique mené sur corpus a, d'une part, fait émerger l'importance et la complexité de traitement de l'unité 'jour' et, d'autre part, permis de montrer la difficulté à catégoriser les expressions temporelles qualifiées de calendaires d'un point de vue aspectuel, l'opposition classique entre ponctuel et duratif ne fonctionnant pas correctement car, intrinsèquement, une EC relève de ces deux aspects. Plusieurs implantations de calendriers particuliers, dont le calendrier chinois de la dynastie Qing (1644–1911), ont validé la pertinence du modèle calendaire utilisé. Une implantation informatique d'un module de repérage des EC dites simples du point de vue de leur expression dans le texte (sans dictionnaire des termes du domaine) et du point de vue calculatoire (car non indexicales) a été réalisée ainsi qu'une extraction du système calendaire propre au corpus de dépêches (Aziez *et al*., 2003). D'autres travaux ont enrichi cette première analyse des EC et ont donné lieu à une évaluation du repérage de ces expressions dans des corpus différents (un corpus de romans et un corpus de biographies) (Boulanger *et al*., 2005 ; Christova *et al*., 2006). Actuellement, nous poursuivons le travail effectué en intégrant les éléments qui concernent la catégorisation aspectuelle des EC selon des principes opératoires. Ces principes sont testés dans le cadre du développement d'une application, la lecture assistée de biographies, qui nécessite la construction d'une représentation intermédiaire des textes que nous nommons « vue calendaire » (voir l'exemple présenté dans la partie 6.4).

6. Vers une lecture assistée de biographies

La lecture d'un texte imprimé est relativement contrainte par le support matériel sur lequel est inscrit ce texte. S'il est toujours possible pour le lecteur de « survoler » le texte en recherchant des indices lexicaux ou typographiques qui le guident dans une lecture qui ne serait plus linéaire, force est de constater que ses possibilités de lecture sont de fait limitées. Des outils d'aide à la recherche d'information pertinente, comme les index, permettent en partie de pallier ce handicap. Le passage à la numérisation a ouvert un grand nombre de possibilités en permettant de placer des liens entre des parties d'un texte. Néanmoins cette lecture hypertextuelle a trouvé ses limites, soulignées par de nombreux auteurs comme (Cotte, 2004).

Notre approche se focalise sur un type de lecture pour un « genre » textuel, les biographies ou les articles de presse qui dressent des portraits de personnages. Ces textes sont caractérisés entre autre par l'utilisation d'un dispositif qui consiste à introduire des « unités textuelles » en rupture avec la description strictement chronologique des événements relatifs à un personnage. Il ne s'agit pas ici de discuter ou de décrire précisément ce dispositif mais plutôt de montrer que l'utilisation d'outils de repérage d'EC tels que présentés ci-dessus, intégrés dans une plate-forme de navigation textuelle, permet de proposer au lecteur différents parcours de lecture, ici une lecture linéaire croisée à une lecture chronologique. Nous allons tout d'abord décrire quelques caractéristiques des textes que nous nous proposons de traiter, puis nous présenterons le système de repérage des EC et les annotations qu'il produit. Nous illustrerons notre démarche en utilisant un texte écrit par Daniel Rondeau publié dans la rubrique « Horizons portrait » du journal *Le Monde* paru dans l'édition du 6 octobre 2005 et intitulé « *Villepin pile et face* ».

6.1. *Repérage et annotation des expressions calendaires cadratives pour une lecture assistée de biographies*

Le repérage et l'annotation des EC peut être réalisé classiquement par l'utilisation d'un logiciel fondé sur l'utilisation d'automates à états finis de reconnaissance. Dans le cadre de cet article, nous nous appuierons essentiellement sur (Christova *et al.*, 2006). Nous limiterons ensuite ces EC à celles dites « cadratives » : relativement à l'application visée (la lecture assistée de textes sous forme d'une navigation textuelle), ces dernières sont en effet plus pertinentes car permettant de naviguer à travers des « blocs » textuels facilement identifiables. Seront dites « cadratives » les EC instaurant un cadre pour un bloc textuel donné, selon la méthodologie exposée dans (Charolles, 1997).

6.1.1. *Repérage des expressions calendaires*

Les EC peuvent être catégorisées selon deux paradigmes : (i) un premier que nous qualifions de « paradigme référentiel » ; (ii) un deuxième que nous qualifions de « paradigme aspectuel » (voir 6.1.2).

Seul le paradigme référentiel est convoqué dans le cadre du repérage des EC (expressions en relation immédiate ou médiate avec un calendrier). Suivant ce paradigme, classiquement retenu dans la littérature sur le sujet, les EC peuvent alors être de deux types, absolues ou relatives et, dans ce dernier cas, soit déictiques soit anaphoriques (textuelles ou extra-linguistiques). Le travail d'analyse linguistique de (Christova *et al.*, 2006) a été réalisé à partir d'un corpus de biographies[9] et de

[9] Trois premiers « sous-corpus » ont été considérés : le corpus 1 contenait de très courtes biographies (3 à 14 lignes), la vie de la personne concernée étant décrite de manière totalement chronologique grâce à des périodes explicites comme « en 1949 », « le 3 janvier », etc. Le corpus 2 était constitué de trois articles de même type mais nettement plus longs (30 à 50 lignes, soit environ une page chacun), et contenant quelques dates

manière à permettre le repérage d'expressions temporelles à même de s'inscrire dans l'objectif de génération du système calendaire propre au corpus (expressions pouvant être insérées dans un calendrier) ; des critères de classification opératoires et des termes utiles pour repérer les expressions sélectionnées en vue de construire les patrons d'extraction en ont été déduits[10].

Dans ce travail, ont été retenues :

(i) les EC indiquant une date ou une durée absolues ;

(ii) les EC à valeur déictique, nécessitant, pour être interprétées, la connaissance de la date de rédaction de l'article ou de la date effective du moment d'énonciation dans le cas d'une citation.. Le repérage de ces expressions est en effet pertinent, dans la mesure où elles peuvent être aisément situées sur un calendrier quand le moment d'énonciation auquel elles renvoient est connu, sachant que, pour de nombreuses biographies (corpus 1 et 2, faisant toutefois peu usage d'expressions déictiques) et pour certaines citations, ce moment n'est pas connu ;

(iii) les EC anaphoriques textuelles, dont l'antécédent doit se trouver dans le texte, sachant que ces dernières peuvent également poser des difficultés d'interprétation, l'antécédent pouvant se trouver à bonne distance de l'expression anaphorique (parfois même dans le paragraphe précédent) ;

(iv) les EC anaphoriques fondées sur les connaissances du monde. Toutes ces expressions peuvent correspondre à une « date » ou à une « durée » précises mais aussi être plus ou moins « floues » quant à la zone temporelle qu'elles désignent ; cela dépend des marqueurs qu'elles comprennent, tels que par exemple un article déterminant, une locution prépositionnelle ou un connecteur (sachant qu'en l'absence de tels marqueurs, les indices typographiques jouent un rôle prépondérant).

Valeur			**EC**
Absolue			*Le 14 février 2003 ; Début 1944 ; En 1977 ; Sur les années 40 ; 1970-1975 ; Juin 40 ; En 1972-1973 ; À l'aube des années 1980 ; Dès juin 1940 ; Depuis 1980 ; Un soir de l'été 2004 ; Au début de l'année 2005 ; A partir de 1958 ; Jusqu'en juillet 1962 ; À dater de novembre 1975 ; Quelques jours avant le 21 avril 1997 ; Du printemps 1991 à décembre 1994 ; Depuis le début des années 60 ; La veille de Noël 1994 ; …*
Relative	*Anaphorique*	*Textuelle*	*Le même jour ; Pendant trois ans ; Cinq ans auparavant ;Trois semaines plus tôt ; L'année suivante ; Le même jour ;Cinq ans auparavant ; À*

anaphoriques. Le corpus 3 était restreint à la seule biographie de De Villepin (« *Villepin pile et face* ») précédemment citée, d'environ 320 lignes (8 pages), rédigée dans un ordre peu chronologique, avec de nombreux retours en arrière et un emploi fréquent d'EC relatives.

[10] Nous renvoyons à (Christova *et al*., 2006) pour la présentation des taux de rappel et de précision.

			l'automne de cette année-là ; L'année suivante ; Le lendemain ; Trois semaines plus tôt ; Ce matin-là ; Quelques jours plus tard ; Dès cet instant ; …
		Extra-linguistique	*Pendant l'Occupation ; Dix ans après sa mort ; Durant sa détention ; Tout au long de sa vie ; Dès l'âge de dix-neuf ans ; À 18 ans ; Deux jours avant son arrivée rue de Varenne ; …*
	déictique		*Le 21 juin dernier ; Il y a encore quelques années ; Il y a six mois ; Jeudi dernier ; …*

Tableau 1. *Catégorisation des EC suivant un paradigme référentiel*

Le tableau 1 reprend une partie des EC repérées dans le cadre de ce travail. Certaines d'entre elles qui figuraient initialement dans les expressions retenues ont été éliminées par nos soins car non valides selon nous au regard des critères que nous avons affinés à la suite de ce travail. On notera ainsi dans ce tableau que, concernant les EC absolues, elles se terminent toutes par la spécification de l'année (en rapport avec le corpus qui inscrit d'emblée dans ce grain de par la date de parution du type j/m/a). Quant aux EC relatives, il est notable que leur caractère anaphorique ou déictique est le plus souvent indécidable, du moins localement, c'est-à-dire sans prise en compte du co-texte[11]. Ainsi seules les EC contenant des marqueurs tels que, par exemple, 'plus tôt', 'avant', 'même' ou 'là' invitent nécessairement à une interprétation anaphorique, et seules les EC contenant des marqueurs tels que, par exemple, 'il y a' ou 'dernier' invitent nécessairement à une interprétation déictique. Aussi, nous n'avons répertorié dans le tableau que les expressions relatives qui pouvaient être interprétées de manière sûre comme anaphoriques ou déictiques. Elles comprennent en particulier celles considérées comme non ambiguës dans la biographie de De Villepin citée précédemment.

6.1.2. *Catégorisation aspectuelle des expressions calendaires*

Repérer puis catégoriser les EC en expressions relatives ou absolues permet de décider si un calcul doit être mis en place (dans le cas des relatives) ou non (dans le cas des absolues) pour pouvoir alors les inscrire dans la représentation calendaire du texte (ou du corpus de textes).

Mais cette distinction ne suffit pas pour être à même de représenter leur sémantique et, au-delà, pour pouvoir envisager de représenter le parcours dans le calendrier qu'appréhende un lecteur : il faut en effet savoir quelles *zones temporelles* elles désignent dans ce calendrier. Ces zones vont dépendre des marqueurs autres que les grains qu'elles comprennent : articles déterminants, locutions

[11] Ainsi, par exemple, des expressions comme « *le 2 janvier* », « *après le 29 mai* » ou « *en cette deuxième semaine de janvier* » (expressions de "dates" ou de "durées" sans précision de l'année) sont, sans prise en compte du co-texte, ambiguës entre lecture anaphorique et lecture déictique ; de même pour des expressions comme « *durant le septennat précédent* », « *durant l'année suivante* » ou « *en quelques mois* ».

prépositionnelles, connecteurs, quantificateurs, etc. Les zones désignées pourront alors être précises ou au contraire « floues », être perçues comme des « points » ou des intervalles, comprendre ou non les bornes initiale et/ou finale. Or ceci relève directement de la catégorie de l'aspect.

C'est pourquoi nous nous proposons ici d'aborder le problème de la catégorisation aspectuelle de ces expressions ainsi que celui de leur représentation. En cela, notre démarche va au-delà du simple étiquetage de ce type d'informations, comme proposé par exemple dans le métalangage standard TimeML (Pustejovsky *et al.*, 2003) ou dans (Muller *et al.*, 2004).

Le plus souvent, les EC sont catégorisées en EC « ponctuelles » et « duratives », ou encore en « EC-dates » et « EC-durées » (*cf.* Muller *et al.*, 04) selon qu'elles désignent respectivement un point (point d'ancrage de l'événement afférent comme dans l'exemple « *Le 2 septembre 2002, l'usine a explosé* ») ou un intervalle (celui dans lequel s'inscrit l'événement afférant comme dans l'exemple « *À partir de mars 2003, les usines ont repris leur activité* »). Selon nous, cette distinction n'est pas pertinente : tout grain conduit à considérer une zone temporelle et la représentation sous forme d'intervalle ou de point n'est qu'une question d'échelle ou encore de grain d'observation (l'EC « *Le 2 septembre 2002* » peut tout autant être appréhendée comme désignant un intervalle ou un point).[12] Sont aussi exclues en général les expressions (ou du moins classées dans une catégorie « fourre-tout ») les EC comme « *au début des années 80* » (étiquetée comme « absolue de forme particulière » chez (Muller *et al.*, 2004) par exemple) ou comme « *quelques années plus tard* » et « *ces dernières années* » (auxquelles on attribue le plus souvent l'étiquette « flou »).

Selon nous, le fait que des critères aspectuels clairs (ou du moins suffisamment définis) ne soient pas véritablement considérés dans ce type de travaux provient de deux raisons qui sont étroitement liées. La première tient au fait que la plupart de ces systèmes n'ont comme visée ultime que le simple repérage/étiquetage (« sémantique ») des EC en vue de les associer aux événements afférents comme dans les exemples donnés ci-dessus (l'analyse se situe donc au niveau de la proposition et non du texte, c'est seulement dans un second temps que ces systèmes vont alors chercher éventuellement à ordonner temporellement les événements entre eux[13]). La seconde raison tient au fait qu'ils ne cherchent pas à *représenter* le texte - d'un point de vue ici strictement calendaire -, ce qui impliquerait non seulement de chercher à ordonner temporellement les EC entre elles mais aussi de montrer le parcours effectué par un lecteur sur l'axe calendaire et de préciser alors les unités

[12] D'un point de vue plus pratique, cette distinction repose souvent sur le fait qu'il faille réaliser une (dans le cas d'EC « ponctuelles ») ou deux (dans le cas d'EC « duratives ») opérations de normalisation *i.e.* de transformation des EC en leurs valeurs positionnables sur un calendrier.

[13] Pour notre part, comme cela sera présenté plus loin, nous chercherons à étiqueter des « blocs » de propositions et non des propositions isolées. Pour cela, nous emploierons le terme « d'EC cadrative » (*cf.* partie 6.1.3).

calendaires perçues selon un certain point de vue (dans leur intériorité, dans leur globalité ou fermeture… ces termes relevant directement de la catégorie de l'aspect). Hormis que ceci nous pose un problème sur le plan théorique[14], c'est précisément pour cette dernière raison que nous avons dû aborder cette problématique d'une catégorisation aspectuelle suffisamment fine des EC et représentable explicitement. Nous proposons de catégoriser aspectuellement les EC à l'aide des notions de point et d'intervalle (ouvert, fermé, semi-ouvert à droite ou semi-ouvert à gauche) ainsi que de voisinage. Ces notions vont nous permettre d'appréhender la (ou les) partie(s) de la zone temporelle correspondant à un grain calendaire en fonction des expressions dans lesquelles apparaissent ces grains (avec présence ou non d'articles définis ou indéfinis, de prépositions, etc.).

Dans le tableau 2, nous représentons de façon figurative l'analyse aspectuelle que nous faisons des EC – seulement cadratives (*cf.* 6.1.3 pour une définition des EC cadratives) – tirées du texte « *Villepin pile et face* ». Considérant que les référents des EC relatives ont été repérées, nous répertorions l'ensemble des EC de ce texte sous une forme absolue (nous indiquons entre parenthèses les EC initiales). Les représentations figuratives qui illustrent leur signification aspectuelle seront reprises quand nous donnerons une vue calendaire de ce texte (*cf.* partie 6.4). Elles sont fondées sur une taxonomie aspectuelle des EC qui est en cours de validation sur d'autres textes de notre corpus.

EC cadratives du texte	**Grain**	**Représentations figuratives**
À l'aube des années 80	Décennie	
En 1977 ; en 2002 ; en 1992 ; en 1995[15]	Année	
Depuis 1980 ; dès 1980 (= dès cet instant)	Année	
Au début de l'année 2005	Zone_ds_année	
À l'automne de l'année 1995 (= à l'automne de cette année-là)	Saison	

[14] Dans le champ plus large de l'analyse de la temporalité linguistique dans laquelle les catégories de temps et d'aspect sont indissociables, il nous semble en effet nécessaire de chercher à analyser la sémantique de tous les types d'expressions temporelles.

[15] Une expression comme « *De 1974 à 1975* » donnerait lieu à une représentation figurative où figurent explicitement les bornes initiale et finale de l'intervalle ainsi défini.

Depuis juillet 2004	Mois	
Quelques mois après l'automne 1995 (= quelques mois plus tard)	Mois	
Au mois d'avril 2005 (= il y a six mois)	Mois	
Durant la semaine 22 de 2005 (= trois semaines plus tôt)	Semaine	
Le 21 juin 2005 (= le 21 juin dernier)	Jour	
Quelques jours avant Noël 1994 (= à la veille de Noël 1994)[16]	Jour	
Quelques jours après le 14 février 2003 (= quelques jours plus tard)	Jour	
Le 15 septembre 2005 (= le 15 septembre dernier) ; le 14 février 2003 ; le 14 février 2003 (= le même jour) ; le 2 janvier 2005 (= le 2 janvier) ; le 20 septembre 2005 (= le 20 septembre dernier) ; le 29 septembre 2004 (= jeudi dernier)	Jour	
Après le 29 mai 2005 (= après le 29 mai)	Jour	
Un soir de l'été 2004, le matin du 21 juin 2005 (= ce matin-là)	Partie_journée	

Tableau 2. *Catégorisation des EC suivant un paradigme aspectuel*

6.1.3. *Repérage et annotation des expressions calendaires cadratives dans les textes*

Nous cherchons à étiqueter non pas des propositions mais des « blocs » de propositions. Ceci est théoriquement mieux fondé du fait qu'une EC valide (ou inscrit dans une zone temporelle donnée) bien souvent pas seulement une

[16] Entre plusieurs interprétations possibles d'une EC, nous avons systématiquement choisi la plus étendue. Pour l'EC « *À la veille de Noël* 1994 », nous n'avons pas retenu l'interprétation correspondant à « *au 24 décembre* 1994 » mais l'interprétation correspondant à « *quelques jours avant Noël 1994*».

proposition mais plusieurs propositions. C'est pourquoi nous ne cherchons pas à étiqueter toutes les EC d'un texte mais seulement les EC dites « cadratives » dans le cadre de notre projet de lecture assistée. Cette dénomination procède d'une analyse en termes de « cadres temporels » instaurés par des adverbiaux temporels ; ce type de cadres ressort des « cadres de discours » de (Charolles, 1997).[17]

Le texte « *Villepin pile et face* » tiré du journal *Le Monde* avec lequel nous allons illustrer notre représentation est composé de 42 paragraphes (noté P1 à P42). Une des caractéristiques importantes de ce texte tient à ce qu'un grand nombre de paragraphes débute par des EC cadratives. Néanmoins, certains paragraphes contiennent plusieurs EC cadratives, ce qui nous a obligé, pour ces cas, à délimiter des segments textuels plutôt que des paragraphes. Nous avons ainsi identifié automatiquement 28 segments textuels qui débutent par une EC comme par exemple : « *Le 21 juin dernier* », « *Ce matin-là* », « *À l'aube des années 1980* », « *À l'automne de cette année-là* ». Il faut préciser que le texte présente de nombreuses « ruptures chronologiques » mises en place par l'auteur du texte, au sens où l'ordre linéaire ne correspond plus à l'ordre chronologique. Par ailleurs, certaines ruptures n'ont pas été identifiées car elles ne sont pas introduites par une expression cadrative mais par la sémantique d'un verbe associée à un adverbe (« *J'ai reparlé récemment de cette période* »). Enfin, la sémantique verbale est parfois utilisée par l'auteur pour faire s'écouler le temps linéairement au cours de la lecture d'un ou de plusieurs paragraphes consécutifs (« *Je l'ai alors accompagné pendant deux mois* »), ce qui correspond à la relation de narration (*cf.* Le Daoulec *et al.*, 2005).

```
<UT Type="Cadre calendaire" Nro="1">
        <UT Type="EC" Nro="1">
         <Attribut Nom="Grain">Jour</Attribut>
        <Attribut Nom="Etendu">Point</Attribut>
        <Attribut Nom="Type">Déictique</Attribut>
        <Attribut Nom="Déplacement">Moins</Attribut>
        <Attribut Nom="Position">Calculable_Relative</Attribut>
        <Attribut Nom="Zoom">Nul</Attribut>
        <Chaîne>Le 21 juin dernier</Chaîne>
```

Figure 5. *Annotation de l'EC cadrative « le 21 juin dernier »*

Nous avons défini un ensemble d'attributs qui annotent les expressions à repérer. À titre d'exemple, nous donnons dans la figure 5[18] les annotations, qui sont conformes à la DTD de Navitexte (Couto 2006), de l'EC cadrative « *Le 21 juin*

[17] Voir par exemple l'analyse des adverbiaux temporels de (Le Draoulec *et al.*, 2005).

[18] Il convient de souligner que la représentation d'un texte proposée par NaviTexte repose sur un seul élément de base (l'UT), mais que cet élément est typé (*cf.* section 7 et Couto *et al.*, 2006) et que les éléments peuvent être emboîtés. Ainsi, dans la figure 5, l'UT typée « cadre calendaire » englobe d'autres UT, de types différents (propositions, segments, etc.) du texte, qui sont sous la portée de cette UT.

dernier ». Les attributs sont définis en relation avec l'ensemble des informations (référentielles et aspectuelles, *cf.* attributs Grain, Etendu et Type) dont on dispose sur l'EC ainsi qu'avec le type de cheminement opéré dans le calendrier selon la formalisation proposée en partie 4.2.1[19] (*cf.* attributs Déplacement, Position et Zoom).

6.2. *Vue calendaire d'un texte : exemple*

On appelle calendrier du texte l'ensemble des éléments calendaires qui interviennent dans le texte. Nous nous sommes restreints ici aux EC cadratives. En général, deux cadres temporels de même grain qui se suivent (en effaçant les cadres intermédiaires qui sont de granularité plus fines) portent sur des segments différents. C'est effectivement ce que l'on observe dans le texte analysé à une exception près. Par exemple, les deux cadres initiés par l'EC « *en 2002* » ne sont pas contigus. Entre les deux s'insèrent les cadres initiés par les EC « *en 1992* » et « *à la veille de Noël 1994* ». En revanche, pour la suite d'EC « *en 1995* », « *l'automne de cette année-là* », « *en 1995* », on est en présence de deux cadres identiques contigus. En effet, « *l'automne de cette année-là* » est un sous-cadre de 1995, ce qui rend contigu les deux cadres 1995. Nous avons émis l'hypothèse qu'il devait exister une référence calendaire en dehors de 1995 entre les deux occurrences de l'EC « *en 1995* ». C'est effectivement le cas dans le texte. La lecture des paragraphes concernés [P30, P33] nous révèle la présence d'une autre EC, encastrée au sein d'un discours rapporté qu'elle date : « Que fait le pouvoir ?, me dit Villepin *un matin de mars 1997* ? S'enferme-t-il dans une pièce… » Cette hypothèse nécessite bien évidemment une validation sur corpus. Par ailleurs, nous sommes confortés dans la nécessité de devoir repérer les discours rapportés et leurs estampilles temporelles pour affiner la description calendaire des textes.

La figure 6 illustre la vue calendaire de la biographie de De Villepin. Les EC cadratives sont représentées les unes par rapport aux autres suivant la représentation des systèmes calendaires adoptée. Sur ce système calendaire, l'élément datant l'article est estampillé par 'N'. C'est l' « origo » du référentiel propre à l'article (ici *le 6 octobre 2005*). Toutes les expressions déictiques sont calculées par rapport à N ou un de ses extraits (*6 octobre 2005* pour les jours, *octobre 2005* pour les mois, *2005* pour les ans). La borne ouverte d'un intervalle semi-ouvert à droite ou ouvert coïncidera toujours avec 'N'. Chaque expression anaphorique est reliée à son référent par un trait vertical (en pointillé). Chaque EC cadrative est située alors dans ce calendrier, en suivant leur ordre dans le texte par une lecture verticale de haut en bas. À gauche de chaque EC est indiqué le numéro de la proposition suivi du numéro du paragraphe où elle apparaît.

Un point signale un positionnement précis comme pour « *Le 21 juin dernier* » ou « *en 2002* ». Une ellipse représente une situation imprécise, qui peut être ramenée à la notion usuelle de voisinage (ex. : *« À la veille de Noël », « à l'aube des*

[19] Ici, on remonte dans les jours et le zoom est nul. On en déduit une position sur le calendrier.

années 80 », « quelques jours plus tard » ou « quelques mois plus tard »). Un voisinage doit être positionné topologiquement par rapport à sa référence. Souvent il s'agit de positionner l'ellipse par rapport à un point, comme pour « *À la veille de Noël* » qui jouxte Noël à sa gauche ou « *quelques jours plus tard* », expression pour laquelle l'ellipse sera placée à la droite de la référence sans la toucher. Mais parfois, il est nécessaire de transformer le point qui représente le segment référence en un intervalle, pour y placer un voisinage. C'est le cas dans l'analyse de « *à l'aube des années 80* » pour laquelle « *à l'aube* » désigne un voisinage autour de la borne initiale de la décennie. Un segment orienté décrit une durée. Dans le texte, toutes les durées sont produites par le marqueur « *depuis* » c'est-à-dire un segment dont le début est un point du système calendaire, et la fin le 'N'. Les mouvements à l'intérieur du système calendaire sont décrits par le parcours visuel 'gauche vers droite', 'droite vers gauche' ou stationnaire de lecture de haut en bas. L'exemple montre *a priori* un parcours chronologique hiératique de la progression textuelle.

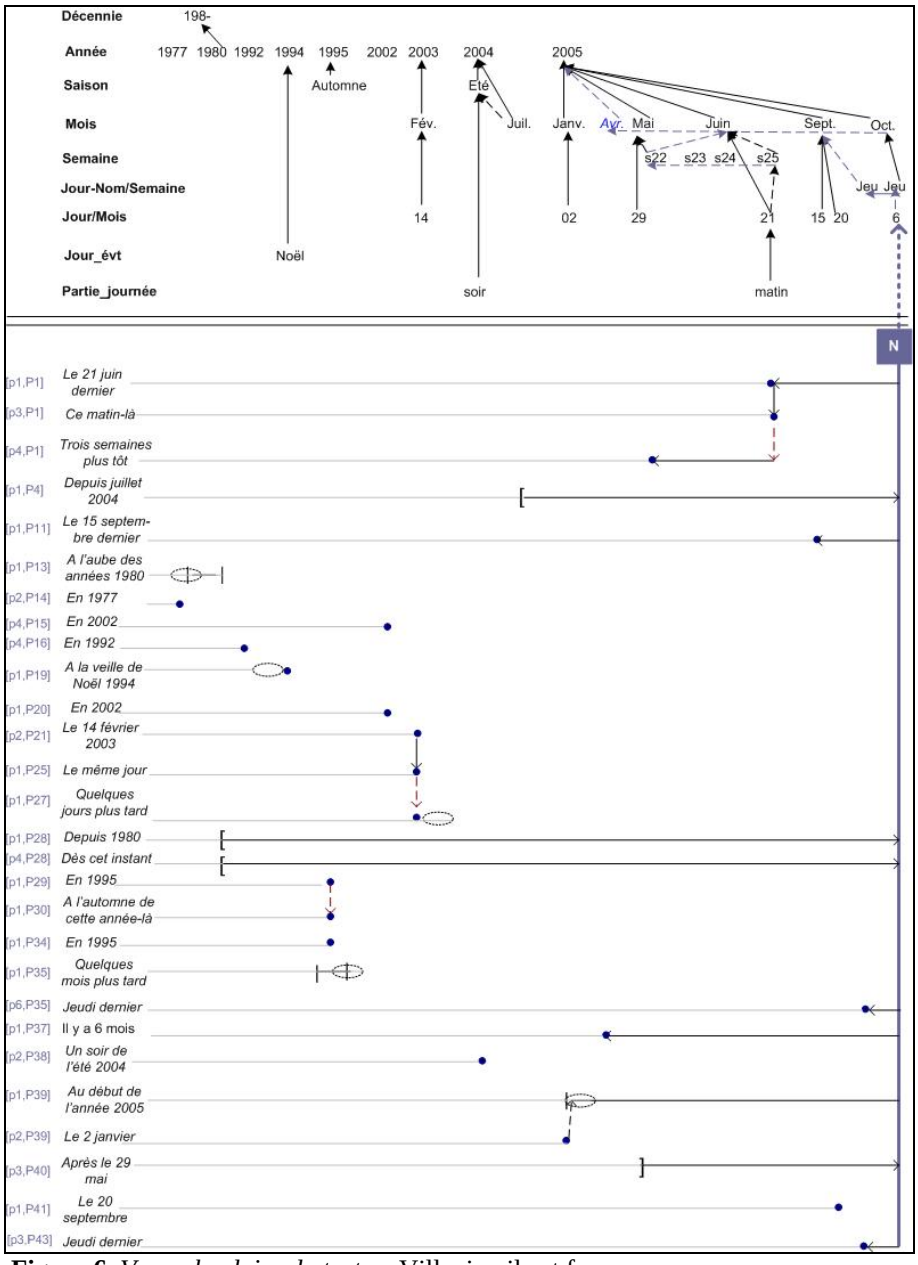

Figure 6. *Vue calendaire du texte «* Villepin pile et face *»*

7. Conclusion et perspectives

Nous avons entrepris d'utiliser la représentation calendaire proposée pour développer un système d'aide à la lecture de biographies qui s'appuie sur les fonctionnalités offertes par la plate-forme de navigation textuelle *NaviTexte* (Couto *et al.*, 2006). Rappelons brièvement les principes d'encodage d'un texte proposés par cette plate-forme. La représentation du texte, décrite dans un format standard XML, se divise en deux parties : le *Corps*, où les unités textuelles, significatives pour la tâche sont délimitées, et la *Tête*, où s'expriment les relations non hiérarchiques entre ces mêmes unités. Dans le *Corps*, l'élément de base du modèle est l'*Unité Textuelle* (UT) typée, ce qui permet d'incorporer de nouveaux éléments textuels de manière simple. Ces principes d'annotation sur lesquels s'appuie *NaviTexte* sont classiquement ceux proposés par les standards tels que ceux de la TEI. Dans le cadre de notre travail, les cadres temporels instaurés par des EC cadratives sont annotés dans le *Corps* comme UT de type « cadre calendaire ». Par ailleurs, un objet *Séquence* est déclaré dans la *Tête* et ordonne chronologiquement ces UT. Il convient de remarquer qu'un objet *Séquence* ne permet que l'ordonnancement des segments correspondants à des dates différentes, en d'autres termes il n'est possible de déclarer qu'un ordre total, ce qui est insuffisant pour notre travail car plusieurs segments peuvent référer à des périodes qui se chevauchent comme par exemple « *À la fin des années 1970* » et « *À l'aube des années 1980* ». C'est pourquoi nous travaillons à la représentation de structures sous formes de S_langages (Schwer, 2004 ; Battistelli *et al.*, 2004).

Dans *NaviTexte*, la navigation est conceptualisée comme une opération reliant une UT *source* avec une UT *cible*. Une opération de navigation est définie comme une opération qui cherche l'UT cible à partir de l'UT source, en vérifiant les différentes conditions, exprimées dans le langage SEXTANT (Couto *et* Minel, 2006) et en suivant l'orientation et l'ordre spécifiés par le type d'opération. Il est ainsi possible de spécifier que la recherche de la cible ne se fait pas en suivant l'ordre des UT dans le texte (ordre narratif), mais un ordre (déclaré par l'intermédiaire d'un objet *Séquence*). C'est cette fonctionnalité que nous utilisons pour la lecture d'une biographie dans l'ordre chronologique.

Une première maquette nous a permis de tester différents scénarios d'usage et de préciser la finalité de l'étiquetage. Nous cherchons à étiqueter, non pas des propositions, mais des « blocs » de propositions. Ceci nous distingue des approches telles que par exemple celle adoptée par (Muller *et al.*, 2004) ou dans le cadre du projet TimeML (Pustejovsky et *al.*, 2003). Il nous semble en effet théoriquement mieux fondé de pratiquer ainsi du fait qu'une EC valide bien souvent, non pas une proposition, mais plusieurs propositions d'un texte. C'est pourquoi, dans le cadre de ce projet, nous avons choisi d'étiqueter seulement les EC dites « cadratives » selon une méthodologie définie par (Charolles, 1997). La représentation calendaire d'un texte permet ensuite d'inscrire des dates (« précises » ou « floues ») relevées dans le texte (calculées dans le cas des EC relatives) mais aussi des périodes dont nous

explicitons si les bornes initiales et/ou finales sont exclues on non (qu'elles soient explicites ou non). La représentation calendaire d'un texte permet alors plusieurs lectures de ce texte (ou parcours dans le calendrier du texte) :

(i) une lecture linéaire (appelée « lecture narrative » dans l'interface graphique et donc visualisable selon un axe vertical) ;

(ii) une lecture chronologique (selon un ou plusieurs axes horizontaux, le nombre de ces axes étant fonction des effets de zoom spécifiés et donc du nombre de grains employés dans un texte) ;

(iii) une lecture libre (initiée par des opérations de navigation textuelle comme par exemple « Aller à l'événement calendaire suivant de même grain»).

Il convient enfin de noter que l'ordonnancement chronologique des UT de type « cadre calendaire » a été réalisé manuellement. Néanmoins, un processus de calcul automatique est en cours de réalisation. Ce calcul se fonde non pas sur des tables de transitivité de relations binaires à la (Allen, 1984) mais sur le calcul algébrique des S-langages, que nous avons déjà expérimenté (Battistelli *et al.*, 2004) et qui permet de traiter directement des relations n-aires sur des chaînes de dates (Jungjaryannon *et al.*, 2002) pour construire les ordonnancements temporels des segments textuels.

Remerciements

Le logiciel NaviTexte est financé par le programme ECOS-Sud U05H01.

8. Bibliographie

Allen J. « Towards a General Theory of Action and Time », *Artificial Intelligence 23*, p. 123-154, 1984.

Aziez A., Porquet T., Santorum L., Signourel O., Repérage et étiquetage d'expressions temporelles en vue de leur exploitation dans un système calendaire, rapport de stage annuel, M2 Professionnel *Ingénierie de la Langue pour la Gestion Intelligente de l'Information*, Université Paris 4, 2003.

Barzilay R., Elhadad N., McKeown K., « Sentence Ordering in Multidocument Summarization », *Actes de HLT'01*, 2001.

Battistelli D., Chagnoux M., Desclés J.-P., « Référentiels et ordonnancements temporels dans les textes », *Cahiers Chronos*, Amsterdam/Atlanta, Rodopi, 2006.

Battistelli D., Minel J.-L., Picard E., Schwer S. R. « Temporalité linguistique et S-langages », *Actes de TALN'04*, Fès, Maroc, p. 33-38, 2004.

Bechet G., Clerin-Debard F., Enjalbert P., « A qualitative Model for Time Granularity », *Computational Intelligence*, Vol. 16 (2), p. 137-175, 2000.

Bettini C., Jajodia S., Wang S., *Time granularities in Databases, Datamining, and Temporal Reasoning*, Springer, 2000.

Bilhaut F., « The Linguastream Platfom », *Actes de 19th Spanish Society for Natural Language Processing Conference (SEPLN)*, Alcalá de Henares, Spain, p. 339-340, 2003.

Boulanger V., Kang J., Repérage d'expressions temporelles dans le système Unitex, rapport de stage annuel, M2 Professionnel *Ingénierie de la Langue pour la Gestion Intelligente de l'Information*, Université Paris 4, 2005.

Chandra R., Segev A., Stonebracker M., « Implementing Calendars and Temporal Rules in Next Generation Databases », *Actes de I. C. on Data Engineering*, p. 264-273, 1994.

Charolles M., « L'encadrement du discours - Univers, champs, domaines et espace », *Cahiers de recherche linguistique, LANDISCO*,vol 6, Université Nancy 2, p. 1-73, 1997.

Christova V., De Vismes O., Repérage d'expressions calendaires et relations de dépendance, projet annuel, M2 Professionnel *Ingénierie de la Langue pour la Gestion Intelligente de l'Information*, Université Paris 4, 2006.

Cotte D., « Leurres, ruses, désorientation dans les écrits de réseau : la métis à l'écran. », *Communication et langages*, n° 139, p. 63-74, 2004.

Couto J., Minel J.-L, « Navigation textuelle : représentation des textes et des connaissances.», *TAL* n° 47/1, 2006.

(DA) Dictionnaires d'autrefois, http ://colet.uchicago.edu/ cgibin/dico1look.pl/

Ferro L., Gerber L., Mani I., Sundheim B., Wilson G., « TIDES Standard for the Annotation of Temporal Expressions », http ://www.mitre.org/work/tech\-papers/tech\-papers\-04/ferro\-tides/, 2003

Filatova E., Hovy E., « Assigning Time-Stamps to Event-Clauses », Actes de *Workshop on Temporal and Spatial Information Processing, ACL'2001*, p. 88-95, 2001.

Hitzeman J., Moens M., Grover C., « Algorithms for Analyzing the Temporal Structure of Discourse », *Actes de EACL'95*, p. 253-260, 1995.

International Standard ISO 8601, interchange formats –Information interchange – Representation of dates and times, 1997.

Jungjariyanonn S., Schwer R. S., « Extended Boolean Computations », Actes de *Workshop on Spatial and Temporal Reasoning, ECAI'15*, Lyon, p. 63-68, 2002.

Le Draoulec, A., Péry-Woodley, M.-P., « Encadrement temporel et relations de discours », *Langue Française* 148, p. 45-60, 2005.

Littre P.-E., *Dictionnaire de la langue française*, Tomes 1 à 7, Encyclopaedia Britannica France, 1994.

Mani I., « Recent Developments in Temporal Information Extraction », *Actes de RANLP'03*, 2004.

Mani I., Wilson G., « Robust Temporal Processing of News », Actes de 38ème ACL, p. 69-76, 2000.

Maurel D., Reconnaissance de séquences de mots par automate, adverbes de date du Français, Thèse de Doctorat, Université Paris 7, 1989.

Muller P., Tannier X., « Une méthode pour l'annotation de relations temporelles dans des textes et son évaluation », *Actes de TALN'04*, Fès, Maroc, 2004.

Niezette M., Stevenne J.-M., « An efficient Symbolic Representation of Periodic Time », Actes de *First Conf. on Information and Knowledge Management*, p. 280-290, 1991.

Pustejovsky J., Castano J., Ingria R. , Saurı R., Gaizauskas R., Setzer A., Katz G., « TimeML : Robust Specification of Event and Temporal Expressions in Text », *Actes de IWCS-5 Fifth International Workshop on Computational Semantics*, 2003

Schilder F., Habel Ch., « From Temporal Expressions to Temporal Information : SemanticTagging of News Messages. », *Actes de ACL'01, Workshop on temporal and spatial information processing*, p. 65 -72, 2001.

Schwer R. S., « Formalizing Calendars with the Category of Ordinals », *Applied Intelligence*, Vol 17 (3), p. 275-295, 2002.

Schwer R. S., « Traitement de la temporalité des discours : une Analysis Situs », *Cahiers Chronos*, Amsterdam/Atlanta, Rodopi, 2006.

Schwer R. S., Vauzeilles., « Calendars inside the Framework of Finite Ordinals Category », *Proceedings of ECAI-98 Workshop on Spatial and Temporal Reasoning*, 1998.

Setzer A., Gaizauskas R., « Annotating Events and Temporal Information in Newswire Texts », *Actes de 2ème LREC*, p. 64-66, 2000.

Song F., Cohen R. « Tense Interpretation in the Context of Narrative », *Actes de 9ème AAAI*, p. 131-136, 1991.

TERQAS, Time and Event Recognition for Questions Answering Systems, an ARDA Workshop on Advanced Question Answering Technology, 2002, http ://www.timeml.org/terqas/

Text Encoding Initiative, http ://www.tei-c.org

TLFI, Trésor de la Langue Française informatisé, http ://atilf.atilf.fr/Dendien/scripts/tlfiv5/

Vazov N., « A System for Extraction of Temporal Expressions from French Texts », *Actes de TALN'2001*, p. 315-324, 2001.

Annexe

A1. Au sujet de la norme ISO

Comme toute norme, la norme ISO est au mieux un bon compromis. En tant qu'elle traite des calendriers pour les applications industrielles, elle est une bonne illustration de la pratique des calendriers dans ce contexte. Son but est la représentation des dates dans le calendrier grégorien, des horaires et des périodes de temps. Ce n'est pas la partie spécification des représentations numériques des informations temporelles qui retient ici notre attention mais les unités retenues, et la classification proposée des « dates ».

Cette classification repose sur deux unités de base : 'an' et 'jour'. L'unité pivot est le 'jour'. C'est un système d'horo-datation. À la question « *Quand ?* » est répondue une date si elle situe un événement sur toute la ligne temporelle avec une unité au moins égale au 'jour', un horaire si la réponse situe un événement à l'intérieur d'une journée. Ainsi, on peut dater à n'importe quel niveau de granularité, la date des apparitions des espèces animales relèvera plutôt des siècles, voire des millénaires, unités à ajouter dans ce cas. En revanche, l'horaire correspond à l'unité la plus fine du système ; pour ISO, c'est la seconde. Trois types de dates sont à considérer : (i) la *date* en général, exprimée par combinaison d'unités 'siècle', 'an-calendaire', 'mois-calendaire', 'semaine-calendaire', 'jour-calendaire', ou simplement 'jour de l'année' ; (ii) la *date calendaire*, jour particulier de l'année-calendaire, identifié par son nombre ordinal à l'intérieur d'un "mois-calendaire" lui-même dans une année ; (iii) la *date ordinale*, jour particulier de l'année-calendaire, identifié par son nombre ordinal à l'intérieur d'une année.

La définition d'ISO 8601 du 'jour calendaire' est une période de 24 heures commençant à 0000 et finissant à 2400 (qui est égal au début de 0000 du jour suivant) ; la définition du 'jour' est une unité de temps de 24 heures. Puis, une note précise qu'un jour calendaire est souvent appelé jour. On trouve cette même note en ce qui concerne 'mois'/'mois calendaire', 'semaine'/'semaine calendaire' et 'an'/'an calendaire' : l'attribut 'calendaire' définit une période de temps (entité ordinale), son absence une unité (entité cardinale). Période et unité sont définies comme des périodes relatives : un mois est une unité d'un nombre variable de jours (28, 29, 30 ou 31) ; un mois calendaire se définit comme entité intermédiaire entre an et jour : douzième partie d'an calendaire et ensemble d'un certain nombre de jours ; une semaine est une unité de sept jours ; une semaine calendaire est une période de sept durée jours[20] au sein d'un an calendaire sauf pour la première et la dernière de l'année ; un an est une unité dont la égale une année calendaire. Un an calendaire est la période de temps cyclique dans un calendrier qui est nécessaire à une révolution de la terre autour du soleil (approximée en un nombre entier de jours).

[20] Pour la date ordinale, les semaines commençant et finissant l'année peuvent être tronquées, la semaine 1 comprenant un jeudi (il peut donc y avoir une semaine 0). Une semaine ordinaire commence un lundi, est identifiée par son numéro ordinal dans l'année.

Pour les unités infra-jour, ISO se fonde sur le Système International des Unités (SI). L'unité de base du système de mesure du temps est la seconde. La minute est définie comme 60 secondes, l'heure comme 60 minutes et le jour comme 24 heures. Le système calendaire global pris en compte par ISO est représenté dans la figure 2[21].

Il ressort de cette description la force de « l'ambiguïté » langagière entre unité et période, et la particularité de l'unité 'jour', comme élément pivot.

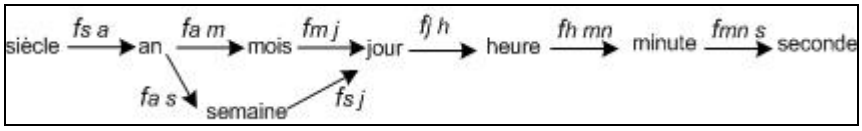

Figure A1 : *Le système horo-dateur de ISO 8601*

Le cas de minuit

Si les cloches qui sonnent les heures marquent une transition entre deux périodes consécutives d'une heure, l'expression « *Rendez-vous à 8 h* » est interprétée comme le début de la huitième heure et non la fin de la septième. Tel n'est pas le cas de *minuit*, qui garde son rôle de cloche. La norme ISO-8601 accepte deux représentations de *minuit*, l'une comme 'début du jour' (00 :00 :00), l'autre comme 'fin du jour' (24 :00 :00). ISO-8601 considère en effet le jour calendaire comme un intervalle fermé sur l'ensemble discret de la C-chronologie des secondes, dont les bornes sont partagées entre deux jours consécutifs, l'écriture de cette borne dépendant de l'intervalle dans lequel on la capte. Cette représentation ne correspond donc pas à une partition de la ligne temporelle mais à un recouvrement, contrairement au modèle calendaire de (Schwer *et* Vauzeilles, 1998 ; Schwer 2002), qui associe à chaque occurrence d'unité un intervalle semi-ouvert à droite.

Ce modèle distingue deux parcours différents pour J : 24 :00 :00 et J+1 : 00 :00 :00. Le second fait un pas sur la C-chronologie Jour, puis descend sans déplacement horizontal sur les C-chronologies plus fines, le premier descend sur la C-chronologie Heure, se déplace de 24 occurrences puis descend sur les unités plus fines. Il y a bien égalité des localisations, mais non égalité des parcours depuis J : 00 :00.

[21] Les morphismes $f_{x\,y}$ se composent ainsi : $f_{x\,y} \circ f_{y\,z} = f_{x\,z}$. Par exemple $f_{a\,m} \circ f_{m\,j} = f_{a\,j}$, ce qui permet de passer d'une échelle à une autre en « oubliant » les étapes intermédiaires.