Proposition de caractérisation et de typage des expressions temporelles en contexte

Maud Ehrmann, Caroline Hagège

Proposition de caractérisation et de typage des expressions temporelles en contexte

Maud Ehrmann, Caroline Hagège
Xerox Research Center Europe - XRCE
6, Chemin de Maupertuis, 38240 Meylan
{Maud.Ehrmann, Caroline.Hagege}@xrce.xerox.com

Résumé. Nous assistons actuellement en TAL à un regain d'intérêt pour le traitement de la temporalité véhiculée par les textes. Dans cet article, nous présentons une proposition de caractérisation et de typage des expressions temporelles tenant compte des travaux effectués dans ce domaine tout en cherchant à pallier les manques et incomplétudes de certains de ces travaux. Nous explicitons comment nous nous situons par rapport à l'existant et les raisons pour lesquelles parfois nous nous en démarquons. Le typage que nous définissons met en évidence de réelles différences dans l'interprétation et le mode de résolution référentielle d'expressions qui, en surface, paraissent similaires ou identiques. Nous proposons un ensemble des critères objectifs et linguistiquement motivés permettant de reconnaître, de segmenter et de typer ces expressions. Nous verrons que cela ne peut se réaliser sans considérer les procès auxquels ces expressions sont associées et un contexte parfois éloigné.

Abstract. Temporal processing in texts is a topic of renewed interest in NLP. In this paper we present a new way of typing temporal expressions that takes into account both the state of the art of this domain and that also tries to be more precise and accurate that some of the current proposals. We explain into what extent our proposal is compatible and comparable with the state-of-the art and why sometimes we stray from it. The typing system that we define highlights real differences in the interpretation and reference calculus of these expressions. At the same time, by offering objective criteria, it fulfils the necessity of high inter-agreement between annotators. After having defined what we consider as temporal expressions, we will show that tokenization, characterization and typing of those expressions can only be done having into account processes to which these expressions are linked.

Mots-clés : Temporalité, typage et caractérisation des expressions temporelles.

Keywords : temporal processing, temporal expressions characterization and typing.

1 Introduction

Le travail que nous présentons s'inscrit dans la perspective de l'analyse temporelle des textes, qui comprend généralement les étapes suivantes :

– reconnaissance et caractérisation des expressions temporelles
– résolution référentielle et normalisation des expressions temporelles
– reconnaissance des événements (avec temps et aspect associé)

– mise en relation événement-expressions temporelles ou événement-événement
– inférence temporelle

La définition de ces étapes est arbitraire et, si elle s'avère utile pour structurer nos discours sur l'analyse temporelle, il importe d'avoir à l'esprit le fait qu'elles sont nécessairement interdépendantes. Nous adhérons à la position de (Gosselin, 1996) qui indique que " *... les marques temporelles et aspectuelles se répartissent sur divers éléments de l'énoncé (le verbe, le temps verbal, les compléments du verbe, les circonstanciels, les constructions syntaxiques etc.) qui paraissent interagir les uns avec les autres de telle sorte que la valeur de certains marqueurs semble ne pouvoir être fixée indépendamment du calcul global de la valeur du tout* ". Dans notre proposition, nous nous intéressons seulement à la première des tâches mentionnées ci-dessus, tout en gardant à l'esprit qu'elle ne peut pas être dissociée des autres.

Plusieurs travaux de TAL s'intéressent au traitement de la temporalité dans les textes. On citera la référence que constitue la norme TimeML (Saurí *et al.*, 2006), laquelle propose un ensemble d'annotations des expressions temporelles et des procès. Cette norme est actuellement adaptée pour le français (Bittar, 2008) et plus ou moins fidèlement observée dans plusieurs travaux pour le traitement automatique de la temporalité, parmi lesquels (Parent *et al.*, 2008) pour le français et, pour l'anglais, (Hagège & Tannier, 2008) ainsi que d'autres systèmes ayant participé à la campagne TempEval[1]. (Battistelli *et al.*, 2006) se distingue de cette mouvance TimeML en proposant une caractérisation fine d'une sous-classe d'expressions temporelles du français (les expressions calendaires) à l'aide d'opérateurs. Cette approche apparaît comme complémentaire de celle adoptée dans TimeML puisque, d'une part, elle n'envisage qu'un sous-ensemble des expressions temporelles et que, d'autre part, elle complémente et précise, pour ce sous-ensemble, le typage de plus haut niveau proposé par TimeML. Il s'avère cependant que de nombreuses difficultés persistent relativement à la délimitation et au typage des expressions temporelles. Dans cet article, nous proposons un système de typage des expressions temporelles (ET dans la suite du document) visant à pallier certains problèmes que nous avons pu relever dans la norme existante (TimeML), tout en cherchant à rester compatible avec les travaux actuellement menés dans le domaine.

Dans la première section, nous présentons les principaux éléments ayant motivés notre travail ainsi que nos objectifs. Nous détaillons ensuite notre définition et notre typage des ET. Le typage proposé a pour objectif du contribuer au projet plus ambitieux de la mise en place d'un système d'annotation automatique de la chronologie des procès apparaissant dans les textes. Nous verrons qu'un des points clefs de notre travail de caractérisation est que celle-ci ne peut se faire sans une prise en compte d'un contexte parfois étendu.

2 Motivations et objectifs

Deux points essentiellement motivent notre proposition de caractérisation et de typage des ET. Le premier est relatif au fait qu'il est, en pratique, souvent difficile de typer des expressions temporelles. En effet, si des expressions comme *le 15 janvier 2008*, *mardi prochain* ou encore *pendant 2 jours* sont considérées sans grande discussion comme des dates (pour les deux premiers) et une durée, il en est d'autres plus difficiles à typer, comme dans les exemples suivants (en gras) :

(1) ***Pendant ces quelques jours****, Lyon vit aux couleurs du Sirha, grand messe de la Gastronomie !*

[1] `http://timeml.org/tempeval`

(2) *L'entrée éventuelle de la Turquie dans l'UE, c'est* ***au moins dans quinze ans****.*
(3) ***Le matin****, il prend son petit-déjeuner et part travailler*.

Dans bien des cas, les expressions temporelles demandent, pour leur caractérisation, la connaissance d'un contexte allant bien au-delà de la simple expression. La solution à de telles difficultés pourrait apparaître dans les normes existantes (TimeML et TIDES) or, et c'est le deuxième élément motivant notre proposition, il est possible d'observer certaines imprécisions dans ces dernières. Au regard du typage des ET tout d'abord : TimeML impose la détermination d'un type, pouvant prendre les valeurs DATE, TIME, DURATION et SET, mais le guide ne donne pas de critères précis pour ce typage (Saurí *et al.*, 2006). En effet, chaque type reçoit une description non définitoire ainsi qu'une courte liste d'exemples non contextualisés pour la plupart. Cette description est insuffisante pour l'annotation de cas "difficiles" trouvés dans les textes. Un autre type d'imprécision observable dans TimeML et TIDES concerne la segmentation des ET : les ET de type *entre le 3 janvier et le 5 janvier* sont segmentées en deux parties (Saurí *et al.*, 2006; Ferro *et al.*, 2005) et, autre exemple, TIDES considère l'expression *8 :00 pm. Friday* comme un seul segment mais l'expression *at 8 :00 p.m. on Friday* comme deux segments. Cela est dû au fait que ces normes ont adopté comme règle d'or de ne pas intégrer les introducteurs d'expressions temporelles dans les ET ; nous nous distinguons de cette position (voir ci-après la section 3.2). Enfin, dans TimeML, certains cas d'annotation ne font pas une distinction claire entre annotation pour la caractérisation et annotation des étapes de calculs. Par exemple, dans *John left 2 days before yesterday*, une étape d'annotation intermédiaire considère *2 days* comme une DURÉE et *yesterday* comme une DATE, avant l'adoption d'une annotation finale normalisant l'ensemble de l'expression *2 days before yesterday* sous forme d'une balise *TIMEX3* vide. Ces annotations "intermédiaires" relevant du calcul et non de la caractérisation n'ont à notre sens pas lieu d'être dans un guide d'annotation.

Ces observations nous ont conduit à vouloir proposer un typage des ET fondé sur des critères rigoureux tout en restant le plus possible compatible avec les normes actuelles. Ce typage serait intégrable dans des campagnes d'annotation d'entités nommées (Ester II et HAREM[2]), qui considèrent de plus en plus l'annotation des ET comme une extension naturelle de l'annotation de la catégorie DATE (traditionnellement présente dans le jeu d'étiquettes standard des entités nommées) mais dans lesquelles la complexité du travail d'annotation des ET est souvent sous-estimée.

3 Définition des expressions temporelles

Une ET correspond à toute expression linguistique faisant référence à une " étendue de temps " et respectant les critères d'indentification, de délimitation et de segmentation présentés ci-après.

3.1 Critères pour l'identification

Les critères que nous définissons pour l'identification des ET sont les suivants : une suite linguistique est une ET si elle satisfait conjointement le critère 1, au moins un des sous-critères listés dans le critère 2, et le critère 3[3].

[2]Respectivement, Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques (`http://www.afcp-parole.org/ester/`) et Campagne d'évaluation des EN du portugais (`http://www.linguateca.pt/HAREM/`).

[3]Ces critères sont pour partie inspirés de (Hagège *et al.*, 2008).

Critère 1 : une ET permet, en contexte, de répondre à une des questions suivantes : *(prep) Quand ?*, *(prep) Combien de temps ?*, ou *A quelle fréquence ?*

Critère 2 : une ET est constituée d'un ou d'une suite de syntagmes noyaux[4]. Chacun de ces syntagmes noyaux doit contenir au minimum une des unités lexicales listées ci-dessous et qui fonctionnent comme tête syntaxique de ce syntagme :

2. 1 : un patron numérique " temporel " caractéristique, tel que *20/02/2009* ou *8 :00 EST*. Notons qu'une partie de ce patron peut être omis (ex : *Il arrive le 3*).

2. 2 : une unité de mesure temporelle (seconde, jour, année, etc.) ou un adverbe en *-ment* dérivé de ces expressions (*quotidiennement*, *mensuellement*, etc.)

2. 3 : un substantif nommant un élément calendaire (*lundi*, *mars*, etc.), une saison (*été* etc.) ou correspondant à une fête du calendrier de référence (*Noël* etc.)

2. 4 : un nom désignant un moment particulier de la journée (*matinée*, *soir*). On adjoint également des noms génériques de période temporelle comme *époque*, *moment*, *instant*, etc.

2. 5 : un indexical " temporel ", soit un adverbe de temps simple (non dérivé) ou un nom " situant les faits dans la durée par rapport au moment de la parole ou un autre repère " (Grevisse & Goosse, 1986). Par exemple, *jadis*, *hier* pour les adverbes et *veille* et *surlendemain* pour les noms.

2. 6 : un groupe prépositionnel complément d'un nom événementiel formé de <NUM+Ntemp> (*un voyage de 5 jours*). Il convient d'être attentif au caractère événementiel du nom ; en cas d'hésitation, nous proposons le test suivant : remplacer le groupe prépositionnel par une subordonnée relative introduite par " qui <forme du verbe *durer*> <NUM> <Ntemp> ". De la sorte, " 4 ans " est bien une ET dans *un projet de 4 ans* , mais non dans *un enfant de 4 ans* (**un enfant qui a duré 4 ans*).

2. 7 : une expression de fréquence de type *de temps en temps*, *quelques fois*, *fréquemment* et faisant référence à une fréquence.

2. 8 : une expression construite avec un présentatif (*il y a*, *cela fait*) et un des éléments listés ci-dessus (il y a cinq ans)

Pour chacun des points listés ci-dessus, nous avons établi des listes fermées issues de (Grevisse & Goosse, 1986) et d'exemples rencontrés dans les textes. Ces listes ne sont pas données ici faute de place mais sont partie intégrante d'un guide d'annotation et absolument indispensables pour l'application de ce critère.

Critère 3 : une ET ne peut pas se paraphraser en "l'événement qui s'est produit (prep)+ET ". En cas d'ambiguïté, c'est la lecture événementielle qui prime.

L'association de ces trois critères permet d'identifier de manière précise un ensemble de suites linguistiques correspondant à des ET ; ces dernières se trouvent, le plus souvent, dans une position adverbiale. Le critère 3 permet notamment de ne pas considérer comme ET des expressions répondant aux critères 1 et 2 mais dont la référence est événementielle (usage métonymique d'une ET en tant que nom d'événement). Ainsi, pour les exemples suivants :

(4) *Après le 11 Septembre 2001, le Congrès américain a adopté plusieurs lois*.
(5) *Après les attentats du 11 Septembre 2001, le Congrès américain a adopté plusieurs lois*.

[4] On utilise syntagme noyau comme traduction de *chunk*.

l'expression " Après le 11 Septembre 2001 " du premier exemple ne sera pas annotée comme ET puisque, même si elle respecte les critères 1 et 2, elle est en contexte paraphrasable par *Après l'événement qui s'est produit le 11 Septembre 2001*. Dans le deuxième cas, " du 11 Septembre 2001 " est annotée comme une ET : elle respecte les critères 1, 2 et 3 (ici l'expression ne dénote pas l'événement mais bien la date à laquelle se sont produits les attentats).

Nous souhaitons apporter quelques précisions concernant des cas potentiellement " incertains " :
- **expressions génériques** : non respect du critère 1 pour " hiver " dans *L'hiver est rude dans cette région* ; il ne s'agit pas d'une ET puisque la question permettant d'extraire l'expression à partir de l'exemple est *Qu'est-ce qui est rude dans cette région ?*
- **expressions figées** : en dépit du respect des critères 1 et 2 et 3, nous choisissons de ne pas considérer comme ET des expressions du type *l'heure du crime* ou *le jour du seigneur* et ce en raison de leur sens non compositionnel. Des expressions telles que *l'heure du déjeuner* ou *l'heure de la pause*, dotées d'un sens compositionnel, sont en revanche des ET.
- **expressions événementielles**[5] : ces expressions peuvent ou non obéir aux critères 1, 2 et 3. Par exemple *Fête des Mères* ne respecte pas le critère 2 alors que *Journée Internationale de la Femme* le respecte. Dans tous les cas cependant, ces expressions (qui correspondent à des événements) ont la forme syntaxique NP1 prep NP2 et on peut leur appliquer le test suivant : " expression *célèbre* NP2 ". (ex. " La Journée Internationale de la Femme célèbre la femme ", " le Bicentennaire de la Révolution célèbre la Révolution ", etc.) On choisit, même si ces expressions ont une forte connotation temporelle, de ne pas les inclure dans les ET[6].

3.2 Critères pour la délimitation

Une fois identifiées, il importe de délimiter les ET. Pour ce qui est de la borne gauche des ET, nous considérons *tous* les dépendants syntaxiques à gauche de la tête de l'ET (telle que définie dans le critère 2 section 3.1) comme faisant partie intégrante de l'ET. Ce choix est différent de ce qui est pratiqué dans TimeML. Il nous semble en effet important de conserver des informations capitales concernant aussi bien le typage des ET (*il y a trois ans* vs. *pendant trois ans*) que le calcul de la référence temporelle (*ce lundi* vs. *le lundi*). Cela permet par ailleurs de dissocier le problème de l'analyse syntaxique (avec notamment la difficile analyse des rattachements prépositionnels) de celui de l'analyse des ET. Ainsi, dans la phrases suivantes, l'ET est constituée de la suite en gras :

(6) ***Depuis bientôt plus d'un an****, les débats autour de la classification d'âge pour les livres jeunesse en Angleterre ont fait rage.*

Cherchant à conserver le plus d'information possible, cette large délimitation à gauche des expressions temporelles est compatible avec les travaux de (Battistelli *et al.*, 2006). Concernant la borne droite des ET, nous considérons tous les dépendants à droite s'il s'agit d'adjectifs, d'adverbes, de prépositions ou de noms faisant partie des listes définies par le critère 2. Nous considérons ainsi les expressions suivantes :

(7) *Il part* ***au printemps de l'année prochaine****.*
(8) ***Au troisième jour*** *de sa visite d'état en Chine, que faut-il conclure ?*[7]
(9) *900 passagers bloqués* ***6 heures durant****.*

[5] Une amorce de réflexion sur les " dates événementielles " est presente dans (Battistelli *et al.*, 2006). Cette question a par ailleurs été évoquée lors de la soumission du projet ANR FilTempo lors de discussions entre les partenaires (plus particulièrement Delphine Battistelli de MoDyCo et Xavier Tannier du Limsi).

[6] On considerera donc *La journée de la visite du maire* comme ET car le test mentionné ne peut être appliqué.

[7] Notons que ce critère de borne droite impose de considérer seulement l'unité " l'heure " dans *l'heure du déjeuner*

3.3 Critères pour la segmentation des ET complexes

Les critères que nous avons adoptés pour la segmentation des ET sont introduits dans (Hagège & Tannier, 2008). Définis pour l'anglais, ils sont, ainsi que nous avons pu le vérifier, parfaitement applicables aux ET du français. Nous les présentons ci-dessous et les illustrons par des exemples.

Une ET complexe doit être segmentée en ET minimales si et seulement si les deux critères suivants sont vérifiés simultanément :

Critère 1 (syntaxique) : toutes les combinaisons procès-ET minimale sont valides syntaxiquement.

Critère 2 (sémantique) : la combinaison procès-ET complexe implique logiquement toutes les combinaisons procès-ET minimale.

Dans les exemples suivants, nous montrons le résultat de l'application des critères de segmentation sur les ET complexes marquées en gras.

(10) *Il est parti* ***pendant deux jours avant Noël***.

Les combinaisons *Il est parti pendant deux jours* et *Il est parti avant Noël* sont toutes deux syntaxiquement valides (critère 1 vérifié). Par ailleurs, si nous admettons que la valeur de vérité de (10) est VRAI, alors les valeurs de vérité des deux combinaisons procès-ET minimale qui sont *Il est parti pendant deux jours* et *Il est parti avant Noël* sont également VRAI (critère 2 est vérifié). Dans ce contexte, nous devons donc segmenter la suite *pendant deux jours avant Noël* en deux ET, *pendant deux jours* d'une part et *avant Noël* d'autre part. Considérons un autre exemple :

(11) *Il est tombé* ***deux jours avant Noël***.

La combinaison **Il est tombé deux jours* n'est pas syntaxiquement valide même si *Il est tombé avant Noël* l'est. Le critère 1 n'étant pas vérifié, on ne peut, dans ce contexte, segmenter l'expression. Pour illustrer l'importance de la considération du contexte, examinons à présent l'exemple suivant :

(12) *Il est parti* ***deux jours avant Noël***.

Les suites *Il est parti deux jours* et *Il est parti avant Noël* sont valides syntaxiquement : ici le critère 1 s'applique. Le critère 2 n'est, en revanche, pas vérifié. En effet, admettons que *Il est parti deux jours avant Noël* soit VRAI et considérons ici que la phrase est paraphrasable par *Il est parti le 23 décembre* (hors de tout contexte, interprétation privilégiée même si une ambiguité existe). *Il est parti avant Noël* est alors également VRAI, mais nous ne pouvons pas inférer que *Il est parti deux jours* soit VRAI. Nous devons donc considérer la suite comme une et une seule ET[8].

4 Caractérisation des ET et proposition de typage

Nous proposons un ensemble de types pour caractériser les ET. Chaque type est déterminé selon des critères précis que nous pensons pouvoir être des critères intersubjectifs. Avant de présenter ces types, nous souhaitons souligner deux points. Le premier est que toute ET est représentée

ou *l'heure de la pause*, laquelle est ensuite analysable en fonction de l'événement " déjeuner " ou " pause " auquel elle se rapporte.

[8] Notons que les expressions complexes segmentables d'après nos critères correspondent aux cas où il est possible linguistiquement d'avoir le déplacement dans la phrase d'une des sous-expressions.

par un intervalle ou ensemble d'intervalles (au sens mathématique) qui peuvent être bornés ou non bornés et extrêmement petits. Nous préférons ne pas travailler à la fois avec des points et des intervalles, et ce afin de ne pas devoir présupposer d'une granularité minimum de l'étendue temporelle. Ainsi, il n'y a plus lieu de distinguer un type DATE d'un type HEURE ainsi que le fait TimeML. Le second point est relatif à l'objectif du typage des expressions temporelles, lequel n'est pas considéré comme une fin en soi. Notre but est de pouvoir ordonner temporellement selon un ordre partiel des procès apparaissant dans les textes. Le typage des expressions temporelles est un des éléments qui permettra de spécifier et d'ancrer sur une ligne temporelle les procès auxquels elles sont rattachées. La notion d'ancrage sur le calendrier et de grandeur de l'intervalle correspondant aux ET est donc cruciale pour nous.

4.1 Typage de premier niveau

Le critère permettant de distinguer à quel type une ET doit être associée est celui de l'**ancrage** possible de cette expression dans un calendrier et du nombre d'occurrences de cet ancrage. Nous proposons, en plus de cette notion d'ancrage, un test linguistique aidant à la catégorisation : l'utilisation d'une interrogative formée à partir du procès auquel l'ET est rattachée et qui permet d'extraire l'ET. Ce test des interrogatives doit être ordonné strictement, selon l'ordre indiqué dans le tableau 1 ci-dessous, présentant les différents types que nous considérons.

Type	Représentation	Ancrage sur ligne temporelle	Interrogative
DATE	Intervalle	1 ancrage unique	*(prep) Quand ?*
DURÉE	Intervalle	pas d'ancrage	*(prep) Combien de temps ?*
FRÉQUENCE	Ensemble d'intervalles	ancrage multiple	*A quelle fréquence ?* ou *Quand ?*

TAB. 1 – Expressions temporelles : typage de premier niveau.

DATE Le critère principal pour déterminer si une ET est de type DATE est celui de la possibilité d'ancrage de l'expression sur une ligne temporelle divisée selon un certain calendrier. Si cet ancrage est possible et s'il est unique, alors nous avons affaire à une expression de type DATE. Sous le type de premier niveau DATE sont regroupées des expressions comme *le 23 janvier 2003* mais aussi *depuis 2 mois* dans un contexte comme *Lia est partie* ***depuis deux mois***. En effet, il est possible, d'une part, de poser la question " Depuis quand Lia est-elle partie ? " (correspondant à la réponse " depuis deux mois ") et, d'autre part, de placer sur un calendrier l'intervalle correspondant à l'ET, moyennant la connaissance du moment de l'énonciation, cet intervalle n'étant pas borné à droite. Dans cette catégorie, nous avons également des expressions telle que *entre le 2 et le 3 janvier* (*Elle arrivera* ***entre le 2 et le 3 janvier***) qui, selon nos critères de segmentation doit être considérée comme une seule expression, constitue un intervalle ancrable de manière unique sur une ligne temporelle et peut répondre au test de l'interrogative en *(prep) Quand ?*. Nous nous distinguons donc ici clairement de l'annotation de TIDES et de TimeML sur cet exemple. Nous considérons également comme DATE une expression comme *toute la journée* dans le contexte du titre du journal *Libération* du 10 février 2009 suivant : *Grève SNCF : Trafic perturbé* ***toute la journée***. Il y a en effet ancrage de l'ET puisque *toute la journée* correspond ici à la date de parution du journal d'où le titre est extrait, soit le 10 février 2009. Enfin, dans l'exemple *il est arrivé* ***à midi***, l'expression en gras est également de type DATE car nous pensons qu'il n'y a aucune bonne

raison de distinguer les ET dont la granularité serait supérieure ou égale au jour des ET dont la granularité serait l'heure du point de vue de la caractérisation.

DURÉE Les ET de type DURÉE sont également représentées par un intervalle borné au moins d'un coté. Cependant, à la différence des DATES, l'intervalle représentant l'ET **n'est pas ancrable** sur une ligne de temps dans le contexte procès-ET. Un exemple typique de durée est " pendant deux mois " dans *Elle a travaillé à la Poste pendant deux mois*. En effet, l'ET permet bien de délimiter une " étendue de temps " de 2 mois, mais rien, ni dans l'expression elle-même, ni dans le contexte de son association avec le procès, ne nous donne un indice pour ancrer cet intervalle sur une ligne temporelle. Le test de l'interrogative dans ce cas prototypique est bien celui indiqué dans le tableau, l'exemple permet en effet de répondre à la question " Combien de temps a-t-elle travaillé à la Poste " mais en aucun cas à la question " Quand a-t-elle travaillé ? ". Autre exemple d'ET de type DURÉE :

(13) *Elle a été mariée* ***trois mois***.

FRÉQUENCE Il s'agit ici d'ET qui définissent non plus un intervalle mais un ensemble d'intervalles temporels, avec dans ce cas un ancrage multiple. Ainsi, " le dimanche " dans *Le Dimanche il a pour habitude de se promener dans les bois*, est un cas de fréquence ; l'ET ne définit pas un dimanche particulier, mais bien un ensemble de dimanches. Quelques exemples d'ET de type FRÉQUENCE :

(14) *Il va à la campagne* ***tous les week-ends***.
(15) *Elle assure une permanence* ***un dimanche par mois***.

4.2 Typage de deuxième niveau

Au-delà du typage de premier niveau présenté ci-avant, il est possible de raffiner les types DATE et DURÉE selon le degré de précision de l'ancrage sur la ligne temporelle et suivant le type de référence nécessaire pour le calcul de cet ancrage. Le tableau 2 synthétise les sous-types de DATE et de DURÉE que nous avons considérés[9].

Les dates absolues sont des dates pour lesquelles un positionnement sur le calendrier est possible en considérant la seule ET en dehors de tout contexte. On estime que l'on a positionnement sur le calendrier à partir du moment où la mention de l'année, au moins, est présente dans la date. Les dates relatives à l'énonciation sont celles pour lesquelles il est nécessaire de connaître le moment de l'énonciation pour pouvoir les positionner sur le calendrier. Les dates relatives à une référence textuelle sont celles qui sont positionnables connaissant un moment de référence dans le texte. Dans ce cas des dates relatives à une référence textuelle, il est possible de distinguer deux sous-types. Le premier correspond aux dates anaphoriques, pour lesquelles un positionnement temporel ne sera possible que par le biais d'une situation décrite par le texte. Le second cas correspond aux dates incomplètes. Ces dates sont des dates comprenant un élément calendaire d'indication de jour et/ou de mois mais pas d'année. Leur repérage est également effectué par un contexte présent dans le texte. Il s'agit en général d'un renvoi à une autre ET mentionnée auparavant[10]. Enfin, les dates indéterminées[11] qui sont, en tant que dates, ancrables sur une ligne temporelle, mais avec

[9] Nous retrouvons dans nos sous-types des notions présentes dans TimeML comme le type de référence des expressions référentielles. Nous complétons TimeML en introduisant deux types de référence textuelle et les dates intdéterminées.

[10] Cette distinction est présente dans (Gosselin, 1996).

[11] Nous utilisons également ici la terminologie de Gosselin.

Sous-type de DATE	**Type d'ancrage**	
Date absolue	Directement calculable à partir de l'expression. ex : *Il est parti **en juin 2004***.	
Date relative à l'énonciation	Calculable connaissant la date du moment de l'énonciation ex : *Il viendra **demain***.	
Date relative à référence textuelle	Calculable à partir d'un temps de référence dans le texte.	
	dates anaphoriques *le lendemain* (anaphore), *le 3ème jour de grève* (cataphore)	dates incomplètes *le 3 juin*
Dates indéterminées	il y a ancrage mais il reste indéterminé. ex : ***Un jour**, elle partira*.	
Sous-type de DURÉE	**Valeur de l'intervalle**	
Durée déterminée	Valeur connue. ex : *Il a travaillé **pendant deux heures***.	
Durée indéterminée	Valeur indéterminée. ex : *Il a été malade **pendant un certain temps***.	

TAB. 2 – Expressions temporelles : typage de deuxième niveau.

un fort degré d'indétermination qu'aucun contexte ne peut compléter.

Nous n'avons considéré que deux sous-types de DURÉE. Les durées déterminées, pour lesquelles on peut déterminer la longueur de l'intervalle en fonction d'une unité du calendrier, et les durées indéterminées, pour lesquelles la longueur de l'intervalle reste vague et indéterminée.

5 Illustration et conclusion

Nous illustrons la nécessité de considérer l'ET dans son contexte pour pouvoir la caractériser convenablement par les exemples suivants.

(16) *Mon fils de 28 mois a de gros coups de fatigue **le matin**.*
(17) ***Un matin**, il prit son balluchon et quitta la maison.*
(18) *L'hospitalisation a duré **un matin**.*
(19) ***Le matin** il avait pris son petit déjeuner dans la hâte.*
(20) ***Le matin** de son départ, elle pleura toutes les larmes de son corps.*
(21) ***Le matin** est un moment propice à la réflexion*

Le premier exemple est un cas de FRÉQUENCE, il y a en effet ancrage multiple (répétition des coups de fatigue sur la ligne du temps). En (17), il s'agit d'une DATE indéterminée : il y a bien un ancrage unique sur la ligne du temps pour ce procès via l'ET, cet ancrage reste vague. En (18) nous avons une DURÉE : la question ici est *combien de temps ?* et non *quand ?* et il n'y a pas d'ancrage du procès auquel l'ET est associée. L'exemple suivant est un cas de DATE à reférence textuelle anaphorique (la référence n'est pas explicite dans la phrase mais est donnée par un contexte antérieur). (20) illustre lui aussi un cas de DATE à référence textuelle anaphorique mais, ici, la référence est explicite (son départ) et il s'agit d'une cataphore. Enfin, le dernier exemple montre un cas d'emploi générique de *Le matin* ; les questions *quand ?* et *combien de temps ?* ne peuvent être posées.

Les différentes interprétations de ces ET qui en surface sont identiques ou très proches sont motivées par plusieurs facteurs linguistiques comme la détermination, l'aspect verbal du procès auquel est associée l'ET ou encore la sémantique lexicale du procès. La combinaison de tous ces éléments

devra être prise en considération pour un traitement complet de la temporalité.

Nous avons par ailleurs tenté de mettre en application les directives proposées dans cet article en les appliquant à l'annotion des ET de deux textes. Le premier est une transcription manuelle de l'oral provenant de la campagne Ester II[12], le deuxième est un article de Wikipédia relatant une expédition polaire. Pour ces deux textes, les deux auteurs ont annoté de manière indépendante les expressions temporelles selon la typologie présentée. Pour les deux textes, 175 expressions sont en surface de potentielles expressions temporelles (comportent des éléments lexicaux mentionnés dans le critère 2 de reconnaissance des ET). Sur ces 175 expressions, 147 sont annotées comme ET par les deux annotateurs, 11 sont annotées comme n'étant pas des ET par les deux annotateurs (soit donc un accord dans plus de 90% des cas). 16 expressions sont annotées comme ET par les deux parties mais avec des différences de type et enfin, une seule expression est annotée comme ET par une partie et non ET par l'autre. Pour l'ensemble des cas litigieux (17 cas au total), les deux annotateurs sont parvenues à un accord après discussion.

Pour conclure, nous pensons donc que cette proposition de caractérisation et de typage permet d'effacer certaines zones d'ombres rencontrées sur le terrain lors de l'annotation des expressions temporelles. Elle peut être adoptée dans le cadre de campagnes d'annotation d'entités nommées et constitue un premier pas vers la tâche plus ambitieuse de l'analyse automatique de la temporalité dans les textes, à laquelle nous nous sommes attelées pour le français.

Références

BATTISTELLI D., MINEL J.-L. & SCHWER S. (2006). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *Traitement Automatique des Langues*, p. 11–37.

BITTAR A. (2008). Annotation des informations temporelles dans des textes en français. In *Actes de RECITAL 2008*, Avignon.

FERRO L., GERBER L., MANI I., SUNDHEIM B. & WILSON G. (2005). *TIDES 2005 Standard for the Annotation of Temporal Expressions*. Rapport interne, MITRE.

GOSSELIN L. (1996). *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*. Louvain-la-Neuve : Duculot.

GREVISSE M. & GOOSSE A. (1986). *Le Bon Usage*. Louvain-la-Neuve : Duculot, 1993 edition.

HAGÈGE C., BAPTISTA J. & MAMEDE N. (2008). Proposta de anotação e normalização de expressões temporais da categoría TEMPO para o HAREM II. In *Actes de Encontros do Segundo HAREM*, Aveiro, Portugal.

HAGÈGE C. & TANNIER X. (2008). XTM : A robust temporal text processor. In *Actes de CICLing 2008*, Haïfa, Israël.

PARENT G., GAGNON M. & MULLER P. (2008). Annotation d'expressions temporelles et d'événements en français. In *Actes de TALN*, Avignon.

SAURÍ R., LITTMAN J., KNIPPEN B., GAIZAUSKAS R., SETZER A. & PUSTEJOVSKY J. (2006). *TimeML Annotation Guidelines Version 1.2.1*.

[12] http://www.afcp-parole.org/ester/